British Journal of Political Science (2020), page 1 of 18
doi:10.1017/S0007123420000071

British Journal of Political Science

ARTICLE

Addressing Monotone Likelihood in Duration Modelling of Political Events

Noel Anderson[1]*, Benjamin E. Bagozzi[2] and Ore Koren[3]

[1]Department of Political Science, University of Toronto, Canada, [2]Deparment of Political Science and International Relations, University of Delaware, USA and [3]Department of Political Science, Indiana University, Bloomington, IN, USA
*Corresponding author. E-mail: noel.anderson@utoronto.ca

(Received 1 May 2019; revised 30 October 2019; accepted 27 January 2020)

Abstract

This article provides an accessible introduction to the phenomenon of monotone likelihood in duration modeling of political events. Monotone likelihood arises when covariate values are monotonic when ordered according to failure time, causing parameter estimates to diverge toward infinity. Within political science duration model applications, this problem leads to misinterpretation, model misspecification and omitted variable biases, among other issues. Using a combination of mathematical exposition, Monte Carlo simulations and empirical applications, this article illustrates the advantages of Firth's penalized maximum-likelihood estimation in resolving the methodological complications underlying monotone likelihood. The results identify the conditions under which monotone likelihood is most acute and provide guidance for political scientists applying duration modeling techniques in their empirical research.

Keywords: duration models; monotone likelihood; penalized maximum likelihood; external intervention; state partition

Duration models have become ubiquitous in the political science literature – and for good reason. Many topics in empirical and policy research have at their heart questions about the time it takes until an event occurs. How long will a war last? When will a regime stabilize or transition? When will the tenure of a political office terminate? Duration models, which generate estimates of the conditional probability of an event occurring over time, are ideal for studying these questions (see Box-Steffensmeier and Jones 2004). For that reason, recent studies have applied duration modeling strategies to topics as diverse as regime duration (Gates et al. 2006; Svolik 2008), the length of inter- and intra-state wars (Balch-Lindsay and Enterline 2000; Bennett and Stam 1996; Cunningham 2011), treaty ratification (McKibben and Western 2020), legislative position taking (Box-Steffensmeier, Arnold and Zorn 1997; Kropko and Harden 2020), the fate of political leaders (Licht 2017; Omgba 2009), and the effectiveness of international peacekeeping missions (Fortna 2008), among others.

As duration modeling has become more popular, researchers have developed new tools to facilitate and improve their application in empirical political science research. Recent work has contributed to substantive improvements in the simulation of duration data (Harden and Kropko 2019), the prediction of duration dynamics (Chiba, Metternich and Ward 2015) and the interpretation of results (Jones and Metzger 2019; Kropko and Harden 2020; Licht 2011). In this article, we add to this growing body of literature by highlighting a methodological challenge that has to date escaped scrutiny in the political science literature on duration models – monotone likelihood.

Monotone likelihood occurs when covariate values are monotonic when ordered according to failure time. In practice, this most often occurs when a value of a covariate, or a linear

combination of covariates, uniquely correspond(s) to all termination events. At face value, this might imply a substantively important relationship between variables, but from a methodological perspective it introduces mathematical complications that can lead to infinite coefficient estimates and standard errors for a particular sample. This is problematic precisely because, as Heinze and Schemper (2001, 114) aptly put it, '[i]n general, one does not assume infinite parameter values in underlying populations'. Covariates, in other words, must covary. Yet in practice, monotone likelihood generates such arbitrarily large (or small) parameter estimates that meaningful and reliable substantive interpretation of results becomes impossible. For example, in empirical applications detailed below, we demonstrate that monotone likelihood can generate effect sizes that imply durations that are thousands, millions, and even billions of times longer (or shorter).

To address these issues, this article provides an accessible introduction to the phenomenon of monotone likelihood in duration modeling, with applications in political science research. We explain the nature of the problem and describe an easy-to-apply corrective known as Firth's penalized maximum-likelihood estimation. While political scientists have discussed this correction in logit and multinomial logit settings (Cook, Hays and Franzese 2020; Cook, Niehaus and Zuhlke 2018; Rainey 2016; Rainey and McKaskey, Forthcoming; Zorn 2005), it has not – to our knowledge – been applied to duration modeling within political science research. This is a notable omission, as the problems we highlight are especially acute in duration frameworks, where researchers frequently model time-varying data that often multiplies by several factors the number of observations for each unit (including the number of censored observations), and which makes these zero events subject to serial correlations (Box-Steffensmeier and Jones 2004, 95–118). Using simulation techniques as well as applications drawn from the literature, we illustrate the advantages of Firth's penalized maximum-likelihood estimation in this context and provide guidance for researchers who encounter monotone likelihood in their duration modeling applications.

The article proceeds as follows. The following section provides a general statement of the problem and identifies its consequences. The subsequent section overviews Firth's penalized maximum-likelihood estimation and explains how it corrects for monotone likelihood. Next, the results of a large number of Monte Carlo experiments are reported that (1) evaluate the causes of monotone-likelihood issues, (2) assess the potential for penalized maximum-likelihood estimation to address them and (3) compare the performance of a standard Cox proportional hazard model to that of a Cox proportional hazard model with Firth's penalized maximum-likelihood correction. In the penultimate section, the empirical implications of monotone likelihood – and the ameliorative properties of penalized maximum likelihood – are demonstrated with applications drawn from the existing literature. The final section concludes by summarizing our main findings and identifying fruitful paths for future research.

Statement Of The Problem

In duration modeling, monotone likelihood occurs when 'at each failure time, the covariate value of the failed individual is the largest of all covariate values in the risk set at that time or when it is always the smallest' (Heinze and Schemper 2001, 144; also see Tsiatis 1981). Put differently, monotone likelihood is a property of samples where covariate values are monotonic when ordered according to failure time. Table 1 provides a stylized example. Here, X_1 and X_2 are monotonic continuous covariates: they take on values that are always the largest (X_2) or always the smallest (X_1) of all covariate values in the risk sets $R(t_f) = \{t_1, t_2, t_3, t_4, t_5\}$.[1] X_3 provides another

[1]To fix these ideas, it is helpful to walk through the case of the monotonic covariate X_2, which takes values that are always the largest of all covariate values in the risk set. Assuming all cases terminate during the period of observation, there are five individuals included in the risk set $R(t_f)$ at the start of the observation period. When the first individual approaches termination at time t_1, it is the first individual that is taking on the largest value of X_2 (8). After the first individual terminates, they

Table 1. Monotonicity in continuous and dichotomous predictors

$R(t_f)$	t_1	<	t_2	<	t_3	<	t_4	<	t_5	
X_1	1		2		3		4		5	Monotonic
X_2	8		3		1		0		−2	Monotonic
X_3	1		1		1		0		0	Monotonic
X_4	5		3		4		2		1	Non-monotonic
X_5	1		0		1		1		1	Non-monotonic

example of monotonicity, this time in a binary predictor variable, where the values $X_3 = 1$ for the failed individuals at $R(t_f) = t_1, t_2, t_3$ are the largest values of X_3 in the ordered risk set. X_4 and X_5 then represent non-monotonic ordering of failure times for continuous and binary covariates, respectively.

Because the predictors X_1, X_2 and X_3 are monotonic, the (partial) likelihood of one's estimated duration model will be monotone for the coefficient estimates associated with these predictors. Consequently, while the likelihood function will converge to a finite value, estimates of β_1, β_2, and β_3 will diverge to positive or negative infinity. For an estimator that does not account for this sample property, the net result will be biased, model-dependent estimates with arbitrarily large (or small) parameter values.[2]

In practice, monotone likelihood most often occurs when a value of a covariate, or a linear combination of covariates, uniquely correspond(s) to all termination events – a phenomenon akin to 'separation' in other binary response models. In their pioneering work on this issue, Albert and Anderson (1984) distinguish between two types of separation: 'complete' and 'quasi-complete'.[3] Translated into a duration modeling framework, complete separation refers to instances in which the values of one or more covariates uniquely correspond to all termination *and* survival events;[4] quasi-complete separation refers to instances in which one value of a covariate uniquely corresponds to all termination events, but not to all survival events.[5] In either case, monotone likelihood arises due to the absence of overlap of failure times between two or more groups. To see this intuitively, consider the case of a quasi-completely separating dichotomous predictor X_D, where all values $X_D = 0$ correspond to the full set of termination events. In such cases, there is no overlap of failure times in the two groups $X_D = 0$ and $X_D = 1$, as there are no termination events associated with the value $X_D = 1$. Were a researcher to order X_D's values according to failure time, the resulting vector would be monotonic (in this case, never increasing, as X_D never varies from 0 when a termination event occurs).

