Effect of Improving Data Imputation and Processing Procedures on ERS Farm Household Income Forecasts

Wei-Yin Loh

University of Wisconsin, Madison

Joint work with David Williams, Daniel Ayasse, Katherine Lim and Christine Whitt (Economic Research Service)

The findings and conclusions in this presentation are those of the authors and should not be construed to represent any official USDA or U.S. Government determination or policy

Motivation

- Mission of USDA's Economic Research Service (ERS): to anticipate trends and emerging issues in agriculture, food, the environment, and rural America and to conduct high-quality, objective economic research to inform and enhance public and private decision making
- Three times each year, ERS produces estimates and forecasts of incomes in order measure and predict well-being of farm households
- Reports on well-being of farm households presented to public and policymakers
- Objective of project: improve estimates of farm household's well-being
- Underlying dataset providing information on farm household well-being is ARMS (Agricultural Resource Management Survey)
- Some variables have high levels of non-response requiring imputation

Current ERS imputation method

- Process uses a number of methods to assign values to missing data
- Main method is imputation with conditional means in cells defined by pairs of demographic variables (e.g., 15 cells defined by Age Class and Education)

	EDUC1	EDUC2	EDUC3
AGE1			
AGE2			
AGE3			
AGE4			
AGE5			

Two approaches to mean estimation

- Let S_1 and S_2 be the subsets of nonmissing and missing y_i, respectively
- Let $\hat{\pi}_i$ be the estimated propensity score (probability that y_i is nonmissing)
- Let $\hat{y}_i$ be the imputed value of y_i if it is missing
- Let w_i be the sampling weight (if any)

1. **Inverse probability weighted estimate** via propensity scores

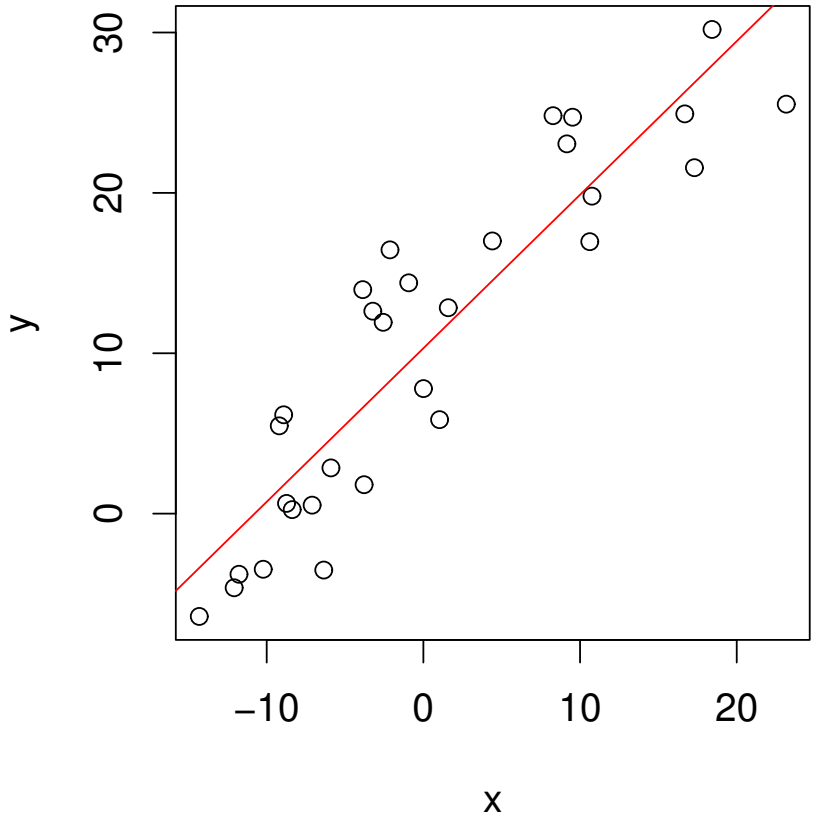
$$\Big(\sum_{i \in S_1} w_i / \hat{\pi}_i \Big)^{-1} \sum_{i \in S_1} w_i y_i / \hat{\pi}_i$$

