Explaining Rare Events in International Relations

Gary King and Langche Zeng

Many of the most significant events in international relations—wars, coups, revolutions, massive economic depressions, economic shocks—are rare events. They occur infrequently but are considered of great importance. In international relations, as in other disciplines, rare events—that is, binary dependent variables characterized by dozens to thousands of times fewer 1's (events such as wars or coups) than 0's (nonevents)—have proven difficult to explain and predict. Though scholars have made substantial efforts to quantify rare events, they have devoted less attention to how these events are analyzed. We show that problems in explaining and predicting rare events stem primarily from two sources: popular statistical procedures that underestimate the probability of rare events and inefficient data-collection strategies. We analyze the issues involved, cite examples from the international relations literature, and offer some solutions.

The first source of problems in rare-event analysis is researchers' reliance on logit coefficients, which are biased in small samples (those with fewer than two hundred observations), as the statistical literature well documents. Not as widely understood is that the biases in probabilities can be substantively meaningful when sample sizes are in the thousands and are always in the same direction: estimated event probabilities are always too small. A separate, often 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

We thank Scott Bennett, Kristian Gleditsch, Paul Huth, and Richard Tucker for data; the National Science Foundation, the Centers for Disease Control and Prevention (Division of Diabetes Translation), the National Institutes of Aging, the World Health Organization, the Center for Basic Research in the Social Sciences for research support; the editors and referees of IO for helpful comments; and Ethan Katz for research assistance. The software ReLogit: Rare Events Logistic Regression, which we wrote to implement the methods discussed in this article, is available at ⟨http://gking.harvard.edu⟩. We have written a companion to this article, titled "Logistic Regression in Rare Events Data," that overlaps this one; it includes complete mathematical proofs, more general notation, and other technical material but fewer examples and less pedagogy; it is available at ⟨http://gking.harvard.edu⟩ and is forthcoming in *Political Analysis.*

International Organization 55, 3, Summer 2001, pp. 693–715

same direction as biases in the coefficients. Applied researchers almost never correct for underestimated event probabilities, even though doing so is easy. These problems will be innocuous in some applications; in others, however, the error can be as large as the reported estimated effects. We demonstrate how to correct for these problems and provide software to make the computation straightforward.

A second and more important source of problems in analyzing rare events lies in how data are collected. Given fixed resources, researchers must weigh the tradeoff between gathering more observations and including better or additional variables. The fear of collecting data sets with no events (and thus with no variation on Y) has led researchers to choose very large data sets with few, and often poorly measured, explanatory variables. This choice is reasonable given the perceived constraints, but far more efficient strategies exist. Researchers can, for example, collect all (or all available) 1's and a small random sample of 0's and not lose consistency or even much efficiency relative to the full sample. This strategy drastically changes the optimal tradeoff between collecting more observations and including better variables by enabling scholars to focus data-collection efforts where they matter most; for example, later in the article we use all national dyads for each year since World War II to generate a data set with 303,814 observations, of which only 0.3 percent, or 1,042 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 data sets containing 5,000–10,000 observations are inconvenient to deal with if one has to collect variables for all the cases.) Moreover, most dyads involve countries with little relationship to each other (such as Burkina Faso and St. Lucia) much less with some actual probability of going to war, and so there is a well-founded perception that much of the data is "nearly irrelevant."[2] Indeed, the data may have very little substantive content, which explains why we can forgo collecting most of these observations without much loss of efficiency. In contrast, most existing approaches in political science designed to cope with this problem, such as selecting dyads that are "politically relevant,"[3] are reasonable and practical approaches to a difficult problem, but they necessarily reframe the question, 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, 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 0's need not be collected or could even be deleted with only a minor effect on substantive conclusions.

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

2. Maoz and Russett 1993, 627.

3. Maoz and Russett 1993.

These procedures enable scholars who wish to add new variables to an existing data-set to save about 99 percent of the nonfixed costs in their data-collection budgets or to reallocate data-collection efforts to generate a larger number of informative and meaningful variables than would otherwise be possible.[4] International relations scholars over the years have given extraordinary attention to issues of measurement and have generated a large quantity of data. By selecting on the dependent variable as we suggest, scholars can build on these efforts; compared with traditional sampling methods the method we propose will increase the efficiency of subsequent data collections by changing the optimal tradeoff in favor of fewer observations and more sophisticated measures that more closely reflect the desired concepts.

This procedure of selecting on Y also addresses a long-standing controversy in the literature on international conflict. Qualitative scholars devote their efforts where the action is (the conflicts) but draw criticism for introducing bias by selecting on the dependent variable. In contrast, quantitative scholars are criticized for spending time analyzing very crude measures on many observations that for the most part contain no relevant information.[5] To a certain degree, the intuition on both sides is correct: the substantive information in the data lies much more with the 1's than the 0's, but researchers must be careful to avoid selection bias. Fortunately, corrections are easily made, and so the goals of both camps can be met.[6]

Logistic Regression: Model and Notation

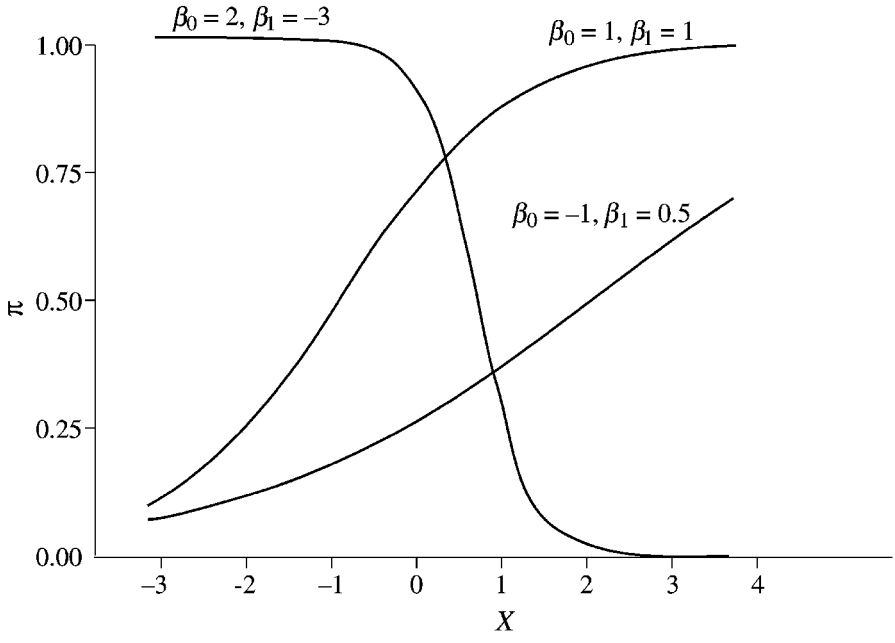
In logistic regression, a single outcome variable, Y_i ($i = 1, \ldots, n$), is coded 1 (for war, for example) with probability π_i, and 0 (for peace, for example), with probability $1 - \pi_i$. Then π_i varies as a function of some explanatory variables, such as X_i for democracy. The function is logistic rather than linear, which means that it resembles an escalator (see Figure 1), and mathematically it is expressed as

$$\pi_i = \frac{1}{1 + e^{-\beta_0 - \beta_1 X_{1i}}}.$$

4. 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.

