Estimating Survival Models When Event-Failure is Overrreported

Benjamin E. Bagozzi,*

November 10, 2016

Abstract

Political scientists often study dependent variables that correspond to the time until an event occurred (or "failed"), otherwise known as "survival data". At times, however, our ability to accurately code an event as having failed is fairly poor. For example, event-failures are often imperfectly identified according to a crude cutoff criteria, ensuring that some observations that are coded as *non-right censored* persist beyond their *recorded failure*. Imperfectly recorded event-failures of this sort actually correspond to right censored events: the researcher can only accurately conclude that the observation lasted up until the recorded failure time. Concluding that the observation actually terminated at that point in time will often be problematic as there is a non-zero probability that the observation persisted past that point. Moreover, if heterogeneity exists among these imperfect codings of event-failures, then survival models will yield biased estimates of parameter effects. To address this problem I develop a new split population survival model that—in a similar fashion to a cure model—explicitly models the misclassification probability of failure (vs. right censored) events. After presenting this framework, I apply the resultant survival model to simulated and published political science data, finding that my approach allows one to account for imperfect detection in failure-events.

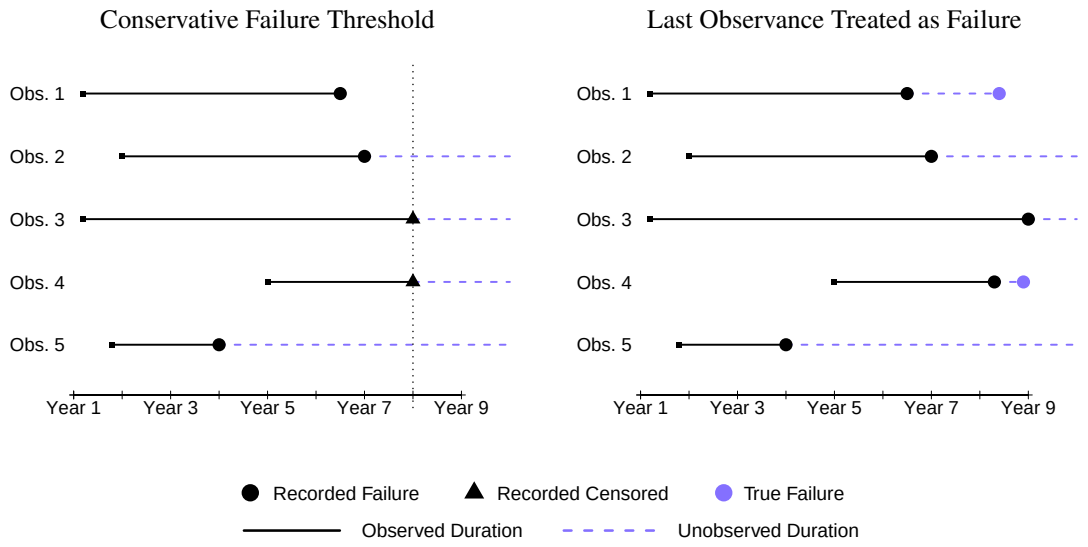
*Department of Political Science, University of Minnesota. emails: bbagozzi@umn.edu

The Problem

I argue that there are instances in social science survival data where (some subset of) the events that are recorded in one's data as "failures," actually persist beyond their recorded failure time. In this manner, these observations' true *censored* values are misclassified as *failed*. I imagine a few possible situations where this arises.

