POST-STRATIFICATION AND EFFICIENT ESTIMATION IN U.S. AGRICULTURAL LABOR SURVEYS

Raj S. Chhikara*, Charles R. Perry, Lih-Yuan Deng*, William C. Iwig, Floyd M. Spears*, and Susan Cowles

Charles R. Perry, National Agricultural Statistics Service
3251 Old Lee Highway, Room 305, Fairfax VA, 22030

Abstract

A multiple frame consisting of both a list and an area frame is used in most agricultural surveys conducted by the National Agricultural Statistics Service of the U.S. Department of Agriculture. The area frame sampling is used to account for the lack of coverage by the list frame sampling. Typically, the area frame estimate of the portion not on the list (NOL) has low precision. To improve upon its precision, certain alternative approaches were investigated. This study of the Agricultural Labor Survey involved post-stratification of all list sample respondents and all NOL sample respondents in California and Florida for the 1991-92 surveys based on some auxiliary data available from the 1991 June Enumerative Survey. A difference estimator was also developed for the NOL portion using the auxiliary variables. No improvement in the precision of the multiple frame estimate was realized under the straightforward application of post-stratification. On the other hand, the difference estimation in conjunction with post-stratification led to some gain in precision and hence proved to be a better approach.

KEY WORDS: Multiple Frame, Post-Stratification, Difference Estimation

1 Introduction

The National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture (USDA) conducts quarterly Agricultural Labor Surveys to estimate the number of hired, self-employed and unpaid workers and associated hours per week and wage rates. Both a list frame and an area frame are used to sample agricultural operations or area segments and a multiple frame estimate is obtained by summing the estimates for the list and area frame components. The area frame component estimate is for the area not overlapped with the list and is labeled as NOL. The estimates are developed at the state or regional level.

The area frame sampling for the NOL is used primarily to compensate for the noncoverage of a target population in the list frame. The NOL sample is a subset of the June Enumerative area frame survey sample. All area frame sample farms are checked against the list frame to determine the nonoverlapping (NOL) sample units. A subsample of these base NOL farms is then used for the quarterly labor surveys. The NOL estimate accounts for a smaller portion (less than 30 percent) of a survey estimate than the list frame. However, it accounts for a substantially larger portion of the variance of a multiple frame estimate. Among other factors, the NOL has relatively a much smaller sample size than the list and this would make the NOL estimator less reliable.

Certain alternative methods have been investigated to improve upon the reliability of the NOL estimator or to develop estimates based only on the list frame. The basic approach is to post-stratify both the list and the NOL sample data by farm type or other available auxiliary information and then obtain a post-stratified estimate for the list and NOL. Since the current list estimator is viewed to be reliable, we only considered developing a new estimator for the NOL component. For details on post-stratification, one may refer to Rumburg, et al. (1993).

Besides the post-stratified estimator, a difference estimator is developed using additional auxiliary variables in conjunction with post-stratification based on farm type. The estimators investigated are described in their general form in the next sec-

* These authors' work was supported in part under a cooperative research program of the United States Department of Agriculture at the University of Houston-Clear Lake.

tion and in their applied form in Section 3. The 1991-92 agricultural survey data from California and Florida are used to compute and evaluate these estimators. The numerical results are given in Section 4 and show that the precision of the post-stratified NOL estimate in the two largest agricultural labor states is less reliable than the current estimator unless the difference estimator is used in conjunction with post-stratification.

2 Estimators

The estimation approach in its most generic form can be formulated as follows: Let a population of N units consist of H strata with N_h units in stratum h, $h = 1, 2, ..., H$. Suppose n_h sample units are selected in stratum h and $n = \sum_{h=1}^{H} n_h$ is the total sample size for the survey. Next, let the sample units be post-stratified into K post-strata determined using some auxiliary information obtained for the sample units during the survey. Suppose n_{hk} is the number of sample units that correspond to stratum h and post-stratum k, and $n_{.k} = \sum_h n_{hk}$, $k = 1, 2, ..., K$. If y_{hi} is the sample response of interest for the ith sample unit in stratum h, then the population total, Y, can be estimated in two ways:

(a) Stratified Estimator

$$\hat{Y}_s = \sum_{h=1}^{H} N_h \bar{y}_h \qquad (1)$$

where $\bar{y}_h = \sum_{i=1}^{n_h} \frac{y_{hi}}{n_h}$.

