Advance Access publication September 4, 2015

Political Analysis (2015) 23:534–549
doi:10.1093/pan/mpv019

Estimating Onsets of Binary Events in Panel Data

Liam F. McGrath

Centre for Comparative and International Studies (CIS) and Institute for Environmental Decisions (IED), ETH Zürich, Switzerland
e-mail: liam.mcgrath@ir.gess.ethz.ch (corresponding author)

Edited by Patrick Brandt

Onsets of binary events are often of interest to political scientists, whether they be regime changes, the occurrence of civil war, or the signing of bilateral agreements, to name a few. Often researchers transform the binary event outcome of interest, by setting ongoing years to zero, to create a variable which measures the onset of the event. While this may seem an intuitive way to go about estimating models where onset is the outcome of interest, it results in two problems that can affect substantive inferences. First, it creates two qualitatively different meanings for a unit time period to have a zero, which estimators are unable to "know." Second, it ignores the possibility that variables may have differing effects upon binary event onsets and durations. This article explores how much this transformation can harm our substantive inferences by analytically demonstrating the resulting bias and the use of Monte Carlo experiments, as well as offering recommendations to avoid these problems. I also conduct a sensitivity analysis on the determinants of civil war onset to examine how substantive inferences are affected by this issue. In doing so, I find that there is considerable difference in the size of estimated coefficients and whether a variable is considered a robust determinant of civil war.

1 Introduction

Many researchers often are interested in the onset of binary event outcomes in political science. These interests span many fields, such as: What determines the onset of civil wars? Under which conditions do countries experience regime change? When do countries decide to sign preferential trade agreements? All of these, and many more, questions deal with the occurrence of a binary outcome, and often use time-series cross-sectional data in order to get leverage upon the answers.

With the widespread use of this form of data, there have also been a variety of ways at which researchers attempt to get at the determinants of onsets. Typically this comes in two forms: researchers either set ongoing years of the event to zero or instead set ongoing years to missing. Table 1 highlights that the most common approach is to set ongoing years to zero.[1] While this may seem like a fairly innocent decision, the choice of transformation has direct consequences upon the reliability of the estimation. Importantly, setting ongoing years to zero contains two features that can be problematic for applied research. First, the ongoing years of the event receive the same value as years when the event does not occur. However, the estimator of choice does not

Author's note: Thanks to Janina Beiser, Kevin Clarke, Patrick Kuhn, Thomas Plümper, Curtis Signorino, Janne Tukiainen, Robert Walker, Julian Wucherpfennig, and Christopher Zorn, the Editor, and the anonymous reviewers for comments and suggestions. Replication materials are available at http://dx.doi.org/10.7910/DVN/DOSJCD. Supplementary materials for this article are available on the *Political Analysis* Web site.

[1] List compiled through searching for articles with the keyword "onset" in the Political Science and International Relations category of the Social Sciences Citation Index, as of August 2013. From this, the first two hundred records were downloaded, dating back to 2005. Of these two hundred, sixty-five of the papers had empirical sections that analyzed the onset of a binary event. I was not able to code the form of the onset variable in twelve cases, due to a lack of description of the variable or replication data. The list of papers, and coding of their analyses, is included in the Supplementary Materials.

© The Author 2015. Published by Oxford University Press on behalf of the Society for Political Methodology.
All rights reserved. For Permissions, please email: journals.permissions@oup.com

https://doi.org/10.1093/pan/mpv019 Published online by Cambridge University Press

Table 1 Recoding binary outcomes in the literature

Onset coding	*Number of papers*
Ongoing set to missing	19 (36%)
Ongoing set to zero	34 (64%)
with lag of binary outcome	*12 (35%)*
without lag of binary outcome	*22 (65%)*
Coding explicitly discussed	36 (55%)

"know" this information.[2] Second, this transformation implicitly assumes that the effect of independent variables upon onset and continuation of binary events is identical. As a result, approximately 65% of papers published examining onsets of binary events potentially face issues with regard to the reliability of the estimates.

This is not to say that all researchers are unaware of the potential pitfalls of setting ongoing years to zero. For example, Hegre and Sambanis (2006) (henceforth referred to as HS) in an extreme bounds analysis of the onset of civil wars, where they set ongoing years to missing, state: "The alternative way is to code periods of ongoing war as 0s (except the year of onset), but countries with ongoing wars may have a systematically different risk of a new war, and we would need to control for that as well as to consider the effects of the ongoing war on the other explanatory variables." However, not all existing work is so upfront with the choice of outcome transformation in estimation. Often it is not even explicitly discussed how the onset variable is generated, with only 55% of articles doing so. This leaves other researchers having to explore replication data to find out that ongoing years have been set to zero.

In this article, I show how transforming a binary dependent variable in this way leads to biased results and poor confidence interval coverage. I also demonstrate that these problems are avoided by following one or a combination of two strategies. First, researchers should set ongoing years to missing, or use the untransformed dependent variable while estimating a first-order Markov transition model. This avoids the two problems of a value of zero in the outcome variable having two different meanings, as well as accounting for differential effects for onsets and durations. Second, if researchers continue to use a dependent variable with ongoing years set to zero, then they must at minimum include a one-period lag of the untransformed variable. Doing so solves the problem of a value of zero in the outcome variable having two different meanings. However, within this specification, it is still possible and desirable to explore possible interactive effects with the one-period lag of the untransformed dependent variable. Implementing these approaches would result in more reliable inferences for the approximately 42% of papers in the literature that follow neither approach.

The article proceeds as follows. In the next section, I discuss the potential problems in setting ongoing event years to zero, by analytically demonstrating the bias this results in. In the third section, I run Monte Carlo experiments to examine how sensitive results are to setting ongoing years to zero, under the conditions where regressors have differing effects on onsets and continuation of binary events. The results show that setting ongoing years to zero results in substantial bias, which is avoided with the use of a first-order Markov model or including a one-period lag of the untransformed variable. In the fourth section, I then replicate a sensitivity analysis of the determinants of the onset of civil wars by HS. While this analysis does not suffer from the problems discussed in this article, as ongoing years are dropped, replicating the analysis by additionally estimating the models with ongoing years set to zero provides a sense of the distribution of how inferences are affected by this transformation. For four of the ten variables that are statistically significant at conventional levels, the classification of statistical significance is dependent upon whether ongoing years are set to zero or not. In addition, there can be large differences in the

[2]As will be discussed in the next section, this is typically not a problem if a one-period lag of the untransformed binary outcome is included in the estimation equation, as in Fearon and Laitin (2003). However, only approximately 35% of articles that transform ongoing years to zero do this.

https://doi.org/10.1093/pan/mpv019 Published online by Cambridge University Press

estimated coefficients due to the choice of setting ongoing years to zero. The final section summarizes these insights and offers guidance for avoiding these issues.

2 Estimating Onsets

A common technique to estimate the effect of variables upon a binary event onset is to transform the binary outcome variable y into a new "onset" variable, here denoted as y'. The transformation takes the form

$$y'_{it} = \begin{cases} 0 & \text{if } y_{it-1} = 1 \\ y_{it} & \text{if } y_{it-1} = 0. \end{cases} \tag{1}$$

This transformation results in the years after the initial onset of the binary event being set to zero, so long as the binary event is still ongoing. Doing so results in a variable that can be considered to measure the onset of the event of interest. Table 2 illustrates the use of this transformation for a unit experiencing binary events over time.

2.1 *What's in a Number?*

To illustrate the problems with this approach, I analytically examine the degree of bias.[3] To do so, I assume a general data-generating process in the form of a first-order Markov transition model (Jackman 2000; Przeworski et al. 2000; Beck et al. 2001; Przeworski and Vreeland 2002):[4,5]

$$y^T_{i,t} = 1\{\alpha + \beta x_{i,t} + \delta y_{i,t-1} + \gamma y^T_{i,t-1} x_{i,t} + \epsilon_{i,t} \geq 0\}. \tag{2}$$

For the simplest case, I start with the assumption that there is no state dependence in the form of the state you are in affecting the baseline probability, $\delta = 0$, and in the form of the independent variable having differing effects given the state inhabited, $\gamma = 0$. This reduces the data-generating process to standard binary dependent variable estimation. Defining the true value of the dependent variable as y^T, the data-generating process is

$$y^T_{i,t} = 1\{\beta x_{i,t} + \epsilon_{i,t} \geq 0\}, \tag{3}$$

where the constant term α is dropped for ease of notation.

