SORT 46 (1) January-June 2022, 137-158 DOI: 10.2436/20.8080.02.121

Estimation of cut-off points under complex-sampling design data

Amaia Iparragirre*[,1], Irantzu Barrio[1,3], Jorge Aramendi[2]
and Inmaculada Arostegui[1,3]

Abstract

In the context of logistic regression models, a cut-off point is usually selected to dichotomize the estimated predicted probabilities based on the model. The techniques proposed to estimate optimal cut-off points in the literature, are commonly developed to be applied in simple random samples and their applicability to complex sampling designs could be limited. Therefore, in this work we propose a methodology to incorporate sampling weights in the estimation process of the optimal cut-off points, and we evaluate its performance using a real data-based simulation study. The results suggest the convenience of considering sampling weights for estimating optimal cut-off points.

MSC: *62J12, 62P25, 62D05*

Keywords: *Optimal cut-off points, complex survey data, sampling weights.*

1. Introduction

Survey data are gaining popularity in a number of fields, including but not limited to, social and health sciences. This type of data is data collected from a finite population, concerned to be studied, by some complex sampling design such as stratification or clustering, among others (Kalton, 1983). One of the differences between complex survey data and simple random samples is that, in the first, each sampled observation has assigned a sampling weight, which indicates the number of units that this observation represents in the finite population. Therefore, the straightforward application of the most

[*] *Corresponding author:* E-mail: amaia.iparragirre@ehu.eus, Address: Departamento de Matemáticas. Facultad de Ciencia y Tecnología. Universidad del País Vasco (UPV/EHU). Barrio Sarriena s/n. 48940 Leioa.

[1] Departamento de Matemáticas, Universidad del País Vasco (UPV/EHU).

[2] Eustat - Euskal Estatistika Erakundea - Instituto Vasco de Estadística.

[3] BCAM - Basque Center for Applied Mathematics, Bilbao, Spain.

Received: September 2021.
Accepted: May 2022.

commonly applied statistical techniques, which are typically designed to be applied to simple random samples, is usually not suitable for complex survey data (Skinner, Holt and Smith, 1989).

In this paper, we focus on the particular case of a binary response variable Y and, specifically, on the logistic regression model to predict Y according to a collection of covariates whose distribution may be discrete or continuous. From a practical point of view, one of the most important characteristics of this kind of model is the support they provide for decision-making, since increasing knowledge about potential predictors helps the decision-making process (Steyerberg, 2008; Baker and Gerdin, 2017). In this context, decisions such as whether or not to recommend a patient to start treatment, or to give a diagnosis about a disease, are based on the individual risk (probability) of event given by the estimates of the logistic regression model. In order to make these decisions, first, for each individual, the predicted probability of event is classified based on a cut-off point. In this way, for example, if the individual's probability of suffering from extreme poverty is greater than the selected cut-off point, he or she is assigned a social benefit, while in contrast, if that is lower no social support is provided (Steyerberg, 2008; Pauker and Kassirer, 1980). Hence, cut-off point estimation is widely employed in practice, in the field of prediction models, especially, but not exclusively, in clinical prediction models (Steyerberg et al., 1999; Chen et al., 2015; Spence et al., 2018).

At this point, the main issue is usually to select a valid cut-off point that will provide the best classification of individuals in practice. Many strategies have been proposed in the literature in order to estimate optimal cut-off points. It should be noted that we can not talk about optimal cut-off points in general terms. In contrast, a cut-off point will or will not be the optimal depending on the objective of a particular study. Therefore, when we talk about selecting an optimal cut-off point, we are talking about selecting the one which satisfies a certain optimality criterion. Hence, as we have mentioned above, different techniques have been proposed to select optimal cut-off points, given a particular criterion. For instance, some of those methods select the optimal cut-off point with the aim of obtaining a certain value of sensitivity/specificity (i.e., probability of classifying correctly an individual with/without the event of interest) or to maximize a function of these two parameters as for example the Youden index (Youden, 1950). Some others select the cut-off point that maximizes some particular indexes, such as Kappa (Cohen, 1960; Greiner, Pfeiffer and Smith, 2000). Greiner (1995, 1996) proposed a method to select the optimal cut-off point that minimizes the error or either maximizes the accuracy of the classification rule. There are some other methods that select optimal cut-off points based on some other criteria related to several parameters such as predicted values (i.e., probability of event/non-event for an individual classified as event/non-event) (Vermont et al., 1991) or prevalence (i.e., the probability of event in the population) (Manel, Williams and Ormerod, 2001), among others. Besides, other methods are based on the analysis of the cost of incorrect and the benefit of correct diagnosis (Swets, 1992; Pauker and Kassirer, 1980; Wynants et al., 2019). An extensive review of those techniques can be found in López-Ratón et al. (2014).

However, those techniques have usually been designed and applied for simple random samples and, as far as we know, there is a lack of proposals to consider complex sampling designs, and in particular sampling weights, throughout the estimation process of optimal cut-off points. It is widely known that when the sampling designs are not considered for the analysis of data derived from complex surveys the variances tend to be underestimated, which can lead to biased estimates of test statistics (Yao, Li and Graubard, 2015; Skinner et al., 1989; Heeringa, West and Berglund, 2017; Binder and Roberts, 2009). In the same way, we believe that sampling weights should not be ignored when estimating optimal cut-off points when working with complex survey data. Therefore, in this work, we propose a methodology to modify some of the methods to select optimal cut-off points of the probability of event in the logistic regression framework that have been previously proposed in the literature, so that they take into account sampling weights in the estimation process. In addition, the performance of the proposed methods is compared to the performance of those which ignore the sampling weights, by means of a simulation study. In particular, we focus on surveys which are based on one-step stratified samples.

The rest of the paper is organized as follows. Section 2 describes the real survey that has motivated this work. Section 3 defines some basic notation that will be used along the rest of the paper. Furthermore, we describe some of the methods that are usually applied in practice to estimate optimal cut-off points of the probability of event in the logistic regression framework and finally we propose a new methodology which takes into account the effect of the sampling weights in the cut-off point estimation process. In Section 4, we describe the simulation process that has been carried out so as to study the performance and effectiveness of the proposed method to incorporate sampling weights into the estimation process of optimal cut-off points and we show the results we have obtained in the mentioned simulation study. The methodology proposed in this work has been applied to real survey data and this application is described in Section 5. Finally, we conclude with a discussion in Section 6.

2. Motivating data set

This work has been motivated by the Survey on the Information Society in Companies[1], which has been designed, conducted and collected by the Official Statistics Basque Office (Eustat). This survey, which is usually denoted as ESIE survey due to its Spanish acronym, is carried out annually among the companies in the Basque Country (BC) in order to collect information about the implementation of New Information and Communication Technology in the companies of the BC. In particular, the information considered in this study is related to the survey carried out in 2010.

The finite population is defined by a total of 14 200 companies, all of which have at least 10 employees. From this population a sample of 2 852 was obtained by means

[1] https://en.eustat.eus/estadisticas/tema_150/opt_1/tipo_7/temas.html

of one-step stratified sampling technique with simple random sampling in each stratum. Strata are defined by means of the combination of three categorical variables: the province where the company is located (3 categories), activity of the company (65 categories) and the number of employees (2 categories). In this way, a total of 390 different strata have been defined. However, it should be noted that in some of these strata there are no units in the population, so in fact we have 325 strata in total ($h = 1,\ldots,H$, where $H = 325$). Once the sample is obtained, a sampling weight is assigned to the companies sampled in each stratum. The sampling weight (w_i, $\forall i \in S$) is calculated per stratum as the total number of companies in the finite population of the stratum (let us denote it as N_h, $\forall h \in \{1,\ldots,H\}$) divided by the number of companies sampled in that stratum (denoted as n_h, $\forall h \in \{1,\ldots,H\}$). In other words, for a unit i sampled from stratum h its sampling weight is computed as follows:

$$w_i = \frac{N_h}{n_h}, \quad \forall i \in S. \tag{1}$$

Each sampling weight indicates the number of companies that this sampled company represents in the finite population.

