Reassessing Schoenfeld Residual Tests of Proportional Hazards in Political Science Event History Analyses[*]

Sunhee Park[†] David J. Hendry[‡]

Published in *American Journal of Political Science*, Vol. 59, No. 4, pp. 1072–1087.

Abstract

An underlying assumption of proportional hazards models is that the effect of a change in a covariate on the hazard rate of event occurrence is constant over time. For scholars using the Cox model, a Schoenfeld residual-based test has become the disciplinary standard for detecting violations of this assumption. However, using this test requires researchers to make a choice about a transformation of the time scale. In practice, this choice has largely consisted of arbitrary decisions made without justification. Using replications and simulations, we demonstrate that the decision about time transformations can have profound implications for the conclusions reached. In particular, we show that researchers can make far more informed decisions by paying closer attention to the presence of outlier survival times and levels of censoring in their data. We suggest a new standard for best practices in Cox diagnostics that buttresses the current standard with in-depth exploratory data analysis.

Keywords: scaled Schoenfeld residuals; proportional hazards assumption; event history analysis; replication; simulation

Running Head: Reassessing Schoenfeld Residual Tests

[*] The authors contributed equally to this work. Author ordering was chosen at random. We are grateful to Janet Box-Steffensmeier, Wendy Cho, Kyle Joyce, Philip Chen, Han Dorussen, Charles Gregory, Zein Murib, Maron Sorenson, and participants in the event history ITV course held at the Ohio State University, the University of Illinois, the University of Minnesota, and the University of Wisconsin for offering helpful comments on earlier versions of this project. And we are indebted to the current and previous editor and anonymous reviewers of the *AJPS* who provided invaluable suggestions during the review process. We also thank Jason Morgan for providing critical suggestions with respect to the simulation algorithm, as well as Kyle Beardsley, Scott Bennett, Kathleen Cunningham, Zachary Elkins, Scott Gartner, Hein Goemans, John Huber, Cecilia Martinez-Gallardo, Michaela Mattes, Maria Victoria Murillo, Stephen Quackenbush, Burcu Savun, Dan Smith, Paul Warwick, and Christopher Zorn for helpful correspondence with respect to replication materials. Thanks also go to all of the authors, too numerous to list here, who have made their data and replication code publicly available. Any and all errors are our own. Replication materials are available from the *AJPS* Dataverse (http://dvn.iq.harvard.edu/dvn/dv/ajps) and from the authors' websites.

[†] Visiting Researcher, Department of Political Science and Government, Aarhus University, Bartholins Allé 7, DK-8000 Aarhus C, Denmark, (sunhee.park@ps.au.dk)

[‡] Assistant Professor, Department of Political Science and Government, Aarhus University, Bartholins Allé 7, DK-8000 Aarhus C, Denmark, (david.hendry@ps.au.dk)

In event history analysis, the researcher's interest is typically in modeling the effects of a set of covariates on the hazard rate of event occurrence. Many of the most commonly used regression models for event history analysis (e.g., the Cox, exponential, Weibull, and Gompertz models) carry the built-in assumption that the effect on the hazard rate of a change in a covariate is constant regardless of when in the process it occurs—the so-called proportional hazards assumption. If the effect of a covariate on the hazard rate differs over the time span under study, using proportional hazards regression techniques will lead to biased coefficient estimates and suboptimal significance tests (Kalbfleisch and Prentice 2002). As political scientists have come to recognize the power of proportional hazards models for examining the duration and timing of political events, they have also become increasingly aware of the need to test and correct for violations of the proportional hazards assumption (thanks in large part to the work of Box-Steffensmeier and colleagues; e.g., Box-Steffensmeier and Jones 2004; Box-Steffensmeier and Zorn 2001).[1]

An active area of research in statistics has developed an array of techniques for detecting nonproportional hazards (Gill and Schumacher 1987; Grambsch and Therneau 1994; Lagakos and Schoenfeld 1984; Ng'andu 1997; Stablein, Carter, and Novak 1981; Therneau, Grambsch, and Fleming 1990; Winnett and Sasieni 2001). One method in particular, developed by Grambsch and Therneau (1994) for use with the Cox proportional hazards model, focuses on plotting methods and tests of linear association to examine trends in covariate-specific scaled Schoenfeld residuals over time. When these tests reveal that the proportional hazards assumption is violated, one common recommendation is to interact the offending covariate(s) with some function of time, and move forward with estimation and interpretation (Box-Steffensmeier and Jones 2004; Box-Steffensmeier and Zorn 2001). Political scientists have begun to take heed of this advice, resulting in generally more sophisticated and nuanced interpretations of the duration and timing of political events (e.g., Box-Steffensmeier, Reiter, and Zorn 2003; Chiozza and Goemans 2004; Licht 2011).

Though the increased application of diagnostic testing for the proportional hazards assumption and the use of time interactions as a corrective technique have improved event history analyses in political science, we argue in this paper that testing for nonproportional hazards is not yet

[1] To demonstrate how important proportional hazards event history analysis has become across subfields of political science, Table SI.A.1 in the Supporting Information presents the results of a content analysis of the *American Journal of Political Science*, *American Political Science Review*, and *Journal of Politics*, identifying all articles published between 1992 and 2012 that employ a proportional hazards model.

a solved problem. When scholars use the residual-based tests mentioned above, they are forced to make a choice about a transformation of the time scale, or to choose no transformation at all. Until now, however, in empirical applications these choices have largely been arbitrary, often left to the default setting of the researcher's chosen statistical software, and almost never reported (but see Box-Steffensmeier and Zorn 2001). The arbitrary nature of this choice would be relatively unimportant if it were not consequential for the conclusions reached. However, we employ replications and simulations to demonstrate that the choices made do in fact impact the ability of the diagnostic tests to detect violations of the proportional hazards assumption, and thus the subsequent choices regarding corrective measures and, ultimately, substantive interpretations.

The simplest summary of our findings is that data structure matters in determining which function of time should be employed in the Grambsch-Therneau tests of proportional hazards. Specifically, researchers must be aware of the presence of outlier survival times (a not uncommon feature of political science data) as well as the level of censoring in their data. Our analysis of replication materials from over a decade of published work reveals that whether or not scholars are aware of the need to make a choice about a transformation of the time scale when performing these tests, untransformed time and the natural log of time have been the applications of choice in political science research. However, our simulations indicate that these may in fact be the least desirable choices for the types of survival distributions most common in political science research, particularly as the level of censoring grows larger. For many common outlier and censoring scenarios, other choices (namely the rank and left-continuous Kaplan-Meier transformations, to be explained in greater detail below) will be superior.

The goal of this paper is to provide applied researchers with additional guidance on appropriate diagnosis of violations of the proportional hazards assumption in applications of the Cox model. We advocate an approach to detecting proportional hazards in which researchers employ the diagnostic procedures developed by Grambsch and Therneau (1994) and recommended by Box-Steffensmeier and colleagues (Box-Steffensmeier and Jones 2004; Box-Steffensmeier and Zorn 2001; Box-Steffensmeier, Reiter, and Zorn 2003), but supplement these techniques with in-depth exploratory data analysis. In a set of simulations and replications of published research, we demonstrate how information about outlier survival times and censoring can be used to choose an appropriate transformation of the time scale during diagnostic testing, without the need for removing or correcting for outlier survival times during estimation of Cox parameters. Our findings show that knowledge of the very basic elements of one's data—particularly with respect to outlier survival times and censoring—will lead to more informed decisions during diagnostic testing, and thus more accurate substantive conclusions.

Detecting Nonproportional Hazards

The importance of the proportional hazards assumption in appropriate specification of a large class of event history models has spawned an extensive literature on appropriate means of testing for nonproportional hazards (see Ng'andu 1997, for a review). As Box-Steffensmeier and Zorn (2001) point out, statistical tests of the proportional hazards assumption fall into three general classes: (1) tests focusing on piecewise estimation of models for subsets of data defined by stratification of time; (2) tests focusing on interactions between covariates and some function of time; and (3) tests based on examinations of regression residuals. A variety of tests have been recommended within each class. In this paper we focus exclusively on a popular diagnostic method falling into the third class that examines the relationship between scaled Schoenfeld residuals and time.[2]

Scaled Schoenfeld Residuals and Proportional Hazards

The basic logic behind scaled Schoenfeld residual tests for proportional hazards is quite intuitive, and can be seen as a natural extension of methods of examining residuals in the linear regression framework.[3] To begin, let $Z_{ij}(t)$ be the jth covariate of the ith unit, where $i = 1, \ldots, n$, $j = 1, \ldots, p$, and the notation indicates that Z_{ij} is allowed to vary as a function of the time scale. Then the Cox proportional hazards model assumes that the hazard rate for the ith individual

[2] We have chosen this particular focus due to its increasing popularity within political science. However, we alert readers to the fact that it is not the only means of testing for proportional hazards, and that its application is limited to the Cox model. For a recent argument showing that nonproportionality can be tested and modeled within the Weibull framework as well, see Zuehlke (2013). We thank an anonymous reviewer for pointing us to this reference.

