IMPUTING HOURLY WAGES IN THE NATIONAL MEDICAL EXPENDITURE SURVEY

Ayah E. Johnson, Steven B. Cohen, and Alan C. Monheit
National Center for Health Services Research and Health Care Technology Assessment

KEY WORDS:

Introduction

The Household Component of the National Medical Expenditure Survey (NMES) was established to provide an assessment of health care utilization, costs, sources of payment, and insurance coverage of the U.S. civilian noninstitutional population. The household component is an 18-month panel survey with 1987 as the reference period, collecting measures on health status, use of health care services, expenditures and sources of payment, insurance coverage, employment, income and assets, as well as demographic information.

Complete nonresponse at the person level for NMES survey data was accounted for by adjusting the sampling weights. However, the proportion of item missing data is a problem for a large number of variables. Logical edits have been implemented and, whenever possible, data has been imputed based on other information provided. The issue being addressed is how to deal with the remaining item nonresponse, given that standard analysis techniques for modeling, or inference, require data sets with complete data profiles. In the absence of this criterion, the current choices are either to impute for missing data or to analyze only cases with complete data.

Imputing for missing data is done to maintain the representativeness of the sample, and to enable users to make national and regional estimates. Imputation also ensures consistency between results of different analyses, reduces nonresponse bias for item nonresponse, and enhances the ability to apply standard analysis techniques to data sets with complete data profiles without loss of sample size. Imputation does introduce a new component of variability, variance due to imputation. This variability can be quantified by implementing multiple imputation for each missing value. In this paper, two methods of multiple imputation are carried out. The first uses the "hot deck" procedure, and the second uses the Bayesian approach (Rubin, 1987). The variance due to imputation is computed, and a comparison is conducted to examine whether there is a significant difference in the variability due to imputation captured by each of these two methods.

Analyzing only cases with complete data can be viewed as analyzing the data as reported by persons participating in the survey. If the assumption is made that respondents are similar in their characteristics to nonrespondents, no bias and no variation due to imputation is introduced in data analysis. Moreover the representativeness of the sample can be maintained by proper weight adjustments. There is some loss of precision depending on the rate of item nonresponse since the sample size gets smaller, but the effort associated with obtaining a set with complete data is simplified.

The assumption that respondents are similar in their characteristics to nonrespondents is not a necessary assumption that can be used for estimating parameters of interest, but it is made to facilitate imputation of data. Although many studies have shown that there is a differential between respondents and nonrespondents, the assumption is very difficult to prove or disprove. It can, however, be inferred based on partial responses. Thus, the choice between imputing or analyzing complete data is a trade-off between additional variability due to imputation versus nonresponse bias and increased variance of the estimators.

The focus of this paper is to investigate and assess the various alternatives for dealing with item nonresponse. We have focused only on one variable: "hourly wage". The sample universe for this analysis is all employed persons in the first wave of the Household component of the National Medical Expenditure Survey (NMES). The imputation techniques that are being evaluated include:

(1) Model-based imputation
(2) Nonresponse adjustments to sample weights
(3) hot deck imputation, and
(4) multiple imputation.

The statistical assessment includes analysis of the effect of imputation on the distribution and on the variability of survey estimates such as means and proportions, and their respective standard errors. There is no assessment of bias. An assessment of bias can be done if one creates a synthetic subset of nonrespondents, imputes for missing data using each of the methods, and compares the imputed values to the observed. This will be done as a follow-up to this study.

Data Base Used for Imputation

The population includes all eligible persons who reported having a job and are not self-employed in the Household Component of NMES. This segment of the data base includes 13,605 eligible persons with positive weights. Of those, 11,614 had reported their hourly wage. The remaining 1,991 persons did not report their hourly wage. Hourly wages were derived based on a sequence of questions on work related patterns including:

1. Is the person employed? If so does he/she work for someone else?
2. How many hours does the person work a week ?
3. How many days a week ?
4. What is the wage rate, and the unit, "PER WHAT" (per hour, per day, per month, per year), associated with the wage rate.

