STATISTICAL PROBLEMS IN BLS COMPENSATION SURVEYS WHEN COLLECTED ESTABLISHMENT DATA DIFFERS FROM THE ASSIGNED DATA

Susan R. Black, Lawrence R. Ernst, and Jason Tehonica, Bureau of Labor Statistics
Susan Black, BLS, 2 Massachusetts Ave. N.E., Room 3160, Washington, DC 20212

Key Words: Sampling; Weighting; Respondent Burden; OCSP; ECI; EBS

1. Introduction

The National Compensation Survey (NCS) is a Bureau of Labor Statistics (BLS) establishment survey program of employee salaries, wages, and benefits. The program is designed to produce data at local levels, broad regions, and nationwide. The NCS will replace three existing BLS survey programs: Employment Cost Index (ECI), Occupational Compensation Survey (OCS) Program and Employee Benefits Survey (EBS). The NCS was developed to expand the data products of the existing compensation programs to eliminate duplicate data collection and processing requirements, reduce respondent burden, develop more efficient and streamlined collection and processing techniques, and to address budget constraints. Cohen (1997) presents a more detailed overview of the NCS program.

Within each geographic PSU, there are two stages of sampling. An establishment sample is chosen during the first stage of selection. At the second stage of selection an occupation sample is selected within each establishment. For local estimates, the weight for each employee in a selected defined occupations is obtained by taking the product of the reciprocal of the probability of selecting the establishment, the reciprocal of the probability of selecting the defined occupation within the selected establishment, and nonresponse adjustment factors for establishment and occupation nonresponse. All of the terms used to describe the weight of each employee are presented in more detail in Black, Ernst, and Tehonica (1997), along with a more detailed description of the sample design. Other weighting factors used to compensate for situations where the unit collected differs in some way from the originally assigned unit will be described in this paper.

This paper will address the statistical problems in the NCS when collected establishment data differs from the assigned data. A large survey program, such as NCS, attempts to have clear and precise data collection procedures set. There are many data collection problems that occur which are out of the hands of the data collector. When a data collector encounters these problems a weight adjustment can be made to lessen the effect on variances and biases of the estimates.

Section 2 covers establishment data collection issues. The most complex issue arises when a respondent provides data for more locations within a survey area than for what is sampled. Also discussed is subsampling a unit that consists of several physical locations or divisions. Adjustment for the situation when a unit is comprised of several locations and one or more of the locations are considered nonrespondents is also covered. Section 3 discusses the subsampling of employees when the selected defined occupation has a large number of incumbents. The occupation selection and weighting issues for establishments that are part of a Central Office Collection (COC) procedure are discussed in Section 4.

Due to space limitations, sections of the paper describing how to modify the variance estimation procedures to reflect the special sampling and weighting required for these situations has been deleted. Also deleted were sections describing future data collection problems arising from multiple panels, update interviews and birth establishments. The complete paper is available from the authors.

2. Establishment Issues

Data collection and weighting for establishments in the NCS can be likened, at least in part, to shooting at a moving target. This is because as a result of issues such as establishments which change ownership, and the inability or unwillingness of the respondent to report data for precisely the unit that it is desired to collect, obtaining unbiased estimates for the data collected can be a daunting and in part an operationally impossible challenge. This section discusses these issues.

Corresponding to each assigned unit is a set of physical locations from which data is actually collected in the initiation interview. We envision that there are six steps in determining this set of physical locations. Some of these steps require modifications in the weights to compensate for changes in the set of locations as we proceed through the steps. Not all the steps are required for each sampled unit. In fact, for the vast majority of sampled units we stop after step 1. Understanding these steps requires being able to distinguish between the following terms: "assigned unit," "modified assigned unit," "subsampled assigned unit," "desired collection unit" and "actual collection unit." Each of these terms is defined in the description of the six steps that we now proceed to present.

