BUREAU OF THE CENSUS

STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR-84/03

DUAL SYSTEM ESTIMATION BASED ON ITERATIVE PROPORTIONAL FITTING

by

Beverley D. Causey

U.S. Bureau of the Census
Washington, D.C. 20233

Recommended by: Paul P. Biemer

Report completed: September 1, 1983

Report issued: January 19, 1984

DUAL SYSTEM ESTIMATION BASED ON ITERATIVE PROPORTIONAL FITTING

Beverley D. Causey

Abstract

For a dual-system match between files from the Current Population Survey and the Internal Revenue Service we obtain population estimates. To minimize the effects of "correlation bias" we form these estimates within cells as narrowly defined as possible; yet in order to enhance stability of estimation we use iterative proportional fitting rather than cell-by-cell estimation, to estimate cell probabilities of match.

Key words: capture-recapture; correlation bias; raking.

1. THE DUAL SYSTEM

The objective of this study is to estimate total U.S. population (ages 15-64) as of February 1978, both in entirety and by sex, race, region, etc. This study is one of several disparate approaches taken at the Census Bureau to this estimation problem, but here we will not be concerned with problems of aggregating a consensus of information based on these. We use a "dual system" based on a match between files for the Current Population Survey (CPS) for February 1978 and the Internal Revenue Service (IRS) for early 1978, corresponding to tax returns for 1977. The study is restricted to persons of ages 15 to 64. In this section we discuss the principle of dual system estimation and why we are led to use the "raking" procedure of Section 2. Section 3 disucsses some details of this particular study. Section 4 gives numerical results, with "jackknife" estimation of variance. Section 5 extends somewhat the ideas of Section 2.

Consider a group "g" of persons such as "female black, age 25-29, living in the West." We want to estimate the size of "g." Let: T denote the (unknown) number of persons in group g; A be the event that a person picked at random from group g is included in the IRS file; B be likewise for the CPS file; Z denote the (known) number of group-g persons in the CPS file; R denote the (known) number of group-g persons in the IRS file; and Y denote the (known) number of persons represented in the CPS file who are also in the IRS file.

We want to estimate T. Apart from the fact that we do not know T, estimates of P(A), P(B), and P(AB) are given by Z/T, R/T, and Y/T. Thus <u>if</u> we assume P(AB) = P(A)P(B), we are led to form the equation Y/T = (R/T)(Z/T); hence we estimate that T = RZ/Y is the number of persons in group g. The aggregation of values of RZ/Y over a series of groups, based on (for example)

sex-race-ethnicity-age-region categories, will yield an estimate of total population for larger groupings, such as all female, all persons 25-29, or the entire U.S.

This is the principle of the dual system. There are some characteristics peculiar to our CPS and IRS files which we will discuss at the end of this section; but for the mement we (continue to) discuss the dual system at a general level. In forming estimates of total population based on aggregation over groups g, we assume that events A and B are independent within each group g. Hence we want to choose the groups g so that the consequences of departure from this assumption are held to a minimum. These "consequences" are best measured in terms of "correlation bias": bias in the estimator of total number of persons, aggregated over groups, that arises from dependence between events A and B <u>within</u> each group g. It may be demonstrated, as by Chandrasekar and Deming (1949), that: (1) if this dependence is positive (as is typical), and (2) if there is positive (weighted) correlation <u>across</u> the groups g between P(A) and P(B), then the <u>magnitude</u> of the inherent correlation bias will be less if we base our estimation of total persons on computing T for the individual g's separately and then aggregating, than if we aggregate the g's before computing T. (Note that if there is positive dependence between A and B, the <u>direction</u> of the correlation bias is toward understatement of population total.) This line of argument leads us to work with groups g as specific and detailed as possible.

(Besides the issue of independence, a lesser reason for using the most detailed groupings is that, if we do not, the results based on the various choices of arbitrarily chosen, less detailed groupings will not in general be mutually consistent; and there is little reason why they should be consistent. For example, our estimates of total number of persons in the West differed

by about 1 million for "g's" based on age, race, and sex, and then based on region and race/ethnicity.)

