Penalized logistic regression to improve predictive capacity of rare events in surveys

Jessica Pesantez-Narvaez and Montserrat Guillen*
(J.P-N) https://orcid.org/0000-0003-3161-7807; (M.G) https://orcid.org/0000-0002-2644-6268
Department of Econometrics, Riskcenter-IREA, University of Barcelona, Av. Diagonal 690, 08034 Barcelona, Spain

Abstract. Logistic regression as a modelling technique of rare binary dependent variables with much fewer events (ones) than non-events (zeros) tends to underestimate their probability of occurrence. The vast literature devoted to the prediction of rare binary data identifies several ways to improve predictive performance by making modifications to the likelihood estimation. We propose two weighting mechanisms for incorporation in a pseudo-likelihood estimation that improve the predictive capacity of rare binary responses in data collected in complex surveys. We multiply sampling weights by specific correctors that lead to lower root mean square errors for event observations in almost all deciles. A case study is discussed where this method is implemented to predict the probability of suffering a workplace accident in a logistic regression model that is estimated with data from a survey conducted in Ecuador.

Keywords: Survey data, sampling design, uncommon events, weighting, pseudo-likelihood.

1. Introduction

Models of binary dependent variables sometimes deal with much fewer events (ones) than non-events (zeros). We address the statistical problem of modelling survey data as in [7], who propose a method to correct the likelihood estimate in logistic regression that seeks to predict rare events.

Examples of phenomena that do not occur very often can be found in all areas, where the percentage of cases of interest falls below 10 or even 5%. In socioeconomic surveys, model rare phenomena could include the estimation of the proportion of workers who changed their job in the week prior to the interview. In health surveys, responses to the use of certain drugs or diseases can also be quite infrequent.

Our aim is to improve the predictive capacity of models for rare phenomena with data collected in a complex sample design. We propose a new method and we also present a case study, in which we analyse survey data to model the occurrence of workplace accidents.

Due to economic and time costs, surveys are usually conducted using complex sampling designs (e.g. stratified, cluster or two-stage sampling) rather than simple random sampling (SRS). Sampling weights are defined to make the sample representative of the population and to avoid selection bias, even if the observations in some survey designs are dependent.

The design effect, which measures the ratio between the variance estimation under a specific sample design and that of an SRS, varies from one survey to another and even varies for each estimator within a given survey. Deviations from SRS are expected to produce a loss of efficiency, but this loss should be kept as low as possible. Sampling errors should be carefully estimated, and inference in general must consider the data collection mechanism.

* Montserrat Guillen. Tel.: +34934037039; Fax: +34934021821; Email: mguillen@ub.edu.

Even when sampling weights are considered in the modelling process, randomness is still influenced by the sampling procedure. [10] demonstrates that the modelling process must take into account sampling weights as well as the random part of the model to obtain the precision of the estimates, and to assess modelling performance.

Apart from the complexity in the way survey samples are obtained, the presence of rare events i.e. binary dependent variables that have few non-zero cases, is quite common in practice. This can represent a challenge for the performance of predictive models, which seek to determine the factors affecting the probability of the rare event. The reason for this is that the small number of observed cases leads to quite unstable model results. [7] prove that binary dependent models, in particular logistic regression, tend to underestimate the event probability for this type of rare event data, and they propose a correction procedure in the usual logistic regression maximum likelihood estimation to manage bias. However, they leave aside rare binary dependent variable modelling prediction as a design-based analysis with sampling weights. Yet, ignoring sampling weights might affect the meaning and precision of the coefficients.

Modifications of the maximum likelihood estimation through weights are not new in the vast literature devoted to generalized linear models. For instance, [13] introduces the quasi-likelihood function, [12] modify the weighted exogenous sampling likelihood function estimator by weighting each observation's contribution to the likelihood. [9] and [1] incorporate weighting mechanisms in the maximum likelihood estimation method.

We propose a statistical procedure that incorporates both approaches. We consider rare events in samples that deviate from SRS and we modify the maximum likelihood estimation to improve the predictive accuracy of the model. Hence, we aim to contribute to the existing literature by proposing a weighting mechanism that can be incorporated in the likelihood estimation, which then naturally becomes a pseudo-likelihood estimation, of a penalized logistic regression model. This mechanism is capable of performing two joint tasks: first, it controls the randomness of a sampling procedure by considering the sampling weighting, stratification or clustering that originates from a complex survey design; and second, it provides the model with greater sensitivity, in order to obtain more accurate predictions of rare events than if only a weighted design-based logistic regression model had been used.

Our motivation for proposing a weighting mechanism is that it allows us to differentiate between the relevance of observations in the sample. In this way, we can avoid the under-representation or over-representation of observations when it comes to estimating choice probabilities from choice-based samples as introduced by [12]. But the mechanism extends this idea further, so that the importance of the observations varies depending on the proximity to the mean value of the response. An adjustment parameter calibrates the impact of the weighting mechanism on the model estimation. In addition, a threshold value is chosen to provide the best predictive performance.