5. See Bueno de Mesquita 1981; Geller and Singer 1998; Levy 1989; Rosenau 1976; and Vasquez 1993.

6. 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 in Bueno de Mesquita and Lalman 1992 (appendix) and in an unpublished paper they cite that has recently become available as Achen 1999.

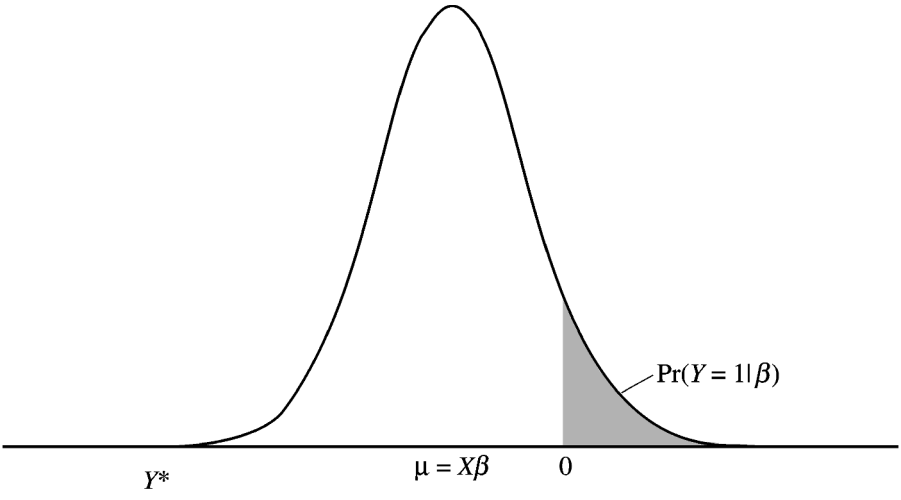


Note: This graph plots the values of an explanatory variable (X) and an outcome variable (π) (for example, the probability of war) for each of several pairs of intercepts and slope coefficients.

FIGURE 1. *Examples of logistic curves*

Another way to define the same model is to first imagine an unobserved continuous variable that represents the propensity of a country to go to war, Y_i^*. We cannot measure this variable directly like we can the presence of war, but it exists and we have some indicators of it (for example, we know that Iran's propensity to go to war today is higher than Barbados', even though we observe both to be in an equivalent state when it comes to the absence of military conflict). Let us assume that Y_i^* follows a logistic distribution, which is a bell-shaped distribution that looks so close to the normal that from a substantive perspective the difference is trivial (mathematically, of course, there is a small difference).

If we observe Y_i^* and want to know the effects of X_i, most political scientists would simply run a regression with Y_i^* as the dependent variable. However, instead of observing Y_i^*, we only observe whether this propensity is greater than some threshold beyond which the country goes to war. For example, if $Y_i^* > 0$, we should see a war $Y_i = 1$; otherwise, if $Y_i^* < 0$, we should observe peace $Y = 0$. This observation mechanism (see Figure 2) turns out to be the chief troublemaker in bias induced by rare events.



Note: The horizontal axis (Y^*) is the unobserved propensity to go to war. The shaded area is the probability that a nation will go to war (π).

FIGURE 2. *Logit observation mechanism*

The coefficients of β are estimated using maximum likelihood.[7] As part of the estimation process, we also get the standard error, which for the estimate of β_1 is approximately the square root of

$$V(\hat{\beta}_1) = \frac{1}{\sum_{i=1}^{n} \pi_i(1 - \pi_i)X_i^2}.$$

In this equation the factor most affected by rare events is $\pi_i(1 - \pi_i)$, which we now interpret. Note that $\pi_i(1 - \pi_i)$ reaches its maximum at $\pi_i = 0.5$ and approaches 0 as π_i gets close to 0 or 1. Most rare-events applications yield very small estimates of π_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 (that is, for which $Y_i = 1$) will usually be larger than among observations for which $Y_i = 0$; the estimate will also be closer to 0.5, because probabilities in rare-event studies are normally very small.[8] The result is that $\pi_i(1 - \pi_i)$ will usually be larger for 1's than for 0's, and so its contribution to the variance (its

7. Which merely gives the values that maximize the likelihood of getting the data we actually observe; see King 1989.

8. Beck, King, and Zeng 2001.

inverse) will be smaller. In this situation, additional 1's will be more informative than additional 0's.

Finally, we note that in logistic regression the quantity of interest is not the raw $\hat{\beta}$ output most computer programs provide but the more direct functions of the probabilities. For example, *absolute risk* is the probability that an event occurs given chosen values of the explanatory variables, $\Pr(Y = 1|X = x)$. The *risk ratio* is the same probability relative to the probability of an event given some baseline values of X, such as $\Pr(Y = 1|X = 1)/\Pr(Y = 1|X = 0)$, the fractional increase in the risk. This quantity is frequently reported in the popular media (for example, the probability of developing some forms of cancer increases by 50 percent if one stops exercising) and is common in many scholarly literatures. Also of considerable interest is the *first difference* (or risk difference), the change in probability as a function of a change in a covariate, such as $\Pr(Y = 1|X = 1) - \Pr(Y = 1|X = 0)$. The first difference is usually most informative when measuring effects, whereas the risk ratio tends to be easier to compare across applications or time periods. Although scholars often argue about the relative merits of each,[9] we suggest that when convenient it is best to report both the risk ratio and first difference or to report the two component absolute risks.

Advantages of Selecting on the Dependent Variable

We first distinguish among alternative data-collection strategies and show how to adapt the logit model for each. We then build on the adapted models to allow rare-event and finite sample corrections. Political scientists understand that selecting a sample of data using a rule correlated with Y causes bias, even after controlling for X.[10] Some scholars know that data containing sample-selection bias can be corrected. What seems essentially unknown in the discipline is that correcting for selection on a binary dependent variable is easily accomplished, requires no assumptions, and can save enormous costs in data collection.

