A Look at CPS Non-Response and Trends in Poverty

Charles Hokayem[1], James P. Ziliak[2], Christopher R. Bollinger[3]

[1]Social, Economic, and Housing Statistics Division, U.S. Census Bureau, Washington, D.C. 20233-8500
[2]Department of Economics and Center for Poverty Research, University of Kentucky, Lexington, KY 40506-0034
[3]Department of Economics, University of Kentucky, Lexington, KY 40506-0034

Abstract
The Current Population Survey Annual Social and Economic Supplement (CPS ASEC) serves as the data source for official income and poverty statistics in the United States. There is a concern that the rise in non-response to earnings questions could deteriorate data quality and distort estimates of income and poverty. The CPS ASEC relies on a hot deck imputation procedure to address non-response. This paper assesses the extent of the bias in poverty rates caused by earnings non-response and the hot deck procedure. We use a dataset of matched CPS ASEC records to Social Security Detailed Earnings Records (DER) to study the impact of earnings non-response on estimates of poverty over the time period 1997-2008. Initial results show substituting DER earnings data for earnings imputed in the CPS ASEC produces poverty rates that are higher than the official poverty rate but not as high as poverty rates produced from completely dropping imputed earners.

Key Words: CPS ASEC, poverty measurement, hot deck imputation, non-response bias, earnings, measurement error

1. Introduction

The accurate measurement of the income distribution and poverty statistics is vital to assessing economic growth, characterizing income inequality, and gauging the effectiveness of the federal safety net. The Current Population Survey Annual Social and Economic Supplement (CPS ASEC) serves as the official source of income and poverty statistics for the United States. CPS ASEC respondents may be reluctant to answer income questions out of concern for response confidentiality, or they may just have insufficient knowledge of the answer (Groves 2001). One method of addressing non-response simply deletes missing data and uses sampling weights to calculate population statistics of interest (Ziliak 2006). An alternative method to address non-response fills in missing data using a matching procedure that relies on matching observations with missing data to observations with complete data based on socioeconomic characteristics (Little and Rubin 2002). This second procedure, referred to as a cell "hot deck," offers the advantage of retaining more observations in the final data set than simply deleting any observation with missing data; however, the hot deck procedure may bias estimates of population statistics. Hirsch and Shumacher (2004) and Bollinger and Hirsch (2006) study the hot deck procedure in a related survey, the CPS Outgoing Rotation Group, and show the hot deck procedure causes earnings regression parameters to be biased. Given the bias in regression parameters there is a possibility the hot deck procedure could bias estimates of statistics derived from income such as poverty rates.

This paper assesses the extent of the bias in poverty rates caused by earnings non-response and the hot deck procedure. We assemble a dataset of matched CPS ASEC records to Social Security Detailed Earnings Records (DER) to study the impact of earnings non-response on estimates of poverty. The CPS ASEC-DER matched data file covers CPS ASEC years 1998-2009, allowing for the systematic study of long term trends in income imputation and poverty rates. We present an initial analysis on the bias in poverty rates by substituting DER earnings for CPS earnings. Our analysis recalculates the official poverty rate based on different assumptions on the availability of DER earnings and imputation status of survey respondents.

2. Literature Review

Several papers examine the effect of measurement error and income imputation on poverty. Chesher and Schluter (2002) provide a theoretical treatment of measurement error on various measures of welfare. Their derivations allow a study of the sensitivity of income inequality and poverty measures to the amount of measurement error in the income distribution. The amount of measurement error is characterized by the degree of measurement error variance. Their simulations comparing income distributions with and without measurement error show measurement error can upwardly bias poverty rates and Gini coefficients. Poverty rates measured in surveys may overstate poverty. Chesher and Schluter apply their method to measuring the degree of this bias to regional poverty in Indonesia.

Nicholas and Wiseman (2009) merge administrative data from the Social Security Administration (SSA) with the CPS ASEC 2003 to study poverty among the entire U.S. population and among the elderly for calendar year 2002. Their analysis uses several SSA files for earnings, Old-Age, Survivors, and Disability Insurance (OASDI) payments (social security), and SSI payments. Wage and salary earnings come from Summary Earnings Record (SER) and Detailed Earnings Record (DER) files; Social Security benefits come from the Payment History Update System (PHUS) file; and SSI payments come from the Supplemental Security Record file. Using administrative records for SSI payments corrects for underreporting of this benefit in the CPS ASEC. Their analysis substitutes administrative earnings for CPS earnings and self-employment income when available, leaving all other sources of income from the CPS. Nicholas and Wiseman develop measures of income that vary on the availability of administrative and CPS data and employ a reweighting adjustment for CPS observations unmatched to the administrative data. Their results confirm that the CPS substantially understates SSI receipt. They find that using administrative data reduces official poverty rates for the entire national population and for the SSI recipient population. The poverty rate for the entire U.S. population falls from 12.1 percent to between 9.3 percent and 11.8 percent while the SSI poverty rate falls from 44.3 percent to between 39.0 and 40.9 percent. Using a relative measure of poverty, half of equivalence-adjusted median income, has a smaller effect on poverty rates.

