Supplemental Appendix For: *Addressing Monotone Likelihood in Duration Modeling of Political Events*

Noel Anderson
Department of Political Science
University of Toronto
noel.anderson@utoronto.ca

Benjamin E. Bagozzi
Deparment of Political Science and International Relations
University of Delaware
bagozzib@udel.edu

Ore Koren
Department of Political Science
Indiana University Bloomington
okoren@iu.edu

Contents

1 Overview **2**

2 Literature Survey **3**

3 Monte Carlo Simulations **8**

3.1 Model Performance with respect to $\mathbf{x_1}$ 10

3.2 Model Performance with respect to $\mathbf{x_2}$ and $\mathbf{x_3}$ 16

3.3 Model Performance with respect to $\mathbf{x_4}$ and $\mathbf{x_5}$ 21

3.4 Additional Model Comparison Metrics 24

3.5 Summary 35

Bibliography **36**

1 Overview

This supplemental appendix proceeds in several parts. In the section immediately below, we present a literature survey that reviews the past 10 years (2008-2018) of publications in the *American Political Science Review*, the *American Journal of Political Science*, and the *Journal of Politics* for the presence of monotone likelihood within published research. Following this, our supplemental appendix then comprehensively presents and discusses a series of Monte Carlo experiments that compare (i) a standard Cox model to (ii) a Firth Cox model under varying levels of censoring, numbers of observations, and degrees of omitted variable bias.

2 Literature Survey

In this section, we seek to evaluate the extent to which monotone likelihood confronts political scientists employing duration models in their empirical research. To do so, we conducted a survey of all articles that employed a duration model between the years 2008-2018 and that appeared in one of three leading political science journals: the *American Political Science Review*, the *American Journal of Political Science*, and the *Journal of Politics*. Articles were identified by searching Google Scholar for the following terms: "competing risks," "cox model," "cox ph," "cure model," "duration model," "event history," "exponential model," "hazard model," "proportional hazards," "survival model," and "weibull model." These searches were implemented separately for each of our three target journals. We then retained only those articles that presented (or discussed the estimation of) a duration model either in their main article or supplemental appendix for the literature survey presented below.

All duration model articles identified via the steps described above are listed in Tables 1-4 below. A total of 57 articles were identified, and these articles were then double-coded by two separate coders (both are authors on the accompanying paper) for the presence of monotone likelihood. Presence was recorded if any of the following conditions were identified: (i) evidence that monotone likelihood affected parameter estimates of at least one independent or control variable, and for at least one reported or discussed duration model, within an article's main paper or supplemental appendix; (ii) evidence that monotone likelihood required authors to drop offending variables; (iii) evidence that monotone likelihood affected model specification; and/or (iv) evidence that monotone likelihood undermined substantive interpretation of results.

When possible, replication data was examined to verify the presence of monotone likelihood. However, because many articles did not have replication data available, our judgments concerning monotone likelihood often had to be made on the basis of the reported parameter estimates within each article's tables. In such instances, careful consideration was given to (i) the size of the reported estimated effect; (ii) the units of measurement and range of the independent or control variable under consideration;[1] (iii) the units of measurement and range of any interacted variables with the variable of interest;[2] and (iv) additional correlates of monotone likelihood, as reported by the original authors (e.g., analyses reporting particularly low Ns, large numbers of dichotomous predictors, extreme imbalance in censored cases, etc.). Where our conclusions based upon these criteria were uncertain, we erred towards coding an article's potential for containing monotone likelihood as "No."[3]

Given the above approach, our final codings of the articles listed below provide a conserva-

[1] We treated unconditional fixed effects as control variables within our assessments of monotone likelihood given their association with separation issues in similar contexts (Cook, Hays and Franzese, 2018).

[2] We erred towards coding an article as "No" (as in, "No evidence of monotone likelihood") when a potential monotone likelihood issue involved a multiplicative interaction, unless their was clear evidence to suggest that (near)infinite estimates arose in the plausible range of the combined interaction effect.

[3] For example, we record Schleiter and Morgan-Jones (2009), Smith and Fridkin (2008), Berlinski, Dewan and Dowding (2010), and Gibler and Tir (2010) as "No" (as in, "No evidence of monotone likelihood") in our monotone likelihood evaluations due to insufficient corroborating evidence, although one unit changes in binary variables within each article exhibit estimated hazard ratios that range from 22-115, implying increases in the hazard of termination ranging from 2,100-11,400%.

tive best guess—based upon two expert coders—of the extent to which monotone likelihood confronts political scientists employing duration analysis. However, we stress two points. First, these cases sometimes rely only upon the parameter estimates reported in article tables, and hence are not definitive. Accordingly, we label these codings as "likely ML" cases, and provide a brief descriptive summary of our rationales for coding each positively coded monotone likelihood instance, alongside our final codings, in Tables 1-4. Second, our codings reflect the problem of monotone likelihood not only in the narrow sense of its effect on reported parameter estimates, but as a broader methodological challenge that has to date often required researchers to employ some of the distasteful remedies we discuss in the accompanying article. Put differently, our survey seeks to assess the extent to which researchers have been forced to grapple with the challenge monotone likelihood poses to empirical research, not whether their primary reported models or parameter estimates are "wrong."

With these caveats in mind, we find that out of the 57 duration model articles appearing in the *American Political Science Review*, the *American Journal of Political Science*, and the *Journal of Politics* from 2008-2018, 8—or 14%—contained strong evidence of monotone likelihood. This represents a substantial share of contemporary duration model research across three of the political science discipline's top journals. What is more, we found that the problem of monotone likelihood has confronted researchers across the empirical subfields of political science—from American politics to comparative politics to international relations—and affects cross-national studies and subnational analyses alike. And importantly, this estimate is likely a conservative one in light of the "file drawer problem:" where researchers simply do not publish or report models that seem otherwise incorrect or hard to understand.

In short, our survey suggests that monotone likelihood regularly confronts researchers employing duration models in their empirical research. This is an important methodological challenge of significance for the political science discipline, broadly conceived.

Table 1: *American Political Science Review*: 2008-2018

Article	**Likely ML**	**Rationale**
Berliner and Erlich (2015)	No	
Cunningham (2011)	No	
Debs and Goemans (2010)	No	
Fortunato and Loftis (2018)	No	
Hollyer, Rosendorff and Vreeland (2015)	No	
Huber and Martinez-Gallardo (2008)	No	
Kokkonen and Sundell (2014)	No	
Lyall (2010)	No	
Schleiter and Morgan-Jones (2009)	No	
Smith and Fridkin (2008)	No	
Svolik (2008)	Yes	The article estimates a series of (split population and standard) log-logistic and Weibull duration models of democratic survival. Exponentiating the log(time) coefficient estimates for the article's binary `monarchy (vs. not independent)` control variable yields expected changes in survival time for countries that were monarchies prior to democratic transitions of between 8.611475e-07 ↔ 1.759916e-09. This implies an almost 100% decrease in survival time, relative to countries that were not independent prior to a democratic transition.

Table 2: *American Journal of Political Science*: 2008-2018

Article	**Likely ML**	**Rationale**
Arias, Hollyer and Rosendorff (2018)	No	
Beardsley (2008)	No	
Berry, Burden and Howell (2010)	No	
Bueno de Mesquita and Smith (2010)	No	
Carpenter et al. (2012)	No	
Darmofal (2009)	No	
Fukumoto (2009)	No	
Gibler and Tir (2010)	No	
Kelley and Simmons (2015)	No	
Knutsen and Nygård (2015)	No	
Laver and Benoit (2015)	Yes	The paper estimates a series of Cox models to study post-electoral cabinet durations in European minority settings. Model 3 in Table 8 incorporates a large number of binary indicator variables for each country in the dataset, including some that are not associated with any termination events. For example, there are no cases where `denmark` = 1 and `termination` = 1. Consequently, this indicator in monotonic when ordered according to failure time, which accounts for its corresponding hazard ratio of 0.0000903.
Leventoğlu and Metternich (2018)	No	
Maltzman and Shipan (2008)	No	
Mattes and Savun (2010)	No	
Ostrander (2016)	No	
Park and Hendry (2015)	No	
Thrower (2017)	No	
Treier and Jackman (2008)	No	

Table 3: *Journal of Politics*: 2008-2018, Part I

Article	**Likely ML**	**Rationale**
Acosta (2014)	No	
Ahlquist (2010)	No	
Baccini and Urpelainen (2013)	No	
Berlinski, Dewan and Dowding (2010)	No	
Boix and Svolik (2013)	No	
Boudreau and MacKenzie (2018)	No	
Camerlo and nán (2015)	No	
Capoccia, Sáez and de Rooij (2012)	Yes	The paper examines the resilience of regional political actors' territorial demands in India. The authors confront monotone likelihood when interacting some covariates (in particular, the `state religion` indicator variables) with $log(t)$, as there are too few cases of termination on the "other minority" group (only 1 case). This results in monotone likelihood when attempting to estimate the model, necessitating that the authors drop the `state religion`$*log(t)$ interaction in their models reported in Table 2 and Figures 1-2.
Curry and Hurwitz (2016)	No	
Davis and Wilf (2017)	No	
Findley and Young (2015)	No	
Flores and Nooruddin (2012)	Yes	The article estimates log-normal survival models for time until (i) economic recovery and (ii) crisis recurrence in post-conflict environments. In the reported Full Model version for economic recovery (Table 2), the combined effect of the `second year*new democracy` interaction when both variables are equal to one implies a 61,400% increase in expected survival time. The article's publicly available replication data and .do file for this specification do not replicate perfectly, with a slightly smaller N than that reported in the main paper. But upon examining the dataset, we find that there are no cases where termination = 1 and `second year*new democracy` = 1. Consequently, this variable is monotonic when ordered according to failure time, resulting in parameter estimates that converge towards positive infinity.
Gauri, Staton and Cullell (2015)	Yes	The article estimates a Cox model of compliance with Costa Rican Supreme Court direct orders in amparo and habeas corpus cases in Table 1. The final model includes fixed effects for 202 distinct public agency defendants. The use of fixed effects inadvertently introduces convergence issues, as a number of the institutional indicator variables that are thereby included in the model are quasi-completely separating. For example, there are no termination events associated with `institution codes 16, 155, 193`, and `200`; consequently, these variable are monotonic when ordered according to failure time (in this case, never varying from 0 when termination = 1.) This results in parameter estimates that converge toward negative infinity, with hazard ratios of 3.52e-13, 3.38e-13, 3.52e-13, and 5.47e-13, respectively.
Golder, Golder and Siegel (2012)	No	
Hassell (2015)	No	
Heersink (2018)	No	

Table 4: *Journal of Politics*: 2008-2018, Part II

Article	**Likely ML**	**Rationale**
Johns and Pelc (2018)	Yes	The article estimates a series of Cox models of challenges to WTO violations. On page 885, the authors note, “We also ensure that the results are robust to modifications similar to those we made to our duration analysis: we add controls for the number of complainants in the dispute and then exclude all multiple complainants disputes altogether.” The former models are not reported in the supplemental appendix, but are included in the article’s publicly available replication files. When replicated, two of these three models report hazard ratios of 5.464326e-08 and 4.149118e-09, and nonrecoverable standard errors for `number of complainants`. The linear combination of this variable and the `legal measure` fixed effects appears to be inadvertently introducing monotone likelihood. In addition, there are no termination events associated with `topics` 3, 5, 9, or 11 (after missing observations are dropped via listwise deletion); consequently, these controls are monotonic when ordered according to failure time (i.e. never varying from 0 when termination = 1).
Kittilson (2008)	No	
Koch and Sullivan (2010)	Yes	The article estimates Cox models of the duration until a major power withdraws from an intervention. In Table 1-2, these models include interactions between `right party executive` and either `executive approval` or `incumbent party support`. Beyond a Table 1 hazard ratio for the `right party executive` interaction component term of 3.8e-08, Figures 1-2 imply marginal effects or 95% CIs whose hazard rates are effectively zero across supported ranges of these interacted variables (e.g., 4.5e-06).
Maeda (2010)	No	
Malesky (2009)	No	
Narang (2014)	Yes	The article estimates Cox models of the duration of peace after civil war. The hazard ratio for `decisive victory` is very small across all models in Table 2 (ex. 0.000593 in Model 1). This variable is interacted with `logged total humanitarian aid` (whose range, mean, and standard deviation are 0 ↔ 9.06, 2.75 and 3.46, respectively). Hence, zero values of `logged total humanitarian aid` are plausible, and the effect of the individual `decisive victory` coefficient can be interpreted directly for cases where `logged total humanitarian aid` = 0. Herein, `decisive victory` = 1 is associated with an over 99.9% decrease in the hazard of peace failing relative to cases where `decisive victory` = 0 when humanitarian aid is absent.
Owsiak and Rider (2013)	No	
Potter (2017)	No	
Reenock, Staton and Radean (2013)	No	
Scherer, Bartels and Steigerwalt (2008)	No	
Wallace (2013)	No	
Wolford (2017)	No	

3 Monte Carlo Simulations

This section fully presents the results from our Monte Carlo experiments. Recall that these experiments assess the performances of (i) a standard Cox proportional hazards (PH) model and (ii) a Cox PH model with Firth's penalized maximum likelihood correction for five varying levels of censoring, at six different sample sizes, and across three different specification scenarios. Specifically, we simulate duration data to match the (nonparametric) Cox model's assumed data generating process (dgp) via the simulation methods and software developed by Harden and Kropko (2018). For each simulated duration dependent variable Y, we draw our explanatory variables $\mathbf{x}$ from $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \mathbf{x_4}, \mathbf{x_5})'$ where $\mathbf{x_1} = 1_{\mathcal{N}(0,1)>1.75}$, $\mathbf{x_2} = 1_{\mathcal{N}(0,1)>1.75}$, $\mathbf{x_3} = \ln(\mathcal{N}(5,1))$, $\mathbf{x_4} = 1_{Uniform[-2.5,12]>11.5}$, and $\mathbf{x_5} = 1_{\ln Uniform[1,100]>1.65}$. Accordingly, this set of explanatory variables satisfies the *root cause* of monotone likelihood: the presence of unbalanced binary covariates. In addition, this set of explanatory variables also ensures that our simulations closely approximate the *types* of variable specifications that commonly arise in political science analyses of duration data, which typically include more than three independent variables that are themselves a mixture of (multiple, imbalanced) binary predictors and (logged) continuous predictors. Our Monte Carlo simulations then independently vary the following three conditions:

1. The level of censoring (c)—i.e., the proportion of all duration cases that exhibit non-terminations within our period of observation—which we set to range across $c = \{0.05, 0.25, 0.5, 0.75, 0.95\}$.

