Recent Modification of Imputation Methods for National Compensation Survey Benefits Data

by Sarah Stafira
Bureau of Labor Statistics

Originally Posted: August 28, 2009

The NCS modified its methodology for imputing benefits data for March 2009 because prior methods allowed for errors in imputed data to be carried forward from one quarter to the next.

Introduction

The Bureau of Labor Statistics (BLS) collects and publishes a variety of data on employee benefits as part of the National Compensation Survey (NCS) program. The NCS modified its methodology for imputing benefits data for March 2009 because prior methods allowed for errors in imputed data to be carried forward from one quarter to the next. This article describes the NCS imputation methodology for benefits data item nonresponse, notes the change to the imputation process that is applied to the data from the March 2009 quarter, and explains why the change is necessary.

Background

The NCS comprises a sample of private industry and State and local government establishments that are selected using a multistage sample design. All sampled establishments are asked to supply data on wages, and a subset also provides data on employer-provided benefits and associated costs. During the initial contact with a sampled establishment, the NCS selects occupations and collects data for these sampled occupations, along with establishment information. The establishment is periodically recontacted to determine if there are any changes in the data collected previously.[1] The NCS is a voluntary survey, so selected establishments can decline to participate or can participate partially by supplying responses to only certain survey items.

As in any sample survey, estimates generated as part of the NCS program are subject to both sampling error and nonsampling error. Sampling error occurs because a sample makes up only a part of the population being studied; different samples of the population can produce different estimates.[2] Standard errors are calculated for benefit estimates to serve as a measure of sampling error. Nonsampling error is error coming from sources other than the sampling process. The primary sources of nonsampling error are survey nonresponse, mistakes in data collection, and data processing errors. Nonsampling error is generally not measured, but the NCS employs procedures to mitigate nonsampling error, such as weight adjustments for nonresponse and quality assurance programs to reduce collection and processing errors.

The NCS has three kinds of nonresponse: establishment, occupational, and item nonresponse. Establishment nonresponse is addressed by adjusting the weights of responding establishments (respondents) that are similar to the nonresponding establishments. Occupational nonresponse is handled in a similar fashion; that is, weights of responding occupations, in the same establishment or another, are adjusted to represent similar occupations for which data were not provided.

Item nonresponse happens when a respondent supplies some, but not all, data for an occupation. For example, a respondent may know the provisions of the retirement plans offered, but be unable or unwilling to supply the percentage of workers who participate in each type of plan. Item nonresponse is addressed with item imputation. Imputation of data is a process by which a missing data element is assigned a value obtained from a responding unit with similar characteristics.

Imputation Methodology For NCS Benefits Data

There are several methods of imputation that can be used to address item nonresponse, including regression modeling, cell mean imputation, and nearest neighbor imputation. The NCS benefits program uses a nearest neighbor, within-cell approach to impute for missing participation, access, and provisions data.[3] In this method, imputation classes, or cells, are formed

based on auxiliary data. The auxiliary data used by the NCS are establishment and occupational characteristics known for all units and include the following:

1. Census region (Midwest, South, Northeast, or West)
2. Two-digit North American Industry Classification System (NAICS) code
3. Full-time/part-time status
4. Union/nonunion status
5. Occupational grouping (based on the Standard Occupational Classification System)
6. Industry grouping (based on the NAICS)
7. Establishment size (based on the establishments number of employees)
8. Selected benefit provisions[4]
9. Ownership (private industry or public sector)
10. Benefit type[5]

An “unusable” unit, or recipient, receives (is assigned) the value for the characteristic of interest from a “usable” unit, or donor, within the same cell that is “nearest” to the unusable unit; “nearness” is defined as the minimum absolute difference in reported employment between the recipient and donors within a cell. If a donor unit is not found when all variables are used to form the imputation cells, then the cells are redefined by disregarding one of the variables, thus expanding the pool of donors. For the NCS, the first variable dropped is census region.

At this point, if a donor unit is not found, then the imputation cells are redefined again by ignoring the two-digit NAICS code. The process of dropping variables to increase the donor pool continues using a predetermined hierarchy until a donor unit is identified. The list of auxiliary characteristics given above provides the order in which the variables are dropped.

