DIVISION OF THE HUMANITIES AND SOCIAL SCIENCES
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA 91125

AN IMPROVED STATISTICAL MODEL FOR MULTIPARTY ELECTORAL DATA

James Honaker
Harvard University

Jonathan N. Katz
California Institute of Technology

Gary King
Harvard University

SOCIAL SCIENCE WORKING PAPER 1111

February, 2001

An Improved Statistical Model for Multiparty Electoral Data*

James Honaker† Jonathan N. Katz‡ Gary King§

Abstract

Katz and King (1999) develop 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 politics 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 where each party pulls support from. Although the mathematics of their statistical model covers any number of political parties, it is computationally very demanding, and hence slow and numerically imprecise, with more than three. The original goal of our work was to produce an approximate method that works quicker in practice with many parties without making too many theoretical compromises. As it turns out, the method we offer here improves on Katz and King's (in bias, variance, numerical stability, and computational speed) even when the latter is computationally feasible. We also offer easy-to-use software that implements our suggestions.

1. INTRODUCTION

We offer a computationally feasible algorithm, and easy-to-use software, that approximates Katz and King's (1999; hereinafter KK) full information maximum likelihood

*An earlier version of the paper was presented at the annual meetings of the American Political Science Association, Washington, D.C., 2000 under the title "A *Practical* Statistical Model for Multiparty Electoral Data". For research support, we gratefully acknowledge the John M. Olin Foundation, the National Science Foundation (SBR-9729884, SBR-9753126, and IIS-9874747), the National Institutes of Aging, and the World Health Organization.

†Department of Government, Littauer Center North Yard, Harvard University, Cambridge Massachusetts 02138; tercer@latte.harvard.edu, http://data.fas.harvard.edu/gov-dept/graduate/tercer).

‡D.H.S.S. (228-77), California Institute of Technology, Pasadena, CA 91125; JKatz@caltech.Edu, http://jkatz.caltech.edu, (626)395-4030).

§Department of Government, Harvard University and the Global Programme on Evidence for Health Policy, World Health Organization; Center for Basic Research in the Social Sciences, 34 Kirkland Street, Harvard University, Cambridge MA 02138; http://GKing.Harvard.Edu, King@Harvard.Edu, (617) 495-2027.

(FIML) model of district-level multiparty electoral data. Our model provides a tool analogous to what least squares regression provides American politics researchers in that two-party system. That is, scholars can use a tool like this with aggregate multiparty electoral data, and explanatory variables also coded at the district-level, to explain, predict, or infer about counterfactuals in real election results.

The main intended 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, more numerical precise and also has lower mean square error. We believe the result is a useful tool for comparative politics research.

We first introduce our notation and the basic KK model (Section 2) and an overview of our proposed approach (Section 3, with technical details in Appendix A). Then, we present Monte Carlo comparisons of KK with our method in three party data (Section 4) and replicate KK's empirical results with our methods (Section 5). We then conduct Monte Carlo experiments in data with more parties than would be computationally feasible under the KK model (Section 6). Finally, we replicate a real empirical article and show how the substantive results change when our improved method is applied (Section 7).

2. NOTATION AND THE FULL INFORMATION APPROACH

Let V_{ij} denote the fraction of the vote in district i ($i = 1, \ldots, n$) for political party j ($j = 1, \ldots, J$). 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, as $V_{i1} = 1 - V_{i2}$. In general, multiparty electoral data are an example of *compositional data*, where the set of variables fall on the simplex, which means that each vote proportion falls between 0 and 1,

$$V_{ij} \in [0, 1] \qquad \text{for all } i \text{ and } j \tag{1}$$

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

$$\sum_{j=1}^{J} V_{ij} = 1 \qquad \text{for all } i. \tag{2}$$

KK provide some graphical tricks with which to understand these constraints and to portray this type of data.

Following Aitchison (1986), KK avoid the complications induced by the constraints by modeling the $J-1$ log ratios of the vote variables $Y_{ij} = \ln(V_{ij}/V_{iJ})$, for $j = 1, \ldots, J-1$. The advantage of this approach is that the set of Y_{ij} variables are 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 showed 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 added a component of the model to cope with partially contested or uncontested district elections. As they do, we set 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). Then, in order to use the information from not fully contested district elections, we assumed 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, is plausible for a wide range of circumstances, and the machinery used to modify the model for it could easily be modified to suit qualitative knowledge in different electoral circumstances. However, some assumption along these lines is always necessary.

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 so ignoring it and listwise deleting these districts say, is not a reasonable option. (Thus, as we explain below, these compositional data are not "missing completely at random" or even "missing at random").

