Beyond the Hazard Ratio: Generating Expected Durations from the Cox Proportional Hazards Model

Online Appendix

Jonathan Kropko Jeffrey J. Harden[*]

May 25, 2017

[*] Department of Politics, University of Virginia (email: jkropko@virginia.edu); Department of Political Science, University of Notre Dame, (email: jeff.harden@nd.edu). Previous versions of this article benefited from presentations at the 2014 Annual Meeting of the Society for Political Methodology, Athens, Georgia. We would also like to thank Justin Kirkland, Bruce Desmarais, Meg Shannon, Fred Boehmke, Anand Sokhey, the editor and three anonymous reviewers for helpful comments. Data replication sets, code, and online appendices are available at http://dataverse.harvard.edu/dataverse/BJPolS and at http://dx.doi.org/10.7910/DVN/ELT9VD.

Appendices

A A Brief Summary of Survival Models

Survival models are designed to provide an explanation for why a particular observation survives for a particular duration of time. A duration, denoted t_i for observation i, can be a patient's lifespan after a diagnosis, the time needed for a negotiation to result in an agreement, or the amount of time that passes before an event like government failure or war, among many other examples. While both parametric survival models and the Cox model have similar purposes, they also exhibit key differences. We briefly review these models in a technical discussion below. See Box-Steffensmeier and Jones[1]—which we rely on extensively for this review—for more details.

A.1 Parametric Survival Models

Survival models improve upon ordinary least squares (OLS) for duration data by allowing for skewed distributions of the durations and by explicitly accounting for the fact that some observations are right censored, meaning that their durations end some time after data collection ends. The likelihood function used by all parametric survival models takes the form

$$L(\theta|\mathbf{t},\mathbf{X}) = \prod_{i=1}^{N} f_i(t)^{\delta_i} \; S_i(t)^{1-\delta_i}, \tag{1}$$

where i indexes observations, θ represents the parameters to be estimated, $\mathbf{t}$ represents the observed durations with t_i referring to the duration of observation i, $\mathbf{X}$ represents the matrix of covariates, N is the sample size, and δ_i is an indicator for the right censored observations.

[1] Box-Steffensmeier and Jones 2004

$f_i(t)$ is the PDF of failure times t and

$$S_i(t) = 1 - \int_0^t f_i(t)\, dt \tag{2}$$

is the survivor function, which represents the probability that observation i survives until time t or later.

An important concept in survival modeling is the hazard function, or hazard rate, defined as the ratio of the failure PDF to the survivor function,

$$h_i(t) = \frac{f_i(t)}{S_i(t)}. \tag{3}$$

The hazard rate represents the relative risk of failure at time t_i conditional on survival until time t_i.[2] Results from survival models are often expressed in terms of the hazard ratio, the ratio of two (actual or hypothetical) observations' hazard rates. The failure and survivor functions are different for each observation. These idiosyncratic functions share the same baseline failure PDF, $f_0(t)$, and the variation across cases is induced by the data.

If a parametric survival model can be reparameterized as

$$\log(t_i) = \alpha + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \sigma\varepsilon_i, \tag{4}$$

then the model has an accelerated failure time interpretation in which it is possible to derive the expected duration and marginal change in duration with respect to a covariate. The exponential, Weibull, and log-normal models can all be interpreted in this way. Specifically,

[2] Box-Steffensmeier and Jones 2004, 15

in equation 4, $\sigma = 1$ for the exponential model, ε_i is distributed by the type-1 extreme-value distribution for the Weibull model and by the standard normal distribution for the log-normal model.[3]

Because parametric survival models provide an analytic function for the hazard rate, the hazard is assumed to follow one of the paths allowed by this functional form. In many cases this assumption is too restrictive. For instance, the exponential model assumes that the hazard is constant, the Weibull model assumes that the hazard rate is monotonically increasing or decreasing over time, and the log-normal model assumes that the hazard rate is either monotonic or unimodal. Many researchers do not wish to assume that the hazard rate follows exactly one of these forms. As a result, the Cox model has gained wide use in the social sciences relative to the parametric survival models.

A.2 The Cox Proportional Hazards Model

The parameters of the Cox model are estimated by maximizing a partial likelihood function. Cox[4] shows that the this estimator converges, in the sample size, to the maximum likelihood estimator. Carrying notation from above forward, the partial likelihood is defined as follows

$$\prod_{i=1}^{N} \left[\frac{\exp(\beta' X_i^{(t_i)})}{\sum_{\forall j \in R_{t_i}} \exp(\beta' X_j^{(t_i)})} \right]^{\delta_i} \tag{5}$$

where β is a vector of regression coefficients to be estimated and δ_i is again an indicator for right censored observations.

The key feature of the estimator is the term in the denominator: $R(t_i)$. This refers to observation i's 'risk set.' An observation's risk set is comprised of the observations that

[3] See Box-Steffensmeier and Jones 2004, 23–32

[4] Cox 1975

experience the event at the same time or after observation *i*. Thus, the partial likelihood estimator is the product of the conditional probability of failure at a certain time, given all the observations that have not yet failed at that time. This allows the method to estimate parameters while relying only on the ranks of the durations—not the actual durations—and thus avoid making an assumption about the baseline hazard.[5]

However, that advantage comes with some costs. If the baseline hazard assumption is correct, a parametric model will be more efficient than the Cox model because the former uses more information from the data.[6] Moreover, coefficient estimates can only be interpreted with respect to the hazard rate; positive coefficients indicate the hazard is rising while negative estimates signify a decrease. For example, exponentiating a coefficient estimate yields the average multiplicative change in the hazard rate—a hazard ratio—for a one-unit increase in the independent variable. Similarly, Box-Steffensmeier and Jones[7] recommend the following formula for computing the percentage change in the hazard rate between two values (X_1 and X_2) of an independent variable

$$\%\Delta h(t) = \left[\frac{\exp(\beta[X_i = X_1]) - \exp(\beta[X_i = X_2])}{\exp(\beta[X_i = X_2])}\right] \times 100. \tag{6}$$

As we state above, these hazard rate-based computations are mathematically sound; researchers who employ them are not doing so in error. However, because the Cox model

[5] This also means the partial likelihood estimator is quite sensitive to model specification issues such as omitted variables and measurement error. However, analysts need not abandon the Cox model due to these problems because they can be resolved with a robust estimator of the partial likelihood. See (Desmarais and Harden 2012).

[6] Box-Steffensmeier and Jones 2004

[7] Box-Steffensmeier and Jones 2004, 60

does not use the actual durations, estimates of expected duration are not readily available. In the main text we contend that the interpretation and communication of substantive results from the Cox model can be improved by adapting the Cox model to estimate expected durations and marginal changes in these durations with respect to a covariate. We show that COX ED fulfills this objective.

B Journal Article Meta Analysis

Our meta analysis examined use of the Cox proportional hazards model and methods for interpreting results in five political science journals from 1996–2015: *American Political Science Review*, *American Journal of Political Science*, *British Journal of Political Science*, *Journal of Politics*, and *International Organization*. We first searched for articles using Google Scholar. Then we coded the articles based on (1) the language used to frame hypotheses and (2) the methods used to interpret Cox model results. We describe the details of these procedures below.

Our central objectives were to assess (1) the type of language researchers typically use to frame their hypotheses when employing the Cox model and (2) the methods they typically use to interpret Cox model results. On the first objective we considered two possible framing styles: a risk frame and a duration frame. A risk frame discusses hypotheses with respect to the risk of event occurrence. For example: 'as X increases, the risk of event Y occurring also increases.' In such a case the researcher is not primarily concerned with duration, but rather focuses on how the covariates make the event more or less likely to occur. In contrast, a duration frame discusses the hypothesis in terms of event time, as in 'as X increases, the number of days until event Y occurs decreases.' In this case the length of time that an event takes is of central importance.

B.1 Searching for Articles

We searched https://scholar.google.com/ for [‘cox proportional hazards’ OR ‘cox model’ OR ‘cox regression’] in each journal listed above, one journal at a time. We set the date range to 1996–2015. We downloaded all of the articles returned by these searches, checked each one to make sure it included analysis with the Cox model, then saved it for coding in the next step. We included any paper that reported the estimation of a Cox model in the main text. This produced 80 total articles.[8] The articles came from four subfields: international relations (42), comparative politics (21), American politics (13), and methodology (4). Additionally, the articles spanned all five journals: 20 from *American Journal of Political Science*, 20 from *Journal of Politics*, 17 from *British Journal of Political Science*, 12 from *American Political Science Review*, and 11 from *International Organization*.

B.2 Coding Hypothesis Text

First, we identified the number of hypotheses in each article and copied the text of those hypotheses. If multiple analyses were presented, we only included hypotheses pertaining to the Cox model(s) reported in the main text. If no hypotheses were presented with the Cox model (i.e., in a descriptive analysis), we used the authors’ descriptions of the model specification (i.e., variables used and purpose of the estimation).

Next, we placed all of the hypotheses’ text into a single string, omitted common English stop words, then counted word frequencies of the remaining words. This produced a list of 1,570 unique words, from which we identified words as either predominantly part of a risk frame and words predominantly used in a duration frame.[9] Our general rule was to

[8] We also searched for articles published since 1990, but found no relevant articles published before 1996.

