EFFECTS OF IMPUTATION ON REGRESSION COEFFICIENTS

Robert Santos, Survey Research Center, University of Michigan

1. Introduction

Most research on imputation in surveys has examined its effects on means, totals, and other univariate statistics (see, for example, Bailar and Bailar, 1978; Ford, 1976; Kalton, 1981). Although regression analysis with missing data has been examined extensively, the effect of imputing missing data on bivariate statistics such as covariances, correlations and regression coefficients has not been widely treated in survey literature. Since regression analysis is frequently used in survey analysis, there is a need to assess the biases of estimated regression coefficients from imputed data.

The purpose of this paper is twofold. First, the large sample biases incurred by estimating regression coefficients with imputed data are examined theoretically in a finite population sampling framework. This investigation studies several imputation techniques and employs two missing data models. Secondly, the paper investigates empirically the impact of regression coefficient estimation from imputed data via a simulation study. Data from the first two waves of the 1978 Income Survey Development Program (ISDP) Research Panel, a prototype panel of the Survey of Income and Program Participation, are used for this study.

Before presenting the results, it is important to emphasize several limitations. The theoretical results assume a simple random sample design and specific nonresponse generating models. The empirical results rely on a specific nonresponse model and all results are obtained via simulation. Thus, interpretations must be made with caution and are suggestive.

2. Theoretical Framework

Consider a finite population of N units, in which each unit would either respond or incur missing data for a specified variate Y with probability 1, if a census was taken of the population. Thus, over repeated trials a unit always responds or always does not. The population may then be partitioned into "response" and "nonresponse" strata. Let R and M (where R+M=N) be the numbers of units in the response and nonresponse strata, respectively, and let the subscripts ´r´ and ´m´ be indicators of the respondent and nonrespondent subpopulations (e.g., $\bar{Y}_r$, S^2_{my}, S_{rxy}, etc.). Finally, assume that N, R and M are large.

Throughout we assume that a simple random sample without replacement (SRS) of size n has been drawn from the population. Three survey variables are to be recorded: X, Y, and Z. The data for X and Z are observed for all units in the sample; however, the values of Y are obtained for r of the n sample units (assume P{r>1}=1); m = (n−r) y-values are missing. Imputation of survey data yields a ´complete´ sample data set with no apparent missing data. Let y* denote the imputed value of the variable Y for all sample units for which Y is missing. Define the variable Y'_i as being the actual value y_i, if the unit is a respondent, and as an imputed value y^*_i, if not. Then the estimator of the element covariance may be expressed as

$$s_{xy} = \Sigma(x_i - \bar{x})y'_i/(n - 1). \qquad (2.1)$$

The regression coefficient of interest in the theoretical portion of this paper is the regression slope obtained from the regression of Y on X, defined by $B_1 = S_{xy}/S^2_x$ and estimated by $b_1 = s_{xy}/s^2_x$. Thus, it will suffice to consider the biases of s_{xy} due to imputation, since s^2_x is based on the total sample.

3. Imputation Schemes

As illustrated below, many imputation schemes may be construed as regression predictions. In the most general setting, the value assigned to a missing datum is determined by the model

$$Y_i = B_0 + \Sigma B_j Z_{ji} + E_i, \qquad (3.1)$$

where the Z_j (j=1,...,p) are available auxiliary variables. The "respondent" sample data are employed in the estimation of the regression coefficients, providing the imputations

$$y^*_i = b_{r0} + \Sigma b_{rj} Z_{ji} + \hat{e}_i \qquad (3.2)$$

where $E^*(\hat{e}_i) = 0$. Here E* denotes expectation over the generation of the estimated residual terms. As we will see later, the residual term plays an important role in retaining the variation found in the respondent data. The subscript r in the above coefficients reminds us that the estimated coefficients are based on the respondent sample data.

"Deterministic" imputations are covered in equation (3.2) by setting $\hat{e}_i=0$ for each unit with missing data. "Stochastic" imputations are readily derived from the regression prediction schemes by simply appending an appropriately determined random residual.

Various randomization schemes can generate the residual errors but presently we will consider only the following scheme: "observed" residuals $\hat{e}_{ri} = y_i - y^*_i$ for respondents will be randomly sampled and appended to the regression predictions.

