Comparing Impact of Alternative Approaches for Item Non-Response in the Job Openings and Labor Turnover Survey[1]

DRAFT, May 20, 2004

Mark Crankshaw
Bureau of Labor Statistics, Statistical Methods Staff, Room 4985,
2 Massachusetts Ave. NE, Washington, DC 20212

Key Words: Sample survey; labor statistics; donors; nearest neighbor imputation; ratio data

The Bureau of Labor Statistics produces monthly estimates for job openings, hires, and separations from the Job Openings and Labor Turnover Survey (JOLTS). The JOLTS survey has implemented a hot-deck nearest neighbor imputation algorithm to account for item non-response. For this survey, the imputed values taken from donors take the form of a ratio. The distribution of these borrowed ratios varies greatly depending on the size of the donor establishment. We wish to examine how well the profile of our imputation donors matches the profiles of reporting establishments of comparable size. If the profiles do not match, we would like to measure the extent of statistical bias. In the absence of other information, we assume non-respondents have approximately the same characteristics as respondents. We wish also to examine variations of our current imputation approach to identify the variation that most reduces bias.

1. Introduction

The Job Opening and Labor Turnover Survey (JOLTS) was initially developed in 1999 by the Bureau of Labor Statistics (BLS) to provide information about labor supply and demand. BLS developed a survey and selected a sample of approximately 16,000 establishments covering all nonagricultural businesses in the public and private sectors for the 50 States and the District of Columbia. The JOLTS survey collects data at the establishment level on Total Employment, Hires, Quits, Layoffs & Discharges, and Other Separations. While other BLS surveys such as the Current Population Survey (CPS) serve as supply-side indicators of labor, the JOLTS survey serves as a demand-side indicator of labor. Prior to the development of the JOLTS survey, there was no demand-side indicator for labor.

Like most surveys, JOLTS Statistical Methods Staff (SMS) had to develop imputation and variance estimation procedures. SMS selected a stratified estimation design with a standard Non-response Adjustment Factor (NRAF) to account for establishment non-response and a hot deck nearest neighbor imputation algorithm to account for data item non-response. Establishment non-response is operationally defined as an establishment that did not report total employment. Item non-response is operationally defined as an establishment that reported total employment but did not report one or more of the other data items. A balanced half sample replication technique was chosen to produce variance estimates.

The JOLTS survey has now produced and published over three years of estimates. This paper will assess how well the hot deck nearest neighbor has performed during that time. It will also suggest some modifications to our current hot deck nearest neighbor imputation algorithm.

The remainder of this paper is outlined as follows: Section 2 discusses JOLTS imputation in theory; Section 3 covers JOLTS imputation in practice; Section 4 details a modified approach to the current JOLTS imputation algorithm; Section 5 compares the current and modified approaches; and Section 6 offers a summary.

2. JOLTS Imputation in theory

In the JOLTS survey, item non-response (that is, when there is only a partial response to the survey) is accounted for using a hot-deck nearest neighbor imputation algorithm. The underlying assumption in this methodology is that individual missing item responses will be similarly distributed as responding units of similar size, industry division, and geographic location.

The hot deck nearest neighbor algorithm is a process in which the imputation recipient (non-respondent) takes on the currently reported value of an imputation donor (respondent) within the same imputation cell. In JOLTS imputation, responses are sorted by strata, and by reported employment within

[1] Opinions expressed in this paper are those of the author and do not constitute policy of the Bureau of Labor Statistics.

[2] I would like to acknowledge John Eltinge and George Stamas for their assistance in the writing of this paper.

strata. Records with responses for the item serve as the donor pool. The donor selected for any given recipient will be the record with a response on the item and the smallest difference in employment from among the donor population. If e_i is employment of the recipient record, then we select as a donor the record *j'* from the donor pool, *D*, such that $|e_i - e_{j' \in D}| < |e_i - e_{j \in D}|$ where $j' \neq j$. Once a recipient record has been paired with its donor, the procedure imputes a value by calculating the item ratio (item response/employment) from the donor and multiplying that ratio by the employment from the recipient record.

There are many advantages to the hot-deck nearest neighbor approach. Among these are that this imputation algorithm preserves the reported distribution of item values and it permits the use of the same sample weights for all sampled items.