[9] This was a subjective assessment, and we encourage interested readers to obtain the replication materials

code any word that related to probability, likelihood, or chance in the risk frame category and any word relating to time in the duration frame category. In all, we coded eight unique words as risk frame words and 68 unique words as duration frame words. Table 1 reports these words and their frequencies.

B.3 Hypothesis Framing Results

The fact that we coded many more words as duration frame words gives some preliminary evidence that authors tend to frame their hypotheses with respect to the time until event occurrence more often than the risk of event occurrence. However, this may be skewed by the possibility that authors simply have more choices when it comes to duration frame words. Looking at word counts of all eight risk frame words and the top eight duration frame words reveals a larger count of risk frame words: 140 instances of risk frame words and 120 duration frame words. Nonetheless, the full count of all the duration frame words we coded is 317—substantially larger than the total count of risk frame words.

We also coded each article as either predominantly using a risk frame, duration frame, or equal use of both frames. We accomplished this in two ways: a count of unique words and a count of total words. First, we counted how many unique words from each frame appeared in the text of the hypotheses. This approach did not give additional weight to the same word appearing more than once. We then coded an article's frame as the type with the most instances of unique words from its list. Second, for each article we counted the total number of instances of words in its hypotheses from each frame (i.e., allowing for repeats of the same word). In both cases if an equal number of risk and duration frame words appeared, we coded the article as equal.

and assess whether they agree with our decisions.

Table 1: Frequency and Frame of Hypothesis Framing Words

Word	Frequency	Coded Frame
likely	77	Risk
risk	26	Risk
hazard	16	Risk
probability	9	Risk
likelihood	8	Risk
propensity	2	Risk
unlikely	1	Risk
odds	1	Risk
time	29	Duration
duration	16	Duration
timing	16	Duration
longer	14	Duration
survival	12	Duration
end	11	Duration
tenure	11	Duration
termination	11	Duration
early	10	Duration
initiator	10	Duration
delay	9	Duration
delays	8	Duration
durable	8	Duration
earlier	8	Duration
first	8	Duration
future	8	Duration
deadline	7	Duration
length	7	Duration
quickly	6	Duration
durability	5	Duration
finite	5	Duration
hazards	5	Duration
inhibit	5	Duration
mortality	4	Duration
periods	4	Duration
process	4	Duration
processes	4	Duration
shorten	4	Duration
times	4	Duration
conclude	3	Duration
past	3	Duration
period	3	Duration
short	3	Duration
shortens	3	Duration
shorter	3	Duration
survive	3	Duration
date	2	Duration
deadlines	2	Duration
fail	2	Duration
failing	2	Duration
failure	2	Duration
terminate	2	Duration
live	2	Duration
long	2	Duration
term	2	Duration
immediately	2	Duration
last	2	Duration
delayed	1	Duration
delaying	1	Duration
durations	1	Duration
indefinitely	1	Duration
longer	1	Duration
longer lasting	1	Duration
preceding	1	Duration
pre deadline	1	Duration
prior	1	Duration
prolonging	1	Duration
remain	1	Duration
remained	1	Duration
remains	1	Duration
retards	1	Duration
shortrun	1	Duration
slow	1	Duration
slowly	1	Duration
survives	1	Duration
temporal	1	Duration
concludes	1	Duration
onset	1	Duration

Using the unique word count, we coded 52 articles as using a duration frame, 15 with a risk frame, and 13 with equal use of both frames. With the total word count these numbers are 49, 22, and 9, respectively. About 43 per cent (34) of the articles use words from both frames, 31 contain no risk frame words, and 15 contain no duration frame words. We also conducted these counts after deleting all of the duration words that appear three or fewer times to check whether these results are driven only by the fact that there may be more choices of duration frame words. In that case 40 articles are still coded with the duration frame using both the unique and total counts.

B.4 Coding the Interpretation Methods

Our second objective was to code which method(s) each article used to interpret the results of the estimated Cox model. This was accomplished by reading the results sections of the articles and identifying each unique method used. We created a total of four categories based on what we found in the text, which we list below. Note that all of the articles discussed the sign and significance of the Cox model coefficient estimates. The categories reflect any interpretation beyond sign and significance. The articles employed an average of 1.35 of these interpretation methods. 56 articles used one method, 20 articles used two methods, and 4 articles employed three different methods.

- *Hazard ratios* (30 articles). This category included any article that reported the exponentiation of one or more Cox model coefficients, as well as a discussion about the resulting multiplicative effect of a one-unit change in the covariate of interest.

- *Changes to the hazard rate* (44 articles). This category included any article that reported a marginal change in the hazard rate (usually expressed as a percentage) corresponding to a substantively interesting change in the value of a covariate.

- *Empirical estimates of the hazard and/or survivor functions* (24 articles). This category included any article that graphically displayed an estimate of the baseline hazard from the model and/or computed the survivor function. Typically this was done for different covariate values to show the effect of changes to the covariate.

- *Only sign and significance of the coefficient estimates* (10 articles). This category included any article that did not report any interpretation of the Cox model other than the sign and significance of the relevant coefficient estimates.

The most important finding from this analysis is the fact that all of the articles that go beyond sign and significance in their interpretation of the Cox model focus on the hazard rate, whether through hazard ratios, changes to the hazard rate, or estimation and graphing of the baseline hazard and/or survivor functions. To further emphasize this point, a few articles in our data did report expected durations, but those estimates came from estimating the model using the Weibull parameterization.[10]

B.5 Meta Analysis Conclusions

This analysis yields two important insights. First, we find that political scientists employing the Cox model over the last 20 years tend to discuss their theoretical expectations in the language of time until event occurrence. Language related to the risk of event occurrence also appears, but it is less common than duration-based framing. Approximately 81 per cent of the articles in our sample contain more duration words or an equal amount of duration and risk words, compared to only 61 per cent containing more or equal risk

[10] e.g., Senese and Quackenbush 2003, 714

words. It is clear that researchers' substantive interests usually center on the duration of some political phenomenon, not just its likelihood of occurring.

This first finding contrasts sharply with the second finding, which is that researchers nearly exclusively rely on interpretation of the hazard rate after estimating the Cox model. We found no instances where researchers generated expected durations from their Cox model estimates. Thus, researchers who employ the Cox model are typically forced to switch the manner in which they discuss their research when moving from hypotheses to results. This provides motivation for our research. COX ED allows researchers to maintain consistency between the language they use to describe their theoretical framework and the language they use to communicate their empirical findings.

C The Relationship Between Hazard and Failure Probability

Here we show proof that the hazard ratio for a proportional hazards model is equal to the multiplicative change in the probability of failure at a particular instant t, conditional on survival until time t (see the discussion in the main text associated with footnote **??**).

Consider an example of a proportional hazards model in which the coefficients are non-zero. Without loss of generality, consider how an observation t_1 in which $X_1 = 1$ and $X_j = 0$ for $j > 1$ compares to a baseline observation t_0 in which all covariates are zero so that the hazard, failure probability density function (PDF), and survivor function for the observations are all equal to the baseline functions. Let β be the coefficient on X_1. The ratio of the hazard functions for each observation is

$$\frac{h_1(t)}{h_0(t)} = \frac{\exp(\beta)h_0(t)}{h_0(t)} = \exp(\beta). \tag{7}$$

Therefore, a one-unit increase in X_1 is associated with a multiplicative increase of $\exp(\beta)$ in hazard. Now consider how the probability of failure between $t = a$ and $t = b$, conditional on $t > a$, compares for each observation. The conditional probability that the baseline observation fails in this interval is given by

$$Pr(a \leq t_0 \leq b | t_0 > a) = \frac{Pr(a \leq t_0 \leq b)}{Pr(t_0 > a)}. \tag{8}$$

The numerator can be calculated from the baseline failure cumulative distribution function (CDF),

$$\begin{aligned} Pr(a \leq t_0 \leq b) \quad &= \int_a^b f_0(t)\, dt \quad = F_0(b) - F_0(a), \\ &= [1 - S_0(b)] - [1 - S_0(a)] \\ &= S_0(a) - S_0(b), \end{aligned} \tag{9}$$

and the denominator is the baseline survivor function $S_0(t)$ at $t = a$. The entire conditional probability is given by

$$Pr(a \leq t_0 \leq b | t_0 > a) = \frac{S_0(a) - S_0(b)}{S_0(a)}. \tag{10}$$

Likewise, the conditional probability that the non-baseline observation fails in this interval is

$$Pr(a \leq t_1 \leq b | t_1 > a) = \frac{S_1(a) - S_1(b)}{S_1(a)}. \tag{11}$$

We can rewrite the numerator of (11) as[11]

$$\begin{aligned} F_1(b) - F_1(a) &= [1 - S_1(b)] - [1 - S_1(a)] \\ &= S_1(a) - S_1(b) \\ &= S_0(a)^{\exp(\beta)} - S_0(b)^{\exp(\beta)}, \end{aligned} \tag{12}$$

and we can rewrite the denominator of (11) as

$$S_1(a) = S_0(a)^{\exp(\beta)}, \tag{13}$$

so that the conditional probability is

$$Pr(a \leq t_1 \leq b | t_1 > a) = \frac{S_0(a)^{\exp(\beta)} - S_0(b)^{\exp(\beta)}}{S_0(a)^{\exp(\beta)}}. \tag{14}$$