Let y' refer to the transformed version of the variable previously discussed, where researchers set continuing years of the event to zero. Due to this transformation, the new data-generating process is

$$y'_{i,t} = \begin{cases} 1\{\beta x_{i,t} + \epsilon_{i,t} \geq 0\} & \text{if } y^T_{i,t-1} = 0 \\ 1\{-\beta x_{i,t} - \epsilon_{i,t} \geq 0\} & \text{if } y^T_{i,t-1} = 1 \quad \& \quad y^T_{i,t} = 1. \end{cases} \tag{4}$$

Given this expression, the true data-generating process can be rewritten in the following latent variable form:

$$y'^*_{i,t} = (1 - (y^T_{i,t-1} y^T_{i,t}))(\beta x_{i,t} + \epsilon_{i,t}) + (y^T_{i,t-1} y^T_{i,t})(-\beta x_{i,t} - \epsilon_{i,t}) \tag{5}$$

$$= \underbrace{\beta x_{i,t} + \epsilon_{i,t}}_{\text{Correctly specified}} \underbrace{-2 y^T_{i,t-1} y^T_{i,t} \beta x_{i,t} - 2 y^T_{i,t-1} y^T_{i,t} \epsilon_{i,t}}_{\text{Omitted variable}}. \tag{6}$$

[3] This discussion of the bias draws heavily on work by Meyer and Mittag (2013) on misclassification (in general) with binary dependent variables.

[4] Jackman (2000) notes that this model, while not commonplace in political science research, is frequently used in other disciplines. Examples are in bio-statistics (Diggle, Liang, and Zeger 1994), applied econometrics (Boskin and Nold 1975; Bane and Ellwood 1986; Barmby 1998), and sociology (Yamaguchi 1991).

[5] The expression contained within the curly braces is evaluated as a logical expression. This allows for a more compact expression of the standard definition of a binary data-generating process, where $y = 1$ if $x_i\beta + \epsilon \geq 0$ and $y = 0$ if $x_i\beta + \epsilon < 0$.

https://doi.org/10.1093/pan/mpv019 Published online by Cambridge University Press

Table 2 Binary event transformation for a given time-series

Unit	Time	y	y'
1	1	0	0
1	2	1	1
1	3	1	0
1	4	1	0
1	5	0	0
1	6	1	1
1	7	0	0
1	8	0	0

Note that the nature of the transformation results in a form of omitted-variable bias, as long as all that is entered into the estimation is $x_{i,t}$. To understand the extent of this bias, we focus on the expression that includes $x_{i,t}$. First, write this omitted variable in the form of a linear projection on $x_{i,t}$:

$$2y^T_{i,t-1}y^T_{i,t}\beta x_{i,t} \;=\; \lambda x_{i,t} + \nu_{i,t}. \tag{7}$$

We then substitute this expression into (6), resulting in

$$y'^{*}_{i,t} \;=\; \underbrace{(\beta - \lambda)}_{\text{Biased coefficient}}\; x_{i,t} + \underbrace{\epsilon_{i,t} - \nu_{i,t}}_{\text{Misspecified error term.}} \tag{8}$$

From this expression, we can sign the bias of the parameter associated with $x_{i,t}$. If there exist observations where $y^T_{i,t-1} = y^T_{i,t} = 1$, then λ will have the same sign as β. In addition, the size of λ is a function of the proportion of observations in the sample that were transformed; that is, the frequency of ongoing years. As a consequence, the transformation of the dependent variable results in attenuation bias for the effect of x_i upon onsets.[6] Note that dropping ongoing years of the binary event, as approximately 36% of papers do, removes observations where the omitted variable term does not equal zero. therefore eliminating the bias.

2.2 *The Issue of State Dependence*

We now move on to seeing how the bias is affected when the data-generating process involves state dependence. In particular, we will focus on the case where the independent variable has a different effect upon onsets and duration, $\gamma \neq 0$.[7] Our new data-generating process is

$$y^T_{i,t} = 1\{\beta x_{i,t} + \gamma y^T_{i,t-1}x_{i,t} + \;\epsilon_{i,t} \geq 0\}. \tag{9}$$

As before, $y'_{i,t}$ is the transformed version of the dependent variable, where

$$y'_{i,t} = \begin{cases} 1\{\beta x_{i,t} + \gamma y^T_{i,t-1}x_{i,t} + \;\epsilon_{i,t} \geq 0\} & \text{if } y^T_{i,t-1} = 0 \\ 1\{-\beta x_{i,t} - \gamma y^T_{i,t-1}x_{i,t} - \;\epsilon_{i,t} \geq 0\} & \text{if } y^T_{i,t-1} = 1 \quad \& \quad y^T_{i,t} = 1. \end{cases} \tag{10}$$

[6]It should be noted at this point that including $y^T_{i,t-1}$ in the estimation equation, as in Fearon and Laitin (2003), reduces this particular form of bias. This is because the omitted variable term includes $y^T_{i,t-1}$, therefore including this in the estimation corrects for this issue. However, this does mean that the parameter associated with $y^T_{i,t-1}$ should not be interpreted in a causal way; that is, as the likelihood of onset given that there is currently an event occurring.

[7]Again, for ease of exposition we omit the constant term α as well as the change in constant term when a country is experiencing a binary event δ.

https://doi.org/10.1093/pan/mpv019 Published online by Cambridge University Press

Now we write this data-generating process in latent variable form:

$$y^{*}_{i,t} = \underbrace{\beta x_{i,t} + \epsilon_{i,t}}_{\text{Correctly specified}} + \underbrace{y^{T}_{i,t-1}(\gamma x_{i,t} - 2y_{i,t-1}y_{i,t}\gamma x_{i,t} - 2y^{T}_{i,t}\beta x_{i,t})}_{\text{Omitted variables including } x_{i,t}} - \underbrace{2y^{T}_{i,t-1}y^{T}_{i,t}\,\epsilon_{i,t}}_{\text{Omitted variable from transformation.}} \tag{11}$$

This expression is similar to before, with the addition of the state-dependent effects. However, unlike before, the bias is more dependent upon features of $y^{T}_{i,t-1}$ and y^{T}_{t}. The expression containing the state-dependent effects can take on the following values dependent upon $y^{T}_{i,t-1}$ and y^{T}_{t}:

$$y^{T}_{i,t-1}(x_{i,t}\gamma - 2y_{i,t-1}y_{i,t}x_{i,t}\gamma - 2y^{T}_{i,t}x_{i,t}\beta) = \begin{cases} 0 & \text{if } y^{T}_{i,t-1} = 0 \\ x_{i,t}\gamma & \text{if } y^{T}_{i,t-1} = 1 \quad \& \quad y^{T}_{i,t} = 0 \\ -x_{i,t}\gamma - 2x_{i,t}\beta & \text{if } y^{T}_{i,t-1} = 1 \quad \& \quad y^{T}_{i,t} = 1. \end{cases} \tag{12}$$

Thus, the form and severity of this bias will depend upon the frequencies of $y^{T}_{i,t-1}$ and y^{T}_{t} in the sample, as well as the values of β and γ.

While we can know the frequencies of $y_{i,t}$ and $y_{i,t-1}$ before estimation we do not know β and γ; thus, we cannot be certain a priori the extent to which inferences will be biased. However, as we know features of the dependent variable, some features of the bias are apparent. In particular, dependent variables that measure rare events will likely have little bias due to the large number of cases where $y^{T}_{i,t-1} = 0$, resulting in a value of zero for the omitted variable expression. This is why setting ongoing years to zero in the context of dyadic interstate war data will lead to little bias.