2. **Missing value imputation** by regression modeling or hot-deck sampling

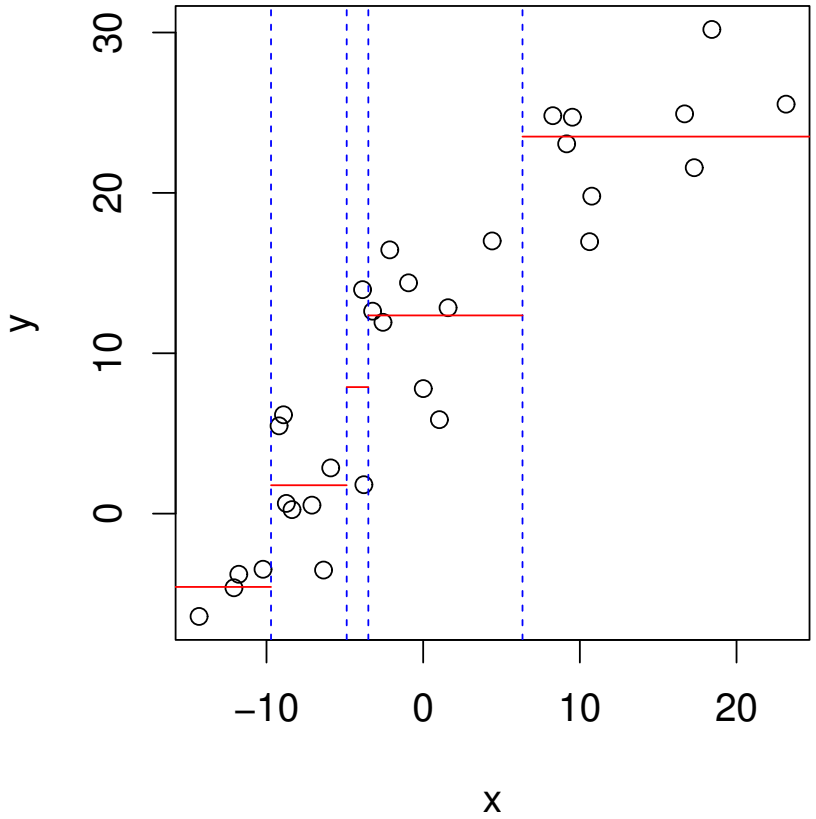
$$\Big(\sum_{i \in S_1 \cup S_2} w_i \Big)^{-1} \Big(\sum_{i \in S_1} w_i y_i + \sum_{j \in S_2} w_j \hat{y}_j \Big)$$

Regression imputation (assumptions: same correct model for missing and nonmissing data)