Note that the substitution $S(t) = S_0(t)^{\exp(X\beta)}$ is predicated on the assumption of proportional hazards. The ratio of the two probabilities is then

$$\frac{Pr(a \leq t_1 \leq b | t_1 > a)}{Pr(a \leq t_0 \leq b | t_0 > a)} = \frac{\frac{S_0(a)^{\exp(\beta)} - S_0(b)^{\exp(\beta)}}{S_0(a)^{\exp(\beta)}}}{\frac{S_0(a) - S_0(b)}{S_0(a)}}$$

[11] We exponentiate the coefficient β twice because we raise the baseline survivor function to the power of the hazard ratio, which is $\exp(\beta)$ (see Box-Steffensmeier and Jones 2004, 65, eq. 4.15).

$$= \frac{S_0(a)^{\exp(\beta)} - S_0(b)^{\exp(\beta)}}{S_0(a) - S_0(b)} \cdot \frac{S_0(a)}{S_0(a)^{\exp(\beta)}}$$

$$= S_0(a)^{1-\exp(\beta)} \cdot \frac{S_0(a)^{\exp(\beta)} - S_0(b)^{\exp(\beta)}}{S_0(a) - S_0(b)}. \tag{15}$$

In order to consider the multiplicative change in the conditional probability of *instantaneous* failure, let $S_0(a) = x$, $S_0(b) = y$, and $\exp(\beta) = \alpha$, and consider the following limit:

$$\lim_{x \to y} y^{1-\alpha} \frac{x^{\alpha} - y^{\alpha}}{x - y}$$

$$= y^{1-\alpha} \lim_{x \to y} \frac{x^{\alpha} - y^{\alpha}}{x - y}.$$

This limit is the definition of the derivative of the function $g(x) = x^{\alpha}$ evaluated at $x = y$, so the limit is equal to $g'(y) = \alpha y^{\alpha-1}$. Substituting for y and α, the instantaneous ratio of probabilities is equal to

$$\lim_{b \to a} \frac{Pr(a \leq t_1 \leq b | t_1 > a)}{Pr(a \leq t_0 \leq b | t_0 > a)} = S_0(b)^{1-\exp(\beta)} \exp(\beta) S_0(b)^{\exp(\beta)-1} = \exp(\beta).$$

D Summary of the Monte Carlo Simulations

Here we summarize an assessment of the performance of the COX ED methods. We use simulated data—in which we know the true relationships—to compare each COX ED approach to three parametric models. Specifically, we evaluate these different methods' capacity to return accurate expected durations for each observation in the data and marginal changes in duration for a unit change in a covariate. We also evaluate COX ED's performance in producing the correct confidence intervals (see Section F).

We simulate the data in two different ways. In one approach we simply draw the event times from a parametric distribution (exponential, Weibull, and log-normal). As we show in Section F, COX ED performs well under this type of data generating process (DGP). However, this approach is not our preferred simulation method. While the process is straightforward, drawing from a parametric distribution artificially inflates the performance of the parametric model. Furthermore, doing so is unrealistic because applied researchers never truly know if their data come from a particular parametric distribution. Accordingly, we also use the 'random spline' DGP, which generates baseline hazards by fitting cubic splines to randomly-drawn points.[12] This produces a variety of shapes, some of which are monotonic or unimodal, but many of which are multimodal. We use the randomly-drawn baseline hazard functions, along with three covariates and true coefficients generated from standard normal distributions, to create simulated durations.

There is not a single obvious method to generate simulated marginal effects, so we employ two similar strategies and present results from both. The first strategy uses the same method as the NPSF approach to COX ED to calculate true marginal effects, the second strategy uses the same method as the GAM approach. See Section E for complete details of this process.

After simulating data with the random spline method, we next estimate each version of COX ED and the exponential, Weibull, and log-normal survival models on the simulated data, and assess how accurately they return the expected durations and marginal changes in duration.[13] We compare competing models because it is difficult to assess absolute per-

[12] Harden and Kropko 2017

[13] We use the survival package in R for model estimation here and in our replication analyses (Therneau 2013).

formance in a simulation setting.[14] We conduct these comparisons with two performance criteria: (1) the root mean square error (RMSE) of each method's expected durations for each observation and (2) the RMSE of each estimator's expected change in duration for a one-unit change in a covariate. In both cases smaller values indicate less error, and thus better performance. We run these simulations in R for 1,000 iterations each with sample sizes of 50, 200, 500, and 1,000.

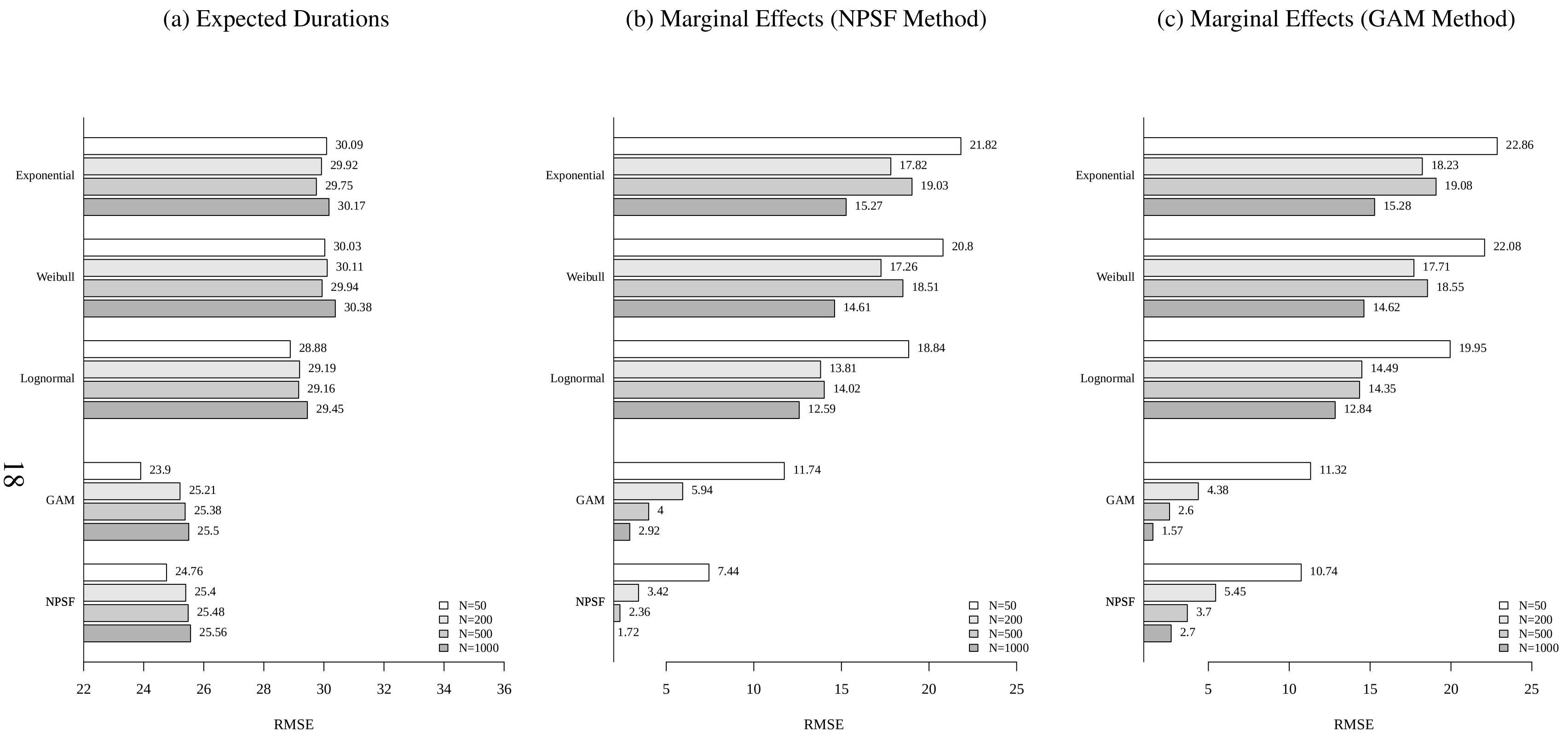
We present the simulation results for expected durations and marginal changes in expected duration in Figure 1. The three parametric models are illustrated at the top of each figure, and the COX ED approaches are below the parametric results. These results are drawn from the DGP with 10 per cent right censoring.[15]

The most important result is that the GAM and NPSF versions of COX ED have lower RMSEs than all three parametric models for both evaluative metrics, and at all sample sizes. Given that the only methods currently used by applied researchers for obtaining expected duration and marginal changes in expected duration use parametric models, this result provides a strong argument for the use of the COX ED procedures. The GAM approach to COX ED returns slightly more accurate expected durations than the NPSF approach, and as expected, each approach generally returns more accurate marginal effects than the other when its method is employed to generate the true marginal effects. However, these differences between the GAM and NPSF approaches are quite small compared to the differences between COX ED and the parametric models. We cannot claim that these results provide definitive evidence for the use of one COX ED approach over the other. That question will

[14] Carsey and Harden 2014

[15] See Section F for results with 5 per cent and 20 per cent right censoring (our conclusions remain the same).

