DISCRETE MEASUREMENT, CONTINUOUS TIME AND EVENT HISTORY MODELING

Bruce A. Desmarais

A thesis submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Political Science in the Department of Political Science.

Chapel Hill
2008

Approved by:

Thomas Carsey

James Stimson

Georg Vanberg

Abstract

Bruce Desmarais: Discrete Measurement, Continuous Time and Event History Modeling
(Under the direction of Thomas Carsey, James Stimson and Georg Vanberg)

Most even history models used in political science assume the time being analyzed is continuous. Discrete measurement causes this assumption to be violated. The violation of this assumption is shown to introduce non-trivial bias to parameter estimates. Analysis of discrete-measured data as interval-censored is shown to greatly reduce this bias. The empirical properties of the bias introduced by discrete measurement and the interval-censoring correction are explored through Monte-Carlo simulations and a replication of the analysis of civil war duration from (Fearon 2004). I also demonstrate that analyzing discrete-measured continuous-time data as interval-censored is a better approach than the discrete-time models proposed in (Box-Steffensmeier and Jones 2004). The conclusion of the analysis is that event-history analysis of continuous-time variables should always be implemented as interval-censored estimation.

Table of Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

δ	Binary Indicator
D/d	True Duration
f()	Probability Density Function
F()	Cumulative Distribution Function
L()	Likelihood Function
λ	Scale Parameter
M/m	Measured Duration
S()	Survival Function
θ	Vector of Parameters

I. Discrete Measurement

Time-to-event models have come to play an important role in empirical political research. The majority of the estimators used in these studies, whether parametric or semi-parametric, assume that the time being modeled follows a continuous distribution. Though this continuity assumption may hold in theory, in practice time is always measured discretely. This discrete measurement means that the value of the temporal variable as measured is indicative of a possible range of the continuous variable. Discrete measurement of a continuous time variable creates an interval-censored duration sample, meaning the true values of the durations are unknown, but it is known what interval in which they lie. Neglect of this sampling limitation causes biased parameter estimates. In this paper I demonstrate this bias, illustrate the use of interval censored survival estimators to correct for this bias, and re-analyze data previously published to demonstrate the effect of the correction.

The Problem

In this section I develop the problem that derives from imposing discrete measurement on a continuous-time process. In the usual manner, the duration that serves as the time under study is constructed by subtracting a start time from an end time. These start and end times are approximated by a discrete measurement process. That is, the researcher has some measurement tool, a calendar or stopwatch for instance, and the time that the tool reads when the event starts and ends are the values assigned to the start and

end times respectively. If the precision of the measurement tool is annual then all events that start between January, 1 2000 and December, 31 2000, and end between January 1, 2001 and December, 31 2001 will be assigned a duration of one (2001-2000 =1). Note that the duration that leads to this measurement of one year could actually be as short as one day (12/31/2000-1/1/2001) or as long as one day short of two years (1/1/2000-12/31/2001). Discrete measurement can cause major distortions related to the ordering of measured durations. Given discrete measurement, if two durations (A and B) are measured to be equal or adjacent, it is possible that A > B, B >A, or B = A. It was demonstrated above that events of different length can be falsely assigned the same duration. The following is an example of how the order of durations can actually be opposite of that measured. Consider one conflict (A) that begins in 1991 and ends in 1992, thus receiving a duration of one, and another (B) that starts in 1991 and ends in 1993, receiving a duration of two. If A starts on 1/1/1991 and ends on 12/31/1992, its daily duration is 730 days, and if B starts on 12/31/1991 and ends on 1/1/1993, it has a daily duration of 367 days. In annual terms, A is only one half the duration of B, but in daily terms, A is nearly twice as long as B. This example demonstrates the most extreme distortion that can be introduced via discrete measurement.

More generally, the discrete measurement of the duration of an event leads to a known interval within which the actual duration falls.[1] This interval arises from the two intervals indicated by the discrete measurement of the start and end times. If an event is measured to start at time t_s, its exact start time is between t_s and t_s +1. If it is measured to

[1] In all of the examples and derivations discussed hereon it is assumed that the precision of measurement has been scaled to the integer level (e.g. if the precision is daily and event that lasts one day is of duration one, not 1/365 years. This assumption retains the generality of results and reduces the amount of jargon in the discussion.

end at time t_e, the actual end time t_e is known to be in t_e and t_e+1. These two intervals can be used to construct the known upper and lower bounds within which the actual duration lies. The longest possible duration, given (t_s, t_e) is given by the latest possible end time minus the earliest possible start time ((t_e+1)-t_s) and the shortest possible duration is given by the earliest end time minus the latest start time (t_e – (t_s +1)). This gives an interval within which the exact duration lies. The interval for the true duration D is given in (1).

$$(t_e - t_s - 1 \le D \le t_e - t_s + 1) \tag{1}$$

On each side of the inequality in (1) is given the expression that amounts to the standard discrete measurement of a duration; that is, the subtraction of the discrete measured start time from the discrete measured end time. Consider the expression (t_e – t_s) to be equal to the discrete measured duration M. Then, given a discrete measurement M, the known interval for D is given in (2).

$$\begin{aligned} &(M - 1 \le D \le M + 1) \\ &\Leftrightarrow D \in (M - 1, M + 1) \end{aligned} \tag{2}$$

The General Solution

Interval censoring is the condition where the value of a variable is known to lie within a specific interval, but its exact value cannot be determined. The likelihood function of discretely measured continuous duration data can be constructed as an interval censored likelihood function. Derived from the interval given in (2), if the measured duration M > 0 the probability of observing the duration M is equal to the probability of the true duration D being between (M-1) and (M+1) (assuming the measurement precision has been scaled to the integer level), and the likelihood of

observing M = 0 is the probability of D being between zero and one. The probability of observing M is given by (3).[2]

$$\begin{aligned} &P(M) = P(D \in (M-1, M+1)) \\ &\Leftrightarrow P(M) = F(M+1) - F(M-1) \end{aligned} \tag{3}$$

