Advance Access publication January 25, 2010

Political Analysis (2010) 18:189–205
doi:10.1093/pan/mpp044

Proportionally Difficult: Testing for Nonproportional Hazards in Cox Models

Luke Keele
Department of Political Science, 2140 Derby Hall, 150 North Oval Mall, Ohio State University, Columbus, OH 43210
e-mail: keele.4@polisci.osu.edu (corresponding author)

The Cox proportional hazards model is widely used to model durations in the social sciences. Although this model allows analysts to forgo choices about the form of the hazard, it demands careful attention to the proportional hazards assumption. To this end, a standard diagnostic method has been developed to test this assumption. I argue that the standard test for nonproportional hazards has been misunderstood in current practice. This test detects a variety of specification errors, and these specification errors must be corrected before one can correctly diagnose nonproportionality. In particular, unmodeled nonlinearity can appear as a violation of the proportional hazard assumption for the Cox model. Using both simulation and empirical examples, I demonstrate how an analyst might be led astray by incorrectly applying the nonproportionality test.

1 Introduction

Political scientists who are interested in estimating event history or time-to-event models have a variety of options available to them. Many of these models require the analyst to make arbitrary decisions about the functional form of the hazard. The Cox model, however, does not assume a functional form for the baseline hazard rate as it is left unspecified in the statistical model. This flexibility has led to the complete dominance of the Cox model in biostatistics and to widespread usage in political science applications. In fact, the authoritative text on event history models in political science recommends that analysts generally fit Cox models (Box-Steffensmeier and Jones 2004).[1]

The Cox model, unlike most parametric survival models, has a variety of accompanying diagnostics. Of these diagnostics, the test for nonproportional hazards is of particular importance. As Box-Steffensmeier and Jones (2004, 131) note: "... whether the proportional hazards assumption holds is arguably the primary concern when fitting a Cox model." Despite the large amount of information available to analysts about Cox model diagnostics, the properties of the test for nonproportional hazards are not fully understood in the political science literature on event history models. Although the standard test for nonproportional hazards is fairly powerful, it also detects a variety of other specification errors. The power of

Author's note: For helpful comments, I thank Jan Box-Steffensmeier, Mark Kayser, Irfan Nooruddin, and the anonymous reviewers. I also thank Hein Goemans for sharing his data.

[1] It is not without some irony that David Cox himself thought analysts should instead be estimating Weibull or other parametric models. He did not fully understand the popularity of the Cox model preferring parametric models which he thought better suited to prediction (Reid 1994, 450).

the test to detect nonproportionality is dependent on the correct specification of the model. Specifically, omitted covariates and interactions as well as nonlinear functional forms can appear as violations of the proportional hazards assumption complicating the detection of nonproportionality. Here, I outline the interaction between specification errors and the proportional hazards assumption for the Cox model.[2] I propose a diagnostic strategy for Cox models that allows analysts to better separate specification errors from the problem of nonproportional hazards. A key aspect of the diagnostic strategy is the use of nonparametric regression techniques for the testing of nonlinear functional forms. The correct diagnostic strategy is important for two reasons. First, it will reduce the bias in the point estimates. Second, it can lead to very different substantive conclusions. For example, nonlinearity and time-varying effects are very different empirical patterns that lead to much altered interpretation of model results. Finally, I use data on international disputes, Food and Drug Administration (FDA) drug approval, and the tenure of political leaders to demonstrate how specification errors affect diagnostics for the Cox model and the conclusions we draw about empirical processes.

2 The Cox Model and Proportional Hazards

In the event history model framework, the outcome variable y_i is the time until the occurrence of some event or ''failure.'' The hazard rate denotes the rate of failure per time unit on the interval $[t, t + \Delta t]$ conditional on survival to or beyond t. This rate can be formally expressed as:

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t | T \leq t)}{\Delta t}. \quad (1)$$

Applied analysts are interested in how a matrix of k covariates, $\mathbf{X}_i$, shift the hazard up or down. We estimate a vector of coefficients $\boldsymbol{\beta}$, which can be transformed to represent how changes in X_i change the hazard rate. There are a variety of ways that we can parameterize the hazard rate for estimates of $\boldsymbol{\beta}$. The Cox proportional hazards model is one widely used statistical model for assessing how a set of covariates alters the hazard rate (Cox 1972). The popularity of the Cox model stems from the fact that we need not assume a specific probability distribution for the hazard rate. Box-Steffensmeier and Jones (2004) recommend that political scientists most often use the Cox model since political theories are rarely specific enough to allow for an a priori choice of the probability distribution for the process in question. The Cox proportional hazards model parameterizes the hazard rate, $h(t)$, in the following way:

$$h(t \mid X_i) = h_0(t) e^{X_i \beta}. \quad (2)$$

In the Cox model, $h_0(t)$ is an unspecified nonnegative function of time called the baseline hazard. $\mathbf{X}_i$ denotes a covariate matrix for subject i where one or more of the covariates may vary over time. Typically, analysts assume the model is linear in the variables—a point I will return to shortly.[3]

[2] Box-Steffensmeier and Jones (2004) briefly discuss the diagnosis of nonlinear function forms, but they do not review the connection between specification errors and nonproportional hazards.

[3] The discussion that follows does not apply to parametric survival models such as the Weibull or log-logistic, but the interaction between specification errors and the proportional hazards assumption does not exist for these parametric models in two ways. First, for proportional hazards models such as the Weibull, there is no method for the detection for nonproportional hazards. Second, for some parametric models, such as the lognormal or log-logistic, there is no assumption of proportional hazards.

