A Preanalysis Plan for Replicating Replications: Estimation Methods for the Cox Model[*]

Jeffrey J. Harden[†] Anand E. Sokhey[‡] Hannah Wilson[§]

June 22, 2018

Contents

1 Introduction

This document describes plans for a methodological replication analysis designed to compare the conventional method for estimating the Cox proportional hazards model to an alternative method that is robust to outliers. It follows the evaluation framework described in the paper "Replications in Context: A Framework for Evaluating New Methods in Quantitative Political Science." Below we provide some brief background on the problem and the methods in question. Then we discuss the details of our planned replication analysis.

The Cox proportional hazards model is a popular choice among several alternatives for the

[*] This plan is associated with the paper titled "Replications in Context: A Framework for Evaluating New Methods in Quantitative Political Science."

[†] Associate Professor, Department of Political Science, University of Notre Dame, 2055 Jenkins Nanovic Halls, Notre Dame, IN 46556, jeff.harden@nd.edu.

[‡] Associate Professor, Department of Political Science, University of Colorado Boulder, 333 UCB, Boulder, CO 80309, anand.sokhey@colorado.edu.

[§] Graduate Student, Department of Political Science, University of Notre Dame, 2060 Jenkins Nanovic Halls, Notre Dame, IN 46556, hwilson2@nd.edu.

analysis of duration data, likely due to its inherent flexibility (Box-Steffensmeier and Jones 2004). However, Desmarais and Harden (2012) note that the method is particularly sensitive to deviations from its assumptions, such as measurement error or omitted variable bias. These problems can easily generate outlying event times (observations with large deviations between the actual rank of the duration and the model's expected failure rank), which bias parameter estimates. Bednarski (1993) presents an alternative estimator of the Cox model that downweights outliers with the goal of reducing this bias. Desmarais and Harden (2012) compare this alternative estimator to the conventional partial likelihood estimation method (Cox 1975) and develop a sample-based test for empirically choosing between the two approaches.

Here we plan to compare Bednarski's (1993) estimator and the conventional partial likelihood approach. Desmarais and Harden (2012) conduct this same comparison, reporting results from five replication studies. We plan to collect a large sample of studies from the population (in this case, political science research that employs the Cox model) and employ our evaluation approach to examine the practical utility of the alternative estimation method. We will examine the full distribution of differences between the two methods and conduct inference regarding the null hypothesis that the alternative estimator is no different, on average, from the conventional approach. Additionally, we will assess the frequency with which Desmarais and Harden's (2012) sample-based test selects the alternative estimator. The results will place the replications of Desmarais and Harden (2012) in better context, which is especially important here because they focus their attention on the replications that point to the alternative method as the better choice.

1.1 Summaries of Methods

Here we briefly summarize the two methods we plan to compare in this replication analysis. See the original sources for more details.

- **PLM**. The PLM method is the original approach to Cox model estimation (Cox 1975). It involves maximizing a "partial" likelihood function that is described as such because only the ranks of the durations are used, not the actual event times. This estimator converges to maximum likelihood as the sample size increases if the model's assumptions are valid (see

Desmarais and Harden 2012, 115).

- **IRR**. If one or more of the Cox model's assumptions are *not* valid, the PLM approach will yield sensitivity in its estimates. The iteratively reweighted robust (IRR) estimator is designed to address this problem (Bednarski 1993). Specifically, it modifies the score of the partial likelihood to downweight outliers, thereby reducing their influence on the final estimates. Outliers in this context are defined as observations with large differences between the expected and actual failure ranks; they often result from misspecification problems such as measurement error or omitted variables (Desmarais and Harden 2012, 115–116). The analyst controls the severity of the downweighting by setting a proportion (e.g., 0.05) of the most extreme outliers to obtain weights of zero.

2 The Evaluation Framework

In our evaluation framework the first step is to preregister the replication plan. This step involves (a) defining the relevant population of studies, (b) listing the target sample from this population, and (c) defining replication quantities of interest. We address each of these items below.

