A Nested Case-Control Study of Childhood Leukemia

{ Courtney Schiffman, Nima Hejazi, and Aurelien Bibaut } University of California, Berkeley

Introduction

Our data set consists of 202 matched subjects, 101 cases (diagnosed with Leukemia during the study) and 101 controls (healthy throughout the study). Beginning in 1980, dried blood spots were collected from babies upon birth. Beginning in 1995, children were enrolled in the study within 72 hrs. of being diagnosed with childhood leukemia. For each case, a matched control was found. In 2015, bloodspots were released from storage for all 202 subjects and analyzed. This resulted in over 600 molecular measures for each subject, along with baseline covariates (e.g., gender, mother's weight) measured previously . The observed data structure is as follows:

$$O = (\tilde{T} = min(T, C), \Delta = I(T < C), W) \quad (3)$$

Results I

Fit a LASSO-penalized Cox Proportional Hazards model for over 680 covariates:

$$1.368787 = exp(\beta) = \quad (4)$$

$$\frac{P(Y(t) = 1|Female, W, Y(t-1) = 0)}{P(Y(t) = 1|Male, W, Y(t-1) = 0)} \quad (5)$$

$$0.7370129 = exp(\beta) = \frac{P(Y(t) = 1|NotBF, W)}{P(Y(t) = 1|BF, W)} \quad (6)$$

Some significant coefficients from the Cox Proportional Hazards LASSO:

Name	Cox coefficient
X51505	1.520002
X566	1.636203
X54009	1.639914

From the logistic LASSO fit, some predicted conditional probabilities of being diagnosed with Leukemia in the first 2 years of life.

P(Y(2)=1 | W)	Strata
0.522562559	♀, non-hispanic, breast-fed
0.473372777	♀, non-hispanic, breast-fed
0.331672394	♂, hispanic, breast-fed
0.006105754	♀, non-hispanic, not breast-fed
0.005526099	♀, non-hispanic, breast-fed

Objectives

Our objectives for this project are to estimate

1. The relative risk associated with gender with LASSO Cox Proportional Hazards:

$$exp(\beta) = \frac{P(Y(t) = 1|A = 1, W, Y(t-1) = 0)}{P(Y(t) = 1|A = 0, W, Y(t-1) = 0)} \quad (1)$$

2. The conditional probability of being diagnosed with childhood leukemia in the first 2 yrs. of life using logistic LASSO regression.

$$logit(P(Y(2) = 1|W)) = W * \beta \quad (2)$$

3. Study the effects of matched case control sampling on Cox Proportional Hazards estimation using a simulated data set.
4. Study the effect of sex on the probability of survival using Longitudinal TMLE.
5. Estimate the analogue to the β in the Cox Proportional Hazards model using the previous Longitudinal TMLE results.

Discussion

Borgan (1995) suggests a modified partial likelihood for a Cox Proportional Hazards model in studies with nested case-control sampling:

$\pi(r|t, i)$ =Probability of sampled risk set r if individual i has an event at time t

$$\lambda_{i,r}(t) = Y_i(t)\alpha_0(t)exp(\beta^T X_i)\pi(r|t, i) \quad (7)$$

The partial likelihood in nested case-control:

$$L_{ncc}(\beta) = \prod_{t_j} \frac{exp(\beta^T X_{i_j})\pi(i_j|t_j, \tilde{R}_j)}{\sum_{l \in \tilde{R}_j} exp(\beta^T X_l)\pi(l|t_j, \tilde{R}_j)} \quad (8)$$

Nested Case-Control studies help to cut down expenses, but there are many complications created:

1. How do we interpret the CoxPH and Logistic Regression results? Are they applicable at some degree to the full population?
2. Do we have measurements of certain covariates for the full cohort?
3. What other options exist for small samples?