[3] The discussion in this section draws substantially on work presented elsewhere (Box-Steffensmeier and Jones 2004; Box-Steffensmeier and Zorn 2001; Grambsch and Therneau 1994; Therneau and Grambsch 2000), but provides a highly condensed argument due to space considerations. Readers interested in more technical detail are referred to Therneau and Grambsch (2000, esp. chs. 4 and 6).

satisfies the following relationship:

$$h_i(t) = h_0(t) \exp\left(Z_i(t)\beta\right), \tag{1}$$

where h_0 is the baseline hazard, $Z_i(t)$ is a $1 \times p$ vector of covariates for unit i, each of which can be time-fixed or time-varying, and β is a $p \times 1$ vector of coefficients. Therneau and Grambsch (2000) set up the rationale for a residual test by introducing an alternative to the Cox model in which the estimated coefficient is also allowed to vary as a function of time. That is,

$$h_i(t) = h_0(t) \exp\{Z_i(t)\beta(t)\}. \tag{2}$$

Examining (1) and (2), when $\beta(t) = \beta$, proportional hazards is implied. An explicit test of this restriction involves analysis of model residuals. Regression models for time-to-event data require more thought about the meaning of a residual because observations may be censored. The Cox model adds an additional complication in that the baseline hazard is not estimated (Cox 1972), and hence the fitted model does not provide a systematic component (Hosmer, Lemeshow, and May 2008). One can, however, derive the score process for each individual unit under study. For each unit i, and for any given time t, the score process is essentially a row vector of differences between the covariate values for individual i and a weighted mean of the covariate values for all individuals at risk at time t. Schoenfeld (1980) proposed a residual derived by summing the score processes over units experiencing the event of interest at each unique event time. The simplest representation of the Schoenfeld residual is one in which there are no tied event times. Following the notation of Therneau and Grambsch (2000), we define the risk score for unit i at time t as $r_i(t) = \exp[Z_i(t)\beta]$, and we define $Y_i(t)$ as an indicator function such that $Y_i(t) = 1$ if unit i is under observation and at risk, and 0 otherwise. Then at the kth event time, t_k, the Schoenfeld residual is given by

$$\begin{aligned} s_k &= Z_{(k)} - \frac{\sum_i Y_i(t_k) r_i(t_k) Z_i(t_k)}{\sum_i Y_i(t_k) r_i(t_k)} \\ s_k &= Z_{(k)} - \bar{z}(\hat{\beta}, t_k), \end{aligned}$$

where $Z_{(k)}$ is the covariate vector of the unit experiencing the event at time k, $\hat{\beta}$ is the estimate of β based on maximization of the partial likelihood function,[4] and $\bar{z}(\hat{\beta}, t_k)$ acts as a weighted mean of the covariate values for all units at risk at time t.[5] Additionally, we can define the weighted variance of Z at the kth event time as $V(\beta, t_k) = \{\sum_i Y_i(t_k) r_i(t_k)[Z_i(t_k) - \bar{z}(\beta, t_k)]'[Z_i(t_k) - \bar{z}(\beta, t_k)]\}/[\sum_i Y_i(t_k) r_i(t_k)]$. Then, scaling the Schoenfeld residuals by the weighted variance of X at the kth event time yields the scaled Schoenfeld residual:

$$s_k^* = V^{-1}(\hat{\beta}, t_k) s_k.$$

Grambsch and Therneau (1994; see also Therneau and Grambsch 2000) show that $E(s_{kj}^*) + \hat{\beta}_j \approx \beta_j(t_k)$. Therefore, the restriction for proportional hazards, $\hat{\beta}_j = \beta_j(t_k)$, implies that $E(s_{kj}^*) = 0$, which occurs if the s_{kj}^* values are a random walk across the time scale.[6] This leads naturally to a calculation of the relationship between s_{kj}^* and t_k, or some function $g(t_k)$, and to plots of $s_{kj}^* + \hat{\beta}_j$ against t_k or $g(t_k)$ as a means of diagnosing and visualizing the presence and nature of any nonproportionality. Therneau and Grambsch (2000) suggest a linear regression of s_k^* on $g(t_k)$, and they motivate their suggested test by appealing to a heuristic approach rooted in generalized least squares. Letting $\bar{g}$ be the mean of $g(t_k)$ and d the number of event times such that $k = 1, \ldots, d$, the least squares slope of such a regression for the jth covariate is given by

$$\tilde{\theta}_j = \frac{\sum_{k=1}^{d} \left(g(t_k) - \bar{g}\right)\left(s_{kj}^* - \bar{s}_j^*\right)}{\sum_{k=1}^{d} \left(g(t_k) - \bar{g}\right)^2} = \frac{\sum_{k=1}^{d} \left(g(t_k) - \bar{g}\right) s_{kj}^*}{\sum_{k=1}^{d} \left(g(t_k) - \bar{g}\right)^2},$$

where the final equality holds because, by definition, $\sum_{k=1}^{d} s_k = 0$. Denote the information matrix of the partial likelihood estimation of β as $\mathcal{I}(\hat{\beta}) \equiv \mathcal{I}$, and note that $\sum_{k=1}^{d} V(\hat{\beta}, t_k) = \mathcal{I}$. Because the values of $V(\hat{\beta}, t_k)$ can become relatively unstable late in the observation time as the

[4] Knowledge of the details regarding estimation of β are useful but not necessary for the discussion that follows. Space considerations prevent us from presenting a full derivation of $\hat{\beta}$. Interested readers are referred to the original paper by Cox (1972) as well as chapter 3 of Therneau and Grambsch (2000) and chapter 4 of Box-Steffensmeier and Jones (2004).

[5] In the case of tied data, the Schoenfeld residual for an event time is given by $s_k = \int_{t_{k-1}}^{t_k} \sum_i [Z_i - \bar{z}(\hat{\beta}, s)] dN_i(s)$, where $N_i(s)$ is the number of observed event times for unit i. However, most computer software for event history analysis simply assumes no ties and returns individual residual values for each unit experiencing the event at a particular time (Therneau and Grambsch 2000).

[6] It should be noted that phenomena other than random walks can also produce this relationship, including certain nonlinear trends. See Keele (2010).

number of units in the risk set diminishes, Therneau and Grambsch (2000) suggest substituting $\sum_{k=1}^{d} V(\hat{\beta}, t_k)$ with its average value, $\mathcal{I}/d$. Letting $\mathcal{I}^{jk} = \mathcal{I}_{jk}^{-1}$ be the (j,k)th element of $\mathcal{I}^{-1}$, and using the result that $\mathrm{Var}(s_k^*) \approx V^{-1}(\beta, t_k)$ (Grambsch and Therneau 1994), we have that $\sum_{k=1}^{d} \mathrm{Var}(s_{kj}^*) \approx d\mathcal{I}^{jj}$. Then the test statistic for the proportional hazards assumption with respect to the jth covariate is given by

$$T_j = \frac{\left[\sum_{k=1}^{d} \left(g(t_k) - \bar{g}\right) s_{kj}^*\right]^2}{d\mathcal{I}^{jj} \sum_{k=1}^{d} \left(g(t_k) - \bar{g}\right)^2} \tag{3}$$

and is asymptotically distributed as $\chi^2(1)$ under the null hypothesis that the relationship between covariate j and the event times follows the proportional hazards assumption.[7]

Carrying this logic forward, Grambsch and Therneau (1994) also suggest a global test for proportional hazards over all p covariates. If we let S be the $d \times p$ matrix of unscaled Schoenfeld residuals, $S^* = dS\mathcal{I}^{-1}$ the matrix of scaled Schoenfeld residuals under the assumption of constant variance, and g^* the $d \times 1$ vector whose kth element is $g(t_k) - \bar{g}$. Then the test statistic for the global test is given by

$$T = \frac{g^{*\prime} S^* \mathcal{I} S^{*\prime} g^*}{d \sum_{k=1}^{d} \left(g(t_k) - \bar{g}\right)^2} \tag{4}$$

and is asymptotically distributed as $\chi^2(p)$ under the null hypothesis that the relationship between the combination of the p covariates and the event times follows the proportional hazards assumption. In summary, the covariate-specific test is a test of the null hypothesis that the impact of covariate j on the hazard rate violates the proportional hazards assumption. The global test, on the other hand, is a test of the null hypothesis that the combined effect of all covariates in the model violates the proportional hazards assumption.

Since their recommendation to political scientists by Box-Steffensmeier and colleagues (Box-Steffensmeier and Jones 2004; Box-Steffensmeier and Zorn 2001; Box-Steffensmeier, Reiter, and Zorn 2003), exploration of scatterplots of s_{jk}^* versus $g(t_k)$, as well as calculations of their correlations and the test statistics presented in (3) and (4) have become the standard means of evaluating the proportional hazards assumption in applications of the Cox model to political phenomena,

[7] Space considerations prevent us from presenting full derivations of the estimator and test statistic. Detailed arguments are presented by Grambsch and Therneau (1994) and Therneau and Grambsch (2000, ch. 6). In Section B of the Supporting Information, we provide a brief sketch of these arguments.

and their application has increased over time (Berlinski, Dewan, and Dowding 2010; Berry, Burden, and Howell 2010; Chiozza and Goemans 2004; Crescenzi 2007; Debs and Goemans 2010; Diermeier and Stevenson 1999; Gibler and Tir 2010; Koch and Sullivan 2010; Leeds and Savun 2007; Maeda 2010; Maltzman and Shipan 2008; Murillo and Martínez-Gallardo 2007; Schleiter and Morgan-Jones 2009).

