Duration Models and Proportional Hazards in Political Science

Janet M. Box-Steffensmeier
Department of Political Science
Ohio State University
Columbus, OH 43210-1373
jboxstef+@osu.edu

Christopher J. W. Zorn
Department of Political Science
Emory University
Atlanta, GA 30322
czorn@emory.edu

Version 1.9: April 20, 1998

Abstract

In recent years political scientists have increasingly adopted a wide range of techniques for modeling duration data. A key assumption of all these approaches is that the hazard ratios (i.e., the conditional relative risks across substrata) are proportional to one another, and that this proportionality is maintained over time. Estimation of proportional hazards (PH) models when in fact hazards are non-proportional results in coefficient biases and decreased power of significance tests. In particular, misspecified PH models will overestimate the impact of variables whose associated hazards are increasing, while coefficient estimates for covariates in which the hazards are converging will be biased towards zero. We investigate the proportionality assumption of two widely used duration models, the Weibull parametric model and Cox's (1972) semiparametric approach, in the context of a duration model of Supreme Court retirements. We address the potential problems with incorrectly assuming proportionality, illustrate a range of techniques for testing the proportionality assumption, and conclude with a number of means for accurately and efficiently estimating these models in the presence of non-proportional hazards.

Paper prepared for presentation at the Annual Meeting of the Midwest Political Science Association, April 23-25, 1998, Chicago, Illinois. Thanks to everybody who helped us. All analyses were performed in Stata 5.0 and S-Plus 4.0; data are available on request from the authors. This is a preliminary version; comments and suggestions are welcome and appreciated.

Introduction

Most models of duration commonly used in political science make the assumption that the hazards over different values of the covariates are proportional. While it is central to accurate estimation and interpretation of these models, the proportional hazards assumption has received little attention by political scientists modeling duration data. This is unfortunate, for a number of reasons. First, absent explicit examinations of the validity of these assumptions, the reader is unable to infer whether or not in fact that assumption holds for any particular analysis. This difficulty is exacerbated by the fact that, in many circumstances, theoretical information and empirical data suggest that the assumption itself is of dubious accuracy. In fact, it is often the case that many substantively interesting hypotheses imply time-dependence or other forms of nonproportionality in the conditional probability of failure.

These considerations are important because, as has been widely shown in the literature on biostatistics, estimation of proportional hazards models with nonproportional hazard rates in the data can result in biased estimates, incorrect standard errors, and faulty inferences about the substantive impact of independent variables (e.g. Kalbfleisch and Prentice 1980; Schemper 1992; Collett 1994). Because of this, it is important that, as political scientists begin to use these models more often in applied research, they take care to examine the extent to which this assumption is consistent with their data, which will reduce the impact that violations have on their results.

The plan of the paper is as follows. We first discuss the nature of the proportional hazards assumption in the abstract, showing what the assumption means for our expectations about the impact of covariates on the conditional hazard. We also discuss forms of nonproportionality, considering examples when we might expect deviations from conditionality to be the norm rather than the exception. We go on to address the impact of nonproportionality in two widely-used models for duration data, Cox's (1972) proportional hazards model and the Weibull model. In the context of each, we discuss tests for proportionality, both graphical and statistical, as well as suggesting methods for dealing with nonproportionality should it arise. In each section, we illustrate these techniques using political science data on Supreme Court careers from 1789 to 1993. We show how substantive considerations can prompt concerns about nonproportional covariates, and how applied researchers can both test and implement remedies for nonproportionality.

The Issue of Proportionality

We first consider a general model of failure time (e.g. Kalbfleisch and Prentice 1980). The dependent variable of interest is the duration until the occurrence of some event, which we refer to generically as a "failure". We may write the conditional probability of failure (i.e., the *hazard rate*) at time t as:

$$h(t) = Prob(T{=}t \mid T{\geq}t) = f(X\beta), \tag{1}$$

a function of a set of covariates **X** and a coefficient vector β.[1] In the context of political science data, the observations in question may consist of cabinets (e.g. King et. al. 1990; Warwick 1992), wars (e.g. Bueno de Mesquita and Siverson 1995; Bennet and Stam 1996; Bennett 1996, 1997) war-dyads (Werner 1998), members of Congress (Box-Steffensmeier 1996; Box-Steffensmeier et. al. 1997; Katz and Sala 1996), or other units of analysis.

For illustrative purposes, consider the widely-researched case of cabinet durations. Suppose we have two types of countries in our data, type A and type B. Let $h_A(t)$ and $h_B(t)$ be the hazards of failure at time t for countries of type A and type B, respectively. If the hazard at time t for a country of Type A is proportional to the hazard at that same time for a country of type B, we can write a model with proportional hazards as:

$$h_A(t) = Ch_B(t) \tag{2}$$

for any non-negative value of t and where C is a constant. The value of C is the hazard ratio; i.e., the ratio of the hazard of failure at any time for a country of type A relative to a country of Type B. If $C < 1$, the hazard of failure at t is smaller for a country of type A, relative to an country of type B, with the result that type A will be expected to experience a cabinet dissolution later than type B. If $C > 1$, type B's cabinets will, on average, be expected to last longer. In either case, the true survivor

[1] See Kiefer (1988), Beck (1997), and Box-Steffensmeier and Jones (1997) for good introductions to hazard rate models.

functions for countries of type A and type B do not cross, but will instead be roughly parallel (Collett 1994, 44-5).

Models that can be characterized by equation (2) are known generally as *proportional hazards* (PH) models. Both the Cox and Weibull models are PH models. Proportional hazards models "assume that the hazard functions of all individuals differ only by a factor of proportionality. (That is, if one individual's hazard rate is 10 times higher than another's at one point in time, it is 10 times higher at all points in time.)" (Chung, Schmidt, and Witte 1991, 71).

There are a number of instances, where, for substantive reasons, we might expect that the PH assumption would not hold. In biomedical research, for example, a common reason for nonproportional hazards is that treatment effects decrease over time as subjects develop resistances to therapies. Thus, the hazard of death or morbidity for a treatment group, initially lower than that for the control, increases as the study wears on, and the two hazard rates converge. Alternatively, hazards may also be diverging, as the impact of a treatment grows more pronounced over time.[2] Finally, in some instances, hazards may actually cross. Collett (1994) gives the example of choosing between traditional drug therapy and surgery in cancer treatments: while the initial risk of the surgery is higher due to complications and other factors, the long-run prognosis of those undergoing surgery is better. Thus initial hazards are higher for surgery patients, but those hazards decline, while those for chemotherapy patients increase, over time.

In political science, we may ". . . expect that the effect of one or more predictor variable on the hazard rate increases or decreases over time. There may be a number of different explanations for such change to occur, including learning effects, shifts in life-course position, maturational changes, and so on" (Teachman and Hayward 1993). Consider an example from research into Congressional careers. A staple of this research has been the contention that, because of their perennial minority status and higher opportunity costs of holding office, Republicans were more likely to serve shorter careers than Democrats (e.g. Bullock 1972; Kiewiet and Zeng 1993). Leaving aside for the moment the obvious difficulty with this argument in the context of the 104th and 105th Congresses, by this logic

[2]It is important to note that, by themselves, converging or diverging hazards are *not* indicative of nonproportional hazards. In fact, converging hazards are required to meet the assumption of proportionality in models where the hazards are decreasing over time. Similarly, in models where hazards are increasing, we would expect proportional hazards to diverge. What is important is the extent to which that convergence or divergence deviates from proportionality.

it is likely that, rather than being a constant effect, the impact of party is a variable one, which waxes and wanes depending on the extent of majority control. Hazards for Democrats and Republicans may converge, diverge, or even (in the case of a change of partisan control) cross. Implicit in this example is our inability to write the hazard functions for the independent variable in question in the form of equation (2). This is particularly true in the case of crossing hazards, which by definition cannot be proportional, but is also the case for hazards that are converging or diverging over time.