Following on from this introduction, this paper is divided in four parts. Section 2 outlines the methodology and the two weighting mechanisms are presented and justified in detail. Three criteria are proposed to find the best predictive model among all possible models by choosing an optimal weight adjustment and a classifying threshold. Section 3 describes the data used herein as an illustrative example. Specifically, we are interested in modelling the occurrence of workplace accidents. Section 4 presents the results and the predictive performance obtained in the case study. Section 5 concludes.

2. Methodology

Let X_{ij} be the data matrix where *i* corresponds to observations (or instances) and *j* corresponds to the independent variables (attributes or features), with *i = 1 ,…, n* and *j = 1,…, k*. There are *n* observations and *k* independent variables. And let Y_i be the binary outcome for observation *i*.

Our goal is to classify observations between the binary outcome Y_i, taking into consideration the covariates X_j.

2.1. Penalized logistic regression and pseudo-likelihood estimation

One supervised method of machine learning is the logistic regression model. [4] and [11] define logistic regression as a predictive method used for binary classification problems which, unlike a linear regression model, provides estimates about the probability of an outcome.

To formally define the penalized logistic regression model, we first introduce the pseudo-likelihood estimation (weighted maximum likelihood) with survey data.

For every instance X_i (row vector of X_{ij}), the outcome response is either Y_i = 1 if the observations belong to a positive class (event) or Y_i = 0 if they belong to a negative class (non-event).

Binary variable Y_i is a Bernoulli trial:

$$Y_i \sim \text{Bernoulli } (Y_i | \, p_i),$$

where p_i is the probability that Y_i equals 1 and is specified as:

$$p_i \; = \; P \; (Y_i =1 \mid \; X_{i1},\ldots, X_{ik} \;) = \frac{e^{\beta_o + \sum_{j=1}^{k} X_{ij}\beta_k}}{1 + e^{\beta_o + \sum_{j=1}^{k} X_{ij}\beta_k}} \, . \quad (1)$$

Conversely, the probability that Y_i equals 0 is $1 - p_i$. Unlike linear regression, logistic regression uses a logit function as the linear predictor, which is the log odds of the positive response, defined as:

$$\eta_i = \log\left(\frac{p_i}{1-p_i}\right) \; = \; \beta_o + \sum_{j=1}^{k} X_{ij}\beta_k \;\; . \qquad (2)$$

Then, the classical likelihood function is the joint Bernoulli probability distribution of observed values of Y_i as follows:

$$l \; (\beta_o, .., \beta_k; \; X_i) = \; \prod_{i=1}^{n} {p_i}^{Y_i} \;\; (1 - p_i)^{1-Y_i}, \qquad (3)$$

Parameter estimates of the classical logistic regression can be found by maximizing the likelihood or log-likelihood function.[2] For reasons of computational convenience, we use the log-likelihood function, which we denote by L for simplicity:

$$L = \sum_{i=1}^{n} ln \; p \; (X_i)^{Y_i} \; + ln \; (1 \; - \; p \; (X_i))^{1-Y_i}. \quad (4)$$

Furthermore, if weights are incorporated in the log-likelihood function (4) then a weighted log-likelihood is obtained:

$$L = \sum_{i=1}^{n} W_i \;\; (ln \; p \; (X_i)^{Y_i} + ln \; (1 \; - \; p \; (X_i))^{1-Y_i}), \quad (5)$$

where W_i represents the weight of the i-th observation. Therefore, estimating the parameter vector becomes a maximization problem whose objective function is the pseudo-likelihood function defined in (5).

Maximization in (5) can be computed with the *survey()* package in R: Partial derivate equations are solved by an iteratively reweighted least squares algorithm, which is a Fisher scoring algorithm (further details can be found in [3]). The *survey()* package created by [10] not only allows the weighting procedure to be incorporated, but it also adapts the penalized logistic regression to complex survey designs in order to provide design-based standard errors. So, if survey data include a stratified and/or a clustered design, the maximization includes the corresponding formulas to find correct sample-based standard errors.

[14] note the importance of weighting the observations from complex samples in order to derive unbiased estimates of population features. Weighting can be used to both guarantee sample representativeness in a modelling process (as noted by [12]) and to control the relevance of observations. Thus, our approach proposes weighting observations not only to correct a survey sample design but also to improve its predictions. This is of particular interest for low frequency events, which are more difficult to predict than high frequency occurrences. Our corrections are introduced in a penalized logistic regression model with a pseudo-likelihood estimation method.

Sample correction and weighting aimed at improving predictive capacity have both been widely discussed in the literature but, to the best of our knowledge, in these discussions they have typically been addressed separately. We aim to study these weighting procedures jointly and define W_i in (5) in accordance with these objectives.