Like Nicholas and Wiseman (2009) Turek et al. (2012) use administrative data from the Social Security Administration to study poverty with a focus on the effects of income imputation in the CPS on poverty. Turek et al. merge earnings information from the Detailed Earnings Record file with the CPS ASEC 2006 (calendar year 2005) to examine the effect of substituting DER earnings for reported CPS earnings on income estimates and number of persons in poverty. Their analysis separates individuals by CPS imputation status: no imputes, item imputes, and whole imputes. Item imputes are individuals who respond to the CPS ASEC supplement but need specific income questions imputed. Whole imputes are individuals who refuse to respond to the CPS ASEC supplement and need the entire supplement, including all income questions, imputed. After substituting DER earnings for CPS earnings, an overwhelming majority of individuals do not change poverty status. The poverty status for 93.7 percent of all individuals does not change. This result holds by all three imputation types: no imputes (94.4 percent), item imputes (92.8 percent), and whole imputes (89.2 percent).

This paper differs from the previous literature using administrative records for poverty measurement in several ways. First, the analysis assembles a matched data set covering a long time period, 1997-2008. Second, the analysis examines trends in non-response and imputation and their impact on poverty rates. Third, while previous analyses study different components of income, this analysis focuses on just earnings imputation since earnings account for over 80 percent of income.

3. The Current Population Survey Hot Deck Procedure

The Census Bureau has used a hot deck procedure for imputing missing income since 1962. The current system has been in place with few changes since 1989 (Welniak 1990). The CPS ASEC uses a variation of the cell hot deck procedure to impute missing income and earnings data. The cell hot deck procedure assigns individuals with missing income values that come from individuals with similar characteristics. The hot deck procedure for the CPS ASEC earnings variables relies on a sequential match procedure. First, individuals with missing data are divided into one of 12 allocation groups defined by the pattern of non-response. Welniak (1990) lists the 12 allocation groups and non-response patterns. Examples include a group that is only missing earnings from longest job or a group that is missing both longest job and earnings from longest job. Second, an observation in each allocation group is matched to another observation with complete data based on a large set of socioeconomic variables, the match variables.[1] If no match is found based on the large set of match variables, then a match variable is dropped and variable definitions are collapsed to be less restrictive. This process of sequentially dropping a variable and collapsing variable definitions is repeated until a match is found. When a match is found, the missing income amount is substituted with the reported income amount from the first available matched record. The missing income amount does not come from an average of the available matched records.

For example, suppose the set of match variables consists of gender, race, education, age, and region where education is defined by less than high school, high school, some college, and college or more. If no match is found using this set of match variables, then

[1] The set of match variables includes gender, race, age, relationship to householder, years of school completed, marital status, presence of children, labor force status of spouse, weeks worked, hours worked, occupation, class of worker, other earnings receipt, type of residence, region, transfer payments receipt, and person status.

the race variable could be dropped and education could be redefined by collapsing education categories to high school or less, some college, and college or more. If no match exists, then region could be dropped to obtain a match. This process of dropping and redefining match variables continues until the only match variable remaining is gender. This sequential match procedure always ensures a match.

4. Data

The data used for the analysis come from the Current Population Survey Annual Social and Economic Supplement (CPS ASEC) for survey years 1998-2009 (reporting income for 1997-2008). The CPS ASEC is matched to the Social Security Administration's Detailed Earnings Record (DER) file. The Detailed Earnings Record file is an extract of Social Security Administration's Master Earning File (MEF) and includes data on total earnings, including wages and salaries and income from self-employment subject to Federal Insurance Contributions Act (FICA) and/or Self-Employment Contributions Act (SECA) taxation. Since individuals do not make SECA contributions if they lose money in self-employment, only positive self-employment earnings are reported in the DER file (Nicholas and Wiseman 2009). The DER file contains all earnings reported on a worker's W-2 forms. These earnings are not capped at the FICA contribution amounts and include earnings not covered by Old Age Survivor's Disability Insurance (OASDI) but subject to Medicare tax. The DER file also contains deferred wages such as contributions to 401(k), 403(b), 408(k), 457(b), 501(c), and HSA plans. The DER file is not a comprehensive source of gross compensation. Abowd and Stinson (2011) describe parts of gross compensation that may not appear in the DER file such as pretax health insurance premiums and education benefits. Workers in the DER file are uniquely identified by a Protected Identification Key (PIK) assigned by the Census Bureau. The PIK is a confidentiality-protected version of the Social Security Number.

