

# A Solution to Separation in Binary Response Models

Christopher Zorn

*Law and Social Science Program, National Science Foundation,  
4201 Wilson Boulevard, Suite 995, Arlington, VA 22230  
e-mail: czorn@nsf.gov*

A common problem in models for dichotomous dependent variables is “separation,” which occurs when one or more of a model’s covariates perfectly predict some binary outcome. Separation raises a particularly difficult set of issues, often forcing researchers to choose between omitting clearly important covariates and undertaking post-hoc data or estimation corrections. In this article I present a method for solving the separation problem, based on a penalized likelihood correction to the standard binomial GLM score function. I then apply this method to data from an important study on the postwar fate of leaders.

## 1 Introduction

The use of generalized linear models (GLMs) for the quantitative analysis of social science data has increased appreciably in the past four decades. This is particularly true of the familiar logit and probit models for dichotomous dependent variables (cf. Aldrich and Nelson 1984), the application of which has grown to become the de facto standard means for estimating regression-like models of binary outcomes.<sup>1</sup> Such models offer a number of advantages over ordinary least squares approaches, and their inclusion in standard software packages is effectively universal.

A common issue in the use of such models for binary responses is that of *separation*: the presence of one or more covariates that perfectly predict the outcome of interest. The simplest example is a  $2 \times 2$  table of  $Y$  and  $X$  with an “empty cell.” As nearly any quantitative analyst can attest, in models where the response variable of interest is dichotomous, separation is a particularly thorny problem, for a host of reasons. From an estimation perspective, separation leads to infinite coefficients and standard errors; perhaps even more important, there is a wide variation in how commonly used statistical software packages address the separation issue. Substantively, separation often forces researchers to make difficult, consequential, and largely arbitrary choices about data, measurement, and

---

*Author’s note:* Thanks to Hein Goemans for making his data available, and to Giacomo Chiozza, Kristian Gleditsch, Georg Heinze, Alastair Smith, and the anonymous reviewers at *Political Analysis* for helpful comments. The usual caveat applies. An earlier version of this paper was presented at the Conference on Bringing Leaders Back into International Relations, School of International Service, American University, April 23–24, 2004, Washington, DC; my thanks to Carmela Lutmar for inviting me. Replication materials are available on the *Political Analysis* Web site.

<sup>1</sup>For example, Gill (2002, p. 212) notes that the likelihood for the binary-response GLM has been “described in every econometric book ever printed (only a slight exaggeration).”

model specification. Moreover, separation occurs most frequently in those circumstances (e.g., small datasets, observational data, and so forth) in which its effects are likely to be especially pernicious.

But perhaps the most challenging aspect of the separation problem is its evasiveness. In fact, the very nature of the problem renders an assessment of its extent almost impossible. Because the heretofore standard social science approaches to dealing with separation involve model specification, it is likely that few authors are willing to acknowledge that separation had factored into the choices of covariates in their analyses. Instead, far more frequently, authors encountering separation in their data modify their models without alerting readers to the fact. In this sense, separation is, in all likelihood, far more widespread than its presence in published (or unpublished) work suggests, a phenomenon analogous to the “file drawer problem” in meta-analysis (cf. Iyengar and Greenhouse 1988).

In this article, I outline the nature of the separation problem, including its technical contours, the difficult choices it forces researchers to make, and its most commonly recommended remedies. I go on to present a recently developed method to solve the separation problem, one rooted in a more general “penalized-likelihood” approach to eliminating small- $N$  bias in maximum likelihood estimates. After discussing this solution, I apply the approach to data from a widely cited study on the postwar fate of leaders and demonstrate how this solution obviates the need for choosing between data manipulation and specification bias.

## 2 The Problem

Consider the archetypical logistic<sup>2</sup> regression model for a binary dependent variable  $Y_i$ ,  $i \in 1, 2, \dots, N$  and a vector of  $k$  covariates  $\mathbf{X}_i$  with corresponding  $k \times 1$  coefficient vector  $\beta$ :

$$Pr(Y_i = 1 \mid \mathbf{X}_i, \beta) \equiv \pi_i = \frac{1}{1 + \exp(-\mathbf{X}_i\beta)}.$$

The log-likelihood for this model is straightforward:

$$\ln L(\beta \mid Y) = \sum_{i=1}^N \left\{ Y_i \ln \left[ \frac{1}{1 + \exp(-\mathbf{X}_i\beta)} \right] + (1 - Y_i) \ln \left[ 1 - \frac{1}{1 + \exp(-\mathbf{X}_i\beta)} \right] \right\}$$

as is the score function:

$$\frac{\partial \ln L(\beta \mid Y)}{\partial \beta} \equiv \mathbf{U}(\beta) = \sum_{i=1}^N \left[ Y_i - \frac{1}{1 + \exp(-\mathbf{X}_i\beta)} \right] \mathbf{X}_i = \mathbf{0} \quad (1)$$

(e.g., McCullagh and Nelder 1989, chap. 4). Equation (1) can be solved for  $\hat{\beta}$  using standard iterative methods (e.g., the method of scoring or Newton-Raphson); inference can then be accomplished by considering the diagonal elements of the inverse of the standard information matrix, evaluated at  $\hat{\beta}$ :

<sup>2</sup>The intuition presented here applies to probit and nearly all other binary-response GLMs as well; for expositional clarity, I will focus on logit for the balance of the paper.

$$\widehat{\mathbf{Var}}(\hat{\beta}) \equiv -[\mathbf{I}(\beta)]^{-1} = -\left\{E\left[\frac{\partial^2 \ln \mathbf{L}(\beta|Y)}{\partial \beta \partial \beta'}\right]\right\}^{-1}. \quad (2)$$

In the context of models for binary outcomes, separation occurs when one or more of a model's covariates perfectly predict the outcome variable  $Y$ . The term is due to Albert and Anderson (1984), who differentiate between “complete” and “quasicomplete” separation, the latter denoting the case in which such perfect prediction occurs only for a subset of observations in the data (see also Lesaffre and Albert 1989). Formally, separation implies the existence of a subvector  $\mathbf{X}_s \subseteq \mathbf{X}$  by which all  $N$  observations can be correctly categorized as either  $Y_i = 0$  or  $Y_i = 1$ .<sup>3</sup> The presence of this subvector causes monotonicity in the (log-)likelihood and results in maximum likelihood estimates  $\hat{\beta}_s$  for the variables in  $\mathbf{X}_s$  that equal positive or negative infinity, and associated standard error estimates that are infinite as well.