2. The number of observations N, which we assign as $N = \{100, 250, 500, 1000, 2000, 5000\}$.

3. The degree of omitted variable bias, where we consistently generate our duration outcome with $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \mathbf{x_4}, \mathbf{x_5})'$, but then separately estimate models based upon one of the three following specifications:

    - $Y = \beta_1\mathbf{x_1}$ (i.e., four omitted variables)
    - $Y = \beta_1\mathbf{x_1} + \beta_2\mathbf{x_2} + \beta_3\mathbf{x_3}$ (i.e., two omitted variables)
    - $Y = \beta_1\mathbf{x_1} + \beta_2\mathbf{x_2} + \beta_3\mathbf{x_3} + \beta_4\mathbf{x_4} + \beta_5\mathbf{x_5}$ (i.e., no omitted variables)

In these manners, our Monte Carlo exercises do not directly modify the severity of monotone likelihood, but rather allow the severity of this problem to arise organically as a combined function of the levels of imbalance in our independent variables, our sample sizes, the levels of censoring considered, and the different rates of omitted variables evaluated.[4] Furthermore, given the N's considered here, we note that our Monte Carlo experiments expand upon the work of Heinze and Schemper (2001), who earlier demonstrated the severity of the monotone likelihood problem within duration analyses of samples encompassing

[4] Because we provide our own (fixed) binary predictors to Harden and Kropko's Cox simulation routines, note that our chosen censoring levels are generated for our survival outcomes by these simulation routines' adjustment of our actual parameter values for each condition. Where we evaluate (in)accuracy in our specific parameter values below, we accordingly use each simulation run's (unique) true parameter values as a baseline for evaluating that simulation's corresponding coefficient estimates.

$N = \{50, 100, 200\}$. We believe that the larger N's that we evaluate below are of particular interest to political scientists, given that duration analyses within political science frequently analyze datasets with $N > 200$.

For each of the $5 \times 6 \times 3 = 90$ different combinations of varying conditions outlined above, we generate 1,000 simulated datasets.[5] For each simulated dataset, we then estimate a standard Cox model and a Cox model with Firth's penalized maximum likelihood estimation approach, both in R, and store our relevant coefficient estimates' root mean squared errors (RMSEs) and nonconvergence rates.[6] We base the latter quantity on instances where a particular model provided an error or warning message either globally or for a given covariate,[7] *or* where a model's coefficient estimate or its standard error was at least 500 times larger (or smaller) than the true value for a given parameter under consideration. Our nonconvergence findings and conclusions are robust to alternate nonconvergence thresholds (e.g., 50 times larger or smaller) and/or to the removal of this threshold criteria altogether. We illustrate the latter point in Figures 10-12 further below, which re-produce our main nonconvergence figures whilst defining nonconvergence based upon instances where a particular model provided an error or warning message either globally or for each chosen covariate.

For the quantities discussed below and in the main paper, we primarily evaluate our nonconvergence criteria in relation to the coefficient estimates and standard errors for $\mathbf{x_1}$, since this was the only explanatory variable that was included in all 90 different experimental conditions under consideration. However, our conclusions with respect to nonconvergence remain unchanged when we alternatively define nonconvergence based upon our parameter estimates and standard errors for $(\mathbf{x_2}, \mathbf{x_3}, \mathbf{x_4}, \mathbf{x_5})'$. We demonstrate the latter point further below in our discussions of the nonconvergence findings for these additional independent variables.

The RMSE measures that we report and discuss below allow us to directly assess the overall accuracy of our coefficient estimates, wherein—by virtue of the RMSE being the square root of the mean squared error (MSE)—it reflects both the bias and variance of each estimate considered here. As a secondary measure of bias and variance, we also directly examined the MSEs for all Monte Carlo comparisons and coefficient estimates presented below. However, given the (near) infinite estimates obtained under monotone likelihood conditions in the standard Cox model context, the Cox model's MSEs were typically many orders of magnitude greater (i.e., worse) than that of the Firth Cox model. This rendered graphical comparisons of Cox MSEs and Firth Cox MSEs less than informative, relative to our RMSE comparisons, leading us to instead favor RMSEs within our primary Monte Carlo assessments. Nevertheless, we report comparable MSE plots (i.e., to our primary

[5] This is consistent with Heinze and Schemper (2001).

[6] We estimate the Cox model using the survival package in R (Therneau, 2015) and estimating the Firth Cox model via the coxphf package (Heinze and Ploner, 2018). During estimation, we conservatively set each model's iteration limit to 500. Our findings and conclusions are comparable when we instead use the default iteration limits for the Cox model and Firth Cox models that we estimate in R (of 20 and 50, respectively), as well as when we alternatively increase each model's iteration limit to 1,000 (or 10,000). For both models, any instances of tied events are handled via Breslow's approximation (Breslow, 1974).

[7] E.g., "simpleWarning in fitter(X, Y, strats, offset, init, control, weights = weights,) : Loglik converged before variable 1 ; beta may be infinite."

RMSE plots) towards the end of this Supplemental Appendix section (Figures 7-9). These Figures demonstrate that our RMSE-based conclusions with respect to accuracy—and the preferability of the Firth Cox model over the standard Cox model—remain unchanged when one considers MSEs in place of RMSEs.

3.1 Model Performance with respect to $\mathbf{x_1}$

We begin by discussing the nonconvergence rates, and RMSE results, obtained for $\mathbf{x_1}$. The $\mathbf{x_1}$-based nonconvergence rates for our Cox and Firth Cox models appear in Figure 1. In this Figure, each column of subfigures depicts a differing level of censoring, each row of subfigures depicts a different level of omitted variable bias, and the x-axis within each subfigure illustrates our nonconvergence results across each N evaluated. The y-axes in Figure 1 then depict the proportion of all simulations that exhibited nonconvergence for a given level of omitted variable bias (row) and level of censoring (column). Turning to Figure 1, one can first observe that when censoring is low-to-moderate (i.e., 5%↔50%), the Cox and Firth Cox models each generally exhibit low nonconvergence—even in the presence of multiple imbalanced binary predictors (i.e,. the overarching conditions of our Monte Carlo simulations). That being said, for low-to-moderate levels of censoring, the Cox model at times exhibits noticeably higher (and hence worse) levels of nonconvergence in comparison to the Firth Cox model. This is most observable in Figure 1 under conditions where $N \leq 500$ and/or under conditions of 50% or 5% censoring.

Under moderate-to-high (i.e., 75% ↔ 95%) levels of censoring in Figure 1, the performance of the Cox and Firth Cox models begin to diverge more markedly. At levels of 75% censoring and $N = 100$, estimating a Cox model instead of a Firth Cox model leads to model nonconvergence (for $\mathbf{x_1}$ or globally) within approximately 35% of all simulations. These levels of nonconvergence are dramatically higher than those obtained for the Firth Cox model (of 1.5%-3.8%) under comparable conditions. Increasing N to 250 at this level of censoring (i.e., 75%) continues to yield substantially higher rates of nonconvergence (4.9%-5.0%) for the Cox model, relative to the Firth Cox model (1.1%-1.5%). Only when $N = 500$ do the nonconvergence rates observed under the 75% censoring condition appear comparable among the Cox and Firth Cox models within the visual plots. However, even in these cases, the Firth Cox model exhibits preferable or equivalent nonconvergence rates to that of the Cox model across all N's considered.

Turning to the 95% censoring cases in Figure 1, we find that monotone likelihood issues become so substantial that the standard Cox model becomes essentially unusable with low-to-moderate N's. For example, at $c = 95\%$ and $N = 100$, the Cox model fails to converge across approximately 90% of all simulations, whereas the Firth Cox model exhibits nonconvergence in only 6.5%-20.6% of these cases. Likewise, when $c = 95\%$ and $N = 250$, we observe nonconvergence rates for the Cox model of 52.9%-53.3%, whereas the Firth Cox model exhibits far more conservative nonconvergence rates (of 1.9%-2.0%). At $c = 95\%$ and $N = \{500, 1000\}$, our findings are similar but at times slightly less pronounced, with 17.1%-54.4% of all simulations failing to converge in the standard Cox case in comparison to nonconvergence rates of 1.1%-2.0% in the Firth Cox case. Even at $N = 2,000$, we continue to find consistently higher levels of nonconvergence in the Cox case (1.9%-2.1%) in comparison to the Firth Cox case (1.1-1.2%) when $c = 95\%$. Based upon these results, we can conclude that one cannot reliably trust the Cox model (i.e., with equal confidence to the Firth Cox model) under conditions of 95% censoring until one's N is larger than 2,000.

What effect does the exclusion or inclusion of additional relevant predictors have on the nonconvergence results discussed above? If we compare the subfigure plots across the rows presented in Figure 1, we find that the Cox and Firth Cox models' relative deficiencies with respect to nonconvergence slightly worsen as one correctly includes additional covariates related to Y in the model. For example, when censoring is equal to 95% and $N = 100$, adding four additional relevant predictors (row 3) to a model that previously only included a single relevant predictor (row 1) increases the nonconvergence rates for $\mathbf{x_1}$ under the Cox model from 89.1% to 89.5%, whereas the nonconvergence rates for the Firth Cox model increase from 6.5%→21.5%). To take another example, for the Monte Carlo simulations employing 95% censoring and $N = 250$, we likewise find that including all relevant predictors (row 3) as opposed to a single relevant predictor (row 1) increases the Cox model's nonconvergence rates for $\mathbf{x_1}$ from 52.9%→53.3%. Under these same conditions, the nonconvergence rates for the Firth Cox model similarly increase (1.9%→2.0%) while still remaining far below the Cox model's nonconvergence rates.

In sum, Figure 1 suggests that a researcher will often encounter nonconvergence when using a Cox model with imbalanced predictors for samples of 100↔500 observations, no matter the level of censoring. For levels of censoring at or greater than 75% of all observations, these threats of nonconvergence increase substantially, and can extend to samples as large as 1000↔2000 observations. Moreover, if one's true dgp includes more than one imbalanced predictor, each of the Cox-based nonconvergence issues mentioned above will paradoxically become *more* acute as one's Cox model properly includes more of these true predictors, further complicating estimation. The Firth Cox model is much less sensitive to these monotone likelihood-induced nonconvergence-issues. Thus, while the Firth Cox model does exhibit some nonconvergence issues of its own,[8] it nevertheless remains a superior choice over the Cox model, especially for samples lower than 1,000 and/or for duration outcomes that exhibit censoring rates of 75% or higher.

[8] Which in these cases are primarily attributable to our treatment of any parameter estimate that is greater than 500 times its true value to be a case of "nonconvergence."

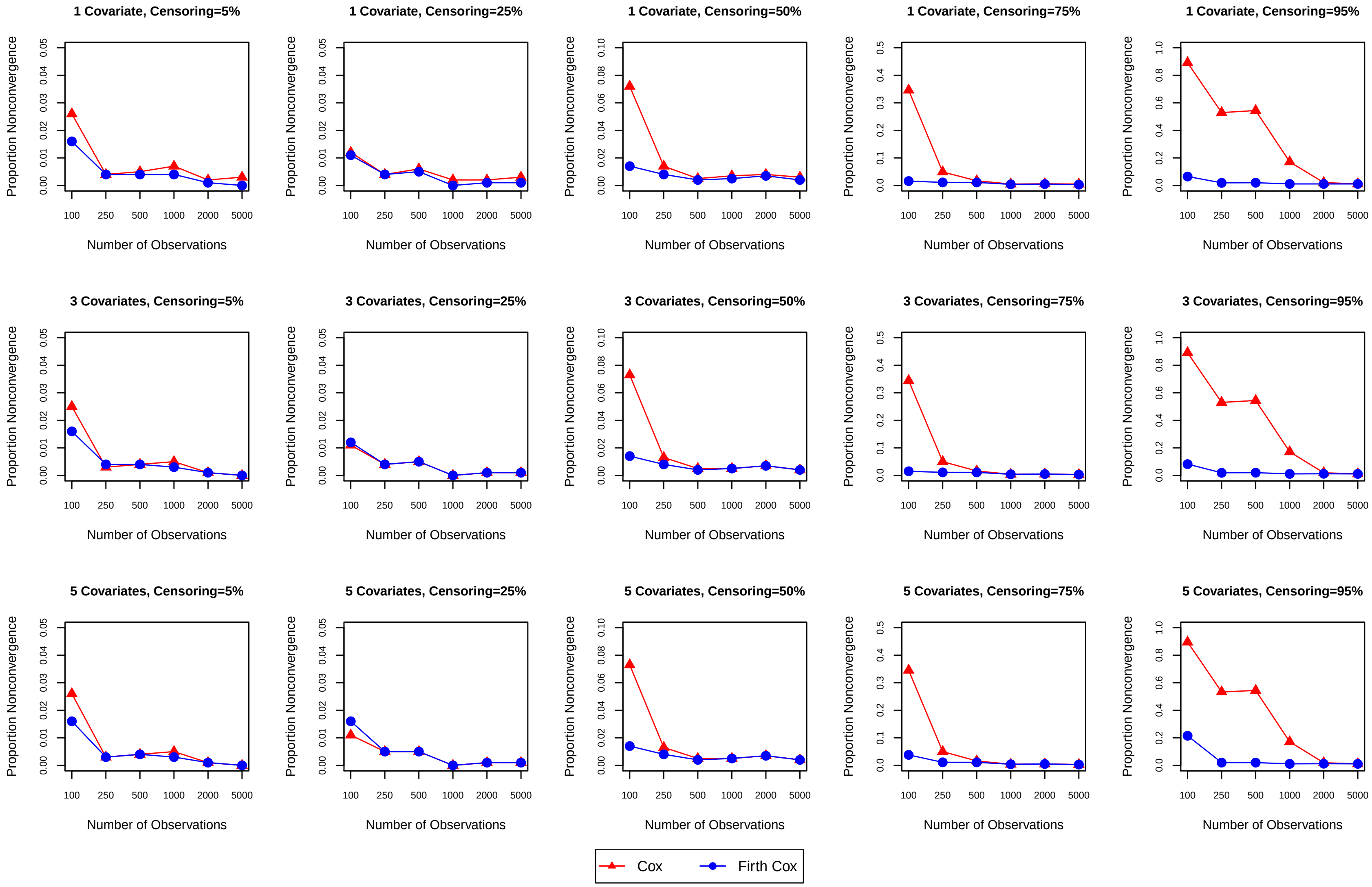

Figure 1: Nonconvergence Levels for $\mathbf{x_1}$ Across All Monte Carlo Experiments

We now turn to the RMSE statistics for $\hat{\beta}_1$ (Figure 2). As above, we present all experimental conditions in this Figure, wherein the columns indicate changes in the level of censoring considered, the rows present changes in the number of explanatory variables included within each estimated duration model, and the values along each x-axis denote the different numbers of observations considered. The y-axes to these plots then present the $\hat{\beta}_1$ RMSE for each model, averaged across all relevant simulations. Starting with columns 1 and 2 of Figure 2, we generally observe no discernable differences in RMSEs between the Cox and Firth Cox estimators under conditions of 5%-25% censoring, though the Firth Cox estimator does consistently have slightly higher levels of accuracy—based upon lower RMSEs—relative to the Cox model in each and every instance. The one exception to the above statement is the case of $N = 100$, wherein the Firth Cox's RMSEs more noticeably outperform those of the Cox model across columns 1-2 of Figure 2. With regards to overall trends in RMSEs across our N's, we can note that accuracy in $\hat{\beta}_1$—as implied by a lower RMSE value—improves for both the Cox model and the Firth Cox model as one's sample size increases from 100 to 5000 observations at these low levels of censoring (i.e., $c = \{5\%, 25\%\}$). We observe similar overall trends under conditions of 50% censoring (column 3) in Figure 2, though here we also find that the Firth Cox model much more markedly outperforms the Cox model—in terms of lower RMSEs—when $N = \{100, 250\}$.

