Regression Tree Models for Analyzing Survey Response

Daniell Toth* Polly Phipps†

Abstract

Modeling various conditional response propensities for a survey based on known unit characteristics or contact history is important when analyzing survey response. One increasingly important technique is to model these conditional response propensities using non-parametric regression tree models. Regression trees provide mutually exclusive cells with homogeneous response propensities that make it easier to identify interpretable associations between class membership characteristics and response propensity, compared to other regression models. We provide examples of how regression trees have been used to analyze survey response, including: gaining insight into how characteristics of sample members are associated with response, incorporation of auxiliary variables and para-data for use in adaptive designs and follow-up procedures, and the identification of auxiliary variables for nonresponse adjustment.

Key Words: adaptive sampling; automatic interaction detection; non-parametric regression; nonresponse; propensity models; survey data.

1. Introduction

Large scale surveys generally have sample units that do not provide the desired data. Such units are considered nonrespondents. The missing information resulting from nonresponse could have a negative impact on the quality of the information the survey provides. Specifically, a biased estimate could result from using data obtained from a survey with missing responses if the responses obtained from responding units tend to differ from those of nonresponding units. For example, the bias for an unadjusted mean estimator $\bar{Y}_R = N_R^{-1} \sum_{i \in N_R} y_i$ is

$$(1 - r)(\bar{Y}_R - \bar{Y}_{R'}), \tag{1}$$

where r is the response rate for outcome variable Y, and $\bar{Y}_R$ and $\bar{Y}_{R'}$ are the mean outcome for respondents and nonrespondents respectively. Thus, the bias of an unadjusted mean estimator using only the data from respondents depends on the difference between the mean of respondents and that of nonrespondents.

If the variable Y is independent of response, the data is said to be missing completely at random (MCAR). In this case, $\bar{Y}_R = \bar{Y}_{R'}$ and so by equation (1), the bias is zero even if the rate of nonresponse $(1 - r)$ is large. In this situation, the unadjusted estimator is unbiased and therefore no adjustment for nonresponse is necessary. This is unlikely to occur in practice. A less aggressive assumption, is that the nonresponse data is missing at random (MAR). That is, Y is independent of response given the auxiliary variables $\mathbf{X}$. In this situation, the estimator is unbiased after being adjusted for the differing values of $\mathbf{X}$. An even weaker assumption is that the missing data due to nonresponse is missing not at random (MNAR). Under this assumption, Y is not independent of the response rate even conditionally given the auxiliary variables $\mathbf{X}$. In this situation $\bar{Y}_R \neq \bar{Y}_{R'}$ even after adjusting for $\mathbf{X}$. However, if the variable of interest Y is associated with $\mathbf{X}$, the bias of the estimator due to nonresponse can be reduced using the auxiliary data.

*Office of Survey Methods Research, Bureau of Labor Statistics, Suite 1950, Washington, DC 20212 (email:toth.daniell@bls.gov)

†Office of Survey Methods Research, Bureau of Labor Statistics, Suite 1950, Washington, DC 20212 (email:phipps.polly@bls.gov)

Therefore, identifying auxiliary variables from the set of known variables $\mathbf{X}$, which are associated with the variable of interest and response propensity is very important for both analyzing and mitigating the effect of nonresponse on the quality of the data (Kreuter et al. 2010). It is also helpful to have a good model of the probability of response given $\mathbf{X}$, to better understand the potential impact of nonresponse on a given survey.

Let $\mathbf{x}_i$, and $R(\mathbf{x}_i)$ be the known auxiliary data and the response indicator given the auxiliary data for unit i, respectively. The probability that unit i responds given the units' known auxiliary data,

$$P(R_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i, y_i) = p(\mathbf{x}_i), \tag{2}$$

is usually modeled using data from the responding units and all known auxiliary data. A good model is also helpful in assessing the likely impact that a change in data collection procedures will have on the response rate of a given survey. The model may incorporate auxiliary data available from the sample frame, as well as data obtained during the attempts to collect the survey data (paradata). This usually provides a large number of variables to potentially include in the model and nonresponse is often determined by a number of interactions between the variables. This could make modeling the response propensity difficult (Schouten 2007).

In this article we review a number of applications of the nonparametric approach, Automatic Interaction Detection (AID), also known as regression trees, used to address survey nonresponse. We give a brief review of how survey nonresponse is currently handled in Section 2. A much more comprehensive review can be found in Brick (2014). In section 3 we show how tree models, in comparison to the more often used logistic regression models, allow for easy identification of characteristic variables that are associated with response propensity, and their coefficients are easy to interpret. A comprehensive review of regression trees and their history is given in Loh (2014). In Section 4 we provide several examples of the use of regression trees to analyze and adjust for survey nonresponses found in the literature and suggest potential future applications.

2. Addressing Survey Nonresponse

In order to mitigate the effects of nonresponse on survey estimates, adjustments are made to the data collection procedure while the data is being collected, or more likely, to the estimator after the data has been collected, or both. Either approach requires identifying auxiliary variables $\mathbf{X}$ that are associated with response indicator variable R and the variable of interest Y (see Kreuter et al. 2010).

