Logistic Regression in Rare Events Data

Gary King
Center for Basic Research in the Social Sciences, 34 Kirkland Street, Harvard University, Cambridge, MA 02138
e-mail: King@Harvard.Edu
http://GKing.Harvard.Edu

Langche Zeng
Department of Political Science, George Washington University, Funger Hall, 2201 G Street NW, Washington, DC 20052
e-mail: lzeng@gwu.edu

We study rare events data, binary dependent variables with dozens to thousands of times fewer ones (events, such as wars, vetoes, cases of political activism, or epidemiological infections) than zeros ("nonevents"). In many literatures, these variables have proven difficult to explain and predict, a problem that seems to have at least two sources. First, popular statistical procedures, such as logistic regression, can sharply underestimate the probability of rare events. We recommend corrections that outperform existing methods and change the estimates of absolute and relative risks by as much as some estimated effects reported in the literature. Second, commonly used data collection strategies are grossly inefficient for rare events data. The fear of collecting data with too few events has led to data collections with huge numbers of observations but relatively few, and poorly measured, explanatory variables, such as in international conflict data with more than a quarter-million dyads, only a few of which are at war. As it turns out, more efficient sampling designs exist for making valid inferences, such as sampling all available events (e.g., wars) and a tiny fraction of nonevents (peace). This enables scholars to save as much as 99% of their (nonfixed) data collection costs or to collect much more meaningful explanatory variables. We provide methods that link these two results, enabling both types of corrections to work simultaneously, and software that implements the methods developed.

Authors' note: We thank James Fowler, Ethan Katz, and Mike Tomz for research assistance; Jim Alt, John Freeman, Kristian Gleditsch, Guido Imbens, Chuck Manski, Peter McCullagh, Walter Mebane, Jonathan Nagler, Bruce Russett, Ken Scheve, Phil Schrodt, Martin Tanner, and Richard Tucker for helpful suggestions; Scott Bennett, Kristian Gleditsch, Paul Huth, and Richard Tucker for data; and the National Science Foundation (SBR-9729884 and SBR-9753126), the Centers for Disease Control and Prevention (Division of Diabetes Translation), the National Institutes of Aging (P01 AG17625-01), the World Health Organization, and the Center for Basic Research in the Social Sciences for research support. Software we wrote to implement the methods in this paper, called "ReLogit: Rare Events Logistic Regression," is available for Stata and for Gauss from http://GKing.Harvard.Edu. We have written a companion piece to this article that overlaps this one: it excludes the mathematical proofs and other technical material, and has less general notation, but it includes empirical examples and more pedagogically oriented material (see King and Zeng 2000b; copy available at http://GKing.Harvard.Edu).

1 Introduction

WE ADDRESS PROBLEMS in the statistical analysis of rare events data—binary dependent variables with dozens to thousands of times fewer ones (events, such as wars, coups, presidential vetoes, decisions of citizens to run for political office, or infections by uncommon diseases) than zeros ("nonevents"). (Of course, by trivial recoding, this definition covers either rare or very common events.) These variables are common in political science and related social sciences and perhaps most prevalent in international conflict (and other areas of public health research). In most of these literatures, rare events have proven difficult to explain and predict, a problem we believe has a multiplicity of sources, including the two we address here: most popular statistical procedures, such as logistic regression, can sharply underestimate the probability of rare events, and commonly used data collection strategies are grossly inefficient.

First, although the statistical properties of linear regression models are invariant to the (unconditional) mean of the dependent variable, the same is not true for binary dependent variable models. The mean of a binary variable is the relative frequency of events in the data, which, in addition to the number of observations, constitutes the information content of the data set. We show that this often overlooked property of binary variable models has important consequences for rare event data analyses. For example, that logit coefficients are biased in small samples (under about 200) is well documented in the statistical literature, but not as widely understood is that in rare events data the biases in probabilities can be substantively meaningful with sample sizes in the thousands and are in a predictable direction: estimated event probabilities are too small. A separate, and also overlooked, problem is that the almost-universally used method of computing probabilities of events in logit analysis is suboptimal in finite samples of rare events data, leading to errors in the same direction as biases in the coefficients. Applied researchers virtually never correct for the underestimation of event probabilities. These problems will be innocuous in some applications, but we offer simple Monte Carlo examples where the biases are as large as some estimated effects reported in the literature. We demonstrate how to correct for these problems and provide software to make the computation straightforward.

A second source of the difficulties in analyzing rare events lies in data collection. Given fixed resources, a trade-off always exists between gathering more observations and including better or additional variables. In rare events data, fear of collecting data sets with no events (and thus without variation on Y) has led researchers to choose very large numbers of observations with few, and in most cases poorly measured, explanatory variables. This is a reasonable choice, given the perceived constraints, but it turns out that far more efficient data collection strategies exist. For one example, researchers can collect all (or all available) ones and a small random sample of zeros and not lose consistency or even much efficiency relative to the full sample. This result drastically changes the optimal trade-off between more observations and better variables, enabling scholars to focus data collection efforts where they matter most.

As an example, we use all dyads (pairs of countries) for each year since World War II to generate a data set below with 303,814 observations, of which only 0.34%, or 1042 dyads, were at war. Data sets of this size are not uncommon in international relations, but they make data management difficult, statistical analyses time-consuming, and data collection expensive.[1] (Even the more common 5000–10000 observation data sets are inconvenient to deal with if one has to collect variables for all the cases.) Moreover, most dyads involve

[1] Bennett and Stam (1998b) analyze a data set with 684,000 dyad-years and (1998a) have even developed sophisticated software for managing the larger, 1.2 million-dyad data set they distribute.

countries with little relationship at all (say Burkina Faso and St. Lucia), much less with some realistic probability of going to war, and so there is a well-founded perception that many of the data are "nearly irrelevant" (Maoz and Russett 1993, p. 627). Indeed, many of the data have very little information content, which is why we can avoid collecting the vast majority of observations without much efficiency loss. In contrast, most existing approaches in political science designed to cope with this problem, such as selecting dyads that are "politically relevant" (Maoz and Russett 1993), are reasonable and practical approaches to a difficult problem, but they necessarily change the question asked, alter the population to which we are inferring, or require conditional analysis (such as only contiguous dyads or only those involving a major power). Less careful uses of these types of data selection strategies by others, such as trying to make inferences to the set of all dyads, are biased. With appropriate easy-to-apply corrections, nearly 300,000 observations with zeros need not be collected or could even be deleted with only a minor impact on substantive conclusions.

With these procedures, scholars who wish to add new variables to an existing collection can save approximately 99% of the nonfixed costs in their data collection budget or can reallocate data collection efforts to generate a larger number of more informative and meaningful variables than would otherwise be possible.[2] Relative to some other fields in political science, international relations scholars have given extraordinary attention to issues of measurement over many years and have generated a large quantity of data. Selecting on the dependent variable in the way we suggest has the potential to build on these efforts, increasing the efficiency of subsequent data collections by changing the optimal trade-off in favor of fewer observations and more sophisticated measures, closer to the desired concepts.

This procedure of selection on Y also addresses a long-standing controversy in the international conflict literature whereby qualitative scholars devote their efforts where the action is (the conflicts) but wind up getting criticized for selecting on the dependent variable. In contrast, quantitative scholars are criticized for spending time analyzing very crude measures on many observations almost all of which contain no relevant information (Bueno de Mesquita 1981; Geller and Singer 1998; Levy 1989; Rosenau 1976; Vasquez 1993). It turns out that both sides have some of the right intuition: the real information in the data lies much more with the ones than the zeros, but researchers must be careful to avoid selection bias. Fortunately, the corrections are easy, and so the goals of both camps can be met.

The main intended contribution of this paper is to integrate these two types of corrections, which have been studied mostly in isolation, and to clarify the largely unnoticed consequences of rare events data in this context. We also try to forge a critical link between the two supporting statistical literatures by developing corrections for finite sample and rare events bias, and standard error inconsistency, in a popular method of correcting selection on Y. This is useful when selecting on Y leads to smaller samples. We also provide an improved method of computing probability estimates, proofs of the equivalence of some leading econometric methods, and software to implement the methods developed. We offer evidence in the form of analytical results and Monte Carlo experiments. Empirical examples appear in our companion paper (King and Zeng 2000b).[3]

[2] The fixed costs involved in gearing up to collect data would be borne with either data collection strategy, and so selecting on the dependent variable as we suggest saves something less in research dollars than the fraction of observations not collected.

[3] We have found no discussion in political science of the effects of finite samples and rare events on logistic regression or of most of the methods we discuss that allow selection on Y. There is a brief discussion of one method of correcting selection on Y in asymptotic samples by Bueno de Mesquita and Lalman (1992, Appendix) and in an unpublished paper they cite that has recently become available (Achen 1999).

2 Logistic Regression: Model and Notation

In logistic regression, a single outcome variable Y_i ($i = 1, \ldots, n$) follows a Bernoulli probability function that takes on the value 1 with probability π_i and 0 with probability $1 - \pi_i$. Then π_i varies over the observations as an inverse logistic function of a vector $\mathbf{x}_i$, which includes a constant and $k - 1$ explanatory variables:

$$\begin{aligned} Y_i &\sim \text{ Bernoulli}(Y_i \mid \pi_i) \\ \pi_i &= \frac{1}{1 + e^{-\mathbf{x}_i \boldsymbol{\beta}}} \end{aligned} \tag{1}$$

The Bernoulli has probability function $P(Y_i \mid \pi_i) = \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i}$. The unknown parameter $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1')'$ is a $k \times 1$ vector, where β_0 is a scalar constant term and $\boldsymbol{\beta}_1$ is a vector with elements corresponding to the explanatory variables.

An alternative way to define the same model is by imagining an unobserved continuous variable Y_i^* (e.g., health of an individual or propensity of a country to go to war) distributed according to a logistic density with mean μ_i. Then μ_i varies over the observations as a linear function of $\mathbf{x}_i$. The model would be very close to a linear regression if Y_i^* were observed:

$$\begin{aligned} Y_i^* &\sim \text{ Logistic}(Y_i^* \mid \mu_i) \\ \mu_i &= \mathbf{x}_i \boldsymbol{\beta} \end{aligned} \tag{2}$$

where $\text{Logistic}(Y_i^* \mid \mu_i)$ is the one-parameter logistic probability density,

$$P(Y_i^*) = \frac{e^{-(Y_i^* - \mu_i)}}{\left(1 + e^{-(Y_i^* - \mu_i)}\right)^2} \tag{3}$$

Unfortunately, instead of observing Y_i^*, we see only its dichotomous realization, Y_i, where $Y_i = 1$ if $Y_i^* > 0$ and $Y_i = 0$ if $Y_i^* \leq 0$. For example, if Y_i^* measures health, Y_i might be dead (1) or alive (0). If Y_i^* were the propensity to go to war, Y_i could be at war (1) or at peace (0). The model remains the same because

$$\begin{aligned} \Pr(Y_i = 1 \mid \boldsymbol{\beta}) = \pi_i &= \Pr(Y_i^* > 0 \mid \boldsymbol{\beta}) \\ &= \int_0^\infty \text{Logistic}(Y_i^* \mid \mu_i) dY_i^* = \frac{1}{1 + e^{-\mathbf{x}_i \boldsymbol{\beta}}} \end{aligned} \tag{4}$$

which is exactly as in Eq. (1). We also know that the observation mechanism, which turns the continuous Y^* into the dichotomous Y_i, generates most of the mischief. That is, we ran simulations trying to estimate $\boldsymbol{\beta}$ from an observed Y^* and model 2 and found that maximum-likelihood estimation of $\boldsymbol{\beta}$ is approximately unbiased in small samples.

The parameters are estimated by maximum likelihood, with the likelihood function formed by assuming independence over the observations: $L(\boldsymbol{\beta} \mid \mathbf{y}) = \prod_{i=1}^n \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i}$. By taking logs and using Eq. (1), the log-likelihood simplifies to

$$\begin{aligned} \ln L(\boldsymbol{\beta} \mid \mathbf{y}) &= \sum_{\{Y_i = 1\}} \ln(\pi_i) + \sum_{\{Y_i = 0\}} \ln(1 - \pi_i) \\ &= -\sum_{i=1}^n \ln\left(1 + e^{(1-2Y_i)\mathbf{x}_i \boldsymbol{\beta}}\right) \end{aligned} \tag{5}$$

(e.g., Greene 1993, p. 643). Maximum-likelihood logit analysis then works by finding the value of $\boldsymbol{\beta}$ that gives the maximum value of this function, which we label $\hat{\boldsymbol{\beta}}$. The asymptotic

variance matrix, $V(\hat{\beta})$, is also retained to compute standard errors. When observations are selected randomly, or randomly within strata defined by some or all of the explanatory variables, $\hat{\beta}$ is consistent and asymptotically efficient (except in degenerate cases of perfect collinearity among the columns in $\mathbf{X}$ or perfect discrimination between zeros and ones).

