UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001

July 10, 2015

2015 AMERICAN COMMUNITY SURVEY RESEARCH AND EVALUATION REPORT
MEMORANDUM SERIES #ACS15-RER-07

MEMORANDUM FOR ACS Research and Evaluation Advisory Group

From: Tommy Wright
Chief, Center for Statistical Research and Methodology

Prepared by: Eric V. Slud
Mathematical Statistics Area
Center for Statistical Research and Methodology

Subject: Impact of Mode-Based Imputation on ACS Estimates

Attached is the final American Community Survey Research and Evaluation report for Impact of Mode-Based Imputation on ACS Estimates. This was a research project to investigate whether there was any evidence that ACS estimates would be strongly affected by an imputation methodology that would take mode of response into account, for ACS outcome variables known to have allocation rates strongly depending on response mode.

If you have any questions about this report, please contact Eric V. Slud, (301) 763-4991.

Attachment

cc: ACS Research and Evaluation Work Group
Howard Hogan (DIR)
Thomas Louis (ADRM)
Ron Jarmin (ADRM)
Yves Thibaudeau (CSRM)
Todd Hughes (ACSO)
Tony Tersine (DSSD)
Beth Tyszka (DSSD)
David Raglin (ACSO)
Gail Denby (ACSO)
Dameka Reese (ACSO)
Agnes Kee (ACSO)
Nate Walters (SEHSD)
Michael Beaghen (DSSD)
Eric Vickstrom (SEHSD)

American Community Survey Research and Evaluation Program

July 10, 2015

Impact of Mode-Based Imputation on ACS Estimates

FINAL REPORT

Eric V. Slud
Center for Statistical
Research & Methodology

United States™ Census Bureau

Contents

Page intentionally left blank

EXECUTIVE SUMMARY

The goal of this research was to examine via data analysis whether American Community Survey estimates with missing survey items imputed by models including response mode as an explanatory variable are markedly different from analogous estimates based on a similar model ignoring mode. To make the results relevant to the question of whether American Community Survey hot-deck imputation methodology should be modified to account explicitly for response mode, the relation between item imputations done via fitted models as opposed to hot-deck was also explored.

Analysis focused on 17 American Community Survey outcome variables, known from previous research to show different frequencies across mode of being missing. All were binary, 8 at housing-unit level and 9 at person-level, and most were defined as indicators that multi-level variables fall in a specified interval of values. American Community Survey data from 2012 were analyzed to fit logistic-regression and regression-tree predictive models for these binary outcomes in terms of other American Community Survey variables not including mode, based on complete-case data with no missing predictors or outcomes. The models were used to decompose the case universe for each outcome into cells (ranging in number from 7 to 13). The cases with missing outcome were imputed in three ways: with 2012 hot-deck values, using the complete-case outcome=1 proportion for the cell containing the case requiring imputation, and using the complete-case outcome proportion for a system of cells cross-classifying the model-based ones by mode.

This research found no evidence that model-based imputation of binary outcomes would be improved by cross-classification of single-outcome imputation cells with mode of response. Data analyses were done to assess the quality of the models fitted; to compare the mode-pooled and mode-crossed cell outcome proportions for the complete cases from 2012 versus the cases with non-missing outcome and at least one missing predictor; and to compare the survey-weighted total estimates for the imputed case-outcomes across the three methods of imputation.

The conclusions of the research can be briefly summarized as follows:

(1) The cases with missing covariates or responses show dramatically different relationships between covariates and outcomes than the complete cases, for almost all outcomes.

(2) Because of (1), the Missing-at-Random assumptions implicitly governing almost all American Community Survey analyses to adjust for item nonresponse are in serious doubt.

(3) Logistic regression models for binary outcomes in terms of plausible predictors, with variables selected by criteria such as BIC, do not easily achieve statistical adequacy when mode is ignored. When mode terms and interactions are included, many such interactions are needed. Since the resulting cells formed by cross-classifying logistic-prediction scores by mode are already quite small, with sizes of hundreds rather than thousands in many cases, such models may be too noisy even for full-year data to be used in imputation applications.

(4) When item imputation is based on moderate numbers (7–13) of mode-pooled cells defined through logistic regression or regression-tree models, and compared with imputations done using mode-crossed cells, the differences in imputed totals behave very differently for housing-unit level than for Person-level outcomes. Housing-unit-level total differences are highly significant, although they amount to at most a few percent of the imputed cases, which in turn range from a few percent up to ten percent of all cases. Person-level outcome totals show virtually no difference between mode-pooled and mode-crossed imputations.

1 INTRODUCTION

Census Bureau research on the American Community Survey (ACS) [1], based on data from 2005, showed that the degree and patterns of item nonresponse are quite different for the different response modes: Mail, Computer-Aided Telephone Interview (CATI), Computer-Aided Personal-Interview (CAPI) with no mailable address, CAPI-subsampled after CATI, and now Internet. Yet current ACS practice is to impute item nonresponses, for households from which a form is accepted as a "response" to ACS, without restriction of imputation donors by mode.

The original ACS imputation process was apparently designed by Don Dalzell following procedures used for the 1990 long form ([2], p. 4 under sub-heading "Edit and Imputation"). So the pooling of imputations for mail returns and personal interviews in the 1990 long-form census was carried over to ACS, apparently from lack of resources or time to modify it [3]. Also according to [3], Chip Alexander and Lynn Weidman distinguished this in internal Census Bureau reports as one of the highest-priority problems with ACS methods, and it appeared on an extensive list of ACS problems prepared by Debbie Griffin as a precursor to developing an ACS research agenda.

The purpose of this project was to understand whether there is a practical need to modify ACS residential-unit hot-deck item allocation (or imputation) procedures to take account of response mode in defining the hot-deck donor pool. However, the ACS hot-deck specifications are sufficiently complicated that the ACS item imputation process would never change to depend explicitly on response mode unless some data analysis showed that this would have a serious impact on reported results, for at least some ACS tabulations of important attributes. This project was proposed as preliminary data-analytic research to investigate the likely effect on ACS data quality of possible changes in the ACS imputation methodology, which is `mode-pooled` in the sense of combining data without regard to response mode. The research reported here aims to inform decisions being made in the near future about imputation methodology as the Census Bureau moves to the new Control and Response Data System (CARDS) that will in the future perform final processing and analysis steps for Census Bureau surveys including the ACS.

Since it is not feasible in a limited project to investigate detailed, fully specified alternatives to the production methods for ACS hot-deck imputation, the effect of mode-based imputations will instead be assessed within the context of a simplified form of model-based or mass imputation. This entails model-building and assessment for key survey item-response variables, based on predictors that are themselves either frame-level housing variables, housing-level survey outcomes, or person-level survey outcomes. It also requires a model-based rule for imputing missing survey outcomes – based on models which either do or do not use response-Mode as a categorical predictor – and a comparison of the survey-weighted estimates derived by standard ACS weighting and estimation methodology from these imputations with those for the current hot-deck imputations. That comparison, with differences measured by the standard errors of the estimates generated from the current hot-deck imputations, will enable us to formulate tentative conclusions about the impact of mode-based versus mode-pooled imputations on ACS survey estimates.

The following Section provides the background to this research in four subsections: first, a review of the literature relevant to ACS allocation by mode, next a brief summary of literature on survey inference under hot-deck imputation, third a discussion of model-based imputation through conditional models for single items in terms of other survey outcomes and background variables, and finally, information on the data materials for this project, the full year's data on ACS residential Housing Units processed within 2012, and the survey outcomes to be investigated.

2 BACKGROUND

2.1 ACS Item Nonresponse and Imputation

Beyond the passage of [2] cited above concerning the similar imputation methodology to that of the 1990 long-form census used in ACS tests and later in ACS production, there seems to be little published literature containing details of the ACS hot-deck imputation methodology. There is very limited discussion of imputation allocation in the ACS Methodology document [4], with references only to Census Bureau subject-definition documentation. Various research and evaluation reports have been issued over the years concerning ACS nonresponse and allocation (item-imputation) rates [5, 6, 7, 8] and summaries of allocation rates by survey variable are now regularly published in online ACS documentation, currently at

`www.census.gov/acs/www/methodology/item_allocation_rates_data/`

The report [1], which distinguishes patterns of response by different modes for ACS variables, guided the choice of ACS outcome variables to study in this research.

2.2 Model-Based versus Hot-Deck Imputation

In many major censuses and surveys, missing data at the unit or item level are imputed. In United States Census Bureau censuses and surveys, this is most often done by hot-deck imputation, in which sampled units are ordered by some frame variables and a unit's missing items are assigned from a unit – usually the nearest unit in the ordered frame list – within the subset of "donors" defined as units sharing certain geographic, frame, and possibly demographic attributes with the unit to be imputed. (In particular, the hot-deck approach ensures that the donor is usually not far separated geographically from the unit with the missing data item, so that neighborhood effects automatically play a role in the assignment, a feature that is hard to reproduce in the covariates incorporated into models.) The imputed data fields are indicated through an imputation flag, but are for most purposes treated as though they had been observed directly. That is, the imputation flags play no role in most direct survey-weighted estimates, although they may play a role in survey variance estimates. While there is high-quality statistical methodological literature on estimation of variances in surveys with hot-deck imputation [9, 10, 11, 12], these methods are not implemented in ACS variance estimation, so that the item allocations are treated as having been made without error. This topic is beyond the scope of the present literature review and project, and here also the variances of estimators will be assessed using replication methods as in ACS, not taking account of the sampling variablility of the imputation model parameters used.

The current ACS hot-deck imputation methodology is documented as part of edit and imputation procedures, differently for each ACS survey variable. Hot-deck donor pools are specified through a series of conditional rules, in specification files that are occasionally updated, stored in ACS shared file directories with access restricted to the U.S. Census Bureau. A mode-based specification would have required defining and implementing in software new rules creating alternative donor pools based on response mode under all of the existing conditions for defining donor pools. This would have been a large job, requiring input from the subject-matter experts who designed the existing allocation rules for all variables studied, as well as new software coding and checking. Accordingly, such an approach was viewed as being beyond the scope of this research.

There is journal literature on model-based imputation, or "mass imputation" as distinguished from imputation at individual case level by methods such as hot-deck imputation. In principle, any model-based estimation (e.g., by cell-based cross-tabulation or some sort of regression modeling) might be used, and in that sense mass imputation is related to models used in small-area estimation [13]. Papers with particular reference to Census Bureau examples are [14, 15] in which loglinear models are proposed for model-assisted imputation. A recent treatment of inference from multiple-frame household surveys using model-based mass imputation is [16].

The basic building block for models of missing variables in terms of predictors that are never missing is a conditional model, often of generalized linear type, possibly with random effects (GLMMs). If the model could be fitted based only on complete-case data, and if one could assume that the same conditional model holds (a *Missing at Random* or MAR assumption) even when the outcome variable is missing, then the conditional model could be used either to simulate multiple possible values of the missing variable or to generate a single best predictor for the missing value according to some criterion.

The use of models for imputation of a missing survey variable is complicated by the strong possibility that for many sampled survey units, at least one desirable predictor variable may also be missing for the same unit. There are three distinct approaches, summarized well in [17]:

(i) monotone missingness means that the pattern of missing data is monotone in the sense that there is an ordering of variables $X_1, \ldots, X_K$ in which X_j is missing only when all of $\{X_i\}_{i=j+1}^K$ are also missing;

(ii) joint multivariate modeling means that all of the relevant outcome variables and predictors that might be missing are treated as joint outcomes for which a joint probability model can be specified; or

(iii) fully conditional specification is the strategy of modeling each relevant outcome variable and predictor that might be missing in terms of all other variables.

Approach (i), where applicable, is a direct extension of the building-block idea allowing missing items to be singly or multiply imputed from a conditional model in terms of non-missing predictors. In the case of monotone missing variables $\{X_j\}_{i=1}^K$, assume there is a block of never-missing covariates Z, and that for each $j \geq 1$ there are sufficiently many cases with $(Z, \{X_i\}_{i=1}^j)$ non-missing to fit a conditional model for X_j given $(Z, \{X_i\}_{i=1}^{j-1})$. Then missing values X_1 would be imputed from the conditional model given Z; missing X_2 from the conditional model given (X_1, Z) where X_1 is already imputed from its conditional model in those cases where it was missing; and inductively, after imputing all missing values among $X_1, \ldots, X_{j-1}$, the value X_j is imputed from its conditional model given Z, $\{X_i\}_{i=1}^{j-1}$. This use of *cascaded models* had been developed in various papers mentioned by [17], leading up to [18].

If missingness is not monotone, but variables $\{X_i\}_{i=1}^K$ can be modeled jointly given a block of never-missing variables Z, then the joint multivariate modeling approach (ii) can be used, as developed particularly in [19].

Finally, in many applications one can fit – for example, from complete cases – a variety of conditional GLMM models for variables X_i given $(Z, \{X_j :\ 1 \leq j \leq K,\ j \neq i\})$. Assuming such a set of full conditional models, of all variables modeled in terms of all others, then under a comprehensive Missing-at-Random (MAR) assumption saying that all these conditional relationships are the same regardless of the missingness status of predictors or outcome variables, one can fill in missing values iteratively using the following idea. Start with some set of imputed values, such as those generated by a hot-deck imputation, for the missing (predictors and) outcome variables. Now successively use the full conditional predictors to impute missing X_i from the conditional model given $(Z, \{X_j :\ 1 \le j \le K,\ j \ne i\})$, first for $i = 1$; next for $i = 2$ with missing X_1 values already replaced by their first conditional-model imputed values; and so on as i ranges up to K, each time imputing X_i from the conditional model $Z, \{X_j :\ 1 \le j \le K,\ j \ne i\}$ with the current set of imputed values substituted for missing values of the predictors. Next continue this process again as i ranges from 1 up through K, always updating missing values of X_j values with their most recent conditional-model imputed values. Repeat the process until approximate convergence of the distribution of survey estimates based on the multiple imputations. Although this idea may have had precursors elsewhere, it seems to have appeared first in [18] under the rubric of successive multiple imputations and to have been cited thereafter as the 'method of chained equations'. See the reports [20, 21] for recent efforts at using conditional models and chained equations as in [18] for the purposes of imputation in Census Bureau surveys.