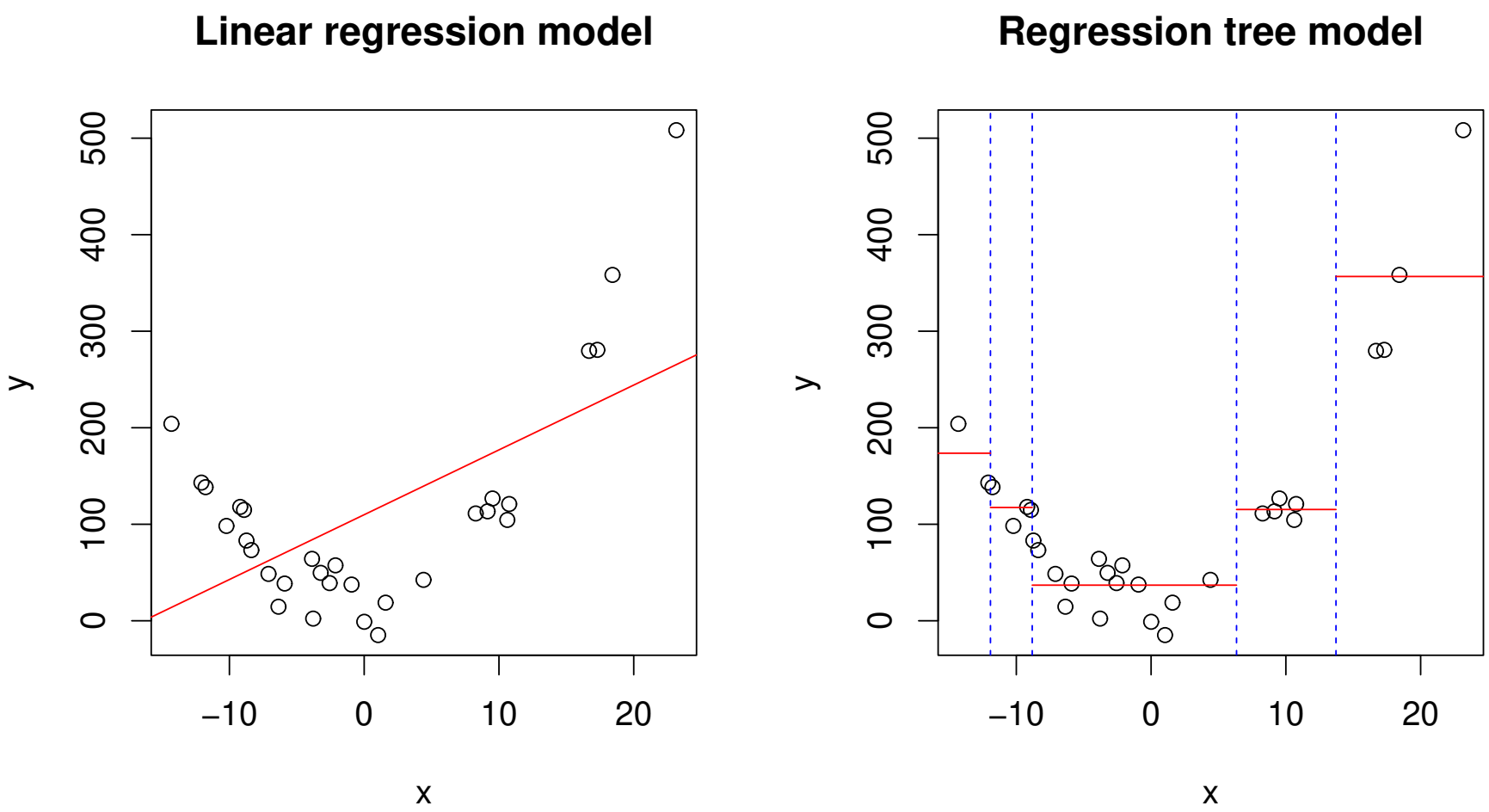
Linear regression model



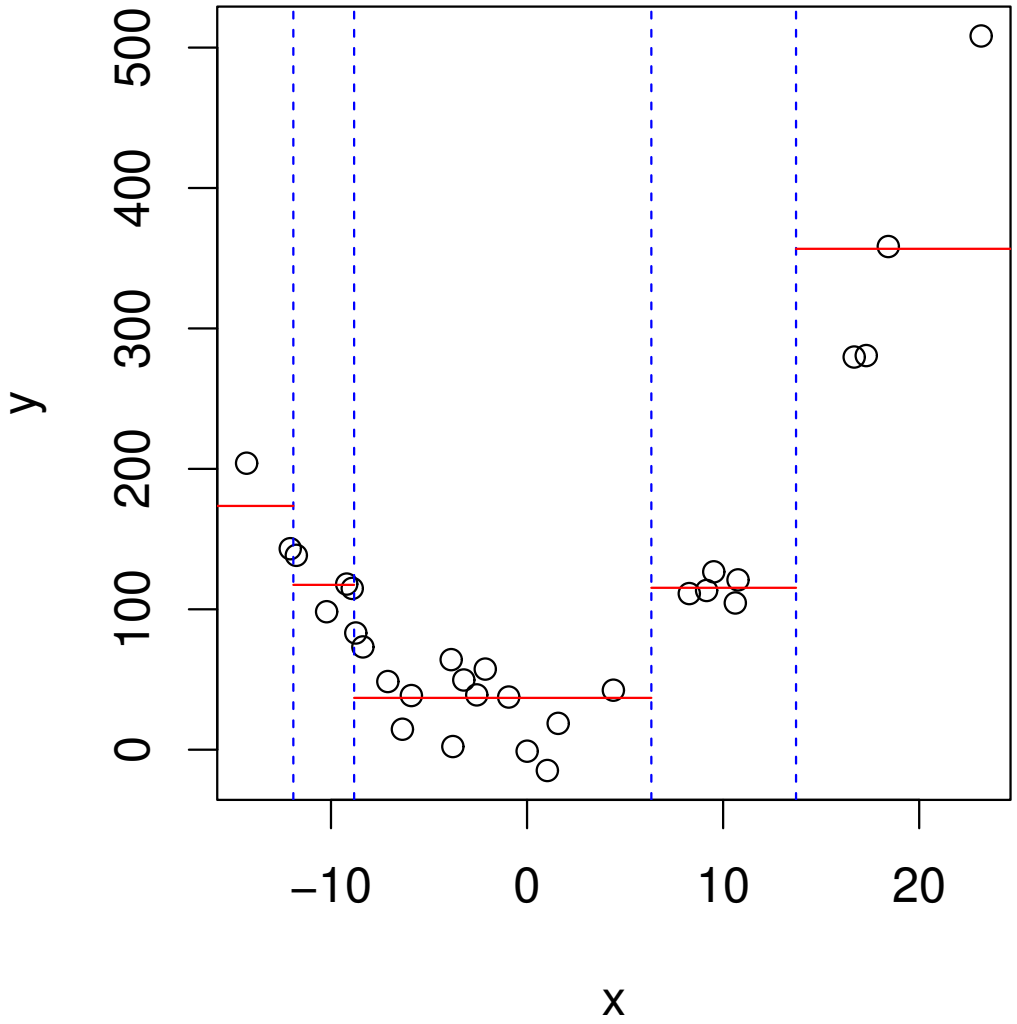
