AN EVALUATION OF ALTERNATIVE USDA AGRICULTURAL LABOR SURVEY ESTIMATORS

Floyd M. Spears[1] and Raj S. Chhikara[1], University of Houston-Clear Lake
Charles R. Perry[2], and Susan Cowles[2], USDA-NASS
Floyd M. Spears, University of Houston-Clear Lake, 2700 Bay Area Blvd., Houston, TX 77058

KEY WORDS: Jackknife, Post-stratification, Variance Estimation

Abstract

This paper summarizes the results of a study of a number of list-based estimators that would minimize the use of area frame sampling for incompleteness of the list frame when estimating hired, self-employed, and unpaid workers from the Quarterly Agricultural Labor Surveys (QLS). The study was based on data from sixteen quarterly labor survey periods during 1992-96. The evaluation compares the currently used direct expansion (DE) estimator and seven alternative estimators. The jackknife procedure is used to evaluate the performance of the these estimators.

1 Introduction

The National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture (USDA) has been investigating estimation approaches that would minimize (or eliminate) the use of area-frame samples in adjusting for incompleteness of the list frame. These investigations were initiated because of the relatively high survey cost and respondent burden associated with the area frame samples and also because of the poor precision of the resulting estimates for the area that is non-overlapping with the list frame (NOL). Seven alternative estimators are considered and compared to the currently used direct expansion (DE) estimator. Five of the alternative estimators rely less on area-frame samples than the DE. Development of these estimators is detailed in Chhikara, et al (1995), Perry, et al (1993),(1997), Rumburg, et al (1993) and Spears, et al (1996).

In Section 2, we discuss the agricultural labor survey data and identify variables of interest and relevant auxiliary variables. In Section 3, we describe the estimators considered in this study. These estimators are compared and evaluated in Section 4 using numerical results obtained from the 1992-96 quarterly labor survey data (QLS).

2 Labor Survey Data

Variables:

The QLS are conducted to estimate various characteristics for the three types of farm labor: hired workers, self-employed workers, and unpaid workers. The survey response variables for all three labor types include {total number of workers employed} and {total hours worked for a specified week}, and for hired workers, the {total wages paid for a specified week}. The variables of interest for each type of workers are (1) the total workers for a specified week and (2) the average weekly hours per worker for a specified week. For hired workers, another variable of interest is (3) the average hourly wage rate per hired worker for a specified week.

The auxiliary information from the JAS and the QLS that was used to develop the regression type estimators for the NOL included: Farm Peak Number of Workers, Farm Type, and Number of Partners. The farm type, a categorical variable, was found to effect an efficient post-stratification of sample data. The peak number of workers was found to have the highest correlation with the hired number of workers. Only the number of partners had any significant correlation with the number of self-employed workers. None of the auxiliary variables were correlated with the number of unpaid workers, so the fixed mean model was used.

Post-Stratification:

The post-stratification by farm-type often resulted in some post-strata that did not contain sufficient data to reliably estimate the regression parameters.

[1] 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.

[2] These authors are with the National Agricultural Statistics Service, 3251 Old Lee Highway, Room 305, Fairfax, VA 22030

This problem was remedied by collapsing farm-types until each post-stratum contained at least 15 sample observations. The collapsing procedure involved initial computation of regression coefficients for all post-strata based on annual JAS (historical) data. If the smallest post-strata had less than 15 sample observations, it was collapsed with the closest post-strata measured in terms of the distance between the regression coefficients. The regression coefficients were then updated and the procedure was repeated until the smallest post-strata had at least 15 sample observations. Evaluations were also made using post-strata with a minimum of 30 and 60 sample observations, but no significant differences were detected from the case with a minimum of 15 sample observations per post-strata.

3 Estimators

Currently, NASS employs a multiple frame estimation methodology that combines separate, independently computed, direct expansion estimates of the list and NOL components into an estimate of the total. A detailed description of the multiple frame DE estimator is given in Kott (1991). The list frame DE component is based on a sample which is large enough to ensure efficiency.

The present study focuses on the development of a more efficient, yet cost effective, estimator of the NOL component than is currently available. The eight estimators evaluated here (including the DE) are grouped into three categories with respect to use of the NOL. Unless otherwise specified, these estimators are combined with the list component DE estimate to arrive at an estimate of the total.

3.1 Quarterly Labor NOL Samples Used

The estimators described in this section require the NOL sample from the current quarter.

