A GENERAL SOLUTION TO NONIGNORABLE MISSING OUTCOMES IN BINARY CHOICE DATA

BRENTON KENKEL[†]

Work in progress. Comments welcome.

ABSTRACT. I provide a set of methods to estimate logistic regression coefficients when there is nonignorable missingness or measurement error in the outcome variable. Instead of requiring knowledge of the process by which the data went missing, these methods estimate bounds on the set of results that could be obtained under any assumption about the source of missingness. This approach is conservative by design, allowing analysts to discover exactly what kinds of results are—and are not—consistent with their data, given the uncertainty induced by missingness. Unlike existing approaches to nonignorable missing data, the estimators presented in this paper do not depend on untestable assumptions, nor do they require application-specific programming. I use simulations to illustrate which applied settings are most favorable to each of the three methods that I develop. I apply the methods to data from two previous studies, Oneal and Russett's (1997) analysis of the liberal peace and Lyall's (2010) analysis of counterinsurgency success, to examine the robustness of their results against potential miscoding of outcomes.

[†]Senior Research Assistant, Department of Politics, 028 Corwin Hall, Princeton University, Princeton, NJ 08544. Email address: bkenkel@princeton.edu.

Date: September 13, 2013.

I am grateful to Jeff Arnold, Matt Blackwell, Bear Braumoeller, Rob Carroll, Kevin Clarke, Hein Goemans, James Honaker, Jeff Marshall, Miguel Rueda, and Curt Signorino for their comments and advice. The errors that remain are, of course, my own. All code used in this paper is available in a public repository at http://github.com/brentonk/idlogit. I gratefully acknowledge funding from the Theory and Statistics Research Lab at the University of Rochester during the writing of this paper.

1. INTRODUCTION

Missing data is a major concern for virtually all quantitative research in political science. It is well known that removing incomplete cases from a dataset is liable to cause bias in statistical analysis, and that multiple imputation is superior to listwise deletion for analyzing data with missingness (King et al. 2001). Nonetheless, multiple imputation is not a catch-all solution to missing data problems. For the results of multiple imputation to be valid, the process that generates missingness must not be directly related to the values of the unobserved variables or to the population parameters that the analyst hopes to estimate. These conditions, which together constitute the *ignorability* assumption, refer to unobservable quantities, making it impossible to test whether they hold in a particular application (King et al. 2001, p. 51). What should an analyst do with missing data if she suspects that these conditions do not hold, or simply wishes to ensure that her results are robust against violations of ignorability?

The traditional approach to nonignorable missingness is to model the process that causes data to go missing. The most common example of this approach is selection models (Heckman 1979; Dubin and Rivers 1989), which are used to correct for bias when the outcome of interest is observed only in a non-random subset of the data. Like multiple imputation, these models are invaluable to social scientists, but are no panacea for the missing data problem. In fact, the main drawback of multiple imputation—its reliance on untestable assumptions—applies with equal force to modeling the missingness process. If the model is specified incorrectly, the estimates will be inconsistent. In applications where theory or substantive knowledge is not strong enough to determine how data go missing, the modeling approach is of limited use. Moreover, even when the missingness-generating process for a particular problem is known in advance, it may be analytically or computationally difficult to incorporate it into an estimator.

This paper presents a way to deal with missing data that does not depend on untestable assumptions or require application-specific programming. I focus on the case of missingness in a binary dependent variable, a common problem in both international relations and survey research, but the general approach to estimation that I employ is more widely applicable. Drawing from the partial identification literature in econometrics (see Manski 2003), I develop a set of methods to estimate bounds on the set of results that could be obtained under *any* assumption about the source of missingness. Rather than specifying a particular missingness model and using it to form an estimate, the methods presented in this paper find every result that every such model would yield. This approach is conservative by design, allowing analysts to discover exactly what kinds of results are—and are not—consistent with their data,

given the uncertainty induced by missingness. The methods here are thus an ideal robustness check for an analyst who wants to confirm that her statistical findings are not driven by an untestable assumption about how data went missing.

I consider a situation in which there is a binary outcome variable that is related to covariates through a standard logistic model. Some observations of the outcome have gone missing via a potentially nonignorable process whose form is unknown to the analyst. Since missingness creates some uncertainty, it is not possible to identify a consistent point estimate—a single set of coefficients that best fits the sample data—without making additional assumptions. That does not mean, however, that all possible coefficient estimates fit the data equally well. It is possible to show that some estimates are better than others (even though there may not be a unique optimum) without imposing any more conditions. In other words, we can rule out some coefficient estimates on the basis of the observed data alone.

To understand the basis of this approach, it may be instructive to consider an example in a simpler setting. Consider the problem of computing a student's grade from five in-class quizzes, each graded out of 100. Suppose the recorded scores are $(80, 85, 80, 95, *)$, where the last is unobserved due to some unknown factor. Given the observed data, what we conclude about the student's final average will depend on what we assume about how the last grade went missing. If it were due to a random error, such as in the gradebook software, it would be reasonable to assume the missing score is similar to the others. On the other hand, if it were missing because of concealment by the student, we might suspect the true score is quite low. What is true in any case is that the student's overall average cannot possibly be lower than 68 or greater than 88; in statistical terms, the sample mean is bounded between 68 and 88. Importantly, there is no assumption about how the final quiz went missing that would lead the instructor to conclude that the student should receive a grade outside this range. The purpose of this paper is to take this basic type of reasoning and apply it in the more complex setting of logistic regression.

I present three closely related methods to estimate bounds on logistic regression coefficients in the presence of nonignorable missing outcomes. Each method has distinct advantages and disadvantages, and the choice of estimator may depend on the goals of the analyst and the sample size of the application. The first estimator, which I call the fill-in method, is the most intuitive. It relies on the fact that there are only finitely many realizations of a set of missing binary outcomes, and involves sampling from these possibilities. This method works best when the amount of missingness is relatively small. The other two methods are adapted from econometric approaches to regression with an interval-measured

outcome (Manski and Tamer 2002; Chernozhukov, Hong and Tamer 2007). One of them, the moment inequality method, is computationally feasible but produces bounds that are too wide on average. This method is ideal as a "hard test" robustness check. The other, the minimum distance method, produces more accurate bounds but is computationally challenging, particularly in large samples, as it requires first-stage nonparametric regression estimates.

The methodology of finding bounds on estimates given relatively weak assumptions is taken from the econometric literature on partial identification (for an overview, see Manski 2003). The seminal contribution in this area is Manski (1990), who shows that it is possible to obtain bounds on estimated treatment effects with no additional assumptions when the outcome of interest is bounded. Similar methods have since been developed for nonresponse in surveys (Horowitz and Manski 1998), missing covariate data in experiments (Horowitz and Manski 2000), covariates that are discretized or measured only in intervals (Manski and Tamer 2002; Magnac and Maurin 2008), and missing treatment status (Molinari 2010). Each of these techniques exploits limits on the level of variation in the missing component to establish bounds on the set of parameter estimates that could be possibly be obtained from the data. Similarly, the methods developed in the present paper rely on the fact that the true value of a missing binary outcome must be either 0 or 1.

There has been some recent interest in partial identification in political science, particularly in the causal inference literature. Hanmer (2007) uses Manski's (1990) method to study the effect of election-day registration on voter turnout, and Glynn (2009) uses it to examine whether oil production and state weakness affect civil war occurrence. Quinn (2009) combines Manskian bounds with informative priors in a method of sensitivity analysis for causal inference from crosstabs; he uses this to assess whether aversion to jury duty affects voter registration. Poast (2010) applies Molinari's (2010) method to estimate the potential range of issue linkage's effect on compliance success in international negotiations, accounting for the fact that linkage attempts may be under-reported in failed negotiations. Imai and Yamamoto (2010) develop nonparametric bounds for treatment effect estimates when there is potentially systematic error in the measurement of treatment status. The work closest to the present paper is Aronow et al. (2013), which develops a double sampling method to deal with nonignorable missing outcomes in experimental data. The methods developed in the present paper, by contrast, apply to settings with observational data; however, unlike in Aronow et al. (2013), they do not allow for narrowing beyond the worst-case bounds.

The remainder of the paper proceeds as follows. In the next section, I lay out the basic model and some formal definitions. Section 3 presents three estimators for logistic regression with nonignorable missing outcomes, including details on computation. Section 4 presents the results of a simulation analysis that compares the performance of the three methods across samples of varying size. Two empirical applications are presented in Section 5. The first replicates Oneal and Russett's (1997) study of the liberal peace and examines whether its results are sensitive to the treatment of continued disputes. The second replicates Lyall's (2010) results about counterinsurgency success, finding additional grounds for Lyall's conclusion that democracy has no consistent effect. I conclude and discuss directions for future research in Section 6.