2.2. Weighting mechanisms

Let SW_i be a vector of sampling weights and PW_i a vector of predictive weights. These two weighting mechanisms are introduced in (5), where W_i is the result of the product between PW_i and SW_i.

The basis for the sampling weights lies in the probability of choosing a respondent. This means that each observation in the sample is given a weight to account for the probability of that observation being selected from the population. For this reason, sampling weights incorporate an expansion factor that is equal to the number of population units represented by each observation in the sample. Sampling weights are defined as follows:

[2] Maximizing a log-likelihood function is equivalent to maximizing a likelihood function.

$$SW_i = \frac{F_{exp\,i} * n}{\sum_{i=1}^{n} F_{exp\,i}}, \qquad (6)$$

where $F_{exp\,i}$ is the vector of expansion factors defined as the inverse of the probability of choosing each observation in the sample.

For the predictive weighting, PW_i, we propose two alternatives, which we call *a* and *b*:

a) $PWa_i = \left|\widehat{Y}_i - \overline{Y}\right|^{\varepsilon}$

b) $PWb_i = \left|\widehat{Y}_i - \overline{Y}^{\varepsilon}\right|$

where $\widehat{Y}_i$ is the vector of estimated probabilities of a simple initial weighted, design-based logistic regression (accounting for SW_i only, where other sample-design features such as stratification and/or clustering would only affect standard errors) and $\bar{Y}$ is the estimated weighted mean response of the dependent variable. Let ε be the adjustment parameter that calibrates the distance between $\widehat{Y}_i$ and $\bar{Y}$ in both alternatives *a* and *b*.

Note that the estimated probabilities $\widehat{Y}_i$ lie between 0 and 1.

- PWa_i differentiates the weight of observations that are located far from the mean. The possible scenarios for selecting the adjustment parameter are:

$\varepsilon = 0$; The maximum pseudo-likelihood estimation remains the same as the weighted design-based model.

$\varepsilon > 0$; The weighting attaches greater importance to the observations whose original predictive value is located far from the mean response.

$\varepsilon < 0$; The weighting gives greater importance to the observations whose original predictive value is located near the mean response.

- PWb_i isolates the estimated probabilities from the mean. The choice of the threshold is usually located near the mean response. Observations whose predicted probability is located near the mean are more likely to be influenced by the choice of the threshold, than those that have a predictive probability that is located far from the mean. This weighting mechanism allows three possible scenarios for selecting the adjustment parameter:

$\varepsilon = 0$; Then $\overline{Y}^{\varepsilon}$ =1 and the predictive weights equal the estimated probability of the non-event, $(1 - \widehat{Y}_i)$.

$\varepsilon > 0$; More weight to the observations which are much greater than the mean and less weight to the observations which are much smaller than the mean.

$\varepsilon < 0$; Less weight to the observations which are located far from the mean and more weight to the observations which are located near the mean.

So far, PW_i and SW_i may have a different scale. While the sampling weights in (6) sum up to *n*, this is not necessarily true of the predictive weights. Therefore, we propose rescaling them and obtaining the new $PWa_i{}'$ and $PWb_i{}'$ as follows:

$$PWa_i{}' = \frac{PWa_i * n}{\sum_{i=1}^{n} PWa_i} \ ; \qquad (7)$$

$$PWb_i{}' = \frac{PWb_i * n}{\sum_{i=1}^{n} PWb_i} . \qquad (8)$$

Then, the two final weights $PSWa_i$ and $PSWb_i$ combining the sampling and predictive weights can be defined as:

$$PSWa_i = SW_i * PWa_i{}' , \qquad (9)$$

$$PSWb_i = SW_i * PWb_i{}' . \qquad (10)$$

2.3. Choosing the adjustment parameter

Three criteria are established for choosing the adjustment parameter to test the predictive performance of each model.

- Receiver operating characteristic (ROC) optimal criterion

[5] propose the ROC curve as a graphical plot that seeks to determine the relationship between sensitivity – i.e. the percentage of true positive values (on the y-axis) – and 1-specificity – i.e. the percentage of false

positive values (on the x-axis). Sensitivity and specificity are measures of the performance of a binary classification method. Sensitivity is a measure of the proportion of actual positives (events) that are correctly identified as such, while specificity is a measure of the proportion of actual negatives (non-events). The ROC curve illustrates the capacity of the logistic regression model, as a particular case of a binary classifier method given a threshold Ψ. The threshold is a fixed value in [0,1], which determines when an estimated probability is large enough for the binary prediction to take the value of 1. The desired model should have a high true positive rate as well as a small false negative rate Therefore, the best prediction model would yield a point on the ROC curve that is as close as possible to the coordinate (0,1).

