APPLICATION OF AN ITERATIVE REGRESSION TECHNIQUE TO A NATIONAL POTENTIAL CROPLAND SURVEY

J. Jeffery Goebel, Iowa State University

The problem of estimation in a sample survey when data are available from outside sources for several characteristics of the population is considered. The control totals are incorporated into the estimation through an iterative generalized least squares regression technique that is applicable for any sampling scheme. We apply the technique to a national survey concerned with estimating the acreage of potential cropland and the difficulties involved in developing the land into cropland. This study was motivated by a desire to establish our future agricultural capabilities and to identify programs that might be required to ensure a supply of food sufficient to meet national and international demands. The multi-stage sampling scheme employed systematic and cluster sampling. The goal of the design was to minimize the variances of the national estimates, subject to a fixed cost restriction and certain accuracy restrictions on regional estimates.

* * * * *

We consider the case where n multivariate and categorical (or discrete) observations are taken--the n observed units are selected by a multi-stage procedure that does not necessarily give each unit an equal probability of being included in the sample. We know the sampling procedure (including the probabilities of selection) and also several of the population figures for some of the categories. In other words, we assume our observations can be represented by an r-dimensional classification, where each dimension represents an attribute and the i-th attribute (or classification) has s_i categories. Let $P_{j_1,j_2,\ldots,j_r}$ be the proportion of the population whose first classification is j_1, whose second classification is $j_2,\ldots$, and whose r-th classification is j_r, where $j_i = 1,2,\ldots,s_i$ and $i = 1, 2,\ldots,r$. Then we know some of the subtotals such as the fraction of the population in category j_1.

It certainly seems desirable to include such information in the inference procedure. This has been recognized for a long time [see, for example, El-Badry and Stephan (1955)], and becomes increasingly relevant as the stockpile of data swiftly grows. Using the extra data should improve the estimates one wishes to derive from the sample. Not only can one significantly reduce the variances (as we have found in a separate study), but also the figures to be presented will match with other published figures.

Our situation is related to contingency table estimation with known column and row totals. Deming and Stephan (1940) were one of the first to consider this situation, where the $P_{i.}$'s and $P_{.j}$'s are specified and estimates for the cell entries are desired.

	1	2	. . .	s	Total
1	$\tilde{P}_{11}$	$\tilde{P}_{12}$	. . .	$\tilde{P}_{1s}$	$P_{1.}$
2	$\tilde{P}_{21}$	$\tilde{P}_{22}$	. . .	$\tilde{P}_{2s}$	$P_{2.}$
⋮	⋮	⋮		⋮	⋮
r	$\tilde{P}_{r1}$	$\tilde{P}_{r2}$	. . .	$\tilde{P}_{rs}$	$P_{r.}$
Total	$P_{.1}$	$P_{.2}$	. . .	$P_{.s}$	$P_{..}$

They assumed the $\{n_{ij}\}$ were multinomially distributed and minimized $S = \sum_{i=1}^{r}\sum_{j=1}^{s} w_{ij}(n_{ij} - n\tilde{P}_{ij})^2$ over $\{\tilde{P}_{ij}\}$ subject to the marginal restrictions $\sum_{i=1}^{r}\tilde{P}_{ij} = P_{.j}$ and $P_{i.} = \sum_{j=1}^{s}\tilde{P}_{ij}$, using $w_{ij} = \frac{1}{n_{ij}}$. They obtained the closed form solution

$$\tilde{P}_{ij} = \hat{p}_{ij}(1 + \lambda_{i.} + \lambda_{.j}) ,$$

where $\hat{p}_{ij} = n_{ij}/n$ and the λ's are obtained from simultaneous equations in the n_{ij}'s and controls; they also gave a quick simple iterative technique for construction of $\{\tilde{P}_{ij}\}$. Stephan (1942) improved on the iterative procedure. J. H. Smith (1947), under the multinomial assumption, developed a maximum likelihood estimator: $\hat{P}_{ij} = \hat{p}_{ij}(a_i + b_j)^{-1}$, where $\{a_i\}$ and $\{b_j\}$ are functions of the controls, $\{P_{i.}\}$ and $\{P_{.j}\}$.