Figure 1: Simulation Results with 10 Per Cent Right Censoring



(a) Expected Durations

(b) Marginal Effects (NPSF Method)

(c) Marginal Effects (GAM Method)

Note: The graphs report RMSE results for the expected durations (panel a), marginal effects using the NPSF method to create the true marginal effect (panel b), and marginal effects using the GAM method to create the true marginal effect (panel c). At each simulation iteration, 10 per cent of the observations are randomly selected to be right-censored. Each simulation contains 1,000 iterations. Within each simulation iteration, we obtain *N* expected durations, but only one marginal change in duration. Each iteration allows us to calculate an RMSE from the *N* expected durations, and we report the mean of these RMSE statistics across simulation iterations. For the marginal effects, we can only calculate an RMSE once all of the simulation iterations are complete. In all cases smaller RMSE values indicate less error, and thus better performance.

likely depend on other factors, such as features of the DGP, the fit of the Cox model, and the relative sparsity of the data.

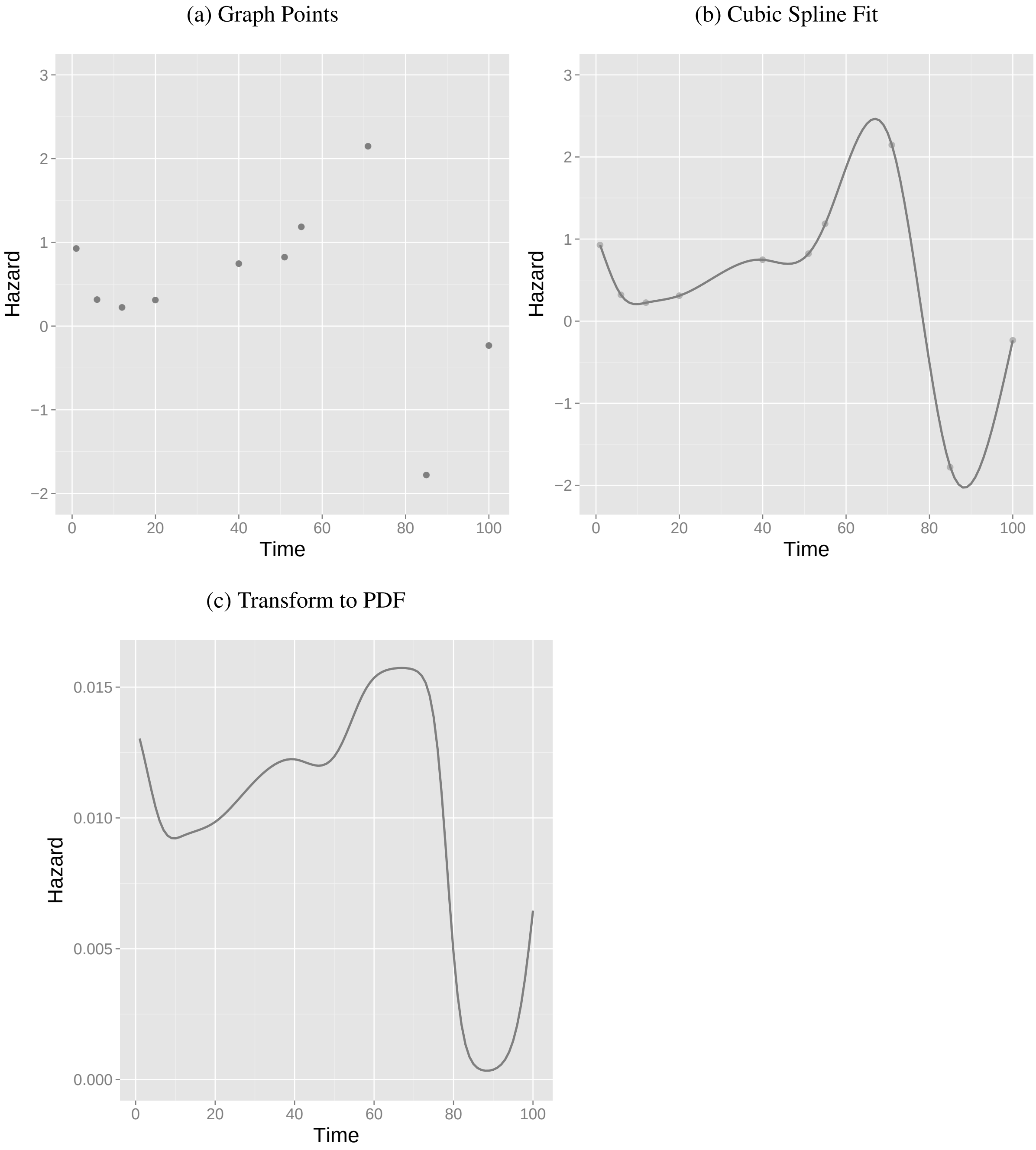
E Simulation Methods

The simulations in this paper present us with three challenges. First, we would like to generate baseline hazard functions that represent a variety of realistic hazard functions that may or may not be monotonic or unimodal. Second, we must draw simulated durations from these hazard functions. Third, we need to simulate marginal changes in these durations as a result of a change in a covariate. We discuss each of these issues in turn below.

E.1 Simulating Baseline Hazard Functions

Many researchers prefer to use the Cox model in order to avoid making an assumption that the baseline hazard follows a particular functional form. In particular, researchers often do not want to assume that the hazard is constant as in the exponential model, monotonic as in the Weibull model, or unimodal as in the log-normal model. Instead of using the assumed distributions of those parametric models, in our main simulations we generate baseline hazards by fitting a cubic spline to randomly selected points according to the following steps (see Section F for simulations with parametric DGPs). Figure 2 illustrates key components of the process.

Figure 2: An Example of the Random Spline DGP



(a) Graph Points

(b) Cubic Spline Fit

(c) Transform to PDF

Note: Panel (a) shows an example of 10 randomly-drawn time points (steps #1–3). Panel (b) gives the cubic spline fit to those points (step #4). Panel (c) shows the transformation from the cubic spline to a valid PDF (step #5).

1. We create a time index that counts integers from 1 to 100. This index serves as the *x*-axis for the randomly generated baseline hazard function.

2. We draw 10 points on this graph, as illustrated in panel (a) of Figure 2. The *x*-coordinates for two of the points are 1 and 100, and we randomly draw *x*-coordinates for the other 8 points without replacement. For example, for the illustration in Figure 2, points are chosen to occur at 1, 6, 12, 20, 40, 51, 55, 71, 85, and 100.

3. We randomly draw the *y*-coordinates for these points from the standard normal distribution.

4. We fit a cubic smoothing spline to the 10 points. Panel (b) of Figure 2 shows an example.

5. Finally, we apply two transformations to this function to produce a valid PDF. First, we pass the *y*-values to the standard normal PDF and take the densities. This transformation ensures that the function is non-negative. Second, we divide the *y*-values by their sum to ensure that the function integrates to 1. This final step in the generation of a baseline hazard function is illustrated in panel (c) of Figure 2.

E.2 Drawing Simulated Durations

Having generated a baseline hazard function, our next challenge is to generate individual durations from this function in a way that depends on covariates. To that end, we follow the functional form of the Cox model by using the following steps:

1. We generate a cumulative baseline hazard function by taking the cumulative sum of the baseline hazard.

2. We then create a baseline survivor function from the formula

$$H_0(t) = -\log(S_0[t])$$

by exponentiating the negative cumulative baseline hazard.[16]

3. We randomly generate three covariates (column vectors of length N denoted X_1, X_2, and X_3), three coefficients (scalars denoted β_1, β_2, and β_3), from the standard normal distribution.[17] We then create a linear predictor $X\beta$ by multiplying

$$X\beta = \begin{bmatrix} X_1 & X_2 & X_3 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}.$$

4. We use the baseline survivor function and the linear predictor to construct the individual-specific survivor functions:[18]

$$S_i(t) = S_0(t)^{\exp(\mathbf{X}_i\beta)}.$$

In other words, we take the baseline survivor function to the power of each element of $\exp(X\beta)$. If, for example, the sample size is 50, then $\exp(X\beta)$ has 50 elements and the baseline survivor function is taken to the power of each of these elements to produce 50 individual-specific survivor functions.

[16] Box-Steffensmeier and Jones 2004, 14

[17] We exclude a constant because the Cox model is formulated without a constant.

[18] Box-Steffensmeier and Jones 2004, 65

5. We subtract each individual-specific survivor function from 1 and we take the first differences to obtain the individual-specific failure PDFs.

6. In order to draw a duration for each observation from each individual-specific failure PDF we multiply each PDF by 1,000 and round every value up. We then expand a list of integers from 1 to 100 by these rounded values. For example, if after multiplying by 1,000 and rounding up the first two values of the PDF become 5 and 20, then the list of integers from 1 to 100 is expanded to produce 5 copies of 1, 20 copies of 2, and so on. Finally, we draw one randomly selected element from this expanded list. The drawn element becomes the duration for the observation.

7. Finally, we randomly set a specified proportion of the observations (5, 10, or 20 per cent) to be right censored. The simulation results reported in the main text use 10 per cent censoring, and the simulation results using 5 per cent and 20 per cent censoring are reported in Section F. The censored observations are chosen at random, uncorrelated with the baseline hazard and with the linear predictor.

E.3 Calculating Simulated Marginal Effects

The procedure described above produces simulated durations drawn from a randomly-generated baseline hazard. The next task is to simulate 'true' values of a marginal effect for a change in a covariate on these durations. Unfortunately, there is not one obvious method for simulating marginal effects. We therefore use two strategies for calculating the marginal effects: one that is similar to the way the GAM approach to COX ED calculates marginal effects, and one similar to the way the NPSF approach calculates marginal effects. In practice, the two strategies yield very similar results.

