Droughts, Land Appropriation, and Rebel Violence in The Developing World[*]

Benjamin E. Bagozzi,[†] Ore Koren,[†] and Bumba Mukherjee[‡]

December 11, 2016

[*] Note: authors appear in alphabetical order.

[†] Department of Political Science, University of Minnesota. Email: bbagozzi@umn.edu koren044@umn.edu.

[‡] Department of Political Science, Penn State University. Email: sxm73@psu.edu.

Split Popn Cox Overrep Failure

Likelihood function

Recall from Ben's "Parametric Zombie Survival Model" that the probability of misclassification (that is, subset of non-censured failure outcomes that are being misclassified) is

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

The unconditional density is thus given by the combination of an observation's misclassification probability and its probability of experiencing an actual failure conditional on not being misclassified,

$$\alpha_i + (1 - \alpha_i) * f(t_i) \tag{2}$$

And the unconditional survival function is therefore

$$(1 - \alpha_i) * S(t_i) \tag{3}$$

where

$$\alpha_i = \frac{\exp(\mathbf{Z}\gamma)}{1 + \exp(\mathbf{Z}\gamma)} \tag{4}$$

The likeihood function of the Parametric Zombie Survival Model is from equation 7 in Ben's document is defined as

$$L = \prod_{i=1}^{N} [\alpha_i + (1 - \alpha_i) f(t_i|X, \beta)]^{C_i} [(1 - \alpha_i) S(t_i|\mathbf{X}, \beta)]^{1 - C_i} \tag{5}$$

And the log likelihood is

$$L = \sum_{i=1}^{N} \{C_i \ln[\alpha_i + (1 - \alpha_i) f(t_i|X, \beta)] + (1 - C_i) \ln[(1 - \alpha_i) S(t_i|\mathbf{X}, \beta)] \tag{6}$$

We extend these definitions and notation from the Parametric Zombie Survival Model to the Cox PH framework. To this end, first note that the conditional hazard function in the Cox PH model in this case is

$$h(t|\mathbf{X}) = h_0(t)e^{\mathbf{X}\beta} \tag{7}$$

where $h_0(t)$ is the unknown baseline hazard hazard function. Using equation (4), the description of α_i and the survival function, failure density and the conditional hazard function (see equation 6) of the Cox model, we can define the likelihood function of the split popn Cox-Overrep Failure (OF) model as follows,

$$\begin{aligned} L = \prod_{i=1}^{N} \left[\alpha_i + (1-\alpha_i)h_0(t_i)e^{\mathbf{X}\beta} \exp\left\{-\exp(\mathbf{X}\beta)\int_0^{t_i} h_0(u)du\right\}\right]^{C_i} \\ \left[(1-\alpha_i)\exp\left\{-\exp(\mathbf{X}\beta)\int_0^{t_i} h_0(u)du\right\}\right]^{1-C_i} \end{aligned} \tag{8}$$

The log-likelihood from this expression is

$$\begin{aligned} \ln L = \sum_{i=1}^{N} \left(C_i \ln\left[\alpha_i + (1-\alpha_i)h_0(t_i)e^{\mathbf{X}\beta} \exp\left\{-\exp(\mathbf{X}\beta)\int_0^{t_i} h_0(u)du\right\}\right]\right) \\ + \left((1-C_i)\ln\left[(1-\alpha_i)\exp\left\{-\exp(\mathbf{X}\beta)\int_0^{t_i} h_0(u)du\right\}\right]\right) \end{aligned} \tag{9}$$

where $\int_0^{t_i} h_0(u)du$ is the cumulative baseline hazard function that can be estimated by a non-parametic step function. To see this, let $u_0 < u_1 < ....u_k$ denote the distinct uncensored observations. If the baseline hazard is constant between these values, then an estimator for the baseline hazard can be defined as

$$\int_0^{t_i} \widehat{h}_0(u)du = \sum_{j=1}^{k} H_j I(u_j \leq t) \tag{10}$$

To simplify notation, let $\Lambda(t) = \sum_{j=1}^{k} H_j I(u_j \leq t)$. Then the likelihood in (8) can be written as

$$L = \prod_{i=1}^{N} \left[\alpha_i + (1-\alpha_i) h_0(t_i) e^{\mathbf{X}\beta} \exp\left\{-e^{\mathbf{X}\beta}\Lambda(t)\right\}\right]^{C_i} \left[(1-\alpha_i) \exp\left\{-e^{\mathbf{X}\beta}\Lambda(t)\right\}\right]^{1-C_i} \tag{11}$$

And the corresponding log likelihood as

$$\ln L = \prod_{i=1}^{N} \big(C_i \ln\left[\alpha_i + (1-\alpha_i) h_0(t_i) e^{\mathbf{X}\beta} \exp\left\{-e^{\mathbf{X}\beta}\Lambda(t)\right\}\right]\big) + \big((1-C_i) \ln\left[(1-\alpha_i) \exp\left\{-e^{\mathbf{X}\beta}\Lambda(t)\right\}\right]\big) \tag{12}$$

While these likelihood and log-likelihood expressions look complicated,it should be(in theory)relatively easy to estimate using logit for α and separately fitting a Cox PH model for the survival part.

Finally, the following is clearly not necessary for the paper – but I think (am not sure about this), the integral in the contribution of either the censored *or* uncensored observations in the likelihood and thus log-likelihood function can be treated as zero. If so, then the expression for the full log likelihood function can be fully derived and written completely more easily after making the necessary substitutions and taking logs. *If* this can be done mathematically, then I can try to differentiate the fully defined log likelihood, solve the score equations and define the "profile likelihood" that only depends on the parameters γ and β. This will allow us to study the full profile likelihood that in any case will (likely) confirm using logit for α and separately fitting a Cox PH model for the survival part. And this in effect means that standard conclusions from asymptotic theory can be used (or cited) from the logit and Cox PH model to demonstrate that the parameter estimates from the split popn Cox-overrep failure are consistent and asymptotically normally distributed.