EMPLOYMENT COST INDEX (ECI)--ESTIMATION PROCEDURES

Douglas A. Wright and Steven F. Kaufman, Bureau of Labor Statistic

I. Introduction

The complexity of the survey design, that is, systematic double sampling with unequal probabilities followed by 2-way controlled selection, in conjunction with the need to address the many problems of nonresponse while making maximum use of the available data, have resulted in a very complex estimation system for the ECI.[1]

On a quarterly basis estimates are published for the following categories: nationwide, 5 Major Industry Divisions (MIDs), 9 Major Occupational Groups (MOGs), 4 regions, union/nonunion, and metro/nonmetro. The estimates currently published include both quarter-to-quarter and year-to-year relatives; indices are calculated but currently not published. By referring to some of the estimates as indices, all that is meant is that these estimates are ratios which relate the current quarter to the base period. The current indices are not pure Laspeyres measures relative to the base period in the sense that constant weights are not associated with each quote over time. The general formula for our estimate of the national index at time t involves the product of a series of quarter-to-quarter relatives, namely

$$\hat{R}_t = \prod_{t=1}^{L} \hat{R}_{t,\ t-1} \quad \text{, where}$$

$$\hat{R}_{t,\ t-1} = \sum_s \sum_k C_{sk} \hat{\bar{X}}_{sk,t} \Big/ \sum_s \sum_k C_{sk} \hat{\bar{X}}_{sk,t-1}, \text{ where}$$

C_{sk} = employment represented by k^{th} sample occupation in the s^{th} 2-digit Standard Industrial Classification (SIC).[2] This measure of employment is based on the 1970 Census of Population.

$\hat{\bar{X}}_{sk,t}$ = estimated total average hourly compensation for the s^{th} SIC, k^{th} occupation at time t.

It may also be noted that $\hat{\bar{X}}_{sk,t}$ is in turn a ratio itself, namely

$$\hat{\bar{X}}_{sk,t} = \frac{\sum_{i(t)\varepsilon N} \delta_{ski}\, w_{ski,t}\, E_{ski}\, \bar{X}_{ski,t}}{\sum_{i(t)\varepsilon N} \delta_{ski}\, w_{ski,t}\, E_{ski}} \quad \text{, where}$$

$\bar{X}_{ski,\ t}$ = reported total average hourly compensation for the i^{th} establishment, k^{th} occupation, s^{th} SIC at time t. It is comprised of two parts: $\bar{X}_{1\ ski,\ t}$ -- average hourly wage and $\bar{X}_{2\ ski,\ t}$ -- average hourly benefit cost. These values may be reported or imputed.

E_{ski} = reported employment for i^{th} establishment, k^{th} occupation, s^{th} SIC at the time of initiation.

$w_{ski,t}$ = 'sample' weight associated with the ski^{th} unit of observation at time t. The weight of a unit may vary over time due to sample supplementation.

δ_{ski} = random variable reflecting the number of times the ski^{th} unit is selected at every stage of sampling except the phase I occupational sample.

$i(t)\varepsilon$ N denotes that the set of establishments being summed over is a function of time t. In particular, the summation is only over the 'active' set of establishments (i.e. respondents and all temporary nonrespondents) having data both in time t and t-1. (This data may be imputed.) The summation here is over all establishments since we are discussing the national index.

This estimate is a product of ratio estimates. In addition, the numerator and denominator of each ratio taken individually involve sums of ratios themselves. The sample weights as noted above can vary over time, but for the relative $\hat{R}_{t,t-1}$ relating time t to t-1, the weights are the same. Also, declining sample may result in some of the C_{sk} being effectively 'zeroed out' for some of the relatives due to lack of sample.