That in rare events data ones are more statistically informative than zeros can be seen by studying the variance matrix,

$$V(\hat{\beta}) = \left[\sum_{i=1}^{n} \pi_i (1 - \pi_i) \mathbf{x}_i' \mathbf{x}_i \right]^{-1} \tag{6}$$

The part of this matrix affected by rare events is the factor $\pi_i(1 - \pi_i)$. Most rare events applications yield small estimates of $\Pr(Y_i = 1 \mid \mathbf{x}_i) = \pi_i$ for all observations. However, if the logit model has some explanatory power, the estimate of π_i among observations for which rare events are observed (i.e., for which $Y_i = 1$) will usually be larger [and closer to 0.5, because probabilities in rare event studies are normally very small (see Beck et al. 2000)] than among observations for which $Y_i = 0$. The result is that $\pi_i(1 - \pi_i)$ will usually be larger for ones than zeros, and so the variance (its inverse) will be smaller. In this situation, additional ones will cause the variance to drop more and hence are more informative than additional zeros (see Imbens 1992, pp. 1207, 1209; Cosslett 1981a; Lancaster and Imbens 1996b).

Finally, we note that the quantity of interest in logistic regression is rarely the raw $\hat{\beta}$ output by most computer programs. Instead, scholars are normally interested in more direct functions of the probabilities. For example, *absolute risk* is the probability that an event occurs given chosen values of the explanatory variables, $\Pr(Y = 1 \mid X = x)$. The *relative risk* is the same probability relative to the probability of an event given some baseline values of X, e.g., $\Pr(Y = 1 \mid X = 1) / \Pr(Y = 1 \mid X = 0)$, the fractional increase in the risk. This quantity is frequently reported in the popular media (e.g., the probability of getting some forms of cancer increase by 50% if one stops exercising) and is common in many scholarly literatures. In political science, the term is not often used, but the measure is usually computed directly or studied implicitly. Also of considerable interest is the *first difference* (or attributable risk), the change in probability as a function of a change in a covariate, such as $\Pr(Y = 1 \mid X = 1) - \Pr(Y = 1 \mid X = 0)$. The first difference is usually most informative when measuring effects, whereas relative risk is dimensionless and so tends to be easier to compare across applications or time periods. Although scholars often argue about their relative merits (see Breslow and Day 1980, Chap. 2; and Manski 1999), reporting the two probabilities that make up each relative risk and each first difference is best when convenient.

3 How to Select on the Dependent Variable

We first distinguish among alternative data collection strategies and show how to adapt the logit model for each. Then, in Section 5, we build on these models to also allow rare event and finite sample corrections. This section discusses research design issues, and Section 4 considers the specific statistical corrections necessary.

3.1 *Data Collection Strategies*

The usual strategy, as known in econometrics, is either *random* sampling, where all observations (X, Y) are selected at random, or *exogenous stratified* sampling, which allows Y to be randomly selected within categories defined by X. Optimal statistical models are identical under these two sampling schemes. Indeed, in epidemiology, both are known under one name, *cohort* (or *cross-sectional*, to distinguish it from a panel) study.

When one of the values of Y is rare in the population, considerable resources in data collection can be saved by randomly selecting within categories of Y. This is known in econometrics as *choice-based* or *endogenous stratified* sampling and in epidemiology as a *case-control* design (Breslow 1996); it is also useful for choosing qualitative case studies (King et al. 1994, Sect. 4.4.2). The strategy is to select on Y by collecting observations (randomly or all those available) for which $Y = 1$ (the "cases") and a random selection of observations for which $Y = 0$ (the "controls"). This sampling method is often supplemented with known or estimated prior knowledge of the population fractions of ones—information that is often available (e.g., a list of all wars is often readily available even when explanatory variables measured at the dyadic level are not). Finally, *case-cohort* studies begin with some variables collected on a large cohort, and then subsample using all the ones and a random selection of zeros. The case-cohort study is especially appropriate when adding an expensive variable to an existing collection, such as the dyadic data discussed above and analyzed below, or Verba and co-workers' (1995) detailed study of activists, each of which was culled from a larger random sample, with very few variables, of the entire U.S. population. In this paper, we use information on the population fraction of ones when it is available, and so the same models we describe apply to both case-control and case-cohort studies.

Many other hybrid data collection strategies have also been tried. For example, Bueno de Mesquita and Lalman's (1992) design is fairly close to a case-control study with "contaminated controls," meaning that the "control" sample was from the whole population rather than only those observations for which $Y = 0$ (see Lancaster and Imbens 1996a). Although we do not analyze hybrid designs in this paper, our view is *not* that pure case-control sampling is appropriate for all political science studies of rare events. (For example, additional efficiencies might be gained by modifying a data collection strategy to fit variables that are easier to collect within regional or language clusters.) Rather, our argument is that scholars should consider a much wider range of potential sampling strategies, and associated statistical methods, than is now common. This paper focuses only on the leading alternative design which we believe has the potential to see widespread use in political science.

3.2 *Problems to Avoid*

Selecting on the dependent variable in the way we suggest has several pitfalls that should be carefully avoided. First, the sampling design for which the prior correction and weighting methods are appropriate requires independent random (or complete) selection of observations for which $Y = 1$ and $Y = 0$. This encompasses the case-control and case-cohort studies, but other endogenous designs—such as sampling in several stages, with nonrandom selection, or via hybrid approaches—require different statistical methods.

Second, when selecting on Y, we must be careful not to select on X differently for the two samples. The classic example is selecting all people in the local hospital with liver cancer ($Y = 1$) and a random selection of the U.S. population without liver cancer ($Y = 0$). The problem is that the sample of cancer patients selects on $Y = 1$ *and* implicitly on the inclination to seek health care, find the right medical specialist, have the right tests, etc. Not recognizing the implicit selection on X is the problem here. Since the $Y = 0$ sample does not similarly select on the same explanatory variables, these data would induce selection bias. One solution in this example might be to select the $Y = 0$ sample from those who received the same liver cancer test but turned out not to have the disease. This design would yield valid inferences, albeit only for the health-conscious population with liver cancer-like symptoms. Another solution would be to measure and control for the omitted variables.

This type of inadvertent selection on X can be a serious problem in endogenous designs, just as selection on Y can bias inferences in exogenous designs. Moreover, although in

the social sciences random (or experimenter control over) assignment of the values of the explanatory variables for each unit is occasionally possible in exogenous or random sampling (and with a large n is generally desirable since it rules out omitted variable bias), random assignment on X is impossible in endogenous sampling. Fortunately, bias due to selection on X is much easier to avoid in applications such as international conflict and related fields, since a clearly designated census of cases is normally available from which to draw a sample. Instead of relying on the decisions of subjects about whether to come to a hospital and take a test, the selection into the data set in our field can often be entirely determined by the investigator. See Holland and Rubin (1988).

Third, another problem with intentional selection on Y is that valid exploratory data analysis can be more hazardous. In particular, one cannot use an explanatory variable as a dependent variable in an auxiliary analysis without special precautions (see Nagelkerke et al. 1995).

Finally, the optimal trade-off between collecting more observations versus better or more explanatory variables is application-specific, and so decisions will necessarily involve judgment calls and qualitative assessments. Fortunately, to help guide these decisions in fields like international relations we have large bodies of work on methods of quantitative measurement and, also, many qualitative studies that measure hard-to-collect variables for a small number of cases (such as leaders' perceptions).

We can also make use of some formal statistical results to suggest procedures for deciding on the optimal trade-off between more observations and better variables. First, when zeros and ones are equally easy to collect, and an unlimited number of each are available, an "equal shares sampling design" (i.e., $\bar{y} = 0.5$) is optimal in a limited number of situations and close to optimal in a large number (Cosslett 1981b; Imbens 1992). This is a useful fact, but in fields like international relations, the number of observable ones (such as wars) is strictly limited, and so in most of our applications collecting all available or a large sample of ones is best. The only real decision, then, is how many zeros to collect in addition. If collecting zeros were costless, we should collect as many as we can get, since more data are always better. If collecting zeros is not costless, but not (much) more expensive than collecting ones, then one should collect more zeros than ones. However, since the marginal contribution to the explanatory variables' information content for each additional zero starts to drop as the number of zeros passes the number of ones, we will not often want to collect more than (roughly) two to five times more zeros than ones. In general, the optimal number of zeros depends on how much more valuable the explanatory variables become with the resources saved by collecting fewer observations. Finally, a useful practice is sequential, involving first the collection of all ones and (say) an equal number of zeros. Then, if the standard errors and confidence intervals are narrow enough, stop. Otherwise, continue to sample zeros randomly and stop when the confidence intervals get sufficiently small for the substantive purposes at hand. For some data collections, it might even be efficient to collect explanatory variables sequentially as well, but this is not often the case.

4 Correcting Estimates for Selection on Y

Designs that select on Y can be consistent and efficient but only with the appropriate statistical corrections. Sections 4.1 and 4.2 introduce the *prior correction* and *weighting* methods of estimation under choice-based sampling. For the past 20 years, econometricians have made steady progress generalizing and improving these methods. However, Hsieh et al. (1985) have shown that two of these econometric methods are equivalent to prior correction for the logit model. In Appendix A, we explicate this result and then prove that the best econometric estimator in this tradition also reduces to the method of prior correction when

the model is logit and the sampling probability, $E(\bar{y})$, is unknown. To our knowledge, this result has not appeared previously in the literature.

4.1 *Prior Correction*

Prior correction involves computing the usual logistic regression MLE and correcting the estimates based on prior information about the fraction of ones in the population, τ, and the observed fraction of ones in the sample (or sampling probability), $\bar{y}$. Knowledge of τ can come from census data, a random sample from the population measuring Y only, a case-cohort sample, or other sources. In Appendix B, we try to elucidate this method by presenting a derivation of the method of prior correction for logit and most other statistical models (although prior correction is easiest to apply to the logit model). For the logit model, in any of the above sampling designs, the MLE $\hat{\boldsymbol{\beta}}_1$ is a statistically consistent estimate of $\boldsymbol{\beta}_1$ and the following corrected estimate is consistent for β_0:

$$\hat{\beta}_0 - \ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right] \tag{7}$$

which equals $\hat{\beta}_0$ only in randomly selected cross-sectional data. Of course, scholars are not normally interested in $\boldsymbol{\beta}$ but rather in the probability that an event occurs, $\Pr(Y_i = 1 \mid \boldsymbol{\beta}) = \pi_i = (1 + e^{\mathbf{x}_i\boldsymbol{\beta}})^{-1}$, which requires good estimates of both $\boldsymbol{\beta}_1$ and β_0. Epidemiologists and biostatisticians usually attribute prior correction to Prentice and Pyke (1979); econometricians attribute the result to Manski and Lerman (1977), who in turn credit an unpublished comment by Daniel McFadden. The result was well-known previously in the special case of all discrete covariates (e.g., Bishop et al. 1975, p. 63) and has been shown to apply to other multiplicative intercept models (Hsieh et al. 1985, p. 659).

Prior correction requires knowledge of the fraction of ones in the population, τ. Fortunately, τ is straightforward to determine in international conflict data since the number of conflicts is the subject of the study and the denominator, the population of countries or dyads, is easy to count even if not entirely in the analysis.[4]

A key advantage of prior correction is ease of use. Any statistical software that can estimate logit coefficients can be used, and Eq. (7) is easy to apply to the intercept. If the functional form and explanatory variables are correct, estimates are consistent and asymptotically efficient. The chief disadvantage of prior correction is that if the model is misspecified, estimates of both β_0 and $\boldsymbol{\beta}_1$ are slightly less robust than weighting (Xie and Manski 1989), a method to which we now turn.