In the survey data considered for this paper, strata sizes in the finite population (i.e., N_h, $\forall h \in \{1,\ldots,H\}$) ranges from 1 to 860, where the median is 12 and the interquartile range $4-44$. An unequal probability sampling design has been applied in the sampling process, in which the probabilities of being sampled from each stratum (i.e., n_h/N_h, $\forall h \in \{1,\ldots,H\}$) range from 0.0391 to 1 (with a median of 0.6667 and an interquartile range of $0.2604-1$). The dichotomous response variable considered for this work indicates whether a company has its own website (1) or not (0). The probability of event in the sample (without considering the sampling weights) is 0.8222, while the weighted estimate of the probability of event (computed by taking into account the number of units that each element represents in the finite population by means of the sampling weights w_i, $\forall i \in S$) is 0.7544.

3. Methods

In this section, first of all, we introduce the basic notation that we will use throughout this document. In addition, we describe some of the methods that are usually applied for estimating optimal cut-off points in this context based on different optimality criteria for simple random samples. Finally, we develop a new estimation method, in which we propose to introduce the sampling weights in these methods so that they are valid in complex design samples.

3.1. Basic notation and preliminaries

Let $\boldsymbol{X} = (X_1,\ldots,X_p)^{\intercal}$ be a vector of p random predictor variables denoting the covariates and Y a random variable denoting the dichotomous response variable. Without loss of generality, and in order to ease the notation, suppose that the covariates $\boldsymbol{X}$ are continuous and the response variable Y takes the value 1 to represent the event or the presence of the characteristic of interest, and 0 otherwise. Let $P(Y=1|\boldsymbol{X})$ represent the conditional probability of event given the vector of covariates $\boldsymbol{X}$. Then, the linear form of the logistic regression model for Y is written as follows:

$$\text{logit}\,(P(Y=1|\boldsymbol{X})) = \ln\left[\frac{P(Y=1|\boldsymbol{X})}{1-P(Y=1|\boldsymbol{X})}\right] = \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}, \tag{2}$$

being $\boldsymbol{\beta} = (\beta_0,\ldots,\beta_p)^{\mathsf{T}}$ the vector of regression coefficients.

Consider $U=\{1,\ldots,N\}$ a finite population of N units. In the context of complex survey data, let S be a sample of n units drawn from the finite population by some complex sampling design. To each sampled observation $i \in S$, a set of values $(y_i,\boldsymbol{x}_i,w_i)$ is associated where each sampling weight w_i indicates the number of units that $i \in S$ represents in the finite population (note that $\sum_{i\in S} w_i = N$) and y_i and $\boldsymbol{x}_i$ indicate the realizations of the variables Y and $\boldsymbol{X}$ for the sampled units, respectively. For each $i \in S$ let us define its probability of event as $p(\boldsymbol{x}_i) = P(Y=1|\boldsymbol{X}=\boldsymbol{x}_i)$, which can be estimated as follows:

$$\hat{p}(\boldsymbol{x}_i) = \frac{e^{\hat{\boldsymbol{\beta}}^{\mathsf{T}}\boldsymbol{x}_i}}{1+e^{\hat{\boldsymbol{\beta}}^{\mathsf{T}}\boldsymbol{x}_i}} \quad (i \in S), \tag{3}$$

where the estimated regression coefficients $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0,\ldots,\hat{\beta}_p)^{\mathsf{T}}$, are usually obtained by maximizing the weighted pseudo-likelihood function, defined as (Binder, 1981, 1983):

$$PL(\boldsymbol{\beta}) = \prod_{i\in S} p(\boldsymbol{x}_i)^{y_i w_i}(1-p(\boldsymbol{x}_i))^{(1-y_i)w_i}. \tag{4}$$

3.2. Optimal cut-off point estimation methods

It is usually very useful in practice to select a cut-off point in order to distinguish between units with and without the event of interest. In our particular case, we are interested in discriminating between units with and without the event of interest based on their estimated probability of event. In this context, one observation $i \in S$ is usually classified as event if its estimated probability of event exceeds a determined threshold c which has been previously selected (Magder and Fix, 2003; Pepe, 2003). The correct classification of an observation with the event of interest is usually denoted as *true positive* (TP), while the correct classification of an observation without the event of interest is commonly denoted as *true negative* (TN). But usually, those classifications are not entirely accurate. Therefore, some of the observations are commonly classified incorrectly: an observation with the event of interest may be classified as non-event (*false negative* (FN)) or an observation without the event of interest may be classified as event (*false positive* (FP)).

Methods of estimation of the optimal cut-off point have been developed in the literature, with the aim of optimizing diverse measures. In particular, many methods consist on the optimization of an objective function of the Receiver Operating Characteristic (ROC) curve, which is a curve that describes the global accuracy of a model (Bamber, 1975; Pepe, 2003). Coming back to our particular case, taking into account that the predicted probabilities range from 0 to 1, the ROC curve of a logistic regression model can be defined as follows (Hosmer and Lemeshow, 2000; Pepe, 2003):

$$ROC(\cdot) = \{(1 - Sp(c), Se(c)),\ c \in (0,1)\}, \tag{5}$$

where $Se(c)$ and $Sp(c)$ are defined as follows and denoted as sensitivity and specificity, respectively:

$$\begin{array}{l} Se(c) = P\left[P(Y=1|\boldsymbol{X}) \geq c|Y=1\right], \\ Sp(c) = P\left[P(Y=0|\boldsymbol{X}) < c|Y=0\right]. \end{array} \tag{6}$$

In practice, following the notation defined so far, assume that to each sampled observation $i \in S$ a set of values $(y_i, \boldsymbol{x}_i, w_i)$ is associated. Suppose that the vector $\hat{\boldsymbol{\beta}}$ is obtained by means of the pseudo-likelihood function in (4) and $\hat{p}(\boldsymbol{x}_i)$ are estimated for $i \in S$ following (3). Let us define the following groups of correctly or incorrectly classified observations, for a specific cut-off point c:

$$\begin{array}{ll} TP_c = \{i \in S : y_i = 1 \text{ and } \hat{p}(\boldsymbol{x}_i) \geq c\}, & TN_c = \{i \in S : y_i = 0 \text{ and } \hat{p}(\boldsymbol{x}_i) < c\}, \\ FP_c = \{i \in S : y_i = 0 \text{ and } \hat{p}(\boldsymbol{x}_i) \geq c\}, & FN_c = \{i \in S : y_i = 1 \text{ and } \hat{p}(\boldsymbol{x}_i) < c\}. \end{array} \tag{7}$$

In addition, let us define an indicator function associated to each of the sets defined in (7) as follows. For example, for the set TP_c:

$$1_{TP_c}(i) = \left\{ \begin{array}{lll} 1 & \text{if} & i \in TP_c, \\ 0 & \text{if} & i \notin TP_c. \end{array} \right. \tag{8}$$

In the same way, indicator functions can be defined as in (8) for the rest of the sets described in (7), which will be denoted as $1_{TN_c}(i)$, $1_{FP_c}(i)$ and $1_{FN_c}(i)$, hereinafter. Then, for a specific cut-off point c, sensitivity and specificity parameters can be estimated based on sample S as follows:

$$\widehat{Se}(c) = \frac{\sum_{i \in S} 1_{TP_c}(i)}{\sum_{i \in S}\left[1_{FN_c}(i) + 1_{TP_c}(i)\right]}, \quad \widehat{Sp}(c) = \frac{\sum_{i \in S} 1_{TN_c}(i)}{\sum_{i \in S}\left[1_{TN_c}(i) + 1_{FP_c}(i)\right]}. \tag{9}$$

For this study, we have selected some of those methods which are based on several optimality criteria related to sensitivity and specificty parameters:

- *Youden* (Youden, 1950; Greiner et al., 2000): This method selects the cut-off point (c^{Youden}) that maximizes the Youden Index, which is defined as the sum of sensitivity and specificity parameters minus one, i.e.,

$$c^{\text{Youden}} = \underset{c \in (0,1)}{\operatorname{argmax}} \left\{ \widehat{Se}(c) + \widehat{Sp}(c) - 1 \right\}. \tag{10}$$

