ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS DE MATHEMATICA
Volume 21, Number 1, June 2017
Available online at http://acutm.math.ut.ee

Downward calibration property of estimated response propensities

Natalja Lepik and Imbi Traat

Abstract. We consider four methods for estimating response propensities: three traditional ones (linear, logistic, probit) and one more recent, a decision tree method. We show that some but not all the methods produce estimates that calibrate sample totals of auxiliary variables down to the response set totals. The downward calibration property reveals interesting relationships between estimated propensities, auxiliary variables, and true response probabilities. However, the property itself does not guarantee more accurate propensity estimation. Our simulation study shows that the accuracy of the estimation method depends primarily on the relationship nature between true response probabilities and auxiliary variables.

1. Introduction

Let $U = \{1, \ldots, N\}$ be a finite population of N units. Let s be a sample of size n drawn by some sampling design, so that the inclusion probability of unit k is a known quantity $\pi_k = \Pr(k \in s)$. The inverse of inclusion probability is called the design weight, $d_k = \pi_k^{-1}$. Let r denote a response set of size m – a subset of s, where study variables are measured:

$$r \subset s \subset U.$$

A common feature of nowadays sample surveys is low response rate, i.e., the size of r is often much less than the sample size. Since response set is usually biased compared to the full sample s, simple estimates computed from r are also biased. Adjustments have to be made to reduce nonresponse bias.

Received October 9, 2016.
2010 *Mathematics Subject Classification.* 62D05.
Key words and phrases. Nonresponse; response propensity; response probability; downward calibration; linear regression; logit; probit; decision tree.
http://dx.doi.org/10.12697/ACUTM.2017.21.07

Formation of r is subject to a response mechanism, unknown to us. Let R_k be the response indicator for the unit k in s. Its value 1 means response ($k \in r$), and the value 0 means nonresponse ($k \in s-r$). Denote the response probability of unit $k \in s$ by θ_k,

$$\theta_k = \Pr(R_k = 1 | k \in s) = E(R_k|s).$$

Here, and later in this paper, whenever the operator $E(\cdot)$ is used, it means the expectation with respect to the response mechanism, unless otherwise stated. Sample s is assumed to be fixed in this paper. If the response probabilities θ_k were known for $k \in r$ then unbiased estimator for the population total $t = \sum_U y_k$ is

$$\hat{t} = \sum_r d_k \theta_k^{-1} y_k,$$

where y_k is the study variable value for unit k, available only in r. The unbiasedness is shown by the usual two-phase reasoning (see, e.g., [8]). Here and elsewhere in the paper, the notation $\sum_A$ means summation over k in the set A. Unfortunately θ_k are not known. It is natural for a statistician to estimate θ_k by $\hat{\theta}_k$ and use $\hat{\theta}_k^{-1}$ as a nonresponse adjustment weight. A review on nonresponse weighting adjustments is given by Brick in [2]. In this paper we do not concentrate on the estimator $\hat{\hat{t}} = \sum_r d_k \hat{\theta}_k^{-1} y_k$, but rather on the properties of $\hat{\theta}_k$.

Usually response is modeled with available auxiliary information. Denote by $\mathbf{x}_k$ a J-dimensional vector of auxiliary variables, known for all $k \in s$. Nowadays sample surveys can obtain auxiliary information from many registers and data files, so that the dimensionality J can be considerably high. The model is built for conditional probability

$$D_k = \Pr(R_k = 1 | \mathbf{x}_k, k \in s),$$

called response propensity (see, e.g., [10]). The response propensity is estimated from the data $(R_k, \mathbf{x}_k)$, $k \in s$ and then used as an estimate for the response probability:

$$\hat{\theta}_k = \hat{\Pr}(R_k = 1 | \mathbf{x}_k, k \in s) = \hat{D}_k.$$

The response propensities are unbiased for the response probabilities θ_k,

$$\theta_k = E(R_k|s) = EE(R_k|\mathbf{x}_k, s) = ED_k,$$

where the first expectation in $EE(\cdot)$ is with respect to the distribution of $\mathbf{x}_k$. Notice that unbiasedness cannot be claimed for the estimated response propensities, i.e., generally $E\hat{D}_k = E\hat{\theta}_k \neq \theta_k$. Many modeling methods are available for binary variable, here R_k. Along classical old statistical methods, like logistic and probit regression, many new methods have been developed.

Computer science has been the driving force here. These algorithmic methods are classified as machine learning or statistical learning methods (see, e.g., [5]). The new methods like decision trees, random forests, support vector machines and others have just started to find their application in various areas of statistics. For example, in sampling theory the decision tree method has very recently been described for nonresponse weighting in [11], and for analyzing nonresponse structure in [6].

Our interest lies in the properties of $\hat{\theta}_k$, obtained by different methods. Primarily we are interested in the property, defined in the next section, which we call downward calibration. The property is known to hold for linear regression (see, e.g., [9]). We show that this property holds also for logistic regression, but in general not for a probit model and not for a decision tree method.

The three classical methods (linear, logistic, probit) produce rather similar estimates $\hat{\theta}_k$. We show that these methods give exactly the same expression for $\hat{\theta}_k$ for special configuration of an auxiliary vector, called the group vector. The group vector allows to simplify many complex expressions involving auxiliary variables and, in this way, do analytical comparisons. The J-dimensional vector $\mathbf{x}_k$ is called a group vector if it contains zeros and only one 1, identifying the group out of J groups where the unit k belongs: $\mathbf{x}_k' = (0, \ldots, 0, 1, 0, \ldots, 0)$ with 1 in position j meaning that k belongs to group j. In practice the group vector can be formed by crossing several categorical variables. For example, crossing variables sex with 2 categories and education with 3 categories a 6-dimensional group vector is received.