The idea behind the iterative chained equations in [18] is similar to the idea of repeated Monte Carlo simulations of posterior densities based on full conditional distributions within the Gibbs Sampler. [17] indicates, however, that there are known theoretical compatibility requirements for convergence in this process, equivalent to the existence of a joint distribution for all predictor and imputed variables. When 'incompatibility' occurs, convergence will happen only in the weaker sense that the cycle of successive conditional distributions stabilizes. However, [17] points out that the survey estimates resulting from imputations obtained after stopping the iterations may still be well-behaved.

A further requirement necessary to make all of the conditional steps of iterative chained equations work is a very stringent comprehensive MAR condition: that is, the conditional distributional relationships in models specified using complete data, must be assumed to hold exactly when either the predictors or outcome variables are missing. In discussing "fully conditional specification", [17] does allude to the necessity of assuming essentially that all of the MAR conditions for separate conditional models hold simultaneously, which (under full-support conditions) requires that the conditional joint density for all sometimes-missing predictors and outcomes given the never-missing covariates is the same for cases with some missing data as for the complete cases. If that were true, then the joint cross-tabulated behavior of the data with some predictors missing using imputed predictor values ought to be the same as for the complete-case data.

We will investigate in the present report the extent to which the modeled relationships among variables within ACS complete cases also persist when some predictors or outcome variables are missing. For some of the conditional models developed, we will see that this last condition is clearly not true, implying either that the current hot-deck imputations are seriously flawed or that the needed MAR-type assumptions are invalid. It is generally hard to distinguish which of these reasons best explains failures of some of the models on missing-data cases.

The discussion about model-based imputation methods of [17], which we have generally followed here, was primarily intended to support multiple imputations derived from Monte Carlo simulations drawn from conditional models. Since our objective here is to work with model-based single imputations, specific rules are needed for defining imputations as conditional predictions to optimize some criterion. Appendices A and C address the choice of predictions derived from one or more conditional predictive models under criteria of maximum correlation or minimum misclassification error on complete cases.

Although GLM's are the conditional models used in many implementations of chained-equation multiple imputation, as in [18], we consider here Classification and Regression Tree (`CART`) models based on *recursive partitioning* (Appendix B) as alternatives. Such models have previously been considered in multiple-imputation settings by [22] and papers referenced there. In this report, `CART` type models are investigated as a way to develop relevant interactions and to define moderate sets of population subgroups for the display of survey outcomes with and without cross-classification by `mode`.

2.3 Data Materials

The analysis undertaken in this project is based on 2012 ACS data, in the form of the 2012 edited swapped data-files for all cases processed toward ACS estimates in calendar 2012, for occupied Housing Units. The variables used in the analysis will include the principal demographic, geographic, housing-type and other control variables currently used in defining hot-deck imputation donor pools; approximately 23 survey outcome variables chosen with reference to [1]; and the allocation flags needed to distinguish in which cases the control and outcome variable values were allocated (hot-deck imputed) as opposed to being reported or assigned. The decision to restrict to 2012 data was made primarily because the project builds on knowledge about Mode item-response-rate differences studied in earlier ACS reports before the Internet mode came on-line, and it would not make sense to use data cutting across the period (April 2013) when recent CATI changes (as described and documented in [23]) first came into effect. In addition, predictive models of imputed outcome means will be explored that make use of explanatory variables derived from the Planning Data Base [24] (PDB), as block-group aggregates of 2010 Census variables. The variables proposed for use in this research are summarized in Table 1.

The final list of outcome variables analyzed below is shorter than the list of housing and person variables in Table 1. The overall list was informed by the research of [1] on patterns of item nonresponse across response modes, with the idea that these missing items must be imputed much more frequently for responses in some modes than others. As will be explained further below, for simplicity of prediction and interpretation of cross-classified outcome patterns, it seemed methodologically clearest to focus on binary outcome variables. Therefore, some purely quantitative variables such as income or yearly taxes were omitted. For others, such as property insurance or time-related variables (year last married, or year house was built), it was easier to construct thresholds from which binary variables could be defined and interpreted. The binary variables were chosen in more than one way, with very different probabilities of a positive outcome. Some of the multicategory discrete variables, like household type, seemed too complex to describe as outcomes, and simpler versions, such as a `spousal` household indicator, were used as predictors but were not chosen as outcomes for lack of clear predictors among other variables on the list. The reduced list involved some arbitrary choices, but made the project more manageable.

The list of binary outcome variables analyzed in this Report is as follows. Housing-Unit-level outcome variables include `FS` (= HU indicator of receiving Food Stamps), `OWN` (= HU-level indicator of being occupied by Owner rather than Renter), `MULTI` (= indicator of HU within a

Table 1: Variables proposed for use in the Mode-Based Imputation Study.

Variable Name	*Brief description*	*Allocation flag*
	HU-level control	
MODE	resp. mode (mail, CATI, CAPI)	*
BLD	type of building (4-val. recode)	FBLD
TEN	tenure (owner/renter status)	FTEN
VACS	vacancy status (binary recode)	FVACS
HHT	household/family type	FSEX, FREL
CCS	MSA status (central, MSA, other)	*
	HU outcome	
FS	yearly food stamp recipiency	FFS
INS	property insurance	FINS
MRG, MRGX	mortgage payment, status	FMRGX
TAX	yearly real estate taxes	FTAX
YBL	year (HU) built	FYBL
FINC	family income	FFINC
CENRACE	race (multiple, census coding)	FRAC
	Person control	
AGE	age	FAGE
HIS	Hispanic origin (binary)	FHIS
MAR	marital status	FMAR
REL	relationship to HU ref. person	FREL
SCHL	educational attainment	FSCHL
SEX	sex	FSEX
	Person outcome	
CIT	citizenship	FCIT
ESR	employment status recode	FESR
JWD	time of departure for work	FJWD
LANX	speak another language at home	FLANX
PA	cash public assistance income	FPA
TI	total income	FTI
HICOV	any health insur. coverage	FHICOV
MARHY	year last married	FMARHY
	Block-group PDB	
MAILRET	2010 census mail-return rate	*
perc.black.cen	percent Black in 2010 census	*
popdensity.cen	pop'n density from 2010 census	*

Multi-unit building), PIns (= yearly Property Insurance payment for property owner, thresholded at 960 or 1310 to make binary variables), MRGX (=indicator of property owner having a mortgage), YBL (=year HU built, with binary indicators for < 1960 or between 1950 and 1979). Person-level outcome variables modeled include MARHY (= year last married, with binary indicators for < 1970 or < 1990), NoHS (recode of educational-level variable SCHL to indicator of never having completed High School), Coll (recode to indicate attainment of 4-year College degree), PostGR (recode of SCHL to indicator of a post-college degree), BLACK (recode of RACE to indicator of Black alone), NILF (indicator for not in labor force, ESR=6), HICOV (indicator of any health insurance coverage), LANX (indicator of language other than English spoken in the home), and JWD (daily departure time for work, thresholded at 0620). The further HU-level variables used in conditional predictive models are: CCS (3-level factor, distinguishing MSA Central City, MSA other, and non-MSA), City indicator (of CCS=1), RACE of reference person (4 levels), HSP (Hispanic origin of reference person), BLD (3- or 4-level code for building type), SPOU (spousal HU, i.e., husband and wife living together), REG (4-level Census region), and DIV (9-level Census groupings of states refining REG). Additional person-level predictors include: AGE in years, RACE (often recoded to indicators WHITE for White alone and BLACK for Black Alone), and HIS (hispanic indicator). Some of the PDB variables were tried in the models but were found not to be useful predictors in the presence of individual HU and person-level predictors and were not retained in any of the final models.

3 RESEARCH QUESTIONS

This research was conducted to answer two broad questions related to the possible impact of taking response mode directly into account for missing items in ACS data. The first is methodological, concerning the adequacy of evaluating imputation methods in a model-based framework.

Q1. Can a model-based imputation rule (without mode as explanatory variable) imitate the current (2012) ACS hot-deck imputation for the important ACS outcome variables listed in Table 1 sufficiently well that the corresponding weighted ACS estimates are close ?

The rationale for studying the impact of mode-based imputation in a model-based framework was to learn whether response-mode would have a strong or even a noticeable predictive effect beyond the covariates used in defining the universe for each variable together with the (mostly housing-level or demographic) covariates used to define hot-deck donor pools. The idea was that a finding of important mode-based effects in a model-based setting would strongly suggest the need to include response-mode explicitly in the specifications for hot-deck donor pools in the future. Beyond that, it was thought useful to learn how closely a model-based imputation would resemble the hot-deck imputations currently produced by ACS. It was known that, at least for some ACS variables, the hot-deck imputations contain a strong geographic component which would be hard to reproduce in a covariate-based model.

The models would have to be assessed at a minimum by examining ACS estimates, nationally and on subdomains, and checking that the mass-imputation based estimates fall within confidence intervals with standard errors of the ACS estimates themselves estimated by ACS successive-difference replicate-weight methodology. The subdomains on which to study the behavior of ACS estimates in this way, should include both geographic (states and larger counties) and demographic (cross-classifications by age groups, race, ethnicity and sex) aggregates.

The second research question is the main one motivating this project.

Q2. Are the ACS estimates based on an imputation model, designed to satisfy the requirements of Q1 without mode as an explanatory variable, markedly different from those based on a similar model explicitly incorporating mode as an explanatory variable ?

While mass imputation (with or without mode pooling) based on cellwise conditional expectations or other model predictions would not directly account for the variability of ACS survey weighted estimates, it would provide two different sources of imputed data from which estimates could be produced. The differences between the mode-pooled and mode-based model estimates can be compared with the margins of error for the survey-weighted ACS estimates (national and subdomain) corresponding to the same outcome variable. Additional subsidiary research questions, not separately stated here because they are contingent on the forms of the best-fitting models that can be found for imputed items, involve the investigation of properties of the model related to their stability with respect to slight changes in the set of selected explanatory variables and interactions chosen. If the answers to the two stated research questions are positive, it would still be necessary to show that the chosen mass-imputation models are reasonable, both by comparison with the existing hot-deck imputation method and by model goodness-of-fit assessments, which might be done either by likelihood-based criteria or cross-validation.

4 METHODOLOGY AND LIMITATIONS

As mentioned above in Section 2.3, the analysis in this report is restricted to binary outcome variables. This restriction results in several kinds of simplification: in the classes of models to be considered; in the interpretation and display of the model results; in the algorithms embodying model predictions as imputations of feasible outcome values; and in the extension of a model omitting mode to a cross-classified predictive model and imputation algorithm including mode. In particular, as discussed in Section 4.2 and Appendix C, imputation for each HU- or person-level outcome variable is then based either on the predicted probability that the outcome is 1 or on thresholding that probability at a computed optimal threshold.

Briefly, the predictive conditional models considered for the binary outcome variables in terms of other variables are of two types: generalized linear models (GLMs) such as logistic regression, and regression trees based on recursive partitioning (CART type models as in [25, 26]), as will be discussed in detail in Section 4.1, where the models are compared in terms of general criteria of fit. Next, in Section 4.2, the predictive models are compared also with respect to criteria of faithfulness of binary classification (on complete classes), and the optimal thresholds for imputation are computed. Further discussion of the models relates to their generalizability to cases in which some of the predictors are missing and have hot-deck imputed values substituted or in which the outcomes themselves are missing and imputed, and this examination of the outcome variables is undertaken in Section 4.3, in the spirit of [6]. Finally, Section 4.4 describes the method used to compare ACS-style survey estimates produced with model-imputed outcomes – separately for models that do and do not take account of mode – versus estimates and standard errors using the current ACS hot-deck-imputed missing values and replicate-weight based standard errors.

4.1 Fitting and Selection of Complete-case Models

This section together with Sections 4.2 and 4.3 primarily address research question **Q1** from Section 3. The fitting and selection of predictive models is done for invidual outcome variables in terms of other variables. Recalling that the goal of the model-fitting is imputation for binary outcome variables, the most important task for the model is to define a set of cells – not too numerous, perhaps 7 to 13, and none very small – in terms of predictor variables other than `mode`, for which the within-cell proportions of 1 outcomes are as different as possible from one another. The outcome variables analyzed, and the variables other than `mode` selected for them in the best models fitted, are displayed in Table 2. The specified forms of the logistic regression models for all of the outcome variables listed in Table 2 are given in Appendix D.

Separate models could have been developed including `mode`. That was not done in this research project because many interactions between `mode` and existing predictors were suggested by early modeling efforts, and the main objective of the analysis was not to develop complicated models with interactions so much as to examine whether outcome rates within population decompositions cross-classified by mode did seriously depend on `mode`. Thus, this section restricts attention only to models defined without explicit reference to `mode`. The complete-case models are viewed as a way of first assessing how predictable each outcome is in terms of a natural set of predictors and their (mostly pairwise) interactions, and then defining a decomposition of the population into covariate-based cells, which are called *nodes* because they often correspond to terminal nodes in a CART-type model. `Outcome=1` rates in these cells are compared with the rates observed in further decompositions of the cell by `mode`.

The strategy used in this project to develop conditional predictive models was to fit models initially using all plausible demographic and related control variables from Table 1 (restricting to housing-level predictors for housing-level outcomes), first by a recursive-partitioning or CART-type tree model using `rpart`, next by logistic-regression models incorporating the variables and some of the interactions appearing in the tree-based models, but generally avoiding interactions of order higher than 2. Within each type of model, we removed variables or interactions that had only very small effects on predictions, or where these effects made a difference only in very small complete-case subgroups. The subgroups considered in this way included the terminal-node groups within the tree-based models, and small quantile-interval groups within logistic regressions. In those few cases where the `rpart` software did not produce any meaningful data splits based on an "Information" criterion – usually because the overall rate of values 1 for a specific outcome variable was small, 0.1 or less in the entire respondent population – the groups were entirely based on (at most 10–12) quantile intervals of fitted probabilities of outcome 1. Whether from tree-based models or logistic regressions, groups of 3000 or less were avoided where possible, both because of their unstable identification within the complete cases, and because the corresponding subgroups of subjects with missing predictors or missing outcomes would then be extremely small (often less than 1000) even before cross-classification by `mode`.

The model-fitting process was far from exhaustive, and was driven by the goal of finding population subdivisions into 7–13 cells each containing at least thousands of single-year ACS complete-class observations. More detailed and predictively successful models could undoubtedly be found, by incorporating higher-order interactions as well as local geographic terms down to state level and below, for example by restricting new model terms to those which improve model

Table 2: Binary outcome variables, universe and mean and number of model coefficients (`df`), and quality metrics for best model fitted. Misclassification-rate (`Miscl`) and Correlation (`Cor`) are the metrics of Appendix C comparing the continuous predictor and binary outcome.

				Model Metrics		
Outcome	Universe	Mean	df	AUC	Miscl	Cor
HU-level						
FS	all	.121	31	.754	.184	.362
OWN	all	.712	27	.820	.245	.634
MULTI	all	.194	25	.776	.176	.662
PIns > 960	OWN=1	.336	80	.645	.411	.284
PIns > 1310	OWN=1	.163	80	.660	.257	.244
MRGX	OWN=1	.604	72	.622	.446	.260
YBL < 1960	BLD$\neq 1$	.339	44	.665	.403	.317
YBL 1950–1979	all	.382	53	.567	.464	.129
Person level						
MARHY< 1970	Ever-married, AGE≥ 15	.198	38	.660	.137	.753
MARHY< 1990	same	.524	38	.804	.306	.622
NoHS	AGE> 15	.148	79	.570	.186	.509
PostGR	AGE> 21	.112	64	.634	.192	.178
BLACK	all	.092	78	.708	.153	.287
HICOV	all	.873	28	.638	.193	.351
LANX	AGE> 4	.167	20	.687	.189	.567
JWD < 620	ESR≤ 4 JWTR$\neq 11$	.212	91	.645	.312	.253
NILF	AGE> 15	.360	64	.584	.299	.594

Source: American Community Survey, 2012.

deviance by at least 15 per degree of freedom. Choosing such a high deviance per model parameter as entry criterion has the effect of limiting model dimension but possibly worsening cross-validated prediction. (This number 15 should be compared with the Bayesian Information Criterion (BIC) – the most severe deviance-penalty term in common use for model selection – equal to natural logarithm of sample size n. Since n could be taken either to be the number ofcomplete cases among approximately 2 million HU's or 5.2 million persons, the BIC would require deviance improvements per degree of freedom of no more than $\log(5.2e6) = 15.5$.) Such additional model terms are not easy to find, although the search for them would be a worthwhile subject of future research if model-based imputation methods are desired for practical implementation.

Table 3: Counts and Occurrence rates for HU level Outcomes within defined universe, in 2012 ACS Data. Case subgroups are: Comp = Complete-cases, MissPrd = cases with incomplete predictors but non-imputed outcomes, MissRsp = cases with hot-deck imputed outcomes. Outcomes with suffix a or b are the two choices used to define indicators, in same order as Table 2.

	COUNTS			RATES		
Outcome	Comp	MissPrd	MissRsp	Comp	MissPrd	MissRsp
FS	1,994,021	143,256	25,704	0.1207	0.2006	0.1149
OWN	2,001,600	137,806	23,575	0.7124	0.6306	0.6528
MULTI	2,001,600	137,544	23,837	0.1944	0.1959	0.5287
PINSa	1,138,943	61,678	328,627	0.3365	0.3165	0.2853
PINSb	1,138,943	61,678	328,627	0.1631	0.1549	0.1338
MRGX	1,416,378	84,571	28,299	0.6045	0.6018	0.7743
YBLa	1,688,274	113,706	225,726	0.3386	0.3517	0.3888
YBLb	1,798,781	121,889	242,311	0.3822	0.3882	0.4151

Source: American Community Survey, 2012.

4.1.1 Housing-level outcomes

The occupied housing-unit outcomes analyzed in this Report are listed in Tables 2 and 3. The variables included as predictors for these, listed in the logistic-regression specifications of Appendix D, are other housing variables together with some demographic variables for the HU reference person when those were found to be useful. Three metrics describing the fit of the models are listed in Table 2, and will be interpreted beginning in Section 5.1. The size of the Complete-case dataset used to fit each model is determined by the universe for the outcome intersected with the cases for which none of the model outcomes or predictors is missing. (Here 'missing' means that the corresponding allocation flag indicated a survey variable that was imputed for a given case.) Table 3 displays for each outcome variable the number of complete cases in this sense, as well as the number of cases for which the outcome-variable was not missing but at least one predictor variable was missing, and those for which the outcome-variable itself was missing. These three sets of cases, respectively denoted Comp, MissPrd, and MissRsp in Table 3, are populations that may be qualitatively different, and they will be compared in Section 5.2 with respect to relationships between predictor and outcome variables, after some preliminary discussion in Section 4.3 about the Rates columns of Table 3, which show the proportion of their cases for which the outcome-variable indicator is 1.

4.1.2 Person-level outcome variables

The person-level outcome variables analyzed in this Report are listed in Tables 2 and 4. The variables included as predictors for these, listed in the logistic-regression specifications of Appendix D, are housing-level variables (including HU and the variable SPOU denoting spousal family structure) together with person-level demographic variables including Age, Race, Hisp, AIAN, REL (relationship to HU reference-person), and indicators of educational attainment (NoHS, Coll, PostGR). Other possible predictor variables, such as personal or family income or poverty indi-

Table 4: Counts and Occurrence rates for Person level Outcomes within defined universe, in 2012 ACS Data. Case subgroups are: `Comp` = Complete-cases, `MissPrd` = cases with incomplete predictors but non-imputed outcomes, `MissRsp` = cases with hot-deck imputed outcomes. Outcomes with suffix `a` or `b` are the two choices used to define indicators, in same order as Table 2.

	COUNTS			RATES		
Outcome	Comp	MissPrd	MissRsp	Comp	MissPrd	MissRsp
MARHYa	2,794,403	68,137	301,395	0.198	0.207	0.182
MARHYb	2,794,403	68,137	301,395	0.524	0.513	0.552
NoHS	3,961,156	84,391	197,094	0.148	0.207	0.289
PostGR	3,629,500	76,088	165,227	0.112	0.089	0.060
BLACK	4,735,122	459,261	97,524	0.092	0.161	0.076
HICOV	4,346,677	317,176	628,054	0.874	0.827	0.933
LANX	4,458,073	343,733	187,781	0.167	0.265	0.177
JWDa	1,899,403	97,942	278,098	0.212	0.235	0.215
NILF	3,765,380	229,040	248,221	0.360	0.393	0.545

Source: American Community Survey, 2012.

cators, were thought to be less reliable in the sense that despite their predictive value they will more frequently be missing than the other predictors variables used. Three metrics describing the fit of the models are listed in Table 2, and will be interpreted beginning in Section 5.1 below. The counts of complete cases, cases with at least one missing predictor but non-missing outcome, and cases with missing outcome are determined as described in Section 4.1.1 above. These counts, along with the corresponding rates of the outcome indicator variable being 1 in these populations, are shown in Table 4. Also for the Person-level outcomes, these populations of `Comp, MissPrd`, and `MissRsp` cases may be qualitatively different, and they will be compared in Sections 4.3 and 5.2.

4.2 Model-based Imputation

The methods described so far, which take no explicit account of `mode` of response, consist of:

(i) the definition for each binary outcome variable of a complete-case population defined by the non-missing (non-allocated) status not only of the outcome variable but also of all predictors used in fitting a best logistic regression model;

(ii) a decomposition into cells of that complete-case population, defined in most cases by a CART model (for which, see Appendix B) on a subset of the predictor variables in the logistic regression, and in other cases (when CART software gives either no splits or too many terminal nodes) by quantile intervals for the logistic-regression model scores.

This Section describes how to use these models and cells to impute the binary outcome variables when they are missing, first by a method that continues to ignore response `mode`, and second by a method that explicitly takes `mode` into account. There are many choices to be made in defining the method of imputation, and a rationale is provided for the method adopted.

The primary methodological issue arising in imputations from a propensity-type model is how to convert an estimated probability into a set of case-level binary decisions. This question arises whether the model does or does not explicitly account for response `mode`. The main methods available are: (a) to assign an outcome to individual cases *at random* and independently with the respective probabilities given by the model, (b) to assign outcomes not to individuals but to aggregates such as strata, in proportions defined by the model (which essentially means to assign fractional propensities rather than binary outcomes to individuals), (c) using the model to provide individual probabilities of binary outcomes, to assign binary outcomes to individuals in such a way as to optimize some decision-theoretic criterion, like correlation or expected weighted squared difference, between individual probability and assigned outcome, and (d) to assign individual binary outcomes randomly according to the model, as in (a), but to do this multiple times independently to account for the uncertainty at the individual level. These methods are all discussed in general references on imputation and missing data, such [19] and the handbook containing [17], although perhaps (c) is covered more thoroughly in references on machine learning and classification, as is discussed in Section 4.2 above and Appendix C below.

In surveys where the numbers of individuals with outcomes needing imputation is large, the Law of Large Numbers makes the difference between methods (a) and (b) not very important. (Note, however, that the method in (b) differs from the *fractional imputation* idea advanced in [12], which is another method that could be added to the list (a)-(d).) On the other hand, some classification and prediction applications require that a single binary decision be made for each individual imputation, and perhaps the same constraint is important in census as opposed to survey applications. In such settings where single imputation is an individual-level prediction problem, methods (c) must be considered, and the methods first considered in this research were of type (c). The intention was first to apply `mode`-free models with propensities constant on large strata defined from covariates in `CART`-type models, and then simply to augment the strata to a full cross-classification by `mode` of the `mode`-free CART groups (called `nodes`) within the complete-cases. However, imputations of type (c) in many of the outcome variables would have the characteristic of imputing all or none of each CART node to have `Outcome=1`, and to have that property persist without drawing any distinctions between the `mode` subgroups of each node. Moreover, logistic-type models which provide distinctions between individual propensities (probabilities of `Outcome=1`) seemed to require extensive `mode`-by-covariate interactions to fit much better than `mode`-pooled models. Instead of careful re-fitting of logistic models for each outcome with appropriate interactions by mode, it was judged simpler to look directly to empirical outcome distributions on `node` by `mode` decompositions of the complete cases, using method (b) to assign propensities rather than $\{0, 1\}$ outcomes to individuals.

In mathematical notation, suppose that for a given $\{0, 1\}$-valued outcome k there are nodes $j = 1, \ldots, J$ with n_j complete cases in the outcome-k sample universe within node group j, of which n_{jm} cases are in response mode $m = 1, 2, 3$, and let r_j and $r_{j,m}$ respectively denote the numbers of outcome-k values 1 within the node-j group and within the node-j by mode m group. (The number $J = J(k)$ of nodes does depend on the specific outcome k.) Then for all cases with missing (i.e. allocated or hot-deck-imputed) outcome in ACS data, within the universe for that outcome, the `mode`-free imputed fraction for cases with covariate-defined `node` j is taken to be

r_j/n_j, and the corresponding imputed fraction for cases in node j and mode m to be r_{jm}/n_{jm}. The assignment of individual cases to node groups is based on survey variables observed or imputed for those cases, according to rules obtained in one of several ways: either the groups are obtained from CART-type regression trees (in some cases by using nodes that were not actually the terminal nodes for the CART models fitted), or by cross-classifying the predictor variables appearing prominently in the CART and logistic models fitted, or by creating quantile intervals from the fitted logistic-regression scores. The variables figuring in the node definitions are listed in Table 11 in Appendix E. As an indication of the types of cells used, the specific cell definitions in terms of survey variables are given in Appendix E for the HU-level outcome variables.

Research Question 2 asks whether the model-based imputation of outcomes are materially different when the model takes direct account of response mode. One way to assess this, strictly within the Complete-case population for a specific outcome (where recall that by Complete we mean that the outcome along with the relevant model-based predictors displayed in Appendix D for that outcome are not missing), is to ask whether the mode-by-(Outcome=1) decomposition of the population is approximately row-column independent within each node group. This is done by calculating for each outcome k the corresponding chi-square test statistic (with $2 \cdot J$ nominal degrees of freedom)

$$X_{RC}^{(k)} = \sum_{j=1}^{J} \sum_{m=1}^{3} \Big(r_{jm} - \frac{r_j}{n_j} n_{jm} \Big)^2 \cdot \frac{n_j}{n_{jm}} \Big\{ \frac{1}{r_j} + \frac{1}{n_j - r_j} \Big\} \tag{1}$$

Here and in later sections where chi-square type statistics are used as metrics of similarity of populations, if those statistics were to be used for formal hypothesis testing it would be proper to follow [28] in modifying the standard chi-square definition to incorporate survey weights. However, chi-squares are used in this report as descriptive tools rather than as test statistics, partly because the selection of the node groups was itself the result of model fitting on the survey data. So the Rao-Scott survey-weighted chi-squares are not used here.

4.3 Comparisons between Complete and Incomplete Cases

Whether imputation is done by the piecewise-constant propensity method just described, or by a more elaborate logistic model with and without mode as a predictive covariate, all of the methods make an assumption that the relation between outcome propensities is the same in the cases where outcome was missing (hot-deck-imputed) as in those cases where the outcome was observed. Following Rubin's missing-data terminology ([19]), this is a *Missing at Random* (MAR) assumption. Some such assumption is needed for models for missing outcomes conditional on covariates to be identifiable from complete-case data, yet this assumption may be incorrect.

The same kind of assumption is implicit in the *hot-deck* imputation methodology currently used by ACS, if we view the similarity of outcomes between 'neighbors' (in a suitably re-ordered frame list) to be a model assumption whose validity does not depend on whether an individual case has a particular outcome observed or missing. It is well known (*cf.* [6]) that the population of nearest-neighbors – the donors of substituted item values under hot-deck imputation – of individuals (or of HU's) with specific missing items is often far different from the general population of individuals (or respectively, of HU's). The uncheckable assumption, analogous to MAR, is that despite this difference, the nearest-neighbor donors are much more similar to the individuals with missing items than is the general population.

In the present research, each specific binary outcome can be compared in three populations with respect to the relation between their covariates (or membership in CART-defined `node` groups) and their (observed or imputed) outcomes. The three populations are: the Complete-cases, i.e., those individuals (or HU's, for HU-level outcome variables) for whom neither the outcome nor the relevant predictive variables are missing; the Missing-Predictor cases for whom the outcome of interest is observed but for whom at least one relevant predictive variable is missing; and the Missing-Response cases, for whom the outcome of interest is observed rather than imputed. In Tables 3 and 4, these three populations are respectively labeled `Comp`, `MissPrd`, and `MissRsp`, and their sizes in the 2012 ACS sample along the rates with which the binary outcome indicators are equal to 1. Considering the sizes of these populations, the outcome occurrence rates can be strikingly different, depending on the outcome-variable. For example, although the hot-deck-imputed rate of usage of Food Stamps is not much different than the rate seen among complete cases, the rate among HU's in which at least one of the key predictors `HSP, RACE, SEX` (of reference person), or `SPOU, OWN, BLD` is missing is much higher. (See Appendix D for the predictors of `FS`.) Compare also the `MRGX=1` outcome rates for the three populations: HU's with missing predictors and observed response have virtually the same rate as the Complete cases, but HU's with missing response are imputed by the hot deck at much higher rates. This is probably well justified, since the geographic grouping of HU types may make owners without mortgages more likely to be neighbors in the frame list and also more likely to provide `MRGX` information in ACS. `MULTI` is another variable where the contrast between Complete-case and Imputed rates is striking, but likely justified.

Since the differences between the three indicated populations are of great interest for what they suggest about the validity of MAR assumptions and hot-deck imputations, we measure their differences in two ways in Section 5.2 below. First, we use a chi-squared test statistic as metric for differences between the distribution across `node` groups of the model-pooled Missing-Predictor or Missing-Response populations as contrasted with the Complete-case population. We also use chi-squared metrics to measure the differences between the way these populations are distributed into `node`-by-`mode` cells. For both of these calculations, summarized in Table 7 below, the precise definitions are as follows. If populations A and B of total size n_A, n_B are decomposed into cells $c = 1, \ldots, C$, with numbers n_{cA}, n_{cB} respectively falling into the c'th cell, then the rates n_{cA}/n_A are used as estimated probabilities of cell-c membership in computing the expected number falling into cell c in population B, yielding the chi-square type statistic

$$d(A,B) \;=\; \sum_{c=1}^{C} \left(n_{cB} \,-\, n_B \, \frac{n_{cA}}{n_A} \right)^2 \Big/ \frac{n_{cA} \cdot n_B}{n_A} \tag{2}$$

regarded as a distance between the two populations. This distance does depend strongly on the number of cells and the size n_B of the second population, so these quantities accompany the metric values in Table 7.

4.4 Evaluation of Survey-weighted Imputed Estimates

The ultimate answer to Research Question 2 requires the comparison of survey-weighted total estimates of ACS outcome variables, between data in which imputations are made by comparable methods either with or without response `mode` as a predictor. The imputation methodology described so far uses models only to define subgroups of the population, called `node` groups, from survey-variable covariates without reference to `mode`, and then further decomposes these groups by `mode` in calculating mode-based imputations for comparison. Although ultimately model-based, this method has been designed to be as close as possible to design-based. The imputation method, described in Section 4.2 above, is synthetic in estimating probabilities of outcome-indicators equal to 1 according to the observed rates of occurrence of outcome 1 in the `node` or `node`-by-`mode` groups.

To compare the mode-pooled versus mode-based imputation methods, initially only the national aggregate survey-weighted totals are calculated. The standard used to assess differences for each outcome variable is the replicate-weight-based standard error calculated on the hot-deck-imputed ACS totals *on the domain of missing-outcome cases falling within the universe for that outcome.* One might instead have used the replicate-weight-based standard errors for the whole-population survey-weighted total. Those standard errors are much larger than the ones based only on the missing-outcome domain. But since the imputations do not affect the survey observations used for outcomes that did not require imputation, it seems more reasonable to assess the extent of imputation differences only on the domain of imputed cases for each outcome.

Beyond the national-level comparisons of imputed-outcome survey-weighted totals, it would make sense to look also for differences in subnational analyses. However, for imputations created according to cells as done here, the greatest differences will occur on the mode-based cells within `node` groups which differ most in outcome-rates from the full `node` group. So in the the present report, imputation methods are compared only based on national survey totals and row-column chi-squares (1) as described in Section 4.2.

4.5 Limitations

The use of the methodology just described to answer the two main Research Questions of this research project has several limitations.

(1) While the goal of the present research is to assess the possible impact of modifying ACS hot-deck imputation methods by restricting donor pools within response mode, the synthetic cell-based imputation scheme implemented here with and without cross-classification by `mode` is very different from hot-deck imputation and so cannot precisely justify conclusions about `mode`-based hot-deck modifications.

(2) The Missing-at-Random assumptions needed for the validity of either model-based or hot-deck imputation may not hold.

(3) Exhaustive model search has not been undertaken, although reasonably good predictive models do appear to have been found for the selected outcome variables studied. Since

predictive models that are sufficiently strong tend to diminish the importance of mode-based cross-classification in the definition of imputation cells, it is possible that a concerted effort to find stronger predictive models for HU-level outcome variables would remove most of the changes in estimates that this research finds may result from mode-based imputation.

(4) The outcome variables studied here have been chosen or constructed to be binary from those known to have mode-based variations in allocation fractions. However, one cannot easily extrapolate results to other outcome variables, or even to the fully quantitative form of survey variables like Earnings or Year Built.

(5) The method by which model-based imputations are done relies on imputation cells defined from covariates some of which can themselves be missing. This may affect the results, either in strengthening or (more likely) weakening the impact of cross-classifying model-based imputation cells by mode. To mitigate this effect of partially imputed covariates defining the cells, a more elaborate 'chained-equation' version ([17, 18]) of the imputation method could have been adopted. This could be done for completeness in further research, although it would be unlikely to change the overall conclusions.

5 RESULTS

5.1 Metrics and Best-fitting models

Logistic regression models were fitted to the HU- and Person-level outcome variables in the ACS 2012 data by the methods described in Sections 4.1, 4.1.1 and 4.1.2. The predictor-variable specifications of the fitted models are given in Appendix D, with the universes and coefficient vector dimensions listed in Table 2. The quality of these models can be assessed in several ways. First, with motivations and formulas given in Appendices A and C, the metric values Area Under ROC Curve (AUC), Misclassification percentage, and Correlation (between predicted probability of Outcome=1 and the Outcome value) were calculated and are displayed in the last three columns of Table 2.

Folklore about AUC's says that values in the range of .7 or .8 are good, and values below .6 are bad. But when the AUC values are compared across the different outcome variables in Table 2, they are seen roughly to be larger when $\min(\text{Mean}, 1 - \text{Mean})$ is smaller. Similarly, misclassification rates for binary outcome variables are roughly increasing with respect to $\min(\text{Mean}, 1 - \text{Mean})$. Correlations do not seem to show a direct relationship with Mean. Any of the three might be an approporiate measure of model quality in an imputation setting, depending on the loss function chosen. The quality of the models is also not directly related to the numbers of terms (df in Table 2) in the model: the interaction terms making up most of the terms in models with large numbers of coefficients actually have only a small incremental effect on predictive accuracy, even when they are highly significant as required by BIC.

Since the purpose of all fitted models in this report is imputation, the models can also be judged according to how well they enable the construction of cells defined from predictor variables with well separated Outcome=1 occurrence rates. Such cells are readily defined from logistic regression models by decomposing the population of cases according to quantile intervals for the

Table 5: Chi-square statistics for `mode`-by-Outcome Independence of Complete-case counts within `node` groups. `CompCt` denotes complete-case count, in thousands. `MARHY` is shortened to `MYR`.

HU items	FS	OWN	MULTI	PINSa	PINSb	MRGX	YBLa	YBLb	
Df's	14	14	26	24	24	18	18	18	
CompCt	1,994	2,002	2,002	1,139	1,139	1,416	1,688	1,799	
χ^2	12234	69016	8538	2120	688.2	6025	3891	2269	
Person items	MYRa	MYRb	NoHS	PostGR	BLACK	HICOV	LANX	JWDa	NILF
Df's	20	20	14	24	20	18	20	20	22
CompCt	2,794	2,794	3,961	3,630	4,735	4,347	4,458	1,899	3,766
χ^2	14.0	29.0	16.6	30.9	25.1	49.5	38.5	19.5	15.3

Source: American Community Survey, 2012.

model-predicted probability of `Outcome=1`. Alternatively, if it is only the cells that are desired, one may directly fit a CART or `rpart` ('recursive partitioning') type model (*cf.* Appendix B). This was done in as many cases as possible, subject to the `R rpart` package [27] converging and providing useful covariate-based splits without generating groups that were too small. Generally, splits resulting in groups of less than 1000 among the complete cases were disallowed.

The `Outcome=1` rates in the separate `node` groups are shown, for all outcome variables except `MULTI` and `PostGR` due to the larger number of `node` groups for those outcomes, in Table 12 of Appendix E. Note that the cells have not necessarily been chosen for overall greatest separation of rates, and that the construction of cells by quantiles of logistic-regression predictors is slightly more effective than `rpart` in achieving better overall separation.

All of the logistic regression models fit well enough to produce chi-square goodness of fit values small enough to be nonsignificant if the degrees of freedom were those of a fully specified model. Since the models did have many estimated coefficients, the effective degrees of freedom are smaller than than the number of node groups minus 1, and by this criterion many of the models do not show adequate fit. However, since the purpose of this research was primarily to examine the imputation differences between the `mode`-pooled and `mode`-crossed cell decompositions, and since fit of models including `mode` would have required many interaction terms between `mode` and other predictor variables, the lack of fit was not pursued further.

Finally, in response to the major question of Research Question 2, the Row-Column Chi-squares are used to examine the hypothesis that the outcome rates are constant over modes within `node` groups, via the chi-square descriptive statistics (1), in Table 5.

The dramatic and immediate result of the chi-square statistics (1) in Table 5 is the order-of-magnitudes difference between the values for the HU-level versus the Person-level outcomes. The Person-level outcomes yield chi-square values that would be either non-significant or only slightly significant, although as mentioned in Section 4.2, the proper way to implement a formal test would have been the Rao-Scott survey-weighted chi-square of [28]. However, despite being applied to a smaller population (HU's as opposed to persons), the chi-square values on HU-level outcomes are all (except for `PINSb`) in the thousands, typically two full orders of magnitude larger than on Person outcomes. Examination of the tables of `Outcome=1` rates in `node` groups

Table 6: Cross-tabulated rates of occurrence of FS=1 and of HICOV=1 on node-by-mode subgroups of the respective complete-case populations of HU's and ACS persons.

FS *counts and rates*

Node	Counts			FS=1 rates		
	Mode 1	Mode 2	Mode 3	Mode 1	Mode 2	Mode 3
1	1,060,791	180,889	180,109	.0527	.0815	.1080
2	295,285	33,624	149,037	.2048	.2946	.2685
3	7,095	1,705	6,121	.2751	.2985	.2815
4	24,851	3,400	24,399	.4290	.4641	.4223
5	4,162	1,466	4,923	.4671	.4816	.4532
6	566	210	1,032	.6078	.5762	.5262
7	5,841	1,839	6,676	.5511	.5269	.5052

HICOV *counts and rates*

Node	Counts			HICOV=1 rates		
	Mode 1	Mode 2	Mode 3	Mode 1	Mode 2	Mode 3
1	55,241	9,288	23,231	.5572	.5734	.5548
2	96,971	15,667	36,787	.6584	.6654	.6606
3	102,561	17,032	41,434	.7504	.7477	.7456
4	247,962	40,749	98,571	.7844	.7888	.7837
5	188,773	32,663	77,590	.8172	.8170	.8135
6	431,432	69,881	165,614	.8434	.8430	.8441
7	441,305	73,360	172,102	.8970	.8981	.8974
8	429,551	72,086	174,123	.9366	.9391	.9381
9	808,368	127,072	297,263	.9498	.9504	.9483

Source: American Community Survey, 2012.

cross-classified by mode, illustrated in Table 6 on complete cases for a single HU outcome (FS) and a single Person-level outcome (HICOV), shows that there are small but real differences of rates for different modes within node of the order of a few percent for the HU outcome, and a remarkable lack of difference for the Person-level outcome. This distinction between the effect of mode on imputation at HU and Person levels shows itself again in the survey-weighted outcomes based on missing-Response cases, in Table 8 in Section 5.3 below.

It was mentioned above in Section 3 that, at least for some ACS variables, the hot-deck imputations contain a strong geographic component difficult to reproduce in a covariate-based model. Indeed, attempts at fitting models with geographic (regional or state-level) intercepts did not produce greater concordance with hot-deck results than the best-fitting models described below in Appendix D. The reason for this is conjectured to be that the truly useful hot-deck geographic effects, for example in such housing variables as indicators of multiple-unit unit dwellings or or property insurance payments about a specified threshold, are local neighborhood effects.

5.2 Results on Subpopulations Defined by Missing Data

Section 4.3 described the idea for comparing the Complete-case and Missing-Predictor populations for each outcome variable, by examining the different proportions of those populations appearing in cells defined either by the covariate-based `node` groups given in Appendix E or by the cross-classification of these `node` groups by `mode`. The results of both kinds of comparisons are given by the chi-square statistics $d(A, B)$ of formula (2). The sizes n_B of the Missing-Predictor population are given in the second column of Table 7, for each outcome. For each outcome variable, the relevant population is the universe for that outcome, given within Table 2.

First, in examining the distribution of the populations into `node` group cells, defined by covariates without regard to `mode`, the proportions of the Complete Case population falling into the `node`-groups j are the ratios n_{jA}/n_A appearing in (2), and the chi-square value $d(A, B)$ is given in column 4 of the Table. All of these chi-square values are large compared to the number of `node`-groups. However, the relatively small values of $d(A, B)$, near 100, for `PINS, MRGX, YBL, MARHY`, indicate that the differences between Complete-case and Missing-predictor population `node`-group proportions for these outcomes are really quite small, and the differences for `PostGR, HICOV, JWD, NILF` and `OWN` are only slightly larger. By contrast, the differences for `FS, NoHS, BLACK` and `LANX` are much larger.

Next consider the comparison of the distribution of the Complete-case and Missing-Predictor populations into cross-classified `node`-by-`mode` groups. The Missing-Predictor population sizes n_B by outcome are as before, but now the `node`-by-`mode` proportions of the Complete-case population are the ratios n_{jA}/n_A. The interesting feature of these chi-squares $d(A, B)$ appearing in the last column of Table 7 as compared with those in column 4 is that the `Mode-crossed` final column is only slightly larger for the Person-level outcomes, but with the sole exception of `PINSa` the HU-level outcomes show a much larger difference between these chi-square columns. This observation is different from the finding in Table 5 that for Person-level outcomes, the `Outcome=1` rates are all essentially the same across `mode` within `node` for Complete cases. Table 7 does not at all reflect `Outcome=1` proportions, but only the proportions of Complete and Missing-Predictor cases falling in covariate-defined subgroups. The near-equality of columns 4 and 6 for Person-level outcomes in Table 7 says that the population distribution difference for Person outcomes is almost completely due to the `node` cells which ignore `mode`, while for HU outcomes (other than `PINSa`), there are additional population differences relating to `mode` within `node`.

5.3 Weighted survey estimates and contrasts

Finally, the survey-weighted ACS totals using three possible imputation rules are displayed and interpreted in Table 8, on the set of cases for each outcome that fall within the universe for that outcome variable and that were allocated (imputed) in 2012 ACS. The three imputation rules, described fully in Section 4.2, are respectively the fractional Complete-case rates of `Outcome=1` for each covariate-defined `node` group, the rates of `Outcome=1` for the `node` group by `mode` combination, and the Hot-Deck imputation as done in ACS 2012. Differences among the survey-weighted totals for the three sets of imputations Table 8 can be compared to the replicate-weight-based standard error estimated by the method used in the ACS.

Table 7: Chi-square Discrepancies between Complete-case and Missing-Predictor Populations

		# Cases	Mode-pooled		Mode-crossed	
	Outcome	MissPrd	Cells	Chi-square	Cells	Chi-square
HU	FS	143,256	7	5707.	21	8665.
	OWN	137,806	7	232.3	21	502.7
	MULTI	137,544	13	423.3	39	631.3
	PINSa	61,678	11	92.9	33	110.8
	PINSb	61,678	11	58.6	33	100.2
	MRGX	84,571	9	92.5	27	189.8
	YBLa	113,706	9	102.0	27	372.4
	YBLb	121,889	9	76.7	27	192.7
Person	MARHYa	68,137	10	124.9	30	138.2
	MARHYb	68,137	10	112.8	30	118.3
	NoHS	84,391	7	3522.	21	3530.
	PostGR	76,088	12	312.3	36	334.1
	BLACK	459,261	10	12150.	30	12205.
	HICOV	317,176	9	235.2	27	238.7
	LANX	343,733	10	6057.	30	6084.
	JWDa	97,942	10	379.1	30	397.6
	NILF	229,040	11	185.8	33	198.7

Source: American Community Survey, 2012.

The glaring results in Table 8 are that the totals of the Hot-Deck imputed outcomes are dramatically different from the totals of the other sets of imputed outcome variables, and also that the totals of the Mode-pooled or Mode-crossed imputed Person-level outcomes are virtually identical. On the other hand, the Mode-pooled and Mode-crossed cell-based total imputations in the HU-level outcomes are clearly different by up to about 1%, an amount which is small relatively but amounts to tens to hundreds of thousands of individuals, and is highly statistically significant. Even this 1% is only relative to the imputed cases, themselves only a small fraction of the numbers of total cases (1 to 2 million) in the 2012 universe for the HU-level outcomes.

The sharp discrepancies between hot-deck and the other imputation methods are consistent with the findings of Section 5.2 that the Complete-case population for each outcome differs markedly from the Missing-Predictors and Missing-Response population. Table 7 shows that the distribution of Complete-cases by `node` group was far from the distribution in the Missing-Predictor case population, and similar chi-square metrics (not shown) could similarly be used to document the difference between Complete and Missing-Response cases. The mode-based and mode-pooled imputed totals were nearly identical for person outcomes, and showed small but highly significant differences in HU level outcomes, not always in the same direction.

The degree of difference between national weighted survey totals of Mode-pooled versus Mode-crossed imputed outcomes is very slight in relative terms, tiny for Person-level outcomes and small but significant for HU-level outcomes among the cases requiring imputation. However, for Table 8 and tables like the first half of Table 6 (shown for `FS`)) and similar tables for other HU-level outcomes, within subnational pockets of the population with larger incidence of missing HU outcomes, the discrepancies of up to a few percent among `Outcome=1` rates across `mode` within `node` groups could be large enough to be noticeable.

Table 8: Survey-weighted Outcome Totals in Missing-Item Cases within the Universe for each Outcome, Imputed by Cell-based method with `Mode` pooled or cross-classified, and by Hot Deck.

	Miss-Rsp					
	Outcome	# cases	Mode-pooled	Mode-crossed	Hot-Deck	SE
HU						
	FS	25,704	229,790	237,834	251,091	4394
	OWN	23,575	856,319	820,113	864,969	8220
	MULTI	23,837	746,929	721,596	1,084,545	12719
	PInsa	328,627	4,122,103	3,710,699	5,645,709	25442
	PInsb	328,627	2,038,684	1,795,918	2,702,354	15345
	MRGX	28,299	918,663	956,011	1,217,538	10834
	YBLa	225,726	5,973,867	6,358,998	5,890,227	20463
	YBLb	242,311	6,868,641	6,534,449	7,439,669	27302
Person						
	MARHYa	301,395	3,582,769	3,582,676	2,697,515	15224
	MARHYb	301,395	6,922,163	6,922,066	8,911,170	34219
	NoHS	197,094	1,950,023	1,950,051	3,065,463	22907
	PostGR	165,227	1,078,684	1,078,722	592,823	10713
	BLACK	97,524	348,576	348,561	376,680	9371
	HICOV	628,054	25,264,785	25,264,956	26,822,927	65767
	LANX	187,781	1,698,556	1,698,566	1,792,469	22963
	JWDa	278,098	1,796,942	1,797,023	3,376,229	20640
	NILF	248,221	4,074,087	4,074,101	6,332,555	30113

Source: American Community Survey, 2012.

6 CONCLUSIONS AND FUTURE RESEARCH

The answers to the Research Questions posed in Section 3 are largely but not completely negative. First, for all of the outcome variables studied in this research, the distribution within covariate-defined (`node`-group) cells of the Hot-Deck imputed ACS cases is quite different from the distribution of the Complete-case and Missing-Predictor (hot-deck-imputed) population. This effect, which provides a strongly negative answer to **Q1**, is much stonger than any of the effects found in this research concerning differences between `mode`-pooled versus `mode`-crossed cell-based imputation. Tables 7 and 8 together document these differences, and call into question any Missing at Random assumptions used here or that might be used in other model-based imputation, but also suggest that the joint hot-deck-imputed outcome distributions based on the hot-deck donor population might be very different in important ways from the (unobservable) joint outcome-variable distributions of the cases with missing outcomes requiring imputation. As a result, ACS estimates of national and subdomain totals of imputed single survey outcomes might be fairly accurate, while survey-based models of relationships between multiple ACS outcome variables – due to the prevalence of cases in which at least one of those outcomes was imputed – might be quite poor, in a way that few ACS users would anticipate.

Question **Q2** in this research project asked whether ignoring mode, as is done in current ACS imputation methodology, might be leading to ACS nonsampling estimation errors. At least with respect to the binary person-level outcome variables studied here, the answer seems to be No: as displayed in the bottom halves of Tables 5, 6 and 8, this research has found no evidence that the model-based imputation of binary person-level outcomes is noticeably affected by cross-classification of single-outcome imputation cells with mode of response. With respect to HU-level outcome variables, the situation is less clear. The top halves of Tables 5 and 6 indicate that the `node-by-mode` distribution within node-groups of `Outcome=1` cases by `mode` shows signficant dependence, and the survey-weighted totals of imputed housing-level outcomes are significantly different in the top half of Table 8 according to whether node-group model-based imputations are done by `mode` subgroup. Although mode-based imputation alters the housing-level imputed item totals significantly, with differences in survey-weighted imputed national totals of tens of thousands to hundreds of thousands, the differences are actually quite small, amounting to at most a few percent of cases imputed, and for each item the cases with missing data requiring imputation are themselves only a few percent of ACS cases.

These and other related conclusions of this research can be briefly summarized as follows:

(1) The populations defined by missing covariates and responses appear to be dramatically different in their relationship between covariates and outcomes from the complete-case population, for almost all outcomes. (The outcomes used in this comparison for missing outcomes are Hot-Deck-imputed ACS values in the 2012 data.)

(2) Because of (1), the Missing-at-Random assumptions implicitly governing almost all ACS analyses to adjust for item nonresponse are in serious doubt.

(3) Logistic regression models to fit binary outcomes in terms of plausible predictors, using variable-selection criteria such as `BIC`, do not easily achieve statistical adequacy, when `mode` is ignored. When `mode` terms and interactions are included, many such interactions are needed. Since the resulting cells formed by cross-classifying logistic-prediction scores by `mode` are already quite small (with sizes of hundreds rather than thousands in many cases, such models may be too noisy even for full-year data to be used in imputation applications.

(4) When moderate numbers (7–13) of `mode`-pooled cells defined through recursive-partitioning (`CART` or `rpart`) models or logistic regression are used in imputation, and compared with imputations done using `mode`-crossed cells, the differences in imputed totals behave very differently for HU-level than for Person-level outcomes. HU-level total differences are large enough to be highly significant, even though they amount to at most a few percent of the imputed cases. The Person-level outcome totals show virtually no difference between `mode`-pooled and `mode`-crossed imputations.

The very different joint behavior of multiple survey variables depending on the missing status of those variables may be partly due to the failure of the Missing at Random (MAR) assumption on which imputation is ultimately based. Without additional information on the true values of missing items, there is no way to check this MAR assumption. One possibkly fruitful avenue for research along this line, suggested to the author by Nathan Walters, may be to use Failed Edit Followup (FEFU) cases in older ACS datasets [8] as a 'truth deck'. That is, by comparing the item values recorded after resolution in FEFU interviews with the values that would have been imputed for those cases based on their pre-FEFU data, one may obtain helpful information relevant to MAR on the accuracy of hot-deck and other imputation methods in the ACS context.

More broadly, ACS imputation research might be dramatically advanced by creating research data-files on past years of ACS data containing for all non-imputed case variables what their imputed values would have been if they were treated as missing, along with the corresponding case data that actually was missing. This would be a computationally intensive endeavor, but would allow for the first time a detailed study of the extent to which donor-derived data from the hot deck differs from the data that ACS measures.

Further research into model-based imputation may be motivated by an undoubted lack of fit of the models arrived at in Appendix D, and also by the large differences between the imputations generated by the models and the ACS hot-deck method. It was mentioned at the end of Section 5.1 that some of the lack of fit and some of the lack of correspondence to hot-deck may both be due to not accounting for local neighborhood effects in the covariates used. For this reason, an interesting topic for future imputation research is to explore whether models including neighbor-averaged results can incorporate the best features of predictive models along with the geographic specificity of hot-deck.

It still seems likely that models including neighborhood-average values of key predictors would not alter the main finding that imputation via response-mode makes only a very small difference in overall ACS estimates. However, model-based imputations may still be considered in the future as a way to improve the joint behavior of multiple imputed characteristics for single cases (housing units or persons).

Acknowledgments

This research has benefited greatly from discussions with Ryan Janicki on modeling strategy, and Ryan did most of the `CART` and logistic regression modeling for `HICOV` and `LANX`. I am also grateful to Darcy Morris for conversations about `rpart` and model-based imputation via thresholding, and to Chandra Erdman, Maria Garcia, Yves Thibaudeau, Martin Klein and Joe Schafer for discussions and literature sources about chained-equation imputation. Finally, I thank Nathan Walters for the suggestion mentioned above to use older FEFU data as a 'truth deck'.

Appendix A. ROC curves and AUC

The Receiver Operating Characteristic or ROC Curve is a method of displaying the performance of a binary classification rule based on data through a statistic or quantitative measure exceeding a threshold. That is, if a quantity S is to be used as a decision rule in the form that $S \geq c$ corresponds to a positive decision to declare that $D = 1$, where conventionally D is the unknown indicator of a state of interest (such as *diseased* members of a population), then the ROC curve is the plot of points $(P(S \geq c|D = 0), (P(S \geq c|D = 1))$ as c varies. Here $P(S \geq c|D = 1)$, called the *Sensitivity* (of classification based on $S \geq c$) or *True-Positive (TP) rate*, is determined from knowledge or estimation of probabilities from a model connecting the random variable S with the binary indicator D for a member of a population. The quantity $P(S < c|D = 0)$ is called the *Specificity* and $1 - P(S < c|D = 0) = P(S \geq c|D = 0)$ the *False Positive (FP) rate* of the classification based on $S \geq c$.

A typical definition for the random score S is the value $P(D = 1 \,|\, X)$ from a probability model based on a vector of covariate observations X on a population member which may be in one of two conditions $D = 0$ or $D = 1$. Based on a training set of observed data $\{(X_i, D_i)\}_{i=1}^{n}$ on (randomly selected) population members, the conditional probability $P(D = 1|X)$ may be estimated as a function (perhaps of an assumed specified form) of X; then for another member of the population for whom X but not D is observed the quantity $S = P(D = 1|X)$ is used in the form $I_{[S \geq c]}$ as a predictor for $D = 1$ after choosing a threshold c. The strength of the predictor S is sometimes expressed through the AUC or "Area Under [the ROC] Curve", which is usually based on estimates of Sensitivity and Specificity (or TP and FP rates) for a set of values c. Conceptually, one may view the area AUC under the ROC curve defined by (FP,TP) points for all thresholds c, as a mathematically defined quantity (integral) summarizing the curve. However, the AUC also has the following interpretation as a probability [20] measuring the quality of threshold-based binary predictors created from the variable S. Suppose that two members of the population are selected at random, respectively one with values (X, S) from the subpopulation satisfying $D = 1$ and the other with values (X', S') from the subpopulation with $D = 0$. Then $AUC = P(S > S')$. This quantity is given by a double-integral expression (the first displayed equation in Appendix C) when the conditional densities of S given $D = 0$ and $D = 1$ exist, but is usually estimated from a few estimated points $(\mathrm{FP}(t_j), \mathrm{TP}(t_j))$ on the ROC curve $(0 \equiv t_0 < t_1 < \cdots < t_J < t_{J+1} = \infty)$, via a trapezoid-rule approximation

$$\widehat{AUC} \;=\; \mathop{}\limits_{j=0}^{J} (\mathrm{FP}(t_{j+1}) - \mathrm{FP}(t_j)) \,\cdot\, (\mathrm{TP}(t_{j+1}) + \mathrm{TP}(t_j))/2 \tag{3}$$

Values of AUC close to 1/2 indicate a very weak predictor S, since it can be seen that a uniformly distributed random variable S independent of D leads to $AUC = 1/2$. A perfect predictor, for which there is some value c so that $S \geq c$ if and only if $D = 1$, would have $AUC = 1$. Generally, values AUC in the range (0.7, 0.9) are considered 'good', although the meaning of 'good' predictions varies with the application. Many other numerical features of the AUC curve can also be used to summarize the quality of prediction. This is a topic actively studied in Machine Learning, cf. [30].