However, onsets of civil war and democracy, while being rare events, differ in a key way. While these onsets can be classified as rare events, the incidences of civil war and democracy are not. This results in a considerable number of observations where the omitted variable expression will not equal zero, due to the presence of ongoing years. Further discussion of this issue is located in the next section.

3 A Monte Carlo Study

Having demonstrated the bias that can arise when researchers transform ongoing years to zero, I now implement a Monte Carlo study. In doing so, we can also learn about how the transformation affects 95% confidence interval coverage and root mean squared error (RMSE), as well as the bias of estimates. To do so, I define the data-generating process as

$$y^{*}_{i,t} = \alpha + \beta x_{i,t} + \delta y_{i,t-1} + \gamma x_{i,t}y_{i,t-1} + \epsilon, \tag{13}$$

where ϵ is drawn from the logistic distribution with mean zero and variance $\pi^2/3$. The outcome variable $y_{i,t}$ then takes a value of one when $y^{*}_{i,t}$ is greater than zero, and the value of zero otherwise.

To explore a wide range of scenarios, I set parameter values in the following ways. The constant term α is taken from the set $\{-5, -4, -3, -2\}$ and the change in intercept, δ, when $y_{i,t-1} = 1$ is taken from the set $\{2, 3, 4, 5, 6\}$. Values for the effect of the independent variable upon onset β and duration γ are taken from the set $\{-2.5, -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2, 2.5\}$. This results in 2420 possible combinations of the parameters, which are the Monte Carlo scenarios. For each scenario, I compute a thousand Monte Carlo iterations, which results in a total of 2.42 million Monte Carlo iterations. I fix the number of units to equal 100 and time periods to equal 40, a common temporal and cross-sectional domain of applied research. From these samples, I exclude scenarios that lead to an average proportion of y greater than 0.25, so as to ensure the experiments are similar to conditions typically faced by applied researchers.[8] It should be noted that the negative conclusions on setting ongoing years to zero become even more severe in experiments with larger proportions of y in the sample than focused on in this section.

[8]This reduces the number of scenarios to 1617. As will be discussed later in the text, graphs of the results for the full set of experiments are included in the Supplementary Materials (Figs. 1–3).

https://doi.org/10.1093/pan/mpv019 Published online by Cambridge University Press

Four models are compared in the Monte Carlo experiments. Model 1 involves estimating a logit model on the transformed onset variable, while also including a cubic polynomial of time since the end of the binary event spell.[9] This is the most commonly used model in the literature when setting ongoing years to zero. Model 2 also involves estimating a logit model on the transformed onset variable; however, the cubic polynomial of time now measures time since last binary event onset. This is not as common an approach compared to Model 1; however, it is included, as there exist papers that do this (e.g., Getmansky 2012). Model 3 estimates the fully interactive first-order Markov transition logit. Finally, Model 4 estimates a logit model on the transformed onset variable while also including a one-period lag of the untransformed dependent variable, as is done by Fearon and Laitin (2003).

I focus on two quantities of interest: bias and confidence interval coverage.[10] Both of these quantities are derived from the general formula

$$\frac{\hat{\theta} - \theta}{\theta} * 100. \tag{14}$$

In the case of bias, $\hat{\theta}$ is equal to the mean estimate of the effect of x upon onsets, β, for each model and θ is the true value of this parameter defined in the experiment.[11] In the case of 95% confidence interval noncoverage, $\hat{\theta}$ is the proportion of 95% confidence intervals that include the true β for each model and $\theta = 0.95$, which is the nominal 95% confidence interval coverage. Therefore, noncoverage is interpreted as the difference in percentage points between the observed 95% confidence interval coverage and the expected 95% coverage.

Table 3 summarizes the mean results of the Monte Carlo simulations.[12] From this initial summary we can see that simply setting the ongoing years to zero results in worse performance compared to the estimation of a first-order Markov model or including a one-period lag of the untransformed dependent variable. The bias for Models 1 and 2 is approximately five times larger in absolute terms than that of Model 3, and also results in considerably worse 95% confidence interval coverage.

While these problems appear small in general, the performance of Models 1 and 2 can be significantly worse, dependent upon features of the data analyzed. Therefore, I move on to showing how aspects such as the proportion of observations where $y = 1$, as well as the number of observations recoded both as a proportion of observations in the sample and of observations where $y = 1$, affect the quality of inference from the models. In doing so, I illustrate cases where simply setting ongoing years to zero comes at a significant inferential cost, as well as cases where there is little harm in doing so.

Figure 1 plots the association between these quantities of interest and the proportion of observations where the dependent variable y receives a value of 1. Unsurprisingly, as the proportion of events in the sample increases, the performance of models that simply set ongoing years to zero decreases. At low levels of events in the sample, particularly when 5% or less of observations of $y = 1$, models that only set ongoing years to zero have similar performance to those that take into account whether the binary event is still ongoing.[13]

[9]While I do not include temporal dependence of this form in the Monte Carlo setup, some researchers have suggested that the inclusion of temporal controls such as those proposed by Beck, Katz, and Tucker (1998) and Carter and Signorino (2010) mitigates the problem of setting ongoing years to zero. For example, Bergholt and Lujala (2012, 152) state that "we include all country-year observations following the conflict onset... To control for the possibility that a country *that is already experiencing conflict*, or that recently endured one, may be more likely to experience another conflict, we include a variable that counts the years since the last year of conflict, as suggested by Beck, Katz & Tucker (1998)" [emphasis added]. However, as noted previously by Beck, Katz, and Tucker (1998, 1272): "If conflicts really are multi-year, we should simply drop all but the first year of the conflict from the analysis."

[10]RMSE was also calculated, but is located in the Supplementary Materials due to issues of space (Figs. 1–3). In general, RMSE is high for all models when there are few observations where $y = 1$, but remains at a lower level for Models 3 and 4 as the proportion of observations where $y = 1$ increases.

[11]This means that experiments where $\beta = 0$ are not included. Examination of cases where $\beta = 0$ shows that there is no bias (in terms of distance) for all models in these cases.

[12]Replication materials are available at http://dx.doi.org/10.7910/DVN/DOSJCD (McGrath 2015).

[13]This suggests that empirical applications using dyadic country-year data that set ongoing years to zero do not suffer from problems. For example, dyadic studies of interstate war since World War 2 have a proportion of observations at war equal to 0.3% (King and Zeng 2001, 694).

https://doi.org/10.1093/pan/mpv019 Published online by Cambridge University Press

Table 3 Summary of the Monte Carlo simulations

Mean	Model 1 zero *t* incidence	Model 2 zero *t* onset	Model 3 Markov	Model 4 zero *w*/lag of incidence
Bias (%)	−1.45	−1.72	0.30	0.50
CI noncoverage (%)	−5.85	−7.16	0.03	0.00
y	0.09	0.09	0.09	0.09
Recoded *y*	0.05	0.05	0.05	0.05