Figure 1 provides a visual representation of the problem using simulated duration data. In the left panel, we plot predicted survival curves from a Cox model that incorporates a single dichotomous covariate, X_a, that is non-monotonic (that is, there is overlap of failure times in the groups $X_a = 1$ and $X_a = 0$). In this panel, the predicted proportion of cases that have not yet terminated gradually declines as a function of X_a and time. In the right panel, we plot predicted survival

are removed from the risk set, leaving only four individuals remaining. In turn, when the second individual approaches termination at time t_2, it is the second individual who is now taking on the largest value of X_2 (3) among all four individuals who remain in the risk set. After the second individual terminates, they too are removed from the risk set, leaving only three individuals. This process continues until all individuals have terminated. In line with the Heinze and Schemper (2001, 144) quotation above, at each failure time, the value of X_2 for the failed individual is the largest of all values of X_2 in the risk set at that time. Thus, when we order X_2 according to failure time, the resulting vector is monotonic.

[2]Specifically, the parameter estimates returned will correspond to the set of extreme, but finite, values for which a model's log likelihood changes by less than a given software program's assigned tolerance threshold (Heinze and Schemper 2001, 114).

[3]See also Zorn (2005), Rainey (2016).

[4]I.e., where all $X_D = 0$ correspond to all termination events *and* all $X_D = 1$ correspond to all survival events, or vice versa.

[5]I.e., where all $X_D = 0$ correspond to all termination events, but not all $X_D = 1$ correspond to all survival events, or where all $X_D = 1$ correspond to all termination events, but not all $X_D = 0$ correspond to all survival events.

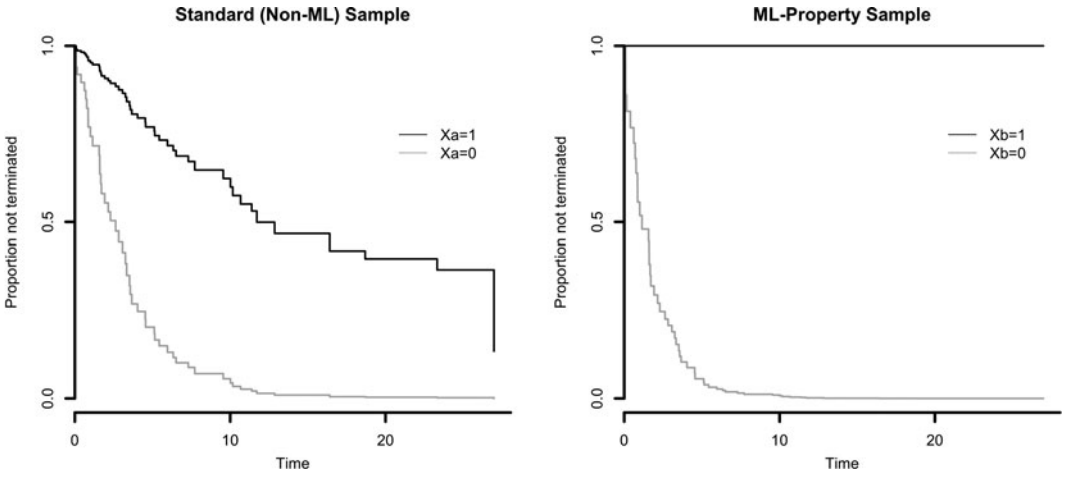

Figure 1. Survival curves for simulated Cox regressions

curves from a Cox model that incorporates a single dichotomous covariate, X_b, that is quasi-completely separating (and thus, monotonic). In this case, all $X_b = 0$ correspond to the full set of termination events, meaning that there are no termination events associated with the value $X_b = 1$, and thus there is no overlap of failure times in the two groups $X_b = 0$ and $X_b = 1$. This causes the survival curve for $X_b = 1$ to become horizontal in orientation. The corresponding Cox model estimate for X_b will diverge to negative infinity.

Importantly, the absence of a quasi-completely separating covariate does not eliminate the risk of monotone likelihood, for two reasons. First, covariates do not need to be quasi-completely separating to be monotonic when ordered according to failure time. Secondly, even when no single covariate is monotonic, a *linear combination* of covariates can be. As above, this leads to a flattening of the likelihood function and an inflation of parameter values towards infinity.

At its root, then, monotone likelihood is the result of highly imbalanced and thus highly predictive covariates. This problem can be decomposed into a number of contributing factors (Bryson and Johnson 1981; Heinze and Schemper 2001; Johnson et al. 1982; Loughin 1998):

- *Dichotomous predictors*. Models that rely on a large number of dichotomous predictors are more likely to be afflicted by monotone likelihood. The categorical nature of these variables restricts their range to just two values, thereby rendering monotonicity more likely. While monotone likelihood may occur with any type of data, it is less common in the case of continuous independent variables.
- *Number of covariates*. As the number of included covariates increases, the probability that the likelihood function will be monotone in at least one of the regression parameters increases for two reasons. First, there are more opportunities for at least one covariate to be monotonic when ordered according to failure time. Secondly, even where no single covariate is monotone, a function of the covariates can be.
- *Small sample sizes*. As sample size decreases, the probability that at least one covariate is monotonic when ordered according to failure time increases as a function of small sample bias.
- *Proportion of censored observations*. Increased censoring generates imbalance by increasing the number of zeros (censored observations) to ones (terminations) on the termination indicator variable. Datasets that incorporate time-varying variables are especially susceptible to this concern owing to their counting process data structure, which breaks up individual cases into multiple observation periods.

The various statistical programs and packages commonly used in the political science literature differ in how they deal with monotone likelihood. Some packages will detect nonconvergence, warn the user and fail to estimate parameter values;[6] others will detect nonconvergence and issue a warning, yet will still estimate parameter values even for those variables afflicted by monotone likelihood;[7] and still others will fail to detect nonconvergence and estimate all parameter values in the model.[8] The parameter estimates returned in these instances will typically correspond to the set of extreme, but finite, values for which changes in a model's log-likelihood fall below a given tolerance threshold, which itself is variable across software and estimation routines.[9] Consequently, in cases of monotone likelihood, a researcher's results will vary as a function of the software program and packages employed.

Faced with estimates that imply fitted probabilities that converge to extreme positive or negative values, researchers are often forced to choose among a range of distasteful remedies. Heinze and Schemper (2001, 114–115), for instance, review some of these issues. First, some researchers simply drop the offending variable from their analysis, introducing *omitted variable bias*. Secondly, researchers may choose to change their modeling strategy, risking *model misspecification* in light of censored or truncated data. Thirdly, researchers may stratify on the offending variable, allowing the hazard to vary by each category of the problematic variable, but consequently *forgoing estimating its own effect* on duration. Fourthly, biased estimates can result in *misinterpretation of substantive effects*, or indeed in a decision to avoid interpreting the results altogether. Fifthly, and perhaps most insidious of all, a failure to understand the underlying causes of monotone likelihood can contribute to the *file drawer problem*, where researchers simply do not publish or report models that seem otherwise incorrect, hard to understand, or that fail to support a theory's proposed hypotheses.