(b) Post-Stratified Estimator

$$\hat{Y}_{ps} = \sum_{k=1}^{K} \hat{N}_{.k} \hat{\bar{Y}}_k \qquad (2)$$

where $\hat{N}_{.k}$ is the estimated size of post-stratum k and $\hat{\bar{Y}}_k$ is an estimator of mean response for post-stratum k. Both $\hat{N}_{.k}$ and $\hat{\bar{Y}}_k$ should be determined using an approach that would yield $\hat{Y}_{ps}$to be an efficient estimator.

In the present study, we obtain $\hat{N}_{.k}$ by summing the weights associated with the n_{hk} sample units for $h = 1, 2, ..., H$, that correspond to post-stratum k. A sample unit weight is inversely proportional to its probability of being selected. The estimator $\hat{\bar{Y}}_k$ is obtained either by the post-stratum sample mean or prediction mean using certain regressors as follows:

$$\hat{\bar{Y}}_k = \frac{1}{n_{.k}} \sum_{i \in U_k} y_{hi} = \bar{y}_{.k}$$

where U_k represents the set of all sample units falling in the kth post-stratum. Accordingly, the post-stratified estimator of Y is given by

$$\hat{Y}_{ps} = \sum_{k=1}^{K} \hat{N}_{.k} \bar{y}_{.k}. \qquad (3)$$

Suppose there are additional auxiliary variables, say $x_1, x_2, ..., x_m$, which are expected to be linearly correlated to the response variable y. Based on a prior or non-overlapping data set, one can estimate the regression equation given by $\hat{y} = b'x$, where b is the column vector of estimated regression coefficients and x is the column vector of values of the auxiliary variables. The estimator $\hat{\hat{\bar{Y}}}_k$ can be obtained by

$$\hat{\hat{\bar{Y}}}_k = \frac{1}{n_{.k}} \sum_{i \in U_k} \hat{y}_i = \bar{\hat{y}}_{.k}$$

and the corresponding post-stratified estimator is

$$\hat{\hat{Y}}_{ps} = \sum_{k=1}^{K} \hat{N}_{.k} \bar{\hat{y}}_{.k} \qquad (4)$$

3 Estimation In the Agricultural Labor Survey

The stratified and post-stratified estimators defined in the previous section are investigated and evaluated using the 1991-92 Agricultural Labor Survey data from California and Florida. Auxiliary data available from the 1991 June Enumerative Survey (JES) for the farm value of sales in 1990 (x_1) and the peak number of workers in 1991 (x_2) are used in addition to the farm type information for developing post-stratification. Farm type refers to the largest source of gross income for the operation, such as cash grains, fruits, livestock, dairy, etc. The data for the variables x_1 and x_2 are used in two different ways: (i) Post-stratification based on farm type was further refined using the data for auxiliary variables x_1 and x_2 and (ii) a least-squares regression fit for the number of hired workers (y) was made in terms of x_1 and x_2 as regressors using the list sample data corresponding to each original post-stratum. Thus, when the two estimators defined in Equations (3) and (4) were computed corresponding to cases (i) and (ii), they would utilize equivalent auxiliary information and be comparable. Specifically, the following three estimators were evaluated and compared for their bias and precision.

3.1 Direct Expansion Estimates

The stratified estimates and their cv's were computed separately for the list and NOL components for each of the four quarters ending in July 1991, October 1991, January 1992 and April 1992. In each case, this estimate is simply an aggregation of the expanded strata sample means as described in Equation (1) and is called the direct expansion estimate. The variance of the direct expansion estimate is that of the usual stratified estimator which is straightforward to estimate. Next, a multiple frame estimate is obtained by adding the list and NOL component estimates.

