Title

Multiple Imputation Methodology for Missing Data, Non-Random Response and Panel Attrition

Author

Brownstone, David

Publication Date

1997-03-01

UCI-ITS-WP-97-4

Multiple Imputation Methodology for Missing Data, Non-Random Response and Panel Attrition

UCI-ITS-WP-97-4

David Brownstone

Department of Economics and
Institute of Transportation Studies
University of California, Irvine
Irvine, CA 92697-5100, U.S.A., dbrownst@uci.edu

March 1997

Institute of Transportation Studies
University of California, Irvine
Irvine, CA 92697-3600, U.S.A.
http://www.its.uci.edu

Revised version of a paper presented at the Theoretical Foundations of Travel Choice Modelling Conference, Stockholm, Sweden, August 7-11, 1996, This revision benefited from comments from the editors, Chandra Bhat, Ami Glazer and especially Kenneth Small. The research described in this paper was partially supported by the University of California Transportation Center.

1. Introduction

Modern travel-behavior surveys have become quite complex; they frequently include multiple telephone contacts, travel diaries, and customized stated preference experiments. The complexity and length of these surveys lead to pervasive problems with missing data and non-random response biases. Panel surveys, which are becoming common in transportation research, also suffer from non-random attrition biases. This paper shows how Rubin's (1987a) multiple imputation methodology provides a unified approach to alleviating these problems.

Before discussing solutions to problems caused by missing data and selection, it is important to recognize that their presence causes fundamental problems with identifying models and even "simple" population estimates. Section 2 reviews this work and stresses the need to make generally untestable assumptions in order to carry out any inference with missing data.

Once some identifying assumptions are made, the most common "method" for handling missing data is to remove observations with any missing data. This method, which Rubin calls complete case analysis, is inefficient, but it is easy to implement with standard statistical packages. Complete case analysis will lead to biased estimates if the process generating the missing data is correlated with the endogenous variables in the model. For example, suppose we estimate a mode choice model from a random sample of a metropolitan area. If respondents who never take the bus are more likely to not respond to questions about bus travel times and costs, then complete case analysis on the original random sample will yield inconsistent estimates. In this example, the missing data mechanism is equivalent to choice-based sampling.

One common solution for missing data is to impute missing values. For the mode choice example given above, this is usually done using zonal network data together with the respondent's reported home and work locations. Unfortunately, most commonly used imputation methods do not preserve the variance of the underlying variable and therefore

produce inconsistent estimates when variables containing imputations are included in models. Even when this problem is avoided, inference is always conditional on the imputed values and therefore ignores errors in the imputation process. Section 3 provides general methods for generating correct imputations

The non-random sample in the above mode choice example was caused by using complete case analysis with a non-random missing data process. Although they only discussed deliberate non-random sampling (e.g. choice-based sampling), Manski and Lerman's (1977) Weighed Exogenous Sample Maximum Likelihood Estimator (WESMLE) provides consistent and asymptotically normal estimates for any sampling procedure. Unfortunately, the WESMLE requires the sampling weights to be known constants. In many cases discussed in Sections 5 and 6 these weights need to be estimated, and it can be quite difficult to modify the WESMLE to properly account for estimation error in the sampling weights. Ignoring this estimation error will lead to downward biased standard errors.

Multiple imputation creates multiple imputed values and weights, and then combines the estimators using each set of values into a final consistent estimator that accounts for the errors in the imputation process. Although multiple imputation estimators are rarely efficient, the method can be applied using standard statistical software packages. Other maximum-likelihood methods require extensive programming and/or computation time. Section 3 reviews multiple imputation methodology, and the following sections discuss applying multiple imputation to common applied transportation problems.

To simplify the exposition, Sections 2 - 4 consider only the case of missing data. Formally, survey nonresponse is just a special case where all of the data for the missing respondents are missing. Similarly, erroneous data (or measurement error) is equivalent to having all observations on the true values missing. Sections 5 and 6 discuss survey nonresponse and panel attrition.