There are two computational issues. First, the KK model computes the likelihood for these observations by treating them as censored and using the assumption above. For example, suppose 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, $L_i^{23} = \Pr(Y_{i1} < 0, Y_{i2} > Y_{i1} \mid \psi)$, wherever the set of parties contesting is $P_i = \{2, 3\}$. The problem is that this likelihood requires an analytically intractable integration.[1] KK computed this integration numerically. This computation was time consuming, but not overwhelmingly so 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 turns out to be a lot of the reason why our method out-performs KK.

A second computational problem with the KK approach is exponentially increasing numbers of combinations of parties that could contest. 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

[1] That is, $L_i^{23} = \int_{-\infty}^{0} \int_{Y_{i1}}^{\infty} \mathrm{T}(Y_{i1}, Y_{i2}|\psi)\, dY_{i2}\, dY_{i1} = \int_{-\infty}^{0} \mathrm{T}(Y_{i1}|\mu_{i1}, \sigma_1, \nu)[1 - \mathrm{F_T}(Y_{i1}|\mu_{2|1}, \sigma_{2|1}, \nu + 1)]\, dY_{i1}$, where $\mathrm{F_T}$ is the cumulative distribution function of the (univariate) t, and $\mu_{2|1}$ and $\sigma_{2|1}$ are the conditional moments defined in KK.

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 1,024 pieces; with $J = 20$, the likelihood has over a million parts.

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. Although they seemed to scale better, the base speed was so slow that the method would likely deter political scientists from using the method. Another approach might be simulating higher dimensional integrals, but this does not seem promising after 5-6 parties or so, nor does it address the combinatorial problem.

3. OVERVIEW OF OUR ALTERNATIVE APPROACH

We now provide an intuitive overview, saving the technical details for Appendix A. 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 enables us to build on the literature on multiple imputation rather than the harder-to-analyze application-specific approaches, as in KK (see King, Honaker, Joseph, and Scheve, 2000).

In districts without all parties contesting, Y_{ij} is missing for all j, since the effective vote would differ for all parties, not only the one not contesting. If these cells in our data matrix were nonignorably missing, 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, Joseph, King, and Scheve, 2000). Then, we would follow the usual strategy of multiple imputation: create M datasets 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. The analyst would then use whatever method they would apply if all the data were observed to each of the M data sets, and then use the usual procedure for combining the results of the separate imputations.

Unfortunately, two problems make this strategy incorrect without some modification. First, we already know 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. Appendix A 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 draw from the posterior distribution with the EMis algorithm from 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. What this procedure does is cause 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 not be as good an 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 1.4. The sampling correction also has a benefit: since it is a two-step approach, and thus any modeling assumptions only affect 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 seems well worth dealing with given the enormous computational savings that result. And, as we show below, our method produces lower mean square error than the KK FIML approach so the compromises, if any, are outweighed by our overall approach.

The results of the imputation algorithm with the simulation-based correction for partially contested districts make available a set of M imputed datasets. 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. Thus, we treat the weights as known conditional on imputation j, and as a result are able to make the analysis non-iterative and indeed no slower than an ordinary weighted least squares analysis. Section 1.5 explains this procedure.

The procedure for scholars wishing to use our procedure is then fairly straightforward. Researchers will use a specially modified version of Amelia, input the observed voting data for all the parties, along with any other covariates to be used in the analysis model or others that might help predict (the 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, with missing values of the covariates (if any) 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 5 or so) analyses on the M sets of data need to be performed separately and results averaged, as in multiple imputation, although software exists that makes this step transparent.

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 $Y_{13}^{\bar{2}}$ 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 $Y_{13}^{\bar{2}}$, as

$$\mathbf{D} = \{Y_{12}^*, Y_{23}^*, Y_{13}^{\bar{2}}, X\} \sim^{iid} \text{t } (\mu, \Psi, \nu). \tag{3}$$

where X are covariates.[2] Then we impose uncontestedness as structural missingness according to one of two rules that we now describe.

The assumption of multiple imputation is that the missingness mechanism is "Missing at Random"(MAR), which means that elements of the dataset 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 "Non-Ignorable"(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

[2]For each simulation, we set the parameters near to that for UK elections case: $\mu = \{0.5, 0.4, -0.7, 0.6, 0.6, 1\}$, $\nu = 5$, and

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

votes, if it had run, than the parties that did field candidates. Of course, if sufficiently informative covariates are included, this NI process might be made MAR. To study these issues, we run Monte Carlos under these two separate data generation processes. In both we draw 100 districts.