The ROC optimal criterion is based on setting all possible adjustment parameters ε in the domain of the penalized logistic regression, considering that for each ε, there is a choice of possible thresholds [0.01, 0.02, …, 0.99]. The best model coordinates in the ROC plot are those with the shortest distance to the point (0,1). All ROC distances to the coordinate (0,1) are computed. Therefore, the ROC optimal criterion is a minimization problem where ε and Ψ have to be found.

- Constrained receiver operating characteristic (C-ROC) criterion

The C-ROC criterion is motivated by a discussion of desirable statistical performance measures of a good predictive model. A good predictive model would be expected to accomplish maximum levels of sensitivity, minimum type I and type II errors or, at least, a minimum type II error.

First, a predictive model with maximum sensitivity is especially important for identifying the true positive rate (Y_i =1), which is the main point of interest for our study. However, finding such a model might imply very low levels of specificity, which might be a disadvantage. Second, a good predictive model can also be expected to have the smallest possible false positive and false negative rates. However, it is far from straightforward to minimize both false positive and false negative rates, because when one is low the other is high. Thus, finding a suitable cut-off threshold for deciding the best predictive model in line with this criterion requires making a compromise. Third, reducing type II errors might be considered dangerous in prediction implementations because, in some cases, the reason for predicting rare events is to prevent them.

Thus, so far, it would seem that the three requirements are all necessary, but that they are not all feasible at the same time. For this reason, taking as our base criterion the ROC analysis described above, we propose using the C-ROC, which comprises the following two steps:

1.- Finding the first adjustment parameters based on the ROC optimal criterion.[3] In order words, this requires ranking the models from best to worst in terms of how well they meet the ROC criterion and selecting the first m ones.

2.- Maintaining the subset based on this previous order and finding the adjustment parameter whose corresponding model has the highest sensitivity value. If values are equal then, once fixed, select the one with the highest specificity.

The goal is to retain the model with the highest levels of sensitivity, reducing a minimum specificity. This is feasible if the adjustment parameters of each predictive model are first sorted according to the ROC criterion.

- Assessing performance with the root mean square error (RMSE)

This is a statistical measure that rates the difference between observed and predictive values: the smaller the RMSE, the better the model's predictive performance. The RMSE is calculated as follows:

$$\text{RMSE} = \left[\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 / \, n\right]^{1/2}, \qquad (11)$$

where $\hat{Y}_i$ is the predicted values from the estimated model. In our application, we have used this criterion only for the subsample of events and (11) was used to analyse predictive performance rather than as a criterion to select the adjustment parameters.

[3] Let m be a positive number. The intuitive idea for m is just how many better models, in terms of the ROC criterion, the analyst is willing to sacrifice in order to opt for a model with a higher sensitivity among those m models. However, m should be small enough to maintain the models' ROC distance as small as possible. Thus, m should be selected from between 2 and 10; nevertheless, the choice depends on the quantity and characteristics of the sample data. In the application shown in this article, m is fixed equal to 6.

3. Data of the illustrative example

We use a workplace accident data set taken from the Ecuadorian National Survey of Employment, Underemployment and Unemployment (ENEMDU) conducted in December 2017 by the *Instituto Nacional de Estadísticas y Censos* (INEC). The data were collected in personal interviews to gather information about the labour market in Ecuador. The survey employs a two-stage sampling design: the first step involves the stratification of 2,586 primary sample units (PSUs) represented as sectors, and the second step involves choosing 12 secondary sample units (SSUs) per every PSU represented as dwellings by a simple probabilistic sampling. The final observation unit is the household (for further details see [6]).

The dataset has 110,283 observations (individuals) and 313 variables. Only the subset of individuals that were employed at the time of the survey was selected. This is a subsample of 31,057 observations.

In the ENEMDU, all members of a dwelling are interviewed and so all the members of a dwelling form a cluster. This means a potentially positive correlation in their answers to the questionnaire. This would imply greater standard errors in the estimated coefficients than if the clustered sampling design was not taken into consideration.

Table 1. Definition of the variables in the dataset

Type	Variable	Description
Dependent	Workplace accident	Binary variable which takes the value of 1 if the employee had a workplace accident and 0 otherwise.
Independent	Age	Continuous numerical variable that represents the employee's age.
	Seniority	Continuous numerical variable that represents the seniority (years) in the current job.
	Men	Binary variable that takes the value of 1 if the employee is a man and 0 if a woman.
	Urban	Binary variable that takes the value of 1 if the employee lives in an urban area, and 0 in a rural area.
	Marital	Categorical variable for marital status: single, married and other.
	Workplace safety training	Binary variable that takes the value of 1 if the employee has received a workplace safety training and 0 otherwise.
	Working hours	Continuous numerical variable that represents the number of working hours per week.