2. The Problem: Nonignorable Missing Outcomes

In this section, I provide formal definitions of the statistical model and the process that determines missingness, and I review some familiar results on when this process can be ignored in the course of estimating regression parameters. There are N units, $i = 1,\ldots,N$, for which the outcome of interest is a binary random variable Y_i. Each unit is also characterized by X_i, a row vector that contains k covariates, including an intercept.[1] The data-generating process is the logistic model,

$$\begin{aligned} Y_i^* &= X_i'\beta + \epsilon_i, \\ Y_i &= \mathbf{1}\{Y_i^* \geq 0\}, \end{aligned} \tag{2.1}$$

where $\epsilon_i \sim F_\epsilon$ and $X_i \perp\!\!\!\perp \epsilon_i$. If F_ϵ is logistic, as I maintain throughout the rest of the paper, the conditional probability of Y_i can be written as the familiar expression

$$\Pr(Y_i = 1 \,|\, X_i) = \frac{1}{1+\exp(-X_i'\beta)} \equiv \Lambda(X_i'\beta). \tag{2.2}$$

In addition, let M_i be an indicator of whether Y_i is unobserved. If $M_i = 0$, the dataset records the true outcome for unit i; if $M_i = 1$, the outcome is "blank," or $Y_i = *$. For ease of exposition, I adopt the convention that the dataset is ordered with incomplete observations first, so that $M_i = 1$ for all $i =$

[1] I assume throughout that there is no missingness in the covariates. A task for future work is to combine the estimators presented here with methods that correct for missingness in the covariates.

$1,\ldots,N_m$ and $M_i = 0$ for all $i = N_m + 1,\ldots,N$. A reasonably general representation of the missingness-generating process can be written as

$$\begin{aligned} M_i^* &= \varphi(X_i, Y_i; \theta) + \nu_i, \\ M_i &= \mathbf{1}\{M_i^* \geq 0\}, \end{aligned} \tag{2.3}$$

where $\nu_i \sim F_\nu$ and θ is a parameter vector. The researcher's goal is to obtain a consistent estimate of the logistic regression coefficients, β. If $N_m = 0$, estimation may proceed as usual via maximum likelihood; otherwise, additional assumptions are necessary in order to obtain a point estimate of the coefficients.

The most common approach to missingness in the political science literature is multiple imputation (MI) of the unobserved values. In general, MI consists of creating multiple datasets, each containing a different draw of the unobserved values from an estimate of their distribution conditional on the covariates and other auxiliary information, and combining the sample estimates obtained separately from each imputed dataset.[2] In order for MI estimates to be consistent, the missingness-generating process must be ignorable (see Rubin 1976, 1987). This requires, first, that the data be *missing at random* (MAR). Under MAR, conditional on the values of the observed data, the probability that a quantity is unobserved does not depend on its own value: holding all else fixed, the distribution of the observed values is the same as that of the unobserved values. In terms of the model presented above, this amounts to φ being constant in Y_i and the independence condition $\epsilon_i \perp\!\!\!\perp \nu_i$. The other condition for ignorability is that the missingness- and data-generating processes have *distinct parameters*. In the present context this implies that no element of θ is a function of β or vice-versa. Because both of the conditions involve unobservables, they cannot be tested in data—it is up to the researcher to decide whether it is plausible to assume ignorability.

The textbook example of nonignorable missingness is nonresponse in public opinion surveys. Take the example of a survey question about a sensitive topic, such as the respondent's feelings toward individuals of a certain ethnic group. A potential example of a MAR violation is if people with negative attitudes are less likely to answer the question. This situation would not violate MAR if both the negative attitudes and the chance of not answering the question are simply related to some external factor. For example, members of extremist parties may be more likely on the whole to have hostile feelings and to refuse to answer a question about ethnic relations. As long as the probability of refusal

[2]For details on multiple imputation and its variants, see Rubin (1987), Rubin (1996), Schafer (1999), King et al. (2001), and Honaker and King (2010).

is no higher among the extremist party members who actually have negative views than among those who do not, MAR may still be satisfied. Conversely, if we could predict the chance of non-response better if we knew a respondent's actual views in addition to his or her party membership, then MAR is violated and the missingness process is nonignorable.

Systematic measurement error poses problems similar to those of nonignorable missingness. Measurement error itself can be thought of as a missingness problem, in that we do not observe the variable of interest but rather some noisy indicator of its value. In fact, if we make no assumption at all about the relationship between the observed and actual values under measurement error, then from a statistical standpoint the situation is indistinguishable from missing data (Blackwell, Honaker and King 2011). In this case, to apply methods such as multiple imputation, one must assume the probability of mismeasurement satisfies the MAR and distinctness assumptions discussed above.

Standard methods for estimation under nonignorability require strong assumptions that, like MAR, typically cannot be tested in data. In order to obtain a point estimate of β when the data are not missing at random, the researcher must fully specify the missingness-generating process and estimate a joint model of missingness and the outcome variable. For the present problem, this would require knowing the functional form of φ and the distribution of ν_i. One could then write the likelihood function of (β, θ) given (Y, M, X) and use maximum likelihood or Bayesian methods for estimation. In the political science literature, selection models comprise the most common (perhaps exclusive) application of explicitly modeling the missingness mechanism. For example, the censored probit model is obtained from the above framework by setting $\varphi(X_i, Y_i; \theta) = X_i'\theta$ and assuming a standard bivariate normal distribution on (ϵ_i, ν_i) (Dubin and Rivers 1989).

For most applied researchers who have to deal with missing outcome data, modeling the missingness-generating process is not a desirable or feasible solution. The model must specify the form of the relationship between the covariates, the true outcome value, and the probability that the outcome goes unobserved. Yet it is impossible to test any assumptions about the relationship between Y_i and M_i, meaning a researcher must rely on strong prior knowledge in order to set up a plausible model. As in any structural modeling situation, the results may be highly sensitive to small perturbations in the data or specification (see, for example, Brandt and Schneider 2004, on censored probit). On the other hand, the best substantively justified functional form for φ is unlikely to be implemented in existing statistical software, and the corresponding likelihood function may be problematic for numerical optimization

or MCMC sampling. The applied researcher may be left with an unpleasant choice between a poorly specified model that is feasible to estimate and a well-specified model that cannot be computed.

In the next section, I describe a universal method to estimate logistic regression models in the presence of unobserved or mismeasured outcomes, i.e., a method that requires no application-specific programming. Instead of using a model or other assumptions to pin down a point estimate, the estimators presented here compute the entire set of results that could follow from any missingness model. The goal is to be as robust as possible, acknowledging the uncertainty that arises when some observations are missing, as opposed to using *a priori* conditions to try to fill in the gaps.

3. A General Solution: Bounding Coefficient Estimates

Nonignorable missingness in binary data may be thought of as a partial identification problem (see Manski 2003). If the chance of missingness depends on the value of the outcome in an unknown way, then the coefficients of the logistic regression equation (2.1) are not point identified. In the terminology of likelihood theory, there may be many sets of coefficients under which the observed data is equally likely, given our uncertainty about the process that generated missingness. However, this does not mean that we cannot learn anything from the data without assuming more about the missingness process. It may be possible to show that some coefficients are incompatible with the data under any model of missingness, just as it was possible to rule out an average below 68 or above 88 in the example in the introduction. The set of coefficients that cannot be ruled out in this way is called the *identified set*. The purpose of the estimators that I present here is to find bounds on this set—i.e., to place limits on what we could conclude from the sample data while accounting for the uncertainty due to missingness.

In fact, the approach I adopt is quite similar to the one used in the introductory example, which relied on the fact that the missing score must lie between 0 and 100. I exploit the fact that $Y_i \in [0, 1]$ in order to establish bounds on the set of logistic regression coefficient estimates that could be obtained from a particular sample under some realization of the missing outcomes. I provide three distinct methods to estimate these bounds from sample data. These techniques require no application-specific assumptions or programming, making them a truly general solution to nonignorability. All three are computationally intensive, but each has other particular advantages and drawbacks. The first two, the "fill in" and moment inequality methods, are always feasible but may respectively underestimate or overestimate the width of the bounds. The final method, minimum distance, is more accurate but

requires some auxiliary nonparametric estimates that may pose formidable computational challenges in large samples.

Before discussing the methods, it should be noted that these estimated bounds are conceptually distinct from confidence intervals. Returning to the example of a student's course grade, we were able to show that the average of the five scores must be between 68 and 88. This is a statement about the sample average, not the population average. The bounds depend only on the fact that each score must lie between 0 and 100; no assumptions about sampling or the population distribution are necessary. The construction of a confidence interval from this data would be a separate, additional step. Just as confidence sets in ordinary estimation situations are intervals that contain the point estimate, in this case they would be a superset of the bounds on the point estimate. The construction of confidence intervals under partial identification is a complex topic (see Imbens and Manski 2004; Chernozhukov, Hong and Tamer 2007; Romano and Shaikh 2008), and will be studied for the logistic regression case in future work.

3.1. **Fill-in method.** Because Y_i is binary, there are only 2^{N_m} possible true realizations of the sample data. If N_m is small enough (typically below 20), it is feasible to compute all of the estimates of β that could be obtained from a particular sample, given some realization of the unobserved outcomes. Such results would give a researcher a full picture of which coefficient estimates could or could not be supported by the sample data, and which results are robust to assumptions about the unobserved outcomes. For example, if the coefficient on a particular covariate is positive in each of the 2^{N_m} replicates, a researcher can be confident that no assumption about the missingness-generating process would change the sign of the point estimate. Unfortunately, the amount of computation required for this approach grows exponentially with the number of missing outcomes. On current hardware, it could take years to calculate the full set of possible estimates when $N_m \geq 30$.

