CLASSIFICATION AND REGRESSION TREES AND FORESTS FOR INCOMPLETE DATA FROM SAMPLE SURVEYS

Wei-Yin Loh, John Eltinge, Moon Jung Cho and Yuanzhi Li

University of Wisconsin-Madison, U.S. Census Bureau,

Bureau of Labor Statistics, and University of Wisconsin-Madison

Supplementary Material

Section S1 gives the definitions and number of missing values of additional predictor variables, Section S2 shows the results for the 2014 CE data, and Section S3 describes the simulation experiments using parametric models.

S1 Definitions of variables

Table 1 gives the name, definition and numbers of missing values of additional predictor variables.

Table 1: Variable definitions; last column gives number of missing values

Name	Definition	#NA
AGE2_	Flag variable for age of spouse	0
AGE_REF	Age of reference person	0
AS_COMP2	Number of females age 16 and over in CU	0
BATHRMQ	Number complete bathrooms in unit	21

Continued on next page

Table 1 – *Continued from previous page*

Name	Definition	#NA
BEDROOMQ	Number of bedrooms in unit	25
BLS_URBN	Urban or rural	0
BUILDING	Type of building (10 unordered values)	0
BUILT	Year range that the property was built	585
CUTENURE	Housing tenure (owned, rented, etc.; 6 values)	0
EARNCOMP	Composition of earners (8 unordered values)	0
EDUC_REF	Education of reference person (9 ordered values)	0
ELCTRCCQ	Amount spent on electricity this quarter	0
ETOTA	Total outlays last and current quarters	0
FAM_TYPE	Relationship of members to ref. person (9 values)	0
FDHOMECQ	Food at home this quarter	0
FDHOMEPQ	Food at home last quarter	0
FEDRFNDX	Federal income tax by all CU members	2530
FEDR_NDX	Flag variable for Federal income tax refund	0
FEDTAXX	Federal income tax pad by all CU members	3752
FFTAXOWE	Estimate Federal tax liabilities	0
FINCBTAX	Amount CU income before tax in past 12 months	0
FINCBT_X	Flag variable for CU income before taxes	0
FRRETIRX	Soc. sec. and railroad retirement inc. of all members	0
FSALARYX	Wage and salary income of all CU members	0
FSTAXOWE	Estimated State tax liabilities for CU	0
FSLTAXX	State and local inc. taxes deducted for all members	0
HEALTHPQ	Health care last quarter	
HLFBATHQ	Number of half bathrooms	23
INC__ANK	Flag variable for percent income rank	0
INC_HRS1	Hours usually worked per week by ref. person	1697
INCOMEY1	Employer from which ref. person got the most earnings	1697
IRAX_	Flag variable for value of all retirement accounts	0
LIQUIDX_	Flag variable for checking, savings, CDs, etc.	0
MARITAL1	Marital status of ref. person (5 unordered values)	0
MISCEQPQ	Miscellaneous household equipment last quarter	0
NETR_NTX	Flag variable for amount of net rental income or loss	0
NO_EARNR	Number of earners	0
NUM_AUTO	Number of owned automobiles	0
OCCUCOD1	Occupation (18 unordered values)	1697

Continued on next page

Table 1 – *Continued from previous page*

Name	Definition	#NA
OWNDWEPQ	Mortgae interest, property tax, repairs on owned dwellings	0
PERSCAPQ	Personal care this quarter	0
POPSIZE	Population size of the PSU	33
PREDRGPQ	Prescription drugs last quarter	0
PSU	Primary Sampling Unit (21 unordered values)	2579
RACE2	Race of spouse	1879
REF_RACE	Race of reference person (6 unordered values)	0
REGION	Region of country (NE, MW, S, W)	33
RENTEQVX	Monthly rent if home rented	660
RETPENCQ	Retirement, pensions, Social Security this quarter	0
RETSURVI	Descriptor variable for income imputation	0
RETSURVI	Imputation descriptor for RETSURVX (3 categories)	0
RETSURVM	Mean income from retirement, survivor, disability pensions	3289
RETSURVX	Amount received in retirement, survivor, or disability pensions	3520
ROOMSQ	Number of rooms in unit	30
ROYESTX_	Flag variable for income from royalty, estates, trusts	0
SEX_REF	Sex of reference person (2 values)	0
SLOCTAXX	Total amount paid for state and local income taxes	3990
SLOC_AXX	Flag variable for SLOCTAXX	0
SLRF_NDX	Flag variable for SLRFUNDX	0
ST_HOUS	Living quarters used as student	0
housing (yes/no)	0	
STATE	State (39 categories)	486
STOCKX	Value of directly-held stocks, bonds, mutual funds	4319
TOTTXPDX	Personal taxes paid by CU in past 12 months	0
TOTXEST	Estimated total taxes paid	0
VEHFINPQ	Vehicle finance charges last quarter	0
VEHQ	Number of owned vehicles	0