Over 99% of the non-reporters did not complete the "PER WHAT" question (part 2 of question 4). 88.4% provided neither a dollar value for wage rate nor the unit of payment (per hour, per year). All non-reporters completed the question on the number of hours they worked per week, but only 7.2% answered the question concerning the number of days per week.

The pattern for nonresponse to these questions shows that most of the nonrespondents did not want to disclose their wage rate but were willing to disclose some information on the number of hours a week they worked. There is 14.63% non-response for this one variable, and the objective is to attempt different imputation techniques and compare them as to the impact on estimates of means and proportions, as well as distributions.

Modeling Hourly Wages and Nonresponse to Wages

Prior to any imputation, one has to model (a) the hourly wage; and, (b) the nonresponse to hourly wage. The models are designed to help identify (a) covariates which explain the hourly wage; and, (b) covariates that can explain the nonresponse to the hourly wages. The two behavioral equations are linked because wages are allowed to alter the propensity to report; and by building a joint model which represents both the regression model to be estimated and the process of determining when the dependent variable is to be observed, we can account for non-randomness of the observed values of the dependent variable. Thus, if [V1,....,Vm] denotes the set of covariates explaining the hourly wage, and [X1,......,Xn] denotes the set of covariates explaining the nonresponse to the hourly wage, then the intersection of these two sets constitutes the new set of variables used for classification and sorting of the population of interest. The classification and the sorting of the population is done to define a pool of donors who are similar in their characteristics to recipients. This new set is used for nonresponse adjustments and for hot deck imputation. For model-based imputation and multiple imputation, in addition to modeling the hourly wage, one has to adjust for selection bias. This selection bias reflects propensity to not report the hourly wage: It is an adjustment to the model which captures the differences between reporters and non-reporters. The selection bias is estimated by modeling nonresponse using ordinary least squares.

Modeling the Hourly Wages

The model used to predict hourly wages is based upon a standard "human capital" earnings function described in detail by Mincer (1974), and Chiswick (1974). The model hypothesizes that an individual's labor market earnings are related to investments in formal schooling, measured by years of schooling completed, and investments in on-the-job training, measured indirectly by years of labor market experience[1]. In addition, the model adjusts for earning differences attributable to geographic differences, measured indirectly by region and urban locality, ethnicity, gender, family structure, marital status, occupation and number of hours of work. The model also includes the square of experience, to accommodate considerations of economic theory which predict that post-schooling investments in human capital decline over the life cycle. Interactions between race and gender, and marital status and gender, are included to capture other possible non-linearities in earning behavior. The explicit list of predictors to the log of hourly wages is listed in Table 1.0.

Empirical results provided in Table 1.0 reveal that the model explains 40% of the variation in the log of hourly wages, which is reasonable for micro household data, and is consistent with applications of earning functions reported by Mincer, 1974, Chiswick, 1974, and Grossman and Benham, 1973.

The model was specified as follows: let Y1 denote the Log(hourly wages), and let (V1,...,Vm) denote the set of covariates (main effects and interactions), then for each person, i, who responded, the log of hourly wage, $Y1_i$, is specified as:

$$Y1_i = \text{Constant} + \beta_1 * V1 + \ldots\ldots + \beta_m * Vm;$$

SURREGR[1], (Holt, 1977) was used to estimate the coefficients of the semi-log model. SURREGR is a computerized procedure that derives weighted least squares estimates of regression parameters, and tests for hypothesis when complex survey data are used.