A natural second-best solution when N_m is too large for a full characterization is to randomly sample the possibilities. The researcher could "fill in" the unobserved values at random, run a logistic regression on the completed dataset, and repeat. Formally, the procedure would be to repeat the following M_1 times:

(1) Draw $Y' = (Y'_1, \ldots, Y'_{N_m})$ from a uniform distribution over $\{0,1\}^{N_m}$

(2) Set $Y = (Y'_1, \ldots, Y'_{N_m}, Y_{N_m+1}, \ldots, Y_N)$ and run a logistic regression of Y on X

(3) Store the coefficient estimates $\hat{\beta}$

The bounds on each coefficient are then its minimal and maximal value among the recorded estimates. This fill-in procedure is not a form of multiple imputation, despite their superficial similarity, as the missing values are not sampled from a conditional distribution, and the coefficients obtained in each iteration are not averaged to obtain a single point estimate.

The obvious appeal of this method is its simplicity. It requires no special numerical methods and is trivial to implement in any software that performs logistic regression. More broadly, the notion of filling in the unobserved outcomes suggests a general heuristic for when the methods set out in this paper may be useful—namely, when an analyst believes it would be useful to know the results under every possible combination of some set of potential outcomes. For example, the methods introduced in this paper could be used as a somewhat conservative approach to measurement error in binary data. If the researcher suspects that the response variable is poorly measured in a subset of units, these methods will yield the set of coefficients that could plausibly be obtained under some assumption about the measurement process. Similarly, if there are two different coding rules for the same binary dependent variable, the techniques presented here can be used to determine how sensitive the estimates are to the choice of coding scheme.

The problem with the "fill in" method is that it will underestimate the range that the coefficients could actually take, especially when N_m is large. Indeed, it should be impossible for this method to produce bounds that are too wide: any coefficient found in the process is, by definition, an estimate that could be produced under some realization of the missing outcomes. However, the probability of actually reaching the boundary via random sampling decreases rapidly with the number of unobserved outcomes. Moreover, because this method is not justified with reference to any asymptotic theory, it is unclear whether or how it could be used to construct confidence regions or perform hypothesis tests. (The same is true of collecting the full set of possible estimates when it is feasible to do so.) The next two methods that I consider are more complex to derive and compute, but have sounder justifications and are based on criterion functions that can be used for inference about population parameters.

3.2. **Moment inequality method (MIM).** The most popular framework for estimation of partially identified models in the econometric literature is the method of moment inequalities (Chernozhukov, Hong and Tamer 2007). This method entails using moment inequality conditions, which are equations of the form

$$\mathbb{E}[f(X_i, Y_i; \beta)] \leq 0, \tag{3.1}$$

to characterize the population identified set. The population identified set refers to the set of parameters that generate the same distribution of observed quantities as the true set of coefficients. If there is no chance of missingness, the identified set is a singleton that contains only the true parameter value; otherwise, if data may go missing by a process unknown to the analyst, the identified set will typically contain many sets of coefficients. Under broad regularity conditions, the identified set can be consistently estimated using the sample analogue of (3.1). This approach is an extension of the generalized method of moments (GMM), and hence is easily applied to many models that arise in the economic literature.

It is straightforward to apply moment inequalities to the problem of unobserved outcomes in binary data. The following moment restriction on the (potentially unobserved) outcome variable is immediate from (2.2):

$$\mathbb{E}[Y_i - \Lambda(X_i'\beta) \,|\, X_i] = 0, \tag{3.2}$$

which implies the unconditional moment restriction

$$\mathbb{E}[h(X_i)(Y_i - \Lambda(X_i'\beta))] = 0 \tag{3.3}$$

for any well-behaved function h.[3] If Y_i were observed for the full sample, a GMM estimator of β could be constructed via the sample analogue of (3.3). The partial identification approach is to derive inequalities from (3.3) that can be expressed in terms of observables. Define the outcome bounds $Y_{0i} = 0 \cdot M_i + Y_i \cdot (1 - M_i)$ and $Y_{1i} = 1 \cdot M_i + Y_i \cdot (1 - M_i)$, so that $Y_{0i} \leq Y_i \leq Y_{1i}$ for each unit i. If $h(X_i) \geq 0$,

$$\mathbb{E}[h(X_i)(Y_{0i} - \Lambda(X_i'\beta))] \leq 0 \leq \mathbb{E}[h(X_i)(Y_{1i} - \Lambda(X_i'\beta))]. \tag{3.4}$$

It is possible that more than one β solves (3.4), in which case the model is not point-identified. That is, more than one set of parameters may generate the same probability distribution over the observed quantities, (X_i, Y_{0i}, Y_{1i}). The purpose of estimation under partial identification is to characterize this set of observationally equivalent parameters.

As in GMM, the estimation principle for moment inequalities is to replace the population expectations with their sample analogues. In a partially identified model, however, the goal is not necessarily to find a single parameter value that best satisfies the sample moment conditions. Instead, we seek the *set* of parameter values that satisfy them (or come closest to doing so). For the problem examined here, the

[3] In what follows, all equalities and inequalities are assumed to apply component-wise if h is vector-valued.

moment inequality estimator is the set of all β that solve the system

$$\begin{aligned}\left(\frac{1}{N}\sum_{i=1}^{N} h(X_i)(Y_{0i} - \Lambda(X_i'\beta))\right)_+ &= 0,\\ \left(\frac{1}{N}\sum_{i=1}^{N} h(X_i)(\Lambda(X_i'\beta) - Y_{1i})\right)_+ &= 0,\end{aligned} \tag{3.5}$$

where $(t)_+ = \max\{t, 0\}$. The set of solutions to (3.5) is the estimate of the identified set.

The first issue to consider in applying the moment inequality estimator is the choice of h. This is analogous to the well-known problem of selecting instruments for models defined by conditional moment inequalities in GMM estimation, where any function of the information set may be a valid instrument (Dominguez and Lobato 2004). For the logistic regression model considered here, an appealing choice is to let $h(X_i) = Z_i$, some positive-valued affine transformation of X_i. An example would be to rescale each covariate to the unit interval, $Z_{ij} = (X_{ij} - \min X_{\cdot j})/(\max X_{\cdot j} - \min X_{\cdot j})$ for each $j = 1, \ldots, k$. Any such selection of $h(X_i)$ makes the inequality conditions (3.5) equivalent to the first-order conditions for maximum likelihood estimation of β when Y_i is fully observed. In other words, as the number of missing outcomes goes to zero, the identified set collapses to the maximum likelihood estimate of the coefficients. A similar choice would be to set

$$h(X_i) = (Z_{i1}, \ldots, Z_{ik}, Z_{i1}^2, \ldots, Z_{ik}^2, \cdots, Z_{i1}^d, \ldots, Z_{ik}^d) \tag{3.6}$$

for some integer $d \geq 1$, where Z_i is defined as before. Greater choices of d imply more restrictions on the identified set, and hence narrower estimated bounds. This approach is used by Cerquera, Laisney and Ullrich (2012) in their simulation study of partial identification estimators in a linear regression model with an interval-measured covariate. Using the same framework, Manski and Tamer (2002, pp. 530–531) show that the estimated bounds from a moment-inequality estimator converge to a superset of the population identified set. Therefore, the fact that the bounds narrow with d is innocuous, and in fact helpful for precise estimation.

The other main issue for estimation of the identified set is computation, as may be difficult to characterize the moment inequality estimator analytically. Estimation therefore requires numerical methods. Standard hill-climbing techniques, which are designed to find a single local maximum of a criterion function, are not as useful for characterizing a set of maximizers. Instead, some sort of grid search

or simulation-based technique is required. I use an approach based on simple random sampling, exploiting the fact that the results from the fill-in method can give us a reasonable idea of the center and spread of the identified set:

(1) Draw M_1 members of the fill-in estimate of the identified set, $\beta^{(1)}, \ldots, \beta^{(M_1)}$
(2) Compute the sample mean $\bar{\beta}$ and covariance matrix $\hat{\mathrm{V}}[\beta]$ of $\beta^{(1)}, \ldots, \beta^{(M_1)}$
(3) For each $i = M_1 + 1, \ldots, M_1 + M_2$:
    (a) Draw a candidate value $\beta^{(i)}$ from $N(\bar{\beta}, \gamma\hat{V}[\beta])$, where $\gamma \geq 1$
    (b) Store $\beta^{(i)}$ as a member of the identified set if it satisfies (3.5); otherwise discard it

Computation is one of the main advantages of the moment inequality method over the fill-in method. In most applied problems we have $N_m \gg k$, so the dimensionality of the search space is much lower. Moreover, unlike the Fisher scoring method typically used to fit logistic regression models, checking the inequality constraints (3.5) requires no costly matrix inversions. However, just as the fill-in method is almost guaranteed to yield bounds that are too narrow, the bounds obtained by moment inequalities will usually be too wide. As mentioned above, this is a consequence of the fact that the model is defined by an infinite number of non-equivalent moment conditions, only finitely many of which can be checked in practice. Although these two methods are imperfect, they are still useful—especially in combination with one another. If even the too-narrow fill-in bounds contain 0 for the coefficient on some covariate, then we know that any conclusion about the direction of the relationship on the basis of the sample data would be sensitive to assumptions about missingness. Conversely, if even the too-wide moment inequality bounds fail to contain 0, a researcher can be confident in the sign of the sample estimate.