The first situation corresponds to overly conservative "failure thresholds." In many studies of survival data, researchers' events of interest don't always have clear end-points (i.e., failures), and the researcher him or herself must establish some sort of threshold criteria to determine whether (and when) a duration observation (or some subset of observations) can be considered "failed"/ended. Often the strategy is to choose a failure-threshold that, if anything, *underestimates* the length of one's actual event, the implicit reasoning being that it is better to be conservative and ensure that coded events end before they truly do than it is to code events as incorrectly persisting beyond their true failures. As an example of this, consider research on civil war duration. Here, researchers typically analyze the durations of rebel-government conflicts, 1946-2004, but record end dates ("failures") for specific conflicts based upon 24-month spells with fewer than 25 battle-deaths per year (UCDP/PRIO). This threshold may be overly conservative, especially for lower-intensity civil wars in remote or poor information environments that persist indefinitely with little actual fighting. I illustrate this in the "Conservative Failure Threshold" Subfigure below. Here, some cases persist beyond the window of time under analysis, and hence are accurately recorded as censored. In addition, some cases are also accurately recorded as failures during one's period of interest. However, an additional subset of recorded failures in the above figure persisted beyond their recorded failure time, due to researchers' overly conservative thresholds for determining failures. This introduces bias if it is random (towards zero), and additional bias in unknown directions if the variation in these underestimates of failures is correlated with commonly studied covariates (as it likely would be for the civil war case, especially with regards to things like terrain, rebel group size, country size, remoteness of civil war region, media attention and press freedoms, and so on).

The second corresponds to very long range historical survival data analyses. E.g., studies that assess the survival lengths of ancient civilizations, the durations of ancient wars or



spells between ancient wars, the impacts of ancient human hunting or settlement patterns on the extinction rates of various animal and plant species, and so on. For each of these studies, the researcher typically does not have data on the precise year for a given subject's *extinction event*, due to substantial time that has passed since the case being studied. Instead, the researcher must make do with the best available proxy of such an event: the *last known* record of a given ancient civilization or social activity being "alive." In the cases of civilizations, for example, this often amounts to the last known carbon dating (or other dating metric) of a societies' material products (e.g., baskets, houses) or members themselves. In this case, one could argue that the resultant survival data would best correspond to the "Last Observance Treated as Failure" Subfigure above: each observation's recorded failure time is an underestimate of that observation's true life-span (i.e., we know with certainty that that observation lasted up until that point, but there is a strong likelihood that it persisted for at least some amount of time past that recorded failure). To the extent that these underestimates of duration are non-random, and are correlated in their severity with commonly studied covariates (e.g., environmental or geographic conditions), bias will arise in survival estimates of these phenomenon.

Thirdly, another possibility where this types of "inflated" death cases may arise in survival data could be in instances where actors self-report their duration of engagement in a given activity (political participation, compliance with a given law, drug usage, making payments on a good, etc.). In a subset of these cases, some (but not all) actors may have strategic reasons to underreport their duration of engagement (i.e., to claim they finished engaging in an activity

when in fact they continued to engage in that activity). Here again then, the recorded "deaths" in one's data will be inflated.

Additional data applications in political and social science may be: Leblang and Bernhard (2000), Hannan and Freeman (1988). Below I derive an adjusted survival model (currently with a cheesy name) that accounts for this potential that some of one's survival events actually "lived on" beyond a researcher's recorded failure time for that event in their data.

The Parametric Zombie Survival Model

Define a general parametric survival model for continuous time duration data, wherein N subjects $i = \{1, 2, ...N\}$ each eventually experience an event of interest. Consistent with extant formulations of such models (Box-Steffensmeier and Jones, 2004), not all subjects need experience the event during a particular sample-period, as some may survive until the end of the sampling window, in which case they are "censored" in their final period of observation ($\tilde{C}_i = 0$ if censored, and 1 otherwise). The duration of interest t is thus assumed to have a probability density function (PDF) of $f(t) = \Pr(T_i = t)$, where T is an observation's duration of time until experiencing the event *or* censoring. The cumulative probability distribution (CDF) of the event (i.e., the probability of the event on or before t) is therefore: $\Pr(T_i \leq t) \equiv F(t) = \int_0^t f(t)\,dt$, where the probability of survival can be defined as $\Pr(T_i \geq t) \equiv S(t) = 1 - F(t)$ under the assumption that all subjects will eventually experience the event of interest at some point in time. Given the PDF $f(t)$ and CDF $F(t)$ defined above, the hazard of an event at t given that the event hasn't occurred prior to that point is $h(t) = \frac{f(t)}{S(t)}$.