Others have used different criteria for constructing estimators. El-Badry and Stephan (1955) derived generalized least squares estimates which are approximately equivalent to the maximum likelihood estimators. Ireland and Kullback (1968) proposed estimating the cell probabilities of the contingency table by using the theory of minimum discrimination information, [i.e., the discrimination information $I(\cdot) = \sum\sum P_{ij} \ln (P_{ij}/\hat{p}_{ij})$ is minimized]. Their estimates are iterative (and shown to converge) and best asymptotically normal. They also discussed higher dimensional (>2) contingency tables with various marginals known.

More recently Chen and Fienberg (1974) obtained iterative MLEs for the case where only one of the classifications is known for some of the observations (and controls are present). One can inject other types of constraints--either instead of the types of controls to be discussed here or in conjunction with them. Bishop, Fienberg, and Holland (1975) in their new book Discrete Multivariate Analysis discuss these additional constraints, plus many other facets of contingency table analysis.

Grizzle, Starmer, and Koch (1969) analyzed categorical data by linear models--hence having all the advantages of using weighted regression (the estimates, testing, etc.). They considered taking $\{n_{i.} : i = 1,2,\ldots,s\}$ samples from s multinomial distributions, each having r categories of response, and gave a noniterative procedure.

Let us now consider regression approaches in the estimation of categorical data. Suppose we have weights, say $\{a_i\}$, for each observation such that the weighted sample mean

$$\bar{y}_w = \frac{1}{n} \sum_{i=1}^{n} a_i y_i$$

is unbiased for $\bar{Y}$ = the population mean of the characteristic of interest, Y, where the $\{a_i\}$ are nonnegative and add to n . If k auxiliary variables are available (they can be either continuous, or "0-1," as in our case) and the population figures (means or totals) are known for them, we can then use the generalized regression estimator

$$\bar{y}_G = \bar{y}_w + (\underset{\sim}{\bar{X}} - \underset{\sim}{\bar{x}}_w)\underset{\sim}{b}_G , \qquad (1)$$

where

$$\underset{\sim}{b}_G = [\Sigma\, a_i \underset{\sim}{Z}_i \underset{\sim}{Z}_i']^{-1} [\Sigma\, a_i y_i \underset{\sim}{Z}_i] ,$$

$$\underset{\sim}{Z}_i' = (1,\ X_{i1} - \bar{x}_{w1},\ X_{i2} - \bar{x}_{w2},\ \ldots,\ X_{ik} - \bar{x}_{wk}),$$

and $\underset{\sim}{\bar{x}}_w$ is the vector of the weighted sample means of the k+1 X's, with all $X_{i0} = 1$. Equation (1) can be written as a linear function of the observed y_i's, say

$$\bar{y}_G = \sum_{i=1}^{n} w_i y_i . \qquad (2)$$

Note that under simple random sampling all the a_i are 1 and $\bar{y}_G$ is simply the ordinary regression estimator.

The actual algorithm we use is slightly different than this form--it includes differences from population means (instead of from the weighted sample means) and an additional initial "weight." The first-stage regression weight is $w_i^{(0)}$, where for i = 1,2,...,n,

$$w_i^{(j)} = a_i\left(\frac{1}{n} + u_i^{(j)}\right)\left(1 + n\,\bar{u}_w^{(j)}\right)^{-1} , \qquad (3)$$

$$u_i^{(j)} = g_i^{(j)}\ (\underset{\sim}{X}_i - \underset{\sim}{\bar{X}})'\ [\underset{\sim}{A}_j]^{\dagger}\ (\underset{\sim}{\bar{X}} - \underset{\sim}{\bar{x}}_w),$$

$$\underset{\sim}{A}_j = \sum_{t=1}^{n} a_t\, g_t^{(j)}\ (\underset{\sim}{X}_t - \underset{\sim}{\bar{X}})(\underset{\sim}{X}_t - \underset{\sim}{\bar{X}})',$$