3.3. **Minimum distance method (MDM).** The final estimation technique that I discuss, the minimum distance method, does not suffer from the over- or under-coverage problems of the previous two. It is based on Manski and Tamer's (2002, p. 532) results on partial identification for regression with an interval-measured outcome. Manski and Tamer characterize the population identified set analytically as the set of β that satisfy

$$\begin{aligned} g_0(X_i) - f(X_i;\beta) &\leq 0, \\ f(X_i;\beta) - g_1(X_i) &\leq 0, \end{aligned} \tag{3.7}$$

for almost all X_i, where $g_0(X_i) = \mathbb{E}[Y_{0i}\,|\,X_i]$, $g_1(X_i) = \mathbb{E}[Y_{1i}\,|\,X_i]$, and $f(X_i;\beta) = \mathbb{E}[Y_i\,|\,X_i]$ under the population parameter β. In the logistic regression case, we have $f(X_i;\beta) = \Lambda(X_i'\beta)$. As an estimator of the identified set from sample data, Manski and Tamer suggest the set of β minimizing the criterion

function

$$Q(\beta) = \frac{1}{N}\sum_{i=1}^{N}\Big\{(f(X_i;\beta) - \hat{g}_0(X_i))^2 \mathbf{1}\{f(X_i;\beta) < \hat{g}_0(X_i)\} + (f(X_i;\beta) - \hat{g}_1(X_i))^2 \mathbf{1}\{f(X_i;\beta) > \hat{g}_1(X_i)\}\Big\} \tag{3.8}$$

where $\hat{g}_0(X_i)$ and $\hat{g}_1(X_i)$ are consistent estimates of $g_0(X_i)$ and $g_1(X_i)$. Manski and Tamer prove that this minimum distance estimate converges to the population identified set defined by (3.7) as $N \to \infty$, so in large samples the bounds should be neither too wide nor too narrow on average.

In some cases, particularly when the number of observations is small, there may be no β such that $Q(\beta) = 0$, in which case there will typically be a unique minimizer. To avoid this sort of degenerate solution, Manski and Tamer propose a modified estimator, the set of β such that $Q(\beta) \leq c_N$. If the critical value c_N is taken from a sequence such that $c_N \to 0$ as $N \to \infty$, the estimator retains its consistency. A greater concern for the applied researcher is data-driven selection of the critical value such that the estimator has reasonable finite-sample properties. In the simulations in the next section, I examine three potential selections of critical value when $Q(\beta)$ has no roots: the minimum, median, and maximum value of $Q(\beta)$ among the coefficients drawn via the fill-in method.

The main obstacle to implementing the minimum distance method is computing the nonparametric functional form estimates $\hat{g}_0$ and $\hat{g}_1$. I show in the Appendix that if ϵ_i and ν_i are independent,

$$\begin{aligned} g_0(X_i) &= F_\nu(-\varphi(X_i, 1; \theta))\Lambda(X_i'\beta), \\ g_1(X_i) &= 1 - F_\nu(-\varphi(X_i, 0; \theta))(1 - \Lambda(X_i'\beta)). \end{aligned} \tag{3.9}$$

Neither of these corresponds to a generalized linear model or even a generalized additive model. Therefore, unless the researcher has strong prior knowledge about the missingness process, a fully nonparametric method is necessary to estimate $\hat{g}_0$ and $\hat{g}_1$. Applicable methods include kernel regression (see Pagan and Ullah 1999, ch. 3) and local logit (Frölich 2006), though typically not the single-index model of Klein and Spady (1993). These estimators are imprecise in small samples due to the curse of dimensionality, but they can be highly costly or infeasible to compute in large samples, especially when using cross-validation to select an optimal bandwidth. One potential solution is to estimate the bandwidth from random subsamples and then rescale to account for the difference in sample size, as proposed by Racine (1993).

Once the estimates $\hat{g}_0$ and $\hat{g}_1$ have been obtained, estimation of the minimum distance bounds is similar to the moment inequality method. I use the following procedure:

(1) Draw M_1 members of the fill-in estimate of the identified set, $\beta^{(1)}, \ldots, \beta^{(M_1)}$

(2) Compute the sample mean $\bar{\beta}$ and covariance matrix $\hat{\mathrm{V}}[\beta]$ of $\beta^{(1)}, \ldots, \beta^{(M_1)}$

(3) For each $i = M_1 + 1, \ldots, M_1 + M_2$:

   (a) Draw a candidate value $\beta^{(i)}$ from $N(\bar{\beta}, \gamma\hat{V}[\beta])$, where $\gamma \geq 1$

   (b) Compute and store $Q(\beta^{(i)})$, given by (3.8)

(4) The estimated identified set is $\{\beta^{(i)} \,|\, Q(\beta^{(i)}) = 0\}$ if this is non-empty; otherwise, it is $\{\beta^{(i)} \,|\, Q(\beta^{(i)}) \leq \hat{c}_r\}$, where $\hat{c}_r$ is the r'th order statistic of $\{Q(\beta^{(i)}) \,|\, 1 \leq i \leq M_1\}$

Like the moment inequality method, this also involves searching over the k-dimensional covariate space rather than the 2^{N_m}-dimensional missing outcome space, meaning it has the same potential computational advantages in large samples. Of course, depending on the number of covariates, this may be offset by the difficulty of nonparametric estimation in the previous step.

4. Simulations

I carry out a Monte Carlo experiment to compare the finite-sample performance of the three methods. The data- and missingness-generating processes are

$$Y_i^* = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i, \tag{4.1}$$

$$M_i^* = \theta_0 + \theta_1 Y_i X_{1i} + \nu_i, \tag{4.2}$$

where $\beta = (1, 0, -1)$, $\theta = (-1.5, 1)$, and ϵ_i and ν_i are each i.i.d. logistic. The missingness-generating process is nonignorable, as the realized value of the outcome variable Y_i enters into the probability that it is unobserved. The covariates are $X_{1i} \sim U(-1, 1)$ and $X_{2i} \sim U(-1, 1)$, making the *ex ante* probability of missingness about 19%.

The first step in the analysis is to compute $\mathscr{B}$, the population identified set. Recall that Manski and Tamer (2002) show that the population identified set is equivalent to the set of solutions to (3.7). These are the sets of coefficients that could, under some missingness model (not necessarily the true one), generate the same distribution over the observable quantities (X_i, Y_{0i}, Y_{1i}) as the true coefficients do. To compute $\mathscr{B}$, I first use the assumed missingness model (4.2) to characterize the functions g_0 and g_1, as given in (3.9). I then perform a grid search over $\Re^3$ to find the points that satisfy (3.7). The identified set itself cannot be expressed in closed form, but its bounds along each dimension are given in the table below.

	lower	upper
β_0	0.41	1.24
β_1	−0.57	0.24
β_2	−1.17	−0.57

Table 1. Bounds of $\mathscr{B}$, the identified set for β.

The goal of the experiment is to characterize how well each of the estimators described in the previous section recovers the true bounds of $\mathscr{B}$. Of particular concern is how sample size affects the performance of the estimators, so the experiment is repeated 100 times each for $N = 100$, 1,000, and 10,000. In each iteration, the bounds on $\mathscr{B}$ are estimated via the following:

- Fill-in method, using 1,000 draws.
- Moment inequality method (MIM) for $d = 1, 10, 20$ powers of X rescaled to the unit interval, using 20,000 draws each.
- Minimum distance method (MDM), using 20,000 draws each. $\hat{g}_0$ and $\hat{g}_1$ are estimated via a locally constant kernel regression.[4] If $Q(\beta) > 0$ for all candidate points, the critical value is chosen using the minimum, median, or maximum of $Q(\beta)$ among the coefficients drawn via the fill-in method.

The candidate points for MIM and MDM are chosen from a multivariate normal distribution centered around the mean of the fill-in results. The variance of the distribution is $\gamma = 9$, 36, and 400 times that of the fill-in results for $N = 100$, 1,000, and 10,000 respectively; these scaling factors were chosen so that about half of the MIM ($d = 1$) candidate points would be accepted.

The results of the simulation are given in Table 2 and illustrated (for the two greater sample sizes) in Figure 1. Our main interest is in how accurately each method characterizes the bounds on β_1 and β_2, the coefficients for the two covariates. First, the fill-in method appears to perform best when there are few observations in the dataset. The mean squared error of the estimated bounds under the fill-in method decreases with N, but this is mainly due to lower variance—the bias actually tends to become worse. In particular, the average estimated bounds on β_1 and β_2 under the fill-in method are closer to the true values when $N = 1{,}000$ than when $N = 10{,}000$. The fill-in bounds appear to narrow with N, and at a certain point they become too narrow. This pattern is likely because of the difficulty

[4] I use the R package **np** (Hayfield and Racine 2008) to estimate the kernel regressions. For the trials with $N = 100$, bandwidths are selected via cross-validation in each iteration in the regression of Y_0 on X. To ensure that $\hat{g}_1(X_i) \geq \hat{g}_0(X_i)$, the same bandwidths are then used to estimate $\hat{g}_1$. In the trials with larger N, to reduce computation time, I use the medians of the bandwidths selected under $N = 100$ after rescaling to account for the convergence rate (Racine 1993; Hayfield and Racine 2008). This procedure is similar to that of Cerquera, Laisney and Ullrich (2012).

