Proper Specification of Nonproportional Hazards Corrections in Duration Models

Shuai Jin and Frederick J. Boehmke

Department of Political Science, University of Iowa, Iowa City, IA 52242, USA. Email: shuai-jin@uiowa.edu

Abstract

Parametric and nonparametric duration models assume proportional hazards: The effect of a covariate on the hazard rate stays constant over time. Researchers have developed techniques to test and correct nonproportional hazards, including interacting the covariates with some function of time. Including this interaction term means that the specification now involves time-varying covariates, and the model specification should reflect this feature. However, in situations with no time-varying covariates initially, researchers often continue to model the duration with only time-invariant covariates. This error results in biased estimates, particularly for the covariates interacted with time. We investigate this issue in over forty political science articles and find that of those studies that begin with time-invariant covariates and correct for nonproportional hazards the majority suffer from incorrect model specification. Proper estimation usually produces substantively or statistically different results.

1 Introduction

Commonly used duration models such as the Cox, exponential, Weibull, and Gompertz models assume proportional hazards (PH),[1] meaning that the relative effect of each covariate remains constant over time. When this assumption does not hold, estimating PH models produces biased parameter estimates and decreases the power of significance tests for all parameters (Box-Steffensmeier and Zorn 2001). A variety of tests has been developed to allow researchers to evaluate the PH assumption and determine whether a correction is warranted. Correcting for nonproportional hazards (NPH) typically involves allowing the effect of a variable to change over the course of an ongoing duration spell through the inclusion of an interaction between the offending covariate and some function of time.

These issues are well known in political science and much headway has been made in the diagnosis, correction, and interpretation of NPH over the last two decades (Therneau and Grambsch 2000; Box-Steffensmeier and Zorn 2001; Box-Steffensmeier, Reiter, and Zorn 2003; Box-Steffensmeier and Jones 2004; Keele 2010; Licht 2011; Park and Hendry 2015). Despite this progress, however, our review of the literature indicates that the primary correction for NPH often gets implemented incorrectly even when diagnosed properly. Specifically, researchers working with time-invariant covariates (TICs) often fail to recognize that interacting a variable with a function of time requires a shift in the data structure and estimation to account for the presence of time-varying covariates (TVCs) to avoid bias from misspecification. Even among studies that originally have TVCs scholars often fail to adjust the underlying time intervals to correctly reflect the scale of the analysis time, which dictates the values of the time variable used for the correction for NPH.

We evaluate the implementation of corrections for NPH in forty-three political science studies. Of the twenty-one studies that implement a correction, twelve had the opportunity to implement

Political Analysis (2017)

DOI: 10.1017/pan.2016.16

Published

Corresponding author
Shuai Jin

Edited by
Justin Grimmer

Authors' note: For helpful comments, the authors thank Jan Box-Steffensmeier, Luke Keele, Amanda Licht, Shawna Metzger and Chris Zorn.

1 Note that not all parametric models assume PH. The exponential, Weibull, and Gompertz models do, while log-logistic, log-normal, gamma, and inverse Gaussian models do not (Collett 1994, Chapters 5 & 6).

the correction incorrectly. Of these twelve studies, at least seven implemented it incorrectly. At a minimum then, at least sixty percent of studies likely reported inaccurate results. We reestimate these models with the appropriate time-varying setup and find broad differences in the estimates and the resulting statistical and substantive conclusions. Before discussing these results, we provide a brief review of the PH assumption, ways to test it, and how to correct for NPH.

2 The Proportional Hazards Assumption

The Cox, exponential, Weibull, and Gompertz models all imply PH, but discussion largely focuses on the semiparametric Cox model.[2] Consider $h_0(t)$, which captures the (nonparametric) common baseline hazard across observations:

$$h_i(t \mid X_i) = h_0(t) \exp(X_i\beta), \tag{1}$$

where $\exp(X_i\beta)$ captures how covariates produce deviations from this baseline; i denotes observations $i = 1, \ldots, n$; t points in time; β the vector of coefficients, and X_i the vector of covariates of length K. Considering the ratio of the hazard function at different values of a covariate, x_k and x_k', shows that the relative effect does not change with t:

$$\frac{h_i(t \mid x_k)}{h_i(t \mid x_k')} = \frac{h_0(t) \exp(X_i\beta)}{h_0(t) \exp(X_i'\beta)}, \tag{2}$$

$$= \exp(\beta(x_k - x_k')). \tag{3}$$

The PH assumption has both substantive as well as statistical relevance since substantive theories often imply nonproportionality. Vermunt (1997, p. 101) states that the PH assumption is unrealistic in most applications. Singer and Willett (1993, p. 186) also warns that "violations of the proportionality assumption are the rule, rather than the exception." A common example comes from biomedical research in which treatment effects for certain therapies decrease over time as subjects develop resistance. Take Collett's (1994) illustration of how the effect of an independent variable may change sign over time in cancer treatments. The initial risk of choosing surgery is higher than choosing drug therapy, but the long-run prognosis of patients undergoing surgery is better; thus the surgery treatment increases the hazard initially but decreases it in the long run.