3. JOLTS Imputation in practice

While the selection of the imputation donor is, in theory, predicated by a pre-specified distance function, in practice, there are conditions under which the calculated donor would not be selected. This occurs when:

1) **A donor has been flagged as an edit failure**. All JOLTS reported data are screened against predefined edit parameters to identify data that may have been misreported or incorrectly recorded. Any record flagged as an edit failure is ineligible to be a donor.

2) **A donor has been flagged as an outlier.** An univariate outlier detection algorithm is used in JOLTS to identify records that may be potential outliers within a particular industry division-size class cell. Records that are flagged as outliers are ineligible to be donors.

3) **A donor has been flagged as a 'D' failure.** It logically follows for JOLTS that reported monthly change in employment should be equal to the difference between reported new hires and total separations. A variable ,'D', is calculated for each JOLTS reporter which is equal to monthly change in employment minus the difference between hires and separations. Record which have a 'D' value that exceeds a preset tolerance (by size class) are flagged as 'D' failures and are ineligible to be donors.

All of these factors decreased the number of eligible imputation donors from which to select. With a limited number of eligible donors in an imputation cell, it is quite possible that a recipient could receive as a donor a record significantly smaller or larger than itself. These real-world situations and constraints brought into question the underlying assumption of the hot deck nearest neighbor approach. Mainly, that individual missing item responses will be similar in profile to responding units of similar size, industry division, and geographic location.

4. A Modified Approach

JOLTS SMS staff recently developed a modified approach to the current hot deck nearest neighbor approach. The purpose of the modification was to reduce, if not eliminate, the above mentioned scenario in which, due to low response rate or the presence of many ineligible donors, the "nearest" neighbor for a recipient unit could be a unit of a significantly different size. To understand the nature of the modification, it is necessary to explain the programming of our nearest neighbor approach.

In our current algorithm, the imputation cell is defined by U.S. Census Region and NAICS Industry Division. A small uniform random number from 0 to .1 is added to reported employment to break "nearness" ties—this is referred to as random reported employment. All units in the sample that are eligible for imputation are sorted within the imputation cell by ascending random employment. Table 1 details a sample imputation cell with reported employment and job openings. When the Flag variable is equal to 'Y' then the record has been flagged as an ineligible donor due to one of the conditions listed in section three.

Table 1. Sample Imputation Cell

Schedule	Random Employment	Job Openings	Flag
001	3.08	0	N
002	5.02	1	N
003	6.01	.	
004	12.00	1	N
005	15.07	3	Y
006	28.06	0	N
007	55.02	2	N
008	79.00	.	
009	79.03	5	N
010	178.04	11	N
011	229.08	8	Y
012	455.02	23	N
013	2764.04	69	Y
014	5684.03	.	
015	7955.00	.	

The current imputation algorithm creates two arrays for all records and for each data variable. The first array records the potential donors for a variable in the downward direction (that is, potential donors with less reported employment than the recipient) while the second array records the potential donors in the upward direction (that is, potential donors with more reported employment than the recipient). The length of the array is set at nine to help reduce using the same donor record twice. Only units that are eligible donors are allowed to fill the array.

The current algorithm begins with the smallest unit in the imputation cell that is in need of imputation (in the case of the sample imputation cell, it begins with schedule 003). Table 2 details the filled in arrays for schedule 003.

Table 2. Filled-in imputation array for Schedule 003

Up Dist	Up Donor	Up Rate	Down Dist	Down Donor	Down Rate
6	004	.083	**1**	**002**	**.200**
22	006	.000	3	001	.000
49	007	.036	3	001	.000
73	009	.063	3	001	.000
172	010	.062	3	001	.000
223	011	.035	3	001	.000
449	012	.051	3	001	.000
449	012	.051	3	001	.000
449	012	.051	3	001	.000

For this smallest unit, the first potential donor in the upward array (if it exists) is compared with the first potential donor in the downward array (if it exists). The potential donor that is closest with respect to random reported employment to the recipient is deemed the nearest neighbor. In the table, Up Dist refers to the distance to the nearest neighbor in the upward direction, while Down Dist refers to the distance to the nearest neighbor in the downwards direction. For Schedule 003 the nearest neighbor that would be selected in the current algorithm is Schedule 002 since Down Dist is smaller than Up Dist. The chosen donor is now ineligible to be a donor for subsequent recipients.