Since the model is fitted using data from respondents only, there is a potential for selection bias (as mentioned above). To adjust for selection bias, a two step estimator is used (Olsen, 1980; Mitchell et al., 1986). The first stage employs ordinary least squares to construct an estimator which accounts for the nonrandomness of the sample, and is denoted by PR1. Thus PR1 is a new regressor which explains the nonresponse and, when included in the regression models, corrects for the possibility of selection bias with respect to the observed dependent variable. Olsen shows in his paper (1980) that this correction leads to a similar correction factor as the Mill's ratio. In the second step of modeling, this additional regressor, PR1, is included in the original model, and the person's log hourly wage is adjusted for selection bias; the new model is then estimated by weighted least squares, SURREGR in this instance. The new model incorporating the selection bias term is then:

$$Y1_i = \text{Constant} + \beta_1 * V1_i + \ldots\ldots + \beta_m * Vm_i + PR1_i$$

and, the sample person weight which is determined by the survey design is adjusted by the selection weight, WGT1, where

$$WGT1 = 1/(\surd PR1_i * (1-PR1_i)).$$

Table 1.0 provides the estimates of the coefficients, the respective standard errors, and the α value indicating the level of significance after the adjustment for selection bias.

A second adjustment was made after predicting the hourly wage using the model. This adjustment was suggested to correct for bias introduced by the retransformation of the "log of hourly" wages to "hourly wages" (see Duan et al., 1982).

Years of education, most occupations, number of hours a week a person works, the North central and the South regions, SMSA, experience, and employment status of the spouse are statistically significant. The variable explaining the experience as well as the interactions of marital status and gender are statistically significant. Gender is included but is not statistically significant, and within the different racial groups, only Hispanics have a statistically significant effect. The selectivity bias introduced in the equation is significant at $\alpha = 0.05$, indicating that the propensity to report has an effect on the log of the hourly wage.

Table 1.0. Parameter Estimates for the Log of Hourly Wages

Variable	Estimate	Standard error		
Constant	1.29	0.15		
Gender	-0.07	0.05	0.15	
Race				
Blacks	-0.02	0.02	0.17	
Hispanics	-0.06	0.02	0.01	*
Non-Whites	-0.06	0.04	0.11	
Years of education				
9-11	0.10	0.03	<0.01	*
12	0.19	0.03	<0.01	*
13-16	0.28	0.03	<0.01	*
16+	0.50	0.03	<0.01	*
Occupations				
Managerial & professional	0.03	0.03	0.22	
Sales	-0.03	0.03	<0.01	*
Administrative	-0.14	0.02	<0.01	*
Precision production	-0.05	0.02	0.06	
Operators, fabricators laborers	-0.15	0.02	<0.01	*
Transportation	-0.12	0.03	<0.01	*
Service occupation	-0.40	0.02	<0.01	*
Laborers/ not Farming	-0.26	0.03	<0.01	*
Farming managers	-0.50	0.14	<0.01	*
Farming operators	-0.45	0.05	<0.01	*
Unknown	-0.18	0.07	0.02	*
No. of hours working				
Full time/one job	0.10	0.02	<0.01	*
Full time/more than a job	-1.68	0.75	0.03	*
Region				
North Central	-0.09	0.02	<0.01	*
South	-0.14	0.02	<0.01	*
West	-0.02	0.02	0.41	
SMSA				
Noncertainty SMSA	-0.13	0.02	<0.01	*
Other SMSA	-0.21	0.02	<0.01	*
No. of children	-0.002	0.01	0.76	
Experience	0.03	0.002	<0.01	*
Spouse Employed	-0.07	0.03	0.02	*
Spouse reported wages	0.02	0.02	0.52	
Marital Status				
Never married	-0.05	0.03	0.05	*
Widowed	-0.01	0.03	0.70	
Divorced	-0.03	0.01	0.26	
Separated	-0.09	0.04	0.02	*
EXPERIENCE^1	-0.0001	<0.0001	<0.01	*
Race * Gender				
White males	0.11	0.03	<0.01	*
Black males	0.01	0.04	0.80	
Marital Status * Gender				
Married male	0.31	0.05	<0.01	*
Male never married	0.11	0.05	0.04	*
Widowed male	0.24	0.101	0.02	*
Separated male	0.25	0.06	<0.01	*
Divorced male	0.33	0.06	<0.01	*
Selectivity Bias	0.71	0.28	0.01	*

$N = 11614$
$R^2 = 0.40$

* Significant at =0.05

Modeling Nonresponse to Hourly Wages

One of the major issues debated when imputation strategies are used is whether non-reporters and reporters have similar characteristics. Lillard et al. (1978) noted that non-reporters on income are either the high earners or those that earn very little, or are below the poverty level. Moreover, they state that those who do not report income are systematically different than those who do. For economists, the important question is whether non-reporters differ in their income from reporters, and whether such differences are fully captured by variables used to define the pool of "donors" in the imputation. Thus modeling and understanding nonresponse is an integral part of imputation even when model-based imputation is used.

