MACS 30250 Methods and Initial Results Assignment

Dr. Richard Evans

TA Zunda Xu

Submitted by Junho Choi

1 Data

For this study, I collected the data at the individual level from the official website of Compassion International.[1] The focus was on the various signals about the potential sponsorship recipients (PR hereforth) can be picked up by the potential sponsors (PS hereforth). These sets of information were then analyzed with the expectation that at least some subset of such information can ultimately impact whether a PS commits to a sponsorship (i.e. provides the match). Of course, it is not to argue that the dataset curated for this study is the entirety of the potential covariates and features affecting the PS's decisions. For instance, there may be information from the images of the PS – such as complexion and subtle facial expressions – that may influence the PS, but are not captured by the dataset. However, I also note that effort was made to capture most of the textual information provided to the PS.

Data collection was conducted daily from April 19, 2019 to May 3, 2019 (initial collection and 14 daily follow-ups) using web scraping through Python's Selenium module.[2] Daily web scraping was made at around 9:00 PM Central Standard Time (CST), for approximately 3 hours each session, with the exception of collections made in May due to network problems, in which the processes started at 11:00 PM CST. The reason for the rather inefficient hours of data collection was due to the structure of the Compassion International's website; information about each individual is stored in each separate web page, rather than the organization having a public list or table of all the individual and corresponding information.

Descriptive statistics of selected variables are listed in Table 1, and they are separated by whether the PR associated with the data were ever matched within the 14-day window of observation. Judging by the difference in means and the t-statistics associated with them, it is observable that there is heterogeneity between the two groups separated by match status.

[1] For the up-to-date list of children who are waiting for sponsorship, I refer to the following web page: `https://www.compassion.com/sponsor_a_child/`

[2] I note that while I certainly wished to gather data for longer window of observation, the official Compassion website reassigned the IDs and web URLs associated with each PR on May 4, 2019, making it unable for me to make updates to the existing dataset.

Among the variables, those located in Panel D perhaps require more attention; these variables are the more-prominently signaled information, in which the official website of Compassion highlights using "badges" or indicators next to the photographs of the PR if applicable.

One characteristic outstanding from the data is that the PR who were successfully matched exhibit qualities indicating higher ability of survival even without the sponsorship. For instance, it is more likely to find those vulnerable to exploitation in the unmatched group than the matched. Another feature that is of the similar brood is that those whose parents are both employed – and therefore having the income to support the PR – are more likely to be found in the matched group than the unmatched.

Another key characteristic from the dataset is that all of the African nations are marked as "AIDS-affected" by Compassion; it can be confirmed that the descriptive statistics for the two separate labels are exactly the same. While it is true that the risk of HIV/AIDS is higher in African nations, it could be an over-generalization to indicate all of such countries as having the same risk of AIDS. For instance, Burkina Faso, in which 10.79% of the overall sample PR are located, has a lower HIV/AIDS adult prevalence rate than, say, Haiti, which is not indicated as AIDS-affected (CIA n.d.). Therefore, additional care may be needed when interpreting the results with respect to the PR in African nations.

2 Methods

In essence, the methods used in this study are variants of models for survival analysis. If one considers being on the list of unmatched PR as continuing to "survive" or being "alive," and being matched as reaching "death" or "failure," it is the researcher's goal to understand what are the covariates influencing or correlated with survival and death (Ciuca and Matei 2010). The predilection for choosing models for survival analyses over others is due to the fact that the data at hand is "right-censored"; in other words, one is not able to see whether an observation continues to survive or reaches death after the window of observation. In addition, it is one of the goals of this research to understand what may affect the duration until a successful match; survival analysis is therefore useful in such a circumstance (Rodriguez 2010).

In the following subsections, I will showcase two methods: Cox proportional hazards model and survival (decision) tree. The former is an example of a linear, (semi-)parametric model where as the latter is an example of a nonparametric model. Both branches of models are used in order to produce a balanced analysis containing interpretability as well as reasonable prediction.