Regression tree model

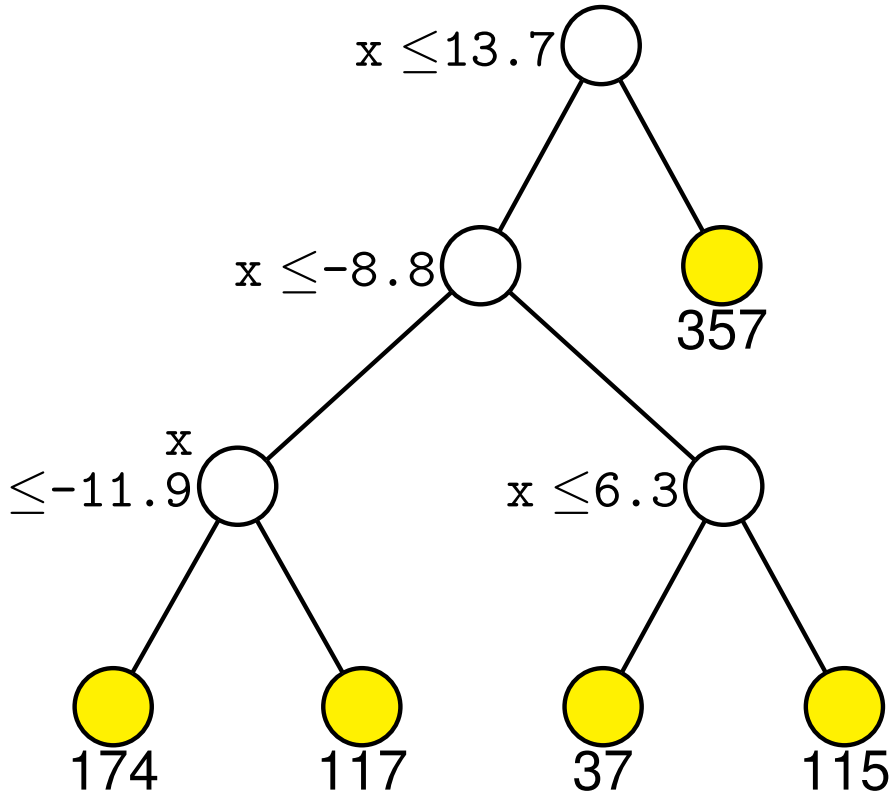


Regression imputation (nonlinear)