Choice of Time Transformation in Scaled Schoenfeld Residual Tests

Though the general increase in awareness regarding the proportional hazards assumption is certainly encouraging, testing for proportional hazards within the general framework put forth above is not yet a solved problem. At least one outstanding issue, which we identify in this paper, is the specification of $g(t)$, the function of time against which to compare the scaled Schoenfeld residuals. The primary purpose of using a transformed version of the time scale rather than its identity for diagnostic testing is to avoid potential problems with outlier survival times for units that experience the event of interest (Therneau and Grambsch 2000). After all, the scaled Schoenfeld residual procedure for detecting nonproportionality is a test of linear association between two variables, subject to all of the well known issues surrounding influential data points (see, e.g., Cook 1979; Cook and Weisberg 1982; Weisberg 2005).[8]

To demonstrate why a transformation of the time scale may be necessary when testing for nonproportionality using the scaled Schoenfeld residual method, Figure 1 presents two examples of scaled Schoenfeld residual plots of single covariates from published Cox specifications in the political science literature. The left panel, from Cunningham (2011), and the right panel, from Maeda (2010), illustrate a relatively common data structure in political science research, namely, long-tailed survival distributions. Specifically, in both datasets, most units experiencing the event have survival times in the lower third of the distribution, and a very small handful of cases have survival times in the upper third of the distribution. If these cases with relatively long survival times also happen to deviate substantially from the typical case with respect to $E\left(s^{*}_{kj}\right)+\hat{\beta}_j$, they have the potential to exert unwarranted influence on the formal test of linear association used to

[8] For a discussion of a different set of issues stemming from the fact that the scaled Schoenfeld residual procedure is specifically a test of *linear* association, see Keele (2010).

identify the covariates that violate the proportional hazards assumption.[9]

[Figure 1 about here.]

One standard corrective technique that has been recommended is the use of a transformation of the time scale to reduce the effect of outliers (Therneau and Grambsch 2000). The criterion for an acceptable candidate transformation function is simply that it maintains the ordering of event times in the empirical distribution. Therefore any monotonic function of time can be used, and standard statistical packages come with a series of built-in choices (Cleves et al. 2010; Therneau 1999). Here, we focus our attention on the four choices available in the most commonly used statistical software packages for event history analysis in political science: identity (untransformed) t, the natural log of t, the rank of t (i.e., the observed event times placed in integer-rank order, $1, 2, \ldots, t$), and the left-continuous version of the Kaplan-Meier survival curve of t (i.e., $1 - KM(t-)$; see Kaplan and Meier 1958 and Therneau and Grambsch 2000).[10,11] In the following sections, we report the results of simulations and replications of published research indicating that the choice of a transformation is consequential.

Simulations

In this section we conduct simulations to demonstrate the performance of the Grambsch-Therneau tests of proportional hazards under the various time transformations available in the most popular statistical software for event history analysis. We took several steps to generate simulated data that mimic the types of data structures frequently encountered by political scientists. First, unlike other event history simulations in political science research, we generated simulated data with

[9] When characterized this way, the first obvious issue that arises is the choice of criteria to determine that a unit deviates substantially from the typical case. In Section C of the Supporting Information, we present an example of a common formal test used to determine the presence of outliers that can be used to aid in decision making.

[10] A derivation of the Kaplan-Meier estimator and additional information about how it is used as a transformation of the time scale are provided in Section D of the Supporting Information.

[11] After identifying all articles using proportional hazards event history models published in the *American Journal of Political Science*, *American Political Science Review*, and *Journal of Politics* published between 1990 and 2012 (see Table SI.A.1), we also searched for all replication materials for those same articles by gathering publicly available information and contacting individual authors. Among the articles for which we were able to obtain replication materials, we found that 100% published after 2000 used either Stata or R for estimation. The `estat phtest` function in Stata (Cleves et al. 2010) and the `cox.zph` function in the Survival package for R (Therneau 1999) each provide the same four options examined here as possible transformations of the time scale. Stata also provides an option to incorporate a user-defined transformation.

time-varying covariates. To our knowledge, all published event history simulations in political science generate units with time-fixed covariates, yet the vast majority of empirical applications examine units with time-varying covariates. Second, since most empirical applications use data with measurements taken at fixed time intervals (e.g., days or years), we ensured that our simulated survival times were a function of covariates that varied at integer-valued steps of the time scale. Third, we used a mean shift model (e.g., Weisberg 2005) to generate outlier survival times, which are a common feature of data used in published research. Finally, we generated survival times as a function of both a binary and a continuous variable, as the vast majority of published event history analyses utilize a mixture of covariate types.

To generate survival data with time-varying covariates that vary at integer-valued steps of the time scale, we used the method presented in Hendry (2014), augmented slightly to include outlier survival times and violations of the proportional hazards assumption. Specifically, the method generates data that follow a Cox model with time-varying covariates by using a transformation of a truncated piecewise exponential random variable.[12] Bounds were chosen so that units would have minimum and maximum survival times of 10 and 150, respectively. Incorporating violations of the proportional hazards assumption, the hazard rate for the ith unit in our simulated data can be presented as the following augmented Cox specification:

$$h_i(t) = h_0(t) \exp\left(\beta_1(t) Z_1(t) + \beta_2(t) Z_2(t)\right)$$

where $h_0(t)$ is the baseline hazard, $t = 1, 2, \ldots, T_i$, where T_i is the survival time for unit i, $Z_1(t) \sim Uniform[-.5, .5]$, and $Z_2(t) \sim Binomial(.5)$. The time-varying form of β_1 and β_2 indicates how violations of the proportional hazards assumption were incorporated. In all of the

[12] All simulations were performed using R 2.15.1 (R Development Core Team 2012) with a *Mersenne-Twister* random number generator on a machine with an Intel Xeon 2.26 GHz processor using Windows 7 64-bit. Piecewise exponential random variables were generated using a suite of functions in the `msm` package (Jackson 2014), truncated piecewise exponential random variables were generated using rejection sampling, Cox parameters were estimated using the `coxph` function with the Efron method for handling tied data, and diagnostic testing of the proportional hazards assumption was performed using the `cox.zph` function in the `survival` package (Therneau 1999). Replication code and all results for the simulations are available from the authors' websites.

simulations presented below, these parameters were defined as follows:

$$\beta_1(t) = \begin{cases} .1, & \text{if } t < 10; \\ 1, & \text{if } 10 \leq t < 15; \\ 2, & \text{if } 15 \leq t < 20; \\ 3, & \text{if } t \geq 20; \end{cases} \qquad \text{and} \qquad \beta_2(t) = \begin{cases} -5, & \text{if } t < 10; \\ -3, & \text{if } 10 \leq t < 15; \\ -1, & \text{if } t \geq 15. \end{cases} \tag{5}$$

In other words, violations of the proportional hazards assumption were incorporated by allowing the Cox parameters to vary at step functions of the time scale, such that changes in covariates have varying impact on the hazard rate depending on when in the process those changes occur.

After generating the simulated data according to the above specifications, we added cases with outlier survival times to each simulated dataset by randomly selecting five units and adding a value to those units' survival times that was randomly chosen to fall between zero and the median survival time. For values of this random draw that are close to the median, this addition of time represents a non-trivial quantity. Furthermore, the choice to add random draws of time to five cases ensures that we achieve a range of outlier patterns across our simulated datasets.[13]

We then chose a censoring distribution by first defining a desired proportion of censored cases (either .5, .25, or .1), and then determining whether the probability of being censored would be uniform across units, or whether units with relatively long or relatively short survival times were more likely to be censored. Specifically, we defined the proportion of censored cases by first generating a vector of censoring indicators. In the case of uniform censoring, indicators were uniformly distributed across cases. For the situations in which relatively long (short) survival times were more likely to be censored, cases in the upper (lower) quartile of survival times were more likely to be censored than cases in the lower (upper) three quartiles.[14] The situation in which units with relatively long survival times are more likely to be censored might represent a common empirical setting in which all units come under observation at the same time, and

[13] In Section F of the Supporting Information, we present a selection of scatterplots of the scaled Schoenfeld residuals versus functions of time, which allows for visualization of outlier survival times. Interested readers can consult the replication materials to produce these same scatterplots for any of the simulations.

[14] A detailed description of the algorithm used to generate the simulated data is presented in Section E of the Supporting Information.

the researcher stops collecting measurements at some defined end time. For instance, in a study of state policy adoption from some predefined starting point to some predefined ending point, those states that never adopt the policy in the specified time frame would both be censored and have the longest survival times. The situation in which units with relatively short survival times are more likely to be censored might represent another common empirical setting consisting of staggered entry of units with a defined end time of observation. For example, in a study of leadership survival during a specific time period, leaders who come into office toward the end of the period would both be censored and have relatively short survival times. Uniform censoring could represent empirical settings in which units have multiple modes of exit from the data, but the researcher is only interested in one mode. For each of these censoring distributions, in addition to a setting without censoring, we generated 1000 simulated datasets with 500 units each, estimated Cox parameters,[15] and performed scaled Schoenfeld residual tests using the four different time transformations discussed previously. Table 1 presents a summary of the performance of the tests.