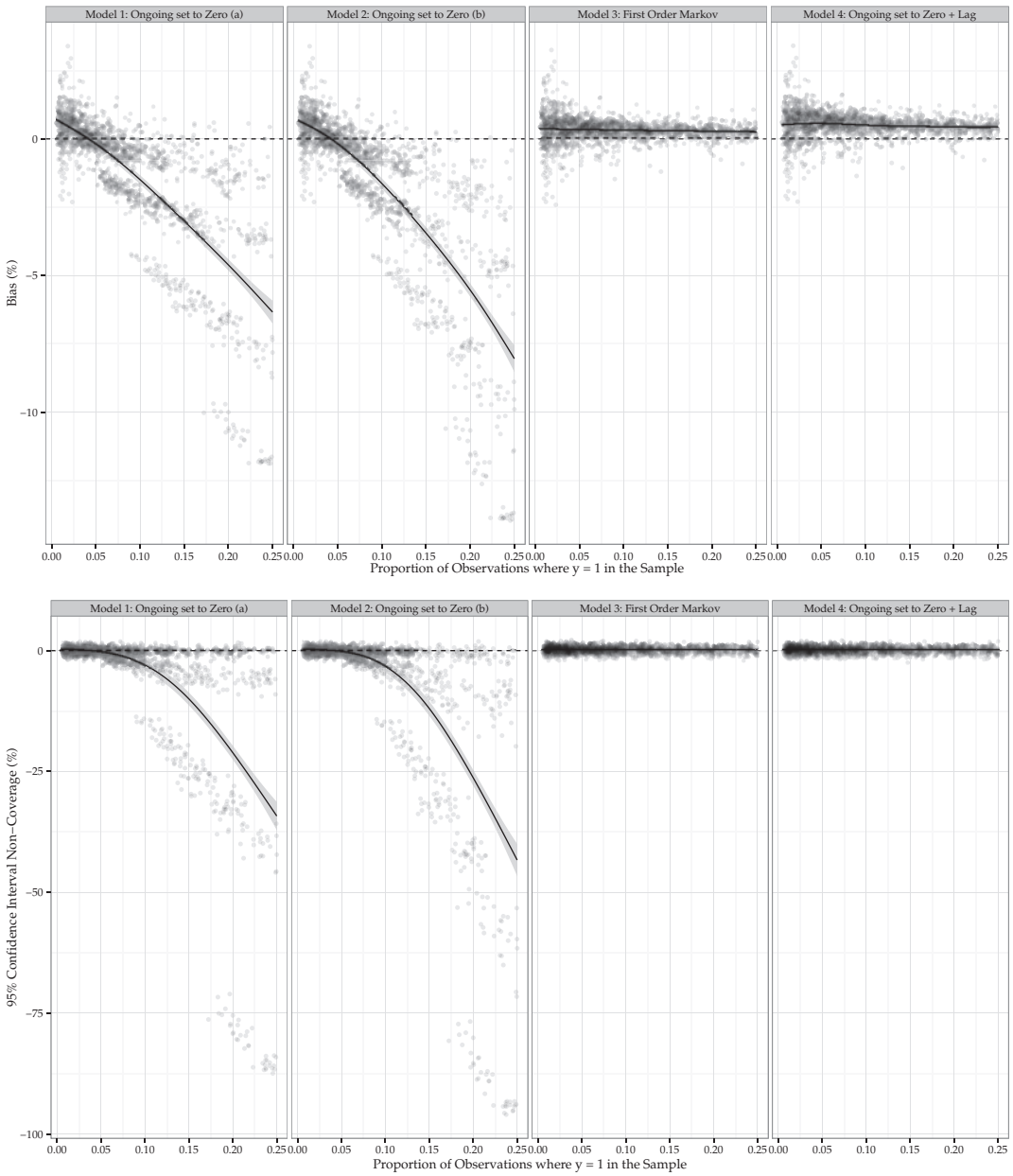


Fig. 1 Bias and confidence interval noncoverage as a function of the proportion of $y = 1$ in the sample. Model 1 is a logit with a cubic polynomial of time since last event, estimated on an outcome variable where ongoing years are set to zero. Model 2 is a logit with a cubic polynomial of time since last onset, estimated on an outcome variable where ongoing years are set to zero. Model 3 is a first-order Markov transition logit, where event incidence is the outcome variable. Model 4 sets ongoing years to zero and includes a one-period lag of the untransformed dependent variable.

https://doi.org/10.1093/pan/mpv019 Published online by Cambridge University Press

However, beyond this point, performance of Models 1 and 2 considerably worsens. For example, when the proportion of observations where $y = 1$ is 15–20% of the sample, there is greater bias and worse confidence interval coverage in Models 1 and 2.[14] In this range, the average bias is approximately 4–5% for Models 1 and 2, which simply set ongoing years to zero. In addition, there is approximately 10–20% less coverage of the estimated 95% confidence intervals than the 95% expected when appropriately constructed. This is in comparison to the first-order Markov transition model (3) and the model that includes a one-period lag of the untransformed dependent variable (4), which both suffer from negligible bias and have a mean 95% confidence interval coverage of that does not differ from 95%.[15]

I now examine how these quantities of interest vary depending upon the number of ongoing years set to zero as a proportion of the entire sample. Figure 2 displays how the proportion of observations recoded is associated with bias and confidence interval coverage. Although performance initially worsens for Models 1 and 2 as the proportion of observations recoded increases, performance unexpectedly increases from approximately 0.15 onward. This counterintuitive non-monotonicity occurs for two reasons. The first reason for this is due to discarding experiments where the average proportion of y in the sample is greater than 0.25. The second reason is that this non-monotonicity occurs due to pooling the results for different values of beta. In order to maintain the focus of the experiments to cases that are typical for political science research, I therefore present these subsequent results by fitting the Loess curves separately for each (absolute) value of beta determined by the experiment.[16]

Figure 3 displays the same results as Fig. 2; however, this time fitting the Loess curves separately for each absolute value of beta. The results show that as the proportion of observations recoded increases, the performance of Models 1 and 2 that solely set ongoing years to zero worsens. In addition, as the effect of x upon the onset of the binary event increases, bias and lack of confidence interval coverage increase. In contrast, Models 3 and 4, which include information on whether the unit is still experiencing the binary event, perform considerably better, in terms of having little bias and appropriate 95% confidence interval coverage.

Figure 4 shows how the performance of models is affected by the number of ongoing events set to zero, as a proportion of the number of observations where the dependent variable equals one.[17] Examining these associations, we can see that both bias and confidence interval coverage worsen as the proportion of the dependent variable increases.

I also subset these results into categories based upon the proportion of observations where $y = 1$ in the data displayed in Fig. 5. This is done to further understand how the number of ongoing years set to zero as a proportion of the number of observations where $y = 1$ affects the performance of estimators, and to see how these two aspects interact with one another. In doing so, we can see that the effect of the proportion of ys recoded conditional upon the frequency of $y = 1$ does not differ considerably across different overall proportions of ys in the sample. Rather, issues of bias and confidence interval coverage seem to be more driven by the overall proportion of observations recoded and of $y = 1$ in the sample, when comparing these associations to Figs. 1 and 3.

In summary, the Monte Carlo estimates offer a number of points to consider when estimating onsets of binary events in time-series cross-sectional data:

- As the proportion of observations that are recoded in the sample increases, models that set ongoing years to zero without incorporating information about whether the binary event is still ongoing perform poorly in terms of bias and confidence interval coverage.

[14]This particular proportion is of interest, as this is typical for data on civil war incidence.
[15]Model 4 is relatively more biased than Model 3, being approximately 50% larger. However, this bias is small in absolute terms, so it is not focused on here.
[16]Figure 1 in the Supplementary Materials plots the associations for the full sample. The Loess curve follows a general negative trend as the proportion recoded increases.
[17]Similar to Fig. 2, the unconditional Loess curve shows the same non-monotonicity for the same two reasons noted before. Therefore, I follow the same approach as in Fig. 3 and estimate separate Loess curves dependent upon the absolute value of beta.

https://doi.org/10.1093/pan/mpv019 Published online by Cambridge University Press