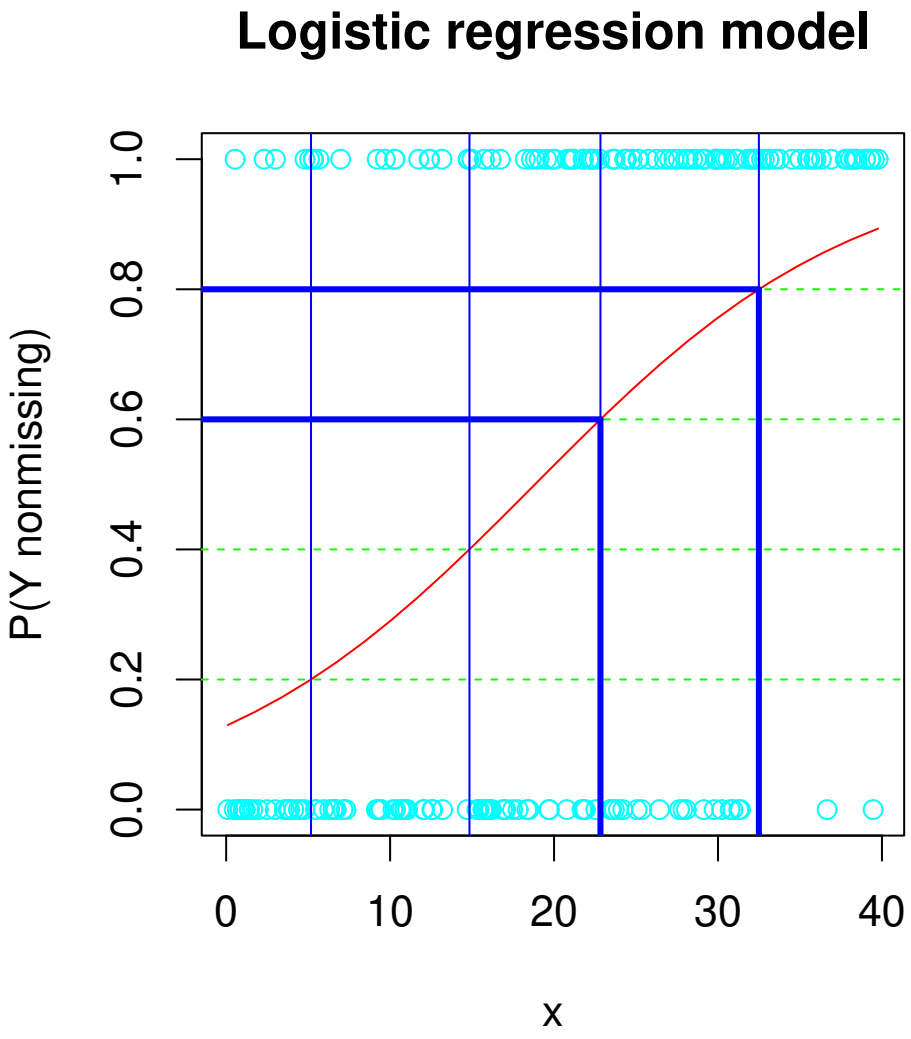


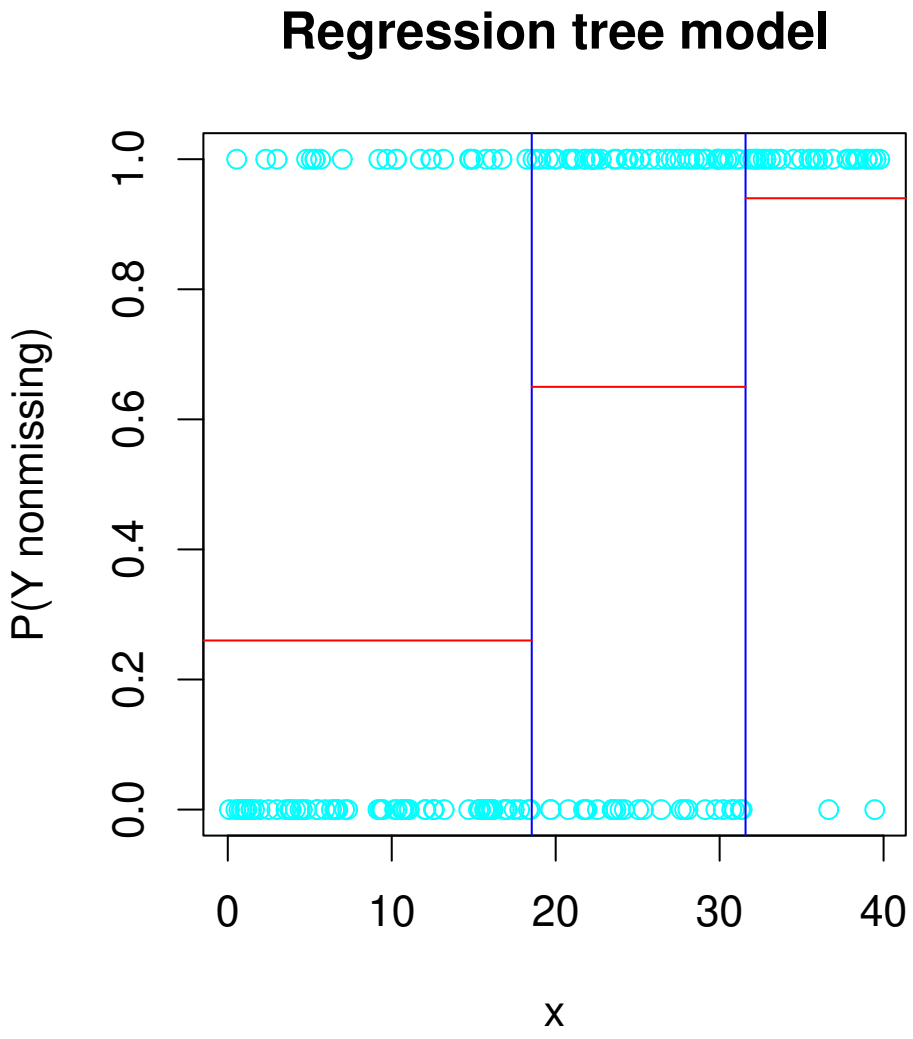
Piecewise-constant regression tree model