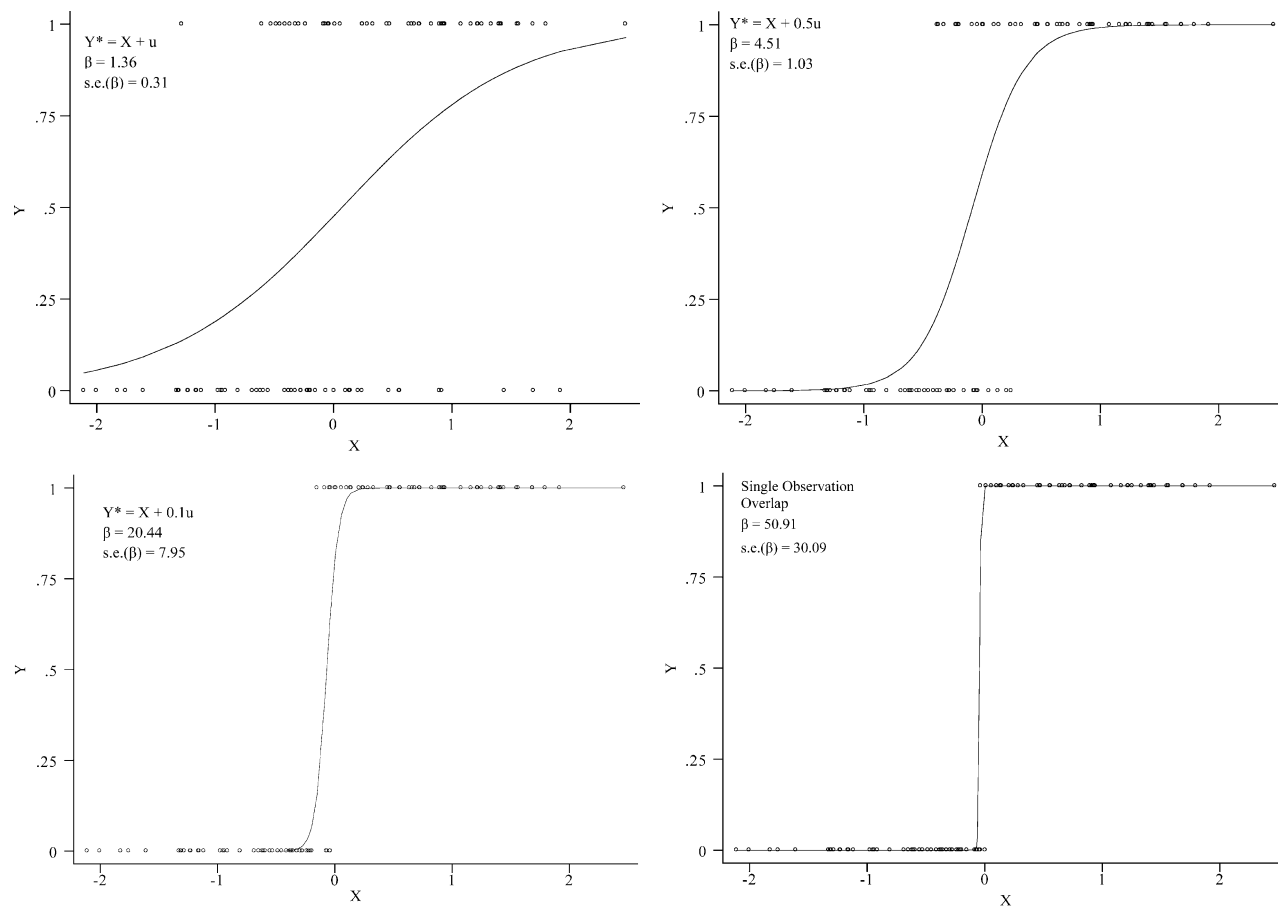
To understand intuitively the relationship between the phenomenon of separation and the resulting infinite parameter estimates, consider the obverse of separation: what Albert and Anderson (1984) refer to as “overlap.” Figure 1 presents results of bivariate logistic regressions on four simulated datasets, each with  $N = 100$ . In the first three panes, the data take the form:

$$\begin{aligned} X_i &\sim i.i.d. N(0, 1) \\ Y_i^* &= X_i + \alpha u_i, \\ u_i &\sim i.i.d. N(0, 1) \\ Y_i &= \begin{cases} 0 & \text{for } Y_i^* \leq 0 \\ 1 & \text{for } Y_i^* > 0 \end{cases} \end{aligned}$$

with  $\alpha = 1.0, 0.5$ , and  $0.1$ , respectively. In the fourth pane, data on  $Y$  were created such that only a single observation “overlapped”; specifically, the observation with  $X_i < 0$  closest to  $X = 0$  had  $Y = 1$ , while the observation with  $X_i > 0$  closest to  $X = 0$  had  $Y = 0$ . The resulting logistic regressions illustrate what occurs as separation grows more severe. On the one hand, coefficient estimates become larger, as the ability to correctly predict  $Y$  on the basis of the values of  $X$  grows stronger. At the same time, however, the estimates of their standard errors also increase. Intuitively, the latter result is due to the fact that the likelihood is almost completely “flat” in the region of the parameter estimate; substantial changes in  $\hat{\beta}$  yield only small differences in  $\ln \mathbf{L}$ , with the result that the diagonal elements of the information matrix are very large. As noted above, in the limit, complete separation corresponds to the ability to predict  $Y$  perfectly from  $X$  and yields likelihoods that are invariant (flat) and parameter estimates and standard errors that are, in theory, infinite in size.