When censoring is moderate-to-high (i.e., $c = \{75\%, 95\%\}$) the consequences of monotone likelihood in Figure 2 become far more pronounced. To this end, the fourth column in Figure 2 clearly demonstrates that the standard Cox model's estimates of β_1 exhibit substantially poorer accuracy than those of the Firth Cox model when $c = 75\%$ and $N = \{100, 250\}$, and exhibit moderately worse accuracy than those of the Firth Cox model when $c = 75\%$ and $N = 500$. For $c = 75\%$, these differences in RMSEs—and hence accuracy—then tend to subside as $N \to 2000$. Turning to the final column of Figure 2, we can observe that when censoring is high (i.e., 95%), inaccuracy in $\hat{\beta}_1$ under the standard Cox model is far more pronounced than was observed in the $c = 75\%$ case, and is now often 5-to-10 times that of the Firth Cox model. The most acute divergences in this regard again arise when N is less than 1,000, though we can note that even with an N of 1,000, the Cox model's RMSEs are still over 5 times larger than those of the Firth Cox model. It is not until we increase N beyond 1,000 that these substantial divergences in RMSEs subside. However, even in the cases of $N = \{2000, 5000\}$, the Firth Cox model's RMSEs for $\hat{\beta}_1$ are consistently 5%-to-37% smaller than those of the standard Cox model.

As one moves down the subfigure rows of Figure 2, we find that the RMSE trends discussed above for $\mathbf{x_1}$ are amplified for the standard Cox model when one includes additional appropriate independent variables within one's duration model specification. That is, adding two, or four, additional covariates—several of which have imbalance themselves—noticeably increases the RMSEs obtained for $\hat{\beta}_1$ within the standard Cox model. This is consistent with the nonconvergence results discussed above, and is less observable for the Firth Cox model in this case. Returning to the standard Cox model results, we can also note that the added inaccuracy in $\hat{\beta}_1$ that arises from increasing the number of included covariates is less substantial than the changes in RMSEs that arise due to the differential values of c and N evaluated in Figure 2.

Altogether, the nonconvergence and RMSE results for $\mathbf{x_1}$ suggest several key points. Firstly, no matter the level of censoring or number of observations, the Firth Cox model always outperforms the Cox model with respect to RMSEs for $\hat{\beta}_1$. This implies that—with respect to accuracy—the Firth Cox model is preferable to the non-Firth Cox model across all sample sizes considered here, and no matter the actual degree of monotone likelihood observed. This conclusion is in line with recent findings for Firth's penalized maximum likelihood estimator in the context of logit models (Rainey and McKaskey, forthcoming), and with past simulation analyses of the Firth Cox model's performance (e.g., Heinze and Schemper, 2001). Secondly, the results discussed above also reveal that monotone likelihood issues are most severe within low sample sizes *and* high censoring, wherein the joint contribution of these two issues to nonconvergence and inaccuracy within the standard Cox model arises most markedly in situations where one's $N \leq 1000$ and censoring is moderate-to-high (e.g., 75-95%). Furthermore, when these estimation issues do arise, adding *additional* (imbalanced) control variables can exacerbate monotone likelihood issues for a given variable of interest.

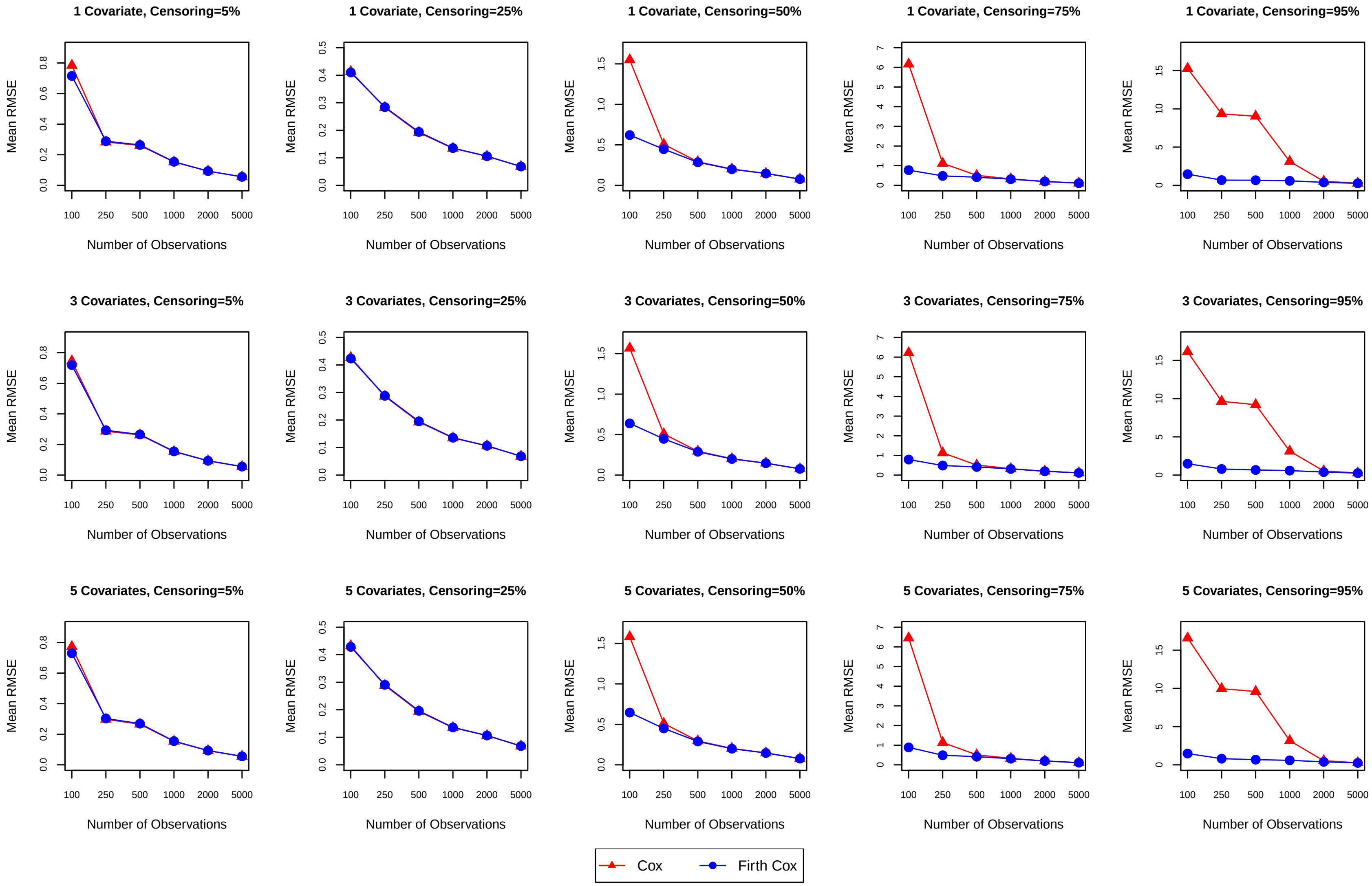

Figure 2: Root Mean Squared Errors for $\mathbf{x_1}$ Across All Monte Carlo Experiments

3.2 Model Performance with respect to $\mathbf{x_2}$ and $\mathbf{x_3}$

We now discuss our Monte Carlo results for $\mathbf{x_2}$ and $\mathbf{x_3}$. Similar to the assessments of $\mathbf{x_1}$ presented above, we first focus these $\mathbf{x_2}$ and $\mathbf{x_3}$ discussions on these two variables' identified rates of nonconvergence (as a share of all relevant simulations), and then turn to an assessment of the RMSEs obtained for these two covariates' respective coefficient estimates. Recall that $\mathbf{x_2}$ and $\mathbf{x_3}$ were operationalized as $\mathbf{x_2} = 1_{\mathcal{N}(0,1)>1.75}$ and $\mathbf{x_3} = \ln(\mathcal{N}(5,1))$ within our Monte Carlo simulations and were only included within $5 \times 6 \times 2 = 60$ of our experimental conditions.[9] We plot these experimental conditions for both covariates of interest in Figure 3 (nonconvergence) and Figure 4 (RMSEs). As before, we present all relevant experimental conditions in these figures. Our different censoring conditions appear across the columns of our subfigures in Figure 4. The relevant specification choices appear along different subfigure rows, now separately for $\mathbf{x_2}$ (rows 1-2) and $\mathbf{x_3}$ (rows 3-4). The x-axis of each subplot then depicts the different numbers of observations that our Monte Carlo experiments evaluated.

As mentioned, the nonconvergence rates for $\mathbf{x_2}$ and $\mathbf{x_3}$ are depicted in Figure 3. With regards to $\mathbf{x_2}$, we find similar patterns of nonconvergence to those noted for $\mathbf{x_1}$. When censoring is low-to-moderate (i.e., 5%↔50%), the Cox and Firth Cox models frequently exhibit comparable levels of nonconvergence. Exceptions to this pattern arise within the $N = \{100, 250\}$ conditions, wherein the standard Cox model exhibits markedly higher rates of nonconvergence in most instances, and especially for $c = \{5\%, 50\%\}$. For moderate-to-high (i.e., 75% ↔ 95%) levels of censoring, and in particular for cases where $N < 1000$, monotone likelihood issues lead to substantial nonconvergence issues for the Cox model, but not for the Firth Cox model. For instance, we find that the Cox model exhibits 8.8%-82.5% nonconvergence when $N = \{100, 250\}$ and $c = \{75\%, 95\%\}$; and 0.9%-32.7% nonconvergence when $N = 500$ and $c = \{75\%, 95\%\}$. By comparison, the Firth Cox model exhibits far lower levels of nonconvergence in each of these circumstances (1.5%-20.7% and 0.7%-1.9%, respectively).

Under these same conditions of 75%-95% censoring, we find that after one's N rises above $N = 1000$, the problems of monotone likelihood largely subside. However, the Cox model continues to exhibit up to double the rate of nonconvergence in comparison to the Firth Cox model in these contexts. For example, at $N = 2000$ and $c = 95\%$, we observe nonconvergence rates of 2.8% for the Cox model, in comparison to nonconvergence rates of 1.4% for the Firth Cox model. The Cox and Firth Cox model's nonconvergence issues for $\mathbf{x_2}$ typically become slightly more substantial when $\mathbf{x_4}$ and $\mathbf{x_5}$ are added to our model specifications, which is consistent with the findings discussed above.

The Cox model also frequently performs worse than the Firth Cox model in nonconvergence for $\mathbf{x_3}$ in Figure 3. However, in this case, both models generally exhibit low levels of nonconvergence across the board. As a result, we can conclude here that the differences in nonconvergence between the Cox model and Firth Cox model for this particular variable are fairly slight and often negligible, no matter the N, specification, or censoring rate considered.[10] This is to be expected given that—unlike $\mathbf{x_1}$ and $\mathbf{x_2}$—$\mathbf{x_3}$ is a continuous predictor

[9] That is, these two independent variables were only included within the Monte Carlo conditions that included three or five independent variables within our estimated Cox and Firth Cox model specifications.

[10] The one exception may be cases of $N = 100$. However, even in these cases, differences in nonconvergence

rather than a(n imbalanced) binary predictor. This finding, in turn, underscores that monotone likelihood based nonconvergence issues are most directly associated with (imbalanced) binary covariates, and are unlikely to arise in contexts where one's covariates are continuous.

do not grow larger than 1%.