(a) $N = 100$.

		β_0		β_1		β_2	
		Avg.	RMSE	Avg.	RMSE	Avg.	RMSE
Fill-in	lower	**0.42**	**0.22**	−0.85	0.54	**−1.41**	**0.51**
	upper	**1.14**	**0.28**	0.34	0.43	**−0.18**	**0.57**
MIM ($d = 1$)	lower	0.23	0.30	−2.19	1.74	−3.01	1.95
	upper	1.65	0.57	1.83	1.69	1.35	2.00
MIM ($d = 10$)	lower	0.25	0.29	−1.60	1.24	−2.15	1.21
	upper	1.55	0.51	0.95	0.94	0.24	1.00
MIM ($d = 20$)	lower	0.28	0.29	−1.41	1.18	−1.90	1.10
	upper	1.46	0.50	0.75	0.91	0.08	0.94
MDM (min)	lower	0.57	0.36	−0.39	0.34	−0.75	0.58
	upper	1.02	0.34	0.15	0.36	−0.22	0.61
MDM (median)	lower	0.47	0.25	**−0.55**	**0.33**	−0.92	0.56
	upper	1.13	0.32	**0.30**	**0.32**	−0.06	0.66
MDM (max)	lower	0.37	0.25	−0.77	0.63	−1.19	0.82
	upper	1.29	0.49	0.48	0.53	0.11	0.84

(b) $N = 1{,}000$.

		β_0		β_1		β_2	
		Avg.	RMSE	Avg.	RMSE	Avg.	RMSE
Fill-in	lower	0.64	0.24	−0.42	0.18	**−0.92**	**0.27**
	upper	0.85	0.39	−0.04	0.30	**−0.54**	**0.10**
MIM ($d = 1$)	lower	0.30	0.12	−1.48	0.92	−2.13	0.97
	upper	1.37	0.16	1.10	0.87	0.68	1.26
MIM ($d = 10$)	lower	0.30	0.12	−1.25	0.70	−1.79	0.64
	upper	1.37	0.16	0.79	0.57	0.18	0.76
MIM ($d = 20$)	lower	0.30	0.12	−1.19	0.65	−1.72	0.58
	upper	**1.36**	**0.15**	0.75	0.54	0.09	0.69
MDM	lower	**0.46**	**0.09**	**−0.45**	**0.15**	−0.88	0.31
	upper	1.07	0.19	**0.22**	**0.09**	−0.25	0.35

(c) $N = 10{,}000$.

		β_0		β_1		β_2	
		Avg.	RMSE	Avg.	RMSE	Avg.	RMSE
Fill-in	lower	0.72	0.31	−0.31	0.26	−0.79	0.38
	upper	0.79	0.45	−0.19	0.43	**−0.67**	**0.10**
MIM ($d = 1$)	lower	0.31	0.10	−1.57	1.01	−2.21	1.04
	upper	1.41	0.17	1.16	0.92	0.76	1.34
MIM ($d = 10$)	lower	0.31	0.10	−1.30	0.74	−1.85	0.68
	upper	1.40	0.17	0.83	0.60	0.19	0.76
MIM ($d = 20$)	lower	0.31	0.10	−1.27	0.71	−1.78	0.61
	upper	1.40	0.16	0.78	0.54	0.10	0.68
MDM	lower	**0.46**	**0.06**	**−0.49**	**0.09**	**−0.98**	**0.19**
	upper	**1.13**	**0.11**	**0.21**	**0.05**	−0.36	0.21

Table 2. Averages and root mean squared error of the estimated bounds of the identified set across 100 Monte Carlo iterations for each N. Entries in bold represent the lowest RMSE across methods for the bound in question. The parenthetical for MIM represents the number of powers of X (rescaled to the unit interval) used to form the matrix of moment conditions. The parenthetical for MDM in (a) represents the rule for selecting the critical value as a quantile of $Q(\beta)$ evaluated over the fill-in set when there are no exact zeroes of $Q(\cdot)$.

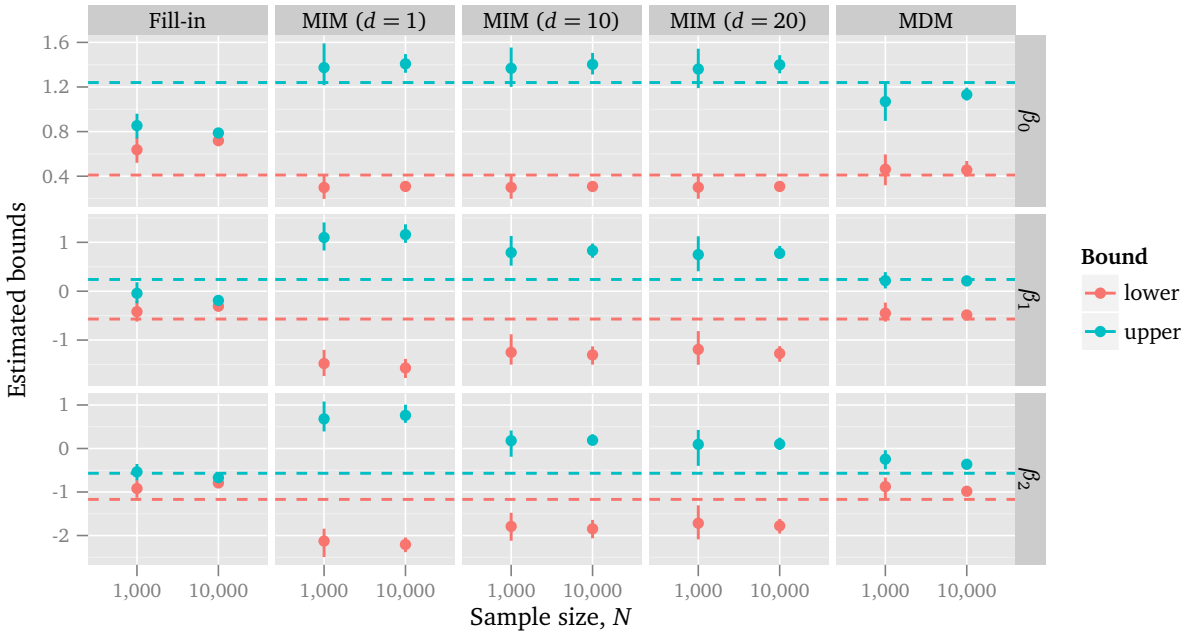

Figure 1. Illustration of the simulation results for $N = 1{,}000$ and $N = 10{,}000$. The dots represent the mean of the sampling distribution; the lines represent the range from the 2.5th to 97.5th percentiles of the sampling distribution. The dotted lines denote the population bounds.

in searching over 2^{N_m} missing outcome values, a problem whose dimension grows exponentially with N. For a fixed number of fill-in iterations, the chance of having a draw whose coefficient is near the boundary of the identified set decreases rapidly with the number of missing outcomes. A potential solution to this "anti-consistent" behavior is to let the number of fill-in draws grow with N, but this would be intolerably costly from a computational standpoint, especially because the cost of running a logistic regression also increases with N. A better approach is to use one of the other two methods, both of which search over spaces whose dimensionality does not increase with N.

As expected, the moment inequality method yields a conservative estimate—bounds that are too wide. In all cases, even with $d = 20$, the average estimated bounds under MIM are wider than those from the fill-in method or (with one exception) MDM. Moreover, the bias of MIM does not appear to improve with sample size. Moving from 1,000 to 10,000 observations, the average estimated bounds under MIM approximately stay the same; if anything, they become even wider. Among the three MIM estimators examined, there appears to be no drawback to using the one with the greatest number of restrictions ($d = 20$), which produces on average the narrowest bounds and lowest mean squared error

of the three. There is some additional computational cost, as the number of moment conditions to be evaluated is $2d + 1$, but in practice the extra calculation time is negligible.

Unlike the other two estimators, the minimum distance method gets close to the bounds of the true identified set when N is sufficiently large. With $N = 10{,}000$ and even $N = 1{,}000$, the average MDM estimates of all six bounds are reasonably near their actual values. By contrast, neither of the other two even yields the right sign for all six bounds. This is perhaps unsurprising, since the MDM estimator is a direct analogue of (3.7), which defines the identified set. In the larger samples, values of β such that $Q(\beta) = 0$ were always found, making it unnecessary to consider the choice of critical value. With $N = 100$, a zero of the criterion function was unavailable in 33% of iterations. The estimates under the "minimum" and "median" rules for choosing the critical value (using the value of $Q(\beta)$ evaluated on the fill-in coefficients) have approximately the same mean squared error; the latter does slightly better, but the number of trials is small and this may simply be due to sampling error. Both methods do better than the "maximum" rule.

