Political Analysis (2024), **32**, 240–255
doi:10.1017/pan.2023.34

ARTICLE

Implementation Matters: Evaluating the Proportional Hazard Test's Performance

Shawna K. Metzger

Department of Political Science, University at Buffalo, Buffalo, NY, USA.

Email: smetzger@buffalo.edu

(Received 14 July 2022; revised 21 March 2023; accepted 5 May 2023; published online 7 November 2023)

Abstract

Political scientists commonly use Grambsch and Therneau's (1994, *Biometrika* 81, 515–526) ubiquitous Schoenfeld-based test to diagnose proportional hazard violations in Cox duration models. However, some statistical packages have changed how they implement the test's calculation. The traditional implementation makes a simplifying assumption about the test's variance–covariance matrix, while the newer implementation does not. Recent work suggests the test's performance differs, depending on its implementation. I use Monte Carlo simulations to more thoroughly investigate whether the test's implementation affects its performance. Surprisingly, I find the newer implementation performs very poorly with correlated covariates, with a false positive rate far above 5%. By contrast, the traditional implementation has no such issues in the same situations. This shocking finding raises new, complex questions for researchers moving forward. It appears to suggest, for now, researchers should favor the traditional implementation in situations where its simplifying assumption is likely met, but researchers must also be mindful that this implementation's false positive rate can be high in misspecified models.

Keywords: duration models; proportional hazards assumption; Monte Carlo simulations

Edited by: Jeff Gill

1. Introduction

Political scientists typically use Grambsch and Therneau's (1994; Therneau and Grambsch 2000) Schoenfeld residual-based test to assess the Cox duration model's proportional hazards (PH) assumption. This assumption states that a covariate x's effect is multiplicative on the baseline hazard, $h_0(t)$. One way proportionality can occur is if x's effect is unconditional on t, a subject's time at risk of experiencing some event. If x's effect *is* conditional on t, it is no longer proportional, as its effect is "time-varying." Failing to account for a covariate's time-varying effect (TVE) produces inefficient estimates, at best, and bias in *all* the covariates' point estimates, at worst (Box-Steffensmeier and Zorn 2001; Keele 2008, 6). Detecting PH violations, then, is a priority for political scientists, given our general interest in explanation and, therefore, accurate estimates of covariates' effects. R's `survival::cox.zph`, Stata's `estat phtest`, and Python's `lifelines.check_assumptions` all currently use Grambsch and Therneau's Schoenfeld-based test (hereafter, "PH test").

Like any specification-related test, the PH test's ability to correctly diagnose PH violations depends on several factors. Examples include the TVE's magnitude, the presence of misspecified covariate functional forms, omitted covariates, covariate measurement error, the number of failures, and sample size (Therneau and Grambsch 2000, sec. 6.6); covariate measurement level (Austin 2018); unmodeled heterogeneity (Balan and Putter 2019); choice of $g(t)$, the function of t on which the covariate's effect is presumed to be conditioned (Park and Hendry 2015); the nature of the PH violation, and the percentage of right-censored (RC) observations (Ng'andu 1997). Each of these affects either the PH test's statistical size or power, impacting the frequency with which we obtain false positives (size) or true positives (power), thereby affecting the test's performance.

New factors affecting the PH test's performance have recently come to light. Metzger (2023c) shows *how* the PH test is calculated also impacts the test's performance. Traditionally, Stata, Python, and R (< `survival` 3.0-10) all compute the PH test using an approximation, which makes certain simplifying assumptions to expedite computation (Metzger 2023c, Appx. A). By contrast, R (≥ `survival` 3.0-10) now computes the PH test in full, using the actual calculation (AC), without any simplifying assumptions.[1] Metzger's (2023c) simulations suggest surprising performance differences between the approximated and actual calculations, with the latter outperforming the former. However, Metzger examines a limited number of scenarios to address her main issues of concern, pertaining to model misspecification via incorrect covariate functional forms among uncorrelated covariates, and leaves more extensive investigations of the calculations' performance differences to future work.

This article uses Monte Carlo simulations to more thoroughly investigate whether the PH test's approximated and actual calculations perform similarly, in general. My simulations show that they do not, but in unexpected ways. Congruent with Metzger (2023c), I find that the AC generally outperforms the approximated calculation when the covariates are uncorrelated, regardless of the amount of right censoring (RC), the way in which RC is induced, the sample size, the PH-violator's time-varying-to-main-effect ratio, or the non-PH-violating covariate's magnitude or dispersion. In these instances, the AC is well sized and well powered, whereas the approximation is also well sized but can be underpowered.

However, in a surprising turn of events, the *approximation* outperforms the AC considerably when the covariates are correlated, even moderately so ($|\mathrm{Corr}(x_1,x_2)| = 0.35$). The AC continues to be well powered, but produces an increasingly large amount of false positives as the correlation's absolute value increases—sometimes as high as 100% of a simulation run's draws. By contrast, the approximation's behavior effectively remains the same as the no-correlation scenario: well sized or very near to it, but sometimes underpowered. These findings have weighty implications because they point to a complex set of trade-offs we were previously unaware of: using an appropriately sized test (the approximation, for the scenarios I check here), while knowing the approximation can also have many false positives in misspecified models (Metzger 2023c), among other potential complications. False positives would lead researchers to include PH violation corrections, likely in the form of a time interaction. Including unnecessary interaction terms results in inefficiency, which can threaten our ability to make accurate inferences (Supplementary Appendix E).

My findings are also weighty because political science applications frequently satisfy the conditions under which the AC is likely to return false positives. I identified all articles using a Cox duration model in eight political science journals across 3.5 years, and examined the correlations between identified PH violators and non-violators.[2] Nearly 87% of the articles have a moderate correlation for at least one violator–non-violator pairing, with an average of 5.15 such pairings per article. By contrast, only ~14% of these articles have easily identifiable features that might prove problematic for the approximation, in theory (fn. 1). To further underscore my findings' implications for political scientists, I also reanalyze a recently published study using the Cox model (Agerberg and Kreft 2020) to show that we reach different conclusions about the authors' main covariate of interest, depending on which PH calculation we use.

I begin by walking through the differences between the PH test's approximated and actual calculations, to provide some sense of why their applied behavior may differ. Next, I describe my simulations' setup. Third, I discuss my simulation results that show the approximation is appropriately sized in far more scenarios than the AC. Fourth, I move to the illustrative application and the different covariate

[1] The change was motivated by the simplifying assumptions' tenability in certain circumstances, particularly for multistate duration models and their signature covariate-by-strata interactions (Therneau 2021, lines 42–45). Competing risks models are a special case of multistate models (Metzger and Jones 2016).

[2] See Supplementary Appendix G for details and a more complete discussion.

https://doi.org/10.1017/pan.2023.34 Published online by Cambridge University Press

effect estimates the two calculations imply. I conclude with a summary and discuss my findings' implications for practitioners.

2. The PH Test Calculation

2.1. Overview

Why might the two calculations perform differently? In short, the approximation makes several simplifying assumptions when calculating one of the formula's pieces.[3]

Grambsch and Therneau's PH test amounts to a score test (Therneau and Grambsch 2000, 132), also known as a Rao efficient score test or a Lagrange multiplier (LM) test. Score tests take the form:

$$LM = U\mathcal{I}^{-1}U', \tag{1}$$

where U is the score vector, as a row, and $\mathcal{I}$ is the information matrix. In a Cox model context, a covariate's entry in the score vector is equal to the sum of its Schoenfeld residuals, making U particularly easy to compute (Therneau and Grambsch 2000, 40, 85). The score test for whether covariate j is a PH violator amounts to adding an extra term for $x_j{*}g(t)$ to the original list of covariates (Therneau 2021), where $g(t)$ is the function of time upon which x_j's effect is potentially conditioned. Usual choices for $g(t)$ include t and $\ln(t)$, but others are possible (and encouraged, in some cases: see Park and Hendry 2015).