2. Identification with Missing Data

When data are missing, the only way to identify population statistics such as means and variances is to make assumptions that determine the distribution of the missing data. Since by definition we don't observe missing data, these identifying assumptions are typically untestable. If one is nonetheless willing to maintain these untestable identifying assumptions, then sections 3-5 of this chapter describe a general methodology for estimation and inference. This section describes work primarily due to Manski (Manski, 1995 and the included references) which shows that even though population statistics are unidentified, it is sometimes possible to bound them between estimable functions.

To illustrate the construction of these bounds, I will closely follow Horowitz and Manski (1995). Assume that each member of the population is characterized by a value of (y, x, and z), where $y \in Y$ is the outcome of interest, $x \in X$ is a covariate vector, and z is a binary variable indicating response to some or all survey questions. Suppose that we are interested in estimating conditional expectations, $E[g(y)|x]$, where $g(\cdot)$ is some specified function. If we observe y only when $z=1$, then:

$$(1) \qquad E[g(y)|x] = E[g(y)|x, z=1]\cdot P(z=1|x) + E[g(y)|x, z=0]\cdot P(z=0|x)$$

The observed sample identifies the mean $E[g(y)|x, z=1]$ and the response probability $P(z=1|x)$, but provides no information about the mean of the unobserved sample $E[g(y)|x, z=0]$. However, this unidentified part must lie between $K_0 \equiv \inf_{y\in Y} g(y)$ and $K_1 \equiv \sup_{y\in Y} g(y)$, which yields the following sharp bounds:

$$(2) \qquad E[g(y)|x, z=1]\cdot P(z=1|x) + K_0\cdot P(z=0|x) \le E[g(y)|x] \le E[g(y)|x, z=1]\cdot P(z=1|x) + K_1\cdot P(z=0|x).$$

These bounds are equivalent to imputing all the missing data equal to the highest and lowest possible values of y. Note that the length of this bound is $(K_1 - K_0)\cdot P(z=0|x)$, so if $g(\cdot)$ is bounded, then decreasing the nonresponse probability will sharpen the bound. One

important case where $g(\cdot)$ is bounded is when it is an indicator function. Horowitz and Manski (1995, Table 1) give an example trying to estimate the employment probability from the 1979 U.S. National Longitudinal Survey of Youth. The nonresponse rate is 18.4% ($P(z=0|x)$), and the employment probability in the observed sample is 78% ($E[g(y)|x, z=1]$, where y is 1 if the respondent is employed and zero otherwise). Since the probability of employment in the missing observations is bounded between 0 and 1, equation (2) shows that the desired overall employment probability lies between 63% and 82%. The length of this bound is much larger than the sampling variation of the probability estimators, and increasing the sample size with the same nonresponse probability will not decrease the width of the interval.

Of course, many researchers are willing to make identifying assumptions about the unobserved mean $E[g(y)|x, z=0]$. A popular assumption, due to Heckman (1976), has $g(y|x)$ follow a normal distribution. The most common identifying assumption is that the nonresponse process is *ignorable* (see Rubin ,1987a, for a formal definition), which here means that the conditional distribution of $y|x$ is the same for both respondents and nonrespondents. As long as no x values for the unobserved sample are outside the range of x values for the observed sample, the assumption of ignorability allows nonparametric identification of $E[g(y)|x]$. One reason why ignorability is so commonly invoked is that without additional information there is no direct evidence to contradict the ignorability assumption (Rubin ,1987a, Section 5.1).

Many other identifying assumptions are possible, but they must all yield estimates within the bounds given by equation (2). Horowitz and Manski (1995) show that weighted means using standard poststratification (or ratio) nonresponse weights (which are estimates of the inverse of the probability of response) may fall outside these bounds. Weighted estimators are only guaranteed to yield sensible estimates if the weights are constructed conditioning on x. However, imputation estimators which explicitly impute missing values and then average over observed and imputed values will always lie inside the bounds given in equation (2).

From a Bayesian perspective, Horowitz and Manski's bounds are similar to Leamer's (1983) "extreme bounds analysis." Leamer derives his bounds by looking at the envelope of the Bayesian confidence regions over all possible prior distributions. In many cases, both Leamer's and Horowitz and Manski's bounds are infinite, and even when they are bounded they are frequently quite wide.