[Table 1 about here.]

Using a threshold p-value of .05, if the Grambsch-Therneau scaled Schoenfeld residual tests are performing as intended, we expect them to correctly detect a violation of proportional hazards in about 950 simulations out of 1000. Looking across the different settings summarized in Table 1, we find that there is actually substantial variation in the desired behavior. As expected, for any given censoring distribution and proportion of cases censored, and for both the covariate-specific and global tests, those that employ the untransformed version of the time scale detect the lowest number of violations among the four choices, and never detect a number of violations within the expected range for the chosen p-value threshold. This should be the case given that we specifically produced data with outlier survival times and the various transformations of the time scale have been suggested as means of correcting for the presence of outliers. Additionally, within each time transformation and censoring distribution, the tests are more likely to detect a violation of the proportional hazards assumption in the case of Z_2 than in the case of Z_1. This is also expected, given that we defined β_2 to vary slightly more dramatically than β_1 (see (5)). Further, for all

[15] Summaries of Cox parameter estimates for the simulated data are not presented due to space constraints. Interested readers should consult the replication materials.

three of the transformed versions of the time scale and for any censoring pattern, the number of violations detected by the global tests always falls within the expected range, with the exception of $\ln(t)$ and a uniform censoring proportion of .5, which is on the margin. And whether censoring is uniform, biased toward long survival times, or biased toward short survival times, the tests are more likely to detect violations when the level of censoring is relatively low. This too is not surprising given that Schoenfeld residuals are defined only at event times.

Interestingly, for each censoring distribution, proportion of cases censored, and type of covariate examined, the rank transformation outperforms all other tests in terms of correctly detecting violations of proportional hazards. In cases of low censoring, the Kaplan-Meier transformation performs just as well or almost as well as the rank transformation, but the difference between the two becomes larger in cases of higher censoring. In fact, in the case of no censoring the results of the tests using the Kaplan-Meier and rank transformations are virtually identical. The divergence between the two as the censoring proportion becomes larger also makes sense, as the left-continuous Kaplan-Meier function is the only transformation examined here that explicitly incorporates information from censored cases (specifically, by incorporating the number of units still in the risk set at each event time; see Section D of the Supporting Information). However, regardless of the censoring distribution, when the proportion of censored cases reaches .5, though the rank transformation detects the most violations, in almost no case does a test detect the number of violations in the expected range for the p-value threshold.

The important implications for political science research come when we combine the relatively strong performance of the tests using the rank transformation with the relatively poor performance of the tests using the log transformation. First, for political scientists who are aware of the problem of outlier survival times for detecting violations of proportional hazards using the Grambsch-Therneau method, the natural log seems to have been the transformation of choice (e.g., Box-Steffensmeier and Zorn 2001; Chiozza and Goemans 2004). The origin of the popularity of this choice likely has to do with the ubiquitousness of the log transformation in a variety of other applications in political science research, including its use as a correction for nonlinearity in bivariate relationships and as a transformation of time when researchers interact covariates with time to explicitly model nonproportionality within the Cox framework (e.g., Box-Steffensmeier and Jones 2004; Box-Steffensmeier and Zorn 2001; Chiozza and Goemans 2004; Debs and Goemans 2010; Licht 2011; Maeda 2010; Maltzman and Shipan 2008). The simulation results presented here suggest that for at least some types of data structures, the log transformation may actually be the least desirable corrective measure, particularly in the presence of moderate to heavy censoring. Second, political scientists who are aware of the need to test for proportional hazards in applications of the Cox model, but unaware of the need to account for outlier survival times, find themselves at the mercy of the default settings of their chosen statistical software. For instance, users of Stata will have identity time as their default (Cleves et al. 2010), while users of the `survival` package in R will have the Kaplan-Meier transformation (Therneau 1999). As we have noted, the use of identity time is generally inappropriate in the presence of outlier survival times, and therefore practitioners with long-tailed survival distributions should strongly consider the value of a time transformation. For relatively low levels of censoring, the choice of a transformation to correct for outliers may only be consequential on the margins. However, having 50% or more of cases censored is a not uncommon feature of political science event history data (e.g., Maeda 2010), and it is therefore an issue of which researchers employing proportional hazards models should be aware.

The evidence from the simulations initially suggests that the tests employing the rank transformation are the most likely to detect violations of proportional hazards in the presence of a small handful of outlier survival times and heavy censoring. However, the recommendation by Grambsch and Therneau (1994; see also Therneau and Grambsch 2000) is that the formal tests of statistical significance be used in conjunction with graphical displays of the relationship between scaled Schoenfeld residuals and time. We take this a step further to argue that examination of graphical displays of scaled Schoenfeld residuals can also be instructive in the decision about an appropriate transformation. For these simulations, examination of plots for each of the covariates across each of the 1000 simulated datasets for each censoring distribution would be infeasible.[16] In addition, though we have taken some effort to generate simulated data that cover a wide range of common empirical circumstances encountered by political scientists, we certainly have not cov-

[16] Interested readers can potentially examine all 20,000 plots (2 covariates × 10 censoring distributions × 1000 simulated datasets) using the replication materials. An example using each of the four time transformations for the case of no censoring is presented in Section F of the Supporting Information.

ered all situations. In the following section, we provide two illustrations from published political science research to demonstrate that the choice of a time transformation is also consequential in published empirical work, as well as to show how to use the graphical displays to make more informed choices about time transformations.

Replications

Using the replication data made available by authors, we performed or reperformed the scaled Schoenfeld residual tests for the proportional hazards models from 19 articles published between 1992 and 2012 in the *American Journal of Political Science*, *American Political Science Review*, and *Journal of Politics.* These articles constituted all published analyses for which we were able to (1) obtain replication materials, and (2) replicate the authors' original published findings.[17] In this section, we present two examples as illustrations. Analyses for the remaining replications appear in Section G of the Supporting Information.

Illustration: Proportional Hazards and Government Agendas

In "The Government Agenda in Parliamentary Democracies," Martin (2004) examines the effect of a variety of factors on the organization of the policy agenda in four European democracies. In the article, the author uses a Cox specification but does not present explicit tests of the proportional hazards assumption. At the time of publication, such tests were not common in political science research. After successfully replicating the author's Cox specification, we performed scaled Schoenfeld residual tests using the time transformations discussed above. Table 2 presents the results.

[Table 2 about here.]

[17] All analyses were performed using the `survival` package (Therneau 1999) in R 2.15.1 (R Development Core Team 2012). In a small number of cases, replication data was either publicly available or transmitted to us by an author or authors, but the original published estimates were unable to be recovered from the information at hand. In some of these cases, the original analysis was performed using Stata, and Stata was able to recover the original findings, while R was unable to produce estimates. In one case, the `survival` package in R was used in the original analysis and we were able to recover the original Cox estimates, but the software was unable to perform the Grambsch-Therneau tests because of sparse matrices induced by a certain pattern of missing data. In certain other cases, replication data was provided, but still not enough information was available to recover the original estimates. For most of the published proportional hazards models from these journals that were not replicated, the reason was that data and replication materials were not made available.

Examining Table 2, it is immediately apparent that the choice of a time transformation matters for determining the set of covariates that are found to violate the proportional hazards assumption. While 9 out of 14 covariate-specific tests reached consistent conclusions across time transformations, discrepancies exist for the other 5 tests. For *Government Issue Saliency* (a continuous variable), and *Industrial Policy* and *Social Policy* (both binary variables), only the test that uses the natural log transformation of the time scale produces a statistically significant test statistic. For *Environmental Policy* (a binary covariate), the test that uses the natural log transformation is the only one that does not produce a statistically significant test statistic. And for the binary variable for *Luxembourg*, the tests using untransformed time and the natural log transformation are in agreement that it does not violate, while the tests using the rank and Kaplan-Meier transformations are in agreement that it does.

Plotting the covariate-specific scaled Schoenfeld residuals against the various time transformations is instructive as to why the discrepancies arise, and offers guidance on which version of the test should be employed. For instance, Figure 2 presents plots of *Government Issue Saliency* against each of the time transformations under consideration. Unlike the presentation for Cunningham (2011) and Maeda (2010) in Figure 1, the data from Martin (2004) do not seem to be long-tailed. A series of formal outlier tests using studentized residuals (not presented) confirms that outlier survival times are not a concern. In other words, exploratory data analysis and formal tests seem to indicate that we should be able to proceed with diagnostic testing using identity time. In fact, the picture in the upper right panel of Figure 2 indicates the perils of an uninformed decision. Rather than mitigating the impact of overly influential cases, the natural log transformation seems to unnecessarily create them. Additionally, it should be noted that no units are censored in these data. Therefore, the appropriate course, we argue, is to apply no transformation at all. In this case, researchers who are unaware of the issue of time transformations in diagnostic testing, and whose chosen statistical software transforms by default (e.g., the `survival` package in R, but not Stata), would come to an erroneous conclusion. Likewise, researchers who are aware of the issue of time transformations and blindly employ the log transformation would be making the least desirable choice. And the main point is that a simple graphical examination of one's data, possibly supplemented with formal tests for outliers, can be very informative on this point.