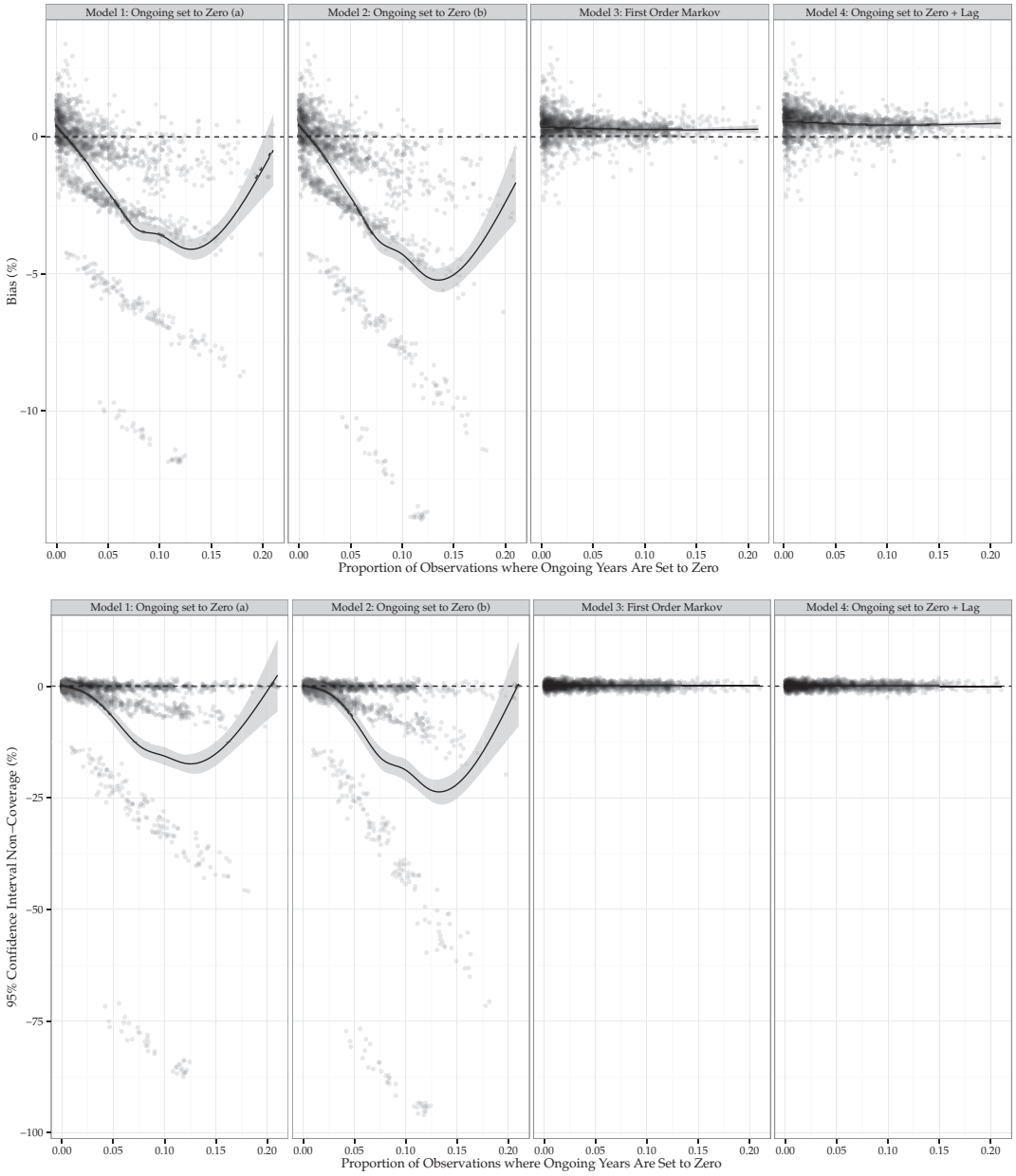


Fig. 2 Bias and confidence interval noncoverage as a function of the proportion of observations of y recoded in the sample. Model 1 is a logit with a cubic polynomial of time since last event, estimated on an outcome variable where ongoing years are set to zero. Model 2 is a logit with a cubic polynomial of time since last onset, estimated on an outcome variable where ongoing years are set to zero. Model 3 is a first-order Markov transition logit, where event incidence is the outcome variable. Model 4 sets ongoing years to zero and includes a one-period lag of the untransformed dependent variable.

- While setting ongoing years to zero and including a one-period lag of the untransformed variable (as in Fearon and Laitin 2003) results in relatively larger bias than estimating a first-order Markov model, the difference is typically small in absolute terms.
- It is safe to estimate models where ongoing years are set to zero if the proportion of observations is small; that is, less than 5%. Typical dyadic time-series cross-sectional data on the onset of interstate war or the signing of preferential trade agreements, for example, tend to have proportions of event years less than 1%. However, beyond this

https://doi.org/10.1093/pan/mpv019 Published online by Cambridge University Press

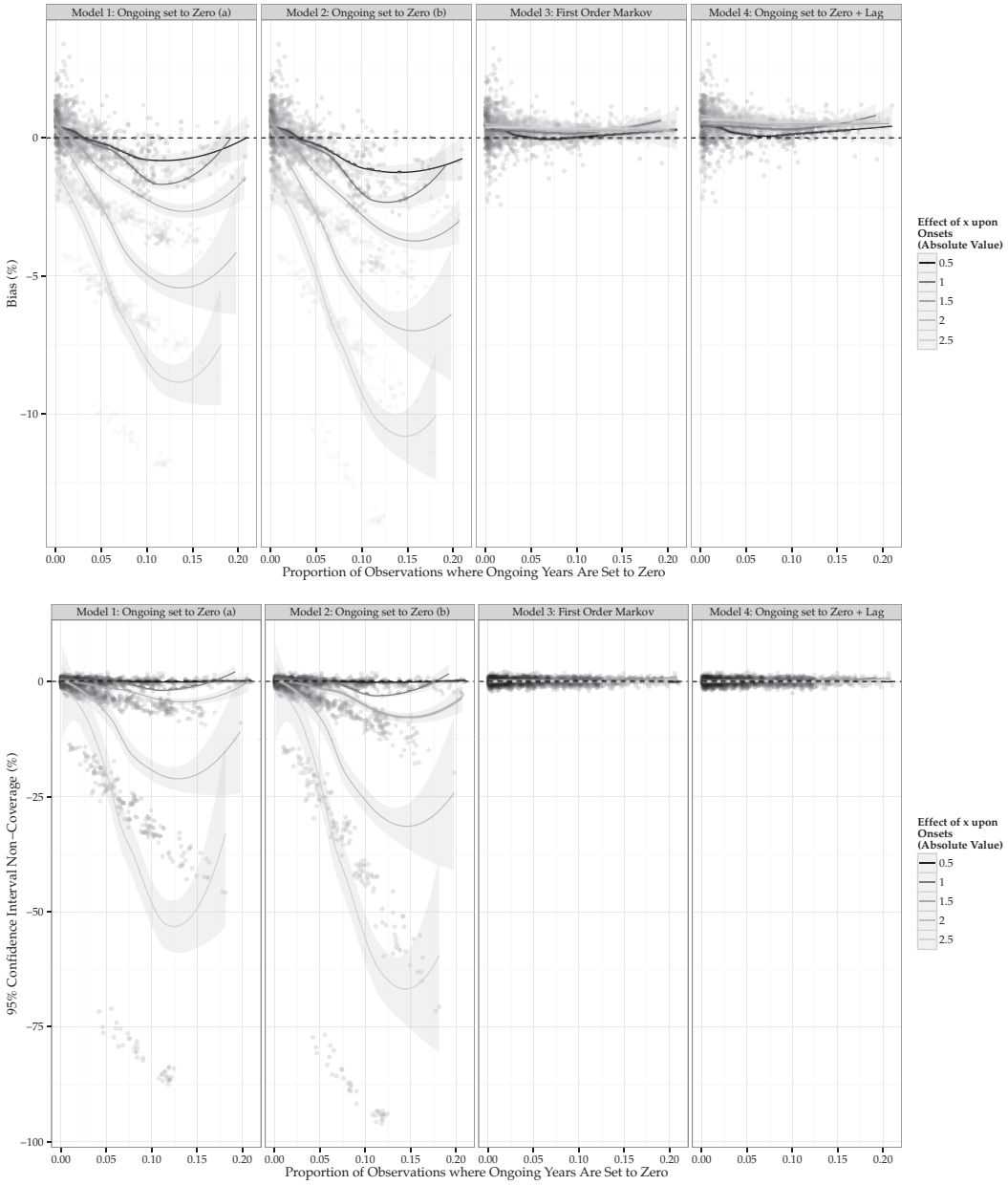


Fig. 3 Bias and confidence interval noncoverage as a function of the proportion of observations of y recoded in the sample conditional upon the absolute value of the coefficient β capturing the effect of x upon onsets.

range there are inferential issues. There is considerable bias when the proportion of observations lies between 15% and 20%, which is typical for data on civil war incidence.