It should be noted that benefit type and ownership are never dropped. In rare situations a recipient will not find a donor, even when the imputation cell is based only on benefit type and ownership. If this happens, the data item will remain missing.

Exhibit 1 shows the selection of a donor unit for a recipient in nearest neighbor, within-cell imputation. Suppose unit R1 is a recipient and units D1, D2, and D3 are donors in the imputation process used to impute missing participation for defined benefit retirement plans. The imputation cell is formed based on the following characteristics: benefit type, ownership, selected benefit provisions, establishment size, industry grouping, occupational grouping, union/nonunion status, full-time/part-time status, two-digit North American Industry Classification System (NAICS) code, and census region. The reported employment of each unit is also given.

Exhibit 1. Donor selection in nearest neighbor, within-cell imputation

Unit	Benefit Type	Owner-ship	Benefit Provision	Establish-ment Size	Industry Grouping	Occupa-tional Grouping	Union / Nonunion	Full-time / Part-time	Two-digit NAICS	Census Region	Reported Employ-ment
R1 (recipient)	Defined benefit	Private industry	No employee contribution required	Less than 100 employees	Wholesale & Retail Trade	Office & Adminis-trative Support Occupations	Nonunion	Full-time	Wholesale Trade	Midwest	50
D1 (donor)	Defined benefit	Private industry	No employee contribution required	Less than 100 employees	Wholesale & Retail Trade	Office & Adminis-trative Support Occupations	Nonunion	Full-time	Wholesale Trade	South	25

Unit	Benefit Type	Owner-ship	Benefit Provision	Establish-ment Size	Industry Grouping	Occupa-tional Grouping	Union / Nonunion	Full-time / Part-time	Two-digit NAICS	Census Region	Reported Employ-ment
D2 (donor)	Defined benefit	Private industry	No employee contribution required	Less than 100 employees	Wholesale & Retail Trade	Office & Adminis-trative Support Occupations	Nonunion	Full-time	Wholesale Trade	West	65
D3 (donor)	Defined benefit	Private industry	No employee contribution required	Less than 100 employees	Wholesale & Retail Trade	Office & Adminis-trative Support Occupations	Nonunion	Full-time	Retail Trade	West	55

There are no donor units that match the recipient unit based on all of the variables used to form the imputation cell. By ignoring census region and only considering the remaining cell variables, there are now two potential donors, specifically D1 and D2. Unit D3 is not a potential donor based on the cell formation including the two-digit NAICS because its value does not match that of R1. In order to determine which unit, D1 or D2, will serve as the donor for R1, the minimum absolute difference in reported employment between the donor and recipient is calculated. The absolute difference in employment between R1 and D1 is 25, while the absolute difference in employment between R1 and D2 is 15. Because 15 is the minimum, D2 is determined to be the donor for R1 because it is the donor "nearest" to the recipient within the cell.

Imputation At Initiation And Update Collection

At initiation--the first collection cycle for the establishment--the nearest neighbor, within-cell imputation methodology is used to fill in missing benefits access, participation, and provisions data, as needed. If there are missing benefits data in subsequent data collection, imputed data will generally be the value from the prior collection period, even if that value was imputed. NCS first introduced the "carrying forward" of prior collected or prior imputed data in the publication *National Compensation Survey: Employee Benefits in Private Industry in the United States, March 2007*.[6] The benefit publications for March 2003 through March 2006 relied solely on nearest neighbor, within-cell imputation.

Consider the example shown in exhibit 2. Unit B is initiated in cycle 1 and is missing a provision related to its life insurance plan. The life insurance plan is known to use a "multiple of earnings" formula that has a maximum payout amount, but the maximum is not known. Nearest neighbor, within-cell imputation is used at initiation to fill in the missing maximum so that the unit can be used in estimation. Unit A, a donor unit, is matched to unit B because they have the same type of life insurance plan, as well as similar establishment and occupational characteristics. All provisions are known for unit A, so the maximum life insurance value coded for unit A is used to fill in the missing life insurance maximum for unit B.

