**A Unified Log-linear Modelling Framework for Post-Stratification and Raking Weights: an Application to the National Asian American Survey (NAAS)**

Prepared by: David Crow, Associate Director
University of California, Riverside
Survey Research Center

July 2009

## Introduction

The University of California, Riverside, Survey Research Center (SRC) developed and calculated survey weights for the National Asian American Survey (NAAS). This technical note enumerates the variables and data sources used in the weighting adjustments, offers a general description of the Iterative Proportional Fitting algorithm developed by Deming and Stephan (1940)—known as "raking"—and details its application to the NAAS data set.

## NAAS Weighting Adjustment Variables

The National Asian American Survey (NAAS) comprises 5,159 observations drawn from a national U.S. sample. The SRC adjusted NAAS sample proportions to population proportions estimated from the 2006-2008 three-year average of the Current Population Survey (CPS) and the 2007 American Community Survey (ACS), both conducted by the U.S. Census Bureau. In consultation with NAAS principal investigators, the SRC defined four categorical weighting adjustment variables:

- **ethnicity**, with seven (7) categories: Indian, Chinese, Filipino, Japanese, Korean, Vietnamese, and Other;
- **gender**, with two (2) categories: male and female;
- **education**, with three (3) categories: less than high school, high school graduate, and at least some college; and
- **nativity/citizenship**, with three (3) categories: foreign-born citizen, foreign-born non-citizen, and native-born citizen.

Weighting adjustments were carried out at six different levels of geographical aggregation:

1) nationally;
2) for each of four **census regions**;
3) for each of nine **census divisions**;
4) for seven **states** with sufficiently large concentrations of Asian Americans (California, Hawaii, New Jersey, New York, Texas, Virginia, and Washington);
5) for 13 Metropolitan Statistical Areas, or **MSAs**, with sufficiently large concentrations of Asian Americans (Washington, D.C., Dallas-Fort Worth,

Atlanta, Seattle, Sacramento, Philadelphia, New York City, San Francisco Bay Area, Honolulu, Chicago, Los Angeles, San Diego, and Houston); and

6) by **destination type** under two definitions in which all MSAs were categorized broadly as "new" or "traditional" destinations, then more finely as one of six types ("emerging", "re-emerging", "pre-merging", "former", "continuous", and "post-WWII").

These geographical categorizations yielded seven weight vectors, two of which correspond to the two definitions of destination type.

Since the number of adjustment cells was very large (7 x 2 x 3 x 3 = 126, multiplied by the relevant number of geographical units, e.g., 126 x 7 = 882 in the case of states), we estimated survey weights using a technique known as sample balancing or "raking", described in the following section.[1] In essence, raking corrects an imbalanced sample by adjusting sample totals (and proportions) to known totals in the population totals using the *marginal* distributions of weighting variables when a full cross-classification of these variables is unobtainable or impractical.

Post-Stratification and "Raking" Weights

The proportion of respondents in a sample with a given set of characteristics may differ from the proportion of people with those characteristics in the population for many reasons. These include sampling error, non-coverage error (when the sampling frame omits some members of the population), non-response error (when members of some subpopulations respond to a survey proportionally more than members of other subpopulations), and the sample design itself (as in the cases of oversampling and stratified samples with unequal sampling fractions). When for any of these reasons (or others) the proportion of survey respondents differ from their known population proportion on one or more characteristics, weighting is often used to adjust sample to population proportions. Auxiliary population information is typically obtained from census data or from other studies.

Post-stratification (PS) is a weighting technique that matches sample proportions to population proportions on characteristics chosen *after* data have been collected (see, e.g., Gelman and Carlin 2002; Kalton 1983: 74-75; and Lohr 1999: 114-115). First, the sample is divided into $H$ mutually exclusive categories, or strata, on the basis of observable characteristics such as gender, ethnicity, and age, where $H$ is the number of cells resulting from a complete cross-classification of the post-stratification variables. Then the proportion of sample members falling in cell $h$ ($p_h = n_h / n$) is multiplied by a weighting factor that makes the sample proportion equal to the proportion in the corresponding population "adjustment" cell ($\pi_h = N_h / N$). The PS weight for all sample units in cell $h$ is given by:

$$w_h = \frac{N_h / N}{n_h / n} = \frac{\pi_h}{p_h},$$

[1] Estimation of raking weights was performed using Stata's user-contributed "survwgt" routine, developed by Nicholas Winter, Economics, University of Virginia. The routine may be downloaded at: http://ideas.repec.org/c/boc/bocode/s427503.html.

where $N_h$ is the population size in cell $h$, $N$ is the total population size, $n_h$ is the number of sample units in cell $h$, and $n$ is the total number of sample units. The weights average to 1:

$$\overline{w}_h = \sum_{h=1}^{H} w_h n_h / n = 1$$

A cell weight over 1 indicates that the sample units in cell $h$ are upweighted to compensate for underrepresention (with respect to the population); conversely, a cell weight under 1 means the sample units in the cell are being downweighted to compensate for *over*representation.

But full post-stratification is unfeasible in many instances. The number of PS cells is often very large and can even exceed the sample size. Sparse data and zero cell counts make estimation of PS weights highly unstable or impossible. Also, complete cross-classifications of the population are often unobtainable and only marginal totals are known. A common solution to this problem is to "post-stratify on the margins", or "rake" the sample to marginal "adjustment" totals (Gelman 2007: 155; see also Deville et al. 1993).

Deming and Stephen (1940) proposed the "iterative proportional fitting" algorithm for estimating raking weights. The algorithm sets initial weighting factors for each cross-classification term to 1 and adjusts the weight factors for the first cross-classification term so that the weighted sample becomes representative with respect to the population (i.e., matches the marginal population distribution) for that term. The resulting weight factors are similarly adjusted for the next cross-classification term. Since this disturbs the representativeness of the other cross-classification terms, the process is repeated sequentially for each cross-classification term until the weighting factors converge (see Bethlehem 2002: 281 for a description of the algorithm).

So, for a two-way table with row variable $I$ and column variable $J$, each cell proportion in sample row $i$ is multiplied by a factor that makes the marginal proportion for that row ($p_{i+} = \sum_{i=1}^{I} p_i$) equal to the row marginal proportion in the population ($\pi_{i+} = \sum_{i=1}^{I} \pi_i$).[2] The resulting sample cell proportions in each column $j$ are then multiplied by a factor that makes the sample column proportions ($p_{+j} = \sum_{j=1}^{J} p_j$) equal to the population column proportions ($\pi_{+j} = \sum_{j=1}^{J} \pi_j$). The sample cell proportions are adjusted to the row marginals again, then the column marginals iteratively until they no longer change appreciably. The repeated lateral and vertical adjustments evoke a "raking" motion that gives its name to the algorithm.

Formally, Deming and Stephen proposed least squares minimization of a sum of residuals:

$$S = \sum (m_h - n_h)^2 / n_h,$$

where $n_h$ is the observed cell frequency in cell $h$, and $m_h$ is the calculated or adjusted cell frequency. Adjusted cell frequencies are constrained such that when summed across rows or columns, the sum is equal to the marginal totals. For the two-way case, the marginal constraints are:

[2] The raking procedure can be performed on cell frequency counts as well as proportions, substituting sample row ($f_{i+}$) and column ($f_{+i}$) marginal totals for sample marginal proportions, and population row ($F_{i+}$) and column ($F_{+j}$) marginal totals for sample marginal proportions. In this case, the resulting cell weights are equivalent to expansion factors. The Stata "survwgt" routine, in fact, rakes to marginal totals.

$$\sum_{j=1}^{J} m_{ij} = m_{i+} \text{ for } i=(1, \ldots, I) \text{ and}$$

$$\sum_{i=1}^{I} m_{ij} = m_{+j} \text{ for } j=(1, \ldots, J).$$

Deming and Stephen derived normal equations that allow for analytic solutions for two- and three-way contingency tables when some or all of the marginal totals are known. However, they proposed the "iterative proportions" algorithm described above, which allowed for much more rapid estimation of $m_i$—often converging after just two or three iterations.

Little and Wu (1991) showed that the raking weights can be estimated in the Generalized Linear Model (GLM) framework. For a two-way contingency table, the model is:

$$f\left(\hat{\pi}_{ij}/p_{ij}\right) = \mu + \mu_i + \mu_j$$

where $p_{ij}$ is the observed sample cell proportion, $\hat{\pi}_{ij}$ is the adjusted cell proportion corresponding to $m_h$ in Deming and Stephen's notation), $\mu$ is the intercept, the $\mu_i$'s are row effects (with ANOVA-type contrasts or dummy variable identifying constraints), and the $\mu_j$'s are column effects (also with identifying constraints). Two of the link functions they propose are the identity link, leading to the least squares estimators proposed by Deming and Stephen, and the natural logarithm, leading to a log-linear model. The raking weights are given by $\left(\hat{\pi}_{ij}/p_{ij}\right)$, and $\hat{\pi}_{ij}$ is estimated such that it is as close to 1 as possible and conforms to the marginal constraints given by Deming and Stephen (reformulated here in Little and Wu's notation):

$$\sum_{i=1}^{I} \hat{\pi}_{ij} = \pi_{i+} \text{ for } i=(1, \ldots, I)$$

$$\sum_{j=1}^{J} \hat{\pi}_{ij} = \pi_{+j} \text{ for } j=(1, \ldots, J).$$

Adjusting cell proportions to simple marginal totals—i.e., of single variables taken one at a time—is the equivalent of the log-linear independence model (subject to the marginal constraints noted above). Estimating raking weights as additive row and column effects "will not work well when interactions exist" (Little and Wu 1991: 88; see also Lang and Agresti 1994). For its part, full post-stratification can be formulated as the saturated model in the log-linear framework. In the two-way case:

$$ln\left(\hat{\pi}_{ij}/p_{ij}\right) = \mu + \mu_i + \mu_j + \mu_{ij},$$

where the last term denotes the interaction effects of row $i$ and column $j$. Extensions to multi-way tables are straightforward.

As noted above, however, full post-stratification is often unfeasible. A compromise solution is to post-stratify to "joint marginals"—that is, the joint probability of two variables assuming given values, independently of the values other variables assume. Post-stratification to joint marginals accommodates interaction effects among the raking variables while obviating the necessity for cell probabilities obtained by full post-stratification, difficult or impossible under conditions of sparse data.

So, for example, in a three way contingency table with variables *I*, *J*, and *K*, the joint marginal probabilities for *I* and *J* sum the joint distribution of all three variables over the values of *K*:

$$\sum_{k=1}^{K} \pi_{ijk} = \pi_{ij+}$$

If we wanted to rake to the joint marginals for each two-way interaction between *I, J*, and *K* ($\pi_{ij+}$, $\pi_{+jk}$, and $\pi_{i+k}$, respectively), the log-linear model is:

$$ln(\hat{\pi}_{ijk}/p_{ijk}) = \mu + \mu_i + \mu_j + \mu_k + \mu_{ij} + \mu_{jk} + \mu_{ik}$$

Note that this is equivalent to the saturated model with the parameter for the highest-order, three-way interaction ($\mu_{ijk}$) set to 0. This formulation of log-linear raking models (and raking models using other link functions in the GLM framework) is highly flexible, since any combination of lower-order interactions may be specified to calculate the raking adjustment weights.

NAAS Weights

The SRC estimating raking weights for the National Asian American Survey (NAAS) using the procedure for post-stratifying to joint marginals described in the preceding paragraphs. Recapitulating, there were four raking variables:

*I* = ethnicity (7 levels)
*J* = gender (2 levels)
*K* = education (3 levels)
*L* = nativity/citizenship (3 levels)

An additional variable, state (*M*), was used to estimate national-level raking weights.

Theory and prior empirical knowledge led the NAAS principal investigators to believe that the gender, education, and nativity composition of the sample would differ across ethnic groups. This belief implies three bivariate interactions: ethnicity x gender (*IJ*), ethnicity x education (*IK*), and ethnicity x nativity (*IL*). In addition, at the national level, the ethnic composition of the sample varies by state, resulting in the bivariate interaction ethnicity x state (*IL*).

*National Weights*

The national-level raking weights to four sets of joint marginals: *IJ, IK*, *IL*, and *IM*. The log-linear formulation of the raking model is:

$$ln(\hat{\pi}_{ijklm}/p_{ijklm}) = \mu + \mu_i + \mu_j + \mu_k + \mu_l + \mu_m + \mu_{ij} + \mu_{ik} + \mu_{il} + \mu_{im}$$

Note that the model includes no three-way or higher-order interactions, and omits most of two-way interactions.

*Census Region and District, MSA, and Destination Type Weights*

Since data in the NAAS sample (and in the CPS and ACS estimated totals) was much sparser at lower geographical levels of aggregation, a simpler model was used to estimate raking weights for these survey subdomains. Cell frequencies were raked to the joint marginal totals of ethnicity x gender (*IJ*), but only to the simple marginal totals for education (*K*) and nativity (*L*). The common model for all these subdomains is:

$$ln(\hat{\pi}_{ijkl}/p_{ijkl}) = \mu + \mu_i + \mu_j + \mu_k + \mu_l + \mu_{ij}$$

Conclusion

Raking is a well-established technique to adjust sample proportions (and frequencies) to known proportions (and totals) in the population when the full cross-classification of the sample is unknown or results in low cell counts. In these cases, post-stratification is impossible or unstable. However, raking to single marginal totals doesn't allow survey researchers for interactions that may exist between the raking variables. A suitable compromise is "raking" or "post-stratifying to joint marginals", the technique described in this note and used to estimate sample weights for the NAAS.

Bibliography


Bethlehem, Jelke G., 2002. "Weighting Non-Response Adjustments Based on Auxiliary Information", in Groves, Robert M., et al., eds., *Survey Nonresponse* (New York: Wiley and Sons), pp. 275-288.

Deming, W. Edwards, and Frederick F. Stephan, 1940. "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known", *The Annals of Mathematical Statistics*, Vol. 11, No. 4, pp. 427-444.

Deville, Jacques, Carl-Erik Sarndal, and Olivier Sautory, 1993. "Generalized Raking Procedures in Survey Sampling", *Journal of the American Statistical Association*, Vol. 88, No. 423, pp. 1013-1020.

Gelman, Andrew, and John B. Carlin, 2002. "Post-Stratification and Weighting Adjustments", in Groves, Robert M., et al., eds., *Survey Nonresponse* (New York: Wiley and Sons), pp. 289-302.

Gelman, Andrew, 2007. "Struggles with Survey Weighting and Regression Modeling", *Statistical Science*, Vol. 22, No. 2, pp.153-164.

Kalton, Graham, 1983. *Introduction to Survey Sampling* (Beverly Hills, CA: Sage Publications).

Lang, Joseph B., and Alan Agresti, 1994. "Simultaneously Modeling Joint and Marginal Distributions of Multivariate Categorical Responses", *Journal of the American Statistical Association*, Vol. 89, No. 426, pp. 625-632.

Little, Roderick J.A. and Mei-Miau Wu, 1991. "Models for Contingency Tables With Known Margins When Target and Sampled Populations Differ", *Journal of the American Statistical Association*, Vol. 86, No. 413, pp. 87-95.

Lohr, Sharon, 1999. *Sampling: Design and Analysis* (Pacific Grove, CA: Duxbury Press).