One can then use these probability statements to define the (log) likelihood for the general parametric model. Note that uncensored observations ($\tilde{C}_i = 1$) provide information on both the hazard of an event, and the survival of individuals prior to that event, whereas censored observations ($\tilde{C}_i = 0$) only provide information on an observation having survived at least until time T_i. Combining each set of observation's respective contributions to the density and survival functions, one can define the general parametric likelihood as:

$$L = \prod_{i=1}^{N} [f(t_i)]^{\tilde{C}_i} [S(t_i)]^{1-\tilde{C}_i} \tag{1}$$

With corresponding log-likelihood:

$$\ln L = \sum_{i=1}^{N} \left\{ \tilde{C}_i \ln \left[f(t_i) \right] + (1 - \tilde{C}_i) \ln \left[S(t_i) \right] \right\} \tag{2}$$

Including covariates in 2 is possible by conditioning the terms of the log-likelihood on $\mathbf{X}$ and its associated parameter vector β:

$$\ln L = \sum_{i=1}^{N} \left\{ \tilde{C}_i \ln \left[f(t_i|\mathbf{X}, \beta) \right] + (1 - \tilde{C}_i) \ln \left[S(t_i|\mathbf{X}, \beta) \right] \right\} \tag{3}$$

Depending on the distributional assumptions one makes with respect to the probability of an event, the above log-likelihood statements can be used in conjunction with any of the commonly used parametric survival models—including the exponential, Weibull, Gompertz, log logistic, log normal, and generalized gamma.

I next alter this general model to account for asymmetric misclassification arising within one's censored and failure observations. In particular, I focus here on the situation where censored cases are misclassified as failed observations, in which case one's observed censoring indicator, C_i accurately records all censored cases $\left(\forall \ (C_i = 0) : (\tilde{C}_i = 0) \right)$ but misclassifies some subset of non-censored failure outcomes as censored $\left(\exists \ (C_i = 1) : (\tilde{C}_i = 0) \right)$. Drawing on the notation used in Box-Steffensmeier and Zorn's (1999) presentation of the split population survival model, one can define the probability of misclassification as:

$$\alpha = \Pr(C_i = 1 | \tilde{C}_i = 0) \tag{4}$$

Which implies that the unconditional density is defined by the combination of an observation's misclassification probability and its probability of experiencing an actual failure conditional on not being misclassified:

$$\Pr(\alpha = 1) + \Pr(\alpha = 0) \Pr(t_i \leq T_i) = \alpha_i + (1 - \alpha_i) * f(t_i) \tag{5}$$

With a corresponding unconditional survival function of:

$$\Pr(\alpha = 0) \Pr(t_i > T_i) = (1 - \alpha_i) * S(t_i) \tag{6}$$

Combining each set of observation's respective contributions to the density and survival functions, one can define the general parametric likelihood for the effects of covariates $\mathbf{X}$ on survival as:

$$L = \prod_{i=1}^{N} [\alpha_i + (1-\alpha_i) f(t_i|\mathbf{X},\beta)]^{C_i} [(1-\alpha_i) S(t_i|\mathbf{X},\beta)]^{1-C_i} \tag{7}$$

With the log-likelihood of:

$$\ln L = \sum_{i=1}^{N} \{C_i \ln [\alpha_i + (1-\alpha_i) f(t_i|\mathbf{X},\beta)] + (1-C_i) \ln [(1-\alpha_i) S(t_i|\mathbf{X},\beta)]\} \tag{8}$$

