A New Method for Weighting Survey Respondents

Feiming Chen
Nielsen
Chicago, IL

Abstract

In sample survey, we often need to adjust the weights of survey respondents to align marginal distributions of the sample to those of the population. Popular methods are poststratification, raking, generalized raking, etc. In this paper we present a new weighting method based on Tikhonov regularization and provide a R package for its implementation.

Keywords: sample survey, poststratification, raking, singular value decomposition, Tikhonov regularization.

1 Introduction

The weighting of survey respondents is often used to improve the accuracy of survey estimates. Typically each respondent is assigned an initial weight that is equal to the inverse of her sampling probability (inverse probability weighting). Further weighting adjustments may follow. In particular, poststratification is frequently used to adjust the weights with the information of known population distributions (such as those of Census). It works by multiplying the original weight w_i of each respondent i, who is in post-stratum h, by an adjustment factor $f_h = p_h/q_h$, where p_h and q_h are the population and sample proportions of the post-stratum h, respectively. This adjustment makes the sample distribution across the post-strata align to the corresponding population distribution. In general, there are two goals in the weight adjustments:

- Want alignment to known quantities (such as poststratification adjustment).
- Want weight adjustment factors to be small in some sense.

When there are multiple category variables to align to (such as age, sex, race, etc.) and we only know their marginal distributions, we can use *raking*, aka Iterative Proportional Fitting (**?**), to successively align marginal distributions of the sample to those of the population until convergence. Each raking step introduces a new multiplicative adjustment factor f_i. However, raking can be slow or fail to converge when there are many categories and thus many empty cells. In this paper we present a non-iterative weighting method based on Tikhonov regularization. It is relatively faster than raking and can give a reasonable result even when the sample is sparse across a large number of post-strata. .

There are many other methods on the modification of the original weights. The most notable one is the generalized raking procedure (**??**), which includes the classic raking method and the generalized least square weighting (GLS) as its special cases. Our approach differs from the above in that we use a different criterion and technique to generate the solution, which is unlikely to be a solution of the generalized raking. Other relevant literature on weighting can be seen in **?**.

2 Model

Suppose we collect a sample of size n. Each survey respondent i is assigned an initial design-based weight w_i, $i = 1, 2, \ldots, n$, which are assumed to be the inverse sampling probabilities. Suppose there are m categorical variables in the survey with known population distributions. We intend to adjust the initial weights such that the sample distribution in each category is as close to the corresponding population distribution as possible. With such adjusted weights, we can obtain an estimate of certain population total in the same manner as that of a Horvitz-Thompson estimator. That is, let y_i be the value of the variable of interest, and $\tilde{w}_i$ be the new weight, then we use

$$\hat{t}_{\tilde{w}} = \sum_{i=1}^{n} \tilde{w}_i y_i \tag{1}$$

to estimate the corresponding population total t. We will show this $\hat{t}_{\tilde{w}}$ is asymptotically equivalent to the Horvitz-Thompson estimator $\hat{t}_w$. For the following discussion, without loss of generality we scale w_i's so that $\sum_{i=1}^{n} w_i = 1$. Let A_j ($A_j \geq 2$) be the number of category levels for category j ($j = 1, 2, \ldots, m$). Then there are $H = \prod_{j=1}^{m} A_j$ post-strata. Each respondent falls into one and only one stratum. Without loss of generality, assume each stratum has at least one respondent. Let h_i be the stratum to which respondent i belongs. Then the total initial weights in stratum h is

$$u_h = \sum_{i:h_i=h} w_i, \quad h = 1, 2, \ldots, H$$

Suppose we adjust each respondent weight w_i by a multiplicative factor f_h (called *weight ratio*), where $h_i = h$. We wish to restrict f_h to a narrow range around one so as to make small weight adjustment. If we write $f_h = 1 + \beta_h$ where $\beta_h \geq -1$, then we want β_h to be close to zero. The total adjusted weights in stratum h is now

$$S_h = \sum_{i:h_i=h} w_i f_h = (1 + \beta_h) u_h, \quad h = 1, 2, \ldots, H.$$

Like the original weight w_i, we require the new weight $w_i f_h$, along with S_h, to sum up to one. This gives

$$\sum_{h=1}^{H} \beta_h u_h = 0, \tag{2}$$

since $\sum_{h=1}^{H} u_h = \sum_{i=1}^{n} w_i = 1$. Let I_{hjk} be a 0-1 indicator variable that is one if stratum h sets category j at level k, where $k = 1, 2, \ldots, A_j$. Let p_{jk} be the population proportion of level k in category j. Note $\sum_{k=1}^{A_j} p_{jk} = 1$ for $j = 1, 2, \ldots, m$. Then the requirement that the sample distribution be close to the population distribution in each category can be written as

$$\sum_{h=1}^{H} S_h I_{hjk} = \sum_{h=1}^{H} (1 + \beta_h) u_h I_{hjk} = p_{jk},$$

which can be rearranged as

$$\sum_{h=1}^{H} \beta_h u_h I_{hjk} = p_{jk} - \sum_{h=1}^{H} u_h I_{hjk}, \quad j = 1, 2, \ldots, m, \quad k = 1, 2, \ldots, A_j. \quad (3)$$