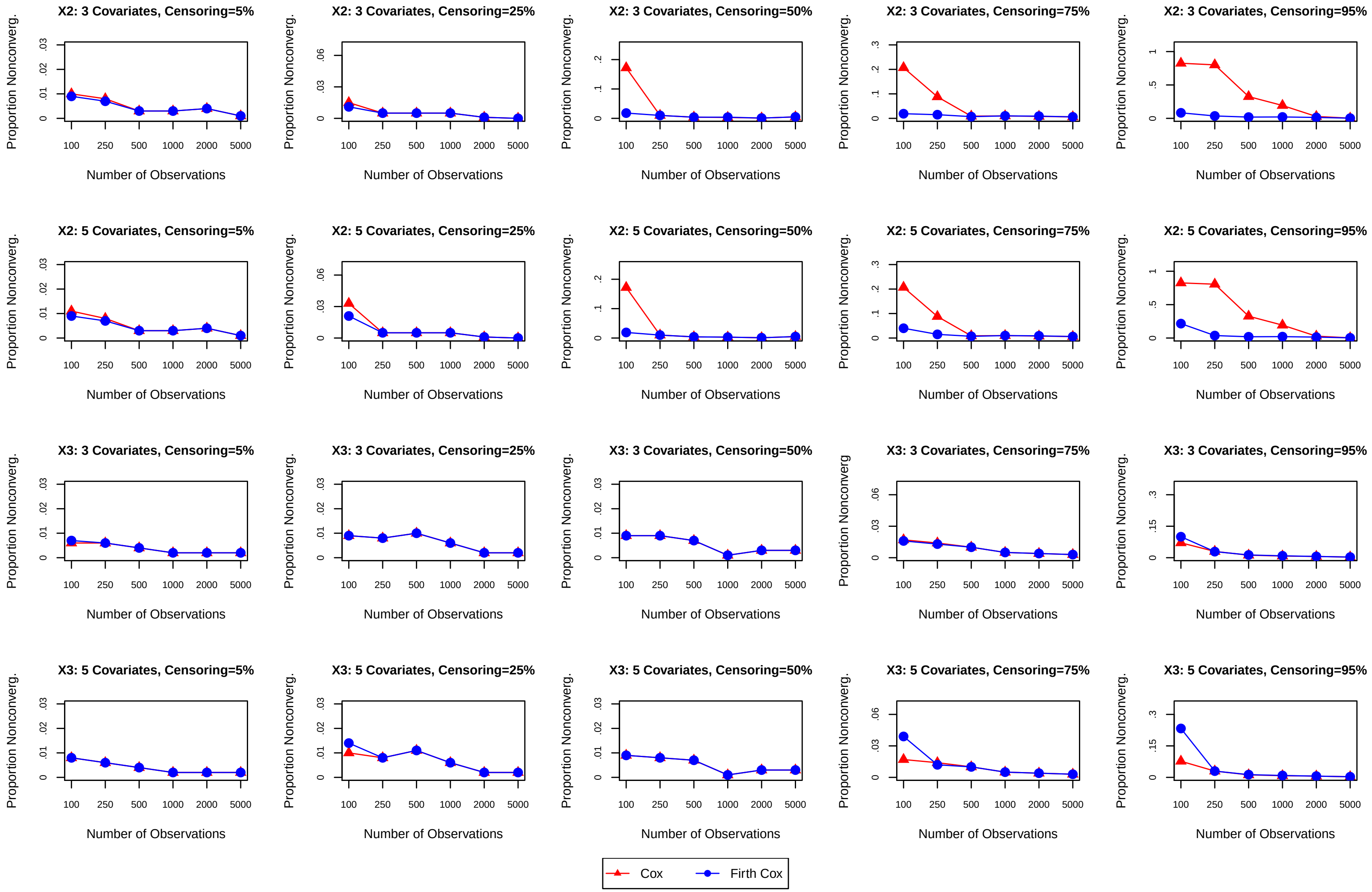

Figure 3: Nonconvergence Levels for $\mathbf{x_2}$ and $\mathbf{x_3}$ Across Relevant Monte Carlo Experiments

Turning to the RMSE statistics for $\mathbf{x_2}$ and $\mathbf{x_3}$, which are reported in Figure 4, we can first note that our Cox and Firth Cox coefficient estimates for β_2 exhibit similar RMSEs across our various experimental conditions when censoring is low-to-moderate. The one exception to this generalization occurs at $N = 100$, where we typically find that the Firth Cox model substantially outperforms the Cox model in accuracy when $c = \{25\%, 50\%\}$. At moderate-to-high levels of censoring (i.e., 75% to 95%), the β_2 RMSEs for the Cox model further indicate that a standard Cox estimator returns coefficient estimates with substantially higher inaccuracy than those of the Firth Cox model in circumstances where one is analyzing a dataset with a low-to-moderate N. This is most noticeable in Figure 4 when $N = \{100, 250\}$, but also consistently arises in cases where $N = \{500, 1000, 2000\}$. Moreover, even among those $\mathbf{x_2}$ Monte Carlo conditions where the Cox and Firth Cox models' RMSEs do not visibly diverge in Figure 4, we find in examining the underlying plotted RMSE values that all RMSEs are slightly lower for the Firth Cox model, in comparison to the Cox model, across all conditions. Thus, with respect to accuracy, the Firth Cox model appears to be a superior choice to that of the Cox model across small to moderate Ns, and no matter the level(s) of underlying monotone likelihood in one's data and models.

When we instead consider $\mathbf{x_3}$, we find that the Cox model *does not* exhibit substantially higher inaccuracies in $\hat{\beta_3}$, in comparison to the β_3 estimates obtained from the Firth Cox model. This generalization holds under virtually all conditions evaluated. The one exception is the $N = 100$, $c = 95\%$ condition, where the Cox model does on average yield noticeably less accurate β_3 estimates, in comparison to the Firth Cox model. With regards to the overall similarity in $\hat{\beta_3}$ accuracy levels for the Cox and Firth Cox models across most conditions, recall that $\mathbf{x_3}$ is continuous and hence does not exhibit imbalance of its own. For such a variable, monotone likelihood problems—and the biases associated with monotone likelihood—are unlikely to arise, thus ensuring that the standard Cox model obtains similar accuracy to the Firth Cox model in all but the most extreme (low information) conditions. That being said, we still do find that the Firth Cox model obtains *slightly* more accurate estimates of β_3 across *all* of the experimental conditions presented in Figure 4, based upon our averaged RMSE values. This reinforces our earlier finding that the Firth Cox model will often slightly improve the accuracy of one's duration model estimates (i.e., over those of the standard Cox model) within low-to-modest sample sizes, even in the absence of monotone likelihood issues.

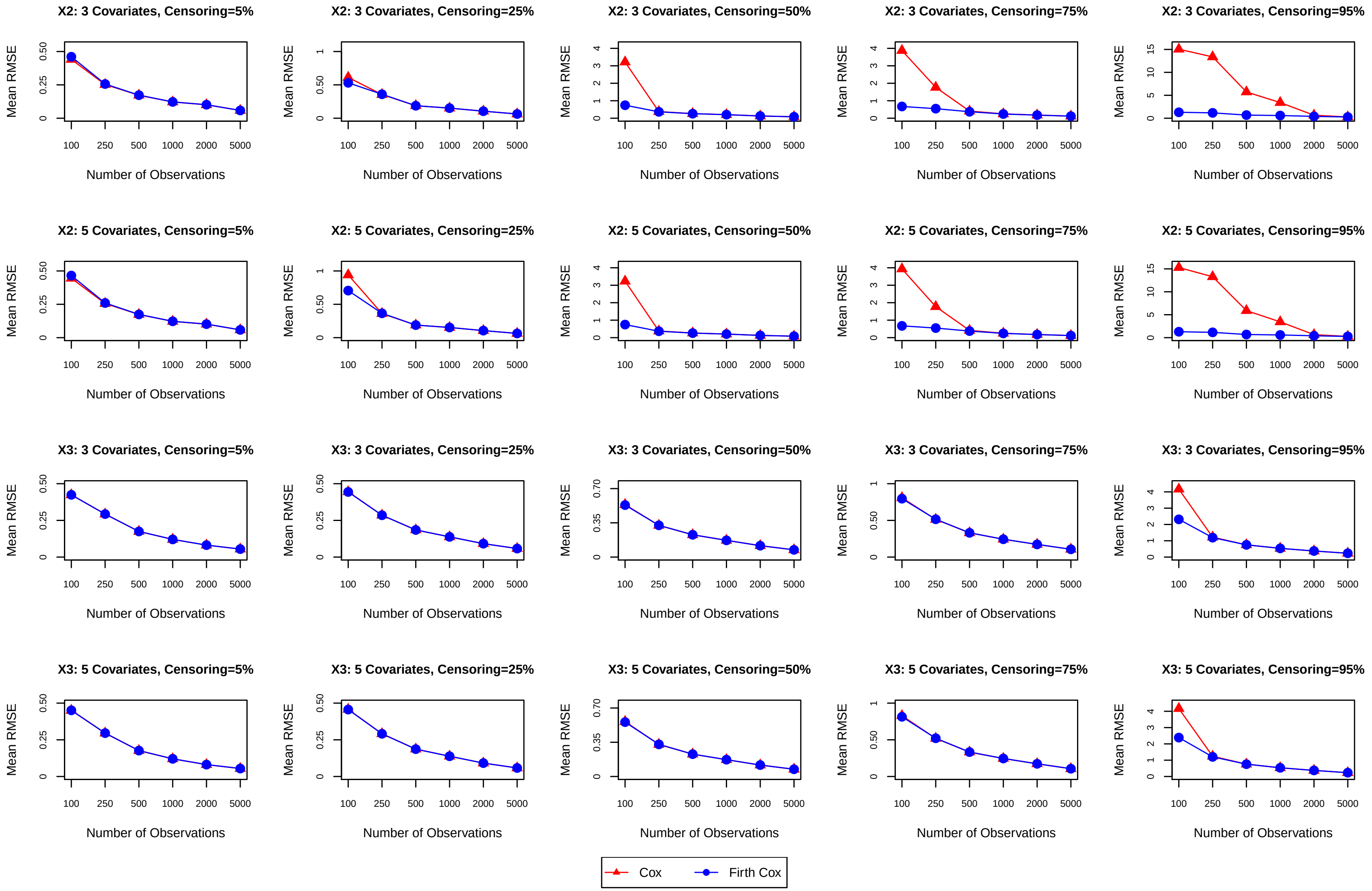

Figure 4: Root Mean Squared Errors for $\mathbf{x_2}$ and $\mathbf{x_3}$ Across Relevant Monte Carlo Experiments

3.3 Model Performance with respect to $\mathbf{x_4}$ and $\mathbf{x_5}$

We now turn to the nonconvergence rates and RMSE values for our final two independent variables of interest: $\mathbf{x_4}$ and $\mathbf{x_5}$. As stated at the outset of our Monte Carlo section, these two binary covariates were drawn according to $\mathbf{x_4} = 1_{Uniform[-2.5,12]>11.5}$ and $\mathbf{x_5} = 1_{\ln Uniform[1,100]>1.65}$ in our Monte Carlo simulations, therein ensuring that each contained a degree of imbalance. While we consistently used these two covariates to generate our duration outcomes across all Monte Carlo experiments, also recall that we only included $\mathbf{x_4}$ and $\mathbf{x_5}$ as predictors within $5 \times 6 = 30$ of our experimental conditions' actual estimated models (i.e., within the conditions that included five variables in the Cox and Firth Cox model specifications). For these conditions, we plot the relevant nonconvergence rates for $\mathbf{x_4}$ and $\mathbf{x_5}$ in Figure 5 and the RMSE values for $\mathbf{x_4}$ and $\mathbf{x_5}$ in Figure 6. As above, these two figures depict all relevant experimental conditions. The different levels of censoring that we consider are depicted across the subfigure columns of Figures 5-6 and the x-axis of each subfigure therein depicts the different N's evaluated. For Figures 5-6, each subfigure row then corresponds to a different covariate considered (i.e., $\mathbf{x_4}$ or $\mathbf{x_5}$), since model specification does not vary in this context.[11]

Turning to the nonconvergence rates for $\mathbf{x_4}$ and $\mathbf{x_5}$ in Figure 5, we find that the standard Cox model exhibits notably higher rates of nonconvergence than does the Firth Cox model when (i) $N = \{100, 250\}$ for virtually all censoring cases and (ii) $N = \{100, 250, 500, 1000, 2000\}$ for the 95% censoring case. At the extremes, we find in these instances that the standard Cox model fails to converge in 4.3%-6.8% of all simulations in the 50% censoring case, in 45.6%-77.0% of all simulations in the 75% censoring case, and in 77.4%-95.1% of all simulations in the 95% censoring case. By comparison, the Firth Cox model exhibits far lower nonconvergence rates (of 1.1%-1.4%, 5.0%-5.0%, and 20.3%-21.4%, respectively) in these same cases. Thus, when censoring is moderate (50%-75%) and one has multiple imbalanced predictors, the Firth Cox model is a far superior choice to the Cox model when N is low, and is practically a necessity when censoring is high (i.e., $c = 95\%$). The RMSEs in Figure 6 reinforce these conclusions. Specifically, Figure 6 demonstrates that substantial biases and inefficiencies in one's Cox model coefficient estimates for $\mathbf{x_4}$ and $\mathbf{x_5}$ arise in the conditions highlighted immediately above, but do not arise in the case of the Firth Cox model. That is, we find that failure to employ penalized maximum likelihood estimation within a Cox model's estimation yields noticeably poorer accuracy in $\hat{\beta}_4$ and $\hat{\beta}_5$ when (i) censoring is moderate-to-high (75%-95%) for most relevant N's and (ii) when N is low, no matter the level of censoring.

[11] That is, we only evaluate $\mathbf{x_4}$ and $\mathbf{x_5}$ under conditions of full model specification in our Monte Carlo experiments.