The algorithm then proceeds to the next smallest unit. Again, the first potential donor in the downward array (if it exists) is compared with the first potential donor in the upward array (if it exists). However, if the first potential donor in either direction of the array is now ineligible, then the second potential donor will be used in the comparison. The algorithm proceeds down the array, if necessary, until all nine potential donors in each direction have been exhausted. In the worst case scenario, the closest of the 9th place donors in the array is chosen as the nearest neighbor by default. In cells where there are many flagged units and few eligible donors it is possible that the 9th place donor may be of a considerably different size to its intended recipient. If, as suspected, the profiles of donors differ across size classes then the profile of donated ratios could differ from the respondents within size class. It is precisely this worst case scenario that the modified version of the imputation algorithm seeks to mitigate.

In the modified algorithm, instead of an array of length nine, an array of length two is used. The algorithm fills the downward array with the two closest eligible donors in the downward direction and it fills the upward array with the two closest eligible donors in the upward direction. Therefore each recipient has at most four potential donors to choose from. Instead of comparing random reported employment to determine the "nearest" neighbor, as in the current algorithm, the modified algorithm randomly selects, with equal probability, from one of the four "nearest" neighbors. Unlike the current algorithm, there is no attempt to make a donor ineligible simply because it has been used before. In the modified approach, the least "nearest" neighbor that can be selected is in the second spot on the array. This is a significant improvement over the current potential of selecting the "nearest" neighbor in the 9th spot on the array.

Table 3. Filled-in imputation array for modified algorithm

Schedule	Donor 1	Donor 2	Donor 3	Donor 4	Ran Num	NN
003	002	001	004	006	.311	001
008	007	006	009	010	.273	006
014	012	010	-	-	.851	010
015	012	010	-	-	.377	012

Table 3 illustrates the filled-in imputation array for the modified approach for the sample imputation cell. The far right column (NN) is the nearest neighbor schedule selected by the algorithm. As an example of how the nearest neighbor is selected consider Schedule 003. This schedule has four possible donors. If the random number were less than or equal to .25, then the first donor would be chosen.

If the random number were greater than .25 and less than .50, then the second donor would be chosen, and so on. Since the random number was .311 then, as shown below, the second donor is chosen as nearest neighbor.

In terms of validating the results of the current and new imputation algorithms, the new algorithm is much easier to validate for two reasons. First, only four possible values are to be selected from in the new algorithm, as opposed to eighteen possible values in the current algorithm. Second, when given the random number used to select from the four, it is easy to determine which donor will be selected. In the current algorithm it is difficult to readily determine the donor selected and the outcome of the comparisons made.

5. Comparing Distributions

Our underlying assumption in the nearest neighbor imputation approach is that individual missing item responses will be similar in profile to responding units of similar industry and size. In order to compare how "successful" our current and modified algorithms perform in holding the underlying assumption it is necessary to develop the profile of responding units by industry and size. All JOLTS respondents and donors from December 2000 to November 2003 were analyzed. From these data a distribution of ratios for both respondents and donors were produced.

Tables 4, 5, and 6 (see Appendix) document the distributions of ratios for job openings, hires, and quits. The distributions are given for the donors selected with the current imputation algorithm, the donors selected with the new imputation algorithm, and the reported values (that is, the pool of all possible donors). The tables are broken down by establishment size class. There are four bands of ratios (that is, variable relative to employment) listed. The first are donors and reporters who reported zero for a given variable. The second band are donors and reporters whose ratio is greater than zero yet less than ten percent. The third band are donors and reporters whose ratio is greater than ten percent yet less than 25 percent, and the fourth and final band are donors and reporters whose ratio is greater than 25 percent.

There are three facts that are readily apparent when the tables data is examined: first, the distribution of JOLTS ratios are not symmetric but are skewed; second, that JOLTS ratios have substantially different distributions by size; and third, that both the current and new imputation algorithms select donors which fit the profile of the distribution of ratios for reporters fairly closely.

