AGRICULTURAL LABOR ESTIMATION USING ONLY LIST-FRAME SAMPLING

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

KEY WORDS: Regression Estimator, Prediction, Post-stratification

Abstract

The National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture utilizes area frame samples to compensate for the lack of coverage by list frame sampling. The area frame samples are considerably smaller than the list frame samples. The cost of observing area samples is prohibitive and the precision of estimates based on the area samples is low. An estimation approach that eliminates or reduces the need for an area frame sample is desirable. A regression estimator that makes use of only the list-frame sample data is developed. Auxiliary information collected by NASS for its annual June Enumerative Survey (JES) is used to post-stratify the list samples of a labor survey as well as the area samples of the JES and then to predict the labor characteristics of interest for the area not overlapping with the list (NOL) using a difference estimator. The procedure is implemented to estimate the number of hired workers, hours worked per week, and the hourly wage rates in all states for each quarterly reporting period during the years, 1992-1993, 1993-1994 and 1994-1995. The resulting estimates are in close agreement with those obtained using the current procedure which requires area-frame sampling.

1 Introduction

In the last few years, the National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture (USDA) has been investigating estimation approaches that would either minimize or eliminate the use of area-frame samples because of their excessive survey cost and respondent burden as well as poor precision resulting from their use for estimation of a characteristic in the area which is nonoverlapping with the list (NOL). The present study is a continuation of an earlier one whose results were reported in Chhikara, et. al (1995). In that study, certain post-stratified and difference estimators were investigated for their performance using the 1991-92 quarterly labor survey data for California and Florida. The difference estimator was shown to have performance similar to that of the currently employed direct expansion estimator. This outcome was viewed to be promising for the development of a "list-only" estimator whereby one does not need to observe the NOL samples.

The basic approach involves post-stratification of sample data for both the current survey period and the annual June Enumerative Survey (JES) based on farm types which is followed by construction of a regression estimator using certain auxiliary variables. Only the list samples are used in obtaining the least-square fits; and these data-fitted equations are then used to predict a characteristic total for the NOL based on its auxiliary information available from the JES. In the present study, three list-based estimators are developed based on this approach. Also considered are two other estimators which are described later.

In Section 2, we explore the agricultural labor survey data and identify relevant auxiliary variables. In Section 3, we describe different estimators that were considered for evaluation in this study. This is followed by their numerical evaluations using the 1992-93, 93-94 and 94-95 quarterly survey data as discussed in Section 4 for the number of hired workers, the hours worked per week and the hourly wage rates. The list-based estimators are compared with the currently used direct expansion estimator which requires the use of NOL sample data.

2 Labor Survey Data

Variables:

The quarterly labor surveys (QLS) are conducted for an estimation of hired workers, self-employed workers and unpaid workers. Presently we restrict ourselves to the case of hired workers and consider the response variables 1) Number of Hired Workers (y_1), 2) Average Weekly Hours (y_2), and 3) Hourly Wage Rate (y_3).

Auxiliary information from the JES and the QLS used in constructing regression estimators for the NOL include: 1) Farm Value of Sales (x_1) 2) Peak Number of Workers (x_2) and 3) Farm Type (x_3).

The farm type is a categorical variable and is utilized for post-stratification of sample data. The other two variables are quantitative and can be used as regressors. Of these two variables, only the peak number of workers appears to be well correlated with the hired number of workers. An examination of the scatter plots and correlation coefficients for the hired number of workers versus the peak number of workers shows that it is appropriate to assume a non-intercept regression model for the hired workers for each of the variables y_1, y_2 and y_3, with peak number of workers (x_2) as the regressor.

Post-Stratification:

The post-stratification by farm-type sometimes gives rise to post-strata that may not have enough sample observations to reliably estimate the regression function in a post-stratum. This problem was remedied by collapsing farm-type post-strata so that there were at least 15 sample observations in each post-stratum. The collapsing procedure involves initial computation of regression coefficients for all post-strata based on annual JES (historical) data. If the smallest post-strata has less than 15 sample observations, it is collapsed with the closest post-strata measured in terms of the distance between the regression coefficients. The regression coefficients are then updated and the procedure is repeated until the smallest post-strata has at least 15 sample observations.