Estimation of proportional hazards models when in fact hazards are non-proportional results in coefficient biases and decreased power of significance tests. In particular, misspecified PH models will overestimate the impact of variables whose associated hazards are diverging, while coefficient estimates for covariates in which the hazards are converging will be biased towards zero (Kalbfleisch and Prentice 1980). Schemper (1992) summarizes the consequences of assuming constant hazard ratios when they are not applicable: "For covariates whose hazard ratios are non-constant over time, the power of corresponding tests decreases because of suboptimal weights for combining the information provided by the risk sets of times where failures occur (Lagkos and Schoenfeld, 1984). For other covariates with constant hazard ratios, testing power declines as a consequence of an inferior fit of the model" (Schemper 1992, 455). He shows that "the relative risk for covariates with hazard ratios increasing over time is overestimated while for covariates with converging hazards, perhaps the most frequent violation, the relative risk is underestimated" (Schemper 1992, 455). In addition, the extent of this bias can be consequential. Gray (1996), for example, finds that when two treatments have overlapping or crossing hazards, the power of models based on the proportional hazards assumption can be reduced by as much as 90 percent.

Despite its importance, the strong proportionality assumption is rarely tested in political science applications. In most cases, we simply do not know if political science data typically violates this assumption. However, the literature has often concluded " . . . that violations of the proportionality assumption are the rule, rather than the exception" (Singer and Willett 1993). We advocate that analysts routinely check this assumption. In the following sections, we examine the importance of the PH assumption for two widely-used duration models: Cox's (1972) proportional hazards model and the parametric Weibull specification for hazards.

Supreme Court Retirements, Critical Nominations, and Proportional Hazards

Throughout the paper, we consider the case of the careers of Supreme Court justices as a running example for demonstrating techniques relating to the proportional hazards assumption. Compared with the extensive work on Presidential and Congressional careers (e.g. Cunliffe 1972; Jones 1994; Wawro 1996), the careers of Supreme Court justices have received relatively little attention (but see Squire 1988; Hagle 1992). This is perplexing, given the vast attention focused on the appointment and confirmation process (e.g. Segal et. al 1992), for it is clear that no such appointment of confirmation can be forthcoming until one of the current occupants of the high Court vacates his or her seat.

We consider data on 107 justices who sat on the U.S. Supreme Court between the Court's founding in 1789 and 1993. Of these, 47 died in office, 52 retired from the Court, and eight remained on the Court at the end of our observation period.[3] We focus here on retirements, and accordingly treat all non-retiring justices as censored throughout our analyses. The dependent variable of interest is thus the hazard of retirement; i.e., the duration of a justice's career. This variable ranges from a low of one term of service on the bench (Justices Rutledge, T. Johnson, Byrnes and Thomas) to the 36 years on the Court served by justice Douglas.

We analyze seven independent variables related to the hazard of retirement. Our primary substantive variable of interest is the impact of "critical nominations" on Supreme Court retirement behavior. We use Van Winkle's (1994) modification to Ruckman's (1993) notion of a "critical nomination". Justices receive a 1 on this variable if either (a) their appointment causes the partisan balance on the Court to shift (e.g. for a Republican, from 5-4 democratic to 5-4 Republican), or (2) if their appointment moves the Court from a 6-3 to a 5-4 margin, or (3) if their appointment moves the Court from a 5-4 to a 6-3 margin. This variable thus depends (in most cases) not only on the partisan balance on the Court, but also on the party of the justice. We expect that justices who are appointed as "critical nominees" are more likely to be sensitive to political considerations regarding their retirement, and, because the composition of the Court changes slowly, will be less likely to retire early than other justices, since to do so is likely to result in another "critical nomination".

Drawing on previous work, we note whether the justice in question is of the same political party as his or her appointing President. As with critical nominations, our theory suggests that justices

[3]We code the retirement of Justice Byron White in 1993, but do not include his replacement, Justice Ruth Bader Ginsburg, in our data.

who differ in party from their appointing President have an incentive to stay in office at least until a different President assumes the presidency; this suggests that such justices will be less likely to retire early in their careers. We also record the age of the justice at his or her nomination, with the expectation that justices appointed at a younger age will serve longer terms and have correspondingly lower hazards of retirement than their older brethren. We also include an indicator for whether a justice was appointed in the 20th century (coded one) or not (coded zero). Because of medical advances and other issues, it has been recognized that more recent justices are more likely to retire from the Court, rather than to die in office. We also include a number of control variables. These include the political party of the justice (coded 0 for Democrats and 1 for Republicans), and indicator variables for Chief Justices and justices appointed from the South. Summary statistics for these variables, as well as for the dependent variable and the censoring indicator, are presented in Table 1.

From the outset, we have substantively-related concerns about the extent to which the proportional hazards assumption accurately characterizes our data. In particular, two variables suggest the potential for differential influence on the hazard rate over time. Most important, the critical nomination variable may have a nonproportional effect. In the first year or two on the Court, a "critical justice" who chooses to retire will likely create another critical nomination; thus, the impact of this variable ought to be large early in a justice's tenure. As time passes, changes in the Court's membership are likely to changed those dynamics, and the impact of how "critical" a justice's initial nomination was wears off. Thus, we might also expect converging hazards for critical and non-critical justices.[4] In addition, the impact of a justice's age at appointment may not be constant over time. In particular, we might expect that the impact of this variable will decrease over a justice's tenure: justices appointed at a relatively young age are significantly less likely to retire soon than their older counterparts, but the influence of this age difference decreases the longer the justice remains on the Court.

Cox's Proportional Hazards Model

[4]A more plausible argument, examined in Zorn and Van Winkle (1997), would include time-varying covariates on such factors as critical nominations and partisan agreement with the current president. However, a time-varying covariate model is no longer one of proportional hazards, and so for illustrative purposes we focus on the case presented here instead.

The proportional hazards model developed by Cox (1972) is a popular and flexible model that does not assume a specific probability distribution for the time until an event occurs. The absence of a need to parameterize time dependency is a significant advantage in most political science applications, since in most cases our theories do not allow us to specify *a priori* what distribution should be used, and in many cases the parameterization chosen can have a large impact on the substantive conclusions drawn.[5]

The hazard rate for the proportional hazards model is:

$$h(t\,|\,x) \;=\; h_0(t)e^{(\beta_k' X_i)} \tag{3}$$

where $h_0(t)$ is the (unspecified) baseline hazard function and X_i are covariates for individual i. Such models are typically estimated via a quasi- or partial-likelihood procedure in which the term for the baseline hazard is treated as a nuisance parameter and integrated out of the likelihood.

The Cox model is a model of proportional hazards; that is, it assumes that the hazard functions of any two individuals with different values on one or more covariates differ only by a factor of proportionality. The baseline hazard rate varies with time but not across individuals so that the ratio of the hazards for individuals *i* and *j* are independent of *t* and are constant for all *t*:

$$\frac{h_i(t)}{h_j(t)} \;=\; e^{\beta(x_j - x_i)} \tag{4}$$

As noted above, estimation of Cox's model in the presence of hazards that do not satisfy the proportionality assumption can result in biased and inefficient estimates of all parameters, not simply those for the covariate(s) in question. As a result of this possibility, it is widely recognized outside the political science literature that procedures to check this assumption are critical.

In the context of the Cox model, three general classes of tests for nonproportionality have been proposed. All can be thought of as variations on a more generalized Cox model which allows hazard ratios to vary over time:

---

[5]If one's theory does tell us what kind of duration dependency to expect, of course, that parameterization should be used (see Bennett 1997; forthcoming).

$$h(t) = h_0(t)\,e^{[\beta_k X_i + \gamma_k X_i g(t)]} \qquad (5)$$

In the specification given in equation (5), the effects of individual covariates are allowed to vary according to some function $g(\cdot)$ of time. The intuition behind many tests for nonproportionality is to test for $\gamma = 0$. As noted there are three general classes of tests that fall into this framework:

- Tests based on changes in parameter values for coefficients estimated on a subsample of the data defined by t,
- Tests based on plots of survival estimates and regression residuals against time, and
- Explicit tests of coefficients on interactions of covariates and time.