Attempting to reduce the chance of nonresponse bias for a survey estimate by changing the collection procedure during data collection is referred to as responsive (or adaptive) data collection. For examples of how adaptive data collection is used in practice, see Mohl and Laflamme (2007) and Wagner et al. (2012), and Section 4.3. Based on models using unit characteristics, the response propensity of each unit is estimated and sample allocation, collection efforts, and other resources are distributed in order to balance the obtained sample. This is undertaken with the idea that the closer the characteristic proportions of the obtained sample matches the characteristic proportions of the population, the less likely the estimates are to be biased (Peytcheva and Groves 2009 and Schouten, Cobben and Bethlehem 2009).

One option for adjusting the dataset to reduce or remove nonresponse bias after data collection is to impute the missing data using a model (implicitly or explicitly) estimated from the observed data (Brick and Kalton 1996). Another option is to adjust the weights of the observed units to account for the units that are missing. For example, the weight for

each observed sample unit is itself weighted by the inverse of the estimated probability to respond for that unit.

This could lead to extreme weights, so more often, weight adjustment is done by using weight adjustment cells. The predicted probability of response is used to divide the sample units into cells of homogeneous response propensities and then each unit in the cell is given the same non-response weight adjustment. These cells are then used in adjusting estimates for nonresponse bias (see Vartivarian and Little 2002). This is the case when either weighting (see for example Kim and Kim 2007) or calibration is used to adjust for nonresponse (Kott and Chang 2010).

A common tool used to model the response propensity is the parametric logistic regression model (Rosenbaum and Rubin 1983; Little 1982). The response propensity for unit i given characteristic variables $\mathbf{x}_i$, is modeled by

$$\hat{p}(\mathbf{x}_i) = (1 + \exp\{-z_i\})^{-1}, \tag{3}$$

where $z_i = \beta\mathbf{x}_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$. In order to produce mutually exclusive response cells for which the propensity is approximately equal, the units are often grouped according to the quantiles of their predicted propensities (see Eltinge and Yansaneh 1996).

For the the logistic regression model assumption to hold, the response rate must be monotonic with respect to the model variables. Let

$$\text{logit}(p(\mathbf{x})) = \log\left(p(\mathbf{x})\right) - \log\left(1 - p(\mathbf{x})\right),$$

then for the logistic model to hold, we can see from equation (3) that

$$\text{logit}(\mathbf{x}) = \beta\mathbf{x}, \tag{4}$$

must be linear. If the modeler fails to find a set of variables that result in a linear relationship, the model will suffer from lack of fit. For a set of variables, $\mathbf{X}$ to satisfy equation (4) adequately, a number of interaction terms are often needed. This can make the resulting model and the formed groups very difficult to interpret.

3. Regression Trees

When the primary goal requires the identification of a set of unit characteristics that partition the units into groups of similar responding establishments, regression trees are very helpful. Tree regression is a nonparametric approach that automatically results in mutually exclusive response cells, $C_1, \ldots, C_{k+1}$, based on unit characteristics. Classes are constructed using characteristic variables known for all sampled units such that each cell contains units with homogeneous propensity scores.

A recursive partitioning algorithm is used to build a binary tree that describes the association between an unit's characteristic variables and its propensity to respond. A recursive partitioning algorithm begins by splitting the entire sample, $\mathcal{S}$, into two subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ according to one of the characteristic variables. The desired value (in this case the proportion of respondents) is then estimated for each subset separately.

At each step, the split that minimizes the loss function being used to evaluate the model is chosen from among all possible splits on the auxiliary variables. For example, if estimated mean squared error is the criteria used to evaluate a model fit, then the split that results in the largest decrease in the estimated mean squared error will be selected. To estimate $p(\mathbf{x},)$ the mean value $p(\mathbf{x})$ for each subset, $\mathcal{S}_j$, $j = 1,\ 2$, is estimated separately, using

$$\Big(\sum_{i\in S} \mathbb{1}_{\{\mathbf{x}_i \in \mathcal{S}_j\}}\Big)^{-1} \sum_{i\in S} R_i \mathbb{1}_{\{\mathbf{x}_i \in \mathcal{S}_j\}}. \tag{5}$$

It is easy to incorporate sampling weights into this estimator to obtain a regression tree that provides design consistent estimates for the response propensity, $p(\mathbf{x})$ (see Toth and Eltinge 2011).

This results in a set of mutually exclusive response cells, $C_1, \ldots, C_{k+1}$, based on unit characteristics, that contain units with homogeneous propensity scores. The resulting tree model can then be written as

$$p(\mathbf{x}) = \mu_1 C_1(\mathbf{x}) + \ldots + \mu_{k+1} C_{k+1}(\mathbf{x}), \tag{6}$$

where C_i is the indicator function of whether a given unit's characteristics designate membership to class i.

For example, if class C_i is defined as satisfying the first d splits and not satisfying the rest, then the estimated propensity of establishments in that class is given by

$$\mu_i = \beta_0 + \ldots + \beta_d,$$

for some set of $\{\beta_j\}_{j=1}^{k}$. More generally, in this form, the response propensity for a given unit is β_0, plus the sum of all the coefficients β_i for which the unit's characteristics satisfy split S_i.

4. Applications of Regression Trees in Survey Nonresponse

We now discuss three areas of application for regression trees concerned with nonresponse: adjusting for nonresponse; analyzing nonresponse; and adaptive data collection.

4.1 Nonresponse Adjustment

One way of dealing with nonresponse (or missing values in general) is to model the underlying probability distribution of the data and impute the missing values. Regression trees are not usually used for imputation because of the potential instability of estimates from a regression tree model estimates and the lack of methods for estimating the standard errors for estimates. However, Borgoni and Berrington (2013) develop a tree based imputation method to impute missing multivariate Y and apply the method to the 1970 British Cohort Study.