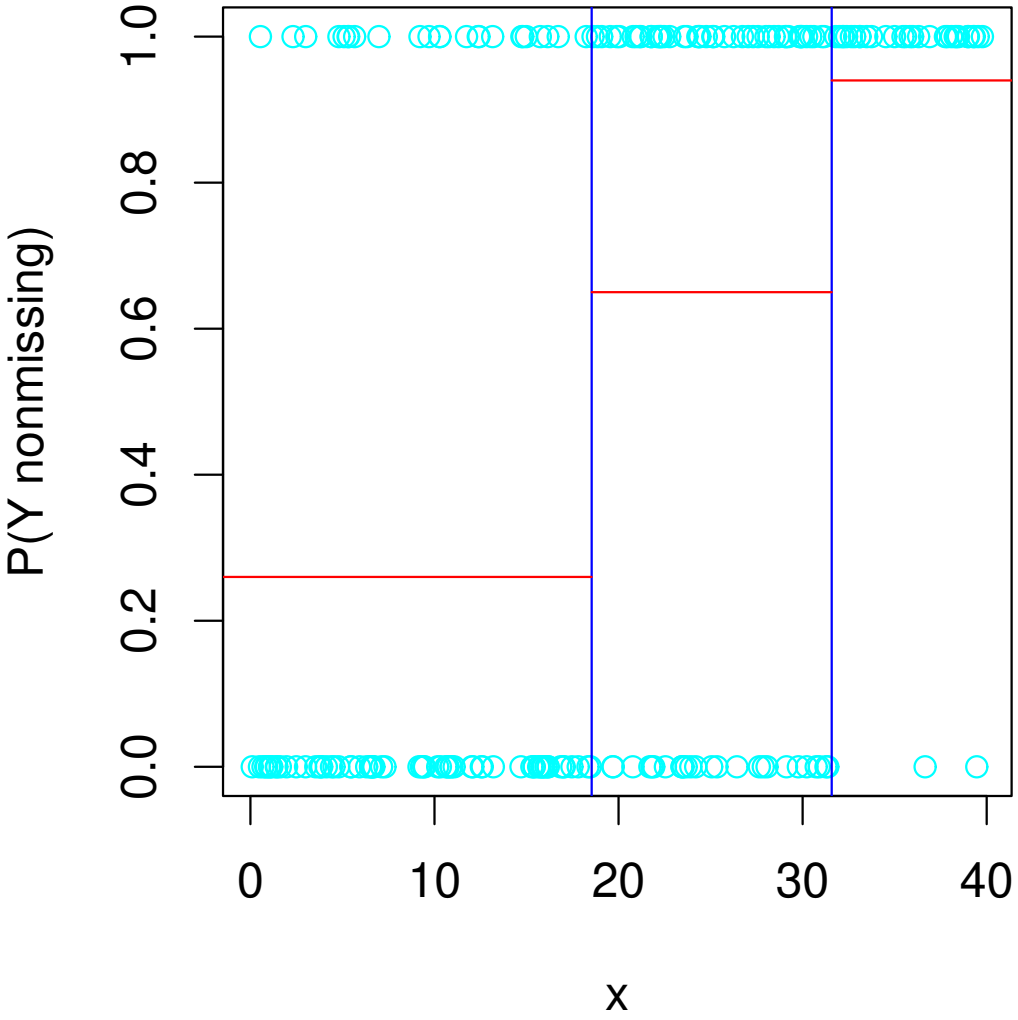


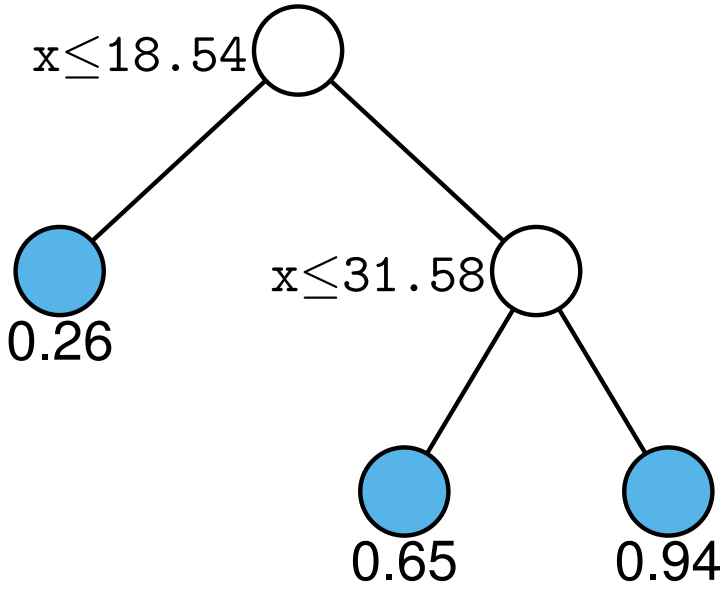
Hot-deck imputation cells using propensity scores fitted by logistic regression and regression tree