Note the right hand side of (3) is the difference between the expected proportion from the population and the observed proportion from the sample. Combining (2) and (3) gives us a linear system

$$Z = X\beta, \quad (4)$$

where, if we let $L = 1 + \sum_{j=1}^{m} A_j$, $\beta = [\beta_1, \beta_2, \ldots, \beta_H]'$,

$$Z = \begin{bmatrix} 0 \\ p_{11} - \sum_{h=1}^{H} u_h I_{h11} \\ \vdots \\ p_{mA_m} - \sum_{h=1}^{H} u_h I_{hmA_m} \end{bmatrix}_{L \times 1},$$

and

$$X = \begin{bmatrix} u_1 & u_2 & \cdots & u_H \\ u_1 I_{111} & u_2 I_{211} & \cdots & u_H I_{H11} \\ \vdots & \vdots & \ddots & \vdots \\ u_1 I_{1mA_m} & u_2 I_{2mA_m} & \cdots & u_H I_{HmA_m} \end{bmatrix}_{L \times H}.$$

The rank of X is at most $L - m$, so (4) is under-determined (in contrast to a usual regression) for solving β. However, since we want β_h's to be close to zero, we can use the so-called *Tikhonov regularization* (**?**) to penalize the least square solution that gives a large norm of β (this technique has also been used in ridge regression). Thus we minimize the object function

$$\chi^2 = ||Z - X\beta||^2 + r^2 ||\beta||^2, \quad (5)$$

where $||\cdot||$ represents the Euclidean norm. The regularization parameter r ($r > 0$) determines the trade-off between minimizing the residual sum of squares and minimizing the norm of the estimate. The minimizer $\hat{\beta}_r$ of (5) can be expressed in terms of the Singular Value Decomposition (SVD) of X (**?**). Let the SVD of X (note $H \geq L - m$) be $X = U\Sigma V'$, where $U = (U_1, \ldots, U_L)_{L \times L}$, $V = (V_1, \ldots, V_L)_{H \times H}$ and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_{L-m})_{L \times L}$ (assume $\mathrm{rk}(X) = L - m$). Then we have

$$\hat{\beta}_r = (X'X + r^2 I)^{-1} X' Z = \sum_{i=1}^{L-m} \phi_i \frac{U_i' Z}{\sigma_i} V_i, \quad (6)$$

where

$$\phi_i = \frac{\sigma_i^2}{\sigma_i^2 + r^2}, \quad (7)$$

which filters out right singular vectors V_i's for which the ratio of *signal* σ_i^2 to *noise* r^2 is much smaller than one. Remember the elements of $\hat{\beta}_r$ should be no less than -1, so our final estimate is $\tilde{\beta}_r = \max(\hat{\beta}_r, -1_H)$, where 1_H is a H-dim vector of one's and the function $\max$ is applied element-wise.

Note the storage space of V is $H \times L$ and the computation time of the SVD is $\mathcal{O}(6HL^2 + 20L^3)$ (**?**, R-SVD, page 254), which is acceptable if H is not too large. For example, if we have 8 categories, each of which has 5 levels, then $L = 41, H = 5^8 = 390625$. The storage of V is about 130MB with single precision (8 bytes) for each element. The computation of SVD takes about 40 seconds on a PC with 1.6GHz CPU.

The choice of the regularization parameter r can be determined via Generalized Cross Validation (GCV) (**?**) as the minimizer of the cross-validated prediction error. Equivalently, r is the minimizer of the GCV function

$$\mathcal{G} = \frac{||y - X\tilde{\beta}_r||^2}{(H - \sum_{i=1}^{L-m} \phi_i)^2},$$

where the denominator is the square of the effective number of degrees of freedom and ϕ_i's are the filter factors in (7). We use a golden section search to select the minimizer.

In practice, people may want as few zero weights as possible. Thus we may increase the selected r to reduce the number of elements of $\hat{\beta}_r$'s hitting the -1 floor, at the expense of compromising the marginal fitting result somewhat. Alternatively post hoc we can reset those small weight ratios to be at least as large as a specified non-zero number (say 0.01).

Although H is a large number when there are many categories, in reality not all post-strata are non-empty. Thus we can significantly reduce the number of elements in β (hence the size of X) by dropping the β_i's whose corresponding post-strata are empty, as shown in the second example in section 3.

Because of the explicit solution of $\hat{\beta}_r$ in (6), it is straightforward to observe that as sample size increases, Z goes to $\mathbf{0}$ while other terms in (6) are bounded, so $\hat{\beta}_r$ goes to $\mathbf{0}$ too. In more exact terms, assume the finite population size N, along with the sample size n, tends to infinity. Also assume that

1. $Z \to \mathbf{0}$ in design probability.
2. $Z = O_p(n^{-1/2})$.