The following discussion describes several common imputation techniques:

a. Do Nothing (DN). For completeness we consider the popular "imputation procedure" which rejects all incomplete cases in the sample and calculates estimators from the r remaining units.

b. Grand Mean (GM). This imputation scheme assigns the overall observed mean of the respondents to all missing data. The tacit model is $Y_i = B_0 + E_i$, with imputations made by specifying $b_{r0} = \bar{y}_r$. This scheme and the Do Nothing technique do not utilize the auxiliary information, Z, specified in the problem formulation (Section 2).

c. Cell Mean (CM). This imputation scheme partitions the sample into nonempty imputation cells determined by the auxiliary data, Z, available for all sample units. Each sample unit in class h with missing data is imputed the value $\bar{y}_{rh} = \Sigma\, y_{rhi}/r_h$. Such a scheme implicitly assumes a deterministic regression model of the form $Y_i = B_0 + \Sigma\, B_j D_{ji} + E_i$, where the variates D_{ji} are dummy variables corresponding to the H classes.

d. Random Cell Imputation (RI). Instead of imputing the class mean, one might assign the value recorded from a randomly selected respondent within the same imputation cell. For m_h nonrespondents in cell h, random cell imputation involves taking a SRS size m_h from the r_h observed units. The y-values of the selected units are randomly imputed to the missing data. Adjustments to this procedure may be made when $m_h > r_h$, although this situation is highly undesirable. The RI scheme employs respondent residuals $\hat{e}_{rhi}$ which are selected at random within imputation classes. That is, if unit i in class h has a missing y-value, it is imputed by $\bar{y}_{rh} + \hat{e}_{rh}$, where $\bar{y}_{rh}$ denotes the respondent mean from the imputation cell to which unit i belongs and $\hat{e}_{rh}$ is a randomly chosen respondent residual taken from the same cell.

e. Simple Regression Prediction Imputation (RG). This imputation method utilizes the observed data (z_i, y_i), to estimate regression coefficients, yielding $b_{ro} = \bar{y}_r - b_{r1}\bar{z}_r$ and $b_{r1} = s_{rzy}/s^2_{rz}$; predicted values $\hat{y}_i = b_{ro} + b_{r1}z_i$ are imputed to the missing values. (See Buck,1960.)

f. Simple Regression Prediction Plus Random Residual (RR). Instead of imputing missing values directly from the estimated regression line, one might wish to randomly disperse the regression predictions by adding a random residual error term with zero mean. The imputations are defined by $y^*_i = \hat{y}_i + \hat{e}_i$ where $\hat{y}_i = b_{ro} + b_{r1}z_i$ and $\hat{e}_i$ is randomly sampled from the "observed" residuals $\hat{e}_{ri} = y_i - \hat{y}_i$.

4. Missing Data Models

The biases of element covariance and variance estimators obtained from imputed data depend on the nature of the missing data. While general formulae for the biases can be derived, there are no clear conclusions without making assumptions about the relationships between the respondent and nonrespondent populations. For these reasons, two alternative missing data models will be assumed, and the corresponding biases of the various imputation schemes in estimating S_{xy} and S^2_y will be investigated within the context of each model.

First, the data will be assumed to be "missing at random," the respondents being a SRS of size R from the population. For very large values of R and M, univariate and bivariate subpopulation parameters can be assumed equal, i.e., $\bar{Y}_r = \bar{Y}_m = \bar{Y}$, $S_{rxy} = S_{mxy} = S_{xy}$, etc. Although the assumption of randomly missing data across the total population is usually untenable, the use of this model provides useful insights into the effects of the various imputation schemes.

A less restrictive nonresponse model assumes that the data are randomly missing (in the sense described above) within certain subclasses of the population. This is the second model which we will consider. The classes within which the data are missing (at differing rates) are assumed to be the imputation classes defined by the Cell Mean Imputation technique. Thus, within the subgroups formed by the variate Z, the population parameters for respondents and nonrespondents are approximately equal: $\bar{Y}_{rh} = \bar{Y}_{mh} = \bar{Y}_h$, $S_{rxyh} = S_{mxyh} = S_{xyh}$, etc., for $h = 1,\ldots,H$.

