Why Do We Give?

Urgency and Stability in Child Sponsorship Programs

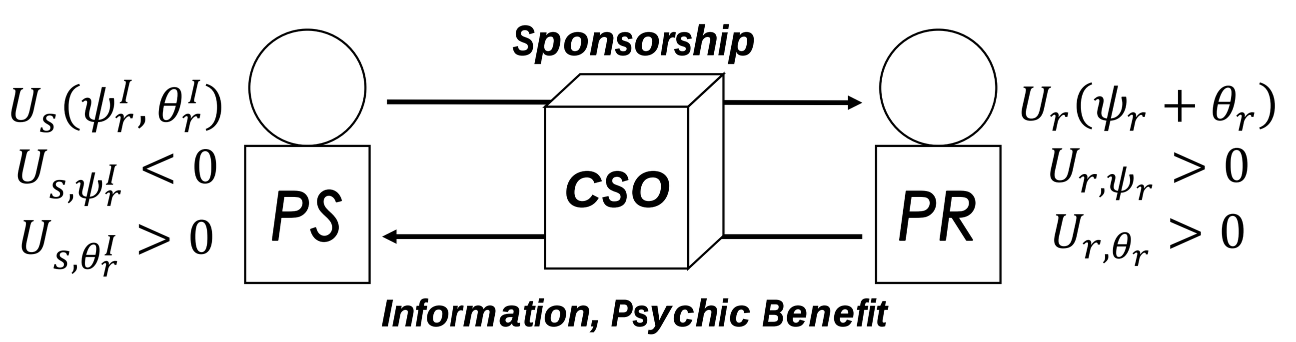
Junho Choi (Computational Economics)

MASTERS IN COMPUTATIONAL SOCIAL SCIENCE
THE UNIVERSITY OF CHICAGO

Research Question

Are matches in a child sponsorship program (CSP) made based solely on how **urgently** potential sponsors need help, or also on how **stable** the matches can be?

Motivation (and how CSP works)



U: utility function
ψ: instability
θ: urgency
I: info provided by CSO
s, r: sponsor and recipient

As the welfare (of sponsors [PS] and recipients [PR]) maximizer, is it in the child sponsorship organization (CSO)'s best interest to relay ψ and θ as is?

Literature and Theory

Warm-glow model of giving

Andreoni (1989): motivation behind giving not just to accomplish a common objective, but also for something personal ***(warm glow)***

Myerson-Satterthwaite Impossibility

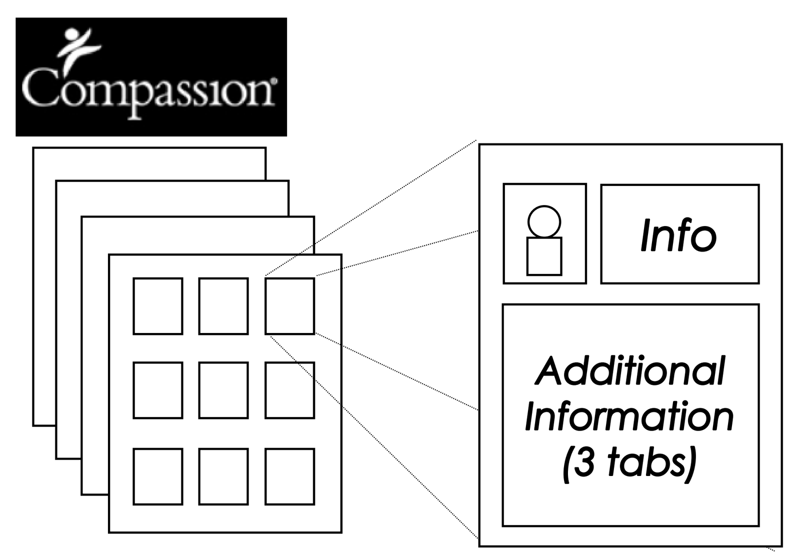
Using *Myerson and Satterwaite (1983)*: truth-telling can conflict with Pareto efficiency, even in CSP matchings

References

Andreoni, James. 1989. "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence." Journal of Political Economy 97 (6): 1447-1458.