- *MaxProdSpSe* (Lewis et al., 2008): This method selects the cut-off point c that maximizes the product between sensitivity and specificity parameters, i.e.,

$$c^{\text{MaxProdSpSe}} = \underset{c \in (0,1)}{\operatorname{argmax}} \left\{ \widehat{Se}(c) \cdot \widehat{Sp}(c) \right\}. \tag{11}$$

- *ROC01* (Metz, 1978; Vermont et al., 1991): This method selects the cut-off point c that minimizes the distance between the ROC curve and the point (0,1), i.e.,

$$c^{\text{ROC01}} = \underset{c\in(0,1)}{\text{argmin}} \left\{ (\widehat{Se}(c)-1)^2 + (\widehat{Sp}(c)-1)^2 \right\}. \tag{12}$$

- *MaxEfficiency* (Greiner, 1995, 1996): This method selects the cut-off point c that maximizes the efficiency or, in other words, minimizes the error, i.e.,

$$c^{\text{MaxEff}} = \underset{c\in(0,1)}{\text{argmax}} \left\{ \widehat{p}_Y \widehat{Se}(c) + (1-\widehat{p}_Y)\widehat{Sp}(c) \right\}, \tag{13}$$

  where $\widehat{p}_Y$ is the estimated prevalence which is calculated as follows:

$$\widehat{p}_Y = \frac{1}{n}\sum_{i\in S}\left[1_{FN_c}(i) + 1_{TP_c}(i)\right]. \tag{14}$$

3.3. Cut-off point estimation proposal with sampling weights

Although sensitivity and specificity parameters, as well as the prevalence, can be estimated by expressions (9) and (14) in any kind of data, including complex survey data, these expressions have been defined in a simple random sampling scenario. However, in complex survey data each of the sampled units has a sampling weight associated, which indicates the importance of each of them within the sample. Thus, the influence of all sampled units is not uniform. Therefore, we believe that the estimates obtained by means of the above-mentioned formulas may be misleading for complex survey data and they should be pondered, so that they incorporate the sampling weights. In this way, instead of the number of correct or incorrect classifications in sample S, it should be considered the number of units that these correctly or incorrectly classified observations represent in the finite population. For this reason, we propose to consider the sampling weights w_i to estimate sensitivity $(\widehat{Se}_w(c))$ and specificity $(\widehat{Sp}_w(c))$ parameters as follows:

$$\widehat{Se}_w(c) = \frac{\sum_{i\in S} w_i \cdot 1_{TP_c}(i)}{\sum_{i\in S} w_i \cdot [1_{FN_c}(i) + 1_{TP_c}(i)]}, \quad \widehat{Sp}_w(c) = \frac{\sum_{i\in S} w_i \cdot 1_{TN_c}(i)}{\sum_{i\in S} w_i \cdot [1_{TN_c}(i) + 1_{FP_c}(i)]}. \tag{15}$$

where the indicator functions are the ones described in (8).

In addition, note that sampling weights should also be considered to estimate the prevalence $(\widehat{p}_{Y,w})$:

$$\widehat{p}_{Y,w} = \frac{1}{N}\sum_{i\in S} w_i \cdot \left[1_{FN_c}(i) + 1_{TP_c}(i)\right]. \tag{16}$$

Therefore, we propose to estimate the optimal cut-off points based on the modified parameters of sensitivity $(\widehat{Se}_w(c))$ and specificity $(\widehat{Sp}_w(c))$ when working with complex survey data, i.e.:

$$c_w^{\text{Youden}} = \underset{c\in(0,1)}{\text{argmax}} \left\{ \widehat{Se}_w(c) + \widehat{Sp}_w(c) - 1 \right\}, \tag{17}$$

$$c_w^{\text{MaxProdSpSe}} = \underset{c\in(0,1)}{\text{argmax}} \left\{ \widehat{Se}_w(c) \cdot \widehat{Sp}_w(c) \right\}, \tag{18}$$

$$c_w^{\text{ROC01}} = \underset{c\in(0,1)}{\text{argmin}} \left\{ (\widehat{Se}_w(c)-1)^2 + (\widehat{Sp}_w(c)-1)^2 \right\}, \tag{19}$$

$$c_w^{\text{MaxEff}} = \underset{c\in(0,1)}{\text{argmax}} \left\{ \widehat{p}_{Y,w}\widehat{Se}_w(c) + (1-\widehat{p}_{Y,w})\widehat{Sp}_w(c) \right\}. \tag{20}$$

4. Simulation study

This section describes the simulation process developed in this work and the scenarios that have been drawn. The results obtained in this simulation study are also presented in this section.

As stated above, the aim of this work is to study the influence of sampling weights in the estimation process of optimal cut-off points for the methods described in Section 3.2. Since the decision of which optimal cut-off point estimation method to use in practice depends on the research of interest, the objective of this work is not to compare the behaviour of the methods among them, but to compare the estimates that we obtain for each of these methods when sampling weights are considered or not in the estimation of sensitivity and specificity parameters.

In addition, we study the impact that the proposed estimators have in the estimation of the probability of event in the finite population. Therefore, a theoretical finite population is required, in which the response variable is known for all the units in the finite population. Thus, a pseudo-population has been generated based on real survey data. The real survey on which this pseudo-population is based is described in Section 2 and the process followed to generate it is explained in detail in Appendix A. The pseudo-population sampling process, which is replicated several times in the simulation study, is also based on the same real-life survey. This sampling process is described in Appendix B.

4.1. Scenarios and set up

Let $U = \{1,\ldots,N\}$ be the pseudo-population generated by following the steps described in Appendix A to which $\{(y_i,\boldsymbol{x}_i)\}_{i=1}^N$ are assigned. From this pseudo-population, a total of $R = 500$ samples have been obtained and the sampling weights have been assigned to the sampled units by the sampling process described in Appendix B. The optimal cut-off points estimation methods that have been applied in this study are the ones described in Section 3.2, i.e., $m \in \{$Youden, MaxProdSpSe, ROC01, MaxEfficiency$\}$.

The steps that have been followed in the simulation study are described below. For $r = 1,\ldots,500$:

Step 1. Draw a sample $S^r \subset U$ by one-step stratification with simple random sampling without replacement in each stratum (Appendix B, mimicking the sampling process carried out for the real-life dataset described in Section 2).

Step 2. Fit the logistic regression model to S^r and estimate $\hat{\boldsymbol{\beta}}^r$ by (4).

Step 3. For $i \in S^r$, estimate $\hat{p}^r(\boldsymbol{x}_i)$ by means of $\hat{\boldsymbol{\beta}}^r$ following (3).

Step 4. Estimate the optimal cut-off points, $c^{m,r}$ (see (10), (11), (12), (13)) and $c_w^{m,r}$ (see (17), (18), (19), (20)) for each method m.

As mentioned above, the selection of the optimality criteria for selecting the cut-off points is based on the particular goal of each study. Therefore, our goal is not to compare the performance of the described methods between them. That is, the aim is not to compare the performance of a method $m \in \{$Youden, MaxProdSpSe, ROC01, MaxEfficiency$\}$, to the rest of the methods, but to compare the cut-off points selected by means of the method m when sampling weights are considered or not in the estimation process. Thus, we define the difference and absolute difference between weighted and unweighted cut-off points as follows:

$$\text{Diff}^{\,m,r} = c^{m,r} - c_w^{m,r} \quad \text{and} \quad \text{AbsDiff}^{\,m,r} = |c^{m,r} - c_w^{m,r}|\,. \tag{21}$$

In addition, we would also like to regard the impact that may have the decision to select weighted or unweighted optimal cut-off points in the classification of all the units in the finite population. Thus, we continue with the simulation study as follows:

Step 5. For $i = 1, \dots, N$ calculate $\hat{p}^r(\boldsymbol{x}_i)$ by means of $\hat{\boldsymbol{\beta}}^r$ (**Step 3.**) following (3).