MAR: 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 are then made partially contested with the third party not running. Different simulations are run, with value of N varying from 0 to 50. This can be thought of as the third party having the same information as us and making the best guesses about where they will fail and withdrawing from those based on their guesses, where N is a function of their resources.

NI: The actual effective votes are drawn, and the N districts with the lowest effective votes for the third party are then made partially contested with the third party not running. The value of N is varied again 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.[3] 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 the graph 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 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 (dots and dashes) and without including the observed vote $Y_{13}^{\bar{2}}$ in the imputation stage (dotted). In the MAR graph, neither of these omissions hurt our method much, and so the main improvement would appear to be the algorithm that does the imputation. This is a important improvement over the KK model that we believe is due primarily to numerical stability. That is, if our computers had infinite precision, FIML should do as well as our method. Our algorithm has several features designed to improve numerical stability, chief among these is estimation of ν which is known to be difficult and not globally concave. Our method is not likely to fix this entirely (nor is any other), but it appears to be a substantial improvement.

In the NI graphs, the rejection sampler appears to be having a large effect, since removing it (i.e., as shown by the dashed and dotted line) increases the MSE. It is still

[3]We 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 Figure 2, which we discuss below.

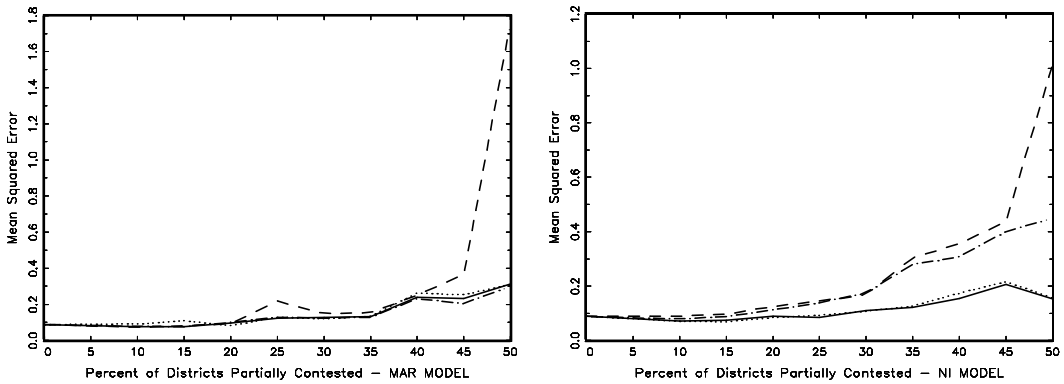

Figure 1: *Mean Square Error Comparisons. MSE is plotted for four methods of estimation under MAR (for the left graph) and NI (for the right graph) data generation processes. The dashed line is the FIML model of KK. The solid line the imputation model of this paper, while the dotted line is the imputation model without* $Y_{13}^{\bar{2}}$*, and the line with dots and dashes is our method without the rejection sampler.*

below KK, but not by much until very high levels of contestedness. However, removing $Y_{13}^{\bar{2}}$ (the dotted line) has very little effect and so we infer that this feature — which was the only bit of information used in our model but not in the KK model — does not account for the vast majority of the improvements over KK.

To explore our hypothesis of improved numerical stability, we also look at the distributions of the individual coefficients. For the constant term (in the left graphs) and the first coefficient (for the right graphs), and for MAR (at the top) and NI (at the bottom), Figure 2 gives box plots for the distribution of our Monte Carlo simulations for each level of uncontestedness, with the KK model boxplots in **bold**. In each graph, the truth is labeled with a horizontal dotted line.

For 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 interesting are the outliers in the KK model but not in our method. These can be seen most clearly in all but the bottom left graph. Indeed, only in the NI constant term (the 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 verified that this is only small sample result, that is 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. 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 in order to have faster convergence and the ability to automate the program so that it would be easy to use.