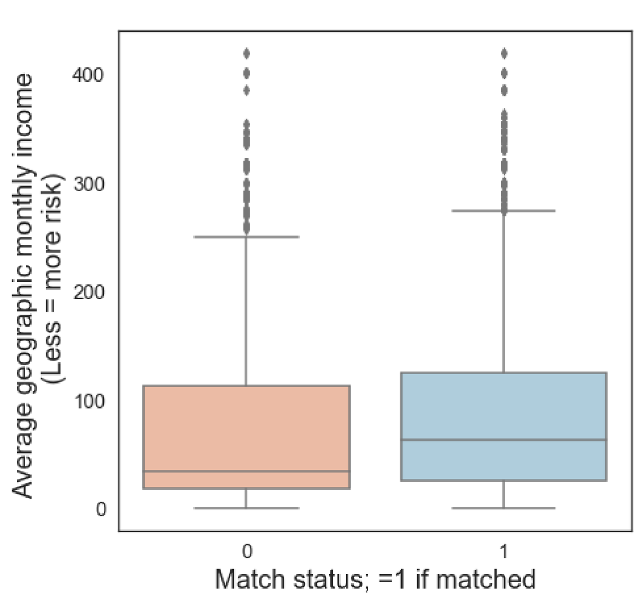
Myerson, Roger B., and Mark A. Satterthwaite. 1983. "Efficient Mechanisms for Bilateral Trading." Journal of Economic Theory 29 (2): 265-281.

Data and Preparation

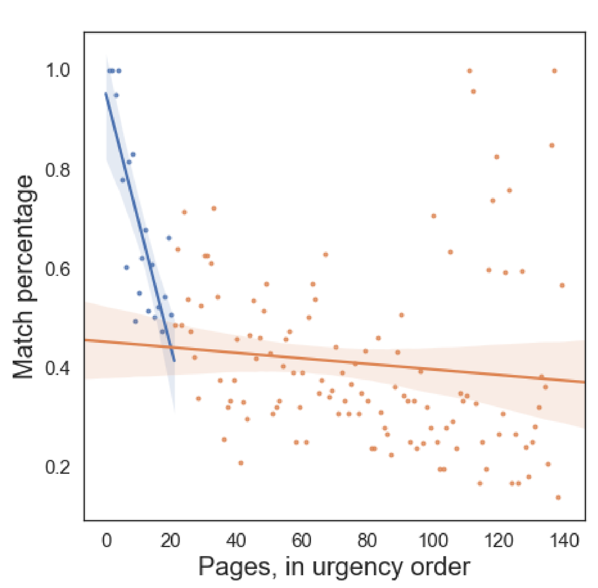


Most textual data scraped using Python, from Compassion International website
N=9518, 15-days-worth of data
(Data collected from April 18, 2019 to May 3, 2019)

Initial Glance at Data



Geographic income by match status



Match percentage by urgency (pages)

Data Issues

(1) Continents / regions as covariates difficult as all obs. from *Africa*: also coded as **AIDS area**

(2) Graphic information (i.e. photographs of PR) unincorporated in the dataset

(3) **Urgency**: not only how dire the need for help, but also how long exposed to the PS on the waiting list