In the case of a binary covariate  $X_s$ , the intuition is similar: situations in which  $X$  and  $Y$  do not overlap correspond to “empty cells” in the implied  $2 \times 2$  table formed by the two variables. *Complete* separation corresponds to the case in which only the two opposing diagonal cells of the table contain data; in such circumstances,  $Y$  can be perfectly predicted by  $X_s$  for all the observations in the data (that is, where all  $X_s = 0$  correspond to  $Y = 0$  and  $X_s = 1$  to  $Y = 1$ , or vice versa). This leads to two results. First, there is no variance left to be explained in  $Y$  by the model's other covariates, so the corresponding parameter

<sup>3</sup>An alternative interpretation is given by Agresti (2002, p. 195), who notes that, when this is the case, “a hyperplane can pass through the space of predictor values such that on one side of that hyperplane  $Y = 0$  for all observations, while on the other side  $Y = 1$  always.”



**Fig. 1** Actual and predicted values, simulated logistic regressions.

estimates for the remaining covariates will be zero. Second, because the likelihood is flat, the diagonal elements of (2) will be infinite in size, thus yielding infinite standard error estimates as in the continuous data example above.

In contrast, *quasicomplete* separation occurs when only one cell of the implied  $2 \times 2$  table of  $X_s$  and  $Y$  is “empty.” Under such conditions, the parameter estimate for the separating variable  $X_s$  (and its standard errors) will also be infinite in size, but the model’s other covariates may remain relatively unaffected. As we will see below, such cases are far more common than those of complete separation in the observational data commonly used by political scientists.<sup>4</sup> At the same time, it is important to note that in both such circumstances standard statistical software packages will often fail to alert the analyst to the presence of the problem. In fact, as we will see below, the actual values for  $\beta$  obtained in the presence of complete or quasicomplete separation are almost completely a function of the researcher’s (typically) arbitrary choice of convergence criteria for the estimation routine.

### 3 Dealing with Separation

The crux of the problem, then, is summarized by Heinze and Schemper (2002, 2409): “In general, one does not assume infinite parameter values in underlying populations.”<sup>5</sup> The problem of separation has long been recognized in studies of generalized linear models; for example, both Wedderburn’s (1976) and Silvapulle’s (1981) existence results require the absence of complete separation. Somewhat more recently, Albert and Anderson (1984) present a general discussion and typology of separation, along with a discussion of possible strategies for its detection and amelioration. At the same time, the issue of separation has received only marginal attention in the vast majority of widely used texts on the subject of models for limited dependent variables. Neither Maddala (1983) nor Long (1997) raises the issue of separation in any form. Greene (2003, §21.4.6) addresses the issue only in passing, in his discussion of log-linear models for contingency tables. Similarly, McCullagh and Nelder mention separation in the context of estimation, noting only that “failure to converge is rarely a problem unless one or more components of  $\beta$  are infinite, which usually implies that some of the fitted probabilities are either zero or one”; they go on to note that “abnormal convergence means that the log likelihood is either very flat or, more likely, has an asymptote” (1989, p. 117).<sup>6</sup>

As a practical matter, separation forces the analyst to choose from a number of problematic alternatives for dealing with the problem. The most widely used “solution” is simply to omit the offending variable or variables from the analysis. In political science, this is the approach taken in a number of studies in international relations (e.g., Krain 1997, Table 3; Gibler and Vasquez 1998, Table 4; Reiter 1999, pp. 382–383), comparative politics (e.g., Wibbels 2000, Table 3), and American politics (Peterson and Wrighton 1998, Table 2). It is also the dominant approach in sociology (e.g., Hill 2000, Table 6; Rotolo 2000, p. 1145), economics (Eisenberg et al. 1997, Table 4; Jianakoplos and Menchik 1997, Table 6; Cameron 2000, n. 27), and the other social sciences, and it is the recommended method in a few prominent texts in statistics and econometrics (e.g., Davidson and MacKinnon 1993, p. 521). Of course, this alternative is a particularly

<sup>4</sup>Moreover, as discussed by Lesaffre and Albert (1989), separation can also occur as the result of a linear combination of a model’s covariates.

<sup>5</sup>This section draws on Heinze and Schemper (2002), who in turn borrow substantially from Firth (1993).

<sup>6</sup>Arguably the best textbook discussion of the problem is that in Venables and Ripley (2002, pp. 198–199), though even that treatment stops short of offering a clear remedy for the problem.

unattractive one; omitting a covariate that clearly bears a strong relationship to the phenomenon of interest is nothing more than deliberate specification bias.<sup>7</sup>

A second alternative is to modify the data in order to eliminate the separation. Clogg et al. (1991) suggest an approach whereby the researcher supplements the  $N$  observations in the data with additional “artificial” data across the various patterns of (categorical) covariates, and then conducts the analysis in the usual fashion on the resulting data (see Clogg et al. 1991 for details). In addition to the ad hoc nature of this solution, both Heinze and Schemper (2002, pp. 2413–2414) and Galindo-Garre et al. (2005) demonstrate conclusively via Monte Carlo simulations that Clogg et al.’s approach is inferior to other available alternatives; accordingly, I do not discuss it any further here.

Yet another alternative is the use of exact logistic regression. First suggested by Cox (1970), exact logistic regression is based on the same idea as exact inference in  $2 \times 2$  contingency tables. Specifically, inference is based on “exact permutational distributions of the sufficient statistics that correspond to the parameters of interest, conditional on fixing the sufficient statistics of the remaining parameters at their observed values” (King and Ryan 2002, p. 164; see also Mehta and Patel 1995; Collett 2002, chap. 9). Importantly, exact logistic regression estimates are obtainable even in the presence of empty cells and complete separation. But while this approach is thus more attractive than the others discussed so far, it remains problematic. In particular, because the conditional distributions of sufficient statistics requires summing over discrete patterns of covariate values, relatively sparse data and/or small numbers of observations in particular patterns of categorical covariates often lead to degenerate estimates, and the inclusion of continuous covariates nearly always does so. Accordingly, this makes the exact method less attractive for researchers who typically combine continuous and categorical explanatory variables in their analyses. In addition, this suggests that the exact approach is most likely to break down in precisely those conditions in which separation is most common—that is, when  $N$ s are small and/or data are sparse.