The last of these methods also provides a way of explicitly modeling time-dependence in the Cox model, and is thus also useful as a means of dealing with nonproportionality, should it arise. We address each of these approaches in turn below.

Nonproportionality Tests Based on Piecewise and Stratified Regressions

The nature of the time-dependence specified in equation (5) implies that covariates will have differential impacts on the hazard rate, depending on the time in the event history. Probably the simplest possibility in this light is to treat $g(\cdot)$ as a step function at some point in the process τ, taking on a value of 0 for all points in time prior to that and 1 after:

$$\begin{aligned} g(t) &= 0 \quad \forall\ t \le \tau, \\ &= 1 \quad \forall\ t > \tau. \end{aligned}$$

Under this view, a natural test for nonproportionality is to estimate separate Cox regressions for different values of $g(t)$, that is, for observations whose survival times fall above and below some predetermined value.

This test is suggested in Schemper (1992) and Collett (1994), and provides a simple first pass at the issue of proportionality. The specification above represents this test in its simplest form; the data may be divided into as few or many separate time periods as are reasonable and suggested by theory. In some cases, the data itself will suggest likely values for the break points, such as congressional redistricting or critical election eras; otherwise, medians or quartiles may be used. By examining the sensitivity of parameter estimates to estimation over different subsets of the data, one can implicitly test the hypothesis that each covariate's impact remains relatively stable over the durations under study.

An alternative means of dealing with nonproportionality is to stratify the data by the covariate of interest. Under stratification, the impact of the remaining independent variables on the conditional hazards is assumed to be constant across strata, but separate baseline hazards are estimated for the *j* different groups:

$$h(t) = h_{0j}(t)e^{(\beta_k X_i)} \tag{6}$$

One benefit of stratification is that it allows straightforward comparisons of model fit and parameter sensitivity to the stratification technique. A major drawback of the stratification approach, as noted by Schemper (1992), is that stratification on a variable of interest prevents estimation of the impact of that variable on the hazard rate. Because the Cox model factors out the baseline hazard, separate baselines for different covariate values are not reported. As a result, the stratification approach is more useful as a diagnostic technique than as a means of dealing with the presence of nonproportionality in substantive variables.

Column 1 of Table 2 presents results of estimating a Cox proportional hazards model on Supreme Court retirements including the aforementioned seven covariates. Coefficient estimates reflect the impact of a given independent variable on the hazard rate; thus, positive values imply higher hazards and correspondingly shorter durations. We obtain statistically significant positive results for the political party, age, and 20th century variables; higher values on each result in increased hazards of retirement from the Court. The result for critical nominations, while in the expected direction, fails to reach statistical significance ($p = 0.07$, one-tailed). Given our expectations about the converging hazards on this variable, it is possible that this result is due to nonproportionality; accordingly, we begin to examine that possibility here.

We next divided the data according its median survival time (T = 15) and ran separate, piecewise Cox regressions on the two subsets of the data; these estimates are presented in columns 2 and 3 of Table 2. The results of this initial test are intriguing. The model appears to do a substantially better job of explaining retirements in justices who serve longer tenures; the model for short-term justices does not predict better than a null model. We also note substantial variation in the impact of three coefficients. First, the age variable is nearly seven times as large in the latter group as in the former, and is strongly statistically significant only in the longer-tenure group. Likewise, we see a complete reversal of sign in the variable for critical nominations: surprisingly, the negative impact of that variable appears to be driven by later, rather than earlier, retirements, though neither coefficient attains statistical significance in the piecewise models. Finally, and unexplainedly, we note a large positive impact of the South variable in justices serving shorter tenures. These results suggest that, for justices serving less than the median time on the Court, a background in the South more than triples the odds of a retirement relative to that of northern justices; a statistically insignificant, negative impact is present in justices serving longer terms.

These results suggest the possibility of some degree of time-dependence in these three variables. To investigate this possibility further, we estimated stratified Cox models by each of these three variables. For the continuous age variable, we stratified into two groups according to the median age (54 years). As noted above, stratification prevents estimation of the coefficient for the variable in question; we examine these models largely to assess the sensitivity of the other variables to stratification, and to compare model fit. The results of these models are presented in columns 4-6 of Table 2.

Turning first to the model with stratification by age, we note that the model does not fit as well as that which includes age as an explanatory variable. Part of this is undoubtedly due to the loss of information accompanying the dichotomization of the age variable, but it also suggests that, rather than simply acting on the baseline hazard, age is better modeled as directly influencing the hazard rate itself. The models with stratification on South and critical nominations provide better explanations of the data than that stratified on age. But it is also the case that we see a greater degree of coefficient variability across these models relative to the basic Cox regression in column 1. The variable for critical nominations is marginally statistically significant in both stratified models in which it appears, but other coefficients (notably that for political party) show relatively wide swings in value across the different models. Clearly, stratification provides only minimal information about the nature of nonproportionality in the effect of an independent variable, a conclusion concurring with that of Schemper (1992). We therefore turn to alternative techniques for analyzing proportional hazards in Cox models.

Graphical Tests for Proportional Hazards in the Cox Model

Yet another set of approaches to detecting a violation of the proportional hazards assumption in the Cox model is examination of graphical plots. These graphical methods fall into two general categories: log-log plots following estimation of a Cox model, and plots of various residuals against some function of time to assess proportionality. Both of these approaches are widely used in biostatistics and have received a great deal of attention in recent years (e.g. Chen and Wang 1991; Deshpande and Sengupta 1995; Grambsch et. al. 1995; Sleeper and Harrington 1990).

Log-log plots were first suggested by Kalbfleisch and Prentice (1980). They are based on the relationship:

$$ln(-ln(S(t))) = -ln(\lambda) + p\,ln(t) \tag{7}$$

where S(t) is the estimated survival function and p is a scale parameter indicating the extent to which the baseline hazard rate is dependent on time. The assumption of proportional hazards implies that, for different values of a covariate, separate plots of $ln(-ln(S(t)))$ against $ln(t)$ should be parallel. Lines which are diverging, converging or crossing suggest time-varying effects of the covariate in question, and are a signal of violation of the proportional hazards assumption. These log-log plots have become one of the most widely-used methods for assessing the assumption of proportional hazards, largely because they are relatively simple to generate and interpret. At the same time, this approach has come under substantial criticism (e.g. Chastang 1983; Schemper 1992) for its failure to consistently and correctly diagnose instances of nonproportionality.

Because of the deficiencies of the log-log plots, analysts have developed other graphical means of examining proportional hazards. Some of the most important are those based on residuals of the Cox model. The general idea of examining residuals in a duration model is to compare the observed and predicted durations and do so by interpreting the Cox model as a special case of a more general "counting process". This conceptualization was first suggested by Andersen and Gill (1982), and has been widely adopted in the literature on biostatistics (see, generally, Fleming and Harrington 1991). In this conceptualization, "each subject in the data is treated as one observation in a (very slow) Poisson process. A censored subject is thought of not as *incomplete data*, but as one whose event count is still zero" (*Guide to Statistics* 1997). The counting process approach, while analytically difficult, provides a very general, unified way of conceptualizing event history models.

Fleming and Harrington (1991, 163-97) derive residuals for the Cox proportional hazards model by considering the model in the framework of counting processes. The Cox model is shown to be a special case of a more general multiplicative intensity counting model. They define the estimated martingale residual M_i for the non-time-varying Cox model as:

$$M_i = \delta_i - \int_0^{Y_i} \exp(\hat{\beta}' X_i)\, d\hat{\Lambda}_0(s) \tag{8}$$

where δ_i is a censoring indicator and Λ_0 is the estimated (unspecified) baseline hazard. These residuals may be interpreted as the difference between the observed failure of the individual and the conditionally predicted failure. These residuals have several properties reminiscent of OLS residuals: for example, $\Sigma M_i = 0$, and $\mathrm{Cov}(M_i, M_j) = 0$ asymptotically.

