EXPERIMENTS WITH MART, AN AUTOMATED MODEL BUILDING IN SURVEY RESEARCH: APPLICATIONS TO THE NATIONAL SURVEY OF PARENTS AND YOUTHS

Paul Zador, David Judkins, Barnali Das, Westat
Paul Zador, Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: MART, Ordinal Logit, Propensity Scoring, Nonresponse Adjustment, Imputation, Automated Model Building, National Survey of Parents and Youths

1. Background

The National Survey of Parents and Youth (NSPY) represents a major component in the evaluation of an on-going national media campaign designed to reduce illicit drug use among youth. In-person surveys covering items on substance abuse, parenting practices, and awareness of anti-drug media advertising are conducted with up to two youths and one adult per household. NSPY is organized into six-month long data collection rounds. Semi-annual reports are published after each wave (Hornik et al, 2000; Hornik et al, 2001). NSPY reports are on a very tight schedule, with just seven weeks for preparing analytic data sets. Allowing two weeks for cleaning data, only five weeks are left for repeatedly performing three types of modeling tasks: weighting, imputation and the preparation of counterfactual projections. Counterfactual projections are used in NSPY analyses to estimate the direct short term *causal* effect of the media Campaign, as described in Section 4. The evaluation uses counterfactual projections to assess confounder-adjusted Campaign impact estimates from exposure-outcome relationships.

In view of this tight schedule, we gave high priority to automating the preparation of analytic files to the maximum extent possible in a way that minimized human review and intervention, worked within the available budget, and remained data-sensitive. In the past, some automated systems were developed at the U. S. Census Bureau for the Current Population Survey. The software systems developed there in the 1960s and 1970s performed nonresponse adjustment and imputation automatically to produce unemployment reports in 15 days after the end of data collection. While these programs were fast, they were not very data sensitive relying on *a priori* models and the subsequent pooling of adjustment/imputation cells with inadequate sample sizes (U.S. Bureau of the Census, 1978, pages 57-58). Past proposals to adopt more data-sensitive approaches have been rejected because they were thought to excessively extend the weighting and imputation process. After a review of several approaches to automation, we opted for exploring and applying methods based on Multiple Additive Regression Trees, or MART for short – which is the experimental software developed by Friedman (1999) that he discussed in the prior talk.

MART was used to model categorical response variables in terms of predictor vectors with both numeric and categorical components, $\mathbf{x} = \{x_1, ..., x_n\}$. MART maps each predictor vector $\mathbf{x}$ into a vector $\mathbf{p} = \{p_1, ..., p_L\}$ of response probabilities. While MART's underlying theory is not simple (see Hastie, Tibshirani and Friedman, 2001), in applications MART functions similarly to ordinal logistic regression (McCullagh, 1980) or classification tree methods such as CHAID patterned after the original SEARCH program by Morgan and Sonquist (1963). There were several features that made MART an attractive option:

- Numeric and categorical predictors are allowed, both with missing values;
- Estimates are invariant to monotone transformations of predictor variables;
- Non-additive relationships are captured automatically;
- Complex interactions among predictors are captured automatically; and
- Runs relatively fast.

We present results for (1) nonresponse propensity models; (2) ad exposure models to be used to impute missing recall information for specific TV and radio ads that were randomly excluded from surveys; (3) exposure propensity models for generating counterfactual weight projections; and (4) an experiment to fine tune MART via cross-validation.

2. Nonresponse (NR) Adjustment

Standard procedure for NR adjustment is to form a partition of the dataset and calculate NR-adjusted weights by the formula:

$$w_{Ri} = w_{Bi}\varphi_i \sum_c \delta_{ci} \frac{\sum_j \delta_{cj} w_{Bj}}{\sum_j \delta_{cj}\varphi_j w_{Bj}} = w_{Bi}\varphi_i \sum_c \delta_{ci} \frac{1}{\lambda_c}, \quad (1)$$

where δ_{ci} indicates membership in cell c of the partition, φ_i indicates response status, w_{Bi} and w_{Ri} are pre- and post-NR adjustment weights, and λ_c is the weighted empirical response rate for cell c of the

partition. The partition constitutes a model for response status. Under the assumption that the propensity to respond is uniform within each cell, adjusted weights are asymptotically unbiased as the number of cases in each cell increases. Since given a correct partition, weights only increase variance, some practitioners (e.g., Little and Vartivarian, 2001) prefer to drop the prior weights from the adjustment. However, retaining weights results in robust estimation of model parameters (Holt, Smith and Winter, 1980) a finding that should apply to cell-specific response propensities as much as to other model parameters, and, most importantly, it assures consistency in the sense that the sum of the adjusted weights in each cell is equal to the sum of prior weights in each cell; i.e.,