Heeding our advice, the researcher using a $p < .05$ confidence level would ultimately conclude that *Government Issue Divisiveness* and *Environmental Policy* are offending covariates that require the use of a corrective technique in Cox estimation.

[Figure 2 about here.]

The choice of identity time in this particular case brings up an additional issue, however, that is worthy of further discussion.[18] Namely, though the effects of *Government Issue Divisiveness* and *Environmental Policy* are both found to violate the proportional hazards assumption according to the covariate-specific tests, the global test using untransformed time is not statistically significant using conventional p-value thresholds. Based on our analyses of replication materials, this is a not uncommon situation with Cox specifications in published political science research, occurring in 20% of the scaled Schoenfeld residual tests presented in the main text and Supporting Information, including one of the tests in the Bennett (1997) illustration presented in the next subsection. In general, when this occurs, only a small minority of covariates in the given model exhibit statistically significant covariate-specific tests, and the global test comes relatively close to conventional thresholds for statistical significance. In the test using identity time in Table 2, for instance, 2 of 14 covariate-specific tests indicate violations, and the p-value for the global test is .067.[19]

With respect to this discrepancy, Box-Steffensmeier, Reiter, and Zorn (2003) pointed out that there is no clear guidance in the literature about the dominance of the global or covariate-specific tests in making decisions about violations of the proportional hazards assumption. Over a decade later, our reading of the literature indicates that this statement is still accurate. Like the graphical techniques recommended by Grambsch and Therneau (1994), the global tests can and should be used and reported in order to paint an overall picture of the degree to which a particular specification adheres to the proportional hazards assumption. However, given the current state of the literature, like Box-Steffensmeier, Reiter, and Zorn (2003), we argue that when researchers

[18] We thank an anonymous reviewer for suggesting this discussion.

[19] Across all of the scaled Schoenfeld residual tests presented in the main text and the Supporting Information that exhibit at least one statistically significant covariate-specific test simultaneously with a non-significant global test, the global tests have $p < .1$ in about 34% of cases, $p < .2$ in about 63% of cases, and $p < .3$ in about 72% of cases.

are making decisions about appropriate model specification, they should not discount indicators of covariate-specific nonproportionality, even in the face of a null result for the global test.

Illustration: Proportional Hazards and Alliance Duration

In "Testing Alternative Models of Alliance Duration, 1816-1984" Bennett (1997) combines hypotheses drawn from several different theories of alliance duration into a single modeling framework. The author tests this series of hypotheses using a Weibull model. Though the tests for proportional hazards discussed here are not relevant within the Weibull framework, Weibull is nonetheless a proportional hazards model. Following the approach taken by Box-Steffensmeier, Reiter, and Zorn (2003), we reexamine Bennett's analysis within the Cox framework, but revise and extend their work by discussing the importance of the choice of a time transformation for scaled Schoenfeld residual tests. Table 3 presents the results of these diagnostic tests using the various time transformations discussed above.

[Table 3 about here.]

Just as in Martin's (2004) analysis, the results in Table 3 indicate that the choice of a time transformation is consequential for Bennett's data as well. Specifically, all three of the tests that use transformed versions of the time scale indicate that *Symmetry* and *War Termination* (both binary variables) violate the proportional hazards assumption, while the test that uses identity time does not. Further, only the test employing the Kaplan-Meier transformation of the time scale indicates that *Mutual Threat* violates the proportional hazards assumption. The global test also exhibits discrepancies. Once again, a plot of the scaled Schoenfeld residuals against time will be instructive as to the appropriate choice. Take, for instance, Figure 3, which plots *Symmetry* against time and the various transformations of time. Unlike the data examined by Martin (2004), the graphical display of Bennett's data suggests the presence of outlier survival times. This is confirmed by tests of studentized residuals (not presented). Therefore, the preliminary evidence in front of us suggests the need for a transformation of the time scale.

[Figure 3 about here.]

For *Symmetry*, the tests using transformed time are all in agreement that the covariate violates the proportional hazards assumption. But examining Table 3 shows that there is not across-the-board agreement about the offending covariates in the model. Therefore, the choice of a specific time transformation will be consequential for the conclusions reached. Supplementing the information from the graphical displays and formal tests with the intuition garnered from our simulations, we examined the level of censoring in Bennett's data and found that about 45% of cases are censored (113 out of 207 units experience the event).[20] From our simulation findings about the performance of the tests under various censoring distributions, we would recommend the test using the rank transformation for these particular data. Taking that advice, a researcher would conclude that *Symmetry* and *War Termination* violate the proportional hazards assumption, and that corrective measures should be taken for these two variables in Cox estimation.

This finding is particularly instructive given the previous replication of these data by Box-Steffensmeier, Reiter, and Zorn (2003). Specifically, in that replication the authors use untransformed time to perform the scaled Schoenfeld residual tests, and, as our results show, they conclude that no specific covariate or the model as a whole violates the proportional hazards assumption. Their replication was an extremely thoughtful exercise in which they demonstrated to researchers how nonproportional hazards could be of substantive interest for many questions in international relations, and how the Grambsch-Therneau method could be used to adjudicate between competing hypotheses. Because the specific hypothesis that they used as an illustration was with respect to *Democracy*, their choice of identity time does not affect their specific conclusion in that instance. However, our argument and our simulation findings suggest that it is erroneous to simply use untransformed time in this particular case. The results for two covariates, as well as the global test, suggest that nonproportionality is still a concern. And the graphical displays and the level of censoring suggest a specific choice for a time transformation that will lead to more accurate conclusions.

[20] Bennett (1997), unlike many authors, reports this figure in the original paper.

Conclusions

As applications of event history analysis in political science research have grown in number, scholars have become increasingly cognizant of the need to examine the underlying assumptions of their chosen statistical models. Researchers working within the Cox framework take advantage of the model's flexibility in that it does not require the practitioner to specify a theoretical functional form for the baseline hazard (Cox 1972), which becomes an advantage because substantive theory is often not developed enough to make strong *a priori* claims about baseline levels of risk in the absence of covariate effects (Box-Steffensmeier and Jones 2004). The Cox model does, however, carry the assumption of proportional hazards, and if this assumption is not met, point estimates and tests of statistical significance will be misleading. The growth in awareness among political scientists of the need to test for proportional hazards within the Cox framework, largely driven by the work of Box-Steffensmeier and colleagues (Box-Steffensmeier and Jones 2004; Box-Steffensmeier and Zorn 2001; Box-Steffensmeier, Reiter, and Zorn 2003), has led to far more thoughtful applications of the Cox model to questions of the duration and timing of political events (Chiozza and Goemans 2004; Licht 2011).

However, we have argued in this paper that the standard statistical test for the proportional hazards assumption in Cox applications in the political science literature requires more thought than it has been given until now. Specifically, the Grambsch-Therneau method of examining the relationship between scaled Schoenfeld residuals and time requires researchers to make a choice about whether to use identity time or a transformation of the time scale. Using simulations and replications, we have shown that this choice will often have an impact on the decisions that researchers make in empirical analyses. We argue that researchers can make far more informed choices about diagnostic testing for proportional hazards by using very basic knowledge about their data that often goes overlooked.

Our suggested course of action with respect to best practices for researchers employing the Cox proportional hazards model is as follows. First, before any modeling occurs, practitioners must determine the levels and patterns of censoring in their data. Though one of the advantages of the Cox model is its ability to easily incorporate censored data, the presence of heavy censoring has been shown to affect inference for certain quantities of interest to practitioners (e.g., the Kaplan-Meier estimate of the survival curve; Rupert G. Miller 1983). More importantly, however, identification of censored and uncensored cases, combined with the qualitative knowledge of a subject-matter expert in a particular subfield, can potentially lead to reexamination of theories and empirical strategies. For instance, an underlying assumption of the basic Cox model presented here is that if indefinite observation was possible, all units would eventually experience the event of interest. Identification of patterns, levels, and identities of censored cases may lead a researcher to question this assumption, and instead conclude that an alternative empirical strategy may be more appropriate (e.g., the use of a so-called "cure" model in which some units are allowed to be unsusceptible to the event; Farewell 1982; Findley and Teo 2006; Svolik 2008).

Second, practitioners must determine whether cases with outlier survival times are present. Though we do not take a stance in this paper on appropriate methods of outlier detection, at a minimum we argue that practitioners should engage in some amount of exploratory analysis of survival times to identify cases that could be exhibiting unwarranted influence over both diagnostic tests and modeling choices. Pre-estimation, this can be accomplished with simple univariate summary measures such as histograms; post-estimation, one can use the scaled Schoenfeld residual plots discussed here. And again, identification of outlier cases carries vast potential to guide subsequent theoretical development and empirical modeling in unanticipated ways. Importantly, investigation of censoring patterns and outlier survival times are relatively simple steps that most researchers already know that they should be taking, but that are likely often neglected in the drive toward multivariate modeling.