Where F() is the cumulative distribution function of the true event times D. Where θ is a vector of parameters to be estimated, the likelihood function for discretely measured continuous time duration data, derived from (3) is given in (4).

$$L(M \mid \theta) = \prod_i F(m_i + 1 \mid \theta) - F(m_i - 1 \mid \theta) \tag{4}$$

The likelihood function where the discrete-measured durations are treated as exact measurements is given in (5), where f() is the PDF of the assumed distribution of the durations. The parameter vector that maximizes (5) is different from that which maximizes (4).

$$L(M \mid \theta) = \prod_i f(m_i \mid \theta) \tag{5}$$

Midpoint imputation occurs when it is assumed that the value of a variable is the midpoint of the interval within which it is known to lie. Since observing a duration M amounts to gaining the knowledge that the true duration D is between M-1 and M+1, assuming the value is actually M (the midpoint between M-1 and M+1) amounts to midpoint imputation. The literature that deals with the general case of interval censoring in general has addressed the use of midpoint imputation in the case of imprecise measurement of a duration. This work has demonstrated that non-trivial bias is introduced to interval censored estimation in event-history models, both parametric and

[2] When the measured duration is zero, it is known that the even begins and ends in the same measurement interval, and thus the interval within which the duration is known to lie is (0,1).

semi-parametric, when midpoint imputation is utilized (Goggins and Finkelstein 2000; Odell, Anderson and D'Agostino 1992; Goggins, Finkelstein, Schoenfeld and Zaslavsky 1998; Kim 1997). The current analysis demonstrates that techniques developed to deal with interval-censored data can help to mitigate limitations imposed by discrete measurement.

A Case Study; the Exponential Distribution

To provide a simple analytic demonstration of bias in parameter estimation due to midpoint imputation, here I study the likelihood function and maximum likelihood estimator for the exact and interval-censored estimation of the scale parameter of the exponential distribution. The bias in using the exact estimator (midpoint imputation) when interval-censoring is present, is given as the difference between the MLE's under interval censoring and exact estimation. The PDF of the exponential distribution is given in (6).

$$\lambda e^{-\lambda t} \tag{6}$$

Where (t) is the duration and $\lambda > 0$ is the scale parameter to be estimated. The likelihood function is the product over the PDF evaluated at each duration in the sample, and follows directly from (6).

$$L(T \mid \lambda) = \prod_i \lambda e^{-\lambda t_i} \tag{7}$$

The CDF of the exponential distribution is given in (8).

$$1 - e^{-\lambda t} \tag{8}$$

Scaling the measurements to the integer level of precision, the interval-censored likelihood function is derived by substituting (8) for F() in (4).

$$L(T \mid \lambda) = \prod_i (1 - e^{-\lambda(t_i+1)}) - (1 - e^{-\lambda(t_i-1)}) \tag{9}$$

The MLEs for (7) and (9) are derived in Appendix I. Comparison of these demonstrates that the parameter estimated from (7) does not equal that estimated from (9). The difference between the estimated parameters is the bias introduced by using the exact likelihood function (7), when the interval-censored likelihood function (9) is appropriate. The MLEs of (7) and (9) are given in (10) and (11) respectively.

$$\lambda_{exact} = \frac{1}{\mu} \tag{10}$$

$$\lambda_{IC} = \frac{1}{2}\ln(\frac{2}{\mu - 1} + 1) \tag{11}$$

Where μ is the sample mean, the bias introduced by using midpoint imputation is given by the difference between the MLE's.

$$\lambda_{exact} - \lambda_{IC} = \frac{1}{\mu} - \frac{1}{2}\ln(\frac{2}{\mu - 1} + 1) \tag{12}$$

The difference between the two estimators (12) is not zero, and thus bias is introduced via midpoint-imputed estimation of the scale parameter of the exponential distribution with data that is interval-censored through discrete measurement. Unfortunately, many of the more common distributions used to model duration data (weibull, lognormal, gamma etc.) do not permit a closed form derivation of the interval-censored MLE, though the exponential is not completely without application as it is a special case of both the weibull and gamma distributions (Box-Steffensmeier and Jones 1997).

Right Censoring and Time-Varying Covariates

The likelihood functions in (4), (5), (7) and (9) are appropriate for estimating the parameters in a sample where all of the observations end and all of the covariates included in the model are constant over time. A major advantage of event-history modeling is that it is possible to analyze observations that either do or do not fail by some observed time (M) (Box-Steffensmeier and Jones 1997). Observations that have not failed at the time of observation are said to be right-censored. The analysis of right-censored observations permits the inclusions of covariates that change over time for each observation (time-varying covariates). For instance, if one were modeling the survival of democratic states, an event-history model would allow the annual survival of a state to be conditioned on that state's annual gross domestic product.