To specifically assess whether x_j is a PH violator using the full score test, the expanded U vector's dimensions, U_j^{E}, are $1 \times (J+1)$, where J is the number of covariates in the original model. The $(J+1)$th element contains the score value for the additional $x_j{*}g(t)$ term, calculated by multiplying x_j's Schoenfeld residuals from the original Cox model by $g(t)$, then summing together that product. With a similar logic, the expanded $\mathcal{I}$ matrix for testing whether x_j is a PH violator ($\mathcal{I}_j^{\mathrm{E}}$) has dimensions of $(J+1) \times (J+1)$. It is a subset of the full expanded information matrix ($\mathcal{I}^{\mathrm{E}}$), which is equal to (Therneau 2021, lines 23–33):

$$\mathcal{I}^{\mathrm{E}} = \begin{pmatrix} \mathcal{I}_1 & \mathcal{I}_2 \\ \mathcal{I}_2' & \mathcal{I}_3 \end{pmatrix} \qquad \begin{aligned} \mathcal{I}_1 &= \sum \widehat{V}(t_k), \\ \mathcal{I}_2 &= \sum \widehat{V}(t_k)\, g(t_k), \\ \mathcal{I}_3 &= \sum \widehat{V}(t_k)\, g^2(t_k), \end{aligned}$$

where k is the kth event time $(0 < t_1 < \cdots < t_k < t_K)$ and $\widehat{V}(t_k)$ is the $J \times J$ variance–covariance matrix at time t_k from the original Cox model. We obtain $\mathcal{I}_j^{\mathrm{E}}$ by extracting the rows and columns with indices 1: J and $j + J$ from $\mathcal{I}^{\mathrm{E}}$. This amounts to all of $\mathcal{I}_1$ and the row/column corresponding to x_j in the matrix's expanded portion.[4]

2.2. Implementation Differences

In a basic Cox model with no strata,[5] the biggest difference between the two calculations originates from $\mathcal{I}^{\mathrm{E}}$. The approximated calculation makes a key simplifying assumption about $\widehat{V}(t_k)$: it assumes that $\widehat{V}(t_k)$'s value is constant across t (Therneau and Grambsch 2000, 133–134). The approximation also uses the *average* of $\widehat{V}(t_k)$ across all the observed failures (d), $\overline{V} = d^{-1}\sum \widehat{V}(t_k) = d^{-1}\mathcal{I}_1$, in lieu of $\sum \widehat{V}(t_k)$, because $\widehat{V}(t_k)$ "may be unstable, particularly near the end of follow-up when the number of

[3] For more details, see Metzger's (2023c) Appendix A and the sources therein, as well as this article's Supplementary Appendixes A and B.

[4] The global test statistic takes the same general form except it uses U^{E}, the expanded score vector with *all* J expanded terms, and *all* of $\mathcal{I}^{\mathrm{E}}$ (Therneau and Grambsch 2000, 134).

[5] In the presence of strata, the two calculations have more potential places of divergence. The approximation's simplifying assumptions about $\mathcal{I}^{\mathrm{E}}$ are more tenuous (Therneau and Grambsch 2000, 141–142), to the point that Therneau and Grambsch suggest a tweak in how practitioners use the test (Metzger and Jones 2021). This tenuousness is one of Therneau's motivations for shifting `survival::cox.zph` to the actual calculation (see fn. 1).

https://doi.org/10.1017/pan.2023.34 Published online by Cambridge University Press

subjects in the risk set is not much larger than [$\widehat{V}(t_k)$'s] number of rows" (Therneau and Grambsch 2000, 133–134).

As a consequence of these simplifying assumptions:

1. $\mathcal{I}^{\mathrm{E}}$'s upper-left block diagonal ($\mathcal{I}_1$) is always equal to $\overline{V} = \sum \widehat{V}(t_k)/d$ for the approximation, after the $\overline{V}$ substitution. By contrast, it equals $\sum \widehat{V}(t_k)$ for the AC.
2. $\mathcal{I}^{\mathrm{E}}$'s block off-diagonals ($\mathcal{I}_2$) are forced to equal 0 for the approximation. For the AC, they would be nonzero ($= \sum \widehat{V}(t_k)\, g(t_k)$).
3. $\mathcal{I}^{\mathrm{E}}$'s lower-right block diagonal ($\mathcal{I}_3$) is equal to $\overline{V}\sum g^2(t_k) \equiv \sum \widehat{V}(t_k)\, d^{-1} \sum g^2(t_k)$ for the approximation (Therneau 2021, lines 38–41), after the $\overline{V}$ substitution. By contrast, $\mathcal{I}_3$ would equal $\sum \widehat{V}(t_k)\, g^2(t_k)$ for the AC.

Supplementary Appendix A provides $\mathcal{I}^{\mathrm{E}}$ for both calculations in the two-covariate case, to illustrate.

Consider the difference between the test statistic's two calculations for covariate x_j in a model with two covariates ($J = 2$).[6] For the approximation, it is equal to (Therneau and Grambsch 2000, 134):

$$T_j^{apx} = \frac{\left\{\sum_k s^*_{j,k}\left[g(t_k) - \overline{g(t)}\right]\right\}^2}{d\widehat{V}_{\widehat{\beta}_j} \sum_k \left(\left[g(t_k) - \overline{g(t)}\right]^2\right)}, \tag{2}$$

where $s^*_{j,k}$ are the scaled Schoenfeld residuals[7] for x_j at time k and $\widehat{V}_{\widehat{\beta}_j}$ is $\widehat{\beta}_j$'s estimated variance from the original Cox model.[8]

If we rewrite the approximation's formula using unscaled Schoenfelds, to make it analogous to the AC's formula:

$$T_j^{apx} = \frac{\left\{\sum_k \overbrace{\left[d\left(\widehat{V}_{\widehat{\beta}_j} s_{j,k} + \widehat{\mathrm{Cov}}_{\widehat{\beta}_j,\widehat{\beta}_{\neg j}} s_{\neg j,k}\right) + \widehat{\beta}_j\right]}^{s^*_{j,k}}\left[g(t_k) - \overline{g(t)}\right]\right\}^2}{d\widehat{V}_{\widehat{\beta}_j} \sum_k \left(\left[g(t_k) - \overline{g(t)}\right]^2\right)}, \tag{3}$$

where $s_{j,k}$ is the unscaled Schoenfeld residual for covariate j at time k and $\neg j$ refers to the other covariate in our two-covariate specification.

By contrast, the AC for x_j when $J = 2$ will equal:

$$T_j^{act} = \frac{\left\{\left[\sum_k \widehat{V}(t_k, x_j) \sum_k \widehat{V}(t_k, x_{\neg j})\right] - \left[\left(\sum_k \widehat{\mathrm{Cov}}(t_k, x_j, x_{\neg j})\right)^2\right]\right\}\left\{\sum_k s_{j,k}\left[g(t_k) - \overline{g(t)}\right]\right\}^2}{\left|\mathcal{I}_j^{\mathrm{E}}\right|}, \tag{4}$$

where the various $\widehat{V}$s and $\widehat{\mathrm{Cov}}$ refer to specific elements of $\widehat{V}(t_k)$, the time-specific variance–covariance matrix, and $\left|\mathcal{I}_j^{\mathrm{E}}\right|$ is $\mathcal{I}_j^{\mathrm{E}}$'s determinant.[9] $\left|\mathcal{I}_j^{\mathrm{E}}\right|$ has $J + 1$ terms; when $J = 2$, it equals (before demeaning $g(t_k)$ [fn. 8]):

[6] The approximation's algebraic formula is the same regardless of J's value. The same is not true for the AC.

[7] Under the approximation's simplifying assumption, for a specific t_k, $s^*_k = ds_k\widehat{V}\left(\widehat{\beta}\right)$ (Therneau and Grambsch 2000, 134). R and Stata calculate their scaled Schoenfelds using this formula. Without the simplifying assumption, $s^*_k = s_k\widehat{V}^{-1}(t_k)$ (Therneau and Grambsch 2000, 131).