In the face of these unappealing alternatives, it is interesting to note that commercial software makers take widely varying approaches to the issue of separation. At one extreme, some software packages (e.g., Stata) are aggressively proactive, automatically omitting variables and dropping observations from the analysis when quasicomplete separation is present and simply failing to provide any estimate at all when separation is complete. At the other extreme (characterized by S-Plus/R), the software estimates parameters as usual and leaves it to the researcher to detect the presence of separation through an examination of the estimated coefficients and standard errors. In this latter case, the actual estimates obtained depend strongly upon the convergence criteria chosen by the analyst.<sup>8</sup>

To illustrate these different approaches and the effect they can have on one’s results, consider the data presented in Table 1, which tabulates the political party affiliation of the 24 justices appointed to the U.S. Supreme Court since World War II and the political party of their appointing president. Not surprisingly, there is a strong relationship; no Democratic president since World War II has appointed a Republican justice, while only two Democratic justices (William Brennan and Lewis Powell) were appointed by

<sup>7</sup>Another alternative is to collect additional data, in the hope that doing so will eliminate the problem. While this is in many respects a good solution, it is not always feasible, particularly in observational or other nonexperimental contexts.

<sup>8</sup>A middle-ground approach, taken by SAS and some others, is to calculate the (theoretically infinite) estimates but warn the researcher that separation is present; S-Plus and R also issue warnings in the case of complete separation, but not when separation is quasicomplete.

**Table 1** U.S. Supreme Court appointments, 1946–2004

<i>Justice's Political Party</i>	<i>Appointing President's Political Party</i>		
	<i>Democratic</i>	<i>Republican</i>	<i>Total</i>
Democratic	9	2	11
Republican	0	13	13
Total	9	15	24

*Note.* Pearson  $\chi^2_1 = 17.02$  ( $p < .001$ ).

Republican chief executives. Were one interested in modeling the political party affiliation of high court appointees, the corresponding affiliation of his or her appointing president would thus be a strong candidate as an explanatory variable.

Table 2 presents results of estimating a logistic regression of the effect of presidential party identification on justice identification for the data in Table 1, using two commonly used software packages. For the analyses conducted using S-Plus, results corresponding to the selection of four different convergence criteria ( $\tau = 10^{-3}$ ,  $10^{-5}$ ,  $10^{-7}$ , and  $10^{-9}$ ) are shown; these values correspond to the minimum change in the log-likelihood necessary for convergence to be considered as having been achieved. Several patterns are apparent. First, as noted above, Stata omits the offending variable and drops the nine cases in which there is no variation on  $Y$ . In doing so, the software reports a message indicating that “ $X \neq 1$  predicts failure perfectly;  $X$  dropped and 9 observations not used.” By contrast, S-Plus retains the *Republican President* variable, but the estimates for its influence on the appointee’s party identification, as well as those for its standard error, increase steadily as the convergence tolerance is decreased. Moreover, the reported significance of  $\hat{\beta}$  decreases steadily as the tolerance decreases, despite the fact that the variable is clearly influential; this is consistent with Heinze and Schemper’s (2002) assertion that simply assigning a high value to  $\hat{\beta}$  is also an insufficient solution to the separation problem.

#### 4 A Penalized Likelihood Approach to Separation

In an important paper, Firth (1993) suggests a method for eliminating the well-known small-sample bias in maximum likelihood estimation. The intuition of Firth’s approach is to introduce a bias term into the standard likelihood function that itself goes to zero as  $N \rightarrow \infty$  but that for small  $N$  operates to counteract the  $O(N^{-1})$  bias present there. The result is a penalized likelihood:

$$\mathbf{L}(\beta | Y)^* = \mathbf{L}(\beta | Y) |\mathbf{I}(\beta)|^{\frac{1}{2}} \quad (3)$$

with corresponding log-likelihood:

$$\ln \mathbf{L}(\beta | Y)^* = \ln \mathbf{L}(\beta | Y) + 0.5 \ln |\mathbf{I}(\beta)| \quad (4)$$

and score equation:

$$\mathbf{U}(\beta)^* = \mathbf{U}(\beta) + 0.5 \operatorname{tr} \left\{ \mathbf{I}(\beta)^{-1} \left[ \frac{\partial \mathbf{I}(\beta)}{\partial \beta} \right] \right\} \quad (5)$$

**Table 2** Stata and S-Plus results, logistic regression

<i>Variable</i>	<i>Stata</i>	<i>S-Plus</i> , $\tau = 10^{-3}$	<i>S-Plus</i> , $\tau = 10^{-5}$	<i>S-Plus</i> , $\tau = 10^{-7}$	<i>S-Plus</i> , $\tau = 10^{-9}$
(Constant)	1.872 (0.760)	−8.202 (12.219)	−13.203 (148.83)	−17.203 (1099.75)	−22.203 (13397.66)
Republican president	(dropped)	10.074 (12.243)	15.075 (148.84)	19.075 (1099.75)	24.075 (13397.66)
<i>N</i>	15	24	24	24	24
Iterations to convergence	0	7	12	16	21

*Note.* Response variable is one for GOP appointees, zero for Democrats.  $\tau$  indicates convergence criterion (change in log-likelihood) for each model.

where  $\mathbf{I}(\beta)$  is again the information matrix in (2), evaluated at  $\hat{\beta}$ . Firth (1993) demonstrates that, for a broad class of generalized linear models, this penalized likelihood is both asymptotically consistent and eliminates the usual small-sample bias found in MLEs, and that penalized-likelihood estimates exist in numerous situations in which standard likelihood-based estimates do not, including that of complete separation in binary-response models. Finally, Firth notes that, for exponential-family GLMs with canonical link functions, a Bayesian interpretation of this correction is the application of Jeffreys' (1946) invariant prior.<sup>9</sup>