Exhibit 2. Example of nearest neighbor, within-cell imputation at initiation (cycle 1)

Prior to imputation:					
Cycle	**Unit**	**Type of Life Insurance**	**Is there a maximum?**	**Maximum value?**	
1	A (donor)	Multiple of Earnings Formula	Yes	$70,000	
1	B (recipient)	Multiple of Earnings Formula	Yes	Unknown	
After imputation:					
Cycle	**Unit**	**Type of life insurance**	**Is there a maximum?**	**Maximum value**	**Source of maximum value**

Prior to imputation:					
1	A (donor)	Multiple of Earnings Formula	Yes	$70,000	Collected
1	B (recipient)	Multiple of Earnings Formula	Yes	$70,000	Imputed from Unit A

Exhibit 3 shows the imputation of unit B at update collection by carrying forward the prior imputed data. At the next collection period, cycle 2, the respondent is still unable or unwilling to supply the maximum value associated with the life insurance plan for unit B. Because the provision was previously imputed to be $70,000, unit B retains or carries forward the maximum of $70,000.

Exhibit 3. Example of imputation at update collection by carrying forward prior imputed data (cycle 2)

Prior to imputation:					
Cycle	**Unit**	**Type of Life Insurance**	**Is there a maximum?**	**Maximum value?**	
2	B (recipient)	Multiple of Earnings Formula	Yes	Unknown	
After imputation:					
Cycle	**Unit**	**Type of life insurance**	**Is there a maximum?**	**Maximum value**	**Source of maximum value**
2	B (recipient)	Multiple of Earnings Formula	Yes	$70,000	Imputed from Unit B, Cycle 1 (carried forward)

Effect Of Nonsampling Error On Imputation

In addition to nonresponse, collection and processing errors are sources of nonsampling error that impact all surveys. Errors can occur when an interviewer fails to ask for all data items or incorrectly records a data element. Also, a respondent may misunderstand the survey question and supply an incorrect answer. To limit the number of data errors, the NCS program has a number of quality assurance programs in place, including computer edits of data, systematic review of collection units, and data collection reinterviews. Also, data collectors are extensively trained so that high standards in data collection are maintained.

Data errors in collection impact survey estimates by increasing the amount of nonsampling error. These data errors affect the imputation process when the erroneous data are assigned to a recipient record. With the application of the “carry forward” methodology in the benefit portion of the NCS program, there is the potential for data errors to remain in the imputed data, cycle after cycle, even if the data on the collected unit are corrected. Without some kind of change to the imputation methods, data errors on imputed records could be repeated in the data for the rest of the time the establishment is in the survey.

Consider the example discussed previously in which unit A, the donor used at initiation, had a maximum life insurance amount of $70,000. At update collection in cycle 3, it is discovered that the life insurance maximum amount is really $700,000, not $70,000. The data coder corrects the collected data on unit A, but due to the imputed data, the maximum amount for unit B will continue to be $70,000 because the prior imputed value of $70,000 is carried forward. Exhibit 4 illustrates this scenario.

Exhibit 4. Example of imputation at update collection by carrying forward prior imputed data (cycle 3), original donor unit is corrected

Prior to imputation:					
Cycle	**Unit**	**Type of Life Insurance**	**Is there a maximum?**	**Maximum value**	
3	A (donor)	Multiple of Earnings Formula	Yes	$700,000	
3	B (recipient)	Multiple of Earnings Formula	Yes	Unknown	
After imputation:					
Cycle	**Unit**	**Type of life insurance**	**Is there a maximum?**	**Maximum value**	**Source of maximum value**
3	A (donor)	Multiple of Earnings Formula	Yes	$700,000	Collected
3	B (recipient)	Multiple of Earnings Formula	Yes	$70,000	Imputed from Unit B, Cycle 2 (carried forward)

Minimization Of Nonsampling Error In The March 2009 Quarter

To address the potential problem of carrying forward erroneous data, a change in the imputation methodology was needed, especially given that the NCS has greatly expanded the number of detailed estimates available. Percentile estimates (10^{th}, 25^{th}, 50^{th} — median, 75^{th}, 90^{th}) of a given quantity, such as the maximum value of multiple of earnings formula life insurance plans, are of particular risk of nonsampling error because coding errors are often found among the extreme values or outliers, which could show up in the 10^{th} or 90^{th} percentiles.

To help minimize nonsampling error, the NCS has conducted additional reviews of the collected benefits data over the last several quarters. Also, the NCS modified its computer programs that assign missing participation, access, and benefit provisions data, starting with data collected and updated in the March 2009 quarter. These computer programs were changed so that all data items for a recipient unit were imputed using the nearest neighbor, within-cell methodology. That is, no prior imputed or collected data were carried forward for recipients of benefits participation, access, and provisions imputation.