3 List-only Estimators

Several list-only estimation procedures were examined in terms of their precision. The estimates were developed based on the stratified and post-stratified estimation approaches as described next.

The basic approach to estimation 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:

(1) Stratified Estimation

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

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

(2) Post-Stratified Estimation

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

where $\hat{N}_{.k}$ is the estimated size of post-stratum k and $\hat{\hat{Y}}_k$ is an estimator of mean response for post-stratum k.

Both $\hat{N}_{.k}$ and $\hat{\hat{Y}}_k$ 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 $\hat{\hat{Y}}_k$ is obtained either by the post-stratum sample mean or prediction mean using certain regressors as follows:

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

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, $x_1, x_2, ..., x_m$, 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 $\hat{\hat{Y}}_k$ can be obtained by

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

and the corresponding post-stratified estimator is

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

The following five list-based estimates were computed on the basis of the stratified and post-stratified estimation approaches discussed above:

(a) Direct Expansion (DE)

(b) Difference (Diff)

(c) Predicted NOL (PNOL)

(d) Post-Stratified Area Frame (PSAF)

(e) Post-Stratified Multi-Frame (PSMF)

In each of cases (a)-(c), a separate NOL estimate is made which is subsequently added to the current estimate for the list component in order to estimate the total for a characteristic of interest. The NOL estimates in the first three cases are computed using regression estimates. The first two cases require NOL sampling in the July quarterly labor survey. Case (c) does not require NOL sampling from the quarterly labor surveys.

Computation of NOL Estimates:

The list-frame sampling units from the labor survey were used along with auxiliary variables from the June Enumerative Survey (JES) to obtain regression equations for the response variables of interest (number of hired workers, average weekly hours and wage rates) for each of the post-strata. These data-fitted equations were used to predict the mean response for the NOL sample units using the auxiliary variables from the JES.

The NOL component is predicted by appropriately expanding each sample unit prediction and aggregating the expanded predictions across all NOL sample units. This NOL prediction for the quarter period of interest is denoted by

$$\hat{Y}_{N,P} = \sum_{i \in U_J} \hat{y}_i e_{J,i} \qquad (5)$$

where U_J denotes the set of all NOL sample units in the JES, $\hat{y}_i$ denotes the predicted response of the ith NOL sample unit from the JES using the data-fitted equations from the current list sample, and $e_{J,i}$ denotes the expansion associated with the ith NOL sample unit from the JES.

(a) The list-based direct expansion estimate makes use of the NOL estimate obtained for the July labor survey. It is simply an aggregation of the expanded strata sample means for the stratified estimator given in Equation (1). Denoting it as $\hat{Y}_{N,\mathrm{DE}}$ the DE estimator for the NOL is given by

$$\hat{Y}_{N,\mathrm{DE}} + (\hat{Y}_{N,P} - \hat{Y}_{N,P_{\mathrm{Jul}}}) \qquad (6)$$

(b) The list-based difference estimate involves the use of the NOL regression estimate for the quarter period in July obtained using the difference estimation procedure given in Equation (8) of Chhikara, et. al (1995). Denoting it by $\hat{Y}_{N,\mathrm{Diff}}$, the NOL difference estimate is given by

$$\hat{Y}_{N,\mathrm{Diff}} + (\hat{Y}_{N,P} - \hat{Y}_{N,P_{\mathrm{Jul}}}) \qquad (7)$$

where $\hat{Y}_{N,P}$ and $\hat{Y}_{N,P_{\mathrm{Jul}}}$ are the predicted NOL for the current and July quarterly labor survey periods, respectively.

(c) The predicted NOL estimate for current quarterly period is given by $\hat{Y}_{N,P}$ as defined in Equation (5).