The likelihood-functions presented above need to be augmented to accommodate right-censored observations. When a right-censored observation (at time M) is included in a sample it contributes the information that its measured duration is greater than M, since it has not failed by time M. The probability that a duration D is greater than M can be expressed as one minus the probability that D is less than or equal to M. The probability that D is less than or equal to M is the cumulative distribution (F()) of D evaluated at M. One minus the cumulative distribution is the survival distribution (Box-Steffensmeier and Jones 1997). The contribution of a right-censored observation, observed at time M, to the likelihood function is the survival distribution (S()) evaluated at time M. Where θ is a vector of parameters to be estimated, the contribution of a right-censored observation to the likelihood function is given in (13).

$$S(M \mid \theta) = 1 - F(M \mid \theta) \tag{13}$$

This contribution is integrated with the likelihood contribution of an uncensored observation in (5) to create the likelihood function that accommodates both right-censored and uncensored observations. Where δ is an indicator that assumes a value of one if the observation fails and zero otherwise, the likelihood function that accommodates right-censoring is given in (14).

$$L(M \mid \theta) = \prod_i f(m_i \mid \theta)^{\delta_i} S(m_i \mid \theta)^{1-\delta_i} \tag{14}$$

This likelihood needs to be altered further to accommodate discrete measurement. With discrete measurement, there are no uncensored observations. An observation is either right or interval-censored. It was developed above that when an observation is observed to fail at time M with discrete measurement, its exact duration is known to fall within (M-1) and (M+1). This uncertainty exists because the exact observed time in the life of the event under study is known to lie in (M-1, M+1), but is not known exactly. If the end of the event does not occur at time M, it is known that its end occurs at some observation point between (M+1) and ∞. Thus, given that an observation does not fail by time M, it is known that its failure occurs only as soon as (M+1). If the duration were recorded at (M+1), from (2) it is known that the actual duration would be in (M, M+2). Therefore, if an event is known to not end by discrete-measured time M, its true duration is known to be at least M. This leads to a convenient result; the survival distribution of a right-censored, exactly-measured variable is equivalent to the survival distribution of a right-censored, discrete-measured variable. Therefore, the likelihood function (15) for a sample of right and interval-censored discrete-measured observations is created by combining (13) and (4), where δ is again the indicator of failure.

$$L(M \mid \theta) = \prod_i (F(m_i + 1 \mid \theta) - F(m_i - 1 \mid \theta))^{\delta_i} S(m_i \mid \theta)^{1-\delta_i} \quad (15)$$

To provide an example of the likelihood function with accommodation for right-censored observations, I again use the exponential distribution. F() and f() for the exponential distribution are given in (8) and (6) respectively. Substituting these into (14) and (15) gives the likelihood function for exponentially distributed uncensored and right-censored observations (16) and the likelihood for interval-censored and right-censored discrete-measured observations (17).

$$L(M \mid \theta) = \prod_i (\lambda e^{-\lambda m_i})^{\delta_i} (e^{-\lambda m_i})^{1-\delta_i} \quad (16)$$

$$L(M \mid \theta) = \prod_i (e^{-\lambda m_i - 1} - e^{-\lambda m_i + 1})^{\delta_i} (e^{-\lambda m_i})^{1-\delta_i} \quad (17)$$

The above derivations demonstrate analytically that event-history models are biased due to discrete measurement and that this bias is efficiently corrected. I note the efficiency of this correction because the interval-censored estimator does not require any more parameters to be estimated than the midpoint imputed estimator. The empirical properties of interval-censoring through discrete measurement have yet to be explored. Through a simulation study and a replication of a previously published analysis, the next two sections explore the problems that discrete measurement imposes upon more common estimators used in political science. In section IV, I discuss the application of interval-censoring methods to discrete-time models.

II. Simulation Study

Simulation Design

In this section I compare the performance of midpoint imputation and interval censored estimation, also performing pseudo-continuous estimation as a baseline for comparison.[3] The three different sampling/estimation techniques are compared within the context of three models common to the use of duration models in political science; the Cox proportional-hazards, the weibull (Accelerated failure time metric), and the lognormal models (Regan 2002; Bennett and Stam 1996; Bolks and Al-Sowayel 2000; Kadera, Crescenzi and Shannon 2003). The Cox PH model is the standard semi-parametric model employed in political-science. Researchers are drawn to this model because no assumption regarding the distribution of failure-times is made. The only assumption is that covariates affect the duration multiplicatively. The weibull and lognormal models are popular parametric estimators that rely upon distributional assumptions, but have the strength that they explicitly model duration dependency and allow for precise predicted durations. The weibull model allows for the hazard rate to be either monotonic increasing or monotonic decreasing in time, and the lognormal model estimates a hazard rate that is non-monotonic in time. The simulations executed are applicable to much of the empirical political science literature.

[3] The pseudo-continuous variates are generated at double-byte precision, which means values are precise to sixty-four digits, as close as most statistical software can come to exact accurracy.

For each model, a distribution conditional upon a single, standard-normally-distributed covariate is generated. The pseudo-continuous estimation is performed on this simulated distribution. Further manipulation is needed to produce the discrete-measured variable. Discrete measurement is simulated by adding a uniform(0,1) disturbance to this simulated distribution and rounding the sum down to the nearest integer.[4] The integer can be thought of as the day, month, year etc. at which the precision of the measurement tool is calibrated, and the decimal values can be thought of as the partial day, month year etc. that are lost through discrete measurement (e.g. if the measurement precision is annual, any event that occurs between the beginning 1991 and 1992 is measured as 1991). The uniform disturbance is the simulated start time within the first interval. Midpoint imputed estimation proceeds by analyzing the rounded variable, and interval censored estimation analyzes the dependent variable as if it lies within an interval of two, centered around the rounded variable.[5] Table 1 gives the details of the simulated distributions. The parameter values chosen are intended to impose duration dependence on the data and produce distributions that resemble daily, monthly and annual measures, with values centered around 1000, 300, and 30 respectively.

Table 1) Simulated Distributions

Model	Simulated Model	Parameter Value (β)
Cox-PH	$Y \sim \text{Weibull}(\rho=3.5, \lambda = e^{3.5(-.002+.03x)})$	0.3
Weibull AFT[6]	$Y = e^{-2+2x+6\sigma}$	2.0
Lognormal[7]	$Y = e^{-2+2x+5\pi}$	2.0

[4] Values of the simulated distribution plus the uniform disturbance below one are rounded to one.