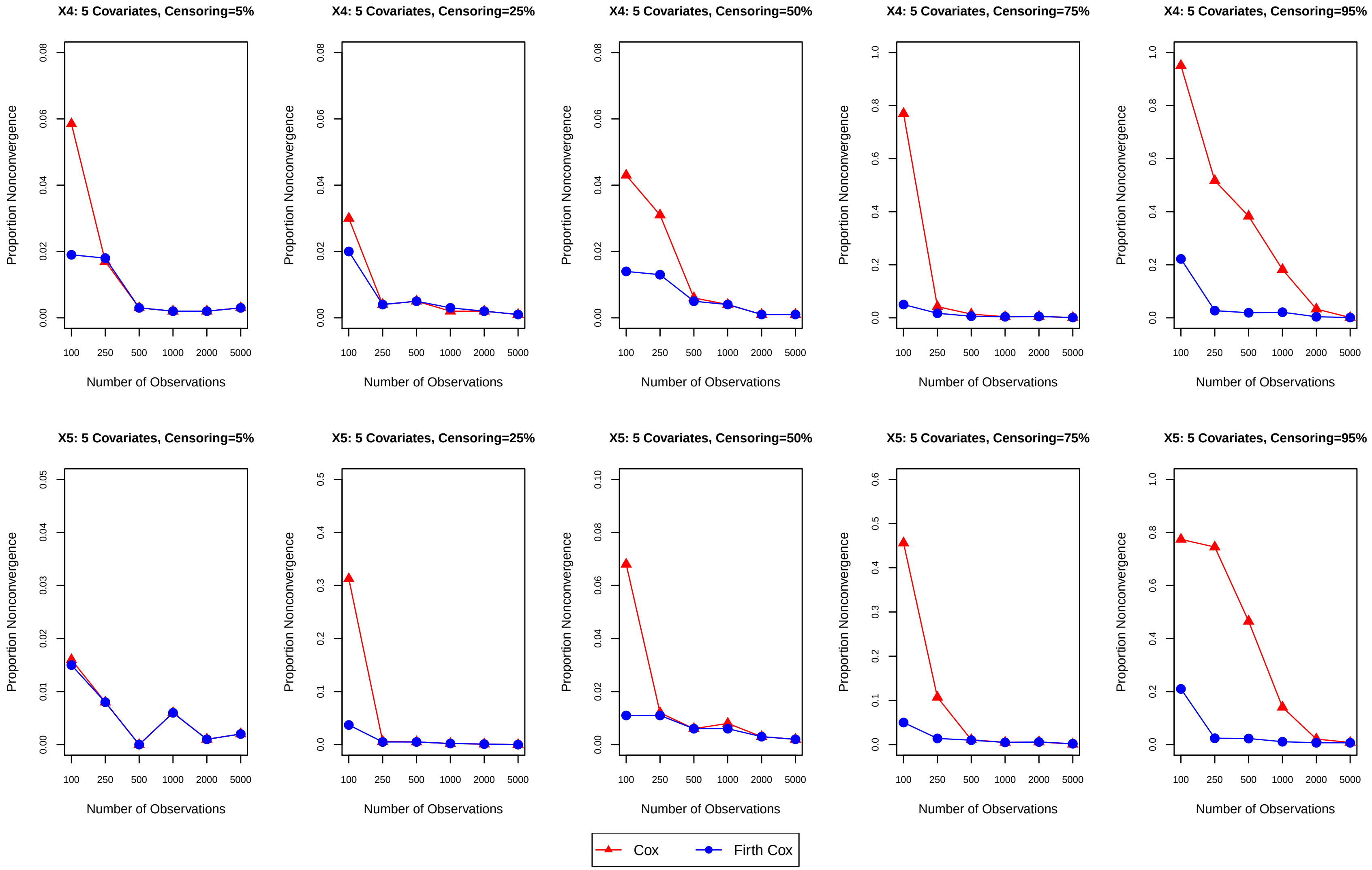

Figure 5: Nonconvergence Levels for $\mathbf{x_4}$ and $\mathbf{x_5}$ Across Relevant Monte Carlo Experiments

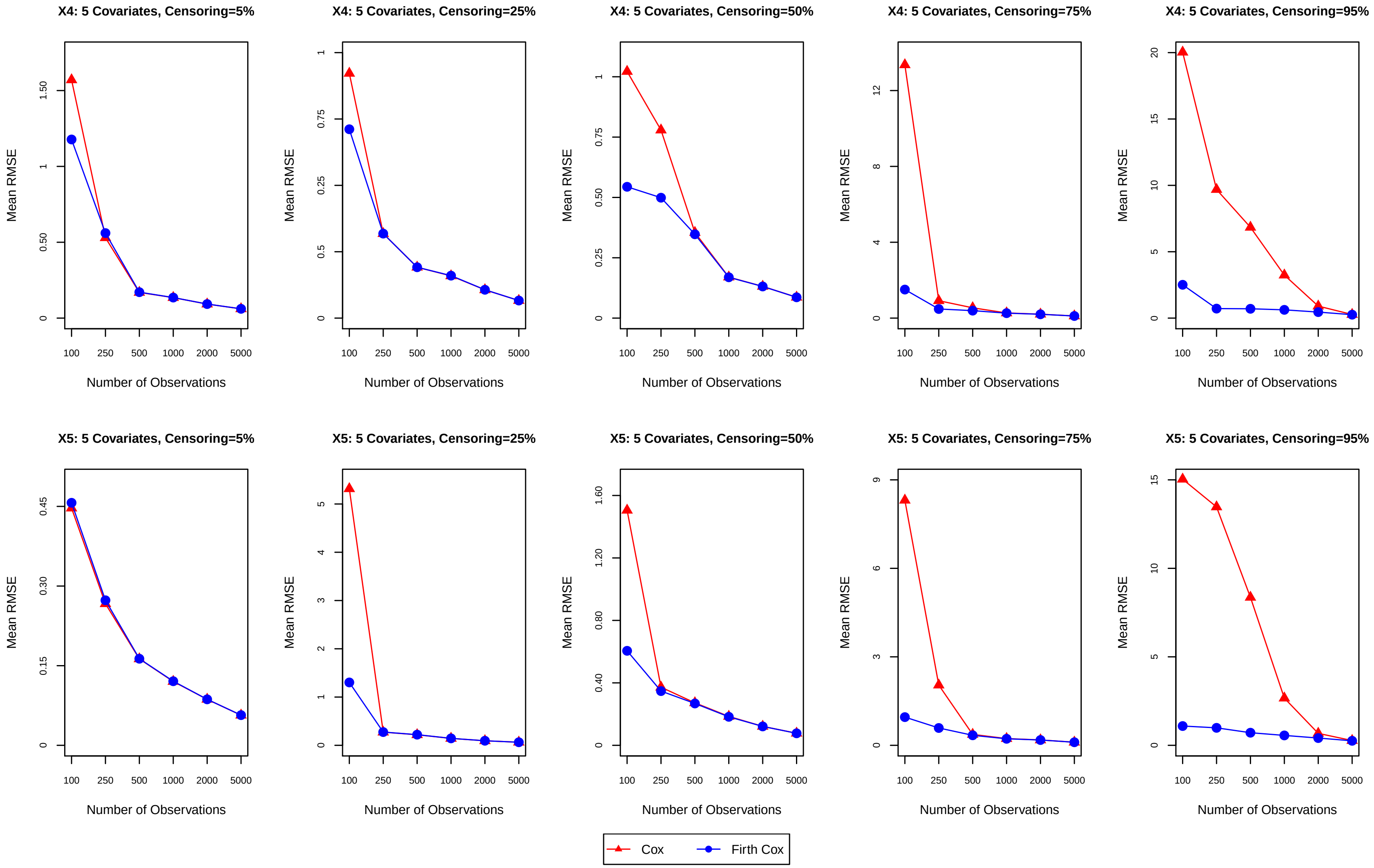

Figure 6: Root Mean Squared Errors for $\mathbf{x_4}$ and $\mathbf{x_5}$ Across Relevant Monte Carlo Experiments

3.4 Additional Model Comparison Metrics

This section presents a number of additional summary quantities for the Monte Carlo comparisons discussed above. First, as mentioned at the outset of the Monte Carlo section of this appendix, recall that we additionally examined the MSEs for all Monte Carlo comparisons and coefficient estimates as a secondary measure of bias and variance (i.e., accuracy) to that of the RMSE. As noted previously, the (near) infinite estimates obtained under monotone likelihood conditions for the standard Cox model oftentimes led the Cox model's MSEs to be so many orders magnitude greater (i.e., worse) than those of the Firth Cox model that comparing these quantities via plots was uninformative. This led us to favor RMSEs within our primary Monte Carlo assessments above. Yet, because the MSE has been favored as a direct summary quantity in similar Monte Carlo analyses in the past (e.g., Rainey and McKaskey, forthcoming), we report comparable MSE plots (i.e., to our primary RMSE plots) in Figures 7-9 below.

Turning to Figures 7-9, we can observe similar overall trends to those identified within our earlier RMSE discussions. Across our four imbalanced predictors (i.e., $\mathbf{x_1}$, $\mathbf{x_2}$, $\mathbf{x_4}$ and $\mathbf{x_5}$) we find that the Firth Cox model consistently exhibits superior accuracy in our β estimates, as measured by lower MSE values. Moreover, when one's N is low—no matter the level of censoring—the Firth Cox model generally recovers estimates with substantially superior accuracy, such that plotted MSEs for the Firth Cox model are often 5-15 times smaller than those of the Cox model. When N is low and censoring is (also) high (e.g., 75%-95%), we further find that the Firth Cox model's levels of accuracy in $\hat{\beta}$—as measured via MSEs—are often well over 30 times (and often over 100 times) smaller than those of the Cox model. In these cases, the Cox model does not appear comparable to the Firth Cox model in accuracy until one's N rises above 2,000. Lastly, we can note that the MSE results presented in 7-9 also underscore several of the related findings discussed above, most notably in demonstrating again that these divergences in accuracy become more acute as one adds additional (imbalanced) covariates to one's model specification.

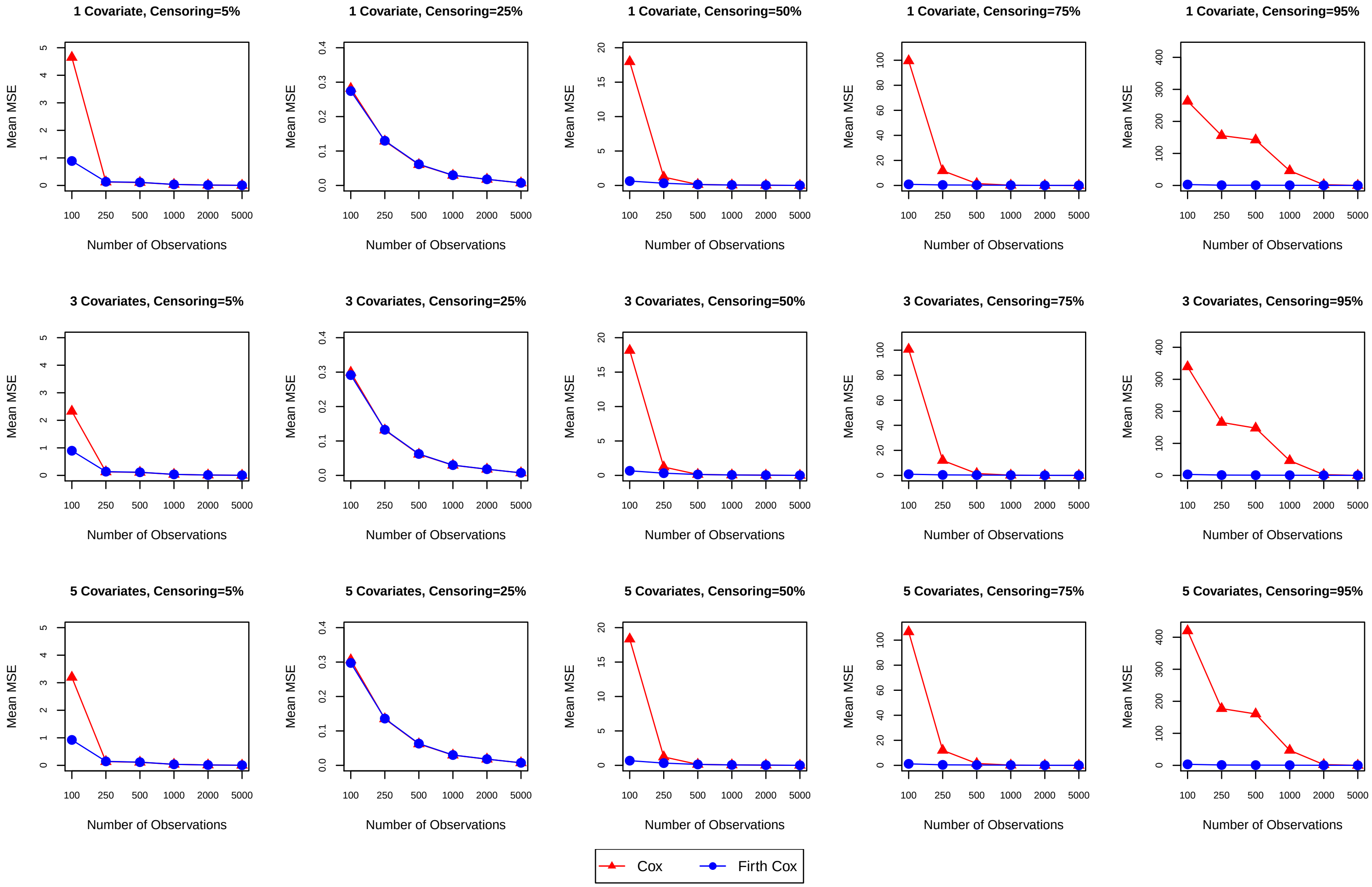

Figure 7: Mean Squared Errors for $\mathbf{x_1}$ Across All Monte Carlo Experiments

