A Heuristic Method for Weighting Survey Respondents

Feiming Chen
Spectra Marketing Systems
Chicago, IL
feimingchen@yahoo.com

August 10, 2006

1 Introduction

The weighting problem: a biased (e.g. convenience) sample needs to have its respondent weight adjusted so that it is more representative of the population.

In the sense that the marginal distributions of some variables fit more closely to those from a more precise source.

2 A simple illustration

The adjustment amounts to multiply the weight of each `male` respondent by 5/8, and multiply the weight of each `female` respondent by 5/2 (there is also a final scaling so that total weights is equal to the population total).

3 General Requirements

- Want alignment of the marginal distributions for multiple categorical variables (or more general linear constraints on the weights).
- Want weight adjustment factors to be small in some sense.

We want to satisfy the above two types of constraints simultaneously.

	Male	Female
Population Distribution	50%	50%
Sample Distribution (before adjustment)	80%	20%
Sample Distribution (after adjustment)	50%	50%

Table 1: A simple example of post-stratification

4 Literature Review

- Post-stratification: divide sample, then reweight using population frequency.
- *Raking* (Iterative Proportional Fitting) [1]
- Adjustment with propensity score [5]: logistic regression to generate propensity score as sampling frequency of subclass. Need a standard sample.
- Generalized raking [2]: constrained minimization that controls the size of the weight ratios. Flexible constraints.

5 Motivation and background

- Try to find an alternative to IPF in order to better control weight range and reduce computing time.
- Little research and not aware of other weighting methods at the time. :-(
- Start from brute force formulation and hope to solve it.

6 The basic idea

Set up a system of equations expressing the alignment constraints. Solve for the weight adjustment factors via *Tikhonov regularization* (analogous to ridge regression), which provides control of the size of the factors.

7 The weighting procedure

- Divide the sample into H post-strata. We shall adjust the weight of each respondent in stratum h, by a multiplicative factor of $f_h = 1 + \beta_h$ ($\beta_h \geq -1$). We desire β_h to be close to zero.
- The sum of new weights should be equal to the sum of the original weights. Let β be a vector of β_h's. That translates to:

  $$u'\beta = 0 \tag{1}$$

- Set up other linear constraints, such as the alignments to known marginal counts. Each constraint translates to:

  $$x'\beta = p - q = y \tag{2}$$

  where p is the expected count from the population and q is the observed count from the sample.

- Combining (1) and (2) gives us a linear system
$$X\beta = Y, \tag{3}$$
which is typically under-determined (in contrast to regression). The row dimension of X is the total number of linear constraints (let it be L). The column dimension of X is H, the number of post-strata.

8 Estimation via Tikhonov regularization

- Since we want β_h's to be close to zero, we can penalize the least square solution that gives a large norm of β by the so-called *Tikhonov regularization* [3]. That is, we minimize
$$\chi^2 = ||Y - X\beta||^2 + r^2||\beta||^2, \tag{4}$$
where r ($r > 0$) is a regularization parameter.
- Using the Singular Value Decomposition (SVD):
$$X = U\Sigma V',$$
the regularized estimate is
$$\begin{aligned}\hat{\beta}_r &= (X'X + r^2 I)^{-1}X'Y \\ &= \sum_{i=1}^{L} \phi_i \frac{U_i'Y}{\sigma_i} V_i,\end{aligned}$$
where
$$\phi_i = \frac{\sigma_i^2}{\sigma_i^2 + r^2},$$
which filters out V_i's for which the ratio of *signal* σ_i^2 to *noise* r^2 is much smaller than one.
- The final estimate is:
$$\tilde{\beta}_r = \max(\hat{\beta}_r, -1_H).$$
In practice, we can set lower and upper bounds (say [0.5,2]) to further restrict the range of weight ratios.

9 Computational resource

- Storage of V is $H \times L$.
- Computation of the SVD is $\mathcal{O}(6HL^2 + 20L^3)$.

Acceptable if H is not too large.

E.g. If we have 8 categories, with 5 levels each, then $L = 41, H = 5^8 = 390625$. V takes 130MB (8 bytes real). The SVD takes about 40 seconds on a PC with 1.6GHz CPU.

Size (np)	Tenure (ten)		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

Table 2: Household Counts by Tenure (ten) and Household Size (np) in Florida (Source: ACS PUMS 2004, SF1 2000)

10 Choice of the regularization parameter r

Determined via Generalized Cross Validation (GCV) as the minimizer of the GCV function [4]:

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

We use a golden section search method to select the minimizer.

11 A Toy Example

- US census data (PUMS and SF1) from `http://dataferrett.census.gov`.
- Two categories: Tenure (2 levels) and Household Size (5 levels)
- Task: adjust the PUMS household weight to fit to marginal counts.

12 A Large Example

- A survey with 7 categories and (5, 8, 6, 4, 4, 2, 5) levels. Only 1119 post-strata (other strata are empty) out of 38,400 possible ones.
- No original weights. Use raw counts.

13 Drawbacks

- No theoretical validation yet in terms of bias and variance.
- Unlike IPF, it cannot strictly maintain odds-ratios in the crosstab.
- It cannot fit to marginal counts closely for large problems unless the original weights are “good” enough.

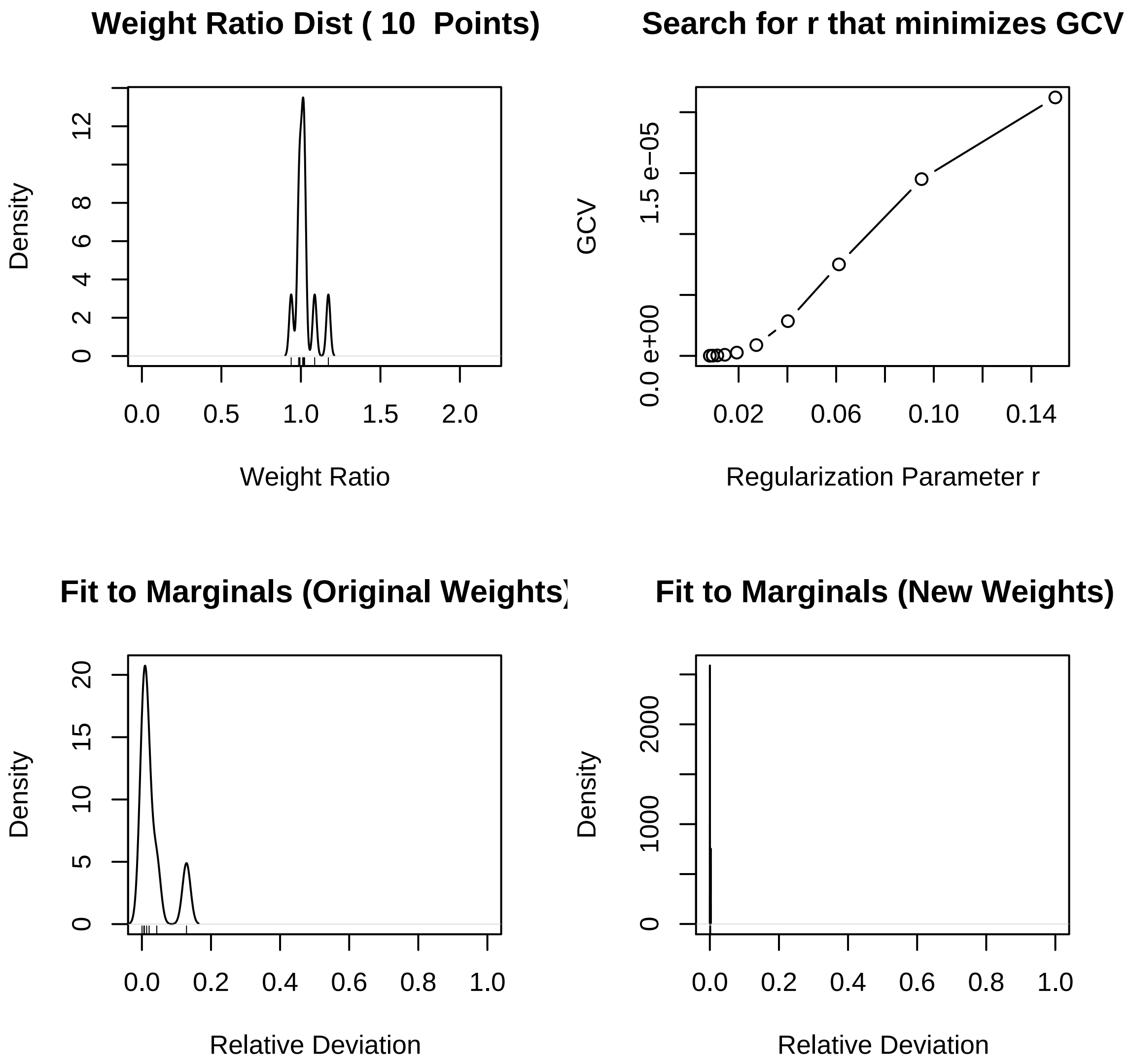

Figure 1: Diagnostic Plots For New Weights

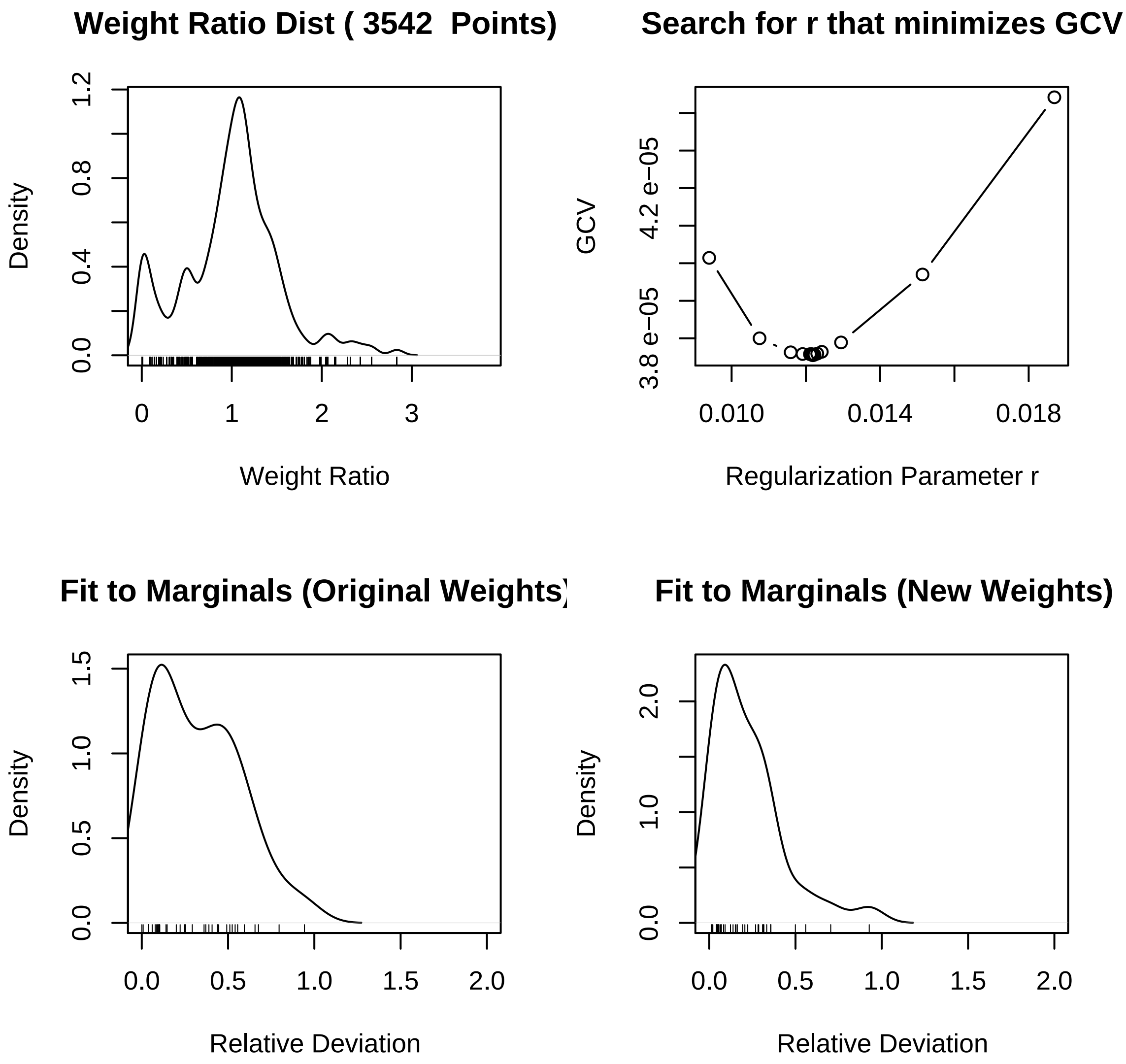

Figure 2: Diagnostic Plots For New Weights

- The range of weight ratios does not seem to be significantly narrower than other methods.

14 Summary

- An experimental method for reweighting survey data.
- Non-iterative method. But SVD may be computational intensive for large crosstabs.
- Can incorporate any linear constraints.
- R package `reweight` for download at:
  `http://www.r-project.org`.

15 Reference

References

[1] W. E. Deming and F. F. Stephan. On a least squares adjustment of a sample frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11:427–444, 1940.

[2] Jean-Claude Deville, Carl-Erik Sarndal, and Olivier Sautory. Generalized raking procedures in survey sampling. *JASA*, 88(423), 1993.

[3] G. H. Golub, P. C. Hansen, and D. P. O'Leary. Tikhonov regularization and total least squares. *SIAM Journal on Matrix Analysis and Applications*, 21:185–194, 2000.

[4] G. H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Tchnometrics*, 21(2):215–223, 1979.

[5] Paul R. Rosenbaum. Model-based direct adjustment. *JASA*, 82(398):387, 1987.