In political science, violations of the PH assumption occur frequently. Licht (2011, p. 228) writes that "essentially, the nature of the political processes of learning, institutionalization, strategic developments, and information transmission which interest us as a discipline are likely to produce frequent violations of the proportional hazards assumption." One example of institutionalization would be that the effects of the size of alliances on the duration of alliances decrease over time, as illustrated in Box-Steffensmeier and Zorn (2001). Larger alliances last longer because they are more vague in purpose and they demand less from their members (Bennett 1997). However, when alliances have survived for long periods of time, they tend to be self-perpetuating (Zorn 2000).

Scholars should therefore test for violations of the PH assumption (Box-Steffensmeier and Jones 2004, p. 132), since misspecified models result in biased parameter estimates and decreased power for statistical tests for all parameters. Schemper (1992, p. 455) notes that this misspecification affects the standard errors for incorrectly specified covariates whose effect actually changes over time due to the incorrect weighting of information at failure times, which in turn spills over to reduce testing power for correctly specified covariates whose effect does not change over time.

2 Note that even for parametric models that do not imply the PH assumption, one would still want to take our recommendations into account when explicitly modeling changes in the effect of a variable vary over time.

Therneau and Grambsch (2000, pp. 132–133) and also Box-Steffensmeier and Zorn (2001, p. 975) mainly recommend three methods to test for nonproportionality in a Cox model: Piecewise regression, Schoenfeld residuals, and explicitly including interactions of covariates and time. Subsequent studies have furthered our understanding of these tests. For example, Keele (2010) warns that tests for PH, including those using the Schoenfeld residuals, can provide false positives due to other model specification issues such as omitted variables, interactions, and nonlinear functional forms.

Much less attention has been paid to testing for PH in parametric models, but the same issues arise. The options for evaluating the PH assumption are more limited but similar. Again, one can divide up the time scale and fit separate Weibull models to the observations that fail in each period (Collett 1994; Box-Steffensmeier and Zorn 2001). A drawback of this piecewise regression approach is the choice of how to divide the time scale can have a significant impact on the results. Zuehlke (2013) suggests modeling the shape parameter as a function of covariates and performing a likelihood ratio test of the null hypothesis that p is constant.

3 Correcting for Nonproportional Hazards

Therneau and Grambsch (2000, pp. 145–147) introduce four strategies to correct for NPH. Box-Steffensmeier and colleagues recommended including an interaction between the offending covariate(s) and some function of time, most commonly the natural log of time (Box-Steffensmeier and Jones 2004, p. 136). More recent work in political science examining options for the best function of time concludes that untransformed time and the natural log of time are not as desirable as the rank or left-continuous Kaplan–Meier transformations of time (Park and Hendry 2015). Whatever form of time one uses, one must then account for the interactive effect to conduct statistical tests and perform substantive evaluation of the effect of variable of interest since its effect changes over time. These can both be accomplished using visualization as recommended by Licht (2011).

Not only does including the interaction terms change the interpretations of the substantive effects, but it also changes the specification and setup of the models. The inclusion of time as a covariate requires data arranged for TVCs, which can necessitate a change in the data structure if the original data only contains TICs. Further, when the original model already has TVCs, researchers need to check whether the data configuration matches the scale of duration time used to create the interaction. For example, if the outcome variable captures days until failure and the initial set of covariates varies on an annual basis, then the data need to be converted from annual to daily observations so that the interaction with time can change on the same scale as the outcome. Failing to do so introduces misspecification into the models. Intuitively, retaining the TIC setup leads the researcher to include the dependent variable on the right hand side in the correction terms for covariates that exhibit NPH. Moving to TVCs changes this to include the value of the covariate interacted with the value of time at every point at which the observation might fail rather than just with the observed failure time. While the early literature made this clear — Collett (1994, p. 194) warns that models including such interactions "can not be fitted by treating [the interaction] in the same manner as time-invariant covariates in the model" and Therneau and Grambsch (2000, p. 147) mentions the interaction "would usually vary continuously in time, which presents a computational challenge" — researchers commonly fail to make this adjustment, often with dramatic consequences for their analysis (as we show shortly).

Fortunately, the solution is easy to implement. Researchers merely need to ensure that they set up their data for TVCs with the scale of the dependent variable matching the time interval between subsequent observations in the data for each spell. Thus, if the dependent variable marks days, then the data should reflect daily observations for each spell. For the Cox model, of course, one only needs to split spells at all observed failure times since it ignores information from times

Table 1. Illustration of how to reconfigure data to correct for NPH.

Original Data

```
stset t, fail(failure) id(caseid)
```

Case ID	X_1	Failure	t_0	t
1	3.0	1	0	37.4
2	13.3	1	0	22.4
3	33.9	1	0	17.9
4	9.1	1	0	2.3

Original Data with Incorrect Adjustment for NPH

```
generate lnt_X1 = ln(t)*X1
```

Case ID	X_1	Failure	t_0	t	lnt_X1
1	3.0	1	0	37.4	10.9
2	13.3	1	0	22.4	41.3
3	33.9	1	0	17.9	97.8
4	9.1	1	0	2.3	7.7

Correctly Reconfigured Data

```
stsplit, at(failure)
stset t, fail(failure) id(caseid)
drop lnt_X1
generate lnt_X1 = ln(t)*X1
```

Case ID	X_1	Failure	t_0	t	lnt_X1
1	3.0	0	0.0	2.3	2.5
1	3.0	0	2.3	17.9	8.7
1	3.0	0	17.9	22.4	9.3
1	3.0	1	22.4	37.4	10.9
2	13.3	0	0.0	2.3	11.3
2	13.3	0	2.3	17.9	38.4
2	13.3	1	17.9	22.4	41.3
3	33.9	0	0.0	2.3	28.7
3	33.9	1	2.3	17.9	97.8
4	9.1	1	0.0	2.3	7.7

Notes: Running Stata's st split command to reconfigure the data is one way to correct for NPH. The other way is to use the tvc option for the Cox estimator in Stata. Above this, we show the Stata code to properly reconfigure the data.[3]

without failures. The time variable used to correct for NPH can then be created and the interaction generated after reorganizing the data. Table 1 shows an example of improper and proper data configurations. The original data structure in the top panel only has TICs. The variable X_1 violates the PH assumption. The incorrect modification to adjust for NPH, shown in the middle panel, involves a single command interacting X_1 with $\ln(t)$. By not reconfiguring the data for TVCs this treats the interaction term as a TIC even though time clearly varies over time. In contrast, the bottom panel shows the data reconfigured for TVCs with the interaction generated afterward.

3 Suites of commands for estimating duration models often make this even easier, for example one can just run Stata's `st split` command and then generate the time interactions or use the `tvc` option for the Cox estimator (which, counterintuitively, does not account for time vary covariates, rather it allows for time-varying effects of variables with the function of time specified by `texp(exp)`).

Because we estimate a Cox model, we only need observations at observed failure times; were this a parametric model, we would need them at all points of the underlying time scale. Analysis, testing, and interpretation can then proceed as usual with TVCs. In the context of the parametric Weibull estimator, Yamaguchi (1991) and Box-Steffensmeier and Zorn (2001) suggest the same approach of including interaction terms to correct for nonproportionality (though see Zuehlke (2013)), meaning that one must again make sure the data vary at the same level of time as the outcome variable.

4 Replications

To evaluate the extent and consequences of incorrect specification of corrections for NPH in political science we attempted to identify as many relevant studies as possible. We started with Park and Hendry's (2015) collection of data from all articles using PH models and published between 1992 and 2012 in the *American Journal of Political Science*, *American Political Science Review*, and *The Journal of Politics*.[4] In addition to the twenty-nine articles from their database, we identified another fourteen studies between 2001 and 2010 that supplied replication data in journals such as *International Studies Quarterly*, *Legislative Studies Quarterly*, and *Journal of Peace Research*.[5] We examined these articles to see if they: (1) included a test for the PH assumption; (2) implemented a correction if needed; and (3) configured their data correctly in the adjusted analysis. We further categorized the ones that did not appropriately reconfigure their data into those that did not have TVCs before the correction and those that did but on a different scale from the outcome variable.[6] Here we report the results of this accounting and provide representative examples to illustrate the possible consequences of incorrectly implementing the NPH correction.[7] For those interested in seeing the consequences with a known data generating process, our supplemental appendix presents the results of a Monte Carlo analysis that also illustrates the bias from misspecifying the correction.

Table A1 in our supplemental appendix reports the results for these forty-three studies. Thirty-one of them reported testing for NPH and of those thirty-one, twenty-one implemented a correction. Twelve studies had the potential for incorrect implementation. Seven of the ten for which we could obtain replication data[8] failed to appropriately reconfigure the data, including three of four that started without TVCs and four of six that had TVCs that were not on the same scale as the outcome variable.

To determine the consequences of implementing the NPH correction without the proper TVC setup, we redid the analysis for all seven studies using the authors' posted data. We reestimated all of the reported results and then reanalyzed the data after appropriate restructuring. We report the full results in our supplemental appendix.[9] Table 2 summarizes the differences for the offending variables. Change in slope indicates the coefficient of the interaction term flips sign. Change in sign means the sign of the combined effect changes. When the coverage of the zero line of the 95% confidence interval changes, a change in significance happens. Change in magnitude records an observable difference in the magnitude of the effect. Alternatively, the substantive conclusion remains the same means that we detected no important change in the statistical and substantive conclusions. We intend this as a rough accounting of the general differences between the features of the results to give a sense of what we found given the subjective nature of many of these conclusions, e.g., does the range of significant effects over time change enough to

4 Park and Hendry (2015) identified seventy articles in total. We evaluated twenty-nine of the seventy articles, including the nineteen articles replicated by Park and Hendry.
5 Among the twenty-nine articles from Park and Hendry (2015), only nine of them perform a correction for NPH.
6 A list of all studies and their categorization can be found in the supplemental appendix.
7 For replication materials, see Jin and Boehmke (2016).
8 Replication data could not be located for Box-Steffensmeier *et al.*'s (2003) replication of Werner (1999) and Bennett (1997).
9 Of the seven studies that did NPH models wrong, two were replication studies that followed the same mistakes as the original studies. We did not present the results of these two studies, since the mistakes were the same.

Table 2. Summary of replication and reanalysis results.

Changes in the effects of the variables offending the PH assumption	
Change in slope	22
Change in sign	13
Change in significance	22
Change in magnitude	23
No big change	20
Total number of variables	51

Notes: The categories are not mutually exclusive. The categorization is based on authors' judgments. See text for more details.

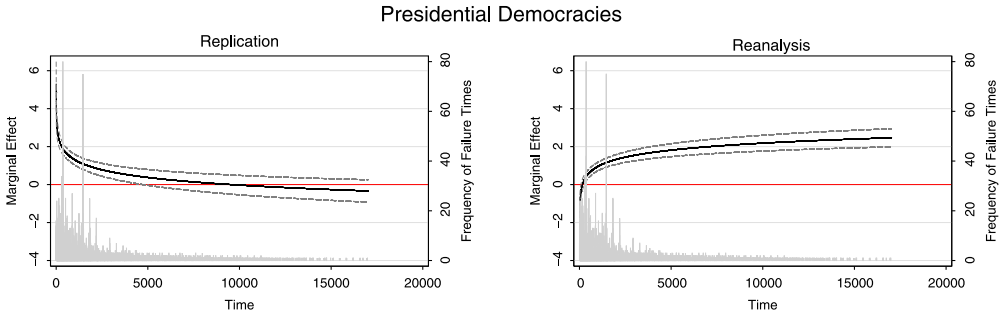

Figure 1. Tenure of political leaders (*CG 2004*).

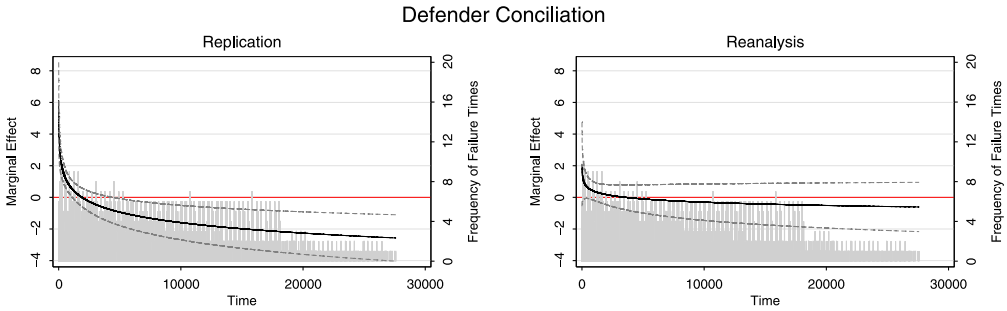

Figure 2. Repetitive military challenges (*Grieco 2001*).

call the difference meaningful? As the table shows, over half the variables exhibit some change that implies a different substantive or statistical conclusion. Nearly half of the effects show a change in sign or slope over time and nearly as many show important differences in magnitude or significance.