1. Determination of the "assigned unit." The frame from which the first stage sample units or "establishments" are selected within each sample geographic area for the NCS is a set of UDB numbers which are in one-to-one correspondence with the set of UI reporting numbers. For the most part, a UDB number represents a physical location for a company. However, sometimes a UDB number represents multiple locations for a company. In either case, an "assigned unit" for each selected UDB number is considered to be the set of physical locations corresponding to the UDB number, based on the report to the UI that was used in the frame construction. The same assigned establishment weight is given to each physical location that is part of an assigned unit, namely the reciprocal of the probability of selection of the corresponding UDB number. This weight is modified by an establishment nonresponse adjustment procedure that adjusts the assigned weights using weighting class cells formed by assigned employment and assigned industry

2. Determination of the "modified assigned unit." For the most part we seek to collect data for the set of physical locations that constitute the assigned sample unit. There are the following two exceptions to the rule of attempting to collect data from precisely the locations that are part of the original assigned unit. An original sample assigned unit as modified by these exceptions is known as a "modified assigned unit."

A. Any location that is part of the assigned unit but is outside of the sample geographic PSU is excluded from the modified assigned unit.

B. Any location within the sample PSU that reports to the UI under the number corresponding to the sampled UDB number that is in business at the time the data is collected but began business too late to be included the original assigned sampled unit, is included in the modified assigned unit.

In addition to excluding locations that are outside of the sample PSU, locations that have gone out of business or outside of the industrial scope of the survey are not included in the data collection process. These locations may be identified at various points in the six-step process. The assigned weight described in step 1 for a unit also applies to each location that is part of the modified assigned unit.

3. Subsampling of the modified assigned unit. Occasionally for reasons of respondent or interviewer burden it is necessary to subsample the set of locations that comprise the desired collection unit. Data is then collected from the subsampled locations only. To do this, the current total employment, known as the reported employment, is obtained for each location that comprises the modified assigned unit. A pps subsample of locations is selected, with the reported employment as the measure of size. The original assigned weight for the modified assigned unit, as modified by the establishment nonresponse adjustment factor, is multiplied by the reciprocal of the probability of the location being retained in the subsampling process to obtain the weight for each retained location after this step. We designate the set of locations remaining after this step as the "subsampled assigned unit."

4. Splitting of the subsampled assigned unit into "desired collection units." Sometimes the locations remaining after step 3 must be partitioned or split prior to PSO sampling. Typically this occurs when the assigned unit consists of multiple locations of a company, but the data required to perform the PSO sampling is kept separately at each location. It can also occur if some of the locations after step 3 have been sold to a different company, in which case the data for these locations would have to be collected from the new owners. Note that sometimes elements of the split unit can consist of more than one physical location, for example, when more than one location is sold to a company. In any case, each element of the split unit will be known as a "desired collection unit." The splitting step by itself requires no modification of establishment weights since it does not alter the set of locations which are to contribute to the estimates, just where the collection is to take place. Also note that when subsampling locations as noted in step 3 is required, splitting will always be done. Each element of the split for a subsample must consist of a single physical location. This is because each location of the subsample generally has a different weight after step 3 and by splitting the subsample into single physical locations we avoid the problem of creating a desired collection unit consisting of multiple locations with different weights.

5. Adjustment for nonresponse after splitting. When step 4 is required and some, but not all of the desired collection units resulting from the splitting become nonrespondents, a special nonresponse adjustment

factor is used. The weight of each respondent desired collection unit associated with the assigned unit is multiplied by this factor. The numerator of this factor is the weight of the original assigned sample unit after establishment nonresponse adjustment times its assigned employment. The denominator is obtained by multiplying the weight after step 4 for each location that is part of a responding desired collection unit by its reported employment and summing the result over all such locations. Note that because assigned employment is used in the numerator and reported employment in the denominator, it is possible that this computation could result in a factor less than 1, in which case we set the factor to be 1. (Unfortunately, we are forced to use assigned employment in the numerator, since we do not collect reported employment for nonresponding locations, and also must use reported employment in the denominator, since assigned employment does not exist for individual locations when the assigned unit consists of multiple locations.) This nonresponse adjustment for splits has also been called an adjustment for "collected less than assigned. This situation is also one of the cases of the "documentation adjustment factor," which is described in the next step.

6. Determination of "actual collection units." An "actual collection unit" is the set of locations from which data is actually collected corresponding to each responding desired collection unit. The main reason why an actual collection unit may differ from a desired collection unit is the inability or unwillingness of a respondent to separate company data for locations that are in sample desired collection units from those that are not. More specifically is the case where there are separate UDB numbers for each location in a sample PSU and the respondent will only give us combined data for all of their locations in the sample PSU. This includes both data from sampled and nonsampled locations, where each sampled location corresponds to a separate sample desired collection unit. Another case for which actual collection unit would differ from a desired collection unit would occur when a sample UDB number corresponds to multiple locations and the respondent is unable to exclude locations acquired through change in ownership.