The decision tree method is a non-parametric method. It divides sample s into subgroups by the values of auxiliary variables, and estimates response probabilities by the response proportions in these groups. An advantage of the tree method is its ability to capture nonlinear dependence between propensity to respond and auxiliary variables.

The results of the paper are illustrated in a simulation study, carried out on real data. The two different response mechanisms in a fixed sample s are considered; the true response probabilities depending either approximately linearly, or nonlinearly, on the auxiliary variables $\mathbf{x}_k$. The four methods (linear, logistic, probit, decision tree) and two combined methods are considered. The combined methods remodel the decision tree output with logistic regression. The downward calibration property of $\hat{\theta}_k$ and its performance in estimation of true θ_k is illustrated.

2. Downward calibration and interpretation

Definition 2.1. We say that propensity estimates $\hat{\theta}_k$, $k \in s$, have downward calibration property if the following holds:

$$\sum_s d_k \hat{\theta}_k \mathbf{x}_k = \sum_r d_k \mathbf{x}_k \quad (J \text{ equalities}). \tag{1}$$

The calibration takes place downward since the bigger set s includes a smaller set r, and the weights $\hat{\theta}_k$ applied to the design-weighted total in s calibrate it to the design-weighted total in r. From this point of view, the usual calibration in sample surveys is an upward calibration, from r to s or even to U.

Note that if the calibration property (1) holds for $\mathbf{x}_k$, then it holds for any linear transformation $A\mathbf{x}_k$.

Proposition 2.1. *The following holds for the sums with true θ_k:*

$$E\left(\sum_r d_k\right) = \sum_s d_k \theta_k, \tag{2}$$

$$E\left(\sum_r d_k \mathbf{x}_k\right) = \sum_s d_k \theta_k \mathbf{x}_k. \tag{3}$$

Proof. The proof is straightforward by using $E(R_k|s) = \theta_k$, and noting that $\sum_r d_k = \sum_s d_k R_k$ and $\sum_r d_k \mathbf{x}_k = \sum_s d_k R_k \mathbf{x}_k$. □

The proposition says that $\sum_r d_k$ and $\sum_r d_k \mathbf{x}_k$ are unbiased for the respective sums with true probabilities.

The downward calibration property reveals a relationship between estimates $\hat{\theta}_k$ and true θ_k.

Proposition 2.2. *If $\hat{\theta}_k$, $k \in s$, satisfy the downward calibration property* (1)*, then*

$$E\left(\sum_s d_k \hat{\theta}_k \mathbf{x}_k\right) = \sum_s d_k \theta_k \mathbf{x}_k$$

Proof. The proof follows from (1) and (3). □

Let now $\mathbf{x}_k$ be such that for a constant vector $\boldsymbol{\mu}$,

$$\boldsymbol{\mu}' \mathbf{x}_k = 1, \ \forall k. \tag{4}$$

The condition is not too restrictive. For example in the models with intercept, the first variable in $\mathbf{x}_k$ is a constant 1, and then $\boldsymbol{\mu} = (1, 0, \ldots, 0)$. For $\mathbf{x}_k$ as a

group vector, $\boldsymbol{\mu}$ is the vector of ones. For such $\mathbf{x}_k$, the downward calibration property (1) states that

$$\sum_s d_k \hat{\theta}_k = \sum_r d_k. \tag{5}$$

Proposition 2.3. *If the downward calibration* (1) *holds, and* $\mathbf{x}_k$ *has property* (4)*, then*

$$E\left(\sum_s d_k \hat{\theta}_k\right) = \sum_s d_k \theta_k.$$

Proof. The proof follows from (5) and (2). □

Note that the unbiasedness of $\sum_s d_k \hat{\theta}_k$ for $\sum_s d_k \theta_k$ does not imply that $\hat{\theta}_k$ is unbiased for θ_k.

The above results can be stated for the weighted averages, just by dividing sums by $\sum_s d_k$. We state the following corollary.

Corollary 2.1. *The generalized response rate* $\sum_r d_k / \sum_s d_k$ *is unbiased for the weighted average of true response probabilities,*

$$E\left(\frac{\sum_r d_k}{\sum_s d_k}\right) = \frac{\sum_s d_k \theta_k}{\sum_s d_k}.$$

If the downward calibration and the property (4) *for* $\mathbf{x}_k$ *hold, then*

$$\frac{\sum_s d_k \hat{\theta}_k}{\sum_s d_k} = \frac{\sum_r d_k}{\sum_s d_k},$$

and further, the average of estimates $\hat{\theta}_k$ *is unbiased for the average of true probabilities,*

$$E\left(\frac{\sum_s d_k \hat{\theta}_k}{\sum_s d_k}\right) = \frac{\sum_s d_k \theta_k}{\sum_s d_k}.$$

For self-weighting designs when d_k is constant for each k, we get from (5) that sample sum of propensity estimates is the number of respondents m and generalized response rate is just the ordinary response rate m/n.