While the complete set of plots appears in our supplemental appendix, we illustrate some of the more dramatic differences here. Figure 1 comes from Chiozza and Goemans's (2004) examination of how regime type, conflict and its outcomes affect the tenure of political leaders. In the base models, they include TVCs measured annually, whereas failure time is measured in days. The original results indicate that "... leaders of presidential democracies face significantly higher risks of losing office than leaders who rule over autocratic regimes. ... but the higher hazards of losing office tend to dissipate over time." However, we find the exact opposite results: As shown in Figure 1, at the very beginning, if the leaders of presidential democracies face lower hazards of losing office than autocratic leaders, then the hazards become higher and higher over time.

Figure 2 illustrates a case with differences in significance. This comes from Grieco (2001), which examines the correlates of repetitive military challenges. The dependent variable captures the duration between the end of a given militarized conflict and the beginning of the next conflict in which the two parties retain their roles as challenger and defender. All covariates violate the PH assumption. The results and interpretations of all covariates in the original analysis change dramatically when we implement the correction with TVCs — a large number of covariates lose their statistical significance with the appropriate adjustment.

5 Conclusion

Political science is well aware of the issues related to the PH assumption. The discipline has made progress in diagnosis, correction, and interpretation of NPH models. However, researchers frequently err when implementing the correction. Interacting a variable with time means that one now has TVCs, and the structure of the data must change correspondingly. Misapplication of the correction for NPH can lead to dramatic changes in significance and even sign. Researchers should take care to ensure that their analyses have the proper configuration so as to generate meaningful results.

Supplementary material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2016.16.

References

Bennett, D. Scott. 1997. Testing alternative models of alliance duration, 1816–1984. *American Journal of Political Science* 41(3):846.

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(4):972–988.

Box-Steffensmeier, Janet M., Dan Reiter, and Christopher Zorn. 2003. Nonproportional hazards and event history analysis in international relations. *The Journal of Conflict Resolution* 47(1):33–53.

Chiozza, Giacomo, and H. E. Goemans. 2004. International conflict and the tenure of leaders: Is war still "*ex post*" inefficient? *American Journal of Political Science* 48(3):604–619.

Collett, David. 1994. *Modelling survival data in medical research*. London: Chapman & Hall.

Grieco, Joseph M. 2001. Repetitive military challenges and recurrent international conflicts, 1918–1994. *International Studies Quarterly* 45(2):295–316.

Jin, Shuai, and Frederick J. Boehmke. 2016. Replication data for: Proper specification of nonproportional hazards corrections in duration models. doi:10.7910/DVN/TVNVN3, Harvard Dataverse, V1 [UNF:6:A97gMCZdKECmi612WeSucA==].

Keele, Luke. 2010. Proportionally difficult: Testing for nonproportional hazards in Cox models. *Political Analysis* 18(2):189–205.

Licht, Amanda A. 2011. Change comes with time: Substantive interpretation of nonproportional hazards in event history analysis. *Political Analysis* 19(2):227–243.

Park, Sunhee, and David J. Hendry. 2015. Reassessing Schoenfeld residual tests of proportional hazards in political science event history analyses. *American Journal of Political Science* 59(4):1072–1087.

Schemper, Michael. 1992. Cox analysis of survival data with nonproportional hazard functions. *Journal of the Royal Statistical Society. Series D (The Statistician)* 41(4):455–465.

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 and Behavioral Statistics* 18(2):155–195.

Therneau, Terry M., and Patricia M. Grambsch. 2000. *Modeling survival data: extending the Cox model*. New York: Statistics for Biology and Health Springer.

Vermunt, Jeroen. 1997. *Log-linear models for event histories*. Thousand Oaks, CA: SAGE Publications, Inc.

Werner, Suzanne. 1999. The precarious nature of peace: Resolving the issues, enforcing the settlement, and renegotiating the terms. *American Journal of Political Science* 43(July):912–934.

Yamaguchi, Kazuo. 1991. *Event history analysis*. Thousand Oaks, CA: SAGE.

Zorn, Christopher J. W. 2000. Modeling duration dependence. *Political Analysis* 8(4):367–380.

Zuehlke, Thomas W. 2013. Estimation and testing of nonproportional Weibull hazard models. *Applied Economics* 45(15):2059–2066.