Direct Expansion:

The NOL component estimator is,

$$\hat{Y}_{\text{nol,de}} = \sum_{h=1}^{H} \sum_{i \in s_h} w_i y_i \qquad (1)$$

where w_i is the expansion factor and y_i is the observed value for the ith sample unit in area stratum h; and s_h denotes the set of its sample units.

Reweighted Expansion:

The NOL component estimator is,

$$\hat{Y}_{\text{nol,rewt}} = \sum_{h=1}^{H} \left(\sum_{i \in s_h} w_i \left(\frac{\sum_{j \in S_h} w_i}{\sum_{j \in s_h} w_i} \right) a_i y_i \right) \qquad (2)$$

where w_i is the first phase expansion factor for the ith sample unit of the second phase farm labor stratum h, y_i is observed value for this sample unit, and a_i is the adjustment due to tract to farm ratio, nonresponse, and data adjustment as a result of farm partnership, etc. S_h is the set of original sample units and s_h is the set of sub-sample units in stratum h.

If, on the other hand, the tract to farm ratio is considered as part of the first phase expansion, then it leads to slightly different values for the w_i and the a_i, and thus, another NOL component estimate.

Difference Estimator:

The NOL component estimator is,

$$\hat{Y}_{\text{nol,diff}} = \hat{Y}_{\text{pnol(100)}} + \hat{D} \qquad (3)$$

where $\hat{Y}_{\text{pnol(100)}}$ is the predicted NOL using the auxiliary data for the 100% NOL sample from JAS and the regression coefficients determined from the quarterly labor list samples, and

$$\hat{D} = \hat{Y}_{\text{pnol(40)}} - \hat{Y}_{\text{nol,de}}$$

is the difference between the predicted and actual estimates computed from the NOL samples for the current quarter.

3.2 July Labor NOL Samples Used

The estimators described in the section adjust the July NOL estimate to arrive at the current NOL estimate; thus, no NOL samples are required in the current quarter.

List-Based Direct Expansion:

The NOL component estimator is,

$$\hat{Y}_{\text{nol,lbde}} = \hat{Y}_{\text{nol,de(J)}} + (\hat{Y}_{\text{pnol(L)}} - \hat{Y}_{\text{pnol(J)}})$$

where $\hat{Y}_{\text{nol,de(J)}}$is the direct expansion for July of the current year and ($\hat{Y}_{\text{pnol(L)}}$-$\hat{Y}_{\text{pnol(J)}}$) is the difference between the predicted NOL for the current quarter and July of the current year.

List-Based Difference:

The NOL component estimator is,

$$\hat{Y}_{\text{nol,lbdi}} = \hat{Y}_{\text{nol,diff(J)}} + (\hat{Y}_{\text{pnol(L)}} - \hat{Y}_{\text{pnol(J)}})$$

where $\hat{Y}_{\text{nol,diff(J)}}$is the difference estimate for July of the current year and ($\hat{Y}_{\text{pnol(L)}}$-$\hat{Y}_{\text{pnol(J)}}$) is the difference between the predicted NOL for the current quarter and July of the current year.

Ratio:

The NOL component estimator is,

$$\hat{Y}_{\text{nol,ratio}} = \hat{Y}_{\text{list,de}} \left(\frac{\hat{Y}_{\text{nol,de(J)}}}{\hat{Y}_{\text{list,de(J)}}} \right)$$

where $\hat{Y}_{\text{list,de}}$is the current direct expansion estimate of the list and $\hat{Y}_{\text{nol,de(J)}}$and $\hat{Y}_{\text{list,de(J)}}$are the July direct expansion estimates of the NOL and list, respectively.

3.3 No NOL Samples Used

The estimators described in this section do not require quarterly NOL samples.

Predicted NOL

An estimator for total is obtained by adding the current list estimate to the predicted NOL estimate for the current period.

The NOL component estimator is,

$$\hat{Y}_{\text{nol,pnol}} = \hat{Y}_{\text{pnol(100)}}$$

where $\hat{Y}_{\text{pnol(100)}}$, the predicted NOL using regression coefficients from the current list survey and the 100% NOL samples from the JAS, is defined in Eq. 3.