More often, regression trees have been used to form nonresponse adjustment cells. A chi-square loss function with the Rao-Scott correction is often used for the regression tree construction algorithm (CHAID) in order to form response adjustment cells for survey nonresponse (Rao and Thomas 2003). There are many examples in the literature where the CHAID algorithm has been used for selecting the variables and forming nonresponse adjustment cells. For example Hogan et al. (2013) use CHAID to adjust for nonresponse in the Program for International Assessment of Adult Competencies Survey; Roth, Montaquila and Chapman (2006) apply it to the National Household Education Survey; Rizzo, Kalton and Brick (1996) apply it to the Survey of Income and Program Participation; Wun et al. (2004) apply it to the Medical Expenditure Panel Survey; Van de Kerckhove, Krenzke, and Mohadjer, L. (2009) apply it to the 2003 Adult Literacy and Lifeskills Survey; and Göksel, Judkins and Mosher (1992) use an AID algorithm for producing response cells for the National Survey of Family Growth (NSFG).

An area for potential future innovation is to use different loss functions to build the regression tree. For example, Schouten and de Nooij (2005) propose a new loss function designed specifically to produce response adjustment cells that minimize nonresponse bias after adjustment. Crafting a loss function tailored to the specific application could lead

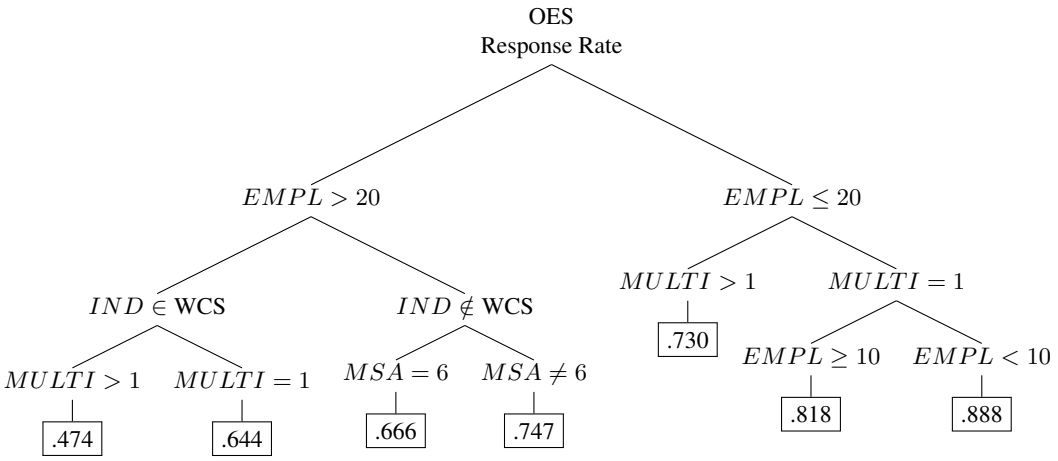

Figure 1: This displays the regression tree estimating an establishment's propensity to respond to the OES for a given set of characteristics. This model was estimated using the May 2006 OES data. The top value in the box is the estimated response rate for the May 2006 data used to build the regression tree model.

to more efficient tree growing procedure and give more meaningful results. For example, it has been suggested (Phipps and Toth 2012 among others) that the resulting tree nodes could be used in an adaptive data collection procedure. Instead of minimizing the MSE of the obtained data, a loss function designed to identify the cells with the highest bias after adjusting for nonresponse could be used. The obtained tree using this loss function could potentially better inform a nonrespondent followup sample.

Another application where a tailored loss function could be helpful is the use of a regression tree model for collapsing strata in an over stratified design. One could build a regression tree on the strata indicators using a loss function that minimizes the bias of a variance estimate for each node, then prune (collapse) the splits that reduce the bias the least. There seems to be a number of potential applications for regression trees with a customized loss function when analyzing and adjusting survey nonresponse that have not been explored.

4.2 Nonresponse Analysis

In this application the analyst is concerned, not with adjusting the data or removing potential bias from the data due to nonresponse, but rather with understanding the patterns of nonresponse. For example, they would like to identify characteristic groups that have higher than average nonresponse and/or higher potential risk of bias.

4.2.1 Easily Interpretable Nonparametric Model for Nonresponse

An example of the use of tree regression to identify groups based on establishment characteristics is the nonresponse analysis for the BLS Occupational Employment Statistics (OES) survey conducted by Phipps and Toth (2012). Figure 1 shows the results of a regression tree model of response propensities using May 2006 semi-annual panel data. Nine characteristics of business establishment sample members are included in the analysis. Four of the characteristics have a significant impact on the propensity to respond. The splits include: employment size, industry, multi (whether the establishment was part of a multi-establishment), and size of metropolitan area.

The results indicate that small, single unit establishments are most likely to respond (82% and 89%) in comparison to small multi-establishment units (73%); establishments

with larger employment have lower response rates, particularly in white-collar service industries (WCS; finance, information and professional and business services) that are multi-establishments (48%); and large establishments in other industries located in the largest metropolitan areas also have lower response rates (67%). These results are easily explained to and understood by survey programs and sponsors. In comparison, the OES model using logistic regression with both continuous and categorical variables and numerous interactions,

$$\begin{aligned} \text{logit}(p(x)) = \quad & \beta_0 + \beta_1 \log((EMPL) + \beta_2 IND + \beta_3 MSA + \beta_4 MULTI \\ & + \beta_5 \log(EMPL) * IND + \beta_6 \log(EMPL) * MSA \\ & + \beta_7 IND * MSA + \beta_8 IND * MULTI \\ & + \beta_9 \log(EMPL) * IND * MSA \end{aligned} \tag{7}$$

is much more difficult to interpret. The model used here came from a stepwise logistic regression procedure.