- While RMSE is similarly large for all models when the proportion of $y = 1$ in the sample is low, RMSE decreases faster and remains smaller when estimating a first-order Markov model or including a one-period lag of the untransformed dependent variable.[18]

[18]See Figs. 1–3 in the Replication Material.

https://doi.org/10.1093/pan/mpv019 Published online by Cambridge University Press

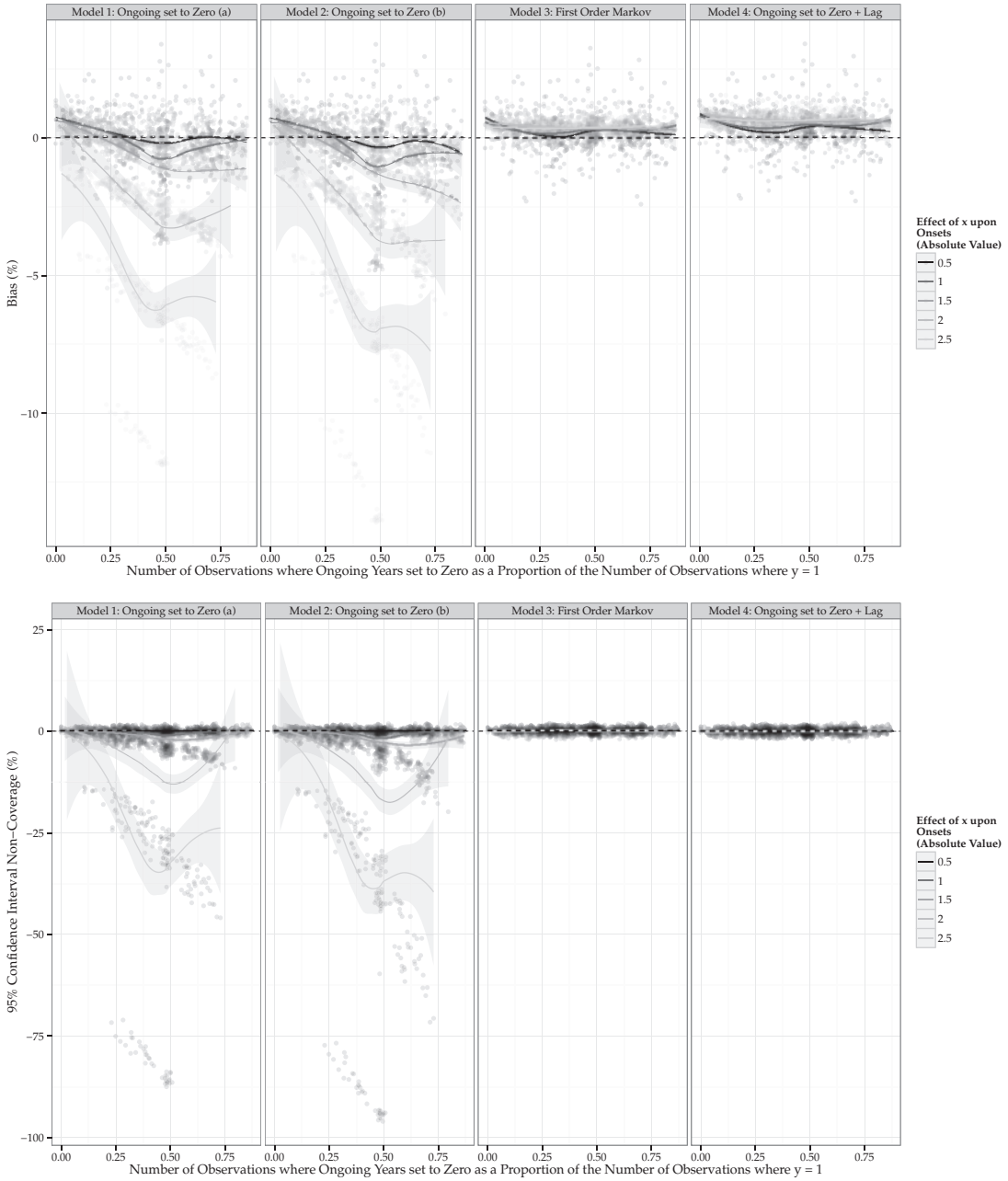


Fig. 4 Bias and confidence interval noncoverage as a function of the number of observations of y recoded as a proportion of the frequency of y in the sample conditional upon the absolute value of the coefficient β capturing the effect of x upon onsets.

4 Replication—Sensitivity Analysis of the Onset of Civil War

To demonstrate the consequences of setting ongoing years to zero within typical empirical analyses, I conduct a replication of Hegre and Sambanis (2006) (henceforth referred to as HS). HS conduct a sensitivity analysis of the determinants of the onset of civil wars, in a similar way to Sala-I-Martin (1997). While HS are correct in setting ongoing years to missing, thereby avoiding the issues raised in this article, I extend their analysis to the estimation of models when ongoing years are set to zero as well as a first-order Markov transition model.

https://doi.org/10.1093/pan/mpv019 Published online by Cambridge University Press