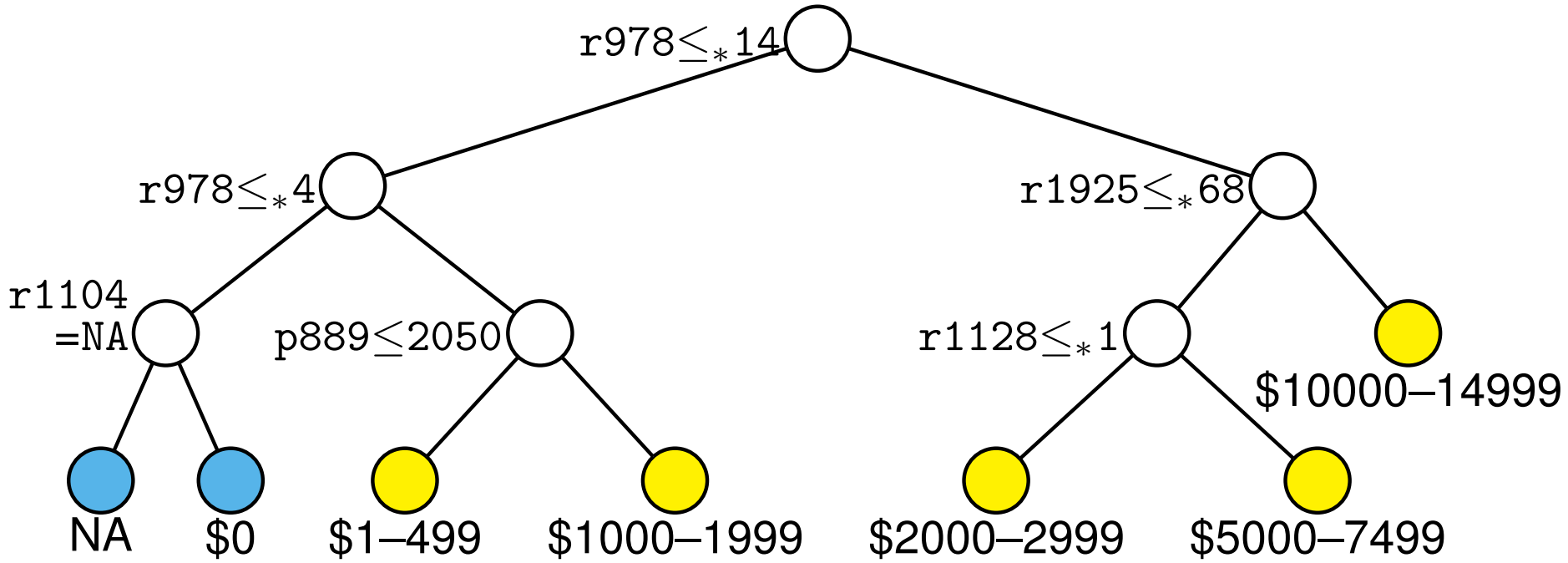


Piecewise-constant propensity score tree model





GUIDE regression tree for predicting r981 (gain/loss on sale of capital assets)