Using the earlier example, exhibit 5 provides an example of imputation at update collection in which no prior data were carried forward. It shows that at update collection in Cycle 4, when all recipients are imputed using nearest neighbor, within-cell imputation, the maximum value assigned to unit B is no longer $70,000.

Exhibit 5. Example of imputation at update collection using nearest neighbor, within-cell imputation (cycle 4)

Prior to imputation:					
Cycle	**Unit**	**Type of Life Insurance**	**Is there a maximum?**	**Maximum value**	
4	A (donor)	Multiple of Earnings Formula	Yes	$700,000	
4	B (recipient)	Multiple of Earnings Formula	Yes	Unknown	
After imputation:					
Cycle	**Unit**	**Type of life insurance**	**Is there a maximum?**	**Maximum value**	**Source of maximum value**
4	A (donor)	Multiple of Earnings Formula	Yes	$700,000	Collected
4	B (recipient)	Multiple of Earnings Formula	Yes	$700,000	Imputed from Unit A

Conclusion

The NCS benefits imputation methodology for the imputation of missing access, participation, and provisions data included a process that carried forward collected or imputed data from previous collection cycles. This proved to be a potential source of additional nonsampling error. To address this problem, the NCS modified its imputation methodology for the March 2009 quarter so that prior imputed or prior collected data are not carried forward. The BLS is committed to publishing accurate, timely, and relevant data; as such, the NCS program not only strives for accuracy of its collected data through validation, but also through regular evaluation of its methods, including imputation techniques, to find ways to improve the quality of its published data. The first benefits estimates using the modified imputation methodology were published in July 2009.[7]

Sarah Stafira
Mathematical Statistician, Statistical Methods Group, Office of Compensation and Working Conditions, Bureau of Labor Statistics.
Telephone: (202) 691-6146; E-mail: Stafira.Sarah@bls.gov.

Notes

1 Generally, sampled establishments remain in NCS sample for five years before being replaced by a new panel. For more information on the sample selection process, see Larry Ernst, Christopher Guciardo, Chester Ponikowski, and Jason Tehonica, "*Sample Allocation and Selection for the National Compensation Survey*," Proceedings of the Section on Survey Research Methods, 2002, American Statistical Association, available online at: http://www.bls.gov/osmr/pdf/st020150.pdf. Additional information can also be found in the *BLS Handbook of Methods*, Chapter 8, National Compensation Survey, Description of the Survey, available online at: http://www.bls.gov/opub/hom/homch8_b.htm.

2 *BLS Handbook of Methods*, Chapter 8, National Compensation Survey, Reliability of Estimates, available online at: http://www.bls.gov/opub/hom/homch8_d.htm.

3 Separate imputation processes are used to impute for missing access, missing participation, and missing benefit provisions, as needed. Access is a measure used to indicate whether employees have a benefit plan available for their use while participation is used to indicate the percentage of those employees who actually participate in the plan. Benefit provisions are characteristics or features of a benefit plan such as the type of life insurance or the employee contribution requirement of a defined benefit retirement plan.

4 Benefit provisions data are used to define the cells if they are known for recipients. For example, in participation imputation for life insurance, if the type of life insurance plan (for example, a multiple of earnings formula) is known for a recipient, then it will be used in forming the imputation cell. If the type of life insurance is not known for the recipient, then this variable is not used in the formation of the imputation cell.

5 For a more comprehensive description of the imputation of benefits data in the NCS, see James A. Buszuwski, Daniel J. Elmore, Lawrence R. Ernst, Michael K. Lettau, Lowell G. Mason, Steven P. Paben, and Chester H. Ponikowski, "*Imputation of Benefit Related Data for the National Compensation Survey*," Proceedings of the Section on Survey Research Methods, 2003, American Statistical Association, available on the internet at http://www.bls.gov/osmr/abstract/st/st030190.htm.

6 See National Compensation Survey: Employee Benefits in Private Industry in the United States, March 2007, Summary 07-05.

7 See the BLS Economic News Release, *Employee Benefits in the United States, March 2009*, at http://www.bls.gov/news.release/ebs2.nr0.htm.