One collateral effect of the current form of the index is that some subindices may appear inconsistent with the national index. Due to this consideration, there are now plans to modify the current estimation procedure, and these plans have been scheduled for future implementation. The proposed form of the estimate is $\hat{R}'_t$ =

$$\frac{\sum_\alpha C_\alpha \left(\hat{\bar{X}}_{1\alpha,0} \prod_{t=1}^{L} \hat{R}_{1\alpha;t,t-1} + \hat{\bar{X}}_{2\alpha,0} \prod_{t=1}^{L} \hat{R}_{2\alpha;t,t-1}\right)}{\sum_\alpha C_\alpha \left(\hat{\bar{X}}_{1\alpha,0} + \hat{\bar{X}}_{2\alpha,0}\right)},$$

where α denotes an estimation cell which is more detailed than the current SIC/occupation cell and which would not cross over any of the subindex categories (for example: SIC/occupation/region/bargaining status/metro status).

C_α = estimated 1970 Census of Population employment represented by the α^{th} cell. This estimate is based on the base period sample and employment counts from the 1970 Census at the 2-digit SIC/occupation level.

$\hat{\bar{X}}_{1\alpha,0}$ = estimated average hourly wage for the α^{th} cell at time 0 .

$\hat{\bar{X}}_{2\alpha,0}$ = estimated average hourly total benefit cost for the α^{th} cell at time 0 .

$$\hat{R}_{1\alpha;t,t-1} = \frac{\sum_{i(t)\varepsilon\alpha} \delta_{\alpha i}\, w_{\alpha i,t}\, E_{\alpha i}\, \bar{X}_{1\alpha i,t}}{\sum_{i(t)\varepsilon\alpha} \delta_{\alpha i}\, w_{\alpha i,t}\, E_{\alpha i}\, \bar{X}_{1\alpha i,t-1}}$$

is the relative to move wages ,

where $\delta_{\alpha i}$, $w_{\alpha i,t}$, $E_{\alpha i}$, $\bar{X}_{1\alpha i,t}$, and $\sum_{i(t)\varepsilon\alpha}$ are defined in a fashion analogous to that for the current index except that the estimation cell is α rather than sk. ($\hat{R}_{2\alpha;t,t-1}$ is the relative to move total benefits and is defined analogously to $\hat{R}_{1\alpha;t,t-1}$.)

One distinction between the two estimates is that the estimation cells of the proposed estimate are smaller. The primary distinction, however, is that the relatives, $\hat{R}_{1\alpha;t,t-1}$ and $\hat{R}_{2\alpha;t,t-1}$, used to move the average hourly wage and average hourly benefit cost, respectively, are formed at the cell level in the proposed estimate rather than at the national level as is true for the current estimate. To see this we can rewrite the current estimate as

$$\hat{R}_t = \frac{\sum_s \sum_k C_{sk} \hat{X}_{sk,0} \left(\prod_{t=1}^{L} \hat{R}_{t,t-1} \right)}{\sum_s \sum_k C_{sk} \hat{X}_{sk,0}}$$

This relative has the same general form as the proposed estimate except that the relative $\hat{R}_{t,t-1}$ is at the national level. A second (obvious) difference is that a single relative is used to move total compensation in the current estimate as opposed to separate relatives for wages and benefits in the proposed estimate. The proposed estimate leads to an index which does not have the consistency problems of the current index. (This will be described in detail in a later section.)

II. Data Collection

To describe the estimation system it is necessary to first discuss the variety of data inputs.

A. Wages

For wages, the data collected is the same for both the base period and the quarterly update,the average hourly wage rate of employees currently matching the occupational definition. The field agent also reports the status of the nonrespondent: whether he is in a strike situation, closed seasonally, or a temporary nonresponse for some other reason. These status codes are later used in the imputation and estimation.

B. Benefits

The collection of benefit data can be broken up into two sections: 1) benefit initiation and 2) benefit update. These sections are briefly described.

Initiation

The benefit initiation is done through a personal visit. The field agent is supposed to collect 1) benefit existence information and 2) for benefits that exist, data elements from which cost figures will be calculated.

The benefit existence information is obtained first before any of the specific data elements are collected. The reasons for this are 1) the benefit existence information is very easy to obtain in comparison to the specific data elements and 2) this information can be used during imputation for those benefits where no specific data elements have been supplied. Therefore, if the respondent refuses to cooperate any further during the request for specific data elements, at least some information will have been obtained for each benefit, and this will be useful during imputation.

After the field agent collects the existence information, he must collect data elements for each benefit in the respondent's benefit package. The collection of these data elements is a complex task due to both the diversity of the types of benefit plans and the different ways each plan can be reported. The field agent first identifies the type of plan that exists. From this information, a decision is made as to the best way to report the plan. The options are 1) an expenditure, 2) a rate (e.g. 5% of gross earnings), or 3) practice and usage (e.g. 2 weeks vacation for employees with over a years service and 30 incumbents in that category). All data elements are sent to the national office where they are converted to a cents/hour worked figure (if they are not already in that form).

Quarterly Update

After benefits have been initiated, a shuttle form is mailed to each respondent on a quarterly basis. The form requests a description of all changes in benefit practices, including new benefits and dropped benefits. The respondent returns this form to its regional office. The regional offices are responsible for obtaining 1) new data elements that reflect the benefit changes and 2) updated existence information if necessary. Once the new data elements are received by the national office new cents/hour worked figures reflecting both benefit practice and wage changes are calculated.

III. Monitoring Incoming Data

Once the data has been collected, it can be entered into the estimation system in many different forms. To control data consistency and accuracy, consistency edits and screening programs are utilized prior to estimation. There are consistency edits for both wages and benefits. For wages, the establishment employment, current quarter average hourly rate, and publication characteristics (union/nonunion, metro/nonmetro, etc.) are reviewed if there have been any marked changes. For benefits, owing to the variety of reported data, there are a multitude of edit checks. There are edits for valid status codes and checks on valid ranges of hours related data. There is a validation of new benefit entries and a review for internal consistency among benefits (e.g. Social Security and Federal Railroad Retirement cannot both apply to the same occupation).

In addition to the consistency edits, there are also screening programs for both wages and benefits. These have not been implemented as yet but will be in the near future. The general procedure for screening wages is based on standardizing the natural log of the quarterly wage change in predetermined industry/occupation/region cells and comparing it to a predetermined value in each cell (which is a function of the occupational employment). This latter value is predetermined so as to provide approximately 50 quotes each quarter for verification.

The benefit screening program calculates each benefit cost as a percentage of total compensation within a Major Industry Division/Major Occupational Group and ranks them from smallest to largest. Then each observation beyond the Kth largest and smallest percentile will be screened out for further review. Also, the quarterly changes are ranked within the Major Industry Division/Major Occupation Group, and some fixed percent of the outliers will be screened out for review.

Since respondents do not always have records of benefit costs or, if they do have them, the records may not be organized by occupation (our unit of observation), data estimates are sometimes provided. Data collectors have been provided with a set of codes to describe the nature of the data they obtain: data obtained directly from primary sources for the unit of observation, data obtained from primary sources for a reporting unit broader than the unit of observation, data obtained from secondary records, data obtained by a method inconsistent with ECI methodology but which is reproducible, etc. This information will be analyzed along with other data to evaluate the quality of the estimates.

IV. Estimation

A. Allocation Module

Once the benefit costs are available for all responding establishments, the next step is to make sure the cost figures are all in a similar format. This means that each cost figure must represent the cost of one and only one benefit within an establishment/occupation. This is not automatically true because the respondent is given the option of reporting cost figures that represent costs for more than one benefit item (collapsed benefits). This is an important option because the respondent does not always have cost figures available for each benefit separately. For example, if a respondent receives both life and health insurance from the same insurance company, the insurance company may only charge the respondent one price for both plans. In this situation, individual cost data does not exist; therefore, the benefit system must be able to handle collapsed data. The following problems arise from collapsed benefits:

1) Since imputations for missing data are done benefit by benefit, these collapses cannot be used in any of the imputation calculations, at least not in their collapsed form. Since there are a significant number of benefit collapses (a crude approximation is 10% of all benefit data), not using the collapsed data could add bias to the ECI.

2) The respondent always has the option of reporting 'data not available' for a given quarter. If two benefits that are collapsed together in a prior quarter both receive a status code of 'data not available' for the current quarter, imputational difficulties can develop unless the computer system can keep track of the varying pattern of benefit collapses from one quarter to another by cross-checking collapsed benefits for all possible situations. This procedure, however, seems to be too complex to implement efficiently. Besides, it does not solve the first problem. The allocation of collapsed benefits is relatively simple and solves both problems.

The main idea behind the allocation procedure is to calculate an average level for each benefit. Only benefit costs that aren't involved in any collapsing (i.e. benefit costs that are reported individually) can be used in these calculations. The collapsed data can then be allocated proportionately to these levels. Since the distribution of benefits will vary for each benefit, the basic estimation cells will have to be collapsed until enough data is available to calculate reliable estimates.

The basic assumption made with this procedure is that the actual cost proportions of individual benefits involved in collapses are the same as the cost proportions of individual benefits not involved in any collapses. If this is true, this process should be relatively unbiased. In any event, the allocation does increase the number of cost figures used in the imputation, thereby reducing the variance (and hopefully the total error).

B. Imputation

Inherent in making estimates for any voluntary survey is the problem of what to do about nonresponse. Practically, adjusting for nonresponse can be accomplished by either applying a factor to the respondent's data or by imputing actual values to the nonrespondents. The desire in either case is to reduce the total mean square error. Nonresponse falls into 2 categories: nonresponse in the base period and nonresponse on an ongoing quarterly basis. Since the imputation for wages and benefits differs somewhat, they will be treated separately.

1. Wages

There has been no base period adjustment for nonresponse per se because it was considered desirable to study the data first before recommending a meaningful nonresponse cell and nonresponse adjustment. (Nonresponse adjustments need not be made based on strata used for selection nor at the estimation cell level). This data has now been analyzed, and analysis reveals that there is a high correlation between the size of an establishment, response rate, and reported level of quarter-to-quarter change. In particular, establishments with small employment (1-7 employees) had a very low response rate (21%) and the largest number of no changes quarter-to-quarter. Establishments from larger size classes had response rates which were generally in the range 90-95%. (Other characteristics analyzed, region and Major Industry Division, did not show the marked relationship that size class did). A nonresponse adjustment making use of this data will be implemented in the future.

Once data is being received for a unit of observation, it may occur that there is a temporary nonresponse, seasonal nonresponse, or that the schedule is still pending at the close of collection. In this situation the previous quarter's data (reported or imputed) is moved by the average wage change of a matched sample of establishment/occupations at some collapsed cell level. The cells for imputation are predetermined and follow a prescribed sequence (described in the next section) based on satisfying certain criteria. The criteria for determining the collapse level is that cells are collapsed (starting with the most detailed cell) until the weighted employment of the respondents reaches some acceptable percentage of the weighted employment of all 'active' schedules in the cell. An establishment/occupation can receive imputed data for no more than 4 consecutive quarters before being reviewed. This allows us the flexibility of using a respondent's data even if he occasionally is a nonresponse for one reason or other.

2. Benefits

After the collapsed benefits have been allocated, the imputation process is begun. All imputations are made to fill in any gaps in a respondent's benefit data (i.e. a nonresponse adjustment). There are no problems with respondents who supply all necessary wage and benefit data. A number of respondents, however, will only partially respond. Furthermore, since wages represent approximately 75% of total compensation, a decision was made that no wage data would be excluded from the total compensation indices because of a lack of benefit data. This is done to maximize the use of available data and hopefully reduce the mean square error. Therefore, in order to obtain total compensation figures for all wage respondents, imputations must be made (when necessary) to obtain a total benefits figure. Adding wages to the total benefit figures yields total compensation. The total compensation figures can then be used to calculate total compensation indices.