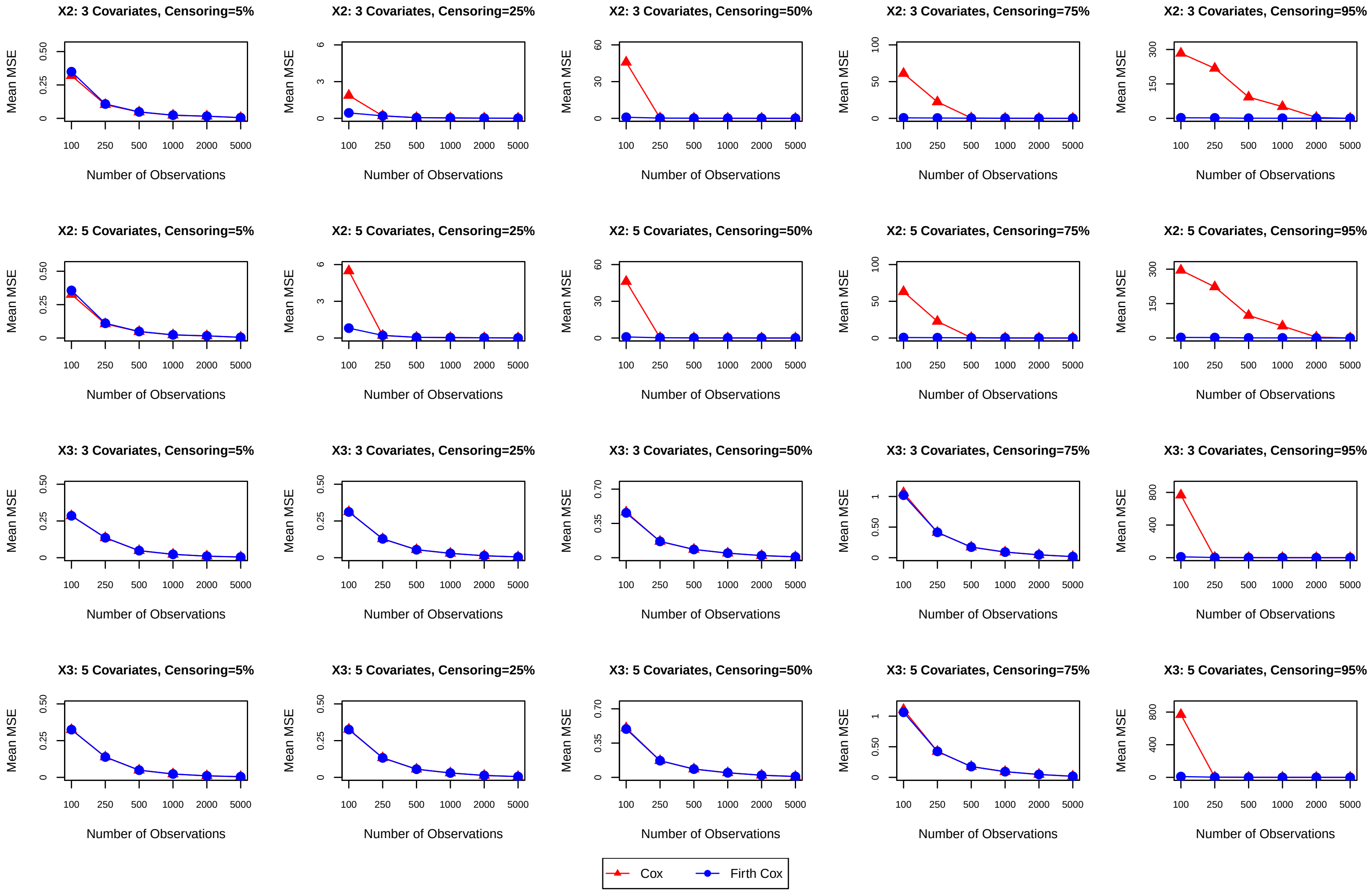

Figure 8: Mean Squared Errors for $\mathbf{x_2}$ and $\mathbf{x_3}$ Across Relevant Monte Carlo Experiments

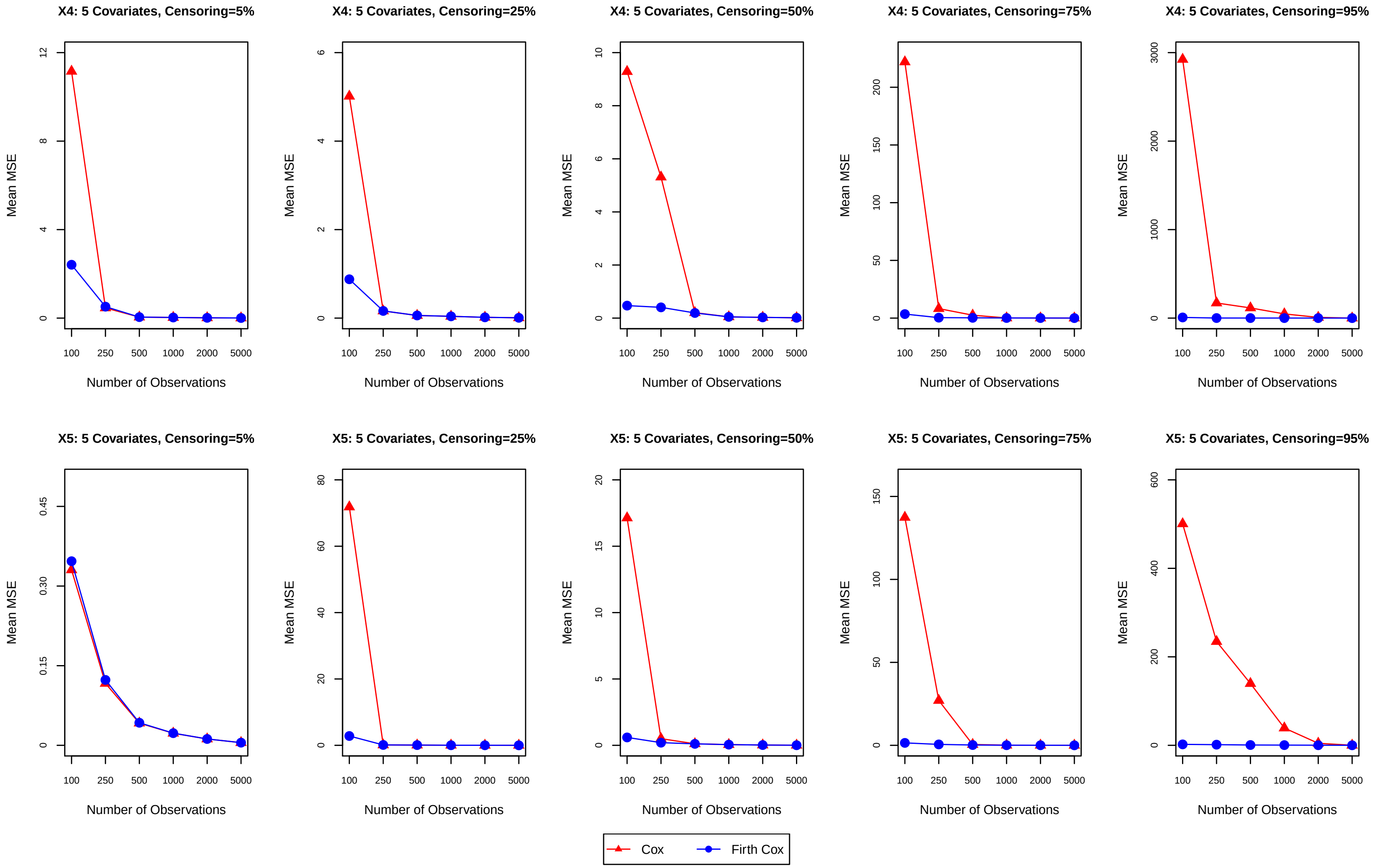

Figure 9: Mean Squared Errors for $\mathbf{x_4}$ and $\mathbf{x_5}$ Across Relevant Monte Carlo Experiments

Second, recall that our Monte Carlo experiments reported nonconvergence rates based upon instances where a particular model provided an error or warning message either (i) globally or for a given covariate, *or* (ii) where a model's coefficient estimate or its standard error was at least 500 times larger (or smaller) than the true value for a given parameter under consideration. To further demonstrate that our primary nonconvergence findings and conclusions are robust to the latter threshold-choice, we re-create our non-convergence plots after removing the "500 times larger/smaller" threshold entirely[12] in Figures 10-12. For these Figures, we maintain the same y-axes ranges to those reported for our primary Nonconvergence results (i.e., in Figures 1, 3, and 5) to facilitate direct comparisons.

If we ignore instances of nonconvergence that arise due to our coefficient estimates or standard errors being at least 500 times larger (or smaller) than a true parameter value, we find highly comparable results to our main nonconvergence findings across Figures 10-12. All nonconvergence rates—for both the Cox and Firth Cox models—exhibit slight decreases relative to our main nonconvergence results. As a result of these decreases, our Firth Cox models' nonconvergence rates are effectively zero across virtually all conditions for most covariates reported in Figures 10-12. On the other hand, the Cox models' nonconvergence rates in Figures 10-12 oftentimes remain quite high. For instance, we continue to find Cox model nonconvergence rates of 20% (10%) or higher for covariates X_1, X_2, X_4, and X_5 when censoring is at 75% and N is equal to 100 (250). We likewise continue to find Cox model nonconvergence rates of roughly 40% or higher for covariates X_1, X_2, X_4, and X_5 when censoring is at 95% and N is less than or equal to 500—and Cox nonconvergence rates greater than 10% in these instances when $N = 1000$. These findings reinforce our earlier conclusions. No matter the nonconvergence criteria used, the Cox model exhibits substantial nonconvergence issues under conditions of high censoring, imbalanced predictors, and low N's, and non-negligible nonconvergence issues even when one or more of these conditions are relaxed. The Firth Cox model addresses these issues, and exhibits virtually zero instances of nonconvergence[13] under these same conditions.

[12] That is, when defining nonconvergence for a given covariate based upon a covariate-specific or global error or warning message.

[13] When nonconvergence is defined based upon covariate-specific or global model errors.

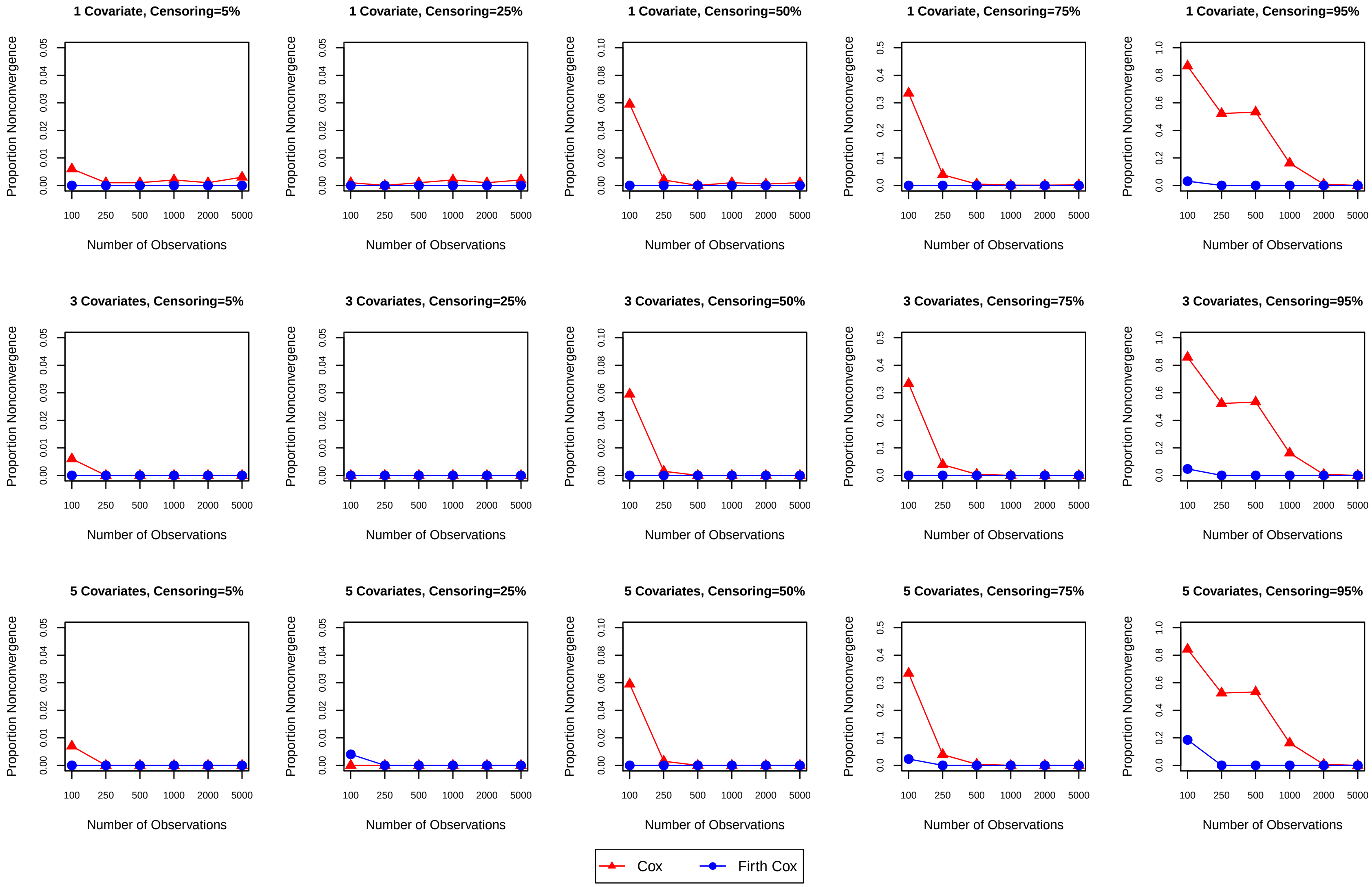

Figure 10: Nonconvergence Levels for $\mathbf{x_1}$, With Nonconvergence Defined Only by Model Warnings & Errors

