Classification and Regression Trees and Forests for Incomplete Data from Sample Surveys

MoonJung Cho, Wei-Yin Loh, John Eltinge, and Yuanzhi Li

BLS

Motivation

Analysis of sample survey data often requires adjustments for **missing** values.

Standard adjustments rely on auxiliary variables for **both** responding and non-responding units.

Motivation

Their application can be challenging when the **auxiliary** variables are numerous and are themselves subject to **incomplete-data** problems.

Motivation

Performance depends on the **number** of X variables and their **incomplete-data patterns**.

This paper shows how classification and regression trees and forests can be applied to these cases.

The U.S. Consumer Expenditure (CE) Quarterly Interview Survey

A nationwide household survey conducted by the U.S. Bureau of Labor Statistics.

It collects information on consumers' **expenditures** and **incomes** as well as **characteristics** of the consumers.

BLS

Goal

Estimate the population mean of **INTRDVX**, the amount of interest and dividend income received during the past 12 months.

High rate of item missingness

CE Public-Use Microdata (2013)

- **Consumer Units** (CUs) are roughly equivalent to households.
- Excluded CUs for which INTRDVX codes were "**validly missing**" or "**topcoded**."
- Remaining 4609 CUs:
  1771 **missing** and 2838 **non-missing** INTRDVX

BLS

Missingness in Predictor Variables

Potential **predictor variables** were themselves subject to relatively high item-missingness rates.

- **630** predictor variables available
- **124** variables have missing values
- **67** variables have more than 95% values missing

BLS

Adjusting for Missingness

Form cells to have
common response propensity π or
common mean of Y

Bias under stochastic response model
(Kalton & Maligalig 1991)

$$B\left(\hat{\bar{y}}_{\pi}\right) \doteq \frac{1}{N\bar{\pi}} \sum \left(\pi_i - \bar{\pi}\right)\left(y_i - \bar{Y}_U\right)$$

BLS

Classification and Regression Trees and Forests

Classification trees and forests
to estimate the unit-level **propensity** for item missingness and obtain inverse probability weighted (**IPW**) estimates.

Regression trees and forests
to estimate conditional **means** in **adjustment cells** defined by the nodes of the trees.

Features of GUIDE

Generalized, Unbiased, Interaction Detection and Estimation by W-Y Loh

For the best split variable, **first selects** an X variable, then finds the best split on the **selected X**.

For **missing** values in the X variables, it creates a **missing level** to use in the chi-square tests for variable selection.

INTRDVX_, a flag variable for INTRDVX, is a dependent variable.

Traditional methods of obtaining
the estimated **probability** that y_i is responding,
are difficult to apply due to the **many** X variables
and the large numbers of **missing** values in X.

Classification trees and forests
to estimate the unit-level **propensity**
for item missingness
and obtain inverse probability weighted (**IPW**)
estimates.

Inverse Probability Weighted (IPW) Estimate *(Little, 1986)*

$$\left(\sum_{i \in S_R} \hat{\pi}_i^{-1} w_i\right)^{-1} \sum_{i \in S_R} \hat{\pi}_i^{-1} w_i y_i$$

where

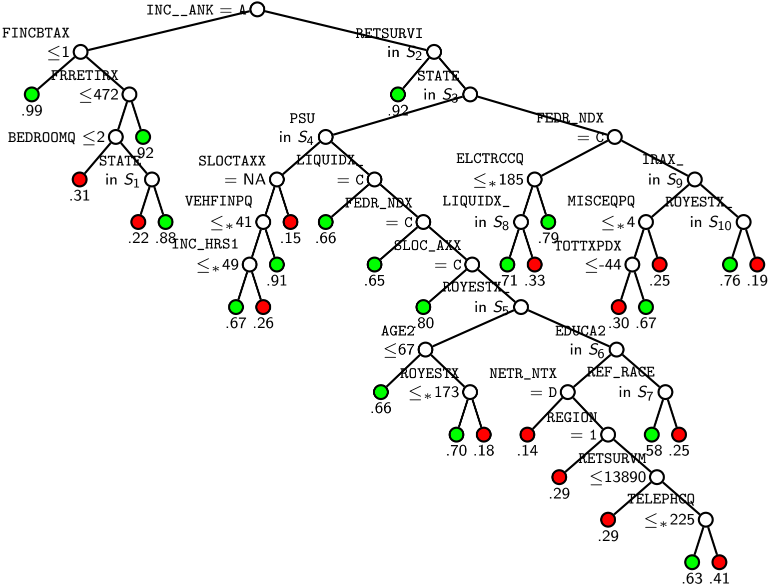
S_R is the sample subset of responding y_i,