Major reasons for nonresponse to wage or income questions are generally: (1) a demand for privacy; (2) a fear for governmental uses of the data, particularly for income tax rates; or, (3) a price of time for completing the survey. These cannot be modeled directly, and surrogate measures as covariates are used to quantify them. Since response/nonresponse is a binary variable, stepwise logistic regression was used as the empirical model.

Let $R1_i$ denote whether the ith person provides the hourly wage, $Y1_i$, or not. Then,

$R1_i = 1$ when $Y1_i > 0$, and
0 otherwise.

Let $(X_1,\dots\dots,X_n)$ be the set of covariates, then

$$\text{Log}\,\{P[R(i)]/[1-P[R(i)]]\} = B_0 + B_1 X_1 + \dots\dots\dots + B_n X_n.$$

This model views the outcomes of the dependent variable as a probabilistic event, and the coefficients of the model as the marginal changes in probability associated with each of the independent variables.

Table 2.0 summarizes the estimates of the coefficients and their respective standard errors. The significance levels for these values are all less than 0.0001, thus the estimates are highly significant. One significant result of this model is the fact that, if a spouse did not report wages there is a high likelihood that the respondent did not report wages. This trend indicates the possibility that for married couple there was one proxy respondent. Also, persons from service occupations were less likely to respond to the hourly wage question.

Table 2.0 Modeling Nonresponse using Logistic Stepwise Procedure

Variable	Estimate	Standard Error
Constant	2.30	0.0008
Spouse reported wages	2.61	0.0010
Spouse employed	-1.54	0.0008
Gender	-0.5230	0.0007
Northeast	-0.4353	0.0007
Race (black/ not black)	-0.46	0.0009
No. of children	0.18	0.0003
Experience	-0.01	0.0002
Service Occupations	0.31	0.0010
Gender * Divorced	0.35	0.0022
Divorced persons	0.15	0.0015

Imputation Of Hourly Wages Background

Imputation strategies include logical imputes of the data, mean value imputes--replacing missing values by means, cold deck and hot deck imputation, model-based and multiple imputation. The mean value imputes or the cold deck approach are not considered for this analysis because of the need to estimate the wage distribution nationally, using donors from the same data base.

Comparison of the Different Imputation Strategies

Imputation strategies should incorporate knowledge accumulated during the data collection, and try to preserve certain properties already existent in the data base. Those include:

- o Preserving the distribution of hourly wages.
- o Providing a mechanism to compute sampling errors which reflects the fact that some data has been imputed.
- o Avoiding extrapolation beyond the reach of the data.
- o Accounting for contextual knowledge about variables in the data base.
- o Maintaining the representativeness of the sample, and reducing the nonresponse bias.

In order to assess the different imputation techniques we have assumed that model-based imputation is a standard to which other techniques are compared. That decision was based on previous research by Lillard et al. (1982), who advocated model-based imputation as opposed to hot decking, mainly because it maximizes the use of contextual knowledge of responses and nonresponse in the data base. The comparisons are designed to examine the differences among means and proportions obtained after implementing each imputation strategy. Also, in order to measure the relative increase in variance due to imputation, we compared variances of estimators computed after each of the three imputation strategies, to the variance obtained after the model-based imputation is implemented. Estimation of bias or mean squared error (MSE) will be done as a follow-up paper, and those statistics may indicate which of the imputation strategies are coming closer to the truth.

