Comparing Partial Likelihood and Robust Estimation Methods for the Cox Regression Model*

August 15, 2011

Abstract

The Cox proportional hazards model is ubiquitous in time-to-event studies of political processes. Plausible deviations from correct specification and operationalization caused by problems such as measurement error or omitted variables can produce substantial bias when the Cox model is estimated by conventional Partial Likelihood Maximization (PLM). One alternative is an iteratively-reweighted robust (IRR) estimator, which can reduce this bias. However, the utility of IRR is limited by the fact that there is currently no method for determining whether PLM or IRR is more appropriate for a particular sample of data. Here we develop and evaluate a novel test for selecting between the two estimators. Then we apply the test to political science data. We demonstrate that PLM and IRR can each be optimal, that our test is effective in choosing between them, and that substantive conclusions can depend on which one is used.

Keywords: Event History Modeling · Cox Proportional Hazards Model · Partial Likelihood Maximization · Iteratively-Reweighted Robust Estimation · Cross-Validation

*Simulation code, the appendix, and R code for implementing the methods proposed here will be made available online if this manuscript is accepted for publication. Replication code of the substantive applications will be made available with permission from the original author(s). Any content in the appendix can be moved to the main text at the discretion of the editors and reviewers.

1 Introduction

The Cox proportional hazards model is a popular method for studying time-to-event data in political science.[1] This popularity stems, in part, from the fact that the Cox model does not require the full specification of the distribution of the durations under study. However, as is the case with many regression routines, Cox regression coefficients are biased under plausible violations of the identifying assumptions, such as contamination from measurement error in the covariates and/or omitted variables (Bednarski 1993). This is problematic because these characteristics are likely to be present to some degree in any empirical application. Researchers are often able to formulate reasonably accurate theoretical models. Yet, even if the theoretical model is accurate, the availability of data and imprecision of measurement tools typically cause operationalizations to fall short of the theoretical model. For example, although central concepts in political science such as ideology or level of democracy are challenging to measure, plausible operationalizations are often, with good reason, included in models of political processes. Furthermore, scientific progress is iterative, with new variables regularly entering into explanations of political phenomena. Thus, it is unrealistic to assume that no important factors will be discovered beyond the current specification of a model. In this paper, we review an implementation of the Cox model that can mitigate the impact of these problems, and propose a sample-based test that researchers can use to determine whether this robust method out-performs the standard approach to estimating the Cox model.

Though specification issues appear in many types of statistical analyses, they are particularly consequential in the Cox model. The innovation offered by Cox (1972) is a regression model which (a) only requires assumptions about the covariate specification to identify the model, and (b) converges, in the sample size, to the maximum likelihood estimator when those assumptions are correct. This means, for instance, that the effect of the number of political parties at the

[1] We use the terms "Cox proportional hazards model," "Cox regression," and "Cox model" interchangeably. A JSTOR search for these terms returns more than 150 studies that use the Cox model in political science journals since 2000, including more than 50 in four of the discipline's leading journals: *American Political Science Review*, *American Journal of Political Science*, *Journal of Politics*, and *International Organization*.

bargaining table on the time taken to form a state's coalition government can be estimated without considering the complicated dynamics in the bargaining process that are *common* among states. However, since all of the identifying information is contained in the covariate specification and measurement, the estimates are extremely sensitive to even slight deviations from the assumed covariate specification (i.e., the assumed model). Contrast this with ordinary least squares, for example, which is unbiased if omitted variables are unrelated to those included in the model. As such, assumptions are of no less importance in the Cox model than in fully parametric models. In the Cox model, the reliance on assumptions manifests completely in the covariate specification rather than being shared between the covariate specification and the distributional assumption.

This sensitivity leads to the question of how to address common specification problems with the Cox model. Generally speaking, either specification and measurement must be improved or the estimation method used must be less sensitive to these problems. In this paper, we focus on an estimation method that is more robust to deviations from the assumptions in the Cox model. We review and extend the iteratively-reweighted robust (IRR) method of estimation developed by Bednarski (1993). IRR is potentially beneficial because it is less biased than the conventional partial likelihood maximization (PLM) method under deviations from the assumed model/measurements. Additionally, it has not yet been adopted in political science. Thus, one goal of our study is to highlight this usefulness to the discipline.

However, beyond simply importing a new method, our main contribution is a novel hypothesis test—the cross-validated median fit (CVMF) test—for determining whether PLM or IRR should be used in a given application of the Cox model. We show below that either method can provide the more accurate estimates (as determined by bias and/or efficiency) in a single sample of data.[2] Because there is no way to determine *a priori* which estimator to choose, researchers have no guidance with respect to which results—PLM or IRR—to trust in drawing substantive inferences. We propose a solution to this problem. Rather than uniformly adopting one or the other, researchers

[2]The properties of any statistical estimator refer to its performance on average, over an infinite number of samples. In a single sample, any estimator can provide more accurate estimates than another. Our test provides within-single-sample guidance in selecting between PLM and IRR.

should use the CVMF test to select the appropriate estimation method for their sample of data.

We begin with an introduction of IRR as an extension of PLM, emphasizing how IRR is less sensitive to specification and measurement errors than PLM. Next, we develop our CVMF test. Then we validate the test with Monte Carlo simulations. We show that, under several types of simulated contamination conditions, the method selected by the CVMF test is the method that produces a coefficient closer to the true parameter. We then apply our test to applications of the Cox model in published research in political science. We illustrate several types of replication results based on the test's selection (PLM or IRR) and the implications of PLM versus IRR for substantive conclusions (less support for the original hypotheses, more support, or mixed results). We find that the test selects IRR in some cases and PLM in others, and that substantive inferences often depend upon the choice between the two estimators. From this, we conclude that both the IRR method and our test can be beneficial to researchers using the Cox model in political science.

2 Robust Estimation of the Cox Proportional Hazards Model

The ubiquitous approach to estimating the parameters of the Cox model is to select the parameters that maximize the partial likelihood. To understand the improvement offered by IRR, it is important to review the mechanics of PLM. Consider a sample of N event times $\boldsymbol{y}$ and indicators of event occurrence $\boldsymbol{\delta}$ with no ties $\{\boldsymbol{y}, \boldsymbol{\delta}\} = \{(y_1, \delta_1), (y_2, \delta_2) \ldots, (y_n, \delta_n)\}$.[3] Let $\boldsymbol{x}_i^{(t)}$ be a vector of covariates for observation i at time t, and $\boldsymbol{R}_t$ be the risk set at time t—the set of observations such that $y_i \geq t \; \forall \; i \in \boldsymbol{R}_t$. The partial likelihood is

$$\prod_{i=1}^{N} \left[\frac{\exp(\boldsymbol{\beta}' \boldsymbol{x}_i^{(y_i)})}{\sum_{\forall j \in \boldsymbol{R}_{y_i}} \exp(\boldsymbol{\beta}' \boldsymbol{x}_j^{(y_i)})} \right]^{\delta_i}, \tag{1}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients. The PLM method finds $\bar{\boldsymbol{\beta}}$ to maximize the partial likelihood. Cox (1975) shows that the PLM converges, in the sample size, to the maximum likeli-

[3]The extensions of IRR to handle tied durations are equivalent to those developed for PLM, which are thoroughly reviewed by Box-Steffensmeier and Zorn (2002). Consequently, we do not discuss them here.

hood estimator—a property that holds when the assumptions of the Cox model are valid (Bednarski 1993).

Nonetheless, many have found that PLM is sensitive to plausible violations of these assumptions, with the most commonly examined being covariate measurement error, omitted covariates, and heterogeneous effects of the covariates (Reid and Crépeau 1985; Arjas 1988; Bednarski 1993; Minder and Bednarski 1996). In the IRR approach, introduced by Bednarski (1993), a new objective function is derived by modifying the partial likelihood to reduce the influence of deviations from the assumptions in the Cox model. IRR contrasts with techniques that explicitly model the deviations from the classic Cox regression model. For example, it is common that either the baseline hazard or the effects of covariates on the hazard vary in a systematic manner, such as over time (Box-Steffensmeier and Zorn 2001), after each repeated event (Box-Steffensmeier and Zorn 2002), over groupings of the observations (Cai, Sen, and Zhou 1999), and across the observations in the study (Longini and Halloran 1996). Researchers often address these issues by explicitly modeling the heterogeneity in the data with "frailty" models (i.e., mixed effects models for event history data, see Box-Steffensmeier and Jones 2004). In the event that this specification of the heterogeneity component of the model is correct and none of the other model assumptions are violated, frailty modeling is likely more efficient than IRR.

However, it is unlikely that a researcher knows the exact nature of the heterogeneity, and so frailty modeling presents another potential specification problem for the researcher. In contrast, IRR addresses the problem without adding additional structure to the model. Rather than explicitly modeling the deviation from the classical assumptions, the IRR method is designed to be less biased under a very broad class of deviations. Most importantly, the current study provides a means of adjudicating between these competing strategies. The test we present below is sufficiently general that it could be used to compare a frailty Cox model to a standard Cox model estimated with IRR.[4]

Measurement error, omitted variables, and functional form misspecification all represent dis-

[4]A fruitful area of future research would be to explore the possibilities and benefits of estimating various heterogeneous effects models with IRR.

tinct challenges to empirical research. Bednarski (1989) identifies a single, cross-cutting manifestation of such problems in the implementation of the Cox model via the conventional PLM. Specifically, he shows that all of these deviations result in disproportionately influential right-tail outliers, or observations that last significantly longer than they are predicted to last. This produces bias in model results, and thus leads to incorrect inferences. Intuitively, these specification issues represent a failure of the model to reflect the real process generating the data. For instance, measuring a covariate with error means the true effect of that covariate on the outcome is not correctly captured in the model. Similarly, an omitted variable means that there is variation in the dependent variable that cannot be adequately explained through the operationalized model. Moreover, since there is no error term or auxiliary parameter (e.g., variance term) in the Cox model, it cannot "account" for observations that, due to real-world complexity in the data generating process, depart from the estimated failure ratios. This can be seen most clearly in predictions from the model that diverge markedly from the actual outcomes, such as a war that, in reality, lasted twice as long as the median war in the sample, but was predicted to be only half as long as the median.[5]

Of course, it is important to note that specification errors cause outliers in both directions. Bednarski (1989) focuses on right-tail outliers because long event times exercise a disproportionate influence on the PLM estimator. Left-tail outliers—those with unexpectedly small values of y_i—need not be considered because they only have a small relative effect on the estimates (Bednarski 1989; Minder and Bednarski 1996). This is due to the use of risk sets (i.e., $\boldsymbol{R}_{y_i}$) in constructing the PLM. Equation 1 shows that observation i contributes to the partial likelihood (1) when observation i fails and (2) whenever an observation fails before i. Suppose observation j fails before i. Then i is said to be in j's "risk set"—the set of observations that are at risk of failing when j fails. Since observations that last longer are in more risk sets and longer-lasting observations contribute more to the risk sets in which they are included, right-tail outliers exercise substantially more influence over the partial likelihood than do left-tail outliers.

[5]This emphasis on differences between reality and prediction are important. Simply observing a large value on the dependent variable is not enough to label that case as an outlier. Indeed, if that case's covariate values produce a prediction of a long duration, then it is not an outlier.

This property of the partial likelihood calculation is illustrated below in Table 1. The hypothetical example gives the actual duration, predicted duration, and the risk sets in which each observation is included. Observations a and e exhibit the same difference between observed and predicted value (6), but, as can be seen in the third column, the right-side outlier e is included in five times as many risk-sets as is a by virtue of e's being a larger duration.

[Insert Table 1 here]

The robustness of IRR comes from trimming the influence of these right-tail outliers (though not necessarily eliminating them completely). The iterative component of IRR derives from the fact that estimates of the regression coefficients are necessary to identify the expected relative durations. The first step with IRR is identifying outlying values, which are defined as those observations with long times-to-event and high hazards of event occurrence. The "outlyingness" of the i^{th} observation increases with either $\exp(\boldsymbol{\beta}'\boldsymbol{x_i})$ or y_i, holding the other constant. This means that, given a certain value of the hazard of event occurrence, a greater outlyingness penalty accrues with each time unit that goes by without the event occurring.

The IRR is derived by directly modifying the score of the partial likelihood with a function $A(y_i, \boldsymbol{x}_i)$ of the covariates and durations. The function must be smooth, bounded, non-negative, and 0 for large values of y and $\boldsymbol{\beta}'\boldsymbol{x}$ (Minder and Bednarski 1996). The score of the partial likelihood, derived from Equation 1, is

$$\sum_{i=1}^{N} \left[\boldsymbol{x}_i - \frac{\sum_{\forall j \in \boldsymbol{R}_{y_i}} \boldsymbol{x}_j \exp(\boldsymbol{\beta}'\boldsymbol{x}_j^{(y_i)})}{\sum_{\forall j \in \boldsymbol{R}_{y_i}} \exp(\boldsymbol{\beta}'\boldsymbol{x}_j^{(y_i)})} \right] \delta_i. \tag{2}$$

The score is equal to $\mathbf{0}$ when the partial likelihood is maximized. The IRR estimates, within a single iteration, are derived by solving the modified score

$$\sum_{i=1}^{N} A(y_i, \boldsymbol{x}_i) \left[\boldsymbol{x}_i - \frac{\sum_{\forall j \in \boldsymbol{R}_{y_i}} A(y_j, \boldsymbol{x}_j) \boldsymbol{x}_j \exp(\boldsymbol{\beta}'\boldsymbol{x}_j^{(y_i)})}{\sum_{\forall j \in \boldsymbol{R}_{y_i}} A(y_j, \boldsymbol{x}_j) \exp(\boldsymbol{\beta}'\boldsymbol{x}_j^{(y_i)})} \right] \delta_i = \mathbf{0}. \tag{3}$$

In Bednarski (1993) and Minder and Bednarski (1996), $A(y_i, \boldsymbol{x}_i) = M - \min[M, y_i \exp(\boldsymbol{\gamma}'\boldsymbol{x}_i)]$,

where M is taken to be a sample percentile (typically between the 80$^{\text{th}}$ and 95$^{\text{th}}$) of $y_i \exp(\boldsymbol{\gamma}'\boldsymbol{x}_i)$.[6] Therefore, each observation's influence on the model estimates is decreasing in the measure of outlyingness, with the largest $(100 - M)\%$ of observations not influencing the estimation.[7] The iterative component of IRR comes in the selection of $\boldsymbol{\gamma}$. In practice, at the first iteration, $\boldsymbol{\gamma}$ is set equal to the PLM, and iteratively updated by setting it equal to the current IRR estimates in the next iteration, cycling through a fixed number of iterations.[8] Minder and Bednarski (1996) find that the IRR estimates stabilize (i.e., $\gamma \approx \hat{\beta}$) after three iterations.[9]

Each iteration of IRR estimates are computed using anywhere from 80% to 95% (i.e., $M\%$) of the sample. This is appropriate if the downweighted observations depart from the generating process of the rest of the data due to measurement or specification errors. However, if these observations are consistent with the proportional hazards process represented by the specification and operationalization of the model, this downweighting reduces the sample size with no apparent benefit. Thus, as Bednarski (1993) shows, *when the assumptions of the Cox model hold, IRR is less efficient than PLM*. For example, in a simulation experiment, Bednarski (1993) shows that the variance of the IRR estimate can be 77% larger than that of the PLM when there are *no* violations of the assumptions in the Cox model. However, he also shows that this disparity can be reversed

[6]Here we use the 95$^{\text{th}}$ percentile. In our applications below, we do not observe substantively meaningful variance in the estimates for $M \in \{.8, .9, .95, .99\}$, but if one were to observe significant differences resulting from M, the CVMF test could be used to arbitrate between estimates with different values of M. Automated, data-driven selection of M would serve as an excellent topic for future research, but we view this as beyond the scope of the current analysis.

[7]Minder and Bednarski (1996) show that $A(y_i, \boldsymbol{x}_i) = M - \min[M, \hat{\Lambda}(y_i)\exp(\boldsymbol{\gamma}'\boldsymbol{x}_i)]$—where $\hat{\Lambda}$ is an estimate of the cumulative hazard function and M is taken to be a percentile of $\hat{\Lambda}(y_i)\exp(\boldsymbol{\gamma}'\boldsymbol{x}_i)$—is more robust to complicated baseline hazard functions. At the time, the computational burden of implementing this more robust version of A was prohibitive. Modern computing capabilities make it feasible, so we use $A(y_i, \boldsymbol{x}_i) = M - \min[M, \hat{\Lambda}(y_i)\exp(\boldsymbol{\gamma}'\boldsymbol{x}_i)]$ throughout the remainder of this paper.

[8]If the analyst preferred, the number of iterations could be allowed to vary, and a stopping rule could be based on a measure of convergence in the estimates. However, Minder and Bednarski (1996) find that this is not necessary. We extended the R code in the `coxrobust` package (Bednarski and Borowicz 2006) to continue iterating until convergence, and were not able to generate conditions under which the estimates in the fourth and third iterations differed.

[9]A derivation of the covariance estimate for the IRR is given in Theorem 4.3 of Bednarski (1993), but in the interest of space, we do not reproduce it here.

when as few as 5% of the observations are contaminated by measurement error. This illustrates that neither method is universally the better choice.

The relative performance of IRR to PLM depends on properties of the sample that are likely unknown. This presents a clear problem in applied research. Given a sample of data and a specification of the model, it is important to determine which estimator more closely characterizes the data generating process of theoretical interest. We introduce the CVMF test to allow researchers to determine which method provides a better fit to the majority of their data. When the PLM provides a better fit to the majority of the observations in the sample, it is clear that a handful of outliers are not driving the PLM, and IRR is inferior. However, when the PLM only fits a minority of the observations better than IRR, this is evidence that the benefits from the downweighting in IRR will be realized. We show below that the CVMF test, on average, selects the estimator that produces the coefficient estimates closer to the true coefficient values.

3 Comparing PLM and IRR Estimates

Bednarski (1993) and Minder and Bednarski (1996) both find that IRR exhibits less bias than PLM in the face of sufficient deviations from the assumptions of the Cox model. However, those studies also show that if the assumptions hold, PLM is more efficient than IRR. Thus, if an analyst's goal is to produce coefficient estimates as close to the true parameter values as possible, neither method can be adopted as strictly optimal. An open problem in an applied setting is to determine whether PLM or IRR provides the estimates that are closest to the truth (whether via less bias, greater efficiency, or both), given a particular sample of data.[10] We address that problem here by providing a measure that can be used to asses the relative fit of the PLM and IRR estimates.

The CVMF test we present can be used for the pairwise comparison of Cox regression estimates derived by PLM, IRR, or any other method. We design the test as an outlier-robust method of determining which estimator provides the better fit to the data, and so it is especially appropriate

[10]Though bias is often seen as the most important problem in choosing an estimator, efficiency can be just as misleading. Given two unbiased estimators, the more efficient is preferable because in a single sample of data an inefficient estimate is the more likely of the two to be further from the true parameter, even if the inefficient estimator averages to the true parameter in repeated samples.

for comparing conventional and robust estimates. Specifically, our test indicates whether there is a significant difference between the median fit to the observations in the sample provided by the estimates of the two methods. We then show later through simulation that the CVMF test is a valid indicator of whether PLM or IRR produces coefficient estimates closer to their true values.

3.1 The CVMF Test

To construct the test, we need (1) a selection framework that is resistant to the influence of outliers, and (2) a common metric on which to compare the IRR and PLM estimates. Yu and Clarke (2011) show that the median loss criterion—the practice of selecting the estimator that offers the maximum median fit—is highly resistant to outliers. Following the second statement of the first definition of the median loss criterion given by Yu and Clarke (2011), we take the IRR estimator $\tilde{\boldsymbol{\beta}}$ to be better than the PLM estimator $\bar{\boldsymbol{\beta}}$ given the true parameter $\boldsymbol{\beta}$, the random variable X from which the sample is drawn, and a fitness function f (i.e., a negative loss function) iff

$$\text{med}_X \left[f(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}) \right] \geq \text{med}_X \left[f(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}) \right], \tag{4}$$

where med_X is the median operator. If f is defined observation-wise, such that it is possible to compute $f_i(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta})$ and $f_i(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta})$ for each $i \in \{1, 2, \ldots, N\}$ observation in the data, then the condition given in Equation 4 can be tested using a paired sign test (Clarke 2007). Let

$$I_i = \begin{cases} 1 & \text{if} \quad f_i(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}) > f_i(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}) \\ 0 & \text{if} \quad f_i(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}) \leq f_i(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}). \end{cases}$$

Then the test statistic for the paired sign test is

$$T = \sum_{i=1}^{N} I_i \tag{5}$$

Given a sample of N conditionally independent observations, T is used to test the null hypothesis

$$H_0 : \text{med}_X \left[f(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}) - f(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}) \right] = 0.$$

This is a nonparametric test in that the distribution of $f(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}) - f(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta})$ need not be known (Clarke 2007). Under H_0, T has a binomial distribution with N trials and probability of success equal to 0.5.[11] Assuming a two-tailed test at significance level α, IRR is selected if T exceeds the 1-$\alpha/2$ quantile of the null binomial distribution, and PLM is selected if T is below the $\alpha/2$ quantile. Values of T within the $100(1-\alpha)\%$ centered confidence interval lend less support for selecting one model over the other.

To complete the derivation of the test, we need to identify f—a measure of fit that can be compared observation-wise using the IRR and PLM estimates. An important feature of the median fit criterion discussed by Yu and Clarke (2011) is that it is less sensitive to misspecification of the objective function than is the expected (i.e., mean) loss criterion. This property of the median fit criterion corresponds exactly to the problem with PLM addressed by IRR. The sum (i.e., $N \times$ mean) log-partial likelihood is the criterion used to select the PLM estimates. Incorrect specification or operationalization results in a misspecification of the partial likelihood, which in turn produces outliers and induces bias in the PLM. The median of the log-partial likelihood, according to the median loss theory of Yu and Clarke (2011), will be less sensitive to the misspecification of the partial likelihood than is the mean log-partial likelihood. Thus, if the IRR estimates improve the median fit by trimming the influence of outliers, then the median log-partial likelihood computed using the IRR estimates will be higher than the median log-partial likelihood computed using the PLM estimates. Equivalently, if outliers are not the determinants of the differences between the

[11] At first glance, it may appear that we are simply stating a justification for the use of the Clarke (2003, 2007) test with the Cox model. However, there are some subtle, yet critical differences. Most importantly, the Clarke test is designed to compare two different models, with different covariate specifications and/or distribution families, fit to the same data. Our test is oriented towards selecting between two different estimators of exactly the same model on the same data. Also, the Clarke test is designed to compare two models estimated by maximum likelihood, whereas our test can be used to compare any two estimation methods.

PLM and IRR, and as a result there is no justification for using IRR, then the median log-partial likelihood of the PLM will not be increased by the downweighting of outliers in IRR.

We take f to be the observation-wise contribution to the log-partial likelihood. However, it is not possible to derive an additive contribution of a single observation to the log-partial likelihood defined in Equation 1. Verweij and Houwelingen (1993) present a cross-validation method for computing an observation's contribution to the log-partial likelihood. We use this measure.[12] An observation's contribution to the cross-validated log-partial likelihood is defined using an iterative, leave-one-out procedure, mechanically similar to the computation of the outlier identification statistic Cook's D (Cook 1977). Let $\hat{\boldsymbol{\beta}}_{-i}$ be the estimate of the regression coefficients computed by either PLM or IRR (i.e., $\bar{\boldsymbol{\beta}}_{-i}$ or $\tilde{\boldsymbol{\beta}}_{-i}$ respectively) when observation i is left out of the sample and $\hat{\boldsymbol{\beta}}$ be the estimate of the regression coefficients when the model is estimated on the full sample. Then the log-partial likelihood contribution of observation i, denoted f_i, is given by

$$f_i(\hat{\boldsymbol{\beta}}) = \sum_{k=1}^{N} \ln\left(\left[\frac{\exp(\hat{\boldsymbol{\beta}}'_{-i}\boldsymbol{x}_k^{(y_k)})}{\sum_{\forall j \in \boldsymbol{R}_{y_k}} \exp(\hat{\boldsymbol{\beta}}'_{-i}\boldsymbol{x}_j^{(y_k)})}\right]^{\delta_k}\right) - \sum_{k\neq i} \ln\left(\left[\frac{\exp(\hat{\boldsymbol{\beta}}'_{-i}\boldsymbol{x}_k^{(y_k)})}{\sum_{\forall j \neq i \in \boldsymbol{R}_{y_k}} \exp(\hat{\boldsymbol{\beta}}'_{-i}\boldsymbol{x}_j^{(y_k)})}\right]^{\delta_k}\right).$$

Thus, the CVMF test statistic is

$$\text{CVMF} = \sum_{i=1}^{N} \mathbf{1}\left[f_i(\tilde{\boldsymbol{\beta}}) > f_i(\bar{\boldsymbol{\beta}})\right],$$

which, under H_0, has a binomial distribution with N trials and probability of success of 0.5.

3.2 Implementing the CVMF Test

Our implementation of the CVMF test builds upon the availability of the PLM estimator in the R package `survival` (Therneau and Lumley 2009) and the availability of the IRR estimator in the R package `coxrobust` (Bednarski and Borowicz 2006). We have programmed the CVMF

[12] We thank a helpful reviewer for noting that one arrives at the PLM by restricting the weighting function in IRR to unity. This suggests it might be possible to derive a hypothesis test oriented towards this restriction. However, we cannot simply apply a likelihood ratio test to compare IRR and PLM, since IRR does not constitute a maximum likelihood estimator.

test in the R language, and made this software available in the supplementary materials on the *Political Analysis* website, as well as on the websites of both authors. Also in the supplementary materials is an appendix that describes the syntax for computing PLM estimates, IRR estimates, and the CVMF test using our software. The appendix starts from the assumption that the reader is familiar with the syntax for estimating the Cox model in Stata (i.e., no familiarity with R), and describes how to move the data into R and estimate an equivalent model with PLM and IRR.

As noted above, leave-one-out cross-validation is used to compute the CVMF statistic. One important practical consideration is that the computation of the CVMF statistic can be time consuming. Given n observations, n sets of IRR estimates and PLM estimates must be computed to calculate the CVMF test statistic. In our experience, computation time has not been prohibitively long—a few minutes for artificial datasets with 1,000 observations—but this could become burdensome if a researcher is working with a dataset with tens or hundreds of thousands of observations.

4 Simulation Study

The CVMF test provides a straightforward way to determine whether the median fit of IRR is better than that of PLM. In this section, we use simulation to assess whether selecting the estimator with the highest median fit corresponds to selecting the estimator which is closest to the true parameter values. Our Monte Carlo study carries two objectives: (1) to demonstrate the differences in performance between PLM and IRR under the contamination conditions mentioned previously, and (2) to evaluate the effectiveness of our test in selecting between PLM and IRR in a given sample. In these simulations we focus on the measurement error, omitted variable, and heterogeneous effects deviations from the correct specification and operationalization. In addition, because the first objective has been addressed in previous simulation studies (e.g., Minder and Bednarski 1996), we present those results in the appendix to conserve space.

4.1 Design

All of the generated times are drawn from an exponential distribution conditional on covariates, with the proportional hazard link function given in Box-Steffensmeier and Jones (1997).[13] In order to observe sample size-based differences in performance, we consider sample sizes of $N = 100$ and $N = 500$. In all of the cases, there is a single covariate in the *estimated* model, with a true parameter value of $\beta = 1$.[14] The covariate included in the model has a standard normal distribution.

We simulate three sets of contamination conditions: measurement error, omitted covariate and heterogeneous effects. In the measurement error condition, we add noise, drawn from a normal distribution with zero mean, to the covariate included in the model. This simulates random measurement error from the use of a covariate that acts as an imperfect "proxy" for a concept. We vary the degree of contamination by repeating with variances in measurement error of 0.1, 0.25, and 0.5. In the omitted covariate condition, we draw both covariates from a bivariate normal distribution. The effect of the omitted covariate is exactly opposite the included one at $\beta = -1$. The variance of the omitted variable is set at 0.5, and its correlation with the included covariate is set to 0, 0.25, and 0.5.[15] Following Longini and Halloran (1996), in the heterogeneous effects condition the effect of the covariate is drawn from a gamma distribution with unit mean. This is synonymous with an effect that varies in intensity throughout the population, but does not change sign. The variance of

[13] We performed the simulation study with a two-parameter Weibull distribution and scale (i.e., duration dependence) parameter set at 0.75, 1, and 1.25, but the results varied negligibly with the scale parameter, so in the interest of space we only report the results with scale $= 1$, which reduces to an exponential distribution.

[14] Minder and Bednarski (1996) consider the case of multiple covariates and show that IRR performs better than PLM under the contamination conditions we consider in our study. The case of a single covariate may seem overly simplistic, but it provides us with a more defensible method of evaluating the performance of our CVMF test. Unlike the typical hypothesis testing case, the two *models* we are comparing are exactly the same, but the estimation methods are different. Thus, we need to identify the *correct* selection on an estimate-by-estimate basis. In the single parameter case, selecting the estimator that gives the estimate closest to the truth is equivalent to making the correct selection. This definition is invariant to symmetric distance metrics, but would depend upon an arbitrary choice among distance measures if multiple covariates were included.

[15] We set the variance of the omitted variable at 0.5 to balance the performance of PLM and IRR. If we equate the omitted variable's variance with the included covariate's, IRR almost always provides the more accurate estimate of the regression coefficient.

the effect is set at 0.75, 1, and 1.25 in the three sub-conditions. We ran each of these 18 cases for 500 iterations.[16]

4.2 Results

The first set of results, presented in the appendix, generally confirms previous simulation work on PLM and IRR (Minder and Bednarski 1996). In particular, under the contamination conditions that we consider, IRR out-performs PLM in terms of bias and mean squared error. However, although contamination is likely to be present in most applied research, these results do not imply that IRR is always better than PLM. In any application the degree of assumption violation is not known. Thus, even though it is reasonable to expect contamination from specification issues, it is *not* advisable to simply use IRR unconditionally to address the problem, because the contamination may not be severe enough to warrant the downweighting of observations. Indeed, in additional simulations (not shown) we found that PLM can still perform better than IRR under very small amounts of contamination. This emphasizes the need for the CVMF test.

With this in mind, the next question to address is whether our CVMF test is effective in choosing correctly between PLM and IRR in the simulations. Here we consider the better estimator to be the one that produces the coefficient estimate closest to the truth. Importantly, the test has a difficult standard to meet in our evaluation. The presence of contamination will render IRR the better estimator *on average* across many samples. However, in a single sample of data—even one drawn from contaminated conditions—it is still possible (though less likely) that PLM could produce a coefficient estimate closer to the true value simply due to chance. Because applied researchers only work with one sample of data, our test must be able to account for the potential peculiarities of a single sample in identifying the better estimator.

Figure 1 presents the test-related results, aggregated over the within-contamination-condition parameter values (for space considerations). The top row (panels a-c) present results with $N = 100$

[16]Specifically, we ran a total of $2N$s $\times$ (3 measurement error parameter values + 3 omitted variable parameter values + 3 heterogeneous effect parameter values) = 18. Multiplying by 500 produces a total of 9,000 iterations. All simulations were run in R (R Development Core Team 2011) with the `survival` (Therneau and Lumley 2009), `coxrobust` (Bednarski and Borowicz 2006), and `mvtnorm` (Genz et al. 2008) packages.

and the bottom row (panels d-f) present N = 500 results. Each point represents a single iteration in the Monte Carlo study. The x-axes in these plots give the difference in the absolute IRR error and the absolute PLM error such that positive values represent instances in which PLM provides an estimate closer to the truth and negative values indicate that IRR is closer to the true value than PLM.[17] The y-axes plot the CVMF test statistic value, with larger values favoring PLM. The areas marked with diagonal lines indicate regions in which the test statistic is statistically significant at the 0.05 level. Points in the white space represent correct selections by our test while points in the gray space represent incorrect selections. The percentage of points occupying each error-difference/test-selection region are given on the plots.

[Insert Figure 1 here]

Note that the test statistic generally points in the correct direction, with the overwhelming majority of points falling in the white regions of the plots. This means that, in most cases, the estimation method that produces the higher median fit value in the CVMF test is also the method that produces a coefficient estimate closest to the true parameter. The test performs quite consistently across all conditions, with approximately 90% of all test statistics pointing in the correct direction when N = 100, and 99% pointing in the correct direction when N = 500. In addition, between 30 and 50% of these correct selections are statistically significant in the N = 100 simulations, and virtually all test statistics point in the correct direction and are significant in the N = 500 simulations.

Most importantly, in cases where the CMVF statistic is significant, the test nearly always points in the correct direction; there are very few instances where the test is incorrect and statistically significant. This good performance is consistent across the three types of contamination that we imposed. Overall, the results show that the CVMF test represents an effective method of identifying the better estimator in a sample-based comparison of PLM and IRR. The next step is assessing the applicability of the test and IRR to data used in political science research.

[17] Absolute error is the absolute difference between the coefficient estimate and the true value.

5 Applications to Time-to-Event Data in Political Science

The Monte Carlo results illustrate the negative consequences of the PLM estimation method when the identifying assumptions of Cox regression do not hold, and that the IRR method can often (but not always) produce coefficient estimates closer to the true parameter under these conditions. Most importantly, the results from the simulations show that our CVMF test is effective in identifying which method produces a better estimate for a given sample. Consequently, our next consideration is the applicability of the test to data in political science. Though we do not know the true parameters in these cases, we can use the CVMF test and infer that the better-fitting method's estimates are likely closer to the true parameters. To highlight the broad applicability of our test in the discipline, we re-analyze examples of research using Cox regression (with PLM) from three subfields: comparative politics, American politics, and international relations.

We show several types of replication outcomes with respect to the test's selection and the implications of IRR versus PLM for substantive conclusions. Table 2 lists six of these types: two possible results of the test (selection of PLM or IRR) and three possible results of the IRR method (less support for the original hypotheses than PLM, more support, or mixed results).[18] We focus on three of these outcomes below. Specifically, we show examples in which (1) the test selects IRR and IRR shows less support for the original hypotheses, (2) the test selects IRR and IRR shows more support, and (3) the test selects IRR and IRR shows mixed results.[19] Two more replications highlighting a fourth outcome—a selection of PLM and IRR showing less support—are presented in the appendix to conserve space.[20] Overall, these replications show that PLM and IRR can each

[18]This is not an exhaustive list of potential outcomes. Regarding the test results, we consider a p-value of 0.05 or less as evidence in favor of a particular estimation method (though in the Box-Steffensmeier et al. [1997] replication the p-value is 0.06). In addition, it is possible for the test to select neither method (see the replication of Hartzell and Hoddie 2003 below). Finally, another possible outcome is that inferences from the PLM and IRR estimates are the same.

[19]Additionally, in cases where the test selects IRR, we identify some observations that IRR completely downweights due to outlyingness as a check on the face validity of the test (see notes 22, 24, and 28). Recall that outlyingness is determined by divergence between the second-to-last IRR iteration hazard-based prediction for an observation's failure time rank and that observation's actual failure time rank.

[20]We were able to replicate each model exactly as reported in the original articles.

be an optimal method in political science, that our CVMF test can be used to choose between them, and that substantive conclusions can depend on which one is used.

[Insert Table 2 here]

5.1 IRR Selected, IRR Less Support

Building upon Diermeier and van Roozendaal (1998), Martin and Vanberg (2003) study coalition bargaining over government formation. They utilize data from 203 bargaining situations in 10 European countries during the post-World War II era. Here we re-analyze the Martin and Vanberg (2003) specification, but assess both sets of theoretical expectations.[21] The dependent variable is the duration, in days, of the coalition bargaining process for each situation. Martin and Vanberg (2003) include the covariates from the Diermeier and van Roozendaal (1998) model as well as their own variables. *Post-Election* is an indicator for whether bargaining began immediately after an election or during the legislative session. The authors expect this variable to produce a negative coefficient (i.e., a reduction in the hazard of coalition formation). They reason that bargaining after an election creates more uncertainty because party leaders know less about one another than they do after repeated interaction. To test their expanded theory, Martin and Vanberg (2003) add *Range of Government*, a measure of ideological distance between the parties in the coalition, a count of the number of parties involved in the process (*Number of Government Parties*), and an interaction between *Number of Government Parties* and the length of time bargaining has transpired (*ln[Time]*). They expect these variables to exert a negative effect on the risk of a bargaining situation concluding.

The authors estimate a Cox model with the PLM method. However, the CVMF test selects the IRR method as the better-fitting estimator at a statistically significant level ($p < 0.05$). Specifically, of the 203 observations in the sample, 67 produce larger PLM cross-validated log-partial likelihood values while 136 produce larger IRR cross-validated log-partial likelihood values.[22] Figure

[21] See the appendix for another extension of this model proposed by Golder (2010).

[22] Observations that are given no weight include Norway in 1963 (prediction: 9$^{\text{th}}$ shortest bargaining time, actual: 42$^{\text{nd}}$ shortest) and Italy in 1987 (prediction: 19$^{\text{th}}$ shortest bargaining time, actual: 44$^{\text{th}}$ shortest).

2 shows the differences between the two estimation techniques. Panel (a) plots standardized coefficients and 95% confidence intervals for the PLM and IRR estimates and panel (b) illustrates the interactive effect of *Number of Government Parties* with *ln(Time)*.

[Insert Figure 2 here]

Beginning with panel (a), notice that the coefficient estimate on *Post-Election* is negative and statistically significant with PLM, providing support for the authors' theory regarding uncertainty. However, the IRR estimate of that coefficient is only about half the magnitude of the PLM estimate and is not statistically significant. Thus, IRR provides no support for the hypothesis that bargaining after an election lengthens the formation process. Next, note that, in line with the Martin and Vanberg (2003) expectation, both methods produce a negative coefficient on *Range of Government*, though neither estimate is statistically significant at the 95% level.

Panel (b) provides a detailed look at the second addition of Martin and Vanberg (2003): the role of coalition size. That graph shows the change in the logged hazard rate for a one party increase in *Number of Government Parties* across the observed range of *ln(Time)*. The gray line shows the PLM estimate and the black line shows the IRR estimate. Recall that the authors expect this effect to be negative. The graph shows that with PLM it is initially positive, but becomes negative after approximately 16 days, a length that a majority (60%) of the cases in the sample reach.[23]

However, this is not the case with IRR. With that method, the effect of *Number of Government Parties* does not become negative until after about 24 days (about 40% of the sample) and is only negative *and* statistically significant at the 95% level after 33 days (30% of the sample). In fact, with the IRR method, the effect of of *Number of Government Parties* is positive and statistically significant (counter to the hypothesis) for a larger portion of the sample (40%) than it is negative and statistically significant (30%). In short, the IRR method, which is selected as the better choice for this model by our test, provides less support for the theoretical expectations outlined by Diermeier and van Roozendaal (1998) and Martin and Vanberg (2003).

[23]The effect is negative and statistically significant with PLM after 25 days (40% of the sample).

5.2 IRR Selected, IRR More Support

Box-Steffensmeier, Arnold, and Zorn (1997) examine the timing of House Members' position-taking on the North American Free Trade Agreement (NAFTA) during the first year of the Clinton presidency. The authors construct a model of position announcement that identifies constituents, organized interests, and policy leaders as key determinants of when a member of Congress takes a stand. The dependent variable in this analysis is the number of days after August 11, 1992 that a House member took either a "yes" or "no" position on NAFTA. The authors include several covariates to explain this outcome that represent constituency, interest group, institutional, and individual factors.

One key set of variables is the proportion of union members in a Congressperson's district and its interaction with *Ideology*, an indicator for conservative members according to 1993 Chamber of Commerce ratings. The authors' expectation of a signaling process predicts that the effect of *Union Membership* is positive for liberal House members. Liberals in highly-unionized districts enjoy agreement between their own preferences and the signal from their constituencies, and, as a result, should take a position earlier. In contrast, liberals representing districts with less of a union presence receive cues that contradict their preferences, leading to a position-taking delay. For conservatives, they expect the opposite relationship: "[t]hose from low-union districts receive a constituency signal consistent with their personal preference, while those from high-union districts find their preference at odds with that of constituents" (Box-Steffensmeier, Arnold, and Zorn 1997, 329). This amounts to the expectation that the effect of *Union Membership* should move in opposite directions for liberals compared to conservatives (i.e., a positive coefficient on *Union Membership* and a negative coefficient on *Ideology* $\times$ *Union Membership*).

The authors estimate a Cox model with the PLM method. However, the CVMF test selects the IRR method as the better-fitting estimator at a statistically significant level ($p = 0.06$).[24] Figure 3 shows the differences between the two estimation techniques. Panel (a) plots standardized coeffi-

[24]Outlying event times that are eliminated with IRR include Jay Kim (R-CA), who is predicted to be the 121st House member to take a position, but actually was the 391st and Rick Santorum (R-PA), who is predicted to be the 212th position-taker, but actually was one of the last (428).

cients and 95% confidence intervals for the PLM and IRR estimates and panel (b) illustrates the interactive effect of *Union Membership* with *Ideology*.[25]

[Insert Figure 3 here]

The coefficient plot in panel (a) shows several key differences between the two estimation techniques. For instance, while the PLM method produces statistically significant coefficients on *Labor Contributions* and *Republican Leadership*, the IRR estimates of those effects are not significant at the 95% level. Moving to the estimates on *Union Membership* and its interaction with *Ideology*, note that both methods produce the hypothesized positive coefficient on *Union Membership* and negative coefficient on *Ideology* $\times$ *Union Membership*, but also that the IRR estimates are larger in magnitude than those of PLM.

Panel (b) illustrates the substantive implications of IRR for this interactive effect. That plot shows the percentage change in the hazard rate for a one standard deviation increase in *Union Membership* for both values of *Ideology*. Consistent with expectations, the PLM estimates (left) show a positive change in the hazard rate for liberals (about 20%) and a negative change for conservatives (-7%). Box-Steffensmeier, Arnold, and Zorn (1997) interpret this as a "meaningful difference in the influence of organized labor…between ideologically opposed members" (332). However, note that the effects are much stronger in magnitude when IRR is used compared to PLM. The IRR effect for liberals is a 26% increase in the hazard rate while the IRR effect for conservatives is -16%.[26] Thus, the better-fitting IRR estimates shows considerably more support for the authors' original expectations than does PLM.

5.3 IRR Selected, IRR Mixed Results

A key area of investigation on civil wars centers on why some peace settlements endure for

[25] Although a proper specification usually requires all components of an interaction term (see Brambor, Clark, and Golder 2006), the authors do not include *Ideology* in the model as a linear term because the Cox model does not estimate an intercept (Box-Steffensmeier, Arnold, and Zorn 1997, 332). Our conclusions are unaffected by this choice and so we maintain the original specification.

[26] In addition, the PLM confidence intervals overlap while the IRR confidence intervals do not. However, a difference of means test shows that the difference between liberals and conservatives is statistically significant with both estimators ($t = 2.04$ for PLM and $t = 2.55$ for IRR).

many years while others collapse, sometimes soon after agreement. Hartzell and Hoddie (2003) describe peace failure as a commitment problem—without safeguards in place to force cooperation, neither side has incentive to uphold an agreement. They expect that the duration of peace between civil war adversaries is primarily affected by the presence of "power-sharing institutions," or stipulations in a peace agreement requiring that power be shared by competing groups in the transitional government. Mattes and Savun (2010) add to this theoretical framework in identifying another key element of peace agreement bargaining: informational asymmetries. They contend that the two sides in conflict have incentives to withhold private information regarding their power and resolve to fight, even as a peace agreement is signed. This uncertainty, in turn, leads to a lack of trust, which can cause a breakdown of peace.

The two sets of authors analyze data from domestic conflicts as defined by the Correlates of War project ending with a negotiated settlement or truce. Hartzell and Hoddie (2003) examine 38 cases from 1945–1998 and Mattes and Savun (2010) extend the data to 2005 (51 cases). The dependent variable is the number of months that peace endured following the signing of an agreement. We focus on the independent variables capturing the authors' central hypotheses.[27] Specifically, *Power-Sharing Provisions* is a count of the number of dimensions (political, territorial, military, and economic) on which an agreement requires power sharing, *Third Party Guarantees* is a binary indicator for whether a third party (another state or an international organization) agreed to enforce the peace, and *Uncertainty-Reducing Provisions* is a count of the number of provisions included in the agreement designed to encourage information transparency between the two sides (see Mattes and Savun 2010, 518). Both sets of authors expect each variable to exert a negative effect on the risk of peace failure, though the third variable (*Uncertainty-Reducing Provisions*) is only included in the Mattes and Savun (2010) model.

Both sets of authors show support for their expectations with a Cox model estimated by the PLM method. In the first (Hartzell and Hoddie 2003), the CVMF test selects neither PLM nor IRR as the better-fitting estimator at a statistically significant level ($p = 0.56$). In other words, the

[27] We use the variable names given by Mattes and Savun (2010).

null hypothesis that the two methods provide equal fit to the data cannot be rejected. However, in the Mattes and Savun (2010) model the test selects IRR over PLM ($p < 0.05$).[28] As in the last examples, this holds consequences for substantive conclusions. Figure 4 plots standardized coefficients and 95% confidence intervals for the PLM and IRR estimates of the Hartzell and Hoddie (2003) model (panel a) and Mattes and Savun (2010) model (panel b).

[Insert Figure 4 here]

Taken together, the original PLM results show support for both expectations. *Power-Sharing Provisions* and *Third Party Guarantees* produce negative coefficients in each model, and those in Hartzell and Hoddie (2003) reach statistical significance at the 95% level. Though the confidence intervals are slightly wider in panel (b), Mattes and Savun (2010) also label *Power-Sharing Provisions* and *Third Party Guarantees* as significant in their own model with one-tailed hypothesis tests ($p < 0.10$). Moving to the next hypothesis, the coefficient on *Uncertainty-Reducing Provisions* in panel (b) is also negative as expected and significant at the 95% level. Thus, both accounts are supported in these data. Hartzell and Hoddie (2003) find that power sharing is important to maintaining peace and Mattes and Savun (2010) demonstrate that reducing uncertainty is also crucial for agreement success.

However, the IRR results suggest a more one-sided conclusion in favor of Mattes and Savun (2010). Notice that, with IRR, the main coefficients of interest remain signed in the expected direction, but magnitude and significance change in both panels. In the Hartzell and Hoddie (2003) model *Power-Sharing Provisions* and *Third Party Guarantees* produce coefficients of slightly different magnitude compared to the PLM estimates and confidence intervals that are wider, though both remain significant at the 90% level. The larger difference appears in the Mattes and Savun (2010) model. In that case, both *Power-Sharing Provisions* and *Third Party Guarantees* drop in

[28]Recall from the simulation results that the CVMF test is better at discriminating between the two methods as the sample size increases. In this particular case, the outlying observations in the Mattes and Savun (2010) model that are completely downweighted include two new additions not in Hartzell and Hoddie (2003): Sudan (1983–2002) and Angola (1998–2001). The Mattes and Savun (2010) PLM model (51 cases) predicts these two observations to be the 5th and 7th shortest peace times, respectively, when they actually were the 36th and 17th shortest peace times.

magnitude from the PLM results (and considerably so for the latter). Furthermore, neither effect is significant at any conventional level with IRR ($p = 0.39$ and $p = 0.61$, respectively).

The biggest difference, however, is shown in the IRR estimate for *Uncertainty-Reducing Provisions*. The coefficient on that variable substantially increases in magnitude between the PLM and IRR estimates and is significant at the 95% level in each case. While Mattes and Savun (2010) report that an increase from no provisions to one provision decreases the hazard of peace failure by 46%, these results suggest the effect is actually a decrease of 65%. In other words, their original PLM result considerably *underestimates* the impact of *Uncertainty-Reducing Provisions*. Overall, this replication shows mixed results. While the work of Hartzell and Hoddie (2003) and Mattes and Savun (2010) suggests that variables reducing both the commitment problem *and* uncertainty contribute to the durability of civil war peace, we only find support for the role of the latter when using the CVMF test and the IRR method that it selects.

6 Conclusions

Event history models are an essential part of the empirical political scientist's toolkit, and chief among these estimators is the Cox proportional hazards model. Though comprehensive in the distributions it can accommodate, the standard PLM approach to estimating the Cox model is sensitive to deviations from its assumptions regarding the regression function (i.e., correct specification and perfect measurement). We show here that the IRR method can be more robust to these deviations than PLM. Specifically, IRR reduces the influence of outlying event times that are generated by several issues that originate in the specification of the regression function, such as measurement error or omitted covariates. However, PLM often provides more accurate estimates than IRR when the model/data do not depart markedly from the identifying assumptions.

Most importantly, until now applied researchers have had no method of assessing which of these two methods is best in a given sample of data. To this end, we introduce a sample-based test that can be used to determine whether PLM or IRR provides better fit to the data at hand. Our CVMF test substantially increases the utility of both methods in applied research. Rather than blindly choosing one or the other, researchers can use the test to make a decision between PLM

and IRR that is firmly grounded in statistical theory.

Our replication analyses reveal that the use of the test holds the potential to significantly enhance inference in applications of the Cox model. In some cases the conventional PLM method is the appropriate choice according to our test. Thus, as shown in the appendix, we support the original findings of Martin (2004) and Golder (2010). However, the CVMF test can be used as statistical justification for using PLM in these instances rather than simply adhering to the default settings of statistical software. In other instances IRR fits the data better, but this does not necessarily imply reduced support for researchers' hypotheses. In fact, IRR may actually produce stronger support, as in our replication of Box-Steffensmeier, Arnold, and Zorn (1997). Thus, our test and IRR are not simply a "robustness check," but a useful set of tools for understanding political phenomena.

The possibility of poorly-predicted outlying event times is a concern for any researcher, because issues such as measurement error or omitted covariates are likely present to some degree. For example, the incremental progress of science constantly produces new variables to consider, such as the addition by Mattes and Savun (2010) of *Uncertainty-Reducing Provisions* to the model of civil war settlement. Thus, it is unrealistic to assume there are no variables omitted from a given model. Additionally, given the difficult task of measuring such elusive concepts as *Ideology* (Box-Steffensmeier, Arnold, and Zorn 1997) and *Level of Democracy* (Hartzell and Hoddie 2003; Mattes and Savun 2010), empirical applications in political science are subject to some level of measurement error. Researchers using the Cox model to study a process where (1) there are open questions about the variables to include in the model and in what functional form or (2) some variables are measured with error or not at all should consider both the PLM and IRR estimators in analyzing their data. In doing so, our CVMF test provides a simple approach to making statistically justified decisions regarding which method best minimizes the risks to inference, leading to more accurate substantive conclusions.

References

Arjas, Elja. 1988. "A Graphical Method for Assessing Goodness of Fit in Cox's Proportional Hazards Model." *Journal of the American Statistical Association* 83(401): 204–212.

Bednarski, Tadeusz. 1989. "On Sensitivity of Cox's Estimator." *Statistics and Decisions* 7(3): 215–228.

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

Bednarski, Tadeusz, and Filip Borowicz. 2006. `coxrobust`*: Robust Estimation in Cox Model*.

Box-Steffensmeier, Janet M., and Bradford S. Jones. 1997. "Time is of the Essence: Event History Models in Political Science." *American Journal of Political Science* 41(4): 1414–1461.

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

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

Box-Steffensmeier, Janet M., and Christopher Zorn. 2002. "Duration Models for Repeated Events." *Journal of Politics* 64(4): 1069–1094.

Box-Steffensmeier, Janet M., Laura W. Arnold, and Christopher J. W. Zorn. 1997. "The Strategic Timing of Position Taking in Congress: A Study of the North American Free Trade Agreement." *American Political Science Review* 91(2): 324–338.

Brambor, Thomas, William Roberts Clark, and Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14(1): 63–82.

Cai, Jianwen, Pranab K. Sen, and Haibo Zhou. 1999. "A Random Effects Model for Multivariate Failure Time Data from Multicenter Clinical Trials." *Biometrics* 55(1): 182–189.

Clarke, Kevin A. 2003. "Nonparametric Model Discrimination in International Relations." *Journal of Conflict Resolution* 47(1): 72–93.

Clarke, Kevin A. 2007. "A Simple Distribution-Free Test for Nonnested Hypotheses." *Political Analysis* 15(3): 347–363.

Cook, R. Dennis. 1977. "Detection of Influential Observation in Linear Regression." *Technometrics* 19(1): 15–18.

Cox, David R. 1972. "Regression Models and Life-Tables." *Journal of the Royal Statistical Society. Series B (Methodological)* 34(2): 187–220.

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

Diermeier, Daniel, and Peter van Roozendaal. 1998. "The Duration of Cabinet Formation Processes in Western Multi-Party Democracies." *British Journal of Political Science* 28(4): 609–626.

Genz, Alan, Frank Bretz, Torsten Hothorn, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, and Fabian Scheipl. 2008. `mvtnorm`*: Multivariate Normal and t Distributions*.

Golder, Sona N. 2010. "Bargaining Delays in the Government Formation Process." *Comparative Political Studies* 43(1): 3–32.

Hartzell, Caroline, and Matthew Hoddie. 2003. "Institutionalizing Peace: Power Sharing and

Post-Civil War Conflict Management." *American Journal of Political Science* 47(2): 318–332.

Longini, Ira M., and M. Elizabeth Halloran. 1996. "A Frailty Mixture Model for Estimating Vaccine Efficacy." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 45(2): 165–173.

Martin, Lanny W. 2004. "The Government Agenda in Parliamentary Democracies." *American Journal of Political Science* 48(3): 445–461.

Martin, Lanny W., and Georg Vanberg. 2003. "Wasting Time? The Impact of Ideology and Size on Delay in Coalition Formation." *British Journal of Political Science* 33(2): 323–344.

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.

Minder, Christopher E., and Tadeusz Bednarski. 1996. "A Robust Method for Proportional Hazards Regression." *Statistics in Medicine* 15(10): 1033–1047.

R Development Core Team. 2011. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reid, Nancy, and Helene Crépeau. 1985. "Influence Functions for Proportional Hazards Regression." *Biometrika* 72(1): 1–9.

Therneau, Terry, and Thomas Lumley. 2009. `survival`*: Survival Analysis, Including Penalised Likelihood.*

Verweij, Pierre J. M., and Hans C. Van Houwelingen. 1993. "Cross-Validation in Survival Analysis." *Statistics in Medicine* 12(24): 2305–2314.

Yu, Chi Wai, and Bertrand Clarke. 2011. "Median Loss Decision Theory." *Journal of Statistical Planning and Inference* 141(2): 611–623.

Table 1: Hypothetical Time-to-Event Data and Predictions

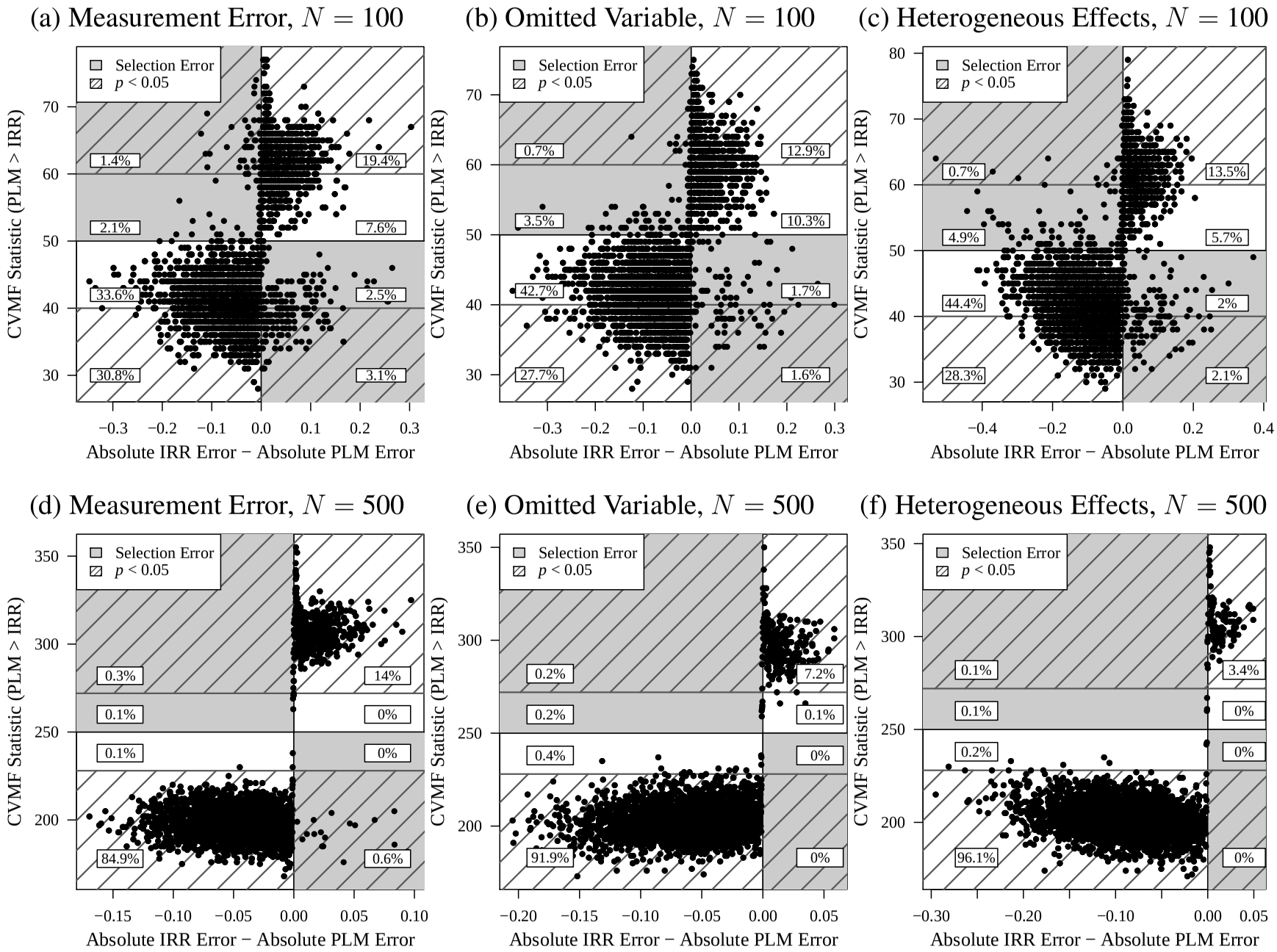
Observation	a	b	c	d	e	f
Actual Duration	1	4	3	2	7	4.5
Predicted Duration	7	5	6	3	1	2
Risk Set	$\{a, b, c, d, e, f\}$	$\{b, e, f\}$	$\{b, c, e, f\}$	$\{b, c, d, e, f\}$	$\{e\}$	$\{e, f\}$
In Risk Sets	$\{a\}$	$\{a, b, c, d\}$	$\{a, c, d\}$	$\{a, d\}$	$\{a, b, c, d, e, f\}$	$\{a, b, c, d, f\}$

Note: Columns report six hypothetical observations in a time-to-event data set. Rows give the observation's actual duration, predicted duration, risk set, and the risk sets in which each observation is included. Notice that observations a *and* e *exhibit the same difference between observed and predicted value (6), but, as can be seen in the third column, the right-side outlier* e *is included in five times as many risk-sets as is* a *by virtue of* e*'s being a large duration.*

Table 2: Types of Replication Examples

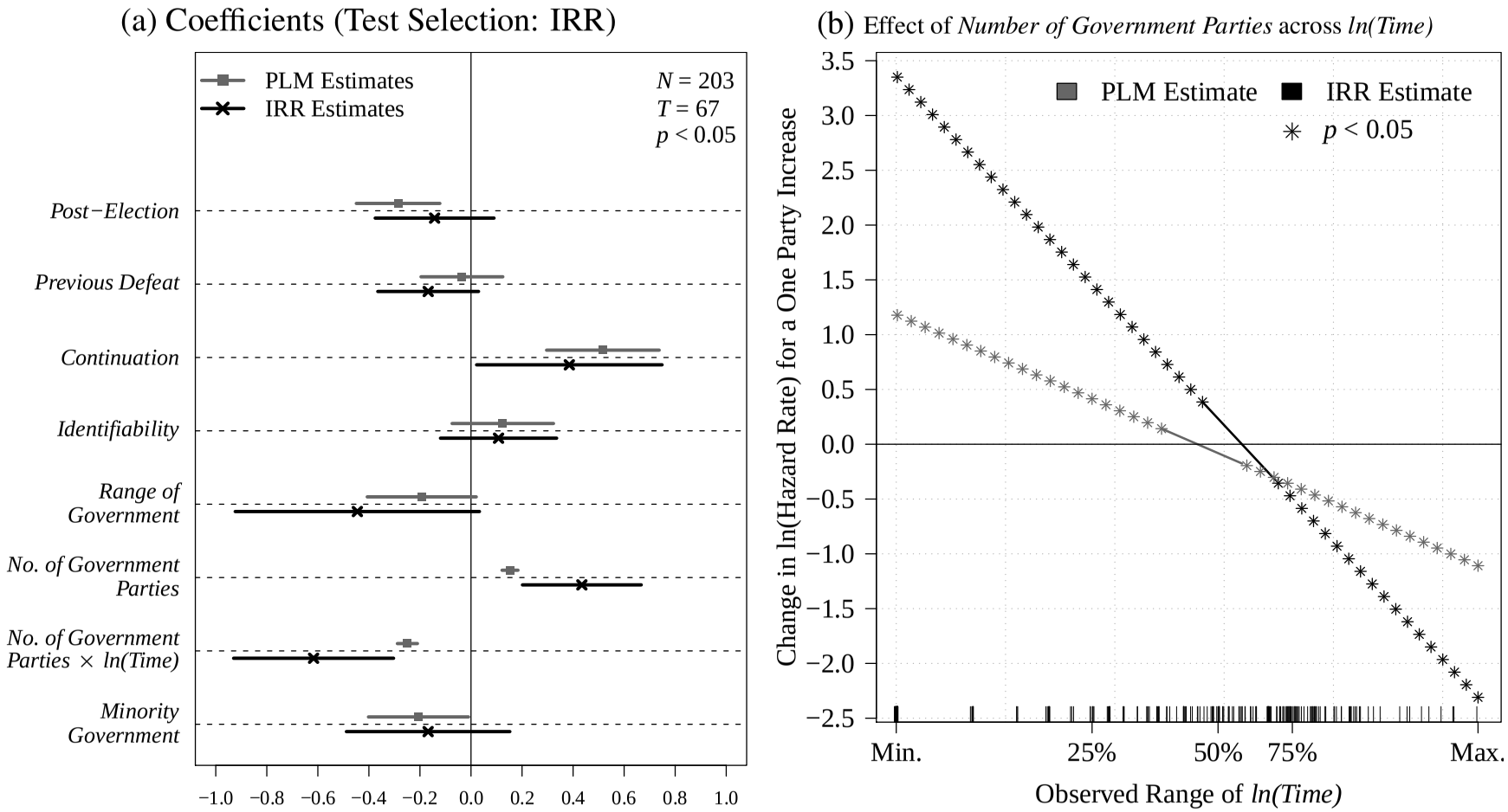
	PLM Selected	IRR Selected
IRR Less Support	Martin (2004), Golder (2010)	Diermeier and van Roozendaal (1998)/Martin and Vanberg (2003)
IRR More Support	*Not Shown*	Box-Steffensmeier, Arnold, and Zorn (1997)
IRR Mixed Results	*Not Shown*	Hartzell and Hoddie (2003)/Mattes and Savun (2010)

Note: Cell entries report classification of each replication example by the selection of the CVMF test (columns) and the implications of IRR versus PLM for the original authors' main hypotheses (rows). As shown below, the CVMF test selects neither method in the Hartzell and Hoddie (2003) model. The replication of Martin (2004) and Golder (2010) are presented in the appendix.



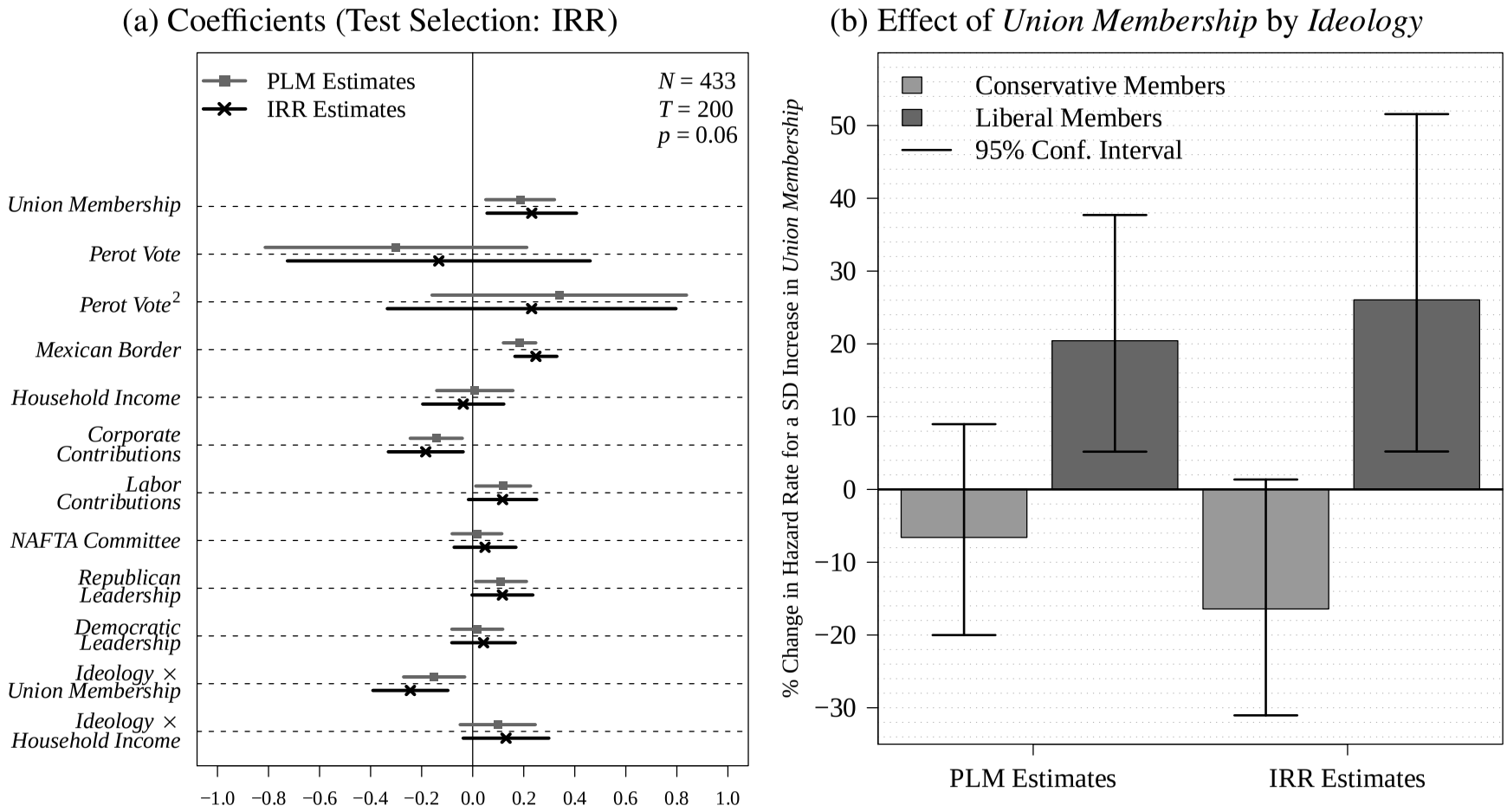
Note: The graphs present the test results for each contamination condition at N = *100 (panels a-c) and* N = *500 (panels d-f). The x-axes give the difference between the absolute IRR and PLM errors (the absolute error being the absolute difference between the estimate and true value); positive values indicate PLM produced an estimate closer to the truth while negative values favor IRR. The y-axes plot the CVMF test statistic value (*T*), with larger values favoring PLM. The diagonal lines indicate regions in which* T *is statistically significant (*p < *0.05), and the gray-shaded areas denote selection errors (e.g., choosing PLM when IRR is actually better). The percentage of points occupying each region are also given. Overall, these plots show that the CVMF test performs well, with over 90% correct selections when* N = *100 and 99% when* N = *500.*

Figure 1: CVMF Test Performance in the Monte Carlo Simulations



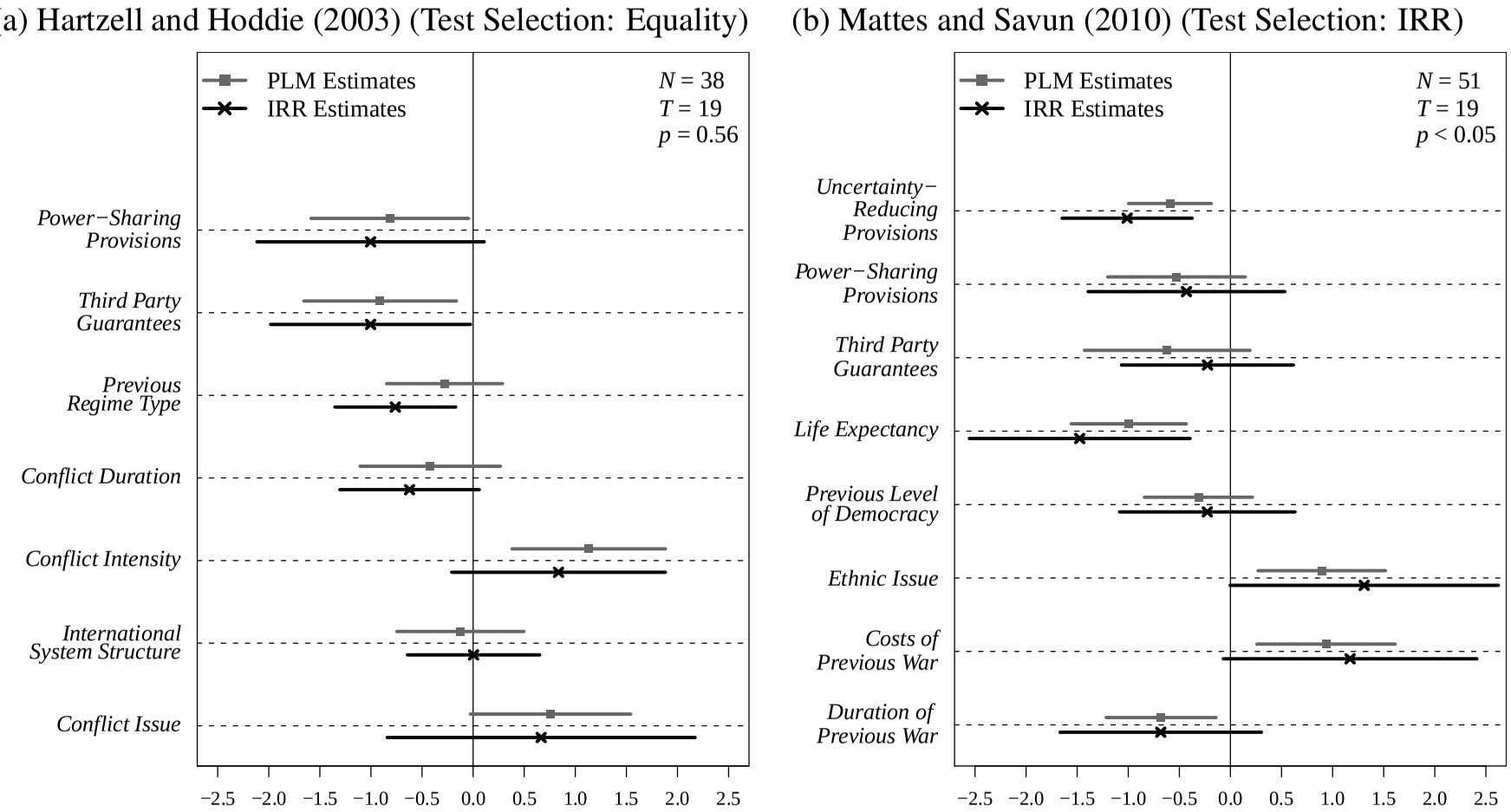
Note: The graph in panel (a) reports standardized coefficient estimates and 95% confidence intervals from the PLM and IRR methods. T *is a count of the number of observations for which the PLM cross-validated log-partial likelihood value is greater than that of IRR. In this case, the PLM values are greater for 67 of 203 observations, which corresponds to a selection of IRR as the better-fitting method* ($p < 0.05$). *Notice that, in line with the theoretical expectations of Diermeier and van Roozendaal (1998), the coefficient on* Post-Election *is negative and statistically significant with the PLM method, but drops in magnitude and becomes nonsignificant with IRR. Panel (b) plots the PLM (gray line) and IRR (black line) estimates of the change in the logged hazard rate for an increase of one party in the coalition across the range of* ln(Time). *Note here that the effect is negative (as expected by Martin and Vanberg 2003) for a majority of the sample (60%) with PLM, but only 40% of the sample with IRR.*

Figure 2: Re-analysis of Effects of Size and Ideology on Bargaining Duration (Martin and Vanberg 2003, Table 2)



Note: The graph in panel (a) reports standardized coefficient estimates and 95% confidence intervals from the PLM and IRR methods. T *is a count of the number of observations for which the PLM cross-validated log-partial likelihood value is greater than that of IRR. In this case, the PLM values are greater for 200 of 433 observations, which corresponds to a selection of IRR as the better-fitting method (*p = *0.06). Notice that both methods produce the expected positive coefficient on* Union Membership *and negative coefficient on* Ideology × Union Membership*, but that the IRR estimates are considerably larger in magnitude than those of PLM. Panel (b) plots the PLM (left) and IRR (right) estimates of the percentage change in the hazard rate for a standard deviation increase in* Union Membership *for both values of* Ideology*. Note that the hypothesized difference between liberals (dark gray) and conservatives (light gray) is larger in magnitude with the IRR estimates than it is with those of PLM.*

Figure 3: Re-analysis of Factors Influencing the Timing of Position Taking by U.S. House Members on NAFTA (Box-Steffensmeier, Arnold, and Zorn 1997, Table 2)



Note: The graphs report standardized coefficient estimates and 95% confidence intervals from the PLM and IRR methods for the Hartzell and Hoddie (2003) model (panel a) and Mattes and Savun (2010) model (panel b). T *is a count of the number of observations for which the PLM cross-validated log-partial likelihood value is greater than that of IRR. In the Hartzell and Hoddie (2003) model the PLM values are greater for 19 of 38 observations, which corresponds to a failure to reject the null that PLM and IRR fit equally well. In the Mattes and Savun (2010) model the PLM values are greater for 19 of 51 observations, which corresponds to a selection of IRR as the better-fitting method (*$p < 0.05$*). Notice that, with IRR,* Power-Sharing Provisions *and* Third Party Guarantees *are negative (as expected), but with much wider confidence intervals in both models. Furthermore, the expected negative effect of* Uncertainty-Reducing Provisions *in the Mattes and Savun (2010) model increases in magnitude and remains significant with IRR.*

Figure 4: Re-analysis of the Stability of Negotiated Civil War Settlements (Hartzell and Hoddie 2003, Table 2; Mattes and Savun 2010, Table 2)

Appendix to "Comparing Partial Likelihood and Robust Estimation Methods for the Cox Regression Model"

Contents

1 Implementing the CVMF Test in R

The IRR estimator and CVMF test are currently only available in R. However, Stata users can follow these steps to use the methods in R. First, for a Cox model estimated in Stata with outcome variable `y` and predictors `x1` and `x2`, save the data in a .dta file:

```
stset y, failure(fail)
stcox x1 x2
save "example.dta"
```

Then, in R, the data can be imported into R as follows:

```
library(foreign)
example <- read.dta("example.dta")
```

Next, read in the accompanying script file "CVMF.R" to load the `CVMF()` function, which estimates the model with PLM and IRR and performs the CVMF test. Note that doing this requires

the `survival` and `coxrobust` packages. Finally, write out the model in R's syntax and feed it into the `CVMF()` function. Here we assign this to an object called `results`. The argument `trunc` is the IRR truncation parameter. The default is 0.95.

```
source(CVMF.R)
form <- Surv(y, event = fail) ~ x1 + x2
results <- CVMF(formula = form, data = example, trunc = .95)
```

1.1 An Example

The file "CVMF.R" contains a basic example of how to use the `CVMF()` function. First, it begins by setting the seed, then creating two independent variables, one of which contains measurement error (`x2e`).

```
## Set the seed for replication purposes
 set.seed(12345)
#
# Create two covariates with measurement error in the second
 x1 <- rnorm(100)
 x2 <- rnorm(100)
 x2e <- x2 + rnorm(100, 0, 0.5)

## Create the dependent variable with the unobserved x2
## Each coefficient has a true value of 1
 y <- rexp(100, exp(x1 + x2))
 y <- Surv(y)
#
## Put the observed variables into a data frame
 dat <- data.frame(y, x1, x2e)
#
## Define the formula
 form <- y ~ x1 + x2e
```

Next, the CVMF test is performed and the results are stored in the object `results`. The selection of the CVMF test is automatically written out on the screen.

```
results <- CVMF(formula = form, data = dat)
IRR supported with a two-sided p-value of 0.021
```

You can also look at the PLM, IRR, and CVMF results in detail by using the dollar sign ($) after the name of the results object.[1] Notice that in this case both of the IRR coefficient estimates are closer to the truth (1) than are the PLM estimates, but the PLM standard errors are smaller than the IRR standard errors.

```
## Take a look at results
results$irr

Call:
coxr(formula = formula, data = data, trunc = trunc)

Partial likelihood estimator
     coef exp(coef) se(coef)        p
x1  0.925      2.52    0.126 2.43e-13
x2e 0.784      2.19    0.122 1.53e-10

Wald test=69.1 on 2 df, p=9.99e-16

Robust estimator
     coef exp(coef) se(coef)        p
x1  0.964      2.62    0.226 1.95e-05
x2e 0.834      2.30    0.193 1.52e-05

Extended Wald test=21.9 on 2 df, p=1.73e-05

## Now the test
 results$cvmf

        Exact binomial test

data:  sum(Cvll.r > Cvll.c) and n
number of successes = 62, number of trials = 100, p-value = 0.02098
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5174607 0.7152325
sample estimates:
probability of success
                  0.62
```

[1]Note that the IRR output automatically reports both PLM and IRR results.

2 Additional Monte Carlo Results

Here we present the first aspect of the Monte Carlo study: a comparison of the relative performance of PLM and IRR under the contamination conditions imposed. These results are presented in Figure A.1. The first row (panels a-c) presents the absolute value of the bias in the PLM estimates divided by the absolute value of the bias with IRR. The second row (panels d-f) gives the mean squared error (MSE) of the PLM estimates divided by the IRR MSE.[2] In all six graphs, values greater than 1 indicate that IRR performance is better (smaller absolute bias or smaller MSE) than that of PLM.

[Insert Figure A.1 here]

Note first that when measurement error is introduced to the covariate by adding noise (panels a and d), IRR outperforms PLM in terms of bias and MSE in all cases. The bias of PLM ranges from 110% to 130% of that of IRR and decreases as the variance of the measurement error increases. Sample size has little effect on the relative bias, but the MSE of PLM is higher than IRR MSE in all cases, and especially with the larger sample size (ranging from 105% to 140% of IRR MSE). In all cases with an omitted variable (panels b and e), the bias and MSE of IRR are lower than those of PLM, though the two methods' performances become similar as the correlation between the omitted and included variables increases. In addition, the relative MSE of PLM to IRR is greater when the sample size is 500.

Specification error introduced via heterogeneous effects (panels c and f) presents a similar result. The absolute bias and MSE of PLM are always higher than the bias and MSE of IRR. Also, there does not appear to be a relationship between the variance of the effect and the relative performance of PLM to IRR. Overall, Figure A.1 supports the claims and findings of earlier work on this topic, showing that considerable gains in estimator performance are possible with IRR when the assumptions underlying the Cox model are violated through specification problems.

[2]Recall that the true parameter value, β, is set to 1. Bias is computed as $\bar{\hat{\beta}} - 1$, where $\bar{\hat{\beta}}$ is the average estimate of the regression coefficient over the 500 iterations from the respective case in the Monte Carlo study, and the mean squared error is computed as $\frac{1}{500}\sum_{i=1}^{500}\left(\hat{\beta}_i - 1\right)^2$.

3 Additional Replications: PLM Selected, IRR Less Support

Here we describe two additional replications. In both cases, the CVMF test selects PLM and IRR shows less support for the original hypotheses.

3.1 Martin (2004)

Martin (2004) examines how government coalitions organize the legislative agenda through an analysis of the sequence and timing of government bills introduced to the legislature. Martin tests his expectations on an original dataset covering government bills introduced in four parliamentary democracies. Specifically, the dependent variable is the number of days between government formation and the introduction of a bill to the legislature in Belgium, Germany, Luxembourg, and Netherlands during the 1980s and 1990s. One main independent variable of interest is a measure of each bill's saliency to the coalition (*Government Issue Saliency*), which is derived from the expert survey data of Laver and Hunt (1992). His theory predicts that, for a bill of average divisiveness, *Government Issue Saliency* shortens the time-to-introduction (a positive effect on the hazard rate), but also that this effect declines over time. Martin reasons that as the end of the parliamentary term draws closer, "it is less likely that *any* type of legislation will be able to make it all the way through the legislative process, 'attractive' or otherwise." (2004, 455, emphasis in original). Accordingly, *Government Issue Saliency* is interacted with the log of the time remaining in the parliamentary term (*ln[CIEP]*).

Martin estimates a Cox model with the PLM method, which, according to the CVMF test, is the better-fitting estimator at a statistically significant level ($p < 0.05$). Thus, we confirm the findings he reports. Nonetheless, we present the differences between the two techniques to highlight the test's usefulness. Panel (a) of Figure A.2 plots standardized coefficients and 95% confidence intervals from the PLM and IRR estimates and panel (b) illustrates the changing effect of *Government Issue Saliency* across the time remaining in the parliamentary term (*ln[CIEP]*).[3]

[Insert Figure A.2 here]

[3]Policy area and country fixed effects are also included in the specification, but not shown.

The coefficient plot in panel (a) only shows the effects of the key variables at one point in time (no time remaining in the term), and so it is primarily useful for gauging an initial sense of the variance associated with each estimate. In this regard, note that the PLM confidence intervals are much smaller than those of IRR for all six variables shown. This is consequential for hypothesis testing at the 95% level on *Opposition Issue Divisiveness*, for which the PLM estimate is significant but the IRR estimate is not. Moving to the effect of saliency over time, note that both methods produce a significant negative coefficient on *Government Issue Saliency*, but a positive coefficient on *Government Issue Saliency* $\times$ *ln(CIEP)*, suggesting that the effect of saliency becomes positive when there is more time remaining in the term.

This effect is shown in detail in panel (b). That plot shows the percentage change in the hazard rate for a one standard deviation increase in *Government Issue Saliency* for a bill of average divisiveness across the range of time left in the parliamentary term. Consistent with expectations, the PLM estimate (gray line) shows a positive change in the hazard rate when there is between 1,400 and about 600 days left, and a negative effect after 600 days (though between about 800 and 400 days the effect is not significant). This shows support for the theory: when there is sufficient time remaining, an increase in saliency to the coalition partners increases the chance of a bill being introduced, all else equal. But as the term draws to a close, the effect weakens, and eventually becomes significantly negative, supporting "the idea that coalition members are less concerned about introducing important legislation late in the parliamentary term, when it is less likely that any bill will make its way through the entire policymaking process" (Martin 2004, 457).

In contrast, note that with the IRR estimates, the effect is weaker, is only positive for the first 400 days, and is never positive and statistically significant. In short, with IRR, there is much less support for the expectation. This illustrates the usefulness of the CVMF test. As stated previously, the test selects PLM at a statistically significant level, which provides evidence in favor of using that method instead of IRR. Thus, a more complete analysis of these data would result in the same conclusions, but with stronger justification for the chosen modeling technique.

3.2 Golder (2010)

Golder (2010) picks up where the work of Diermeier and van Roozendaal (1998) and Martin and Vanberg (2003) leaves off in the study of Western European government formation duration. Rather than focusing exclusively on uncertainty (Diermeier and van Roozendaal 1998) or on the additive effects of uncertainty and complexity (Martin and Vanberg 2003), she considers the possibility of an interactive effect between the two. Using a new data set that includes 17 democracies from Western Europe from 1944–1998, Golder (2010) hypothesizes that bargaining complexity "should lead to increasing delays in the government formation process as uncertainty increases" (12). As in the other two studies, Golder measures uncertainty with an indicator for whether bargaining occurred *Postelection* (high uncertainty) or during the interelection period (low uncertainty) and measures complexity as the effective number of *Legislative Parties* and *Ideological Polarization* between the parties.

Like the previous studies, Golder (2010) estimates a Cox model with the PLM method. The dependent variable is the number of bargaining days to government formation. According to the CVMF test, PLM is the better-performing estimator at a statistically significant level ($p < 0.05$). Thus, as in the Martin (2004) example we confirm the findings Golder (2010) reports. Note that this is a change from the Martin and Vanberg (2003) model, in which IRR was chosen. This suggests that the interactive effect included Golder (2010) may represent a crucial omitted variable in the Diermeier and van Roozendaal (1998) and Martin and Vanberg (2003) versions of the model, which led to the bias-reducing selection of IRR. Once that variable is included, however, the Golder (2010) model can benefit from the added efficiency of PLM.

We again present the differences between the two techniques to highlight the CVMF test's usefulness. Panel (a) of Figure A.3 plots standardized coefficients and 95% confidence intervals from the PLM and IRR estimates and panel (b) illustrates the interactive effect of *Legislative Parties* with the level of uncertainty.

[Insert Figure A.3 here]

Note first from panel (a) that Golder's main theoretical expectation is supported with the PLM

results. Specifically, the coefficients on *Legislative Parties* $\times$ *Postelection* and *Ideological Polarization* $\times$ *Postelection* are negative and statistically significant at the 95% level. All else equal, the effect of increasing complexity (more parties or more ideological distance between parties) exerts a stronger negative effect on the hazard rate of government formation during a postelection period (more uncertainty) than during an interelection period (less uncertainty). However, notice that those same coefficients are both nonsignificant and smaller in magnitude when estimated with IRR. In that case, the effects of *Legislative Parties* and *Ideological Polarization* during a postelection period cannot be statistically distinguished from the effects during an interelection period.

Panel (b) illustrates this in more detail. That plot shows the percentage change in the hazard rate for a one standard deviation increase in *Legislative Parties* for both levels of uncertainty. Consistent with expectations, the negative effect on the hazard rate is stronger in magnitude during the postelection period than during the interelection period. In addition, the confidence intervals show that the difference between those two periods is just on the edge of statistical significance with PLM (t = 1.95). In contrast, that difference is not statistically significant when IRR is used (t = 1.19). In short, this is another example of divergence between PLM and IRR. However, recall that the CVMF test selects PLM at a statistically significant level, which provides evidence in favor of using PLM instead of IRR. Thus, as in the Martin (2004) model, a more complete analysis of these data would ultimately result in the same conclusions, but with stronger justification for the chosen modeling technique.

References

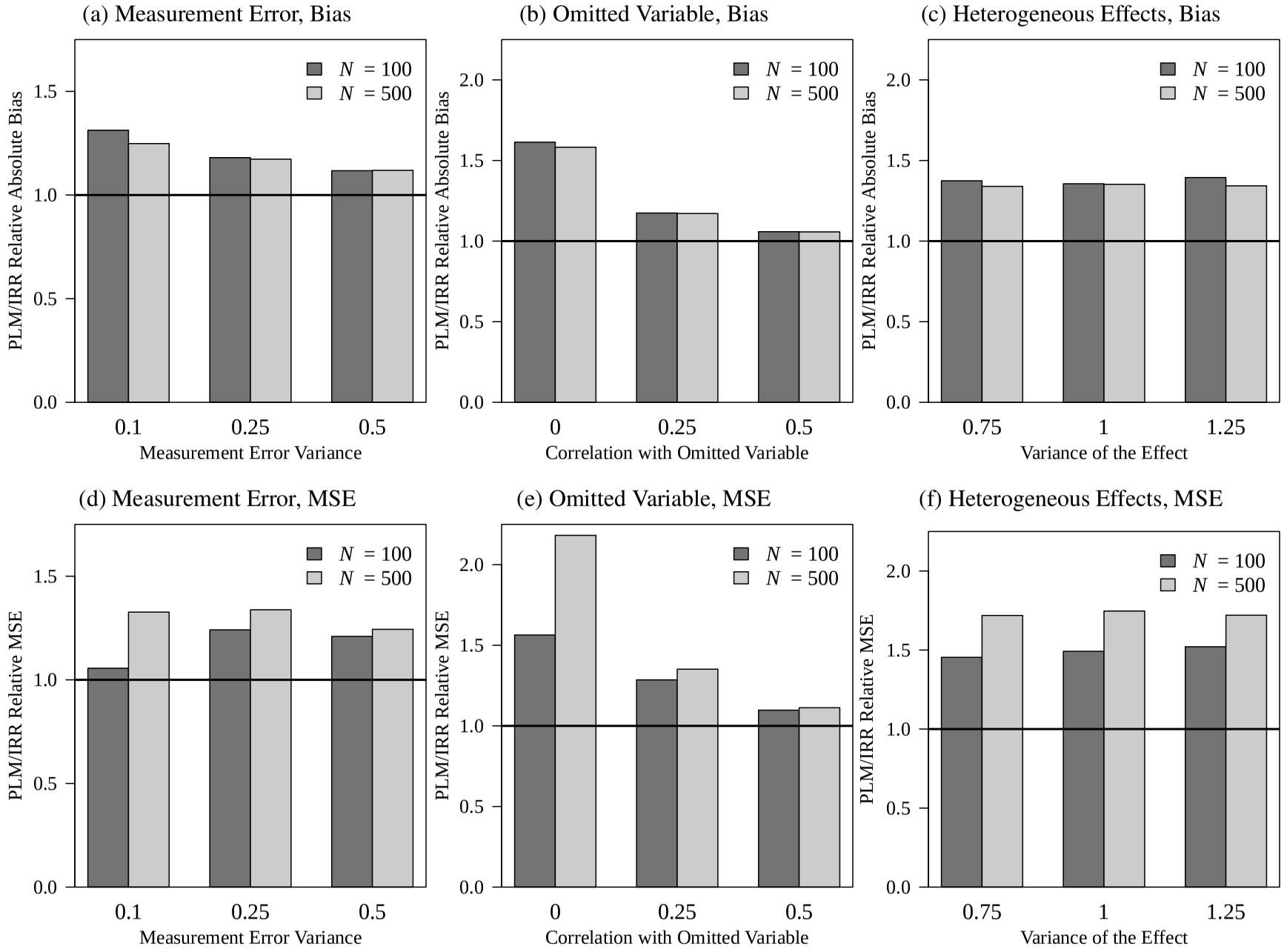
Diermeier, Daniel, and Peter van Roozendaal. 1998. "The Duration of Cabinet Formation Processes in Western Multi-Party Democracies." *British Journal of Political Science* 28(4): 609–626.