[8] $g(t_k)$ is eventually demeaned because it makes certain portions of the calculation more numerically stable without affecting the final answer (Therneau 2021, lines 34–35).

[9] Supplementary Appendix B explains the AC formula's origins.

https://doi.org/10.1017/pan.2023.34 Published online by Cambridge University Press

$$\begin{aligned}\left|\mathcal{I}_j^{\mathrm{E}}\right| = &\left\{\left(\sum_{k=1}^{K}\widehat{V}\left(t_k,x_j\right)\right)\left(\left[\sum_{k=1}^{K}\widehat{V}\left(t_k,x_{\neg j}\right)\sum_{k=1}^{K}\widehat{V}\left(t_k,x_j\right)g^2\left(t_k\right)\right]\right.\right.\\ &\qquad\left.\left.-\left[\left(\sum_{k=1}^{K}\widehat{\mathrm{Cov}}\left(t_k,x_j,x_{\neg j}\right)g\left(t_k\right)\right)^2\right]\right)\right\}\\ &+\left\{\left(\sum_{k=1}^{K}\widehat{\mathrm{Cov}}\left(t_k,x_j,x_{\neg j}\right)\right)\left(\left[\sum_{k=1}^{K}\widehat{\mathrm{Cov}}\left(t_k,x_j,x_{\neg j}\right)g\left(t_k\right)\sum_{k=1}^{K}\widehat{V}\left(t_k,x_j\right)g\left(t_k\right)\right]\right.\right.\\ &\qquad\left.\left.-\left[\sum_{k=1}^{K}\widehat{\mathrm{Cov}}\left(t_k,x_j,x_{\neg j}\right)\sum_{k=1}^{K}\widehat{V}\left(t_k,x_j\right)g^2\left(t_k\right)\right]\right)\right\}\\ &+\left\{\left(\sum_{k=1}^{K}\widehat{V}\left(t_k,x_j\right)g\left(t_k\right)\right)\left(\left[\sum_{k=1}^{K}\widehat{\mathrm{Cov}}\left(t_k,x_j,x_{\neg j}\right)\sum_{k=1}^{K}\widehat{\mathrm{Cov}}\left(t_k,x_j,x_{\neg j}\right)g\left(t_k\right)\right]\right.\right.\\ &\qquad\left.\left.-\left[\sum_{k=1}^{K}\widehat{V}\left(t_k,x_{\neg j}\right)\sum_{k=1}^{K}\widehat{V}\left(t_k,x_j\right)g\left(t_k\right)\right]\right)\right\}. \end{aligned} \tag{5}$$

2.3. *Implications*

Equations (3) and (4) diverge in two major places. Both manifest in the AC (Equation (4)):

1. The additional, non-Schoenfeld term in the numerator (shaded light gray);
2. A substantially more complex denominator. The AC's denominator is one consequence of $\mathcal{I}_2 \neq 0$, as Supplementary Appendix B explains. Additionally, $g(t)$ only appears *inside* the k-summations involving $\widehat{V}\left(t_k\right)$ for the AC's denominator, which stems from $\mathcal{I}_3 \neq \sum \widehat{V}\left(t_k\right) d^{-1} \sum g^2\left(t_k\right)$.

T_j is distributed asymptotically χ^2 when the PH assumption holds (Therneau and Grambsch 2000, 132), meaning T_j's numerator and denominator will be identically signed.

Understanding when each calculation is likely to be appropriately sized (few false positives) and appropriately powered (many true positives) amounts to understanding what makes T_j larger. A higher T_j translates to a lower p-value, and thus a higher chance of concluding a covariate violates PH, holding T_j's degrees of freedom constant. The key comparison is the numerator's size relative to the denominator. Specifically, we need a sense of (1) when the numerator will become larger relative to the denominator and/or (2) when the denominator will become smaller, relative to the numerator.

However, the numerator's and denominator's values are not independent within either calculation. Moreover, the numerator and the denominator do not simply share one or two constituent quantities, but *several* quantities, often in multiple places (and sometimes transformed), making basic, but meaningful comparative statics practically impossible within a given calculation, let alone comparing across calculations. This interconnectivity is one reason I use Monte Carlo simulations to assess how each calculation performs.

The additional term in T_j^{act}'s numerator hints at one factor that may make the calculations perform differently: the correlation among covariates. $\widehat{\mathrm{Cov}}\left(t_k,x_j,x_{\neg j}\right)$ appears in the AC for $J = 2$, both in the numerator's non-Schoenfeld term (Equation (4), light gray shading) and all three terms in the denominator.[10] $\widehat{\mathrm{Cov}}\left(t_k,x_j,x_{\neg j}\right)$ is equal to (Therneau and Grambsch 2000, 40):

$$\widehat{\mathrm{Cov}}\left(t_k,x_j,x_{\neg j}\right) = \left[\frac{\sum_{r\in R(t_k)}\left\{\exp(XB)x_j x_{\neg j}\right\}}{\sum_{r\in R(t_k)}\exp(XB)}\right] - \left[\frac{\sum_{r\in R(t_k)}\left\{\exp(XB)x_j\right\}}{\sum_{r\in R(t_k)}\exp(XB)} \times \frac{\sum_{r\in R(t_k)}\left\{\exp(XB)x_{\neg j}\right\}}{\sum_{r\in R(t_k)}\exp(XB)}\right], \tag{6}$$

[10] The time-specific variance–covariance matrix, of which $\widehat{\mathrm{Cov}}\left(t_k,x_j,x_{\neg j}\right)$ is one element, does not appear in the approximation because of its simplifying assumption. The simplifying assumption's equivalent, $\widehat{\mathrm{Cov}}_{\widehat{\beta}_j,\widehat{\beta}_{\neg j}}$, *does* appear in the approximation, and this quantity is equal to the sum of all the individual $\widehat{\mathrm{Cov}}\left(t_k,x_j,x_{\neg j}\right)$s. However, $\widehat{\mathrm{Cov}}_{\widehat{\beta}_j,\widehat{\beta}_{\neg j}}$ acts as a scaling factor inside the approximation's Schoenfeld-related term, for $\neg j$'s Schoenfeld only (Equation (3)). By contrast, it contributes toward the overall weight for the AC's entire Schoenfeld term (Equation (4)), giving rise to the two calculations' potentially different behavior from $\mathrm{Corr}\left(x_j,x_{\neg j}\right)$.

where $r \in R(t_k)$ represents "observations at risk at t_k^-" and XB is the at-risk observation's linear combination. Correlated covariates would impact $x_j x_{\neg j}$'s value, which eventually appears in both bracketed terms. Generally speaking, as $|\text{Corr}(x_j x_{\neg j})|$ increases, $|x_j x_{\neg j}|$ increases, thereby increasing $\left|\widehat{\text{Cov}}(t_k, x_j, x_{\neg j})\right|$'s value.