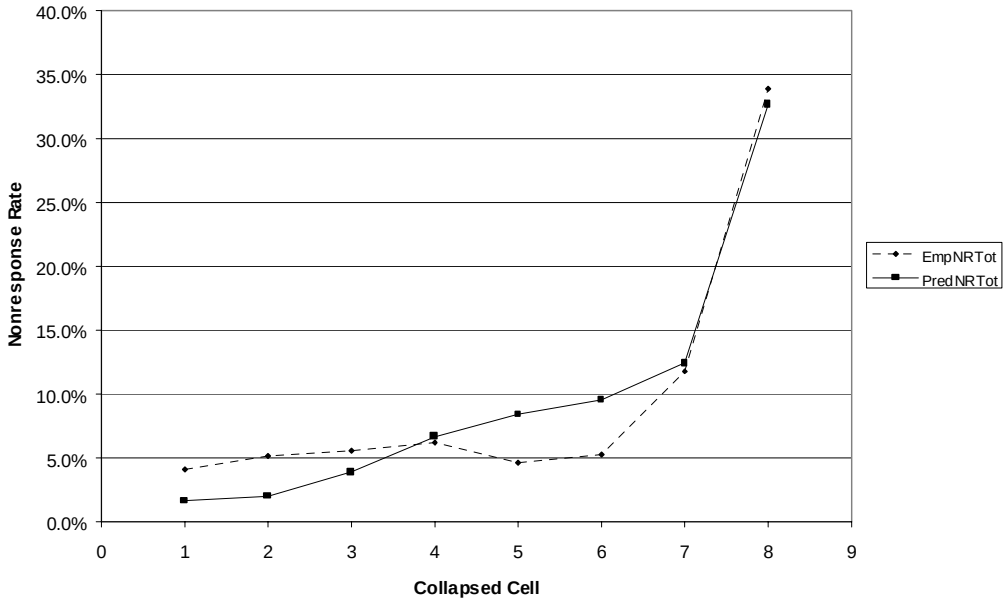
$$\sum_i w_{Ri}\delta_{ci} = \sum_i w_{Bi}\delta_{ci} \;\forall c\,. \qquad (2)$$

Consistency is important for estimates of finite population counts as well as being useful for quality control of the weighting process.

Standard procedures partition observations prior to or during data collection, and then collapse cells with thin samples. *A priori* collapsing rules can be applied automatically without any information from the sample or by an analyst applying general collapsing rules in terms of cell sample sizes and cell response rates. The first variant is fast but is data-blind, the second is slow but is data-sensitive. Both procedures tend to have *ad hoc* elements, though the first is a *priori ad hoc*, and the second is *post-hoc*. The present work continues the tradition of Göksel, Judkins, and Mosher (1992); Judkins and Lo (1993) to make NR adjustment data-sensitive yet analyst independent. These authors used CHAID and logistic regression for partitioning. CHAID tended to result in overfit, while logistic regression is insensitive to interactions and nonlinear terms and both procedures require some analyst intervention. We viewed MART as a solution that might overcome all the disadvantages of earlier methods.

We used a version of the well-known Pool-Adjacent-Violators algorithm (Barlow et al, 1972) to group MART-predicted probabilities into nonresponse adjustment cells that were in sort on predicted nonresponse probabilities. This step is similar to the use of logistic regression in Judkins and Lo (1993). Unfortunately, the empirical response rates turned out to exhibit unacceptably large variation across the cells. The cells with low predicted response rates actually had empirical response rates of λ_c = zero, resulting in infinite adjusted weights. We treated complete identification of nonrespondents as resulting from overfit and adopted an *ad hoc* procedure to deal with it. After considerable experimentation, we found that a partition developed by applying MART and then MUD – the name we gave to our partitioning algorithm – to a 50 percent training sample gave reasonable empirical response rates when it was applied to the full sample **(see Figure).** Although we have not tested the MART-MUD algorithm against CHAID or logistic regression, we are pleased with the ability of the algorithm to accept a large number of covariates and operate with little human intervention.