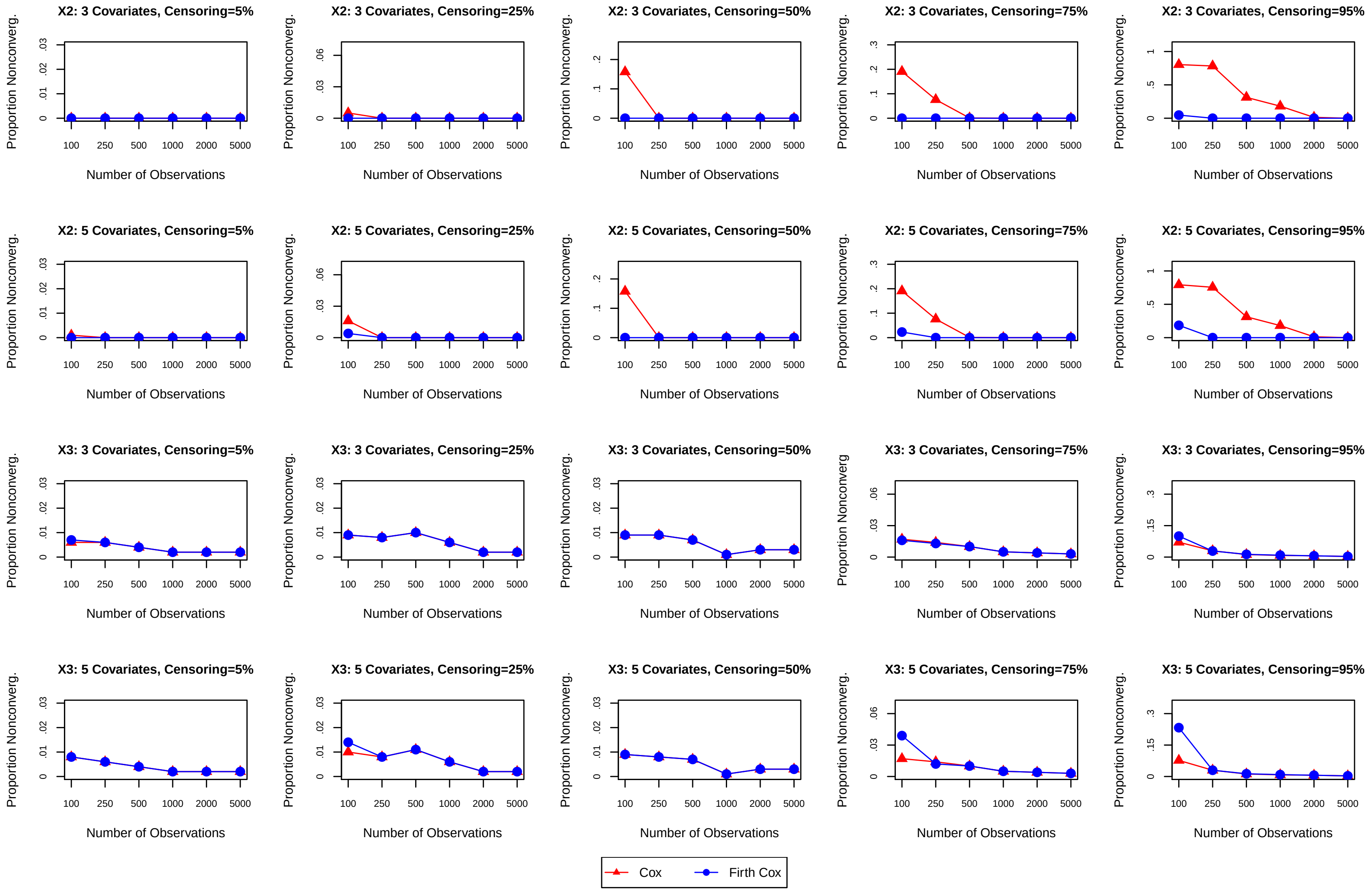

Figure 11: Nonconvergence Levels for $\mathbf{x_2}$ and $\mathbf{x_3}$, With Nonconvergence Defined Only by Model Warnings & Errors

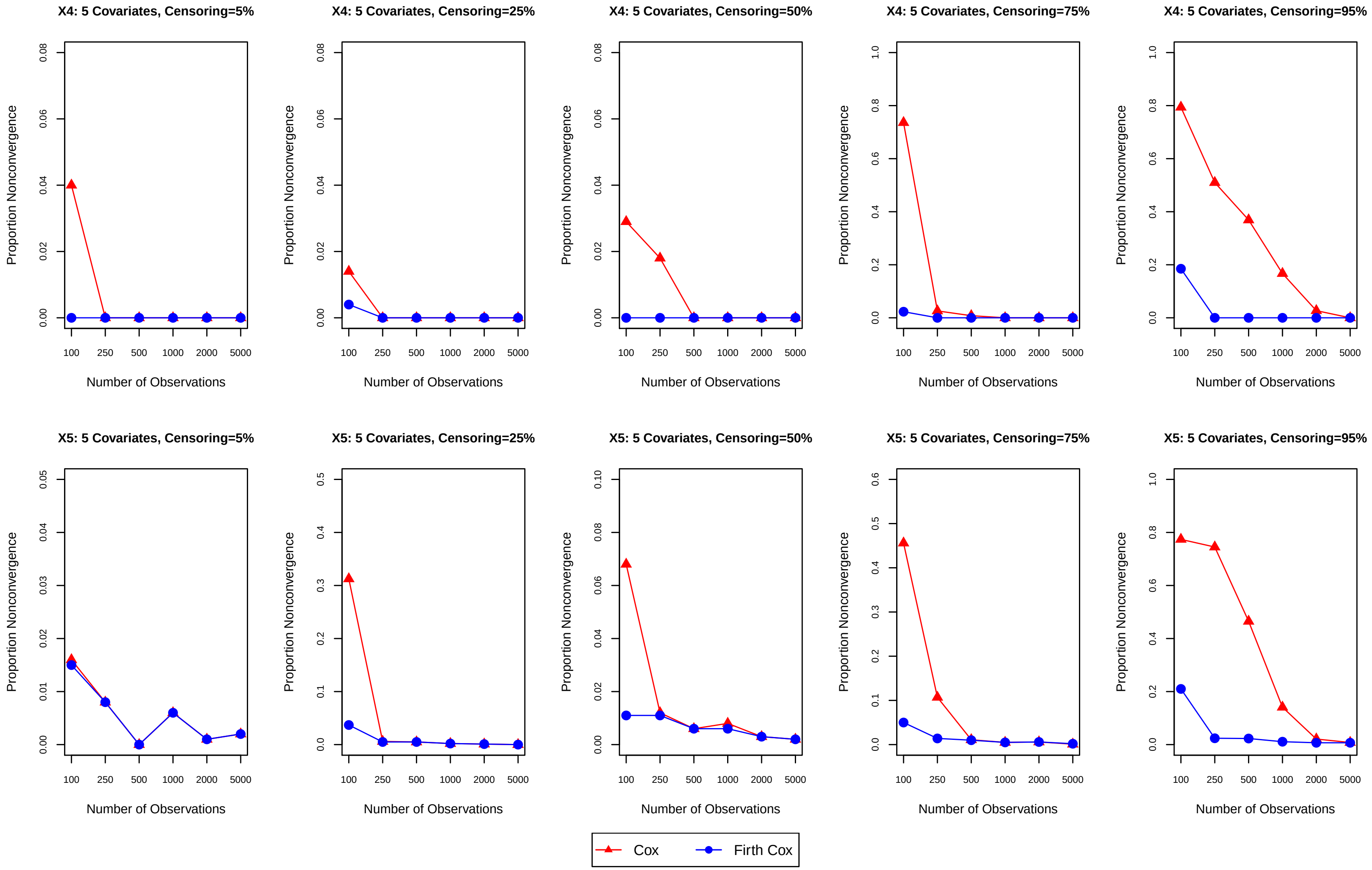

Figure 12: Nonconvergence Levels for $\mathbf{x_4}$ and $\mathbf{x_5}$, With Nonconvergence Defined Only by Model Warnings & Errors

As a final quantity of interest, we also examine the relationship between (i) our binary predictors' *imbalance ratios* and (ii) the nonconvergence rates obtained for each of these predictors. For our four *binary* predictors $(\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_4}, \mathbf{x_5})'$, we calculate imbalance ratios as follows. First, for each Monte Carlo simulation, and for only those observations where a binary predictor was equal to one, we tabulate the number of instances where $Y =$ Censored and where $Y =$ Failed. We then take the ratio of these cases, and normalize this quantity for each N considered. As such, a covariate with an imbalance ratio of 0:100 corresponds to a covariate whose "1"-cases always saw Y fail, whereas a covariate with an imbalance ratio of 100:0 indicates a covariate whose "1"-cases always saw Y exhibit censoring. A binary predictor with an imbalance ratio of 50:50 would instead reflect a predictor that saw an equal number of failures and censoring cases among its "1"-cases in the dataset. After calculating this ratio for reach binary predictor, and for each simulated dataset within our Monte Carlo simulations, we average this quantity for each covariate across all 1,000 Monte Carlo simulations within a given Monte Carlo condition. We then compare this averaged quantity to each binary covariate's nonconvergence rates, which are again averaged across all simulated datasets within a given Monte Carlo condition.

For each Monte Carlo condition considered, we plot these quantities across the full range of imbalance cases that were obtained in our simulations—for all four binary predictors—in Figure 13. We include dashed vertical lines at the 0:95 and 95:0 thresholds to distinguish cases of low-to-moderate imbalance from cases of high imbalance. We then also re-plot these quantities for only the high-imbalance cases (i.e., the cases with an imbalance ratio of 95:0 or greater) within Figure 14. The x-axes to these figures reflect the range of imbalance considered, whereas the y-axes report the corresponding nonconvergence rates. The plotted points then depict a given covariate's levels of imbalance and model nonconvergence for a single Monte Carlo condition, separately for our estimated Firth Cox and Cox models.

Turning to Figure 13, we can first observe that the various conditions evaluated across our Monte Carlo simulations together ensure that we observe imbalance ratios that range across a relatively large share of the possible spectrum of (im)balance. However, most cases in our simulations cluster at the high end of imbalance (i.e., from 95:5 to 100:0) by virtue of the fact that our binary predictors were generally drawn to be imbalanced toward the 95:5 to 100:0 range by design. Looking across all levels of (im)balance in Figure 13, we can note that while the Cox and Firth Cox models often obtain similar levels of (low) nonconvergence at most levels of (im)balance, the Cox model at times exhibits noticeably higher nonconvergence rates across each and every level of (im)balance observed. When imbalance is moderate, the aforementioned divergences in nonconvergence only appear to arise within a small number of Monte Carlo conditions (most often those conditions where five covariates are included, or N is low), and often only encompass differences in nonconvergence rates of 3%-5%. However, even when a covariate's imbalance is relatively moderate and close to 75:25, we observe at least one set of Monte Carlo conditions where the Cox model fails to converge in almost 80% of all simulations—a nonconvergence rate that is nearly ten times the comparable nonconvergence rate for the comparable Firth Cox model. These divergences become even more frequent and pronounced towards the high end of imbalance in Figure 13, which we depict in further detail in Figure 14.

Turning to Figure 14, we find that—at levels of high imbalance—the Cox model exhibits nonconvergence rates that are frequently several times larger than those of the corresponding

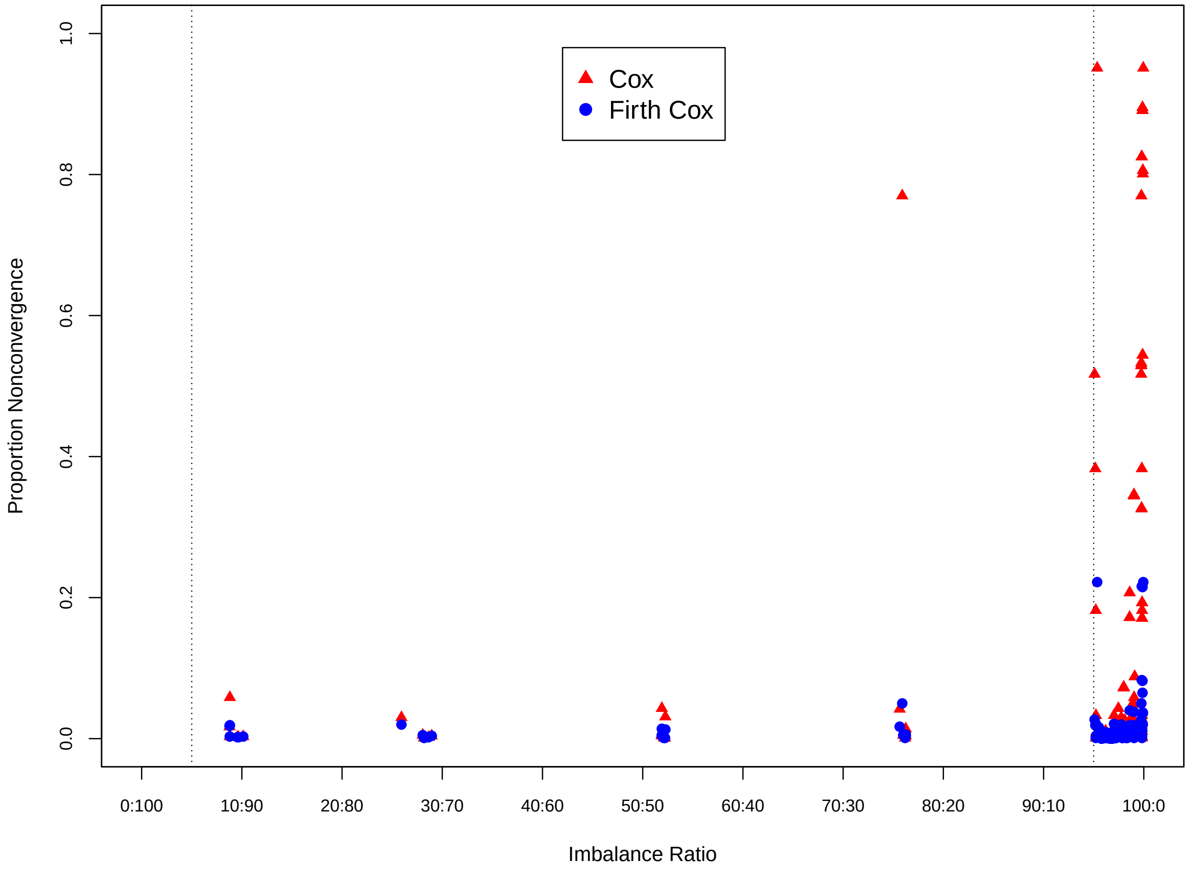

Figure 13: Relationship Between Imbalance Ratio and Nonconvergence for $\mathbf{x_1}$, $\mathbf{x_2}$, $\mathbf{x_4}$, & $\mathbf{x_5}$

Firth Cox model. Moreover, oftentimes the Cox model's nonconvergence encompasses at least 50% of all simulations within a given Monte Carlo condition—which is a threshold that is never crossed in the Firth Cox model averages. In Figure 14, these divergences in imbalance-based nonconvergence rates between the Cox and Firth Cox model appear to arise most systematically at levels of imbalance of 98:2 or higher. Even so, we can also observe multiple instances where the Cox model exhibits uniquely high nonconvergence rates for imbalance ratios closer to 95:5. Hence, for researchers analyzing duration data with binary predictors, no level of balance can guarantee that monotone likelihood issues *will not* arise, whereas levels of high imbalance—i.e., 95:5 or higher (or conversely, 5:95 or lower)—often guarantee that monotone likelihood issues *will* arise. In each instance, the Firth Cox model appears to be a superior alternative to that of the Cox model for estimation and inference.