E.3.1 The GAM Method

We begin by taking the exponentiated linear predictor ($\exp[X\beta]$) using the true coefficient values that we obtained in section E.2. We then rank these values from the smallest to the largest, breaking ties randomly. Next we estimate a GAM that regresses the simulated durations on these ranks.[19] We then create two new datasets, one in which the first covariate is set to 0 for every observation, $X_1 = 0$, and one in which the first covariate is set to 1 for every observation, $X_1 = 1$, and we compute the 'true' exponentiated linear predictor for each of these new datasets. We take the median of each of vector of new exponentiated linear predictors, append these medians to the vector of exponentiated linear predictors from the simulated data, and compute ranks for this augmented vector. We then use the GAM to predict durations out of sample for this vector. The difference in predicted duration for the observations from the data with $X_1 = 1$ and $X_1 = 0$ is the simulated marginal effect.

E.3.2 The NPSF Method

As with the GAM method of calculating a simulated marginal effect, we set the first covariate in $X\beta$ equal to 1 for every observation, then by setting it equal to 0 for every observation and compute two new vectors of exponentiated linear predictors. For each observation in each of the two vectors, we calculate a survivor function using

$$S_i(t_0) = S_0(t)^{\exp(X_{i,0}\beta)} \quad \text{and} \quad S_i(t_1) = S_0(t)^{\exp(X_{i,1}\beta)}, \tag{16}$$

[19] We allow the number of knots in the GAM to be determined automatically via generalized cross-validation, which is invoked in the `gam()` function in the mgcv library in R by setting the `k` parameter equal to -1. The number of knots in the GAM is a tunable parameter in the COX ED package.

where $S_0(t)$ is the baseline survivor function, $X_{i,0}$ is the covariate data for individual i in which the first covariate is fixed to 0, and $X_{i,1}$ is the covariate data for individual i in which the first covariate is fixed to 1. The result is two hypothetical survivor functions for each observation: $S_i(t_0)$ if the first covariate is 0 and $S_i(t_1)$ if the first covariate is 1. For each observation and each hypothetical survivor function, we compute expected durations using

$$E(t) = \int_0^T S_i(t)\, dt, \tag{17}$$

which we approximate using a right Riemann sum. We thus have two hypothetical expected durations for each observation. We take the differences of these expected durations and report the median across observations as the simulated marginal effect.

E.4 The Simulated Dataset

The simulated durations and the generated covariates together form a simulated dataset. We fit a Cox model (and COX ED) to the simulated data along with an exponential, Weibull, and log-normal survival model. For each model, we compute the expected duration of each observation and compare these estimates to the true durations using an RMSE statistic. To evaluate the estimated marginal effects for the parametric survival models, we use the fitted models to predict expected durations out of sample by setting the first covariate to 1 for every observation and again setting that covariate to 0 for every observation. We subtract the values with $X_1 = 0$ from those with $X_1 = 1$ and save the median of the differences. We then compare those medians to the true marginal effects discussed in Section E.3 using another RMSE statistic.

F Additional Simulations

In this section we report the results of several auxiliary simulations. In these simulations we alter the amount of right censoring in these simulated datasets to demonstrate that the results we report in the main text are not conditioned on the amount of censoring. Next, we report coverage probability statistics to assess the methods' capacity to produce the correct confidence intervals. Finally, we compare COX ED to parametric survival models by generating survival data from the parametric DGPs.

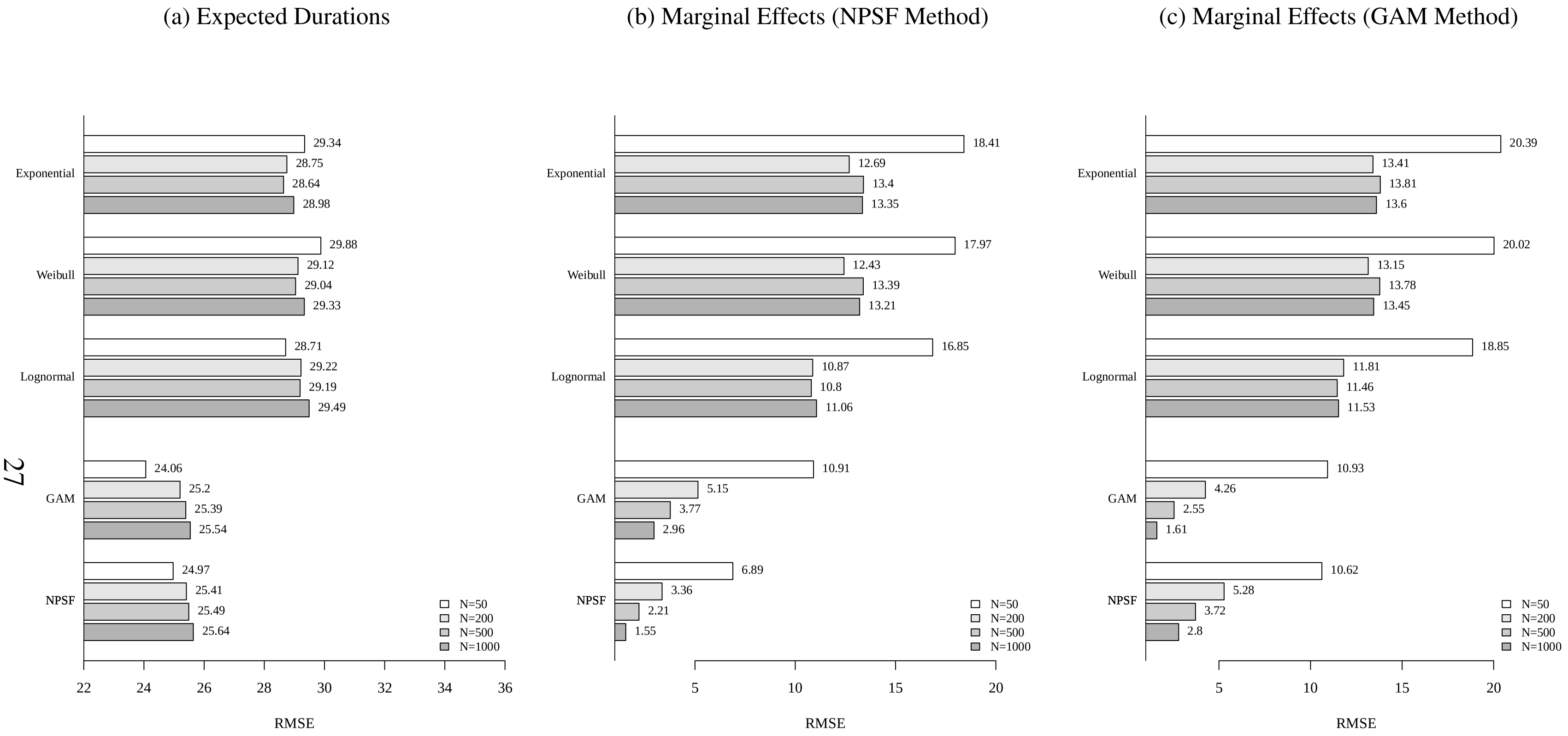
F.1 Varying Right Censoring

Figures 3 and 4 show the results for both approaches to COX ED and all three parametric models in the case of 5 per cent and 20 per cent censoring, respectively. Like the previous results, each figure contains three barplots. The left-hand plot illustrates the RMSE for each method in returning accurate expected durations. The center plot illustrates the RMSE for each method in returning the simulated marginal effects derived using the NPSF strategy, and the right-hand plot illustrates the RMSE for each method in returning the simulated marginal effects derived using the GAM strategy.[20]

For all three levels of right censoring, we find that the GAM and NPSF versions of COX ED have lower RMSEs than the parametric models for expected durations, and that all three versions of COX ED have lower RMSEs than the parametric models for marginal changes in expected duration. We again see that GAM is the strongest approach for expected durations, that NPSF is the strongest for marginal effects calculated with the NPSF strategy, and that GAM is the strongest for marginal effects calculated using a GAM. These

[20] See Section E.3 for a discussion of these two strategies for computing the simulated marginal effect.