MART and MUD on 50% training sample, applied to full sample



3. Ad Imputation

Campaign exposure was estimated by total respondent recall based on all anti-drug ads aired by the media campaign just prior to interviews. To estimate recall for a single ad, respondents were shown the ad in full length on a laptop computer and were asked to say how often they had seen it in recent months. Because of time limitations, this procedure was performed only for random ad samples – for the rest, responses were imputed. We compared results from hot-deck imputation and imputation based on MART-predicted probabilities.

Hot-deck imputation was implemented by using proprietary software called WESDECK [Winglee, Ryaboy and Judkins, 1993]. MART-based imputation was performed by drawing samples from the unique response probability distributions estimated by MART. MART had two a priori advantages over WESDECK, it could use more predictors, and predictors were allowed to have missing values. (A missing value would have to be treated as a valid level by WESDECK, and this tends to yield poor results.)

MART and hot-deck performance were compared on the basis of seven ads targeted to youth with no missing responses (see below). As a simple form of cross-validation, the data were randomly split into four roughly equal parts by ad, and models were trained on three random quarters for imputing observations in the omitted fourth. Both procedures produced an imputed value for every observation. With the MART-based imputation, a single model was constructed for all ads pooled in order to 'borrow-strength' from the possible similarity of ad-exposure/ad response relationships across the ads with 0-1 indicator variables included among predictors to account for the possibly unique effect of individual ads.

Response frequencies were grouped into five levels (e.g., 0, 1, 3, 7.5 and 12.5) that reflected the number of times ads were seen per month. Graphical comparisons suggested that MART slightly outperformed WESDECK for most ads. Average squared misclassification errors were computed using the formula

$$\lambda = \sum_i \sum_j p_{ij} (O_i - P_j)^2 \qquad (3)$$

where O_i and P_j respectively represent the *i*-th actual and *j*-th imputed response values, and p_{ij} is the proportion of times when an actual response at level *i* is imputed at level *j*. The table below shows the results by ad.

In 5 of the 7 comparisons, MART outperformed the hot deck in terms of the squared misclassification error. For two of the ads, 5 and 18, this was not so. Subsequent investigations showed that for ads 5 and 18, an unusually large proportion of the responses fell in category 0 – meaning that, *in our sample,* few people ever saw those ads. This was not so for the rest of the ads. The hot deck imputed only actual values for any ad. This prevented hot deck from wandering very far from actual observed values. However, the way we used MART for imputing ads, this was not so. In our 'stacked analysis', we pooled data across all ads to strengthen the model, but apparently, we did not restrict the MART-predicted probabilities sufficiently to prevent the imputation of values that never actually occurred, although could have occurred, in a larger data set. Although, for a couple of anomalous advertisements, which did not conform to the general pattern, pooling data across the ads proved to be a sub-optimal way of using MART, in general, data pooling considerably improved performance.

Average squared imputation error in recall frequency by ad

Ad	Hot deck	MART
ad2	31.76	28.02
ad5	13.95	17.81
ad7	24.09	21.30
ad10	27.01	26.33
ad11	28.85	27.04
ad18	5.44	10.02
ad19	31.31	27.30

4. NSPY Contest Between MART and Ordinal Logistic Regression for Modeling Exposure Propensity

In NSPY, a three- or four-level ordinal variable for exposure is used to estimate exposure-related outcome-changes after appropriate adjustment for the potential confounding due to pre-campaign respondent characteristics. The overall objective is to measure short term direct exposure effects. We opted to control for bias due to potential confounding by propensity scoring of exposure. The method of propensity scoring has been developed by Rosenbaum and Rubin (1983), Joffe et al, (1999) and Imbens (2000), and for our application, it includes the following steps:

1. Estimate exposure propensity models in terms of covariates.
2. Develop counterfactual (CF) weights by expanding sampling weights to reflect exposure propensity differences among individuals with the same actual exposure.

3. Average outcomes in the CF world by exposure level using the CF weights.
4. Compare CF outcome averages across exposure level.

Two methods competed for estimating exposure propensities in Step 1: ordinal logistic regression (Joffe et al, 1999) and a new approach based on MART. The method requiring ordinal logistic regression (OLR) was implemented by Itzhak Yanovitsky and Elaine Zanutto from the University of Pennsylvania; the MART approach was implemented by the authors of this paper.