Let now $\mathbf{x}_k$ be the group vector. It divides the sample s into J non-overlapping and exhaustive groups s_j with equal $\mathbf{x}_k$ inside the groups. Similarly, it divides r into groups r_j. In group vector case, the downward calibration property (1) takes the form

$$\sum_{j=1}^{J}\sum_{s_j} d_k\hat{\theta}_k\mathbf{x}_k = \sum_{j=1}^{J}\sum_{r_j} d_k\mathbf{x}_k,$$

or, alternatively,

$$\sum_{s_j} d_k\hat{\theta}_k = \sum_{r_j} d_k,\ j = 1,\dots J.$$

Since $\hat{\theta}_k = \hat{\Pr}(R_k = 1|\mathbf{x}_k, k \in s)$ is computed for given $\mathbf{x}_k$, it is constant inside the group, i.e., $\hat{\theta}_k \equiv \hat{\theta}_j$ for $k \in s_j$. Next, we summarize the result.

Proposition 2.4. *For the group vector case, and for the methods with downward calibration property, the estimated response propensities are generalized response proportions (rates) in the groups, i.e., for each unit in s_j,*

$$\hat{\theta}_j = \frac{\sum_{r_j} d_k}{\sum_{s_j} d_k}. \tag{6}$$

It is easy to check that inverted $\hat{\theta}_j$ in (6) are calibration weights in the traditional (upward) sense for a group vector $\mathbf{x}_k$,

$$\sum_{r} d_k\hat{\theta}_k^{-1}\mathbf{x}_k = \sum_{j=1}^{J}\hat{\theta}_j^{-1}\sum_{r_j} d_k\mathbf{x}_k = \sum_{s} d_k\mathbf{x}_k,$$

and the estimator $\hat{\hat{t}} = \sum_r d_k\hat{\theta}_k^{-1}y_k$ is the calibration estimator for $t = \sum_U y_k$. But in general the inverted propensity estimate $\hat{\theta}_k^{-1}$ does not have the calibration property.

More general $\mathbf{x}_k$-vectors (not necessarily group vectors), can be also used to divide s into J non-overlapping and exhaustive groups s_j with $\mathbf{x}_k$ satisfying certain criterion inside the groups. For example, the decision tree method creates such groups by certain optimization algorithm, and computes propensities as response proportions (6) inside the groups, the same for each unit in the group. In this situation, the bias statements for $\hat{\theta}_j$ are not so clear, because for given s the partition into s_j is random, it depends on the realized response set r. But, anyway, the following holds for the sum of weighted propensity estimates:

$$E\left(\sum_{j=1}^{J}\hat{\theta}_j(\sum_{s_j} d_k)\right) = E\left(\sum_r d_k\right) = \sum_s d_k\theta_k.$$

Not all estimation methods obey downward calibration property. Let $\hat{\theta}_k$ be estimated by such a method. If the downward calibration is the aim, then an appropriate re-modeling method can be applied on R_k with $\hat{\theta}_k$ as a covariate. Due to (1), the resulting new propensity estimates $\hat{\theta}_{k,new} = \hat{\Pr}(R_k = 1|\hat{\theta}_k, s)$ will be related to the original ones by

$$\sum_s d_k\hat{\theta}_{k,new}\hat{\theta}_k = \sum_r d_k\hat{\theta}_k.$$

Similarly, the extended auxiliary vector $(\hat{\theta}_k, \mathbf{x}_k)$ can be used for the secondary modeling.

3. Linear, logistic and probit methods

In this section we consider three parametric methods for estimating response propensities – linear, logistic, and probit methods. It has been shown earlier, e.g., in [9] that downward calibration property, although not named so, holds for the propensity estimates $\hat{\theta}_{k,lin}$ received from the linear regression. We show that the downward calibration property holds for the propensity estimates $\hat{\theta}_{k,log}$ computed from the logistic regression, in spite of the fact that no analytical form exists for $\hat{\theta}_{k,log}$. We also show that, in general, the property does not hold for the probit model. The logistic regression and probit models are well explained in [1].

3.1. Linear method. In linear modeling, R_k is assumed to depend linearly on $\mathbf{x}_k$. Regression coefficients $\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_J)$ are estimated by minimizing the design-weighted sum of squares

$$Q(\boldsymbol{\beta}) = \sum_s d_k(R_k - \boldsymbol{\beta}'\mathbf{x}_k)^2.$$

The design weight $d_k = 1/\pi_k$ expresses the importance of unit $k \in s$. The minimum over $\boldsymbol{\beta}$ is received for

$$\hat{\boldsymbol{\beta}} = \left(\sum_s d_k\mathbf{x}_k\mathbf{x}_k'\right)^{-1}\sum_s d_k\mathbf{x}_k R_k.$$

With $\hat{\boldsymbol{\beta}}$ we get the prediction for R_k given $\mathbf{x}_k$, it is the estimated response propensity:

$$\hat{\theta}_{k,lin} = \hat{\boldsymbol{\beta}}'\mathbf{x}_k. \tag{7}$$

It is easy to check by inserting (7) into (1) that downward calibration holds for $\hat{\theta}_{k,lin}$:

$$\sum_s d_k \hat{\theta}_{k,lin} \mathbf{x}_k = \sum_r d_k \mathbf{x}_k.$$

The drawback of the linear method is that $\hat{\theta}_{k,lin}$ may sometimes obtain improper values, negative or greater than one. The positive side is the explicit formula for $\hat{\theta}_{k,lin}$ which allows further analytical studies.

Since the downward calibration holds, the propensity estimates for the group vector $\mathbf{x}_k$ are response proportions in groups, $\hat{\theta}_{k,lin} = \hat{\theta}_j$, $k \in s_j$, where $\hat{\theta}_j$ is in (6). On the other hand, (7) says for the group vector $\mathbf{x}_k$ that

$$\hat{\beta}_j = \hat{\theta}_{k,lin},\ k \in s_j.$$

Consequently, estimated regression coefficients are also response proportions (6) in the groups.

3.2. Logistic method. For logistic modeling, we assume $R_k \sim B(1, D_k)$, where $D_k = P(R_k = 1|\mathbf{x}_k, s)$ has the form

$$D_k = \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_k)}{1 + \exp(\boldsymbol{\beta}'\mathbf{x}_k)}. \tag{8}$$

Proposition 3.1. *Modeling R_k on the auxiliary vector $\mathbf{x}_k$ with logistic regression method, the maximum likelihood estimates of the response propensities D_k, denoted by $\hat{\theta}_{k,log}$, satisfy the downward calibration property:*

$$\sum_s d_k \hat{\theta}_{k,log} \mathbf{x}_k = \sum_r d_k \mathbf{x}_k. \tag{9}$$

Proof. We have data $(R_k, \mathbf{x}_k)$ with weights d_k, $k \in s$. With the usual interpretation of d_k (it shows how many population units the sample unit k represents), the following likelihood function is natural:

$$L(\boldsymbol{\beta}) = \prod_s D_k^{d_k R_k} (1 - D_k)^{d_k(1-R_k)}.$$

We maximize the log-likelihood

$$l(\boldsymbol{\beta}) = \sum_s [d_k R_k \log(D_k) + d_k(1 - R_k)\log(1 - D_k)]. \tag{10}$$

Matrix differentiation gives

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_s \left[\frac{d_k R_k}{D_k} \frac{\partial D_k}{\partial \boldsymbol{\beta}} - \frac{d_k(1 - R_k)}{1 - D_k} \frac{\partial D_k}{\partial \boldsymbol{\beta}} \right], \tag{11}$$

where

$$\frac{\partial D_k}{\partial \boldsymbol{\beta}} = D_k(1 - D_k)\mathbf{x}_k. \tag{12}$$

Inserting (12) into (11) and equating to 0 gives the equations

$$\sum_s [d_k R_k - d_k D_k]\mathbf{x}_k = 0. \tag{13}$$

A numerical solution for D_k is the estimated propensity $\hat{\theta}_{k,log}$, satisfying (13), where we then observe the downward calibration property (9). □

Similarly to the linear regression case, we have for the group vector $\mathbf{x}_k$ the propensity estimates as response proportions in groups, $\hat{\theta}_{k,log} = \hat{\theta}_j$, $k \in s_j$, where $\hat{\theta}_j$ is given by (6). The estimated regression coefficients follow from (8), which now has the form $D_k = \frac{\exp(\beta_j)}{1+\exp(\beta_j)}$, $k \in s_j$. Replacing D_k by $\hat{\theta}_{k,log} = \hat{\theta}_j$, we get

$$\hat{\beta}_j = \log \frac{\hat{\theta}_j}{1 - \hat{\theta}_j},\ k \in s_j,$$

where $\hat{\theta}_j$ is given by (6).

3.3. Probit method. For probit modeling, we also assume that $R_k \sim B(1, D_k)$, where $D_k = P(R_k = 1|\mathbf{x}_k, s)$. But now, instead of being the logistic function (8), D_k is the probit function:

$$D_k = \Phi(\boldsymbol{\beta}'\mathbf{x}_k),$$

where $\Phi(\cdot)$ is the standard normal distribution function.

Proposition 3.2. *Modeling R_k on the auxiliary vector $\mathbf{x}_k$ with probit method, the response propensities D_k, $k \in s$, are estimated by $\hat{\theta}_{k,prb} = \Phi(\hat{\boldsymbol{\beta}}'\mathbf{x}_k)$, where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate of $\boldsymbol{\beta}$. The estimated propensities $\hat{\theta}_{k,prb}$ satisfy*

$$\sum_s d_k \hat{\theta}_{k,prb} \hat{C}_k \mathbf{x}_k = \sum_r d_k \hat{C}_k \mathbf{x}_k, \tag{14}$$

where

$$\hat{C}_k = \frac{\phi(\hat{\boldsymbol{\beta}}'\mathbf{x}_k)}{\Phi(\hat{\boldsymbol{\beta}}'\mathbf{x}_k)[1 - \Phi(\hat{\boldsymbol{\beta}}'\mathbf{x}_k)]}$$

with $\phi(\cdot)$ being the density function of standard normal distribution.

Proof. The log-likelihood is again given by (10), and the derivative with respect to $\boldsymbol{\beta}$ by (11). A difference from the logistic method comes in the term

$$\frac{\partial D_k}{\partial \boldsymbol{\beta}} = \frac{\partial(\Phi(\boldsymbol{\beta}'\mathbf{x}_k))}{\partial \boldsymbol{\beta}} = \phi(\boldsymbol{\beta}'\mathbf{x}_k)\mathbf{x}_k. \tag{15}$$

Inserting (15) into (11) and equating it to zero gives

$$\sum_s [d_k R_k - d_k D_k]\mathbf{x}_k C_k = 0, \tag{16}$$

where $C_k = \phi(\boldsymbol{\beta}'\mathbf{x}_k)/[\Phi(\boldsymbol{\beta}'\mathbf{x}_k)[1-\Phi(\boldsymbol{\beta}'\mathbf{x}_k)]]$. A solution for $\boldsymbol{\beta}$ gives a solution for D_k which is the estimated response propensity $\hat{\theta}_{k,prb}$. With solutions in (16) we recognize the relationship (14), and the proposition is proved. □

We see that the downward calibration property does not generally hold for $\hat{\theta}_{k,prb}$. There is an additional factor $\hat{C}_k$.