5. Covariance Biases

Theorem 1. Assuming the formulation of the missing data problem in Section 2, and excluding terms of order $O(n^{-1})$, the biases of s_{xy} as defined in (2.1) for the various imputation schemes are:

$$\text{Bias}(s_{xyDN}) = 0, \tag{5.1}$$

$$\text{Bias}(s_{xyGM}) = -\bar{M}S_{xy}, \tag{5.2}$$

$$\text{Bias}(s_{xyCM}) = \text{Bias}(s_{xyRI}) = -\bar{M}\,\Sigma\, W_h S_{xyh} \tag{5.3}$$

$$\begin{aligned}\text{Bias}(s_{xyRG}) &= \text{Bias}(s_{xyRR}) \\ &= -\bar{M}S_xS_y(1 - R^2_{xz})^{1/2}(1 - R^2_{yz})^{1/2} \\ &\quad \cdot R_{xy.z}\end{aligned} \tag{5.4}$$

where $\bar{M} = M/N$, $W_h = N_h/N$ and $R_{xy.z}$ is the partial correlation between X and Y adjusted for Z. Derivations are given in Santos (1981).

The DN scheme yields an approximately unbiased estimator of S_{xy}; the present missing data model implies that the complete data are a random sample of a reduced size. Imputation of the grand mean estimates the covariance of the nonrespondents by zero and hence yields estimators of S_{xy} with a negative relative bias of $\bar{M}$. The CM scheme, on the other hand, estimates the covariance of the nonrespondents by the "between class" component of the covariance. This accounts for the negative relative bias of $\bar{M}p$, where p is the proportion of the covariance attributed to the "within class" covariance. The Bias(s_{xyCM}) vanishes to terms $O(n^{-1})$ when the imputation classes account for all of the covariance between X and Y (i.e., when $S_{xyh} = 0$ for $h = 1,\ldots,H$). The Bias(s_{xyRI}) is identical to that of the Cell Mean Imputations; the reason is due to an average zero residual and an uncorrelated (with X) error term.

The relative bias for the Simple Regression Prediction (or Regression Prediction Plus Random Residual) vanishes when either $|R_{xz}| = 1$, $|R_{yz}| = 1$, or when $R_{xy} = R_{xz}R_{yz}$ (i.e., $R_{xy.z} = 0$). Thus, s_{xyRG} and s_{xyRR} are approximately unbiased when the regression of Y on X is used for imputations or when Z and X are highly correlated. Also, they are roughly unbiased for S_{xy} when either Z and Y are highly correlated or when $R_{xy.z} = 0$. The former condition arises when Y and Z are linearly related. The latter condition is fulfilled when the XY correlation can be fully explained by the XZ and YZ correlations.

Theorem 2. Suppose that the data for Y are randomly missing within the cells (determined by the auxiliary variable Z) defined by the Cell Mean Imputation scheme. Furthermore, assume the missing data problem outlined in Section 2. Using the notation of Theorem 1, and excluding terms of order $O(n^{-1})$ or smaller,

$$\begin{aligned}\text{Bias}(s_{xyDN}) = -\bar{M}\ \Sigma\,(W_{mh} - W_{rh})[S_{xyh} \\ + (\bar{X}_h - \bar{X})(\bar{Y}_h - \bar{Y})] \\ - \bar{M}^2(\bar{X}_r - \bar{X}_m)(\bar{Y}_r - \bar{Y}_m).\end{aligned} \qquad (5.5)$$

$$\begin{aligned}\text{Bias}(s_{xyGM}) = -\bar{M}\ \Sigma\ W_{mh}[S_{xyh} + (\bar{X}_h - \bar{X})(\bar{Y}_h - \bar{Y})] \\ - \bar{R}\bar{M}^2(\bar{X}_r - \bar{X}_m)(\bar{Y}_r - \bar{Y}_m);\end{aligned} \qquad (5.6)$$

$$\text{Bias}(s_{xyCM}) = \text{Bias}(s_{xyRI}) = -\bar{M}\ \Sigma\ W_{mh}S_{xyh}; \qquad (5.7)$$

$$\text{Bias}(s_{xyRG}) = \text{Bias}(s_{xyRR}) = (B_{ryx} - B_{yx})S_x^2, \text{ for } X = Z; \qquad (5.8)$$

Here, $W_{rh} = R_h/R$, $W_{mh} = M_h/M$, and $\bar{R} = R/N$.