One potential drawback of the martingale residuals defined in this fashion is that they are badly skewed. In the context of the Cox model, martingale residuals are lower unbounded but bounded from above by one. Therneau et. al. (1990) suggest a transformation of the residuals, based on the idea of deviance residuals common in generalized linear models, which leads to the residuals being symmetrically distributed around zero (e.g. McCullagh and Nelder 1989). They suggest using *deviance residuals*, defined as:

$$d_i = sign_{M_i}[-2(M_i + \delta_i \log(\delta_i - M_i))] \tag{9}$$

This formulation inflates the martingale residuals close to one, while reducing the magnitude of very large negative values.

Finally, Therneau et. al (1990) note that, from the derivative of the partial likelihood in Cox's model with respect to each coefficient, we can derive "score residuals":

$$L_i = \int_0^{\infty} [x_i(t) - \bar{x}(t)]\, dM_i(t) \tag{10}$$

where $\bar{x}(t)$ is the weighted mean of covariate **x** over the risk set at time t, with weights corresponding to $e^{\beta X}$. These score residuals are specific to each covariate, and within a covariate, sum to zero across observations. These score residuals are closely related to the partial residuals introduced by Schoenfeld (1982), which have been widely used to test the proportional hazards assumption. The

Schoenfeld residuals for each covariate *k* are simply the cross-observation sums of the efficient score residuals:

$$s_{kt} = \sum L_i(t) \tag{11}$$

Fleming and Harrington suggest that martingale, deviance and score residuals may be used for a range of useful purposes, including assessing the extent to which a model conforms to the assumption of proportional hazards. More recent work by various authors (e.g. Grambsch and Therneau 1994; Grambsch, Therneau and Fleming 1995) has extended these techniques considerably. The Schoenfeld residuals are particularly important in testing for the proportional hazards assumption, in two ways. First, if the assumption holds, the Schoenfeld residuals should be a random walk over the range of survival times; that is, there should be no relationship between an observation's residual for that covariate and the length of its survival time. Conversely, if proportional hazards does not hold, the fitted (PH) model will underestimate the hazard during those periods where the hazards are divergent and overestimate it when they are convergent. So if, for example, the hazards for a particular covariate are converging over time, the model will underestimate the impact of that variable for small *t*, and overestimate it for large *t*, a fact that will be reflected in the residuals for that covariate. This intuition has also led to Therneau et. al.'s (1990) test, which uses the maximum of the absolute value of the summed (over time) Schoenfeld residuals as a test for nonproportionality.

A second, related test for proportional hazards is to calculate the correlation between the Schoenfeld residuals for a particular covariate and the rank of the survival time (Harrell 1986). A variation of this test, proposed by Grambsch and Therneau (1994), involves examining the "rescaled" residuals, defined for the *k*th covariate as:

$$s_{kt}^{*} = \hat{\beta}_k + s_{kt} V_{kt}^{-1} \tag{11}$$

where V_{kt}^{-1} represents the contribution to the information matrix. Grambsch and Therneau (1994) show that a smoothed plot of s_{kt}^{*} against survival times gives a direct estimate of $\hat{\gamma}$ in equation (5); in other words, it amounts to a direct test of nonproportionality. In addition, they suggest a global test for nonproportionality, based on the aggregated (across covariates) covariance between the unscaled Schoenfeld residuals and the rank of survival time.

Here, we consider two types of graphical methods, as well as two residual-based tests, for nonproportionality. We first examined *ln*(-*ln*(S)) plots for the three variables that showed evidence of nonproportionality in the previous section; these plots are presented in Figures 1.1 - 1.3. Figure 1.1 suggests age at nomination may have a nonproportional impact on retirement; hazards appear to first converge, then diverge. It is difficult, however, to interpret precisely the nature of the nonproportionality from this plot alone. A similar result is obtained for the South indicator in Figure 1.2: once again, the divergence of the two lines suggests that the assumption of proportionality is suspect with respect to this variable. The most dramatic evidence of nonproportionality appears in Figure 1.3. While the data on critical nominations are relatively sparse, the hazards demonstrate clear convergence over time: predicted survival for "critical nominees", initially longer than that for other justices, decreases at a more rapid rate than that for noncritical appointments.

While the results of the log-log plots thus serve to confirm the suspicions derived from the piecewise models, we must interpret these results with care. The tendency of log-log plots to inaccurately diagnose nonproportionality, particularly in the presence of additional covariates, has been widely remarked upon (e.g. Chastang 1983). For this reason, as well as to demonstrate some more recently developed techniques for assessing nonproportionality, we compare the conclusions drawn from the graphical results in Figures 1.1 - 1.3 with those illustrated by the graphical approach of Grambsch and Therneau (1994). Accordingly, we plot the rescaled Shoenfeld residuals against survival times. As discussed above, a trend in this plot indicates that the PH Cox model is systematically over- or under-predicting the actual hazards at particular time points, and provides strong evidence of nonproportionality. We supplement this graphical approach with statistical tests for nonproportionality based on these residuals; we use the Harrell (1986) correlational test for individual variables, as well as calculating Grambsch and Therneau's (1994) global test for nonproportionality.

Figures 2.1 - 2.3 present the residual plots. Each is a graph of the variable-specific rescaled Schoenfeld residuals for each observation against that observation's survival time. The solid lines are fitted natural splines (*d.f.* = 4), while the dotted lines represent two-standard-deviation confidence intervals. Figure 2.1 indicates a slight positive slope to the residuals for the age variable; the coefficient estimates for age in the basic Cox model in Table 2 appear to be a good fit at earlier retirements, but mildly overpredict survival for longer tenures. The results for the South variable, presented in Figure 2.2, show the opposite effect. The slope of the fitted spline is negative, suggesting a tendency for the Cox model to overpredict the impact of the South variable in earlier time periods, and underpredict for longer durations, a finding consistent with the piecewise regressions in Table 2. Finally, the plots for the critical nominations variable, in Figure 2.3, illustrate a very slight, positive trend in the residuals; again, this is consistent with the results in Table 2, indicating only slight nonproportionality in this variable.

In all three cases, the plots of the rescaled Schoenfeld residuals correspond more closely to the statistical results described in the earlier section than do the inferences one makes from the log-log plots in Figures 1.1 - 1.3. These results receive additional confirmation by the results of the Harrell and Grambsch-Therneau tests. Table 3 reports the estimated values for Harrell's ρ's (i.e., the correlation between the unscaled Schoenfeld residuals and the rank of survival time) as well as his chi-square test for the significance of that relationship. The correlation for Southern justices is strongly statistically significant, confirming the results in Figure 2.2. Likewise, the values of rho for political party and age at nomination are marginally statistically significant. The result for age is also largely confirmatory, but that for political party is unexpected. Previous log-log plots (not shown) did not suggest the presence of any nonproportionality for the political party variable, and the coefficient on this variable shows only minor changes across the two piecewise regressions in Table 2. Conversely, counter to the finding in Figure 1.3, the estimated rho for critical nominations is relatively small, and essentially insignificant. We believe that these findings illustrate the superiority of the residual-based plots over the much more common log-log plots as a graphical means of assessing the nonproportionality assumption.

Finally, the global test for nonproportionality is also marginally statistically significant, suggesting (as is clear from the individual test results) that one or more of the variables in the model exhibits substantial nonproportionality. Because of the fact of nonproportionality in our model, we next consider means of estimating Cox regression models in the presence of nonproportionality.