Deciding on a logistic model in this situation, with many continuous and categorical variables and interaction terms, is non-trivial. In an attempt to fit the logistic model using the OES survey, it was determined that the assumption of log-linearity for multi-establishment units indicates a lack of fit. Figure 2 shows that the assumption of linearity is plausible for single establishments $(MULTI = 1)$, but seems invalid for multi-establishment units $(MULTI > 1)$.

Similarly, Yu and colleagues (2007) find that SAT scores have a significant positive effect on response using logistic regression, but when students are partitioned by score distribution, the response pattern changes, with high and low scoring groups having higher response than the middle scoring group. This shows that lack of fit with logistic models pose a potential problem when specified vectors fail to include predictors that fully account for curvature or interactions. Regression trees have an advantage in that they automatically account for non-linear relationships.

Figures 3 – 5 shows tree model results for the BLS Job Openings and Labor Turnover Survey (JOLTS) for different phases of the data collection process, including locating an establishment or address refinement, requesting participation enrollment; and the data collection once a potential respondent has agreed to participate (Earp, Toth, Phipps, and Oslund 2013). These models use auxiliary variables similar to the OES models, allowing survey managers to compare nonresponse across surveys. The address refinement model (Figure 3) indicates that federal government establishments (12.1%) and large private establishments in the trade, transportation and utilities industry (14.8%) are the establishments most likely to be nonrespondents during this phase of collection as they are difficult to locate. In contrast at the enrollment phase (Figure 4), nonresponse is more likely among private- compared to public-sector establishments, especially those with a larger number of employees (14.8%). At the data collection phase (Figure 5), similar to OES, WCS establishments are the most likely to be nonrespondents, especially those with a large (> 180) number of employees (40.8%).

These results are of interest to survey programs and to BLS as an agency as they easily identify units that are more likely to be nonrespondents and warrant additional data collection. The San Francisco region of the BLS has used tree models to identify hard to collect units for survey managers as they decide how to allocate workload across field interviewers. The National Agricultural Statistics Services also has used tree models to analyze nonresponse and to identify hard to reach units prior to data collection (McCarthy and Jacob 2009).

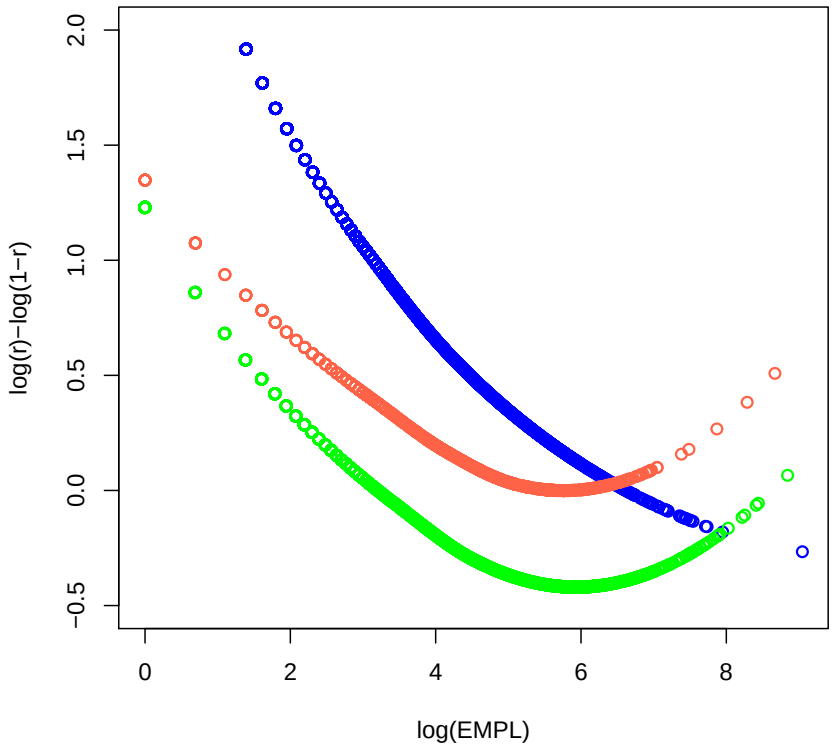

Figure 2: The logit function $\text{logit}\big(p(x)\big) = \log\big(p(x)\big) - \log\big(1-p(x)\big)$ for the smoothed response rates r by the log transformed establishment size. This is displayed for establishments in the professional and business services industry category located in an MSA with over a million people. The circles represent the log-odds ratio by log size for establishments with $MULTI = 1$, the triangles are establishments with $MULTI = 2$, and the diamonds are establishments with $MULTI \geq 3$. The response rate by transformed establishment size is estimated by a loess smoother.