For $\mathbf{x}_k$ as a group vector, the equality (14) can be written in the form

$$\sum_{j=1}^{J} \sum_{s_j} [d_k R_k - d_k \hat{\theta}_{k,prb}]\mathbf{x}_k \hat{C}_k = 0. \tag{17}$$

Since $\mathbf{x}_k$ is a constant vector in s_j having 1 in the position j and zeros elsewhere, also $\hat{C}_k$ as well $\hat{\theta}_{k,prb}$ are constants in s_j, denoted respectively by c_j and $\hat{\theta}_{j,prb}$. The vector equality (17) breaks down to J equalities

$$c_j \sum_{s_j} [d_k R_k - d_k \hat{\theta}_{j,prb}] = 0, \ j = 1, \ldots, J. \tag{18}$$

Finally, we see from (18) that the estimated propensities are just generalized response rates in the groups, $\hat{\theta}_{j,prb} = \hat{\theta}_j$, $j = 1, \ldots, J$, given by (6). In the group vector case, also the coefficients β_j have simple estimates. Since $\hat{\theta}_{k,prb} = \hat{\theta}_{j,prb} = \Phi(\hat{\beta}_j)$ for $k \in s_j$, one has

$$\hat{\beta}_j = \Phi^{-1}(\hat{\theta}_j), \ j = 1, \ldots, J,$$

where $\hat{\theta}_j$ is in (6).

Considering all three methods, we can say that for a group vector case, they all produce the same propensity estimates, equal to the generalized response rates in the groups, though the regression coefficients $\hat{\beta}_j$ are different for each method.

3.4. Decision tree method. The decision tree method is a non-parametric method. Depending on the type of the modeled variable, there are regression trees and classification trees (see, e.g., [5]). In nonresponse case we model the categorical, more precisely, the binary R_k, and thus have the classification tree. The tree is grown using data $(R_k, \mathbf{x}_k)$, $k \in s$. In each step one variable from auxiliary vector $\mathbf{x} = (x_1, \ldots, x_J)$, say x_j is chosen, and a split is made $x_j < a$ and $x_j \geq a$. In the first step, this divides all units of s into 2 nodes s_1 and s_2 (subsets). In each of these nodes the response proportion is found by $\hat{\theta}_{s_j} = \sum_{s_j} d_k R_k / \sum_{s_j} d_k$; the same for all units in s_j, $j = 1, 2$. The splitting

point a is chosen from the purity of the node criterion. The Gini index G_{s_j} or the deviance criterion D_{s_j} is used (see [4]). The often default choice is D_{s_j} which for binary R_k takes the form

$$D_{s_j} = -\hat{\theta}_{s_j} \log \hat{\theta}_{s_j} - (1 - \hat{\theta}_{s_j}) \log(1 - \hat{\theta}_{s_j}).$$

It gives small values for the pure nodes, i.e., for those which consist prevalently from respondents or from non-respondents. The smallest possible D_{s_j} is searched for choosing the splitting point and the variable to split. In the next step the nodes from the first splitting are split again by a new choice of a variable and a splitting point. The procedure runs until some stopping rule is fulfilled. Sometimes it is a small value d of D_{s_j}. The new node is created if the within node deviance is at least d times of that of the root node. After stopping, in each final node the response proportions are computed and taken as the estimated response probabilities $\hat{\theta}_{k,tree}$ equal for each unit k in that node (or alternatively said, equal for a configuration of the auxiliary vector $\mathbf{x}$ realized for that node).

The tree method has some advantages compared to other modeling methods. It decides automatically which variables and which interactions enter the model.

4. Simulation set-up

In this section we study numerically downward calibration property of the linear, logistic, probit and decision tree models. In addition, we consider combined modeling, where the quantities $\hat{\theta}_{k,tree}$ estimated by the decision tree method are taken as covariates for the logistic regression model. Two cases for subsequent logistic regression are considered, $\hat{\theta}_{k,tree}$ as the only covariate, and $(\hat{\theta}_{k,tree}, \mathbf{x}'_k)$ as the covariate vector.

The six methods are compared in a simulation study. The downward calibration property is characterized by the calibration distance. The accuracy and relative bias of the estimated response propensities are also measured. 1000 response sets are simulated from a fixed sample s. In every simulation step (response set) i, $i = 1, 2, \ldots, 1000$, and for all methods we compute the following quantities:

- calibration distance $CD_i = \sqrt{\boldsymbol{v}'_i \Sigma_s^{-1} \boldsymbol{v}_i}$, where $\boldsymbol{v}_i = \sum_s d_k \hat{\theta}_k \mathbf{x}_k - \sum_r d_k \mathbf{x}_k$ and $\Sigma_s = \sum_s d_k \mathbf{x}_k \mathbf{x}'_k$ ($\boldsymbol{v}_i = \mathbf{0}$ for $\hat{\theta}_k$ with property (1));
- accuracy $Q_i = \sqrt{\frac{1}{n} \sum_s (\frac{\hat{\theta}_k - \theta_k}{\theta_k})^2}$;
- relative bias $RB_i = \frac{1}{n} \sum_s \frac{\hat{\theta}_k - \theta_k}{\theta_k}$.

For final illustration we use means and standard deviations of these measures over all simulations.