Notwithstanding these distasteful remedies, monotone likelihood remains a common problem in political science research. To evaluate the extent to which it affects the existing literature, we conducted a survey of all articles that employed a duration model between the years 2008–2018 and that appeared in one of three leading political science journals: the *American Political Science Review*, the *American Journal of Political Science* and the *Journal of Politics*.[10] Of the fifty-seven articles we identified, 14 per cent contained strong evidence of monotone likelihood. This represents a substantial share of contemporary duration model research across three of the discipline's top journals over the past decade. What is more, we found that monotone likelihood can be identified across the empirical subfields of political science – from American politics to comparative politics to international relations – and affects cross-national studies and subnational analyses alike. And importantly, this estimate is likely a conservative one in light of the file drawer problem noted above, whereby many studies afflicted by monotone likelihood are never published. Thus, our survey suggests that monotone likelihood regularly confronts researchers employing duration models in their empirical research. In what follows, we introduce a corrective for this methodological challenge and illustrate its advantages using a combination of mathematical exposition, Monte Carlo simulations and empirical applications.

[6]The flexsurv package in R, for example, will warn users of a failure in the optimization routine and will not estimate a model under monotone likelihood.

[7]The coxph package in R, for example, will often (though not always) detect nonconvergence and issue a warning. Parameter values are still estimated for the model.

[8]The streg function in Stata, for example, will often estimate model parameters without detecting nonconvergence and without warning users.

[9]For example, in Stata, tolerance is set to 1×10^{-4} (when an estimator is programmed via ml) or 1×10^{-6}, whereas in R most relevant packages use 1×10^{-9} as the tolerance threshold.

[10]Details of the survey are reported in full in the Appendix.

A SOLUTION TO MONOTONE LIKELIHOOD

In recent years correctives have been developed for monotone likelihood – namely, the application of penalized maximum-likelihood estimation approaches to Cox regression models. In particular, a procedure developed by Firth (1993) to reduce the bias of maximum-likelihood estimates under monotone-likelihood conditions has been shown to solve the problem in a range of Cox regression applications (Heinze and Dunkler 2008; Heinze and Schemper 2001), as well as for estimators designed for other limited dependent variables (Cook, Niehaus and Zuhlke 2018; Rainey 2016; Zorn 2005). With the application of penalized maximum-likelihood estimation, finite parameter estimates of constant and time-dependent effects can be obtained even in the presence of monotone likelihood.

In formal terms, monotone likelihood can generate bias in any estimated parameters (denoted generally as $\hat{\theta}$) when the score function at the true $\hat{\theta}$ is unbiased (that is, $E[U(\theta)] = 0$) but its curvature is not linear in θ (that is, the rate of change of the function $U(\theta)^{''} \neq 0$). Given that $E[U(\theta)] = 0$, any corresponding duration model's estimation function[11] that is linear in θ will converge to a finite value; this is precisely what the researcher wants when estimating a (generalized) linear model. However, because the function's *curvature* is not linear in θ, some parameter values will be severely biased upward (if $U(\theta)^{''} > 0$) or downward (if $U(\theta)^{''} < 0$).

Consequently, when estimating θ under conditions of monotone likelihood, the researcher introduces positive or negative bias into the score function. This arises even if the function is otherwise unbiased with respect to identification. And notably, because the rate of change of the score function's curvature is itself not linear, such bias will be increasing to infinity. This, in turn, can lead to at least one parameter estimate converging to $-\infty/+\infty$ and a model that arbitrarily over- or under-estimates the true covariate effects (Firth 1993, 28; Heinze and Schemper 2001, 114).

To ameliorate the impact of monotone likelihood in affected samples, one can introduce suitable *corrective bias* into the biased (partial) likelihood estimate $\hat{\theta}$ (Firth 1993). In contrast to other plausible correctives – such as removing the offending variable or arbitrarily trimming down the coefficient size – this approach involves using the function's (that is, $U(\theta)$) own estimates to identify an *informed* correction that is *proportional* to the size of the uncorrected estimator's bias under monotone-likelihood conditions. Mathematically, this correction is expressed as follows. First, assume a standard Cox partial log-likelihood function:

$$\ln L(\beta) = \sum_{i=1}^{m} \left\{ X_i\beta - \ln \left[\sum_{h \in R_j} \exp(X_h\beta) \right] \right\} \tag{1}$$

where R_j denotes all observations at risk at time t_j, i denotes the N distinct event times, X_i denotes the covariates for observations that experienced a termination event at t_j, X_h denotes the covariates for all observations at risk, and β encompasses the associated coefficients (see Box-Steffensmeier and Jones 2004, 52). Now, recall that under monotone-likelihood conditions, our partial-likelihood estimate $\hat{\theta}$ – or more specifically, our β parameters for the offending variable(s) therein, which we denote as β_r – is biased. The purpose of the *corrective bias* approach is to provide an informed correction that is proportional to the size of the bias in the estimator caused under monotone-likelihood conditions. Following Firth (1993), we term the new corrective function $U(\beta_r)^*$, and express it as:

$$U(\beta_r)^* \equiv U(\beta_r) + \alpha_r = 0 \tag{2}$$

[11] E.g., the Cox partial log-likelihood (see Equation 1).

where $U(\beta_r)$ is the *uncorrected* partial likelihood-estimated covariate, and α_r is:

$$\alpha_r = 0.5\ trace\left\{\mathbf{I}(\beta)^{-1}\left[\frac{\partial \mathbf{I}(\beta)}{\partial \beta_r}\right]\right\} \tag{3}$$

In Equation 3, $\mathbf{I}(\beta)^{-1}$ is the inverse information matrix evaluated at each β, or the variance–covariance matrix for $\hat{\beta}$. The term in brackets $[\cdot]$ is the derivative of the information matrix with respect to β_r: $(\partial/\partial\beta_r)\{0.5\ log|i(\beta)|\}$. Accordingly, α_r can be estimated using the Newton–Raphson algorithm or other standard (partial) likelihood optimization routines, where each step is based on the current value of $\hat{\beta}$.[12] In simple terms, this correction 'pushes against' the original (uncorrected estimator) bias. Thus, if $U(\beta_r)$ has a positive bias, the score function will be shifted downward by α_r at each point; if $U(\beta_r)$ has a negative bias, the score function will be shifted upward.

Modification of the score function in this manner directly addresses the aforementioned challenge of arbitrarily large (or small) parameter values in monotone-likelihood contexts through its assurance of finite parameter estimates (Heinze and Schemper 2001, 115). To see the latter point, note that because this informed correction builds on the uncorrected function's own variance–covariance matrix, it is *proportional to the size of the bias in the uncorrected estimator*. This quality provides an important advantage over uninformed corrections, which are likely to over- or under-estimate the bias in the estimation results under monotone-likelihood conditions. Specifically, as $\beta_r \to \pm\infty$, the highest/lowest observed value of a covariate x_k in each risk set gets *weighted more heavily* compared with other covariates, as long as there are at least k distinct failure times. This helps to ensure that even if the (log) likelihood L is monotone, the determinant of $\mathbf{I}(\beta)$ still approaches zero, which in turn means that the penalized maximum likelihood $L^{\star}$ is guaranteed to attain a finite, bias-corrected value of $\hat{\beta}$. Indeed, as first illustrated by Firth (1993) in a general context and by Heinze and Schemper (2001) in a duration model context, $L^{\star}$ is asymptotically consistent and is otherwise resistant to the small-sample biases and arbitrary parameter estimates that arise in (partial) likelihood estimation. In that regard, Firth's (1993) correction lowers estimation bias empirically, regardless of issues of substance (for example, 'what is an intervention?') or thresholds of 'plausibility/implausibility', which are inherently subjective to a particular application.

For the remainder of this article, we accordingly assess the benefits of penalized maximum likelihood in Cox regression applications – hereafter referred to as the 'Firth Cox' (Heinze and Dunkler 2008; Heinze and Schemper 2001). The Cox model is generally viewed as the default choice among contemporary political scientists for applied duration modeling (Box-Steffensmeier and Jones 2004; Ruhe 2018, 91), and can be reformulated and interpreted as an exponential family model (Heinze and Schemper 2001; McCullagh and Nelder 1989, 429). The latter is appealing in this case in that it ensures that the bias-reducing properties of the Firth correction – as originally established for generalized linear models (Firth 1993) – will carry over to the Cox context (Heinze and Schemper 2001, 115). This property has been empirically confirmed for the Firth Cox model by both Heinze and Schemper (2001, 117) and Heinze and Dunkler (2008, 6464), and thus serves as an additional justification for using the Firth Cox model in these contexts.