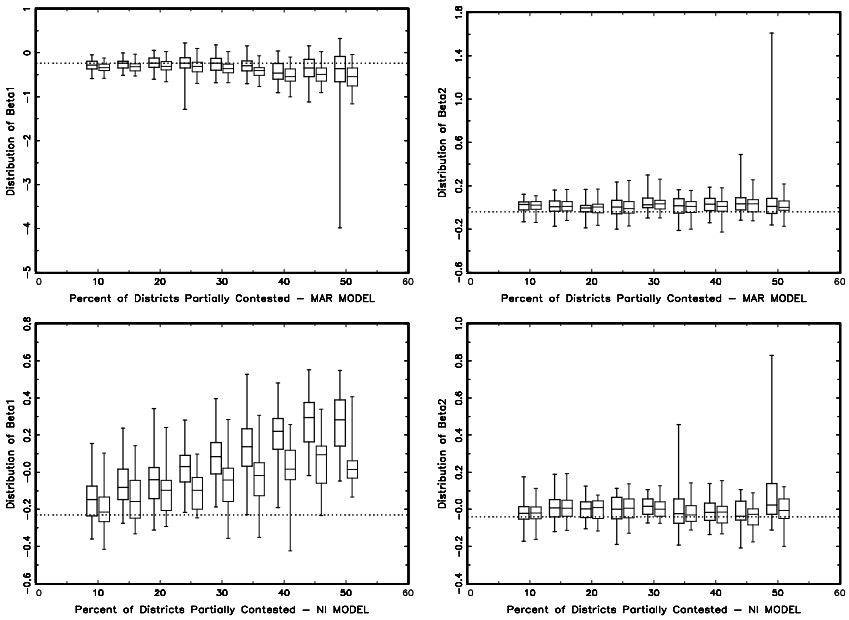

Figure 2: *Distribution of Coefficients Across Monte Carlo Simulations. For the constant term (labeled beta1, in the left graphs) and the coefficient on the first explanatory variable (labeled beta2, in the right graphs), each box plot gives the distribution of the estimates for a given level of uncontestedness. The truth is labeled with a horizontal dotted line. Boxplots for the KK model are in* ***bold*** *and ours are in normal.*

5. 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 Figure 3, 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 and those in the background (above) the effects we calculate with our method. The vertical distance of each arrow above the line indicates the advantage of running an incumbent, as compared to a 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. More remarkable 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.

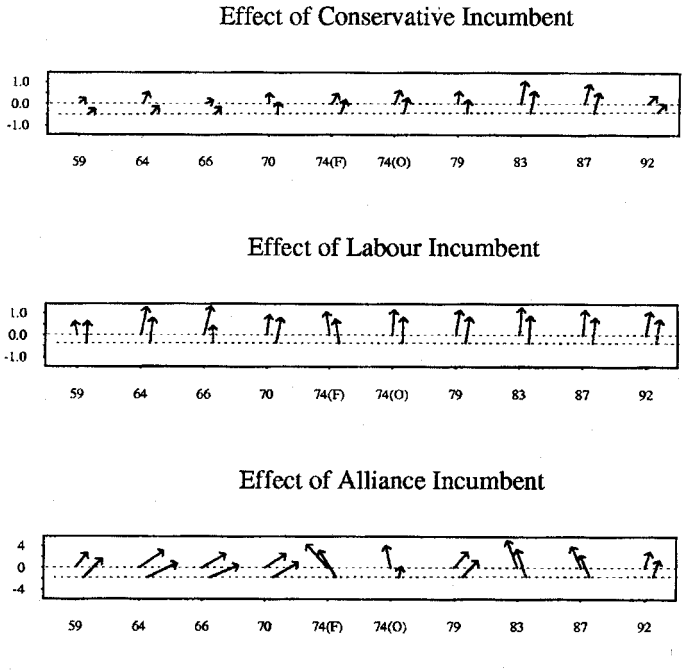

Figure 3: *The foreground arrows represent the effects of incumbency in the original model, while the background arrows are the effects calculated with our imputation model. The vertical distance of each arrow represents the advantage to that party of running an incumbent, while the direction shows from which party the support is being drawn.*

6. MONTE CARLO EVALUATION WITH MANY PARTIES

We also provide an example of estimates from a simulated party system with greater than three parties. In this simulation, there are five parties and four covariates $X_1, \ldots, X_4$, 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 in the analysts model. The covariates are assumed to be completely observed, although if there were scattered MAR missingness that would pose no problem to the imputation model. The unique rows of the matrix of missingness patterns, R_2 is as follows.

$$R_2 = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

The first four columns of 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

5, respectively, does not contest. Columns 11 and 12 are log-vote ratios for when neither party 4 nor 5 contest, and the final four columns correspond to the fully observed covariates. The first row of 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 two, 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 parties 4 and 5 do not contest.

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 the partially contested districts, that is running the model only on the districts for which the actual vote shares of all parties were known. Then we imputed the effective votes for the partially contested districts with our algorithm and again computed this coefficient.

The small vertical bar in each panel of Figure 4 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 zero). 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 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.

7. 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 run our model on twelve 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 the differences between the results from the two methods by a broad overview and then a more direct reanalysis of the central question in their paper. Thus, Figure 5 plots estimates from our method (horizontally) by OLS (vertically) for all coefficients (translated to first differences from our method so as to be directly comparable) from all equations. The dashed 45° line marks the place where we would find equality

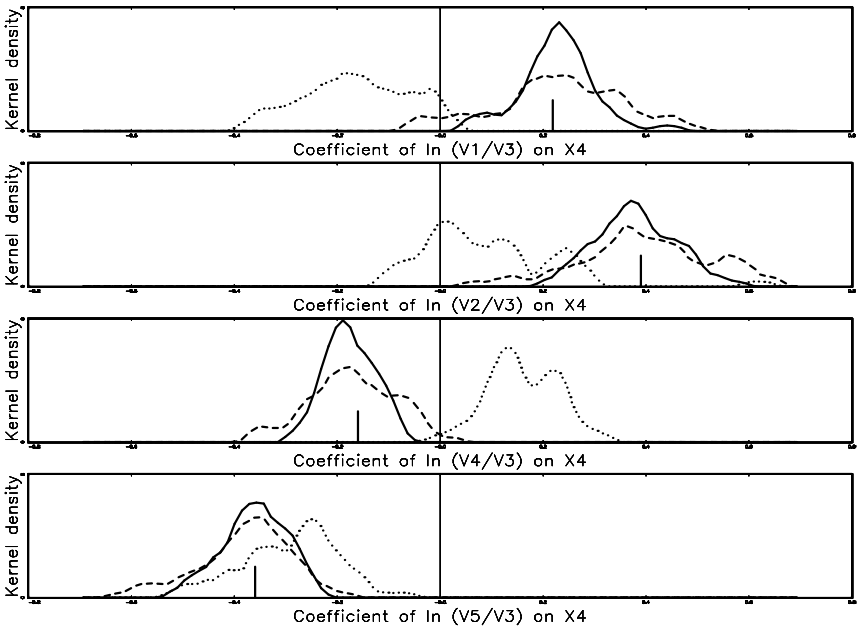

Figure 4: *Distribution over Monte Carlo Simulations. Kernel Density Plots of a coefficient in fifty simulations of a five-party race applied to the effective votes (solid line), listwise deletion on the partially contested districts (dotted line), and our method (dashed). The true coefficient is represented with a small vertical bar. Note that our method is approximately unbiased with only slightly higher variance than the method applied to the effective vote.*

between the two methods. When one method finds a first difference of zero, 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 zero 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.

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 post-Communist country showing real economic growth brought about by their recommended dramatic "shock therapy" transition and privatization. The Polish example was held up to show other post-Communist states as a model and incentive for economic reform. The electorate of Poland seemed to show they were not as happy with these changes as the economists, given that they put the Communists back in power. One response of the economists was that "shock therapy" should have been implemented earlier