Before discussing problems of estimation, we briefly indicate how one might measure the inherent correlation bias for a population in hand broken into G groups "g." Let g = 1, ..., G enumerate the groups; let number of group-g persons be denoted: in total by $N_g(>0)$, event-A by N_{g1}, event-B by N_{g2}, events A and B by N_{g0}. In practice we of course do not have N_g in hand; in this paragraph we are only evaluating a known population. Letting $P_{gh} = N_{gh}/N_g$ for h=0,1,2, we would use the ratio of

$$\Sigma N_g(P_{g0} - P_{g1}P_{g2}) \qquad (1.1)$$

to the total $\Sigma N_g P_{g0}$. If we let $N_{.0}$ denote <u>total</u> number of persons in both files (A and B), independent of how the groups are formed, we obtain the measure

$$1 - (\Sigma N_{g1}N_{g2}/N_g)/N_{.0}. \qquad (1.2)$$

For a given population we would want to compute the value of (1.2) for different groupings g - remembering that in practice such groupings would be formed without knowledge of the values N_g. Typically, we think, this value will be positive, and smallest for the most detailed grouping. Ideally, of course, we should have a value 0.

Hence we will work with specific and detailed groups. In the study at hand we will, as already indicated, cross-classify by sex-race-ethnicity-age-region. We would cross-classify in even more detail (e.g., urban vs. rural, central city vs. suburb) if the information were readily available. An obvious difficulty, however, is that the denominator Y in RZ/Y will become small and subject to large rel-variation, or even zero and thus useless; thus precise estimation of a true "T," group by group, becomes difficult or impossible. Thus we will proceed as follows. Let p denote the probability P(A|B);

a natural estimator of p is $\hat{p} = Y/Z$, and our estimator of T, as above, is $R/\hat{p}$. We will thus, in Section 2, develop an alternative, always nonzero estimator of p which provides stability of estimation and yet permits us to consider groups g as specific and detailed as possible. The idea of Section 2 is suited particularly to our CPS and IRS files; in Section 5 we extend this idea.

Wittes (1970) examines bias in the dual system estimates that arises when, although the probabilities of inclusion in each data source differ by group ("stratum"), the estimates are not group-based. She shows that even with only two groups underlying the probability structure, the bias in the estimated total can be quite substantial. For the group-based estimator Cowan (1982) considers biases that arise from misclassification of population units into groups. Pollock (1982) examines, as we do here, the relationship between probability of inclusion in data source and variables exogenous to the dual system procedure - using a logit model that allows probability of inclusion to vary from group to group.

We have referred to a "CPS file" and an "IRS file." Our "IRS file," from which we obtain "R," is actually an approximate 20 percent random sample from the full IRS file; accordingly, we multiply each R by 4.999947. The CPS file is based on a relatively small sample of households. With each person there is associated a sampling weight corresponding to reciprocal of probability of inclusion in sample (usually about 1400 to 1800). For Y and Z, used in forming $\hat{p}$, we use, instead of counts of numbers of persons, sums of weights associated with these persons; thus we are attempting to depict an (estimated) population that our sample CPS file represents.

2. RAKING

We now will use iterative proportional fitting, or "raking" (Ireland and Kullback 1968) to obtain our alternative $\hat{p}$.

In our study we classified persons in four ways, as already indicated:

(A) Sex: (1) male, (2) female.

(B) Race-ethnicity: (1) Spanish surname (based on a list prepared as a separate project at the Census Bureau on the basis of geographic incidence); for other surnames, (2) white, (3) black, (4) other.

(C) Age: (1) 15-24, (2) 25-34, ..., (5) 55-64.

(D) Region: (1) Northeast, (2) North Central, (3) South, (4) West.

Also, we counted as "included in CPS" only those persons for which a valid social security number (SSN) could be found: without such a number, there is nothing to match a CPS person to the IRS file even if that person is present there. For persons in CPS we make a fifth distinction:

(E) whether a valid SSN was (1) available without search, or (2) found only after search.

For the CPS consider a 5-way table of (weighted) counts Z_{ijkhm}, equal to the sum of weights for persons falling into category i for classification A (e.g., i = 2 for female), j for B, k for C, h for D, and m for E. Let Y_{ijkhm} denote the same, except only for persons matched to IRS.