Firth outlines a range of applications of his penalized-likelihood approach to generalized linear models. In the context of binary logistic regression, the information matrix  $\mathbf{I}(\beta)$  is equal to:

$$\mathbf{I}(\beta) = \mathbf{X}'\mathbf{W}\mathbf{X}$$

where:

$$\mathbf{W} = \text{diag} \left\{ \left[ \frac{1}{1 + \exp(-\mathbf{X}_i\beta)} \right] \left[ 1 - \frac{1}{1 + \exp(-\mathbf{X}_i\beta)} \right] \right\} \equiv \text{diag}[\pi_i(1 - \pi_i)].$$

This gives the penalized-likelihood correction in (5) a particularly simple form. More specifically, in the logit context, the Firth correction amounts to a modification of the standard score equation (1) to:

$$\mathbf{U}(\beta | Y)_{\ell}^* = \sum_{i=1}^N (Y_i - \pi_i) X_i \left( 1 + \frac{h_i}{2} \right) + \sum_{i=1}^N (1 - Y_i - \pi_i) X_i \left( \frac{h_i}{2} \right) = \mathbf{0} \quad (6)$$

where the  $h_i$  are the diagonal elements of the penalized-likelihood version of the standard “hat” matrix  $\mathbf{H}$ :

$$\mathbf{H} = \mathbf{W}^{\frac{1}{2}} \mathbf{X} (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W}^{\frac{1}{2}}$$

<sup>9</sup>A full exposition of the Jeffreys prior is beyond the scope of this article; however, the use of Jeffreys' prior for Bayesian estimation of GLMs has been widely recommended in the statistics literature (cf. Kass 1989; Ibrahim and Laud 1991). In the context of logistic regression, good discussions of Jeffreys' prior include Poirier (1994) and Brown et al. (2001).



(cf. McCullagh and Nelder 1989, Eq. 12.3). As with (1), estimation of  $\hat{\beta}$  in (6) can be accomplished through application of standard Newton–Raphson or quasi–Newton methods to the modified score function, and standard mechanisms for obtaining standard error estimates (e.g., as the square roots of the diagonal elements of  $-\mathbf{I}(\hat{\beta})^{-1}$ ) are also available. Firth (1993, p. 30) notes that the effect of small–sample bias in logistic regression is to bias the estimates  $\hat{\beta}$  away from zero; concomitantly, the second term in (4) is maximized at  $\pi_i = 0.5$ —that is, when  $\beta = 0$ —and therefore the correction shrinks the estimate  $\hat{\beta}$  back toward this value. Because they are shrunk toward zero, penalized–likelihood estimates will typically be smaller in absolute value than standard MLEs, though their standard errors will also be reduced, yielding similar inferences about the significance of parameter estimates for those parameters whose MLE is finite.

In the case of a binary logit model with a single dichotomous covariate, the penalized–likelihood correction has an especially simple interpretation. There, Firth’s approach corresponds to adding 0.5 to each cell of the implied  $2 \times 2$  table, an approach that has long been advocated as a solution to the empty cell inference problem in such tables (e.g., Haldane 1956; Blalock 1979). More generally, the penalized–likelihood approach can be thought of as “splitting each original observation  $i$  into two new observations having response values  $Y_i$  and  $1 - Y_i$  with iteratively updated weights  $1 + h_i/2$  and  $h_i/2$ , respectively” (Heinze and Schemper 2002, p. 2412). Doing so ensures that parameter estimates exist and are nondegenerate for all  $\mathbf{X}$ ; thus, a valuable attribute of the penalized–likelihood approach is its ability to yield consistent parameter estimates in the presence of complete or quasicomplete separation. At the same time, Firth’s consistency results ensure that as  $N \rightarrow \infty$  the penalized–likelihood estimates converge to the MLEs under the usual regularity conditions.

Evidence to date strongly suggests that the penalized–likelihood approach to estimating logit models in the presence of separation is uniformly superior to its alternatives. For example, in a recent study, Galindo–Garre et al. (2005) demonstrate via Monte Carlo simulations that the Firth/Jeffreys approach is superior both to the ad hoc correction of Clogg et al. (1991) and to Bayesian approaches that adopt other uninformative conjugate (Dirichlet) or normal priors, particularly in models that are not fully saturated factorial designs and/or in small samples.

But while penalized–likelihood estimates have a number of attractive properties, it is also important to bear in mind that the resulting penalized profile likelihoods for the coefficients are often asymmetrical; intuitively, this is because the estimates themselves are close to boundary conditions. The practical implication of this is that the usual inferences based on Wald–type statistics can be misleading. Heinze and Schemper (2002) recommend use of penalized–likelihood–ratio tests in lieu of standard Wald tests, together with visual examination of each parameter’s profile likelihood (that is, the shape of the  $[\log-]$ likelihood in the region of the maximum  $\hat{\beta}$ ; cf. McCullagh and Nelder 1989, §7.2.4). It is well known (e.g., Barndorff–Nielsen and Cox 1994) that profile likelihood–based confidence intervals tend to provide more accurate coverage than their asymptotic (normal–based) analog in small samples and/or when parameters are influenced by boundary conditions; Heinze and Schemper (2002) demonstrate that this is often the case in Firth–corrected logistic regression models as well, particularly vis-à-vis coefficient estimates for variables in which separation is present.

In summary, Firth’s penalized–likelihood approach offers an especially attractive alternative to applied researchers faced with the problem of separation in logit, probit, and other binary–response GLMs. In particular, Firth’s method prevents researchers from being forced either to omit manifestly important covariates from their models or to engage

in post-hoc data manipulation in order to obtain parameter estimates for those covariates. Most attractively, Firth's approach is available in a range of commonly used packages for social science data analysis and so is easily implemented using standard software.<sup>10</sup> In the section below, I demonstrate the application of this method to an existing study in which the presence of separation caused the analyst to omit a key independent variable.