(d) The post-stratified area frame estimate is computed using Equation (3) by employing only the area frame data of the JES.

In case (e), the total of a labor characteristic is directly estimated as follows:

(e) The multi-frame post-stratified estimate is obtained by combining the list and NOL sample data for determining $\hat{N}_{.k}$ for each post-stratum and expanding directly the post-stratum sample means. This estimate can be written as a weighted estimator given by

$$\hat{Y}_{ps,wt} = \sum_{k=1}^{K} \hat{N}_{.k} \hat{\bar{y}}_{k,wt} \qquad (8)$$

where

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

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

4 Empirical Results

The list-only estimators discussed in Section 3 were investigated and evaluated using the 1992-93, 93-94 and 94-95 quarterly Agricultural Labor Surveys from states in each of the 17 agricultural regions in the United States. Estimates of the number of hired workers, hours worked per week and the hourly wage rates were computed first at the regional level and then aggregated to the national level. The national estimates were compared to national estimates obtained using the current NASS direct expansion estimator in order to better understand how each estimator performs (Figure 1). The direct expansion estimator was chosen for comparison since it is a currently employed procedure that utilizes the area-frame sampling. It involves an aggregation of the expanded strata sample means as a stratified estimator described in Equation (1).

Among the five list-based estimators considered, the post-stratified multipl frame estimates are consistently smaller than the current direct expansion estimates across all quarters in the case of hired number of workers. All estimators do equally well except the list-based direct expansion in the last two quarters in estimating average hours worked. For estimating the wage rates, the list-based direct expansion and the list-based difference estimate do not compare with the current estimator as well as the other estimators.

The performance of the list-based estimators as compared to the current direct expansion estimator was quantified with the Relative Mean Difference (RMD) and the Relative Root Mean Squared Error (RRMSE). The Relative Mean Difference,

$$\mathrm{RMD}(\hat{Y}) = \frac{\sum_{i=1}^{12}(\hat{Y}_i - \hat{Y}_{DE,i})}{\sum_{i=1}^{12} \hat{Y}_{DE,i}},$$

measures the average bias of the estimator $(\hat{Y}_i)$ relative to the current direct expansion estimator $(\hat{Y}_{DE,i})$ for the 12 quarterly surveys for which estimates were computed. The Relative Root Mean Squared Error measures the variability of the survey estimates with respect to the direct expansion estimate for the 12 quarterly surveys for which estimates were computed and is given by:

$$\mathrm{RRMSE}(\hat{Y}) = \frac{\sqrt{\frac{1}{12}\sum_{i=1}^{12}(\hat{Y}_i - \hat{Y}_{DE,i})^2}}{\frac{1}{12}\sum_{i=1}^{12} \hat{Y}_{DE,i}}.$$

The RMD and RRMSE were computed for each of the list-only estimates of the number of workers, the hours worked per week and the hourly wage rates at the national level. The results are displayed in Figure 2.

The post-stratified multiple frame estimator performs the worst in estimating each labor characteristic. For estimating the number of hired workers, the post-stratified area frame estimator has very little bias. However, it does not perform as well as the list-based direct expansion, difference and predicted NOL estimators do in estimating the other two characteristics. The list-based direct expansion and difference estimators have the smallest overall variability from the currently made direct expansion estimate. For estimating the average weekly hours worked, the list-based difference estimator has the smallest bias and variability measures. For estimating the average wage rates, all of the estimators have relatively little bias (less than 3 percent). The list-based predicted NOL estimator has the smallest measure of overall variability.

The list-based difference estimator has the most robust performance when all three labor characteristics are considered with the predicted NOL estimator only slightly less robust. Since the predicted NOL estimate requires no area frame NOL sampling in any quarterly labor survey, its choice is the most meritorious as a list-only estimator.

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, SRB Research Report No. SRB-93-04, Washington, D.C., July 1993.

[3] 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.