Third, once practitioners have decided to use the Cox proportional hazards model with a particular set of data, it is critical that they evaluate the proportional hazards assumption. To do so, we advocate the use of the Grambsch-Therneau scaled Schoenfeld residual tests, supplemented by knowledge about censoring and outlier survival times garnered from the previous steps, in order to make appropriate choices regarding a transformation of the time scale. Specifically, when a researcher finds that outliers are not a feature of her data, she should proceed with the tests using untransformed time. If, however, outliers are a potential issue, the researcher should use a transformation of time. And the results of our simulations indicate that the rank transformation will often be the best choice. With low levels of censoring, we found that the rank and Kaplan-Meier transformations performed about equally well, and that both outperformed the natural log. As the level of censoring increased, however, whatever the censoring pattern, the rank transformation began to perform substantially better than either the Kaplan-Meier or natural log transformations. Most importantly, regardless of the ultimate decision regarding a transformation of the time scale, the researcher has the ability to explore all of the potential choices graphically, as we have done here.

And finally, upon detection of violations, as suggested by others (e.g., Box-Steffensmeier and Zorn 2001; Box-Steffensmeier and Jones 2004), we recommend that researchers interact the offending covariate(s) with some function of time to model the nonproportionality and proceed with estimation and interpretation (e.g., Chiozza and Goemans 2004; Maeda 2010; Licht 2011). Furthermore, researchers should not discount covariate-specific tests that indicate nonproportionality, even in the presence of a global test that fails to reject the null hypothesis of proportionality (Box-Steffensmeier, Reiter, and Zorn 2003).

The simulations and replications that we have presented here admittedly only scratch the surface of the possible set of scenarios that political scientists may encounter when analyzing event history data. Future research will be needed to investigate issues left unaddressed by this study. Regardless, our broader conclusion should be uncontroversial. That is, researchers employing the Cox proportional hazards model should engage in certain basic techniques of exploratory data analysis—namely, investigation of censoring and outliers—in order to make more informed decisions about detecting and correcting for violations of the proportional hazards assumption. The work of Box-Steffensmeier and colleagues (Box-Steffensmeier and Zorn 2001; Box-Steffensmeier, Reiter, and Zorn 2003; Box-Steffensmeier and Jones 2004) has taken the discipline a long way with respect to appropriate diagnostics and model specifications within the Cox framework. But we argue that supplementing the existing standards with some basic preliminary steps can lead to a new standard for best practices.

References

Bennett, D. Scott. 1997. "Testing Alternative Models of Alliance Duration, 1816–1984." *American Journal of Political Science* 41(3): 846–878.

Berlinski, Samuel, Torun Dewan, and Keith Dowding. 2010. "The Impact of Individual and Collective Performance on Ministerial Tenure." *Journal of Politics* 72(2): 559–571.

Berry, Christopher R., Barry C. Burden, and William G. Howell. 2010. "After Enactment: The Lives and Deaths of Federal Programs." *American Journal of Political Science* 54(1): 1–17.

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

Box-Steffensmeier, Janet M., and Christopher J. W. Zorn. 2001. "Duration Models and Proportional Hazards in Political Science." *American Journal of Political Science* 45(4): 972–988.

Box-Steffensmeier, Janet M., Dan Reiter, and Christopher J. W. Zorn. 2003. "Nonproportional Hazards and Event History Analysis in International Relations." *Journal of Conflict Resolution* 47(1): 33–53.

Chiozza, Giacomo, and H. E. Goemans. 2004. "International Conflict and the Tenure of Leaders: Is War Still 'Ex Post' Inefficient?" *American Journal of Political Science* 48(3): 604–619.

Cleves, Mario A., William W. Gould, Roberto G. Gutierrez, and Yulia V. Marchenko. 2010. *An Introduction to Survival Analysis Using Stata.* 3rd ed. College Station, TX: Stata Press.

Cook, R. Dennis. 1979. "Influential Observations in Linear Regression." *Journal of the American Statistical Association* 74(365): 169–174.

Cook, R. Dennis, and Sanford Weisberg. 1982. *Residuals and Influence in Regression.* New York: Chapman and Hall.

Cox, D. 1972. "Regression Models and Life Tables." *Journal of the Royal Statistical Society, Series B* 34(2): 187–220.

Crescenzi, Mark J. C. 2007. "Reputation and Interstate Conflict." *American Journal of Political Science* 51(2): 382–396.

Cunningham, Kathleen Gallagher. 2011. "Divide and Conquer or Divide and Concede: How Do States Respond to Internally Divided Separatists?" *American Political Science Review* 105(2): 275–297.

Debs, Alexandre, and H.E. Goemans. 2010. "Regime Type, the Fate of Leaders, and War." *American Political Science Review* 104(3): 430–445.

Diermeier, Daniel, and Randy T. Stevenson. 1999. "Cabinet Survival and Competing Risks." *American Journal of Political Science* 43(4): 1051–1068.

Farewell, V. T. 1982. "The Use of Mixture Models for the Analysis of Survival Data with Long-Term Survivors." *Biometrics* 38(4): 1041–1046.

Findley, Michael G., and Tze Kwang Teo. 2006. "Rethinking Third-Party Interventions into Civil Wars: An Actor-Centric Approach." *Journal of Politics* 68(4): 828–837.

Gibler, Douglas M., and Jaroslav Tir. 2010. "Settled Borders and Regime Type: Democratic Transitions as Consequences of Peaceful Territorial Transfers." *American Journal of Political Science* 54(4): 951–968.

Gill, Richard, and Martin Schumacher. 1987. "A Simple Test of the Proportional Hazards Assumption." *Biometrika* 74(2): 289–300.

Grambsch, Patricia M., and Terry M. Therneau. 1994. "Proportional Hazards Tests and Diagnostics Based on Weighted Residuals." *Biometrika* 81(3): 515–526.

Hendry, David J. 2014. "Data Generation for the Cox Proportional Hazards Model with Time-dependent Covariates: A Method for Medical Researchers." *Statistics in Medicine* 33(3): 436–454.

Hosmer, Jr., David W., Stanley Lemeshow, and Susanne May. 2008. *Applied Survival Analysis.* Hoboken, NJ: John Wiley & Sons.

Jackson, Christopher. 2014. "Multi-State Modelling with R: The msm Package." Available from: `http://cran.r-project.org/web/packages/msm/vignettes/msm-manual.pdf` (Accessed on 30 April 2014).

Kalbfleisch, John D., and Ross L. Prentice. 2002. *The Statistical Analysis of Failure Time Data.* Second ed. Hoboken, NJ: John Wiley & Sons.

Kaplan, E.L., and Paul Meier. 1958. "Nonparametric Estimation from Incomplete Observations." *Journal of the American Statistical Association* 53(282): 457–481.

Keele, Luke. 2010. "Proportionally Difficult: Testing for Nonproportional Hazards in Cox Models." *Political Analysis* 18(2): 189–205.

Koch, Michael T., and Patricia Sullivan. 2010. "Should I Stay or Should I Go Now? Partisanship, Approval, and the Duration of Major Power Democratic Military Interventions." *Journal of Politics* 72(3): 616–629.

Lagakos, S. W., and D. A. Schoenfeld. 1984. "Properties of Proportional-Hazards Score Tests Under Misspecified Regression Models." *Biometrics* 40(4): 1037–1048.

Leeds, Brett Ashley, and Burcu Savun. 2007. "Terminating Alliances: Why Do States Abrogate Agreements?" *Journal of Politics* 69(4): 1118–1132.

Licht, Amanda A. 2011. "Change Comes with Time: Substantive Interpretation of Nonproportional Hazards in Event History Analysis." *Political Analysis* 19(2): 227–243.

Maeda, Ko. 2010. "Two Modes of Democratic Breakdown: A Competing Risks Analysis of Democratic Durability." *Journal of Politics* 72(4): 1129–1143.

Maltzman, Forrest, and Charles R. Shipan. 2008. "Change, Continuity, and the Evolution of the Law." *American Journal of Political Science* 52(2): 252–267.

Martin, Lanny W. 2004. "The Government Agenda in Parliamentary Democracies." *American Journal of Political Science* 48(3): 445–461.

Murillo, Maria Victoria, and Cecilia Martínez-Gallardo. 2007. "Political Competition and Policy Adoption: Market Reforms in Latin American Public Utilities." *American Journal of Political Science* 51(1): 120–139.

Ng'andu, N. H. 1997. "An Empirical Comparison of Statistical Tests for Assessing the Proportional Hazards Assumption of Cox's Model." *Statistics in Medicine* 16(6): 611–626.

R Development Core Team. 2012. *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing.

Rupert G. Miller, Jr. 1983. "What Price Kaplan-Meier?" *Biometrics* 39(4): 1077–1081.

Schleiter, Petra, and Edward Morgan-Jones. 2009. "Constitutional Power and Competing Risks: Monarchs, Presidents, Prime Ministers, and the Termination of East and West European Cabinets." *American Political Science Review* 103(3): 496–512.

Schoenfeld, D. 1980. "Chi-Squared Goodness-of-Fit Tests for the Proportional Hazards Regression Model." *Biometrika* 67(1): 145–153.

Stablein, D. M., W. H. Carter, Jr., and J. W. Novak. 1981. "Analysis of Survival Data with Nonproportional Hazard Functions." *Controlled Clinical Trials* 2(2): 149–159.

Svolik, Milan. 2008. "Authoritarian Reversals and Democratic Consolidation." *American Political Science Review* 102(2): 153–168.