3.2 Post-Stratified Estimates

A multiple frame post-stratified estimate is obtained by combining the list and NOL sample data for each post-stratum and expanding directly the post-stratum sample means. Here, the combined (i.e., direct post-stratified multiple frame) estimate makes use of the sample observations obtained from both NOL respondents and list respondents in each post-stratum to estimate its mean. (A list-only post-stratified estimate is obtained by using only list frame data to estimate the post-stratum means, but it is not included in the present discussion.) These estimates use a weighted estimator of Y given by

$$\hat{Y}_{ps,wt} = \sum_{k=1}^{K} \hat{N}_{.k}\hat{\bar{y}}_{k,wt} \tag{5}$$

where $\hat{N}_{.k}$ is estimated from the June Enumerative Survey and

$$\hat{\bar{y}}_{k,wt} = \frac{\sum_{i\in U_k} w_i y_i}{\sum_{i\in U_k} w_i};$$

where w_i is the weight of the ith sample reporting unit that falls in post-stratum k.

The variance of this estimator is not easily tractable as it involves estimates of post-strata weights and their mean responses. A large sample variance formula can be derived using a Taylor series approximation as described in Perry, et al. (1993). We made use of the approximate variance formula given there to compute the standard error of a post-stratified estimate.

3.3 Difference Estimates

The least-square regression fits were obtained using the list sample data in each original post-stratum as stated in case (ii) in the introduction of this section. The linear regression equations for the different post-strata were then used to predict the mean response for all NOL sample units identified in the base June Enumerative Survey. The total number of hired workers for the NOL component was predicted by appropriately expanding each sample unit prediction and aggregating the expanded predictions across all NOL sample units. This NOL prediction is denoted by

$$\hat{Y}_o = \sum_{i\in U_J} \hat{y}_i e_{J,i} \tag{6}$$

where U_J denotes the set of all NOL sample units in JES, and $e_{J,i}$ denotes the expansion associated with the ith NOL sample unit in JES.

Clearly, $\hat{Y}_o$ is expected to be a biased estimator of Y since the regression fits based on the list sample data may not appropriately represent the NOL. An adjustment to $\hat{Y}_o$ is made based on the differences between the observed and predicted responses for the NOL sample units acquired in a quarterly survey (which are a subsample of the NOL units identified in June). This adjustment is an estimate of the bias and is obtained by

$$\hat{D} = \sum_{i\in U_L} (y_i - \hat{y}_i) e_{L,i} \tag{7}$$

where U_L is the set of NOL samples observed in a quarterly labor survey and $e_{L,i}$ is the expansion for the ith observed sample unit of the survey. The NOL component estimate is obtained by adjusting the NOL estimate $\hat{Y}_o$ by the estimated bias $\hat{D}$ so that the estimate is given by

$$\hat{Y}_{\text{NOL}} = \hat{Y}_o + \hat{D}. \tag{8}$$

Obviously, the correction for bias makes $\hat{Y}_{\text{NOL}}$unbiased.

A multiple frame estimate of Y is obtained by summing the list and NOL component estimates:

$$\hat{Y}_D = \hat{Y}_{\text{List}} + \hat{Y}_{\text{NOL}} \tag{9}$$

Since the current list component estimate is reliable, the estimate $\hat{Y}_{\text{List}}$was computed using the stratified estimator as defined in Equation 1. The variance of $\hat{Y}_D$ was approximated for the large sample case by considering $\hat{Y}_{\text{List}}$and $\hat{Y}_{\text{NOL}}$to be independent.

4 Empirical Results

All three estimates and their standard errors were computed using the 1991-92 survey data as discussed in Section 3. These estimates were compared with the published Agricultural Board estimates.

Table 1 lists the estimated NOL total number of hired workers and the corresponding cv's for the difference estimator and direct expansion estimator for California and Florida. Except for California in January and April of 1992 and Florida in July of 1991, the cv of the difference estimate is smaller than that of the direct expansion estimate.

In general, the difference estimator outperformed both the direct expansion and the multiple frame post-stratified estimates. Specifically, Rumburg et al. (1993) showed that the direct expansion estimator provides higher precision than the post-stratified estimator using data from 12 monthly 1991-92 surveys. This study shows the difference estimator performs favorably with the Board estimates and had substantially higher precision than the post-stratified estimates. Table 2 lists the estimated relative efficiency for each of the two alternative estimators. These numerical results clearly indicate that the difference estimator significantly outperforms the post-stratified estimator, and that it is more efficient than the current direct expansion estimator.