4.2 *Weighting*

An alternative procedure is to weight the data to compensate for differences in the sample ($\bar{y}$) and population (τ) fractions of ones induced by choice-based sampling. The resulting *weighted exogenous sampling maximum-likelihood estimator* (due to Manski and Lerman 1977) is relatively simple. Instead of maximizing the log-likelihood in Eq. (5), we maximize

[4] King and Zeng (2000a), building on results of Manski (1999), modify the methods in this paper for the situation when τ is unknown or partially known. King and Zeng use "robust bayesian analysis" to specify classes of prior distributions on τ, representing full or partial ignorance. For example, the user can specify that τ is completely unknown or known to fall with some probability to lie only in a given interval. The result is classes of posterior distributions (instead of a single posterior) that, in many cases, provide informative estimates of quantities of interest.

the weighted log-likelihood:

$$\begin{aligned}\ln L_w(\boldsymbol{\beta} \mid \mathbf{y}) &= w_1 \sum_{\{Y_i=1\}} \ln(\pi_i) + w_0 \sum_{\{Y_i=0\}} \ln(1-\pi_i) \\ &= -\sum_{i=1}^{n} w_i \ln\left(1 + e^{(1-2y_i)\mathbf{x}_i\boldsymbol{\beta}}\right) \end{aligned} \tag{8}$$

where the weights are $w_1 = \tau/\bar{y}$ and $w_0 = (1-\tau)/(1-\bar{y})$, and where

$$w_i = w_1 Y_i + w_0(1 - Y_i) \tag{9}$$

One perceived disadvantage of this model has been that it seemed to require specialized software for estimation. However, the alternative expression in the second line of Eq. (8) enables researchers to use any logit package, since the weight, w_i, appears in one term. All researchers need to do is to calculate w_i in Eq. (8), choose it as the weight in their computer program, and then run a logit model (our software will do this automatically).

Weighting can outperform prior correction when both a large sample is available and the functional form is misspecified (Xie and Manski 1988). Weighting is asymptotically less efficient than prior correction, an effect that can be seen in small samples (see Scott and Wild 1986; Amemiya and Vuong 1987), but the differences are not large. Since misspecification is such a common part of social science analysis, one would think that weighting would normally be preferred. However, two more serious problems limit its application. First, the usual method of computing standard errors is severely biased. Second, rare event, finite sample corrections, which work without modification for prior correction, have not been developed for weighting. We discuss remedies for both problems below, which we feel in most cases makes weighting preferable when information about τ is available.

5 Rare Event, Finite Sample Corrections

In this section, we discuss methods of computing probability estimates that correct problems due to finite samples or rare events. We take the models in Section 4 as our starting point and discuss only estimators that are statistically consistent. Let $\mathbf{x}_0$ be a $1 \times k$ vector of chosen values of the explanatory variables. The nearly universal method used for computing the probability, given $\mathbf{x}_0$, is a function of the maximum-likelihood estimate, $\hat{\boldsymbol{\beta}}$,

$$\Pr(Y_0 = 1 \mid \hat{\boldsymbol{\beta}}) = \hat{\pi}_0 = \frac{1}{1 + e^{-\mathbf{x}_0\hat{\boldsymbol{\beta}}}} \tag{10}$$

and is thus statistically consistent.

Unfortunately, the method of computing probabilities given in Eq. (10) is affected by two distinct problems in finite samples of rare events data: First, $\hat{\boldsymbol{\beta}}$ is a biased estimate of $\boldsymbol{\beta}$. Second, even if $\hat{\boldsymbol{\beta}}$ were unbiased, $\Pr(Y_0 = 1 \mid \hat{\boldsymbol{\beta}})$ would still be, as we show below, an inferior estimator of $\Pr(Y_0 = 1 \mid \boldsymbol{\beta})$. We discuss these two problems and review or develop appropriate corrections in Sections 5.1 and 5.2 respectively. We also consider modifications for both cohort and choice-based sampling designs.[5]

[5] We analyze the problem of absolute risk directly and then compute relative risk as the ratio of two absolute risks. Although we do not pursue other options here because our estimates of relative risk clearly outperform existing methods, it seems possible that even better methods could be developed that estimate relative risk directly.

5.1 *Parameter Estimation*

We know from the statistical literature that the usual estimate of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, is biased in finite samples and that less biased and more efficient methods are available. This knowledge has apparently not made it to the applied literatures (as noted by Bull et al. 1997); at least part of the reason is that the statistical literature does not include studies of the effects that rare events have in greatly magnifying the biases. This situation has led some to downplay the effects of bias; for example, Schaefer (1983, p. 73) writes that "sample sizes above 200 would yield an insignificant bias correction."

Finite sample bias amplified by rare events is occasionally discussed informally in the pattern recognition and classification literatures (Ripley 1996) but is largely unknown in most applied literatures and, to our knowledge, has never been discussed in political science. The issue is not normally considered in the literatures on case-control studies in epidemiology or choice-based sampling in econometrics, although these literatures reveal a practical wisdom given that their data collection strategies naturally produce well-balanced samples.[6]

Our results show that, for rare events data, $\Pr(Y = 1)$ is underestimated, and hence $\Pr(Y = 0)$ is overestimated. To see this intuitively, and only heuristically, consider the simplified case with one explanatory variable illustrated in Fig. 1. First, we order the observations on Y according to their values on X (the horizontal dimension in Fig. 1). If $\beta_1 > 0$, most of the zeros will be to the left and ones will be to the right, with little overlap. Since there were so many zeros in the example, we replaced them with a dotted line fit to the density representing $X \mid Y = 0$ (such as by using a histogram of the X values in each group). The few ones in the data set appear as short vertical lines, and the distribution from which they were drawn appears as a solid line (representing the density of $X \mid Y = 1$). [As drawn, $P(X \mid Y = 0)$ and $P(X \mid Y = 1)$ are normal, but that need not be the case.] Although the large number of zeros allows us to estimate the dotted density line essentially without error, any estimate of the solid density line for $X \mid Y = 1$ from the mere five data points will be very poor and, indeed, systematically biased toward tails that are too short. To see this,

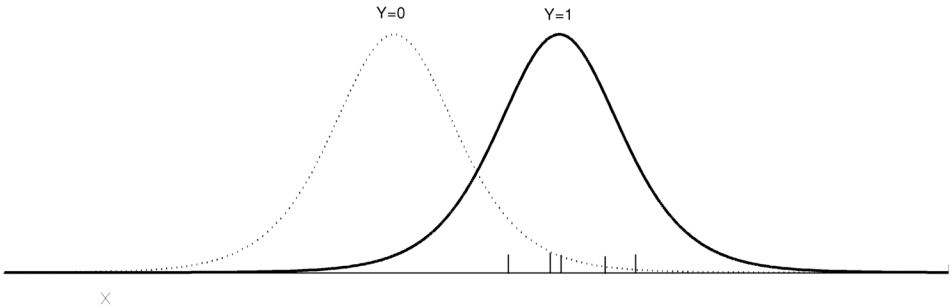

Fig. 1 How rare events bias logit coefficients. Observations are arrayed horizontally according to the value of X, where $\beta_1 > 0$. The few $Y = 1$ observations appear as short vertical lines, along with the (solid) line for the density from which they were drawn. The many $Y = 0$ observations do not appear but their density appears as a dotted line. Because the zeros density will be better estimated than the ones density, the cutting point that best classifies zeros and ones (which is roughly related to β_1) will be too far to the right since no information exists about the left end of the solid density.

[6] "Exact" tests are a good solution to the problem when all variables are discrete and sufficient (often massive) computational power is available (see Agresti 1992; Mehta and Patel 1997). These tests compute exact finite sample distributions based on permutations of the data tables.

consider finding a cutting point (value of X) that maximally distinguishes the zeros and ones, i.e., by making the fewest mistakes (zeros misplaced to the right of the cut point or ones to the left). This cutting point is related to the maximum-likelihood estimate of β and would probably be placed just to the left of the vertical line farthest or second farthest to the left. Unfortunately, with many more zeros than ones, $\max(X \mid Y = 0)$ [and more generally the area in the right tail of $P(X \mid Y = 0)$] will be well estimated, but $\min(X \mid Y = 1)$ [and the area in the left tail of $P(X \mid Y = 1)$] will be poorly estimated. Indeed, $\min(X \mid Y = 1)$ will be systematically too far to the right. (This is general: for a finite number of draws from any distribution, the minimum in the sample is always greater than or equal to the minimum in the population.) Since the cutting point is a function of these tails [which, roughly speaking, is related to $\max(X \mid Y = 0) - \min(X \mid Y = 1)$], it will be biased in the direction of favoring zeros at the expense of the ones and so $\Pr(Y = 1)$ will be too small.[7]

We begin with McCullagh and Nelder's (1989) analytical approximations, but we focus on rare events. We then extend their work some by using their procedures to derive a correction that covers not only the usual logit case, which they discussed and of course can also be used with prior correction as in Section 4.1, but also the weighted model in Section 4.2. As Appendix C demonstrates, the bias in $\hat{\boldsymbol{\beta}}$ can be estimated by the following weighted least-squares expression:

$$\text{bias}(\hat{\boldsymbol{\beta}}) = (\mathbf{X'WX})^{-1}\mathbf{X'W}\xi \tag{11}$$

where $\xi_i = 0.5Q_{ii}[(1+w_1)\hat{\pi}_i - w_1]$, Q_{ii} are the diagonal elements of $\mathbf{Q} = \mathbf{X}(\mathbf{X'WX})^{-1}\mathbf{X'}$, and $\mathbf{W} = \text{diag}\{\hat{\pi}_i(1-\hat{\pi}_i)w_i\}$. This expression is easy to estimate, as it involves running a weighted least-squares regression with $\mathbf{X}$ as the "explanatory variables," ξ as the "dependent variable," and $\mathbf{W}$ as the weight. The bias-corrected estimate is then $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \text{bias}(\hat{\boldsymbol{\beta}})$. (Thus, we use the circumflex $\hat{\boldsymbol{\beta}}$ to refer to the MLE, and the tilde $\tilde{\boldsymbol{\beta}}$ to denote the approximately unbiased estimate of $\boldsymbol{\beta}$). Appendix C also approximates the variance matrix of $\tilde{\boldsymbol{\beta}}$ as a multiple of the usual variance matrix, $V(\tilde{\boldsymbol{\beta}}) = (n/(n+k))^2 V(\hat{\boldsymbol{\beta}})$. A key point is that since $(n/(n+k))^2 < 1$, $V(\tilde{\boldsymbol{\beta}}) < V(\hat{\boldsymbol{\beta}})$, and so we are in the happy situation where reducing bias also reduces variance.

Although the bias correction is easy to use, it is not as straightforward to understand. To provide a better analytical understanding, and to show how it operates under rare events, we have derived a simple expression in a special case. The idea, based on our simulation studies, is that the bias term appears to affect the constant term directly and the other coefficients primarily as a consequence (unlike the sampling designs in Section 4, these corrections affect all the coefficients). Thus, we consider the special case with a constant term and one explanatory variable, and with β_0 estimated and $\beta_1 = 1$ fixed: $\Pr(Y_i = 1) = 1/(1 + e^{-(\beta_0 + X_i)})$. For this case, Appendix D provides a rough approximation for the bias in $\hat{\beta}_0$, where $\bar{\pi} = (1/n)\sum_{i=1}^{n} \pi_i$, as

$$E(\hat{\beta}_0 - \beta_0) \approx \frac{\bar{\pi} - 0.5}{n\bar{\pi}(1-\bar{\pi})} \tag{12}$$

[7] More formally, suppose $P(X \mid Y = j) = \text{Normal}(X \mid \mu_j, 1)$, for $j = 0, 1$. Then the logit model should classify an observation as 1 if the probability is greater than 0.5 or equivalently $X > T(\mu_0, \mu_1) = [\ln(1-\tau) - \ln(\tau)]/(\mu_1 - \mu_0) + (\mu_0 + \mu_1)/2$. A logit of Y on a constant term and X is fully saturated and hence equivalent to estimating μ_j with $\bar{X}_j$ (the mean of X_i for all i in which $Y_i = j$). However, the estimated classification boundary, $T(\bar{X}_0, \bar{X}_1)$, will be larger than $T(\mu_0, \mu_1)$ when $\tau < 0.5$ (and thus $\ln[(1-\tau)/\tau] > 0$), since, by Jensen's inequality, $E[1/(\bar{X}_0 - \bar{X}_1)] > 1/(\mu_1 - \mu_0)$. Hence, the threshold will be too far to the right in Fig. 1 and will underestimate the probability of a one in finite samples.