Equation (5.8) concerns the special case where the cells within which the data are assumed to be missing at random and the predictor auxiliary variable are the same. The regression imputation schemes will yield unbiased estimators of S_{xy} if the respondent and population regression slopes coincide. The expression for the bias when a different variate, Z, determines the randomly missing data classes is complex and hence not included.

The bias of the Cell Mean (and Random Cell) Imputation is analogous to that attained when the data were completely missing at random. The difference here is the weights, W_{mh}, instead of W_h.

The Grand Mean assignments show the dangers of ignoring the response model when adjusting the data. Assuming all of the components in (5.6) are of the same sign, the $\text{Bias}(s_{xyGM})$ will be larger than that of the imputation class schemes. The last term in equation (5.6) reflects the differing response rates within the subgroups assumed to have data missing at random.

The Do Nothing technique yields a bias which is a function of the differences between the respondent and nonrespondent subclass weights, $W_{mh} - W_{rh}$. When the response rates, R_h, differ highly between the missing data model's subgroups, the bias will tend to be large. However, when the response rates are fairly similar, the bias will be reduced substantially. Note that the bias is zero when the response rates among the missing data model's subgroups are uniform, i.e. $\bar{R}_h = \bar{R}$. This corresponds approximately to randomly missing data across the total population.

6. Element Variance Biases

The previous section demonstrated the biases of estimating the element covariance (using the standard formula) for several imputation schemes. We now briefly consider the element variance estimator.

Theorem 3. Assume the missing data problem outlined in Section 2; let s_y^2 be the element variance based on the imputed data set (analagous to 2.1). Also, suppose that the data are missing at random. Using the notation of the previous section and ignoring terms of order $O(n^{-1})$ or smaller,

$$\text{Bias}(s_{yDN}^2) = \text{Bias}(s_{yRI}^2) = \text{Bias}(s_{yRR}^2) = 0 \qquad (6.1)$$

$$\text{Bias}(s_{yGM}^2) = -\bar{M}S_y^2 \qquad (6.2)$$

$$\text{Bias}(s_{yCM}^2) = -\bar{M}\ \Sigma W_h S_{yh}^2 \qquad (6.3)$$

$$\text{Bias}(s_{yRG}^2) = -\bar{M}(1 - R_{yz}^2)S_y^2 \qquad (6.4)$$

The "deterministic" imputation strategies all understate the true variance. The regression prediction imputation yields a bias equal to $-\bar{M}$ times the residual variance obtained from the regression of Y on Z. Thus, the Simple Regression Prediction scheme understates the true variance by $\bar{M}(1 - R_{yz}^2)$ relative to S_y^2. Since the Cell Mean and Grand Mean techniques are special cases of the general regression prediction techniques their relative biases are equal to $-\bar{M}$ times the "proportion of variance unexplained" by the regression model.

An important reason for adopting a randomized imputation technique (RI and RR) is that the randomization scheme can be devised to retain the observed variation in the data. If the regression predicts well, the residuals will be small and the estimators s_{yRR}^2, s_{yRI}^2 will not differ substantially from their respective counterparts, s_{yRG}^2 and s_{yCM}^2. However, if the correlation between Z and Y is low or moderate, addition of the residuals to the regression predictions will increase the variation of the imputed data, yielding improved estimates of the element variance as seen in equation (6.1).

Theorem 4. Suppose that the data for Y are randomly missing within cells defined by the Cell Mean Imputation scheme. Also, assume the missing data problem as defined in Section 2. Using the notations of the previous theorems, and ignoring terms of order $O(n^{-1})$ or smaller,

$$\begin{aligned}\text{Bias}(s_{yDN}^2) = -\bar{M}\ \Sigma\ (W_{mh} - W_{rh})[S_{yh}^2 + (\bar{Y}_h - \bar{Y})^2] \\ - \bar{M}^2(\bar{Y}_r - \bar{Y}_m)^2\end{aligned} \qquad (6.5)$$

$$\text{Bias}(s_{yGM}^2) = -\bar{M}S_{my}^2 - \bar{R}\bar{M}(\bar{Y}_r - \bar{Y}_m)^2 \qquad (6.6)$$

$$\text{Bias}(s_{yCM}^2) = -\bar{M}\ \Sigma\ W_{mh}S_{yh}^2 \qquad (6.7)$$

$$\text{Bias}(s_{yRI}^2) = 0 \qquad (6.8)$$

Here, $S_{my}^2 = \Sigma\ W_{mh}[S_{yh}^2 + (\bar{Y}_h - \bar{Y}_m)^2]$.