Mean Hourly Wages Using Different Imputation Strategies

The mean hourly wage is computed, and the standard errors are computed for each set of imputed values using the Taylor linearization procedure, which accounts for the complex survey design of the NMES. Table 3.0 summarizes the mean hourly wage, the respective standard error, and the overall variance; it also provides the variance of the estimators and an estimate of the overall standard errors for the hot-deck and the Bayes multiple imputation techniques. These standard errors incorporate the variance within and between the three imputed values. A Z-test was performed to test whether there is a statistically significant difference between the mean hourly wage using the model-based imputation, and each of the remaining imputation strategies: (1) nonresponse adjustment; (2) multiple imputation using the hot deck; and (3) multiple imputation using Bayes. In these comparisons we assumed that the model-based approach to be a standard.

Two additional measures are provided in Table 3.0[2]: RV(1) and RV(2). RV(1) is the ratio of the variance of the mean hourly wage using one of the three imputation techniques, to the variance of the mean hourly wage obtained when the model-based approached is applied (variance for the same estimator). RV(2) is the ratio of variance of the estimated mean hourly wage using the multiple imputes obtained from the Bayesian approach, to the variance of mean hourly wage obtained using the multiple hot deck imputation. The column providing the "mean value" in the table gives the mean of the imputed values.

From the results summarized in Table 3.0 it can be seen that the mean hourly wage is relatively stable except for the case when the Bayesian approach is used; even then, the difference is less than one dollar an hour. The mean is only slightly higher when comparing means obtained using the hot deck multiple imputes to the model-based imputes: \$9.5649 compared to \$9.4075; a difference of, \$0.1574. The mean using the nonresponse adjustment is \$9.552 which is very close to the one obtained from the multiple hot deck. The mean using the Bayes approach is highest at \$10.0489. It is \$0.6414 higher than the one from the model-based approach, and \$0.4840 higher than the mean obtained using multiple hot deck. Thus, the maximum difference in the mean hourly wage is at 64 cents, which can yield to a difference of \$1331.2 in the annual wages for a full time employee. When comparing the different means, only the mean resulting from the Bayesian imputes is significantly different from the mean computed after a model-based imputation; the others are not statistically different.

The standard errors are stable, and they increase when the mean increases, accounting for more variability introduced by the imputation technique. There is an increase of 8% in the variance after "hot decking" over the variance estimated after using the model-based strategy. The increase in variance of the mean hourly wage obtained after Bayes imputes is 29%, and the increase in variance after the nonresponse adjustments is 35%. The ratio of variances of the two multiple imputes, Bayes versus hot deck, shows an increase of 19% in variance if the Bayes imputation is used. These results support other findings indicating that model-based imputations yield smaller standard errors. Until one examines some measures of bias one cannot conclude whether this is due to an underestimate of the variance, or this results in a more precise estimate of the variance. This result also indicates that the increase in variance is most notable when implementing nonresponse adjustments, thus increasing the potential for nonresponse bias.

Mean Hourly Wage by Occupation Using Different Strategies.

Table 4.0[2] summarizes the results of comparisons among the various imputation strategies. The estimates being compared are the mean hourly wages within twelve occupation subgroups. A Z-test for the difference of the mean hourly wage is computed assuming that the mean hourly wage after model-based imputation is a standard. None of Z-tests were significant except when using the Bayesian multiple imputation approach. Although the difference between the model-based and the multiple imputation using Bayes is statistically significant, the actual magnitude of the difference in means is less than one dollar, indicating there any imputation strategy would have resulted in a comparable mean hourly wage when examining the different imputation subgroups.

For the nonresponse adjustments there is an increase of 10% to 50% in the variance over the model-based imputation, with a four fold difference for the group that has not reported its occupation. The minimum increase in variance is, for the "operatives", 10%, followed by the "service workers" at 18%. A lower increase in variance is detected when "hot decking": 4% to 22%, except for a single occupation group where the variance decreases by 4%, the "operatives". The highest increase in variance is exhibited for the "farm laborers and foremen" at 32%. When using Bayes, that increase ranges between 1% and 222%. The largest increase in the variance is for the group which has not reported its occupation. When examining the ratio of hot-deck variances versus the Bayes variance, the increase indicates that the Bayes yields, in most cases more conservative estimates, since the variances are higher.