Table 2: Estimates (weighted by FINLWT21) and computation times of mean INTRDVX for 3 sets of predictor variables for the 2014 CE data. Results for AIPW, AMELIA and MICE are based on 5 multiple imputations. AIPW and AMELIA did not yield any results for 573 variables after more than 6 months.

	19 variables		49 variables		573 variables	
	Est.	Sec.	Est.	Sec.	Est.	Sec.
AIPW	2200	208	2139	55887	-	-
AME	2219	212	2154	71168	-	-
GCT	2243	8	2027	16	2057	240
GCF	1853	95	1846	168	1994	2527
GRT	2168	8	2187	16	2146	229
GRF	2182	201	2129	320	2124	2261
DRT	2287	16	2155	32	2159	469
DRF	2312	209	2121	336	2120	2501
GMICE	2241	51	2178	663	2204	67676
MICE	2325	421	Fail	-	Fail	-
RCT	2059	1	2040	1	1905	16
RRT	2050	1	2052	1	2015	6
SIM	2059	-	2059	-	2059	-

S2 2014 CE data

The 2014 CE data have fewer variables than the 2013 data. We again employ three nested sets of predictors variables to estimate the population mean of INTRDVX. The first is the same set of 19 used in the 2013 data. The second and third sets contain 49 and 573 predictors, respectively.

The estimates of mean INTRDVX are shown in Table 2 and graphed in Figure 1. The SIM estimate is \$2059, which is \$159 higher than that for the 2013 data. Again MICE works only with the smallest set of variables. Most of the methods yield estimates within one standard error of SIM; the exceptions are AMELIA, MICE and GMICE (all with 19 predictors).

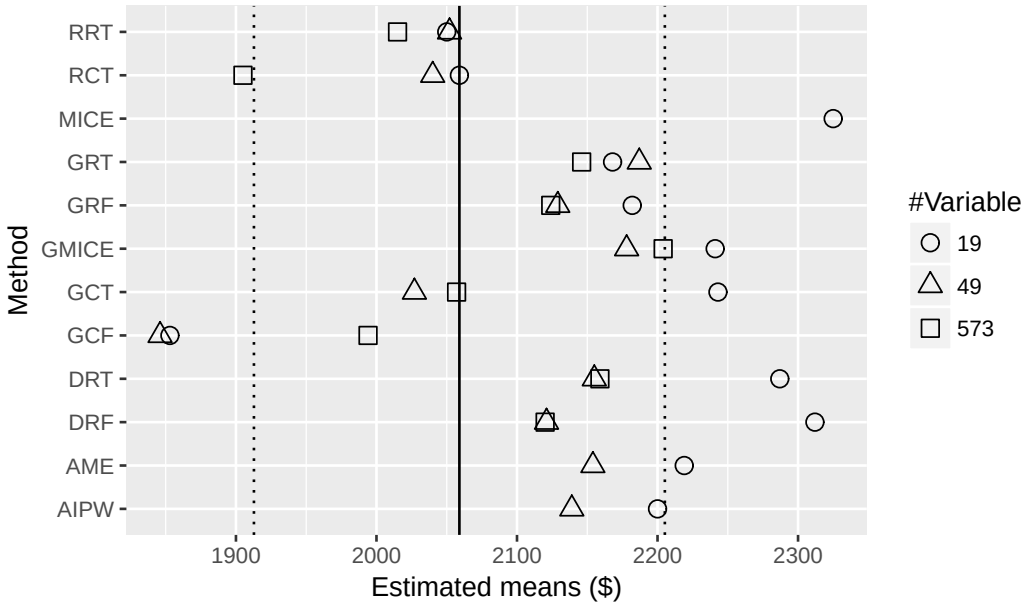

Figure 1: Estimates of mean INTRDVX (weighted by FINLWT21) for 2014 CE data using 19, 49 and 573 predictor variables. The solid line marks the value of the SIM estimate of $2059 and the dotted lines mark SIM plus and minus one standard error of $181 (calculated from balanced repeated replication weights).

S3 Simulations with parametric models

The 6 of the 19 X variables in the 2013 CE data without missing values were used to generate Y and its missing values. The variables are $X_1 =$ `AGE_REF`, $X_2 =$ `BLS_URBN`, $X_3 =$ `EDUC_REF`, $X_4 =$ `NO_EARNR`, $X_5 =$ `NUM_AUTO`, and $X_6 =$ `ETOTA`. Following are the steps in each simulation trial.

1. Let $\mathcal{P}_1$ be the CE dataset with only variables (`INTRDVX`, $X_1, X_2, \ldots, X_{19}$). Let π denote the probability that `INTRDVX` is non-missing and fit the logistic regression model $\log(\pi/(1-\pi)) = \gamma_0 + \sum_{i=1}^{6} \gamma_i X_i$ to $\mathcal{P}_1$. Let $\hat{\gamma}_0, \hat{\gamma}_1, \ldots, \hat{\gamma}_6$ denote the estimated coefficients.

2. (a) For the constant mean model, replace `INTRDVX` in $\mathcal{P}_1$ with Y, an independent normally distributed variable with mean 0 and variance equal to the variance of the non-missing `INTRDVX` values in $\mathcal{P}_1$.

   (b) For the linear model, fit `INTRDVX` $= \beta_0 + \sum_{i=1}^{6} \beta_i X_i + \epsilon$ to the subset of observations in $\mathcal{P}_1$ with complete responses in (`INTRDVX`, $X_1, \ldots, X_6$). Let the least squares estimates of the regression coefficients be $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_6)$ and the residual variance be $\hat{\sigma}^2$. For each unit in $\mathcal{P}_1$, replace `INTRDVX` with $Y = \hat{\beta}_0 + \sum_{i=1}^{6} \hat{\beta}_i X_i + \varepsilon$, where ε is independent normal with mean zero and variance $\hat{\sigma}^2$.

Compute the mean μ of Y in $\mathcal{P}_1$.

3. Make some Y in $\mathcal{P}_1$ missing according to the model $\log(\pi/(1-\pi)) = \hat{\gamma}_0 + \sum_{i=1}^{6} \hat{\gamma}_i X_i$.

4. Make the values in $X_1, \ldots, X_6$ each independently MCAR with probability 0.05. Denote the resulting dataset by $\mathcal{P}_2$.

5. Draw a 10% random sample without replacement from $\mathcal{P}_2$ and apply each method to it to estimate μ.

The results are shown in Figure 2.

Highly correlated predictors

We used 4 transportation variables in the 2013 CE data to simulate observations with highly correlated predictor variables. Tables 3 and 4 give the names and definitions of the variables and their correlations. The steps in each simulation trial are the same as those above, except that the 4 variables replace $X_1, X_2, \ldots, X_6$ and there are no other X variables. The biases and RMSEs are shown in Figure 3 with 2-SE bars.

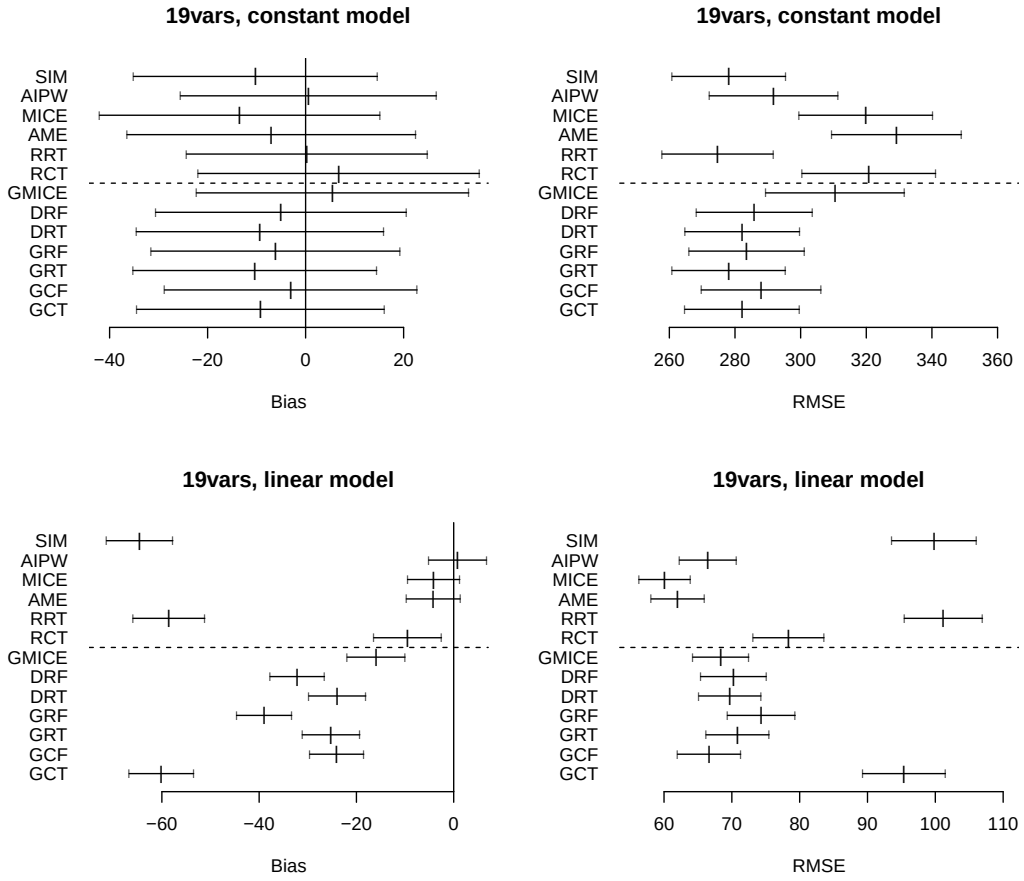

Figure 2: Bias and RMSE (with 2-SE bars) for constant (upper) and linear (lower) regression models with 19 X variables, 10% sampling fraction, logistic missing propensity in Y, and 5% MCAR in each X variable. Dashed lines separate GUIDE from non-GUIDE methods.

Table 3: Four highly correlated transportation expenditure variables

PUBTRAPQ	Public and other transportation last quarter
TFAREP	Trip expenditures last quarter on transportation, including airfare, intercity bus, train, and ship
TTRANPRP	Total trip expenditures on transportation last quarter including airfare, local transportation, tolls and parking fees, and car rentals
TTRNTRIP	Trip expenditures last quarter for public transportation, including airfares

Table 4: Correlation matrix of 4 predictor variables

	PUBTRAPQ	TTRANPRP	TTRNTRIP	TFAREP
PUBTRAPQ	1.0000	0.9865	0.9950	0.9943
TTRANPRP	0.9865	1.000	0.9912	0.9906
TTRNTRIP	0.9950	0.9912	1.0000	0.9995
TFAREP	0.9943	0.9906	0.9995	1.0000

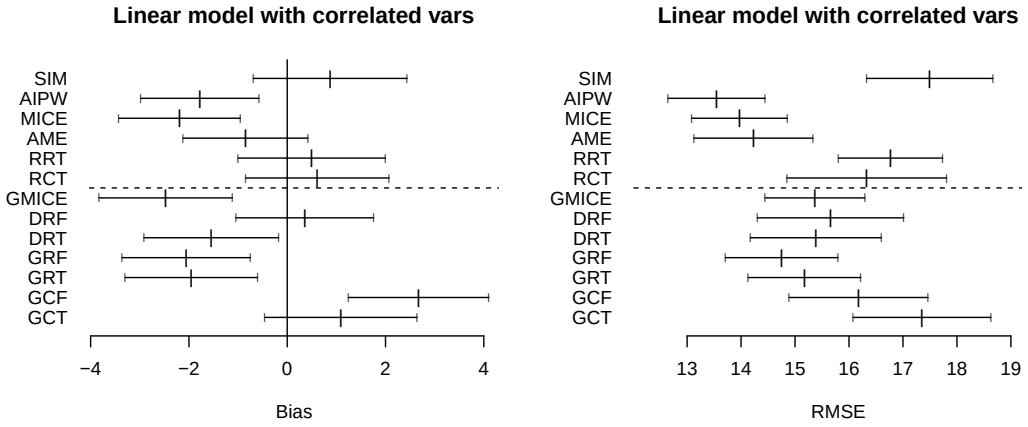

Figure 3: Bias and RMSE (with 2-SE bars) for linear regression model with 4 highly correlated X variables, 10% sampling fraction, logistic missing propensity for Y, and 5% independently MCAR for each X. Dashed lines separate GUIDE from non-GUIDE methods.