Since $\bar{\pi} < 0.5$ in rare events data, the numerator, and thus the entire bias term, is negative. This means that $\hat{\beta}_0$ is too small and, as a result, $\Pr(Y = 1)$ is underestimated, which is consistent with what we argued intuitively above and show via Monte Carlo experiments below. The denominator is also informative, because it shows that as n gets large the bias vanishes, which is one way of proving consistency in this special case. Finally, a key result is that the factor $\bar{\pi}(1-\bar{\pi})$ in the denominator shows that the bias is amplified in applications with rarer events (i.e., as $\bar{\pi}$ approaches zero).[8]

5.2 *Probability Calculations*

This section concerns estimating the probability π in Eq. (1). Since $\tilde{\boldsymbol{\beta}}$ is less biased and has smaller variance, and hence has a smaller mean square error, than $\hat{\boldsymbol{\beta}}$,

$$\tilde{\pi}_0 = \Pr(Y_0 = 1 \mid \tilde{\boldsymbol{\beta}}) = \frac{1}{1 + e^{\mathbf{x}_0 \tilde{\boldsymbol{\beta}}}} \tag{13}$$

is usually preferable to $\hat{\pi}$ [from Eq. (10)]. However, $\tilde{\pi}$ is still not optimal because it ignores the uncertainty in $\tilde{\boldsymbol{\beta}}$ (e.g., Geisser 1993; King et al. 2000). This uncertainty can be thought of as sampling error or the fact that $\tilde{\boldsymbol{\beta}}$ is estimated rather than known, and it is reflected in standard errors greater than zero. In many cases, ignoring estimation uncertainty leaves the point estimate unaffected and changes only its standard error. However, because of the nature of π as a quantity to be estimated, ignoring uncertainty affects the point estimate too.

Indeed, ignoring estimation uncertainty generates too small an estimated probability of a rare event (or in general an estimate too far from 0.5). This can be seen intuitively by considering the underlying continuous variable Y^* that the basic model assumes to be logistic. Under the model, the probability is the area to the right of the threshold [the dark shaded area to the right of zero under the dotted curve in Fig. 2, which illustrates Eq. (4)], an area typically less than 0.5 in rare events data. The problem is that ignoring uncertainty about $\boldsymbol{\beta}$ leads to a distribution that has too small a variance and, thus (with rare events), too little area to the right of the threshold. Adding in the uncertainty increases the variance of the distribution, and the area to the right of the threshold, and thus makes the probability larger (closer to 0.5). For example, in Fig. 2 the additional variance is illustrated in the change from the dotted to the solid density, and hence the increase in the area to the right of the zero threshold [from the dark shaded area marked $\Pr(Y_i = 1 \mid \tilde{\boldsymbol{\beta}})$ to the total shaded area, marked $\Pr(Y_i = 1)$].

Thus, instead of conditioning on an uncertain point estimate with $\tilde{\pi}$, we should be conditioning only on known facts and averaging over the uncertainty in $\tilde{\boldsymbol{\beta}}$ as follows:

$$\Pr(Y_i = 1) = \int \Pr(Y_i = 1 \mid \boldsymbol{\beta}^*) P(\boldsymbol{\beta}^*) d\boldsymbol{\beta}^* \tag{14}$$

[8] An elegant result due to Firth (1993) shows that bias can also be corrected during the maximization procedure by applying Jeffrey's invariant prior to the logistic likelihood and using the maximum posterior estimate. We have applied this work to weighting and prior correction and run experiments to compare the methods. Consistent with Firth's examples, we find that the methods give answers that are always numerically very close (almost always less than half a percent). An advantage of Firth's procedure is that it gives answers even when the MLE is undefined, as in cases of perfect discrimination; a disadvantage is computational in that the analytical gradient and Hessian are much more complicated. Another approach to bias reduction is based on jackknife methods, which replace analytical derivations with easy computations, although systematic comparisons by Bull et al. (1997) show that they do not generally work as well as the analytical approaches.

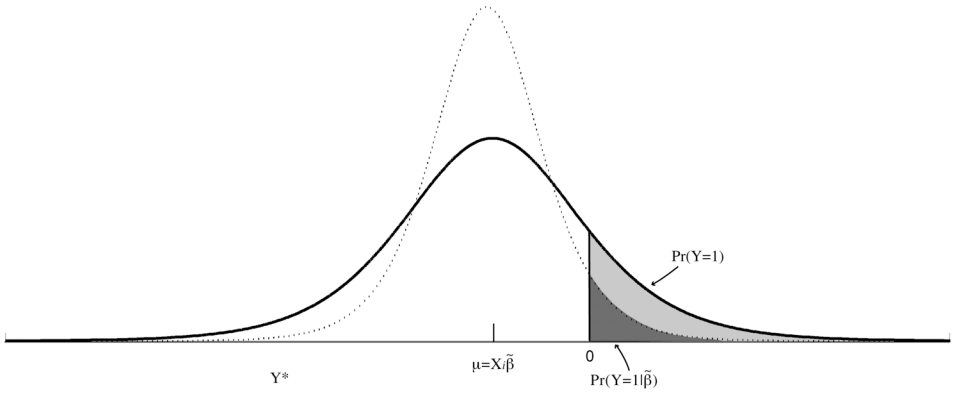

Fig. 2 The effect of uncertainty on probabilities. Although the dotted density (which does not reflect uncertainty in $\boldsymbol{\beta}$) has a smaller variance than the one drawn with a solid line (which has the uncertainty about $\boldsymbol{\beta}$ added in), the mean μ stays the same in both. However, the probability, the shaded area to the right of the zero threshold in the two curves, differs.

where $\boldsymbol{\beta}^*$ is the integration dummy, and to summarize estimation uncertainty $P(\cdot)$ we take the Bayesian viewpoint and use the posterior density of $\boldsymbol{\beta}$ Normal $[\boldsymbol{\beta} \mid \tilde{\boldsymbol{\beta}}, V(\tilde{\boldsymbol{\beta}})]$ (although it will turn out that we will not need this normality assumption). The estimation uncertainty $P(\cdot)$ can also be thought of from a frequentist perspective as the sampling distribution of $\tilde{\boldsymbol{\beta}}$ so that Eq. (14) is the expected value $E_{\tilde{\beta}}[\Pr(Y_i = 1 \mid \tilde{\boldsymbol{\beta}})]$, which is an estimate of $\pi_i = \Pr(Y_i = 1 \mid \boldsymbol{\beta}) = 1/(1 + e^{-\mathbf{x}_i\boldsymbol{\beta}})$.

Equation (14) can be computed in two ways. First, we could use simulation (see Tanner 1996; King et al. 2000): take a random draw of $\boldsymbol{\beta}$ from $P(\boldsymbol{\beta})$, insert it into $[1 + e^{-\mathbf{x}_i\boldsymbol{\beta}}]^{-1}$, repeat, and average over the simulations. Increasing the number of simulations enables us to approximate $\Pr(Y_i = 1)$ to any desired degree of accuracy.

A second method of computing Eq. (14) is through an analytical approximation we have derived. It is more computationally efficient than the simulation approach, is easy to use, and helps illuminate the nature of the correction. This result, proven in Appendix E, shows that Eq. (14) may be approximated without simulation as

$$\Pr(Y_i = 1) \approx \tilde{\pi}_i + C_i \tag{15}$$

where the correction factor is

$$C_i = (0.5 - \tilde{\pi}_i)\tilde{\pi}_i(1 - \tilde{\pi}_i)\mathbf{x}_0 V(\tilde{\boldsymbol{\beta}})\mathbf{x}_0' \tag{16}$$

Standard errors or confidence intervals can easily be computed as part of the simulation in the first approach or by simulating each component of C_i in the second.

These expressions have several intuitive features that help in understanding problems induced by ignoring uncertainty in $\tilde{\boldsymbol{\beta}}$. First, the correction factor C_i, as expected, is zero if the uncertainty in $\tilde{\boldsymbol{\beta}}$, $V(\tilde{\boldsymbol{\beta}})$, is a matrix of zeros, and it grows as the uncertainty grows. Second, in the presence of some uncertainty, the direction of the bias is determined by the first factor in C_i, $(0.5 - \tilde{\pi}_i)$. When $\tilde{\pi}_i < 0.5$, as is usually the case for rare events, the correction factor adds to the estimated probability of an event. Hence, using $\tilde{\pi}_i$ alone generally underestimates the probability of an event.

The logic of the improved estimator in Eq. (15) can be thought of as Bayesian but not completely so since $\boldsymbol{\beta}$ is estimated via $\tilde{\boldsymbol{\beta}}$ [and $V(\tilde{\boldsymbol{\beta}})$]. If prior information is available on the

logit coefficients, β, we might normally prefer a full Bayesian estimation at the first stage as well. However, in the common situation where prior information is unavailable or difficult to elicit or formalize, Bayesian estimation with uninformative priors is equivalent to traditional logit analysis. And from the usually preferred mean square error perspective, using $\tilde{\beta}$ strictly dominates $\hat{\beta}$, which has larger variance and bias. This approach is consistent with those using Bayesian methods to derive methods with good frequentist properties (see also Smith 1998).

The evidence in Section 6 indicates that our estimator in Eq. (15) has a smaller mean square error than other estimators of π_i and, by this standard, is therefore superior. However, like most Bayesian estimators, it is not unbiased. Indeed, since the contrast between the different methods of inference in this case is especially striking and thus instructive, consider what an approximately unbiased estimator would look like. First, recall that a deterministic function of an unbiased estimator is not necessarily unbiased. (For example, the sample mean $\bar{y}$ is an unbiased estimate of a population mean μ, but $1/\bar{y}$ is not an unbiased estimate of $1/\mu$.) Thus, because of the nonlinearity of the logistic functional form, even though $E(\tilde{\beta}) \approx \beta$, $E(\tilde{\pi}_i)$ is not approximately equal to π_i. In fact, by interpreting Eq. (14) as an expected value over $\tilde{\beta}$, we can write $E_{\tilde{\beta}}(\tilde{\pi}_i) \approx \pi_i + C_i$, and the correction factor can be thought of as a bias term. Thus, surprisingly, subtracting the correction factor $(\tilde{\pi}_i - C_i)$ is approximately unbiased, but adding it $(\tilde{\pi}_i + C_i)$ produces a better estimator by reducing the mean square error.[9]

We denote $\tilde{\pi}_i - C_i$ as the *approximate unbiased* estimator and $\tilde{\pi}_i + C_i$ [in Eq. (15)] as the *approximate Bayesian* estimator. In the vast majority of applications, the approximate Bayesian estimator is preferable, although the unbiased estimator might be preferred in specialized situations, such as if one has a large set of small-n studies to be combined, as in a meta-analysis. (For this reason, we include both in some of our Monte Carlo studies below.) We do not see much justification for using the traditional ML method [$\hat{\pi}_i$ in Eq. (10)], except perhaps in situations where the variance matrix of the coefficients is nearly zero or about 50% of observations are ones. In these situations, the benefits of our approach will be relatively minor and might be outweighed by the slightly higher computational costs of our recommended approach.

6 Analyses

We use empirical analyses and Monte Carlo experiments in this section to clarify the conditions under which switching to our recommended approach generates differences substantial enough to warrant the extra effort (Section 6.1). (It is worth noting that the effort involved is quite minor, as the corrections are fairly simple.) We then demonstrate that the coefficients (Section 6.2) and probabilities (Section 6.3) computed under our recommended approach are superior to the traditional maximum-likelihood analysis of the logistic regression model.

6.1 *When Does It Make a Difference?*

In this section, we consider separately the correction for rare events, and we quantify when our recommended approaches make a material difference. Our companion paper offers a simulation analysis based on real data that shows how selection on Y works. Sections 6.2

[9] Deriving $\tilde{\pi}_i - C_i$ as an approximately unbiased estimator involves some approximations not required for the optimal Bayesian version derived in Appendix E. The problem is that instead of expanding a random π_i around a fixed $\tilde{\beta}$ as in the Bayesian version, we now must expand a random $\tilde{\pi}_i$ around a fixed β. Thus, to take the expectation and compute C_i, we need to imagine that in the correction term, $\hat{\pi}_i$ is a reasonable estimate of π_i in this context. This is obviously an undesirable approximation but it is better than setting it to zero or one (i.e., the equivalent of setting $C_i = 0$), and as our Monte Carlos show below, $\tilde{\pi}_i - C_i$ is indeed approximately unbiased.

and 6.3 then discuss interactions between the two corrections, which result primarily from the better balanced, but smaller, samples generated from choice-based sampling.

With Monte Carlo experiments, we now quantify the conditions under which our finite sample and rare events corrections are large enough to counterbalance the extra effort involved in implementing them. We focus here only on full cohort studies, and leave for subsequent sections the combination of endogenous sampling and finite sample, rare events corrections.

We first generated n observations from a logistic regression model with a constant and one explanatory variable drawn from a standard normal density, for fixed parameters β_0 and $\beta_1 = 1$. For each i, we drew a random uniform number u and assigned $Y_i = 1$ if $\pi_i < u$ and $Y_i = 0$ otherwise. We set the sample size to

$$n = \{100, 200, 500, 1000, 2000, 3000, 4000, 5000, 10{,}000, 20{,}000\}$$

and intercept to

$$\beta_0 = \{-7, -6, -5, -4, -3, -2, -1, 1\}$$

These values of $\boldsymbol{\beta}$ generate $\mathbf{y}$ vectors with the percentages of ones equaling $(100 \times \bar{y})\% = \{0.15, 0.4, 1.1, 2.8, 6.9, 15.6, 30.4, 50\}$ respectively. We excluded experiments with both very small percentages of ones and small sample sizes so as to avoid generating $\mathbf{y}$ vectors that are all zeros. This mirrors the common practice of studying rarer events in larger data sets. For each of these experiments, we computed the maximum difference in absolute risk by first taking the difference in estimates of $\Pr(Y = 1 \mid X = x)$ between the traditional logit model and our preferred approximate Bayesian method, for each of 31 values of x, equally spaced between -5 and 5, and then selecting the maximum. We also computed one relative risk, where we changed X from -1 to 1: $\Pr(Y = 1 \mid X = 1)/\Pr(Y = 1 \mid X = -1)$. The pair of X values, $\{-1, 1\}$, defines a typical relative risk that might be computed in examples like this, since it is at plus and minus one standard deviation of the mean of X, but it is of course neither the maximum nor the minimum difference in relative risk that could be computed between the two methods.

Finally, for each Monte Carlo experiment, we computed the maximum absolute risk and the relative risk averaged over 1000 simulated data sets. We have repeated this design with numerous other values of n, β_0, and β_1, and explanatory variables in different numbers and drawn from different (including asymmetric and partially discrete) densities. We also computed different absolute and relative risks. These other experiments led to similar conclusions as those presented here.

We summarize all this information in several ways in the subsequent sections and begin here with the simple graphical presentation in Fig. 3, with the maximum absolute risk in Fig. 3a and the relative risk in Fig. 3b. The horizontal axis in both figures is the percentage of ones in the sample, with data sets that have the rarest events at the left in the figure. For visual clarity, the horizontal axis is on the original logit scale, so that labeled percentages are $(100 \times \bar{y})\%$ but the tick marks appear at values of β_0. In Fig. 3a, the vertical axis is the maximum difference in absolute risk estimated by the two methods in percentage points. It is presented on the log scale, also for visual clarity. In Fig. 3b the vertical axis is the absolute difference in the percentage relative risk, again on the log scale. One line is given for each sample size.

Several conclusions are apparent from Fig. 3. First, as can be seen by comparing the different lines in either graph, the smaller the sample size, the higher the line and thus the larger the effect of our method. Second, since each line slopes downward, the rarer the events

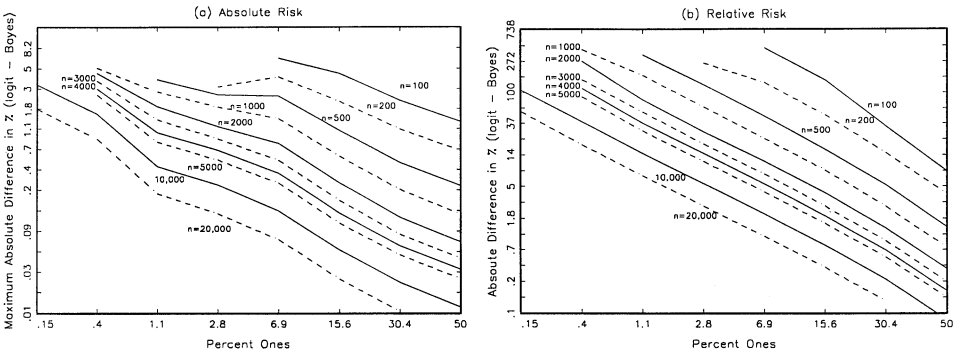

Fig. 3 Logit–Bayesian differences in (a) absolute risk and (b) relative risk as a function of sample size and rareness of events. The higher up each point appears in the graph (due to a smaller n or rarer events), the larger the difference our suggested method makes. The axes are labeled in percentages but on logit (for the horizontal) or log (for the vertical) scales to make the graph easier to read.

in a data set, the larger is the effect of switching methods. Clearly sample size and rareness of events are exchangeable in some way, as both measure the quantity of information in the data.

Finally, we examine the specific numerical values, but to understand these numbers, it is important to appreciate that what may seem like small values of the probabilities can have overwhelming importance in substantive analyses of genuine rare events data. For example, if a collection of 300,000 dyads witnesses a 0.001 increase in the probability of war, that can be catastrophically important because it means about 300 additional wars and a massive loss of human life. If the probability of contracting a particular fatal disease increases from 0.0001 to 0.001, it can mean huge numbers of additional deaths. Relative risks are typically considered important in rare event studies if they are at least 10–20%, but, of course, they can range much higher and have no upper limit. In Bennett and Stam's (1998b, Table 4) extensive analysis of conflict initiation and escalation in all dyads, for example, a majority of the 63 relative risks they report has absolute values of less than 25%.[10]

By these comparisons, the numerical values on the vertical axes in Fig. 3a are sizable and those in Fig. 3b are very large. For a sample with 2.8% ones, the difference between the methods in relative risk is about 128% for $n = 500$. This means that when the logit model estimate of a treatment effect (i.e., of the effect of a given change in X) is to increase the risk of an event by (say) 10%, the improved method's estimate is that the effect of the treatment will increase the risk by 128% on average. This is a very substantial difference. In the same circumstances, the difference between the methods in relative risk is 63% for $n = 1000$ and 28% for $n = 2000$. For 1.1% ones, our preferred method differs from logit on average by 332% for $n = 500$, 173% for $n = 1000$, and 78% for $n = 2000$. These differences are well above many of the estimated relative risks reported in applied literatures.

For absolute risk, with 2.8% ones, the difference in the methods is about 3% for $n = 500$, 2% for $n = 1000$, and 1% for $n = 2000$. With 1.1% ones, the difference between the logit and the Bayesian methods in absolute risk is about 4% for $n = 500$, 3% for $n = 1000$, and

[10] We translated the different format in which Bennett and Stam (1998b) report relative risk to our percentage figure. If r is their measure, ours is $100 \times (r - 1)$.

2% for $n = 2000$. These differences in absolute risk are larger than the reported effects for many rare events studies. The main exceptions are for those studies able to predict rare events with high levels of accuracy (so that estimates of π_i are large when $Y_i = 1$). Of course, Fig. 3 reports the average differences in absolute and relative risk between logit and our preferred method; the real effect in any one application can be larger or smaller.

Figure 3 also demonstrates that no sample size is large enough to evade finite sample problems if the events are sufficiently rare. For example, when $n = 20{,}000$ and 0.15% of the sample is ones, the difference between the existing methods and our improved methods is 1.8% in absolute risk and 53.5% in relative risk.

6.2 *Coefficients*

In this section, we study the properties of the coefficients and standard errors of logistic regression with and without our corrections, and for both cohort and case-control designs. To do this, we begin with the Monte Carlo methods described in Section 6.1, with $\beta_0 = -4$ (i.e., about 2.8% ones) and $n = 1000$, and then successively drop {0, 0.225, 0.45, 0.675, 0.9} fractions of observations with zeros. Since it has been well studied (Xie and Manski 1989), we omit the analysis of prior correction and weighting under model misspecification (which is known to favor weighting).

Although our ultimate goal is to reduce the mean square error, we focus here on bias since these coefficient bias corrections also reduce variance. Figure 4 presents one summary of the results. Biases for the intercept are given in Figs. 4a and b, and for the slope in Figs. 4c

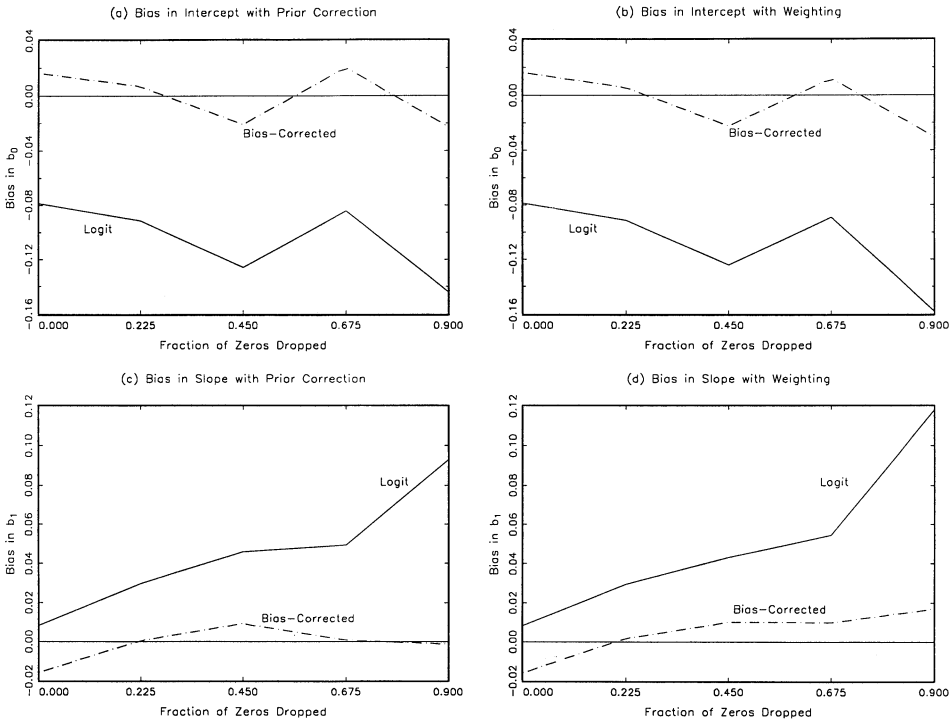

Fig. 4 Correcting bias in logit coefficients.

and d. Figures 4a and c display prior correction analyses, whereas Figs. 4b and d give weighting results. Since the horizontal axis in all figures is the fraction of zeros dropped, the leftmost point (at 0) is the same for both graphs within each row. The vertical axis of all four graphs is the degree of (positive or negative) bias, averaged over the 1000 simulations. The horizontal line in each figure marks the point of zero bias.

The results in Fig. 4 show overall that the logit line is more biased than the bias-corrected line, with a pattern very similar for prior correction and weighting. For the intercept, logit is below the zero bias line, a pattern that we see consistently in these and other simulations. Substantively, this pattern confirms the theoretical result that logit coefficients underestimate the probability of rare events. In addition, as more zeros are dropped, the bias increases, in part because the sample size used in the estimation is also dropping. In part to "compensate" for the bias in the intercept [i.e., since the ML solution constrains $\bar{y} = (1/n)\sum_{i=1}^{n} \hat{\pi}_i$], the bias in the slope is in the opposite direction. This result is typical but not universal, because more complicated situations can occur with more explanatory variables. Of course, the key result of importance in Fig. 4 is that the corrected line always stays fairly close to zero, and, crucially, this is true even for the version we designed to work with weighting methods in Figs. 4b and d. As the fraction of zeros dropped increases, the sample becomes better balanced but smaller, which results in more bias in logit but no appreciable change for the corrected versions.

We also examine, in Fig. 5, biases in the standard errors through the same Monte Carlo experiment. Since the biases in standard errors for the intercept and slope were about the same size, we averaged the biases and present only two graphs, Fig. 5a for prior correction and Fig. 5b for weighting. Also, the graphs for logit and our corrected versions are almost identical, and so we present only the former. For prior correction, we get the expected result that the true standard deviation across simulations is always approximately the same as the usual method based on the information matrix (unlabeled between the two other lines) and also nearly the same as that based on White's heteroskedasticity-consistent variance matrix.