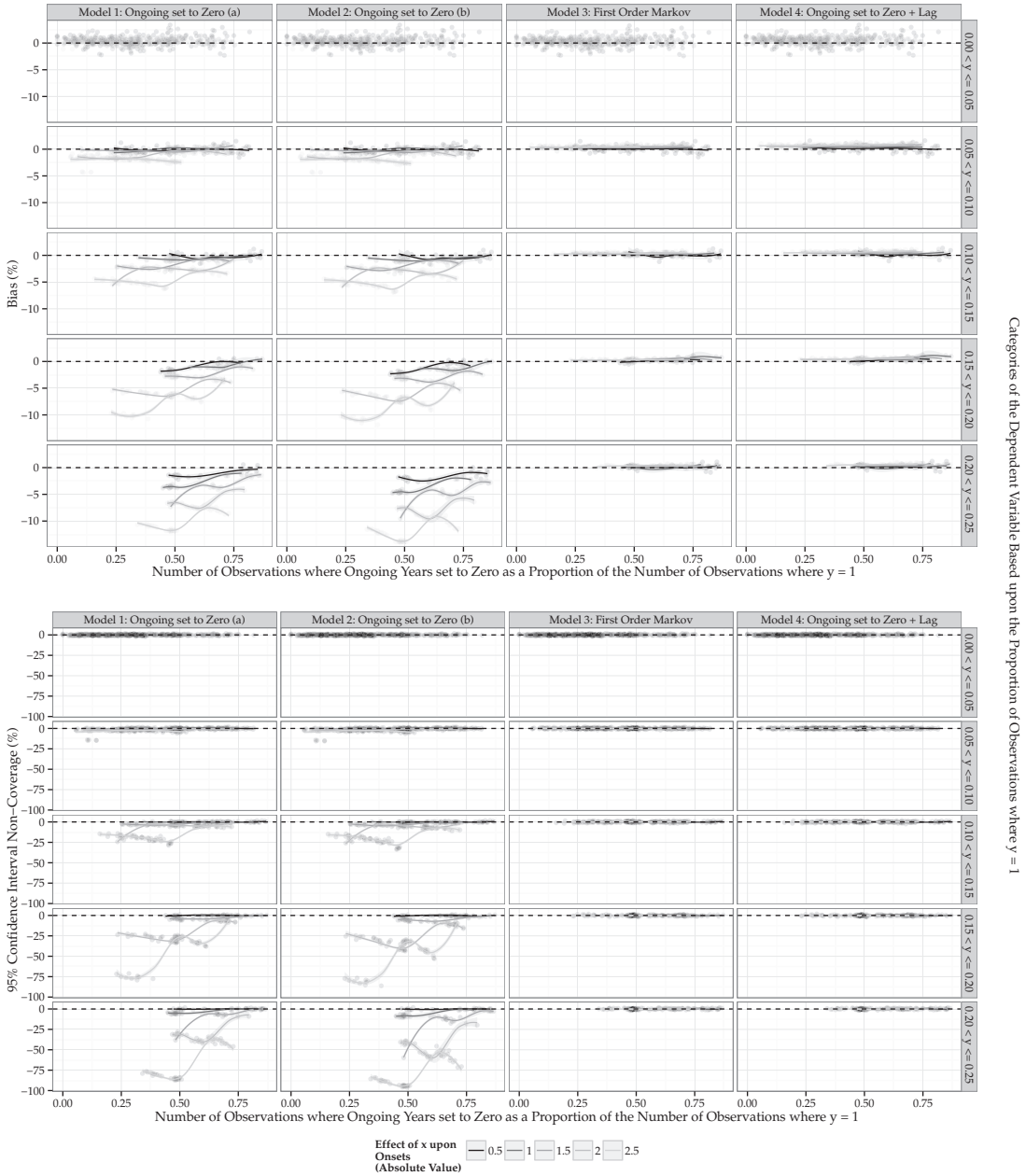


Fig. 5 Bias and confidence interval noncoverage as a function of the number of observations of y recoded as a proportion of the frequency of y in the sample. This is further conditioned upon the absolute value of the coefficient β capturing the effect of x upon onsets.

Performing such a replication, rather than that of a single study, allows examination of the broader effect of setting ongoing years to zero. This sensitivity analysis is able to give us some sense of the distribution of cases where choosing to set ongoing years to zero leads to different inferences, which is not possible to do with the replication of a single empirical analysis. Therefore, we can tentatively say to what extent results in the literature are dependent upon the choice of coding for the binary dependent variable. HS conduct their sensitivity analysis in the following way.[19]

[19]The structure and notation of this discussion closely follow HS, for ease of comparison.

https://doi.org/10.1093/pan/mpv019 Published online by Cambridge University Press

M models are estimated and operationalized as

$$\gamma_j = \alpha_j + \boldsymbol{\beta}_{yj}\mathbf{y} + \beta_{zj}z_j + \boldsymbol{\beta}_{xj}\mathbf{x}_j + \epsilon, \tag{15}$$

where γ is the dependent variable, $\mathbf{y}$ is a vector of three variables that appear in every model,[20] z is the variable of interest, and $\mathbf{x}$ is a vector of three variables taken from the set χ of variables of interest.

While HS follow the approach of Sala-I-Martin (1997), there are some notable differences motivated by the subject of interest. First, each variable of interest z is placed into a category determined by the theoretical concept it seeks to measure. For example, the polity index is included in the "level of democracy" category. The category of a given variable of interest determines which variables can be included in the vector of three variables that are also included in the model. Only variables that are from a different category to that of the given variable of interest are allowed to be included in the vector of three control variables. Continuing with the example, this means that when the polity index is the variable of interest, other variables in the "level of democracy" category such as the measure of democracy used in Przeworski et al. (2000) are excluded from being included in the vector of three control variables.[21] Second, HS weight the estimates from each model by McFadden's Pseudo-R^2, rather than by the log-likelihood, as is the case with Sala-I-Martin (1997).

Following HS, the vector of three variables are GDP per capita, population size, and time since last conflict. I differ from HS by using a cubic polynomial of time since last conflict, instead of their monotonic decay function of time, to ensure greater comparability with current empirical approaches in the literature which allow for non-monotonic effects of time since last conflict. In addition, HS include both GDP per capita and population as their natural logarithm, which is the same here.

The procedure then takes the following form:

1. Choose a variable z from the set of variables of interest χ.
2. Calculate all unique three element vectors $\mathbf{x}$ from the remaining variables of interest that are not of the same category as z.
3. Randomly sample without replacement five thousand of the three element vectors.[22]
4. For each of these vectors, estimate the model outlined in equation (15) for the dependent variables.
5. Store the estimated coefficient, standard error, and p value for z, as well as McFadden's Pseudo-R^2.
6. Repeat same process for the next variable of interest.

From this I focus on two quantities of interest relevant to researchers. The first is the weighted mean of β_{zj} coefficients for the variables of interest in χ. This weighted mean is computed by weighting each of the five thousand estimated coefficients by McFadden's Pseudo-R^2. The second is the non-normal p value for each of these variables.[23] This is computed by similarly using a weighted mean of all of the five thousand estimated p-values, with weights defined by the values of McFadden's Pseudo-R^2 for each model. These quantities of interest allow for comparing how both the substantive and statistical significance of variables is affected by whether researchers set ongoing years to zero or instead account for ongoing years by setting them to missing or using a first-order Markov model.

[20]For this replication I use five variables, as will be subsequently discussed in the main text. This is due to using a cubic polynomial of time, as suggested by Carter and Signorino (2010), instead of a decay function of time, as it allows for non-monotonic hazard rates, which is the most common approach in the literature. The spline approach of Beck, Katz, and Tucker (1998) also allows for non-monotonic hazard rates.

[21]Categories are located in Table 1 in the Supplementary Materials.

[22]This sampling is performed due to the (lack of) availability of computational resources to both perform this replication and the Monte Carlo analysis. Nevertheless, results are consistent with those of HS.

[23]This approach does not rely on the assumption that the distribution of the estimates of β_{zj} is normal. Inspection of the distributions of estimates shows that the distributions are skewed and/or are non-monotonic either side of the point of maximum density of the distribution, implying non-normality.

https://doi.org/10.1093/pan/mpv019 Published online by Cambridge University Press

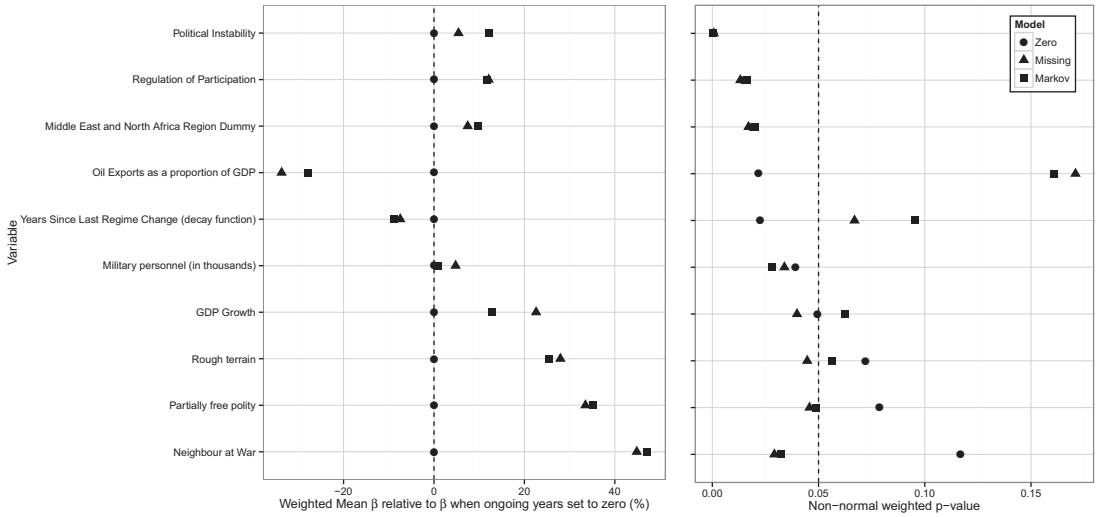


Fig. 6 Comparison of the effects of variables upon the onset of civil war and their statistical significance, between models where ongoing years are set to zero, ongoing years are dropped, and a first-order Markov model.

4.1 *Results of the Replication*

In presenting the results of the replication, I focus on variables that are statistically significant, with a weighted p value of less than 0.05, in at least one of the models.[24] There are ten variables out of the full set of eighty-eight that are found to have robust effects upon the onset of civil war, given the chosen threshold for statistical significance. Figure 6 plots the size of the coefficient capturing the onset effect of independent variables for both the model setting ongoing years to missing and the first-order Markov model, relative to the coefficient estimated when setting ongoing years to zero. In addition, the weighted p values for these variables for all three models are included.