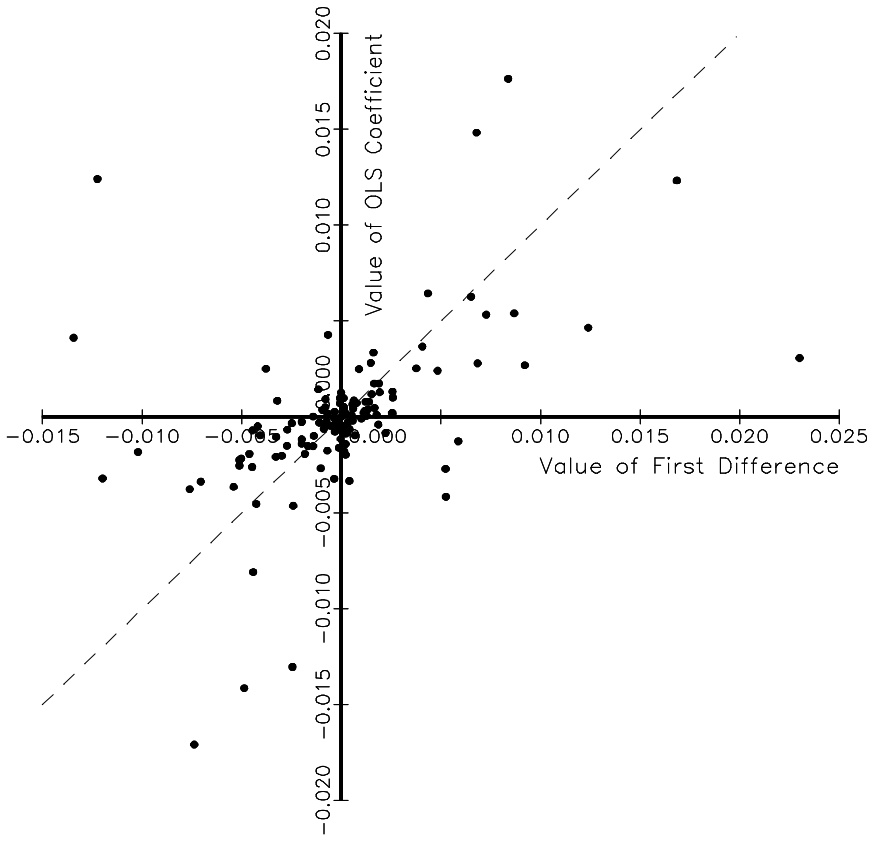

Figure 5: *Plot of First Differences Computed by OLS and Our Method. Each point is one regression coefficient, and the distance from the 45 degree line indicates how far apart the estimates from the two methods are.*

and quicker; as a result, the reasoning goes, growth would have been quick enough and large enough so 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 faired had growth been more profound (a cumulative of 15 percent instead of the 10 percent actually seen since the end of the recession brought on by transition to a market economy) and unemployment had not risen by 2.3 percent? They find that the main reform party, the Democratic Union (UD), would have gained a couple points, but not nearly enough to close the ten point gap between themselves and the chief post-Communist 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 growth in each voivodship (a district in Poland) by five percent, and reducing unemployment by 2.3 percent. The differences are shown in the last two columns of Table 7. Our method agrees with their main finding that the reform party would not have gained control had the economy been much stronger, but very interestingly find that this does shuffle votes profoundly among the post-Communist parties, with the PSL almost replacing the SLD

as the largest vote share recipient. According to our results, in contrary to OLS, the question the election answered was which post-Communist party would win, not how the reform party would fair. That the state of the economy was of such importance to the battle between the two largest post-Communist 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.)

Party	*Acronym*	*Block*	1993 Vote Share	Change in Vote Share	OLS Pred. Change
Democratic Left Alliance	SLD	post-communist	20.41	−1.16	−1.41
Polish Peasant Party	PSL	post-communist	15.40	+3.25	−0.58
Democratic Union	UD	pro-reform	10.59	−0.92	+0.86
Nonpty. Block to Sup. Reform	BBWR	pro-reform	5.41	−0.05	+0.78
Union of Labor	UP	anti-reform	7.28	−0.79	−0.20
Confed. for an Indepndt. Pol.	KPN	anti-reform	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 Dem. Congress	KLD	center-right	3.99	−0.65	0.00
Pol. Peas. Pty. - Party Alliance	PL		2.37	−0.30	−0.53
Party X	PX		2.74	−0.30	+0.01
Solidarity	SL		4.90	+0.22	−0.49
Other			10.35	+0.83	+0.67
		Total	100.00	0.00	0.34

Table 1: The Effects of a Better Economy: Comparing Results from OLS to Our Method

8. 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, and scales 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 of course providing imputations of missing values in the original dataset. A side benefit of our approach is to reduce greatly numerical instability relative to the KK model, producing a method that is not only applicable to a much wider range of political systems but also has less bias and inefficiency than the KK approach.

Appendix A. Technical Details of Our Alternative Algorithm

We begin by describing the existing EMis algorithm for multivariate normal data (Section 1.1) and briefly summarize some useful properties of the multivariate t density (Section 1.2). We then summarize changes in the EMis algorithm we made to accommodate t-distributed data (Section 1.3) and incorporate constraints for partially contested or uncontested seats (Section 1.4). We also summarize the use of the t regression analysis model (Section 1.5) and discuss how long the algorithm takes (Section 1.6).

1.1. *The EMis algorithm*

The EMis (Expectation Maximization with importance resampling) multiple imputation algorithm is an alternative to data augmentation (Schafer 1998) often employed for multiple imputation. The EMis algorithm is as follows: (1) Calculate the maximum posterior of the data using the EM algorithm; (2) Estimate the variance of this point estimate in the space of the sufficient statistics; (3) Construct an approximating distribution of the posterior likelihood of the sufficient statistics; (3) Importance resample m sets of sufficient statistics from this approximating distribution using the actual posterior likelihood. (5) Impute the missing values, D_{mis}, using each of the above samples to create m completed datasets.

The EM algorithm (Orchard and Woodbury 1972; Dempster et. al 1977; McLachlan and Krishnan 1996) is an increasing popular approach to finding maximum likelihood estimates of systems that are intractable or highly complicated analytically. EM is an iterative deterministic algorithm which under given regularity conditions increases the likelihood of its parameter estimates monotonically on every iteration. It is an integral part in the EMis multiple imputation algorithm presented in King et al. (1998). The EM algorithm has also seen use in political science by Lewis (1998), Bailey (1998), and Jackman (2000).