Metrics for prediction quality, based on estimates of misclassification probability or correlation between predictor and D, are discussed in Appendix C, along with the choice of optimal classifiers.

Appendix B. CART and Recursive Partitioning

There are many applications, with imputation a particularly salient one, where it is crucially important to determine a not-too-numerous set of subgroups or population cells, determined by population (predictive) covariates V_i, which have the property that a response variable of interest behaves as differently as possible in different cells and roughly homogeneously within cells. If the response variable Y_i of interest is binary, which is the case of greatest interest in this report, that means simply that the probability of the response being 1 is as different as possible across cells, in some sense. This might be accomplished through a conventional (e.g., generalized-linear) predictive model for Y_i in terms of V_i, in which case the response-probabilities $s_i = P(Y_i = 1 \mid V_i)$ would be well-separated into covariate-defined groups by ordering and grouping, for example defining groups $i \in G_j$ from the $j = 1, \ldots, J$ successive $[(j-1)/J,\, j/J)$ quantile intervals of the predicted responses s_i themselves. However, there is a class of so-called 'recursive partitioning' methods that directly split the data $(Y_i, V_i,\ i = 1, \ldots, n)$ into subgroups according to thresholds applied to k'th coordinate values of the predictor vectors V_i. These methods, initiated by [25], are explained in [36] and `R` package `rpart` implementing them is introduced in [26]. The methods allow either continuous or discrete response variables, and in the binary-response setting here can be viewed as a cell-based alternative to logistic regression, in which the cells are defined from covariates by successive intersection of the data subsets for which single coordinates of the covariates lie above or below specified thresholds. The CART or `rpart` methods are in this sense intrinsically hierarchical and differ in flavor from GLM in allowing some cells to be defined from multiway intersections of conditions based on different covariate entries, while other cells are defined from fewer covariates. Thus, CART or `rpart` models selectively incorporate simultaneous effects from many covariate entries that in a GLM model setting would require a high-order multiway interaction, even while allowing only very low-order interactions among other covariates. This is an attractive feature: GLM models with some variables figuring in high order interactions, but other variables not figuring in interactions at all, are rather hard to build and are generally not considered in systematic model selection strategies. (They may arise in pruned-back highly overparameterized models with high-order interactions, a selection strategy with flaws of its own.)

Appendix C. Prediction-Model Metrics

As in Appendix A, assume that data (X_i, S_i, D_i) on a sampled population, $i = 1, \ldots, n$, are been used to estimate probability models $P(S \,|\, D = d)$ and $P(D = d \,|\, X)$ for $d = 0, 1$, where X_i is a vector of covariate data and S_i is a scalar continuous-valued function of X_i, possibly depending on unknown (estimated) parameters θ. In this setting, we describe alternative measures of quality of S for predicting D, to contrast with the measure AUC defined in appendix A as

$$AUC = \int \int_{\{s > s'\}} p(s|D = 1)\, p(s'|D = 0)\, ds\, ds'$$

The measures described here are simpler: they are the *weighted minimum misclassification probability* for D based on S, and the *maximum correlation* between $P(D = 1|S)$ and a function of S. Each of these will be defined as a characteristic of the true joint probability distribution of (D, S), which may be initially unknown but may under some assumptions be consistently estimated from large-sample (S_i, D_i) data.

Our purpose in presenting prediction metrics is to explain how they lead to optimal choices for binary predictions $\eta(S) = 0, 1$ for D in terms of S. As mentioned in Appendix A, this can be done by selecting the threshold c associated with a single point (FP, TP) on the ROC curve and defining $\eta(S) = I_{[S \geq c]}$. There are many ways in which this has been done, including fixing the FP rate $P(S \geq c \,|\, D = 0)$, or fixing the TP rate $P(S \geq c \,|\, D = 1)$, or fixing the point of tangency to the ROC curve (usually assumed concave or made into an ROC Convex Hull curve, cf. [30, 31]) with a specified positive slope τ (which corresponds to minimizing the loss-function $FP + \tau \cdot (1 - TP)$), or finding the point on the curve which is closest to the point $(0, 1)$. All of these methods and more are compared in the paper [33].

An illustrative ROC curve, with points defined according to some of their specific features, is shown in Fig. 1 on page 36 under the hypothetical logistic-normal model (6) considered below.

If a binary-valued function $\eta(S)$ is used to predict D, then the misclassification probability is defined as $P(\eta(S) \neq D) \;=\; E|\eta(S) - D|$

$$= \int \quad P(D = 1|S = s)\,(1 - \eta(s)) \;+\; P(D = 0|S = s)\,\eta(s) \quad p(s)\, ds$$

which is obviously made smallest by the choice $\eta(s) = I_{[P(D=1|S) \geq 1/2]}$. A weighted version of this quality metric, in which the weight w multiplies probabilities for the false-positive misclassifications $\eta = 1, D = 0$, while probabilities of false-negative events $\eta = 0, D = 1$ are recorded with unit weight: $P(\eta(S) = 0, D = 1) + wP(\eta(S) = 1, D = 0)$

$$= \int \quad P(D = 1|S = s)\,(1 - \eta(s)) \;+\; w\, P(D = 0|S = s)\,\eta(s) \quad p(s)\, ds$$

In the weighted case, the metric is made smallest by the choice $\eta(S) = I_{[P(D=1|S)/(1-P(D=1|S)) \geq w]}$. In the special case where $w = E(D)/(1 - E(D))$, the optimal classification rule is $\eta(S) = I_{[P(D=1|S) \geq E(D)]}$.

If $S = P(D = 1|X)$, then it is easily checked by repeated conditioning that $P(D = 1|S) = S$. (This is the basic idea in the original Rubin and Rosenbaum propensity score paper [32].) Thus the minimum w-weighted misclassification probability is equal to $P(D=1, S < w/(1+w)) + P(D=0, S \geq w/(1+w))$ and would be estimated by

$$n^{-1} \overset{n}{\underset{i=1}{}} \quad D_i \, I_{[S \geq w/(1+w)]} \, + \, w \, (1 - D_i) \, I_{[S \leq w/(1+w)]}$$

The correlation between a binary-valued predictor $\eta(S)$ and D, also called the *Matthews Correlation Coefficient* in the machine-learning literature, is

$$E \quad \eta(S) \, (D - E(D)) \Big\} \Big/ \Big[E \quad \eta(S))(1 - E(\eta(S)) \quad E(D) \, (1 - E(D)) \Big]^{1/2} \tag{4}$$

The choice of function $\eta(s)$ maximizing this correlation, which has not to the author's knowledge been derived explicitly in published literature, can be accomplished in two steps. First suppose that $\eta(s)$ is chosen subject to a fixed value

$$r \; = \; E(\eta(S)) \; = \qquad \eta(s) \, p(s) \, ds$$

Assume for simplicity that that the conditional probability $P(D=1|S)$ is a continuous random variable, which is not at all a restrictive condition if S itself is continuously distributed. Then it is easy to see that, subject to the restriction $E(\eta(S)) = r$, the function $\eta(s)$ maximizing (4) is the one maximizing $E(\eta(S) \, P(D=1|S)) - r \, P(D=1)$, i.e., the one maximizing $E(\eta(S) \, P(D=1|S)$, which is precisely the indicator

$$\eta(s) \; = \; I_{[P(D=1|S=s) > q(r)]} \quad , \qquad P(\, P(D=1|S) > q(r) \,) \; \equiv \; r$$

Thus, the correlation maximizer is the indicator that $P(D=1|S)$ exceeds its $1-r$ quantile $q(r)$, and the correlation-maximizing threshold r is the one maximizing

$$E \quad P(D=1|S) \, (I_{[P(D=1|S) > q_r]} - r) \quad / \sqrt{r(1-r)} \tag{5}$$

Choice of a binary-classification metric, and of the threshold for binary classifiers $\eta(S) = I_{[S \geq c]}$, is an important topic in machine learning literature. See `www.cellprofiler.org/CPmanual/ApplyThreshold.html` for references to papers on "Maximum Correlation Thresholding". The paper [33] compares a number of threshold-based measures of quality for quantitative predictors S in binary classification, while [34] discusses and contrasts error-rate optimization with criteria of effectiveness of classifications based on AUC and other aspects of the ROC curve.

Universal use of a threshold 0.5 for the estimated scores $S = P(D=1|X)$ is not sensible when $E(D) = P(D=1)$ is much smaller than 0.5, since with small $E(D)$ it may be very rare to find $P(D=1|X) > 0.5$. (This point is argued persuasively in [31].) Thus, the unweighted misclassification-probability metric is unsuitable in a discussion of classifications for a number of different outcomes D with $P(D=1)$ ranging from small values to values in the range (0.3, 0.7). Since the overall proportion $n^{-1} \sum_{i=1} D_i \; \approx \; P(D=1)$ will generally be known approximately in a binary classification problem, and since a strong classification model will show a concentration of the histogram for $P(D=1|S)$ near $E(D)$, it makes more sense to try to optimize thresholds

Table 9: Calculated Thresholds and Correlations for Prediction Scores based on Logistic-Normal Model (6). `r*` is the maximizer r^* of (5), `maxcor` the correlation (4) at r^*, and `q(r*)` the threshold. `ed.cor` is (4) for $r = E(D)$, and `corDS` the correlation between D, S.

μ	σ	r*	maxcor	q(r*)	E(D)	ed.cor	corDS
-2.9444	0.5	0.2748	0.0936	0.0663	0.0555	0.0765	0.1182
-2.9444	1	0.1545	0.2189	0.1271	0.0727	0.209	0.2742
-2.9444	1.5	0.1331	0.3616	0.2181	0.1006	0.3584	0.4348
-2.1972	0.5	0.304	0.1254	0.1256	0.1089	0.1129	0.1572
-2.1972	1	0.2105	0.2693	0.1990	0.1339	0.2633	0.3305
-2.1972	1.5	0.2	0.4066	0.2819	0.1679	0.4047	0.4805
-0.8473	0.5	0.4067	0.1793	0.3253	0.3096	0.1768	0.2216
-0.8473	1	0.3709	0.3365	0.3733	0.331	0.3355	0.4027
-0.8473	1.5	0.371	0.4588	0.4125	0.3539	0.4585	0.5321
0	0.5	0.5	0.1919	0.5	0.5	0.1919	0.2363
0	1	0.5	0.3497	0.5	0.5	0.3497	0.4166
0	1.5	0.5	0.4683	0.5	0.5	0.4683	0.5414

for weighted misclassification probabilities with weights w near $P(D=1)/(P(D=0)$. When the weight w takes exactly this value, the optimal threshold for $S = P(D=1|X)$ is $P(D=1)$.

It often turns out that the optimal correlation-maximizing threshold is near $E(D)$ when that number is small, or in other words, that $r = E(D)$ approximately maximizes (5). However, there are also enough cases where they are different to make this criterion of threshold choice worth a separate discussion. This can be seen numerically by considering the *logistic-normal regression model* where vector covariates X are such that

$$P(D=1|X) \;=\; e^{\beta' X}/(1+e^{\beta' X}) \;\equiv\; \texttt{plogis}(\beta' X)\;, \qquad \beta' X \sim \mathcal{N}(\mu,\sigma^2) \tag{6}$$

holds roughly, with approximately normal distribution of $\beta' X$ in the population, for some μ, σ. In this setting, the quantity (5) can be maximized numerically, and Table 9 displays the comparison between $E(D)$ and the correlation-maximizing threshold for an array of different parameters (μ, σ). (The correlations exhibited there are as defined in (4.) The computations for this Table were done in `R` using Gaussian quadratures and the functions `integrate` and `optim`. Similar computations (not shown), done analogously when D instead follows a *probit regression model*,

$$P(D=1|X) \approx \Phi(\beta' X)\;, \qquad \beta' X \sim \mathcal{N}(\mu,\sigma^2) \tag{7}$$

led to similar results, which are displayed in Table 10. Both of these Tables exhibit a final column of correlations $\text{cor}(S, D)$ which by definition is larger than the maximum correlation (4) achievable with a binary-valued function $\eta(S)$, usually about 20% larger.

In Section 4.2 earlier in this report, we relied on the reasoning of this Section to justify the use of the binary imputation method $\eta(S) = I_{[S \geq q(r^*)]}$ where $q(r) \equiv F_S^{-1}(1-r)$ and r^* is the maximizer of (5) defined in terms of the logistic model (6), with $S_i = P(D=1|X_i)$ predicted on the whole population from a parametric statistical model, and with μ, σ^2 estimated as the sample mean and variance of the population-wide set of predictors *logit*(S_i). Another possibility,

Table 10: Calculated Thresholds and Correlations for Prediction Scores based on Probit-Normal Model (7). `r*` is the maximizer r^* of (5), `maxcor` the correlation (4) at r^*, and `q(r*)` the threshold. `ed.cor` is (4) for $r = E(D)$, and `corDS` the correlation between D, S.