Table 7 (see Appendix) gives a similar analysis of the distribution of JOLTS ratios by NAICS industry division in less fine detail. The table details the percentage of ratios that fall below 10 percent for donors selected with the current imputation algorithm, the modified imputation algorithm, and for reporters. The closer the imputation algorithm is to the percentage of reporters with ratios of less than ten percent, the more likely that the selected donors have the appropriate mix of small ratios (those under 10%) and large ratios (those over 10%). Again, it can be seen that the current and modified approach do an equally fair job of matching the profile of donors to that of the reporters in the overwhelming majority of industries. Given that the current imputation algorithm does not outperform the modified algorithm to any great extent, and that the modified program is much easier to validate, a case can be made that the modified imputation algorithm is an improvement over the current imputation algorithm.

6. Summary

An analysis of the distribution of JOLTS ratios was made. It was found that the distribution of JOLTS ratios is highly skewed and varies significantly by establishment size class. The current JOLTS imputation algorithm was compared with a modified imputation approach. Both the current and modified imputation approaches were matched against all possible donors. Both approaches select donors that fit the profile of all possible donors. The advantage of the modified approach is that it is easier to validate.

References

Chen, J. and Shao, J. (1999) "Jackknife Variance Estimation for Nearest Neighbor Imputation", in 1999 *American Statistical Association Papers and Proceedings.*

Clark, K. and Hyson, R. (2001) "New Tools for Labor Market Analysis: JOLTS", in *Monthly Labor Review*, December 2001.

Eltinge, J. , Groves, R., Dillman, D., Little, R., (2002) "Survey Nonresponse", New York, Wiley & Sons.

Fay, R., (1999) "Theory and Application of Nearest Neighbor Imputation in Census 2000", in *1999 American Statistical Association Papers and Proceedings.*

Little, R., (2002) "Statistical Analysis with Missing Data", New York, Wiley & Sons.

Shafer, J. (1997) "Analysis of Incomplete Multivariate Data", London, Chapman & Hall.

Appendix

Table 4. Distributions of Job Openings to Employment ratios by Size Class

			Ratio of Job Openings to Total Employment			
Type	**Size Class**	**Units**	**0**	**[0-.10]**	**[.10-.25]**	**.25+**
Current	1	17433	93.7%	2.0%	2.2%	2.0%
New	1	17433	93.2%	2.5%	2.2%	2.1%
Reported	1	35031	94.4%	1.5%	2.1%	2.0%
Current	2	36375	80.0%	12.7%	5.8%	1.4%
New	2	36375	79.8%	12.9%	5.8%	1.4%
Reported	2	68614	81.0%	11.8%	5.7%	1.4%
Current	3	39712	54.2%	41.2%	3.9%	0.6%
New	3	39712	53.7%	41.8%	3.8%	0.6%
Reported	3	67999	55.9%	39.4%	3.9%	0.9%
Current	4	24644	28.8%	67.0%	3.8%	0.4%
New	4	24644	28.5%	66.4%	3.7%	0.5%
Reported	4	39374	30.8%	64.9%	3.8%	0.5%
Current	5	17215	14.4%	81.2%	4.0%	0.3%
New	5	17215	13.6%	82.1%	4.1%	0.3%
Reported	5	25341	14.8%	80.4%	4.4%	0.5%
Current	6	9383	8.6%	87.5%	3.9%	0.1%
New	6	9383	5.7%	90.5%	3.8%	0.1%
Reported	6	10434	5.9%	89.6%	4.3%	0.2%

Table 5. Distributions of Hires to Employment ratios by Size Class

			Ratio of Hires to Employment			
Type	**Size Class**	**Units**	**0**	**[0-.10]**	**[.10-.25]**	**.25+**
Current	1	15587	89.8%	3.5%	3.8%	2.8%
New	1	15587	89.8%	3.5%	3.8%	2.8%
Reported	1	35547	90.1%	3.3%	3.8%	2.8%
Current	2	32644	68.4%	19.4%	9.5%	2.8%
New	2	32644	68.4%	19.5%	9.4%	2.8%
Reported	2	71300	68.9%	18.9%	9.4%	2.8%
Current	3	35628	37.8%	52.8%	7.5%	1.9%
New	3	35628	38.0%	52.9%	7.2%	1.9%
Reported	3	74245	39.6%	50.9%	7.6%	2.0%
Current	4	22146	20.6%	73.9%	4.3%	1.2%
New	4	22146	20.7%	74.2%	4.0%	1.1%
Reported	4	44533	23.0%	71.2%	4.3%	1.5%
Current	5	15369	8.0%	88.6%	2.4%	0.9%
New	5	15369	8.3%	88.6%	2.2%	0.9%
Reported	5	30401	10.1%	86.0%	2.7%	1.3%
Current	6	7135	2.7%	96.0%	1.0%	0.2%
New	6	7135	2.5%	96.4%	1.0%	0.2%
Reported	6	13845	3.9%	94.6%	1.3%	0.3%