The resultant biases for the simple regression imputations are incomplete and thus are not included. Except for the unbiasedness of the Random Cell Imputation scheme, the results of Theorem 4 mimic those of Theorem 2 for covariances.

7. Empirical Investigation.

Two small scale studies were conducted to examine the effects of estimating regression coefficients using imputed data. The first investigation imputes missing hourly rate of pay for those jobs paid hourly in the July wave of the 1978 Income Survey Development Program panel; the second study involves quarterly wage/salary income as reported from records in the April wave of the 1978 ISDP panel. Both investigations are confined to the area sample portion of the survey. Data sets were constructed by taking all the cases with valid data for the variable concerned and deleting some of the recorded values (using a specified nonresponse generating model), then applying several imputation schemes to form an "adjusted" data set. The imputed item is the dependent variable in all regressions below. Results are compared to the regressions using the real values in place of the imputed ones.

The nonresponse models used to create missing data mirror the actual patterns of missing data observed in the full sample. For hourly wage, data were randomly deleted (at differing rates) within 4 cells defined by crossing levels of sex and interview status (self report v. proxy). Cells defined by a crosstabulation of household income and interview status were utilized in deleting quarterly earnings. Full details of the methodology may be found in Kalton (1981). For hourly wage, nonresponse rates ranged 3.6% to 16.5% in the four cells, averaging 10.1%, while nonresponse rates for quarterly earnings ranged from 45.2% to 78.9%, averaging 55.0%. In order to reduce the variability due to randomly determining cases as "missing," the deletion process was replicated ten times, producing ten replicate simulation data sets. The imputations were performed separately within each replicate data set; regression statistics were then derived separately and averaged.

Table 1 presents estimated element variances for several imputation procedures. Presented as percentages of the true value, the quantities appear for the set of sample cases with missing data and the set of all sample cases. The first column reveals a 14-60% understatement of the hourly wage element variance within the nonresponse stratum (i.e., s^2_{my}) for imputation schemes c-f. Of course, the Grand Mean imputation exhibits no variability in the nonresponse stratum. The Cell Mean Imputation, which employs 8 imputation classes defined by the crosstabulation of union membership, occupation, and industry, displays the next most severe bias. The cell means accounted for 40% of the variation for hourly wage in the nonresponse stratum and hence the scheme understates the variance by 60%. The regression prediction imputation, which imputes a predicted value from the regression of hourly rate of pay on union membership, occupation, age, sex and hours worked per week, underestimated s^2_{my} almost as badly. (The regression employed in the above multiple regression imputation scheme produced an R^2 statistic of 0.53 among the responders.) The "stochastic" imputation schemes, d and f, show modest decreases in the magnitude of the bias of s^2_{my}, yet their biases remain substantial; the random cell imputations perform best, producing a lower estimate of s^2_{my} by about 14%. The regression/residual scheme retains a negative bias of about 17%.

A possible explanation can be given as to why a sizeable element variance bias in the stochastic imputation schemes occurred. First, the data are not missing at random across the total population. In consequence, a zero bias would be expected (for the Random Cell Imputations) if the cells used to delete data and those used in the Random Cell Imputations coincided. However, this was not the case. The upshot is a nonnegligible bias. The use of imputation cells which differ from the missing data model's cells was intentional, since in reality one never knows with certainty the true nonresponse model. The second column of Table 1 shows the effect of imputation on s^2_y, the overall element variance. Note that all estimates understate s^2_y by about 2-12%. The Do Nothing approach exhibits a relative bias of -2.5%. Apart from the Grand Mean imputations, the Cell Mean and Regression Prediction schemes retain the worst biases, with estimates of s^2_y about 7% below the true value. The Do Nothing and stochastic imputations perform best, yielding an overall negative bias of about 2% of the true element variance.