Given the maximum posterior estimate of the parameters $\hat{\theta} = \vec{(}\hat{\mu}, \hat{\Sigma})$ one computes its variance $V(\hat{\theta})$ after reparameterizing to unbounded scales using the log for the standard deviations and the Fisher's z for the correlations. For small dimensions the variance can be computed with the negative of the inverse of the Hessian; for moderate dimensions the outer product of the gradient; and for large dimensions the variance of simulations from some appropriate Markov chain run in the vicinity of the maximum can be used. Since calculation of the variance is not effected by dependency in these draws, and thus the typical autocorrelation checking do not need to be decided by user monitoring, this is more easily automated than typical MCMC methods. Although each of these three methods is less accurate than the previous one, our analyses convince us that they serve quite well to create an approximating distributing for θ. From this approximating distribution one then uses an acceptance-rejection algorithm by keeping draws of $\tilde{\theta}$ with

probability proportional to the "importance ratio" — the ratio of the actual posterior to the asymptotic normal (or multivariate t) approximation, both evaluated at $\tilde{\theta}$ — and discarding the rest. Without priors, the importance ratio is $L\left(\tilde{\theta} \mid D_{obs}\right) / N\left(\tilde{\theta} \mid \hat{\theta}, V(\hat{\theta})\right)$.

In importance resampling, one often wants the approximating distribution (also known as the covering distribution) to have thicker tails then the true distribution to increase confidence that the distribution is being properly approximated everywhere. This can be done by multiplying the variance computed above by some common factor (generally 1.2–1.5 is used as a rule of thumb), or covering a normal with a t distribution with low degrees of freedom. A useful diagnostic can be extracted from the fact that the larger the ratio of the number of draws from the approximating distribution to the number of acceptances needed (here, m, the number of imputed datasets) the better the approximation.

1.2. *Some Useful Properties of the t Distribution*

We begin with a brief summary of properties of the multivariate t distribution that we use below. If Y_i is distributed:

$$Y_i \sim^{ind} \mathrm{N}(\mu, \Psi/u_i) \tag{4}$$
$$u \sim^{iid} \chi^2_\nu/\nu \tag{5}$$

where $\nu > 0$, then Y is distributed as

$$Y \sim^{iid} \mathrm{t}\ (\mu, \Psi, \nu). \tag{6}$$

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

$$L(\mu, \Psi, \nu|Y, u) = L_N(\mu, \Psi|Y, u) + L_G(\nu|u), \tag{7}$$

where

$$L_G(\nu|u) = -n\ \ln\ \left(\Gamma\left(\frac{\nu}{2}\right)\right) + \frac{n\nu}{2}\ \ln\ \left(\frac{\nu}{2}\right) + \frac{\nu}{2}\sum_{i=1}^{n}\Big(\ln(u_i) - u_i\Big). \tag{8}$$

1.3. *An EMis Algorithm for t Distributed Data*

We followed the framework of the EMis algorithm 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 non-iterative 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 equation 4. We take the

vector of weights u to be an additional variable (completely unobserved) to be imputed in the dataset, and the degrees of freedom ν 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 1998), except that the sums and sums of squares and cross-products computed for the M-step need to be appropriately weighted by u.

We began with an EM algorithm for t distributed data but found convergence to be extremely slow. Similar to Lange et.al. (1989), we found results were actually faster by running separate EM algorithms each conditional on some value of ν over a grid of ν values. To speed up convergence we turned to the ECME algorithm (Liu 1994, Liu and Rubin 1994) to find the MLE of θ, a description of which follows.

The E-step of 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[4]. The vector of weights u^{t+1} is similarly created from the expectation:

$$E(u_i^{t+1}) = \frac{p_i + \nu^{(t)}}{\delta_{i,obs}^{(t+1)} + \nu^{(t)}} \tag{9}$$

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

$$\delta_{i,obs}^{(t+1)} = (Y_{i,obs} - \mu_{i,obs})' \Psi_{i,obs}^{-1} (Y_{i,obs} - \mu_{i,obs}) \tag{10}$$

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 ν. Then maximize the L-function (the constrained actual log-likelihood) over $\theta_2 = \nu$ given $\theta_1 = (\mu, \Psi)$. To do this a one dimensional search is implemented over equation 8. 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 $\theta_2 = \nu$ given $\theta_1 = (\mu, \Psi)$ instead, but in most problems this approach leads to much more rapid convergence over ν (Liu 1994). Indeed, with the t distribution, since the likelihood is separable by equation 7, if we

[4]Two families of EM algorithms are possible. In one, the completed data is stored (Beale and Little, 1975) in the other the sufficient statistics (sums, sums of squares, and sums of cross products) are stored (Dempster, Laird, and Rubin 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.

maximized the Q-function rather than the L-function we would have again 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 Poisson distribution on $(\nu - 2)$ with mean $(\hat{\nu} - 2)$. 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 the paper.