The test dataset comprised about 1,500 observations and 266 predictors. We used default MART parameters but set the loss function to penalize misclassifying a level 1 (3) response as a level 3 (1) response more than misclassifying it as a level 2 response.

As before, MART-predicted probabilities proved inconsistent. We developed a procedure that iterated between *consistency* and *renormalization* steps until convergence was established. The consistency step ensured the validity of (2) above, the renormalization assured that propensities summed to 1. Counterfactual projection weights based on MART were then calculated by adjusting weights inversely to propensities

$$w_i^{CF} = \sum_k \frac{\delta_{ik} w_i}{\hat{p}_{ik}} . \qquad (4)$$

With stepwise ordinal logistic regression, counterfactual weights were derived using a 5-level respondent partition that was obtained by grouping respondents after they were sorted on the final model's $X\hat{\beta}$ where X denotes the covariate vector, and $\hat{\beta}$ the parameter estimate. Each cell of the partition included about 20 percent of the total sample.

Based on statistical theory, good propensity models *balance* all available covariates across exposure levels (Rosenbaum and Rubin, 1983). Operationally, this is infeasible to assess with even a small number of covariates since all order interations and arbitrary transformations must be considered. However, we did assess balance in the limited sense of requiring the constancy of counterfactual confounder mean projections across propensity classes for a few binary and continuous predictors. The constancy requirement is intuitive: if a method is to eliminate bias from exposure-outcome relationships, it should leave little variation in the counterfactual means of confounders across exposure levels. We assessed constancy using two criteria, one involved a highly approximate chi-square significance test, the other the informal inspection of the relative range of the three counterfactual means computed as

$$RR = \frac{\max_k(Y_{Ck}) - \min_k(Y_{Ck})}{\sum y_i w_i} . \qquad (5)$$

In the final analysis, OLR was selected in favor of MART for exposure propensity modeling in NSPY. We note that MART-predicted propensities had to be made consistent in a separate step (see above). In terms of statistical performance, MART and OLR were judged to perform equally well when compared on balance, overfit, and the separation of propensities. There was also a tie on timing, in part because of the additional step required to achieve consistency with MART, and in part because in the test data set, OLR models worked adequately even without interactions. The decisive factor in favor of OLR was that it was assumed to give a higher level of comfort to social scientists in the target audience.

5. Fine Tuning of MART Parameters and Measuring Performance via Calibration Plots

Finally, we used a NSPY test data set with 2,996 observations to assess the calibration properties of MART models with different tuning parameters (see Table 1) for a 3-level response variable from 120 predictors. In the context of weather forecasting, Dawid (1982 and 1985) called a forecaster well-calibrated if, of those events to which the forecaster assigns a probability of say 30 percent, the long-run proportion that actually occurs also turns out to be near 30 percent. In the same spirit, one can say that a procedure for predicting class membership probabilities from covariate vector x in the form $p(c \mid x)$ is well-calibrated if a fraction of about p of events with predicted probability p actually occur (Venables and Ripley, 2000).

Table 1. Model parameters

Parameter	Number of levels	Level
Tree size	2	3, 5
Learning rate	3	0.01, 0.08, 0.25
Sampling fraction	3	0.5, 0.60, 0.75
Cost matrix	2	1: 0, 1, 1 2: 0, 1, 4 1, 0, 1 1, 0, 1 1, 1, 0 4, 1, 0

For each parameter combination, one model was fitted using the full data set. To cross-validate these estimates, the full data set was split at random into ten roughly equal parts, and models trained on each of the ten sets of nine random parts were used to 'cross-predict' estimates for the corresponding omitted random tenth. Full sample estimates and cross-predicted estimates were also derived for stepwise logistic regression.

Following Venables and Ripley, we examined the (smoothed) observed proportion of correct predictions as a function of the corresponding estimated probabilities. For well-calibrated predictions, observed proportions and estimated probabilities would be identical, and observed proportions, when plotted on the vertical axis against estimated proportions on the horizontal axis, would range along the 45 degree line. Departure from well-calibration was quantified by the mean squared difference between actual lines and the 45 degree line. When based on cross-predicted probabilities, actual lines depict calibration (Venables and Ripley, 2000) and the mean squared difference is calibration error. When based on predicted probabilities, the corresponding terms are pseudo-calibration and pseudo-calibration error. Similar concepts apply to models estimated from stepwise logistic regression. When calibration and pseudo-calibration lines are displayed together with 45 degree lines, for well-calibrated probabilities all lines overlap. If calibration curve soars above the 45 degree line, then the rank order for the predicted probabilities contains more information than is being captured in the predicted probabilities themselves. We refer to this unexpected phenomenon in the current context as overshrinking.