Step 6. For $i = 1, \dots, N$ classify each unit as event or non-event based on $\hat{p}^r(\boldsymbol{x}_i)$. Define two estimated responses ($\hat{y}_i^{m,r}$ and $\hat{y}_{w,i}^{m,r}$) for each unit based on the cut-off points $c^{m,r}$ and $c_w^{m,r}$ (selected in **Step 4.**) as follows. For each method m and $i = 1, \dots, N$:

$$\hat{y}_i^{m,r} = \begin{cases} 1 & \text{if} \quad \hat{p}^r(\boldsymbol{x}_i) \geq c^{m,r}, \\ 0 & \text{if} \quad \hat{p}^r(\boldsymbol{x}_i) < c^{m,r}, \end{cases} \quad \text{and} \quad \hat{y}_{w,i}^{m,r} = \begin{cases} 1 & \text{if} \quad \hat{p}^r(\boldsymbol{x}_i) \geq c_w^{m,r}, \\ 0 & \text{if} \quad \hat{p}^r(\boldsymbol{x}_i) < c_w^{m,r}. \end{cases}$$

Finally, in order to account for the error that may be introduced in the classification of the units in the finite population by the selected optimal cut-off points, one more parameter is defined. The error is estimated by comparing the prevalence estimated by means of the estimated responses (**Step 6**) to the true prevalence in the finite population. We split the finite population U in K disjointed subsets of the same size where $U = U_1 \cup \ldots \cup U_K$. We repeat this process $L = 10$ times, where for each $l = 1, \dots, L$, $U = U_1^l \cup \ldots \cup U_K^l$. In this way, we get $L \times K$ subsets from U and the prevalence will be estimated in each one of these subsets. Let us define the following indicator functions:

$$1_{U_k^l}(i) = \begin{cases} 1 & \text{if} \quad i \in U_k^l, \\ 0 & \text{if} \quad i \notin U_k^l, \end{cases} \quad \text{for} \quad l = 1, \dots, L \quad \text{and} \quad k = 1, \dots K. \tag{22}$$

We denote as global mean squared error (GMSE) of the prevalence with $L = 10$ replicates the following parameters:

$$\begin{aligned}
\text{GMSE}^{m,r} &= \frac{1}{L \times K} \sum_{l=1}^{L} \sum_{k=1}^{K} \left(\frac{\sum_{i=1}^{N} \hat{y}_i^{m,r} \cdot 1_{U_k^l}(i)}{\sum_{i=1}^{N} 1_{U_k^l}(i)} - \frac{\sum_{i=1}^{N} y_i \cdot 1_{U_k^l}(i)}{\sum_{i=1}^{N} 1_{U_k^l}(i)} \right)^2, \\
\text{GMSE}_w^{m,r} &= \frac{1}{L \times K} \sum_{l=1}^{L} \sum_{k=1}^{K} \left(\frac{\sum_{i=1}^{N} \hat{y}_{w,i}^{m,r} \cdot 1_{U_k^l}(i)}{\sum_{i=1}^{N} 1_{U_k^l}(i)} - \frac{\sum_{i=1}^{N} y_i \cdot 1_{U_k^l}(i)}{\sum_{i=1}^{N} 1_{U_k^l}(i)} \right)^2.
\end{aligned} \tag{23}$$

Different number of subsets have been selected in order to evaluate the impact the sample size of each subset may have: $K \in \{1, 10, 100, 500\}$. In addition, we considered the GMSE evaluated considering the H strata as the subsets where U_h, $\forall h = 1, \ldots, H$ indicates the subset corresponding to stratum h and $U = \bigcup_{h=1}^{H} U_h$:

$$\begin{aligned}
\text{GMSE}_h^{m,r} &= \frac{1}{H} \sum_{h=1}^{H} \left(\frac{\sum_{i=1}^{N} \hat{y}_i^{m,r} \cdot 1_{U_h}(i)}{\sum_{i=1}^{N} 1_{U_h}(i)} - \frac{\sum_{i=1}^{N} y_i \cdot 1_{U_h}(i)}{\sum_{i=1}^{N} 1_{U_h}(i)} \right)^2, \\
\text{GMSE}_{w,h}^{m,r} &= \frac{1}{H} \sum_{h=1}^{H} \left(\frac{\sum_{i=1}^{N} \hat{y}_{w,i}^{m,r} \cdot 1_{U_h}(i)}{\sum_{i=1}^{N} 1_{U_h}(i)} - \frac{\sum_{i=1}^{N} y_i \cdot 1_{U_h}(i)}{\sum_{i=1}^{N} 1_{U_h}(i)} \right)^2,
\end{aligned} \tag{24}$$

where,

$$1_{U_h}(i) = \begin{cases} 1 & \text{if} \quad i \in U_h, \\ 0 & \text{if} \quad i \notin U_h, \end{cases} \qquad \text{for} \quad h = 1, \ldots, H. \tag{25}$$

This simulation study has been carried out by means of the statistical software R. In particular, some functions of the R package `OptimalCutpoints` (López-Ratón et al., 2014) have been modified in order to incorporate an argument that provides us with the option to consider sampling weights in the estimation process of the optimal cut-off points for the described methods.

4.2. Results

In this Section we show the results obtained in the simulation study described in Section 4.1. Figures 1, 2, 3 and 4 depict the box-plots of unweighted and weighted estimates of the optimal cut-off points and the results of the parameters Diff and GMSE (see (21) and (23)) for Youden, MaxProdSpSe, ROC01 and MaxEfficiency methods, respectively. Numerical results of the simulation study are summarized in Table 1.

In general, except for the MaxEfficiency method, the results suggest that the optimal cut-off point estimates differ when sampling weights are ignored or considered in the estimation process. The difference has always been positive (i.e. the unweighted estimates have been greater than the weighted ones), except in the MaxEfficiency method where both positive and negative differences have been observed. For this reason, the mean and standard deviation of the difference and absolute difference parameters are equal for all the methods except for MaxEfficiency (see Table 1). The error generated and accounted in terms of GMSE described in (23) decreases considerably when sampling

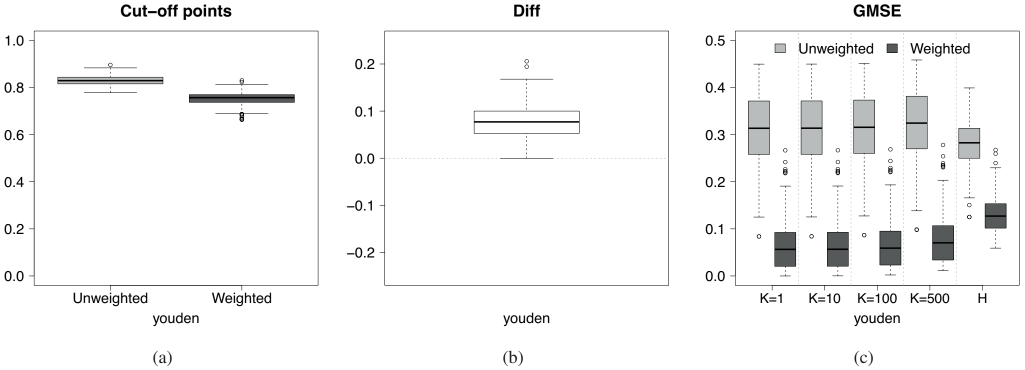
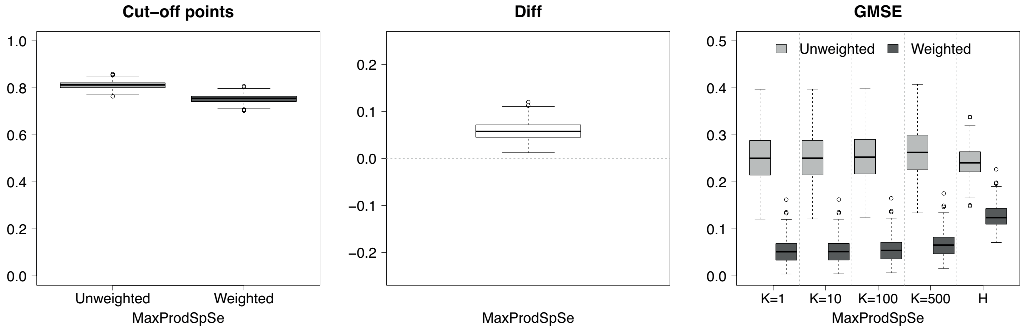