Table 1 records the definitions of the variables in the data set, and Table 2 shows the descriptive statistics of this data set. Overall, employees who declared that they had suffered a workplace accident represent 3.11% of the total, which means the occurrence of such events is quite rare. The mean age of workers who had suffered a workplace accident is 3 years more than that of those who had not suffered an accident. Among male employees, 4.09% had suffered a workplace accident, while only 1.80% of women had. Rural workers present a slightly higher rate of workplace accidents (3.28%) than urban workers (2.98%). Married employees had a higher workplace accident rate with respect to single workers and those of other marital status. Finally, the number of weekly working hours under Ecuadorian law is fixed at 40 (Art. 47 of the Ecuadorian labor code). Workers who exceed this limit by 2 hours are more likely to suffer a workplace accident than workers whose average weekly working hours are 38.

Additionally, employees who had received workplace safety training presented a higher rate of accidents (5.21%) than employees who had not received such training (2.49%). This result may be due to the fact that workers in dangerous work places tend to receive more workplace safety training than others. Finally, the mean number of years of seniority is higher among workers who had suffered workplace accidents than those who had not.

4. Results

This section presents the results of the logistic regression with sampling weights and two estimated penalized logistic regression models based on weighting mechanisms $PSWa_i$ and $PSWb_i$ for each of the criterion proposed in Section 3.

Table 2. Descriptive statistics of the workplace accident data set

Variables		No Workplace Accident (Y=0)	Workplace Accident (Y=1)	Total
Age (years)		36.78	39.57	36.87
Seniority in establishment (years)		8.08	9.23	8.11
Sex	Woman	13,145 (98.20%)	241 (1.80%)	13,361
	Man	17,021 (95.91%)	726 (4.09%)	17,696
Area	Rural	12,252 (96.72%)	416 (3.28%)	12,634
	Urban	17,914 (97,02%)	551 (2.98%)	18,423
Marital Status	Single	9,617 (97.91%)	205 (2.09%)	9,801
	Married	17,761 (96.35%)	672 (3.65%)	18,389
	Other	2,788 (96.87%)	90 (3.13%)	2,867
Workplace safety training	Yes	6,696 (94.79%)	368 (5.21%)	7,064
	No	23,396 (97.51%)	598 (2.49%)	23,993
Working hours		38.17	42.02	38.29
Total		30.091 (96.89%)	966 (3.11%)	31,057

Note: Unweighted estimates are presented for continuous variable means, while for categorical variables the frequencies are presented. Row percentages are shown in parentheses.

Table 3. Statistical predictive performance measures

Statistical predictive performance measures of the weighted design-based logistic regression model						
	Sensitivity (%)	Specificity (%)	Accuracy (%)	ROC criterion distance		Ψ
	56.522	66.458	66.114	0.549		0.03
Statistical predictive performance measures obtained using *PSWa*						
Order	Sensitivity (%)	Specificity (%)	Accuracy (%)	ROC criterion (distance)	ε	Ψ
1º	59.731	63.743	63.619	0.542	0.05	0.03
2 º	59.524	63.966	63.828	0.542	0	0.03
3 º	**60.870**	**62.354**	**62.308**	**0.543**	**0.4**	**0.03**
4 º	60.663	62.471	62.414	0.544	0.35	0.03
5 º	59.110	64.145	63.989	0.544	-0.1	0.03
6 º	59.938	63.215	63.113	0.544	**0.15**	0.03
Statistical predictive performance measures obtained using *PSWb*						
Order	Sensitivity (%)	Specificity (%)	Accuracy (%)	ROC criterion (distance)	ε	Ψ
1º	**59.834**	**63.923**	**63.796**	**0.540**	**0.6**	**0.03**
2 º	59.524	63.999	63.860	0.542	-0.3	0.03
3 º	59.524	63.999	63.860	0.542	-0.25	0.03
4 º	59.524	63.993	63.854	0.542	-0.75	0.03
5 º	59.524	63.993	63.854	0.542	-0.7	0.03
6 º	59.524	63.993	63.854	0.542	-0.65	0.03

Note: Models that meet the C-ROC criterion are bold character when only the first six models are considered.

Table 3 shows the predictive performance measures of three types of model: the first is a simple weighted design-based logistic regression model where only the SW_i, sampling weight mechanism is used, as well the sampling design. The second is the model estimated using $PSWa_i$, and the third is the model estimated using $PSWb_i$. For the second and third model types, we present the first six models that best meet the ROC optimal criterion.

The results in Table 3 for the ROC criterion show that the adjustment parameter with the lowest ROC distance is $\varepsilon = 0.05$, a threshold Ψ = 0.03 and a sensitivity that equals 59.731%, when the weighting mechanism $PSWa_i$ is used in the predictive modelling. The lowest ROC distance when $PSWb_i$ is used is obtained for the adjustment parameter $\varepsilon = 0.6$, a threshold Ψ = 0.03 and a sensitivity that equals 59.834%.