The Census Bureau's Center for Administrative Records Research and Applications (CARRA) matches the DER file to the CPS ASEC. Since the CPS does not currently ask respondents for a Social Security Number, CARRA uses its own record linkage software system, the Person Validation System, to assign a Social Security Number.[2] This assignment relies on a probabilistic matching model based on name, address, date of birth, and gender (NORC 2011). The Social Security Number is then converted to a Protected Identification Key. The Social Security Number from the DER file received from SSA is also converted to a Protected Identification Key. The CPS ASEC and DER files are matched based on the Protected Identification Key and do not contain the Social Security Number.

5. Analysis

A worker can appear multiple times per year in the DER file if they have several jobs. The DER file is collapsed into one earnings observation per worker per year by aggregating total compensation (Box 1 of W-2), SSA covered self-employment earnings (SEI-FICA), and Medicare covered self-employment earnings (SEI-MEDICARE) across all employers. DER earnings are defined as the sum of total compensation plus the

[2] The final year the CPS collected respondent Social Security Number is CPS survey year 2005 (calendar year 2004). Beginning with survey year 2006 (calendar year 2005), all respondents were assigned a Social Security Number using the Person Validation System.

maximum of SSA covered self-employment income or Medicare covered self-employment:

DER Earnings = (Box 1 of W-2) + max(SEI-FICA,SEI-MEDICARE)

In this way DER Earnings is most compatible with the CPS earnings. CPS earnings (PEARNVAL) cover earnings from all wage and salary jobs (WSAL-VAL), business self-employment (SEMP-VAL), and farm self-employment (FRSE-VAL). The CPS total personal income variable (PTOTVAL) used to determine poverty consists of adding a person's total earnings (PEARNVAL) to a person's total other income (POTHVAL):

PTOTVAL=PEARNVAL+POTHVAL

The analysis replaces the portion of total personal income due to earnings (PEARNVAL) with DER Earnings while keeping income from other sources the same (no change in POTHVAL). Replacing earnings income differs by imputation status and the availability of DER Earnings. An individual's imputation status is determined by having either wages and salary from longest job imputed (I-ERNVAL) or wages and salary from other jobs imputed (I-WSVAL). Poverty is recalculated based on two types of substitutions:

Method 1: Replace CPS Earnings (PEARNVAL) with DER Earnings for all persons with a DER match regardless of imputation status and use CPS Earnings for persons without a DER match

Method 2: Replace CPS Earnings (PERNVAL) with DER Earnings for ONLY those persons with imputed earnings and a DER match and use CPS Earnings for persons without a DER match. Use imputed earnings for persons with no CPS Earnings and no DER match.

6. Results

Table 1 shows the results of matching CPS ASEC and DER files for CPS survey years 1998-2009.[3] The table displays the person count based on the CPS ASEC person file, the number of earners, the number of matched records, and the match rate. The match rate is defined as the number of earners matched to a DER record divided by the total number of earners. The match rates range from 66 percent to 85 percent. The table also shows the imputation rate among earners. The rate of imputed earnings begins at 16 percent for 1998, rises to 21 percent for 2003-2005, and falls to 19 percent for 2009. The remainder of the table shows how match rates differ by imputation status among earners. Individuals with no imputed earnings are more likely to have a matched DER record. All counts and rates are unweighted. Figure 1 plots the overall match rate for earners and the match rate for earners by imputation status.

Table 2 shows the effect of imputation and replacing CPS earnings with DER Earnings on the official poverty rate. Poverty rates are weighted using the March supplement person weight. Column 1 shows the official poverty rate over the time period while column 3 shows the official poverty rate after dropping individuals with imputed earnings. This comparison gives a sense of the bias introduced by the imputation

[3] The matched data for CPS survey year 2001 do not include the SCHIP sample expansion. Matched data for survey years after 2001 include the SCHIP sample expansion.

process. Columns 5 and 6 give the difference from the official poverty rate and a test for equality to the official poverty rate at the 10 percent level of significance.[4] Excluding imputed earners from the poverty calculation raises the poverty rate across all years by an average of 0.7 percentage points. Column 7 shows the poverty rate after replacing CPS earnings with DER earnings for all persons regardless of imputation status (Method 1). Comparing this poverty series to the official poverty series for all years excluding 1999-2000 and 2008 still raises the rate but by a smaller amount (average of 0.3 percentage points).