That being said, we note that monotone-likelihood issues can equally impact parametric duration models, such as the Weibull. Accordingly, in the ensuing sections we report empirical models based not only upon the standard Cox model and the Firth Cox model, but also upon the Weibull model. The latter application illustrates that the biases in non-corrected estimators under conditions of monotone likelihood in duration analysis are relevant to parametric and semi-parametric models alike.

[12] A Bayesian interpretation of this correction, which involves finding the mode of the posterior distribution, relies on using the Jeffreys (1946) invariant prior.

Monte Carlo Simulations

To evaluate the causes of monotone likelihood – and to assess the potential for penalized maximum-likelihood estimation to address monotone-likelihood issues – we conducted a large number of Monte Carlo (MC) experiments. These experiments compare the performance of a standard Cox proportional hazard model to that of a Cox proportional hazard model with Firth's penalized maximum-likelihood correction (Firth Cox). We expand upon and improve past simulation studies of monotone likelihood in three ways.

First, we examine the problem under a wider range of conditions: five varying levels of censoring,[13] at six different sample sizes,[14] across three different specification scenarios.[15] Importantly, our MC experiments are designed to ensure we do not directly modify the severity of monotone likelihood, but rather allow the severity of this problem to arise organically as a function of (imbalanced) independent variables that are themselves a product of varying sample sizes, changing levels of censoring and different rates of omitted variables.

Secondly, unlike previous studies, we simulate duration data to match the (nonparametric) Cox model's assumed data-generating process via the methods recently developed by Harden and Kropko (2019), including a mixture of continuous and imbalanced binary predictors.

Finally, we assess the performance of the Cox and Firth Cox models across our resulting ninety distinct combinations of varying conditions in terms of (1) the proportion of relevant simulations that saw nonconvergence due to infinite or near-infinite parameter estimates and (2) our parameter estimates' root mean squared errors (RMSEs), averaged across all simulations. When doing so, we maintain consistent iteration limits on the Cox and Firth Cox models during each simulation run, which provides us with conservative estimates of the latter model's abilities to overcome monotone-likelihood issues. Together, these MC experiments provide to our knowledge the most comprehensive assessment of the performance of Firth's correction to duration modeling to date.

Figure 2 summarizes the results of our MC experiments, reporting nonconvergence rates and RMSEs for the estimates associated with a simulated binary independent variable of interest, X_1.[16] Each column of subfigures depicts a different level of censoring; the x-axes vary the N evaluated; and the y-axes depict either the proportion of nonconvergence obtained across all relevant simulations (subfigure row 1) or the RMSEs for $\widehat{\beta_1}$ (subfigure row 2). We plot the relevant values obtained for our Cox (triangles) and Firth Cox (circles) models within each subfigure after averaging over the distinct levels of omitted variable bias evaluated. The latter condition was collapsed in Figure 2 for summary purposes; it is presented in disaggregated fashion in the Appendix. Note that the scale of nonconvergence and RMSEs across these different specifications increases markedly as the level of censoring increases, as denoted by the differing y-axis values on these plots.

We find that analyses afflicted by monotone likelihood will frequently encounter nonconvergence and inaccurate parameter estimates when using a standard Cox model, but not when using the Firth Cox model. This is especially the case for samples of 100–500 observations, regardless of the level of censoring. However, for levels of censoring at or greater than 75 per cent – common in duration model applications with time-varying covariates – the aforementioned threats to accuracy and convergence with the standard Cox model increase substantially, and can persist in samples as large as 1,000–2,000 observations. For instance, Figure 2's subfigure columns 4 and 5 (75–95 per cent censoring) demonstrate that, using samples of 100–500 observations, researchers will on average encounter nonconvergence rates of 40 per cent for the standard Cox model, compared to nonconvergence rates of 3 per cent for the Firth Cox model. Under

[13] I.e., the proportion of all duration cases that exhibit non-terminations within our period of observation – which we set to range across $c = \{0.05, 0.25, 0.5, 0.75, 0.95\}$.

[14] The number of observations, N, which we assign as $N = \{100, 250, 500, 1{,}000, 2{,}000, 5{,}000\}$.

[15] The degree of omitted variable bias, where we consider four, two and zero omitted variables.

[16] Our full MC results are reported in the Appendix.

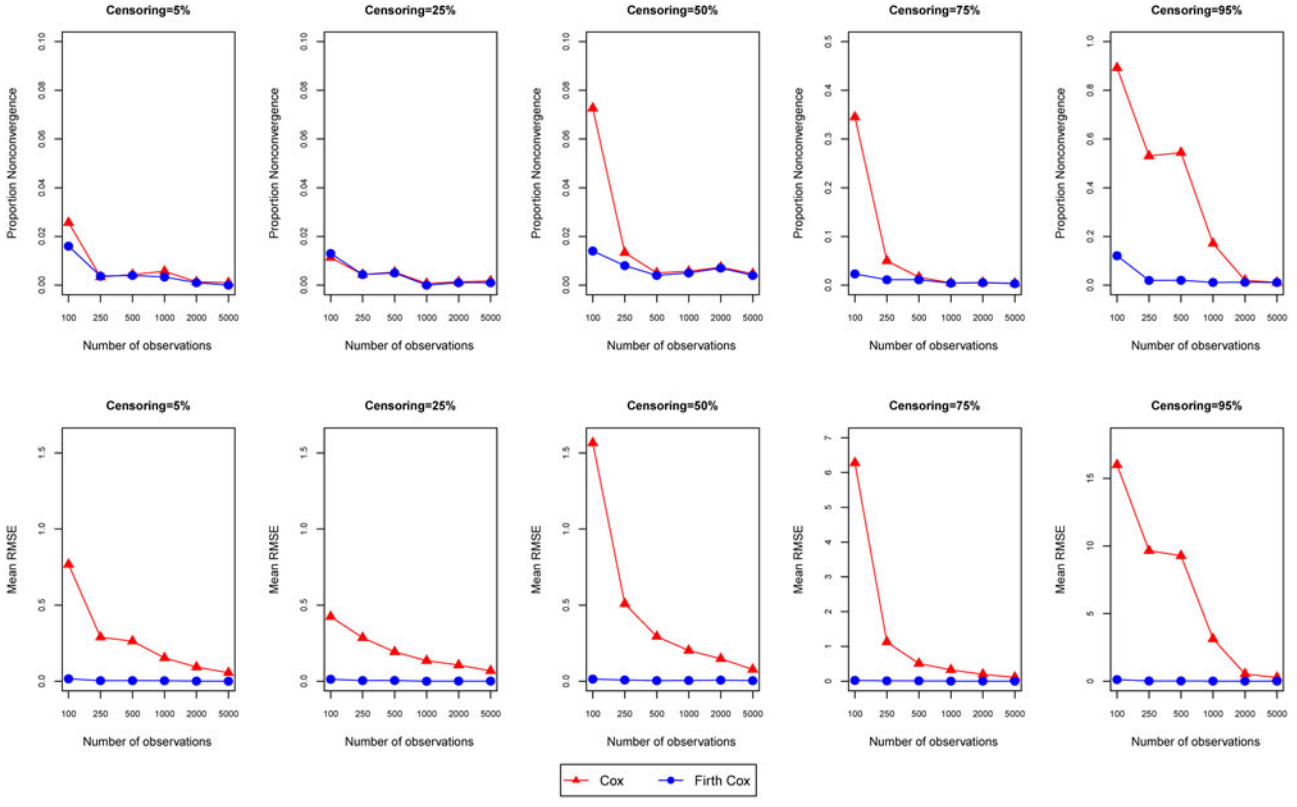

Figure 2. Nonconvergence and RMSEs for X_1 across all Monte Carlo experiments

these same conditions, the Cox model $\widehat{\beta}_1$ RMSEs are on average 250 times larger than those obtained using the Firth Cox model. At samples of 1,000–2,000 and 95 per cent censoring, the Firth Cox model continues to recover $\widehat{\beta}_1$s that are 165 times more accurate than those of the Cox model, with nonconvergence rates that are 8.4 percentage points lower than the Cox model. Our additional MC assessments (reported in the Appendix) reaffirm each of these findings.

In short, monotone likelihood arising from imbalanced predictors leads to inaccuracy and nonconvergence in parameter estimates for the Cox model when the number of observations is lower than 1,000. At levels of 75–95 per cent censoring, the Cox model becomes practically unusable for small-to-moderate sample sizes, exhibiting nonconvergence due to (near-)infinite parameter estimates in a substantial share of all such simulations. By contrast, the Firth Cox model is much less sensitive to these monotone-likelihood challenges, and in most instances recovers parameter estimates that are several orders of magnitude more accurate than those of the Cox model. Indeed, even in those cases where the Cox and Firth Cox models exhibit similar nonconvergence rates, we find that the latter *always* obtain lower RMSEs across all parameters and experimental conditions evaluated, and further confirm these findings for mean squared errors in the Appendix. In line with recent logit-based findings (Rainey and McKaskey Forthcoming), these results suggest that penalized maximum-likelihood corrections may be preferable to standard Cox estimation frameworks in regards to the accuracy of parameter estimates when using low-to-moderate sample sizes, irrespective of the level of censoring or the degree of imbalance in one's binary predictors.

Empirical Applications

In this section we demonstrate the empirical implications of monotone likelihood – and the ameliorative properties of Firth Cox models – with applications drawn from the existing literature.

Application 1: External Intervention and Conflict Duration

In a widely cited article, Regan (2002) examines the effect of external intervention on the duration of civil war. He develops a theory that highlights interveners' manipulation of domestic combatants' costs of fighting and expectations of victory, and hypothesizes that opposing interventions prolong the expected duration of a conflict. To test this proposition, he compiles a novel dataset of external intervention in intrastate conflicts fought between 1944 and 1999. The dataset includes 150 conflicts, of which 119 terminated during the period of observation.[17] Each conflict is measured at a monthly resolution and coded across a range of variables recording various characteristics of the conflict. A set of dichotomous indicators capture whether the conflict received military and/or economic interventions, and if so, whether those interventions supported opposing domestic combatant forces.[18] The resulting cross-national, time-series dataset contains monthly observations recorded for all time-varying variables (n = 13,048).

Table 2 reproduces the main findings of the article. Model 1 replicates the corresponding model from the study, which employs a Weibull duration model with an accelerated failure time (AFT) parameterization. Model 2 reports a modified version of the original specification, dropping a conflict intensity control. Model 3 employs a standard Cox model, and Model 4 adopts a Firth Cox approach.

[17]In the dataset, intrastate conflict is defined as 'armed combat between groups within state boundaries in which there are at least 200 fatalities'. See Regan (2000, 21).

[18]We refer readers to Regan (2002, 66–67) for a description of the included covariates. Note that opposing interventions are those that take place in an offsetting sequence and/or where multiple interventions support opposing actors; military interventions involve one of six categories (troops, naval support, equipment or aid, intelligence or advisors, air support or sanctions); and economic interventions involve the use of economic instruments to affect the balance of power between domestic combatants.

We successfully replicate the original study.[19] However, our results suggest that substantial nonconvergence issues afflict the estimated parameter values under both the Weibull AFT and standard Cox specifications. Indeed, Models 1–3 estimate extreme coefficient values for the variables recording opposing interventions, military interventions, economic interventions and biased interventions on the side of rebel forces (represented in the original study as the 'target' variable). The magnitude of the substantive effect of each of these variables is untenably large.

Consider first the Weibull AFT model estimates. Given the model's accelerated failure time parameterization, a positively signed coefficient implies longer expected durations. Substantive interpretation is facilitated by exponentiating the raw coefficients, which provides relative expected durations given a one-unit increase in x_i. Thus Model 1 estimates that conflicts that experience opposing interventions are over 83,000 times longer than those that do not. Likewise, conflicts that are subject to military interventions are estimated to be over 82,000 times longer, and those that experience economic interventions are estimated to be over 3,500,000,000 times longer. However, the expected duration of conflicts that receive biased interventions on the side of rebel forces is less than 0.0001 per cent that of conflicts that do not (that is, $e^{-10.986}$). All four variables are statistically significant at conventional levels ($p \leq 0.01$), but the magnitude of the substantive effects estimated by the Weibull AFT model strain credulity.

Similar results are identified in Model 2, which drops a conflict intensity control.[20] The magnitude of the substantive effects for opposing interventions, military interventions, economic interventions and biased interventions on the side of rebel forces are estimated to grow even larger in this model, with time ratios converging towards infinite values. These estimates suggest considerable bias owing to monotone-likelihood conditions, with parameter values that are more a function of mathematical complications than empirical evidence.

Model 3 demonstrates that the extreme estimates we identify are not unique to duration models that adopt a Weibull AFT specification. This model re-runs the previous analysis, employing a standard Cox model. Note that because Cox models are expressed in terms of a hazard rate, interpretation of estimated coefficients is opposite that of Weibull AFT models (that is, a positively signed coefficient implies *shorter* expected duration). Substantive interpretation is again facilitated by exponentiating the raw coefficients, which can then be interpreted as hazard ratios. To calculate the effect of a one-unit change in x_i, one subtracts 1 from the reported hazard ratio and multiplies by 100 to recover the percent change in the hazard of conflict termination. Thus, Model 3 estimates that opposing interventions, military interventions and economic interventions all decrease the hazard of conflict termination by over 99.999 per cent. That is, these variables are estimated to virtually eliminate the likelihood of conflict termination. Biased interventions on the side of rebel forces, however, are estimated to increase the hazard of conflict termination by over 4,000,000 per cent – that is, they are estimated to virtually guarantee the end of fighting. All four variables are again statistically significant at conventional levels ($p \leq 0.01$), but the size of the estimated substantive effects is impossibly large.

To explore the extent to which monotone likelihood is responsible for these extreme results, Table 3 provides an overview of imbalance on the problematic variables discussed above. We find that while 297 observations record the occurrence of opposing interventions, no observations record a conflict termination during a month when one of these opposing interventions took place. Notably, this is powerful evidence in support of Regan's hypothesis that opposing

[19]While we were unable to replicate the results in R due to convergence issues, we successfully replicated the results in Stata SE, Version 11.2.

[20]The conflict intensity indicator included in the original dataset and code, labeled avemnth, appears to be a transformed measure of a conflict's total duration rather than a measure of its intensity (the average number of casualties per month of the conflict). Indeed, the correlation between avemnth and a conflict's total duration is 0.9998. Consequently, including avemnth as a right-hand side variable generates considerable model instability owing to misspecification and measurement error. Given that conflict intensity is also captured by the fatalities variable – a time-invariant covariate that records a conflict's total death toll – we drop the avemnth variable in Models 2–4.

Table 2. Replication and extensions of Model 1, Table 1 as reported in Regan (2002, 69)

Variables	Model 1 Weibull AFT	Model 2 Weibull AFT	Model 3 Cox	Model 4 Firth Cox
Opposing	11.336***	21.451***	−17.028***	−1.635**
	(1.028)	(1.303)	(0.459)	(0.691)
	8.377×10^{4}	2.070×10^{9}	4.025×10^{-8}	0.195
Military intervention	11.326***	19.339***	−15.357***	0.780
	(1.007)	(1.503)	(0.744)	(0.897)
	8.292×10^{4}	2.510×10^{8}	2.141×10^{-7}	2.182
Economic intervention	21.997***	39.806***	−31.867***	0.346
	(1.965)	(2.610)	(1.227)	(1.384)
	3.573×10^{9}	1.939×10^{17}	1.450×10^{-14}	1.414
Time × force	0.018	0.022	−0.021	−0.008
	(0.017)	(0.021)	(0.017)	(0.034)
	1.019	1.023	0.979	0.992
Use of force	−0.533	−1.060	0.791	0.720
	(0.797)	(1.437)	(0.890)	(1.203)
	0.587	0.346	2.207	2.054
Homogeneity	−0.005	−0.007	0.005	0.005
	(0.003)	(0.008)	(0.005)	(0.006)
	0.995	0.993	1.005	1.005
Fatalities	0.000***	0.000**	−0.000**	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
	1.000	1.000	1.000	1.000
Intensity	0.041***			
	(0.006)			
	1.042			
Ethnoreligious	0.309**	0.697**	−0.421**	−0.411**
	(0.122)	(0.309)	(0.189)	(0.204)
	1.362	2.007	0.656	0.663
Time × government	−0.001	0.007	0.006	0.007
	(0.017)	(0.024)	(0.020)	(0.025)
	0.999	1.007	1.006	1.007
Time × opposition	−0.023***	−0.030***	0.024**	0.016
	(0.005)	(0.006)	(0.009)	(0.027)
	0.978	0.971	1.024	1.016
Target (gov./opp.)	−10.986***	−19.467***	15.253***	−0.731
	(1.125)	(1.471)	(0.831)	(0.929)
	0.000	0.000	4.211×10^{6}	0.481
International org.	0.042	0.765	−1.519	−0.379
	(0.420)	(1.003)	(1.280)	(0.981)
	1.043	2.150	0.219	0.684
Intervention	0.086	0.618*	−0.378*	−0.380*
	(0.129)	(0.365)	(0.226)	(0.230)
	1.090	1.854	0.685	0.684
Conflict Episodes	150	150	150	150
Terminations	119	119	119	119
Observations	13,048	13,048	13,048	13,048

Note: Models 1–3 report variable coefficients, with robust standard errors clustered on conflict in parentheses, followed by exponentiated coefficients. Model 4 reports Firth Cox penalized maximum-likelihood estimates, with bootstrapped standard errors in parentheses, followed by exponentiated coefficients. *$p \leq 0.10$; **$p \leq 0.05$; ***$p \leq 0.01$. All significance tests are two-tailed.

interventions prolong a conflict's expected duration. But from a methodological perspective, this means there is no overlap of failure times between conflicts that experience an opposing intervention and those that do not. Indeed, the variable is quasi-completely separating: all values opposing = 0 correspond to all termination events. Consequently, it is monotonic when ordered according to failure time, resulting in the inflated coefficient estimates reported in Table 2. This not only impedes meaningful interpretation of the effect of opposing interventions; it also induces considerable instability in parameter estimates of the model's other covariates.

Table 3. Imbalance of problematic variables of interest

	Termination = 0	Termination = 1	Imbalance ratio
Opposing = 1	297	0	297:0
Economic intervention = 1	132	0	132:0
Military intervention = 1	897	9	100:1
Target (gov./opp.) = 1	977	9	109:1

The same problem that afflicts the opposing interventions indicator can be identified for the other variables with extreme coefficient estimates. For example, while 132 observations record economic interventions, no observations record an economic intervention in the final month of a war. Here again, there is no overlap of failure times between conflicts that experience an economic intervention and those that do not;[21] inflated parameter values are the inevitable result. The values of the variables recording military interventions and biased interventions do not uniquely correspond to all termination events, but here too significant imbalance introduces convergence issues. Thus, while a total of 906 observations record military interventions, only nine observations saw a conflict end during one of the conflict-months experiencing these military interventions. Similarly, while a total of 986 observations record biased interventions on the side of rebel forces taking place during ongoing fighting, only nine observations record biased interventions on the side of rebel forces during the final month of a war. Given the degree of imbalance on both predictor variables, their linear combination with other covariates in the model induces monotone likelihood. The net result is extreme parameter values in the Weibull AFT and standard Cox models.

More generally, while the dataset contains a large number of observations (n = 13,048), it is also characterized by a high level of censoring. Both conditions are a function of its conflict-month data structure, which records covariate values for individual conflicts for every 1-month period.[22] This high-resolution data structure has some important strengths; for example, it enables fine-grained measurement of relevant variables over time and provides increased statistical power by increasing the number of observations included in the dataset. Yet this data structure risks introducing covariate imbalance, especially for dichotomous variables, by increasing the number of censored observations recorded on the conflict termination indicator variable. To see this, consider that of the 150 conflicts recorded in the dataset, only 119 terminated during the period of observation. In practice, this means that of the 13,048 total observations in the dataset, only 119 are coded as a conflict termination event (that is, where 0 = censored and 1 = terminated). The net result is a censoring level in excess of 99 per cent. Notably, such a high level of censoring is not uncommon: especially in duration models with time-varying covariates, higher-resolution coding procedures and data structures go hand in hand with higher levels of censoring.

To ameliorate the problem of monotone likelihood, Model 4 re-runs the analysis using a Firth Cox model, which implements penalized maximum likelihood. With this correction, the model's estimated coefficients for the imbalanced predictors are both plausible and in line with other studies. Consider the results identified for opposing interventions – the variable of interest that serves to test Regan's hypothesis that countervailing interventions are associated with longer conflict durations. The model estimates these interventions to be associated with an 80.5 per cent decrease in the hazard of conflict termination relative to conflicts that did not experience opposing interventions – a statistically significant result ($p \leq 0.01$). While this estimate remains substantively large, it is in line with studies that report similar effect sizes when troops are simultaneously provided to both government and rebel forces in a conflict (Balch-Lindsay and

[21]I.e., all values economic intervention = 0 correspond to all termination events.

[22]There is considerable variation in the duration of conflicts included in the dataset. For example, while the shortest conflict lasts just 1 month, the longest spans 616 months.

Enterline 2000), or when civil wars become afflicted by competitive intervention (Anderson 2019).

The indicator variables recording military interventions, economic interventions and biased interventions on the side of rebels flip signs and are no longer statistically significant, but the estimated coefficients for all three measures no longer take on extreme values. The application of penalized maximum-likelihood estimation eliminates the convergence issues associated with these variables, providing finite parameter estimates even when confronted by the problems of monotone likelihood identified above.

In sum, we find support for Regan's hypothesis that opposing interventions generate longer civil wars. While Weibull AFT and standard Cox models estimate unrealistically large effect sizes, we correct for the problem of monotone likelihood and identify estimates that are both plausible and in line with other studies. Our application of the Firth Cox approach highlights the value of penalized maximum-likelihood estimation, which we find offers significant advantages for scholars encountering nonconvergence issues and inaccurate parameter estimates when employing duration models with imbalanced predictors.

Application 2: Partition and Peace Duration

In an important article on the prospects for peace following state partition, Tir (2005) examines the conditions under which countries avoid domestic conflict following their division into rump and secessionist states. Developing an argument that links partition-related factors to post-partition outcomes, he hypotheses that conflict is more likely in partitioned countries that emerge from a violent partition process (for example, Bosnia/Serbia), whereas peace is more likely in partitioned countries that emerge from a peaceful partition process (for example, Czech Republic/Slovakia). To test this proposition, he compiles a dataset of all cases of state partition that occurred during the twentieth century. Adopting a country-year format, the dataset records observations for both rump and secessionist states for each year following their partition through to 1996. A binary variable records whether the partition process was peaceful or violent; additional variables record whether the partition was ethnically based and whether the post-partition state was the secessionist (as opposed to the rump) country. Finally, a battery of controls captures the state's level of ethnic heterogeneity, its level of economic development, the coherence of its political regime and its population size.[23]

As Tir's research question seeks to identify the determinants of post-partition peace duration, the dataset records as a case of 'failure' those country-years that experience the termination of a peace episode (that is, the onset of domestic conflict). While the original study employs two measures of domestic conflict – the occurrence of armed conflict and the onset of civil war – we focus on the latter dependent variable, which draws on data compiled by the Correlates of War project (Sarkees 2000; Small and Singer 1982).[24] In total, there are 1,532 years of observation that capture forty-nine peace episodes, of which eighteen experienced peace terminations during the period of observation.

Table 4 reproduces and extends the findings of the article. Model 5 reports the original specification, which employs a standard Cox model. The results suggest that considerable nonconvergence issues afflict the parameter estimates of the independent variable of interest – the indicator for a peaceful partition process. In particular, the model estimates that peaceful partitions decrease the hazard of peace termination by over 99.999 per cent; that is, peaceful partition processes are found to virtually guarantee the continuation of peace in rump and secessionist states. These results are strongly statistically significant ($p \leq 0.01$) and represent powerful

[23]We refer readers to Tir (2005, 553–554) for a description of the included covariates. Note that regime coherence is a binary measure that captures whether the state's political regime is either democratic or authoritarian; both regime types are considered to be 'coherent' compared to anocracies. For ease of interpretation of the results, we rescale the ethnic heterogeneity measure in the models reported below.

[24]In this dataset, civil wars are defined as sustained conflicts taking place within the boundaries of an internationally recognized state that cause at least 1,000 battle-related fatalities.

Table 4. Replication and extensions of Model 4, Table 1 as reported in Tir (2005, 556)

Variables	Model 5 Cox	Model 6 Firth Cox	Model 7 Firth Cox
Peaceful partition	−19.489***	−2.459**	−2.537*
	(0.613)	(1.231)	(1.785)
	0.000	0.086	0.079
Ethnic partition	−0.534*	−0.540	−0.689
	(0.400)	(1.447)	(1.358)
	0.586	0.583	0.502
Secessionist country	−1.516**	−1.490	−1.545
	(0.727)	(1.250)	(1.473)
	0.219	0.225	0.213
Ethnic heterogeneity	−0.001	−0.016	−0.034
	(0.039)	(0.078)	(0.106)
	0.999	0.984	0.967
Ethnic heterogeneity squared	−0.000	0.000	0.000
	(0.000)	(0.001)	(0.001)
	1.000	1.000	1.000
Economic development			−0.817
			(1.196)
			0.442
Regime coherency	−0.924***	−0.939	−0.605
	(0.378)	(0.767)	(1.058)
	0.397	0.391	0.546
Population size	0.348**	0.253	0.110
	(0.208)	(0.436)	(0.394)
	1.417	1.288	1.116
Peace Episodes	49	49	36
Terminations	18	18	18
Observations	1,532	1,532	1,436

Note: Model 5 reports variable coefficients, with robust standard errors clustered on conflict in parentheses, followed by exponentiated coefficients. Models 6 and 7 report Firth Cox penalized maximum likelihood estimates, with bootstrapped standard errors in parentheses, followed by exponentiated coefficients. *$p \leq 0.10$; **$p \leq 0.05$; ***$p \leq 0.01$. In keeping with the original study, all significance tests are one-tailed.

evidence in support of Tir's hypothesis. However, the estimated effect sizes, which converge toward negative infinity, undermine substantive interpretation of the results.

Is monotone likelihood to blame for the extreme effect sizes? Table 5 provides a descriptive overview of the peaceful partition indicator's degree of imbalance. We find that while a total of 485 observations record country-years that followed a peaceful partition, none of these observations saw the onset of a civil war.[25] In other words, there is no overlap in failure times between cases that did and did not experience a peaceful partition, as there are no termination events associated with the value peaceful partition = 1. Consequently, the indicator variable is monotonic when ordered according to failure time. This accounts for the extreme parameter values estimated by the model.

To solve the problem of monotone likelihood in this case, Model 6 re-runs the analysis employing a Firth Cox model. The penalized maximum-likelihood estimation ameliorates convergence issues. This model estimates that peaceful partitions are associated with a 91.4 per cent decrease in the hazard of peace termination ($p \leq 0.05$). To be sure, this remains a large effect size. However, it is now both more reasonable and substantively meaningful. While monotone likelihood resulted in almost infinite effect size estimates when employing a standard Cox model, the use of penalized maximum likelihood corrects for this problem and enables the estimation of finite parameter values even in the face of highly imbalanced covariates.

[25]Note that the full dataset has four cases in which peaceful partition = 1 and termination (i.e., civil war onset) = 1; however, due to listwise deletion resulting from missing data, these cases are dropped when the model is estimated, resulting in the imbalance reported in Table 5.

Table 5. Imbalance of problematic variable of interest

	Termination = 0	Termination = 1	Imbalance Ratio
Peaceful partition = 1	485	0	485:0

Model 7 extends the analysis further, incorporating a measure of the partitioned state's level of economic development. This variable was excluded in the original article's estimated model owing to the rarity of peace terminations and missing data issues, which left insufficient variance for the joint estimation of the peaceful partition indicator and the economic development variable.[26] As noted above, dropping problematic variables is one of a number of distasteful remedies scholars have in the past been forced to use when encountering monotone likelihood in their empirical research. We demonstrate the advantages of penalized maximum-likelihood estimation by (re)incorporating the economic development indicator into a fully specified model.

Our Model 7 results show that convergence issues need not force researchers to risk omitted variable bias when encountering monotone likelihood: Firth's corrective ensures finite parameter estimates can be obtained even when insufficient variance undermines the joint estimation of covariates under a standard maximum-likelihood approach. We find that the estimate of the peaceful partition indicator is robust to the incorporation of the economic development indicator, though there is greater uncertainty around this estimate in the fully specified model.

In sum, we find evidence in support of Tir's hypothesis that sustained peace is more likely in partitioned countries that emerge from a peaceful partition process. While standard Cox model estimates suffer from monotone-likelihood problems that complicate substantive interpretation of the results, we show that Firth Cox models can ameliorate this issue to render parameter estimates both meaningful and plausible. Further, while Cox models require that some variables be omitted owing to insufficient variance, we demonstrate that Firth Cox models enable researchers to avoid this distasteful remedy and estimate fully saturated models even under low-variance conditions. Here again, our results underscore the value of penalized maximum-likelihood estimation when encountering monotone likelihood in duration modeling of political events.

Conclusion

Political scientists now widely understand the problem of (quasi-)complete separation within (multinomial) logit models (Cook, Hays and Franzese 2020; Cook, Niehaus and Zuhlke 2018; Rainey 2016; Zorn 2005). While such separation problems – and the monotone-likelihood challenges that arise from them – are also commonplace in duration models, this issue has not yet been widely recognized within political science. Indeed, this issue is likely even more acute in the latter setting due to the time-dependent nature of duration data. We rectify this deficiency by outlining the problem of monotone likelihood within duration modeling and by detailing a readily accessible solution: the application of Firth's penalized maximum-likelihood approach to Cox model estimation. We then evaluate the extent of the monotone-likelihood problem for duration analyses in political science via a series of simulations and replications, before demonstrating the substantial advantages of the Firth Cox model in each context.

In brief, our simulations expand upon and improve past studies to illustrate that monotone likelihood can consistently arise in samples as large as 1,000–2,000, especially where censoring is 75 per cent or higher. They likewise demonstrate that monotone likelihood issues will

[26]See Tir (2005, 555, footnote 14).

commonly produce nonconvergence and (near-)infinite parameter estimates within standard Cox models when samples are less than or equal to 500 observations, regardless of the level of censoring. Lastly, our simulations also suggest that the Firth Cox model exhibits superior accuracy to the Cox model for all parameters of interest, regardless of the conditions considered.

Our empirical applications further illustrate the pernicious implications of monotone likelihood, even for studies that employ far larger samples than those mentioned above. Precisely because time-varying political science duration setups essentially guarantee *extremely* high proportions of censoring across all observations, researchers must consider monotone likelihood when employing *any* duration model setup. Our applications also demonstrate that researchers need not risk omitted variable bias when encountering monotone likelihood in their empirical research. With the application of penalized maximum-likelihood estimation, fully saturated models can be estimated even in low-variance conditions.

Altogether, our findings dovetail nicely with past political science research into Firth's correction, especially those that (1) highlight the benefits of Firth's correction for separation problems in binary response models (Rainey 2016; Zorn 2005) and (2) underscore the small sample size advantages of Firth's correction in the logit context more generally (Rainey and McKaskey Forthcoming).

These contributions notwithstanding, future research should work to extend the monotone-likelihood solution evaluated here in at least three ways. First, and in line with Rainey's (2016) recent discussion of Firth's penalization within the logistic regression context, more work should be done to explore and evaluate the potential implications of one's choice of prior within the Firth Cox setup. Secondly, while estimation routines are available for the Firth Cox model, such programs do not exist for many parametric duration models that are commonly used in political science, including the Weibull, Gompertz and log-logistic models. Developing accessible software and code for the application of Firth's penalization to the latter models will further expand political scientists' toolboxes for the accurate modeling of duration outcomes. Finally, political scientists have begun to widely apply split-population duration models to relevant social outcomes given these models' ability to accommodate mixtures of 'at risk' and 'immune' populations (for example, Bagozzi et al. 2019; Beger et al. 2017; Svolik 2008). Such multi-equation models are likely to be particularly sensitive to monotone-likelihood issues, suggesting that Firth's correction may be an especially promising default approach for estimation in this context.

Supplementary material. Data replication sets are available at the Harvard Dataverse at: https://doi.org/10.7910/DVN/OLMVP5. The online appendix is available at: https://doi.org/10.1017/S0007123420000071.

Acknowledgements. We wish to thank Daina Chiba, Justin Esarey, Jonathan Golub, Dotan Haim, Håvard Hegre, Benjamin Jones, Shawna Metzger, George Yin, participants of the 2019 Dickey Center Conference at Dartmouth, and two anonymous reviewers for their advice, comments, and suggestions on previous drafts and presentations of this article. All errors remain our own. Bagozzi's research was supported by the National Science Foundation under Grant No. DMS-1737865.

References

Albert A and Anderson JA (1984) On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71**(1), 1–10.

Anderson N, Bagozzi, B, Koren, O (2020) "Replication Data for: Addressing Monotone Likelihood in Duration Modeling of Political Events", https://doi.org/10.7910/DVN/OLMVP5, Harvard Dataverse, V1

Anderson N (2019) Competitive intervention, protracted conflict, and the global prevalence of civil war. *International Studies Quarterly* **63**(3), 692–706.

Bagozzi BE, Joo MM, Kim B and Mukherjee B (2019) A bayesian split population survival model for duration data with misclassified failure events. *Political Analysis* **27**(4), 415–434.

Balch-Lindsay D and Enterline AJ (2000) Killing time: the world politics of civil war duration, 1820–1992. *International Studies Quarterly* **44**(4), 615–642.

Beger A et al. (2017) Splitting it up: the spduration split-population duration regression package for time-varying covariates. *The R Journal* **9**(2), 474–486.

Bennett DS and Stam AC (1996) The duration of interstate wars, 1816–1985. *American Political Science Review* **90**(2), 239–257.
Box-Steffensmeier JM, Arnold LW and Zorn C (1997) The strategic timing of position taking in congress: a study of the north American free trade agreement. *American Political Science Review* **91**(2), 324–338.
Box-Steffensmeier JM and Jones BS (2004) *Event History Modeling: A Guide for Social Scientists*. New York: Cambridge University Press.
Bryson MC and Johnson ME (1981) The incidence of monotone likelihood in the Cox model. *Technometrics* **23**(4), 381–383.
Chiba D, Metternich NW and Ward MD (2015) Every story has a beginning, middle, and an end (but not always in that order): predicting duration dynamics in a unified framework. *Political Science Research and Methods* **3**(3), 515–541.
Cook SJ, Hays JC and Franzese RJ (2020) Fixed effects in rare events data: a penalized maximum likelihood solution. *Political Science Research and Methods* **8**(1), 92–105.
Cook SJ, Niehaus J and Zuhlke S (2018) A warning on separation in multinomial logistic models. *Research & Politics* **5**(2), 1–5.
Cunningham DE (2011) *Barriers to Peace in Civil War*. Cambridge: Cambridge University Press.
Firth D (1993) Bias reduction of maximum likelihood estimates. *Biometrika* **80**(1), 27–38.
Fortna VP (2008) *Does Peacekeeping Work? Shaping Belligerents' Choices after Civil War*. Princeton, NJ: Princeton University Press.
Gates S et al. (2006) Institutional inconsistency and political instability: polity duration, 1800–2000. *American Journal of Political Science* **50**(4), 893–908.
Harden JJ and Kropko J (2019) Simulating duration data for the Cox model. *Political Science Research and Methods* **7**(4), 921–928.
Heinze G and Dunkler D (2008) Avoiding infinite estimates of time-dependent effects in small-sample survival studies. *Statistics in Medicine* **27**(30), 6455–6469.
Heinze G and Schemper M (2001) A solution to the problem of monotone likelihood in Cox regression. *Biometrics* **57**(1), 114–119.
Jeffreys H (1946) An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London: Series A (Mathematical and Physical Sciences)* **186**(1007), 453–461.
Johnson ME et al. (1982) Covariate analysis of survival data: a small-sample study of Cox's model. *Biometrics* **38**(3), 685–698.
Jones BT and Metzger SK (2019) Different words, same song: advice for substantively interpreting duration models. *PS: Political Science & Politics* **52**(4), 691–695.
Kropko J and Harden JJ (2020) Beyond the hazard ratio: generating expected durations from the Cox proportional hazards model. *British Journal of Political Science* **50**(1), 303–320.
Licht AA (2011) Change comes with time: substantive interpretation of nonproportional hazards in event history analysis. *Political Analysis* **19**(2), 227–243.
Licht AA (2017) Hazards or hassles: the effect of sanctions on leader survival. *Political Science Research and Methods* **5**(1), 143–161.
Loughin TM (1998) On the bootstrap and monotone likelihood in the Cox proportional hazards regression model. *Lifetime Data Analysis* **4**(4), 393–403.
McCullagh P and Nelder JA (1989) *Generalized Linear Models*, 2nd Edn. London: Chapman and Hall.
McKibben HE and Western SD (2020) Reserved ratification: an analysis of states' entry of reservations upon ratification of human rights treaties. *British Journal of Political Science* **50**(2), 687–712.
Omgba LD (2009) On the duration of political power in Africa: the role of oil rents. *Comparative Political Studies* **42**(3), 416–436.
Rainey C (2016) Dealing with separation in logistic regression models. *Political Analysis* **24**(3), 339–355.
Rainey C and McKaskey K (Forthcoming) Estimating logit models with small samples. *Political Science Research and Methods*.
Regan PM (2000) *Civil Wars and Foreign Powers: Outside Intervention in Intrastate Conflict*. Ann Arbor: University of Michigan Press.
Regan PM (2002) Third-party interventions and the duration of intrastate conflicts. *Journal of Conflict Resolution* **46**(1), 55–73.
Ruhe C (2018) Quantifying change over time: interpreting time-varying effects in duration analyses. *Political Analysis* **26**(1), 90–111.
Sarkees MR (2000) The Correlates of War data on war: an update to 1997. *Conflict Management and Peace Science* **18**(1), 123–144.
Small M and Singer JD (1982) *Resort to Arms: International and Civil Wars, 1816–1980*, 2nd Edn. Beverly Hills, CA: Sage.
Svolik MW (2008) Authoritarian reversals and democratic consolidation. *American Political Science Review* **102**(2), 153–168.
Tir J (2005) Dividing countries to promote peace: prospects for long-term success of partitions. *Journal of Peace Research* **42**(5), 545–562.
Tsiatis AA (1981) A large sample study of Cox's regression model. *The Annals of Statistics* **9**(1), 93–108.
Zorn C (2005) A solution to separation in binary response models. *Political Analysis* **13**(2), 157–170.

Cite this article: Anderson N, Bagozzi BE, Koren O (2020). Addressing Monotone Likelihood in Duration Modelling of Political Events. *British Journal of Political Science* 1–18. https://doi.org/10.1017/S0007123420000071