This comparison indicates that, although the mean hourly wage computed after implementing each imputation strategy is not always statistically different in any of the comparisons conducted, the impact of the choice of imputation strategy is most apparent when computing the variance of the mean. The decision of which imputation to use can be led by the decision of how conservative the estimate should be.

Comparison of the Distributions of Hourly Wage

The distribution of the hourly wage is obtained after each imputation technique has been executed. In this analysis, the distribution after executing the model-based imputation is assumed to be a standard and all other distributions were compared to it. There is a statistically significant difference in the proportion of the population obtained in a subgroups after the nonresponse adjustments have been completed:

- o The population subgroup earning an hourly wage that is below the minimum hourly wage, $0.01 - $3.50 an hour
- o The population subgroup earning above the minimum hourly wage, between $5.01 - $7.50 an hour,
- o The population subgroup located in the beginning of the upper tail of the distribution, those earning between $20.01 - $50.00 an hour.

These differences exist, and are statistically significant after "hot decking" and after imputing using Bayes theory. It is possible that, for these three subgroups of the population, the model-based approach results in an under estimation of the proportion of the population. When using the multiple Bayes approach there is one additional subgroup where the difference in proportion is statistically significant, the group earning between $7.51 and $10 an hour. Also, the proportions for those earning less than $3.50 an hour are not statistically significantly different.

The increase in variability of the proportions within subgroups, RV(1), is generally greater than 1, indicating that the model-based approach yields smaller variances, regardless of the estimators of interest. The increase in variance is most pronounced when using the multiple Bayesian approach. The increase in variance is the highest at both tails of the distribution. The increase in variance is most pronounced for the subgroup of the population earning between $50 and $75 an hour. The ratio of the variance is a relative measure, the actual difference between the variances is not large.

Conclusions

In this paper we have summarized the results of imputing hourly wages for NMES using the various imputation strategies described in the literature. The results of this exercise indicate only that at,15% item nonresponse, the various imputation strategies yield comparable means and proportions. Larger differences are apparent when using the multiple Bayes approach.

The various imputation strategies yield comparable means and proportions. Larger differences are apparent when estimating means than when estimating proportions. Thus, each of these imputation techniques is designed is preserve the distribution of the variable. On the other hand, the variances associated with these estimators show a higher degree of stability within an imputation technique for the estimates of proportion than for the estimates of the means.

Therefore these imputation techniques do preserve the distribution of the variable being analyzed. The two imputation techniques that provided a mechanism for computing variance due to imputation are the multiple hot-deck and the multiple Bayes. For the hot-deck, the variance due to imputation for "hourly wage", with a 15% item nonresponse is very small, in some cases non-existent. For the Bayes approach the variance due to imputation is large and very pronounced in the estimates of the mean. It is not clear whether this is an estimate due to imputation that is not captured by the other methods, or whether this is an art effect of the imputation technique used. These results cannot be generalized for any other variable, or at any other level of nonresponse.

The log hourly wage model accounted for the contextual knowledge about variable in the data base. This model, as well as the model explaining the patterns of nonresponse was the foundation for developing all imputation procedures. Thus, a careful specification upfront of the models may be a worthwhile investment, which can lead to a cleaner data base.

In terms of costs and other practical consideration, it is not clear which of the methods is more expensive. The element that is clear is the increase in the complexity of the data base, and the increased effort required for analyzing multiply-imputed data bases. For a large and complex survey such as NMES, one has to have an overwhelming justification for carrying observations, and the number of potential variables to be included in any one analysis, can be exorbitant.

In following research (to be submitted at the winter meeting) we will address the issue of which of these techniques lead us closer to the "truth". In order to do that a subgroup of the respondents will be selected and converted to "nonrespondents", and their values will be imputed. The imputed values will be compared to the reported values and estimates of bias as well as mean squared errors will be produced. An additional dimension will be added to this research, addressing the issue that is not discussed openly in the literature: At what point we cannot impute because there is not sufficient information to be used for reliable imputation. The proportion of nonrespondents will be set at different rates, imputation will be conducted, and again, bias and mean squared error computed.