[5] In the case that the simulated discrete-measured variable equals zero, the interval is (0,1)

[6] σ is a type-one extreme value disturbance

[7] π is a normal disturbance with mean = 0 and SD =1.

In the simulation study, each model is simulated one thousand times under each sampling condition to ascertain the sampling distribution of the coefficient on the covariate (x). The simulation is performed for each model on a sample of one hundred and a sample of one thousand in order to compare the effect of sample size on the relative properties of the sampling schemes. Figure 1 provides a simple diagram of the study organization. This study is executed for the Cox PH, weibull and lognormal models. Table 2 presents descriptive statistics for the simulated distributions.

Simulation Results

Table 3 presents summary results from the eighteen simulations.[8] The pseudo-continuous results represent the performance of the respective models under the best possible conditions for the sample sizes simulated. The standard for comparison of the performance of the estimators is the root mean squared-error RMSE(β) (Casella and Lehmann, 1999). This statistic approximates, on average, how close the parameter estimated is to the true parameter, meaning the smaller the RMSE, the better the overall performance of the estimator.[9] In each simulation, based on the RMSE, the interval-censored estimator (ICE) outperforms the midpoint imputed estimator (MIE). The RMSE of the MIE ranges between 119% and 673% of the RMSE of the ICE. The difference in the RMSE of the ICE and MIE is larger in the samples of one thousand, indicating that as the sample size increases, the correction for interval censoring due to discrete

[8] The simulations were executed in R 2.6.1. See the appendix for a sample of the simulation code.

[9] Where n is the number of estimated parameters, e is the value of the estimated parameter, and p is the true value of the parameter, the RMSE of an estimator is computed as:

$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{(e_i - p)^2}$$

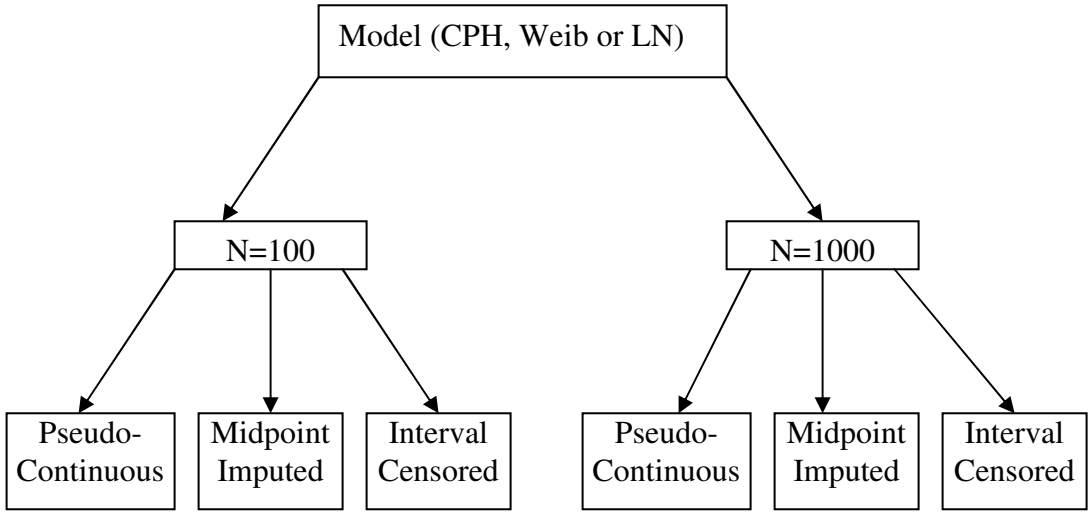
measurement becomes more important in terms of retrieving the true parameter. For instance, in the Cox-PH simulation with sample size equal to 100, the RMSE of the midpoint-imputed estimator is 0.11 and that of the interval-censored estimator is 0.092. The corresponding values for the sample size of 1000 were 0.105 and 0.028 respectively. In terms of the RMSE, the ICE is clearly superior to the MIE.

Table 2) Simulated Distribution Descriptive Statistics

Distribution	Mean	Median	SD	Unique Values
Weibull*	24.12	16.99	19.15	100
Weibull* Discrete	24.2	17	19.1	48
Weibull AFT	271.72	195.47	241.36	100
Weibull AFT Discret	271.82	195	241.37	89
Lognormal	1263.86	1018.75	801.17	100
Lognormal Discrete	1263.89	1018	801.18	97

Descriptive statistics are estimated for a distribution from a single run of the simulation. Statistics shown are for the sample size of 100. These statistics change little for the sample size of 1,000. * Distribution generated for estimation with Cox PH model.

Figure 1) Simulation Design



An alternative measure of estimator performance is bias. In the six model/sample size combinations, the mean estimate produced by the pseudo-continuous estimator is different from the true parameter (based on a 0.05 level two-tailed z-test) in one simulation, the ICE differs in two simulations, and the MIE is different from the true parameter in every simulation. The bias of both the midpoint-imputed and interval-censored estimators is worst in the lognormal model. In the sample size of 100, the bias in the ICE and MIE are 0.101 and -1.29 respectively, and the corresponding values for the sample size of 1000 are 0.12 and -1.29 respectively. To place the bias in relative terms, on average the distance between the mean of the ICE estimates and the true value of the parameter is 4.07% of the distance between the mean of the MIE estimates and the true parameter value. Overall, the potential for bias due to discrete measurement is much less with ICE than with MIE.

Another characteristic of the estimators that is clearly demonstrated by the simulations is that the variance of the ICE is greater than the variance of the MIE. The variance of the ICE is on average 1.96 times that of the MIE. In every simulation an F-Variance-Ratio test rejects the null hypothesis of equal variances in favor of the ICE having the larger variance. The larger variance in the ICE accurately represents the uncertainty regarding the value of the dependent variable. The simulations indicate that the variance in the MIE is always lower than the variance in the ICE, and is statistically significantly lower than the variance of the pseudo-continuous estimator in five of the six simulations. What this means is that use of the midpoint-imputed estimator produces biased parameter estimates and exaggerated confidence in those estimates. To

demonstrate the practical applications of interval-censored estimation, in the following section, I use interval censored estimation on a duration dataset from a previously published analysis of discrete-measured data.

Table 3) Simulation Results by Sample Size and Estimator

	Sample Size =100				Sample Size =1000			
	Cox PH Model (β = 0.3)							
	Mean(β)	Bias(β)	SD(β)	RMSE(β)	Mean(β)	Bias(β)	SD(β)	RMSE(β)
Continuous	0.307*	0.007	0.0034	0.086	0.302	0.002	0.033	0.026
MIE	0.195*	-0.105	0.075	0.11	0.195*	-0.105	0.024	0.105
ICE	0.303	0.003	0.118	0.092	0.299	-0.001	0.035	0.028
	Weibull AFT (β = 2.0)							
	Mean(β)	Bias(β)	SD(β)	RMSE(β)	Mean(β)	Bias(β)	SD(β)	RMSE(β)
Continuous	1.99	-0.01	0.634	0.5	1.99	-0.01	0.187	0.15
MIE	1.07*	-0.93	0.405	0.93	1.13*	-0.87	0.134	0.86
ICE	2	0	0.694	0.55	1.97*	-0.03	0.219	0.18
	Lognormal (β = 2.0)							
	Mean(β)	Bias(β)	SD(β)	RMSE(β)	Mean(β)	Bias(β)	SD(β)	RMSE(β)
Continuous	1.99	-0.01	0.483	0.385	2	0	0.162	0.13
MIE	0.714*	-1.286	0.262	1.286	0.712*	-1.288	0.081	1.28
ICE	2.101	0.101	0.725	0.572	2.12*	0.12	0.21	0.196

Mean(β) is the average estimated parameter. Bias(β) is the difference between the average estimated parameter and the parameter's true value. SD(β) is the standard deviation of the sample of estimated parameters. RMSE(β) ia the root-mean-squared-error of the estimated parameters. MIE = Midpoint Imptuted Estimator, and ICE = Interval Censored estimator *Mean parameter estimate is different from the true parameter at the 0.05 significance level

III. The Duration of Civil War

A very common application of duration models in political science is in the analysis of the duration of war/peace (Balch-Lindsay and Enterline 2000; DeRouen and Sobek 2004; Hegre 2004; Cunningham 2006). In this section I replicate the statistical analyses from a study of the duration of civil war (Fearon 2004), using both the original midpoint-imputed estimator and an interval-censored estimator. In the original analysis, Fearon (2004) used the weibull AFT model to analyze the impact of a number of covariates on the annual duration of civil war. Table 4 describes the covariates used in the estimation.[10]

Table 4). Description of the Predictors of Civil War Duration

Variable	Description
Coup/Revolution	Civil war is the result of a coup or revolution
Eastern Europe	War involves an Eastern European country
Not Contiguous	Civil war is between two parties that are not territorially contiguous (such as a colonial revolt)
Sons of the Soil	Dispute involves territorial or natural resource claims
Ethnic Fractionalization	Level of ethnic diversity in the country in which the war takes place

Table 5 presents the estimates using midpoint imputation and interval-censoring for models I, II, and IV from the original Fearon (2004) article. For each model, the first column reports the midpoint imputed estimates (MIE), the second the interval-censored

[10] See original article for the theoretical justification for the inclusion of the covariates

estimates (ICE) and the third column gives the percentage difference between the ICE and the MIE. All of the interval-censored parameter estimates and standard errors differ from the midpoint-imputed estimates. The coefficients have a straightforward interpretation; they represent the expected change in the natural log of the duration of civil war due to a one unit increase in the independent variable. The differences in the MIE and ICE coefficients and standard errors for the Eastern Europe covariate in model IV are typical of the differences across parameters and models. The ICE coefficient (-0.375) is 7% larger (in absolute value terms) than that in the MIE (-0.349) and the ICE standard error (0.29) is 17% larger than that in the MIE (0.247). Most of the coefficients and all of the standard errors are biased towards zero in the midpoint imputed estimates, but the bias in the standard errors is much larger than that in the coefficients. The percentage difference between the interval censored and midpoint imputed estimates of the covariate effects varies between 5% and 15% and the difference in the standard errors of these effects ranges between 15% and 21%. A final comparison of the midpoint imputed and interval censored estimates in table 5 shows that the interval censored estimator produces a considerably smaller log-likelihood value than the midpoint imputed estimator (approximately 65% lower). It cannot be interpreted that the higher log-likelihood is indicative of a better fitting model. This difference is a result of the increased (yet false) certainty about the value of the dependent variable and thus the parameters that best describe its distribution in the midpoint imputed case. The same false certainty that causes the downward bias in the standard errors of the parameter estimates causes upward bias in the log-likelihood.

Quantitative studies in political science primarily use critical values of Z or T statistics to test for the presence of the effect of a covariate on a dependent variable of interest. Table 6 in Appendix III presents the Z-statistics for the parameters estimated in the MIE and ICE models of civil war duration as well as the percentage difference between the two. In all cases, the Z-statistic is lower in the ICE than the MIE (a decrease of 10% on average for the covariate-effect parameters), supporting the conclusion that midpoint impuation biases results in the direction of a type I inference error (e.g. concluding there is an effect when one does not exist). In terms of covariate effects, there is one parameter that is significant at the 0.05 level in the MIE and not in the ICE in one model. In model II, the effect of contiguousness between civil war combatants is statistically significantly negative at the 0.05 level (two-tailed) in the MIE and not in the ICE. Another substantive conclusion regarding the duration of civil war is changed by the implementation of interval-censored estimation. The weibull model estimates a parameter ($\ln(\rho)$) that describes the way in which the risk of a civil war ending changes over time (duration dependency). When this parameter is statistically significantly greater (less) than zero, the risk of a civil war ending is monotonically increasing (decreasing) over time. If this parameter is not significantly different from zero, the risk of civil war ending does not change over time, and the distribution of the duration of war reduces to an exponential distribution. Over the three MIE models, this parameter is estimated to be around 0.19, and is significantly greater than zero in each model, suggesting that the risk of a civil war ending is increasing over time. This parameter is not significantly different from zero in any of the interval-censored models, implying both that the risk of a civil war ending does not change over time and that the duration of civil war is exponentially

distributed. In this particular case, the use of MIE produces falsities in our substantive understanding of the duration of civil war.

Table 5) Comparison of Midpoint-Imputed and Interval-Censored Estimators of the Duration of Civil War

	Model I			Model II			Model IV		
Independent Variable	Midpoint Imputed	Interval Censored	%Difference	Midpoint Imputed	Interval Censored	%Difference	Midpoint Imputed	Interval Censored	%Difference
Coup or Revolution	-1.14	-1.29	13%	-1.06	-1.2	13%	-1.17	-1.31	12%
	0.213	*0.256*	20%	*0.218*	*0.261*	20%	*0.224*	*0.27*	21%
Eastern Europe	-1.11	-1.16	5%	-1.13	-1.19	5%	-1.09	-1.15	6%
	0.263	*0.314*	19%	*0.262*	*0.313*	19%	*0.265*	*0.316*	19%
Non-Contigious	-0.38	-0.396	4%	-0.526	-0.567	8%	-0.349	-0.375	7%
	0.235	*0.275*	17%	*0.26*	*0.305*	17%	*0.247*	*0.29*	17%
Sons of the Soil	1.13	1.22	8%	1.15	1.24	8%	1.13	1.23	9%
	0.293	*0.343*	17%	*0.291*	*0.341*	17%	*0.293*	*0.343*	17%
Contraband	0.941	1.03	9%	0.943	1.03	9%	0.944	1.03	9%
	0.341	*0.401*	18%	*0.337*	*0.396*	18%	*0.34*	*0.4*	18%
Ethnic War				0.436	0.5	15%			
				0.321	*0.376*	17%			
Log (Lagged Population)							-0.024	-0.016	-33%
							0.065	*0.075*	15%
Constant	2.35	2.29	-3%	2.13	2.04	-4%	2.57	2.45	-5%
	0.123	*0.145*	18%	*0.196*	*0.229*	17%	*0.636*	*0.745*	17%
Ln(p) (Duration Dependence)	0.188	0.029	-85%	0.198	0.038	-81%	0.191	0.031	-84%
	0.076	*0.086*	13%	*0.076*	*0.087*	14%	*0.076*	*0.087*	14%
Log-Lilkelihood	-160.536	-251.377		-159.63	-250.5		-160.47	-251.35	
N(ended)	128(103)	128(103)		128(103)	128(103)		128(103)	128(103)	

Coefficient in Weibull (Accelerated Failure Time) regression reported with standard errors in italics. Model numbers correspond to the models in (Fearon 2004). % Difference = (ICE-MIE)/(MIE)

IV. Discrete Time or Discrete Measurement of Time?

In this section I present an argument that discrete-time duration models should never be used for discretely measured continuous time models, but rather the data generated by discrete measurement of temporal variables should be treated as interval-censored continuous data. In their book on event history modeling, Box-Steffensmeier and Jones (2004) (BSJ) present a detailed chapter on the analysis of discrete-time. One motivation given for the use of discrete time models is the discrete measurement of truly continuous time. Specifically, BSJ present the monthly duration of cabinet governments as an appropriate sample for discrete-time survival analysis. The approach to discrete-time models they present is to use one of the familiar binary-choice models (logit, probit, log-log etc.) to model the hazard rate (the simultaneous rate of failure at a given time). This involves keeping each event in the dataset with a value of zero for the dependent variable for each period up to the period where the event occurs, at which time the dependent variable assumes a value of one and the observation is subsequently removed from the dataset.[11] If the variable under analysis is a discrete-measured continuous-time duration and not a truly discrete variable, this approach cannot accurately estimate the hazard rate. The hazard rate of a duration (d) is given by (18).

$$h(d) = \frac{f(d)}{S(d)} \tag{18}$$

[11] The observations are subject-time much like in a panel model, except here the subjects are removed from the dataset once the event occurs.

Where f() is the PDF and S() is the survival distribution or (1-CDF). The reasoning behind estimating the hazard rate via a binary choice model is that when the dependent variable is equal to one it is known that the event occurred at the time that corresponds with the dependent variable being equal to one, and it is known that the event did not occur prior to that (e.g the probability of failure at time M, due to the structure of the dataset, is conditioned by the knowledge that the duration is at least M) (Box-Steffensmeier and Jones 2004). The information conveyed when the dependent variable is equal to one at measured time (m) in a binary choice subject-time model is expressed in (19).

$$\frac{\Pr(M = m)}{\Pr(M \geq m)} \tag{19}$$

If measurement is exact (19) is equivalent to (18) and the binary choice model can successfully estimate the hazard rate. If discrete measurement is used, the probabilities in (19) are not equivalent to those that comprise the hazard rate in (18). Under discrete measurement the probability in the numerator of (19) is given by (3), and that in the denominator is given by (13). The probability of a duration being equal to a discrete-measured time M that is estimated with a discrete choice model is constructed by dividing (3) by (13) and is given in (20).

$$\frac{F(M+1) - F(M-1)}{S(M)} \tag{20}$$

Subtracting (18) from (20) gives a factor by which the probability estimated in a binary-choice model with discrete measurement differs from the hazard rate (21).

$$\frac{1}{S(M)}(F(M+1) - (F(M-1) + f(M))) \tag{21}$$

To give an example of this difference I again use the exponential distribution. Substituting (6) and (8) into (18) gives the target quantity (22), (the true hazard rate).

$$\frac{\lambda e^{-\lambda M}}{e^{-\lambda M}} = \lambda \tag{22}$$

Then substituting (6) and (8) into (21) gives the actual quantity estimated with discrete measurement (23).

$$\frac{e^{-\lambda(M-1)} - e^{-\lambda(M+1)}}{e^{-\lambda M}} = e^{\lambda} - e^{-\lambda} \tag{23}$$

The difference between the actual quantity estimated under discrete measurement and the true hazard rate of the continuous distribution is given in (24).

$$e^{\lambda} - (e^{-\lambda} + \lambda) \tag{24}$$

This expression does not reduce to zero and thus there is a non-zero difference between the probability estimated with a binary choice, subject-time model and the true hazard rate.

Given well developed methods for dealing with interval-censored data, if the researcher believes that the dependent variable is a continuous-time event-history variable, there is no reason to avoid the estimators, both parametric and semi-parametric, that are traditionally used to model such variables. Interval-censored data can be accommodated in the Cox proportional hazards model (Pan 1997), and in all of the familiar parametric event-history models (Kim 1997). Since discrete-measured data is simply a special case of interval-censored data, a researcher can still use estimators developed for continuous-time to analyze discrete-measured data without sacrificing methodological rigor or legitimacy. Moreover, since all data are in fact

discrete-measured, interval-censored estimation with continuous-time models is the appropriate estimation choice in all applied settings.

V. Conclusion

Discrete measurement introduces bias to duration models when the discrete measured variable is treated as the exact continuous time value. Limited measurements indicate intervals within which a continuous time variable lies. Discrete measurement leads to an interval-censored variable. Statistical techniques for optimal parameter estimation under conditions of general interval-censoring are already well developed for both the Cox proportional hazard model and parametric estimators. In this study I demonstrate how duration models can be improved by analyzing the discrete measured variable as a systematically interval-censored variable. Through simulations, the superiority of the interval-censored to the midpoint-imputed estimator, in terms of both root mean-squared-error and bias, is established. Replications using an important study on the duration of civil war (Fearon 2004) demonstrate that the use of midpoint imputation can bias results in the direction of a type I inference error.

Continuous-time duration models should always be implemented as interval-censored estimators. First, at least for now, all measurements of duration are discrete, so the possibility that discrete measurement introduces bias to parameter estimation always exists. Second, the correction for bias due to discrete measurement is efficient in that it does not require the estimation of additional parameters. Third, the alternative to continuous-time models; so called discrete-time models (Box-Steffensmeier and Jones 2004), are shown in the current analysis to not be adaptable to discrete measurement. Lastly, many statistical software packages including

STATA™ and R are capable of estimating interval-censored duration models.[12] The analyst only needs to implement the estimation accounting for the specific interval censoring created by discrete measurement.

[12] In Stata the program INTCENS estimates interval-censored parametric duration models. There is currently no Stata program that estimates an interval-censored Cox-PH model. In R, the Survival package estimates parametric interval-censored duration models, and the INTCOX program estimates the interval-censored Cox –PH model.

Appendix I. Sample Simulation Code

The following R code is that used to perform the interval-censored Cox Proportional Hazard simulation for the sample size of 1,000.

1) Create a vector to store the simulation results.

```
 cox_ic1000 <- numeric(1000)
 for (i in 1:1000) {
```

2) Generate the independent variable

```
 x <- rnorm(1000)
```

3) Generate the dependent variable

```
 y <- (-log(1-runif(1000)))^3.5/(exp(-.002+.3*x))^(3.5)
```

4) Generate the discrete measured dependent variable

```
 tt <- floor(y+runif(1000))
```

5) Generate the lower-bound of the interval

```
 for (n in 1:1000) {
 if(tt[n] ==0) tl[n] <- .0000000001 else tl[n] <- tt[n]-.9999999999
  }
```

6) Generate the upper bound of the interval

```
 for (m in 1:1000) {
 if(tt[m] ==0) tu[m] <- 1 else tu[m] <- tt[m]+1
  }
 k <- data.frame(cbind(x,tl,tu))
```

7) Estimate and store results

```
surv <-intcox(Surv(tl,tu,type="interval2")~x, data=k)
 cox_ic1000[i] <- surv$coefficients["x"]
}
```