Figures 1 and 2 show the ROC representation of all possible models based on weighting alternatives *a* and *b* respectively; thus, every dot represents a model.

When $PSWa_i$ is used under the C-ROC criterion the best adjustment parameter is $\varepsilon = 0.4$ and a threshold Ψ = 0.03, being among the *six* best models according to the ROC criterion. In this case, the highest sensitivity value is 60.87%. Note we ignore the first two models with a better ranking under the ROC criterion because of their lower sensitivity values (59.731 and 59.524%, respectively). When $PSWb_i$ is used, the best sensitivity of the six models corresponds to an adjustment parameter $\varepsilon = 0.6$ and a threshold Ψ = 0.03. Here the ROC criterion leads to a highest sensitivity value of 59.834%.

Note that the adjustment parameter ε is jointly chosen with Ψ (among all the possible values for Ψ). All the optimal combinations have a threshold Ψ = 0.03 in the subset of models obtained when using $PSWb_i$ and $PSWa_i$, even when all other possibilities were considered. In the weighted design-based logistic regression model (first row of Table 3), a threshold Ψ = 0.03 was set because this value is the mean of the dependent variable.

Table 4. RMSE results of the estimated models when $Y_i = 1$

	Intervals	Weighted design-based model	PSWa (ε =0.4 and Ψ = 0.03)	PSWb (ε =0.6 and Ψ = 0.03)
RMSE $_1$	[0.005;0.012]	0.99039	0.99008	0.99046
RMSE $_2$	(0.012;0.015]	0.98674	0.98643	0.98674
RMSE $_3$	(0.015;0.018]	0.98372	0.98341	0.98370
RMSE $_4$	(0.018;0.021]	0.98080	0.98006	0.98099
RMSE $_5$	(0.021;0.025]	0.97698	0.97386	0.97707
RMSE $_6$	(0.025;0.029]	0.97255	0.97152	0.97269
RMSE $_7$	(0.029;0.034]	0.96890	0.96908	0.96909
RMSE $_8$	(0.034;0.041]	0.96260	0.95959	0.96226
RMSE $_9$	(0.041;0.057]	0.95190	0.94667	0.95105
RMSE $_{10}$	(0.057;0.163]	0.92421	0.91959	0.92285

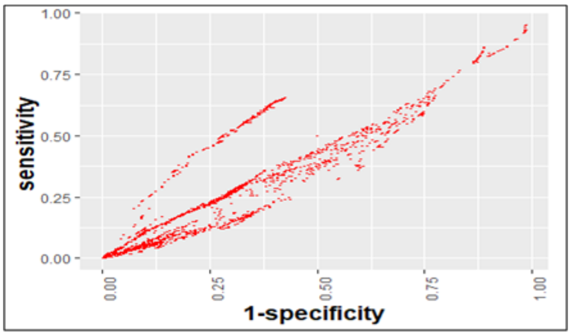

Fig. 1. The classification performance (sensibility and 1-specificity) of the estimated weighted logistic regressions with $PSWa_i$ when ε varies.

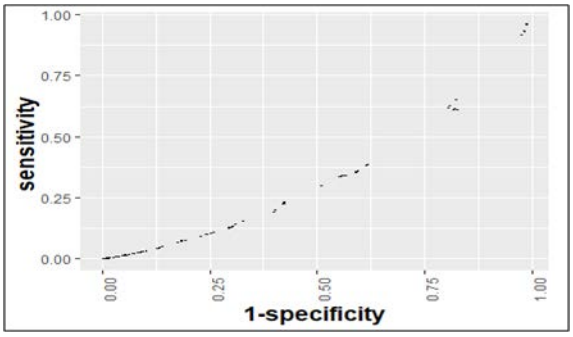

Fig. 2. The classification performance (sensibility and 1-specificity) of the estimated weighted logistic regressions with $PSWb_i$ when ε varies.

Thus, having selected the best adjustment parameters and thresholds that fulfil the proposed C-ROC criterion when using $PSWa_i$ and $PSWb_i$, we can conclude that the $PSWa_i$ with ε =0.1 and Ψ = 0.03 has the highest sensitivity and, thus, gives the best predictive performance in terms of the ROC criterion.

In Table 4, the RMSE was calculated for the lowest (RMSE1) to the highest (RMSE10) decile of predictions based on the best adjustment parameters under the C-ROC criterion solely for employees that had suffered a workplace accident ($Y_i = 1$).

Under RMSE criterion, the model estimated using $PSWa_i$, has smaller RMSE values than those of the other two models in Table 4. Although the improvement appears quite small, it is important to note that in this example only 3.11% of employees suffered an accident, which means this event is extremely rare. When we improve the sensitivity by only a few percentage points we obtain a significant impact on the global prediction performance, as events classed as workplace accidents might be hard to predict.