Overall, the minimum distance method seems to be the best choice for estimating the bounds of the identified set in this case. It is clearly the best method in large samples, and the mean squared error of the "minimum" and "median" variants is competitive with that of the fill-in method with $N = 100$. The main caveat about these results is that the covariate structure (two i.i.d. uniform random variables) is relatively favorable to MDM. Even with 1,000 observations, the covariate space is likely to be densely populated, making it easy to obtain credible nonparametric estimates of $g_0(X_i)$ and $g_1(X_i)$. It remains to be seen whether the advantages of MDM hold up when there is more variance in the regressors or simply more regressors.

5. Applications

In this section, I apply the methods developed earlier in the paper to reanalyze two prominent studies of conflict.

5.1. **Continued disputes and the liberal peace.**

The first application I consider is Oneal and Russett's (1997) study of the liberal peace—specifically, whether their results are sensitive to their assumptions about ongoing disputes. Oneal and Russett analyze a set of $N = 20{,}990$ politically relevant dyad-years from 1950 to 1985. The outcome variable is the occurrence of a militarized international dispute between the two states. Oneal and Russett treat the onset of conflict and its continuation exactly the same, coding $Y_i = 1$ for each year that a dyad either entered or continued a militarized dispute. The

947 cases coded as disputes represent 405 onsets and 542 ongoing disputes. In later work, Oneal and Russett write that the continuation of a dispute should be treated the same as an active renewal, because rational leaders can decide at any time to change their policy (Russett and Oneal 2001, pp. 309–310). Other scholars, such as Bennett and Stam (2004, pp. 53–54), argue that dispute onset and continuation occur by separate political processes. In this case, the presence of an ongoing dispute can be thought of as a censoring factor, preventing us from observing the true value of interest: would these two states have entered a dispute in this year if they had had the opportunity to do so? Since states in an ongoing dispute do not have this opportunity, Bennett and Stam drop these cases from their dataset.

But if conflict onset is indeed censored in observations with an ongoing dispute, making it a type of missing data, the appropriate solution is not to remove these observations from the data. Listwise deletion causes bias unless the values are *missing completely at random* (MCAR), meaning the chance of missingness is unrelated to all other observed information (King et al. 2001). In this application, MCAR would require that the probability of an ongoing dispute be unrelated to all other variables. It is trivial to show that this does not hold; for example, ongoing disputes are more common between contiguous states, as illustrated in the following cross-tabulation.

		Ongoing dispute	
		0	1
Contiguity	0	14,112 (98.4%)	226 (1.6%)
	1	6,336 (95.2%)	316 (4.8%)

Cross-tabulation of contiguity and the occurrence of an ongoing dispute. The χ^2 statistic is 180.7 ($p < 0.01$).

The other traditional approaches to missing data are also unpalatable in the case of ongoing disputes. Both of the necessary conditions for ignorability are questionable here. First, MAR requires that knowing what would have happened in the censored cases—i.e., whether disputes would have began in the absence of an ongoing one—would not allow us to better predict the occurrence of an ongoing dispute in the first place. For example, suppose we could know whether the U.S. and Japan would have gone to war in 1942, had they not already been at war then. In order for MAR to hold, knowing this counterfactual should not allow us to better predict the existence of an ongoing dispute between U.S. and Japan in 1942. Second, the distinct parameters condition means the probability of an ongoing dispute cannot be a function of the parameters that determine the likelihood of conflict onset. A dispute must begin in order to continue, making this condition dubious. However, if the censoring process is nonignorable, then we must model it in order to obtain a point estimate. In this setting, that means we

must correctly model the determinants of dispute continuation—a formidable task in its own right—in order to estimate the determinants of dispute onset. The bounding approach taken in this paper is straightforward by comparison.

Treating the outcome as missing in cases of ongoing disputes, I use the bounding methods presented above to replicate Oneal and Russett's baseline model, "Equation 1" (p. 278).[5] This model contains six covariates, which I briefly review here; see Oneal and Russett (1997, pp. 273–277) for full details:[6]

Democracy$_L$: The lower of the two Polity III scores in the dyad, scaled down by a factor of 10 so that its range is $[-1, 1]$.

Growth$_L$: The lower of the two countries' average GDP growth over three years.

Alliance: An indicator of whether the two countries are either allied with each other or both allied with the United States.

Contiguity: An indicator of whether the two countries are contiguous.

Capability Ratio: The ratio of the stronger state's Composite Index of National Capabilities to the weaker state's, scaled down by a factor of 100.

Dependence$_L$: The lower of the two countries' bilateral-trade-to-GDP ratios, scaled up by a factor of 100.

The large sample size makes the fill-in method inappropriate, but also hinders computation of a kernel regression for the minimum distance method. Therefore, I apply the moment inequality method, using specifications of degree $d = 1$, 2, and 3. I compute the estimates via random sampling as described in Section 3.2.[7] It is important to remember that this method is especially conservative, in that it produces bounds that are too wide on average, making it a hard test for robustness. If the estimated bounds on a coefficient do not include 0, that is a strong sign that the direction of the estimate is robust across assumptions about missingness; however, if the bounds do include 0, that is only a weak sign of non-robustness.

The results of the analysis are plotted in Figure 2. For four of the six variables, including the crucial joint democracy measure, the bounds do not cross 0. Therefore, the results of the reanalysis demonstrate strongly that the negative relationship between joint democracy and conflict, as well as that of

[5] This is also the model replicated by Beck, Katz and Tucker (1998) in their analysis of duration dependence.

[6] Scalings were applied so that the covariates would be of roughly the same magnitude, aiding numerical stability.

[7] The number of candidate points examined and the number found to be in the estimated identified set are as follows: $d = 1$ used 100,000 candidates and 21,982 members; $d = 2$ used 500,000 candidates and 1,293 members; $d = 3$ used 1,000,000 candidates and 539 members.

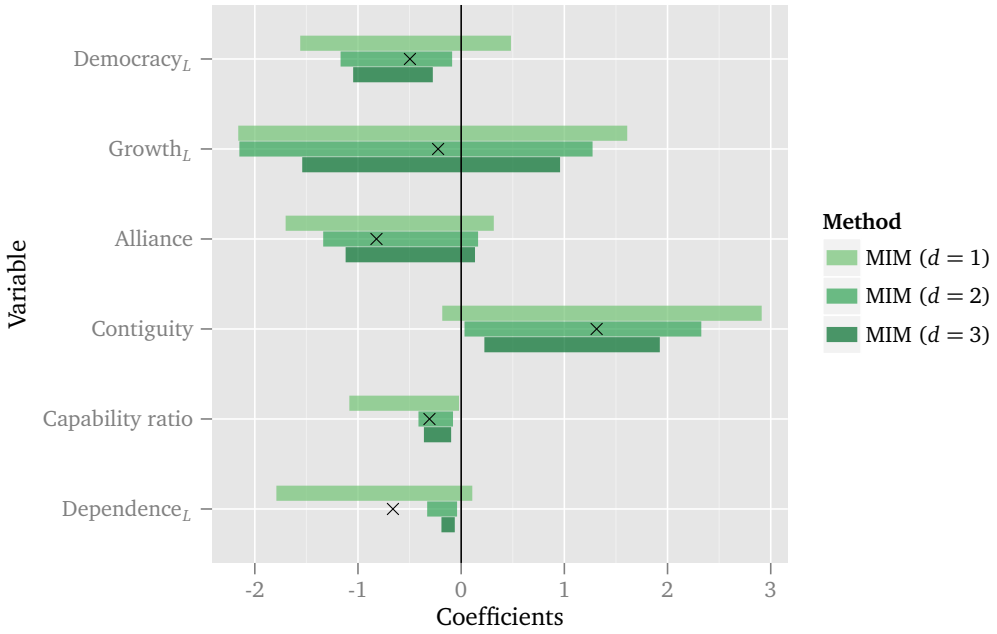

Figure 2. Estimated bounds in the reanalysis of Oneal and Russett's (1997) data. Values marked with an "x" are the estimated coefficient from the original analysis, with all ongoing disputes coded as 1s.

trade dependence, holds across any possible assumption about continued disputes. This result is impressive, seeing as there are more cases coded as missing (542) than as conflicts (405). On the other hand, the results at least suggest that the apparent pacific effect of economic growth may be sensitive to how continued disputes are modeled. In fact, the upper bound of the coefficient on Growth$_L$ in the fill-in estimate (not plotted) is -0.02; since the fill-in bounds are typically too narrow in such a large sample, it is reasonable to infer that the sign of the coefficient for growth is not robust.

The results of the bounds analysis can also be used to address the magnitude of substantive effects. As an example, I examine the relationship between Democracy$_L$ and the predicted probability of a militarized dispute. I take 50 coefficient vectors at random from the estimated identified set (MIM, $d = 3$) and plot the predicted probability of a dispute across the range of Democracy$_L$, holding all other covariates at their medians.[8] The results are shown in Figure 3, as well as the estimated relationship under Oneal and Russett's original model. The reduction in the chance of a dispute due to moving from a fully autocratic dyad (Democracy$_L = -1$) to a fully democratic dyad (Democracy$_L = 1$) ranges from 1.2 percentage points to 4.8 percentage points. The original coefficients imply a reduction of 3.2 percentage points, firmly within the middle of this range. Since the MIM bounds are conservative, it appears that

[8] I sample from the estimated identified set so that the figure will be legible.