Random selection is desirable because the selection rule is known to be independent of all other variables (as long as the sample size is large enough) and so cannot cause bias. Similarly, *exogenous stratified sampling*, which allows Y to be randomly selected within categories defined by X (such as in a random sample of democracies and a random sample of nondemocracies), is desirable for the same reasons. The usual statistical models are optimal under both sampling schemes. Indeed, in epidemiology, random selection and exogenous stratified sampling are both known as *cohort* studies (or *cross-sectional* studies, to distinguish them from panel studies).

When one of the values of Y is rare in the population, analysts can save considerable resources in data collection by randomly selecting within categories of

9. See Breslow and Day 1980, chap. 2; and Manski 1999.
10. For example, King, Keohane, and Verba 1994.

Y. This is known in econometrics as *choice-based* or *endogenous stratified* sampling and in epidemiology as a *case-control* design.[11] This sampling design is also useful for choosing qualitative case studies.[12] 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 the available information on 1's (for example, 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 a subsample using all the 1's and a random selection of 0's. The case-cohort study is especially appropriate when an expensive variable is added to an existing collection, such as the dyadic data discussed earlier and analyzed later. Imagine measuring strategic misperception; for 300,000 dyads, this would be impossible, but much could be learned from a small case-control design.

Many other hybrid data-collection strategies have also been tried and might be useful in international relations. For example, Bruce Bueno de Mesquita and David Lalman's data-collection design is fairly close to a case-control study with "contaminated controls," meaning that the "control" sample is taken from the whole population rather than only from those observations for which $Y = 0$.[13] Although we do not analyze hybrid designs in this article, our view is *not* that pure case-control sampling is appropriate for all political science studies of rare events (one could argue, for example, that 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, we argue that scholars should consider a much wider range of potential sampling strategies and associated statistical methods than are now common. We focus here only on the leading alternative design, which we believe has the potential for widespread use.

Data-collection designs that select on Y can be efficient but are valid only with the appropriate statistical corrections. In the next section we discuss the method of prior correction for estimation under choice-based sampling (we discuss other estimation methods in our companion article). For the past twenty years econometricians have made steady progress in generalizing and improving these methods. However, David Hsieh, Charles Manski, and Daniel McFadden have shown that two of these econometric methods are equivalent to prior correction for the logit model.[14] We recently explicated this result and showed that the best econometric estimator in this tradition also reduces to the method of prior correction under the logit model.[15] We later summarize problems to avoid in designing choice-based samples.

11. Breslow 1996.
12. King, Keohane, and Verba 1994.
13. See Bueno de Mesquita and Lalman 1992; and Lancaster and Imbens 1996a.
14. Hsieh, Manski, and McFadden 1985.
15. King and Zeng forthcoming.

Correcting a Case-control Analysis

Prior correction is the easiest method of correcting a logistic regression in a case-control sampling design. The procedure is to run a logistic regression and correct the estimates based on external information about the fraction of 1's in the population, τ, and the observed fraction of 1's in the sample (or sampling probability), $\bar{y}$. Knowledge of τ can be derived from census data, a random sample from the population measuring Y only, a case-cohort sample, or other sources. In studies of international conflict, we typically have a census of conflicts even if we are analyzing only a small subset, and so τ is generally known.[16]

In any of the sampling designs discussed earlier, the logit model slope coefficients, $\hat{\beta}_1$, are statistically consistent estimates of β_1, and so correction is unnecessary. That this result holds for case-control designs under the logit model seems magical, and it is widely seen as such even by those steeped in the mathematics. The result holds only in the popular logistic regression model, not in probit or linear regression. Although the slope coefficients are consistent even in the presence of selection on Y, they are of little use alone: scholars are not typically interested in β_1 but rather in functions of the probability that an event occurs, $\Pr(Y_i = 1|\beta) = \pi_i = 1/(1 + e^{-\beta_0 - X_i\beta_1})$, which requires good estimates of both β_1 and β_0.[17] Correcting the constant term is easy; instead of the estimate produced by the computer, $\hat{\beta}_0$, use the following:

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

A key advantage of prior correction is its ease of use. Any statistical software that can estimate logit coefficients can be employed, and Equation (1) is easy to apply to the intercept. If the functional form and explanatory variables are correct, estimates are consistent and asymptotically efficient.

Problems to Avoid

Selecting on the dependent variable in the way we suggest has several pitfalls that should be carefully avoided. First, prior correction is appropriate for sampling designs that require independent random (or complete) selection of observations for which $Y_i = 1$ and $Y_i = 0$. Both the case-control and case-cohort studies meet this requirement. However, other endogenous designs, such as hybrid approaches and

16. For methods for the case where τ is unknown, see King and Zeng 2000.

17. 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 in the special case of all discrete covariates and has been shown to apply to other multiplicative intercept models. See Bishop et al. 1975, 63; and Hsieh, Manski, and McFadden 1985, 659.

approaches that employ sampling in several stages using nonrandom selection, require different statistical methods.

Second, when selecting on Y, care must be taken not to select on X differently for the two samples. A classic example of this is selecting all people in the local hospital with liver cancer ($Y_i = 1$) and selecting a random sample of the U.S. population without liver cancer ($Y_i = 0$). The problem is not recognizing the implicit selection on X; that is, the sample of cancer patients selects on $Y_i = 1$ *and implicitly* on the inclination to seek health care, find the right medical specialist, have the right tests, and so on. Since the $Y_i = 0$ sample does not similarly select on the same explanatory variables, the data would induce selection bias. One solution in this example might be to select the $Y_i = 0$ 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.

Inadvertently selecting on X can be a serious problem in endogenous designs, just as selecting 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 caused by selection on X is much easier to avoid in applications such as international conflict and related fields, since a clearly designated census of cases from which to draw a sample is normally available. 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.[18]

A third problem with intentionally selecting on Y is that valid exploratory data analysis can be hazardous. In particular, one cannot use an explanatory variable as a dependent variable in an auxiliary analysis without taking special precautions.[19]

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

We can also use formal statistical results to suggest procedures for determining the optimal tradeoff between collecting more observations and using better variables. First, when 1's and 0's are equally easy to collect and an unlimited number of each are available, an "equal-shares sampling design" (that is, $\bar{y} = 0.5$) is

18. Holland and Rubin 1988.
19. Nagelkerke et al. 1995.

optimal in a limited number of situations and close to optimal in many.[20] This is a useful fact, but in fields like international relations, the number of observable 1's (such as wars) is strictly limited, so in most applications it is best to collect all available 1's or a large sample of them. The only real decision then is how many 0's to collect as well. If collecting 0's is costless, we should collect as many as we can get, since more data are always better. If collecting 0's is not costless but not (much) more expensive than collecting 1's, we should collect more 0's than 1's. However, since the marginal contribution to the explanatory variables' information content for each additional 0 starts to drop as the number of 0's passes the number of 1's, we will not often want to collect more than (roughly) two to five times more 0's than 1's. In general, the optimal number of 0's depends on how much more valuable the explanatory variables become with the resources saved by collecting fewer observations.

Finally, a useful practice is to proceed sequentially. First, collect all the 1's and (say) an equal number of 0's. If the standard errors and confidence intervals are narrow enough, stop; otherwise, continue to sample 0's randomly and stop when the confidence intervals become sufficiently small for the substantive purposes at hand. For some samples, it might even be efficient to collect explanatory variables sequentially as well, but this is not often the case.

Rare-event Corrections

In this section we discuss methods of computing probability estimates that correct problems resulting from rare events or small samples (or both). Using these methods is easy; instead of logit, you would use our ReLogit software. Instead of logit coefficients, you get bias-corrected coefficients; instead of probabilities, you get relative risks; and first differences computed on the basis of logit coefficients result in better estimates from relogit runs. The procedure is as easy to use as logit, and in Stata it is virtually the same. When the results make a difference, our methods work better than logit; when they do not, these methods give the same answer as logit. There does not appear to be a cost to switching, although only in some data with either rare events or small samples does it make much of a difference.

The usual method of estimating a probability is to run a logit, record the estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$, choose a value for the explanatory variable (or variables) X, and plug these into the following equation:

$$\hat{\pi} = \frac{1}{1 + e^{-\hat{\beta}_0 - \hat{\beta}_1 X}}. \tag{2}$$

20. See Cosslett 1981; and Imbens 1992.

Unfortunately, two distinct problems in data with rare events or small samples affect this method of computing probabilities. First, the coefficient estimates are biased. Second, even if you use our unbiased version of logit coefficients, plugging these into Equation (2) would still be an inferior estimator of the probability. The details of our corrections to each problem are given in our companion paper. In this section, we use simple relationships and graphs to provide some intuition into the problems and corrections.

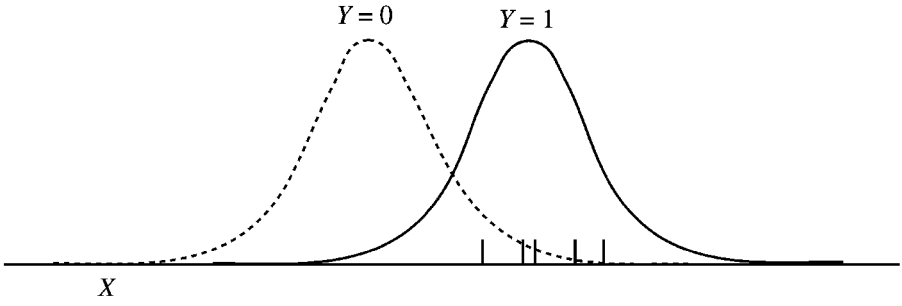
Parameter Estimation

We know from the statistical literature that logit slope and intercept estimates are biased in finite samples, and that less biased and more efficient methods are available (unlike the sampling designs mentioned earlier, these corrections affect all the coefficients). This knowledge has apparently not made it into the applied literatures partially because the statistical literature does not include studies showing how rare events greatly magnify the biases. This omission has led to conclusions that downplay the effects of bias; Robert L. Schaefer, for example, argues that "sample sizes above 200 would yield an insignificant bias correction."[21]

Finite-sample bias amplified by rare events is occasionally discussed informally in the pattern-recognition and classification literatures,[22] but it 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, though these literatures reveal a practical wisdom given that their data-collection strategies naturally produce samples where the proportion of 1's is about half.

Our results show that, for rare-events data, the probability of war, $\Pr(Y = 1)$, is underestimated, and hence the probability of peace, $\Pr(Y = 0)$, is overestimated. To grasp this intuitively, consider a simplified case with one explanatory variable, as illustrated in Figure 3. First, we order the observations on Y_i according to their values on X_i (the horizontal dimension in Figure 3). If $\beta_1 > 0$, and so the probability of $Y_i = 1$ is positively related to the value of X_i, most of the 0's will be to the left and the 1's will be to the right with little overlap. Since there were so many 0's in the example, we replaced them with a dashed line that represents the density of X in the $Y = 0$ group. The few 1's in the data set appear as short vertical lines, and the distribution from which they were drawn appears as a solid line that represents the density of X in the $Y = 1$ group. (As drawn, the distributions of X for 0's and 1's are normal, but that need not be the case.) Although the large number of 0's allows us to estimate the dashed density line using a histogram essentially without error, any estimate of the solid density line for X when $Y = 1$ from the mere five data points will be very poor and, indeed, systematically biased toward tails that are

21. Schaefer 1983, 73.
22. Ripley 1996.



Note: Observations are arrayed horizontally according to the value of X, where $\beta > 0$. The short vertical lines represent the few $Y = 1$ observations; the solid line represents the density from which they were randomly drawn. The many $Y = 0$ observations are not shown, but the solid line represents their density. The cutting point that best classifies 0's and 1's (and is related to β_1) will be too far to the right because the density of 0's will be better estimated than the density of 1's and no information exists about the left end of the density line.

FIGURE 3. *How rare events bias logit coefficients*

too short. To conceptualize this, consider finding a cutting point (value of X) that maximally distinguishes the 0's and 1's, that is, by making the fewest mistakes (0's misplaced to the right of the cut point or 1's to the left). This cutting point is related to the logistic regression 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 0's than 1's, the maximum value of X in the $Y = 0$ group, $\max(X|Y = 0)$ (and more generally the area in the right tail of the density of X in that group, $P(X|Y = 0)$, will be well estimated, but $\min(X|Y = 1)$ (and the area in the left tail of $P(X|Y = 1)$ will be poorly estimated. Indeed, $\min(X|Y = 1)$ 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 $\max(X|Y = 0) - \min(X|Y = 1)$), it will be biased in the direction of favoring 0's at the expense of 1's, and so $\Pr(Y = 1)$ will be too small.

To explain what the rare-events correction is doing, we have derived a simple expression in a special case. The bias term appears to affect the constant term directly and the other coefficients primarily as a consequence; therefore, we consider the special case with a constant term and one explanatory variable and with β_0 estimated and $\beta_1 = 1$ fixed. For this case, we find that the estimated intercept is larger on average than the true intercept by approximately $(\bar{\pi} - 0.5)/[n\bar{\pi}(1 - \bar{\pi})]$, where $\bar{\pi}$ is the average of all the probabilities in the data. Since $\bar{\pi} < 0.5$ in rare-events data, the numerator, and thus the entire expression, is negative. This means $\hat{\beta}_0$ is too small, and consequently $\Pr(Y = 1)$ is underestimated. The

denominator is also informative, because it shows that as n (the number of observations) gets large, the bias vanishes. Finally, the factor $\bar{\pi}(1 - \bar{\pi})$ in the denominator gets smaller as events become rarer. Since this factor is in the denominator, the entire equation for bias is amplified with rarer events (that is, as $\bar{\pi}$ approaches 0).

Probability Calculations

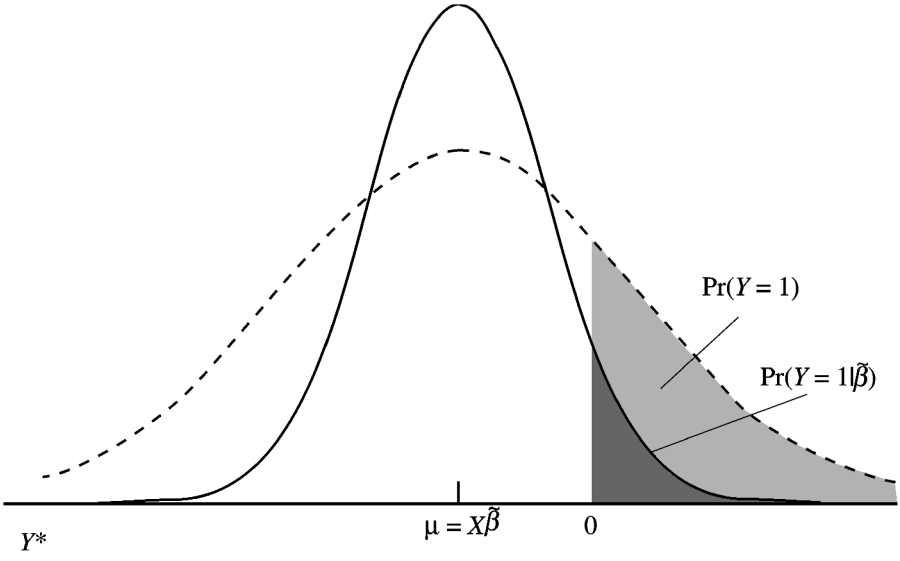
If $\hat{\beta}$ are all the coefficients from the usual logistic regression, and we define $\tilde{\beta}$ as the bias-corrected version outlined earlier, we can compute a probability by plugging this bias-corrected version into the same equation,

$$\tilde{\pi} = \Pr(Y = 1|\tilde{\beta}) = \frac{1}{1 + e^{-X_0\tilde{\beta}}}.$$

It is true that $\tilde{\pi}$ is preferable to $\hat{\pi}$, but $\tilde{\pi}$ is still not optimal because it ignores the uncertainty in $\tilde{\beta}$.[23] This uncertainty can be thought of as sampling error or as the fact that $\tilde{\beta}$ is estimated rather than known. 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 by plugging in an estimate of β generates too small an estimated probability of a rare event (or in general an estimate too far from 0.5). This can be understood intuitively by considering the underlying continuous variable Y^* that the basic model assumes to be logistic (that is, pretty close to normal). In the model shown in Figure 4 the probability is the area to the right of the threshold (the dark shaded area to the right of zero under the solid curve), an area typically less than 0.5 in rare-events data. The problem is that ignoring uncertainty about β leads to a distribution whose variance is too small and thus (with rare events) has too little area to the right of the threshold. To see what happens when adding in the uncertainty, imagine jiggling around the mean of the logistic distribution, noted on Figure 4 as $\mu = X\tilde{\beta}$, and averaging the blur of distributions that results. Since *IO* is not yet capable of showing animated graphics, we present only the result: In Figure 4 the additional variance is illustrated in the change from the solid to the dashed density line; that is, the variance of the distribution increases when we include the uncertainty in the estimate $\tilde{\beta}$. As a result of this increased variance, the mean stays in the same position, but the area to the right of the 0 threshold has increased (from the dark-shaded area marked $\Pr(Y = 1|\tilde{\beta})$ to both shaded areas, marked $\Pr(Y = 1)$). This means that including uncertainty makes the probability larger (closer to 0.5).

23. For example, King, Tomz, and Wittenberg forthcoming.



Note: Although the solid density (which does not reflect uncertainty in β) has a smaller variance than the dotted line (which has the uncertainty about β added in), the mean, μ, stays the same in both. However, the probability, the shaded area to the right of the 0 threshold in the two curves, differs.

FIGURE 4. *The effect of uncertainty on probabilities*

Thus, both the change to unbiased logit coefficients and the change to the improved method of computing probabilities lead to an increase in the estimated probability compared with the traditional methods based on logistic regression. The effect is largest when events are rare or when the sample size is small, whether or not events are rare.

When Does It Make a Difference in Practice?

In this section, we consider separately the corrections for selecting on Y and for rare events, and we quantify when our recommended approaches make a material difference.

Selecting on Y in Militarized Interstate Dispute Data

To demonstrate the advantages of case-control sampling we compiled a data set on international conflict with dyad-years as observations fairly typical of those in the

literature.[24] The outcome variable is coded 1 if the dyad was engaged in a "militarized interstate dispute," and 0 otherwise. The explanatory variables include those typically used in this field, including whether the pair of countries includes a major power, are contiguous, are allies, and have similar foreign policy portfolios. Also included are the balance of military power, the number of years since the last dispute, and the nation in the dyad having the minimum and maximum degrees of trade dependency and democracy.

Table 1 presents five analyses. The first column shows a traditional logistic regression on all 303,814 observations. The next column shows prior correction applied to data with 90 percent of the 0's dropped, leaving only $n = 31{,}319$ observations. In the last column 99 percent of the 0's have been dropped, resulting in only $n = 4{,}070$ observations. For simplicity, and because of the large n, we ignore the finite sample and rare-events corrections for now.

The numerical results in the 90-percent column are fairly close to those in the full sample, with standard errors that are only slightly higher. The analysis with 99 percent of the 0's dropped is also reasonably close to the full sample, but as expected the results are more variable. That the standard errors are slightly larger reflects, and predicts, this fact. Of the forty-four coefficients in the four subsample columns, only a few numbers vary enough to make much of a substantive difference in interpretation (such as indicating changes in significance or magnitude). Of course, since deleting too many observations can decrease efficiency too much and make standard errors unacceptably large, one must always consider the tradeoff between the efficiency gained by including more observations and the resources saved by using better variables. Collecting all dyads, as is now the universal practice, rarely represents the optimal tradeoff, but the decision depends on the goals of the particular application. Be aware that in interpreting Table 1, we expect some variation across the rows due to random sampling. For example, if these columns were analyses of independent random samples of the same size from the same population, we would expect considerable variation across each row (with samples of about 1,000, as is common, the 95-percent confidence interval would be ± 6 percent).

The explanatory variables chosen for this application reflect approximately the state of the art in this field but predict international conflict only very weakly. As such, a key concern among researchers has been to find more meaningful and reliable measures, but the sheer magnitude of the data-collection task effectively dictates far simpler, more arbitrary measures. Using a case-control design, a researcher with a fixed budget can measure much more sophisticated explanatory variables. One way to think about the effort being saved or redirected would be to imagine collecting data for all 303,814 observations but having seventy-five times as many researchers available to collect it (303,814/4,070). (This calculation

24. We began with all dyad-years from 1946 to 1992 available in Tucker 1998; merged variables from Bennett and Stam 1998a,b; and computed years since the last dispute using the program by Tucker 1999 and using measures of alliance portfolios in Signorino and Ritter 1999.

TABLE 1. *Estimating the same parameters without 99 percent of the data*

Explanatory variables	*Full sample*[a]	*Prior correction (90% of 0's dropped)*[b]	*Prior correction (99% of 0's dropped)*[b]
Contiguous	3.56	3.55	3.96
	(0.09)	(0.10)	(0.12)
Allies	−0.27	−0.21	−0.28
	(0.09)	(0.12)	(0.15)
Foreign policy	0.23	0.45	0.38
	(0.20)	(0.24)	(0.26)
Balance of power	1.00	0.96	1.13
	(0.13)	(0.15)	(0.19)
Max. democracy	0.20	0.13	0.17
	(0.06)	(0.07)	(0.08)
Min. democracy	−0.18	−0.07	−0.06
	(0.06)	(0.07)	(0.08)
Max. trade	0.05	0.05	0.05
	(0.01)	(0.01)	(0.01)
Min. trade	−0.07	−0.07	−0.08
	(0.01)	(0.01)	(0.01)
Years since dispute	−0.11	−0.10	−0.09
	(0.01)	(0.01)	(0.01)
Major power	1.31	1.81	2.2
	(0.09)	(0.12)	(0.14)
Constant	−6.78	−6.91	−7.14
	(0.23)	(0.26)	(0.30)
n	303,814	31,319	4,070

Note: Numbers in parentheses are standard errors.
[a]Logistic regression coefficients based on a full sample.
[b]Logistic regression coefficients after prior corrections on data with 90 and 99 percent of $Y_i = 0$ observations randomly dropped.

exaggerates the savings for types of variables that are easier to measure in regional or language clusters and for which an alternative method of selecting on Y might be helpful.) Variables that were previously infeasible to include using quantitative methods but might now be worth collecting include trade commodity data (since the aggregate figures do not distinguish between types of products and services), measures of leaders' perceptions based on survey data or historical work, more meaningful measures of physical proximity between countries than contiguity, and information on the process of strategic interaction as crises unfold.[25] Collecting data on each of these variables can be expensive; by reducing the number of observations and thus the amount of data needed, we should be able to learn a great deal more about international conflict than was previously possible.

25. Signorino 1999.

The Effects of Rare-event and Small Sample Corrections

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 case-control sampling and the small-sample or rare-events corrections. Our method is to simulate artificial data sets from a world whose characteristics we know (because we created it) and are similar to real-world data. We then run our methods and the standard methods and see when the differences are large enough to worry about.

We first generated n observations from a logistic regression model with a constant and one explanatory variable, for fixed parameters β_0 and $\beta_1 = 1$. We set the sample size to

$$n = \{100, 200, 500, 1{,}000, 2{,}000, 3{,}000, 4{,}000, 5{,}000, 10{,}000, 20{,}000\}$$

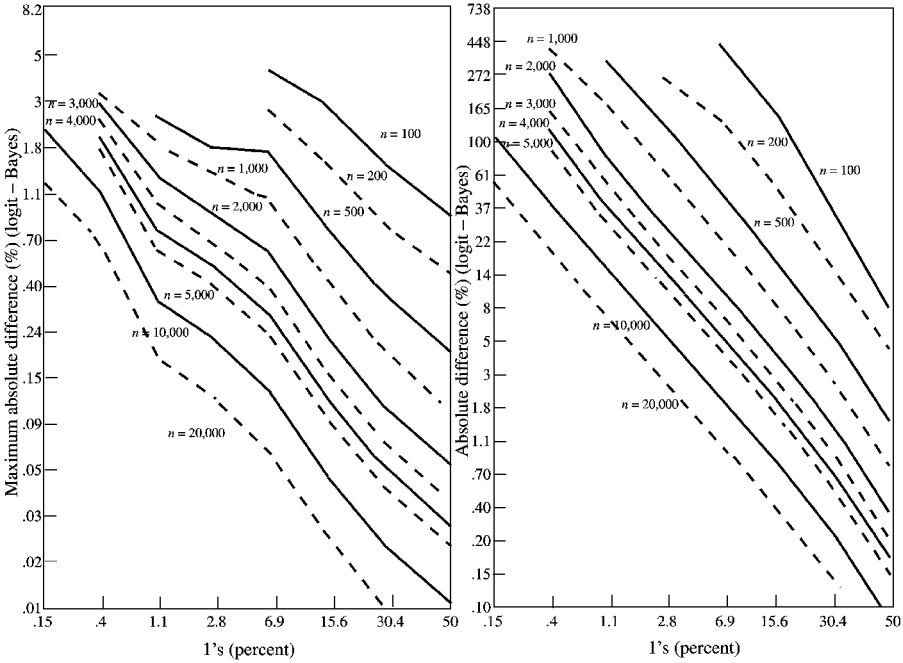
and the intercept to

$$\beta_0 = \{-7, -6, -5, -4, -3, -2, -1, 0\}.$$

These values generate Y vectors with the percentages of 1's 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 1's and small sample sizes to avoid generating Y vectors that are all 0's. This mirrors the common practice of studying rarer events in larger data sets. For each of these experiments, we computed the maximum difference in probability by first taking the difference in estimates of $\Pr(Y = 1|X = x)$ between the traditional logit model and our preferred "approximate Bayesian" method, for each of thirty-one 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: $\Pr(Y = 1|X = 1)/\Pr(Y = 1|X = -1)$. The pair of X values, $\{-1, 1\}$, defines a typical relative risk that might be computed in examples like this, since it is at ± 1 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 1,000 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 conclusions similar to those presented here.