We proceed to describe the general weighting methodology used to handle these types of situations. Note that it is possible for two or more sample desired collection units to correspond to the same actual collection unit. For example, this would be the case if each location of a company corresponded to a separate UDB number and two or more of these UDB numbers were selected with the respondent providing combined data for all locations in the PSU. When there is only a single sample desired collection unit corresponding to an actual collection unit which includes additional locations, the collection situation is referred to as "collected more than assigned." When two or more sample desired collection units are included in an actual collection unit, it is referred to as a "merger." However, the same general weighting methodology is used in both cases to weight an actual collection unit that is a union of desired collection units, and we proceed to describe this methodology. We will assume at first in this description that all desired collection units are respondents, that is neither whole establishment nonresponse adjustment nor the weighting adjustment in step 5 for nonresponse after splitting are needed. We then explain the modifications required when there are nonresponding desired collection units.

Let $Y = \sum_{i=1}^{N} Y_i$ be a parameter of interest, where Y_i is the value for the i-th actual collection unit in a population consisting of N actual collection units. (We assume conceptually that there is a unique actual collection unit corresponding to each desired collection unit, whether the desired collection unit arises from a sampled assigned unit or only nonsampled assigned units. We of course only know the desired collection units and the corresponding actual collection units that are associated with sampled assigned units.) Let $\hat{Y}_i$ be an unbiased estimator of Y_i, that is $E(\hat{Y}_i) = Y_i$. Let w_i, the weight of the i-th actual collection unit, be a random variable which is independent of $\hat{Y}_i$ and which satisfies

$$E(w_i) = 1, \tag{1}$$

and let

$$\hat{Y} = \sum_{i=1}^{N} w_i \hat{Y}_i \,. \tag{2}$$

Then, as observed in Ernst (1989), $\hat{Y}$ is an unbiased estimator of $\hat{Y}$, since

$$E(\hat{Y}) = \sum_{i=1}^{N} E(w_i)E(\hat{Y}_i) = \sum_{i=1}^{N} Y_i = Y.$$

We will use a special case of this result as follows. Let N_i denote the number of desired collection units corresponding to the i-th actual collection unit. Let W_{ij} denote the common weight associated with each location within the j-th desired collection unit of the i-th actual collection unit after step 3. That is, W_{ij} is the reciprocal of the probability that the locations in this desired collection unit are in a subsampled assigned unit. Then let $w_{ij} = W_{ij}$ if desired collection unit ij is a sample unit after step 3 and $w_{ij} = 0$ otherwise; let c_{ij}, $j = 1, \ldots, N_i$, denote a set of constants satisfying

$$\sum_{j=1}^{N_i} c_{ij} = 1; \tag{3}$$

and let

$$w_i = \sum_{j=1}^{N_i} c_{ij} w_{ij} . \tag{4}$$

Then w_i clearly satisfies (1) since

$$E(w_{ij}) = 1, \quad j = 1, \ldots, N_i . \tag{5}$$

Furthermore, although (1) is satisfied for any set of c_{ij}'s satisfying (3), for variance purposes we would like to reduce the variability of w_i, and consequently equalize the values of $c_{ij} W_{ij}$, $j = 1, \ldots, N_i$. Unless step 3, the subsampling step, is needed, this requires that $c_{ij} = p_{ij} / \sum_{k=1}^{N_i} p_{ik}$, where p_{ij} is the probability of selection of the assigned unit associated with the desired collection unit ij. Now, provided each of the assigned units associated with an actual collection unit are noncertainty units from the same sampling stratum, p_{ij} is proportional to the assigned employment, denoted A_{ij}, and hence

$$c_{ij} = A_{ij} / \sum_{k=1}^{N_i} A_{ik} . \tag{6}$$

Now A_{ij} is known for each sampled assigned unit ij that is associated with actual collection unit i. However, the assigned employment for nonsampled assigned units that are associated with this actual collection unit may not be known. Obtaining them would require extracting data from the UDB and, furthermore, there may be problems matching some assigned units to the UDB. Consequently instead of using (6) we let

$$c_{ij} = R_{ij} / R_i , \tag{7}$$

where R_i is the employment reported by the respondent for the i-th actual collection unit during data collection and R_{ij} is the reported employment for the j-th desired collection unit within the i-th actual reported unit. Note that R_{ij} need only be obtained for sampled desired collection units, since $c_{ij} w_{ij} = 0$, for all nonsampled desired collection units.