Main results are as follows. Stepwise logistic regression resulted in smaller pseudo-calibration error (0.033) than any of the MART models (>0.06). The calibration error of stepwise logistic regression was bracketed by the calibration errors of MART models (0.008-0.03). Depending on tuning parameters, MART models may overfit, overshrink, or be well-calibrated, or are essentially error-free. The stepwise model slightly overfits.

6. Summary and Discussion

This paper described current applications of MART to NSPY data for nonresponse adjustment and exposure imputation. In both applications, the software performed adequately after the development and implementation of special additional procedures. Thus, in neither application would MART be viewed as providing purely ‘off-the shelf’ solutions.

As a third potential application to NSPY data, MART-based models to be used for propensity modeling of exposure were compared to models derived using stepwise ordinal logistic regression. In terms of statistical performance, the two methods performed equally well. However, MART-based estimates for exposure propensity required special post-processing to assure that counterfactual projection weights derived from the propensities satisfied consistency requirements. In the absence of evidence for better statistical performance by MART-based propensity estimates, the more familiar statistical approach was selected for implementation.

In the course of developing and testing these applications, we noted that MART-predicted probabilities were typically not well-calibrated. A cross-validation experiment was then performed to examine calibration in relation to MART’s tuning parameters. The results confirmed, that for a wide range of tuning parameters, MART-predicted probabilities are indeed not well calibrated in the sense of Dawid.

It is not yet clear whether the proper choice of tuning parameters might not resolve the calibration problems we encountered in these applications. The survey data set that served as the basis for these analyses was medium sized when it comes to data mining software. There were about 1,500-3,000 observations but more than 100 predictor variables. Moreover, as it is often the case for social surveys, statistical relationships tend to be somewhat weak and strong high-level interaction are rare. Work on optimizing MART tuning parameters has so far been performed using data sets with observation/predictor ratios well above the 3000/100-1500/260 ratio range for the NSPY data (Hastie et al., 2001), and often strong interactions were suspected or introduced to document MART’s ability to identify them. Our experiments suggest that tuning parameters suggested on the basis of past experimentation may not be optimal for the type of survey data that were analyzed in this paper.

It would be of great practical help to survey data analysts to have ready access to automated model-building techniques. Our experiments with MART-based models indicate that although using MART can result in acceptable model quality, considerable additional work will be needed to develop guidance on tuning parameter settings appropriate for working with survey data.

7. References

Barlow, R.E., Bartholmew, D.J., Bremner, J.M., and Brunk, H.D. (1972). *Statistical Inference under Order Restrictions: The Theory and Application of Isotonic Regression*. New York: John Wiley & Sons.

Dawid, A.P. (1982). The Well-calibrated Bayesian (with Discussion). *Journal of the American Statistical Association*, **77**, 605-613. Reprinted in *Probability Concepts, Dialogue and Beliefs*, edited by O.F. Hamouda and J.C.R. Rowley. Edward Elgar Publishing, Ltd. (1997), 165-173.

Dawid, A.P. (1985). Calibration-based Empirical Probability (with Discussion). *Ann. Statist.*, **13**, 1251-1285. Reprinted in *Probability Concepts, Dialogue and Beliefs*, edited by O.F. Hamouda and J.C.R. Rowley. Edward Elgar Publishing, Ltd. (1997), 174-208.

Friedman, J.H. (1999). *Tutorial Getting Started with MART in SPLUS*. [online]. Available: http://www-stat.stanford.edu/~jhf/. [2001, October 18].

Göksel, H., Judkins, D.R., and Mosher, W.D. (1992). Nonresponse Adjustments for a Telephone Follow-Up to a National In-Person Survey. *Journal of Official Statistics*, **8**, 417-433.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer-Verlag.

Holt, D., Smith, T.M.F., and Winter, P.D. (1980). Regression Analysis of Data from Complex Surveys. *Journal of the Royal Statistical Society*, A, **143**, 474-487.

Hornik, R., Maklan, D., Cadell, D., Judkins, D., Sayeed, S., Zador, P., Southwell, B., Appleyard, J., Henessy, M., Morin, C., and Steele, D. (2000). *Evaluation of National Youth Anti-Drug Media Campaign: Campaign Exposure and Baseline Measurement of Correlates of Illicit Druig Use from November of 1999 through May of 2000*. Rockville, MD: Westat, Inc.