Results II

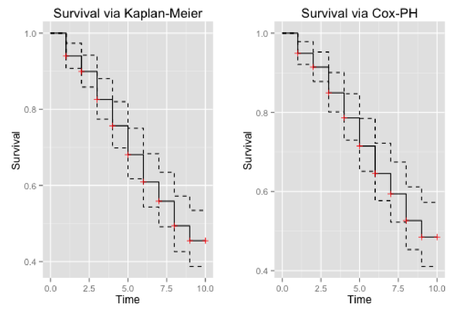

Figure 1: Simulated Case-Control

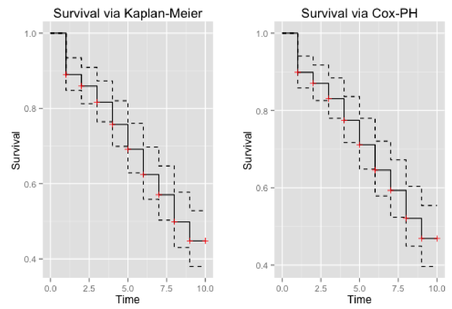

Figure 2: Simulated Matched Case-Control

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
sex	-0.05	0.95	0.22	-0.23	0.82
income	-0.14	0.87	0.05	-2.65	0.01
brfdur	-0.03	0.97	0.01	-3.79	0.00

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
sex	0.11	1.11	0.22	0.47	0.64
income	-0.03	0.97	0.06	-0.60	0.55
brfdur	-0.03	0.97	0.01	-3.48	0.00

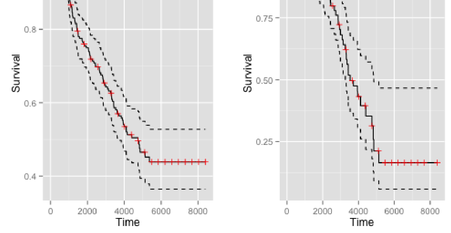

Figure 3: Survival curve estimates for our data

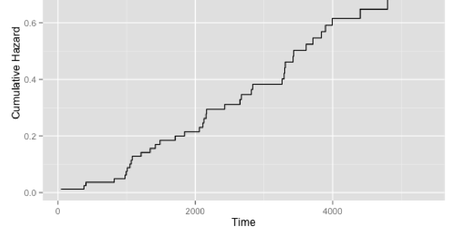

Figure 4: Cumulative hazard estimate for our data

Results III

We use longitudinal TMLE to estimate the sex specific mean survival outcomes $\Psi_1(P_0) = E[E(Y(T) = 1|A = 1, W)]$ and $\Psi_0(P_0)E[E(Y(T) = 1|A = 0, W)]$ and then the analogues of β for sex in Cox, defined as $\log \frac{\log E[E(Y(T)=1|A=1,W)]}{\log E[E(Y(T)=1|A=0,W)]}$. To do so, we reformulate the data structure under the form of the counting processes $N_1(t) = I(\tilde{T} \leq t)I(\Delta = 1)$ and $N_2(t) = I(\tilde{T} \leq t)I(\Delta = 0)$ where $\tilde{T} = min(T, C)$ and $\Delta = I(T < C)$.

We discretized the period 1980-2008 in 30 time points. We get a DAG analogous to that in Stitelman et al. We used the package `ltmle`. We used observations weights to implement the case-control weighted TMLE. We also performed longitudinal TMLE without observation weighting. We obtained the following numerical values.

weighting	$\hat{\Psi}_{0,n}$	$\hat{\sigma}_{\hat{\Psi}_{0,n}}$	$\hat{\Psi}_{1,n}$	$\hat{\sigma}_{\hat{\Psi}_{1,n}}$	$\log \frac{\log \hat{\Psi}_{1,n}}{\log \hat{\Psi}_{0,n}}$
Yes	1.037947E-4	1.350893E-3	1.054344E-4	1.654062E-3	-0.001710174
No	0.4815406	0.04765275	0.495612	0.0550974	-0.04021241