In examining Fig. 6, we see that four of the ten variables are classified as either statistically significant or insignificant, dependent upon whether ongoing years are set to zero. Oil exports as a percentage of GDP (`oil`) and a decay function of years since last regime transition (`progrexc`) are classified as statistically significant when setting ongoing years to zero, but not when setting ongoing years to missing or estimating a first-order Markov transition model. In contrast, whether a neighbor is at war in a given year (`nat_war`) and whether there is a partially free polity (`partfree`) are found to be statistically insignificant when setting ongoing years to zero, yet are statistically significant when dropping ongoing years of conflict or estimating a first-order Markov model. Comparing dropping ongoing years of conflict to the estimation of a first-order Markov model, both have similar p values, and when they do differ (for instance in the case of `gdpgrowth`), the difference is small (approximately 0.01–0.02). Even adopting a weaker criteria for statistical significance such as 0.10 would still result in two of these ten variables' classification of statistical significance to be dependent upon whether ongoing years are set to zero or not.

Turning to substantive effects of the variables, there are also stark differences between whether or not ongoing years are set to zero. In the cases where p values were noticeably different, there is also a considerable difference in the size of coefficients. The coefficient for whether a neighbor is at war is approximately 40% larger if ongoing years are dropped or a first-order Markov model is estimated, compared to a model where ongoing years are set to zero. A similarly large difference is found when looking at the effect of being an oil exporter, with the coefficient being approximately 30% smaller when not setting ongoing years to zero. Again, the coefficients when dropping ongoing years and estimating a first-order Markov model are similar, with only small differences between them.

[24] Results for all variables are located in the Appendix. While there are numerous issues with the use of p values as a mode of inference, they are, nonetheless, the dominant measure used by applied researchers to test hypotheses of the effects of variables of interest.

https://doi.org/10.1093/pan/mpv019 Published online by Cambridge University Press

To summarize, replicating the sensitivity analysis of HS finds that for a significant proportion of variables, being classified as a robust determinant of the onset of civil wars is dependent on whether ongoing years of conflict are set to zero or not. In addition to impacting statistical significance tests, the choice of transformation also leads to considerable changes in the substantive impact of variables, with some coefficients changing by 20% or more, depending on the model estimated. As approximately 42% of the research surveyed for this article simply set ongoing years to zero, reducing this percentage would improve the inferences found in the literature, given these findings with actual data as well as the Monte Carlo evidence.

5 Conclusion

This article has shown how the seemingly intuitive idea of creating an binary onset variable, where ongoing years of the event are equal to zero, from a binary event outcome can cause unintended harm in time-series cross-sectional data. Importantly, some degree of bias occurs regardless of features of the independent variables, apart from when variables have no effect upon the onset of the binary event.

Thankfully, there are fairly simple means by which to better estimate these processes. Monte Carlo analysis has shown that a simple first-order Markov transition model is better able to recover the effect of variables on onsets. While one would think that setting ongoing years of a binary event to zero is unproblematic in the case of no state dependence, there still exists bias even if the variable of interest has equal effects on the onset and continuation of an event. It is also important to note that the inclusion of variables to account for temporal dependence, as recommended by Beck, Katz, and Tucker (1998) and Carter and Signorino (2010), does not account for the bias induced by transforming the dependent variable.

The insights of the analytical and Monte Carlo demonstrations of bias are illustrated by a replication of a sensitivity analysis of the determinants of the onset of civil wars by HS. I extend their analysis, which drops ongoing years of conflict, by also estimating models where ongoing years are set to zero and a first-order Markov model. Doing so provides an indication of the distribution of inferences that are affected by choosing to set ongoing years to zero. The statistical significance for four of the ten variables considered as robust determinants is dependent upon whether ongoing years are set to zero. Furthermore, there can be considerable differences in the size of the estimated coefficients as a result of whether ongoing years are set to zero or not. This suggests potential inferential issues for the approximately 42% of research surveyed for this article that simply set ongoing years to zero.

Moving forward, researchers should be more aware of how a simple transformation can seriously affect substantive inferences, and follow the recommendations regarding specification and dependent variable coding offered in this article. At a minimum, researchers should include a one-period lag of the untransformed dependent variable, as in Fearon and Laitin (2003). Yet in doing so, researchers should keep in mind that it is not correct to interpret the associated parameter in any causal way; it is simply an adjustment to "inform" the estimator of the recoding in the dependent variable. In addition, researchers should take care to examine whether there are state-dependent effects for independent variables. While it may seem slightly more complex, it is the case that standard binary estimators are nested within the first-order Markov transition models. As such, there is much to learn from moving beyond homogeneity by default, by testing rather than assuming that variables have identical effects upon onsets and durations.

Conflict of interest statement. None declared.

References

Bane, Mary Jo, and David T. Ellwood. 1986. Slipping into and out of poverty. *Journal of Human Resources* 21:2–23.

Barmby, Tim. 1998. The relationship between event history and discrete time duration models: An application to the analysis of personnel absenteeism. *Oxford Bulletin of Economics and Statistics* 60:261–65.

https://doi.org/10.1093/pan/mpv019 Published online by Cambridge University Press

Beck, Nathaniel, Jonathan N. Katz, and Richard Tucker. 1998. Taking time seriously: Time-series-cross-section analysis with a binary dependent variable. *American Journal of Political Science* 42(4):1260–88.

Beck, Nathaniell, David Epstein, Simon Jackman, and Sharyn O'Halloran. 2001. *Alternative models of dynamics in binary time-series–cross-section models: The example of state failure*. Working paper.

Bergholt, D., and P. Lujala. 2012. Climate-related natural disasters, economic growth, and armed civil conflict. *Journal of Peace Research* 49(1):147–62.

Boskin, M. J., and F. C. Nold. 1975. A Markov model of turnover in aid to families with dependent children. *Journal of Human Resources* 10:476–81.

Carter, David B., and Curtis S. Signorino. 2010. Back to the future: Modeling time dependence in binary data. *Political Analysis* 18(3):271–92.

Diggle, Peter, Kung-Yee Liang, and Scott L. Zeger. 1994. *Analysis of longitudinal data*. Oxford: Oxford University Press.

Fearon, James D., and David D. Laitin. 2003. Ethnicity, insurgency, and civil war. *American Political Science Review* 97(1):75–90.

Getmansky, Anna. 2012. You can't win if you don't fight: The role of regime type in counterinsurgency outbreaks and outcomes. *Journal of Conflict Resolution* 57:709–34.

Hegre, Havard, and Nicholas Sambanis. 2006. Sensitivity analysis of empirical results on civil war onset. *Journal of Conflict Resolution* 50:508–36.

Jackman, Simon. 2000. *In and out of war and peace: Transitional models of international conflict*. Working paper.

King, Gary, and Lanche Zeng. 2001. Explaining rare events in international relations. *International Organization* 55(3):693–715.

McGrath, Liam F. 2015. Replication data for: Estimating onsets of binary events in panel data. Harvard Dataverse, V1 [UNF:6:QIfNzWwGaK+slGPMJKjf+w==] http://dx.doi.org/10.7910/DVN/DOSJCD.

Meyer, Bruce, and Nikolas Mittag. 2013. *Misclassification in binary choice models*. Working paper.

Przeworski, Adam, and James Raymond Vreeland. 2002. A statistical model of bilateral cooperation. *Political Analysis* 10(2):101–12.

Przeworski, Adam, Michael E. Alvarez, Jose Antonio Cheibub, and Fernando Limongi. 2000. *Democracy and development: Political institutions and well-being in the world, 1950–1990*. Cambridge, UK: Cambridge University Press.

Sala-I-Martin, Xavier X. 1997. I just ran two million regressions. *American Economic Review* 87(2):178–83.

Yamaguchi, Kazuo. 1991. *Event history analysis. Vol. 28 of Applied Social Research Methods Series*. Newbury Park, CA: Sage.

https://doi.org/10.1093/pan/mpv019 Published online by Cambridge University Press