Column 11 shows the poverty rate after replacing CPS earnings with DER earnings for only those persons with imputed earnings (Method 2). Again, the poverty series is higher than the official poverty series, but only for 2001-2008, by an average of 0.3 percentage points. The earlier years, 1997-2000, are not statistically different at the 10 percent level of significance. Figure 2 plots each series and shows the effects of dropping imputed earners and replacing CPS earnings with DER earnings by each Method. Figure 3 plots the difference in each series from the official poverty rate.

7. Conclusion and Future Work

This paper uses a unique dataset of administrative earnings data matched to the CPS ASEC to study the effects of earnings imputation on poverty measurement. Initial results show substituting earnings data for earnings imputed in the CPS ASEC produces poverty rates that are higher than the official poverty rate but not as high as poverty rates produced from completely dropping imputed earners. Future work will further assess data quality by comparing the income responses in the CPS ASEC to the administrative earnings records to provide a sense of how closely aligned the sources are. The analysis so far has focused on the entire population. Future work will also provide estimates of poverty by various demographic groups. A technical issue to address is adjusting poverty rates for unmatched DER records. Table 1 shows not all CPS records are matched to a DER record and match rates vary by imputation status, suggesting the matched sample may differ from the sample used for the official poverty rate. This matched sample should be adjusted for the probability of selection. In this context, the adjustment can be accomplished by multiplying the survey weights by the inverse probability of being matched. With the new weights, unmatched CPS observations can be dropped from the analysis.

Acknowledgements

We would like to thank Martha Stinson, Graton Gathright, and Gary Benedetto for their help in understanding the Detailed Earnings Record files. We also thank Trudi Renwick, Chuck Nelson, Ed Welniak, Adam Bee, and session participants at the 2012 Joint Statistical Meetings for helpful comments. The views expressed in this research, including those related to statistical, methodological, technical, or operational issues, are solely those of the authors and do not necessarily reflect the official positions or policies of the Census Bureau. The authors accept responsibility for all errors. This paper is released to inform interested parties of ongoing research and to encourage discussion of

[4] All comparative statements in this paper have undergone statistical testing, and, unless otherwise noted, all comparisons are statistically significant at the 10 percent significance level.

work in progress. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone more limited review than official publications.

References

Abowd, John and Martha Stinson. 2011. "Estimating Measurement Error in SIPP Annual Job Earnings: A Comparison of Census Bureau Survey and SSA Administrative Data." Working Paper, May 2011.

Bollinger, Christopher and Barry Hirsch. 2006. "Match Bias from Earnings Imputation in the Current Population Survey: The Case of Imperfect Matching." Journal of Labor Economics, 24(3), pp. 483-519.

Chesher, Andrew and Christian Schluter. 2002. "Welfare Measurement and Measurement Error." Review of Economic Studies, 69, pp. 357-378.

Groves, Robert. 2001. Survey Nonresponse. New York: Wiley.

Hirsch, Barry and Edward Schumacher. 2004. "Match Bias in Wage Gap Estimates Due to Earnings Imputation." Journal of Labor Economics, 22(3), pp. 689-722.

Little, Roderick and Donald Rubin. 2002. Statistical Analysis With Missing Data. New York: Wiley.

Nicholas, Joyce and Michael Wiseman. 2009. "Elderly Poverty and Supplemental Security Income" Social Security Bulletin, 69(1), pp. 45-73.

NORC at the University of Chicago. 2011. "Assessment of the US Census Bureau's Person Identification Validation System." Final Report presented to the US Census Bureau.

Turek, Joan, Kendall Swenson, Bula Ghose, Fritz Scheuren, and Daniel Lee. 2012. "How Good Are ASEC Earnings Data? A Comparison to SSA Detailed Earnings Records" Presented at 2012 Federal Committee On Research Methodology Research Conference.

Welniak, Edward J. 1990. "Effects of the March Current Population Survey's New Processing System on Estimates of Income and Poverty." US Census Bureau, Washington, DC, 1990.

Ziliak, James P. 2006. "Understanding Poverty Rates and Gaps: Concepts, Trends, and Challenges." Foundations and Trends in Microeconomics 1: 127–199.