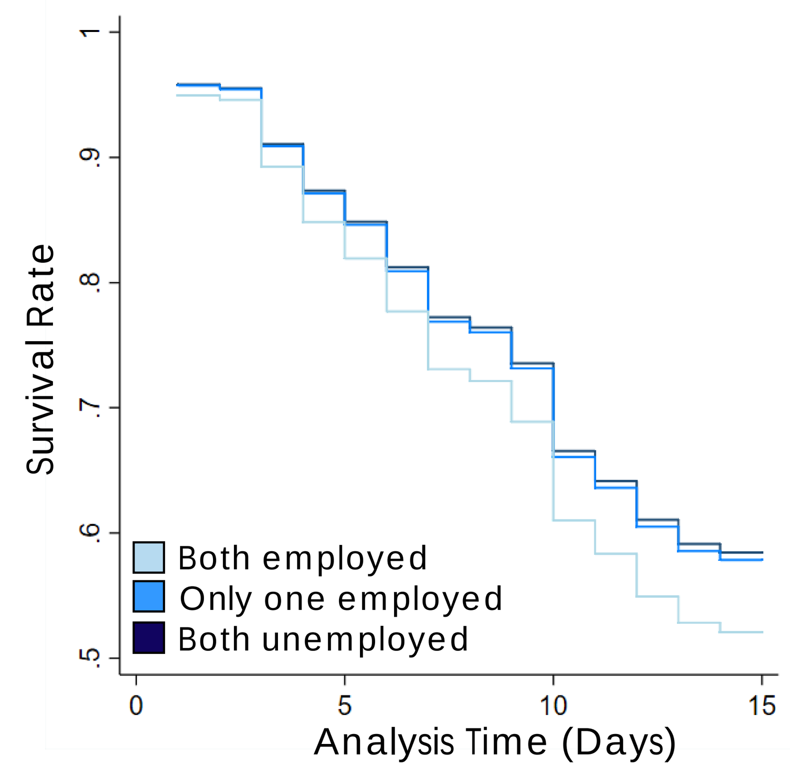
Methodology and Identification

ID 1. Survival analysis and multiple testing

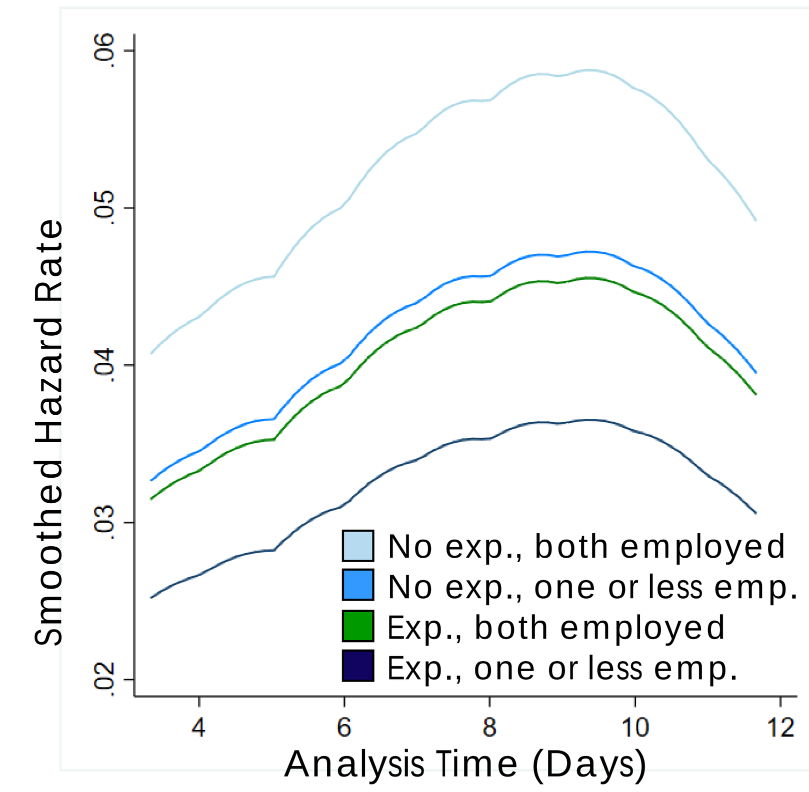
ID 2. Feature importance and prediction by ML