To decrease the mean square error in the imputation process, all available data must be used. Since the benefit existence information is usually available, imputations can be made benefit by benefit.

The types of benefit existence information are given below:

1) Benefit practice exists; data is available and is wage related.[3]

2) Benefit practice exists; data is available and is nonwage related.

3) Benefit practice exists; data is not available and is wage related.

4) Benefit practice exists; data is not available and is nonwage related.

5) Benefit practice does not exist.

6) It is unknown whether the benefit practice exists or not.

Using the benefit existence information, imputations are made in the following manner.

Base Period Imputations

As mentioned earlier, all wage respondents will have a total benefit cost figure. This benefit cost will either be: 1) totally supplied by respondent, 2) partially supplied by respondent and partially imputed **for** or 3) totally imputed for. During the base period all imputations will be average benefit levels. There are three possible situations where levels are imputed for a given benefit:

1) Where a wage related benefit plan exists and data is not available, the average cost, based only on those plans in the imputation cell for which data is wage related and available, is imputed.

2) Where a nonwage related benefit plan **exists**, but data isn't available, the average cost, **based** on only those plans in the imputation cell for which data is nonwage related and available, is imputed.

3) Where it is not known whether the benefit practice exists or not, the average cost, based not only on all plans in the imputation cell for which the benefit practice exists but also on those cases for which no plan exists (zero cost), is imputed.

The distribution of each benefit is different. This implies that calculating imputation levels across fixed cells will yield reliable estimates for some benefits and poor estimates for others. To compensate for the distributional difference, the estimation cells will be collapsed until a reliable estimate can be calculated.

Ongoing Benefit Imputation

During subsequent quarters, the nonresponse imputations can be grouped into five categories:

1) wage related quarter-to-quarter average benefit changes (movements).

2) nonwage related quarter-to-quarter average movements.

3) wage related average levels when the benefit practice is known to exist.

4) nonwage related average levels when the benefit practice is known to exist.

5) overall average levels when it is not known whether the benefit practice exists or not.

The wage and nonwage related average movements are calculated over available wage and nonwage related benefit plans, respectively, based on only those benefit plans which exist for both quarters. The average levels are calculated in a manner similar to the base period average levels.

The use of both movements and averages makes this imputation more complicated than the base period imputation. The complication comes from the fact that zero is a valid benefit cost. This implies that the movements may be of an indeterminate form (i.e. $\frac{o}{o}$ or $\frac{c}{o}$, where c is some positive value). When the movement is of the form $\frac{o}{o}$, then the prior quarter's data is either moved laterally if the benefit is nonwage related or moved by the national quarter-to-quarter wage relative if the benefit is wage related. When the movement is of the form $\frac{c}{o}$, then a wage or nonwage related level, whichever is appropriate, is the imputed value for a 'data not available' benefit. When the movement is of the form c_2/c_1 ($c_1 > 0$), it is applied to the prior quarter's benefit cost, and the resulting quantity is used as the current quarter imputed value. This assumes that the prior quarter cost is not zero. If the prior quarter cost is zero, then the appropriate wage or nonwage related current quarter level is used as the imputed value.

When the existence information is not known or when the existence information is known but data has not been available for a specified number or consecutive quarters, then the imputation is the same as the base period imputation where the existence information is not known.

All movements and levels are calculated over appropriate collapsed cells. The collapse cells for benefits are slightly different than those for wages and are described in the next section. The collapse criteria in terms of an acceptable 'weighted response rate' are also somewhat different. Collapsing may introduce some bias, but the mean square error is hopefully reduced.

C. Collapse Cells

It is sometimes advisable to collapse a cell having a low response rate with a cell which is similar with respect to the variable being measured but has a high response rate. In this way the mean square error is hopefully reduced by making the decrease in variance greater than the corresponding increase in bias. For this reason collapse cells and criteria for collapsing have been established for the ECI. The collapse cells are based on grouping cells which are homogenous with respect to the variable of interest. A sequence is also established in which the cells which are collapsed together are progressively less homogeneous. The sequence of collapses for the wage data is SIC/occupation, SIC/Major Occupational Group (MOG), cluster of related SICs/MOG, subindustry (e.g. durable goods)/MOG and Major Industry Division (MID)/MOG. If the cells have been collapsed to the MID/MOG level and the criteria for an acceptable level of response has still not been satisfied, then for certain MOGs there are additional collapses: Professionals with Managers, NonTransport Operatives with Transport Operatives, Laborers with Service Workers, Sales Persons with Clerks, and Craftsmen with NonTransport Operatives. (The last two collapses are only employed if Sales and Craftsmen, respectively, do not meet the collapse criteria--not vice versa). Once these have been made, then the industry collapse from SIC to cluster, cluster to subindustry, and subindustry to MID is repeated. (The collapse across MOGs has rarely been used and will not be one of the allowable collapses in the future.) The collapse pattern for benefits is similar to that for wages except that 1) certain establishment/occupation benefits are identified as being wage related while others as being nonwage related, and imputation is always done separately for these two categories and 2) the collapsing of MOGs is not allowed. In forming the collapse cells and sequence, an attempt was made to group cells with similar wage and benefit levels and changes.

Since collapse cells are formed without regard to region, union/nonunion status, and metro/nonmetro status, lack of sufficient sample results in collapsing across these (publication) characteristics. In practice, however, only approximately 6% of the 'active' establishment/occupation quotes are imputed for on a quarterly basis; therefore, the effect on the published data is probably not large.

D. Subindices

Under the current estimation formula there is sometimes an inconsistency between the national index and the subindices. By inconsistency, what is meant is that the national index may fall outside the range of the subindices. For example, the metro and nonmetro subindices between December 1976 and December 1977 were 6.9% and 6.7%, respectively, while the national index was 7.0%. To see how this can occur, the current construction of the subindices must first be described.

The 1970 Census of Population employment counts are generally not available for the subindex categories (e.g. union and metro); therefore, these quantities must be estimated from the sample. This is done independently for each subindex characteristic (e.g. union/nonunion status) by multiplying the estimated proportion of employment at the SIC/occupation level within each subcategory (as determined from the sample) by the Census of Population employment represented by that SIC/occupation. Average hourly compensation bills based on those units having the subindex characteristic are also calculated at the SIC/occupation cell level.

These estimated Census employments are then combined with estimated average hourly compensation bills to form the quarter-to-quarter relative (for the subindex characteristic). The subindices are compounded in a manner analogous to that for the national index. That is, the subindex for time t, say for the metro subindex, is the product of quarter-to-quarter relatives, namely

$$\hat{R}^{M}_{t} = \prod_{t=1}^{L} \hat{R}^{M}_{t,t-1} \quad ,\text{where}$$

$$\hat{R}^{M}_{t,t-1} = \sum_{s}\sum_{k} C^{M}_{sk} \hat{\bar{X}}^{M}_{sk,t} \;/\; \sum_{s}\sum_{k} C^{M}_{sk} \hat{\bar{X}}^{M}_{sk,t-1}$$

and C^{M}_{sk}, $\hat{\bar{X}}^{M}_{sk,t}$ are defined in a similar fashion to the national index except that here the Census employment reflects only metropolitan establishments as does the average wage.

If there is varying nonresponse over time, so that wage bills do not cancel, the national index need not fall within the range of the subindices. The reasons for this are that 1) the national wage bill does not equal the sum of the subindex wage bills and 2) the relative to move the subindices is not the same as the relative to move the national index. For example, assume we are talking about the metro/nonmetro subindices and the national index for time t=2, where $R_2^{Metro} = (X_1^{Metro}/X_0^{Metro})(X_2^{Metro}/X_1^{Metro*})$ and $R_2^{NonMetro} = (X_1^{NonMetro}/X_0^{NonMetro})(X_2^{NonMetro}/X_1^{NonMetro*})$ are the corresponding subindices formed from the estimated wage bills and $R_2 = (X_1^{National}/X_0^{National})(X_2^{National}/X_1^{National*})$ is the national index, and where corresponding wage bills for time t=1, say X_1^{Metro} and X_1^{Metro*}, are not equal because they are calculated from a different set of respondents. Then R_2 does not necessarily fall in the interval between R_2^{Metro} and $R_2^{NonMetro}$.

To eliminate this inconsistency, it is necessary to utilize the same relative for updating subindices as is used for the national. The implication of this is that the ideal estimation cell is the Major Industry Division/Major Occupational Group/region/union-nonunion status/metro-nonmetro status cell level since only at this level is the relative both meaningful and consistent for all publication categories simultaneously. The relative may be calculated at a broader level, but it should be applied to each cell separately.

E. Discussion of Sample Weights

The general form of the estimate of the index for time t as discussed in the introduction is

$$\hat{R}_t = \prod_{t=1}^{L} (\sum_{s}\sum_{k} C_{sk} \hat{\bar{X}}_{sk,t} \;/\; \sum_{s}\sum_{k} C_{sk} \hat{\bar{X}}_{sk,\ t-1}).$$

If there is no nonresponse over time, this becomes $\hat{R}_t = \sum_s \sum_k C_{sk} \hat{\overline{X}}_{sk,t} / \sum_s \sum_k C_{sk} \hat{\overline{X}}_{sk,0}$.

Now, what we are trying to estimate is

$$R_t = \sum_s \sum_k E_{sk} \overline{X}_{sk,t} / \sum_s \sum_k E_{sk} \overline{X}_{sk,0}, \text{ where}$$

E_{sk} is 1970 Census employment for the s^{th} SIC, k^{th} occupation, and

$$\overline{X}_{sk,t} = \sum_i E_{ski} \overline{X}_{ski,t} / \sum_i E_{ski} = X'_{sk,t} / E'_{sk}$$

(E_{ski} and $\overline{X}_{ski,t}$ have already been defined).

Our estimator of R_t, displaying the 'indicator' functions δ_{ski} and δ_{sk} and the weights w_{ski} and w_{sk}, is $\hat{R}_t$ =

$$\frac{\Sigma\Sigma\delta_{sk} w_{sk} E_{sk} (\Sigma_{sk}\delta_{i} w_{ski} X'_{ski,t} / \Sigma_{i}\delta_{ski} w_{ski} E_{ski})}{\Sigma\Sigma\delta_{sk} w_{sk} E_{sk} (\Sigma_{sk}\delta_{i} w_{ski} X'_{ski,0} / \Sigma_{i}\delta_{ski} w_{ski} E_{ski})},$$

where δ_{sk} is the number of times the k^{th} occupation in the s^{th} SIC was selected in the Phase I occupational sample and w_{sk} is the corresponding weight, $X'_{ski,t} = E_{ski} \overline{X}_{ski,t}$ is the 'total compensation' for the ski^{th} unit of observation at time t, and δ_{ski} and w_{ski} have already been defined.

Let us first consider $\hat{X}'_{sk,t} = \sum_i \delta_{ski} w_{ski} X'_{ski,t}$. Assume we wish to determine the w_{ski} such that $E(\hat{X}'_{sk,t}) = X'_{sk,t}$. To do this, we employ a sequence of conditional expectations. It is clear that taking the expection of δ_{ski} and setting $w_{ski} = 1/E(\delta_{ski})$ will result in an unbiased estimate of $X'_{sk,t}$ since $E(\hat{X}'_{sk,t}) = E(\sum_i \delta_{ski} w_{ski} X'_{ski,t}) =$

$E(\sum_i \delta_{ski} (1/E(\delta_{ski})) X'_{ski,t}) = \sum_i X'_{ski,t} = X'_{sk,t}$ (Note: Here the summations are over all establishments in the universe so that δ_{ski} is the only random variable). The difficulty with this approach is that at some stage of taking conditional expectations one is left with a product of random variables which is difficult, if not impossible, to simplify into terms involving only population parameters, without making further assumptions. This situation is true for the original sample and is complicated by the addition of certain supplements which are dependent on the original sample.

Two approaches can be taken to this problem:

1) Make assumptions about the correlation of dependent variables, in particular, that they are uncorrelated since they are based on a large initial refinement sample followed by a sizable Phase I subsample. The result is stable estimates of our universe values. Then our sample estimates of these universe values can be substituted to provide the estimated weights.

2) Start with the assumption that the weight to be derived is a random variable. Then, if we determine w_{ski} such that $E(\delta_{ski} w_{ski}) = 1$, $\hat{X}'_{sk,t}$ will be an unbiased estimate of $X'_{sk,t}$. This concept of a weight is more general and encompasses the situation in which the weight is a constant equaling $1/E(\delta_{ski})$. With this approach, no additional assumptions are necessary.

Weights based on these two approaches are quite similar in appearance; however, their variance properties need to be studied. For the most general situation in which the original sample was followed by a dependent supplement, the weight (representing all stages of selection except the Phase I occupational sample) of the ski^{th} probability unit selected in the final sample is $w_{ski} = B_{si}(1/(F_{si}P_sD_{ski} + F'_{si}T_i))$, where

B_{si} = inverse of the expected value of the product of the number of times the i^{th} sample establishment in the s^{th} SIC was selected in the initial refinement sample and the number of times it was selected in the Phase I establishment sample.

F_{si} = expected number of times the i^{th} sample establishment in the s^{th}SIC was selected in the original Phase II establishment sample (given all samples from the preceding stages).

P_sD_{ski} = expected number of times the k^{th} sample occupation in the i^{th} sample establishment in the s^{th} SIC (where P_s patterns were designated) was selected for the 2-way controlled selection (given all samples from the prior stages). Here, D_{ski} is the expected number of times the ski^{th} unit is selected when a single pattern has been designated.

F'_{si} = expected number of times the i^{th} establishment in the s^{th} SIC was selected in the Phase II establishment sample of the supplement (given all samples from the prior stages).

T_{si} = factor which reflects the dependency between the original and the supplemental sample. For a sample unit selected in the supplement, it only remains in the supplement if it was not already selected in the original.

The weight w_{ski} and the corresponding δ_{ski} also apply to the estimate of E'_{sk}, namely $\hat{E}'_{ski} = \Sigma\, \delta_{ski} w_{ski} E_{ski}$.

w_{sk} can be derived similarly; i.e. $w_{sk} = 1/E(\delta_{sk}) = I_{sk} / E_{sk}$, where I_{sk} is the Phase I occupational sampling interval for the s^{th} SIC, k^{th} occupation.

Then, $\delta_{sk} w_{sk} E_{sk} = \delta_{sk} I_{sk}$, and letting $C_{sk} = \delta_{sk} I_{sk}$, we finally obtain our current estimate

$$\hat{R}_t = \Sigma\Sigma C_{sk} (\hat{X}'_{sk,t} / \hat{E}'_{sk}) / \Sigma\Sigma C_{sk} (\hat{X}'_{sk,0} / \hat{E}'_{sk})$$

FOOTNOTES

[1] General Survey Design Aspects of the ECI, Easley Hoy, 1978 ASA Proceedings, San Diego, CA.

[2] 2-digit Standard Industrial Classification is a more detailed breakdown of the economic activities in a Major Industry Division. Throughout this paper any reference to Standard Industrial Classification is to the 2-digit Standard Industrial Classification unless otherwise stated.

[3] A benefit is wage related when the benefit cost increases as the wage increases.