The last two columns of Table 1 show the severe biases incurred for estimating the element variance of quarterly earnings when the rate of nonresponse is high ($\bar{m}$ = 55%). For the imputation of quarterly earnings, the cell mean technique and the random cell imputation scheme employed 8 cells defined by the crosstabulation of sex, household income in March (<\$900, $\geq$\$900) and work status (full-time, part-time). The regression prediction used the regression of quarterly earnings on age, hours worked per week, occupation, household income, sex and whether job was held full quarter or not, yielding an R^2 statistic of 0.58 for the respondent data. The biases of s^2_{my}, s^2_y are varied, with the deterministic schemes (b,c,e) understating the true values severely and their randomized counterparts (d,f) understating the true values slightly. The Do Nothing technique estimated s^2_y best, yielding a relative bias of -0.7% while the stochastic schemes (d,f) incurred a slightly larger negative relative bias.

The results of Table 1 support the theoretical findings of Theorems 3 and 4. If the data for Y were missing at random (even though they are not), we would expect the relative bias of the Grand Mean scheme to be about $-\bar{M}$; this is approximately true in Table 1. Also, we would expect the Do Nothing and the stochastic imputations to yield approximately unbiased estimates of S^2_y. This is approximately true, but not quite; perhaps the reason is that the data are missing at random within cells and not missing at random across the total sample.

Table 2 gives the regression slopes for simple regressions of hourly wage on union membership and sex, respectively, and for the multiple regression of quarterly earnings on age and number of hours worked per week and the simple regression of quarterly earnings on weekly wage. Union membership (Column 1) is interesting because it was used in all imputation schemes (but not in the generation of missing data). The union membership coefficient is understated for schemes a, c-f by 2-3% of the true value. The understatement of the regression coefficient suggests a reduction of the element covariance in the imputed data set. The Grand Mean Imputation incurs a negative bias of about 14% for the regression slope.

The second column is of interest because sex was used to create missing data; furthermore, sex was used in the regression imputations (e,f) but not in the other schemes. The impact of this latter point shows clearly in the better estimation of the sex regression coefficient for schemes e, f compared to techniques b,c and d. Procedures a, e and f yield the least biased regression statistics.

The third and fourth columns (Regression Analysis 3) present the regression slopes which correspond to a *multiple regression* of quarterly earnings on age and number of hours worked per week. Age was utilized only in procedures e and f; hours worked[1] was used in all imputation schemes (except a). Note the severe understatement of the age coefficient in the grand mean and imputation class techniques (b,c,d). Comparing this to the hours worked coefficient (c,d), one might conject that failure to include age in the imputation class scheme incurs a large relative bias. The relatively good estimation of the age coefficient for the regression imputations tends to support this claim. The Do Nothing technique produces multiple regression coefficients which are within 6% of the true values, while the regression imputations (e,f) yield coefficients within 4% of the true values.

The last column of Table 2 shows a simple regression of quarterly earnings on weekly wage, a variable constructed by multiplying hourly wage by hours worked per week. The regression is based on the subclass of persons paid an hourly wage. Unlike the other variables, weekly wage is highly correlated with quarterly earnings (r = .903). Note that the weekly wage coefficient is substantially understated for procedures b-f by 18%-54%. The previous results suggest that s_{xy} is deflated due to the fact that weekly wage was not utilized in the imputations. The Do Nothing approach provides the best estimates of the weekly wage coefficient.

The weekly wage coefficients in Table 2 are intuitively consistent in the following sense: If the correlation between weekly wage and quarterly earnings was perfect (i.e. $r_{xy} = 1$), then the biases of the regression coefficient should approximately mimic the biases of the element variances in Table 1. This, indeed, is roughly the case.

The constant terms in the regression analyses (not shown) were estimated well (within 1-2% of the true value) for all imputation schemes when the dependent variable was hourly wage. However, the quarterly earnings regression analyses yielded extremely biased estimated constant terms. The discrepancies in the constant term biases between regressions with hourly wage and quarterly earnings as dependent variables can be attributed to the disparity in nonresponse rates and the superior estimation of regression slopes for hourly wage regressions over those for quarterly earnings. See Santos (1981) for further details.

Overall, the empirical results of Table 2 support the theoretical findings of Theorems 1 and 2. Theorems 1 and 2 indicate that the biases of the regression slopes will be equal for the deterministic regression prediction schemes and their randomized counterparts. This is confirmed by the empirical results. (Compare the regression slopes between procedures c and d, or between procedures e and f in Table 2.) Furthermore, the bias of the element covariance under the assumption of randomly missing data implies that the maximum relative bias of the general regression prediction imputation is $-\bar{M}$. (This occurs, for instance, when s_{mxy} is estimated by zero.) The regression slope biases adhere to these bounds.