Post-Stratified - Total

The post-stratified area frame estimator does not use the current list estimate, but obtains directly an estimate of the total.

$$\hat{Y}_{\text{psaf}} = \sum_{k=1}^{K} \hat{N}_{.k} \hat{\bar{y}}_k$$

where $\hat{N}_{.k}$ is the estimated size for post-stratum k using the JAS data and

$$\hat{\bar{y}}_k = \frac{\sum_{i \in U_k} w_i y_i}{\sum_{i \in U_k} w_i}$$

where w_i is the weight of the ith labor list sample unit that falls in post-stratum k and y_i is its observed value, and U_k is the set of labor list sample units in that post-stratum.

4 Empirical Jackknife Results

The estimators considered in this study were evaluated using the QLS data from 1992-96 from each of the 17 agricultural regions in the U.S. Estimates of the labor items of interest (described in Section 2) were computed at the regional level and then aggregated to the U.S. level. Estimates from the alternative estimators were compared to the corresponding DE estimates.

4.1 Alternative Estimators Compared to Direct Expansion

The relative mean deviation (R-MD),

$$\text{R-MD}(\hat{Y}) = \frac{\sum_{i=1}^{16} (\hat{Y}_i - \hat{Y}_{DE,i})}{\sum_{i=1}^{16} \hat{Y}_{DE,i}}, \qquad (4)$$

and the relative root mean squared deviation (R-RMSD),

$$\text{R-RMSD}(\hat{Y}) = \frac{\sqrt{\frac{1}{16} \sum_{i=1}^{16} (\hat{Y}_i - \hat{Y}_{DE,i})^2}}{\frac{1}{16} \sum_{i=1}^{16} \hat{Y}_{DE,i}}, \qquad (5)$$

were used for performance criteria. The R-MD measures the average deviation of an estimator from the corresponding DE, relative to the average DE across the 16 quarterly surveys for which estimates were computed. The relative root mean squared deviation measures the variability of the deviation of survey estimates from the DE for the same 16 quarterly surveys.

The R-MD and R-RMSD were computed for each estimator and item estimated at the regional and national levels. The U.S. level results, summarized in Table 1, were used to draw the following conclusions.

- The reweighted estimates are almost the same as the DE at the U.S. level.

- The difference estimates of total workers are on the average higher than DE estimates, with R-RMSD of 3.6 percent for the hired and more than 8 percent for the self-employed and unpaid workers. Estimates for the two estimators are similar for weekly hours and wage rates, with R-RMSD less than 1 percent.

- For the three list-based estimators (LBDE, LBDiff and ratio), estimates of total workers are on the average higher than the DE estimates. This may be expected in the case of the LBDE and LBDiff which incorporate input from the difference estimator, which gives rise to estimates that are themselves higher than the DE estimates. The ratio estimator has the smallest R-MD and R-RMSD among the three list-based estimators.

- The predicted NOL estimates compare favorably to the DE for all three items in the case of hired workers, as do the post-stratified estimates in the case of weekly hours and wage rates for hired workers. However, these two estimators are completely unreliable for the other two types of worker. Overall, these two estimators have R-RMSD substantially higher than other estimators.

4.2 Jackknife Evaluations of Bias and Variance

Each estimator was evaluated for its bias and variance estimated using a jackknife procedure. The jackknife replicates were developed taking into account the sample design and the sampling weights were re-calculated for each jackknife replicate based on the new sample size. A set of 15 jackknife estimates of each item were made for each estimator by applying the estimation process to each of the 15 sets of jackknife replicate data.

For an evaluation of bias, each estimator was compared to the DE estimator in the following manner.

An estimate of bias is obtained by

$$\hat{\text{Bias}} = \hat{Y}_{\text{est}} - \hat{Y}_{\text{de}},$$

where $\hat{Y}_{\text{est}}$is the estimate using a particular estimator and $\hat{Y}_{\text{de}}$is the corresponding estimate for the DE. Next a 95% confidence interval for bias is computed by

$$\hat{\text{Bias}} \pm t_{14,0.025}\hat{\text{SD}}$$

where

$$\hat{\text{SD}} = \sqrt{\frac{14}{15}\sum_{i=1}^{15}(D_i - D)^2},$$

and $D_i = \hat{Y}_{J,i} - \hat{Y}_{\text{de},i}$, $i = 1, 2, \ldots, 15$, $D = \hat{Y}_{\text{est}} - \hat{Y}_{\text{de}}$, and $\hat{Y}_{J,i}$ is the ith jackknife estimate for the estimator being evaluated and $\hat{Y}_{\text{de},i}$ is the DE estimate corresponding to that jackknife replication.

The bias estimates and confidence intervals both computed relative to the DE for the various estimators are depicted for total number of hired workers in Figure 1. The following conclusions are drawn from these results.