μ	σ	r*	maxcor	q(r*)	E(D)	ed.cor	corDS
-2.9444	0.5	0.0564	0.0922	0.0157	0.0042	0.0687	0.1203
-2.9444	1	0.0377	0.3138	0.1218	0.0187	0.3024	0.3860
-2.9444	1.5	0.0643	0.5156	0.2531	0.0512	0.512	0.5953
-2.1972	0.5	0.123	0.1585	0.0529	0.0247	0.1345	0.2002
-2.1972	1	0.0951	0.3881	0.1874	0.0601	0.3796	0.4643
-2.1972	1.5	0.1297	0.5625	0.3065	0.1115	0.5602	0.6396
-0.8473	0.5	0.3306	0.2708	0.2650	0.2243	0.2646	0.3304
-0.8473	1	0.3094	0.4822	0.3633	0.2745	0.4806	0.5598
-0.8473	1.5	0.3329	0.616	0.4210	0.3192	0.6156	0.6890
0	0.5	0.5	0.2952	0.5	0.5	0.2952	0.3580
0	1	0.5	0.5	0.5	0.5	0.5	0.5774
0	1.5	0.5	0.6257	0.5	0.5	0.6257	0.6977

when datasets are large, is to maximize the correlation (5) using quantile estimators based on the (complete-case) population of model-based predictors $S_i = P(D_i = 1|X_i)$, instead of using a logit-normal or probit-normal model to define the quantiles.

The reasoning given here to support a flexible threshold for predictors S_i suggests also for CART models that the node-based piecewise-constant value of $P(D = 1|X)$ ought not to be mapped to $0, 1$ via the universal threshold $1/2$ as is typically done in `rpart` software, but perhaps ought to be thresholded either at $E(D)$ or at a value chosen adaptively by maximizing (5).

C.1. Illustrative ROC and Thresholds under Logistic-Normal Model

We compare and illustrate several different methods of choosing thresholds c for a binary classification rule $\eta(S) = I_{[S \geq c]}$ when the model (6) connecting S and D holds. In that case, $ED = P(D = 1) = \int \texttt{plogis}(\mu + \sigma z)\, \phi(z)\, dz$, and in terms of the standard normal density $\phi(z)$ and $\text{logit}(t) \equiv \log(t/(1-t))$, the ROC curve consists of points (FP(t), TP(t)) defined by

$$\begin{array}{c} \text{FP}(t) \\ \text{TP}(t)) \end{array} \quad = \quad \begin{array}{c} P(S \geq t \,|\, D = 0) \\ P(S \geq t \,|\, D = 1) \end{array} \quad = \quad {}_{t}^{1} \quad \begin{array}{c} 1/\{y\, ED\} \\ 1/\{(1-y)\,(1-ED)\} \end{array} \quad \phi(\frac{\text{logit}(y) - \mu}{\sigma})\, \frac{dy}{\sigma} \tag{8}$$

In that case, the slope of the ROC curve at the point $(\text{FP}(t), \text{TP}(t))$ is

$$\frac{d\,\text{TP}(t)}{dt} \Big/ \frac{d\,\text{FP}(t)}{dt} = \frac{1 - ED}{ED} \cdot \frac{t}{1-t} \tag{9}$$

which is an increasing function of t, and therefore a decreasing function of FP(t), and the ROC curve is strictly concave. The threshold corresponding to the slope $(1 - ED)/ED$ is 0.5, and the threshold ED corresponds to ROC slope 1. Finally, it is easy to see from (9) under this

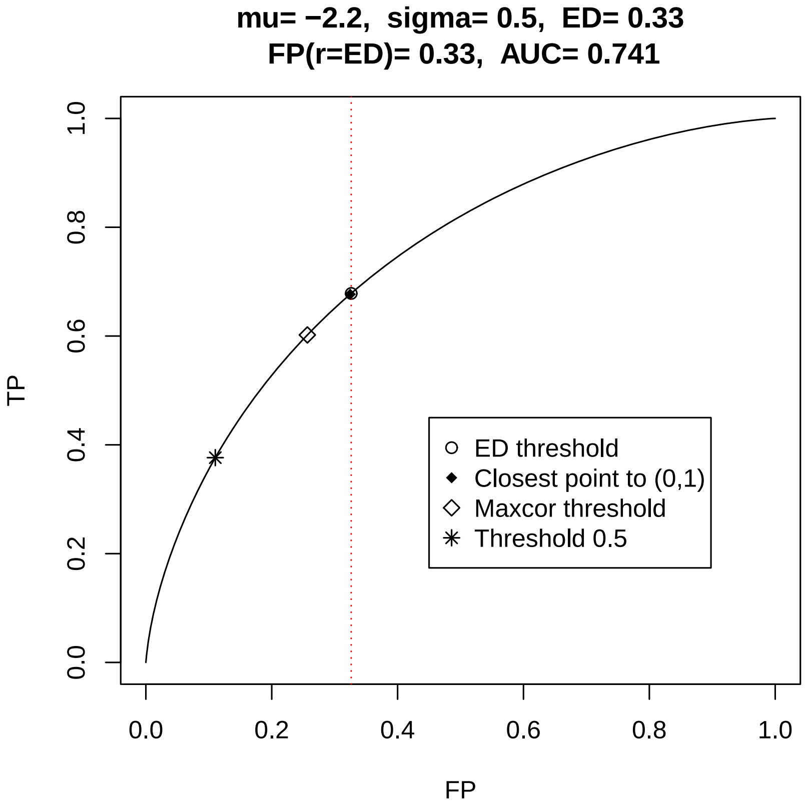

Figure 1: Illustrative ROC curve labeled with points at thresholds 0.5, q(r*) (called the 'Maxcor threshold'), E(D) and the closest point to (0,1), for the logistic-normal case in the eighth line of Table 9. NB: The ED threshold and point closest to (0,1) fall nearly on top of one another.

model that at the point on the ROC curve uniquely minimizing the distance to $(0, 1)$, t satisfies

$$\mathrm{FP}(t)/(1 - \mathrm{TP}(t)) \;=\; \{(1 - ED)/ED\} \;\cdot\; t/(1 - t) \tag{10}$$

Figure 1 illustrates the ROC curve labeled by the points with thresholds 0.5, ED, and $q(r^*)$, and the point closest to $(0, 1)$ for the logistic-normal model with the choice $(\mu, \sigma) \;=\; (-0.847,\, 1.0)$.

Other logistic-normal examples, not pictured, were computed with different choices of μ, σ. They show varying AUC values, such as those in Table 9, and a variety of relative positions of the ROC points associated with thresholds, r^* and $E(D)$. In all cases the point associated with threshold $E(D)$ was near, but not identical, to the ROC point closest to $(0, 1)$. This is apparently a feature of the logistic-normal family of examples, not of ROC curves in general, since the ROC curves based on examples of the probit-normal models (7) do not share it.

Appendix D. Specifications of Best Fitted Models

The following model-statement calls in R exhibit the variables and interactions used in the best-fitting logistic regression models for all of the outcome variables studied in this report, which are listed in order in Tables 3 and 4.

```
 FS ~ City + factor(RACE) + HSP + SPOU + OWN + SEX + factor(BLD) + factor(DIV) +
   City:factor(RACE) + City:HSP + City:SPOU + City:OWN + City:factor(BLD) +
   factor(RACE):HSP + factor(RACE):SPOU + factor(RACE):OWN + factor(RACE):
   factor(BLD) + HSP:SPOU + HSP:OWN + HSP:factor(BLD) + SPOU:OWN +
   SPOU:factor(BLD) + OWN:factor(BLD)

OWN ~ City + factor(RACE) + HSP + SPOU + factor(BLD) + City:HSP + + City:SPOU +
   City:factor(RACE) + City:factor(BLD) + factor(RACE):HSP + factor(RACE):SPOU +
   factor(RACE):factor(BLD) + HSP:SPOU + HSP:factor(BLD) + SPOU:factor(BLD)

I(BLD>2)~ factor(CCS) + factor(REG) + OWN + SPOU + factor(CCS):OWN + factor(CCS):
   factor(REG) + factor(CCS):SPOU + factor(REG):OWN + factor(REG):SPOU + OWN:SPOU

I(PIns>960) ~ factor(CCS) + factor(RACE) + HSP + factor(BLD) + SPOU + factor(REG)
   + factor(DIV) + factor(CCS):factor(RACE) + factor(CCS):HSP + factor(CCS):SPOU
   + factor(CCS):factor(DIV) + factor(CCS):factor(BLD) + factor(RACE):HSP +
   HSP:SPOU + factor(RACE):factor(REG) + HSP:factor(REG) + HSP:factor(BLD) +
   SPOU:factor(REG) + SPOU:factor(BLD) + factor(DIV):factor(BLD)

For I(PIns>1310), which is the PINSb model outcome, variables are the same as
for PINSa, which was I(PIns>960).

I(Mrgx==1) ~ factor(CCS) + factor(RACE) + HSP +SPOU + factor(DIV) + factor(BLD) +
   factor(CCS):factor(RACE) + factor(CCS):HSP + factor(CCS):SPOU + factor(CCS):
   factor(REG) + HSP:SPOU + HSP:factor(REG) + HSP:factor(BLD) + factor(CCS):
   factor(BLD) + factor(RACE):factor(DIV) + factor(RACE):factor(BLD) +
   SPOU:factor(REG) + SPOU:factor(BLD) + factor(REG):factor(BLD)

I(YBL <= 3) ~ factor(DIV) + (factor(CCS) + factor(RACE) + SPOU + factor(REG) +
   BLD)^2 - factor(RACE):SPOU

I(YBL%in%3:5) ~ factor(DIV) + factor(CCS) + factor(RACE) + SPOU + factor(REG) +
   factor(BLD) + factor(CCS):SPOU + SPOU:factor(REG) + SPOU:factor(BLD) +
   factor(REG):factor(BLD) + factor(CCS):factor(RACE) + factor(CCS):factor(REG) +
   factor(CCS):factor(BLD) + factor(RACE):factor(REG) + factor(RACE):factor(BLD)

I(MARHY<1970) ~ factor(CCS) + BLACK + WHITE + factor(REGION) + factor(BLD) +
   factor(AGE.OLD) + factor(CCS):factor(REGION) + factor(CCS):factor(BLD) +
   BLACK:WHITE + BLACK:SEX + WHITE:SEX + BLACK:factor(BLD) + WHITE:factor(BLD) +
   factor(BLD):SEX + factor(REGION):factor(BLD)
```

```
For I(MARHY<1990), which is the MARHYb model outcome, variables
are the same as for MARHYa, which was I(MARHY<1970).

NoHS ~ factor(CCS) + factor(BLD) + factor(REGION) +BLACK + HIS + WHITE + OWN
   + I(AGE<18) + I(AGE==18) + I(AGE>70) + factor(CCS):factor(BLD) +
   factor(CCS):HIS + factor(CCS):I(AGE<18) + factor(CCS):I(AGE==18) +
   factor(CCS):I(AGE>70) + factor(BLD):BLACK + factor(BLD):WHITE + factor(BLD):HIS
   + factor(BLD):OWN + factor(BLD):I(AGE<18) + factor(BLD):I(AGE==18) +
   factor(BLD):I(AGE>70) + factor(REGION):BLACK + factor(REGION):WHITE +
   factor(REGION):HIS + factor(REGION):OWN + factor(REGION):I(AGE<18) +
   factor(REGION):I(AGE == 18) + factor(REGION):I(AGE>70) + BLACK:HIS + BLACK:OWN
   + WHITE:HIS + BLACK:I(AGE<18) + BLACK:I(AGE==18) + BLACK:I(AGE>70) + WHITE:OWN
   + WHITE:I(AGE<18) + WHITE:I(AGE==18) + HIS:OWN + WHITE:I(AGE>70) + HIS:I(AGE<18)
   + HIS:I(AGE==18) + HIS:I(AGE>70) + OWN:I(AGE<18) + OWN:I(AGE==18) + OWN:I(AGE>70)

PostGR ~ factor(REGION) + BLACK + WHITE + HIS + factor(BLD) + OWN + I(AGE>40) +
   I(AGE>70) + SPOU + factor(REGION):BLACK + factor(REGION):WHITE +
   factor(REGION):HIS + BLACK:HIS + factor(REGION):OWN + factor(REGION):I(AGE>40) +
   BLACK:factor(BLD) + BLACK:OWN + BLACK:I(AGE>40) + WHITE:HIS + WHITE:OWN +
   HIS:OWN + WHITE:I(AGE>40) + WHITE:I(AGE>70) + OWN:SEX + OWN:REL + HIS:factor(BLD)
   + HIS:I(AGE>40) + HIS:I(AGE>70) + factor(BLD):OWN + factor(BLD):I(AGE>40) +
   factor(BLD):I(AGE>70) + OWN:I(AGE>40) + OWN:I(AGE>70) + I(AGE>40):SPOU +
   WHITE:SPOU + factor(BLD):SEX + factor(BLD):REL

BLACK ~ factor(CCS) + factor(REGION) + HIS + OWN + AGECAT10 + SEX + I(BLD==2) +
   SPOU + NoHS + Coll + PostGR + HIS:OWN + HIS:SEX + HIS:AGECAT10 + HIS:I(BLD==2)
   + HIS:SPOU + HIS:NoHS + HIS:Coll + HIS:PostGR + OWN:AGECAT10 + OWN:SEX +
   OWN:I(BLD==2) + OWN:SPOU + OWN:NoHS + OWN:Coll + OWN:PostGR + AGECAT10:SEX +
   AGECAT10:I(BLD==2) + AGECAT10:SPOU + AGECAT10:NoHS + AGECAT10:PostGR +
   SEX:I(BLD==2) + SEX:SPOU + SEX:NoHS + SEX:Coll + SEX:PostGR + I(BLD==2):SPOU +
   I(BLD==2):NoHS + I(BLD==2):Coll + I(BLD==2):PostGR + SPOU:NoHS + SPOU:Coll
    + SPOU:PostGR + NoHS:Coll + NoHS:PostGR + Coll:PostGR

HICOV ~ SEX + HIS + factor(REL) + factor(CCS) + OWN + AIAN + SPOU + factor(BLD) +
   I(AGE<16) + I(AGE>64) + SEX:SPOU + HIS:OWN + HIS:factor(REL) + HIS:AIAN +
   HIS:SPOU + factor(REL):OWN + OWN:SPOU + AIAN:SPOU + factor(REL):factor(BLD)

LANX ~ SEX + HIS + factor(REL) + factor(CCS) + OWN + AIAN + SPOU + SEX:SPOU +
   HIS:factor(REL) + HIS:OWN + HIS:AIAN + HIS:SPOU + factor(REL):OWN + OWN:SPOU
   + AIAN:SPOU

I(JWD < 620) ~ SEX + AGE + HIS + REL + WHITE + BLACK + NoHS + PostGR + Coll +
   factor(BLD) + OWN + AIAN + SPOU + SEX:AGE + SEX:HIS + SEX:REL + SEX:WHITE +
   SEX:NoHS + SEX:PostGR + SEX:Coll + SEX:factor(BLD) + SEX:OWN + SEX:AIAN +
   SEX:SPOU + AGE:HIS + AGE:REL + AGE:WHITE + AGE:BLACK + AGE:NoHS + AGE:PostGR +
   AGE:Coll + AGE:factor(BLD) + AGE:OWN + AGE:AIAN + HIS:REL + HIS:WHITE +
   HIS:BLACK + HIS:NoHS + HIS:Coll + HIS:SPOU + HIS:factor(BLD) + REL:WHITE +
   REL:BLACK + REL:NoHS + REL:Coll + REL:factor(BLD) + REL:OWN + REL:AIAN +
   REL:SPOU + WHITE:NoHS + WHITE:PostGR + WHITE:Coll + BLACK:NoHS + BLACK:PostGR
   + BLACK:Coll + NoHS:OWN + BLACK:factor(BLD) + NoHS:factor(BLD) + PostGR:SPOU +
   Coll:OWN + Coll:factor(BLD) + factor(BLD):OWN + OWN:SPOU + factor(BLD):AIAN +
   factor(BLD):SPOU + AIAN:SPOU

I(ESR == 6) ~ AGE + I(AGE^2) + NoHS + PostGR + Coll + SEX + HIS + REL + WHITE +
   BLACK + factor(BLD) + OWN + AIAN + SPOU + SEX:OWN + SEX:AIAN + SEX:SPOU +
   SEX:HIS + SEX:REL +  SEX:WHITE + SEX:BLACK + SEX:factor(BLD) + HIS:WHITE +
   HIS:REL + HIS:BLACK + HIS:factor(BLD) + HIS:OWN + HIS:SPOU + REL:WHITE +
   REL:BLACK + REL:factor(BLD) + REL:OWN + REL:SPOU + WHITE:BLACK + WHITE:
   factor(BLD) + WHITE:OWN + WHITE:AIAN + WHITE:SPOU + BLACK:factor(BLD) +
   BLACK:AIAN + BLACK:SPOU + factor(BLD):OWN + OWN:AIAN + factor(BLD):AIAN +
   factor(BLD):SPOU + AIAN:SPOU
```