Therneau, Terry M. 1999. "A Package for Survival Analysis in S." Available from: `http://www.mayo.edu/research/documents/tr53pdf/DOC-10027379`. (Accessed on 30 April 2014).

Therneau, Terry M., and Patricia M. Grambsch. 2000. *Modeling Survival Data: Extending the Cox Model.* New York: Springer.

Therneau, Terry M., Patricia M. Grambsch, and Thomas R. Fleming. 1990. "Martingale-Based Residuals for Survival Models." *Biometrika* 77(1): 147–160.

Weisberg, Sanford. 2005. *Applied Linear Regression.* 3rd ed. Hoboken, NJ: John Wiley & Sons.

Winnett, Angela, and Peter Sasieni. 2001. "Miscellanea: A Note on Scaled Schoenfeld Residuals for the Proportional Hazards Model." *Biometrika* 88(2): 565–571.

Zuehlke, Thomas W. 2013. "Estimation and Testing of Nonproportional Weibull Hazard Models." *Applied Economics* 45(15): 2059–2066.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

A. Content Analysis

B. Estimators and Test Statistics for Grambsch-Therneau Tests

C. Determination of Outlier Survival Times

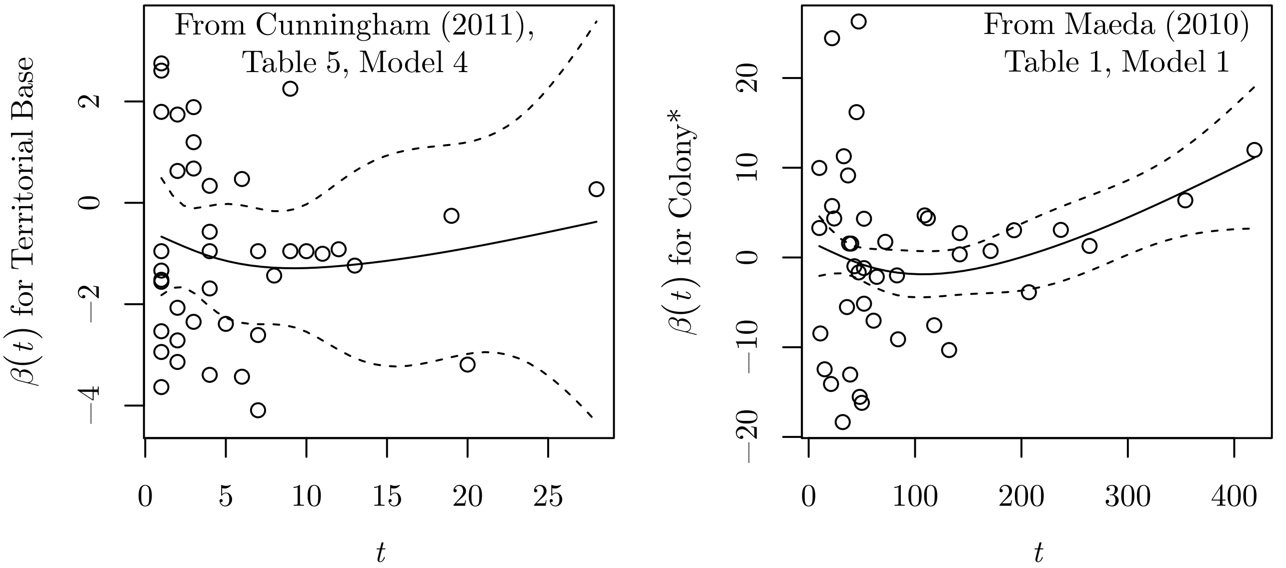
D. The Kaplan-Meier Transformation of the Time Scale

E. Algorithm Used to Generate Simulated Data

F. Example Scaled Schoenfeld Residual Plots from Simulations

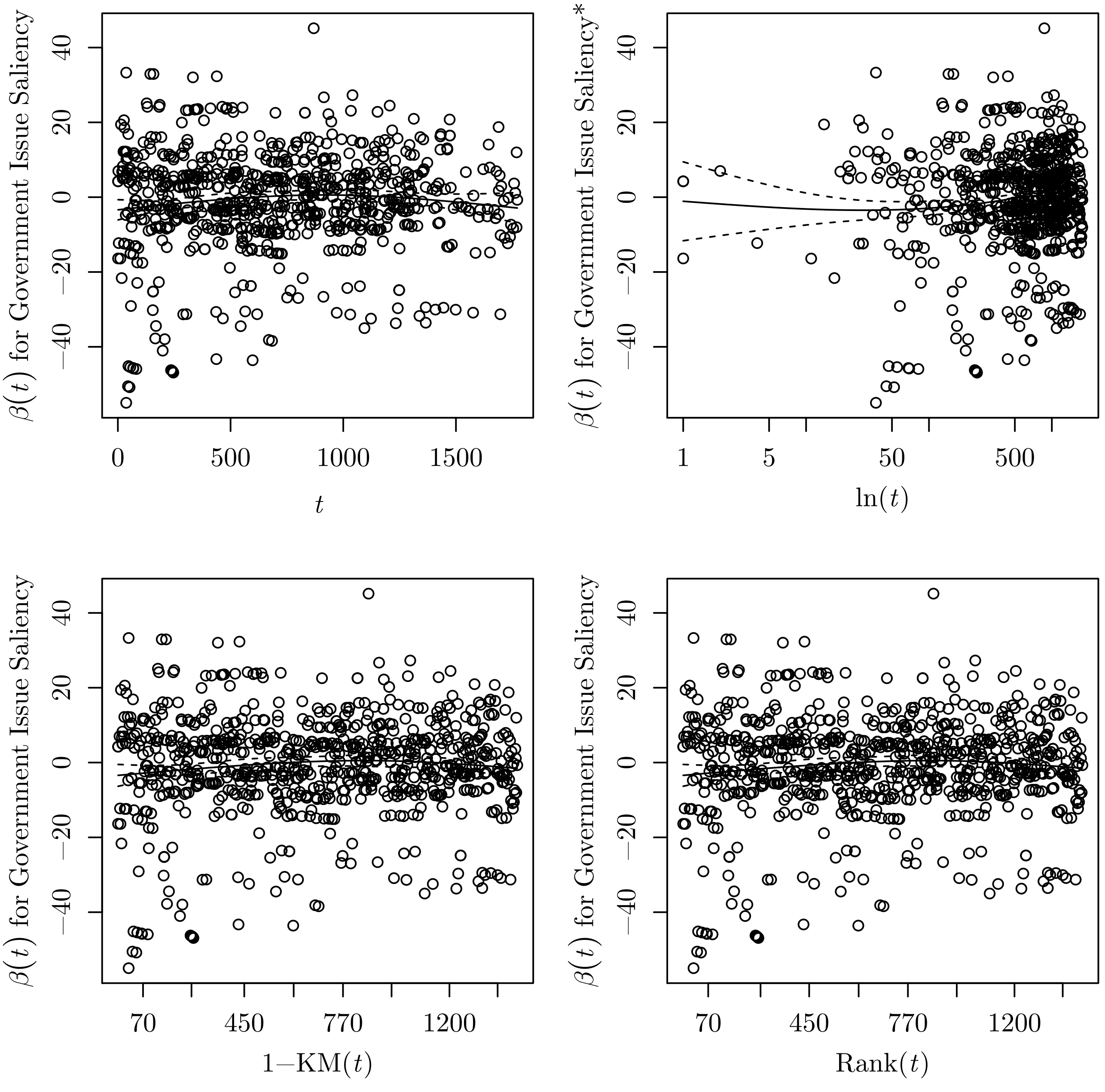
G. Illustrations of Proportional Hazards Tests from Published Political Science Research

Figure 1: Long-tailed Survival Distributions from Published Political Science Research



Note: Cunningham (2011) models the number of years between concessions made by states to self determination (SD) movements. Maeda (2010) models the number of months until the transition from democracy to nondemocracy. For further information, consult the original articles. Solid line is a smoothing spline. Dashed lines represent a ± 2-standard error confidence band.
*Statistically significant test statistic indicating covariate violates the proportional hazards assumption, $p < .05$.