The results are substantially different for weighting, as Fig. 5b shows that the usual information matrix method of computing standard errors is heavily biased with larger fractions of zeros dropped. That the usual method of computing standard errors is incorrect is discussed by Manski and Lerman (1977) and Xie and Manski (1989, Appendix), although the extent of the problem has not, to our knowledge, been demonstrated before. The problem is explained by the information matrix equality not holding under choice-based sampling. Since the other regularity conditions for ML hold, the general asymptotic variance matrix (equivalent to what is known as White's estimator) is available, which also appears in Fig. 5

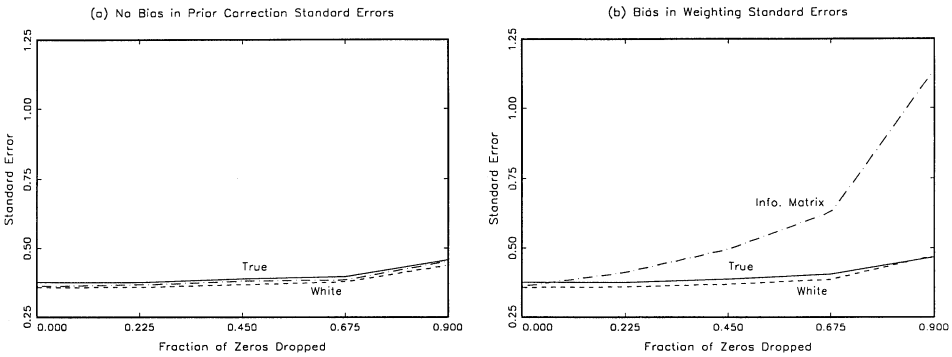

Fig. 5 Correcting bias in standard errors.

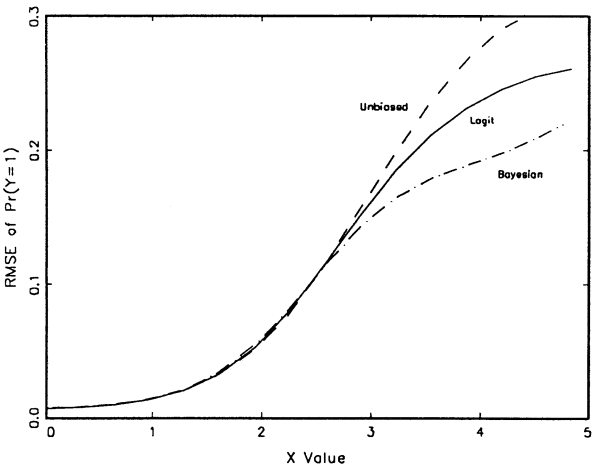

Fig. 6 RMSE in probability estimates: full sample.

as a dashed line (see Davidson and MacKinnon 1993, 263ff). Note how the dashed line in Fig. 5b closely approximates the solid (true) one. From here on, therefore, we use White's standard errors with the weighted estimator.

6.3 *Probabilities*

We now turn to an evaluation of estimates of $\Pr(Y = 1)$ with the same Monte Carlo experiments as described in Section 6.2. We focus here explicitly on the root mean square error (RMSE), since bias and variance are not simultaneously minimized by changes in probability estimates.

We begin with Fig. 6, which plots the RMSE (vertically) as a function of the value of X (horizontally) for three estimators of $\Pr(Y = 1)$, the traditional logit model, our preferred approximately Bayesian method, and the approximately unbiased approach. This is for cohort data without subsampling. In the left half of the graph, the three methods produce

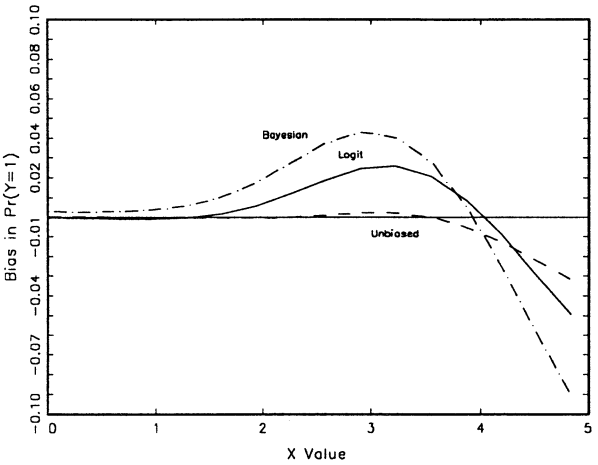

Fig. 7 Bias in probability estimates: full sample.

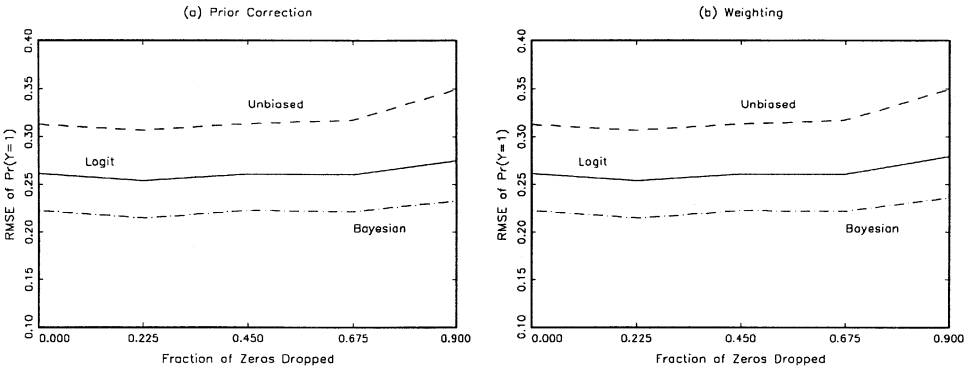

Fig. 8 RMSE of probability estimates: subsampled data.

about the same answers, but in the right half the Bayesian method clearly dominates the other two, with the unbiased method sacrificing the most in RMSE.

In many other similar analyses that we have run, the approximate Bayesian method has the lowest RMSE whenever the RMSE among the three methods differs to any significant degree. When the three are very close in RMSE (as on the left in Fig. 6), our recommended approach is normally better, and although sometimes points can be found where it does very slightly worse, we have not found a case where this makes a substantive difference. For all practical purposes, the approximate Bayesian method would appear to dominate the traditional logit and the approximately unbiased approaches.

Although we follow standard practice and would choose estimators based primarily on the RMSE, it is instructive to analyze the biases in this unusual situation where the three estimators are so algebraically similar. Figure 7 gives bias results in the same fashion as Fig. 6. It shows that the unbiased estimator is indeed closest to being unbiased. The Bayesian estimator has the largest bias for much of the graph, which of course is counterbalanced by a sufficiently lower variance so as to produce the lower RMSE result in Fig. 6.

We also present weighting and prior correction methods applied to subsampled data. Figure 8 plots the RMSE (vertically) by the fraction of zeros dropped (horizontally). For all ranges of subsampling, the Bayesian estimate has a lower RMSE than logit or the unbiased estimator. Virtually the same pattern appears for prior correction as for weighting.

Finally, we briefly evaluate relative risk, as defined in Section 6.1 as $\Pr(Y = 1 \mid X = 1)/\Pr(Y = 1 \mid X = -1)$. We present RMSE results in Fig. 9. This figure demonstrates that the same insights that apply to absolute risks also apply to relative risks: the Bayesian

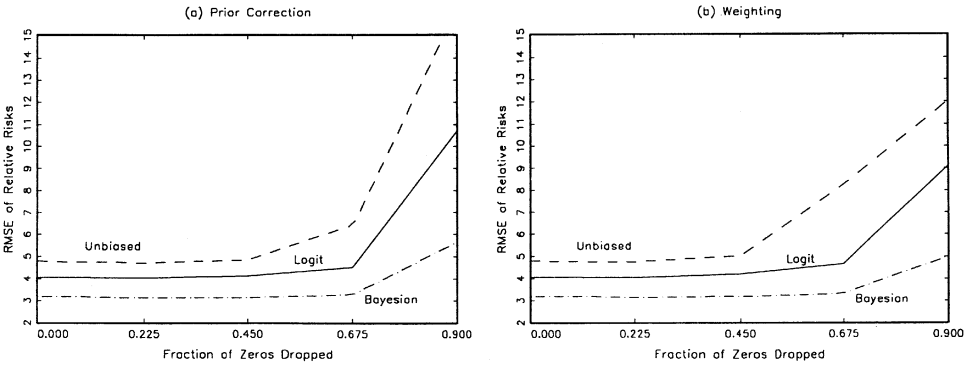

Fig. 9 RMSE of relative risk estimates: subsampled data.

estimator has the lowest RMSE, followed by the logit estimator, followed, finally, by the approximately unbiased approach. Thus, whether judged by absolute or relative risk, our approximate Bayesian estimator seems superior to the traditional approach based on the logit model or the approximately unbiased alternative.

7 Concluding Remarks

When analyzing rare events data, the finite sample properties of the quantities to be estimated may be worth some attention, especially since the rareness of events stands alongside the number of observations in constituting the amount of information in the data. We suggest methods with a lower mean square error and which, by increasing the probability of an event, could make a difference in much applied empirical research. The effects of these methods will be largest when the number of observations is small (under a few thousand) and the events are rare (under 5% or so). Typically, since when larger sample sizes are available, scholars take advantage of the extra information by studying even rarer events, the results in this paper will likely apply to at least some part of most rare event analyses. For example, in international conflict studies, scholars are usually interested in the occurrence of war in addition to the more commonly analyzed, and much larger category of, "militarized interstate disputes." With the additional information brought in by these methods, in combination with more flexible and highly interactive functional forms (Beck et al. 2000), perhaps the quantitative study of war will become more feasible. In addition, models with larger numbers of parameters, such as time-series cross-sectional models with many dummy variables, or neural network models, will likely generate bigger effects.

We also describe methods that enable one to reduce, or redirect, very large fractions of resources available for data collection. Since the resulting samples, with all available ones and a small fraction of zeros, are often fairly small, and because the fraction of ones in these populations is typically also small, we have adapted these methods so that their estimates can be simultaneously corrected for both selection on Y and problems due to finite samples and rare events. When the researcher is confident of the functional form and explanatory variables, prior correction is called for; otherwise, our corrected version of weighting with rare event corrections would seem preferable.

Appendix A: The Equivalence of Prior Correction, Conditional Maximum Likelihood, and Generalized Method of Moments

In this Appendix, we review some newer econometric methods for choice-based samples and prove that Manski and McFadden's (1981; see also Amemiya and Vuong 1987) conditional maximum-likelihood estimator is identical to prior correction (see Section 4.1) when the model is logistic. This was first proven by Hseih et al. (1985). We also prove here, apparently for the first time, that Imbens' (1992; see also Cosslett 1981a, b; Lancaster and Imbens 1996a, b) generalized method of moments estimator is equivalent to prior correction when the functional form is logistic and the sampling probability, $E(\bar{y})$, is unknown.

In exogenous sampling, the likelihood is $P(Y, X \mid \beta) = P(Y \mid X, \beta)P(X)$, but $P(X)$ is not a function of β and so can be dropped when maximizing the likelihood. Matters are not so simple in the full information likelihood analysis of choice-based samples, which involves maximizing

$$P(Y, X \mid \beta) = P(X \mid Y, \beta)P(Y) = \frac{P(X, Y \mid \beta)\bar{y}}{P(Y \mid \beta)} = \frac{P(Y \mid X, \beta)P(X)\bar{y}}{P(Y \mid \beta)} \qquad (17)$$

where $P(Y \mid \beta) = \int P(Y \mid X, \beta) P(X) dX = \tau$ serves as a constraint on $P(X)$ when τ is known. Since X is implicated in this denominator, which involves β, $P(X)$ must be estimated along with β. This means that one needs to maximize the likelihood over all possible parameters β and all possible probability densities $P(X)$.

The problem of estimating $P(X)$ seemed intractable at first (Manski and Lerman 1977), but Manski and McFadden (1981) proposed a *conditional maximum-likelihood* estimator by conditioning Eq. (17) on X. This estimator is consistent and asymptotically normal, more efficient than weighting (Amemiya and Vuong 1987), but not fully efficient in all cases, because it excludes information about $P(X)$ contained in both $P(X)$ and $P(Y \mid \beta)$. We show here that it is equal to prior correction (and hence is fully efficient) in the special case of logit.