Hornik, R., Maklan, D., Judkins, D., Cadell D., Yanovitzky, I., Zador, P., Southwell, B., Mak, K., Das, B., Prado, P., Barmada, C., Jacobson, J., Morin, C., Steele, D., Baskin, R., and Zanutto, E. (2001). *Evaluation of National Youth Anti-Drug Media Campaign: Second Semi-Annual Report of Findings*. Rockville, MD: Westat, Inc.

Imbens, G.W. (2000). The Role of the Propensity Score in Estimating Does-Response Functions. *Biometrika*, **87**, 706-710.

Joffe, M.M. and Rosenbaum, P.R. (1999). Invited commentary: Propensity Scores. *American Journal of Epidemiology*, **150**, 327-333.

Judkins, D. and Lo, A. (1993). Components of Variance and Nonresponse Adjustment for the Medicare Current Beneficiary Survey. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 820-825.

Little, R.J. and Vartivarian, S. (2001). Don't Weight the Rates in Nonresponse Weights! *Proceedings of the Section on Survey Research Methods of the American Statistical Association*.

McCullagh, P. (1980). Regression Models for Ordinal Data. *Journal of the Royal Statistical Society*, B, **42**, 109-142.

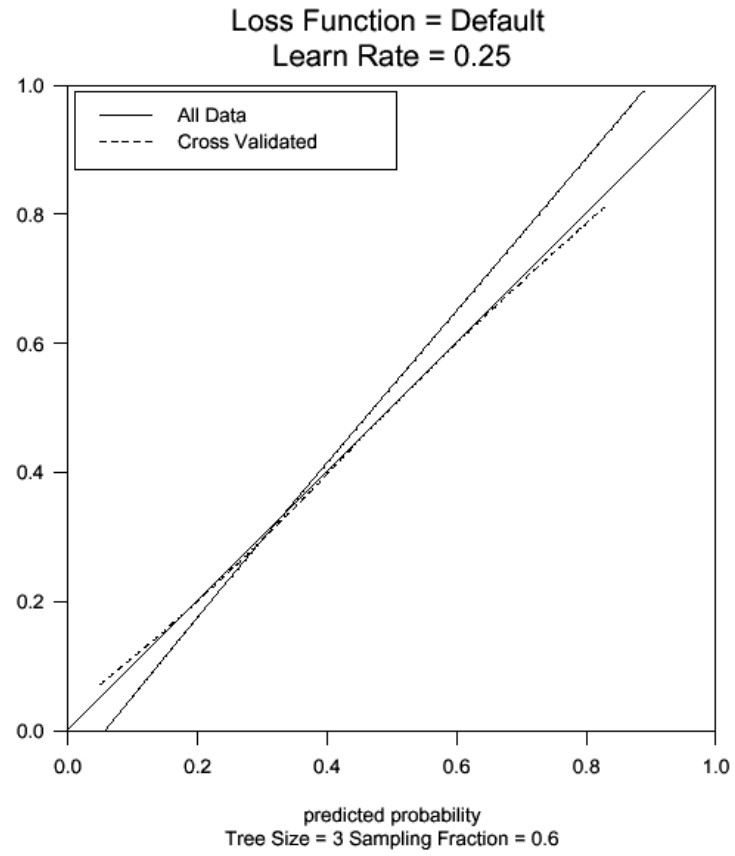
Morgan, J.N. and Sonquist, J.A. (1963). Problems in the Analysis of Survey Data and a Proposal. *Journal of the American Statistical Association*, 58, 415-435.