A key assumption of the Cox model is that the hazard rates for two observations are proportional to one another and that proportionality is maintained over time. More formally, the relative hazard for any two observations i and j must obey the following relationship:

$$\frac{h_0(t)e^{X_i\beta}}{h_0(t)e^{X_j\beta}} = \frac{e^{X_i\beta}}{e^{X_j\beta}}.$$

Box-Steffensmeier and Zorn (2001) review how the assumption of proportional hazards may be violated in political science contexts, and they outline strategies for the diagnosis and corrections of nonproportional hazards. Correcting for nonproportional hazards is critical since it can lead to biased parameter estimates and the power of statistical tests will decline (Therneau, Grambsch, and Fleming 1990). I, next, outline the standard diagnostic for nonproportional hazards.

Although there are visual methods for the diagnosis of nonproportional hazards, the standard diagnostic is a residuals-based test. (See Box-Steffensmeier and Zorn [2001] and Box-Steffensmeier and Jones [2004] for a review of visual methods of diagnosis.) As is true for generalized linear models, there are multiple ways to define the residuals for a Cox model. For the Cox model, one can define Schoenfeld residuals, martingale residuals, deviance residuals, and score residuals. (See Box-Steffensmeier and Jones [2004] and Therneau and Grambsch [2000] for details on the residual processes for Cox models.) Harrell (1986) first suggested calculating the correlation between the Schoenfeld residuals for each covariate and the rank of the survival time as a diagnostic for nonproportional hazards. Therneau, Grambsch, and Fleming (1990) devised a global test based on the absolute value of the summed Schoenfeld residuals. Grambsch and Therneau (1994) modified these tests by using scaled Schoenfeld residuals. Their modified test based on the scaled Schoenfeld residuals has both a global and covariate specific form. The test statistic for the covariate specific test is as follows. First, define s_k^* as the scaled Schoenfeld residuals for covariate k, g_k as the time scale, and $\bar{g}$ as the average time scale.[4] The test statistic for the Therneau and Grambsch nonproportionality test is

$$T_k = \frac{\left\{\sum(g_k - \bar{g})s_k^*\right\}^2}{dI_k \sum(g_k - \bar{g})^2}, \tag{3}$$

where I_k is the information matrix elements for covariate k and d are the event times. This test statistic follows a χ^2 distribution with 1 df for the covariate-specific version of the test. The test statistic can be interpreted as a measure of the correlation between the covariate-specific residual and event times. Test statistics that exceed 5% critical values are viewed as evidence that the nonproportional hazards assumptions has been violated. Box-Steffensmeier and Zorn (2001) provide a demonstration of this test with political science applications; this test is also implemented in a variety of statistical software applications and is the standard diagnostic for nonproportional hazards in the Cox model. If a covariate fails the test, a common solution is to include in the model an interaction between the covariate and some function of time, typically the natural log of time.

Importantly, this test detects a number of specification errors besides nonproportionality. That is, a significant result for the Therneau and Grambsch nonproportionality test

[4]The scale for time is either linear or log-linear.

might be indicative of a number of model failures. The correct interpretation of a significant test result is not evidence that the hazards are not proportional but instead that if the covariate-dependent specification is correct, then there is evidence that the hazards are not proportional. The critical point is that when the model is misspecified, the Therneau-Grambsch nonproportional hazards test may yield a false-positive result (Therneau, Grambsch, and Fleming 1990; Grambsch and Therneau 1994; Therneau and Grambsch 2000).

What specification errors can lead to mistaken signs of nonproportional hazards? The first is simply the omission of an important covariate. This can take two forms: omitting a covariate or omitting an important interaction among included predictor variables. Second, incorrect functional form for a covariate can also lead to a positive test result for nonproportional hazards. That is including a covariate as linear when its effect is nonlinear may also lead to signs of hazards that are not proportional. Finally, use of the proportional hazards model when a different survival model is appropriate can lead to a significant test result. Here, a parametric model such as the log-logistic that does not assume proportionality may be more appropriate (Therneau and Grambsch 2000). This suggests that some care must be taken with the model specification before testing the proportional hazards assumption. This might seem an obvious point; all analysts desire to use the best possible specification. Of course, if the analyst discovers an important omitted variable, this covariate should be included before using the Therneau and Grambsch test. Care should also be taken to consider possible interactions. There are algorithms that will systematically search for interactions among right-hand side variables (Harrell 2001). Such specification searches are rightly criticized as being ad hoc as they are closely related to stepwise regression algorithms. Analysts should instead use theory to motivate interactions and include them on the right-hand side of the model when necessary. Although urging analysts to correctly specify their models tends to be vacuous advice, testing for nonlinear functional forms is one obvious way to improve model specification that is frequently overlooked. Testing for the correct functional form of the covariate is done by either including polynomial functions of variables or using a nonparametric method such as splines.[5] Given the ease of testing for nonlinear functional forms, it behooves analyst to correct functional forms before testing for nonproportional hazards. Testing for nonlinear functional forms with splines, for example, does not appear to be widespread. I conducted a *JSTOR* search of all the political science articles that used splines since 1995. The search returned a significant number of articles. The search did not return any articles that tested for covariate functional form in a Cox model with splines. Although quadratic terms are occasionally included, often continuous variables are included as linear as a matter of course. Although analysts should test for the correct functional form in any regression model, the interaction between functional form and the proportional hazards test lends extra urgency to the modeling of any nonlinearity.

The interaction between specification and nonproportionality suggests a sequence for the diagnostic process:

- First, the analyst should decide on a final specification taking care not to omit important covariates or relevant interactions.

[5]Many texts also outline a graphical approach to the diagnosis of nonlinear functional forms (Hosmer and Lemeshow 1999; Therneau and Grambsch 2000; Box-Steffensmeier and Jones 2004). Under this method, martingale residuals from a null model are plotted against each covariate and a scatterplot smoother is added to the plot. If the scatterplot smoother fit is nonlinear, this provides evidence of a nonlinear functional form. As Therneau and Grambsch (2000) outline, this method often fails if the various predictors are correlated.

- Second, the correct functional form should be found for continuous covariates using either polynomials or splines.
- At this point, the standard Therneau-Grambsch nonproportionality test should be applied to the model. Significant results should lead the analyst to include an interaction between the offending variable and some function of time. The Therneau and Grambsch nonproportionality test should then be repeated.
- If nonproportional hazards still appear to be present, the analyst at this point may wish to consider a parametric model that does not assume the hazards are proportional.