4.2.2 Regression Trees in Linear Form

One particularly convenient form for a regression tree model is a linear function of the splits. Any resulting tree model (6) can be cast as a linear regression of the form

$$p(x) = \beta_0 + \beta_1 S_1(x) + \ldots + \beta_k S_k(x), \tag{8}$$

where S_i for $i = 1 \ldots k$ are indicator functions representing each split. That is, $S_1(\mathbf{x}) = 1$, if $\mathbf{x} \in S_1$, where S_1 is a subset of the population defined by splits on $\mathbf{X}$.

For example, the analysis of nonresponse for the OES survey resulted in the tree given in Figure 1, can be written as the series of indicator functions on splits and corresponding coefficients given in Table 1. Figure 6 gives the tree shown in Figure 1 with color splits matching the color of their corresponding rows in Table 1. Each row gives a split of the tree and its corresponding coefficient on the function indicator of whether the establishment has the defined characteristic or not. In this form, the coefficients

$$(\beta_0, \beta_1, \ldots \beta_k) \tag{9}$$

are interpreted as the association between a specific characteristic with a unit's propensity to respond.

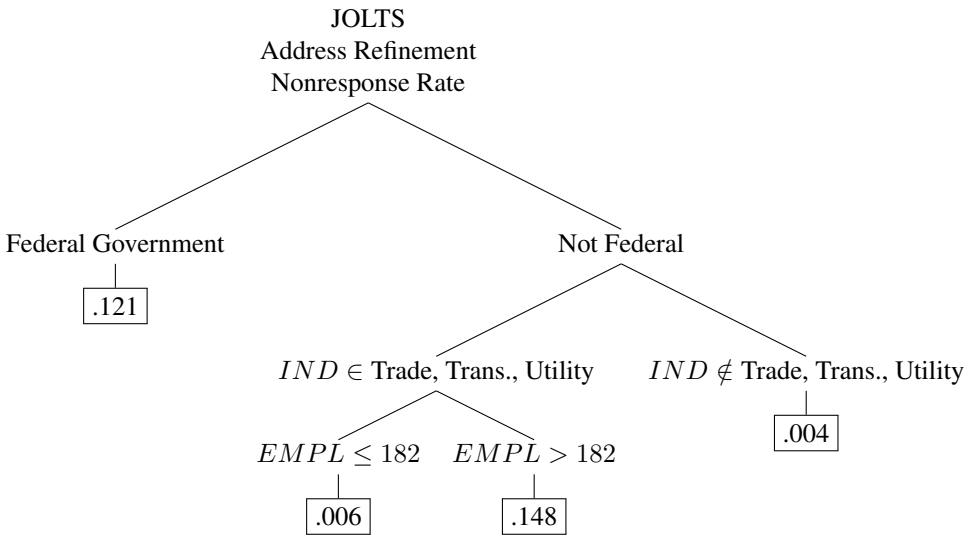

Figure 3: This displays the regression tree estimating the probability that an establishment will not respond during the address refinement phase of the JOLTS survey for a given set of characteristics. This model was estimated using the July 2012 JOLTS data, where overall 1.5% of establishments that made it to this phase did not respond. The value in the box is the estimated rates of nonresponse for the given group of establishments.

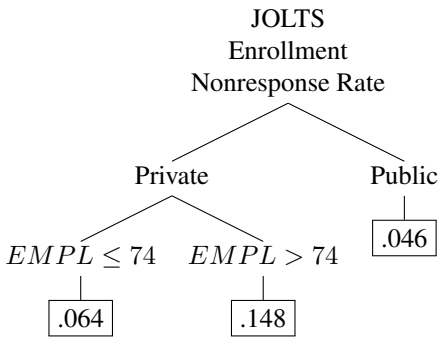

Figure 4: This displays the regression tree estimating the probability that an establishment will not respond during the enrollment phase of the JOLTS survey for a given set of characteristics. This model was estimated using the July 2012 JOLTS data, where overall 9.0% of establishments that made it to this phase did not respond. The value in the box is the estimated rates of nonresponse for the given group of establishments.

Split	May Response Coefficient
1	0.8883
$EMPL > 20$	-0.1411
$EMPL > 20$ & $IND \in$ WCS	-0.1036
$EMPL > 20$ & $IND \in$ WCS & $MULTI > 1$	-0.1691
$EMPL > 20$ & $IND \notin$ WCS & $MSA = 6$	-0.0810
$EMPL \leq 20$ & $MULTI > 1$	-0.1579
$EMPL \leq 20$ & $MULTI = 1$ & $EMPL \geq 10$	-0.0707

Table 1: The splits represented as indicator functions along with their coefficients in the linear model

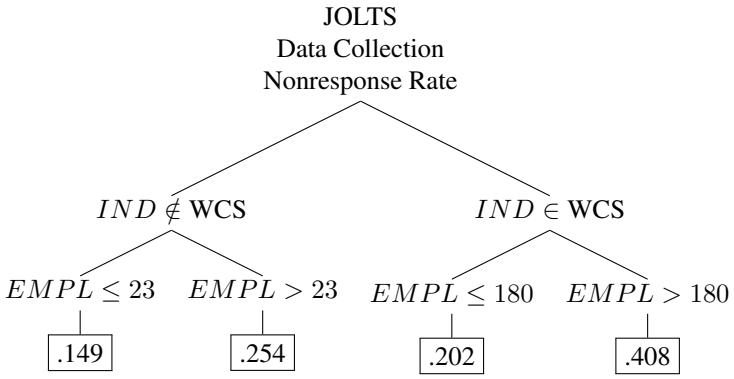

Figure 5: This displays the regression tree estimating the probability that an establishment will not respond during the collection phase of the JOLTS survey for a given set of characteristics. This model was estimated using the July 2012 JOLTS data, where overall 22.7% of establishments did not respond during this phase. The value in the box is the estimated rate of nonresponse for for the given group of establishments.