2.1 Cox Proportional Hazards Model

Cox proportional hazards model is a linear model, in which the covariates are used to predict the conditional (on the covariates) hazard rate at time t (Cox 1972). Hazard rate at time t of observation i, denoted $\lambda_i(t)$, refers to the probability that the said observation will reach death by period $t+1$. Denoting the conditional hazard rate (at time t for observation i) given X_i (the vector of covariates) as $\lambda_i(t|X_i)$, the model for Cox proportional hazards model is as follows:

$$\lambda_i(t|X_i) = \lambda_0(t)\exp\left(X_i'\beta\right)$$

in which $\lambda_0(t)$ is the baseline hazard function at time t. The model is "proportional" in the sense that the conditional hazard rate is proportional to $\lambda_0(t)$. In addition, the reason for calling the Cox model a "semi-parametric" is due to the fact that functional form for the said baseline hazard function is not specified, and is rather fit using the data. Parametric versions of the proportional hazards model include Weibull and exponential proportional hazards model, in which the model explicitly assumes forms for $\lambda_0(t)$ (Rodriguez 2010).

2.2 Survival Decision Tree

Just as in the linear (semi-)parametric models, survival analysis using nonparametric models focuses on hazard rates conditional on covariates. This subsection, focused on survival decision tree, will elaborate on how the said model can be thought of as a regular decision tree classification with information gain splitting criterion where the multiple classes are "no failure" and periods until failure (Kuhn 2014).

Using the conditional hazard function $\lambda_i(t|X)$, we can write the log-likelihood function of observing i at period $t=T$ with the status θ as follows:

$$L_i(T,\theta_i|X_i) = \log\Big[\Big(\prod_{t=1}^{T-1}\lambda(t|X_i)\Big)\Big(\lambda(T|X_i)^{\theta_i}\{1-\lambda(T|X_i)\}^{(1-\theta_i)}\Big)\Big] \tag{1}$$

in which $L_i(\cdot,\cdot|\cdot)$ is the conditional log-likelihood function. Status θ_i is either 0 or 1, where

1 refers to failure or death. The problem is learning what the true form of conditional hazards function is, which is very unlikely to be actually observed. Therefore, we can have estimations or guess of the function $\lambda(\cdot|\cdot)$, denoted $\hat{\lambda}(\cdot|\cdot)$. Then our estimated log-likelihood function is simply $L(\cdot, \cdot|\cdot, \lambda = \hat{\lambda})$.

Using the logarithm transformation, (1) with $\lambda = \hat{\lambda}$ becomes:

$$L_i(T, \theta_i|X_i, \hat{\lambda}) = \Big[(1-\theta_i)\log\{1-\lambda(T|X_i)\} + \theta_i \log\lambda(t|X_i) + \sum_{t=1}^{T-1} \log\lambda(t|X_i)\Big] \tag{2}$$

and we average this over $i = 1, 2, \ldots, N$ (i.e. N observations) for the implementation of maximum log likelihood. Let us write the said average as $A(T, \theta|X, \hat{\lambda}) = \frac{1}{N}\sum_{i=1}^{N} L_i(T, \theta|X_i, \hat{\lambda})$. But notice that if one considers status of death or survival for each time period as a dummy variable, one can rewrite $A(T, \theta|X, \hat{\lambda})$ as follows:

$$A(T, \theta|X, \hat{\lambda}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_i} \Big[(1-\theta_{it})\log\{1-\lambda(t|X_i)\} + \theta_{it}\log\lambda(t|X_i)\Big] \tag{3}$$