Figure 1. *Box-plots of the results obtained for the Youden method across* $R = 500$ *samples: (a) unweighted and weighted estimates of the optimal cut-off points, (b) differences between unweighted and weighted estimates (Diff), and (c) GMSE produced by the unweighted and weighted estimates for* $K \in \{1, 10, 100, 500\}$ *and* H*.*

weights are taken into account. In addition, similar results have been obtained for different $K \in \{1, 10, 100, 500\}$ values, which indicates that the difference between estimated and true prevalence is similar in smaller homogeneous subsets and in the total population. However, it could be observed that the average of GMSE becomes slightly greater as the number of subsets K increases (for both, weighted and unweighted estimates), indicating that the differences between the estimated and true prevalence tend to be a little bit greater in smaller subsets. When considering the strata as non-homogeneous subsets defined by the H strata of the population, the GMSE obtained as described in (24) with the weighted estimates is still smaller than with the unweighted ones. However, the difference between weighted and unweighted GMSE is slightly smaller for the non-homogenous partition than for homogeneous partitions. We believe that the reason is that the difference obtained between estimated and true prevalence differs depending on the number of individuals sampled in each strata, being increased in very small strata. Note that if the population size of a particular stratum is 1 then the error in this stratum is 0 (if the unit is classified correctly) or 1 (otherwise). This is not common when working with homogeneous strata where in all the randomly selected subsets the difference between estimated and true prevalence seem to be similar (results not shown). In addition, note that even though strata are of different sizes, the stratum size is not taken into account when computing the GMSE parameter. Below, the behaviour of each of the methods that have been studied throughout this work will be analysed one by one.

The optimal cut-off point estimated by the Youden method in this simulation study, is 0.8304 on average when sampling weights are not taken into account while the weighted estimates are smaller on average (0.7524), with standard deviations of 0.0208 and 0.0277, respectively. The difference among the unweighted and weighted estimates is on average 0.0780 with a standard deviation of 0.0343 (see Figure 1). The smallest difference observed among the unweighted and weighted estimates is 0 while the largest difference is 0.2057, with a median of 0.0771. The impact of the differences between these estimates



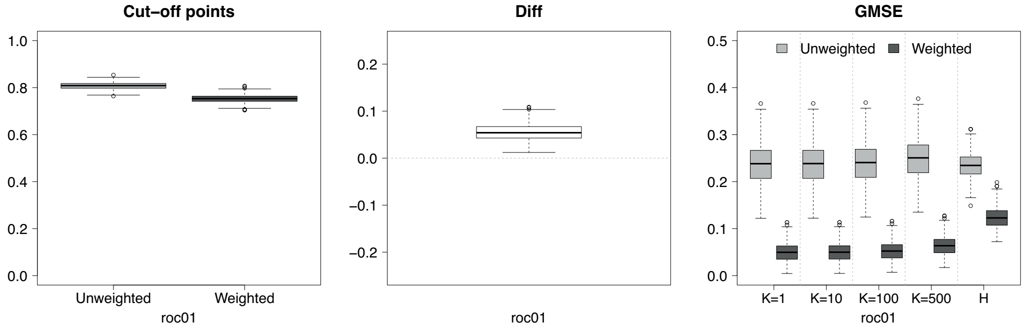
(a) (b) (c)

Figure 2. *Box-plots of the results obtained for the MaxProdSpSe method across $R = 500$ samples: (a) unweighted and weighted estimates of the optimal cut-off points, (b) differences between unweighted and weighted estimates (Diff), and (c) GMSE produced by the unweighted and weighted estimates for $K \in \{1, 10, 100, 500\}$ and H.*

in the total population was measured by the GMSE parameter. In terms of GMSE, the error produced by means of the weighted estimates in the finite population is more or less 5 times smaller than the error produced by means of the unweighted estimates on average. The standard deviation is also smaller for the weighted estimates. For example, for $K = 1$ the GMSE of the unweighted estimates is 0.3110 on average with a standard deviation of 0.0747, while the GMSE of the weighted estimates is 0.0630 on average with a standard deviation of 0.0503. When the GMSE is computed over the $H = 325$ strata, the GMSE turns out to be 0.1298 and 0.2809, for weighted and unweighted estimates, respectively.

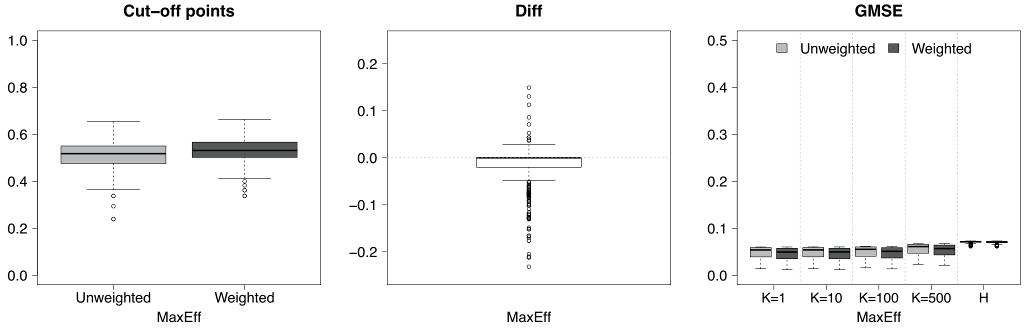
The unweighted estimates obtained by the MaxProdSpSe method are again greater than the weighted ones, being on average 0.8117 and 0.7534, respectively (see Figure 2). The difference between those estimates is 0.0584 on average with a standard deviation of 0.0190. The smallest difference observed among the unweighted and weighted estimates is 0.0121 while the largest difference is 0.1198, with a median of 0.0573. GMSE becomes again 5 times smaller when sampling weights are considered in the estimation process and the standard deviation of the weighted estimates is half of that of the unweighted ones. For example, for $K = 100$ the GMSE is reduced from 0.2532 to 0.0556 on average when considering sampling weights, being the standard deviations of 0.0708 and 0.0342, respectively. The GMSE measured over the different strata for weighted and unweighted estimates is 0.1261 and 0.2425, respectively.

For the ROC01 method weighted estimates are also lower than the unweighted ones (0.7526 and 0.8078 on average, respectively) and the standard deviations are slightly greater (0.0174 and 0.0151, respectively) (see Figure 3). The smallest difference observed among the unweighted and weighted estimates is 0.0121 while the largest difference is 0.1088, being the median of 0.0540 and the average of 0.0552 with a standard deviation of 0.0166. The error generated by the weighted estimates in the finite population is again lower than the error produced by the unweighted estimates in terms of GMSE.



(a) (b) (c)

Figure 3. *Box-plots of the results obtained for the ROC01 method across* $R = 500$ *samples: (a) unweighted and weighted estimates of the optimal cut-off points, (b) differences between unweighted and weighted estimates (Diff), and (c) GMSE produced by the unweighted and weighted estimates for* $K \in \{1, 10, 100, 500\}$ *and* H*.*



(a) (b) (c)

Figure 4. *Box-plots of the results obtained for the MaxEfficiency method across* $R = 500$ *samples: (a) unweighted and weighted estimates of the optimal cut-off points, (b) differences between unweighted and weighted estimates (Diff), and (c) GMSE produced by the unweighted and weighted estimates for* $K \in \{1, 10, 100, 500\}$ *and* H*.*

For example, for $K = 10$, the error obtained by the weighted estimates is 0.0507 on average with a standard deviation of 0.0210, while for the unweighted estimates the error is 0.2368 on average with a standard deviation of 0.0462. The GMSE computed over the different strata takes the value of 0.1245 and 0.2340 for weighted and unweighted estimates, respectively.