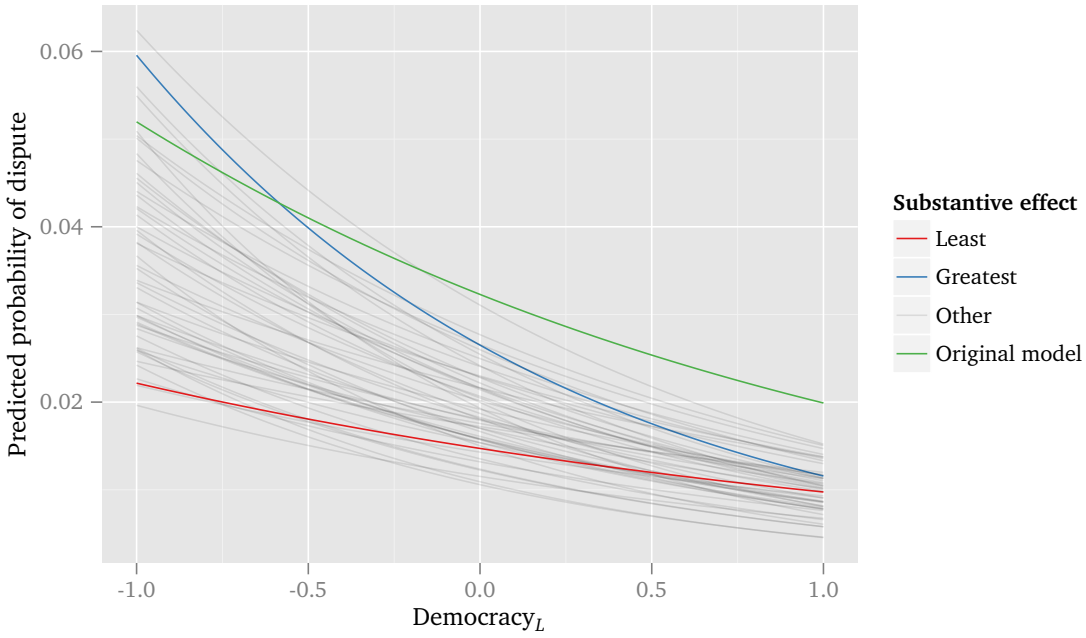

Figure 3. The estimated relationship between Democracy$_L$ and the chance of a militarized international dispute in a "median" observation according to 50 randomly sampled sets of coefficients from the MIM ($d = 3$) estimates. The estimates giving the least and greatest difference in the probability of conflict are highlighted.

uncertainty due to missingness can affect the estimated substantive effect of joint democracy by more than 2 percentage points.

5.2. **"Draws" in counterinsurgency outcomes.** The next study that I examine is Lyall's (2010) analysis of whether democracies are systematically less likely to defeat insurgencies than other types of regimes are. Lyall collects new data on counterinsurgency to critically assess previous claims that "democracies are uniquely deficient when fighting counterinsurgency (COIN) campaigns" (p. 167). Across a variety of model specifications, Lyall finds that the effect of democracy on the chance of defeating an insurgency is negative but not statistically discernible from zero.

One of the issues in Lyall's analysis is that the outcome of a counterinsurgency effort is sometimes ambiguous. He finds that at least 39 of the counterinsurgencies in his sample ($N = 286$) ended in a "draw," where neither the government nor the rebel group fully achieves its aims. Lyall employs multiple specifications to deal with the draws, including using ordered logistic regression models and dropping the observations that end in draws from the dataset. I consider an alternative approach—treating all outcomes as binary, but supposing that the "draws" are potentially measured with error.[9] I

[9]Specifically, I allow for measurement error in the 39 "restrictive draws," while coding the other 16 as wins, as in Lyall's Model 2.

then use the partial-identification estimators to find bounds on the set of results that could be obtained, under any assumption about what the true outcome was in each case coded as a draw. Specifically, given the small sample size, I use both the fill-in method and MDM.[10] I use the same set of covariates as in Lyall's Models 1–3, which I briefly review here; see Lyall (2010, pp. 176–179) for full details:

Democracy: An indicator of whether the state's Policy IV score exceeds 7.

Mechanization: An index ranging from 0 to 4 that measures the mechanization of the state's army.

External support: A sum of indicators of whether the rebels received outside aid and whether they had a sanctuary in a nearby state.

Occupation: An indicator of whether the counterinsurgent state is an external occupying power.

Elevation: The altitude (in meters) of the area where conflict occurred.

Capital distance: The distance (in logged kilometers) between the state capital and the site of conflict.

State power: The state's Composite Index of National Capabilities score.

New state: An indicator of whether the insurgency began within two years of the state entering the international system.

Cold War: An indicator of whether the insurgency took place in 1946–1989.

The results of the analysis are plotted in Figure 4. The main result accords with Lyall's own findings: we cannot rule out the possibility that democracy has no effect, or even a positive effect, on counterinsurgency success. Although the conclusion is the same as in the original paper, its basis is quite distinct. In the original analysis, the sample estimate of the coefficient on democracy is negative in every specification. Therefore, to ground the claim that there is no systematic relationship between democracy and counterinsurgency success, Lyall notes that the sample estimate is never statistically discernible from zero at even the 0.1 level. Since the dataset constitutes the population of counterinsurgencies from 1800 to 2005 (except those ongoing as of when the study was conducted), this use of significance tests is problematic; there is no inference to be made to a larger population. The results from the bounding approach, by contrast, show that a null or positive effect cannot be ruled out even *within the sample*. Due to the uncertainty that arises from not being able to measure the outcome in the "draw" cases,

[10] I generated 10,000 draws for the fill-in method and sampled 100,000 candidate points for MDM. No zeroes of the criterion function (3.8) were found, I used the "median" rule (see Section 3.3) to select a critical value, resulting in an estimated identified set containing 7,153 coefficient vectors.

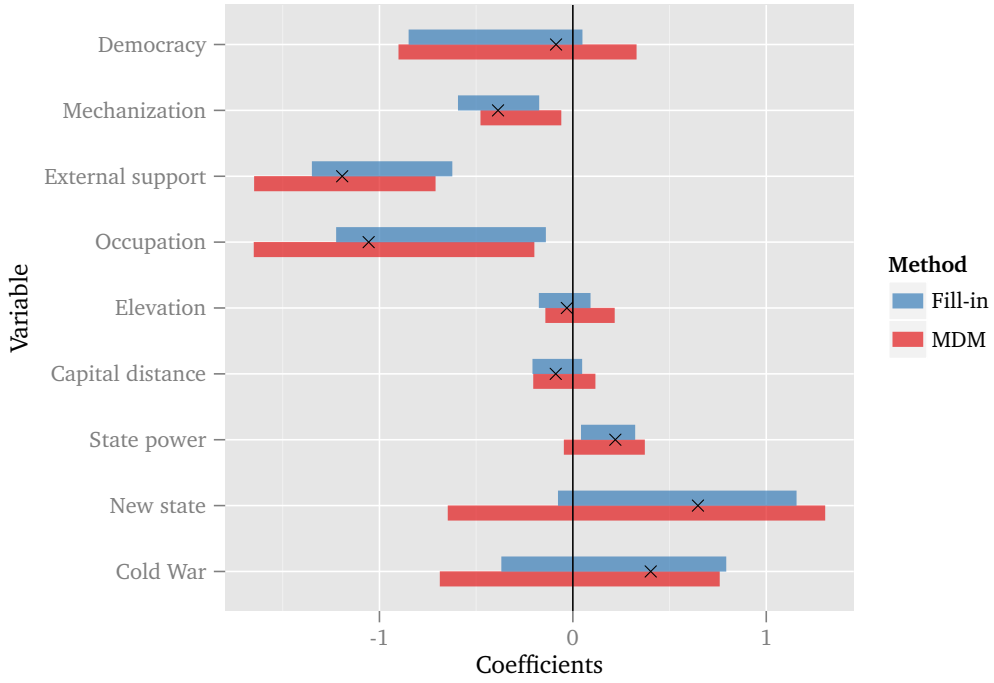

Figure 4. Estimated bounds in the reanalysis of Lyall's (2010) data. Values marked with an "x" are the estimated coefficient from the original analysis, with the 39 cases with a "draw" outcome excluded from estimation.

we cannot pin down whether democracy is positively or negatively associated with counterinsurgency success in the given data.

It is a striking coincidence that the other results also line up with the conclusions of the original paper, even though the bounding methods only consider in-sample uncertainty. Specifically, the three covariates for which the bounds on the corresponding coefficient do not include zero—mechanization, external support, and the occupation indicator—are the same whose coefficients are statistically discernible from zero in Lyall's Models 1–3. In other words, regardless of which of the 39 uncertain outcomes were truly victories or defeats, we can conclude that these three factors are negatively associated with counterinsurgency success.