First, denote the functional form for prior correction (from Section 4.1) as $\pi_{(\mathrm{pc})i} = [1 + e^{-\boldsymbol{x}_i\boldsymbol{\beta} - \ln(w_0/w_1)}]^{-1}$, where $w_1 = \tau/\bar{y}$ and $w_0 = (1-\tau)/(1-\bar{y})$. The likelihood function for the constrained maximum-likelihood (CML) estimator can be written in our notation and simplified as

$$L_{\mathrm{cml}} = \prod_{i=1}^{n} \left[\frac{\pi_i / w_1}{\pi_i / w_1 + (1-\pi_i)/w_0} \right]^{y_i} \left[\frac{(1-\pi_i)/w_0}{\pi_i / w_1 + (1-\pi_i)/w_0} \right]^{1-y_i} \tag{18}$$

$$= \prod_{i=1}^{n} \left(\pi_{(\mathrm{cml})i}\right)^{y_i} \left(1 - \pi_{(\mathrm{cml})i}\right)^{1-y_i} \tag{19}$$

That this likelihood is equivalent to that under prior correction can be proven by rearranging the functional form as follows:

$$\pi_{(\mathrm{cml})i} = \frac{\pi_i / w_1}{\pi_i / w_1 + (1-\pi_i)/w_0} = \left[1 + \left(\frac{1-\pi_i}{\pi_i} \right) \left(\frac{w_1}{w_0} \right) \right]^{-1} \tag{20}$$

$$= \left[1 + e^{-\boldsymbol{x}_i\boldsymbol{\beta} - \ln(w_0/w_1)} \right]^{-1} = \pi_{(\mathrm{pc})i} \tag{21}$$

Cosslett (1981a, b) improves on CML by parameterizing $P(\mathbf{X})$ with a set of weights at each of the n points of support (the weights together defining a simplex) and maximizing Eq. (17) directly. He then sequentially maximizes the weight parameters along with β, resulting in his asymptotically efficient *pseudo-maximum-likelihood* estimator, but this method is very difficult computationally. Imbens (1992; see also Lancaster and Imbens 1996a, b), in what now appears to be the state of the art, proposes a semiparametric generalized method of moments estimator that is consistent and as efficient as Cosslett's but poses fewer computational burdens. By deriving the first-order conditions of the log-likelihood in Eq. (17), Imbens demonstrates that the weights can be written as an explicit function of the other parameters and the data and, hence, substituted out. He then reinterprets the equations in a generalized method of moments framework, which he uses to prove that the estimator is consistent and asymptotically efficient.

Imbens' estimator has four moment equations. He drops the fourth because it is orthogonal to the others. In our logit model, the first moment is $\psi_1 = E(\bar{y}) - y_i$ but when, as usual, $E(\bar{y})$ is unknown, and hence $\bar{y}$ is substituted, $\sum_{i=1}^{n} \psi_1 / n = \bar{y} - \sum_{i=1}^{n} y_i / n = 0$, and so we find that ψ_1 can be dropped as well. The remaining two moments, in our notation and with $E(\bar{y})$ unobserved, are

$$\psi_2 = \tau - \frac{\pi_i}{\pi_i / w_1 + (1-\pi_i)/w_0} \tag{22}$$

$$\psi_{3k} = X_{ik}\left[y_i - \frac{(\pi_i/w_1)}{\pi_i/w_1 + (1-\pi_i)/w_0}\right] \tag{23}$$

where $k = 1, \ldots, K$ indexes elements of ψ_3 and columns of $\mathbf{x}_i$. In the case of logit, $\bar{\psi}_{31} = \sum_{i=1}^{n} \psi_{31}/n$ (corresponding to the constant term, $X_{i1} = 1$) is a linear function of $\bar{\psi}_2 = \sum_{i=1}^{n} \psi_2/n$: $\bar{\psi}_2/w_1 = \bar{\psi}_{31}$. In cases like this, Imbens (1992, p. 120) drops ψ_{31}, but we instead drop ψ_2, which is informationally equivalent. This leaves only ψ_3, which Imbens shows is equivalent to the moments of CML in general and, as we have shown above, is also equal to the moments of prior correction in our case.

Appendix B: The Consistency of Prior Correction

In this Appendix, we derive the method of prior correction described in Section 4.1, beginning with a generic statistical model and specializing in four steps until we reach our logistic regression special case [and hence derive Eq. (7)]. In its most general formulation in Section B.1, prior correction is consistent but not necessarily feasible to apply. Fortunately, in the logit special case discussed in Section B.4, prior correction is consistent, fully efficient, and easy to apply; it gives estimates equivalent to maximizing the full information likelihood in Eq. (17) (Manski and Lerman 1977).

B.1 *In General*

Suppose X, Y are random variables with density $P(X, Y)$ (representing the full sample as in a case-cohort study) and x, y are random variables with density $P(x, y)$ (representing a sample with all ones and a random selection of zeros from X, Y). The density $P(x, y)$ is defined by subsampling such that $P(x \mid y) = P(X \mid Y)$, although the marginals $P(x)$, $P(y)$, and $P(y \mid x)$ do not necessarily equal $P(X)$, $P(Y)$, and $P(Y \mid X)$, respectively. The goal of the analysis is inference about $P(Y \mid X)$, which we express as

$$P(Y \mid X) = P(X \mid Y)\frac{P(Y)}{P(X)} = P(y \mid x)\left[\frac{P(Y)}{P(y)}\frac{P(x)}{P(X)}\right] \tag{24}$$

The general claim is that we can estimate $P(Y \mid X)$ with an iid sample drawn either from $P(X, Y)$ [or $P(Y \mid X)$] or from $P(x, y)$ [or $P(y \mid x)$] with a correction by multiplying the result by the last, bracketed term in Eq. (24). To show this, let D and d be random samples of size n from $P(X, Y)$ and $P(x, y)$, respectively. Then as $n \to \infty$,

$$P(Y \mid X, D) = P(X \mid Y, D)\frac{P(Y \mid D)}{P(X \mid D)} \xrightarrow{d} P(X \mid Y)\frac{P(Y)}{P(X)} = P(Y \mid X) \tag{25}$$

but $P(y \mid x, d) = P(x \mid y, d)P(y \mid d)/P(x \mid d) \overset{d}{\not\to} P(Y \mid X)$ (where $\xrightarrow{d}$ and $\overset{d}{\not\to}$ denote convergence and nonconvergence in distribution, respectively). However, letting $A_y = P(Y \mid D)/P(y \mid d)$ be a function of y and $B = P(x \mid d)/P(X \mid D) = [\sum_{\text{all } y} P(y \mid x, d)A_y]^{-1}$ be a constant normalization factor,

$$\begin{aligned} P(y \mid x, d)A_y B &= P(x \mid y, d)\frac{P(y \mid d)}{P(x \mid d)}A_y B = P(x \mid y, d)\frac{P(Y \mid D)}{P(X \mid D)} \\ &\xrightarrow{d} \frac{P(X \mid Y)P(Y)}{P(X)} = P(Y \mid X) \end{aligned} \tag{26}$$

since $P(x \mid y, d) \xrightarrow{d} P(x \mid y) = P(X \mid Y)$, $P(Y \mid D) \xrightarrow{d} P(Y)$, and $P(X \mid D) \xrightarrow{d} P(X)$. Thus, the corrected subsampled distribution is consistent for the distribution of interest, since $P(y \mid x, d)A_y B \xrightarrow{d} P(Y \mid X)$, for any data collection strategy that selects on Y or $Y \mid X$ (so long as it does not truncate any value), or on X, but not on $X \mid Y$.

B.2 *Finite Discrete Choice Models*

Finite discrete choice models (such as logit, probit, ordinal models, multinomial logit or probit, nested multinomial logit, neural network classification problems, etc.) specify $\Pr(Y = j \mid X)$ for $j = 1, \ldots, J$ with J finite. Letting $\Pr(Y = j \mid D) = \tau_j$, which is assumed known, and $\Pr(y = j \mid d) = \bar{y}_j$ be either known or estimated from the observed sample, the correction factors are $A_j = \tau_j/\bar{y}_j$ and $B^{-1} = \sum_{j=1}^{J} P(y = j \mid x, d)\tau_j/\bar{y}_j$. Then the sample estimate is

$$P(y = j \mid x, d)A_j B = \frac{P(y = j \mid x, d)\tau_j/\bar{y}_j}{\sum_{k=1}^{J} P(y = k \mid x, d)\tau_k/\bar{y}_k} \xrightarrow{d} P(Y = j \mid X) \tag{27}$$

B.3 *Binary Models*

In binary models, such as logit, probit, scobit, neural network classification, etc., $\Pr(Y = 1) = \tau$, and $\Pr(y = 1) = \bar{y}$, and so the correction factors are $A_1 = \tau/\bar{y}$, $A_0 = (1-\tau)/(1-\bar{y})$, and $B^{-1} = \Pr(y = 1 \mid x, d)\tau/\bar{y} + [1 - \Pr(y = 1 \mid x, d)](1 - \tau)/(1 - \bar{y})$. Hence

$$\begin{aligned} P(y = 1 \mid x, d)A_1 B &= \frac{P(y = 1 \mid x, d)\tau/\bar{y}}{P(y = 1 \mid x, d)(\tau/\bar{y}) + [1 - P(y = 1 \mid x, d)](1 - \tau)/(1 - \bar{y})} \\ &= \left[1 + \left(\frac{1}{P(y = 1 \mid x, d)} - 1\right)\left(\frac{1 - \tau}{\tau}\right)\left(\frac{\bar{y}}{1 - \bar{y}}\right)\right]^{-1} \end{aligned} \tag{28}$$

B.4 *Logistic Regression (and Neural Networks)*

Finally, in the logit model if $\Pr(y = 1 \mid x, d) = 1/(1 + e^{-\mathbf{x}_i\boldsymbol{\beta}})$, then

$$P(y = 1 \mid x, d)A_1 B = \left[1 + e^{-\mathbf{x}_i\boldsymbol{\beta} + \ln[(\frac{1-\tau}{\tau})(\frac{\bar{y}}{1-\bar{y}})]}\right]^{-1}, \tag{29}$$

which demonstrates that the MLE of β_1 need not be changed, but the constant term should be corrected by subtracting out the bias factor, $\ln[((1 - \tau)/\tau)(\bar{y}/(1 - \bar{y}))]$.

Equation (29) also applies to any other model with a logit output function. For example, for a feed forward neural network model with a logit output function (as in Beck, King, and Zeng, 2000), only the constant term in the hidden neuron-to-output layer needs be corrected. This can be done subtracting the same bias factor as in binary logit from the constant term.

Appendix C: Logit Bias Corrections with Optional Weights

We now prove the bias correction in Eq. (11). McCullagh and Nelder (1989, Sect. 15.2) show that the bias may be computed for any generalized linear model as $(\mathbf{X'WX})^{-1}\mathbf{X'W}\xi_i$, where the first factor is the Fisher information matrix and $\xi_i = -0.5(\mu_i''/\mu_i')Q_{ii}$, where μ_i is the inverse link function relating $\mu_i = E(Y_i)$ to $\eta_i = \mathbf{x}_i\boldsymbol{\beta}$, μ_i' and μ_i'' are the first and second derivatives of μ_i with respect to η_i, and Q_{ii} are the diagonal elements of $\mathbf{X}(\mathbf{X'WX})^{-1}\mathbf{X'}$.

The key to our derivation is that the weighted likelihood in Eq. (8) can be made equivalent to the unweighted likelihood in Eq. (5) by changing the probability function to $\Pr(Y_i) = \pi^{w_1 Y_i}(1-\pi_i)^{w_0(1-Y_i)}$. Then $\mu_i = E(Y_i) = [1/(1+e^{-\eta_i})]^{w_1} \equiv \pi_i^{w_1}$, and hence $\mu_i' = w_1\pi_i^{w_1}(1-\pi_i)$, $\mu_i'' = w_1\pi_i^{w_1}(1-\pi_i)[w_1 - (1+w_1)\pi_i]$, and $\xi_i = 0.5Q_{ii}[(1+w_1)\pi_i - w_1]$.

We then derive $\mathbf{W}$ from the information matrix for the log-likelihood in Eq. (8):

$$-E\left(\frac{\partial^2 \ln L_w(\boldsymbol{\beta} \mid \mathbf{y})}{\partial\beta_j \partial\beta_k}\right) = \sum_{i=1}^{n} \pi_i(1-\pi_i)\mathbf{x}_j w_i \mathbf{x}_k' = \{\mathbf{X'WX}\}_{j,k} \tag{30}$$

and so $\mathbf{W} = \text{diag}\{\hat{\pi}_i(1-\hat{\pi}_i)w_i\}$.

Finally, to compute the variance matrix of the bias term, we use McCullagh and Nelder's (1989, p. 457) rough approximation for small $\boldsymbol{\beta}$, $[n/(n+k)]\hat{\boldsymbol{\beta}} \approx \tilde{\boldsymbol{\beta}}$, and so $V(\tilde{\boldsymbol{\beta}}) \approx (n/(n+k))^2 V(\hat{\boldsymbol{\beta}})$.