Where α can then be estimated via a binary-response function such as a probit, complementary log-log, or logit (defined below):

$$\alpha_i = \frac{\exp(\mathbf{Z}\gamma)}{1+\exp(\mathbf{Z}\gamma)} \tag{9}$$

Interestingly, if one were to define a probability of non-misclassification $(\delta = 1-\alpha)$ and substitute this quantity into Equation 8, the log-likelihood would be defined as:

$$\ln L = \sum_{i=1}^{N} \{C_i \ln [(1-\delta_i) + \delta_i f(t_i|\mathbf{X},\beta)] + (1-C_i) \ln [\delta_i S(t_i|\mathbf{X},\beta)]\} \tag{10}$$

which is symmetric to the log likelihood of the cure (i.e., split-population) survival model:

$$\ln L = \sum_{i=1}^{N} \left\{\tilde{C}_i \ln [\delta_i f(t_i|\mathbf{X},\beta)] + (1-\tilde{C}_i) \ln [(1-\delta_i) + \delta_i S(t_i|\mathbf{X},\beta)]\right\} \tag{11}$$

This implies that the general properties of the cure survival model also hold for the misclassification survival model, including (i) the reduction of the latter to a normal parametric model whenever $\delta = 1$ or $\alpha = 0$ and (ii) parameter identification even in the case where identical covariates are included in $\mathbf{Z}$ and $\mathbf{X}$.[1] Furthermore, note that the cure models are typically conceived as models for observations where some subset of the population is cured from ever experiencing the event of interest. By contrast, the survival model presented in 10 is in fact a model

[1] See Box-Steffensmeier and Zorn (1999, 5) for a discussion of these properties in the context of the cure model.

for instances where some subjects are observed as having failed or experienced the event of interest, even though they in actuality "live on" past their observed-failure point. Accordingly, I term the model presented in 10 as the parametric zombie survival model, as its usefulness applies to situations where one believes that one's observed event-failures to be contaminated with latent zombie-cases of this sort.

The Parametric Zombie Survival Model with Time Varying Covariates

Survival models, including the parametric versions and cure model referenced above, can typically be relaxed to include time-varying covariates by re-coding one's survival data such that an observation's within time periods of observation receive a unique start and stop time t_i, which are then coded as censored $\tilde{C}_i = 0$. This simple adjustment allows one to estimate the usual survival models, with no modification to the model itself, on such transformed data. I *think* this should work for the above Zombie Survival model as well, however, one alternative adjustment could also be useful here (I don't think it is, but haven't fully made up my mind), which I describe in detail immediately below.

While a time varying version of the parametric zombie survival model begins by requiring the same transformation to one's survival data, the zombie model also requires an additional adjustment. To see why, recall that the mixtures proposed above *only* arise for misclassification among final period-censored and event-experiencing cases. Hence, it would be non-realistic to expect that intermediate censored-coded cases (i.e., those that do not experience the event for which we observe subsequent information) to be misclassified. This therefore means one must adjust the zombie model further to accommodate for this fact. In a similar manner to extant latent class models (Holm, Jaeger and Pedersen, 2009) the added benefit of this adjustment, and the zombie model's application to time varying data more generally, is that the added structure in the data and model ensures the time varying zombie survival model is is more strongly identified than the otherwise fragile identification setting of non-time varying zombie survival model.

To do so, define a new indicator variable Q_i such that $Q_i = 1$ for an observation-period that is potentially misclassified (i.e., is censored due to existing until the end of the period

of observation or is recorded as terminated) and $Q_i = 0$ for an individual-period that is not potentially misclassified (i.e., is observed as both not experiencing the event not censored due to one's window of analysis closing). Accordingly, we can use Q_i to convert the formulas 7-8 to the likelihood and loglikelihood of the time varying zombie model as so:

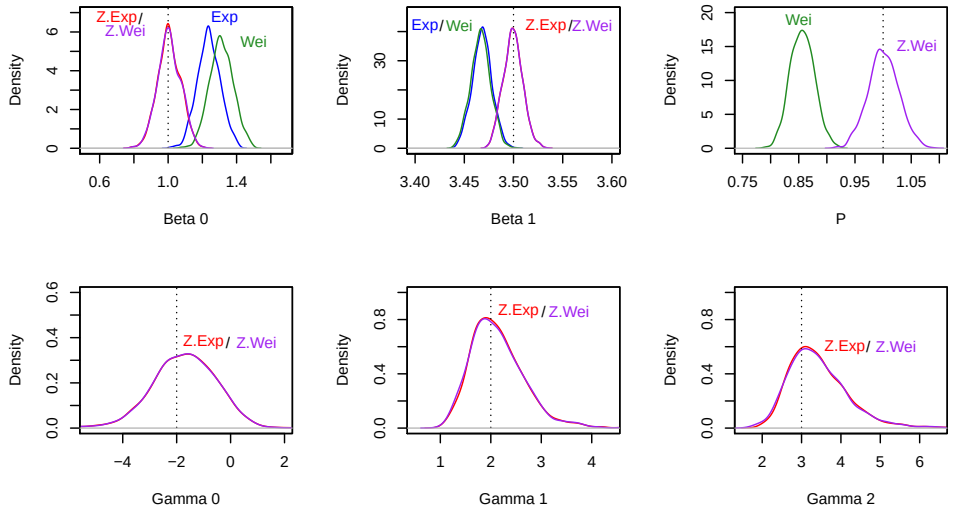
$$L = \prod_{i=1}^{N} [\alpha_i + (1-\alpha_i) f(t_i|\mathbf{X},\beta)]^{Q_i C_i} [(1-\alpha_i) S(t_i|\mathbf{X},\beta)]^{Q_i(1-C_i)} [S(t_i|\mathbf{X},\beta)]^{(1-Q_i)(1-C_i)} \quad (12)$$

With the log-likelihood of:

$$\begin{aligned} \ln L = \sum_{i=1}^{N} \Big(& Q_i C_i \ln [\alpha_i + (1-\alpha_i) f(t_i|\mathbf{X},\beta)] + Q_i(1-C_i) \ln [(1-\alpha_i) S(t_i|\mathbf{X},\beta)] \\ & + (1-Q_i)(1-C_i) \ln [S(t_i|\mathbf{X},\beta)] \Big) \end{aligned} \quad (13)$$

Monte Carlos

Mixture D.G.P.

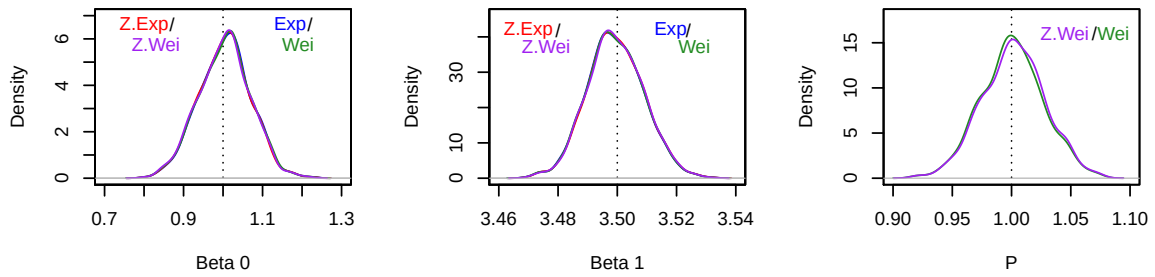


- Simulated 1,000 datasets ($n = 1,000$) with a zombie Weibull (z.Weibull) distributed dep. variable ($p = 1$, $\alpha = 5\%$) arising from $(\mathbf{1}, \mathbf{x_1})$ and $(\mathbf{1}, \mathbf{z_1}, \mathbf{z_2} \equiv \mathbf{x_1})$.

- Then compared exponential, Weibull, z.exponential, and z.Weibull estimates.

- Exponential/Weibull noticeably more biased than z.exponential/z.Weibull.

Non-Mixture D.G.P.



- Simulated 1,000 datasets ($n = 1,000$) with a Weibull distributed dependent variable ($p = 1$) arising from $(\mathbf{1}, \mathbf{x_1})$.
- Then compared exponential, Weibull, z.exponential, and z.Weibull estimates.
- Each model performs comparably, although the z.exponential and z.Weibull models have non-negligible convergence problems ($\approx 27\%$ of sims).