More broadly, each formula provides guidance as to which features of the data-generating process (DGP) might be useful to vary across the different simulation scenarios. Consider the pieces that appear in either equation:

- $\widehat{V}(t_k)$. In the AC, the individual elements of $\widehat{V}(t_k)$ appear in both the numerator and the denominator (e.g., $\widehat{\text{Cov}}(t_k, x_j, x_{\neg j})$, as previously discussed for the *correlation among covariates*). In the approximation, $\widehat{V}(t_k)$ appears only indirectly via $\widehat{V}\left(\widehat{\beta}\right)$, the model's estimated variance–covariance matrix, as $\widehat{V}\left(\widehat{\beta}\right) = \mathcal{I}^{-1}$ and $\mathcal{I} = \sum \widehat{V}(t_k)$. Portions of $\widehat{V}\left(\widehat{\beta}\right)$ appear in the approximation's numerator, as part of the scaled Schoenfeld calculation ($\widehat{V}_{\widehat{\beta}_j}$, $\widehat{\text{Cov}}_{\widehat{\beta}_j, \widehat{\beta}_{\neg j}}$), and in its denominator ($\widehat{V}_{\widehat{\beta}_j}$).
- $\sum_{r \in R(t_k)} \exp(XB)\theta$, where θ is a generic placeholder for a weight,[11] appears in multiple places in both calculations: namely, within the formula for $\widehat{V}(t_k)$'s individual elements and within the unscaled Schoenfeld formula. $\exp(XB)$ is an at-risk observation's risk score in t_k, meaning its (potentially weighted) sum speaks to the total amount of weighted "risk-ness" in the dataset at t_k.[12] The *riskset's general size* at each t_k, then, is relevant.
- $\exp(XB)$ also suggests that the *covariates' values*, along with their *respective slope estimates*, are of relevance. Additionally, the covariates are sometimes involved with the weights (see fn. 11), producing another way in which their values are relevant.
- t, the duration. It ends up appearing demeaned in both calculations, $g(t_k) - \overline{g(t)}$ (see fn. 8). The demeaning makes clear that t*'s dispersion* is relevant.
- Only observations experiencing a failure are involved in the final steps of the $\widehat{V}(t_k)$ and Schoenfeld formulas, implying the *number of failures* (d) is relevant.

3. Simulation Setup

I use the `simsurv` package in R to generate my simulated continuous-time durations (Brilleman *et al.* 2021).[13] All the simulations use a Weibull hazard function with no strata, a baseline scale parameter of 0.15, and two covariates: (1) a continuous, non-PH-violating covariate ($x_1 \sim \mathcal{N}$) and (2) a binary, PH-violating covariate ($x_2 \sim \text{Bern}(0.5)$). x_2's TVE is conditional on $\ln(t)$. Making the PH violator a binary covariate gives us a best-case scenario, because others' simulations suggest that the Schoenfeld-based PH test's performance is worse for continuous covariates than for binary covariates (Park and Hendry 2015).

I design my simulations to address whether there are performance differences between the approximated and actual PH test calculations in a correctly specified base model, where x_1 and x_2 are the only covariates.[14] I vary a number of other characteristics that can impact the PH test's performance, per Section 1's discussion. Some of the characteristics' specific values are motivated by existing duration model-related simulations. In total, I run 3,600 different scenarios, derived from all permutations of the

[11] $\theta = \{1, x_j x_{\neg j}, x_j, x_{\neg j}\}$ appear in different portions of the $\widehat{V}(t_k)$ and/or s_k formulas (e.g., Equation (6)).

[12] $\exp(XB)$ is always nonnegative, but θ can be negative. Thus, each additional observation at risk at t_k does not necessarily produce a larger value of $\sum_{r \in R(t_k)} \exp(XB)\theta$.

[13] See Metzger (2023a,b) for replication materials.

[14] I use the phrase "base model" to acknowledge it is the model we would first estimate to test for PH violations. In terms of matching the true DGP, it is not the outright correct model (Supplementary Appendix E) because it lacks any PH-violation corrections.

characteristics I list in Supplementary Appendix C.[15] The results section's discussion focuses primarily on five of these characteristics:

- Three Weibull *shape parameter (p)* values $\{0.75, 1, 1.25\}$, producing scenarios with decreasing, flat, and increasing baseline hazards, respectively. $p = 1$ matches Keele (2010) and Metzger (2023c). Varying p impacts t's dispersion by affecting how quickly subjects fail. Higher shape values reduce t's dispersion, all else equal.
- Two *sample sizes* $\{100, 1{,}000\}$. The first matches Keele (2010) and Metzger (2023c). I run $n = 1{,}000$ to check whether the $n = 100$ behavior persists when the PH test's asymptotic properties are likely in effect.
- Five *levels of correlation* between the two covariates $\{-0.65, -0.35, 0, 0.35, 0.65\}$. I use the `BinNor` package to induce these correlations (Demirtas, Amatya, and Doganay 2014).[16] I run both positive and negative correlations to verify that the behavior we observe is independent of the correlation's sign, as the formulas suggest. The results are indeed roughly symmetric for the scenarios I run here. Therefore, I only report the positive correlation results in text, but the supplemental viewing app (see fn. 15) has the graphs for both.
- Two *RC patterns*. In one pattern, I randomly select rc% subjects and shorten their observed duration by (an arbitrarily selected) 2%. In the second, I censor the top rc% of subjects such that their recorded durations are at the $(100 - rc\%)$th percentile. The first ("random RC") corresponds to a situation where subjects become at risk at different calendar times, whereas the second ("top rc%") corresponds to a situation where all subjects become at risk at the same calendar time, but data collection ends before all subjects fail. For two otherwise identical scenarios (including d's value), the top rc% pattern gives me another way to affect t's dispersion without impacting other quantities in either formula, because t's highest observed value is restricted to its $(100 - rc\%)$th percentile.
- Three *RC percentages* (rc%) $\{0\%, 25\%, 50\%\}$. The 25% matches Keele (2010), Metzger (2023c), Park and Hendry's (2015) moderate censoring scenario, and is near Ng'andu's (1997) 30% scenario. The 50% matches Park and Hendry's (2015) heavy censoring scenario and is near Ng'andu's (1997) 60% scenario. Manipulating rc% allows me to vary d across otherwise comparable scenarios.

As Supplementary Appendix C discusses, I also vary the pattern regarding x_2's effect (specifically, the ratio of x_2's TVE to its main effect), the recorded duration's type, x_1's mean, and x_1's dispersion.

For each of these 3,600 scenarios, I estimate a correctly specified base model to determine whether PH violations exist, as discussed previously. I then apply the two PH test calculations and record each calculation's p-values for every covariate. I report the PH tests' p-values for $g(t) = \ln(t)$ from both calculations, to match the DGP's true $g(t)$.[17,18]

In the ideal, I would run 10,000 simulation draws for each of the 3,600 scenarios because of my interest in p-values for size/power calculations (Cameron and Trivedi 2009, 139–140). However, the estimating burden would be prohibitive. Additionally, while I am interested in seeing how each calculation performs against our usual size/power benchmarks, my primary interest is comparing how the calculations perform *relative to one another*. Having fewer than 10,000 draws should affect both calculations equally, provided any imprecision is unaffected by any of the calculations' performance

[15] The simulation results for all 3,600 scenarios are viewable in a supplemental viewing app accessible through the replication materials.

[16] The actual correlation in a draw's dataset may not equal the specified correlation, similar to `MASS::mvrnorm(empirical=FALSE)`'s behavior. The actual correlation's mean within a scenario closely matches its specified correlation (see viewing app's "Empirical Correlations" tab [fn. 15]).

[17] The viewing app also contains graphs of the p-values' distribution ("p-values: Distributions" tab [fn. 15]).

[18] I use 25 of these scenarios to begin exploring how the final model estimates, with PH corrections, behave (Supplementary Appendix E).

differences (i.e., the simulations might give an imprecise estimate of statistical size, but both calculations would have the same amount of imprecision). Nonetheless, I compromise by running 2,000 simulations per scenario.

4. Simulation Results

The key quantity of interest is the rejection percentage ($\hat{r}_p$), the percent of p-values < 0.05, from the PH test for each calculation–covariate pairing within a scenario.[19] For x_1, the non-PH violator, this value should be 5% or lower, corresponding to a false positive rate of $\alpha = 0.05$. For PH-violating x_2, 80% or more of its PH test p-values should be less than 0.05, with 80% representing our general rule of thumb for a respectably powered test.[20] Our first priority typically is evaluating whether a statistical test's calculated size matches our selected nominal size, α. Our second priority becomes choosing the best-powered test, ideally *among those with the appropriate statistical size* (Morris, White, and Crowther 2019, 2088)—a caveat that will be relevant later.