Taking all the results from the previous criteria, the weighting mechanism $PSWa_i$ is the best in terms of improving a model's predictive performance. This does not mean that $PSWb_i$ is not a suitable weighting mechanism; but, due to the type of exogenous variables in the model and the frequency of the dependent variable, $PSWa_i$ is more effective.

Figures 3, 4 and 5 show the predictions of the workplace accident and no workplace accident observations for each model (weighted design-based model, alternative *a* and alternative *b* with their optimal ε and Ψ). The proposed weighting mechanisms improve the predictive performance without producing abrupt or incoherent results. This outcome is also supported in Appendix 1, where the model parameter estimates are presented. In fact, all three figures seem to have a similar density distribution.

Figure 6 presents the histogram of predictions for observations that are equal to 1 for all three methods. Alternatives a (pink) and b (green) are located to the right of the histogram of predictions for the weighted design-based model (blue). It seems that alternatives a and b have a greater frequency of predictions equal to 1 for the observations that lie closer to the mean (0.031) and to the right of the figure. This seems to indicate that the predictive performance is improved, in the sense that it is more likely to detect cases Y_i = 1 under alternative a than it is in the other cases.

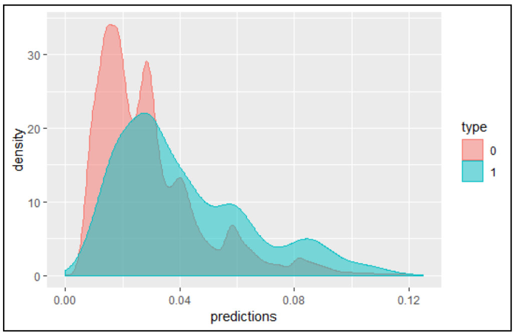

Fig. 3. Predictions obtained by the weighted model colored by Y_i = 1 and Y_i = 0.

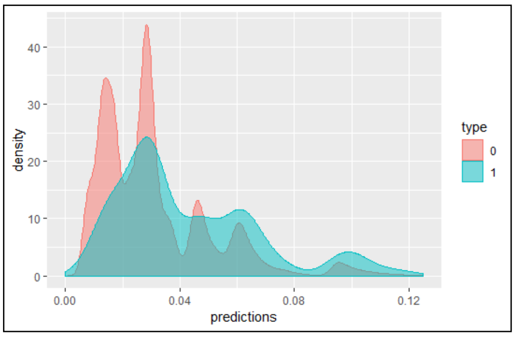

Fig. 4. Predictions obtained by the weighted model with *PSWa* (ε = 0.4) colored by Y_i = 1 and Y_i = 0.

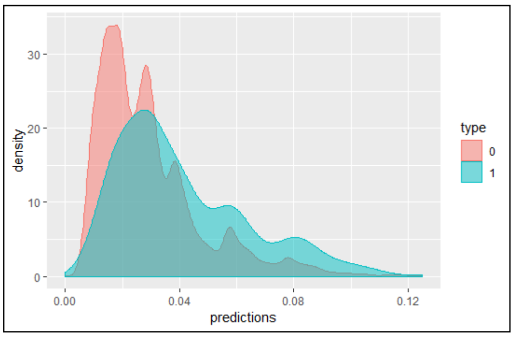

Fig. 5. Predictions obtained by the weighted model with *PSWb* (ε = 0.6) colored by Yi = 1 and Yi = 0.

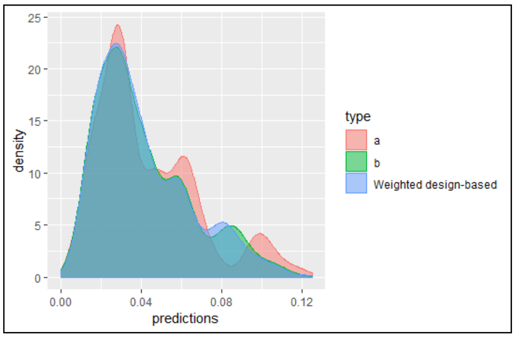

Fig. 6. Predictions for the observations that are equal to 1 of the unweighted model, alternatives *a* and *b*.

5. Conclusions

Our main conclusion is that the methods proposed can improve the predictive performance of logistic regression classifiers in survey data and that this is specially so for most deciles of the predictive distribution. We have compared two weighted procedures with the baseline model and shown that the choice of a specific weighting parameter, together with that of the threshold, leads to better accuracy than that obtained with the weighted design-based logistic regression model.

Moreover, we have proposed the ROC optimal criterion and the C-ROC optimal criterion as alternatives for measuring the predictive performance of a weighted estimation. Their standard procedures can be replicated in similar cases that seek to predict rare binary events.

We have found evidence that predicting the outcome response for respondents of a survey asked whether or not they had suffered a workplace accident can be improved for these individuals in all deciles of the prediction. This means that *PSWa* is able to predict individuals whose characteristics lie farther from the mean values. This result shows that the discrimination capacity can be improved by underweighting or overweighting observations, even if they already carry a sample weight.