Let $Y_{ij...} = \Sigma_k \Sigma_h \Sigma_m Y_{ijkhm}$, and let $Y_{i.k..}$, $Y_{.jk..}$) etc. be defined similarly. By means of raking we in effect fit a set of factors $f_{ij...}$, $f_{i.k..}$, etc. - a factor defined for each of the $\binom{5}{2} = 10$ pairwise combinations of A, B, C, D, and E; these factors are chosen (in a way to be soon discussed) so that if we set

$X_{ijkhm} = Z_{ijkhm} f_{ij...} f_{i.k..}$, multiplied by 8 similar factors,

we have $X_{ij...} = Y_{ij...}$, etc., where $X_{ij...} = \Sigma_k \Sigma_h \Sigma_m X_{ijkhm}$, etc. This method provides, by criteria of information theory, the closest link between the table values Z_{ijkhm} and the desired marginal totals $Y_{ij...}$, etc. The sampling error in the resulting entries X_{ijkhm} will be tied to that of

the marginal totals $Y_{ij...}$, etc., rather than to individual Y_{ijkhm}; at the same time we are able to take into account "pairwise interactions" between the "factors" A, B, C, D, and E if we view the situation in terms of a loglinear model:

$$\log(X_{ijkhm}/Z_{ijkhm}) = a_{ij...} + ..., \text{ etc., for } Z_{ijkhm} > 0.$$

(Depending on circumstances one may at time wish to exclude some pairwise interactions, or even include some three-way.) Thus we set:

$$r_{ijkhm} = f_{ij...}f_{i.k..} \text{ multiplied by 8 similar factors,} \tag{2.1}$$

in such a way that we have $X'_{ij...} = Y_{ij...}$, where

$$X'_{ij...} = \Sigma_k \Sigma_h \Sigma_m Z_{ijkhm}r_{ijkhm}. \tag{2.2}$$

etc. without forming X_{ijkhm} as such. We may use "r" as "$\hat{p}$" for group ijkhm. Computationally, the raking procedure for our particular problem with $\binom{5}{2}$ pairwise combination, is defined by:

$$\begin{aligned} r^{(0)}_{ijkhm} &= 1, \\ r^{(1)}_{ijkhm} &= r^{(0)}_{ijkhm}\, Y_{ij...}/X'^{(0)}_{ij...} \;\forall\; i, j, ..., \\ r^{(10)}_{ijkhm} &= r^{(9)}_{ijkhm}\, Y_{...hm}/X'^{(9)}_{...hm} \;\forall\; h, m,; \end{aligned} \tag{2.3}$$

we then generate $r^{(20)}_{ijkhm}$ likewise from $r^{(10)}_{ijkhm}$, etc. with $\lim_{n\to\infty} r^{(10n)}_{ijkhm}$ equal to the desired r_{ijkhm}. Routinely, the desired convergence is obtained (for all practical purposes) in a very few cycles of (2.3).

With our marginal total $Y_{ij...}$ always positive, we will always have $r_{ijkhm} > 0$ always even when $Y_{ijkhm} = 0$. Having obtained r_{ijkhm}, we may also go through the same exercise except with Y_{ijkhm} replaced by

$$Y^*_{ijkhm} = Z_{ijkhm} - Y_{ijkhm}, \tag{2.4}$$

the sum of weights of CPS cases <u>not</u> matched to IRS. We obtain a corresponding set of fitted probabilities r^*_{ijkhm}. For all ijkhm we may use

both "r" and "1-r^*" as estimators of "p", the probability of match; our first approach was to form a weighted average of these two estimators, the weights independent of ijkhm. But at the suggestion of R. Fay of the Census Bureau, we adopted a more straightforward approach (for which the results appeared to be little changed). Let r/r^* be estimator of $p/(1-p)$, the odds of match to IRS; we then obtain $\hat{p}$, the estimator of p itself, to be

$$\hat{p}_{ijkhm} = r_{ijkhm}/(r_{ijkhm} + r^*_{ijkhm}). \tag{2.5}$$

As discussed in Section 1, our final estimator of number of persons in group G is $R/\hat{p}$. However, R, for IRS, is defined for all ijkh, without the E-classification. Thus $\hat{p}$ must be formed and used for all ijkh, as a peculiarity of this particular study. We use the natural estimator

$$\hat{p}_{ijkh} = \left(\sum_{m=1}^{2} Z_{ijkhm}\, \hat{p}_{ijkhm}\right)/\left(\sum_{m=1}^{2} Z_{ijkhm}\right) \tag{2.6}$$

and estimate

$$T_{ijkh} = R_{ijkh}/\hat{p}_{ijkh}. \tag{2.7}$$

The category "m=2" (SSN found only after search) is relatively sparse; yet by always combining it in (2.6) with the copious category "m=1," we avoid instability in the denominator (2.7). In the event that we have zero denominator and numerator in (2.6), we replace Z_{ijkhm} in (2.6) by $Z_{...m}$; this replacement is of little consequence, since R_{ijkh} is very small, or zero, whenever this step is needed.

Again - instead of just 5 factors (i,j,k,h,m) we'd <u>like</u> to have as many factors as possible. Even if the total number of cells in our complete-factorial framework becomes astronomical, we need only to work computationally with the (more manageable) nonempty cells.

ratios of sums of CPS weights, as already discussed, there is no major need, arising from exclusion of panel i, for adjustment of these weights. Then, based on jackknifing (Miller 1974),

$$(1/8)\ \Sigma\ (F_i - F)^2 \qquad (4.1)$$

estimates Var(F) (and is associated with a t-distribution with 7 d.f. to construct a confidence interval for the true population size). Furthermore, the estimated F may then itself be corrected by subtracting from it the amount

$$(7/8)\ \Sigma\ (F_i - F). \qquad (4.2)$$

We use this correction in the results which we present, but its effects are minor.

Table 1 shows estimators of population, with corresponding estimated standard deviations in parentheses underneath, for the 4 regions (D) corresponding to rows, and the 4 race-ethnicity categories (B) corresponding to columns.

5. EXTENSION

We know how many CPS persons are matched to IRS; and we thus know, at least for 20 percent of the IRS, how many IRS persons are matched to CPS. We do not, however, have CPS weights for all the persons in our (20 percent) IRS file. Thus in Section 2 we (for each G) estimated T by $R/\hat{p}$ with R number of IRS persons and $\hat{p}$ = (estimated) P(A|B). Apart from this problem of weights we might well have interchanged the roles of CPS and IRS and let R be number of CPS persons and $\hat{p}$ = (estimated) P(B|A). Thus we would obtain two equally plausible estimators of T; one might work with the average of these two estimators (which we suspect, will usually not differ greatly from each other). Also, to ensure evenhanded treatment of our two files, we might (1) let $\hat{p}^*$ denote the product of (estimated) P(B|A) as obtained in this manner; and (2) estimate T by $Y/\hat{p}^*$, with Y the number of persons in both files.

1. Estimators by Region and Race/Ethnicity

	Spanish Surname	White	Black	Other	All
Northeast	1300067.64 (18222.31)	28389483.50 (62880.65)	3183947.37 (55356.99)	369673.72 (9819.74)	33243173.75 (122984.94)
North Central	731857.07 (23189.17	34338027.00 (82560.47)	3504817.12 (46899.36)	247809.47 (13536.70)	38822506.50 (119093.78)
South	2188280.25 (22344.36)	36114711.50 (101677.30)	8530798.12 (47274.84)	373485.31 (12553.20)	47207274.00 (138212.30)
West	2973470.97 (25673.44)	20950512.00 (70565.00)	1465630.95 (20593.30)	1518573.16 (36454.03)	26908187.00 (107719.60)
All	7193675.94 (64001.72)	119792735.00 (256011.19)	16685194.25 (105467.46)	2509541.62 (52650.75)	146181i.00 (362709.95)

REFERENCES

Chandrasekar, C., and W.E. Deming (1949), "On a Method of Estimating Birth and Death Rates and the Extent on Registration," Journal of the American Statististical Association, 44, 101-115.

Cowan, C. D. (1982), "Modifications to Capture-Recapture Estimation in the Presence of Errors in the Data," unpublished report, Statistical Methods Division, U.S. Census Bureau.

Ireland, C.T., and S. Kullback (1968), "Contingency Tables with Given Marginals," Biometrika, 55, 179-188.

Miller, R.G. (1974), "The Jackknife - A Review," Biometrika, 61, 1-15.

Pollock, K.M. (1982), "The Use of Auxiliary Variables in k-Sample Capture-Recapture Experiments," unpublished report, Department of Statistics, North Carolina State University.

Wittes, J.T. (1970), "Estimation of Population Size: The Bernoulli Census," unpublished Ph.D. dissertation, Department of Statistics, Harvard University.