Table 11: Numbers of imputation cells for ACS outcomes, and the variables used to define them.

Outcome	#cells	Variables used to define cells
FS	7	OWN, RACE, SPOU, BLD, City
OWN	7	BLD, SPOU, City, HSP
MULTI	13	OWN, CCS, SPOU, REG
PINS	11	BLD, SPOU, REG, CCS, RACE
MRGX	9	OWN, BLD, CCS, SPOU
YBL	9	BLD, REG, SPOU, CCS
MARHY	10	AGE, SEX, BLD
NoHS	7	AGE, WHITE, OWN
PostGR	12	OWN, WHITE, AGE
BLACK	10	AGE (no splits in CART!)
HICOV	9	quantiles of GLM (see App.D)
LANX	10	quantiles of GLM (see App.D)
JWD	10	SEX, Coll, AGE, BLD, REL
NILF	11	quantiles of GLM (see App.D)

Appendix E. Cells Used for Model-Based Imputation

For each outcome, the universe (given in Table 2) is decomposed into a set of 7–13 cells to be used in imputation with or without cross-classification by mode of response. The number of cells and variables used defining the decomposition are given in Table 11. Those cell decompositions with variables listed were defined from CART model hierarchies, and the others were defined from quantile-intervals of logistic regression model predictors, either because CART model cells were numerous or because no CART splits could be found (when outcome rate was close to 0 or 1). The variables entering the logistic regressions used to create quantile intervals from model predicted values can be read off from the model specification in Appendix D above. R code defining mode-pooled imputation cells for the HU-level outcome variables is as follows:

```
NodFS = ifelse(OWN==1, 1, ifelse(RACE!=2, 2, ifelse(SPOU==1, 3,
            ifelse(BLD==3, 4, ifelse(City==0, ifelse(BLD==2, 5, 6), 7) ) )))

OWNgp = ifelse(BLD==3, 1, ifelse(SPOU==1, 2,
               ifelse(City==1, ifelse(RACE!=1, 3, ifelse(HSP==1, 4, 5)),
                    ifelse(HSP==0, 6, 7)    )))

MultNod = ifelse(OWN==1, 1, (CCS-1)*4 + SPOU*2 + pmin(REG,2)+1)

PInsNod = ifelse(BLD==1, 2, ifelse(SPOU<0.5, 6, ifelse(REG!=3,
             ifelse(CCS==3, 28, ifelse(REG!=1, 58,
                 ifelse(CCS==2, 118, ifelse(BLD==2, 238, 239) ) ) ),
       ifelse(CCS==3, 30, ifelse(BLD==3, 62, ifelse(RACE!=1,126,127) )) )))

MrgxNod = ifelse(OWN<.5, 1, 2 + 4*(BLD>1) + 2*(CCS<3) + SPOU)

YBLnod = ifelse(BLD==1, 1, ifelse(REG>2, 2, (REG==1)*4 + SPOU*2 + (CCS==1) + 3))
```

E.1. Outcome Proportions by Node Group

Table 12 contains Counts and Outcome=1 rates for 15 of the 17 outcome variables studied. The only ones omitted are MULTI and PostGR, because they had more than 11 node groups and a similar pattern of separation between the Outcome=1 rates in the separate node groups. Note that the separations in rates are systematically largest in those instances (HICOV, LANX, NILF) where the node groups are created from quantiles of logistic regression predictors.

Table 12: Summary of Counts (in thousands) and `Outcome=1` rates in App. E `Node` groups

	Node	1	2	3	4	5	6	7	8	9	10	11
FS	Rate	.063	.231	.280	.428	.463	.558	.527				
	Count	1,422	478	14.9	52.6	10.6	1.8	14.4				
OWN	Rate	.156	.901	.621	.646	.787	.794	.651				
	Count	389	948	53.5	14.3	106	462	28.9				
PINSa	Rate	.087	.274	.291	.421	.360	.266	.403	.448	.523	.470	.517
PINSb	Rate	.024	.125	.117	.204	.161	.149	.206	.241	.312	.250	.351
PINS	Count	63.9	389	120	48.3	246	3.8	91.7	23.0	136	13.1	4.1
MRGX	Rate	.281	.423	.256	.395	.417	.561	.584	.710			
	Count	22.7	22.1	27.1	22.8	119	209	377	617			
YBLa	Rate	*	.239	.443	.540	.352	.474	.507	.658	.412		
YBLb	Rate	.291	.392	.416	.397	.364	.401	.398	.332	.385		
YBL	Count	111	883	153	61.1	231	43.5	101	70.6	145		
MYRa	Rate	.009	.263	.386	.470	.530	.599	.618	.615	.721	.828	
MYRb	Rate	.362	.817	.818	.857	.866	.879	.867	.867	.913	.938	
MARHY	Count	1,930	119	95.0	32.6	25.1	48.0	39.2	81.4	138	286	
NoHS	Rate	.964	.450	.101	.191	.349	.502	.340				
	Count	124	52.5	3,270	441	45.0	16.3	11.9				
BLACK	Rate	.023	.041	.048	.049	.051	.058	.083	.123	.165	.295	
	Count	687	684	253	179	144	912	525	464	404	483	
HICOV	Rate	.558	.660	.749	.785	.816	.844	.897	.937	.949		
	Count	87.8	149	161	387	299	667	687	676	1,233		
LANX	Rate	.077	.121	.175	.176	.171	.108	.189	.185	.213	.726	
	Count	3,213	367	38.2	70.2	32.1	5.5	11.7	54.3	131	535	
JWDa	Rate	.146	.158	.188	.286	.297	.265	.368	.381	.354	.449	
	Count	909	327	87.7	77.7	69.7	23.5	360	11.9	3.4	28.8	
NILF	Rate	.033	.090	.128	.168	.196	.227	.288	.425	.667	.848	.968
	Count	377	375	187	379	188	377	376	564	378	375	189

References

[1] M. Joshipura, 2005 Respondent characteristics evaluation (2008), *ACS Memo. Series* **ACS 08-RER-03** `Downloads/library/2008/2008_Joshipura_01.pdf`
NB: ACS URL's in `Downloads/` or `methodology/` directory at `http://www.census.gov/acs/www/`

[2] C. Alexander, The American Community Survey design issues and initial test tesults (1996). `Downloads/library/1997/1997_Alexander_02.pdf`

[3] Lynn Weidman (2013), personal communication.

[4] ACS Design & Methodology, Ch. 10, Data preparation and processing for housing units and group quarters (2014). `methodology/methodology_main/`

[5] A. Tersine, Item non-response: 1996 American Community Survey (1998). Presented at 1998 ACS Symposium. `Downloads/library/1998/1998_Tersine_01.pdf`

[6] T. Leslie, D. Raglin, and E. Braker, Can the American Community Survey trust using respondent data to impute data for survey nonrespondents? Are non-respondents to the ACS different than respondents? (2003) `Downloads/library/2003/2003_Leslie_01.pdf`

[7] S. Clark, American Community Survey item nonresponse rates: mail versus internet (2014), **ACS 14-RER-04**. `Downloads/library/2014/2014_Clark_01.pdf`

[8] S. Clark, Evaluation of the effect on item nonresponse of changes to the Failed Edit Follow-up Operation (2014), **ACS 14-RER-26**. `Downloads/library/2014/2014_Clark_03.pdf`

[9] J.N.K. Rao and J. Shao, Jackknife variance estimation with survey data under hot deck imputation (1992), *Biometrika* **79**, 811-822.

[10] J.N.K. Rao and J. Shao, Modified balanced repeated replication for complex survey data (1992) , *Biometrika* **86**, 403-415.

[11] D. Rubin, **Multiple Imputation for Nonresponse in Surveys.** John Wiley, 1987.

[12] W. Fuller and J.-K. Kim, Fractional hot deck imputation (2004), *Biometrika* **91**, 559-578.

[13] J.N.K. Rao, **Small Area Estimation.** Wiley-Interscience, 2003.

[14] Y. Thibaudeau, Model explicit item imputation for demographic categories (2002), *Survey Methodology* **28**, 135-143.

[15] Y. Thibaudeau, E. Slud and A. Gottschalk, Modeling log-linear conditional probabilities for prediction in surveys (2014), Census Bureau preprint.

[16] J. Chipperfield, J. Chessman, and R. Lim, Combining household surveys using mass imputation to estimate population totals (2012), *Austral. & New Zeal. Jour. Statist.* **54**, 223-238.

[17] S. van Buuren, Fully Conditional Specification (2014). Ch. 13, pp. 267-294 in: *Handbook of Missing Data Methodology*, eds. G. Molenberghs et al. CRC Press, Taylor and Francis.

[18] T. Raghunathan, J. Lepkowski, J. Van Hoewyk, and P. Solenberger, A multivariate technique for multiply imputing missing values using a sequence of regression models (2001), *Survey Methodology*, **27**(1):85-95.

[19] J. Schafer, **Analysis of Incomplete Multivariate Data.** CRC/Chapman & Hall, 1997.

[20] M. Garcia, C. Erdman, and B. Klemens, Multiple imputation methods for imputing earnings in the Survey of Income and Program Participation (2014). *Proc. U.N. Econ. Commission for Europe Conf. of European Statisticians*, Paris, April 2014.

[21] G. Benedetto and M. Stinson, Testing new imputation methods for earnings collected by the Survey of Income and Program Participation (2009), Census Bureau preprint.

[22] L. Doove, S. v. Buuren and E. Dusseldorp, Recursive partitioning for missing data imputation in the presence of imputation effects (2014), *Comp. Statist. & Data Anal.* **72**, 92-104.

[23] D. Griffin, Effect of changing call parameters in the American Community Survey's computer assisted telephone interviewing operation (2013), *Memo. Series* **ACS13-RER-17**. `Downloads/library/2013/2013_Griffin_03.pdf`

[24] 2012 Planning Database Documentation, US Census Bureau (2013). `http://www.census.gov/research/data/planning_database/2012/`

[25] L. Breiman, R. Olshen, J. Friedman and C. Stone, **Classification and Regression Trees.** Wadsworth, 1984.

[26] T. Therneau and E. Atkinson, An introduction to recursive partitioning using the RPART routines (2013), Mayo Foundation preprint, online documentation for `rpart` in `R`.

[27] R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org/`.

[28] J. N. K. Rao and A. Scott, The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence of two-way tables (1981), *Jour. Amer. Statist. Assoc.* **76**, 221-230.

[29] Wikipedia article, Receiver Operating Characteristic (2015), `http://en.wikipedia.org`

[30] P. Flach, ROC Analysis, article in *Encyclopedia of Machine Learning*, eds. C. Sammut and G. Webb, Springer, 2010.

[31] F. Provost and T. Fawcett, Robust classification for imprecise environments (2001), *Machine Learning* **42**, 203-231.

[32] P. Rosenbaum and D. Rubin, The central role of the propensity score in observational studies for causal effects (1983), *Biometrika* **70**, 41-55.

[33] E. Freeman and G. Moisen, A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa (2008), *Ecol. Modeling* **217**, 48-58.

[34] C. Cortes and M. Mohri, AUC Optimization versus Error-rate Minimization, in: *Advances in Neur. Info. Process. Sys.* (NIPS 2003), vol. 16, Vancouver, Canada, MIT Press, 2004.

[35] P. Baldi, S. Brunak, et al., Assessing the accuracy of prediction algorithms for classification: an overview (2000), *Bioinformatics Review* **16**, 412-424.

[36] T. Hastie, R. Tibshirani and J. Friedman, **The Elements of Statistical Learning**. 2nd ed. Springer, 2009.