r889. estimated market value of farm crops owned and stored

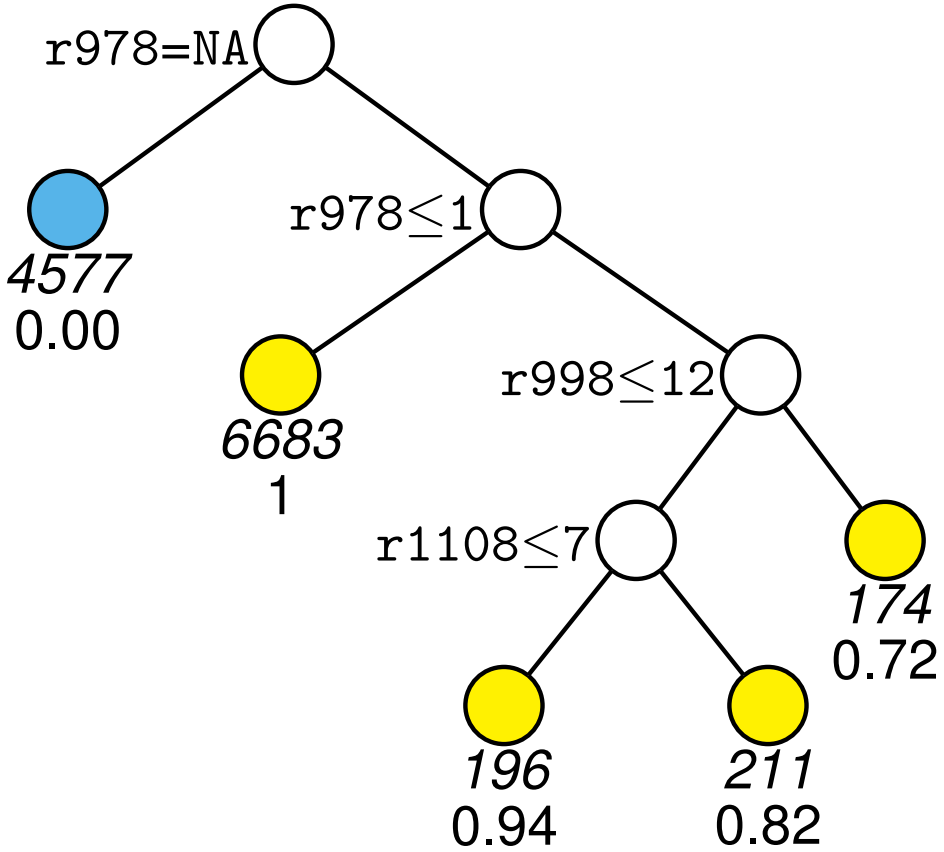
r978. total proceeds from the sale of farm and non-farm capital assets

r1104. rent payments for principal producer's dwelling

r1128. number people living in household between 18–64 with health insurance

r1925. age of Person 1

Propensity score tree for P(r981 = nonmissing)



r978. total proceeds from the sale of farm and non-farm capital assets

r998. income from public sources (e.g., social security, veteran's benefits)

r1108. health and/or dental insurance costs

Differences between methods

	Current	GUIDE	Change
	Median dollars per household		
Farm income	-1,198	306	1,504
Off-farm income: Total	67,873	54,650	-13,223
Off-farm income: Earned Income	32,428	30,000	-2,428
Off-farm income: Unearned Income	31,057	18,900	-12,157
Total household income	80,060	71,618	-8,442
	Mean dollars per household		
Farm income	25,566	28,288	2,722
Off-farm income: Total	96,688	87,810	-8,878
Off-farm income: Earned Income	63,530	58,803	-4,727
Off-farm income: Unearned Income	33,158	29,006	-4,152
Total household income	122,255	116,098	-6,157

Farm household incomes may be overestimated using current method

Farm household incomes exceeded nonfarm households beginning 1990's (Key et al., 2017)

Advantages of GUIDE

1. Automatically selects predictor variables to form “cells” for imputation
2. Does not impute missing values in predictor variables
3. Accepts interval-coded variables
4. Imputation cells explicitly described by decision tree diagrams

References

Key, N., Prager, D., and Burns, C. (2017). Farm household income volatility: An analysis using panel data from a national survey. Technical Report ERR-226, U.S. Department of Agriculture, Economic Research Service.

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

Loh, W.-Y. (2014). Fifty years of classification and regression trees (with discussion). *International Statistical Review*, 34:329–370.

Loh, W.-Y. (2023). GUIDE website.
`https://pages.stat.wisc.edu/~loh/guide.html`.

Loh, W.-Y., Eltinge, J., Cho, M. J., and Li, Y. (2019). Classification and regression trees and forests for incomplete data from sample surveys. *Statistica Sinica*, 29:431–453.