2.1 Defining the Population

The relevant population for this replication analysis is any political science study that employs the Cox model. This is a broad definition, covering all empirical subfields of the discipline. For simplicity, we limit our scope to political science, but our definition could cover other fields as well. We do not include studies that *only* employ parametric duration models. While the data in these studies *could* be used to estimate the Cox model, we exclude them because those researchers chose a different methodological strategy. Studies that employ parametric duration models *and* the Cox model are included in the population.

2.2 The Target Sample

The large size of the population of studies defined above necessitates that we draw a sample of studies with which to conduct our replication analysis. This sample should be representative of the population, so ideally we would randomly sample from the entire population. However, a simple random sample is not feasible because we do not have a list of every study that belongs in the population and replication data is not available for all studies. Accordingly, we first selected a set of journals, then randomly sampled studies from those journals. We had two goals in mind when choosing journals: covering a representative collection of empirical political science and selecting journals for which we knew replication data are publicly available.

Specifically, we began with the list of 27 journals that signed the Journal Editors' Transparency Statement (JETS) as part of the discipline-wide Data Access & Research Transparency (DA-RT) initiative in 2014.[1] This choice ensured that we would be able to find studies with available replication data. To obtain representative coverage of the empirical subfields of political science, we chose the four most prominent general interest journals from the JETS list, which also happen to be among the most prominent journals in the discipline: *American Political Science Review* (APSR), *American Journal of Political Science* (AJPS), *Journal of Politics* (JOP), and *British Journal of Political Science*. We also selected three subfield journals: *State Politics & Policy Quarterly* (SPPQ, American politics), *Comparative Political Studies* (CPS, comparative politics), and *Journal of Peace Research* (JPR, international relations).

Next we accessed each journal's public replication data repository and determined the time period for which data were available for all articles published that year (as of June 22, 2018). We report this information in the first two columns of Table 1. To generate the sampling frame, we first searched each journal in Table 1 in the years listed using Google Scholar. We used the search string [Cox proportional hazards model].[2] The results—comprised of a total of 33 studies—are listed in the spreadsheet file accompanying this document. Next we read each article and denoted in the file whether it is relevant (i.e., belongs in the population). To make this determination we examined each article's text for positive evidence of the use of the Cox model for at least one model mentioned in the article.[3] For instance, word searches for "Cox" typically pointed us to phrases

[1] See the statement here: https://www.dartstatement.org/2014-journal-editors-statement-jets. By signing JETS, an editor publicly declared several commitments in support of a transparent research process, including easy access to replication data as a requirement for publication.

[2] We also tried "Cox model" alone and with "Cox proportional hazards model". The search string we chose appeared to be the most effective at minimizing false positive and eliminating false negatives.

[3] The model itself could be reported in the article or an appendix, but we required that it be a reported model

in the text or tables of results presenting Cox model estimates. This process yielded our target sample: 24 relevant studies that we could potentially replicate (see Table 1). We plan to attempt to replicate the entire target sample after depositing this document. We will report the final number we were able to replicate in our results, including explanations for any studies that we failed to replicate or otherwise omitted. We will also re-replicate the five replication studies presented in Desmarais and Harden (2012) for the sake of comparison.[4]

Table 1: Time Periods of Full Data Availability in Journal Replication Data Repositories

Journal	Website	Years	Search Hits	Relevant Studies
APSR	https://dataverse.harvard.edu/ dataverse/the_review	2017–2018	1	1
AJPS	https://dataverse.harvard.edu/ dataverse/ajps	2013–2018	7	4
JOP	https://dataverse.harvard.edu/ dataverse/jop	2015–2018	5	4
BJPS	https://dataverse.harvard.edu/ dataverse/BJPolS	2015–2018	5	2
SPPQ	https://dataverse.unc.edu/ dataverse/sppq	2015–2018	2	1
CPS	http://journals.sagepub.com/ home/cps	2014–2018	2	2
JPR	https://www.prio.org/Data/ Replication-Data/	2000–2018	11	10

Note: Cell entries report replication data information for each journal that we sampled. The search result information is current as of June 22, 2018.