Finally, in contrast to the results obtained by the rest of the methods, for the MaxEfficiency method no significant differences are observed among the unweighted and weighted estimates. Optimal cut-off point estimates throughout the $R = 500$ samples are quite similar in terms of mean and standard deviation. The average of the unweighted estimates is of 0.5106 while for the weighted estimates the average is of 0.5297. The standard deviation of the weighted estimates (0.0522) is slightly lower than the standard deviation of the unweighted estimates (0.0579). The smallest absolute difference

Table 1. *Average (mean) and standard deviation (sd) of the a) unweighted and weighted optimal cut-off points, b) difference (Diff) and absolute difference (AbsDiff) among them and, c) GMSE produced by the unweighted and weighted optimal cut-off points when classifying units in the finite population for* $K \in \{1, 10, 100, 500\}$ *and H across* $R = 500$ *samples for all the methods considered.*

		Youden	**MaxProdSpSe**	**ROC01**	**MaxEff**
		Mean (sd)	**Mean (sd)**	**Mean (sd)**	**Mean (sd)**
Cut-off points	**Unweighted**	0.8304 (0.0208)	0.8117 (0.0157)	0.8078 (0.0151)	0.5106 (0.0579)
	Weighted	0.7524 (0.0277)	0.7534 (0.0183)	0.7526 (0.0174)	0.5297 (0.0522)
Diff		0.0780 (0.0343)	0.0584 (0.0190)	0.0552 (0.0166)	-0.0191 (0.0456)
AbsDiff		0.0780 (0.0343)	0.0584 (0.0190)	0.0552 (0.0166)	0.0232 (0.0436)
GMSE (K=1)	**Unweighted**	0.3110 (0.0747)	0.2509 (0.0525)	0.2366 (0.0440)	0.0482 (0.0132)
	Weighted	0.0630 (0.0503)	0.0530 (0.0243)	0.0505 (0.0198)	0.0454 (0.0136)
GMSE (K=10)	**Unweighted**	0.3112 (0.0762)	0.2511 (0.0544)	0.2368 (0.0462)	0.0483 (0.0140)
	Weighted	0.0632 (0.0509)	0.0532 (0.0253)	0.0507 (0.0210)	0.0456 (0.0144)
GMSE (K=100)	**Unweighted**	0.3131 (0.0899)	0.2532 (0.0708)	0.2390 (0.0642)	0.0496 (0.0211)
	Weighted	0.0656 (0.0566)	0.0556 (0.0342)	0.0531 (0.0307)	0.0469 (0.0211)
GMSE (K=500)	**Unweighted**	0.3219 (0.1361)	0.2628 (0.1203)	0.2488 (0.1153)	0.0556 (0.0419)
	Weighted	0.0764 (0.0791)	0.0667 (0.0621)	0.0642 (0.0594)	0.0530 (0.0413)
GMSE (H)	**Unweighted**	0.2809 (0.0470)	0.2425 (0.0325)	0.2340 (0.0270)	0.0706 (0.0022)
	Weighted	0.1298 (0.0377)	0.1261 (0.0250)	0.1245 (0.0235)	0.0701 (0.0025)

observed among the unweighted and weighted estimates is 0 while the largest absolute difference is 0.2318. In particular, in more than 50% of the cases the difference between weighted and weighted estimates is 0. The difference of the error produced by those estimates in the finite population is also negligible. For $K = 1$ for example, the GMSE produced by the unweighted estimates is on average of 0.0482 with a standard deviation of 0.0132, while the average of GMSE of the weighted estimates is 0.0454 with a standard deviation of 0.0136. The GMSE calculated over the $H = 325$ strata, is 0.0701 for weighted estimates and 0.0706 for unweighted estimates.

5. Application to a real survey data

The methodology proposed in Section 3 could be applied to real-world surveys. In particular, for illustration purposes, we have applied this methodology to the ESIE survey data described in Section 2.

In this case, the response variable Y in which we are interested in indicates the availability of the website for each company: it takes the value $y_i = 1$ if a company has its own website and $y_i = 0$ otherwise. Assume that the goal is to estimate the probability

Table 2. *Optimal cut-off point estimates obtained by means of Youden, MaxProdSpSe, ROC01 and MaxEfficiency methods, considering or not the sampling weights.*

	Youden	**MaxProdSpSe**	**ROC01**	**MaxEff**
Unweighted	0.7998	0.7998	0.7998	0.3882
Weighted	0.7518	0.7518	0.7470	0.3882

of event for Y of the companies in the finite population. Thus, we want to fit a logistic regression model to our sample. Four categorical variables that are also available in the finite population will be used as predictors: X_1 (which indicates the province where the company is located, in 3 categories), X_2 (indicates the activity of the company, in 9 categories), X_3 (indicates the ownership of the company, in 7 categories) and X_4 (indicates the number of employees of the company, in 4 categories). In this way, a logistic regression model was fitted to the sample considering these four covariates, the regression coefficients where estimated and $\hat{p}(\boldsymbol{x}_i)$ where calculated for each sampled unit.

We have applied the methods described in Section 3 for the selection of optimal cut-off points, which have been estimated by both, ignoring and considering sampling weights. The results are shown in Table 2. It can be observed that the unweighted and weighted estimates differ when Youden, MaxProdSpSe and ROC01 methods are applied, which is in line with the results obtained in the simulation study. In particular, the unweighted estimates are greater than the weighted estimates, which are similar to the ones observed in Section 4.2 (see Table 1). The unweighted and weighted estimates obtained by means of the MaxEfficiency method are equal, which is also in line with the results observed in the simulation study. Those estimates obtained by the MaxEfficiency method are lower than the average of the estimates obtained in the simulation study. However, it should be noted that this may be justified by the large standard deviation observed previously for the cut-off points estimated by means of the MaxEfficiency method (see Figure 4 and Table 1).

6. Discussion

In this work, a methodology has been proposed for estimating optimal cut-off points of the probability of event in the logistic regression framework cons

idering sampling weights in the estimation process. In particular, we have focused on data derived from complex sampling designs. For this purpose, four optimal cut-off point estimation methods (which are denoted as Youden, MaxProdSpSe, ROC01 and MaxEfficiency (López-Ratón et al., 2014)) have been selected and modified in order to incorporate sampling weights in the estimation process. These four methods have been selected for being the ones most commonly applied in the literature. In particular, the so widely used `pROC` package in R (Robin et al., 2011) has incorporated the Youden and ROC01 methods for the estimation of optimal cut-off points. All these methods are based on different optimality criteria that are related to sensitivity and specificity parameters. Therefore, we propose a methodology for considering sampling weights in the estimation process of sensitivity and specificity parameters, as well as in the estimation of prevalence, in order to estimate optimal cut-off points based on these parameters by taking into account the sampling weights. A simulation study has been carried out in order to analyse the behaviour of both methodologies by comparing the optimal cut-off point estimates obtained by means of the above-mentioned methods when sampling weights are considered or ignored in the estimation process. The error that those estimates generate in the estimation of the probability of event of interest in the finite population has also been analysed in this simulation study. In particular, we considered the GMSE in order to evaluate the behaviour of the prevalence once the cut-off point was estimated, by comparing it with the true prevalence. We also considered it interesting to study the differences in estimating sensitivity and specificity based on the cut-off points estimated with and without sampling weights. However, in this case, the theoretical value of these parameters in the population are unknown and therefore the comparison is not so direct. Even so, we have observed (results not shown) that the differences are in line with those observed when studying the GMSE.