Table 1: Match Rate and Imputation Rate

Calandar Year	Person Record Count	Earners	Matched Records	Match Rate (Earners)	Earnings Imputation Rate	IF IMPUTED and EARNER: Matched If Imputed	Total Imputed	Match Rate	IF NOT IMPUTED and EARNER: Matched If Not imputed	Total Not Imputed	Match Rate
1997	131,617	69,573	53,005	71%	16%	6,111	11,329	54%	43,193	58,244	74%
1998	132,324	70,218	49,474	66%	18%	5,873	12,363	48%	40,173	57,855	69%
1999	133,710	71,783	50,661	66%	17%	6,079	12,492	49%	41,071	59,291	69%
2000	218,269	69,040	51,311	68%	20%	7,255	13,771	53%	39,916	55,269	72%
2001	217,219	113,577	89,543	73%	20%	12,983	22,534	58%	69,388	91,043	76%
2002	216,424	111,698	84,692	70%	21%	12,510	23,097	54%	65,235	88,601	74%
2003	213,241	109,672	74,585	62%	21%	10,340	22,649	46%	58,111	87,023	67%
2004	210,648	108,120	71,632	61%	21%	10,057	22,296	45%	55,812	85,824	65%
2005	208,562	107,532	100,013	85%	19%	15,631	20,016	78%	75,786	87,516	87%
2006	206,639	106,738	99,633	85%	20%	16,145	20,853	77%	74,524	85,885	87%
2007	206,404	107,038	99,217	84%	20%	16,201	21,174	77%	74,100	85,864	86%
2008	207,921	107,134	98,764	84%	19%	15,086	20,014	75%	74,981	87,120	86%

Sources: U.S. Census Bureau, Current Population Survey, 1998-2009 Annual Social and Economic Supplement. For information on sampling and nonsampling error, see <www.census.gov/apsd/techdoc/cps/cpsmar10.pdf>.
Social Security Administration, Detailed Earnings Record, 1997-2008.

Table 2: Poverty Rates Based on Differing Assumptions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Calendar Year	Official Poverty Rate(%)	Std. Error	Poverty Rate After Dropping Imputed Earners (%)	Std. Error	Difference from Official Rate	Test for Equality	Poverty Rate Using DER Earnings for all Persons (%) (Method 1)	Std. Error	Difference from Official Rate	Test for Equality	Poverty Rate Using DER Earnings for Persons with Imputed Earnings (%) (Method 2)	Std. Error	Difference from Official Rate	Test for Equality
1997	13.3	0.211	13.8	0.225	0.6	*	13.4	0.212	0.2	*	13.4	0.212	0.1	
1998	12.7	0.206	13.3	0.222	0.6	*	12.9	0.208	0.2	*	12.9	0.207	0.2	
1999	11.9	0.199	12.4	0.213	0.5	*	11.9	0.199	0.1		11.9	0.199	0.1	
2000	11.3	0.193	12.0	0.211	0.7	*	11.3	0.194	0.0		11.4	0.194	0.1	
2001	11.7	0.139	12.5	0.152	0.8	*	12.1	0.141	0.4	*	12.0	0.141	0.3	*
2002	12.1	0.140	12.9	0.154	0.8	*	12.4	0.142	0.3	*	12.4	0.142	0.3	*
2003	12.5	0.142	13.3	0.154	0.8	*	12.7	0.143	0.3	*	12.7	0.143	0.2	*
2004	12.7	0.142	13.6	0.155	0.8	*	13.0	0.143	0.3	*	12.9	0.143	0.2	*
2005	12.6	0.141	13.3	0.152	0.7	*	13.0	0.143	0.4	*	12.9	0.142	0.3	*
2006	12.3	0.139	13.1	0.151	0.8	*	12.7	0.141	0.4	*	12.7	0.141	0.5	*
2007	12.5	0.139	13.3	0.152	0.9	*	13.0	0.141	0.5	*	13.0	0.141	0.5	*
2008	13.2	0.142	14.0	0.154	0.8	*	13.2	0.142	0.0		13.6	0.143	0.3	*

Standard errors are estimated using generalized variance parameters. *p<0.10
Sources: U.S. Census Bureau, Current Population Survey, 1998-2009 Annual Social and Economic Supplement. For information on sampling and nonsampling error, see <www.census.gov/apsd/techdoc/cps/cpsmar10.pdf>.
Social Security Administration, Detailed Earnings Record, 1997-2008.