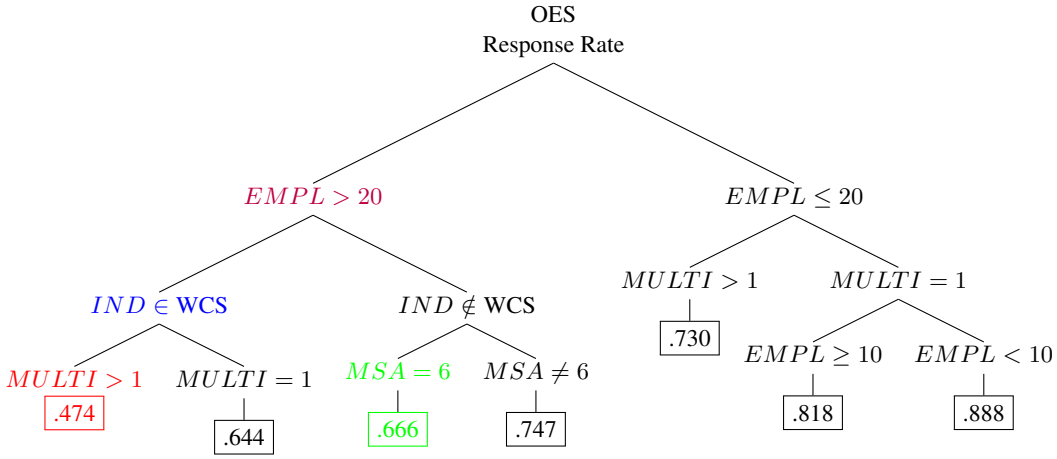

Figure 6: The regression tree used to model response propensity to the OES survey. The splits are colored to correspond to their corresponding coefficient given in Table 1.

Split	May Response Response Coefficient	May Wage Wage Coefficient
1	0.8883	8261
$EMPL > 20$	-0.1411	-970
$EMPL > 20$ & $IND \in$ WCS	-0.1036	4818
$EMPL > 20$ & $IND \in$ WCS & $MULTI > 1$	-0.1691	1298
$EMPL > 20$ & $IND \notin$ WCS & $MSA = 6$	-0.0810	1706
$EMPL \leq 20$ & $MULTI > 1$	-0.1579	3394
$EMPL \leq 20$ & $MULTI = 1$ & $EMPL \geq 10$	-0.0707	-559

Table 2: Results from the recursive partitioning of the OES mail survey data. Column 1 displays coefficients for the set of splits estimating response propensity for May 2006. Column 2 displays November 2006 response coefficients, based on the May tree model. Column 3 displays coefficients for the tree model to estimate May 2006 average wage per employee.

4.2.3 Bias Analysis

This parametric form of a regression tree has some potentially useful applications. For example, using an available proxy for the variable of interest, new model parameters (9) are estimated replacing the response indicator with this proxy. This results in a model of the proxy variable conditioned on the splits from the response propensity model. The potential bias of an estimator can be evaluated by looking at the resulting coefficients of this model.

This type of analysis was done by Phipps and Toth (2012) for the OES survey using available administrative record wage data for every establishment in the sample as proxy for occupational wage data collected by OES. Table 2 shows the estimated coefficients for response indicator and for the wage proxy variable. This shows which of the characteristic variables used in the model, that are known to be associated with nonresponse, are also associated with the variable of interest.

Another advantage of using a proxy, rather than the actual variable of interest, is that a comparison between respondents and nonrespondents on the value of the proxy variable can done for each of the partition groups. Since these groups were identified by the regression tree for nonresponse, we already know that they are associated with differing propensities of response. Figure 7 shows a graphical display of the difference in the reported wage variable between respondents and nonrespondents within the groups identified by the regression tree in the OES analysis. A difference within these groups suggests that weighting, using these response cells is likely to be inadequate for eliminating bias from the estimate (Horton, Toth and Phipps 2014).

4.3 Adaptive Data Collection

Difficulties in achieving and maintaining response rates are leading survey organizations to focus on potential respondents and how to direct additional data collection efforts most efficiently. One possible application of tree regression is to identify units to focus efforts on during data collection. The trees natural structure that partition the units and automatic selection of variables allows for the easy identification of units that are more likely to be nonrespondents and warrant additional collection effort.

For example, tree models using JOLTS, identify public- versus private-sector ownership characteristics as important to response at the early stages of collection, and industry type at later stages. If outcome variables, such as turnover, were identified as related to nonresponse, the JOLTS would be a possible candidate for an adaptive design procedure. They may monitor the rates of response in each of the identified cells and adjust efforts and resources to achieving higher response rates in cells as they are identified.