Although we used the same auxiliary information in the development of the post-stratified and difference estimators, the two estimates performed quite differently. Apparently, the use of variables x_1 and x_2 when categorized do not improve upon the post-stratification made based on the farm type, whereas x_1 and x_2 as regressors seem to be well correlated with the response variable within each farm type post-stratum. This was verified when we examined the results of regression analysis performed for each of the response variables, (i) number of hired workers, (ii) wages paid and (iii) number of hours worked, using the list samples in each state. Table 3 lists the R^2 values obtained for California. In most cases, the values of R^2 are significantly high and are indicative of the usefulness of variables x_1 and x_2 as predictors.

The results in Table 3 show that seasonality is an important factor in prediction of the response variable as a linear function of the farm value sales and the peak number of hired workers. In each case, the R^2 values are fairly consistent with that expected for a farm type in season. For example, the correlation is high in July, moderate in October and low in January for vegetable and fruit operations as one would expect.

5 Further Research

The post-stratified and difference estimators are being further investigated using the 1993-94 and 1994-95 agricultural labor survey data. An initial evaluation basically confirms the previous conclusion showing a poor performance of the post-stratified estimator, but the difference estimator providing estimates that are slightly better than the direct expansion estimates. As a result, the difference estimation procedure is being utilized to develop list-only estimates whereby one does not need to observe for the NOL samples. The results of this evaluation study will be reported at a later date.

References

[1] Perry, Charles; Chhikara, Raj; Deng, Lih-Yuan; Iwig, William and Rumburg, Scot. Generalized Post-stratification Estimators in the Agricultural Labor Survey, SRB Research Report No. SRB-93-04, Washington, D.C., July 1993.

[2] Rumburg, Scot; Perry, Charles; Chhikara, Raj S. and Iwig, William C. Analysis of a Generalized Post-Stratification Approach for the Agricultural Labor Survey. SRB Research Report No. SRB-93-05, July 1993.

Table 1: Total Number of Hired Workers for NOL

	Direct Expansion Estimator		Difference Estimator	
Period	Estimate	CV	Estimate	CV
	California:			
Jul 1991	43128	33.2	48244	27.4
Oct 1991	38988	40.4	33156	38.1
Jan 1992	34694	45.3	15486	47.2
Apr 1992	27006	32.8	29078	58.4
	Florida:			
Jul 1991	4072	34.4	9802	81.4
Oct 1991	17900	77.3	14002	42.0
Jan 1992	13201	70.2	10934	34.2
Apr 1992	14404	63.0	12308	29.7

Table 2: Estimated Relative Efficiency[‡] of Multiple Frame Post-stratified and Difference Estimators Relative to Direct Expansion Estimator for Estimation of Total Number of Hired Workers

State	Post-stratified Estimator	Difference Estimator
California	0.67	1.07
Florida	0.32	1.20

[‡] For the post-stratified estimator, the estimated relative efficiency is computed by squaring the ratio of the average cv's obtained from the twelve monthly surveys, whereas for the difference estimator, the average cv's are based on the four quarterly surveys during 1991-92.

Table 3: R^2 for List Samples in California Using Farm Value of Sales (x_1) and Peak Number of Workers (x_2) as Independent Regressors

	July 1991			October 1991			January 1992		
Farm type	Hired Workers	Wages Paid	Hours Worked	Hired Workers	Wages Paid	Hours Worked	Hired Workers	Wages Paid	Hours Worked
Grain	0.96	0.91	0.96	0.67	0.51	0.70	0.84	0.83	0.84
Cotton	0.91	0.93	0.90	0.84	0.79	0.80	0.59	0.36	0.29
Other	0.82	0.80	0.79	0.69	0.68	0.62	0.79	0.75	0.75
Vegetables	0.83	0.72	0.76	0.53	0.48	0.47	0.22	0.18	0.26
Fruits	0.86	0.90	0.90	0.66	0.71	0.69	0.32	0.24	0.23
Nuts	0.37	0.40	0.44	0.93	0.71	0.75	0.86	0.79	0.71
Nursery	0.96	0.92	0.86	0.91	0.77	0.76	0.53	0.34	0.35
Livestock	0.81	0.74	0.89	0.50	0.46	0.51	0.42	0.40	0.35
Poultry	0.87	0.88	0.89	0.97	0.97	0.97	0.98	0.99	0.99
Dairy	0.88	0.85	0.75	0.98	0.95	0.91	0.72	0.74	0.70