Now, although R_i is always known for any responding actual collection unit, it is possible that the respondent will not be able to provide the values of R_{ij} for sample desired collection units, in which case in place of (7) we could use

$$c_{ij} = A_{ij} / R_i . \tag{8}$$

With this value of c_{ij}, (3) does not necessarily hold since $\sum_{j=1}^{N_i} A_{ij} \neq R_i$ in general. However if the time lag is relatively short between the time of frame construction and the time of initial data collection, A_{ij} and R_{ij} generally do not differ by much provided desired collection unit ij consists of the same locations as the associated assigned unit.

We have discussed several possible values for c_{ij}. The value of c_{ij} that we have been using in the NCS in the case when $w_{ij} > 0$ for all j is (6). In this case c_{ij} is known as the merge adjustment factor. Otherwise, we have been using (8). In this case c_{ij} is also known as an adjustment for "collected more than assigned" or a case of the "document adjustment factor," with the other case described in Step 5. In the case when $w_{ij} > 0$ for at least two j's, the summation in (4) has also been described as the final weighting step for merges. The reason that we have not been using (7) is the difficulty in obtaining R_{ij} from a respondent who cannot separate out the data.

3. Occupational Issues

Except in the case of COCs, which are described separately in the next section, once it has been determined for which actual collection units PSO sampling will be done through the six step process described in the previous section, the selection of defined occupations for the initiation interview themselves entail relatively few data collection issues that require sampling and weighting modifications. This is because, unlike the case for establishment sampling where the actual collection units can be quite different in certain situations from the original assigned units, the selection of employees from an employee list is relatively straightforward. There is, of course, some occupational nonresponse, which requires two stages of occupational nonresponse adjustment as described in Black, Ernst, and Tehonica (1997). In this section we discuss only one occupation selection issue for the initiation interview, the subsampling of departments for certain occupations.

In Step 3 of Section 2 a procedure was presented for reducing respondent or interviewer burden by subsampling locations. We describe here another burden reducing procedure that is used at the occupational selection level. Ideally when an actual collection unit consists of multiple locations and/or departments, and an occupation is selected, data should be collected for all employees in the occupation in the collection unit. Occasionally, it is impossible to obtain data for an entire collection unit when an occupation is selected that is a dominant occupation in the unit. The respondent burden and collection burden can be overwhelming in some situations. For example, an elementary teacher in a school district or a nurse in a hospital could be quite burdensome in large school districts and hospitals. In these cases, methods of subsampling the occupation to particular departments or locations are used.

If the respondent is willing to give information by department or location, collection of the data occurs for the department or location in which the occupations that were selected reside. For example, there are 10 schools within a school district and elementary teacher is sampled 4 times. Each selected elementary teacher resides in a different school. The collection for elementary teacher is limited to the 4 schools. No additional weighting adjustment is needed since the final employee weight or individual weight will take the number of employees collected for into account during its computation. That is, the reciprocal of the probability of selection of the defined occupation, which is a key component of this weight, is the PSO sampling interval divided by the number of employees in the defined occupation. If data is only collected for teachers in a specific school, for example, then that becomes the defined occupation from a sampling and weighting perspective.

4. COC Establishments

Another data collection problem in NCS is the collection of sampled units that belong to a large company which has a policy that requires collection of data for their establishments be done from a single respondent. This single respondent is normally located at the central office for the company. Conducting PSO separately at each sampled COC establishment of a company becomes burdensome to the respondent that can jeopardize cooperation from these types of companies. Because COC establishments are found in most of the NCS PSU samples, it is imperative that we get cooperation from these types of companies.

In consultation with regional office collection staff the following alternate method of collection for the COCs was proposed. Since each of the companies was willing to provide national employment counts on all occupations it was determined that PSO could be conducted on this national data for each company to create a fixed job list for its establishments. The size of the occupation sample is determined on a case by case basis taking into account the number of establishments nationwide, the employment count nationwide, and the number of occupations nationwide. This type of occupation selection is known as Central PSO (CPSO).