In general, the results suggest the convenience of incorporating sampling weights into the estimation process of optimal cut-off points. For three out of the four methods studied, estimates obtained differ depending on whether the sampling weights were considered or not. Furthermore, it can be observed that the error in the estimates of the response variable obtained by taking into account sampling weights is much smaller than that generated by the estimates obtained by ignoring them for the units in the finite population. Although the cut-off point estimates may not seem very different from each other in some cases, it is observed that the effect of applying one or the other estimate for the classification of units in the population is considerable. In our opinion, the reason for this is that a large amount of individuals of the finite population (specifically, more than 20% of all the units on average) has estimated probabilities which range in the interval defined by the unweighted and weighted estimates and thus, choosing the unweighted cut-off point leads to misclassify a larger number of units in the finite population.

Nevertheless, the results related to the MaxEfficiency method appear to be different compared to Youden, MaxProdSpSe and ROC01. In general, in the results obtained using this method, there are no great differences between the estimates obtained by ignoring or considering the sampling weights, and furthermore, in most cases, the two estimates coincide. Therefore, the errors generated in the population by these estimates are also similar and there are no significant differences among them. Hence, we can say that, at least under the scenario we have worked on, there is no difference among the unweighted and weighted estimates obtained by the MaxEfficiency method. However, we believe that this could be due to a particular characteristic of the scenario in which we have worked and not a specific property of the method itself. Specifically, we believe that differences among those estimates obtained by using or not sampling weights could occur when there are also significant differences between unweighted and weighted estimates of the prevalence, which is not the case in the scenario that has been studied. In particular, the unweighted estimate of the prevalence is 0.8330 on average in the simulated samples, while the weighted estimate is 0.7552. Due to the properties of the efficiency function, we believe that different cut-off point estimates may be obtained for this method when one of the prevalence estimates (either weighted or unweighted) is greater than 0.5, while the other is smaller (results not shown). Nevertheless, studying the mathematical properties of this behaviour is part of a further research, which is out of the scope of this paper.

Finally, we would like to comment on the limitations of this study. First of all, it should be noted that we have conducted this simulation study based on a real survey data. Therefore, the effect that the sampling technique chosen may have on the differences between weighted and unweighted optimal cut-off point estimates remains to be studied as further work. For example, it should be mentioned that in this study we have only analysed the effect of the sampling weights obtained by means of one-stage stratification. Data derived from other sampling techniques such as clustering or two-stage sampling have not been considered. It would also be interesting to study the behaviour of the studied methods under non-informative complex sampling designs. Secondly, it would be interesting to analyse and compare the behaviour of the methods that have been studied throughout this document in different scenarios, for instance, with different prevalence values. Nevertheless, it should be noted that as the simulation study we have used is based on a real survey, the prevalence of the scenario we have analysed was also described by the observed data.

In conclusion, in this work we have implemented four of the most commonly used optimal cut-off point estimation methods, which are implemented in diverse software. Out of these four methods, in three of them the use of sampling weights highly improve the results, while in the fourth, the results do not differ whether you use the sampling weights or not. Therefore, our recommendation is to incorporate the sampling weights in the estimation process of optimal cut-off points when working with data derived from complex sampling designs. However, it should be noted that if one is interested in applying other methods, different from those studied throughout this paper, it should be considered whether it is appropriate or not the use of sampling weights in each particular case.

Acknowledgment

This work was financially supported in part by grants from the Departamento de Educación, Política Lingüística y Cultura del Gobierno Vasco [IT1294-19] and by the Spanish State Research Agency through BCAM Severo Ochoa excellence accreditation [SEV-2017-0718] and through project [PID2020-115882RB-I00 / AEI / 10.13039/5011 00011033] funded by Agencia Estatal de Investigación and acronym “S3M1P4R”; Basque Government through the [BERC 2018-2021]; the SPRI-Basque Government through the Elkartek project 3KIA [KK-2020/00049]. The work of AI was supported by grant [PIF18/213] and [PES18/26] from the Universidad del País Vasco UPV/EHU.

We would like to acknowledge the Official Statistics Basque Office (Eustat) for providing us with the ESIE survey data. We also gratefully acknowledge María Xosé Rodríguez Álvarez for helping us incorporate the sampling weights into the `OptimalCutpoints` R package functions.

Conflict of interest

The authors declare that there are no conflicts of interest.

References

Baker, T. and Gerdin, M. (2017). The clinical usefulness of prognostic prediction models in critical illness. *European Journal of Internal Medicine*, 45:37–40.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12(4):387–415.

Binder, D. A. (1981). On the variances of asymptotically normal estimators from complex surveys. *Survey Methodology*, 7(2):157–170.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51(3):279–292.

Binder, D. A. and Roberts, G. (2009). Design- and model-based inference for model parameters. *Handbook of Statistics*, 29:33–54.

Chen, J.-Y., Feng, J., Wang, X.-Q., Cai, S.-W., Dong, J.-H., and Chen, Y.-L. (2015). Risk scoring system and predictor for clinically relevant pancreatic fistula after pancreaticoduodenectomy. *World Journal of Gastroenterology*, 21(19):5926–5933.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Filella, X., Alcover, J., Molina, R., Giménez, N., Rodríguez, A., Jo, J., Carretero, P., and Ballesta, A. M. (1995). Clinical usefulness of free PSA fraction as an indicator of prostate cancer. *International Journal of Cancer*, 63(6):780–784.

Greiner, M. (1995). Two-graph receiver operating characteristic (TG-ROC): a Microsoft-EXCEL template for the selection of cut-off values in diagnostic tests. *Journal of Immunological Methods*, 185(1):145–146.

Greiner, M. (1996). Two-graph receiver operating characteristic (TG-ROC): update version supports optimisation of cut-off values that minimise overall misclassification costs. *Journal of Immunological Methods*, 191(1):93–94.

Greiner, M., Pfeiffer, D., and Smith, R. (2000). Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive Veterinary Medicine*, 45(1-2):23–41.

Hanley, J. A. and Mcneil, B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36.

Heeringa, S. G., West, B. T., and Berglund, P. A. (2017). *Applied Survey Data Analysis (2nd ed.)*. Chapman and Hall/CRC.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*. Wiley New York.

Kalton, G. (1983). *Introduction to Survey Sampling*. Thousand Oaks, CA: Sage.

Lewis, J. D., Chuai, S., Nessel, L., Lichtenstein, G. R., Aberra, F. N., and Ellenberg, J. H. (2008). Use of the noninvasive components of the Mayo score to assess clinical response in ulcerative colitis. *Inflammatory Bowel Diseases*, 14(12):1660–1666.

López-Ratón, M., Rodríguez-Álvarez, M. X., Cadarso-Suárez, C., Gude-Sampedro, F., et al. (2014). OptimalCutpoints: an R package for selecting optimal cutpoints in diagnostic tests. *Journal of Statistical Software*, 61(8):1–36.

Lumley, T. and Scott, A. (2015). AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology*, 3(1):1–18.

Magder, L. S. and Fix, A. D. (2003). Optimal choice of a cut point for a quantitative diagnostic test performed for research purposes. *Journal of Clinical Epidemiology*, 56(10):956–962.

Manel, S., Williams, H. C., and Ormerod, S. J. (2001). Evaluating presence–absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, 38(5):921–931.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8(4):283–298.

Pauker, S. G. and Kassirer, J. P. (1980). The threshold approach to clinical decision making. *New England Journal of Medicine*, 302(20):1109–1117.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, New York.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyse and compare ROC curves. *BMC Bioinformatics*, 12(77).

Rutter, C. M. and Miglioretti, D. L. (2003). Estimating the accuracy of psychological scales using longitudinal data. *Biostatistics*, 4(1):97–107.

Skinner, C. J., Holt, D., and Smith, T. F. (1989). *Analysis of Complex Surveys*. John Wiley & Sons.

Spence, R. T., Chang, D. C., Kaafarani, H. M., Panieri, E., Anderson, G. A., and Hutter, M. M. (2018). Derivation, validation and application of a pragmatic risk prediction index for benchmarking of surgical outcomes. *World Journal of Surgery*, 42(2):533–540.

Steyerberg, E. W. (2008). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer Science & Business Media.

Steyerberg, E. W., Marshall, P. B., Jan Keizer, H., and Habbema, J. D. F. (1999). Resection of small, residual retroperitoneal masses after chemotherapy for nonseminomatous testicular cancer: a decision analysis. *Cancer*, 85(6):1331–1341.

Swets, J. A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist*, 47(4):522–532.

Vermont, J., Bosson, J., Francois, P., Robert, C., Rueff, A., and Demongeot, J. (1991). Strategies for graphical threshold determination. *Computer Methods and Programs in Biomedicine*, 35(2):141–150.

Wynants, L., van Smeden, M., McLernon, D. J., Timmerman, D., Steyerberg, E. W., Van Calster, B., et al. (2019). Three myths about risk thresholds for prediction models. *BMC Medicine*, 17(192).

Yao, W., Li, Z., and Graubard, B. I. (2015). Estimation of ROC curve with complex survey data. *Statistics in Medicine*, 34(8):1293–1303.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1):32–35.

A. Generation of the pseudo-population

This section describes the process of generating the pseudo-population that has been implemented in the simulation study described in Section 4. The pseudo-population has been generated based on a real survey data, which is described in Section 2. Let us denote as S_{ESIE} the original survey sample and U_{ESIE} the real finite population of size N (note that $S_{\text{ESIE}} \subset U_{\text{ESIE}}$). It should be noted that some information of the finite population U_{ESIE} and the real sample S_{ESIE} is known for us. In particular, for the N units in the finite population the values for the vector of covariates $X_1,\dots,X_p$ are known, i.e. $\left\{(x_{1j},\dots,x_{pj})\right\}_{j\in U_{\text{ESIE}}}$. In addition to the values for the covariates, the values of the response variables $Y_1,\dots,Y_q$ are also known for the units in the sample, i.e. $\left\{(y_{1j},\dots,y_{qj},x_{1j},\dots,x_{pj})\right\}_{j\in S_{\text{ESIE}}}$. In the ESIE survey, a total of H strata have been defined (i.e., $\{1,\dots,H\}$) combining information of three categorical design variables, which will be denoted as X_1, X_2 and X_3. Therefore, the finite population can be partitioned in subsets defined by means of these strata, i.e., $U_{\text{ESIE}} = \bigcup_{h=1}^{H} U_{\text{ESIE},h}$. $\forall h \in \{1,\dots,H\}$ let us indicate as N_h the size of stratum h in the finite population U_{ESIE} ($U_{\text{ESIE},h}$) and as n_h the size of this stratum in the sample S_{ESIE}. Then, the sampling weight associated to a unit $j \in S_{\text{ESIE}}$ from stratum h is the following:

$$w_j = \frac{N_h}{n_h}. \tag{26}$$

Our goal is to generate a pseudo-population (U) based on the known real ESIE survey data, for which all the information of the covariates $X_1,\dots,X_p$ and the response variables $Y_1,\dots,Y_q$ will be available. This new pseudo-population U will be the same size as the true ESIE population (N). In order to ease the notation, the variable names of the pseudo-population are the same as in the real finite population and the units of the real ESIE population will be denoted as $j \in U_{\text{ESIE}}$ while the units that are artificially generated for the pseudo-population will be denoted as $i \in U$.

Several dichotomous response variables are available in the original survey (being the response variable Y, which we have applied in the simulation study, one of them). All possible combinations of these response variables have been examined. For instance, assuming that $Y_1,\ldots,Y_q$ are all the response variables that are available in the survey (where $Y \in \{Y_1,\ldots,Y_q\}$), for some unit $j \in S_{\text{ESIE}}$: $\boldsymbol{y}_j = (y_{1j},\ldots,y_{qj}) = \alpha$, $\forall\alpha \in \{\alpha_1,\ldots,\alpha_A\}$, where $\{\alpha_1,\ldots,\alpha_A\}$ is the set of all of possible combinations of the responses. For each stratum (i.e., $\forall h \in \{1,\ldots,H\}$) and for each possible combination of the responses (i.e., $\forall\alpha \in \{\alpha_1,\ldots,\alpha_A\}$) we generate $N_{h,\alpha}$ units in the pseudo-population (U) as:

$$N_{h,\alpha} = \sum_{j\in S_{\text{ESIE}}} w_j \cdot 1_{U_{\text{ESIE},h}}(j) \cdot [\boldsymbol{y}_j = \alpha], \tag{27}$$

where,

$$1_{U_{\text{ESIE},h}}(j) = \begin{cases} 1, & \text{if } j \in U_{\text{ESIE},h}, \\ 0, & \text{if } j \notin U_{\text{ESIE},h}, \end{cases} \tag{28}$$

and

$$[\boldsymbol{y}_j = \alpha] = \begin{cases} 1, & \text{if} \quad (y_{1j},\ldots,y_{qj}) = \alpha, \\ 0, & \text{if} \quad (y_{1j},\ldots,y_{qj}) \neq \alpha. \end{cases} \tag{29}$$

In this way, $N_{h,\alpha}$ is the number of units of the pseudo-population U in stratum h, which take the values of responses $(y_{1j},\ldots,y_{qj}) = \alpha$. Once we repeat the process for $\forall h \in \{1,\ldots,H\}$ and $\forall\alpha \in \{\alpha_1,\ldots,\alpha_A\}$ a pseudo-population of $N = \sum_{h\in\{1,\ldots,H\}} \sum_{\alpha\in\{\alpha_1,\ldots,\alpha_A\}} N_{h,\alpha} = \sum_{j\in S_{\text{ESIE}}} w_j$ units will be generated with the information of response variables (Y, among others) and strata (hence, information of the design variables X_1, X_2 and X_3 will also be generated). Note that the pseudo-population U has been created in such a way that has the same number of individuals N as the ESIE finite population U_{ESIE}.

Finally, we generate the rest of the covariates as follows. $\forall s \in \{4,\ldots,p\}$ assume that X_s is a categorical variable with a total of D categories: $\{1,\ldots,D\}$. Then, for each unit generated in the pseudo-population ($\forall i \in U$) from stratum h, we generate $x_{si} \in \{1,\ldots,D\}$ following a categorical distribution (i.e., $x_{si} \sim Cat(\pi_{s1},\ldots,\pi_{sD})$) where the probability corresponding to each category $d \in \{1,\ldots,D\}$ is calculated as follows based on the known ESIE finite population U_{ESIE}.

$$\pi_{sd} = \frac{\sum_{j\in U_{\text{ESIE}}} 1_{U_{\text{ESIE},h}}(j) \cdot [x_{sj} = d]}{\sum_{j\in U_{\text{ESIE}}} 1_{U_{\text{ESIE},h}}(j)}, \quad \forall d \in \{1,\ldots,D\}, \tag{30}$$

where $1_{U_{\text{ESIE},h}}(j)$ is defined in (28) and,

$$[x_{sj} = d] = \begin{cases} 1 & \text{if} \quad x_{sj} = d, \\ 0 & \text{if} \quad x_{sj} \neq d, \end{cases} \quad \forall j \in U_{\text{ESIE}} \text{ and } \forall d \in \{1,\ldots,D\}. \tag{31}$$

In this way, the pseudo-population has been generated with the response variable of interest Y, the vector of covariates $\boldsymbol{X}$ and the strata.

B. Pseudo-population sampling process

The pseudo-population generated following the steps described in Appendix A, has been sampled by one-step stratified sampling with simple random sampling without replacement in each stratum, in the same way as the real survey data described in Section 2.

In the sampling process, first, we identify how many units have been sampled from a stratum h ($\forall h \in \{1,\ldots,H\}$) in the real survey sample S_{ESIE} (let us denote this amount as n_h). Then, we sample randomly n_h units from stratum h of size N_h from the pseudo-population U. In this way, repeating the process for $\forall h \in \{1,\ldots,H\}$ we sample a total of n units (where $n < N$) to the sample $S \subset U$.

Finally, sampling weights are assigned to each sampled unit as follows. For $\forall i^* \in S$ (assume that $i^* \in h$ ($\forall h \in \{1,\ldots,H\}$)), then:

$$w_{i^*} = \frac{N_h}{n_h}. \tag{32}$$