## 5 An Example: The Fate of Leaders

In an influential study, Goemans (2000) examines the fate of leaders following major international crises. Among other things, he argues that leaders in "mixed" regimes will be more likely to be punished (that is, exiled, imprisoned, or killed) following even relatively minor military losses, but that only "disastrous" losses will result in punishment to either democrats or dictators. Goemans examines this hypothesis using data on the postwar fates of 204 leaders between 1816 and 1975; his primary variable of interest is whether ( $= 1$ ) or not ( $= 0$ ) a leader was punished (exiled, imprisoned, or killed) following the war. He models this outcome as a function of five covariates, four of which are a combination of two factors: regime type ("mixed" or non-mixed, the latter including both democratic or autocratic states) and the magnitude of the loss ("small" or "big," with winners as the omitted reference category).<sup>11</sup>

Goemans's fifth covariate is whether ( $= 1$ ) or not ( $= 0$ ) the leader in question faced a postwar regime change imposed by one or more foreign powers. He notes that for a host of reasons leaders who are overthrown by foreign powers are particularly likely to face punishment; in fact, all 22 of the leaders in his data subjected to such a regime change were exiled, imprisoned, or killed, while only 27 (or 14.8 percent) of the other 182 leaders in his data faced such punishment. The data thus present a clear example of quasicomplete separation, with the result that it is impossible to estimate the influence of this important covariate on the quantity of interest using conventional means. Goemans's solution is twofold: first, he estimates a model with the *Foreign-Imposed Regime Change* variable omitted; second, he substitutes an alternative coding of the variable of interest,<sup>12</sup> which did not result in separation being an issue. The first option is clearly suboptimal, since it raises the strong possibility of specification bias. Likewise, his second solution, while not an unreasonable approach, is not a general one, since such alternative measurements of important quantities of interest are often neither so readily available nor so easily justifiable.

Here I reanalyze Goemans's analysis, first examining the results when the offending variable is included (with a resulting infinite coefficient estimate) and then applying Firth's penalized-likelihood approach to the data. Table 3 presents the results of maximum-likelihood (MLE) and maximum penalized-likelihood (MPLE) estimates of the effects of regime types, war losses, and foreign-imposed regime change on the postwar punishment of leaders.<sup>13</sup> Not surprisingly, standard logistic regression shows the marginal influence of *Foreign-Imposed Regime Change* to be effectively infinite; in particular, those results suggest that the log-odds of punishment for such leaders are roughly 8,400,000,000 times greater than those who are removed without foreign intervention. At the same time, the

<sup>10</sup>For a review of software available for fitting Firth's penalized-likelihood logistic regression model, see the appendix.

<sup>11</sup>See Goemans (2000, pp. 564–565) for details of his data selection and coding rules.

<sup>12</sup>Specifically, the operationalization used in Werner (1996), in which a single leader subject to a foreign-imposed regime change was not coded as having been punished.

<sup>13</sup>Data and R commands to replicate these analyses are available on the *Political Analysis* Web site. Note that I had no difficulty replicating the results in Goemans (2000, Table 4), though for brevity I omit those results here.

**Table 3.** MLE and MPLE estimates of postwar leader punishment

Variable	MLEs		MPLEs	
	$\hat{\beta}$	Odds Ratio	$\hat{\beta}$	Odds Ratio
(Constant)	-2.955 (0.459)	—	-2.865 (0.438)	—
Other small loser	0.851 (0.659)	2.34	0.845 (0.629)	2.33
Other big loser	3.360 (1.022)	28.8	3.198 (1.003)	24.5
Mixed-regime small loser	2.693 (0.622)	14.8	2.614 (0.607)	13.7
Mixed-regime big loser	3.243 (0.891)	25.6	3.115 (0.877)	22.5
Foreign-imposed regime change	22.852 (4840.2)	$8.40 \times 10^9$	5.493 (1.507)	243.0

Note.  $N = 204$ . Response variable is one for punished leaders, zero if not. Standard errors are in parentheses.

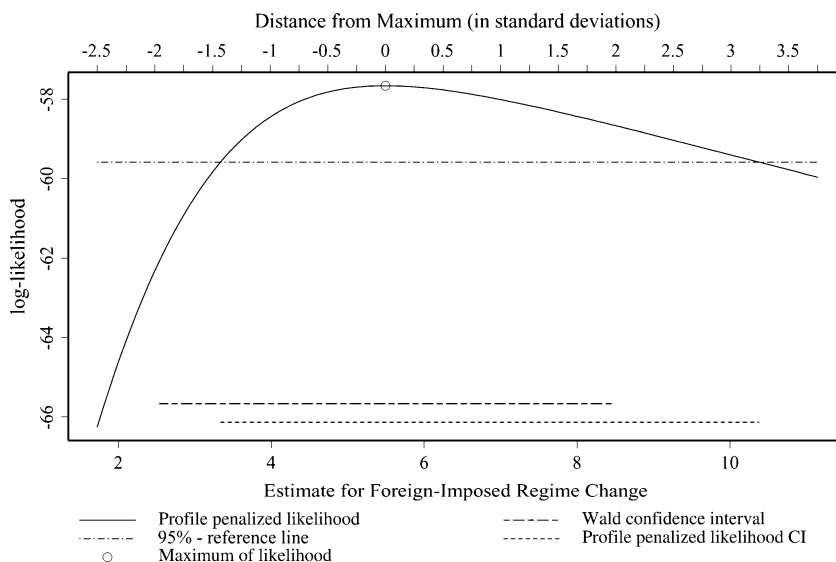
estimated coefficient is far smaller than its standard error, and the bounds on the confidence intervals for the estimated odds ratio exceed the precision of the software.

The corresponding penalized-likelihood results are provided in columns three and four of Table 3 and present a far more credible picture of the influence of foreign-imposed regime change on postwar leaders' fates. The coefficient estimate, while large, is nonetheless reasonable, and the corresponding odds ratio is also well within the bounds of plausibility. Likewise, the estimated coefficient is substantially larger than its estimated (profile penalized likelihood-based) standard error, as we would expect.

As noted above, Heinze and Schemper (2002) suggest that analysts adopting Firth's approach examine the profile likelihood in the region of the maximum, in order to determine the accuracy of Wald-type test statistics. Accordingly, Figure 2 plots the profile penalized likelihood for the effect of *Foreign-Imposed Regime Change* across a range of possible  $\beta$ s, along with a 95 percent confidence line. The lower lines in the graph represent confidence intervals, with the longer-dashed line denoting that based on asymptotic (Wald) statistics and the shorter-dashed line denoting the profile penalized-likelihood interval.

Figure 2 reveals several important phenomena. First, as one would expect, the profile likelihood is asymmetric; specifically, it is right-skewed, with a correspondingly wider range of plausible values to the right of the point estimate  $\hat{\beta}$  than to the left. This is unsurprising; given the nature of the data (that is, given the separation of  $Y_i$  on this covariate and the resulting  $\beta = \infty$ ) we would expect that the "true" value of  $\beta$  would be more likely to be higher than its estimate than lower.<sup>14</sup> Conversely, the Wald-based interval is symmetrical around the point estimate, demonstrating that such asymptotic-based statistics are likely to be inaccurate in the presence of separation. Figure 2 also shows that the profile penalized-likelihood confidence interval is somewhat wider than its asymptotic counterpart, suggesting that the possibility of Type I error is decreased by the former approach.

<sup>14</sup>Note that similar plots for the other four covariates show no such asymmetry, a fact consistent with the supposition of Heinze and Schemper (2002) that it is the covariates responsible for the separation that are most likely to yield misleading Wald-based confidence intervals. Those plots are available from the author upon request.



**Fig. 2** Profile likelihood for *Foreign-Imposed Regime Change*, by  $\hat{\beta}$ .

## 6 Conclusion

Many, if not most, quantitative social scientists have, at one time or another, encountered separation and empty cells in the course of their research. Separation is an especially vexing issue, arising most often in situations in which data are sparse and strong relationships are present. Moreover, to date methods for ameliorating separation have been singularly unappealing: neither omission of key covariates from explanatory models nor ex-post data manipulation ought to rank highly on the list of approaches for dealing with such issues.

The penalized-likelihood method proposed by Firth (1993) provides a simple, valid, easy-to-implement solution to the separation problem. It involves neither arbitrary data manipulation nor complicated modifications to otherwise standard models, it does not alter the interpretation of those models in any way, and it is available in a host of existing software packages. It has a straightforward Bayesian justification as Jeffreys' invariant prior for the binomial logit model. Perhaps best of all, Firth's approach is asymptotically equivalent to (optimal) maximum-likelihood methods in large samples and is superior to them in small samples—precisely the situations in which separation is most likely to be a concern. Thus, while standard logistic regression should and will remain a central part of any applied social scientist's quantitative tool kit, Firth's penalized-likelihood model is an exceptionally attractive alternative when such standard models break down.

At the same time, two caveats are in order. First, a number of authors have urged caution in the use of Jeffreys' prior, particularly in instances in which the model in question has large numbers of nuisance parameters (e.g., Ibrahim and Laud 1991). Poirier (1994), for example, notes that Jeffreys' prior should not be used for conditional logit models, due to that model's inclusion of large numbers of nuisance parameters.<sup>15</sup>

<sup>15</sup>More generally, the use of Jeffreys' prior is appropriate only in those circumstances in which the asymptotic bias present is of order  $O(N^{-1})$ ; in situations in which the information per parameter fails to grow at a uniform rate, this may or may not be the case. I thank an anonymous reviewer for pointing this out.

Second, as noted above, researchers should take care in making inferences when adopting the penalized-likelihood approach. In particular, analysts should take care to examine the profile likelihood in the neighborhood of the maximum for estimates where separation is an issue and to report profile likelihood-based confidence intervals and significance tests when that likelihood is asymmetric. Failure to do so could result in inaccurate inferences about model parameters.

#### Appendix: Software for Estimating Firth's (1993) Penalized-Likelihood Logistic Regression

Several routines for estimating the penalized-likelihood model described here have been written for a number of widely used software packages. Firth's own *brlr* package, written for R, shares much of its syntax with the widely used *glm* package for the same platform. Additionally, Georg Heinze and Meinhard Ploner (2003, 2004) have authored two packages for estimating these models. Their *logistf* library operates under both S-Plus and R, and was used to estimate the results presented here. That library also includes two additional functions that are useful for diagnostic purposes: *logistf*test performs penalized-likelihood-ratio tests on subsets of model coefficients, while *logistfplot* automatically generates plots of the profile penalized likelihoods of the sort shown in Figure 2. In addition, Heinze and Ploner have also authored the SAS macro *FL*, which will estimate penalized-likelihood logistic regression models in that package as well.

#### References

- Agresti, Alan. 2002. *Categorical Data Analysis*, 2nd ed. New York: Wiley.
- Albert, A., and J. A. Anderson. 1984. "On the Existence of Maximum Likelihood Estimates in Logistic Regression Models." *Biometrika* 71(1):1-10.
- Aldrich, John H., and Forrest D. Nelson. 1984. *Linear Probability, Logit, and Probit Models*. Newbury Park, CA: Sage.
- Barndorff-Nielsen, O. E., and David R. Cox. 1994. *Inference and Asymptotics*. London: Chapman and Hall.
- Blalock, Hubert M. 1979. *Social Statistics*, rev. 2nd ed. New York: McGraw-Hill.
- Brown, Lawrence D., T. Tony Cai, and Anirban DasGupta. 2001. "Interval Estimation for a Binomial Proportion." *Statistical Science* 16(2):101-133.
- Cameron, Lisa J. 2000. "Limiting Buyer Discretion: Effects on Performance and Price in Long-Term Contracts." *American Economic Review* 90(March):265-281.
- Clogg, Clifford C., D. B. Rubin, N. Schenker, B. Schultz, and L. Weidman. 1991. "Multiple Imputation of Industry and Occupation Codes in Census Public-Use Samples Using Bayesian Logistic Regression." *Journal of the American Statistical Association* 86:68-78.
- Collett, Dave. 2002. *Modeling Binary Data*, 2nd ed. London: Chapman and Hall.
- Cox, David R. 1970. *Analysis of Binary Data*. New York: Wiley.
- Davidson, R., and J. G. MacKinnon. 1993. *Estimation and Inference in Econometrics*. New York: Oxford University Press.
- Eisenberg, Theodore, John Goerdt, Brian Ostrom, David Rottman, and Martin T. Wells. 1997. "The Predictability of Punitive Damages." *Journal of Legal Studies* 26(June):623-661.
- Firth, David. 1993. "Bias Reduction of Maximum Likelihood Estimates." *Biometrika* 80(1):27-38.
- Galindo-Garre, Francisca, Jeroen K. Vermunt, and Wicher P. Bergsma. 2005. "Bayesian Posterior Estimation of Logit Parameters with Small Samples." *Sociological Methods and Research*: forthcoming.
- Gibler, Douglas M., and John R. Vasquez. 1998. "Uncovering the Dangerous Alliances, 1495-1980." *International Studies Quarterly* 42(December):785-807.
- Gill, Jeff. 2002. *Bayesian Methods: A Social and Behavioral Sciences Approach*. Boca Raton, FL: Chapman & Hall.
- Goemans, Hein. 2000. "Fighting for Survival: The Fate of Leaders and the Duration of War." *Journal of Conflict Resolution* 44(October):555-579.
- Greene, William H. 2003. *Econometric Analysis*, 5th ed. Upper Saddle River, NJ: Prentice-Hall.
- Haldane, John Burden Sanderson. 1956. "The Estimation and Significance of the Logarithm of a Ratio of Frequencies." *Annals of Human Genetics* 20(July):309-311.
- Heinze, Georg, and Meinhard Ploner. 2003. "Fixing the Nonconvergence Bug in Logistic Regression with SPLUS and SAS." *Computer Methods and Programs in Biomedicine* 71:181-187.

- Heinze, Georg, and Meinhard Ploner. 2004. "Technical Report 2/2004: A SAS Macro, S-Plus Library and R Package to Perform Logistic Regression Without Convergence Problems." Section of Clinical Biometrics, Department of Medical Computer Sciences, Medical University of Vienna, Vienna.
- Heinze, Georg, and Michael Schemper. 2003. "A Solution to the Problem of Separation in Logistic Regression." *Statistics in Medicine* 21(21):2409–2419.
- Hill, Mark E. 2000. "Color Differences in the Socioeconomic Status of African American Men: Results of a Longitudinal Study." *Social Forces* 78(June):1437–1460.
- Ibrahim, Joseph G., and Purushottam W. Laud. 1991. "On Bayesian Analysis of General Linear Models Using Jeffreys' Prior." *Journal of the American Statistical Association* 86:981–986.
- Iyengar, S., and J. B. Greenhouse. 1988. "Selection Models and the File–Drawer Problem." *Statistical Science* 3:109–135.
- Jacobsen, J. 1989. "Existence and Unicity of MLEs in Discrete Exponential Family Distributions." *Scandinavian Journal of Statistics* 16:335–349.
- Jeffreys, H. 1946. "An Invariant Form of the Prior Probability in Estimation Problems." *Proceedings of the Royal Society of London, Series A* 186:453–461.
- Jianakoplos, Nancy A., and Paul L. Menchik. 1997. "Wealth Mobility." *Review of Economics and Statistics* 79(February):18–31.
- Kass, R. E. 1989. "The Geometry of Asymptotic Inference." *Statistical Science* 4(August):188–219.
- King, Elizabeth N., and Thomas P. Ryan. 2002. "A Preliminary Investigation of Maximum Likelihood Logistic Regression versus Exact Logistic Regression." *American Statistician* 56(August):163–170.
- Krain, Matthew. 1997. "State–Sponsored Mass Murder: The Onset and Severity of Genocides and Politicides." *Journal of Conflict Resolution* 41(June):331–360.
- Lesaffre, E., and A. Albert. 1989. "Partial Separation in Logistic Discrimination." *Journal of the Royal Statistical Society, Series B* 51:109–116.
- Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.
- McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.
- Mehta, C. R., and R. Patel. 1995. "Exact Logistic Regression: Theory and Examples." *Statistics in Medicine* 14:2143–2160.
- Peterson, Geoff, and J. Mark Wrighton. 2000. "Expressions of Distrust: Third–Party Voting and Cynicism in Government." *Political Behavior* 20(March):17–34.
- Poirier, Dale J. 1994. "Jeffreys' Prior for Logit Models." *Journal of Econometrics* 63(August):327–339.
- Reiter, Dan. 1999. "Military Strategy and the Outbreak of International Conflict: Quantitative Empirical Tests, 1903–1992." *Journal of Conflict Resolution* 43(June):366–387.
- Rotolo, Thomas. 2000. "A Time to Join, A Time to Quit: The Influence of Life Cycle Transitions on Voluntary Association Membership." *Social Forces* 78(March):1133–1161.
- Silvapulle, M. J. 1981. "On the Existence of Maximum Likelihood Estimates for the Binomial Response Models." *Journal of the Royal Statistical Society, Series B* 43:310–313.
- Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*, 4th ed. New York: Springer.
- Wedderburn, R. W. M. 1976. "On the Existence and Uniqueness of the Maximum Likelihood Estimates for Certain Generalized Linear Models." *Biometrika* 63(April):27–32.
- Werner, Suzanne. 1996. "Absolute and Limited War: The Possibilities of a Foreign Imposed Regime Change." *International Interactions* 22(1):67–88.
- Wibbels, Erik. 2000. "Federalism and the Politics of Macroeconomic Policy and Performance." *American Journal of Political Science* 44(October):687–702.