Estimating a Split-Population Cox Model using Zero-Inflated Poisson Regression

George W. Williford

williford@uga.edu | georgewilliford.com

The University of Georgia

Standard duration models assume that all individuals in the sample will eventually fail, regardless of whether they are censored. This is often an unjustified assumption that can produce biased coefficient estimates. Split-population models correct this problem by estimating the probability that a unit fails at all, then estimating the hazard rate conditional on the probability of failure. Although parametric split-population models have seen limited applications in political science, I demonstrate that a split-population Cox model can be estimated using a zero-inflated Poisson model based on a convenient equivalence between Cox and Poisson regression. I demonstrate the usefulness of the model using a replication analysis of data on international cease-fires.

1. Motivation

1.1 The Problem of Long Term Survivors

Standard duration models assume that all individuals in the sample will eventually fail.

In many cases, this assumption is unrealistic. For example, many cancer patients won't relapse, and many cease-fires won't break down.

By inflating the size of the survivor function in parametric models or the risk-set in semi-parametric models, including these observations creates downward bias in the coefficient estimates.

1.2 Split-Population Models

Mixture (or split-population) models correct this problem by assuming that the sample is composed of individuals drawn from two different populations, one that is at risk of experiencing the event and one that is not.

Which population an individual belongs to is not directly observable, but can be modeled as a latent variable, δ_i, that determines whether an observation is cured or not using binomial response models:

$$\delta_i = \frac{\exp(\mathbf{z_i}\gamma)}{1+\exp(\mathbf{z_i}\gamma)}. \quad (1)$$

Individuals that are observed to fail contribute to the likelihood function via the failure function weighted by δ_i.

$$\delta_i f(t_j|C=1) \quad (2)$$

Right-censored observations contribute via the survivor function, which is split into two parts: the probability that a given unit never fails $(1-\delta_i)$ and the probability that a unit fails but survives beyond the maximum observed failure time

$$(1-\delta_i)+\delta_i S(t) \quad (3)$$

where $S(t)$ is the survivor function for those observations that do fail.

Combining equations 2 and 3 produces the following likelihood function,

$$L=\prod_{i=1}^{N}[\delta_i f(t_j|C_i=1)]^{C_i}[1-\delta_i+\delta_i S(t)]^{1-C_i} \quad (4)$$

and the corresponding log-likelihood function,

$$\ln L=\sum_{i=1}^{N} C_i[\ln\delta_i+\ln f(t_j|C_i=1)]+(1-C_i)[\ln(1-\delta_i)+\delta_i S(t)]. \quad (5)$$

Parametric split-population models have seen limited application in political science. However, semi-parametric models are often a better choice for various reasons.

2. Semi-Parametric Split-Population Model

A Cox proportional hazards model can be fitted using Poisson regression.

- Assume each individual fails according to a very slow counting (Poisson) process.
- Each individual is observed for a series of intervals, t_j, and failure in each interval is recorded as a dummy variable.

Assuming this dummy variable is distributed according to a Poisson distribution,

$$\Pr(Y_i=y)=\frac{\exp(-\exp(\mathbf{X_i}\beta))\exp(\mathbf{X_i}\beta)^{y_i}}{y_i!}, \quad (6)$$

the probability of observing a zero is given by $\exp(-\exp(\mathbf{X_i}\beta))$, and the probability of observing a one is given by $\exp(-\exp(\mathbf{X_i}\beta))\exp(\mathbf{X_i}\beta)$.

These are identical (up to a constant) with the survivor function and failure density of the Cox model:

$$S(t)=\exp(-\exp(\mathbf{X_i}\beta)\int_0^t h_0(t)dt) \quad (7)$$

$$f(t)=\exp[-\exp(\mathbf{X_i}\beta)]\exp(\mathbf{X_i}\beta)\int_0^t h_0(t)dt) \quad (8)$$

Estimating $\int_0^t h_0(t)dt$ by including a factor variable for the risk set (i.e., a step function) provides parameter estimates identical to the Breslow approximation for the Cox model.

Substituting (7) and (8) into (4) produces a model equivalent to a zero-inflated Poisson.

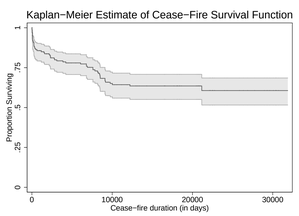
3. Replication Analysis of Data on Interstate Cease-Fires

3.1 Preliminary Diagnostics

Below, I have replicated a study on international ceasefires using data from Lo, Hashimoto, and Reiter (2008).

Out of 188 ceasefires in the dataset, only 56 (roughly 30%) experience a recurrence of conflict.

This can be further illustrated by examining the Kaplan-Meier estimate of failure time. Less than 40% of ceasefires are expected to fail in 80 years:



3.2 Demonstration of Cox-Poisson Equivalence

	Cox	Poisson
Foreign Imposed Regime Change (FIRC)	-164.0***	-164.0***
	(-30.13)	(-30.13)
FIRC * ln(t)	17.09***	17.09***
	(-3.114)	(-3.114)
Capability Change	0.27	0.27
	(-0.371)	(-0.371)
Battle Tide	-0.784**	-0.784**
	(-0.336)	(-0.336)
Third Party Intervention	0.028	0.028
	(-0.37)	(-0.37)
Agreement Strength	0.0117	0.0117
	(-0.0862)	(-0.0862)
One Democracy	0.169	0.169
	(-0.416)	(-0.416)
Two Democracies	7.540**	7.540**
	(-3.09)	(-3.09)
Two Democracies * ln(t)	-1.682***	-1.682***
	(-0.599)	(-0.599)
Tie	0.482	0.482
	(-0.516)	(-0.516)
Battle Deaths (ln)	-0.222**	-0.222**
	(-0.0967)	(-0.0967)
Conflict History	0.373	0.373
	(-0.475)	(-0.475)
Existential Stakes	0.846*	0.846*
	(-0.447)	(-0.447)
Contiguity	7.193	7.193
	(-5.472)	(-5.472)
Contiguity * ln(t)	-0.707	-0.707
	(-0.612)	(-0.612)
Constant		-5.594***
		(-2.075)
N	5524	5524
Log pseudolikelihood	-207.48497	-247.89261
AIC	444.9699	557.7852

3.3 Split Population/Zero-Inflated Poisson Results

	Incidence (Inflation)	Hazard (Poisson)
Foreign Imposed Regime Change (FIRC)	-2.997	-202.9**
	(-2.334)	(-79.86)
FIRC * ln(t)		20.99***
		(-8.134)
Capability Change	-11.70***	-1.652
	(-3.256)	(-2.242)
Battle Tide	2.166*	0.0277
	(-1.296)	(-0.672)
Third Party Intervention	2.223	1.944
	(-2.089)	(-2.658)
Agreement Strength	-0.509	-0.471
	(-0.327)	(-0.527)
One Democracy	-0.947	-0.245
	(-1.633)	(-0.834)
Two Democracies	-4.852	146.5***
	(-4.069)	(-38.18)
Two Democracies * ln(t)		-32.13***
		(-8.269)
Tie	2.04	2.75
	(-2.709)	(-2.962)
Battle Deaths (ln)	0.768	0.266
	(-0.497)	(-0.489)
Conflict History	-0.504	0.0188
	(-3.579)	(-2)
Existential Stakes	-2.136	-0.465
	(-2.586)	(-1.82)
Contiguity	-0.846	3.579
	(-1.622)	(-4.424)
Contiguity * ln(t)		-0.297
		(-0.533)
Constant	-1.732	-6.65
	(-4.379)	(-4.937)
N	5524	
Log pseudolikelihood	-233.1439	
AIC	530.2878	
Vuong Test	z = 1.61, Pr>z=0.0534	

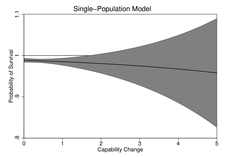
3.4 Differences in Hazard Coefficients between Single and Split Population Models

	Difference in $\hat{\beta}$	Percent Change in $\hat{\beta}$
Foreign Imposed Regime Change (FIRC)	-38.9	23.72
FIRC * ln(t)	3.9	22.82
Capability Change	-1.922	711.85
Battle Tide	0.8117	103.53
Third Party Intervention	1.916	6842.86
Agreement Strength	-0.4827	4125.64
One Democracy	-0.414	244.97
Two Democracies	138.96	1842.97
Two Democracies * ln(t)	-30.448	1810.23
Tie	2.268	470.54
Battle Deaths (ln)	0.488	219.82
Conflict History	-0.3542	94.96
Existential Stakes	-1.311	154.96
Contiguity	-3.614	50.24
Contiguity * ln(t)	0.41	57.99
Constant	-1.056	18.88
Average Difference	4.3907375	1049.75

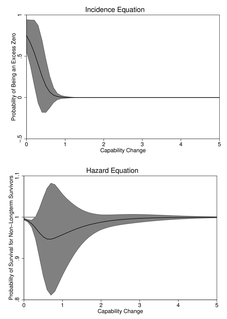
3.5 Differences in Inferences Drawn

- Battle Tide influences *whether* a cease-fire breaks down, but not *when* it does.
- Battle Deaths and Existential Stakes become insignificant.
- Capability Change has a significant effect on whether a break down occurs (versus no effect in the single-population model)

Single Population Model



Split Population Model



4. References

Lo, Nigel, Barry Hashimoto, and Dan Reiter. 2008. Ensuring Peace: Foreign-Imposed Regime Change and Postwar Peace Duration, 19142001. International Organization 62 (4): 717736.

33rd Annual Meeting of the Society for Political Methodology, Rice University, July 2016. Standard errors clustered on conflict in parentheses, *p<0.10, **p<0.05, ***p<0.01. All graphs display 95% confidence intervals.