I report $\hat{r}_p$ along the horizontal axis of individual scatterplots grouped into 3 × 3 sets, where each set contains 45 scenarios' worth of results. The set's rows represent different Corr(x_1,x_2) values, and its columns represent different shape parameter values. Each scatterplot within a set, then, represents a unique Corr(x_1,x_2)–shape combination among a set of scenarios that share the same true linear combination, sample size, recorded duration type, and values for x_1's mean and dispersion. I split each scatterplot into halves and report the results from random RC on the left and top rc% RC on the right, with the halves' dividing line representing 0% of a scenario's p-values < 0.05 ($\hat{r}_p = 0\%$) and the scatterplot's side edges representing $\widehat{r}_p = 100\%$. I use short, solid vertical lines within the plot area to indicate whether a particular covariate's $\widehat{r}_p$ should be low (non-PH violators ⇒ size; closer to halves' dividing line) or high (PH violators ⇒ power; closer to scatterplot's edges). Within each half, I report the three censoring percentages using different color symbols, with darker grays representing more censoring.[21]

I report one of the scatterplot sets in text (Figure 1) to concretize the discussion regarding correlated covariates' effect, as it exemplifies the main patterns from the results.[15] I then discuss those patterns more broadly.

4.1. *Specific Scenario Walkthrough*

Figure 1 shows the simulation results for $x_1 \sim \mathcal{N}(0,1)$ where $XB = 0.001x_1 + 1x_2\ln(t)$, $n = 100$, and the estimated model uses the true continuous-time duration. In general, if the two tests perform identically, the circles (approximation) and triangles (AC) should be atop one another for every estimate–RC pattern–rc% triplet in all scatterplots. Already, Figure 1 makes clear that this is not the case.

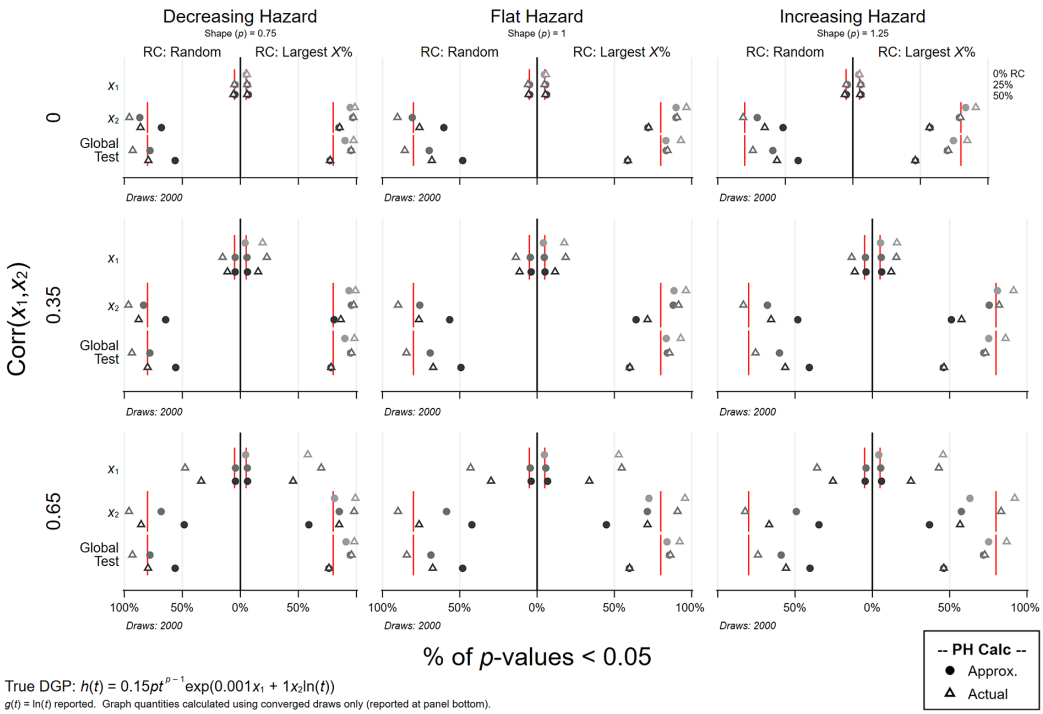
I start by comparing my current results with those from previous work, to ground my findings' eventual, larger implications. Figure 1's top row, second column most closely corresponds to Metzger's (2023c) simulations. This scatterplot, Corr(x_1,x_2) = 0, $p = 1$, with top 25% RC (scatterplot's right half, medium gray points), is analogous to her Section 3.3's "correct base specification" results.[22] My top 25%

[19] $\hat{r}_p$'s 95% confidence interval (CI) will be equal to $\hat{r}_p \pm 1.96\left(\sqrt{\frac{\hat{r}_p(1-\hat{r}_p)}{S}}\right)$ (Morris, White, and Crowther 2019, 2086). If $\hat{r}_p = 5\%$ (for non-PH violators), $S = 2{,}000$ produces a 95% CI of [4.1%, 6.0%]. If $\hat{r}_p = 80\%$ (for PH violators), $S = 2{,}000$ produces a 95% CI of [78.2%, 81.7%].

[20] I use "the calculation's power" as a discussion shorthand, but the type of statistical test or the calculation of that test does not affect statistical power, strictly speaking (Aberson 2019, ch. 1).

[21] The RC pattern is irrelevant for scenarios with 0% RC. I arbitrarily assigned 0% RC to each scatterplot's right half.

[22] There, x_2's distribution is the same as here and the top 25% largest durations are censored, but (a) $x_1 \sim \mathcal{U}[0,1]$, (b) x_1's effect size is 1, not 0.001, and (c) Metzger's linear combinations have either an additional x_1 quadratic term or an additional x_1x_2 interaction. Regarding (b), Figure 1's general patterns are unchanged if I increase x_1's effect size to 1 (see supplemental viewing app [fn. 15]).



True DGP: $h(t) = 0.15pt^{p-1}\exp(0.001x_1 + 1x_2\ln(t))$

$g(t) = \ln(t)$ reported. Graph quantities calculated using converged draws only (reported at panel bottom).

Figure 1. Illustrative simulation results, nonnegative correlations only ($n = 100$).
Negative correlations omitted for brevity; Corr(x_1,x_2) < 0 follow similar patterns as Corr(x_1,x_2) > 0. Vertical lines represent target $\widehat{r}_p$ for a well-sized (x_1) or well-powered (x_2) test.

RC results match Metzger (2023c): both calculations are appropriately sized or close to it (for x_1: 6.5% [approx.] vs. 5.5% [actual]) and both calculations are well powered (for x_2: 90.2% [approx.] vs. 90.6% [actual]). The calculations having similar size and power percentages also mirrors Metzger's (2023c) Section 3.3.

The story changes in important ways once Corr(x_1,x_2) ≠ 0 (moving down Figure 1's columns). Figure 1 shows that the AC performs progressively worse as Corr(x_1,x_2) becomes larger, evident in how the triangles representing non-PH violator x_1's false positive rate move away from each scatterplot's $\hat{r}_p = 0\%$ dividing line. The AC returns an increasingly large number of false positives for x_1 that far surpass our usual 5% threshold, nearing or surpassing 50% in some instances. This means we become more likely to conclude, incorrectly, that a non-PH-violating covariate violates PH as it becomes increasingly correlated with a true PH violator. Despite the AC's exceptionally poor performance for non-violating covariates, it continues to be powered just as well or better than the approximation for PH violators, regardless of |Corr(x_1,x_2)|'s value. These patterns suggest that the AC rejects the null too aggressively—behavior that works in its favor for PH violators, but becomes a serious liability for non-PH violators.

By contrast, correlated covariates only marginally affect the approximated calculation. The approximation has no size issues across |Corr(x_1,x_2)| values—it stays at or near our 5% false positive threshold, unlike the AC. However, it does tend to become underpowered as |Corr(x_1,x_2)| increases, meaning we are more likely to miss PH violators as the violator becomes increasingly correlated with a non-PH violator. While this behavior is not ideal, it suggests that practitioners should be more mindful of their covariates' correlations, to potentially contextualize any null results from the approximation.