$[\underset{\sim}{A}_j]^{\dagger}$ denotes the generalized inverse of $\underset{\sim}{A}_j$, and $\underset{\sim}{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ik})'$ is now k-dimensional. The $\{g_t^{(0)}\}$ are the additional set of weights mentioned above and are commonly all unity. There are some situations where improvements can be made by using nonconstant $g_t^{(0)}$'s. An example is stratified sampling, where the means are known only for the population, and not by strata. The weighted combined estimator given by Cochran (1963) is such that a_t is inversely proportional to the sampling fraction f_h for the stratum and

$$g_t^{(0)} = N_h(1 - f_h)/(n_h - 1),$$

where the t-th observation is from stratum h .

If $|u_i^{(j)}| > M/n$ for any i (where $.1 \le M \le 1$ is fixed), we iterate--do a $(j+1)^{st}$ step. For each distance $d_i^{(j)} = |4n\, u_i^{(j-1)}/(3M)|$ an adjusting weight $g_{(j)i}$ is computed as:

$$g_{(j)i} = \begin{cases} 1 & , \ 0 \le d_i^{(j)} < \frac{1}{2} \\ 1 - \frac{4}{5}(d_i^{(j)} - \frac{1}{2})^2 & , \ \frac{1}{2} \le d_i^{(j)} \le 1 \\ 4/(5d_i^{(j)})^{-1} & , \ 1 < d_i^{(j)} . \end{cases}$$

Then we set $g_i^{(j)} = \prod_{h=0}^{j} g_{(j)i}$ and use line (3) to compute a new set of $w_i^{(j)}$'s .

Iteration will cease when:

(i) the $\{u_i^{(j)}\}$ are such that $|u_i^{(j)}| > M/n$ for all "i," since then the distances $d_i^{(j)}$ are too large and the required limitations on the weights cannot be met; or

(ii) if a specified number of iterations have occurred; or

(iii) if $|u_i^{(j)}| < M/n$ for all "i."

The third condition is the desired reason for ceasing iteration, and the result will be regression weights with the following properties:

$w_i \ge 0$, i = 1,2,...,n;

$(1-M)\ \max_i\{w_i/a_i\} \le (1+M)\ \min_i\{w_i/a_i\}$;

$\sum_{i=1}^{n} w_i \underset{\sim}{X}_i = \underset{\sim}{\bar{X}} = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k)'$;

$$\sum_{i=1}^{n} w_i = 1 .$$

Also, the associated regression coefficients will be best asymptotically normal estimates.

Let us now consider a recently completed national survey which was analyzed using this regression technique. In early 1975 there was considerable national and international concern about the global food situation. An example is the following excerpt from an editorial in the Journal of Soil and Water Conservation: "Crucial to our nation's ability to provide increased food supplies is the availability of our agricultural resource base. ...Now as we look to the future, particularly in terms of world food trade needs and world food security, we must be concerned about the adequacy of our land resource base." The United States Department of Agriculture was receiving numerous requests for data on potential cropland from Congress and from others. Data on current land use and on the potential for new cropland were needed so that the Department of Agriculture could respond accurately to these inquiries. In spring 1975 the Statistical Laboratory at Iowa State University cooperated with the Soil Conservation Service and the Economic Research Service of the United States Department of Agriculture in designing a national sample for this purpose.

At the time of design, the types of data desired for the sample segments of land were:

(i) type of soil;
(ii) 1975 land use--cropland, pasture and range, forest, other land, urban, and water;
(iii) 1967 land use;
(iv) types of development problems that would significantly inhibit development for cropland;
(v) type of development necessary for conversion to cropland;
(vi) the potential for conversion to cropland within the next 10-15 years.