Figure 3: Simulation Results with 5 Per Cent Right Censoring

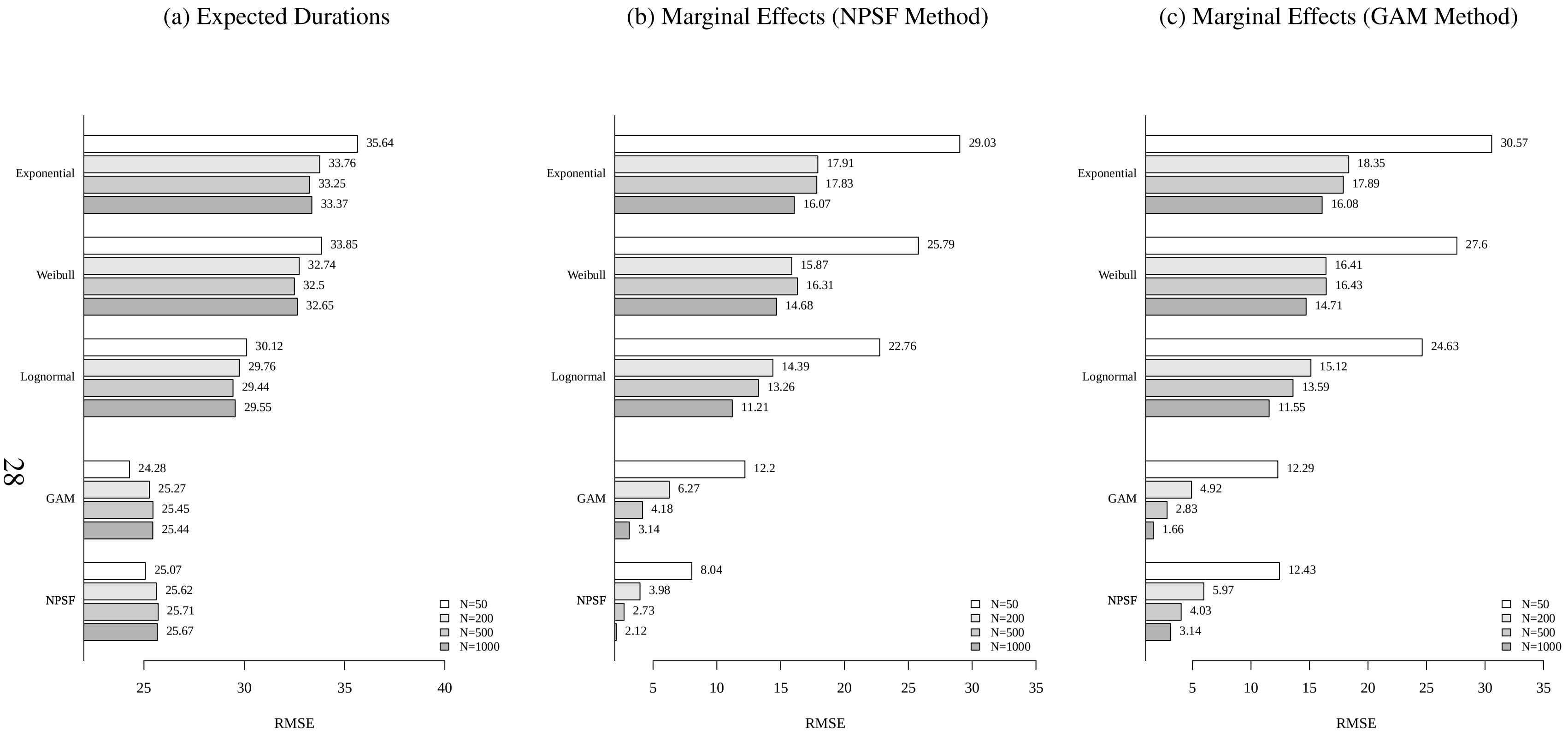

(a) Expected Durations

(b) Marginal Effects (NPSF Method)

(c) Marginal Effects (GAM Method)

Note: The graphs report RMSE results for the expected durations (panel a), marginal effects using the NPSF method to create the true marginal effect (panel b), and marginal effects using the GAM method to create the true marginal effect (panel c). At each simulation iteration, 5 per cent of the observations are randomly selected to be right-censored. Each simulation contains 1,000 iterations. Within each simulation iteration, we obtain *N* expected durations, but only one marginal change in duration. Each iteration allows us to calculate an RMSE from the *N* expected durations, and we report the mean of these RMSE statistics across simulation iterations. For the marginal effects, we can only calculate an RMSE once all of the simulation iterations are complete. In all cases smaller RMSE values indicate less error, and thus better performance.

Figure 4: Simulation Results with 20 Per Cent Right Censoring


(a) Expected Durations

(b) Marginal Effects (NPSF Method)

(c) Marginal Effects (GAM Method)

Note: The graphs report RMSE results for the expected durations (panel a), marginal effects using the NPSF method to create the true marginal effect (panel b), and marginal effects using the GAM method to create the true marginal effect (panel c). At each simulation iteration, 20 per cent of the observations are randomly selected to be right-censored. Each simulation contains 1,000 iterations. Within each simulation iteration, we obtain *N* expected durations, but only one marginal change in duration. Each iteration allows us to calculate an RMSE from the *N* expected durations, and we report the mean of these RMSE statistics across simulation iterations. For the marginal effects, we can only calculate an RMSE once all of the simulation iterations are complete. In all cases smaller RMSE values indicate less error, and thus better performance.

conclusions correspond to the findings reported in the main text.

F.2 Coverage Probability Simuations

In addition to evaluating the COX ED point estimates, we also use simulation to assess whether the methods produce the correct confidence intervals. We accomplish this with coverage probabilities.[21] At each iteration of the simulation, and with each of the COX ED methods, we compute a confidence interval around the estimated marginal effect. After the simulation is complete, we then compute, for each estimator, the proportion of iterations in which the *true* marginal effect is contained in the confidence interval. We compute 95 per cent confidence intervals in all cases, so this proportion should be (close to) 0.95 if the method is accurately capturing the true variation in the marginal effect. Values above 0.95 reflect confidence intervals that are, on average, too large and values below 0.95 indicate overconfidence.

We conduct these simulations with the random spline method described above at the three levels of censoring used previously (5, 10, and 20 per cent). We compute coverage probabilities for the GAM, and NPSF methods using both methods for computing the true marginal effect. We set the number of bootstrap iterations to 200 and the number of simulation iterations to 1,000. These simulations require a considerable amount of computation time because each iteration involves three estimators that each generate 200 bootstrap samples. Accordingly, we limit our sample sizes to $N = 50$ and $N = 200$. Table 2 presents the results.

Overall, the GAM and NPSF methods perform fairly well in these simulations, with GAM performing the best. Using the GAM strategy for the true marginal effect, the GAM

[21] See Carsey and Harden 2014, 92–3

Table 2: Coverage Probability Simulation Results

Method	N	Censoring	Coverage Probabilities	
			GAM DGP	NPSF DGP
GAM	50	5%	0.918	0.916
		10%	0.923	0.934
		20%	0.910	0.918
	200	5%	0.925	0.898
		10%	0.924	0.909
		20%	0.916	0.907
NPSF	50	5%	0.836	0.948
		10%	0.848	0.958
		20%	0.839	0.934
	200	5%	0.782	0.933
		10%	0.804	0.939
		20%	0.799	0.935

Note: Cell entries report simulation coverage probabilities for each estimator's 95 per cent confidence interval after simulating the true marginal effect with the GAM and NPSF DGPs. A value of 0.95 indicates that the method is accurately capturing the true variation in the marginal effect. Values above 0.95 reflect confidence intervals that are, on average, too large and values below 0.95 indicate overconfidence.

method coverage probabilities range between 0.910 and 0.925, while those of the NPSF method range between 0.782 and 0.848. With the NPSF strategy, the NPSF method coverage probabilities are very close to the standard of 0.95, ranging from 0.933 to 0.958. The GAM coverage probabilities are also close, ranging from 0.898 to 0.934.

F.3 Simulations from Parametric Hazard Functions

In the simulations described above, we generate simulated durations from baseline hazard functions that do not follow any particular functional parametric form. We argue that these baseline hazard functions are more realistic than common parametric functions. However, these functions may also favor COX ED over the competing survival models because the exponential, Weibull, and log-normal models are usually misspecified. To compare the estimators under ideal conditions for the parametric models, we simulate the durations from the assumed distribution of each of the parametric models, then compare each model to COX ED (GAM approach).

We conduct three simulations from parametric hazard functions. First, we generate the baseline hazard from an exponential distribution in which the rate parameter is set to $\frac{1}{\exp(X\beta)}$, and we compare the relative performance of the exponential survival model and COX ED. Second, we generate the baseline hazard from a Weibull distribution in which the scale parameter is set to $\exp(X\beta)$ and the shape parameter to 5, and we consider the performance of COX ED relative to the Weibull survival model. Finally, we generate the baseline hazard from the log of the normal distribution with a mean equal to $\exp(X\beta)$ and a standard deviation equal to 1, and we compare COX ED and the log-normal survival model. In all of the parametric simulations we set the proportion of right censoring to 10 per cent of the observations.

In Table 3 we present the results for the three parametric models as ratios over the RMSE for COX ED. Within each simulation iteration we obtain N expected durations from each method, but only one marginal change in duration. Thus, in each iteration we calculate an RMSE for each method from the N expected durations. We report the mean of the ratios of these RMSE statistics across simulation iterations. For the marginal effects, we can only calculate an RMSE for each method once all of the simulation iterations are complete. Thus, we simply report the ratios of those statistics. In all cases ratios that are greater than 1 favor COX ED because the RMSE (or average RMSE) for COX ED is greater than the

RMSE for the parametric model.

Table 3: Comparison of Expected Duration RMSE and Marginal Change in Expected Duration RMSE with the Parametric DGPs

Model	Sample Size	Expected Durations		Marginal Δ
		Average Ratio	% Ratios > 1	Ratio
Exponential	50	1.113	77%	0.856
	200	1.031	67%	0.745
	500	1.004	64%	0.740
	1,000	0.995	56%	0.820
Weibull	50	1.088	69%	0.514
	200	0.953	55%	0.588
	500	0.886	42%	0.420
	1,000	0.864	38%	0.395
Log-normal	50	1.131	96%	0.936
	200	1.080	100%	0.322
	500	1.065	100%	0.450
	1,000	1.054	100%	0.241

Note: Cell entries report the ratio of the parametric models' RMSE to the COX ED (GAM approach) RMSE for each model/sample size combination with the parametric DGPs. Values less than 1 indicate better performance by the parametric models. Values greater than 1 indicate better performance by COX ED. The first two columns of results summarize the ratios of the expected duration RMSEs: the average ratio and the proportion greater than 1. The third column of results gives the ratios of the marginal effect RMSE. The proportion of observations that are right censored is fixed at 10 per cent.