$\hat{\pi}_i$ the estimated probability that y_i is responding,

w_i sampling weight.

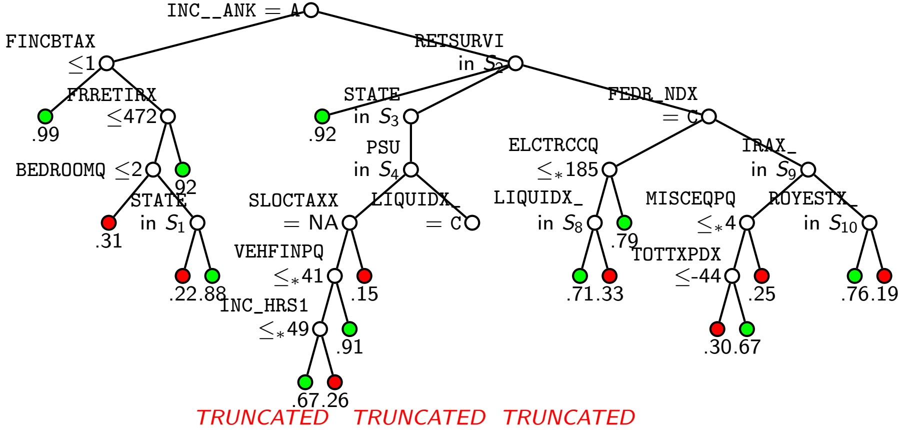
BLS

GUIDE classification tree estimating P(INTRDVX missing)



Enlarged View

GUIDE classification tree estimating P(INTRDVX missing)



We observed

Missing-variable flags
are important predictors of missingness propensity of INTRDVX.

Tree methods can explore the use of **both** observed values and related missing-variable flags.

Regression Trees and Forests

Regression trees and forests are used to model the conditional mean of INTRDVX.

Unlike classification models, the regression tree uses only 2838 CUs with **non-missing** INTRDVX.

Mean Imputation Estimate

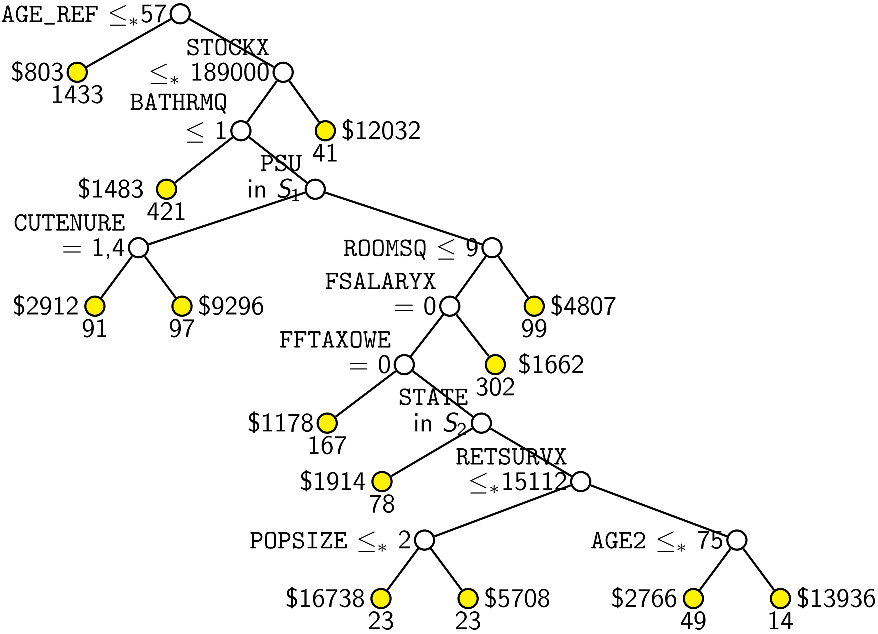
$$\left(\sum_{k\in S} w_k\right)^{-1}\left(\sum_{k\in S_R} w_i y_i + \sum_{j\in S_{NR}} w_j \hat{y}_j\right)$$

where

$S = S_R \cup S_{NR}$,

$\hat{y}_j$ the predicted from X values in S_{NR}.

GUIDE regression tree estimating y_i = INTRDVX



Comparison of Methods

AME	AMELIA imputation (multivariate normal likelihood and EM)
AIPW	IPW using logistic regression with AMELIA for X imputation
MICE	MICE imputation
GMICE	MICE using GUIDE instead of linear and logistic regression
GCT	IPW using GUIDE classification tree
GRT	GUIDE regression tree imputation
GCF	IPW using GUIDE classification forest
GRF	GUIDE regression forest imputation
SIM	Simple estimate ignoring missing responses

Comparison of Methods

Applied methods to 3 nested sets of X variables

- The set of **19** variables for which MICE does not fail
- The set of **52** variables by combining 19 above and the top 20 X variables for predicting INTRDVX_ and INTRDVX
- The full set of **587** variables

Estimates of Mean INTRDVX (SIM = 1900)

	19 variables		52 variables		587 variables	
	Est.	Sec.	Est.	Sec.	Est.	Sec.
AME	2088	139	2184	111068	-	-
AIPW	2055	122	1900	72029	-	-
GCT	1925	8	1946	13	1969	197
GCF	1983	113	1926	173	1914	2028
GRT	2055	8	2010	14	2009	190
GRF	2007	248	1993	360	1944	2030
GMICE	2094	57	2005	434	2002	76874
MICE	2031	430	Fail	-	Fail	-

BLS

Computation Time

- Every method works on the set of 19 variables.
- **MICE** is the slowest for 19 predictors and fails for other two sets.
- **AME** is the second slowest for 19 variables. Computation was terminated for 587 variables.
- Single tree is much faster than forest.

BLS

Estimates of Mean INTRDVX

- **SIM** estimates as $1900 for all three sets.
- Every method works on the set of 19 variables. **MICE** fails for the other two sets
- The estimates range from a low of $1900 (SIM) to a high of $2184 (**AME**, 52 variables)
- Majority of the estimates lie with one **s.e.**($146) of balanced repeated replicate variance estimate and all within two s.e..

BLS

Findings

Classification and Regression Trees and Forests methods are

- often **competitive** with traditional methods in terms of bias and mean squared error for mean estimation.
- **not** limited by sample size.
- **not** hindered or crippled by multicollinearity or quasi-complete separation.
- orders of magnitude **faster** compared to traditional methods.

BLS

Summary

Potential predictor variables were many and were themselves subject to relatively high item-missingness rates.

- Applied **classification** trees to estimate the **propensity** for item missingness, to be used in inverse probability weighting.
- Applied **regression** trees to estimate **conditional means** in adjustment cells defined by the nodes of the trees.

BLS

References

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) *Classification and Regression Trees*. Belmont: Wadsworth

Little, R.J.A. (1986) Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139–157

Loh, W.-Y. (2002) Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, 12, 361–386

Loh, W.-Y. (2009) Improving the precision of classification trees, *Annals of Applied Statistics*, 3, 1710–1737

Loh, W.-Y., Eltinge, J., Cho, M. and Li, Y. (2017) Classification and regression trees and forests for incomplete data from sample surveys, *Statistica Sinica*, in press

MoonJung Cho

cho.moon@bls.gov