Finally, Figure 1 shows these general patterns for both calculations persist across panels. More specifically, the patterns are similar when the baseline hazard is not flat (within the scatterplot set's

Table 1. False positive %: Corr(x_1,x_2) = 0 vs. ≠ 0, n = 100.

	Corr = 0 better? (no. of combos)		Difference in FP% (average)	
Correlation	Approx.	Actual	Approx.	Actual
−0.65	199	360	−0.58	−33.75
−0.35	189	360	−0.57	−8.99
0.35	180	359	−0.57	−9.09
0.65	171	360	−0.54	−33.47

Note: "Better" = lower FP% (x_1). Difference in FP% = (FP% for Corr = 0) − (FP% for this row's correlation) within comparable scenarios; negative values: Corr = 0 performs better by x% percentage points. Number of unique combinations: 360.

rows), for different censoring percentages (within a scatterplot's half), and for different RC types (across a scatterplot's halves, for the same rc%).

4.2. Broader Correlation-Related Patterns: Descriptive

The AC's behavior is the more surprising of the two findings, but similarly as surprising, Figure 1's patterns are not unusual. They are representative of the AC's behavior in nearly all the 1,800 scenarios where n = 100. There are 360 unique combinations of the Weibull's shape parameter (p), x_2's TVE-to-main-effect ratio, recorded duration type, RC pattern, RC percentage, x_1's mean, and x_1's dispersion for n = 100. Of these 360, the AC's false positive rate for |Corr(x_1,x_2)| ≠ 0 is worse than the comparable Corr(x_1,x_2) = 0 scenario in 359 of them (99.7%; Table 1's left half, second column). For the lone discrepant combination,[23] three of the four nonzero correlations perform worse than Corr(x_1,x_2) = 0. Or, put differently: for the AC, out of the 1,440 n = 100 scenarios in which Corr(x_1,x_2) ≠ 0, 1,439 of them (99.9%) have a higher false positive rate than the comparable Corr(x_1,x_2) = 0 scenario. When coupled with the number of characteristics I vary in my simulations, this 99.9% suggests that the AC's high false positive rate cannot be a byproduct of p, the PH violator's TVE-to-main-effect ratio, the way in which the duration is recorded, the RC pattern or percentage, or x_1's magnitude or dispersion.

Other AC-related patterns from Figure 1 manifest across the other scenarios as well. In particular, like Figure 1, the AC's false positive rate gets progressively worse in magnitude as |Corr(x_1,x_2)| increases across all 360 combinations (Table 1's right half, second column). On average, the AC's false positive rate for Corr(x_1,x_2) = 0 is ~9 percentage points lower compared to |Corr(x_1,x_2)| = 0.35 and ~33.6 percentage points lower compared to |Corr(x_1,x_2)| = 0.65.

The AC's most troubling evidence comes from Figure 1's equivalent for n = 1,000 (Figure 2). With such a large n, both calculations should perform well because the calculations' asymptotic properties are likely active. For Corr(x_1,x_2) = 0, this is indeed the case. Both calculations have 0% false positives for x_1 (size) and 100% true positives for x_2 (power), regardless of p, the RC pattern, or the RC percentage (Figure 2's first row). However, like Figure 1's results, the AC's behavior changes for the worst when Corr(x_1,x_2) ≠ 0. It continues to have a 100% true positive rate (Figure 2's last two rows, x_2 triangles), but also has up to a 100% *false* positive rate, and none of its Corr(x_1,x_2) ≠ 0 false positive rates drop below 50% (Figure 2's last two rows, x_1 triangles). Also, like Figure 1, the approximation shows no such behavior for Corr(x_1,x_2) ≠ 0.

These patterns for the AC appear across the other n = 1,000 Corr(x_1,x_2) ≠ 0 scenarios, of which there are 1,440. Corr(x_1,x_2) = 0 outperforms the comparable Corr(x_1,x_2) ≠ 0 scenario in *all* 1,440 scenarios. Figure 2's 100% false positive rate also bears out with some regularity for the AC (330 of 1,440 scenarios [22.9%]); in all 330, |Corr(x_1,x_2)| = 0.65. In the remaining 1,110 scenarios, the AC's

[23] p = 1.25, coerced start–stop duration, 50% RC with random censoring, $x_1 \sim \mathcal{N}(60, 1)$, and x_2's main effect and TVE signed the same. Descriptively, Corr = 0.35 outperforms Corr = 0 by only 0.35 percentage points.

https://doi.org/10.1017/pan.2023.34 Published online by Cambridge University Press

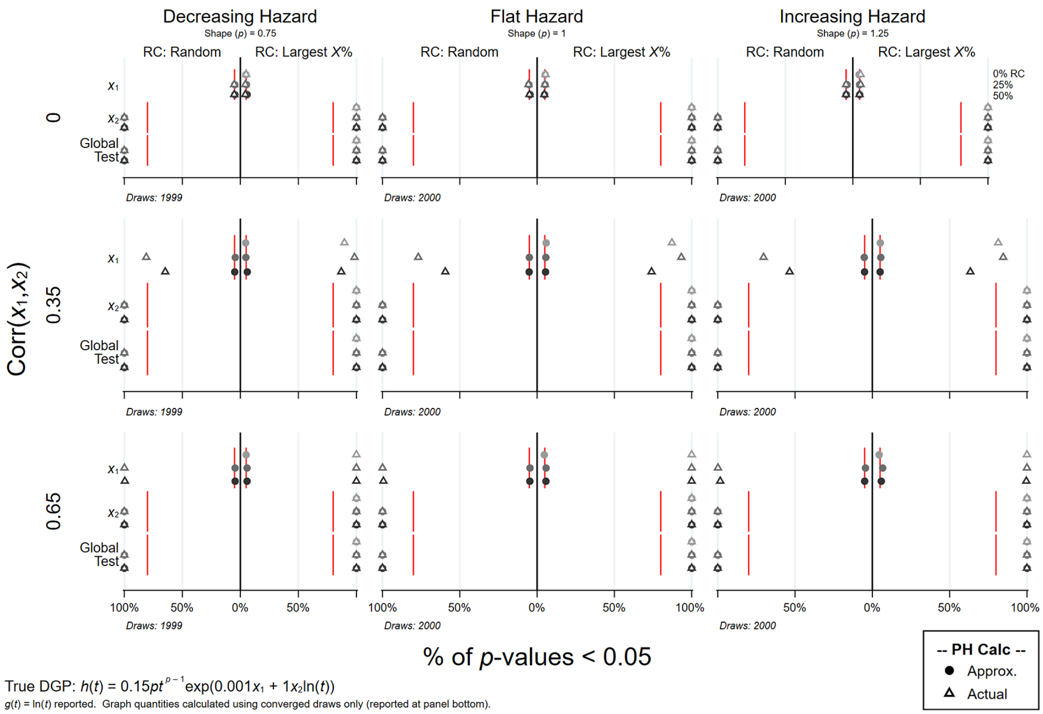


Figure 2. Illustrative simulation results, nonnegative correlations only (n = 1,000).
Negative correlations omitted for brevity; Corr(x_1,x_2) < 0 follow similar patterns as Corr(x_1,x_2) > 0. Vertical lines represent target $\hat{r}_p$ for a well-sized (x_1) or well-powered (x_2) test.

lowest false positive rate is 22.6%. The AC's behavior is so troubling because properly sized tests are typically our first priority in traditional hypothesis testing, as Section 4's opening paragraph discusses. These results indicate that the AC is far from properly sized, whereas the approximation has no such issues. Taken overall, my simulation results for both sample sizes suggest that we should avoid using the AC for situations mimicking the scenarios I examined here, at minimum, if not also more broadly, provided we temporarily bracket other issues that may arise from using the approximation—a theme I return to in my closing remarks.