Figure 1: Match Rate By Imputation Status

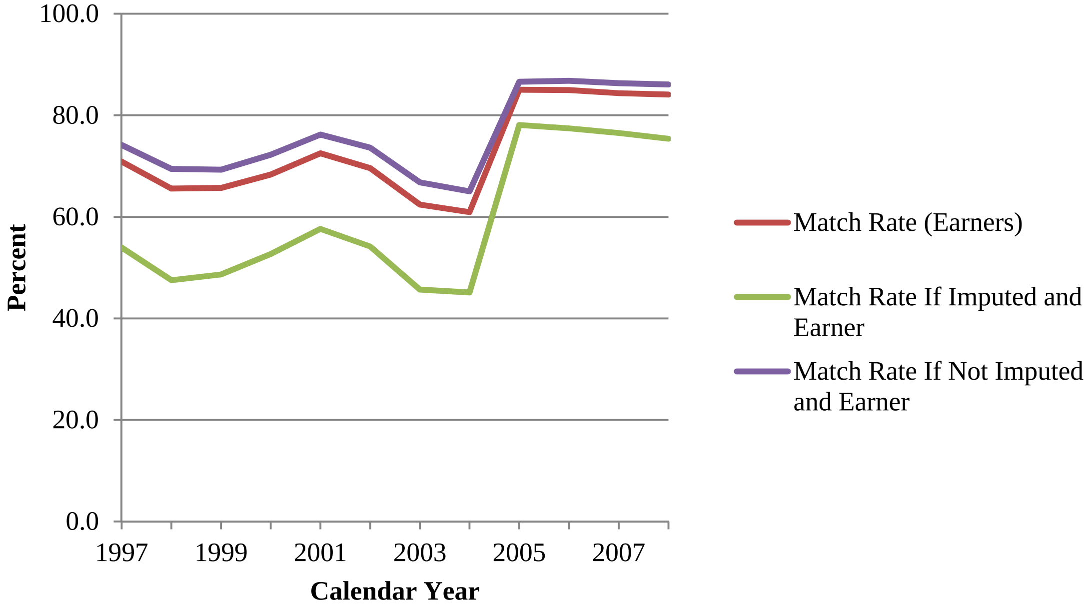

Figure 2: Poverty Rate Based on Differing Assumptions

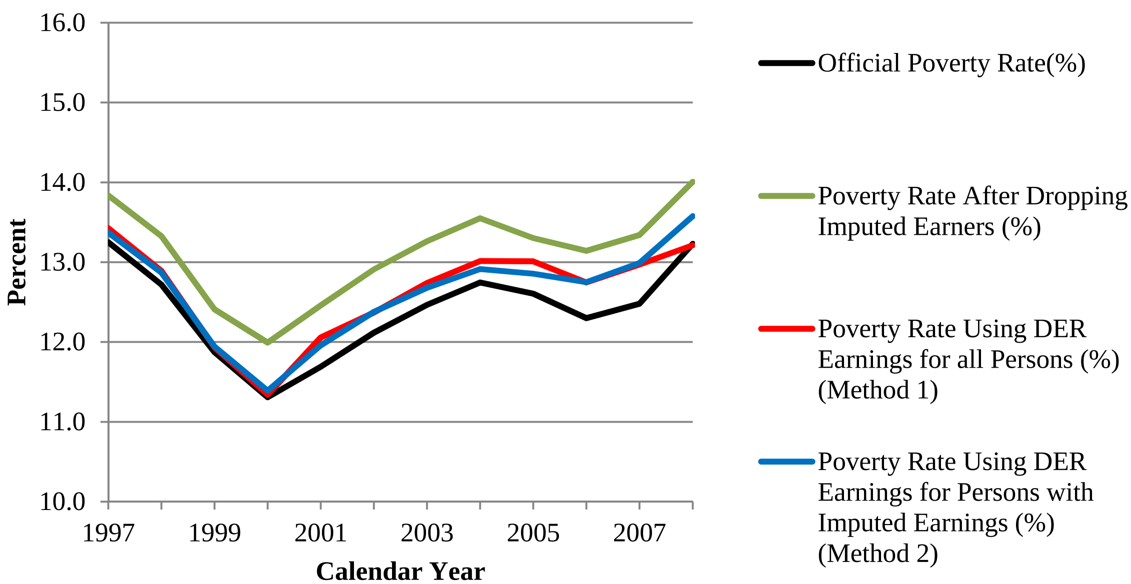

Figure 3: Difference from Official Poverty Rate Based on Differing Assumptions