A major disadvantage of using CPSO is that when the fixed job list is used for each establishment, there is no guarantee that each establishment selected for NCS for the COC will include any of the occupations on the fixed list. This means that the estimated employment for a sample establishment may be 0 or may be many times larger than the PSO employment for the establishment. In NCS, when PSO is done separately at each sample establishment, the sampling and weighting methodology we use guarantees that the estimated employment for the establishment will always equal the total PSO employment, regardless of which occupations are selected. This is not the case when CPSO is conducted without the use of an extra weighting adjustment described below.

Weighting of COC Establishments

As noted in the Introduction, the weight for each employee in a selected job in NCS is obtained by taking the product of the reciprocal of the probability of selecting the establishment, the reciprocal of the probability of selecting the job given that the establishment is selected, and nonresponse adjustment factors for establishment and occupation nonresponse. For COCs there are some differences in these factors and an additional adjustment that is used only for COCs.

The reciprocal of the probability of selecting a COC is no different than for a non-COC establishment, since the COC procedure only impacts the occupation selections. However, during the nonresponse adjustment procedures, the establishments that used the alternate collection procedure, CPSO, are given an establishment nonresponse adjustment factor of 1 and also an occupation nonresponse adjustment factor of 1. They are not put into the nonresponse cells formed for the normal PSO schedules.

For each COC establishment a single employment adjustment factor (EAF) is calculated for the noncertainty sampled occupations collected in the CPSO schedules. This factor adjusts the nationwide based occupation component of the employee weight to account for what is found at an individual establishment so that the occupational component of the employee weights summed over all employees in a sampled job equals the establishment's PSO employment. For non-COC establishments the EAF is not needed since this equality always holds for such establishments without an EAF. For the certainty occupations, the EAF is 1.0000.

To obtain the EAF, first let PSOE denote the PSO employment for a COC establishment. Let n_C, n_S denote the number of certainty and selected noncertainty occupations, respectively, selected from the CPSO list. Let $E_{Ci}, i = 1,\ldots,n_C$, and $E_{Si}, i = 1,\ldots,n_S$, denote the employment in the establishment for the *i*-th certainty occupation and the *i*-th selected noncertainty occupation, respectively. Finally, let $W_{Si}, i = 1,\ldots,n_S$ denote the reciprocal of the probability of selection of the *i*-th selected noncertainty occupation during CPSO. The employment adjustment factor is then

$$EAF = \frac{PSOE - \sum_{i=1}^{n_C} E_i}{\sum_{i=1}^{n_S} \left(W_{Si} \times E_i\right)}$$

The numerator of this fractional factor is the establishment's PSO employment in noncertainty jobs. The denominator is the unadjusted estimate of the number of employees in the establishment in non certainty jobs. Note that if the establishment has no employment in any of the selected noncertainty jobs, then the EAF is not defined since the denominator of the above expression is 0. Also, in some circumstances this factor can be very large which can result in a large increase in the variances of some estimates. We are therefore considering setting a maximum allowable value for this factor. Finally, like other adjustments of this type, this adjustment would introduce a bias into most estimates if they were not already biased.

References

Black, S. R., Ernst, L. R., and Tehonica, J. (1997), "Sample Design and Estimation for the National Compensation Survey," in *Proceedings of the Survey Research Methods Section,* American Statistical Association, to appear.

Cohen, S. H. (1997), "The National Compensation Survey: The New BLS Integrated Compensation Program," in *Proceedings of the Survey Research Methods Section,* American Statistical Association, to appear.

Ernst, L. R. (1989), "Weighting Issues for Longitudinal Household and Family Estimates," in *Panel Surveys,* eds. D. Kasprzyk, G. Duncan, G. Kalton, and M. P. Singh, New York: John Wiley & Sons, pp. 139-159.

Tehonica, J., Ernst, L. R., and Ponikowski, C. H. (1997), "Summary of Estimation and Variance Specifications for the 1996 Albuquerque, NM COMP2000 Test Survey," Bureau of Labor Statistics memorandum to The Record, dated March 3.

Any opinions expressed in this paper are those of the author and do not constitute policy of the Bureau of Labor Statistics.