5. Illustrative Application

The simulations show that the AC is particularly susceptible to detecting violations, with many false positives when true PH violations do exist, but the PH violator(s) are even moderately correlated with non-violators. Political scientists typically correct for PH violations using an interaction term between the offending covariate and $g(t)$. The potential perils of including an unnecessary interaction term are lower than excluding a necessary one, in relative terms. For any model type, unnecessary interactions produce less efficient estimates.[24] This increased inefficiency can take a particular toll in the presence of many such unnecessary interaction terms, which would occur in a Cox model context when a PH test reveals many potential PH violations.

Using the AC to diagnose PH violations for Agerberg and Kreft (2020; hereafter "A&K") illustrates the potential perils of the AC's high false positive rate and its ramifications for inference. A&K's study assesses whether a country having experienced high levels of sexual violence (SV) during a civil

[24] Supplementary Appendix E confirms as much using a small subset of scenarios.

Table 2. Agerberg and Kreft: PH test *p*-values.

Variable	Approx.	Actual
LSVC	0.451	0.073
HSVC*	0.546	0.010
GDPPC (ln)	0.831	0.141
Polity	0.638	0.020
Conflict intensity: Low	0.029	0.035
Conflict intensity: High	0.032	0.021
Peacekeeping operation	0.788	0.083
Foreign aid (ln)	0.109	0.716
Regional quota diffusion	0.252	0.147
Islamic heritage	0.145	0.016
Women's civil liberties	0.959	0.009
Electoral system: PR	0.714	0.726
Electoral system: Mixed	0.502	0.336
Total: no. of viols. ($p \leq 0.05$)	2	6

Note:* = key independent variable. PH test $g(t) = t$, p-value threshold = 0.05 (A&K, Online Appendix B).

conflict ("high SV conflicts" [HSVC]) hastens the country's adoption of a gender quota for its national legislature, relative to non-HSVC countries.[25] They find support for their hypotheses, including the one of interest here: HSVC countries adopt gender quotas more quickly compared to countries experiencing no civil conflict. In their supplemental materials, the authors check for any PH violations using the approximation, with $g(t) = t$. Two of their control variables violate at the 0.05 level (Table 2's "Approx." column), but correcting for the violations does not impact A&K's main findings.

However, a different story emerges if I use the AC[26] to diagnose PH violations.[27] The AC detects six violations in A&K's model—three times as many as the approximation. Importantly, A&K's key independent variable, HSVC, is now a PH violator according to the AC, implying that the effect of high sexual violence during civil conflict is not constant across time. Furthermore, examining HSVC's effect (Gandrud 2015) from a fully corrected model[28] shows that HSVC's hazard ratio (HR) is statistically significant for only $t \in [5,15]$ (Figure 3's solid line).

The t restriction matters because 93% of the countries in A&K's sample become at risk in the same calendar year, meaning HSVC now only affects whether countries adopt a legislative gender quota for a small subset of years in the past (1995–2004) for nearly their whole sample. This conclusion differs from A&K's original findings, which suggested (1) a country having experienced HSVC always increased its chances of adopting a gender quota, relative to countries with no civil conflict, regardless of how long since the country could have first adopted a quota, and (2) this relative increase was of a lesser

[25] See Supplementary Appendix F for additional details about the original study.

[26] As of February 2023, the AC does not incorporate robust or clustered standard errors (SEs) into its computations (https://github.com/therneau/survival/issues/161). A&K's original analysis clusters its SEs on country. The approximated PH test with unclustered SEs identifies no PH violators, suggesting that the AC using unclustered SEs should generally work against identifying violators here.

[27] A&K's duration is miscoded for 10 countries. I use the corrected coding in the rerun; their main results remain at the 0.1 level. Otherwise, I take all of their research design decisions as is.

[28] See Supplementary Appendix F for full regression tables.

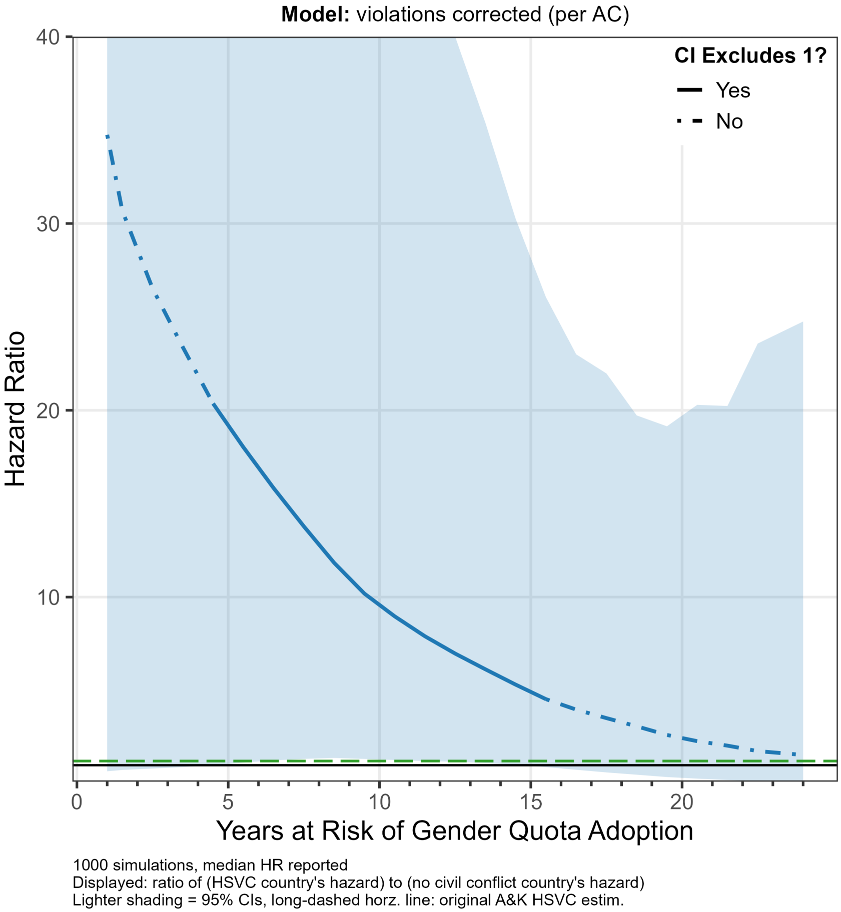


Figure 3. Effect of high sexual violence conflicts across time.

magnitude, evident by the vertical distance between HSVC's estimated HR from the PH-corrected model (Figure 3's solid line) and A&K's original estimated HR (Figure 3, long-dashed horizontal line).

We do not know whether HSVC is a true violator because the data's true DGP is unknown. However, three pieces of evidence suggest that HSVC may be a false positive, albeit not conclusively. First, there is a moderate correlation between HSVC and one of the control variables, "Conflict Intensity: High" (Corr = 0.516), which both the approximation and AC flag as a violator (Table 2). We know the AC is particularly prone to returning false positives in this situation. Second, HSVC's scaled Schoenfeld plot[29] shows no unambiguous trends, as we would expect to see for a PH violator. Finally, a series of martingale residual plots show no clear non-linear trends,[30] ruling out model misspecification from incorrect functional forms, which was Keele's (2010) and Metzger's (2023c) area of focus.

[29] I generate the (unreported) plot using the approximated PH test, as the scaled Schoenfelds use the variance–covariance matrix, and the AC does not currently acknowledge clustered SEs (fn. 26).

[30] Model's XB vs. the model's martingales, each covariate vs. the martingales from an auxiliary model omitting that covariate, and each covariate vs. the martingales from a null model.

https://doi.org/10.1017/pan.2023.34 Published online by Cambridge University Press

6. Conclusion

For Grambsch and Therneau's (1994) test for PH violations, does the way it is calculated affect the test's performance? My Monte Carlo simulations show that the answer is a resounding yes. More importantly, I show that the performance differences are non-trivial. I find that the AC has a high false positive rate in situations where a PH violator is correlated with a non-PH violator, even for correlations as moderate as 0.35. The approximation does not suffer from the same issue, meaning that it has a crucial advantage over the AC, given the importance we place on correctly sized statistical tests in traditional hypothesis testing. From Supplementary Appendix G's meta-analysis, we know moderate correlations are the norm among political science applications, underscoring the potential danger of the AC's behavior.