In conclusion, the empirical results suggest several properties of imputations on regression coefficient estimation. First, one must realize that all imputations will alter the variance-covariance structure of the data set. The actual impact on estimated regression coefficients depends on the *extent* of missing data, the *differences* between the respondent and nonrespondent populations and the *use* of the independent variable(s) in imputing the dependent variable. Deterministic imputations will tend to understate the element variance; their stochastic counterparts will recapture some of the variance. Covariances will tend to be attenuated, especially between imputed variables and items not utilized in the imputation scheme.

A full treatment of the theoretical and empirical results in this paper may be found in Santos (1981).

REFERENCES

Bailar, III, J.C., and Bailar, B.A. (1978). Comparison of two procedures for imputing missing survey values. *Proc. Sect. Surv. Res. Meth., Amer. Stat. Assoc.*, 1978.

Buck, S.F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Jour. Roy. Stat. Soc.*, B, 22, 302-306.

Ford, B. (1976). Missing data procedures: a comparative study. *Proc. Sect. Soc. Stat., Amer. Stat. Assoc.*, 1976, 324-329.

Kalton, G. (1981). Compensating for missing data. Survey Research Center, University of Michigan. (Forthcoming)

Santos, R. (1981). The effects of imputation on complex statistics. Survey Research Center, University of Michigan. (Forthcoming)

[1] Work status, used in the imputation class schemes (c,d), is created from hours worked per week.

Table 1 Element variances of hourly wage and quarterly income by several imputation procedures for the imputed cases only and for all sample units

	Element Variances[1]			
	Hourly Wage		Quarterly Earnings (in millions)	
	Imputed cases only n=105	All sample cases n=1036	Imputed cases only n=220	All sample cases n=400
Real Data[5]	64192 (100%)	52796 (100%)	4.103 (100%)	4.078 (100%)
Adjusted Data:				
a. Do Nothing	—	97.5%	—	99.3%
b. Grand Mean	0%	87.6%	0%	44.6%
c. Cell Mean	39.8%	92.6%	42.4%	68.1%
d. Random Cell Imp.	86.5%	98.3%	96.0%	97.7%
e. Regression Prediction[6]	45.7%	93.3%	54.5%	74.7%
f. Regression/ Residual[6]	82.9%	97.9%	98.2%	98.8%

(See footnotes for Table 2.)

Table 2: Comparison of Regression Slopes in Four Regressions for Several Imputation Schemes where the Dependent Variables are Hourly Wage and Quarterly Earnings

Estimated Regression Slopes[1] (as percentage of true value)

Dependent Variable	Hourly Wage		Quarterly Earnings		
Regression Analysis	1	2	3		4
n	1036	1036	395 of 400[2]		217 of 239[3]
Independent Variable(s)	union mbr.	sex	age	hours worked	weekly wage[4]
Real Data[5] Slopes	155.6 (100%)	83.44 (100%)	29.92 (100%)	82.01 (100%)	0.1262 (100%)
Adjusted Data Slopes:					
a. Do Nothing	97.4%	98.8%	96.6%	105.7%	102.8%
b. Grand Mean	86.4%	88.8%	42.3%	47.8%	46.4%
c. Cell Mean	97.9%	95.8%	52.9%	91.6%	79.6%
d. Random Cell Imputation[6]	97.7%	95.9%	51.5%	91.9%	79.4%
e. Regression Prediction	97.8%	99.7%	97.9%	103.8%	82.3%
f. Regression/Residual[6]	98.0%	99.4%	97.9%	103.6%	81.6%

[1] The regression statistics are averaged over ten replicates of the simulation data set.

[2] Five cases were deleted because of missing data on the independent variables in the "respondent" portion of the sample.

[3] Twenty-two cases were excluded due to missing data on either hourly wage or usual number of hours worked.

[4] Weekly wage is a derived (constructed) variable obtained by taking the product of hourly wage and usual number of hours worked per week. The regression is conducted within the subclass of those with hourly paid jobs.

[5] Except for the "real data" row, all entries are percentages of the parameters to which they refer.

[6] Results for the stochastic imputation schemes are first averaged over ten iterations of the imputation scheme; then they are averaged over ten replicates of the simulation data set.