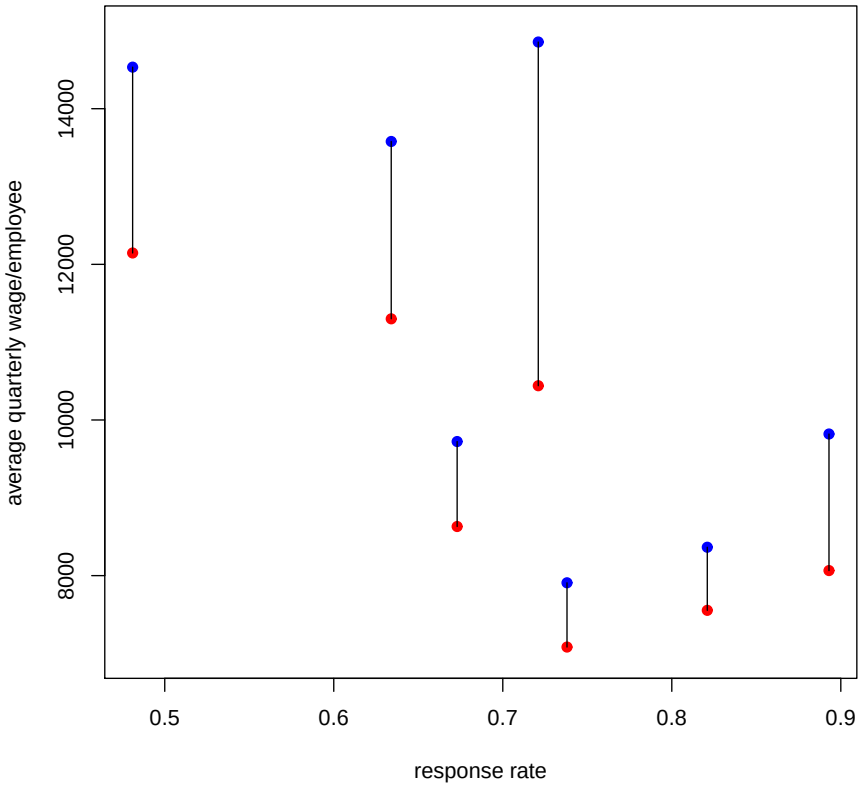

Figure 7: For the seven categories of mail survey-collected establishments defined by the regression tree model the average wage is plotted by response rate. The average wage per employee is given for responding establishments (red) and nonresponding establishments (blue). The line between the two estimates gives a visual representation of the difference between responding and nonresponding establishments within each category. All wage estimates are for the second quarter of 2005 and are produced from the QCEW records.

The OES tree models show that respondents and nonrespondents differ on wages, a major survey outcome. The resulting bias concerns make it a strong candidate for implementing an adaptive design procedure, such as focused contact or nonresponse follow up for groups with low response propensity and high wage differentials. Earp, Mitchell, McCarthy and Kreuter (2012) used tree modeling to evaluate nonresponse bias and direct collection efforts. Using auxiliary data, in this study the authors used an ensemble of classification trees to identify sample units that were likely to be nonrespondents then assigned the units to either a treatment or control group. The treatment group received additional refusal prevention and conversion efforts, such as personal enumeration visits, customized letters, data products, and small incentives.

5. Discussion

Regression trees are a powerful tool to explore survey response issues. Interactions between sample members characteristics often are important in nonresponse models, whether it is between employment size and industry in the case of business surveys or household income and age in household surveys. The automatic interaction detection inherent in trees provides a straightforward method to account for and easily interpret interactions between auxiliary data and paradata and the propensity to respond. Similarly, tree models can be used to identify potential bias and to prioritize cases for collection, or at the end of collection to adjust for nonresponse. While regression trees have not been used extensively to analyze response, direct data collection, or adjust for survey nonresponse, their application holds great potential.

REFERENCES

Brick, M. (2014), "Unit Nonresponse and Weighting Adjustiments a Critical Review," *Journal of Official Statistics,* 29, 329–353.

Brick, M. and Kalton, G. (1996), "Handling Missing Data in Survey Research," *Statistical Methods in Medical Research,* 5, 215–238.

Borgoni, R. and Berrington, A. (2013), "Evaluating a Sequential Tree-Based Procedure for Multivariate Imputation of Complex Missing Data Structures," *Quality & Quantity,* 47, 1991–2008.

Earp, M., Mitchell, M, McCarthy, J.S., and Kreuter, F. (2012), Adaptive Data Collection in Establishment Surveys: Using Proxy Data and Tree Modeling to Identify Likely Nonrespondents and Reduce Bias. *Proceedings of the Fourth International Conference on Establishment Surveys.*
Available at: www.amstat.org/meetings/ices/2012/papers/301898.pdf

Earp, M., Phipps, P., Toth, D., and Oslund, C. (2013), "Identifying and Comparing Characteristics of Nonrespondents throughout the Data Collection Process," *Proceedings of the Government Statistics Section* American Statistical Association, 3370–3379.
Available at: www.bls.gov/osmr/abstract/st/st130090.htm

Eltinge, J. and Yansaneh, I. (1997), "Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. Consumer Expenditure Survey," *Survey Methodology,* 23, 33–40.

Göksel, H., Judkins, D. and Mosher, W. (1992), "Nonresponse adjustment for a telephone follow-up to a national in-person survey," *Journal of Official Statistics,* 8, 417–431.

Hogan, J., Montalvan, P., Diaz-Hoffmann, L., Dohrmann, S., Krenzke, T., Lemay, M., Mohadjer, L., and Thornton, N. (2013), "Program for the International Assessment of Adult Competencies 2012," U.S. Main Study Technical Report (NCES 2014-047). U.S. Department of Education, Washington, DC, National Center for Education Statistics.
Available at http://nces.ed.gov/pubs2014/2014047.pdf

Horton N., Toth, D., and Phipps, P. (2014), "Adjusting Models of Ordered Multinomial Outcomes for Nonignorable Nonresponse in in the Occupational Employment Statistics Survey," *The Annals of Applied Statistics* 8, 956-973.
Available at: arxiv.org/ftp/arxiv/papers/1401/1401.0759.pdf

Kim, J.K. and Kim, J.J. (2007), "Nonresponse Weighting Adjustment Using Estimated Response Probability," *The Canadian Journal of Statistics*, 4, 501–514.