The biggest takeaway from these findings is that practitioners are currently stuck between a rock and a hard place. Both calculations perform adequately when covariates are uncorrelated with one another, but that condition is rarely true in social science applications. Purely on the basis of my simulation results, then, we should favor the approximation.

However, other factors preclude such an easy conclusion. One is a common limitation of any Monte Carlo study: the behavior I find for the approximation is limited in scope to the scenarios I investigated. It may be that, for other scenarios that vary different sets of characteristics, the approximation runs into performance issues similar to the AC. While this may certainly be true, the AC running into such serious performance issues for relatively simple, straightforward DGPs—while the approximation does not—is concerning and is sufficiently notable in its own right. These results also point to a number of related questions worth investigating. As one example, we might ask how the two calculations perform in a model with more than two covariates, and how the correlation patterns among those covariates might matter. The answers would be particularly relevant for applied practitioners.

A second factor is Therneau's main motivation for shifting `survival::cox.zph` from the approximated to actual calculation. His concern was the approximation's simplifying assumption being violated, which is particularly likely in the presence of strata (see fns. 1 and 5). In light of my results, though, violating the approximation's assumption may be the lesser of two evils, if the choice is between that or the AC's exceptionally poor performance for non-PH violators. Future research would need to investigate whether the trade-off would be worthwhile, and if so, under what conditions.

Finally, model misspecification is also a relevant factor. All the models I estimate here involve the correct base specification, with no omitted covariates or misspecified covariate functional forms. However, we know model misspecification can affect the PH test's performance, in theory (Keele 2010; Therneau and Grambsch 2000). Metzger (2023c) examines how both calculations perform in practice with uncorrelated covariates, in both in the presence and absence of model misspecification. She finds that the approximation can have a high false positive rate for some misspecified base models, going as high as 78.3% in one of her sets of supplemental results.[31] Knowing the approximation can suffer from the same performance issues as the AC means we cannot leverage my simulation results regarding the approximation's low false positive rate—the approximation returning evidence of a PH violation does *not* always mean a PH violation likely exists unless practitioners can guarantee no model misspecification exists, which is a potentially necessary, but likely insufficient, condition.

What might practitioners do in the meantime? The stopgap answers depend on the estimated Cox model's complexity, after addressing any model misspecification issues. If the Cox model has no strata and no strata-specific covariate effects, using the approximation is likely the safer bet. If the model has strata, but no strata-specific effects, practitioners can again use the approximation, but only after making the adjustments discussed in fn. 5. In the presence of both strata and strata-specific effects, there is no strong ex ante reason to suspect fn. 5's adjustments would not work, but it is a less-studied situation, traditionally. Future research could probe more deeply to ensure this is the case, especially as competing risks models can fall into this last category.

[31] Scenario 1, 0% RC, binary PH violator. She finds no false positive issues with misspecified base models for the AC in the scenarios she runs—unsurprising, given her simulations' covariates are uncorrelated.

https://doi.org/10.1017/pan.2023.34 Published online by Cambridge University Press

Social scientists' interest in a covariate's substantive effect makes it paramount to obtain accurate estimates of that effect. *Any* covariate violating the Cox model's PH assumption threatens that goal, if the violation is not corrected. I have shown here that successfully detecting PH violations is more fraught than we previously realized when using Grambsch and Therneau's full, actual calculation to test for these violations, rather than an approximation of it. I have suggested some short-term, stopgap solutions, but more research needs to be done to develop more nuanced recommendations and longer-term solutions for practitioners.

Supplementary Material. To view supplementary material for this article, please visit http://doi.org/10.1017/pan.2023.34.

Data Availability Statement. All analyses run using Stata 17.0 MP6 or R 4.1.0. This article's replication code has been published through Code Ocean and can be viewed interactively at https://doi.org/10.24433/CO.0072887.v1 (Metzger 2023a). A preservation copy of the same code and data can also be accessed via Dataverse at https://doi.org/10.7910/DVN/D56UWV (Metzger 2023b).

Acknowledgments. I thank Frederick Boehmke and Benjamin Jones for feedback on earlier drafts, as well as Harvey Palmer and Terry Therneau for general comments. Many of the simulations were run on the High Performance Computing Center at Michigan State University's Institute for Cyber-Enabled Research. Any mistakes are mine alone.

Funding Statement. This work received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing Interest. The author declares no competing interests exist.

References

Aberson, C. L. 2019. *Applied Power Analysis for the Behavioral Sciences*. 2nd ed. New York: Routledge.

Agerberg, M., and A.-K. Kreft. 2020. "Gendered Conflict, Gendered Outcomes: The Politicization of Sexual Violence and Quota Adoption." *Journal of Conflict Resolution* 64 (2–3): 290–317.

Austin, P. C. 2018. "Statistical Power to Detect Violation of the Proportional Hazards Assumption When Using the Cox Regression Model." *Journal of Statistical Computation and Simulation* 88 (3): 533–552.

Balan, T. A., and H. Putter. 2019. "Nonproportional Hazards and Unobserved Heterogeneity in Clustered Survival Data: When Can We Tell the Difference?" *Statistics in Medicine* 38 (18): 3405–3420.

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

Brilleman, S. L., R. Wolfe, M. Moreno-Betancur, and M. J. Crowther. 2021. "Simulating Survival Data Using the `simsurv` R Package." *Journal of Statistical Software* 97 (1): 1–27.

Cameron, A. C., and P. K. Trivedi. 2009. *Microeconometrics Using Stata*. 1st ed. College Station: Stata Press.

Demirtas, H., A. Amatya, and B. Doganay. 2014. "`BinNor`: An R Package for Concurrent Generation of Binary and Normal Data." *Communications in Statistics—Simulation and Computation* 43 (3): 569–579.

Gandrud, C. 2015. "`simPH`: An R Package for Illustrating Estimates from Cox Proportional Hazard Models Including for Interactive and Nonlinear Effects." *Journal of Statistical Software* 65 (3): 1–20.

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

Keele, L. 2008. *Semiparametric Regression for the Social Sciences*. New York: Wiley.

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

Metzger, S. K. 2023a. "Replication Data for 'Implementation Matters: Evaluating the Proportional Hazard Test's Performance.'" Code Ocean. https://doi.org/10.24433/CO.0072887.v1

Metzger, S. K. 2023b. "Replication Data for 'Implementation Matters: Evaluating the Proportional Hazard Test's Performance.'" Harvard Dataverse, V1. https://doi.org/10.7910/DVN/D56UWV

Metzger, S. K. 2023c. "Proportionally Less Difficult? Reevaluating Keele's 'Proportionally Difficult.'" *Political Analysis* 31 (1): 156–163.

Metzger, S. K., and B. T. Jones. 2016. "Surviving Phases: Introducing Multistate Survival Models." *Political Analysis* 24 (4): 457–477.

Metzger, S. K., and B. T. Jones. 2021. "Properly Calculating `estat phtest` in the Presence of Stratified Hazards." *Stata Journal* 21 (4): 1028–1033.

Morris, T. P., I. R. White, and M. J. Crowther. 2019. "Using Simulation Studies to Evaluate Statistical Methods." *Statistics in Medicine* 38 (11): 2074–2102.

https://doi.org/10.1017/pan.2023.34 Published online by Cambridge University Press

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

Park, S., and D. 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.

Therneau, T. M. 2021. "`cox.zph`: `zph.Rnw` Documentation." https://github.com/therneau/survival/blob/f2567b77252ac7935eba0ead364665c654ef28d3/noweb/zph.Rnw.

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

Cite this article: Metzger, S. K. (2024). Implementation Matters: Evaluating the Proportional Hazard Test's Performance. *Political Analysis*, 240–255. https://doi.org/10.1017/pan.2023.34