3 Quantities of Interest

We plan to focus on two replication quantities of interest (RQI) in our evaluation of the different methods. Where applicable, we plan to compute these RQI for all coefficient estimates in the original studies' main regression models of interest. In cases where there are multiple regression models presented, we will identify the main regression by determining which one the original

instead of a vague appeal to robustness.

[4] Throughout this replication analysis we will set the IRR truncation parameter—which defines the proportion of observations with weights of zero—to the software default of 0.05. We will accept any other choices (e.g., the procedure for handling tied durations) that the original studies report.

authors refer to the most in their discussion of results. If more than one model appears to meet this criterion, we will use the first one presented in the text. Additionally, we will record which model we choose in our own replication code.

The first RQI we plan to compute is the ratio of the absolute values of the IRR coefficient estimates to the absolute PLM estimates. We refer to this quantity as the coefficient ratio (CR). It is equal to 1 if the IRR and PLM estimates are equivalent, greater than 1 if the IRR estimate is larger in magnitude, and less than 1 if the PLM coefficient is larger in magnitude. This RQI will provide a measure of the magnitude of the difference between the two estimation methods. We will measure CR at the article-coefficient level; there will be a row in our final dataset for every coefficient estimated in a given study.

Desmarais and Harden (2012) propose a sample-based test called the cross-validated median fit (CVMF) test for choosing between the PLM and IRR methods. They show replication examples in which the test chooses each one. However, their replication analysis does not provide much information on how frequently the IRR method is the superior estimator according to the CVMF test in the relevant population of political science studies that employ the Cox model. By replicating a larger, representative sample from this population, our replication analysis can better assess this question. Thus, the second RQI we will compute is the CVMF test's selection (TS): the estimator selected by the test for the model of interest. This quantity will be measured at the article level with a categorical variable that takes on the values PLM, IRR, or neither. We can then compute proportions of each category across the full sample of replicated studies to assess how often IRR is superior to the PLM method.

3.1 Statistical and Substantive Significance

We plan to conduct what our paper refers to as a full inference replication analysis (FIRA). We will use results from our data to make inferences about the differences between the two estimation methods. This process will involve an assessment of statistical and substantive significance. Our paper recommends declaring our criteria for assessing significance ahead of time. We present those criteria here. Regarding statistical significance, we plan to use the conventional $\alpha = 0.05$ (i.e., 95%

confidence level) threshold for all hypothesis tests.

Our paper recommends using Rainey's (2014) approach to assessing substantive significance. Specifically, we must choose a specific value for each RQI, denoted m, that defines the smallest substantively meaningful average RQI. This value represents how different, on average, the existing and new methods must be such that we believe it would be inadvisable for an applied researcher to ignore the new method as a possible empirical strategy. This value will always be arbitrary to some degree, which is why we declare it in this preanalysis plan to ensure that we do not choose the value based on the results.

For our CR measure, we select $m = \pm 0.10$. In other words, we declare a substantively meaningful difference to be one in which the IRR coefficient estimate is 10% larger or 10% smaller than the PLM estimate. Regarding the TS measure, we select $m = 0.25$. This choice means that we will consider the replication analysis to yield substantive evidence in favor of the utility of the IRR method if it is selected in 25% of the replicated studies.

References

Bednarski, Tadeusz. 1993. “Robust Estimation in Cox’s Regression Model.” *Scandinavian Journal of Statistics* 20(3): 213–225.

Box-Steffensmeier, Janet M., and Bradford S. Jones. 2004. *Event History Modeling: A Guide for Social Scientists*. New York: Cambridge University Press.

Cox, David R. 1975. “Partial Likelihood.” *Biometrika* 62(2): 269–276.

Desmarais, Bruce A., and Jeffrey J. Harden. 2012. “Comparing Partial Likelihood and Robust Estimation Methods for the Cox Regression Model.” *Political Analysis* 20(1): 113–135.

Rainey, Carlisle. 2014. “Arguing for a Negligible Effect.” *American Journal of Political Science* 58(4): 1083–1091.