Our analysis has a number of limitations. First, we might have implemented a cross-validation exercise by leaving part of the sample out of the estimation process. In this way, we could then have tested the model performance on a test sample; however, the proportion of ones in the dependent variable is so small that the test sample presents a serious lack of events (employees with accidents). Second, we deal here with a phenomenon that has a very low frequency because only a small fraction of the respondents suffered a workplace accident. We wonder if the results might differ when analyzing phenomena that are more frequent. However, the method described shows that the score (probability of a response equal to 1) obtained under alternative *a* or *b* provides an index of risk which gives more accurate predictions for workers and that it can serve as a measure of workplace safety. In short, our method can be used to identify those workers at greatest risk of suffering an accident in the workplace.

Further research needs to be dedicated to the definition of combined weights. Here, we have proposed multiplying sampling weights with predictive weights with a previous rescaling. Other alternatives, such as standardization or geometrical averaging, could also be explored.

Acknowledgements

We thank the Spanish Ministry of Economy, FEDER grant ECO2016-76203-C2-2-P. MG thanks the ICREA Academia program.

References

[1] C. Field, B. Smith, Robust estimation: A weighted maximum likelihood approach, International Statistical Review/ Revue Internationale de Statistique (1994), 405-424.

[2] E. Frees, R. Derrig and G. Meyers, Predictive modeling applications in actuarial science. Cambridge University Press, 2014.

[3] P.J. Green, Iteratively reweighed least square for maximum likelihood estimation and some robust and resistant alternatives, Journal of the Royal Statistical Society. Series B (Methodological) (1984), 149-192.

[4] W. Greene, Econometric Analysis, Prentice Hall, New York, 2002.

[5] J. Hanley and B. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1) (1982), 29-36.

[6] INEC, Matriz de Transición Laboral – Documento Metodológico Encuesta de Empleo, Desempleo y Subempleo, December 2017. 2017.

[7] G. King and L. Zeng, Logistic regression in rare events data, Political analysis 9(2) (2001), 137-163.

[8] L. Kish, Survey sampling, John Wiley & sons, New York, 1965.

[9] R. Lenth and P. Green, Consistency of deviance-based M-estimators, Journal of the Royal Statistical Society. Series B (Methodological) (1987), 326-330.

[10] T. Lumley, Analysis of complex survey samples, Journal of Statistical Software 9(1) (2004), 1-19.

[11] P. McCullagh, J.A. Nelder, Generalized Linear Models, Chapman and Hall, New York, 2nd ed, 1989.

[12] C. Manski and S. Lerman, The estimation of choice probabilities from choice based samples, Econometrica: Journal of the Econometric Society (1977), 1977-1988.

[13] R. Wedderburn, Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method, Biometrika 61(3) (1974), 439-447.

[14] C. Winship, L. Radbill, Sampling weights and regression analysis, Sociological Methods & Research 23(2) (1994), 230-257.

Appendix

Table A1 shows the results of the parameter and standard error estimates from the three logistic regression models (weighted design-based standard errors, weighted with *PSWa* and weighted with *PSWb*).

The results show that the coefficients of the weighted *a* and *b* models only change slightly with respect to the base weighted model. Standard errors, which are all design-based, are also similar.

Only the conclusion regarding the significant influence of the number of working days would differ if the *PSWa* weight were implemented.

In this case, we would conclude, therefore, that working hours do not have a significant effect on the probability of suffering a workplace accident.

Table A1. Final results of the estimates from the unweighted model, the model weighted with *PSWa* (ε = 0.4 and *Ψ* =0.03) and the model weighted with *PSWb* (ε = -0.25 and *Ψ* =0.03)

Variables	Weighted		PSWa		PSWb	
Intercept	-3.422	***	-2.880	***	-3.409	***
	(0.291)		(0.390)		(0.282)	
Urban	-0.338	***	-0.496	***	-0.371	***
	(0.095)		(0.123)		(0.100)	
Man	0.678	**	0.772	***	0.696	***
	(0.203)		(0.168)		(0.193)	
Marital (married)	0.428	*	0.525	***	0.457	**
	(0.165)		(0.137)		(0.153)	
Marital (others)	0.256		0.310		0.311	
	(0.251)		(0.278)		(0.290)	
Working hours	0.014	**	0.002		0.014	***
	(0.005)		(0.003)		(0.004)	
Workplace safety	-0.741	***	-0.780	***	-0.748	***
training	(0.114)		(0.163)		(0.115)	
Seniority	0.008		0.010		0.007	
	(0.006)		(0.007)		(0.005)	

The standard errors are shown in parentheses, and the significance of coefficients is given as follows: **.** ,*, **, *** correspond respectively, to the 0.05, 0.01, 0.001, 0 levels of significance.