We summarize the results in Figure 5; Figure 5a shows the maximum absolute risk, and Figure 5b shows the relative risk. The horizontal axis in both figures is the percentage of 1's in the sample, with data sets that have the rarest events at the left



Note: The higher up each point appears in a 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 the logit scale (for the horizontal) and log scale (for the vertical) to make the graph easier to read.

FIGURE 5. *Logit-Bayesian differences in (a) absolute risk and (b) relative risk as a function of sample size and rareness of events*

of the figure. For visual clarity, the horizontal axis is on the original logit scale; labeled percentages are $(100 \times \bar{y})\%$, but the tick marks appear at values of β_0. In Figure 5a, the vertical axis is the maximum difference in absolute risk estimated by the two methods; for visual clarity, it is presented on the log scale. In Figure 5b, the vertical axis is the absolute difference in relative risk, again on the log scale. One line is given for each sample size.

Several conclusions are apparent from Figure 5. First, as can be seen by comparing the lines within either panel, the smaller the sample size, the higher the line, and thus the greater the effect of our method. Second, since each line slopes downward, the rarer the events in a data set, the larger the effect of switching methods. Clearly, sample size and rareness of events are exchangeable in some way since both measure the quantity of information in the data.

We now examine the specific numerical values. 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 shows a 0.001 increase in the probability of war, the finding is catastrophically important because it represents about three hundred additional wars and a massive loss of human life. Relative risks are typically considered important in rare-event studies if they are at least 10–20 percent, but, of course, they can range much higher and have no upper limit. In Scott Bennett and Allan Stam's extensive analysis of conflict initiation and escalation in all dyads, for example, a majority of the sixty-three relative risks they report have absolute values of less than 25 percent.[26]

By these comparisons, the numerical values on the vertical axis of Figure 5a are sizeable and of Figure 5b are very large. For a sample with 2.8 percent 1's, the difference in relative risk between the methods is about 128 percent for $n = 500$. This means that when the logit model estimate of a treatment effect (that is, a given change in X) increases the risk of an event by 10 percent, for example, the estimate in our suggested method will increase the risk by 128 percent on average. This is a very substantial difference. Under the same circumstances, the difference between the methods in relative risk is 63 percent for $n = 1{,}000$, and 28 percent for $n = 2{,}000$. For 1.1 percent 1's, our preferred method differs from logit, on average, by 332 percent for $n = 500$, 173 percent for $n = 1{,}000$, and 78 percent for $n = 2{,}000$. These differences are well above many of the estimated relative risks reported in applied literatures.

For absolute risk, with 2.8 percent 1's, the difference in the methods is about 3 percent for $n = 500$, 2 percent for $n = 1{,}000$, and 1 percent for $n = 2{,}000$. With 1.1 percent 1's, the difference between the logit and Bayesian methods in absolute risk is about 4 percent for $n = 500$, 3 percent for $n = 1{,}000$, and 2 percent for $n = 2{,}000$. 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, Figure 5 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 5 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 percent of the sample are 1's, the difference between the existing methods and our improved methods is 1.8 percent in absolute risk and 53.5 percent in relative risk.

26. Bennett and Stam 1998b, tab. 4. We translated their reported relative risk to our percentage figure—if r was their measure, ours is $100(r - 1)$.

TABLE 2. *Replication of "extended deterrence and the outbreak of war" from Huth 1988*

Change in balance of military forces	*First difference*		*Relative risk*	
	Logit	Bayes	Logit	Bayes
1:4 to 1:1	19.5%	11.9%	40.1%	23.2%
1:1 to 3:1	26.9%	16.8%	39.6%	26.4%

Note: The "first difference" is the difference between two absolute risks as the balance of military forces changes (as indicated in the first column) ($n = 58$).

Rare-event Corrections in Deterrence Outcomes

We illustrate our methods here by reanalyzing Paul Huth's analysis of the determinants of deterrence outcomes.[27] We reproduced his probit analysis (from his Table 1), reran the data with logit (which changed nothing important), and then applied our techniques. Huth's analysis predicts the probability of deterring an aggressor as a function of the military balance of forces, the countries' bargaining behaviors, and their reputation earned in past deterrence episodes. The data include fifty-eight observations and about 58 percent 1's. Although this is not a rare-events situation, recall that a small number of observations is equivalent, and according to Figure 5 Huth's case should be approximately equivalent in terms of the effects of our method to a data set with $n = 1{,}000$ and 2.4 percent 1's, or $n = 2{,}000$ and 1.2 percent 1's.

For simplicity, we focus only on Huth's first two substantive interpretations.[28] Table 2 summarizes our reanalyses for the change in absolute risk (a first difference) and relative risk resulting from a change in the immediate balance of military forces, holding other variables constant at their means. For example, according to the traditional logit approach, a change in the balance of forces from a 1:4 disadvantage for the defender to equality increases the probability of deterrence success by 19.5 percentage points. In contrast, our approximate Bayesian approach gives a much smaller effect of only 11.9 points. (Our approach reduces the effect, rather than increasing it, because $\bar{y} > 0.5$.) Using traditional logit methods to measure relative risk yields an estimate of 40.1 percent, but our method suggests a more modest effect of only 23.2 percent. Similar reductions also occur when changing the balance of military forces from equality to 3:1 for the defender, as in the second row of Table 2. Overall, our approach produces fairly large and substantively meaningful changes.