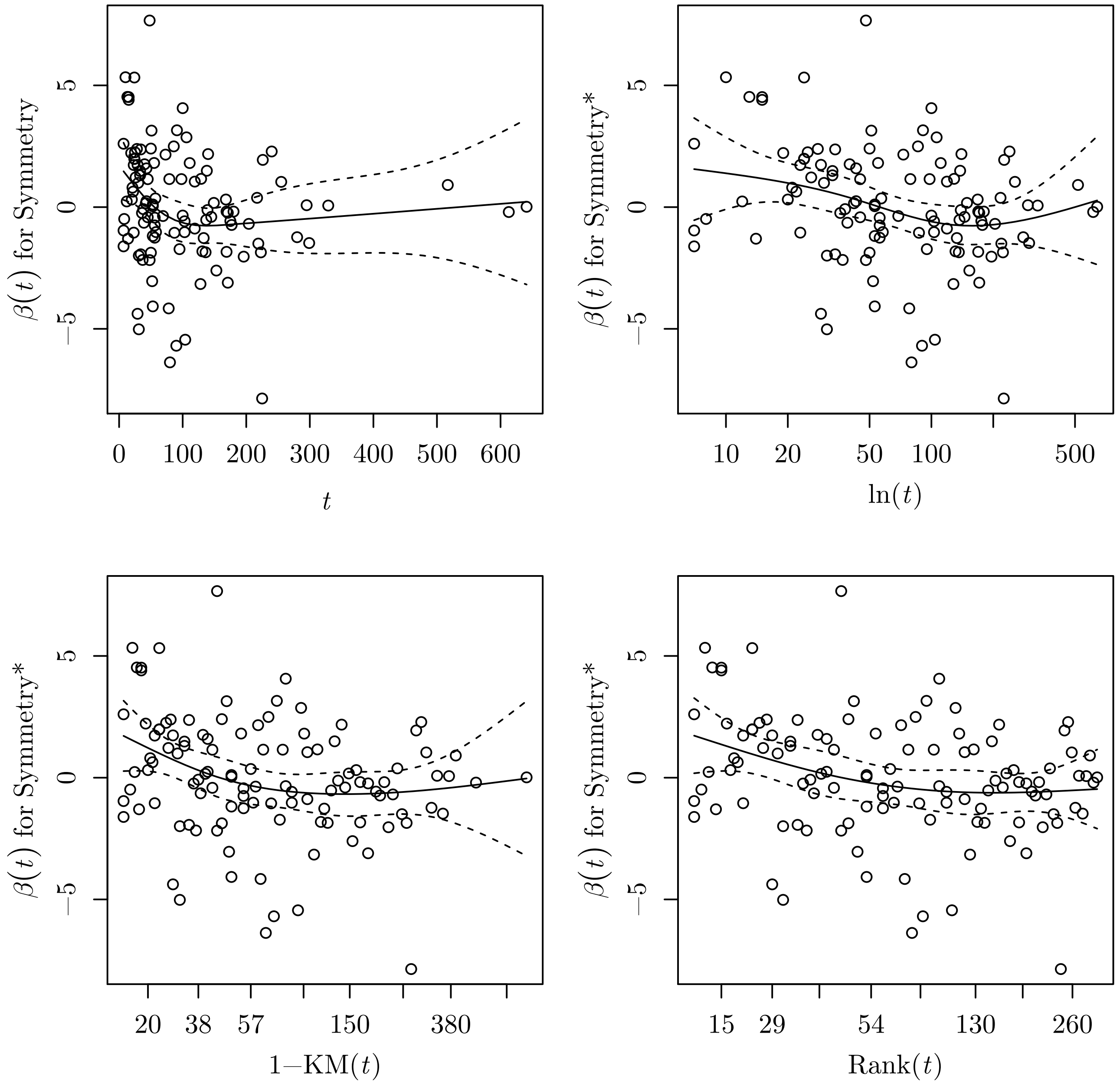
Figure 2: Plot of Scaled Schoenfeld Residuals vs. Time for Four Time Transformations, *Government Issue Saliency*, (Martin 2004, Table 1, Model 1)



Note: Solid line is a smoothing spline. Dashed lines represent a ± 2-standard error confidence band. Consult Martin (2004) for information about covariate.

*Statistically significant test statistic indicating covariate violates the proportional hazards assumption, $p < .05$.

Figure 3: Plot of Scaled Schoenfeld Residuals vs. Time for Four Time Transformations, *Symmetry*, (Bennett 1997, Table 1, Complete Model; as Replicated in Box-Steffensmeier, Reiter, and Zorn 2003)



Note: Solid line is a smoothing spline. Dashed lines represent a ± 2-standard error confidence band. Consult Bennett (1997) for information about covariate.

*Statistically significant test statistic indicating covariate violates the proportional hazards assumption, $p < .05$.

Table 1: Summary of Violations of Proportional Hazards from Scaled Schoenfeld Residual Tests Using Four Time Transformations

Censoring	Proportion		Frequency $p < .05$			
Distribution	Censored	Z	t	$\ln(t)$	$1 - KM(t)$	Rank(t)
None	0.00	Z_1 (Continuous)	633	949	986	987
		Z_2 (Binary)	876	993	1000	1000
		Global Test	949	999	1000	1000
Uniform	0.50	Z_1 (Continuous)	334	710	787	857
		Z_2 (Binary)	573	853	855	939
		Global Test	629	943	958	991
Uniform	0.25	Z_1 (Continuous)	522	891	961	967
		Z_2 (Binary)	751	967	982	992
		Global Test	852	995	1000	1000
Uniform	0.10	Z_1 (Continuous)	589	939	983	985
		Z_2 (Binary)	830	989	997	998
		Global Test	912	1000	1000	1000
Long Biased	0.50	Z_1 (Continuous)	356	730	789	864
		Z_2 (Binary)	565	872	874	943
		Global Test	670	963	979	993
Long Biased	0.25	Z_1 (Continuous)	518	887	949	959
		Z_2 (Binary)	734	967	987	996
		Global Test	836	996	999	999
Long Biased	0.10	Z_1 (Continuous)	594	956	989	990
		Z_2 (Binary)	843	987	997	1000
		Global Test	926	999	1000	1000
Short Biased	0.50	Z_1 (Continuous)	369	710	776	851
		Z_2 (Binary)	547	856	870	953
		Global Test	644	959	975	992
Short Biased	0.25	Z_1 (Continuous)	500	896	963	968
		Z_2 (Binary)	755	971	982	994
		Global Test	834	994	1000	1000
Short Biased	0.10	Z_1 (Continuous)	590	941	987	988
		Z_2 (Binary)	840	985	997	997
		Global Test	919	999	1000	1000

Note: Cell entries are the number of times out of 1000 that a scaled Schoenfeld residual test indicated that the data violates the proportional hazards assumption using a threshold of $p < .05$.

Table 2: Grambsch-Therneau Tests of Proportional Hazards Assumption Using Four Time Transformations (Martin 2004, Table 1, Model 1)

	Time Transformation			
	t	$\ln(t)$	1 - KM(t)	Rank(t)
	ρ	ρ	ρ	ρ
Government Issue Saliency	0.043	0.064*	0.051	0.051
Government Issue Divisiveness	-0.101*	-0.102*	-0.106*	-0.106*
Opposition Issue Saliency	-0.019	-0.028	-0.023	-0.023
Opposition Issue Divisiveness	0.018	-0.001	0.015	0.015
Foreign Policy	0.026	0.018	0.033	0.033
Industrial Policy	0.044	0.064*	0.054	0.054
Social Policy	0.049	0.095*	0.055	0.054
Clerical Policy	0.032	0.048	0.032	0.032
Agricultural Policy	0.008	0.031	0.017	0.017
Regional Policy	0.027	0.033	0.030	0.030
Environmental Policy	0.066*	0.058	0.070*	0.070*
Germany	-0.014	-0.028	-0.015	-0.015
Belgium	-0.023	-0.009	-0.028	-0.029
Luxembourg	-0.060	-0.012	-0.065*	-0.065*
	χ^2	χ^2	χ^2	χ^2
Global Test	22.625	20.420	25.688*	25.705*

Note: Cell entries for the upper panel are Pearson product-moment correlation coefficients with tests of statistical significance based on comparison of the covariate-specific test statistic given in (3) to a $\chi^2(1)$ distribution. Cell entries for the lower panel are the global test statistics given in (4) with tests of statistical significance based on comparison to a $\chi^2(14)$ distribution (Grambsch and Therneau 1994). Covariates whose tests are inconsistent are presented in bold and italics. For more information, consult Martin (2004).

*$p < .05$

Table 3: Grambsch-Therneau Tests of Proportional Hazards Assumption Using Four Time Transformations (Bennett 1997, Table 1, Complete Model; as Replicated in Box-Steffensmeier, Reiter, and Zorn 2003)

	Time Transformation			
	t	ln(t)	1 - KM(t)	Rank(t)
	ρ	ρ	ρ	ρ
Change in Security	-0.025	-0.026	-0.045	-0.049
Alliance Security Improvement	0.115	0.097	0.109	0.098
Mutual Threat	0.093	0.099	0.124*	0.113
Capability Change	-0.077	-0.132	-0.122	-0.137
Symmetry	-0.144	-0.254*	-0.243*	-0.267*
Capability Concentration	-0.105	-0.129	-0.129	-0.133
Democracy (Liberal)	0.063	0.028	0.063	0.045
Polity Change	0.062	-0.003	0.028	0.017
Number of States	-0.064	-0.089	-0.100	-0.106
Wartime	-0.067	-0.031	-0.069	-0.049
War Termination	0.076	0.233*	0.175*	0.217*
	χ^2	χ^2	χ^2	χ^2
Global Test	7.935	20.074*	18.180	21.214*

Note: Cell entries for the upper panel are Pearson product-moment correlation coefficients with tests of statistical significance based on comparison of the covariate-specific test statistic given in (3) to a $\chi^2(1)$ distribution. Cell entries for the lower panel are the global test statistics given in (4) with tests of statistical significance based on comparison to a $\chi^2(11)$ distribution (Grambsch and Therneau 1994). Covariates whose tests are inconsistent are presented in bold and italics. For more information, consult Bennett (1997).

$^*p < .05$