---

[1]There is, currently, no direct measure of labor market experience. Therefore we follow the convention of computing experience as a difference between current age, and years of completed schooling, while adjusting for beginning year of schooling, arbitrarily set at age 6.

[2]All tables are available from the author.

REFERENCES

1. Chiswick, B.R. (1974). Income Inequality. New York: Columbia University press for the National Bureau of Economic Research.
2. Cox, B.G., Cohen, S.B. (1985). Methodological Issues for Health Care Surveys. Marcel Decker, Inc., New York and Basel.
3. Duan, N., Manning, W.G., Jr., Morris, C.N., Newhouse, J.P. (1982). A comparison of Alternative Models for the Demand for Medical care. Rand Health Insurance Experiment Series, pg. 29-33.
4. David M., Little R.J.A., Samuhal, M.E., and Triest R.K. (1986). Alternative methods for CPS income imputation. Journal of American Statistical Association 81, 29-41.
5. Ford, B.L. (1983). An overview of hot deck procedures. Incomplete Data in Sample surveys, Volume 2, Theory and Bibliographies. W.G. Madow, I. Olkin, and D.B. Rubin.
6. Greenlees, W.S., Reece, J.S., and Zieschang, K.D. (1982). Imputation of missing values when the Probability of Response depends on the variable being imputed. Journal of the American Statistical Association, Vol 77, pg. 251-261.
7. Grossman, M. and Benham, L. (1974). Health, Hours and Wages. The Economics of Health and Medical Care. John Wiley and Sons, Inc., New York.
8. Hertzog T.N. (1980). Multiple imputation modelling for individual social security benefit amount, Part II. Proceedings of the Survey Research Methods Section of the American Statistical Association, 404-407.
9. Hertzog T.N., Lancaster C. (1980). Multiple Imputation Modelling for Individual Social Security Benefit Amount, Part I. Proceedings of the Survey Research Methods Section of the American Statistical Association, 398-403.
10. Holt M.M. (1977). SURREGR: Standard Errors of Regression Coefficients from Sample Survey Data. Research Triangle Institute, North Carolina 27709.
11. Kalton G. and Kasprzyk D (1986). The treatment of Missing Survey Data. Survey Methodology, Statistic Canada, Vol 12, No. 1, pp 1-16.
12. Lillard, L., and Smith, J.P., and Welch, F. (1986). What do we really know about wages? The importance of nonreporting and census imputation. Journal of Political Economy 94, 489-506.
13. Little, R.J.A. (1982). Models for Nonresponse in Sample Surveys. Journal of American Statistical Association 77, 237-250.
14. Little, R.J.A., Samuhel M.E., and Triest R.K., (1986). Alternative Methods for CPS Income Imputation. Journal of the American Statistical Association, Vol 81, No. 393.
15. Little, R.J.A., and Schluchter, M.D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. Biometrika 72, 497-512.
16. Mincer.,J. (1974). Schooling Experience and Earning. New York: Columbia University Press for the National Bureau of Economic Research.
17. Mitchell, J.M., and Butler J.S., (1986). Arthritis and the Earning of Men. An Analysis Incorporating Selection Bias. Journal of Health Economics 5, pg 81-98, North Holland.
18. Oh, H. L., and Sheuren, F. J. (1983). Estimating the variance impact of missing CPS income data. Proceedings of the Survey Research Methods Section of the American Statistical Association, 408-415.
19. Olsen, R.J., (1980). Notes and Comments: A Least Squares Correction for Selectivity Bias. Econometrica, Vol 48, No. 7.
20. Neter, J., Wasserman, W., Kutner, M. H. (1983). Applied Regression Models. Richard D. Irwin, Inc. Homewood illinois.
21. Rubin, B. D. (1987). Multiple Imputation for Nonresponse In Surveys. John Wiley and sons.