ID 1-1. Survival Regression Analysis

Cox Proportional Hazards Estimate for Survival Rate by ***Parent Employment***



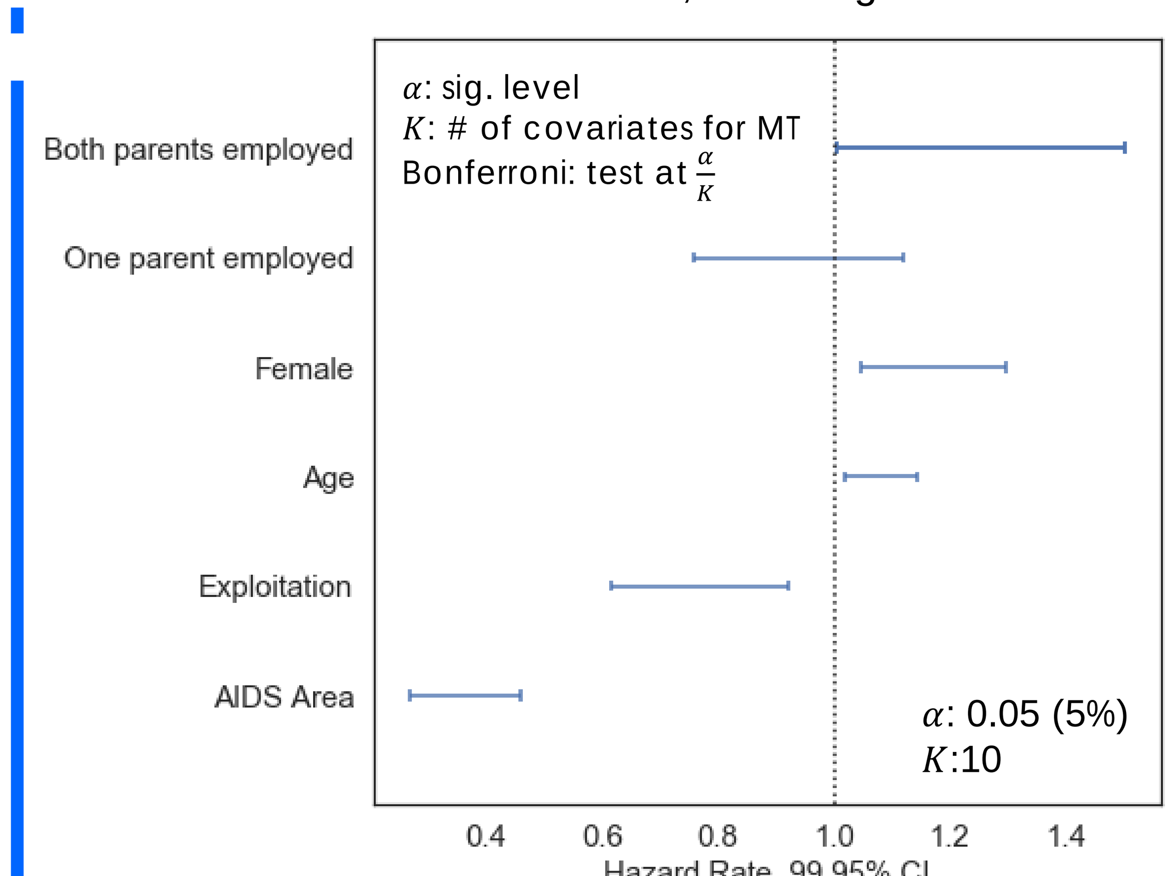
Cox PH Estimate for Hazard Rate by ***Expropriation*** and ***Parent Employment***



**higher hazard / lower survival rate = faster matching*

ID 1-2. Multiple Testing (Bonferroni Correction)

99.5% Confidence Interval Plot for Selected Variables, Cox Regression



*:p<0.01, ^:p<0.1

Variable	Hazard	Robust SE
Urgency	5.4717*	0.5611
AIDS Area	0.3627*	0.0279
Exploitation	0.7678*	0.0442
Age	1.0825*	0.0182
Female	1.1716*	0.5611

Variable	Hazard	Robust SE
One emp.	0.3627	0.9396
Two emp.	1.2546*	0.0713
Educ: unenrolled	0.8290*	0.0495
Educ: pre/kinder	0.7772*	0.0352
Educ: elem/mid	0.8646^	0.0641

Why MT?

(1) Covariates, as signals, are given to PS *at once*; motivation for **multiple testing (MT)**

(2) Sensitivity analysis

Cox PH

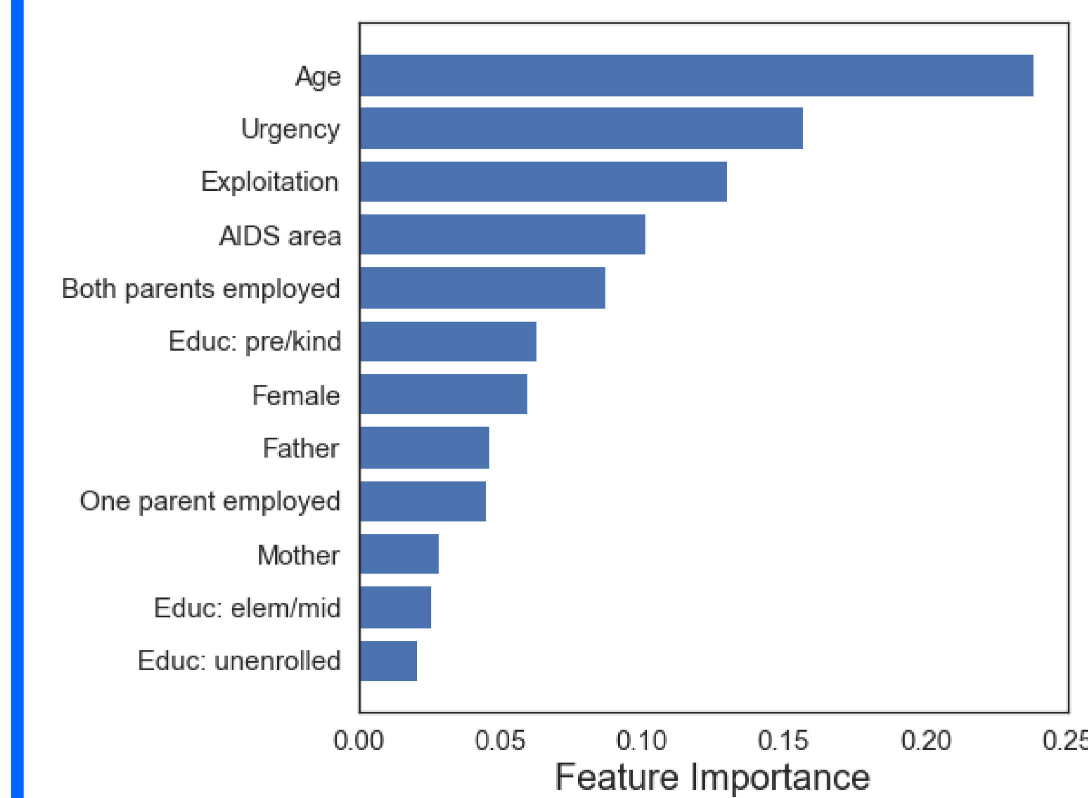
Showcased for semi-parametric baseline hazard function

Initial Results

Covariates alluding to ***instability*** in matching: associated with lower conditional hazard rate even after MT

ID 2-1. Decision Tree & Random Forest

Random Forest, CV for ***Recall***
Plot of Feature Importance



Binary classification

Although multiclass classification possible, **low accuracy of prediction**

Recall optimization

"Out of those that are going to be matched, how many have we correctly identified?" Also did better on **accuracy of prediction**

Initial Results

Accuracy of prediction: ~60%
Feature Importance: similar to regression, but different in ordering by significance

Why Decision Tree / Random Forest?

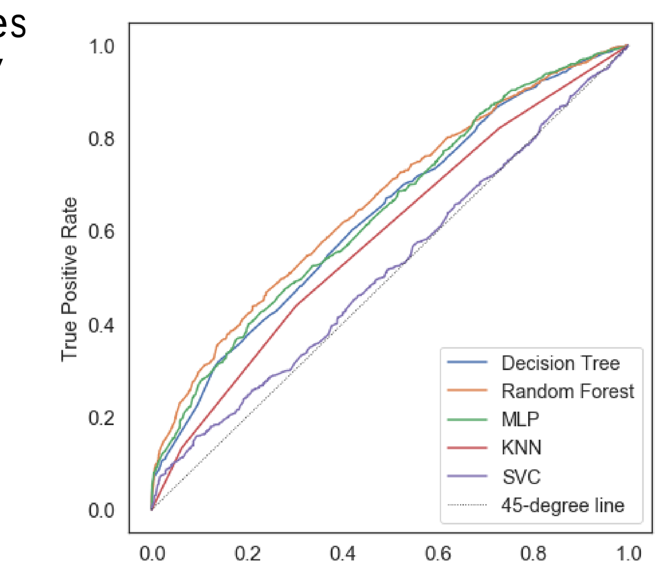
DT / RF classification using **information gain** (entropy) as a metric: a natural extension to MLE using **hazard function** (survival analysis)

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

ID 2-2. Additional Models for Better Prediction

Accuracy and Recall for Matches for the Test Set, Post-HPT and CV

Method	Accuracy	Recall for Matches
DT	0.6067	0.4021
RF	**0.6315**	0.3879
MLP	0.6147	*0.3465*
KNN	0.5819	**0.4388**
SVM	*0.5013*	0.4209



(Left) ROC Curve Comparisons

Initial Results

No method outperforms others by a large margin; *SVC not recommended.*

Discussion of Results and Future Directions

In both (econometric) survival analyses and machine learning cases, covariates indicating **risk in stable matching** (e.g. exposed-to-AIDS area, expropriation): associated with *longer* time to match and vice versa (e.g. parent employment to *shorter* time to match)

Furute Directions: (1) Significance of **urgency** (see *Data Issues*): motivation to use RDD around 180 days of waiting; may require more data points. (2) Longer window of observation may be needed, with some info about the PS.

Acknowledgements

I thank Dr. Richard Evans for insightful guidance and various methods he has taught me through -out the MACSS Perspectives courses.

Contact Information

Junho Choi, MACSS
Computational Economics Concentration
Email: junhoc@uchiago.edu