- The alternative estimators do not exhibit any significant bias for number of hired workers.

- The predicted NOL estimator has positive bias in estimating number of unpaid workers (approximately 30% higher overall). The difference estimator shows bias of at most 10% higher than DE in estimating number of unpaid workers.

- All of the estimators except the reweighted show a slight, consistent bias (though mostly insignificant) in estimating number of self-employed workers. The bias is positive except in the case of the post-stratified estimator which has a negative bias.

Next the jackknife variance is computed in two ways. An estimate of the variance for estimator J, usually based on deviations from the mean, is given by

$$V_{1,J} = \frac{14}{15}\sum_{i=1}^{15}(\hat{Y}_{J,i} - \hat{\hat{Y}}_J)^2. \tag{6}$$

Another estimate of the variance is obtained by considering deviations of the jackknife replicated estimates from the estimate based on the complete sample data. This estimate is given by

$$V_{2,J} = \frac{14}{15}\sum_{i=1}^{15}(\hat{Y}_{J,i} - \hat{Y})^2 \tag{7}$$

where deviations are computed with respect to $\hat{Y}$ which is the estimate based on the complete sample data. The ratio of $V_{1,J}$ to $V_{2,J}$ was close to 1 in most cases, which is indicative of negligible jackknife bias.

The estimated coefficient of variation for an estimator is computed by

$$\hat{\text{CV}}_{2,J} = \frac{\sqrt{V_{2,J}}}{\hat{\hat{Y}}_J}, \tag{8}$$

where $\hat{\hat{Y}}_J$ is the mean jackknife estimate for estimator J. The following conclusions can be drawn from analysis of $\hat{\text{CV}}_{2,J}$.

- $\hat{\text{CV}}_{2,J}$ for the DE is mostly the smallest, though it varies for farm labor item and survey period. The largest value of $\hat{\text{CV}}_{2,J}$ can be as much as two times the smallest $\hat{\text{CV}}_{2,J}$ value for an item estimated across survey periods.

- The $\hat{\text{CV}}_{2,J}$ values for the reweighted, ratio, difference and predicted NOL estimators are fairly consistent with the DE in the case of estimating total number of hired, self-employed and unpaid workers.

Table 1: **R-MD and R-RMSD for Alternative Estimators Relative to DE(US Level)**

				Estimator			
Item Estimated	Diff	LBDE	LBDiff	PNOL	PSAF	RATIO	ReWt
				R-MD(%)			
				Hired			
Total	2.8	0.9	3.4	3.4	-0.3	2.0	0.0
Weekly Hours	0.0	-0.6	-0.2	1.1	2.6	-0.3	-0.0
Wage Rate	-0.1	1.2	0.6	-1.1	0.2	0.3	0.0
				Self-Employed			
Total	7.3	4.2	6.9	11.5	-6.9	2.6	0.0
Weekly Hours	0.7	-5.4	-4.5	18.4	19.5	-0.8	-0.0
				Unpaid			
Total	7.1	0.9	3.4	29.1	10.9	5.8	0.0
Weekly Hours	0.1	-5.7	-4.7	3.2	5.2	-0.4	-0.0
				R-RMSD(%)			
				Hired			
Total	3.6	4.9	5.5	5.4	7.9	4.3	0.2
Weekly Hours	0.3	1.5	0.9	1.4	3.0	0.9	0.1
Wage Rate	0.3	1.9	1.2	1.3	1.8	1.0	0.0
				Self-Employed			
Total	8.3	5.6	7.8	13.2	10.0	3.6	0.7
Weekly Hours	0.9	8.6	7.8	19.6	20.5	2.7	0.5
				Unpaid			
Total	8.2	7.8	8.5	30.7	20.2	7.9	0.6
Weekly Hours	0.6	9.2	8.0	4.2	5.7	2.1	0.3

- The LBDE, LBDiff and post-stratified estimators display no specific pattern in their $\widehat{CV}_{2,J}$ values. Consistency in $\widehat{CV}_{2,J}$ values may exist over time for some items.

References

[1] Chhikara, Raj S., Perry, Charles R., Deng, Lih-Yuan, Iwig, William C., Spears, Floyd M. and Cowles, Susan, "Post-Stratification and Efficient Estimation in U.S. Agricultural Labor Surveys," ASA Proceedings of Survey Research Methods, 1995.