Estimating Cox Models Without Assuming Proportional Hazards

In a recent survey of methods for nonproportional duration data, Schemper examines various estimation approaches in the presence of nonproportional hazards in the Cox model (1992). There, he rejects using the typical (unweighted) Cox model with fixed covariates, since the average hazard ratio of a nonproportional covariate is badly (typically, under-) estimated and *p* values are correspondingly inaccurate. He also rejects stratification by the nonproportional covariate and a Cox model where the nonproportional covariate is replaced with a time dependent term (Lustbader 1980), because of the difficulty in interpreting the hazard ratios.

However, there are two well-known and accepted estimation approaches for dealing with suspected nonproportionality.[6] First, separate Cox models could be used for two distinct time intervals; such a method, where the time interval split is the median survival time, was shown in Table 1, columns 2 and 3 and discussed above. Second, a Cox model with the addition of an interaction effect between the covariate and some function of time is widely used. The additional term can be tested for statistical significance in the typical way, dividing the coefficient by its standard error. This test amounts to an explicit operationalization of Equation (5), and as a result, is a very general way of addressing nonproportionality. In particular, in the case of a binary covariate, this test is a general one, in that it encompasses all possible alternatives to proportional hazards. So, to test if covariate X_2 has a nonproportional effect, one would include in the model an additional variable $X_2 \times ln(\text{TIME})$.[7] Moreover, in addition to amounting to a test for nonproportionality, this approach has the added advantage of explicitly modeling the nature of the nonproportionality, resulting in a more accurately-specified model and greater validity in one's overall results.

To illustrate this approach, we estimated separate Cox regression models for log-time interactions with each of the three potentially non-proportional variables we have discussed to this point: age at nomination, the South indicator, and critical nominations. In addition, we also estimated a model with log-time interactions for all three variables; all of these models are presented in Table 4.

Turning first to the model with the age interaction, we again note strong evidence of nonproportionality in this variable. In the first year of a justice's tenure, each additional year of age increases the justice's hazard of retirement by nearly 60 percent; that effect decreases over time, however, such that, by a justice's nineteenth year on the Court, age at appointment makes essentially

---

[6]A third approach is the weighted Cox model (WCM). The WCM has weighted estimates of the log hazard ratios, which are weighted at the time points where failures occur. Schemper (1992) advises that if one wants more detailed modeling of the time dependence of a covariate's effect then the Cox model with one or more additional time-dependent covariate terms or separate Cox models for distinct time periods should be used (1992, 464) and so we concentrate on these approaches.

[7]Other forms of the interaction are possible, such as $X_2 \times \text{TIME}$ or $X_2 \times (\text{TIME})^2$, which reflects the possible diversity in nonproportionality. Most treatments, however, favor *ln*(TIME) (e.g. Kalbfleisch and Prentice 1980; Collett 1994), and so we use that approach here.

no difference in his or her hazard of retirement.[8] In this model, we also estimate that, after the nonproportional effects of appointment age are controlled for, Republicans have substantially higher hazards of retirement than Democrats; likewise, Southern justices are more likely to retire earlier than their Northern counterparts, though our confidence in this result is marginal (p = .13, two-tailed). the coefficient for critical nominations, while in the expected direction, also fails to reach statistical significance.

We see similar results in the model interacting *ln*T with the South indicator variable. The effect of age at nomination remains statistically significant, as does the initial effect for Southern justices. The interaction, however, is also statistically significant and negative, indicating that, while Southern justices are initially more likely to retire early, that difference disappears by the fifteenth year on the bench. As is the case with age, the hazards for the two types of justices appear to be converging, consistent with the results in Table 2 and Figure 2.2, but contrary to the diverging log-log plots in Figure 1.2. As in the previous model, the variable for critical nominations is negative but statistically insignificant.

Finally, the effect of a critical nomination is also seen to vary over time. The results in column 3 of Table 4 suggest that critical nominees have an initially *higher* probability of retiring than their non-critical counterparts, but that this difference disappears by the thirteenth year, and that after that period the effect of the variable is negative. Estimates of both the direct effect and its interaction are statistically significantly different from zero. The fact that these results indicate convergence of the hazards over time is unsurprising; what *is* surprising is that critical appointees begin with initially *higher* hazards than noncritical ones; we discuss potential explanations for this anomalous finding below.

All three of the models with single log-time interactions represent substantial improvements over the null. Log-likelihood statistics for each, however, suggest that only the first two improve the fit of the model relative to the direct-effects Cox model in Table 2.[9] These model comparisons are also instructive in our discussion of the "full" model with all three log-time interactions, presented in

---

[8]The interactive model assumes that the coefficient on age changes as a function of *ln*T. Since the direct effect is .451/.153 = 2.95 times the size of the interaction, the point at which they "cancel each other out" occurs at T = *exp*(2.95) = 19.1 years.

[9]The values of these test statistics (distributed as $\chi^2(1)$) are 133.64, 32.80, and 3.22, for the three models, respectively.

column 4 of Table 4. There, the most striking results is that, once nonproportionality in the age variable is taken into account, that for the other two variables considered is seen to be small and statistically insignificant. As for the age variables, note that the coefficients are precisely the same as those estimated in the model in column 1. These facts suggest that the model with the age/log-time interaction alone is doing an adequate job of controlling for the nonproportionality in the model overall, a claim reinforced by the lack of improvement in model fit between columns 1 and 4 ($-2 \ln \mathrm{LR} = 0.20$, p = n.s.).[10] We interpret this result as indicating that, for the models with South and particularly the critical nomination interactions, the statistically significant results are due more to correlation between the interaction terms and the dependent variable than any substantial amount of nonproportionality.

This leads to another issue. The full model also demonstrates a potential problem with extensive use of log-time interactions in Cox models: because of the substantial collinearity between the variables and their interactions, and potentially among the interaction terms themselves, models that include a large number of such interactions run the risk of inflated standard errors. We see this in the full model in Table 4; for both direct effects and the interactions, the estimated standard errors are uniformly and substantially larger than in previous models. This fact warns against reflexive use of such interactions to control, *a priori*, for the possibility of nonproportionality. Instead, we recommend that researchers use the aforementioned graphical and statistical tests to determine if, and for what covariates, nonproportionality is an issue in the Cox model, and include log-time interactions only for those variables that are shown to be substantially nonproportional in their effects by these tests.

Whether we base our substantive discussion on the results in column 1 or column 4, we see clear differences between the model estimates that fail to take account of nonproportionality (in Table 2) and those which do. The impact of partisanship, for example, is substantially larger in the latter case than the former: the Cox regression in Table 2 yields a hazard ratio for Republican justices of 1.76 times their Democratic counterparts, while the ratio in Table 4 is 2.37. We also see a correspondingly larger (negative) influence of the variable for region. Conversely, the effects of critical nominations are more attenuated in the models that account for nonproportionality: the critical/noncritical ratio of 0.56 in Table 2 increases to 0.71 in the first model in Table 4, indicating

---

[10]Note that we cannot, at this point, implement the residual-based tests for nonproportionality as a post-model test here. This is because the model is now explicitly nonproportional in $\ln T$; thus, the aforementioned graphical and statistical tests will, correctly, cause us to reject the assumption that the variables interacted with time are proportional in their effect.

that the effect of the variable is smaller in the latter than the former. In all these instances, it is clear that the presence of some degree of nonproportionality has important influence on the estimated influences of the variables in the model.