Then, by noting that U_i and V_i are orthonormal vectors and that ϕ_i/σ_i is bounded, we have

Result 1. $\tilde{\beta}_r$ tends to $\mathbf{0}$ in design probability, and $\tilde{\beta}_r = O_p(n^{-1/2})$.

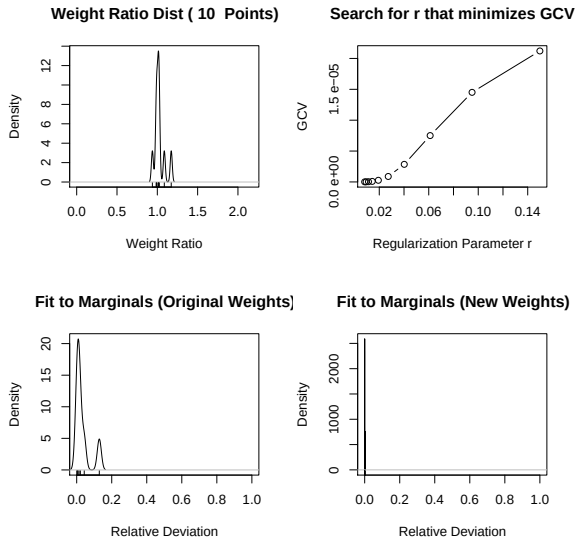
Result 2. The estimator $\hat{t}_{\tilde{w}}$ given by (1) is design-consistent, and $N^{-1}(\hat{t}_{\tilde{w}} - \hat{t}_w) = O_p(n^{-1/2})$.

Note Result 2 holds because of Result 1 and that $\hat{t}_w$ is design-consistent and $N^{-1}(\hat{t}_w - t) = O_p(n^{-1/2})$.

Table 1: Household Counts by Tenure and Household Size in Florida (Source: ACS PUMS 2004, SF1 2000 from US Census Bureau)

Size	Tenure		"True" Marginals
	Owner (1)	Renter (2)	
1 Person (1)	1185571	707097	1687303
2 Person (2)	1955017	568304	2330104
3 Person (3)	695037	356139	977117
4 Person (4)	605659	211830	776458
5+ Person (5)	363346	167639	566947
"True" Marginals	4441799	1896130	6337929

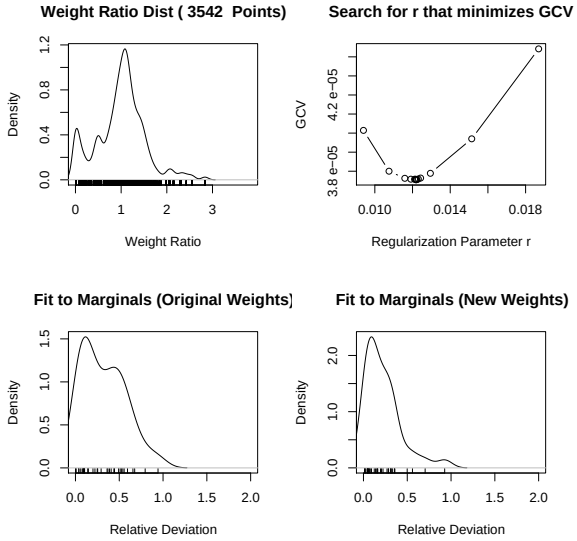
Figure 1: Diagnostic Plots for the Alignment to Two Categories.



3 Two Examples

We have implemented the above method in a R package called `reweight`, which can be downloaded from `http://www.r-project.org`. In the following we use this package to analyze two surveys. Table 1 presents a simple problem of marginal counts alignment for two categorical variables. Figure 1 shows the diagnostic plots for the weight adjustment. Note *Relative Deviation* is the relative deviation of each marginal count under the new weights away from the marginal count under the initial weights. For this "nice" data we can see the method fits to the "true" marginal counts closely. However, for a more complex data, the method performs less well, as shown in Figure 2. It is from a large survey that post-stratifies on 7 categories (the number of category levels are 5, 8, 6, 4, 4, 2, 5, respectively). Note there is only 1119 non-empty strata out of 38400 possible ones.

Figure 2: Diagnostic Plots for the Alignment to Seven Categories.



4 Conclusion

We have presented a new experimental method for reweighting survey data. It works by satisfying a set of constraints while at the same time controlling the size of the weight ratios with the method of Tikhonov regularization. We set the model up to address the common problem of marginal counts alignment in survey data. But in principal it can take any linear constraints. There are two drawbacks of this method when applied to a large survey data with many categories to align to. Firstly it cannot entirely close the gap between the observed marginal counts and the target ones. Secondly the computation of SVD may demand much computer resources. Nevertheless we encourage the reader to try out the R package to see how it performs on your problem.