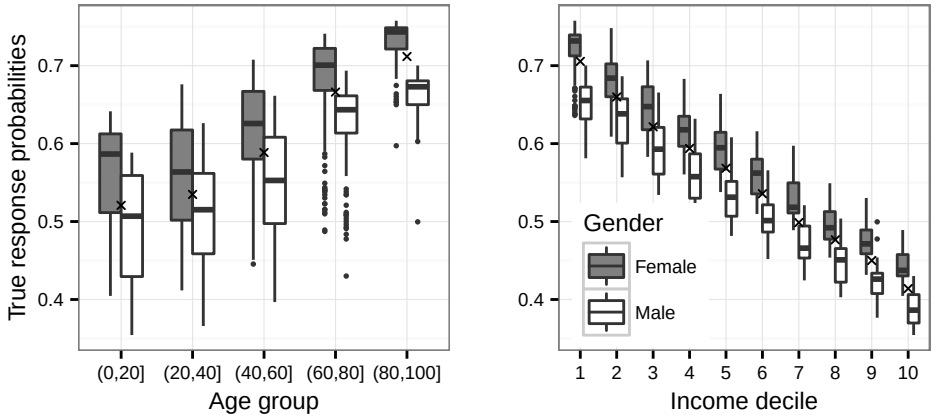
In the simulation experiment we use real Estonian data taken from the European Social Survey (see [3]). Specifically, our vector $\mathbf{x}_k$ consists of four variables: gender (coded as 1 for males and 2 for females), age (measured in full years), and total income of the individual k (expressed as a decile). Those individuals who did not respond to at least one variable were deleted from the data. The final sample consists of 1762 individuals. We assume here equal design-weights for every individual k.

We generate two sets of true response probabilities $\theta_k,\ k \in s$, using different logistic regression procedures. Below we refer to these sets as the linear and the nonlinear case. In obtaining the first set we relate the logit of primary response probabilities θ_k^*, $k \in s$, linearly on age, income and gender:

$$logit(\theta_k^*) = ln\left(\frac{\theta_k^*}{1-\theta_k^*}\right) = -0.1 \cdot income_k + 0.007 \cdot age_k + 0.2 \cdot gender_k.$$

Then, primary response probabilities θ_k^* are normalized and taken as values of θ_k so that the equality $\sum_s \theta_k = m$ holds, where m is the fixed size of the response set.

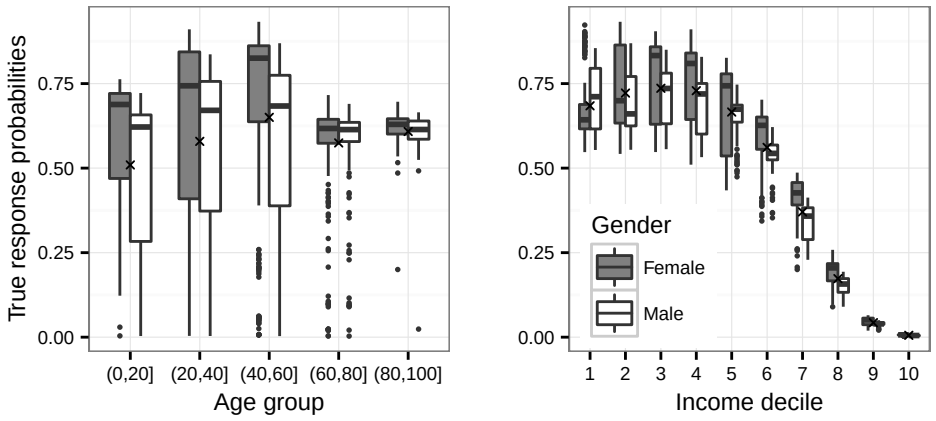
FIGURE 1. Linear case. Distribution of true response probabilities by age group, income decile, and gender.



The distribution of θ_k is depicted on Figure 1. We see that the response probability for female is generally higher than for male, marginal means being 0.63 vs 0.55. Younger people respond with lower probability than older, and average response probability (marked by cross) changes nearly linearly with age groups. Income is also related with response probability, but in decreasing manner: the higher income decile brings lower probability to respond.

The second set of true probabilities θ_k is generated by relating $logit(\theta_k^*)$ non-linearly on age, income and gender. With some manipulation the probabilities θ_k, as seen on Figure 2, are obtained.

FIGURE 2. Nonlinear case. Distribution of true response probabilities by age group, income decile, and gender.



On average (see Figure 2), the response probabilities are lower for the young and elderly people, but are higher for the middle age groups. Individuals with higher income respond with lower probability as for the first set, but the relationship between income and θ_k is non-linear. For this set, the average response probability is also higher for female than for male (0.63 vs 0.56).

Data is treated as a fixed sample s, for which response set of fixed size $m = 1057$ (approximately 60 % of a sample size) is generated in every simulation step i, $i = 1, 2, \ldots, 1000$. Every response set is drawn by the following order sampling method:

- the value u_k is generated for every individual $k \in s$:

$$u_k \sim \frac{U(0,\ 1)}{\theta_k};$$

- then the data is sorted into ascending order by u_k and the first $m = 1057$ units are coded by ones (respondents) and others by zeros (nonrespondents).

As a result, the unit k responds with probability θ_k^0, very close to θ_k. As shown by Rosén in [7], $\lim\limits_{m\to\infty} \dfrac{\theta_k}{\theta_k^0} \to 1$.

5. Simulation results

Simulation results are presented in two tables; for linearly constructed θ_k in Table 1, and for nonlinearly constructed θ_k in Table 2. A more detailed picture is presented on Figure 3. Simulation experiment confirmed that the linear regression and the logistic regression both satisfy the downward calibration property (1). We have $\boldsymbol{v}_i = \mathbf{0}$ for these methods in every simulated response set. Therefore average calibration distance $\overline{CD} = 0$ and standard deviation $s_{CD} = 0$. For the combined methods we also have $\overline{CD} = 0$ and $s_{CD} = 0$.