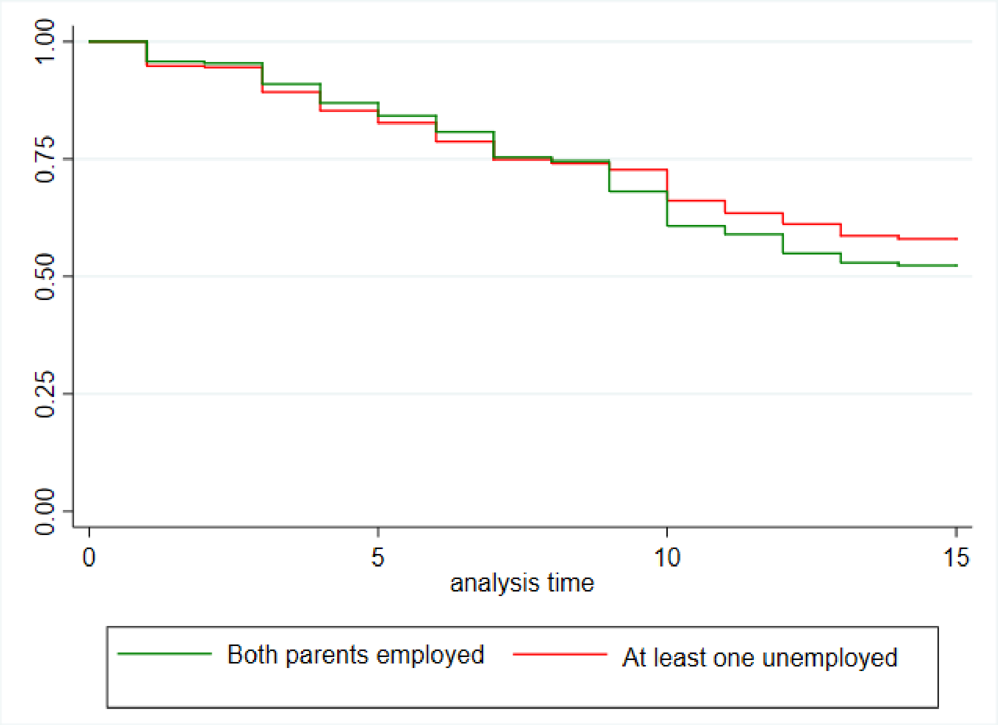
where θ_{it} is binary and equal to 1 if i is still surviving at time t, and 0 if all else. Term (3) is proportional to the information gain splitting criterion, meaning that one can directly use multi-class decision trees for survival analysis as well. Similar exposition can be made for other splitting criteria, and have been used in empirical settings such as medicinal studies (Intrator and Kooperberg 1995; Ture, Tokatli, and Kurt 2009)

3 Initial Results

3.1 Initial Results for Parametric Models

Results of conducting estimation with binary logistic regression and Cox proportional hazards model are produced in Table 2. I note that the logistic regressions, in columns (1) and (2), were conducted in order to buttress the validity of identification when the dependent variable is binary (i.e. matched or not matched). As partially predicted from Section 1, covariates that could indicate that the PR (and PR's family) cannot sustain themselves are negatively correlated with the match status. In both cases of with and without controlling for continents, the employment statuses of guardians are a strong predictor of being matched, where as the risk of exploitation is that of not being matched. Another point of interest is the lost statistical significance on the variable "living with father" when controlling for continents; this may suggest that there is correlation between regions and parenthood tendencies of the PR. That is, for instance, certain regions may be more prone to having single parenthood.

Figure 1: Kaplan-Meier Estimates, Conditioned on Guardian Employment



NOTE: y-axis indicate estimated conditional survival rates, decreasing over time.

As seen from columns (3) and (4) of Table 2, the said results are corroborated and furthered when using Cox proportional hazards model. The output is reported in (average) hazard rate terms. Using the said model, Kaplan-Meier estimates of average conditional survival rates (across window of observation) can be plotted accordingly. An example plot, in which the condition is on both guardians being employed, is given in Figure 1. One caveat, however, is that the said estimates produced from (semi-)parametric models are going to be monotonic in nature and may not capture the true behavior of conditional survival rates across time; this adds motivation for using nonparametric models.

Table 1: Descriptive Statistics by Match Status

	Matched (1)	Not matched (2)	t-statistic (3)
Panel A. Demographics			
Age	4.760	4.587	5.421
Female	47.22	44.77	2.438
Education			
Too young for enrollment	34.70	32.84	1.917
Unenrolled despite of age	13.44	15.33	−2.641
Preschool and kindergarten	33.86	39.00	−5.229
Elementary and middle schools	18.01	12.83	6.952
Panel B. Family Information			
Living with mother	92.57	91.24	2.397
Living with father	69.70	63.59	6.342
Number of siblings	1.558	1.578	−0.638
Guardian employment status			
Both guardians unemployed or unknown	10.07	12.72	−4.090
Only one guardian employed	51.05	53.71	−2.599
Both employed	38.88	33.57	5.383
Panel C. Geographic Information			
Christianity as the dominant religion	66.67	63.38	3.427
Regional average monthly income (USD)	85.847	73.636	7.703
Continent			
Africa	44.29	52.90	−8.589
Asia	10.50	3.91	12.413
South America	20.51	16.43	5.207
North America (Mexico and Caribbean)	9.22	0.05	8.533
Central America	15.49	21.98	−8.361
Panel D. Signaling Information			
Urgency (during the observation window)	4.62	0.11	14.115
Aids-affected area	44.29	52.90	−8.589
Vulnerable to exploitation	35.66	43.63	−8.128
Number of observations	4392	5595	-

Note: Entries for categorical variables are in percentage terms, and those for continuous variables are averaged by group. Education levels are indicated as those that correspond to the U.S. educational system. The dummy variable "Christianity as the dominant religion" equals to 1 if the corresponding PR's nation has 70% or more population designated as Christians, and is considered as Compassion International is a Christian non-profit organization.

Table 2: Linear (Semi-)Parametric Models Results Using the Compassion Dataset

	Logistic		Cox	
	(1)	(2)	(3)	(4)
Urgency	3.878^{***}	3.905^{***}	6.825^{***}	5.760^{***}
	(0.417)	(0.431)	(0.499)	(0.508)
Aids area	-0.147^{*}	-1.096^{***}	0.784^{***}	0.447^{***}
	(0.077)	(0.113)	(0.046)	(0.035)
Exploitation	-0.432^{***}	-0.516^{***}	0.819^{***}	0.785^{***}
	(0.076)	(0.076)	(0.016)	(0.044)
Age	0.102^{***}	0.117^{***}	1.051^{***}	1.061^{***}
	(0.021)	(0.022)	(0.047)	(0.016)
Female	0.139^{***}	0.186^{***}	1.117^{***}	1.150^{***}
	(0.042)	(0.042)	(0.033)	(0.034)
Living with mother	0.003	0.110	1.009	1.096
	(0.072)	(0.074)	(0.054)	(0.062)
Living with father	0.117^{**}	0.064	1.096^{***}	1.051
	(0.048)	(0.048)	(0.038)	(0.037)
One guardian employed	0.010	0.137^{**}	0.968	1.055
	(0.061)	(0.063)	(0.046)	(0.052)
Both guardians employed	0.440^{***}	0.557^{***}	1.288^{***}	1.399^{***}
	(0.067)	(0.069)	(0.065)	(0.073)
Unenrolled	-0.302^{***}	-0.236^{***}	0.850^{***}	0.903^{**}
	(0.068)	(0.069)	(0.042)	(0.045)
Preschool & Kindergarten	-0.265^{***}	-0.355^{***}	0.877^{***}	0.819^{***}
	(0.051)	(0.053)	(0.034)	(0.032)
Elementary or Middle	-0.018	-0.072	1.001	0.956
	(0.068)	(0.090)	(0.061)	(0.061)
Constant	-0.687^{***}	0.108	-	-
	(0.119)	(0.138)		
N	9987	9987	9987	9987
Continent dummies	No	Yes	No	Yes

NOTE: $^{***}: p < 0.01$, $^{**}: p < 0.05$, $^{*}: p < 0.1$. Heteroskedasticity-robust SE in parentheses. Coefficients reported for columns (1) and (2), in which the dependent variable is binary status of being matched within the observation window versus not. Hazard rates reported for columns (3) and (4), in which the dependent variable is the conditional hazard rate; values less than 1 indicate reducing the hazard rate (i.e. correlated with longer period of not being matched). Baseline for education is “too young to be in school”; that for guardian employment is “both unemployed or unknown.”

3.2 Initial Results for Nonparametric Models

For nonparametric models, I produce the feature importance plot for survival tree. This corroborates the parametric result that urgency, age, and guardian employment are important predictors.

4 References

Central Intelligence Agency. n.d. "HIV/AIDS – Adult Prevalence Rate." *The World Factbook.*

Ciuca, Vasilica, and Monica Matei. 2010. "Survival Analysis for the Unemployment Duration." *Proceedings of the 5th WSEAS International Conference on Economy and Management Transformation* 1: 354-359.

Cox, David R. 1972. "Survival Analysis for the Unemployment Duration." *Journal of the Royal Statistical Society, Series B (Methodological)* 34 (2): 187-220.

Intrator, Orna, and Charles Kooperberg. 1995. "Trees and Splines in Survival Analysis." *Statistical Methods in Medical Research* 4: 237-261.

Kuhn, Ben. 2014. "Decision Tree for Survival Analysis," November. `https://www.benkuhn.net/survival-trees`

Rodriguez, Germán. 2010. "Survival Models." *Lecture Notes on Generalized Linear Models*, September. `https://data.princeton.edu/wws509/notes`.

Ture, Mevlut, Fusun Tokatli, and Imran Kurt. 2009. "Using Kaplan–Meier Analysis Together with Decision Tree Methods (C&RT, CHAID, QUEST, C4.5 and ID3) in Determining Recurrence-free Survival of Breast Cancer Patients." *Expert Systems with Applications* 36: 2017–2026.