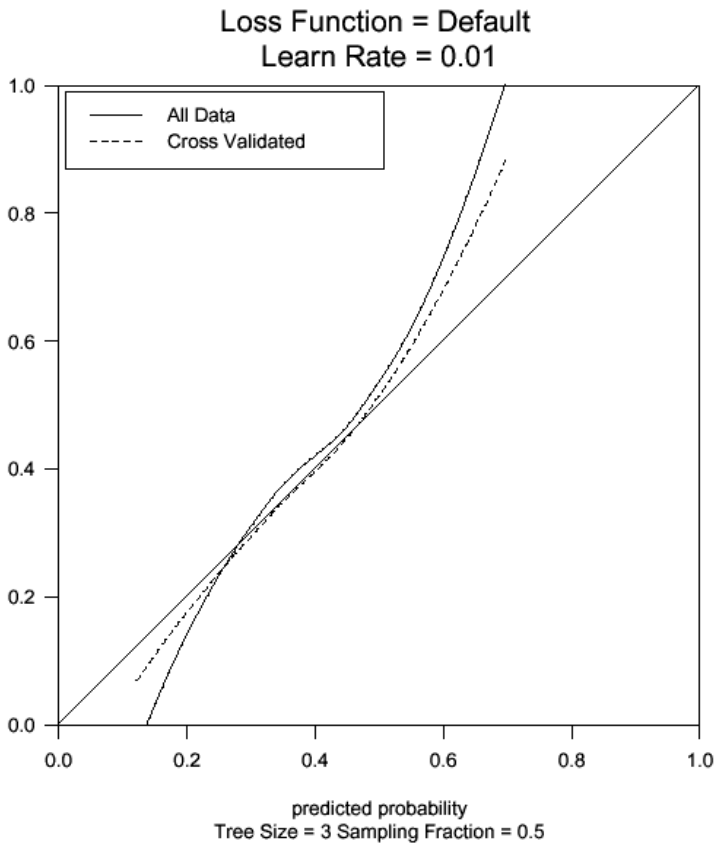
Rosenbaum, P.R. and Rubin, D.B. (1983). The Central Role of the Propensity Score in Observational Studies for Casual Effects. *Biometrika*, **77**, 41-55.

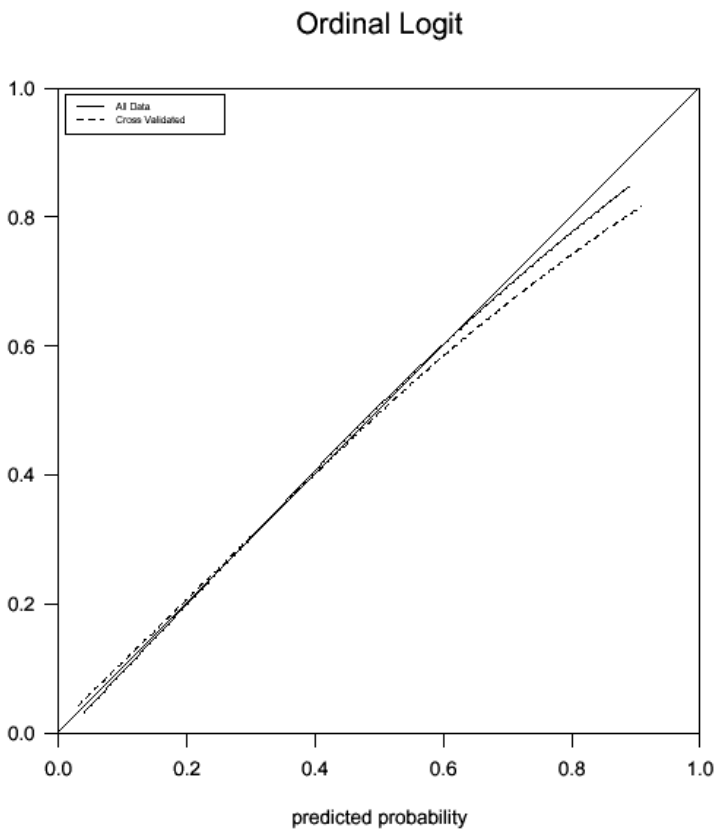
U.S. Bureau of the Census (1978). *The Current Population Survey Design and Methodology* (Technical Paper No. 40). Washington, D.C.: U. S. Government Printing Office.

Venables, W.N. and Ripley, B.D. (2000). *Modern Applied Statistics with S-PLUS*, Third Edition. New York: Springer-Verlag.

Winglee, M., Ryaboy, L., and Judkins, D. (1983). Imputation for the Income and Assets Module of the Medicare Current Beneficiary Survey. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 463-467.


Loss Function = Default
Learn Rate = 0.25
All Data
Cross Validated
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
predicted probability
Tree Size = 3 Sampling Fraction = 0.6


Loss Function = Default
Learn Rate = 0.01
All Data
Cross Validated
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
predicted probability
Tree Size = 3 Sampling Fraction = 0.5


Ordinal Logit
All Data
Cross Validated
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
predicted probability