27. Huth 1988.
28. Huth 1988, 437.

Corrected Forecasts of State Failure

Finally, we include an analysis of data taken from the U.S. government's State Failure Task Force.[29] We use all data for all European nations since the fall of the Soviet Union ($n = 348$). State failure includes the collapse of the authority of the central government to impose law and order, such as occurs during civil war or disruptive regime transitions. As in King and Zeng, the explanatory variables used to explain state failure include the size of the military relative to the size of the total population, population density, legislative effectiveness, democracy (coded into two indicator variables representing autocracy, and partial and full democracy), trade openness, and infant mortality.[30] The fraction of state failures in the data is 1.15 percent.

Our previous results, and those of the State Failure Task Force, indicate that infant mortality is a key indicator of state failure. The assumption is that keeping infant mortality low tends to be an important goal of all governments. Countries that are unable to meet this goal tend to be at greater risk of state failure, whether or not they are democratic, have low military populations, have high trade openness, and so on. We computed the relative risk of state failure from the usual logit model by changing infant mortality from 30 percent of the world median to the world median (or roughly one standard deviation below to one standard deviation above the infant mortality rate for cases with state failure in Europe since 1990). This change, holding other variables constant at their median, yields a relative risk of 33.4. Thus, according to the usual logit model, a nation with infant mortality at 30 percent of the world median level of infant mortality has 33.5 times higher probability of state failure than a nation at the world median. When corrected, our more accurate point estimate of relative risk drops to only 10.2. This estimate has smaller mean square error and less bias. The 90-percent confidence interval is wide but still lower than the standard logit point estimate, ranging from 5.0 to 29.1.

Concluding Remarks

We have discussed how to make the best use of *existing* rare-events data and how to improve data-collection efforts in the future. To improve existing rare-events data, we offered the intuition behind easy-to-use methods for replacing logistic regression technology to reduce bias and increase accuracy with little cost. To improve future data-collection efforts, more essential from the perspective of quantitative studies of international conflict, we offered alternative strategies for collecting data. These strategies will enable scholars to collect much more informative data on a far smaller sample of cases without losing much information needed to make inferences to the entire population of nations or dyads globally.

29. Esty et al. 1998a,b.
30. Ibid.

Analysts have spent considerable resources over the years amassing large and impressive data collections regarding international conflict, but qualitative researchers still rightly complain that these collections miss much of the substance of the problem, code only simplistic variables, and contain relatively little information despite their size. The methods we propose here enable researchers to code far better variables without increasing their costs. Perhaps this will help to narrow the rift in political science between quantitative and qualitative approaches to explaining international conflict.

When analysts use these alternative data-collection strategies, the population fraction of events will still be rare, and the resulting samples will often be fairly small. The methods we propose allow researchers to correct estimates in alternative sampling designs and in rare-events data. The use of both types of corrections will be synergistically more useful than either one alone.

References

Achen, Christopher A. 1999. Retrospective Sampling in International Relations. Paper presented at the 57th Annual Meeting of the Midwest Political Science Association, Chicago.

Beck, Nathaniel, Gary King, and Langche Zeng. 2001. Improving Quantitative Studies of International Conflict: A Conjecture. *American Political Science Review* 94 (1):21–35.

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

———. 1998b. Theories of Conflict Initiation and Escalation: Comparative Testing, 1816–1980. Paper prepared for 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 (433):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, Conn.: Yale University Press.

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

Cosslett, Stephen R. 1981. Efficient Estimation of Discrete-Choice Models. In *Structural Analysis of Discrete Data with Econometric Applications,* edited by Charles F. Manski and Daniel McFadden, 51–111. Cambridge, Mass.: MIT Press.

Esty, Daniel C., Jack Goldstone, Ted Robert Gurr, Barbara Harff, Marc Levy, Geoffrey D. Dabelko, Pamela T. Surko, and Alan N. Unger. 1998. *The State Failure Task Force Report: Phase II Findings.* McLean, Va.: Science Applications International Corporation.

Esty, Daniel C., Jack Goldstone, Ted Robert Gurr, Barbara Harff, Pamela T. Surko, Alan N. Unger, and Robert S. Chen. 1998. The State Failure Project: Early Warning Research for U.S. Foreign Policy Planning. In *Preventive Measures: Building Risk Assessment and Crisis Early Warning Systems,* edited by John L. Davies and Ted Robert Gurr, 27–38. Lanham, Md.: Rowman and Littlefield.

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

Holland, Paul W., and Donald B. Rubin. 1988. Causal Inference in Retrospective Studies. *Evaluation Review* 12 (3):203–31.

Huth, Paul K. 1988. Extended Deterrence and the Outbreak of War. *American Political Science Review* 82 (2):423–43.

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):651–62.

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. 2000. Estimating Risk and Rate Levels, Ratios, and Differences in Case-Control Data. Available at ⟨http://gking.harvard.edu⟩.

———. Forthcoming. Improving Forecasts of State Failure. *World Politics.*

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

King, Gary, Michael Tomz, and Jason Wittenberg. Forthcoming. Making the Most of Statistical Analyses: Improving Interpretation and Presentation. *American Journal of Political Science.*

Lancaster, Tony, and Guido Imbens. 1996a. Case-Control with Contaminated Controls. *Journal of Econometrics* 71 (1–2):145–60.

Levy, Jack S. 1989. The Causes of War: A Review of Theories and Evidence. In *Behavior, Society, and Nuclear War,* vol. 1, edited by Phillip E. Tetlock, Jo L. Husbands, Robert Jervis, Paul C. Stern, and Charles Tilly, 2120–333. New York: Oxford University Press.

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

Manski, Charles F., and Steven R. Lerman. 1977. The Estimation of Choice Probabilities from Choice-based Samples. *Econometrica* 45 (8):1977–88.

Maoz, Zeev, and Bruce Russett. 1993. Normative and Structural Causes of Democratic Peace, 1946–86. *American Political Science Review* 87 (3):624–38.

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 (8):769–75.

Prentice, R. L., and R. Pyke. 1979. Logistic Disease Incidence Models and Case-control Studies. *Biometrika* 66 (3):403–11.

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.

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

Signorino, Curtis S. 1999. Strategic Interaction and the Statistical Analysis of International Conflict. *American Political Science Review* 93 (2):279–87.

Signorino, Curtis S., and Jeffrey M. Ritter. 1999. Tau-b or Not Tau-b: Measuring the Similarity of Foreign Policy Positions. *International Studies Quarterly* 43 (1):115–44.

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

———. 1999. BTSCS: A Binary Time-Series—Cross-Section Data Analysis Utility. Version 3.0.4. Available at ⟨http://www.fas.harvard.edu/~rtucker/programs/btscs/btscs.html⟩.

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