Appendix II. MLE Derivation for the Exponential Distributions

1) Derivation of mle for the exponential distribution:

$$P(T=t)=\lambda e^{-\lambda t}$$

$$\Leftrightarrow L(T\mid\lambda)=\prod_i \lambda e^{-\lambda t_i}$$

$$\Leftrightarrow ll(T\mid\lambda)=\sum_i \ln(\lambda)-\lambda t_i$$

$$\Leftrightarrow ll(T\mid\lambda)=n\ln(\lambda)-\lambda n\mu$$

$$\Leftrightarrow \frac{dll}{d\lambda}=\frac{n}{\lambda}-n\mu=0$$

$$\Leftrightarrow \lambda^*=\frac{1}{\mu}$$

2) Derivation of the mle for the interval censored exponential distribution:

$$P(R=r)=e^{-\lambda(r-1)}-e^{-\lambda(r+1)}=e^{-\lambda(r)}(e^{\lambda}-e^{-\lambda})$$

$$L(R\mid\lambda)=\prod_i e^{-\lambda(r_i)}(e^{\lambda}-e^{-\lambda})$$

$$\Leftrightarrow ll(R\mid\lambda)=\sum_i -\lambda r_i+\ln(e^{\lambda}-e^{-\lambda})$$

$$\Leftrightarrow ll(R\mid\lambda)=-\lambda n\mu+n\ln(e^{\lambda}-e^{-\lambda})$$

$$\Leftrightarrow \frac{dll}{d\lambda}=-n\mu+\frac{2n}{e^{2\lambda-1}}+n=0$$

$$\Leftrightarrow \lambda^*=\frac{1}{2}\ln\left(\frac{2}{\mu-1}+1\right)$$

Appendix III. Replication Z-Statistics

Table 6) Z-Statistics from the Replication Study

	Model I			Model II			Model IV		
Independent Variable	Midpoint Imputed	Interval Censored	%Difference	Midpoint Imputed	Interval Censored	%Difference	Midpoint Imputed	Interval Censored	%Difference
Coup or Revolution	-5.35	-5.04	-6%	-4.86	-4.60	-5%	-5.22	-4.85	-7%
Eastern Europe	-4.22	-3.69	-12%	-4.31	-3.80	-12%	-4.11	-3.64	-12%
Non-Contigious	-1.62	-1.44	-11%	-2.02	-1.86	-8%	-1.41	-1.29	-8%
Sons of the Soil	3.86	3.56	-8%	3.95	3.64	-8%	3.86	3.59	-7%
Contraband	2.76	2.57	-7%	2.80	2.60	-7%	2.78	2.58	-7%
Ethnic War				1.36	1.33	-2%			
Log (Lagged Population)							-0.37	-0.21	-42%
Constant	19.11	15.79	-17%	10.87	8.91	-18%	4.04	3.29	-19%
Ln(ρ) (Duration Dependence)	2.47	0.34	-86%	2.61	0.44	-83%	2.51	0.36	-86%

Under each model the first column gives the z-statistic of the midpoint-imputed parameter estimate, the second column gives the z-statistic for the interval-censored estimate, and the third column reports the percentage change from the midpoint-imputed to the interval-censored estimate.

References

Balch-Lindsay, Dylan; Andrew J. Enterline. 2000. "Killing Time: The World Politics of Civil War Duration, 1820-1992". *International Studies Quarterly*. 44. 4. 615-642.

Bennett, D. Scott; Allan C. Stam III. 1996. "The Duration of Interstate Wars, 1816-1985". *The American Political Science Review*. 90. 2. 239-257.

Bolks, Sean M; Dina Al-Sowayel. 2000. "How Long Do Economic Sanctions Last? Examining the Sanctioning Process through Duration" *Political Research Quarterly*. 53. 2. 241-265.

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

Box-Steffensmeier, Janet; Jones Bradford S. 2004. *Event History Modeling; A Guide For Social Scientists.* Cambridge University Press. NY.

Casella, George; Lehmann, E.L. 1999. *Theory of Point Estimation.* Springer Press. New York, NY.

Cunningham, David E. 2006. "Veto Players and Civil War Duration". *American Journal of Political Science*. 50. 4. 875-892.

DeRouen Jr., Karl R; David Sobek. 2004. "The Dynamics of Civil War Duration and Outcome". *Journal of Peace Research*. 41. 3. 303-320.

Fearon, James D. 2004. "Why Do Some Civil Wars Last so Much Longer than Others?". *Journal of Peace Research.* 41. 3. 275-301.

Goggins, William B.; Dianne M. Finkelstein; David A. Schoenfeld; Alan M. Zaslavsky. 1998. "A Markov Chain Monte Carlo EM Algorithm for Analyzing Interval-Censored Data under the Cox Proportional Hazards Model". *Biometrics*. 54. 4. 1498-1507.

Goggins, William B; Dianne M. Finkelstein. 2000. "A Proportional Hazards Model for Multivariate Interval-Censored Failure Time Data". *Biometrics*. 56. 3. 940-943.

Hegre, Håvard. 2004. "The Duration and Termination of Civil War". *Journal of Peace Research*. 41. 3. 243-252.

Kadera, Kelly M; Mark J.C. Crescenzi; Megan L. Shannon. 2003. "Democratic Survival, Peace, and War in the International System". *American Journal of Political Science*. 47. 2. 234-237.

Kim, Dong K. 1997. “Regression Analysis of Interval-Censored Survival Data with Covariates Using Log-Linear Models”. *Biometrics*. 53. 4. 1274-1283.

Odell, Patricia M; Keaven M. Anderson; Ralph B. D'Agostino. 1992. “Maximum Likelihood Estimation for Interval-Censored Data Using a Weibull-Based Accelerated Failure Time Model”.*Biometrics*. 48. 3. 951-959.

Pan, Wei. 1997. “Extending the Convex Minorant Algorithm to the Cox Model.” *Journal of Computational and Graphical Statistics.* 8. 1. 109-121.

Regan, Patrick. 2002. “Third-Party Interventions and the Duration of Intrastate Conflicts”. *The Journal of Conflict Resolution*. 46. 1. 55-73.