For probit and decision tree methods, the average calibration distance differs from zero, which means that downward calibration property does not hold for these methods. But for the probit model the distance is much smaller than for the decision tree model.

TABLE 1. Linear case. Means and standard deviations of performance measures.

Method	$\overline{Q}$	s_Q	$\overline{RB}$	s_{RB}	$\overline{CD}$	s_{CD}
Linear regression	0.040	0.0138	0	0.0032	0	0
Logistic regression	0.041	0.0139	0	0.0032	0	0
Probit	0.041	0.0139	0	0.0032	0.012	0.0047
Decision tree	0.236	0.0398	0.001	0.0033	0.443	0.2001
Logistic by $\hat{\theta}_{k,tree}$	0.231	0.0401	-0.001	0.0034	0	0
Logistic by $(\mathbf{x}_k, \hat{\theta}_{k,tree})$	0.236	0.0415	0	0.0033	0	0

As we see from Table 1, all six methods are nearly unbiased. The average relative bias of estimated propensities ($\overline{RB}$) is close to zero. The average accuracy ($\overline{Q}$) is near zero for the classical statistical methods. For the decision tree model and for its combinations with logistic regression, the mean accuracy is around 0.23, which says that these methods are 5 times less accurate than the first three methods, for the current set of true probabilities θ_k.

We see from Table 2 that linear, logistic, and probit methods are not able to capture nonlinear structure of response probabilities. The value of average accuracy ($\overline{Q}$) is quite large (around 5.2 – 5.6) for these methods. It

TABLE 2. Nonlinear case. Means and standard deviations of performance measures.

Method	$\overline{Q}$	s_Q	$\overline{RB}$	s_{RB}	$\overline{CD}$	s_{CD}
Linear regression	5.164	0.666	0.956	0.143	0	0
Logistic regression	5.644	0.562	1.083	0.112	0	0
Probit	5.423	0.588	1.033	0.118	0.141	0.017
Decision tree	1.116	0.793	0.041	0.077	0.360	0.161
Logistic by $\hat{\theta}_{k,tree}$	1.805	0.393	0.294	0.056	0	0
Logistic by $(\mathbf{x}_k, \hat{\theta}_{k,tree})$	1.793	0.401	0.291	0.057	0	0

is considerably smaller for the three last methods, where the decision tree model is used. The average relative bias is high, near one, for the classical methods, and it is close to zero for the decision tree method. Thus, decision tree method was the winner here to estimate response probabilities, in spite of the fact that the calibration property did not hold.

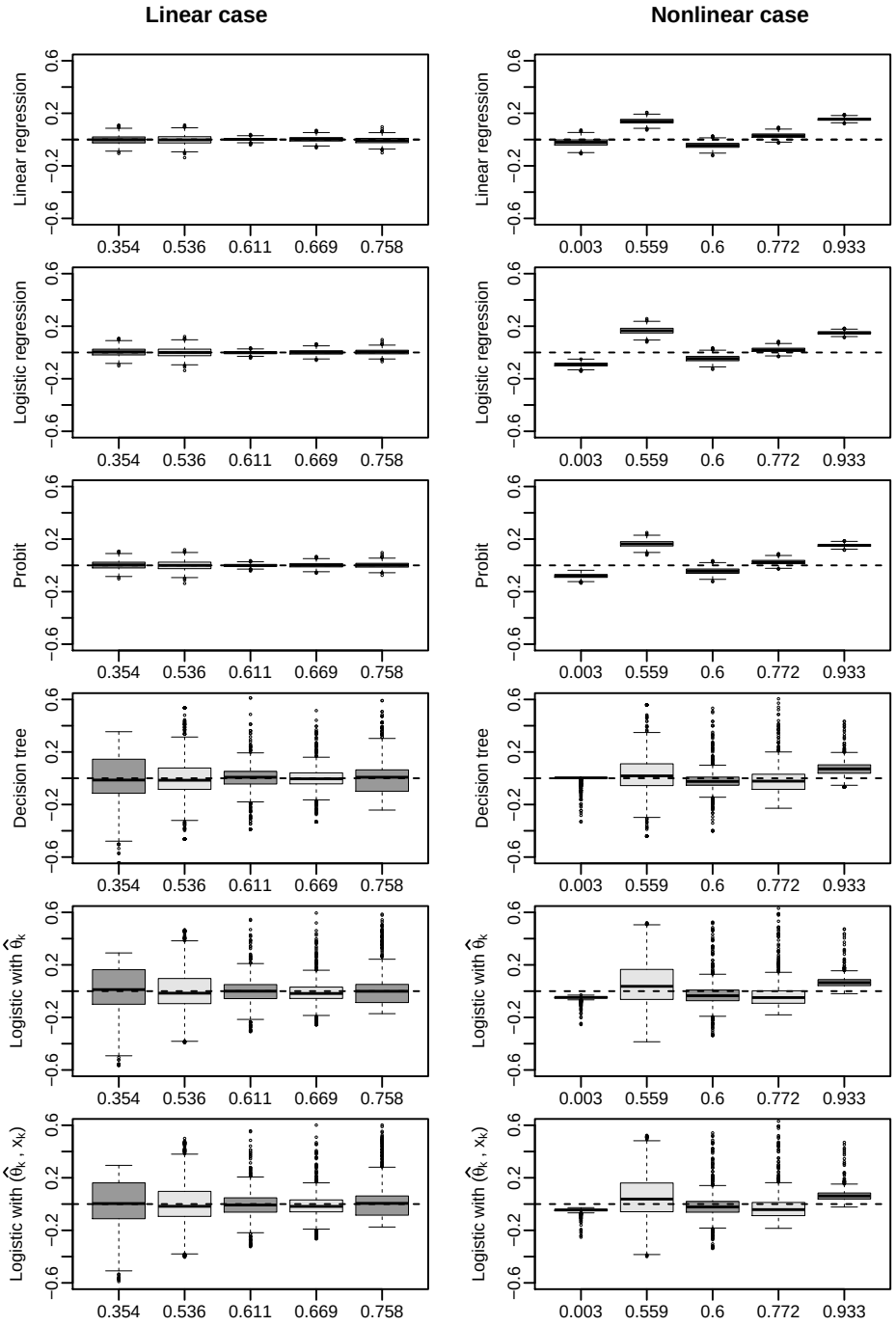
From both tables, we see that combined modeling, where estimates from the decision tree method $\hat{\theta}_{k,tree}$ are used as covariates for the logistic model, makes the downward calibration property to hold true, but does not guarantee more accurate propensity estimates. However, comparing the numbers s_Q, it makes propensity estimates more stable.