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

[3] Perry, Charles, Chhikara, Raj, Spears, Floyd M. and Cowles, Susan, "An Evaluation of List-Only, Reweighted, and Other Estimators for U.S. Agricultural Labor Surveys," USDA-NASS RD Research Report No. RD-97-06, October 1997.

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

[5] Spears, Floyd M., Chhikara, Raj S., Perry, Charles R., Iwig, William C. and Cowles, Susan, "Agricultural Labor Estimation Using Only List-Frame Sampling," ASA Proceedings of Survey Research Methods, 1996.

Relative Bias from DE with 95% Confidence Interval

Hired Workers

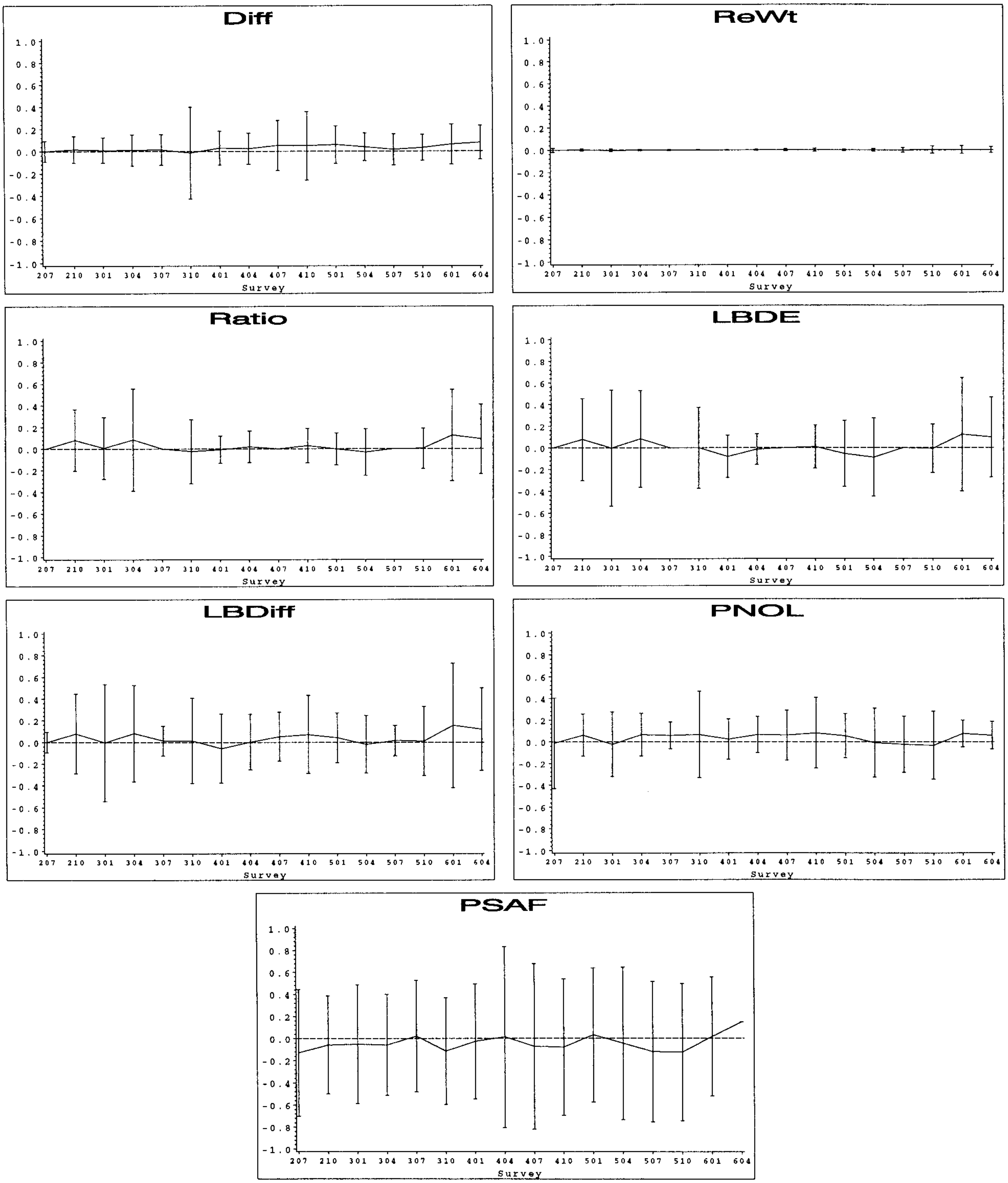

Figure 1: