

## **Estimating Missing Values from the General Social Survey: An Application of Multiple Imputation**

David A. Penn\*

*Middle Tennessee State University, Murfreesboro, TN*

### *Abstract*

*Objectives.* Most researchers who use survey data must grapple with the problem of how best to handle missing information. This article illustrates multiple imputation, a technique for estimating missing values in a multivariate setting. *Methods.* I use multiple imputation to estimate missing income data and update a recent study that examines the influence of parents' standard of living on subjective well-being. Using data from the 1998 General Social Survey, two ordered probit models are estimated; one using complete cases only, and the other replacing missing income data with multiple imputation estimates. *Results.* The analysis produces two major findings: 1) parents' standard of living is more important than suggested by the complete cases model, and 2) using multiple imputation can help to reduce standard errors. *Conclusions.* Multiple imputation allows a researcher to use more of the available data, thereby reducing biases that may occur when observations with missing data are simply deleted.

*Keywords:* subjective well-being, financial well-being, multiple imputation

*JEL Categories:* A10, C42

---

\* Business and Economic Research Center, Associate Professor, Department of Economics and Finance, Middle Tennessee State University, 1301 E. Main St., Box 102, Murfreesboro, TN 37132, email: [dpenn@mtsu.edu](mailto:dpenn@mtsu.edu). SAS coding and data will be shared on request. The paper benefits from the insightful comments and suggestions offered by the reviewers.

## **I. Introduction**

A growing literature suggests that how researchers deal with missing data can affect model estimates and standard errors (Schafer, 1997; Vriens and Melton, 2002; Schafer and Graham, 2002; Raghunathan, 2004; Little and Rubin, 2002). Unfortunately, few studies explicitly discuss this issue. Frequently, researchers delete observations with missing items with no attention to the consequences for model estimates and standard errors. Excluding observations with missing data is hazardous if the deleted observations are different from the observations that remain. To illustrate this point, Table 1 shows the distribution of the variables used in this study, comparing complete observations (no items are missing) with observations for which income is missing. The figures are compiled from the 1998 General Social Survey (GSS). A glance at Table 1 shows that the missing income matters for some variables but not others. For example, the distribution of age is quite different: those who did not respond to the income question (call these the incomplete observations) are substantially older than those who did. In fact, 33.2 percent of respondents with missing income are 65 years old or older, compared with just 17.4 percent of respondents with reported income. Other large differences exist in educational level (those with missing income have less education) and health status (missing income is related to poorer health). Also, respondents with missing income are less likely to be married and more likely to be female.

(Insert Table 1 here)

Of the 1,747 observations pulled from the 1998 GSS, income is missing in 11.7 percent, primarily because of refusals. Since respondents 65 years or older are much more likely to refuse to respond to the income question, deleting observations with missing income eliminates a disproportionate share of the elderly, shifting the age distribution towards younger respondents. The potential for bias introduced by deleting incomplete observations depends on whether, after conditioning on income, health, and other variables, the elderly have different opinions compared with other age groups.

Using data from the GSS, this study illustrates the use of multiple imputation (MI) as a method for estimating missing information. Unlike single-imputation approaches, MI generates several estimates for each missing value, using the variation between estimates as a measure of the uncertainty associated with imputation. MI has received

considerable attention in the statistics and public health literature and is currently receiving attention in the social sciences. MI has been shown to outperform other methods of estimating missing values in simulation studies (Vriens and Melton, 2002), and is generally considered an improvement compared with single imputation and *ad hoc* imputation methods. Models estimated from singly imputed data underestimate standard errors, since the imputed data is treated as if it is known with certainty. By contrast, MI estimates the variation associated with the missing values and incorporates this variation in the model estimates. Accounting for the variation in the missing values constitutes an improvement over single imputation and *ad hoc* methods of missing value estimation.

## **II. Happiness and Subjective Well-Being**

Economists are giving more attention to the study of happiness; Di Tella and MacCulloch (2006), Frey and Stutzer (2002), and Kahneman and Krueger (2006) provide appraisals of recent research in this area. A puzzle that has received substantial attention in the literature is that over time, real median personal income has increased but self-reported happiness has not. As economists tend to associate rising real income with rising welfare, the lack of growth of happiness is problematic. Easterlin (2003) offers a possible explanation: an individual's welfare may depend more on relative income than absolute income. That is, keeping pace with the income of one's family or one's peers is a more important determinant of happiness than are absolute gains in income.

McBride (2001) explicitly tests this relative income notion using data from the General Social Survey (GSS). Specifically, he shows how subjective well-being, measured by the respondent's self-reported level of happiness, depends in part on the respondents' evaluation of his standard of living compared with his parents' standard of living. Respondents who enjoy a standard of living at or above their parents' tend to be more happy, and those with a standard of living worse than their parents are less happy. Of particular interest in this study, McBride eliminates observations from the sample for which values are missing for several variables including education, health status, marital status, income, and parents' standard of living. His analysis was completed with 324 total observations after the eliminations, reducing the sample size by several dozen.

The issue is this: discarding incomplete observations may produce biased regression estimates depending on the characteristics of the deleted data. If the removed observations are a random sub-sample of the data, then deleting the observations will not bias the estimates. If, however, the deleted observations are substantially different from the remaining observations, then bias is much more likely, particularly if the number of deleted observations is relatively large. Thus, regression coefficients that estimate how happiness depends on parents' standard of living could be higher (or lower) depending on the characteristics of the deleted observations.

Does eliminating incomplete observations bias the regression coefficients? This study tests this notion. Using multiple imputation (MI) to estimate missing values, I estimate a simplified version of McBride's model using data from the 1998 General Social Survey. The object of this study is not to precisely replicate McBride's work, but rather to illustrate how missing values can affect regression results. To accomplish this task, I estimate ordered probit models with and without imputed values and compare the results. If the results are substantially different, then deleting the observations may have biased the coefficients.

The remaining sections of the paper are as follows. The next two sections offer an introduction to multiple imputation and discuss the MI software used in this study. Probit regressions are estimated next using the imputed data, and results compared with a regression using complete cases only. The final section offers conclusions.

### **III. Introduction to Multiple Imputation**

Suppose a researcher wishes to use survey data to estimate how consumer spending depends on household income, but finds that responses to the income question are missing for a significant number of observations, probably because of refusals. The researcher could simply discard the observations with missing income and estimate the regression using the complete data only. Or the researcher could use one of many available imputation methods to estimate the missing income data and proceed with the analysis. Choosing between the two approaches depends on the characteristics of the missing income: are the data missing completely at random (MCAR), with the probability of missing information the same for all observations? Or do some types of

households tend to have missing income more often than others; elderly householders, for example, typically refuse to answer the income question more often than younger householders. If the data are missing completely at random, regression coefficients will be unbiased. If, on the other hand, the income data are not MCAR, then complete cases analysis will produce biased regression coefficients, since households included in the analysis could differ in a systematic fashion from those who have been excluded because of missing information.

### *The missing at random assumption*

Data missing completely at random is a special case of a more general category of missingness called *missing at random* (MAR). Data are missing at random if the probability that the variable  $Y$  is missing is not related to the value of  $Y$  itself, after controlling for all other variables in the analysis. To illustrate, suppose our data consist of two variables, educational status and age, with some values of age missing. The MAR assumption holds if the probability that age is missing, given the educational level, does not change as age changes.

The advantage of the MAR assumption is that it allows the analyst to estimate missing values without explicitly modeling the probability that an item is missing. The disadvantage is that for practical purposes the assumption is not testable unless the missing values can somehow be obtained by the researcher. Thus, when data are missing, the researcher may never be certain that the MAR assumption holds. Fortunately, simulation studies show that MI procedures tend to perform well even when the MAR assumption is weak (Shafer and Graham 2002).

The MCAR assumption is more restrictive and harder to satisfy than MAR, but if the researcher is satisfied that the data are indeed MCAR, complete case analysis is appropriate. On the other hand, if the missing data are not a random sample of all the data (MCAR), then analyzing just the complete cases could produce biased estimates. In this case, the researcher should explore methods of estimating the missing values.

Although several *ad hoc* methods are available, the current state of knowledge in statistics points to multiple imputation as the preferred method.<sup>1</sup>

#### **IV. A Brief Description of Multiple Imputation**

In multiple imputation, the researcher generates more than one estimate for each missing value. To illustrate, suppose the researcher's dataset consists of 1,750 observations including 1,500 complete observations and 250 observations with information missing for one or more variables. Multiple imputation proceeds in three steps. First, the researcher estimates multiple (six, for example) separate and distinct values for each missing item in the 250 incomplete observations; the estimated data are then substituted for the missing data to form six complete datasets. Second, regressions or other statistical analyses are performed using each of the six datasets, resulting in six sets of results. Third, the six sets of results are combined into one set of regression coefficients and standard errors. Combining regression coefficients is straightforward; the combined estimate is simply the mean from the six regressions. Combining variances is more involved because now variance includes not only variation occurring *within* each dataset but also variation *between* datasets attributable to the imputations. Little and Rubin (2002) show that combined variance is a weighted sum of the average within-imputation variance and between-imputation variance. The final result is a single set of regression coefficients and standard errors for use by the researcher. All this can be accomplished with currently available software implemented in SAS, SPlus, STATA, IVEware, MICE, and other software packages.

#### **V. Data**

Data for this study are from the General Social Survey (GSS) administered by the National Opinion Research Center at the University of Chicago.<sup>2</sup> The GSS is a personal interview of randomly selected adults in the United States. Nine of the variables downloaded from the 1998 GSS are those used by McBride (2001): happiness, parents'

---

<sup>1</sup> Schafer and Graham (2002) present an overview of the alternative methods available to researchers for estimating missing values.

<sup>2</sup> Data can be downloaded online at <http://webapp.icpsr.umich.edu/GSS/>.

standard of living, educational level, age, marital status, gender, health status, race, and family income. In addition, I downloaded three other variables that may be useful for the purpose of estimating missing values for income, parents' standard of living, and the other items; these are the respondent's working status, occupation (recoded to nine categories), and educational level. Finally, I recoded the income variable from twenty-three categories to four: Less than \$20,000, \$20,000 to \$34,999, \$35,000 to \$74,999, and \$75,000 or more. Categories for variables that will be used in the probit models, along with their means and standard deviations, are presented in Table 2.

(Insert Table 2)

Among the 1,747 total observations downloaded from the GSS<sup>3</sup>, income is missing for 205, parents' standard of living has 26 missing values, 21 are missing for the happiness variable, and smaller numbers of observations are missing for health status and age. At least one item is missing for 13.9 percent of the sample.

## **VI. Implementing Multiple Imputation Using IVEware**

I use IVEware to produce the multiple imputations, a freely available software developed by the Survey Research Center at the University of Michigan. The IVEware imputation model is a regression-based approach; two recent examples of regression-based multiple imputation models include the estimation of missing income in the Consumer Expenditure Survey<sup>4</sup> and a recent study that uses IVEware to multiply impute income in the National Health Interview Survey (Schenker et al: 2006).

IVEware has three important features. First, missing values are estimated using a distributional form appropriate for the variable. To estimate missing values for a continuous variable, for example, a linear regression model is used; when the variable is binary, a logistic model is used, and multi-valued categorical variables are modeled using a generalized logistic model. Second, the IVEware missing data model imputes values using both the fully observed observations and the observations for which values have been imputed, thus putting to use more of the available information. Using a sequential iterative method described below, IVEware merges imputed values back into the

---

<sup>3</sup> The sample used includes all respondents 25 years or older who were asked the parents' standard of living question, a total of 1,747.

<sup>4</sup> A brief explanation can be found at <http://www.bls.gov/cex/csximpute.htm>.

predictor data, making them available to predict missing values for other variables. Third, IVEware, like other MI software, can estimate missing values for several variables simultaneously. In this study, for example, IVEware simultaneously estimates missing values for income, age, parents' standard of living, education, health, and happiness.

The IVEware estimation process proceeds iteratively through multiple rounds. The process begins in round one with the variable that has the smallest amount of missing information. Using a distributional form appropriate to the variable, IVEware uses step-wise regression to generate estimates using just the complete cases as the predictor data. Multiple versions of the missing values are generated by adding residual variance to the model predictions. After filling in the missing values with the imputed values, the variable is merged into the predictor data and the process is repeated now with the variable with the next smallest amount of missing data, and so on until all the variables with missing information have been processed.

Beginning with round two, the regressions are repeated except that now each regression is estimated using *all* the variables: those with no missing data and those with missing values estimated in round one. In additional rounds, the missing value estimates update the predictor data, regressions are re-estimated, and so on, iteratively estimating the regression coefficients and updating the predictor data. The process continues for a predetermined number of rounds, or until the imputations have converged to a stable distribution (Raghunathan et al, 2001).

I chose eleven variables for processing by IVEware: the eight variables that will be used in the probit model (including the dependent variable 'happiness'), plus three variables that are likely to be important predictors of family income: the respondent's educational level, working status, and occupation. In fact, the step-wise regressions selected just these three variables to estimate missing values for income.

## **VII. Estimating the Model of Subjective Well-Being**

Next, I estimate two ordered probit regressions, one using complete cases only and the other a combined estimate from the six sets of imputations. Results from the six regressions are combined into one estimate as explained previously.



The dependent variable is the ‘happiness’ question from the GSS. In this question, the respondent is asked whether he is ‘very happy’, ‘pretty happy’, or ‘not too happy.’ I recode these responses so that ‘very happy’ is assigned a 2, ‘pretty happy’ is a 1, and ‘not too happy’ is a zero.

Dummy variables are coded for all the explanatory variables with one category of income, health, parents’ standard of living, and age left out of the regression. Following McBride, respondents 25 years and younger are excluded. A critical variable is the relative income norm, measured by responses to the question, “compared to your parents when they were the age you are now, do you think your own standard of living now is: much better, somewhat better, about the same, somewhat worse, or much worse?” Parents’ standard of living is coded to four dummy variables, PARSOL1-PARSOL4, capturing responses from ‘much better’ to ‘worse’. Since ‘much worse’ only received 56 responses, this category is combined with ‘somewhat worse’ in a single category, ‘worse’. If the relative income hypothesis is correct, the parents’ standard of living variables will be significantly different from zero and the size of the coefficients will increase in absolute value moving from ‘better’ (PARSOL2) to ‘worse’ (PARSOL4).

## **VIII. Results**

Results for both the multiple imputation model and the complete cases model are presented in Table 3. The multiple imputation model uses all 1,747 observations, while the complete case model is limited to just those with no missing data (1,503 observations). Goodness-of-fit statistics are presented for both models. For both, we can reject the hypothesis that all the regression coefficients are equal to zero.

(Insert Table 3 here)

Reviewing the results, two differences are apparent. First, estimates for ‘Somewhat better than parents’ standard of living’ (PARSOL2) and ‘Same as parents’ standard of living’ (PARSOL3) are substantially larger in absolute value in the multiple imputation model compared with the complete cases model; ‘Same as parents’ standard of living’ (PARSOL3) is 58 percent larger, and ‘Somewhat better than parents’ standard of living’ (PARSOL2) is larger by 59 percent, nearly one standard error. More importantly, the combined effect of the larger MI coefficients and the slightly smaller

standard errors results in smaller p-values compared with the complete cases model. In fact, PARSOL2 and PARSOL3 move from not statistically significant in the complete cases model to significant at the 5 percent level in the multiple imputation model. Thus, based on just the complete cases model, one might be tempted to conclude that the influence of parents' standard of living on happiness is relatively weak, since only one of the three dummy variables has a statistically significant effect. However, when we use multiple imputation to estimate missing values, the results make a stronger case for the importance of the parents' standard of living in support of the relative-income hypothesis.

Other differences are also apparent, particularly regarding the standard errors. The standard errors are smaller in the multiple imputation model, although the absolute magnitude of the difference is not great. The multiple imputation standard errors are affected by two opposing influences. Including data that otherwise would have been eliminated increases the sample size relative to complete cases analysis, thus reducing the standard errors. However, the between-imputation variation adds to the standard errors. In the present case, the balance of the two influences caused a modest decline in the standard errors.

## **IX. Conclusion**

This paper updates a recent study on subjective well-being, using multiple imputation to estimate missing data. MI provides a promising method of incorporating missing values in regression estimates, thereby reducing potential bias caused by deleting observations with missing information. The MI model shows that parents' standard of living is a more important predictor of subjective well-being compared with the complete cases model. Researchers using complete cases only would find weak support for the relative income hypothesis. By contrast, using more of the available data, the multiple imputation model produces stronger conclusions about the influence of parents' standard of living on subjective well-being. Although support for the relative income hypothesis can be found in both models, the evidence is stronger in the multiple imputation model.

Researchers wishing to use multiple imputation could begin by scanning a review article of the available multiple imputation software and techniques by Horton and Lipsitz (2001). Multiple imputation procedures are available in several software

packages including SOLAS, STATA, SAS, SPlus, MICE, IVEware, and Joseph Schafer's online shareware. My feeling on the matter is that IVEware is more flexible than SAS, but SAS is more straightforward when combining the MI regressions. Consequently, I used IVEware to generate the imputations and the SAS PROC MIANALYZE procedure to consolidate the MI regression results.

## References

- Di Tella, Rafael, and Robert MacCulloch. 2006. "Some Uses of Happiness Data in Economics" *Journal of Economic Perspectives* 20:25-46.
- Easterlin, Richard A. 2003. "Explaining Happiness." *Proceedings of the National Academy of Sciences of the United States of America*, September: 11176-11183.
- Frey, Bruno S., and Alois Stutzer. 2002. "What Can Economists Learn from Happiness Research?" *Journal of Economic Literature* 40:402-435.
- Horton, Nicholas J., and Stuart R. Lipsitz. 2001. "Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables." *The American Statistician*, August: 244-254. Article available online at [www.biostat.harvard.edu/~horton/tasimpute.pdf](http://www.biostat.harvard.edu/~horton/tasimpute.pdf).
- Kahneman, Daniel, and Alan B. Krueger. 2006. "Developments in the Measurement of Subjective Well-Being." *Journal of Economic Perspectives* 20:3-24.
- Little, Roderick J.A., and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*, Second Edition, New York: John Wiley & Sons, Inc.
- McBride, Michael. 2001. "Relative-income effects on subjective well-being in the cross-section." *Journal of Economic Behavior & Organization* 45:251-278.
- Ragunathan, Trivellore E. 2004. "What Do We Do with Missing Data? Some Options for Analysis of Incomplete Data." *Annual Review of Public Health* 25:99-117.
- Ragunathan, Trivellore E., James M. Lepkowski, John van Hoewyk, and Peter Solenberger. 2001. "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models." *Survey Methodology* 27:85-95.
- Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. New York: Chapman & Hall/CRC.
- Schafer, Joseph L., and John W. Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods*, 7:147-177.
- Schafer, Joseph L., and Maren K. Olsen. 1998. "Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective" *Multivariate Behavioral Research* 33: 545-571.
- Schenker, Nathaniel, Trivellore E. Ragunathan, Pei-Lu Chiu, Diane M. Makuc, Guangyu Zhang, and Alan J. Cohen. 2006. "Multiple Imputation of Family Income and Personal Earnings in the National Health Interview Survey: Methods and Examples."

National Center for Health Statistics and the University of Michigan. Available online at [www.cdc.gov/nchs/data/nhis/tecdoc.pdf](http://www.cdc.gov/nchs/data/nhis/tecdoc.pdf)

Vriens, Marco, and Melton, Eric. 2002. "Managing Missing Data." *Marketing Research* 14:12-17.

TABLE 1

Distribution of Complete Observations and Observations with Missing Income by  
Variable (percent by variable)

Variable	Complete Observations	Observations with Missing Income
<i>Age</i>		
Less than 36 years	31.4	23.9
26 to 46 years	25.8	13.2
47 to 64 years	25.2	29.3
65 years and older	17.4	33.2
Missing	0.1	0.5
<i>Standard of living compared with parents' standard of living</i>		
Much better	33.1	33.2
Same	30.8	31.2
Somewhat better	21.0	20.5
Worse	14.1	10.7
Missing	1.1	4.4
<i>Education</i>		
Less than high school	14.8	23.9
High school diploma	52.8	49.8
Associate degree	7.7	5.8
Bachelor's degree	16.9	15.1
Graduate degree	7.4	4.9
Missing	0.3	0.5
<i>Happiness</i>		
Very happy	31.8	40.0
Pretty happy	54.9	47.8
Not too happy	12.1	10.7
Missing	1.2	1.5
<i>Health status</i>		
Excellent	30.3	24.4
Good	48.6	49.3
Fair or poor	20.8	25.4
Missing	0.3	1.0
<i>Married</i>	50.0	44.9
<i>Female</i>	54.5	57.5
<i>Black</i>	15.0	13.2
Observations	1,542	205

Source: Compiled from the 1998 General Social Survey

TABLE 2

Means and Standard Deviations of Variables by Happiness (prior to imputations)

Variable	Description	Not too Happy		Pretty Happy		Very Happy		All	
INC1	Less than \$20,000	0.452	(0.500)	0.227	(0.419)	0.155	(0.363)	0.229	(0.420)
INC2	\$20,000-\$34,999	0.226	(0.419)	0.225	(0.418)	0.168	(0.374)	0.207	(0.405)
INC3	\$35,000-\$74,999	0.197	(0.399)	0.331	(0.471)	0.344	(0.475)	0.318	(0.466)
INC4	\$75,000 or more	0.024	(0.154)	0.113	(0.317)	0.190	(0.393)	0.129	(0.335)
PARSOL1	Much better	0.245	(0.431)	0.290	(0.454)	0.431	(0.496)	0.331	(0.471)
PARSOL2	Somewhat better	0.288	(0.454)	0.327	(0.469)	0.293	(0.456)	0.308	(0.462)
PARSOL3	Same	0.197	(0.399)	0.223	(0.416)	0.183	(0.387)	0.209	(0.407)
PARSOL4	Worse	0.245	(0.431)	0.148	(0.356)	0.078	(0.269)	0.136	(0.343)
HEALTH1	Excellent	0.154	(0.362)	0.246	(0.431)	0.434	(0.496)	0.296	(0.457)
HEALTH2	Good	0.385	(0.488)	0.536	(0.499)	0.443	(0.497)	0.487	(0.500)
HEALTH3	Fair or poor	0.457	(0.500)	0.216	(0.412)	0.117	(0.322)	0.213	(0.409)
AGE1	Less than 36 years	0.255	(0.437)	0.279	(0.449)	0.281	(0.450)	0.279	(0.449)
AGE2	36 to 46 years	0.308	(0.463)	0.278	(0.448)	0.248	(0.432)	0.270	(0.444)
AGE3	47 to 64 years	0.221	(0.416)	0.271	(0.445)	0.255	(0.436)	0.257	(0.437)
AGE4	65 years or more	0.216	(0.413)	0.172	(0.377)	0.216	(0.412)	0.193	(0.395)
MARRY	Married	0.226	(0.419)	0.452	(0.498)	0.665	(0.472)	0.494	(0.500)
FEMALE	Female	0.630	(0.484)	0.532	(0.499)	0.550	(0.498)	0.550	(0.498)
BLACK	Black	0.226	(0.419)	0.158	(0.365)	0.101	(0.302)	0.148	(0.355)
Obs.		209		944		573		1,747	

Note: "All" includes 21 observations for which Happiness is missing.

TABLE 3

Ordered Probit Results for the Complete Cases Model and the Multiple Imputation Model

Label	Variable	Complete cases			Multiple imputation		
		Coefficient	Std Error		Coefficient	Std Error	
	Intercept	-0.011	0.128	(0.930)	0.025	0.122	(0.836)
	Intercept2	1.852	0.055	(0.000)	1.813	0.050	(0.000)
INC1	Income less than \$20,000	-0.442	0.118	(0.000)	-0.354	0.114	(0.002)
INC2	Income \$20,000-\$34,999	-0.318	0.109	(0.004)	-0.230	0.102	(0.024)
INC3	Income \$35,000-\$74,999	-0.212	0.097	(0.030)	-0.152	0.092	(0.098)
PARSOL2	Somewhat better than parents' standard of living	-0.108	0.076	(0.159)	-0.172	0.071	(0.016)
PARSOL3	Same as parents' standard of living	-0.128	0.085	(0.133)	-0.203	0.080	(0.011)
PARSOL4	Worse than parents' standard of living	-0.336	0.099	(0.001)	-0.382	0.094	(0.000)
HEALTH2	Good health	-0.354	0.071	(0.000)	-0.356	0.067	(0.000)
HEALTH3	Fair or poor health	-0.828	0.093	(0.000)	-0.852	0.087	(0.000)
AGE2	36 to 46 years	-0.224	0.081	(0.006)	-0.208	0.077	(0.007)
AGE3	47 to 64 years	-0.102	0.085	(0.228)	-0.066	0.079	(0.399)
AGE4	65 years or more	0.274	0.098	(0.005)	0.208	0.089	(0.020)
MARRIED	Married	0.550	0.068	(0.000)	0.530	0.063	(0.000)
FEMALE	Female	-0.034	0.061	(0.576)	-0.047	0.057	(0.407)
BLACK	Black	-0.094	0.087	(0.281)	-0.106	0.082	(0.194)
	Observations	1,503			1,747		
	Log likelihood	-1281.2			-1478.3		
	Pseudo-R <sup>2</sup>	0.11			0.10		
	Likelihood Ratio	287.0			317.4		
	Pr(Likelihood Ratio)	0.000			0.000		

Note: p-values in parentheses. Goodness-of-fit statistics for the multiple imputation model are the median values from the six regressions.