Applications

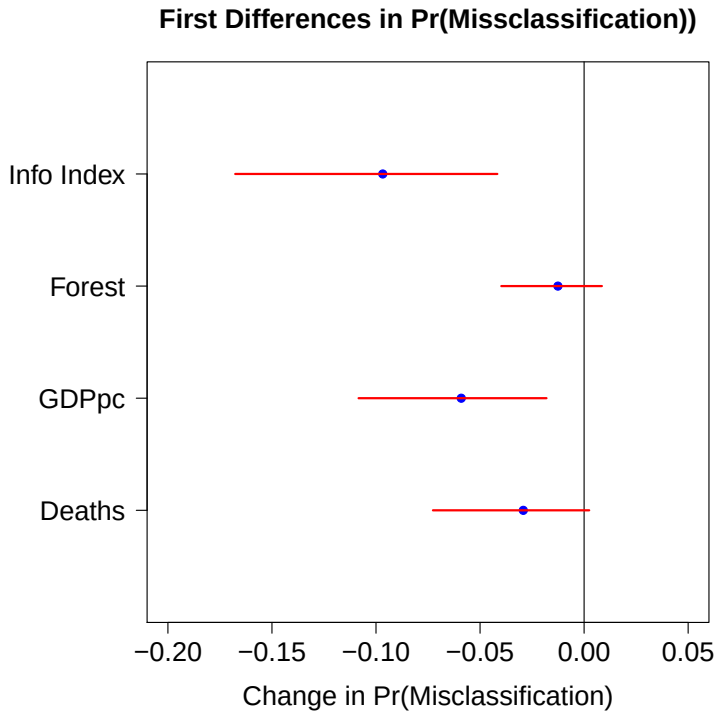
Application 1: Civil War Data

Overview

- E.g., Collier, Hoeffler and Söderbom (2004); de Rouen and Sobek (2004); Buhaug, Gates and Lujala (2009); Cunningham, Gleditsch and Salehyan (2009); Thyne (2012).
- Typically analyze the durations of rebel-government conflicts, 1946-2004.
- End dates ("failures") for specific conflicts are often recorded based upon 24-month spells with fewer than 25 battle-deaths per year.
- This may be overly conservative, especially for lower-intensity civil wars in remote or poor information environments.
- I assess this potential below by replicating Thyne (2012; *Model 1*) using a collection of Weibull and z.Weibull models.

- Thyne's survival covariates were: *info index*, *ln battle deaths*, *ln GDPpc*, *coups*, *ln forest*, *fight for gov.*, & *opposition vetos.*

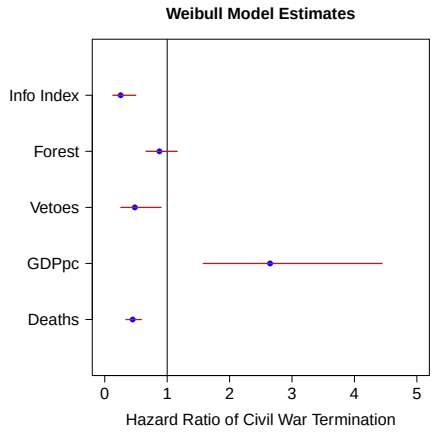
As for misclassification predictors, I estimated multiple z.Weibull models using either a selection of, or all of, Thyne's survival stage covariates as misclassification predictors (i.e., **Z**'s). The first differences in misclassification probabilities for the best fitting specification are presented here, given mean to mean+1SD changes in each misclassification stage covariate. They show that civil wars fought in more economically developed countries, countries with higher political information environments, and (to an extent) more violent civil wars are less likely to be misclassified as having terminated when they (possibly) had not.