List-Based Estimates for United States
Relative to Current Direct Expansion Estimates

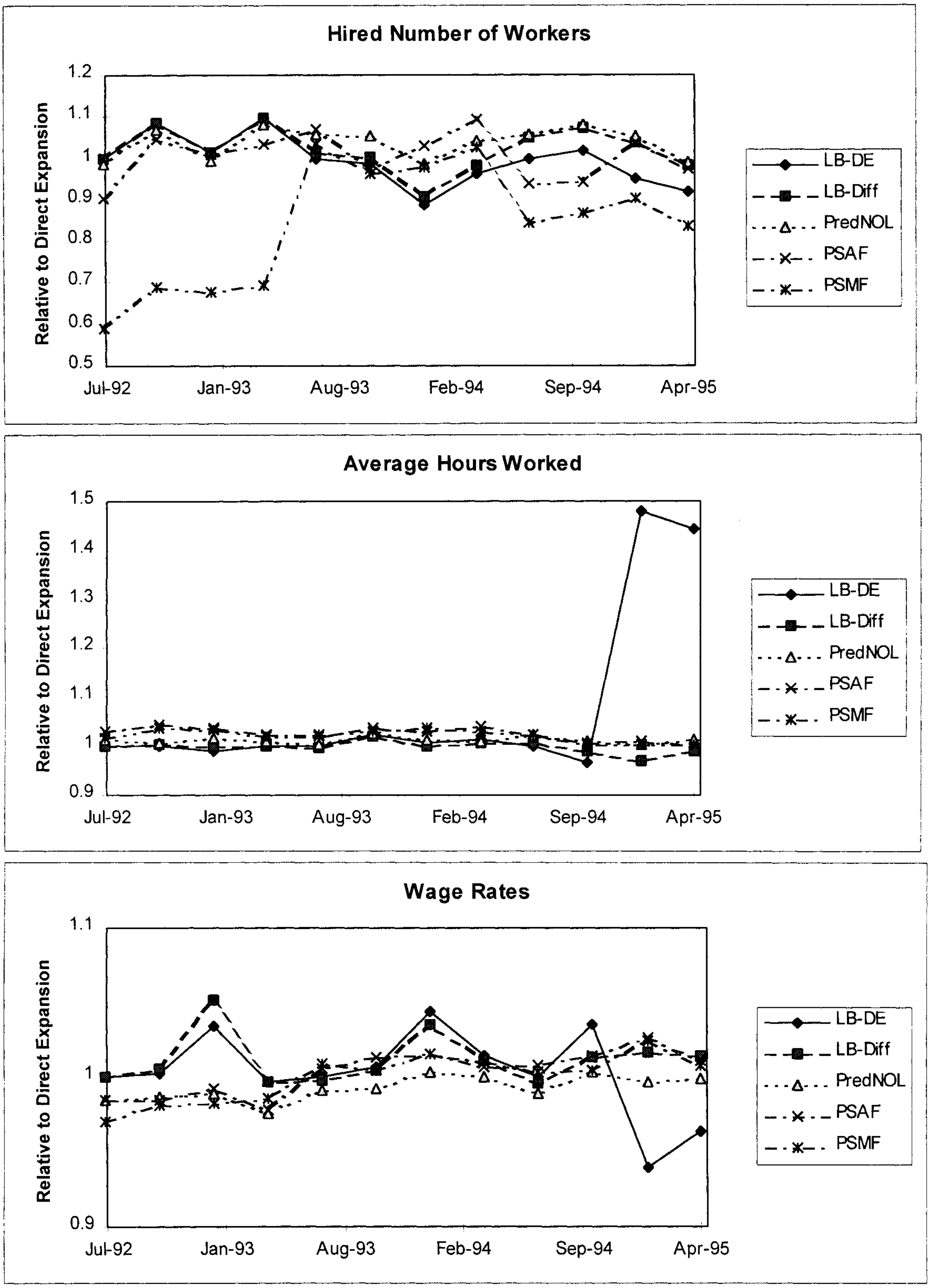

Figure 1.