Next, I present a simulation exercise to demonstrate the interaction between specification and the Therneau-Grambsch test.

2.1 *Simulations*

A simulation demonstrates the need for proper specification before conducting the Therneau-Grambsch test. For the simulation, I start with the following data-generating process (DGP) for the hazard rate:

$$h(t) = h_0(t)\exp(\beta_1 X_1 + \beta_2 X_2 \times \ln(t)).$$

For the above DGP, X_1 is drawn from a uniform distribution and $\beta_1 = 0.1$, X_2 is drawn from a binomial distribution, where the proportion of successes is 0.50 and $\beta_2 = 1$. This component of the DGP should be quite similar to applied data with a mix of continuous and discrete measures. Under this DGP, the effect for X_2 is time varying which induces nonproportional hazards. The DGP also contains several other features common in event history data. The DGP includes a censoring mechanism to induce right censoring. Such censoring is common in event history data as some observations never experience the event in question. For all the simulations, I set the proportion of censored cases to be 0.25. The baseline hazard rate for the DGP is 0.15 and the parametric form of the hazard rate is exponential. I set the number of observations to 100. I altered this DGP in the following two ways:

$$h(t) = h_0(t)\exp(\beta_1 X_1^2 + \beta_2 X_2 \times \ln(t)),$$

$$h(t) = h_0(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2 \times \ln(t) + \beta_4 X_1 X_2).$$

In the first alteration, the DGP now also contains a nonlinear term. In the second DGP, an interaction now exists between the two predictor variables. Using the two altered DGPs, I estimated two incorrectly specified Cox models. In the first, X_1 is included as a linear term. In the second, the interaction between X_1 and X_2 is omitted. I then estimated the models correctly: one taking into account the nonlinear effect of X_1 and one including the interaction. I performed Therneau-Grambsch nonproportionality tests for both the correct specifications and the incorrect specifications. This analysis allows us to infer two things. First, if the misspecification is not due to an omitted time-varying effect, does the nonproportionality test detect it? Second if the specification is correct, does the test then correctly detect nonproportionality? The results are in Table 1.

The evidence in Table 1 is striking. When the specification is incorrect not only due to nonproportionality but also nonlinearity or an omitted interaction, the test indicates that the

Table 1 Grambsch and Therneau nonproportionality tests for simulated data

	Correct specification (test p-value)	*Incorrect specification (test p-value)*
Quadratic model		
X_1	**0.550**	**<.001**
X_2	<.001	<.001
Global test	<.001	<.001
Interactive model		
X_1	**0.606**	**<.001**
X_2	**0.084**	**<.001**
X_1X_2	0.202	—
Global test	**0.247**	**<.001**

Note. p-value less than .05 indicates violation of nonproportional hazards assumption. Cell entries are p-values from the test and not the correlation between residuals and event times. p-value less than .05 indicates violation of proportional hazards assumption. Cell entries are p-values from the test and not the correlation between residuals and event times. Values in bold denote change in inference.

hazards are not proportional for both covariates. Importantly, we know that the hazards are proportional for X_1. When the specification is correct, this is clearly detected by the test in column 1 of Table 1. Once the interaction is omitted or the nonlinearity is left unmodeled, however, we observe in column 2 of Table 1 that the test indicates the effects are time varying when we know this is not the case. Once the specification is corrected either through modeling the nonlinearity or including the appropriate interaction, we also see that the test accurately detects that X_2 has time-varying effects. In the case of the interactive model, the test appears to have low power. This is probably due to the small sample size used for the simulation. This suggests that if the sample size is small, analysts should not be too rigid about using 0.05 as the threshold for the test. The general lesson, however, is clear. Correct specification must precede testing for nonproportionality. Although this might be difficult in terms of finding omitted variables or interactions, all continuous covariates should be tested for nonlinearity. Using the correct testing sequence is also important for substantive interpretation of the model. Distinguishing between an effect that is nonlinear as opposed to time varying is important. The substantive differences in interpretation are considerable. An example helps to clarify the substantive difference. Suppose the one is interested in the length of civil wars and population is one of the covariates. Often we observe nonlinear effects where the effect increases quickly but then levels off or decreases once a threshold is met. It might be the case in this example that as population increases the length of the civil war increases but once the population reaches certain size the association weakens. Contrast that with a time-varying effect for population. Here, we would say that the association between population and the length of the conflict increases or decreases the longer the conflict persists. These are very different empirical patterns that analysts should attempt to distinguish between.

Moreover, analysts should be testing for nonlinearity as a matter of course since the consequences of unmodeled nonlinearity in the Cox model are severe. Past research in biostatistics has documented the effect of failing to model nonlinearity in the variables of a Cox model. Fitting the incorrect functional form to a covariate in a Cox model is a form of misspecification and leads to the same statistical errors: bias and decreased power of tests (Lagakos and Schoenfeld 1984; Struthers and Kalbfleisch 1986; Therneau, Grambsch, and Fleming 1990; Anderson and Fleming 1995). This evidence is based on analytical derivations based on martingale processes but does not give direct evidence

on the magnitude of the problem. LeBlanc and Crowley (1999), however, use simulations to demonstrate that the average model error for a Cox model with an incorrect covariate functional form is *three* times higher than a model that accounts for the nonlinearity.

2.2 *Modeling Nonlinearity*

Before turning to a set of empirical examples, I briefly discuss methods for modeling nonlinearity within Cox models. The simplest method is to fit polynomials to continuous covariates. Here, we would simply place x_i, x_i^2, etc. on the right-hand side of the model. Polynomials, however, may be poor approximations for more complex nonlinear functional forms. Moreover, the fit to the data with polynomials is global and not local which can obscure patterns in the data (Hastie and Tibshirani 1990). Generally, a local form of estimation is preferred to polynomial fits since local estimators provide more flexible nonlinear fits than polynomials (Beck and Jackman 1998; Beck, Katz, and Tucker 1998; Ruppert, Wand, and Carroll 2003; Wood 2006; Keele 2008).

Splines are another method for modeling nonlinearity within regression models.[6] Splines are piecewise polynomial functions that are constrained to join at control points in the data. Typically, the piecewise polynomial functions are cubic and forced to be smooth at the control points, or knots, by forcing the first and second derivatives of the functions to agree at the knots. Semiparametric models combine spline fits with parametric estimates. Such models are often referred to as generalized additive models (GAMs) (Hastie and Tibshirani 1990). Keele (2008) provides a detailed review of splines and their application to social science data. Both Therneau and Grambsch (2000) and Harrell (2001) provide detailed discussions of using splines with Cox models as a method for diagnosing and modeling nonlinearity.

Importantly, a GAM with a spline fit nests linear fits or fits with polynomial transformations as both the linear and quadratic (or cubic) fits are special cases of the spline fit. We can exploit this nested structure of the models to diagnose nonlinear functional forms. A model with a spline fit can be tested against a model with a linear fit using either a Wald or likelihood ratio (LR) test. Testing a model with a spline fit against a linear fit is equivalent to a test with a null of linearity. If the spline model is a statistically significantly better fit to the data, we assume that the functional form is nonlinear. Both the Wald test or LR test are reliable means of assessing whether the functional form for a covariate is nonlinear or can remain in its simpler log-linear form (Hastie and Tibshirani 1990). Statistical software is available for testing nonlinearity within Cox models. In the analysis that follows, I use splines to diagnose nonlinearity.[7] With smoothing splines, the user selects the level of smoothness by selecting the degrees of freedom for each spline. Higher degrees of freedom allow for greater amounts of smoothing. Selection of the smoothing parameter can be done either via visual methods or with the Akaike information criterion (AIC).[8]

[6]There are other methods of local estimation such as local likelihood or kernel density methods (Bowman and Azzalini 1997; Loader 1999). However, neither of these methods, to my knowledge, have been implemented with Cox models.

[7]More specifically, the splines available in software are smoothing splines. Hastie and Tibshirani (1990) show that the fit of smoothing splines is preferable to that of most other splines, and Wood (2006) proves that smoothing splines have the best analytical properties. Moreover, smoothing splines place a penalty on the number of parameters used to avoid overfitting. See Keele (2008) for a review of various spline types.

[8]The best method for selection of the smoothing parameter is generalized cross-validation (GCV; Wood 2006). However, I know of no software that implements spline fits for a Cox model with GCV-selected smoothing. AIC and GCV are the same asymptotically, so this is perhaps the reason why GCV is not included in most software.

3 Empirical Examples

In this section, I present three empirical examples. Each example is from a published paper where a survival model was the mode of empirical analysis. For each example, I replicate the model with a basic specification, a specification that does not include time-covariate interactions. For this specification, I perform the nonproportional hazards tests. I then test for nonlinear covariate functional forms. I then repeat the nonproportional hazards test and report changes in the test. I do not attempt to include any interactions among the variables. Choosing interactions is best done with a theoretical motivation.

3.1 *International Disputes*

The first example is based on a well-known data set on international disputes (Oneal and Russett 1997; Beck, Katz, and Tucker 1998; Reed 2000). This is the same data used by Box-Steffensmeier and Zorn (2001) to demonstrate basic testing methods for nonproportional hazards. The data set is composed of 827 ''politically relevant'' dyads for the period from 1950 to 1985. Each observation is composed of a dyad year, for a total of 20,900 observations with an average of 25.4 years per dyad. The outcome variable is the time until the onset of a militarized interstate dispute between the two nations that make up the dyad. In past work, seven different factors have been identified as important to the risk of a dispute. The seven factors are (1) the level of *democracy* in the dyad, (2) *economic growth*, (3) the presence of an *alliance* between the two nations in the dyad, (4) geographical *contiguity* in the dyad, (5) the ratio of military *capability* between the two nations, (6) the level of intradyadic *trade* measured as a proportion of GDP, and (7) a counter for the number of *previous disputes* within the dyad. All the variables are operationalized as in Beck, Katz, and Tucker (1998) and Box-Steffensmeier and Zorn (2001).[9] Contiguity and previous disputes should increase the risk of a conflict, whereas the rest of the measures should lower the hazard. In this example, there are four continuous covariates: democracy, economic growth, capability ratio, and trade.

As a first step, I estimate a Cox model with the basic specification from Box-Steffensmeier and Zorn (2001) and conduct a Therneau-Grambsch nonproportionality test for this specification. Next, I tested the four continuous covariates for nonlinear functional forms. To do this, I estimate a Cox model with smoothing spline fits for the four variables suspected of nonlinear functional forms and use Wald tests for nonlinearity.[10] The results from the Wald tests are in Table 2. For all four variables, the model fit, according to the Wald test, is better with the spline fit, thus, indicating a ''significant'' amount of nonlinearity.

For the next step in the analysis, I conduct two different Therneau-Grambsch tests for nonproportional hazards. I apply the test to the original specification and to the specification that has been corrected for nonlinear functional forms. This allows the reader to understand whether fitting the nonlinear functional forms alters the test results. The results

[9] Just as in Box-Steffensmeier and Zorn (2001), I omit dyad-years of continuing conflicts as a means of accounting for repeated events. One might also include frailties, but I omit this to ensure comparability across the estimated models.

[10] Fortunately, software for fitting Cox models with spline terms is readily available. S-PLUS, R, and SAS are all capable of fitting such models. The routines in R are identical to those in S-PLUS. I used R to estimate all the models in this article. In R, a Wald test for nonlinearity is reported. The routines in R do not allow for automatic smoothing parameter selection via cross-validation. Currently, AIC is used for automatic selection. In R, nonlinear terms can also be fit in conjunction with stratification and frailty terms. Attempts to include both frailties and nonlinear terms often make convergence more difficult. Both are fairly computationally intensive and to use both asks a lot of the data.

Table 2 Wald nonlinearity tests for international disputes, 1950–85

	χ^2 *Test statistic*	*p-value*
Democracy	17.90	.001
Economic growth	7.78	.051
Capability ratio	134.2	.000
Trade	9.06	.011

Note. Test of spline model against linear fit. *p*-value less than .05 indicates a nonlinear effect.

from the nonproportionality test are in Table 3. With the exception of trade and the capability ratio, all the covariates have time-varying effects on the hazard.[11] Table 3 contains the results from a second Therneau-Grambsch test for nonproportional hazards. Although some of the coefficients still require time-varying terms, there are fewer signs of nonproportional hazards once the nonlinear functional forms have been taken into account. In the first column of Table 3, five of the seven variables showed statistically significant signs of nonproportional hazards, but once spline fits are applied to the four continuous covariates in the analysis, only two of the variables, allies and previous disputes, now require corrections for time-varying effects.

To emphasize how using this different diagnostic strategy alters the substantive form of the model, I fit one final set of models. Table 4 contains a comparison of the final specifications with and without testing for nonlinear functional forms before testing for nonproportional hazards. In the first column of Table 4, we see the specification without testing for the correct functional forms. Under this specification, the correlations between previous disputes and allies and the risk of an international dispute are statistically significant. The second column of Table 4 contains the results from the model fitted after correcting for nonlinearity in the covariate functional forms. As we can see in this example, the correct testing procedure produces a rather different final specification. Again, it is worth

Table 3 Grambsch and Therneau nonproportionality tests international disputes, 1950–85

	Original model (test p-value)	*Spline model (test p-value)*
Democracy	**.007**	**.648**
Economic growth	**<.001**	**.793**
Alliance	<.001	<.001
Contiguity	**<.001**	**.394**
Capability ratio	.121	.509
Trade	.650	.293
Previous disputes	<.001	<.001
Global test	<.001	<.001

Note. p-value less than .05 indicates violation of nonproportional hazards assumption. Cell entries are *p*-values from the test and not the correlation between residuals and event times. The values in bold denote change in inference.

[11]For the variables with spline fits, the Therneau-Grambsch test calculates a test statistic and *p*-value for each knot placement. I report the average *p*-value across the spline fit.

Table 4 Specification comparison with and without spline fits

	Original model[a]	*Spline model*[b]
Democracy	−0.26* (0.15)	—*
Economic growth	−0.29 (1.52)	—*
Alliance	−0.29* (0.19)	−0.22 (0.21)
Contiguity	0.45* (0.19)	0.51* (0.14)
Capability ratio	−0.16* (0.03)	—*
Trade	−2.09 (5.99)	—*
Previous disputes	3.73* (0.17)	3.82* (0.14)
Democracy × ln(time)	0.03 (0.07)	—
Economic growth × ln(time)	−1.41 (0.76)	—
Alliance × ln(time)	0.28* (0.11)	0.26* (0.09)
Contiguity × ln(time)	−0.005 (0.09)	
Previous disputes × ln(time)	−0.97* (0.06)	−0.99* (0.05)

Note. Asterisk represents significance at .05 level for reported coefficient.
[a]Model fit after testing for nonproportional hazards alone.
[b]Model fit after testing for nonlinear functional forms and nonproportional hazards.

emphasizing the different substantive conclusions we would draw from these two specifications. The correct diagnostic strategy leads to three different substantive conclusions. First, for the measure of trade, which is at the heart of substantive concerns; under the incorrect diagnostic strategy, it appears to be unimportant. Under the correct diagnostic strategy, we find a statistically significant but nonlinear association between trade and conflict. Second, under the incorrect specification, the main effect of the alliance variable is statistically significant but under the improved specification focus needs to be placed on the time-varying effect. Third, once economic growth is modeled with a spline, we observe that it decreases the risk of a dispute. Under the incorrect specification, the point estimate for economic growth is not estimated with enough precision for one to conclude that it is correlated with the onset of a dispute. To give the reader a sense of how common this problem might be, I present two more empirical examples.

3.2 *The Politics of Drug Approval*

In the second empirical example, I reanalyze Carpenter's (2002) study of the factors that are related to the time until the FDA grants approval for a new drug. Using a data set for 450 drugs that were subject to FDA review between 1977 and 2000, he finds that the time until approval is unaffected by oversight from Congress and instead is a function of the wealth of the groups representing the disease treated by the drug, media coverage for the disease, and the number of groups that represent the disease. I should note that the analysis, here, is not directly comparable to Carpenter's original analysis. He uses a parametric survival model that is not subject to the nonproportional hazards assumption, while I use a Cox model. The model I estimate also omits the oversight variables reported in Table 2 of Carpenter (2002). Carpenter removes these variables from his later analyses, and I found no evidence that these covariates were ever statistically significant.

The mode of analysis, here, follows that of the previous section. Again, I replicate the original model from the author's analysis. The original specification included a number of continuous covariates. I should note that the author did correct for nonlinearity in that one of the variables was included with a quadratic transformation. There were several

Table 5 Wald nonlinearity tests for FDA Review Duration Models, 1977–2000

	χ^2 *Test statistic*	*p-value*
Incidence of primary indication (per 1000)	7.20	.066
Millions of hospitalization associated with/indication	9.81	.020
Average length of hospitalizations	16.03	.001
National and regional groups	4.69	.200
Nightly TV news disease stories	4.79	.190
Washington Post diseases stories	4.40	.220
Days of congressional hearings on disease	7.02	.071
Order of disease market entry for drug	13.29	.001
FDA drug review staff (full time employees)	17.89	.001

Note. Test of spline model against linear fit. *p*-value less than .05 indicates a nonlinear effect.

other continuous variables, however, that did not have a transformation. To test for nonlinearity, I added spline fits to the continuous variables. Table 5 contains the list of the continuous covariates along with the results from the Wald test comparing the spline fit against the linear fit. For three of the variables, the spline fits as well as a linear fit, while for four other covariates the spline fit is decisively better. Finally, for two of the variables, the *p*-values on the test are slightly above the usual 0.05 threshold. For these two covariates, I plotted the spline estimate. Based on the plots, I modeled the number of days of congressional hearings with a spline, but I left the incidence of primary indication as a linear term.

Again, I conducted the standard Therneau-Grambsch nonproportionality test for both models to document how the diagnostic inferences might be altered depending on whether nonlinearity has been modeled or not. The results of the nonproportional hazards tests for both models are in Table 6. The reader can see in Table 6 that although the global test is not statistically significant, three of the variables clearly appear to need a correction for nonproportionality, whereas a fourth appears to be borderline ($p = .08$).[12] For the Cox model with unmodeled nonlinearity, we find that while the global test does not indicate a problem, three variables violate the proportionality assumption with a fourth being close to the 0.05 threshold. However, once the covariate functional form is correctly modeled, none of the predictors appear to have time-varying effects.

Table 7 contains both a baseline comparison model and a model where I have allowed variables that had significant nonlinear effects to be estimated via smoothing splines. Any variables that did not have a nonlinear effect, I left as linear terms in the model. The model in column 2 of Table 7 is the final model after testing each of the continuous covariates for a nonlinear effect. If the final specification had been based on the incorrect nonproportional hazards test, four log-time interactions would have been included. For example, using the incorrect diagnostic strategy, we would conclude that the number of FDA review staff differed depending on the event time. Instead, we conclude the effect is nonlinear. Examining a plot of the spline fit reveals that increased numbers of FDA review staff increases the likelihood of approval but has diminishing returns above a certain threshold. There are several other differences between the two models. First, once the nonlinearity is fully modeled, the number of national and regional groups no longer has a statistically significant effect on the time until approval. This correlation, however, appears to be

[12]The results reported in the Table 6 for the nonlinear terms are averages across the fitted splines.

Table 6 Grambsch and Therneau nonproportionality tests for FDA Review Duration Models, 1977–2000

	Linear functional form—test p-value	*Nonlinear functional form—test p-value*
Incidence of primary indication (per 1000)	.652	.648
Primary indication is lethal condition	.140	.636
Death rate, primary indication	.294	.585
Primary indication is acute condition	.589	.283
Primary indication results in hospitalizations	.286	.634
Millions of hospitalization associated with/indication	.310	.728
Average length of hospitalizations	**.027**	**.743**
Disease mainly affects women	.551	.944
Disease mainly affects men	.705	.983
Disease mainly affects children	**.049**	**.196**
Orphan drug	**.077**	**.411**
National and regional groups	.889	.478
National and regional groups-squared	.758	–
Nightly TV news disease stories	.846	.755
Washington Post diseases stories	.295	.931
Days of congressional hearings on disease	.119	.651
Order of disease market entry for drug	.851	.703
FDA drug review staff (full time employees)	**.035**	**.337**
Global test	.140	.991

Note. *p*-value less than .05 indicates violation of nonproportional hazards assumption. Cell entries are *p*-values from the test and not the correlation between residuals and event times. Values in bold denote change in inference.

an artifact of unmodeled nonlinearity among the other variables in the model. Second, in the replication model, the number of congressional hearings on the disease is not indicative of the time until a drug will be approved. But that is only true for the linear term. If the association is modeled as nonlinear, we find this covariate is significantly correlated with the time until approval. In general, the more time spent on the drug in congressional hearings, the time to approval declines. Third, the other two variables that were before constrained to be linear in the original model (average length of hospitalizations and order of disease market entry for drug) were not significantly related with the time until approval, but once I model these two variables with splines, the correlations are statistically significant.

This second example has two important implications. First, failing to model nonlinearity in the Cox model can lead to misdiagnosis of nonproportional hazards. In this second example, I found that taking nonlinearity into account completely eliminates the need to respecify the model for this violation of the model assumptions. Second, failure to model nonlinearity can lead to specification errors and incorrect inferences. Here, we find several differences in the basic results of the model once nonlinearity is taken into account. Four covariates that we might have before discounted appear strongly correlated with the time until drug approval. Moreover, another factor, the number of groups that represent a disease appears to be of little importance. Clearly, analysts who use the Cox model need to pay particular attention to modeling of any possible nonlinearity.

Table 7 Cox models of FDA Drug Review, 1977–2000

	Linear functional form	*Nonlinear functional form*
Incidence of primary indication (per 1000)	−0.001 (0.0008)	−0.003* (0.001)
Primary indication is lethal condition	−0.099 (0.226)	0.405^{+} (0.241)
Death rate, primary indication	0.357* (0.252)	0.589* (0.281)
Primary indication is acute condition	0.450 (0.238)	0.088 (0.227)
Primary indication results in hospitalizations	0.170 (0.279)	1.600^{+} (0.950)
Millions of hospitalization associated with/indication[a]	−0.0009 (0.0005)	—*
Average length of hospitalizations[a]	0.212** (0.634)	—**
Average length of hospitalizations × ln(time)[b]	−0.617** (0.218)	—
Disease mainly affects men	0.343 (0.401)	0.718* (0.355)
Disease mainly affects women	0.432 (0.375)	0.019 (0.407)
Disease mainly affects children	−0.086 (1.36)	1.044 (0.845)
Disease mainly affects children × ln(time)[b]	0.249 (0.396)	—
Orphan drug	0.458 (0.785)	0.368^{+} (0.221)
Orphan drug × ln(time)[b]	−0.128 (0.258)	—
National and regional groups	−0.009 (0.009)	−0.004 (0.003)
National and regional groups-squared	0.0004 (0.00005)	—
Nightly TV news disease stories	0.012 (0.016)	−0.033^{+} (0.019)
Washington Post diseases stories	−0.003* (0.002)	0.009** (0.002)
Days of congressional hearings on disease[a]	0.075** (0.023)	—*
Order of disease market entry for drug[a]	−0.007 (0.008)	—*
FDA drug review staff (full time employees)[a]	0.003** (0.002)	—**
FDA drug review staff (full time employees) × ln(time)[b]	−0.018** (0.0006)	—
N	408	408
*ln*L	−1398	−1272.923

Note. Spline terms fit with 4 df. Cell entries are coefficients and asymptotic SEs in parentheses.
[a]Fit with spline in model in the second column.
[b]Interaction omitted from model in the second column.
*$p < .05$, **$p < .01$, $^{+}p < .10$.

3.3 *Tenure of Leaders*

Chiozza and Goemans (2004) examine the correlates of political leaders' tenure. They focus on how conflict and its outcomes influence the length of leaders' time in office, while also controlling for regime type, national characteristics, and economic indicators. They note that several of the covariates fail the standard Therneau-Grambsch nonproportionality test. As a result, they include log-time interactions for the covariates that are flagged by the standard diagnostic. As in the two previous examples, many of the variables are continuous, so one should first test for nonlinearity to improve the specification before conducting the nonproportionality test. Here, I reanalyze their data to compare test results before and after correcting for nonlinear functional forms. First, I replicated the basic model. I then tested each continuous covariate for nonlinearity. The Wald nonlinearity test results are in Table 8. We see from the Wald tests that two of the five continuous measures appear to influence the risk of leader tenure in a nonlinear fashion. Specifically, the measures for trade openness and population appear to alter the risk of tenure termination nonlinearly. Thus, these covariates should be modeled with splines before applying the Therneau-Grambsch test for nonproportional hazards.

Table 8 Wald nonlinearity tests for leaders' office removal

	χ^2 *Test statistic*	*p-value*
Economic development	3.74	.291
Change in economic development	4.43	.220
Trade openness	11.19	.010
Population	16.42	.001
Age	1.49	.690

Note. Test of spline model against linear fit. *p*-value less than .05 indicates a nonlinear effect.

I applied the test for nonproportional hazards to the model with and without the spline fits. The results from both nonproportionality tests are in Table 9. Again, I find that changing the functional form of the model alters the test results. Although many of the predictors still need time-varying effects, both trade openness and population actually have nonlinear effects instead. Once the nonlinearity is taken into account for these variables, they pass the nonproportionality test. This provides additional empirical confirmation that specification testing must precede nonproportionality testing.

Finally, I present what would be the final models based on the incorrect and correct diagnostic procedure in Table 10. The statistical differences between the two models

Table 9 Grambsch and Therneau nonproportionality tests for model of leaders' office removal

	Linear functional form—test p-value	*Nonlinear functional form—test p-value*
Mixed regime	.035	.067
Parliamentary democracy	<.001	<.001
Presidential democracy	<.001	<.001
Civil war	.451	.376
Economic development	<.001	<.001
Change in economic development	.119	.169
Trade openness	**.091**	**.397**
Change in trade openness	.482	.301
Population	**<.001**	**.452**
Age	.009	.010
Previous times in office	.305	.366
Crisis involvement as challenger	.370	.321
Crisis involvement as target	.810	.849
War involvement as challenger	.868	.863
War involvement as target	.395	.338
Win crisis	.779	.766
Lose crisis	.115	.174
Draw crisis	.660	.681
Win war	.883	.859
Lose war	.915	.879
Draw war	.913	.198
Global test	<.001	<.001

Note. p-value less than .05 indicates violation of nonproportional hazards assumption. Cell entries are *p*-values from the test and not the correlation between residuals and event times. The values in bold denote change in inference.

Table 10 Cox models for the determinants of leaders' office removal

	Linear functional form	*Nonlinear functional form*
Mixed regime	4.246** (0.507)	6.650** (0.709)
Mixed regime × ln(time)	−0.482** (0.070)	−0.795** (0.090)
Parliamentary democracy	2.816** (0.555)	6.275** (0.738)
Parliamentary democracy × ln(time)	−0.182* (0.079)	−0.659** (0.100)
Presidential democracy	2.880** (0.675)	5.396** (0.838)
Presidential democracy × ln(time)	−0.266** (0.097)	−0.592** (0.115)
Civil war	0.575** (0.105)	0.457** (0.102)
Economic development	−0.934* (0.0867)	−0.902** (0.078)
Economic development × ln(time)	0.126* (0.013)	0.128** (0.011)
Change in economic development	−0.006* (0.001)	−0.006** (0.001)
Trade openness[a]	−0.029 (0.102)	—**
Trade openness[b] × ln(time)	−0.004 (0.015)	—
Change in trade openness	−0.001* (0.0007)	−0.003** (0.0008)
Population[a]	2.616** (0.108)	—*
Population[b] × ln(time)	−0.394** (0.016)	—
Age	0.182** (0.013)	0.329** (0.011)
Age × ln(time)	−0.026** (0.002)	−0.048** (0.002)
Previous times in office	−0.221** (0.049)	−0.235** (0.050)
Crisis involvement as challenger	−1.115** (0.269)	−1.080** (0.264)
Crisis involvement as target	−0.279 (0.179)	−0.102 (0.173)
War involvement as challenger	−0.106 (0.252)	−0.532** (0.256)
War involvement as target	0.684** (0.229)	0.676** (0.227)
Win crisis	−0.165 (0.232)	−0.483* (0.228)
Lose crisis	−0.683* (0.337)	−0.829** (0.338)
Draw crisis	0.069 (0.227)	0.141 (0.218)
Win war	−0.414 (0.552)	−0.854 (0.589)
Lose war	1.236** (0.296)	1.256** (0.299)
Draw war	−0.288 (0.444)	−0.509 (0.461)
N	9194	9194
lnL	−9215	−9576.1

Note. Spline terms fit with 4 df. Cell entries are coefficients and asymptotic SEs in parentheses.
[a]Fit with spline in model in the second column.
[b]Interaction omitted from model in the second column.
*$p < .05$, **$p < .01$.

are less dramatic than in the two previous examples in that the basic correlational patterns remain the same. There are, however, important differences between the two models since understanding that the effects are nonlinear instead of time varying changes the substance of the model. For example, a time-varying effect for population is a bit nonsensical. Why should the size of the population alter the risk of being removed from office as the tenure of the leader increases? It does make sense, however, to think that leaders of larger countries have longer tenures but that once the population reaches a certain size this effect diminishes. Nonlinearity and time-varying effects are very different empirical patterns, and using the correct functional form can greatly alter the substantive inferences that are drawn from the data.

All three empirical examples provide clear evidence that correcting for nonlinearity alters the nonproportional hazards test results. Although it rarely eliminates the need to include

time-varying interactions, some continuous covariates that might appear problematic no longer do so once the functional form is corrected. In short, correcting the functional form allows for the most direct method to improve the specification without resorting to ad hoc specification searches with little theoretical justification. Although correcting the functional form should be good statistical practice, it is doubly necessary in this instance.

4 Conclusion

The widespread use of event history models has brought increasing sophistication to how these models are applied. Political scientists have moved from simple parametric duration models to more complex models such as those for dealing with competing risks and repeated events (Box-Steffensmeier and Zorn 2002; Gordon 2002). Moreover, the progression from simple to more complex techniques has been quite rapid. The Cox proportional hazards model has gone from rare to widely used. Along with the increased use of these models has come the introduction of diagnostics to ensure that the assumptions that underlie the Cox model are met. One important assumption that must be evaluated is that of proportional hazards. The standard test, the Therneau and Grambsch nonproportional hazards test, sees widespread use as it is easy to conduct and interpret. Unfortunately, application of the standard test requires some care as it is sensitive to several forms of misspecification. Analysts must be aware that omitted predictors, omitted interactions, and nonlinear covariate functional forms can all trigger a significant result from the test. Correcting nonlinear functional forms is an important way to improve the specification before testing for nonproportional hazards. In three different empirical examples, I have demonstrated that correcting the functional form for continuous covariates alters the diagnosis for nonproportionality. Moreover, the importance of the correct testing procedure goes beyond good statistical practice. The substantive difference between a nonlinear effect and a time-varying one is quite large.

References

Anderson, Garnet L., and Thomas R. Fleming. 1995. Model misspecification in proportional hazards regression. *Biometrika* 82:527–41.

Beck, Nathaniel, and Simon Jackman. 1998. Beyond linearity by default: Generalized additive models. *American Journal of Political Science* 42:596–627.

Beck, Nathaniel, Jonathan N. Katz, and Richard Tucker. 1998. Taking time seriously: Time-series-cross-section analysis with a binary dependent variable. *American Journal of Political Science* 42:1260–88.

Bowman, Adrian W., and Adelchi Azzalini. 1997. *Applied smoothing techniques for data analysis*. Oxford, UK: Oxford University Press.

Box-Steffensmeier, Janet M., and Bradford S. Jones. 2004. *Event history modeling: A guide for social scientists*. New York: 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* 45:972–88.

———. 2002. Duration models for repeated events. *Journal of Politics* 64:1069–94.

Carpenter, Daniel P. 2002. Groups, the media, agency waiting costs, and FDA drug approval. *American Journal of Political Science* 46:490–505.

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* 48:604–19.

Cox, D. R. 1972. Regression models and life tables. *Journal of the Royal Statistical Society, Serial B* 34:187–220.

Gordon, Sanford. 2002. Stochastic dependence in competing risks. *American Journal of Political Science* 46:200–17.

Grambsch, Patricia M., and Terry M. Therneau. 1994. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 81:515–26.

Harrell, F. E. 1986. The PHGLM procedure. *SUGI supplemental library user's guide*, ed. Robert P. Hastings. Cary, NC: SAS Institute.

Harrell, Frank E. 2001. *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York: Springer.

Hastie, T. J., and R. J. Tibshirani. 1990. *Generalized additive models*. London: Chapman and Hall.

Hosmer, David W., and Stanely Lemeshow. 1999. *Applied survival analysis regression modeling of time to event data*. New York: Wiley and Sons.

Keele, Luke. 2008. *Semiparametric regression for the social sciences*. Chichester, UK: Wiley and Sons.

Lagakos, S. W., and D. A. Schoenfeld. 1984. Properties of proportional-hazards score tests under misspecified regression models. *Biometrics* 40:1037–48.

LeBlanc, Michael, and John Crowley. 1999. Adaptive regression splines in the Cox model. *Biometrics* 55:204–13.

Loader, Clive. 1999. *Local regression and likelihood*. New York: Springer.

Oneal, John R., and Bruce Russett. 1997. The classical liberals were right: Democracy, interdependence, and conflict, 1950–1985. *International Studies Quarterly* 41:267–94.

Reed, William. 2000. A unified statistical model of conflict and escalation. *American Journal of Political Science* 44:84–93.

Reid, Nancy. 1994. A conversation with Sir David Cox. *Statistical Science* 9:439–55.

Ruppert, David, M. P. Wand, and R. J. Carroll. 2003. *Semiparametric regression*. New York: Cambridge University Press.

Struthers, C. A., and J. D. Kalbfleisch. 1986. Misspecified proportional hazards models. *Biometrika* 73:363–9.

Therneau, Terry M., and Patricia M. Grambsch. 2000. *Modeling survival data: Extending the Cox model*. New York: Springer-Verlag.

Therneau, T. M., P. M. Grambsch, and T. R. Fleming. 1990. Martingale based residuals for survival models. *Biometrika* 77:147–60.

Wood, Simon. 2006. *Generalized additive models: An introduction with R*. Boca Raton, FL: Chapman & Hall/CRC.