Golder, Sona N. 2010. "Bargaining Delays in the Government Formation Process." *Comparative Political Studies* 43(1): 3–32.

Laver, Michael, and W. Ben Hunt. 1992. *Policy and Party Competition*. Ann Arbor, MI: University of Michigan Press.

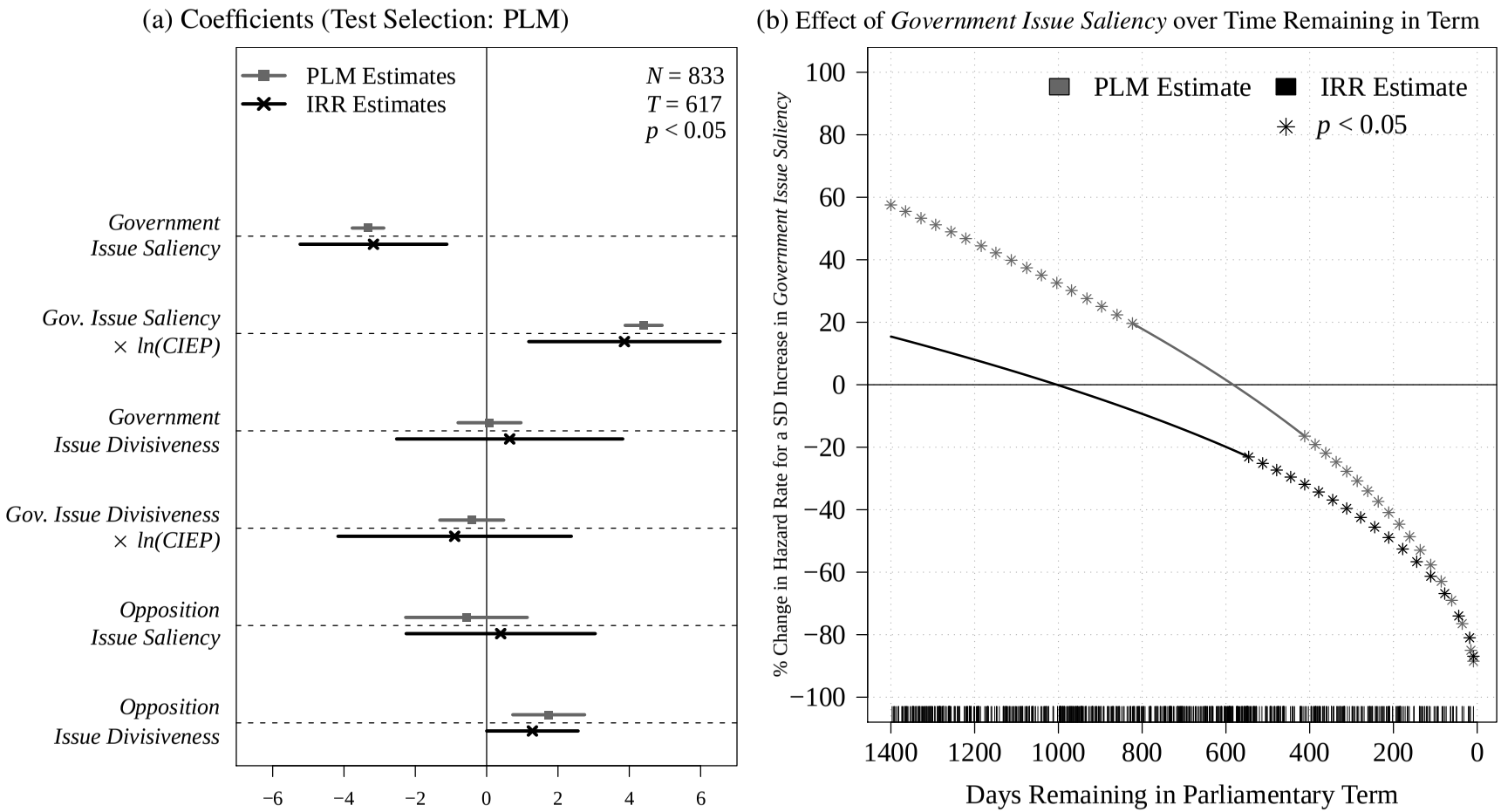
Martin, Lanny W. 2004. "The Government Agenda in Parliamentary Democracies." *American Journal of Political Science* 48(3): 445–461.

Martin, Lanny W., and Georg Vanberg. 2003. "Wasting Time? The Impact of Ideology and Size on Delay in Coalition Formation." *British Journal of Political Science* 33(2): 323–344.



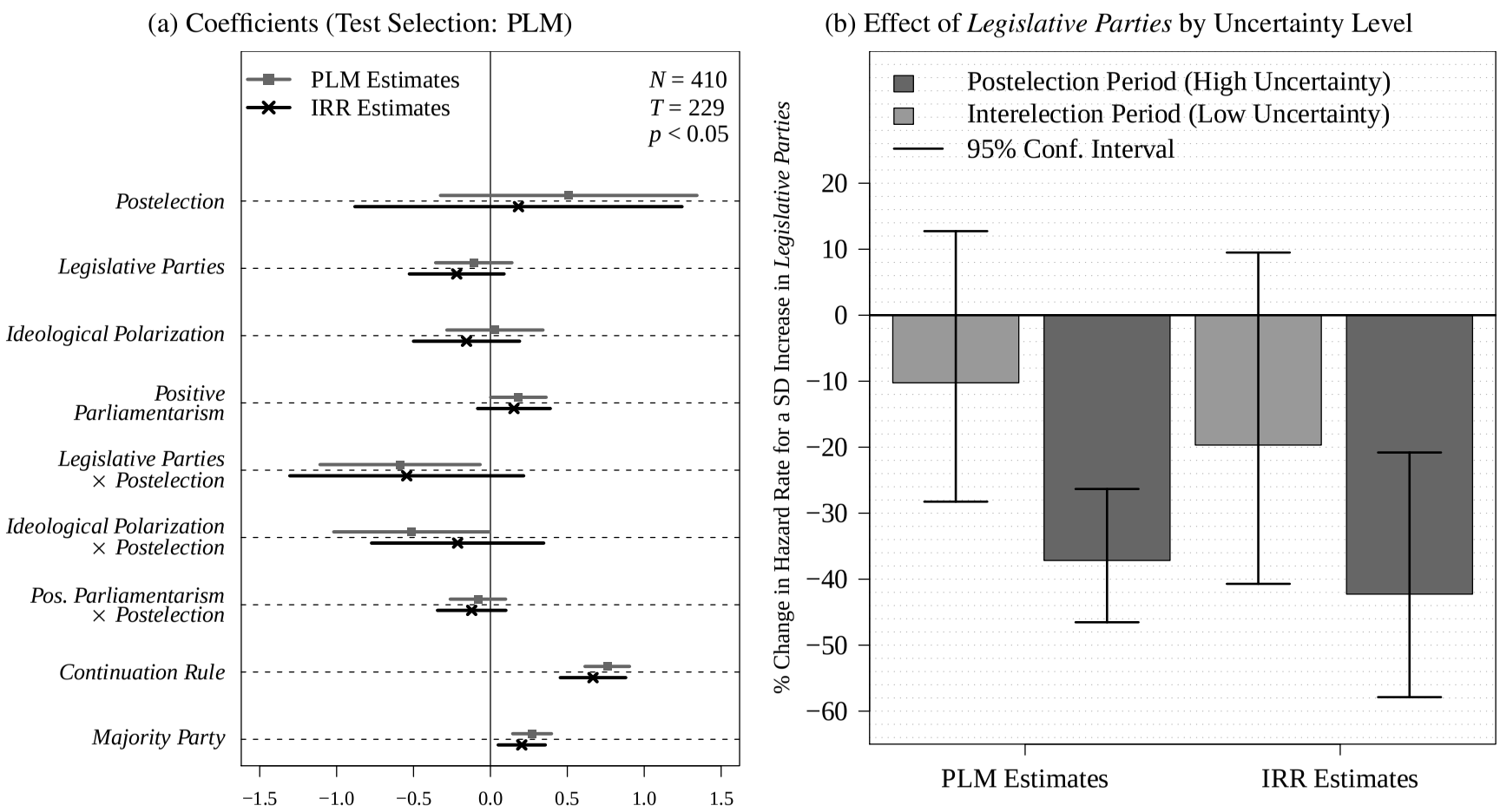
Note: The graphs in panels (a)-(c) present the absolute value of the bias in the PLM estimates divided by the absolute value of the bias with IRR. The graphs in panels (d)-(f) present the mean squared error (MSE) of the PLM estimates divided by the IRR MSE. In all six graphs, values greater than one indicate that IRR performance is better (smaller absolute bias or smaller MSE) than that of PLM. Overall, these graphs show that considerable gains in estimator performance are possible with IRR when the assumptions underlying the Cox model are violated through specification problems.

Figure A.1: Relative Absolute Bias and Relative MSE of PLM and IRR in the Monte Carlo Simulations



Note: The graph in panel (a) reports standardized coefficient estimates and 95% confidence intervals from the PLM and IRR methods. T *is a count of the number of observations for which the PLM cross-validated log-partial-likelihood value is greater than that of IRR. In this case, the PLM values are greater for 617 of 833 observations, which corresponds to a selection of PLM as the better-fitting method (*p < 0.05*). Notice that the confidence intervals are considerably wider when the IRR method is used. This results in a statistically significant estimate for* Opposition Issue Divisiveness *with PLM, but not IRR. Panel (b) plots the percentage change in the hazard rate for a standard deviation increase in* Government Issue Saliency *across the length of the parliamentary term. Note that, as Martin (2004) expects, the effect is positive and statistically significant early in the term with the PLM estimate (gray line), then declines over time. However, this expectation is* not *supported with IRR because the effect is weaker in magnitude and never positive and significant with the IRR estimate (black line).*

Figure A.2: Re-analysis of the Timing of Government Bills (Martin 2004, Table 1)



Note: The graph in panel (a) reports standardized coefficient estimates and 95% confidence intervals from the PLM and IRR methods. T *is a count of the number of observations for which the PLM cross-validated log-partial-likelihood value is greater than that of IRR. In this case, the PLM values are greater for 229 of 410 observations, which corresponds to a selection of PLM as the better-fitting method* ($p < 0.05$). *Notice that, in line with the theoretical expectations of Golder (2010), the coefficients on* Legislative Parties × Postelection *and* Ideological Polarization × Postelection *are negative and statistically significant with the PLM method, but drop in magnitude and become nonsignificant with IRR. Panel (b) plots the PLM (left) and IRR (right) estimates of the percentage change in the hazard rate for a standard deviation increase in* Legislative Parties *for both levels of uncertainty: postelection period (dark gray) and interelection period (light gray). Note that, as Golder (2010) expects, the negative effect on the hazard rate is stronger in magnitude during the postelection period than during the interelection period. In addition, the confidence intervals show that the difference between those two periods is just on the edge of statistical significance with PLM ($t = 1.95$). In contrast, that difference is not statistically significant when IRR is used ($t = 1.19$).*

Figure A.3: Re-analysis of Determinants of the Duration of Government Bargaining Delays (Golder 2010, Table 2)