1.4. *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 directly transform the normal model posterior 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 which did nominate candidates" and thus the effective vote in some district of any party which did not contest in that district must be lower than the effective vote of all other parties which 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 dataset, $j \in 1, \ldots, M$, with sufficient statistics θ_j, each observation is checked as to whether it meets the model constraint[5]. In each round, any observation, y_i, which fails 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 non-increasing 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.

1.5. *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 datasets are imputed, the user can apply whatever model he or she would have used if the dataset had arrived fully observed. In the present application, the researcher can analyze the data as if all parties contested

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

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 only needs to run one (noniterative) weighted least squares analysis, for each imputation, and to average the results as in multiple imputation. Thus, Ameila 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.

1.6. *How long does it take?*

To run the series of ten elections under the original KK FIML approach took 35 minutes. To multiply impute the effective votes (and the missing values in the covariates) for all ten elections, the first stage of our alternative algorithm, took 22 minutes, after which the analysts model must be run on each imputed dataset. Using our augmented weighted least squares approach for all ten takes only a few seconds, so for all practical purposes the time for analysis is essentially the time taken by the imputation model. In practice, researchers are thus asked to invest 22 minutes in imputation time, and can then run as many analysis models as they like, each nearly as quickly as any other regression analysis. In cases with more parties, KK is infeasible but our approach scales up well, approximately as does Amelia.

For a model that cannot be rewritten as Weighted Least Squares, the researcher must balance the additional computational time required to run the original model on each of the M imputed datasets[6] versus the analyst's time in writing and programming a more complicated model. In addition, if there are missing values in the covariates, the imputation approach has the further benefit of increasing efficiency and potentially correcting bias.

Timing is greatly effected by the number of variables in the imputation model, as in the original EMis algorithm (King, Honaker, Joseph, and Scheve, 1999). The number of variables increases with the number of patterns of party contestation across districts. For a given number of (possibly incomplete) covariates, a dataset with a very large number of parties, but where almost all parties contest all districts may take less time to impute the effective vote than a dataset with a small number of parties each of which contest

[6]In the KK example, although we could run WLS we also ran the original maximum likelihood model. To run through the ten imputed elections took only 13.5 minutes, roughly two-fifths the time of the original model, but this needed to be iterated on each of the M datasets, where we chose $M = 10$ for a total of 157 minutes including the time taken by the imputation model.

randomly[7]

References

Aitchison, J. 1986. *The Statistical Analysis of Compositional Data.* London: Chapman and Hall.

Bailey, Michael. 1998. "A Random Effects Approach to Legislative Ideal Point Estimation" Working Paper.

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*, Ser. 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 (May): 514-554.

Katz, Jonathan and Gary King. 1999. "A Statistical Model for Multiparty Electoral Data," *American Political Science Review,* Vol. 93, No.1 (March, 1999): 15-32.

King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 1998. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation" Paper presented at the Annual Meetings of the American Political Science Association, Boston.

Lange, Kenneth L., Roderick J. A. Little, and Jeremy M. G. Taylor. 1989. "Robust Statistical Modeling Using the t Distribution." JASA Vol. 84, No. 408: 881-896.

Lewis, Jeffrey B. 1998. "Estimating Voter Preference Distributions from Individual-Level Voting Data (with application to split-ticket voting.)" Working Paper.

Little, Roderick J. A. 1988. "Robust Estimation of the Mean and Covariance Matrix from Data with Missing Values." Appl. Statist. Vol. 37, No.1: 23-38.

Liu, Chuanhai. 1994. *Statistical Analysis Using the Multivariate t Distribution*

Liu, Chuanhai and Donald Rubin. 1994. "A Simple Extension to EM and ECM with Faster Monotone Convergence." Biometrika.

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* Vol. 47, No.5: 765-785.

McLachlan, Geoffrey and Thiriyambakam Krishnan. 1996. *The EM Algorithm and Extensions* Wiley.

Orchard, T. and Woodbury, M. A. 1972. "A missing information principle: Theory and applications." *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press, 697-715.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data,* Chapman Hall.

[7]For k covariates and n parties, of whom p partially contest some districts, there may be up to $k + \sum_{i=0}^{p} \binom{p}{i}(n-i-1)$ total variables in the imputation model.