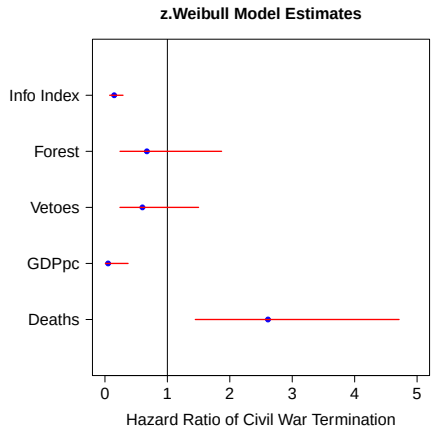


As for the survival stage estimates, the z.Weibull models reveal several unique & counterintuitive survival stage findings. However, these results are not particularly robust across model specifications. Anticipated next steps: improve my misclassification specifications & survival stage interpretations, add a validation component, and explore additional applications.

Application 2: Historical Data

Possible sources to draw upon for civilization survival include: Reuveny (2012), Cioffi-Revilla (1996), Arbesman (2011), Cioffi-Revilla and Landman (1999), Cioffi-Revilla and Lai





(1999), Whittington (1991).

References

Arbesman, Samuel. 2011. "The Life-Spans of Empires." Historical Methods 44(3):127–129.

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

Box-Steffensmeier, Janet M. and Chrisopher Zorn. 1999. "Modeling Heterogeneity in Duration Models." Paper Presented at the 1999 Summer Meeting of The Political Methodology Society.

Buhaug, Halvard, Scott Gates and Päivi Lujala. 2009. "Geography, Rebel Capability, and the Duration of Civil Conflict." Journal of Conflict Resolution 53(4):544–569.

Cioffi-Revilla, Claudio. 1996. "Origins and Evolution of War and Politics." International Studies Quarterly 40(1):1–22.

Cioffi-Revilla, Claudio and David Lai. 1999. "Evolution of Maya Polities in the Ancient Mesoamerican System." International Interactions 26(4):347–378.

Cioffi-Revilla, Claudio and Todd Landman. 1999. "Evolution of Maya Polities in the Ancient Mesoamerican System." International Studies Quarterly 43(4):559–598.

Collier, Paul, Anke Hoeffler and Måns Söderbom. 2004. "On the Duration of Civil War." Journal of Peace Research 41(3):253–273.

Cunningham, David E., Kristian Skrede Gleditsch and Idean Salehyan. 2009. "It Takes Two: A Dyadic Analysis of Civil War Duration and Outcome." Journal of Conflict Resolution 53(4):570–597.

de Rouen, Karl R. and David Sobek. 2004. "The Dynamics of Civil War Duration and Outcome." Journal of Peace Research 41(3):303–320.

Hannan, Michael T. and John Freeman. 1988. "The Ecology of Organizational Mortality: American Labor Unions, 1836-1985." American Journal of Sociology 94(1):25–52.

Holm, Anders, Mads Meier Jaeger and Morten Pedersen. 2009. "Unobserved Heterogeneity in the Binary Logit Model With Cross-Sectional data and Short Panels: A Finite Mixture Approach." Working Paper.

Leblang, David and William Bernhard. 2000. "The Politics of Speculative Attacks in Industrial Democracies." International Organization 54(2):291–324.

Lie, Stein Atle and Rolf W. Lie. 2013. "Changes in survival of cattle (Bos taurus) during Medieval times in two Norwegian cities." Environmental Archaeology 18(2):178–183.

Reuveny, Rafael. 2012. "Taking Stock of Malthus: Modeling the Collapse of Historical Civilizations." Annual Review of Resource Economics 4:303–329.

Thyne, Clayton L. 2012. "Information, Commitment, and Intra-War Bargaining: The Effect of Governmental Constraints on Civil War Duration." International Studies Quarterly 56(2):307–321.

Whittington, Stephen L. 1991. "Detection of significant demographic differences between subpopulations of prehispanic Maya from Copan, Honduras, by survival analysis." American Journal of Physical Anthropology 85(1):167–184.