Later a seventh variate was added:

(vii) whether or not the sample point is prime farmland.

The seven variables are all categorical. Items (i) and (iii) had been previously estimated. Acreages classified by "land capability unit class and subclass" and by "1967 land use" were available for each state from the 1967 Conservation Needs Inventory [see reference (12)]. Therefore, it was decided to use these data as "control" data. Items (ii) and (iv) through (vii) were collected on the sample sites.

The problem of sampling the nation to obtain acreage estimates has been studied--and a sample existed. However, the existing sample was designed to provide answers at the county level. Time and money were not available to study potential cropland in such detail and a new sample was designed. Usable estimates were desired for all 50 states plus Puerto Rico and the Virgin Islands. Costs dictated a national sample of about 500 counties. Each selected county was given roughly the same number of sample points. This was to equalize the work load because regular SCS field personnel were to do the work. Also, it was necessary to specify the number of sample counties per state immediately so that funds could be dispersed. The allocation per state was based upon

(i) the number of counties in the state,
(ii) the size of the state, and
(iii) the acreage in cropland in 1967.

For example, Massachusetts was given five counties, Illinois and Iowa 16 each, while 28 counties were selected in Texas.

A brief synopsis of the sampling scheme follows. The universe for this study consisted of the 1.44 billion acres (in the 50 states, Puerto Rico, and the Virgin Islands) considered to be in inventory for the 1967 Conservation Needs Inventory [see reference (12)]. Inventory acres were basically all rural land, except for federal land not cropped. Within each state (plus Puerto Rico and the Virgin Islands), counties served as clusters and were selected on a systematic basis. Some of the 506 counties included in the Potential Cropland Survey were divided into substrata due to their heterogeneity. Secondary units were then selected within substrata within counties using a systematic scheme. A total of 5,300 secondary units were selected. The secondary units are square areas of land, typically 160 acres in size. The major exceptions are: (i) in the northeastern states, 100 acre units were used; and (ii) in the western and mountain regions, some 40 acre squares were used in irrigated areas, while 640 acre units were used in some nonirrigated areas. The ultimate sampling units were points selected on a two-dimensional systematic basis within secondary units. As discussed, for example, by Strand and Huang (1973), this use of points does not seem to appreciably affect the accuracy of acreage estimates relative to completely observing and mapping the 160 acre secondary units.

For the survey on potential cropland, the regression weights technique was applied to data for each state. Some combining of sample points was done in an effort to reduce computational costs. In effect "new observations" were formed that combined the information by soil grouping (land capability class and subclass), 1967 land use, and substrata within counties. Note that acreages for the soil groupings and 1967 land uses serve as the control figures, while the substrata within counties are geographical regions where homogeniety is expected. The $\{a_t\}$ were constant for each substrata within a county and depended upon the sampling rates.

Two tables of national estimates are included as examples of estimates produced. Of particular interest are: (i) the 7% decrease from 1967 to 1975 in land classified as cropland; and (ii) that cropland acreage could be increased by at least a fourth if all land with high and med-

ium potential was converted to cropland. Variance estimates are being computed for the state, regional, and national estimates derived for this survey.

ACKNOWLEDGMENT

This research was supported in part through Cooperative Agreement 12-10-001-92 between Iowa State University and the Soil Conservation Service, United States Department of Agriculture.

REFERENCES

(1) Bishop, Yvonne M. M.; Fienberg, Stephen E.; and Holland, Paul W. (1975). Discrete Multivariate Analysis. The MIT Press: Cambridge, Mass.

(2) Chen, Tar and Fienberg, Stephen E. (1974). "Two-dimensional contingency tables with both completely and partially cross-classified data." Biometrics 30, 629-641.

(3) Cochran, William G. (1963). Sampling Techniques. John Wiley and Sons: New York.

(4) Deming, W. E. and Stephan, F. F. (1940). "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known." Ann. Math. Statist. 11, 427-444.