Appendix D: Bias in Single Parameter Logistic Regression

In this Appendix, we derive Eq. (12) beginning with McCullagh's (1987, p. 210) general result that

$$E(\hat{\beta}_0 - \beta_0) = -\frac{1}{2n}\frac{i_{30} + i_{11}}{i_{20}^2} + O(n^{-2}) \tag{31}$$

where $i_{30} = E[(\partial L/\partial\beta_0)^3]$, $i_{11} = E[(\partial L/\beta_0)(\partial^2 L/\partial^2\theta^2)]$, and $i_{20} = E[(\partial L/\partial\beta_0)^2]$ are evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$. In our special case, $\pi_i = 1/(1+e^{-(\beta_0+X_i)})$. Then $\partial L/\partial\beta_0 = (1-\hat{\pi}_i)^{Y_i}(-\hat{\pi}_i)^{1-Y_i}$, and $\partial^2 L/\partial^2\beta_0^2 = -(1-\hat{\pi}_i)\hat{\pi}_i$, which by substitution gives

$$E(\hat{\beta}_0 - \beta_0) \approx -\frac{1}{n}\frac{E\left\{(0.5-\hat{\pi}_i)[(1-\hat{\pi}_i)^2 Y_i + \hat{\pi}_i^2(1-Y_i)]\right\}}{(E\left[(1-\hat{\pi}_i)^2 Y_i + \hat{\pi}_i^2(1-Y_i)\right])^2} \tag{32}$$

All the interpretative understanding we wished to convey about this special case is available by studying Eq. (32)—in particular, that the bias is signed (by the first factor in the numerator), reduced as n increases, and amplified when events are more rare (because the denominator is a function of the variance). However, to provide a simpler expression for expository purposes, we also act as if, solely for these interpretative purposes (and not for any empirical analyses, for example), that in this expression $\hat{\pi}_i = \pi_i$. This is obviously a very rough approximation, but the qualitative interpretation remains unchanged. Under this assumption, Eq. (32) simplifies to Eq. (12).

Appendix E: Analytical Approximations for Probability Computations

This Appendix derives Eq. (15) as an approximation to Eq. (14). To apply the integral, which is intractable with direct methods, we first approximate $1/(1+e^{-\mathbf{x}_0\boldsymbol{\beta}})$ by a Taylor series expansion around $\tilde{\boldsymbol{\beta}}$, retaining up to the second-order term. The integral is then easy. Thus,

$$\Pr(Y_0 = 1) \approx \tilde{\pi}_0 + \left[\frac{\partial\pi_0}{\partial\boldsymbol{\beta}}\right]_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}) + \frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})'\left[\frac{\partial^2\pi_0}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}'}\right]_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}) \tag{33}$$

where the second term is $\tilde{\pi}_0(1-\tilde{\pi}_0)\mathbf{x}_0(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})$, the third term is $(0.5-\tilde{\pi}_0)\tilde{\pi}_0(1-\tilde{\pi}_0)\mathbf{x}_0\mathbf{D}\mathbf{x}_0'$, and where $\mathbf{D}$ is $k \times k$ with k, j element equal to $(\beta_k - \tilde{\beta}_k)(\beta_j - \tilde{\beta}_j)'$. Under a Bayesian interpretation, $\tilde{\pi}_0$ and $\tilde{\boldsymbol{\beta}}$ are functions of the data and hence constant but π_0 and $\boldsymbol{\beta}$ are random

variables, and so taking the expectation and making use of the expected bias $\mathbf{b} = E(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})$ and variance matrices $V(\tilde{\boldsymbol{\beta}})$ gives

$$\begin{aligned}\Pr(Y_0 = 1) &= E[1/(1 + e^{-\mathbf{x}_0\boldsymbol{\beta}})] \\ &\approx \tilde{\pi}_0 + \tilde{\pi}_0(1 - \tilde{\pi}_0)\mathbf{x}_0\mathbf{b} + (0.5 - \tilde{\pi}_0)(\tilde{\pi}_0 - \tilde{\pi}_0^2)\mathbf{x}_0[V(\tilde{\boldsymbol{\beta}}) + \mathbf{b}\mathbf{b}']\mathbf{x}_0' \quad (34)\end{aligned}$$

Since $\mathbf{b} \approx \mathbf{0}$, Eq. (34) reduces to Eq. (15).

References

Achen, Christopher A. 1999. "Retrospective Sampling in International Relations," presented at the annual meetings of the Midwest Political Science Association, Chicago.

Agresti, A. 1992. "A Survey of Exact Inference for Contingency Tables (with discussion)." *Statistical Science* 7(1):131–177.

Amemiya, Takeshi, and Quang H. Vuong. 1987. "A Comparison of Two Consistent Estimators in the Choice-Based Sampling Qualitative Response Model." *Econometrica* 55(3):699–702.

Beck, Nathaniel, Gary King, and Langche Zeng. 2000. "Improving Quantitative Studies of International Conflict: A Conjecture." *American Political Science Review* 94(1):1–15. (Preprint at http://GKing.Harvard.Edu.)

Bennett, D. Scott, and Allan C. Stam, III. 1998a. *EUGene: Expected Utility Generation and Data Management Program, Version 1.12*. http://wizard.ucr.edu/cps/eugene/eugene.html.

Bennett, D. Scott, and Allan C. Stam, III. 1998b. "Theories of Conflict Initiation and Escalation: Comparative Testing, 1816–1980," presented at the annual meeting of the International Studies Association Minneapolis.

Breslow, Norman E. 1996. "Statistics in Epidemiology: The Case-Control Study." *Journal of the American Statistical Association* 91:14–28.

Breslow, Norman E., and N. E. Day. 1980. *Statistical Methods in Cancer Research*. Lyon: International Agency for Research on Cancer.

Bueno de Mesquita, Bruce. 1981. *The War Trap*. New Haven, CT: Yale.

Bueno de Mesquita, Bruce, and David Lalman. 1992. *War and Reason: Domestic and International Imperatives*. New Haven, CT: Yale University Press.

Bull, Shelley B., Celia M. T. Greenwood, and Walter W. Hauck. 1997. "Jackknife Bias Reduction for Polychotomous Logistic Regression." *Statistics in Medicine* 16:545–560.

Cordeiro, Gauss M., and Peter McCullagh. 1991. "Bias Correction in Generalized Linear Models." *Journal of the Royal Statistical Society, B* 53(3):629–643.

Cosslett, Stephen R. 1981a. "Maximum Likelihood Estimator for Choice-Based Samples." *Econometrica* 49(5):1289–1316.

Cosslett, Stephen R. 1981b. "Efficient Estimation of Discrete-Choice Models." In *Structural Analysis of Discrete Data with Econometric Applications*, eds. Charles F. Manski and Daniel McFadden. MIT Press. MA: Cambridge.

Firth, David. 1993. "Bias Reduction of Maximum Likelihood Estimates." *Biometrika* 80(1):27–38.

Geisser, Seymour. 1993. *Predictive Inference: An Introduction*. New York: Chapman and Hall.

Geller, Daniel S., and J. David Singer. 1998. *Nations at War: A Scientific Study of International Conflict*. New York: Cambridge University Press.

Greene, William H. 1993. *Econometric Analysis*, 2nd ed. New York: Macmillan.

Holland, Paul W., and Donald B. Rubin. 1988. "Causal Inference in Retrospective Studies," *Evaluation Review* 12(3):203–231.

Hsieh, David A., Charles F. Manski, and Daniel McFadden. 1985. "Estimation of Response Probabilities from Augmented Retrospective Observations." *Journal of the American Statistical Association* 80(391):651–662.

Huth, Paul K. 1988. "Extended Deterrence and the Outbreak of War." *American Political Science Review* 82(2):423-443.

Imbens, Guido. 1992. "An Efficient Method of Moments Estimator for Discrete Choice Models with Choice-Based Sampling." *Econometrica* 60(5):1187–1214.

King, Gary, and Langche Zeng. 2000a. "Inference in Case-Control Studies with Limited Auxilliary Information" (in press). (Preprint at http://Gking.harvard.edu.)

King, Gary, and Langche Zeng. 2000b. "Explaining Rare Events in International Relations." *International Organization* (in press).

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44(2):341–355. (Preprint at http://Gking.harvard.edu.)

Lancaster, Tony, and Guido Imbens. 1996a. "Case-Control with Contaminated Controls." *Journal of Econometrics* 71:145–160.

Lancaster, Tony, and Guido Imbens. 1996b. "Efficient Estimation and Stratified Sampling." *Journal of Econometrics* 74:289–318.

Levy, Jack S. 1989. "The Causes of War: A Review of Theories and Evidence." In *Behavior, Society, and Nuclear War, Vol. 1*, eds. Phillip E. Tetlock, Jo L. Husbands, Robert Jervis, Paul C. Stern, and Charles Tilly. New York, Oxford: Oxford University Press, pp. 2120–2333.

Manski, Charles F. 1999. "Nonparametric Identification Under Response-Based Sampling." In *Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya*, eds. C. Hsiao, K. Morimune, and J. Powell. New York: Cambridge University Press (in press).

Manski, Charles F., and Steven R. Lerman. 1977. "The Estimation of Choice Probabilities from Choice Based Samples." *Econometrica* 45(8):1977–1988.

Manski, Charles F., and Daniel McFadden. 1981. "Alternative Estimators and Sample Designs for Discrete Choice Analysis." In *Structural Analysis of Discrete Data with Econometric Applications*, eds. Charles F. Manski and Daniel McFadden. Cambridge: MA: MIT Press.

Maoz, Zeev, and Bruce Russett. 1993. "Normative and Structural Causes of Democratic Peace, 1946–86." *American Political Science Review* 87(3):624–638.

McCullagh, Peter. 1987. *Tensor Methods in Statistics*. New York: Chapman and Hall.

McCullagh, P., and J. A. Nelder, 1989. *Generalized Linear Models*, 2nd ed. New York: Chapman and Hall.

Mehta, Cyrus R., and Nitin R. Patel. 1997. "Exact Inference for Categorical Data," unpublished paper. Cambridge, MA: Harvard University and Cytel Software Corporation.

Nagelkerke, Nico J. D., Stephen Moses, Francis A. Plummer, Robert C. Brunham, and David Fish. 1995. "Logistic Regression in Case-Control Studies: The Effect of Using Independent as Dependent Variables." *Statistics in Medicine* 14:769–775.

Prentice, R. L., and R. Pyke. 1979. "Logistic Disease Incidence Models and Case-Control Studies." *Biometrika* 66:403–411.

Ripley, Brian D. 1996. *Pattern Recognition and Neural Networks*. New York: Cambridge University Press.

Rosenau, James N., ed. 1976. *In Search of Global Patterns*. New York: Free Press.

Rothman, Kenneth J., and Sander Greenland. 1998. *Modern Epidemiology*, 2nd ed. Philadelphia: Lippincott–Raven.

Schaefer, Robert L. 1983. "Bias Correction in Maximum Likelihood Logistic Regression." *Statistics in Medicine* 2:71-78.

Scott, A. J., and C. J. Wild. 1986. "Fitting Logistic Models Under Case-Control or Choice Based Sampling." *Journal of the Royal Statistical Society, B* 48(2):170–182.

Signorino, Curtis S. 1999. "Strategic Interaction and the Statistical Analysis of International Conflict." *American Political Science Review* 93(2):279–298.

Signorino, Curtis S., and Jeffrey M. Ritter. 1999. "Tau-b or Not Tau-b: Measuring the Similarity of Foreign Policy Positions." *International Studies Quarterly* 40(1):115–144.

Smith, Richard L. 1998. "Bayesian and Frequentist Approaches to Parametric Predictive Inference." In *Bayesian Statistics*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith. New York: Oxford University Press.

Tanner, M. A. 1996. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 3rd ed. New York: Springer-Verlag.

Tucker, Richard. 1998. "The Interstate Dyad-Year Dataset, 1816–1997," Version 3.0. http://www.fas.harvard.edu/~rtucker/data/dyadyear/.

Tucker, Richard. 1999. "BTSCS: A Binary Time-Series–Cross-Section Data Analysis Utility," Version 3.0.4. http://www.fas.harvard.edu/~rtucker/programs/btscs/btscs.html.

Vasquez, John A. 1993. *The War Puzzle*. Cambridge, New York: Cambridge University Press.

Verba, Sidney, Kay Lehman Schlozman, and Henry E. Brady. 1995. *Voice and Equality: Civic Voluntarism in American Politics*. Cambridge, MA: Harvard University Press.

Wang, C. Y., and R. J. Caroll. 1995. "On Robust Logistic Case-Control Studies with Response-Dependent Weights." *Journal of Statistical Planning and Inference* 43:331–340.

Xie, Yu, and Charles F. Manski. 1989. "The Logit Model and Response-Based Samples." *Sociological Methods and Research* 17(3):283–302.