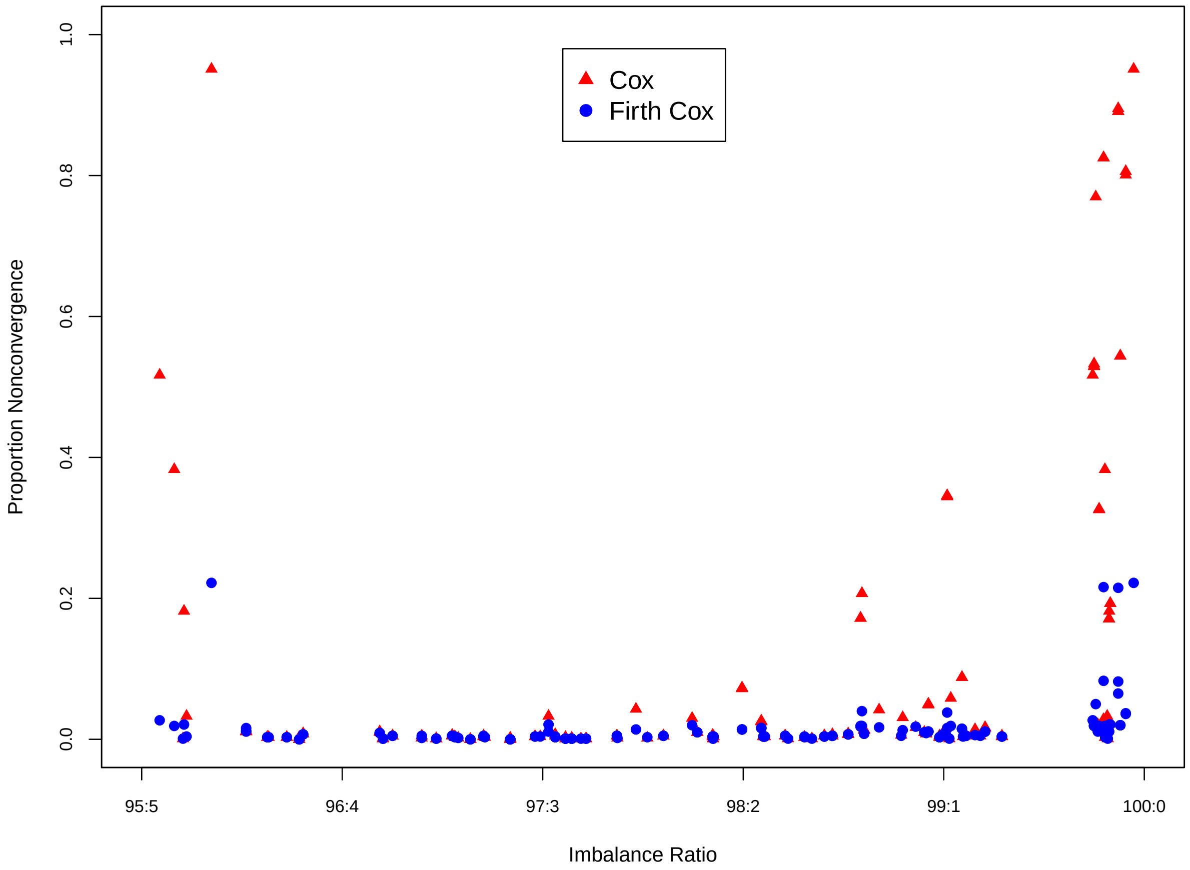

Figure 14: Relationship Between Imbalance Ratio and Nonconvergence for $\mathbf{x_1}$, $\mathbf{x_2}$, $\mathbf{x_4}$, & $\mathbf{x_5}$ Among High-Imbalance Cases Only

3.5 Summary

Our Monte Carlo experiments offer several insights. Assuming the estimation of a duration model with at least one imbalanced binary predictor, we find that monotone likelihood issues are most acute when (i) censoring encompasses approximately 75%-95% of all observations or (ii) $N \leq 1000$. When both of these conditions are present, monotone likelihood problems become so substantial that a third or more of a standard Cox model's estimates will be infinite or near-infinite for at least one coefficient. Adding additional imbalanced predictors in these settings exacerbates these problems, even for one's originally included predictor. In each of these settings, the Firth Cox model consistently outperforms the standard Cox model in terms of both lower levels of nonconvergence and superior levels of accuracy (as measured by RMSE). The Firth Cox model is thus a preferable estimator for duration model applications that include imbalanced predictors, *especially* in circumstances where one's N is moderate-to-low and/or where censoring is moderate-to-high.

We also find that even absent significant monotone likelihood issues, the Firth Cox model is generally (albeit often only slightly) superior to the standard Cox model in terms of accuracy and nonconvergence. Similar to recent findings for the logit estimator (Rainey and McKaskey, forthcoming), this indicates that the Firth Cox model may be preferable to the Cox model in *any* circumstances of low N (i.e., $N \leq 1000$)—that is, even in the absence of imbalanced binary predictors or high censoring. However, it is worth emphasizing that each of the findings and conclusions summarized here are dependent upon the specific choices of sample size, censoring, specification, covariates, and parameter values used above, and thus may not generalize to all real-world settings and contexts.

Bibliography

Acosta, Benjamin. 2014. "From Bombs to Ballots: When Militant Organizations Transition to Political Parties." *Journal of Politics* 76(3):666–683.

Ahlquist, John S. 2010. "Policy by Contract: electoral cycles, parties and social pacts, 1974-2000." *Journal of Politics* 72(2):572–587.

Arias, Eric, James R. Hollyer and B. Peter Rosendorff. 2018. "Cooperative Autocracies: Leader Survival, Creditworthiness, and Bilateral Investment Treaties." *American Journal of Political Science* 62(4):905–921.

Baccini, Leonardo and Johannes Urpelainen. 2013. "International Institutions and Domestic Politics: Can Preferential Trading Agreements Help Leaders Promote Economic Reform?" *Journal of Politics* 76(1):195–214.

Beardsley, Kyle. 2008. "Cooperative Autocracies: Leader Survival, Creditorthiness and Bilateral Investment Treaties." *American Journal of Political Science* 52(4):723–740.

Berliner, Daniel and Aaron Erlich. 2015. "Competing for Transparency: Political Competition and Institutional Reform in Mexican States." *American Political Science Review* 109(1):110–128.

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

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

Boix, Carles and Milan W. Svolik. 2013. "The Foundations of Limited Authoritarian Government: Institutions, Commitment, and Power-Sharing in Dictatorships." *Journal of Politics* 75(2):300–316.

Boudreau, Cheryl and Scott A. MacKenzie. 2018. "Wanting What Is Fair: How Party Cues and Information about Income Inequality Affect Public Support for Taxes." *Journal of Politics* 80(2):367–381.

Breslow, N. 1974. "Covariance Analysis of Censored Survival Data." *Biometrics* 30(1):89–99.

Bueno de Mesquita, Bruce and Alastair Smith. 2010. "Leader Survival, Revolutions, and the Nature of Government Finance." *American Journal of Political Science* 54(4):936–950.

Camerlo, Marcelo and Aníbal Pérez-Li nán. 2015. "Minister Turnover, Critical Events, and the Electoral Calendar in Presidential Democracies." *Journal of Politics* 77(3):608–619.

Capoccia, Giovanni, Lawrence Sáez and Eline de Rooij. 2012. "When State Responses Fail: Religion and Secessionism in India 1952–2002." *Journal of Politics* 74(4):1010–1022.

Carpenter, Daniel, Jacqueline Chattopadhyay, Susan Moffitt and Clayton Nall. 2012. "The Complications of Controlling Agency Time Discretion: FDA Review Deadlines and Postmarket Drug Safety." *American Journal of Political Science* 56(1):98–114.

Cook, Scott J., Jude C. Hays and Robert J. Franzese. 2018. "Fixed effects in rare events data: a penalized maximum likelihood solution." *Political Science Research and Methods* pp. 1–14.

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

Curry, Todd A. and Mark S. Hurwitz. 2016. "Strategic Retirements of Elected and Appointed Justices: A Hazard Model Approach." *Journal of Politics* 78(4):1061–1075.

Darmofal, David. 2009. "Bayesian Spatial Survival Models for Political Event Processes." *American Journal of Political Science* 53(1):241–257.

Davis, Christina L. and Meredith Wilf. 2017. "Joining the Club: Accession to the GATT/WTO." *Journal of Politics* 79(3):964–978.

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

Findley, Michael G. and Joseph K. Young. 2015. "Terrorism, Spoiling, and the Resolution of Civil Wars." *Journal of Politics* 77(4):1115–1128.

Flores, Thomas Edward and Irfan Nooruddin. 2012. "The Effect of Elections on Postconflict Peace and Reconstruction." *Journal of Politics* 74(2):558–570.

Fortunato, David and Matt Loftis. 2018. "Cabinet Durability and Fiscal Discipline." *American Political Science Review* 112(4):939–953.

Fukumoto, Kentaro. 2009. "Systematically Dependent Competing Risks and Strategic Retirement." *American Journal of Political Science* 53(3):740–754.

Gauri, Varun, Jeffrey K. Staton and Jorge Vargas Cullell. 2015. "The Costa Rican Supreme Court's Compliance Monitoring System." *Journal of Politics* 77(3):774–786.

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

Golder, Matt, Sona N. Golder and David A. Siegel. 2012. "Modeling the Institutional Foundation of Parliamentary Government Formation." *Journal of Politics* 74(2):427–445.

Harden, Jeffrey J. and Jonathan Kropko. 2018. "Simulating Duration Data for the Cox Model." *Political Science Research and Methods* pp. 1–8.

Hassell, Hans J. G. 2015. “Party Control of Party Primaries: Party Influence in Nominations for the US Senate.” *Journal of Politics* 78(1):75–87.

Heersink, Boris. 2018. “Trump and the Party-in-Organization: Presidential Control of National Party Organizations.” *Journal of Politics* 80(4).

Heinze, Georg and Meinhard Ploner. 2018. *coxphf: Cox Regression with Firth's Penalized Likelihood*. version 1.13.
URL: *https://CRAN.R-project.org/package=coxphf*

Heinze, Georg and Michael Schemper. 2001. “A Solution to the Problem of Monotone Likelihood in Cox Regression.” *Biometrics* 57(1):114–119.

Hollyer, James R., R. Peter Rosendorff and James Raymond Vreeland. 2015. “Transparency, Protest, and Autocratic Instability.” *American Political Science Review* 109(4):764–784.

Huber, John D. and Cecilia Martinez-Gallardo. 2008. “Regime Type, the Fate of Leaders, and War.” *American Political Science Review* 102(2):169–180.

Johns, Leslie and Krzysztof J. Pelc. 2018. “Free Riding on Enforcement in the World Trade Organization.” *Journal of Politics* 80(3):873–889.

Kelley, Judith G. and Beth A. Simmons. 2015. “Politics by Number: Indicators as Social Pressure in International Relations.” *American Journal of Political Science* 59(1):55–70.

Kittilson, Miki Caul. 2008. “Representing Women: The Adoption of Family Leave in Comparative Perspective.” *Journal of Politics* 70(2):323–334.

Knutsen, Carl Henrik and Håvard Mokleiv Nygård. 2015. “Institutional Characteristics and Regime Survival: Why Are Semi-Democracies Less Durable Than Autocracies and Democracies?” *American Journal of Political Science* 59(3):656–670.

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

Kokkonen, Andrej and Anders Sundell. 2014. “Delivering Stability—Primogeniture and Autocratic Survival in European Monarchies 1000—1800.” *American Political Science Review* 108(2):438–453.

Laver, Michael and Kenneth Benoit. 2015. “The Basic Arithmetic of Legislative Decisions.” *American Journal of Political Science* 59(2):275–291.

Leventoğlu, Bahar and Nils W. Metternich. 2018. “Born Weak, Growing Strong: Anti-Government Protests as a Signal of Rebel Strength in the Context of Civil Wars.” *American Journal of Political Science* 62(3):581–596.

Lyall, Jason. 2010. “Are Coethnics More Effective Counterinsurgents?” *American Political Science Review* 104(1):1–20.

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

Malesky, Edmund. 2009. "Gerrymandering—Vietnamese Style: Escaping the Partial Reform Equilibrium in a Nondemocratic Regime." *Journal of Politics* 71(1):132–159.

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

Mattes, Michaela and Burcu Savun. 2010. "Information, Agreement Design, and the Durability of Civil War Settlements." *American Journal of Political Science* 54(2):511–524.

Narang, Neil. 2014. "Humanitarian Assistance and the Duration of Peace after Civil War." *Journal of Politics* 76(2):446–460.

Ostrander, Ian. 2016. "The Logic of Collective Inaction: Senatorial Delay in Executive Nominations." *American Journal of Political Science* 60(4):1063–1076.

Owsiak, Andrew P. and Toby J. Rider. 2013. "Clearing the Hurdle: Border Settlement and Rivalry Termination." *Journal of Politics* 75(3):757–772.

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.

Potter, Rachel Augustine. 2017. "Slow-Rolling, Fast-Tracking, and the Pace of Bureaucratic Decisions in Rulemaking." *Journal of Politics* 79(3):841–855.

Rainey, Carlisle and Kelly McKaskey. forthcoming. "Estimating Logit Models with Small Samples." *Political Science Research and Methods* .

Reenock, Christopher, Jeffrey K. Staton and Marius Radean. 2013. "Legal Institutions and Democratic Survival." *Journal of Politics* 75(2):491–505.

Scherer, Nancy, Brandon L. Bartels and Amny Steigerwalt. 2008. "Sounding the Fire Alarm: The Role of Interest Groups in the Lower Federal Court Confirmation Process." *Journal of Politics* 70(4):1026–1039.

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

Smith, Daniel A. and Dustin Fridkin. 2008. "Delegating Direct Democracy: Interparty Legislative Competition and the Adoption of the Initiative in the American States." *American Political Science Review* 102(3):333–350.

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

Therneau, Terry M. 2015. *A Package for Survival Analysis in S.* version 2.38. **URL:** *https://CRAN.R-project.org/package=survival*

Thrower, Sharece. 2017. "To Revoke or Not Revoke? The Political Determinants of Executive Order Longevity." *American Journal of Political Science* 61(3):642–656.

Treier, Shawn and Simon Jackman. 2008. "Democracy as a Latent Variable." *American Journal of Political Science* 52(1):201–2017.

Wallace, Jeremy. 2013. "Cities, Redistribution, and Authoritarian Regime Survival." *Journal of Politics* 75(3):632–645.

Wolford, Scott. 2017. "The Problem of Shared Victory: War-Winning Coalitions and Postwar Peace." *Journal of Politics* 79(2):702–716.