For all six methods, using 1000 simulated response sets, boxplots of differences $\theta_k - \hat{\theta}_k$ are presented for selected values of θ_k in Figure 3. The selected values are the five Tukey numbers (minimum, lower quartile, median, upper quartile, and maximum of $\theta_k,\ k \in s$). The horizontal dashed line is drawn at zero and corresponds to the ideal situation, where the true response probability and its estimate are equal.

On the left column of Figure 3, the true θ_k were nearly linearly related to the covariates. We see that in this case all 6 methods estimate unbiasedly θ_k for 5 selected values. The variation is smaller for classical methods.

On the right column of Figure 3, the true θ_k were nonlinearly related to the covariates. We notice that linear regression, logistic regression, and probit models are biased for all selected θ_k. This is in line with results in Table 2, where classical methods had large $\overline{Q}$ and $\overline{RB}$. In contrast to this, the tree method and its combinations perform with smaller biases, but with much higher variance.

FIGURE 3. Distribution of $\theta_k - \hat{\theta}_k$ over simulations for Tukey numbers of θ_k, $k \in s$, on horizontal axes.



6. Conclusions

In this paper we defined the downward calibration property of the estimated response probabilities $\hat{\theta}_k$. The estimates from linear regression are known to have this property. We showed analytically that logistic regression produces estimates with the downward calibration property as well, whereas the probit modeling does not do it. Simulation showed that the decision tree method does not do it either. However, in a special case (auxiliary vector being a group vector) all the methods produce the same $\hat{\theta}_k$, expressed as weighted response proportions in groups.

In a simulation experiment, the accuracy and bias of $\hat{\theta}_k$ with respect to the true θ_k was measured for all the considered methods. This was done under two different response mechanisms. For the linear response mechanism, the classical methods (linear, logistic and probit) performed better than the tree method, whereas for the nonlinear case, the tree method was much less biased and had much better accuracy. We learned that the downward calibration property itself does not guarantee more accurate propensity estimation. The accuracy and bias of the estimation method depend on the relationship nature between true response probabilities and auxiliary variables. Our Figure 3 shows a detailed behavior of $\hat{\theta}_k$ for selected values of the true θ_k for all the methods and for both response mechanisms. An important issue for the future research is the effect of the downward calibration property on the estimator $\hat{\hat{t}} = \sum_r d_k \hat{\theta}_k^{-1} y_k$.

Acknowledgements

This work was supported by the Institutional Research Funding IUT34-5 of Estonia. The authors are thankful for the constructive and competent comments made by the anonymous referee.

References

[1] A. Agresti, *Categorical Data Analysis*, Wiley, New Jersey, 2013.
[2] J. M. Brick, *Unit nonresponse and weighting adjustments: a critical review*, J. Off. Stat. **29** (2013), 329–353.
[3] European Social Survey (2014). *http://www.europeansocialsurvey.org*
[4] T. Hastie, R. Tibshirany, and J. Friedman, *The Elements of Statistical Learning; Data Mining, Inference and Prediction*, Springer, New York, 2011.
[5] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Application in R*, Springer, New York, 2013.
[6] P. P. Phipps and D. Toth, *Analyzing establishment nonresponse using interpretable regrssion tree model with linked administrative data*, Ann. Appl. Stat. **6** (2012), 772–794.
[7] B. Rosén, *On inclusion probabilities for order πps sampling*, J. Statist. Plann. Inference **90** (2000), 117–143.

[8] C. E. Särndal and S. Lundström, *Estimation in Surveys with Nonresponse*, Wiley, Chichester, 2006.
[9] C. E. Särndal, *The 2010 Morris Hansen lecture dealing with survey nonresponse in data collection, in estimation*, J. Off. Stat. **27** (2011), 1–21.
[10] B. Schouten, N. Shlomo, and C. Skinner, *Indicators for monitoring and improving representativeness of response*, J. Off. Stat. **27** (2011), 1–24.
[11] R. Valliant, J. A. Dever, and F. Kreuter, *Practical Tools for Designing and Weighting Survey Samples*, Springer, New York, 2013.

Institute of Mathematics and Statistics, University of Tartu, J. Liivi 2, 50409 Tartu, Estonia
E-mail address: natalja.lepik@ut.ee
E-mail address: imbi.traat@ut.ee