6. Conclusion

I have argued for a general solution to nonignorable missingness in binary data. By taking a partial identification approach and weakening the requirement of obtaining a single point estimate, we can obtain results without having to assume MAR or develop a data-specific technique. I develop three

techniques to estimate the bounds on the identified set, the set of possible coefficient values that generate observationally equivalent distributions over outcomes. Unlike multiple imputation or selection models, these techniques do not depend on untestable assumptions about the process that generates missingness. Through simulations, I have shown that the minimum distance method appears to be most accurate at recovering the bounds of the identified set.

The most obvious area for future work is to implement inferential statistics, such as confidence regions, within the framework of the partially identified logistic regression model. The interpretation and construction of such regions is a major concern of the econometric literature on partial identification (Imbens and Manski 2004; Chernozhukov, Hong and Tamer 2007; Romano and Shaikh 2008, 2010). Another area for additional work is to extend the methods described here to multinomial or ordered discrete choice models. Such techniques could potentially deal both with observations where the outcome is completely missing, and those where the analyst knows only that it falls into some subset of the potential categories. Similarly, it would be useful to integrate partial identification methods for missingness in the covariates (e.g., Manski and Tamer 2002; Magnac and Maurin 2008) with the ones developed in this paper. This would allow for robust analysis of the many datasets in political science that have missingness in both the regressors and the outcome.

Appendix A. Appendix

Here I derive g_0 and g_1. Assume $\epsilon_i \perp\!\!\!\perp \nu_i$, so that $\Pr(M_i = 0 \mid Y_i, X_i) = F_\nu(-\varphi(X_i, Y_i; \theta))$. We then have

$$\begin{aligned}
g_0(X_i) &= \mathbb{E}[Y_{0i} \mid X_i] \\
&= \mathbb{E}[0 \cdot M_i + Y_i \cdot (1 - M_i) \mid X_i] \\
&= \mathbb{E}[Y_i(1 - M_i) \mid X_i] \\
&= \Pr(Y_i = 1, M_i = 0 \mid X_i) \\
&= \Pr(M_i = 0 \mid Y_i = 1, X_i) \Pr(Y_i = 1 \mid X_i) \\
&= F_\nu(-\varphi(X_i, 1; \theta))\Lambda(X_i'\beta)
\end{aligned}$$

and

$$\begin{aligned}
g_1(X_i) &= \mathbb{E}[Y_{1i} \mid X_i] \\
&= \mathbb{E}[1 \cdot M_i + Y_i \cdot (1 - M_i) \mid X_i] \\
&= \mathbb{E}[M_i \mid X_i] + g_0(X_i) \\
&= \mathbb{E}[M_i \mid Y_i = 1, X_i] \Pr(Y_i = 1 \mid X_i) + \mathbb{E}[M_i \mid Y_i = 0, X_i] \Pr(Y_i = 0 \mid X_i) + g_0(X_i) \\
&= (1 - F_\nu(-\varphi(X_i, 1; \theta)))\Lambda(X_i'\beta) + (1 - F_\nu(-\varphi(X_i, 0; \theta)))(1 - \Lambda(X_i'\beta)) + g_0(X_i) \\
&= 1 - F_\nu(-\varphi(X_i, 0; \theta))(1 - \Lambda(X_i'\beta)).
\end{aligned}$$

REFERENCES

Aronow, Peter M., Alan S. Gerber, Donald P. Green and Holger Kern. 2013. "Double Sampling for Missing Outcome Data in Randomized Experiments." Typescript, Yale University.
URL: *http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2305788*

Beck, Nathaniel, Jonathan N. Katz and Richard Tucker. 1998. "Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variable." *American Journal of Political Science* 42(4):1260–1288.

Bennett, D. Scott and Allan Stam. 2004. *The Behavioral Origins of War*. Ann Arbor: University of Michigan Press.

Blackwell, Matthew, James Honaker and Gary King. 2011. "Multiple Overimputation : A Unified Approach to Measurement Error and Missing Data.".

Brandt, Patrick T. and Christina J. Schneider. 2004. "So the Reviewer Told You to Use a Selection Model? Selection Models and the Study of International Relations.".

Cerquera, Daniel, François Laisney and Hannes Ullrich. 2012. "Considerations on Partially Identified Regression Models.".

Chernozhukov, Victor, Han Hong and Elie Tamer. 2007. "Estimation and Confidence Regions for Parameter Sets in Econometric Models." *Econometrica* 75(5):1243–1284.

Dominguez, Manuel A. and Ignacio N. Lobato. 2004. "Consistent Estimation of Models Defined by Conditional Moment Restrictions." *Econometrica* 72(5):1601–1615.

Dubin, Jeffrey A. and Douglas Rivers. 1989. "Selection Bias in Linear Regression, Logit and Probit Models." *Sociological Methods & Research* 18:360–390.

Frölich, Markus. 2006. "Non-parametric regression for binary dependent variables." *The Econometrics Journal* 9(3):511–540.

Glynn, Adam N. 2009. "Does Oil Cause Civil War Because It Causes State Weakness?".
URL: *http://scholar.harvard.edu/aglynn/files/OilWeakStatesCivilWar.pdf*

Hanmer, Michael J. 2007. "An Alternative Approach to Estimating Who is Most Likely to Respond to Changes in Registration Laws." *Political Behavior* 29:1–30.

Hayfield, Tristen and Jeffrey S. Racine. 2008. "Nonparametric Econometrics: The **np** Package." *Journal of Statistical Software* 27(5).

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47(1):153–161.

Honaker, James and Gary King. 2010. "What to Do about Missing Values in Time-Series Cross-Section Data." *American Journal of Political Science* 54(2):561–581.

Horowitz, Joel L. and Charles F. Manski. 1998. "Censoring of Outcomes and Regressors Due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations." *Journal of Econometrics* 84:37–58.

Horowitz, Joel L. and Charles F. Manski. 2000. "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data." *Journal of the American Statistical Association* 95(449):77–84.

Imai, Kosuke and Teppei Yamamoto. 2010. "Causal Inference with Differential Measurement Error: Nonparametric Identification and Sensitivity Analysis." *American Journal of Political Science* 54(2):543–560.

Imbens, Guido W. and Charles F. Manski. 2004. "Confidence Intervals for Partially Identified Parameters." *Econometrica* 72(6):1845–1857.

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.

Klein, Roger W. and Richard H. Spady. 1993. "An Efficient Semiparametric Estimator for Binary Response Models." *Econometrica* 61(2):387–421.

Lyall, Jason. 2010. "Do Democracies Make Inferior Counterinsurgents? Reassessing Democracy's Impact on War Outcomes and Duration." *International Organization* 64(1):167–192.

Magnac, Thierry and Eric Maurin. 2008. "Partial Identification in Monotone Binary Models: Discrete Regressors and Interval Data." *Review of Economic Studies* 75:835–864.

Manski, Charles F. 1990. "Nonparametric Bounds on Treatment Effects." *The American Economic Review* 80(2):319–323.

Manski, Charles F. 2003. *Partial Identification of Probability Distributions*. New York: Springer.

Manski, Charles F. and Elie Tamer. 2002. "Inference on Regressions with Interval Data on a Regressor or Outcome." *Econometrica* 70(2):519–546.

Molinari, Francesca. 2010. "Missing Treatments." *Journal of Business and Economic Statistics* 28(1):82–95.

Oneal, John R. and Bruce M. Russett. 1997. "The Classical Liberals Were Right: Democracy, Interdependence, and Conflict, 1950-1985." *International Studies Quarterly* 41(2):267–293.

Pagan, Adrian and Aman Ullah. 1999. *Nonparametric Econometrics*. Cambridge University Press.

Poast, Paul. 2010. "Does Issue Linkage Work? Evidence from European Alliance Negotiations, 1815 to 1945.".

Quinn, Kevin M. 2009. "What Can Be Learned from a Simple Table? Bayesian Inference and Sensitivity Analyses for Causal Effects from 2 x 2 and 2 x 2 x K Tables in the Presence of Unmeasured Confounding.".

Racine, Jeff. 1993. "An Efficient Cross-Validation Algorithm for Window Width Selection for Nonparametric Regression." *Communications in Statistics: Simulation and Computation* 22(4):1107–1114.

Romano, J P and A M Shaikh. 2008. "Inference for Identifiable Parameters in Partially Identified Econometric Models." *Journal of Statistical Planning and Inference* 138:2786–2807.

Romano, J P and A M Shaikh. 2010. "Inference for the Identified Set in Partially Identified Econometric Models." *Econometrica* 78(1):169–211.

Rubin, Donald B. 1976. "Inference and Missing Data." *Biometrika* 63(3):581–592.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Rubin, Donald B. 1996. "Multiple Imputation After 18+ Years." *Journal of the American Statistical Association* 91(434):473–489.

Russett, Bruce and John Oneal. 2001. *Triangulating Peace: Democracy, Interdependence, and International Organizations*. Norton.

Schafer, Joseph L. 1999. "Multiple Imputation: A Primer." *Statistical Methods in Medical Research* 8:3–15.