Kott, P. and Chang, T (2010), "Using Calibration Weighting to Adjust for Nonignorable Unit Nonresponse," *Journal of the American Statistical Association,* 105, 1265–1275.

Kreuter, F., Olson, K., Wagner, J., Yan, T., EzzatiRice, T. M., CasasCordero, C., Lemay, M., Peytchev, A., Groves, R. M. and Raghunathan, T. E. (2010), "Using proxy measures and other correlates of survey outcomes to adjust for nonresponse: examples from multiple surveys," *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 173, 389–407.

Little, R. (1982), "Models for Nonresponse in Sample Surveys," *Journal of the American Statistical Association,* 77, 237–250.

Loh, W.Y. (2014), "Fifty Years of Classification and Regression Trees," *International Statistical Review,* to appear.

McCarthy, J. and Jacob, T. (2009), "Who are You?: A Data Mining Approach to Predicting Survey Nonrespondents," *Proceedings of the Survey Research Methods Section* American Statistical Association, 5514-5528.

Mohl, C. and Laflamme, F. (2007), "Research and Responsive Design Options for Survey Data Collection at Statistics Canada," *Proceedings of the Survey Research Methods Section* American Statistical Association, 2962 – 2968.
Available at: https://www.amstat.org/sections/srms/proceedings/y2007/Files/JSM2007-000421.pdf

Peytcheva E., Groves R. M. (2009), "Using Variation in Response Rates of Demographic Subgroups as Evidence of Nonresponse Bias in Survey Estimates," *Journal of Official Statistics,* 25, 167–191.

Phipps, P. and Toth, D. (2012), "Analyzing Establishment Nonresponse Using an Interpretable Regression Tree Model with Linked Administrative Data," *The Annals of Applied Statistics,* 6, 772–794.
Available at: http://www.bls.gov/osmr/pdf/st120020.pdf

Rosenbaum, P. and Rubin, D. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika,* 70, 41–55.

Rao, J.N.K., and Thomas, D. (2003), "Analysis of Categorical Response Data from Complete Survey: An Appraisal and Update," in *Analysis of Survey Data*, eds. R. L. Chambers and C. J. Skinner, West Sussex, England: John Wiley and Sonspp, pp. 85–108.

Rizzo, L., Kalton, G. and Brick, M. (1996), "A Comparison of Some Weighting Adjustment Methods for Panel Nonresponse," *Survey Methodology,* 22, 43–53.

Roth, S., Montaquila, J. and Chapman, C. (2006), " Nonresponse Bias in the 2005 National Household Education Surveys Program," Technical Report (NCES 2007-016), U.S. Department of Education, National Center for Education Statistics, Washington, DC: U.S. Government Printing Office.
Available at http://nces.ed.gov/pubs2007/2007016.pdf

Schouten, B. (2007), "A Selection Strategy for Weighting Variables Under a Not-Missing-at-Random Assumption," *Journal of Official Statistics,* 23, 51–68.

Schouten, B., Cobben, F., and Bethlehem, J. (2009), "Indicators for the Representativeness of Survey Response," *Survey Methodology,* 35, 101–113.

Schouten, B. and de Nooij, G. (2005), "Nonresponse Adjustment Using Classification Trees," Discussion paper 05001, Statistics Netherlands,
Available at www.cbs.nl

Toth, D. and Eltinge, J. (2011), "Building consistent regression trees from complex sample data," *Journal of the American Statistical Association,* 106, 1626–1636.
Available at: http://www.bls.gov/osmr/pdf/st100010.pdf

Van de Kerckhove, W., Krenzke, T., and Mohadjer, L. (2009), "Adult Literacy and Lifeskills Survey (ALL) 2003: U.S. Nonresponse Bias Analysis (NCES 2009-063)," National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, DC.
Available at: www.edpubs.org

Vartivarian, S. and Little, R. (2002), "On the Formation of Weighting Adjustment Cells for Unit Nonresponse," *Proceedings of the Survey Research Methods Section* American Statistical Association, 3553–3558.

Wagner, J., West, B., Kirgis, N., Lepkowski, J., Axinn, W., and Ndiaye, S. (2012), "Use of Paradata in a Responsive Design Framework to Manage a Field Data Collection," *Journal of Official Statistics,* 28, 477–499.

Wun, L.M., Ezzati-Rice, T.M., Baskin, R., Greenblatt, J., Zodet, M., Potter, F., Diaz-Tena, N., and Tourzani, M. (2004), "Using Propensity Scores to Adjust Weights to Compensate for Dwelling Unit Level Nonresponse in the Medical Expenditure Panel Survey. Agency for Healthcare Research and Quality Working Paper No. 04004 (Oct).

Yu, C.H., Jannasch-Pennell, S., DiGangi, S., Kim, C., and Andrews, S. (2007), "A data visualization and data mining approach to response and non-response analysis in survey research," *Practical Assessment Research and Evaluation* 12.
Available at: http://pareonline.net/pdf/v12n19.pdf