The expected duration RMSEs show that COX ED (GAM approach) and the exponential model are roughly similar in performance when the DGP is exponential. The expected durations slightly favor COX ED , while the marginal effect RMSE ratios indicate that the exponential model is somewhat better. The Weibull model results show a stronger pattern. When the true DGP comes from the Weibull distribution, the Weibull model outperforms COX ED with respect to recovering marginal changes in duration, and with respect to recovering expected durations when the sample size is 200 or larger. The log-normal results are somewhat different. In that case COX ED produces the smaller expected duration RMSEs across the four sample sizes. However, the marginal effect RMSE ratios show that the log-normal model outperforms COX ED.

Overall, these results show that when the baseline hazard comes from a known distribution (an unlikely situation in applied research), the corresponding parametric survival model's performance relative to COX ED improves. The parametric models consistently recover the marginal effect with lower RMSE. However, in several instances COX ED (GAM approach) still performs nearly equal or better than the parametric models by our RMSE criteria in computing expected durations for each observation. This 'over-performance' of COX ED is evident at the smaller sample sizes. It may be that the the flexibility of the GAM is overfitting the sample when there is less data relative to the parametric assumption, but the parametric fit improves as the sample size increases.

G Additional Replications

Here we present replications of two more published studies from political science: Box-Steffensmeier[22] and Mattes and Savun.[23]

G.1 GAM Approach with TVC: Box-Steffensmeier 1996

Box-Steffensmeier examines whether U.S. House incumbents' ability to raise campaign funds can effectively deter quality challengers from entering the race. The theoretical expectation is that as incumbents raise more money, challengers further delay their decision to run for the incumbent's seat. She employs data on 397 House races in the 1989–90 election cycle to test this hypothesis.

[22] Box-Steffensmeier 1996

[23] Mattes and Savun 2010

The dependent variable in this analysis is the number of weeks after January 1, 1989 when a challenger entered the race. Races in which no challenger entered are coded as the number of weeks after January 1 when the state's primary filing deadline occurred, and are treated as censored. The key independent variable is the incumbent's *War chest*, or the amount of money in millions of dollars that the incumbent has in reserve at a given time. Importantly, this measure updates over the course of five Federal Election Commission (FEC) reporting periods, so it is a time-varying covariate (TVC). The theory predicts a negative coefficient on this variable, which would indicate that as the incumbent raises more money, the hazard of challenger entry declines (and the time until entry increases).

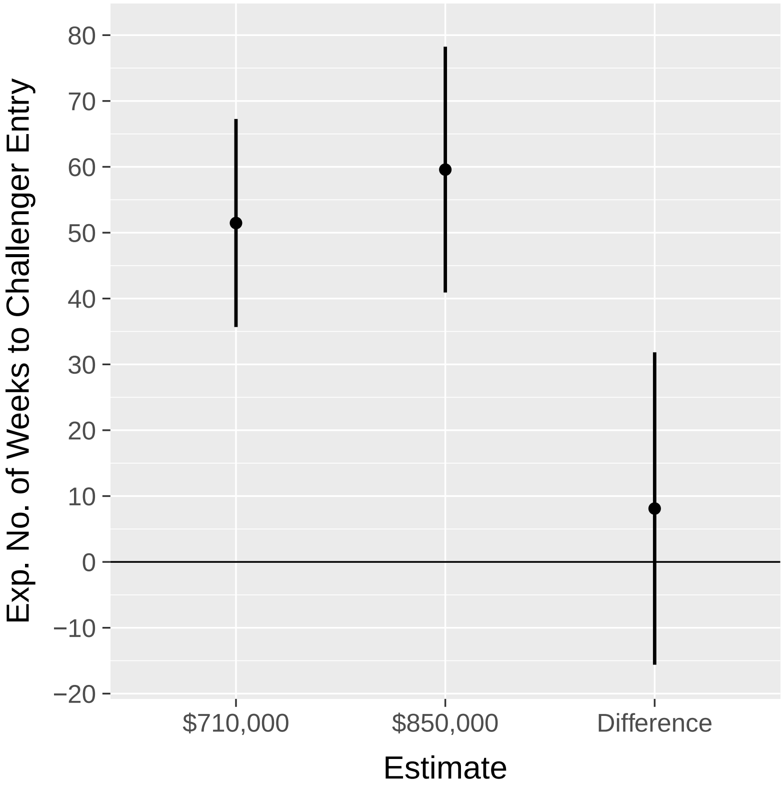
The results of the Cox model provide support. The coefficient on *War chest* is negative and statistically significant. Box-Steffensmeier explains that 'each $100,000 in an incumbent's war chest decreases the hazard of a high quality challenger entering by 16 per cent.'[24] Thus, the data indicate that building a war chest is an effective way to avoid being challenged in an election.

We employed the GAM approach—which can accommodate TVCs[25]—to give an interpretation of these results in terms of the number of weeks a challenger's entry is expected to be delayed with a change in the incumbent's fundraising efforts. We contend that this is more meaningful than the hazard-rate quantities that Box-Steffensmeier reports. Figure 5 gives the expected duration, in weeks, until challenger entry for two values of *War chest*: $710,000 (the median) and $850,000 (the 80th percentile). It also reports the difference between those two estimates.

[24] Box-Steffensmeier 1996, 365

[25] A function in our R package can perform the COX ED procedure using the counting-process data structure that TVCs require.

Figure 5: The Effect of Incumbent *War Chest* on the Expected Time Until Quality Challenger Entry into U.S. House Races (Box-Steffensmeier 1996)



Note: The graph plots the expected number of weeks until quality challenger entry for the two values of an incumbent's *War chest*—$710,000 (median) and $850,000 (80th percentile)—and the difference between the two estimates. Lines indicate 95 per cent confidence intervals.

Figure 5 supports Box-Steffensmeier's assertion that the size of an incumbent's campaign funds corresponds with a delay in challenger entry. All else equal, an incumbent with $710,000 in reserve expects to face a challenger 51 weeks from January 1 while one with $850,000 in campaign money will not be challenged until 59 weeks. However, it is important to note that there is quite a bit of uncertainty around these estimates and so the difference of 8 weeks is not statistically significant.

The relatively large confidence intervals shown in Figure 5 appear due to the fact that only a small number of challengers appear in the data, and so many observations are right censored. As a result, the GAM is fit with only 40 observations (see Figure 7). This highlights a potential drawback of the GAM approach. The GAM relies on the non-censored data, and so it will be estimated with more uncertainty if a large proportion of the observations are censored. However, because the method accounts for uncertainty from the Cox model and the GAM, this makes it susceptible to the more conservative Type II errors: failing to find a significant effect when in truth there is one. In this case, while not statistically significant, the substantive magnitude of an 8-week difference is still noteworthy. A delay of two months over the course of a campaign gives an incumbent a considerable amount of time to generate electoral support without competition.

G.2 GAM Approach with a Small Sample: Mattes and Savun 2010

Our final replication study is Mattes and Savun's analysis of the duration of civil war peace agreements. The central point the authors make is that provisions that require parties to reveal otherwise private military information can greatly increase the endurance of an agreement. Using data covering 51 civil wars from 1945–2005, they quantify the effect of peace agreements with provisions designed to reduce uncertainty between sides on the life of the agreement. These provisions include third-party monitoring, encouraging belligerents to provide troop and weapon information, and third-party verification of information.[26]

The dependent variable is the number of months a peace agreement lasted. Mattes and

[26] See Mattes and Savun 2010, 516–17

Savun model this variable as a function of several covariates: a count of the *Uncertainty-reducing provisions* in the peace agreement and control variables. They hypothesize that '[t]he greater the number of uncertainty-reducing provisions in a civil war agreement, the less likely is the recurrence of civil war between domestic belligerents.'[27] This hypothesis, which is framed in terms of risk, predicts a negative coefficient on *Uncertainty-reducing provisions*, indicating that as the number of provisions increases, the hazard of peace failure declines (longer peace times).

The Cox model results support the authors' hypothesis, producing a negative and statistically significant estimate on *Uncertainty-reducing provisions*. Mattes and Savun report that its effect is 'not only statistically significant but also substantively important.'[28] An increase from zero provisions to one provision corresponds with a 46 per cent drop in the hazard rate of peace failure and an increase from zero to three provisions decreases the hazard rate by 84 per cent. From this, they conclude that provisions that reveal information about warring parties are a useful policy prescription for the international community.

Although the authors frame their hypothesis in the language of risk, we contend that understanding the results in terms of time is still useful.The authors label a drop of 46 per cent in the hazard rate as 'substantively important.' This leads to a key question: what percentage drop would be considered *not* substantively important? Would 10 or 20 per cent be too small to indicate that *Uncertainty-reducing provisions* exerts a meaningful effect? Assessing the magnitude of effects is always arbitrary to some degree, but this issue is compounded when the scale of the effect is not meaningful. It is difficult to state whether a drop of 46 per cent really is 'large' or 'small.' Using the GAM approach, we assess the

[27] Mattes and Savun 2010, 517

[28] Mattes and Savun 2010, 521

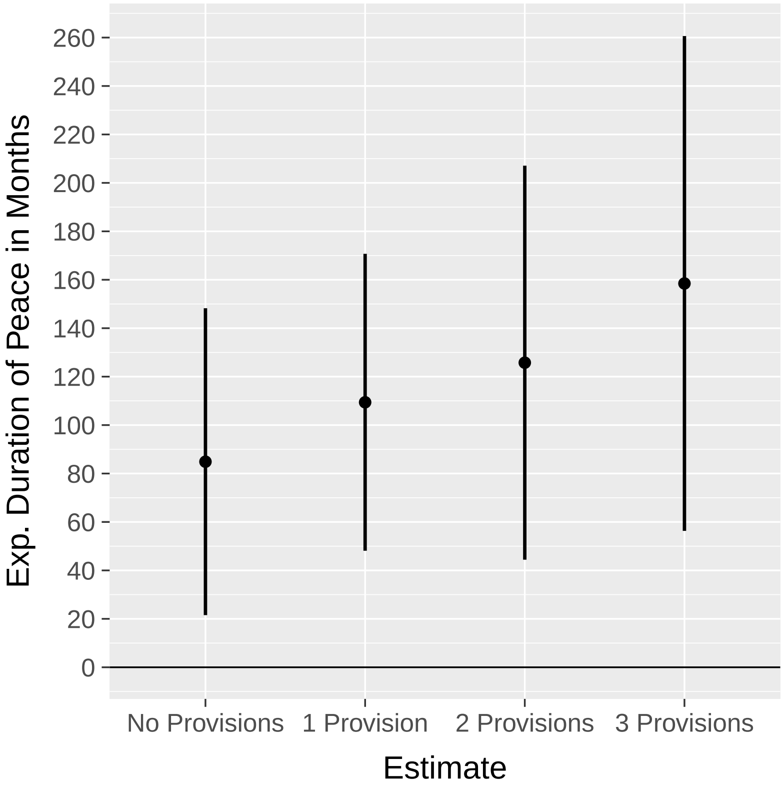
impact of *Uncertainty-reducing provisions* on a much more intuitive quantity: the amount of time peace is expected to last.

Figure 6 supports Mattes and Savun's assertion that *Uncertainty-reducing provisions* exerts a substantively important effect on the duration of civil war peace agreements, though there is a great deal of uncertainty in the estimates. Averaging over the rest of the model, an agreement with no provisions is expected to last about 89 months. Including one provision increases that estimate to about 109, or a gain of 20 months. Moving to two and three provisions brings the estimate to 126 and 158 months, respectively. However, while all of these estimates are statistically significantly different from zero, they are not statistically distinguishable from one another. This is not too surprising given the small sample of 51 cases. More importantly, the data suggest that these estimates are substantively meaningful. The expected difference between a case with no provisions and one with three provisions is 69 months, or the equivalent of moving from the 25th percentile of the observed durations to the 55th. Put differently, it represents almost six additional years of peace. Despite the large confidence intervals, these results indicate that provisions that reduce uncertainty play an important role in the life of peace agreements.

While we reach the same general conclusion as do Mattes and Savun, our analysis using COX ED provides more substantive detail on the effects of *Uncertainty-reducing provisions* on civil war peace duration. This is particularly important given that the authors' research carries important policy implications. They state that '[e]ncouraging the adoption of such uncertainty-reducing provisions in civil war settlements may be a useful policy in the international community's effort to establish peace in civil-war-torn societies.'[29] We suspect

[29] Mattes and Savun 2010, 512

Figure 6: The Effect of *Uncertainty-Reducing Provisions* on the Expected Duration of Civil War Peace Agreements (Mattes and Savun 2010)



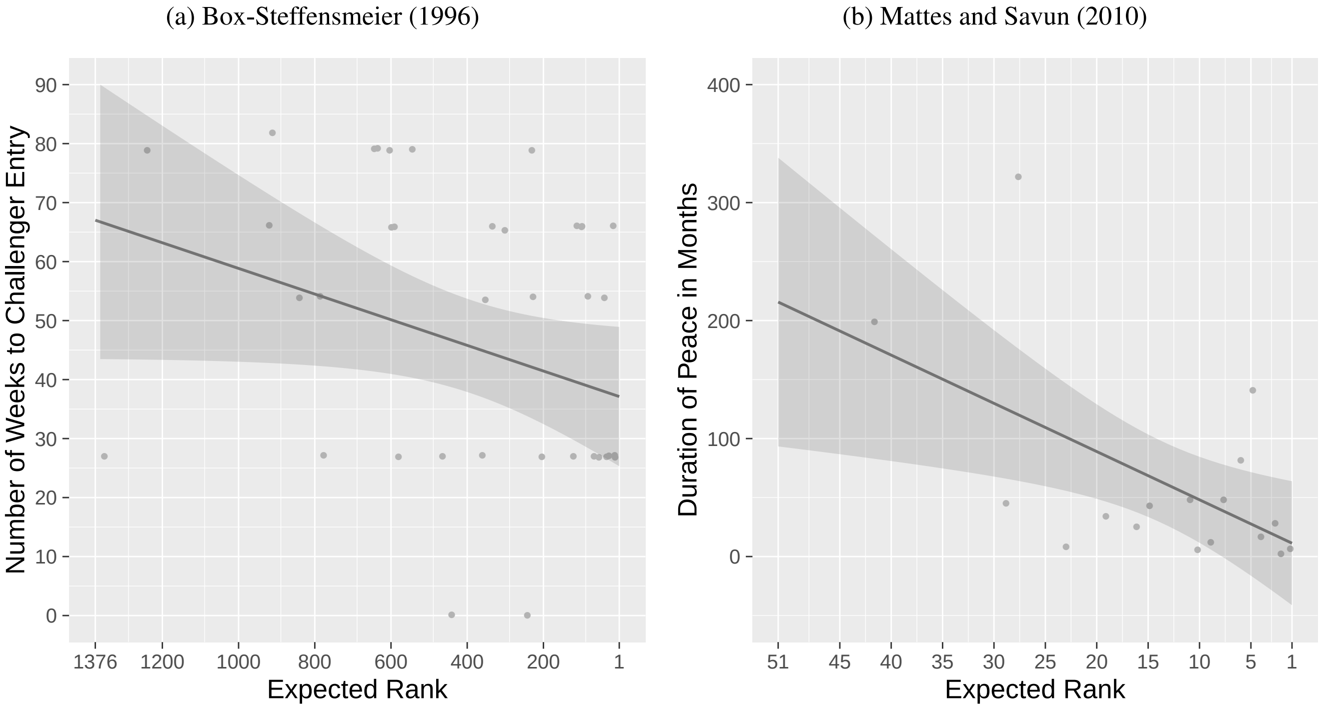
Note: The graph plots the expected peace agreement length, in months, for each observed value of *Uncertainty-reducing provisions*. Lines indicate 95 per cent confidence intervals.

that should political scientists be given the forum to formally make such recommendations, presenting evidence in terms of the expected length of peace agreements rather than relative changes in the hazard rate would be more intuitive to and make a stronger impression on policymakers.

H Replication Model GAM Fits

Figure 7 presents the Cox ED GAM fits for the Box-Steffensmeier[30] and Mattes and Savun[31] replication models. In both graphs the points represent non-censored observations, which are used to fit the GAMs.

Figure 7: Replication Model GAM Fits



(a) Box-Steffensmeier (1996)

(b) Mattes and Savun (2010)

Note: The graphs present the Cox ED GAM fits for the Box-Steffensmeier (1996) and Mattes and Savun (2010) models. Shading indicates 95 per cent confidence intervals.

[30] Box-Steffensmeier 1996

[31] Mattes and Savun 2010

References

Box-Steffensmeier, Janet M. 1996. "A Dynamic Analysis of the Role of War Chests in Campaign Strategy." *American Journal of Political Science* 40(2): 352–371.

Box-Steffensmeier, Janet M., and Bradford S. Jones. 2004. *Event History Modeling: A Guide for Social Scientists*. New York: Cambridge University Press.

Carsey, Thomas M., and Jeffrey J. Harden. 2014. *Monte Carlo Simulation and Resampling Methods for Social Science*. Thousand Oaks, CA: Sage.

Cox, David R. 1975. "Partial Likelihood." *Biometrika* 62(2): 269–276.

Desmarais, Bruce A., and Jeffrey J. Harden. 2012. "Comparing Partial Likelihood and Robust Estimation Methods for the Cox Regression Model." *Political Analysis* 20(1): 113–135.

Harden, Jeffrey J., and Jonathan Kropko. 2017. "Simulating Duration Data for the Cox Model." Working paper.

Mattes, Michaela, and Burcu Savun. 2010. "Information, Agreement Design, and the Durability of Civil War Settlements." *American Journal of Political Science* 54(2): 511–524.

Senese, Paul D., and Stephen L. Quackenbush. 2003. "Sowing the Seeds of Conflict:The Effect of Dispute Settlements on Durations of Peace." *Journal of Politics* 65(3): 696–717.

Therneau, Terry. 2013. "`survival`: A Package for Survival Analysis in S." R package version 2.37-4. http://CRAN.R-project.org/package=survival.