Proportionality in the Weibull Model

The Weibull model is probably the most common duration model used in political science (e.g. Alt and King 1994; Bennett 1997; Bennett and Stam 1996; Bueno de Mesquita and Siverson 1996; Werner 1998). The Weibull model may be written as:

$$h(t|X) = exp(\mathbf{X}_i\beta)p[exp(\mathbf{X}_i\beta)t]^{p-1} \tag{12}$$

where p is an estimate of the "shape" parameter.[11] A shape parameter equal to one denotes constant hazards, $p<1$ indicates that hazards are decreasing, while $p>1$ suggests that hazards are rising over time.[12]

In the context of the Weibull model, the assumption of proportional hazards is equivalent to stating that the shape parameter is equal across different values of the independent variables. That is, "in the Weibull model, the assumption of proportional hazards across a number of groups, g, say, corresponds to the assumption that the shape parameter in the baseline hazard function is the same in each group" (Collett 1994, 195). The Weibull model is a proportional hazards model because the ratio of the hazards for individuals i and j depends only on the covariates and not time:

$$\frac{h_i(t)}{h_j(t)} = \left[\frac{\lambda_i}{\lambda_j}\right]^p \tag{13}$$

---

[11]The popularity of the Weibull over the Cox model is somewhat surprising given that the Cox model imposes less restrictions than the Weibull. Both models make the proportional hazards assumption, but the Weibull model makes parametric assumptions about the hazard rate as well. Moreover, it is generally the case that the Cox model will have superior efficiency relative to the Weibull. The Cox model should not be used, however, with a large number of ties, and the Weibull model can easily take into account unobserved heterogeneity. See Grant and Pellegrini (1998) and Reader (1993) for more about heterogeneity.

[12]Many authors (e.g. Lancaster 1990) discuss the Weibull shape parameter in terms of σ, where $\sigma = 1/p$.

Literature on testing the PH assumption much less well developed for the Weibull than the Cox model. In particular, many of the recently-devised residual-based tests for nonproportionality have yet to be applied in the Weibull context. However, similar to testing proportionality in the Cox model, one can fit a separate Weibull model to each of the g groups (Collett 1994). The values of the statistic $-2\mathit{ln}\mathrm{LR}_g$ for each group should be summed and compared to $-2\mathit{ln}\mathrm{LR}_1$ for the model combining all groups of data and thus assuming a common shape parameter. If the difference is not statistically significant according to a chi-squared distribution on *g-1* degrees of freedom, then the assumption of proportional hazards is justified.

The results of estimating a Weibull model are in Table 5, column 1.[13] We then divide the data into two groups according to the median survival time and reestimate the Weibull model in columns 2 and 3. As in the Cox model estimations, the piecewise model does a better job of explaining retirements in justices who serve longer tenures; the model for short-term does not predict better than a null model. The exact pattern of coefficient changes occur: the increase in the age variable in the long-term compared to short-term group, sign reversal for critical nominations, and a large positive impact of the South variable for the short-term group.[14] Overall, we note that time dependence in the effects of these three variables is again suggested.

We next estimate Weibull models with the addition of an interaction between each of the three potentially non-proportional variables and *ln*(time).[15] The model with the age interaction again shows strong evidence of nonproportionality in this variable. Age at appointment demonstrates converging hazards in the Weibull model, with a large positive impact of age initially, which decreases over the justice's tenure. Similar results are found for the two models interaction log-time with South and critical nominations: in both instances, the models show significant converging hazards for these variables. Overall, as we would expect, the results of these models are similar to those for the Cox model in Table 4. Moreover, as was the case for the Cox regressions, the three models with a single

---

[13]The Weibull is run in log relative hazard form rather than log time form. We use the log relative hazard metric because it is easier to compare to the Cox model output. The likelihood function is the same regardless and it is just a matter of changing interpretation if one changes from the log relative hazard metric to the log expected time metric.

[14]Since the Weibull and Cox models are both proportional hazard models, we should be surprised if the estimates, especially the patterns, were widely different.

[15]See Yamaguchi 1991, 107; Teachman and Hayward 1993, 359.

interaction term all have a statistically significant interaction term and all three models are an improvement over the null model. Log-likelihood statistics for each, however, suggest that only the first two improve the fit of the model relative to the direct-effects Weibull model in column 1. Specifically, the log-likelihood statistics, which are distributed as $\chi^2(1)$, are 199.66, 48.68, and 2.32, respectively.

The full model with all three log-time interactions is presented in the last column of Table 5. Again, the most important result is that after taking into account the nonproportionality in age, the nonproportionality for Southern justices and critical nominations is no longer statistically significant. The coefficients for the age variable in the full model and the age only interaction model (columns 7 and 4, respectively) are again very similar, and there is no statistically significant improvement in model fit for the full model compared to the age only interaction model ($-2ln$LR = 1.68, p=n.s).

Another interesting pattern emerges in the estimates values of the shape parameter p across the various Weibull specifications. All models demonstrate rising hazards, as one would expect for a model of retirement. What is different, and of interest, is the way in which the parameter estimates vary. For example, in the piecewise models, the (significantly) larger hazard for the longer tenured justices suggests that the rising hazards are doing so at an increasing rate, a clear sign of nonproportionality. Likewise, we see the greatest variation from the model in column 1 in the model with log-time interactions with age. Here, as well as in the full model, the estimate of sigma is substantially inflated; note, however, that because of the negative sign on the log-time interaction term(s), the extent of the duration dependence in the hazards is attenuated in the models with those interactions.

In general, the Weibull model results on nonproportionality echo the conclusions from the Cox model: there is strong evidence of nonproportionality, the age variable is the source of that nonproportionality, and missing this fact can distort the influences of the variables in the model. Moreover, the model which controls for these effects does a better job of estimating the effects of the model covariates than that which ignores the nonproportionality. More generally, based on simulation results Teachman and Hayward (1993, 361) show that the exact pattern of results will depend on the underlying or baseline hazard rate, the size of the age coefficient, and the size of the age/log time interaction coefficient since these three factors determine the rapidity of change in the ratio of hazards and how quickly the survival function changes.

Conclusions

Duration models are one of the most rapidly-growing quantitative techniques in political science. As models of durations becoming increasingly common in our discipline, it is important that their properties and underlying assumptions be properly understood and appreciated. The case of proportional hazards is a prime example of this importance. Violations of the proportional hazards assumption have the potential to have widespread and serious consequences for political scientists, for two reasons. First, as noted and shown above, violations of this assumption can have dramatic and detrimental effects on our estimates and therefore on our conclusions. Second, as discussed at the outset of the paper, proportionality in the impact of our variables is quite likely to be the exception rather than the rule. Thus, we are faced with a problem that has the potential to be both serious and widespread in our research.

Mindful of this possibility, we have presented, discussed, and implemented a range of current tests for proportionality in duration models. We concentrated on the two most common models in political science, the Cox and Weibull models, both of which assume that hazards are proportional, and presented both graphical and statistical approaches for detecting and ameliorating deviaitons from proportionality. With respect to the graphical techniques, we show that, in general, the widely used log-log plots for nonproportionality do not perform as well as more recently derived residual-based methods. Likewise, we illustrate that log-time interactions, while useful as both tests and remedies for nonproportionality, must be used sparingly to ensure the precision of one's estimated coefficients. More generally, we hope to raise awareness of this critical, yet often overlooked aspect of hazard rate models.

Bibliography

Alt, James, and Gary King. 1994. "Transfers of Governmental Power: The Meaning of Time Dependence." *Comparative Political Studies* 27(2):190-210.

Andersen, P. K. and R. D. Gill. 1982. “Cox’s Regression Model for Counting Processes: A Large Sample Study.” *Annals of Statistics* 10:1100-1120.

Beck, Nathaniel. 1997. "Modeling Space and Time: The Event History Approach." In E. Scarbrough and E. Tanebaum (Eds.) *Research Strategies in the Social Sciences.*

Bennett, D. Scott. Forthcoming. “Parametric Models, Duration Dependence, and Time-Varying Data Revisited.” *American Journal of Political Science.*

Bennett, D. Scott. 1996. “Security, Bargaining, and the End of Interstate Rivalry.” *International Studies Quarterly* 40(1):157-84.

Bennett, D. Scott. 1997. “Testing Alternative Models of Alliance Duration, 1816-1984.” *American Journal of Political Science* 41(July):846-878.

Bennett, D. Scott and Allan C. Stam III. 1996. "The Duration of Interstate Wars, 1816-1985". *American Political Science Review* 90(June):239-257.

Box-Steffensmeier, Janet M. 1996. "A Dynamic Analysis of the Role of War Chests in Campaign Strategy." *American Journal of Political Science* 40(May): 352-71.

Box-Steffensmeier, Janet M., Laura W. Arnold, and Christopher J. W. Zorn. 1997. "The Strategic Timing of Position-Taking in Congress: A Study of the North American Free Trade Agreement." *American Political Science Review* 91(June):324-38.

Box-Steffensmeier, Janet M., and Bradford S. Jones. 1997. "Time is of the Essence: Event History Models in Political Science. " *American Journal of Political Science* 41(October):1414-61.

Breslow, N. E. 1974. "Covariance Analysis of Censored Survival Data." *Biometrics* 30(March):89-99.

Bueno de Mesquita, Bruce, and Randolph M. Siverson. 1995. "War and the Survival of Political Leaders: A Comparative Study of Regime Types and Political Accountability." *American Political Science Review* 89(June): 841-55.

Bullock, Charles. 1972. “House Careerists: Changing Patterns of Longevity and Attrition.” *American Political Science Review* 66(December):1295-1300.

Callas, Peter W., Harris Pastides, and David W. Hosmer. 1998. “Empirical Comparisons of Proportional Hazards, Poisson and Logistic Regression Modeling of Occupational Cohort Data.” *American Journal of Industrial Medicine* 33(1):33-47.

Chastang, C. 1983. “A Simulation Study in a Two-Covariate Survival Model: Importance of the Proportional Hazards Assumption.” *Controlled Clinical Trials* 4:148.

Chen, Chen-Hsin and P. C. Wang. 1991. “Diagnostic Plots in Cox’s Regression Model.” *Biometrics* 47(September):841-850.

Collett, Dave. 1994. *Modelling Survival Data in Medical Research*. London: Chapman & Hall.

Cox, D. 1972. "Regression Models and Life Tables." *Journal of the Royal Statistical Society, Serial B* 34(n2):187-220.

Cunliffe, Marcus. 1972. *American Presidents and the Presidency*. New York: McGraw-Hill.

Deshpande, Jayant V. and Debasis Sengupta. 1995. “Testing the Hypothesis of Proportional Hazards in Two Populations.” *Biometrika* 82(2):251-61.

Farewell, V. T. and R. L. Prentice. 1980. "The Approximation of Partial Likelihood with Emphasis on Case-Control Studies." *Biometrika* 67(August):273-78.

Gasioroworski, Mark J. 1995. "Economic Crisis and Political Regime Change: An Event History Analysis." *American Political Science Review* 89(December): 882-897.

Gaubatz, Kurt Taylor. 1996. “Democratic States and Commitment in International Relations.” *International Organization* 50(1):109-30.

Grambsch, Patricia M. and Terry M. Therneau. 1994. “Proportional Hazards Tests and Diagnostics Based on Weighted Residuals.” *Biometrika* 81(3):515-26.

Grambsch, Patricia M., Terry M. Therneau, and Thomas R. Fleming. 1995. “Diagnostic Plots to Reveal Functional Form of Covariates in Multiplicative Intensity Models.” *Biometrics* 51(December):1469-82.

Grant, J. Tobin, and Pasquale A. Pellegrini. 1998. “Cosponsorship Coalitions in the U.S. House of Representatives.” Presented at the Annual Meeting of the Midwest Political Science Association, Chicago, IL.

Gray, Todd Edward. 1996. “The Impact of Deviations From the Proportional Hazards Assumption on Power in the Analysis of Survival Data.” Unpublished M.P.H. Thesis, Department of Biostatistics, Rollins School of Public Health, Emory University.

Greene, William. 1997. *Econometric Analysis*, 3d ed. New York: Macmillan Publishing Company.

Hagle, Timothy M. 1993. “Strategic Retirements: A Political Model of Turnover on the United States Supreme Court.” *Political Behavior* 15(1):25-48.

Harrell, F. 1986. *The PHGLM Procedure*. SAS Supplementary Library User's Guide, Version 5. Cary, N.C.: SAS Institute.

Jones, Bradford. 1994. "Congressional Careers: A Discrete Choice Model." Ph.D. dissertation, State University of New York, Stony Brook.

Katz, Jonathan N. and Brian R. Sala. 1996. "Careerism, Committee Assignments, and the Electoral Connection." *American Political Science Review* 90(March):21-33.

Kiefer, Nicholas M. 1988. "Economic Duration Data and Hazard Functions." *Journal of Economic Literature* 26(June):646-79.

Kiewiet, D. Roderick and Langche Zeng. 1993. "An Analysis of Congressional Career Decisions, 1947-1986." *American Political Science Review* 87(December):928-941.

King, Gary. 1989. *Unifying Political Methodology*. New York: Cambridge University Press.

King, Gary, James E. Alt, Nancy E. Burns, and Michael Laver. 1990. "A Unified Model of Cabinet Dissolution in Parliamentary Democracies." *American Journal of Political Science* 34(August):846-71.

Lin, D. Y., L. J. Wei and Z. Ying. 1993. "Checking the Cox Model with Cumulative Sums of Martingale-Based Residuals." *Biometrika* 80(3):557-72.

Peto, R. 1972. "Contribution to Discussion Paper by D. R. Cox." *Journal of the Royal Statistical Society, Serial B* 34(n2):205-7.

Reader, S. 1993. "Unobserved Heterogeneity in Dynamic Discrete Choice Models." *Environment and Planning* A25: 495-519.

Reed, William. 1996. "Alliance Duration and Democracy: An Extension and Cross-Validation of 'Democratic States and Commitment in International Relations'." *American Journal of Political Science* 41(July):1072-78.

Ruckman, P. S. Jr. 1993. "The Supreme Court, Critical Nominations, and the Senate Confirmation Process." *Journal of Politics* 55(3):793-805.

S-Plus 4 Guide to Statistics. 1997. Seattle, WA: MathSoft Data Analysis Products Division.

Schemper, Michael. 1992. "Cox Analysis of Survival Data with Non-Proportional Hazard Functions." *The Statistician* 41:455-65.

Schoenfeld, David. 1982. "Partial Residuals for the Proportional Hazards Regression Model." *Biometrika* 69(1):239-41.

Segal, Jeffrey A., Charles M. Cameron and Albert D. Cover. 1992. "A Spatial Model of Roll Call Voting: Senators, Constituents, Presidents and Interest Groups in Supreme Court Confirmations." *American Journal of Political Science* 36(February):96-121.

Singer, Judith D., and John B. Willett. 1993. "It's About Time: Using Discrete-Time Survival Analysis to Study Duration and the Timing of Events." *Journal of Educational Statistics*, Summer 1993, 18(2): 155-95.

Sleeper, Lynn A. and David P. Harrington. 1990. "Regression Splines in the Cox Model with Application to Covariate Effects in Liver Disease." *Journal of the American Statistical Association* 85(December):941-49.

Squire, Peverill. 1988. "Politics and Personal Factors in Retirement from the U.S. Supreme Court." *Political Behavior* 10:180-190.

Ström, Kaare. 1985. "Party Goals and Government Performance in Parliamentary Democracies." *American Political Science Review* 79(September):738-54.

Teachman, Jay D. And Mark D. Hayward. 1993. "Interpreting Hazard Rate Models." *Sociological Methods and Research* 21(February):340-71.

Therneau, Terry M., Patricia M. Grambsch, and Thomas R. Fleming. 1990. "Martingale-Based Residuals for Survival Models." *Biometrika* 77(1):147-60.

Van Winkle, Steve. 1994. "Supreme Court Nominations, 1789-1993: Reevaluating 'Critical Nominations'." Paper presented at the Annual Meeting of the Midwest Political Science Association, Chicago, IL.

Warwick, Paul. 1992. "Economic Trends and Government Survival in West European Parliamentary Democracies." *American Political Science Review* 86(December):875-86.

Wawro, Gregory. 1996. "Advancement in the House of Representatives." Manuscript: Cornell University.

Werner, Suzanne. 1998. "The Possibility of Recurrent Conflict: A Decision to End the Peace." Working Paper: Department of Political Science, Emory University.

Zorn, Christopher J. W. and Steven R. Van Winkle. 1997. "A Competing Risks Model of Supreme Court Vacancies." Manuscript: Emory University.

Table 1

Summary Statistics for Supreme Court Retirement Data

Variable	Mean	Standard Deviation	Minimum	Maximum
Years on Court	15.48	9.82	1	36
Retirement Dummy	0.49	0.50	0	1
Republican Justice	0.55	0.50	0	1
Partisan Agreement	0.90	0.31	0	1
Age at Nomination	52.66	6.93	32	67
Chief Justice	0.11	0.32	0	1
Southern Justice	0.33	0.47	0	1
20^{th} Century Appointment	0.47	0.50	0	1
Critical Nomination	0.15	0.36	0	1

Note: N = 107 (52 retirements, 55 censored). See text for details.

Table 2

Cox Proportional Hazard Results for Supreme Court Retirement Data

Variables	Cox Regression	Piecewise Cox Regressions		Stratified Cox Regressions		
		$T < 15$	$T \geq 15$	By Median Age	By South	By Critical Nomination
Republican Justice	0.564* (0.332)	0.562 (0.497)	0.287 (0.459)	0.505 (0.323)	0.734** (0.359)	0.602* (0.344)
Partisan Agreement	0.387 (0.474)	1.105 (1.078)	0.167 (0.565)	0.326 (0.475)	0.545 (0.503)	0.334 (0.487)
Age at Nomination	0.081** (0.026)	0.017 (0.035)	0.118** (0.040)	n/a	0.076** (0.026)	0.080** (0.026)
Chief Justice	-0.399 (0.469)	-0.263 (0.635)	-0.313 (0.808)	-0.223 (0.463)	-0.241 (0.477)	-0.404 (0.468)
Southern Justice	0.079 (0.373)	1.189** (0.529)	-0.253 (0.558)	0.130 (0.375)	n/a	0.105 (0.371)
20^{th} Century Appointment	0.530* (0.321)	0.699 (0.450)	0.742 (0.499)	0.751** (0.322)	0.488 (0.319)	0.482 (0.327)
Critical Nomination	-0.583 (0.402)	0.563 (0.808)	-0.428 (0.496)	-0.687* (0.407)	-0.538 (0.404)	n/a
*ln*L	-181.71	-73.16	-75.72	-152.94	-151.95	-161.90
$-2(ln\mathrm{L}_{Null}-ln\mathrm{L}_{Model})$	26.87 ($p < .001$)	7.72 (*n.s.*)	26.35 ($p < .001$)	12.11 ($p = .06$)	26.16 ($p < .001$)	23.68 ($p < .001$)
N	107	52	55	107	107	107

Note: Numbers in parentheses are estimated standard errors. One asterisk indicates $p < .10$, two indicate $p < .05$ (two-tailed). See text for details.

Table 3

Results for Harrell (1986) and Grambsch and Therneau (1994)
Tests for Nonproportionality

Variable	Estimated ρ	χ^2 statistic	p - value
Republican Justice	-0.213	2.24	0.13
Partisan Agreement	-0.117	0.71	0.40
Age at Nomination	0.195	2.16	0.14
Chief Justice	-0.064	0.23	0.63
Southern Justice	-0.281	4.86	0.03
20th Century Appointment	-0.009	0.01	0.94
Critical Nomination	0.142	1.01	0.31
Global Proportionality Test	-	13.20	0.07

Note: Tests based on results of Cox Model in Table 2 (column 1). See text for details.

Table 4

Cox Models with Interactions for Nonproportionality

Variables	Separate Models: Age Interaction	Separate Models: South Interaction	Separate Models: Critical Nomination	Full Model
Republican Justice	0.863**	0.611*	0.472	0.814**
	(0.360)	(0.335)	(0.335)	(0.375)
Partisan Agreement	0.360	0.560	0.434	0.355
	(0.558)	(0.490)	(0.483)	(0.561)
Age at Nomination	0.451**	0.055**	0.078**	0.451**
	(0.066)	(0.026)	(0.027)	(0.069)
Chief Justice	-0.275	0.172	-0.379	-0.263
	(0.526)	(0.483)	(0.471)	(0.540)
Southern Justice	-0.655	5.949**	0.102	-0.695
	(0.431)	(1.183)	(0.377)	(1.640)
20^{th} Century	0.472	0.502	0.580*	0.470
	(0.335)	(0.316)	(0.323)	(0.336)
Critical Nomination	-0.347	-0.385	3.385*	0.541
	(0.438)	(0.412)	(1.878)	(2.198)
Age at Nomination $\times$ *ln*(T)	-0.153**	-	-	-0.153**
	(0.022)			(0.023)
Southern Justice $\times$ *ln*(T)	-	-2.167**	-	0.015
		(0.433)		(0.583)
Critical Nomination $\times$ *ln*(T)	-	-	-1.299**	-0.303
			(0.629)	(0.741)
*ln*L	-114.89	-165.31	-180.10	-114.79
$-2(lnL_{Null}-lnL_{Model})$	160.51	59.67	30.10	160.71

Note: N = 107 for all models. Numbers in parentheses are standard errors. One asterisk indicates $p < .10$, two indicate $p < .05$ (two-tailed). See text for details.

Table 5

Weibull Results for Supreme Court Retirement Data

Variables	Weibull	Piecewise Weibull		Weibull Models with *ln*(Time) Interactions			
		$T < 15$	$T \geq 15$	With Age	With South	With Critical Nomination	Full Model
Constant	-9.597** (1.617)	-7.872** (2.442)	-23.354** (3.332)	-36.729** (4.444)	-11.746** (1.771)	-9.54** (1.635)	-37.910** (4.595)
Republican Justice	0.474 (0.329)	0.538 (0.491)	0.296 (0.452)	0.932** (0.365)	0. 573* (0.326)	0.397 (0.332)	0.836** (0.380)
Partisan Agreement	0.314 (0.460)	1.091 (1.068)	0.130 (0.538)	0.348 (0.491)	0.489 (0.462)	0.326 (0.461)	0.382 (0.496)
Age at Nomination	0.060** (0.024)	0.021 (0.034)	0.111** (0.036)	0.610** (0.072)	0.041* (0.024)	0.056** (0.025)	0.643** (0.078)
Chief Justice	-0.258 (0.466)	-0.248 (0.624)	-0.149 (0.815)	-0.622 (0.496)	0.369 (0.475)	-0.219 (0.469)	-0.735 (0.525)
Southern Justice	0.144 (0.372)	1.179** (0.499)	-0.375 (0.552)	-0.781* (0.428)	8.001** (1.132)	0.145 (0.375)	-1.869 (1.175)
20th Century Appointment	0.618* (0.315)	0.606 (0.433)	0.783 (0.499)	0.616* (0.314)	0.566* (0.309)	0.661** (0.317)	0.672** (0.322)
Critical Nomination	-0.407 (0.378)	0.554 (0.798)	-0.206 (0.473)	-0.295 (0.417)	-0.243 (0.388)	3.31 (2.09)	0.588 (1.871)
Interaction of *ln*(T) and Variable	-	-	-	-0.201** (0.022)	-2.816** (0.397)	-1.203* (0.687)	Age: -0.212** (0.024) South: 0.407 (0.421) CN: -0.295 (0.640)
p	1.728**	1.966**	5.078**	11.460**	2.689**	1.787**	11.862**
lnL	-89.43	-41.53	-10.46	10.40	-65.09	-88.27	11.24
$-2(lnL_{Null}-lnL_{Model})$	21.04**	7.97	29.94**	220.70**	69.72**	23.36**	222.38**
N	107	52	55	107	107	107	107

Note: Numbers in parentheses are estimated standard errors. One asterisk indicates $p < .10$, two indicate $p < .05$ (two-tailed). See text for details.