Table 6. Distributions of Quits to Employment ratios by Size Class

			Ratio of Quits to Employment			
Type	**Size Class**	**Units**	**0**	**[0-.10]**	**[.10-.25]**	**.25+**
Current	1	15551	92.8%	3.0%	2.4%	1.7%
New	1	15551	92.8%	3.0%	2.5%	1.6%
Reported	1	35419	93.3%	2.7%	2.3%	1.6%
Current	2	32581	76.6%	16.4%	5.7%	1.2%
New	2	32581	76.5%	16.5%	5.7%	1.3%
Reported	2	70100	77.5%	15.4%	5.8%	1.3%
Current	3	35555	47.6%	48.6%	3.4%	0.4%
New	3	35555	47.5%	48.8%	3.3%	0.4%
Reported	3	71365	50.0%	46.3%	3.3%	0.4%
Current	4	22105	24.0%	74.1%	1.3%	0.2%
New	4	22105	24.3%	74.2%	1.2%	0.2%
Reported	4	41674	27.2%	71.4%	1.1%	0.2%
Current	5	15330	8.6%	90.7%	0.6%	0.1%
New	5	15330	8.8%	90.4%	0.6%	0.1%
Reported	5	26835	11.1%	89.9%	0.8%	0.2%
Current	6	6984	3.5%	96.1%	0.3%	0.1%
New	6	6984	3.8%	96.9%	0.3%	0.1%
Reported	6	12004	4.0%	95.8%	0.2%	0.1%

Table 7. Distribution of small ratios by Industry

Industry	Job Openings			Hires			Quits		
	Curr	Mod	Rep	Curr	Mod	Rep	Curr	Mod	Rep
Logging & Mining	97.2	97.1	97.2	91.8	91.9	91.8	97.3	97.3	97.1
Construction	95.8	95.9	95.4	83.7	84.4	83.6	94.9	94.9	94.6
Nondurable MFG	97.6	97.7	97.7	94.6	94.7	94.8	97.9	97.9	98.2
Durable MFG	97.7	97.8	97.8	95.8	95.8	95.4	98.4	98.5	98.4
Wholesale Trade	97.4	97.4	97.3	95.2	95.3	95.1	97.9	98.0	97.7
Retail Trade	93.8	94.0	93.6	86.8	87.3	86.6	93.1	93.4	92.7
Transport & Util.	96.3	96.6	96.1	94.4	94.5	94.4	97.7	97.7	97.7
Information	96.3	96.3	95.6	94.9	95.0	94.5	97.9	97.7	97.5
Finance & Insurance	95.7	94.6	94.6	96.4	96.4	96.8	98.5	98.6	98.3
Real Estate & Leasing	94.5	94.6	94.6	90.6	90.9	90.8	94.9	95.3	95.4
Prof. & Business Serv.	93.1	93.4	92.8	89.2	89.7	88.3	95.6	95.6	95.2
Educational Services	97.5	97.7	97.4	94.7	94.9	94.3	98.1	98.2	95.5
Health Care	90.6	90.4	91.0	95.0	95.0	94.5	97.4	97.4	97.0
Arts, Ent., & Rec	93.4	93.3	92.9	83.3	83.6	82.3	93.9	94.1	94.0
Accom. & Food Serv.	89.6	89.8	90.0	79.0	79.6	79.2	86.4	86.7	86.5
Other Services	94.1	93.9	93.8	91.7	91.6	91.4	95.1	94.9	95.0
Federal Govt	97.0	98.4	98.0	99.0	99.2	98.9	99.8	99.8	99.6
State & Local Ed.	99.0	99.0	98.9	98.0	98.0	97.7	99.6	99.6	99.6
State & Local Non-Ed.	94.2	94.2	94.2	97.9	97.9	97.7	99.5	99.5	99.4