Deviations of List-Based Estimates from Current Direct Expansion Estimates at U.S. Level

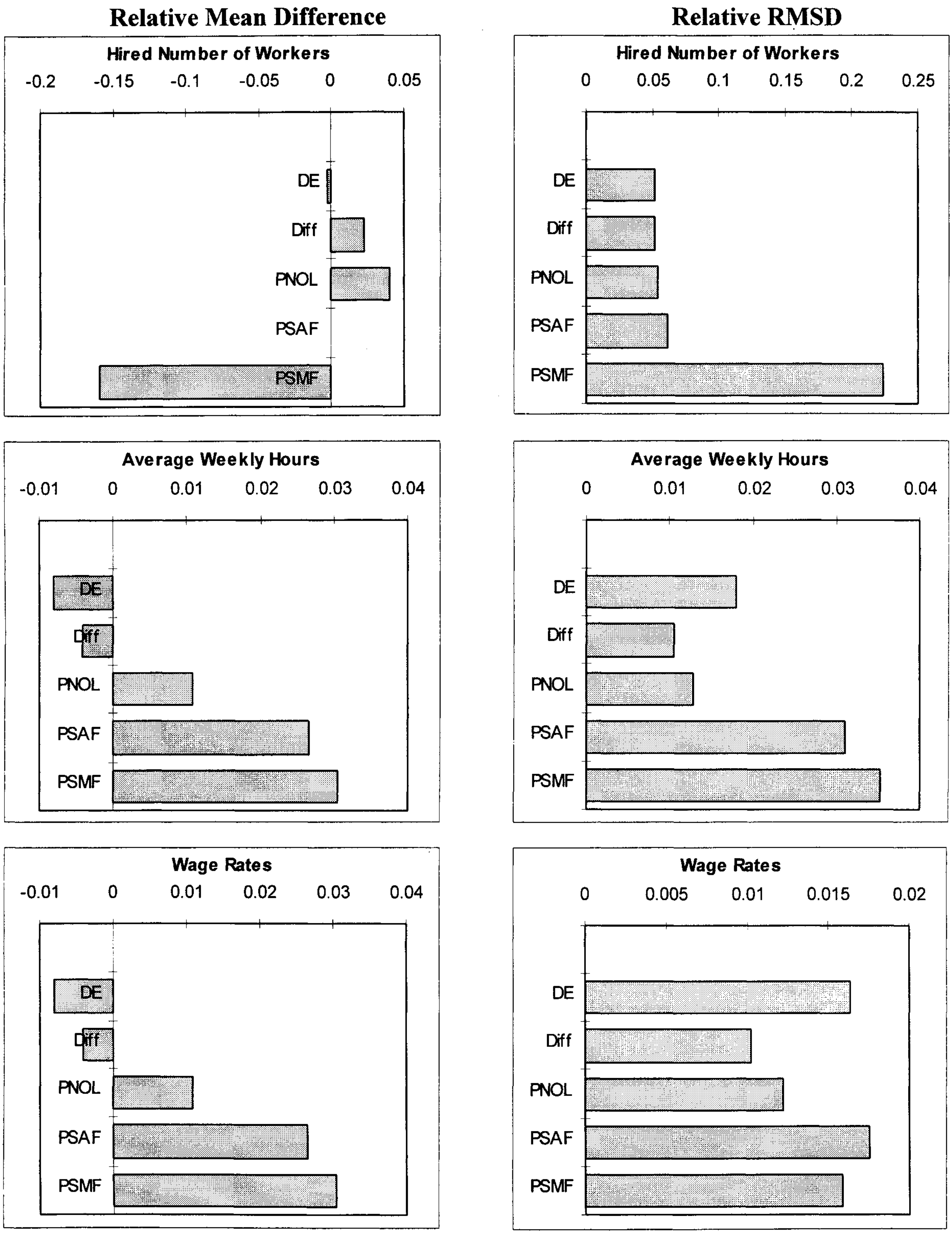

Figure 2.