(5) El-Badry, M. A. and Stephan, F. F. (1955). "On adjusting sample tabulations to census counts." J. Amer. Statist. Assoc. 50, 738-762.

(6) Grizzle, J. E.; Starmer, C. F.; and Koch, G. G. (1969). "Analysis of categorical data by linear models." Biometrics 25, 489-504.

(7) Huang, Elizabeth; Goebel, J. Jeffery; and Fuller, Wayne A. (1974). "Regression M-weights: Computer algorithm." Statistical Laboratory Survey Section, Iowa State University: Ames, Iowa.

(8) Ireland, C. T. and Kullback, S. (1968). "Contingency tables with given marginals." Biometrika 55, 179-188.

(9) Smith, J. H. (1947). "Estimation of linear functions of cell proportions." Ann. Math. Statist. 18, 231-254.

(10) Stephan, F. F. (1942). "An iterative method of adjusting sample frequency tables when expected marginal totals are known." Ann. Math. Statist. 13, 166-178.

(11) Strand, Norman V. and Huang, Her Tzai (1973). "Conservation Needs Inventory: Influence of sample size and form of estimator on sampling variation in selected counties in Iowa, 1957 and 1967." Statistical Laboratory Survey Section, Iowa State University, and Soil Conservation Service: Ames, Iowa.

(12) U. S. Department of Agriculture (1971). Basic Statistics--National Inventory of Soil and Water Conservation Needs, 1967. Statistical Bulletin 461.

TABLE 1: CHANGES IN LAND USE: 1967 VERSUS 1975 (ESTIMATED ACRES x 1,000)

Land use in 1967	Land use in 1975						
	Cropland	Pasture and range	Forest	Other land	Urban	Water	Total
Cropland	351,651	52,884	8,265	12,977	4,846	618	431,241
Pasture and range	31,907	442,352	14,096	14,178	3,211	1,111	506,855
Forest	11,027	62,469	348,681	15,801	4,423	2,152	444,553
Other land	5,832	13,176	4,406	26,874	4,156	2,827	57,271
Total	400,417	570,881	375,448	69,830	16,636	6,708	1,439,920

TABLE 2: 1975 STATUS OF 1967 CNI ACREAGE BY LAND CAPABILITY CLASS AND SUBCLASS (ESTIMATED ACRES x 1,000)

Soil type (class & subclass)	Potential for cropland of 1975 pasture, range, forest, and other land				Cropland	Urban and water	Total
	High	Medium	Low	Zero			
1-	5,091	343	3,002	1,889	33,389	1,301	45,015
2E	19,987	3,750	25,486	8,194	87,594	3,151	148,162
2W	9,545	2,069	13,420	6,243	60,116	1,789	93,182
2S	2,117	268	2,652	1,211	20,451	625	27,324
2C	1,914	730	7,378	920	19,709	55	30,706
3E	17,239	8,605	56,455	16,564	70,351	2,035	171,249
3W	6,022	2,574	26,963	10,194	31,377	2,913	80,043
3S	2,357	1,257	9,034	2,706	11,455	320	27,129
3C	1,215	268	1,787	112	9,662	50	13,094
4E	5,866	4,385	52,243	20,664	29,693	1,795	114,646
4W	1,675	1,569	16,282	11,193	3,968	776	35,463
4S	842	767	12,719	7,181	5,904	728	28,141
4C	0	338	1,062	149	334	0	1,883
5W	476	1,992	14,319	9,779	1,497	892	28,955
5S	5	0	1,927	84	32	28	2,076
6E	3,202	2,394	33,313	123,635	8,794	1,511	172,849
6W	40	153	3,610	7,227	562	103	11,695
6S	442	1,084	21,087	51,850	3,502	444	78,409
6C	231	259	7,793	2,504	362	71	11,220
7-8	0	0	0	312,257	1,665	4,757	318,679
Total	78,266	32,805	310,532	594,556	400,417	23,344	1,439,920