The analysis reviewed in this section implies that we need to be much more humble about our empirical results when there are missing data. A more constructive implication is that researchers involved in collecting data should concentrate more resources on reducing missing data and non-response. Another useful approach is to carry out separate "validation" studies which use intensive interviewing techniques to survey a sample of non-respondents. These validation surveys provide direct evidence about non-respondents to the main survey, so they can be used to identify and measure the unknown $E[g(y)| x, z=0]$. Brownstone and Valletta (1996) show how multiple imputation techniques can be used to combine information from the validation and main surveys to estimate econometric models.

3. Multiple Imputation

Rubin's (1987a) multiple imputation methodology first requires a model for producing proper imputed values. These imputed values must be conditioned on all observed data, and different sets of imputed values must be drawn independently so that they reflect all sources of uncertainty in the imputation model. Although Rubin developed the theoretical properties of this methodology for Bayesian models, Chapter 4 in Rubin (1987a) and Rubin (1996) show that these results apply asymptotically to classical statistical models.

Suppose we are interested in estimating an unknown parameter vector θ. If no data are missing, then we would use the estimator $\tilde{\theta}$ and its associated covariance estimator $\tilde{\Omega}$. If we have a model for predicting the missing values conditional on all observed data, then we can use this model to make independent simulated draws for the missing data. If m independent sets of missing data are drawn and m corresponding parameter and covariance

estimators, $\tilde{\theta}_j$ and $\tilde{\Omega}_j$, are computed, then Rubin's Multiple imputation estimators are given by

(3) $$\hat{\theta} = \sum_{j=1}^{m} \tilde{\theta}_j / m$$

(4) $$\hat{\Sigma} = U + \left(1 + m^{-1}\right) B,$$

where

(5) $$B = \sum_{j=1}^{m} \left(\tilde{\theta}_j - \hat{\theta}\right)\left(\tilde{\theta}_j - \hat{\theta}\right)' / (m-1)$$

(6) $$U = \sum_{j=1}^{m} \tilde{\Omega}_j / m.$$

Note that B is an estimate of the covariance among the m parameter estimates for each independent simulated draw for the missing data, and U is an estimate of the covariance of the estimated parameters given a particular draw. B can also be interpreted as a measure of the covariance caused by the nonresponse process.

Rubin (1987a) shows that for a fixed number of draws, $m \geq 2$, $\hat{\theta}$ is a consistent estimator for θ and $\hat{\Sigma}$ is a consistent estimator of the covariance of $\hat{\theta}$. Of course B will be better estimated if the number of draws is large, and the factor $(1 + m^{-1})$ in equation (4) compensates for the effects of small m. Rubin (1987a) shows that as m gets large, then the Wald test statistic for the null hypothesis that $\theta = \theta^0$,

(7) $$\left(\theta - \theta^0\right)' \hat{\Sigma}^{-1} \left(\theta - \theta^0\right),$$

is asymptotically distributed according to an F distribution with K (the number of elements in θ) and ν degrees of freedom. The value of ν is given by:

(8) $$\nu = (m - 1)(1 + r_m^{-1})^2 \text{ and}$$
$$r_m = (1 + m^{-1}) \text{ Trace}(BU^{-1})/K .$$

This suggests increasing m until ν is large enough (e.g. 100) so that the standard asymptotic Chi-squared distribution of Wald test statistics applies. Meng and Rubin (1992) show how to perform likelihood ratio tests with multiply-imputed data. Their procedures are useful in high-dimensional problems where it may be impractical to compute and store the complete covariance matrices required for the Wald test statistic (equation 7).

The key to successful implementation of multiple imputation is to use a *proper* imputation procedure. The full definition of a proper imputation procedure is given in Rubin (1987a, pp. 118-119). Loosely speaking, if the estimates computed with the true values of the missing data ($\ddot{\theta}$ and $\ddot{\Omega}$) are treated as fixed, then $\hat{\theta}$ and U must be approximately unbiased estimators of $\ddot{\theta}$ and $\ddot{\Omega}$. In addition B must be an approximately unbiased estimator of the variation in $\hat{\theta}$ caused by the non-response mechanism. The safest way to generate proper imputation procedures is to explicitly draw from the (Bayesian) posterior predictive distribution of the missing values under a specific model. Other proper multiple imputation procedures also require no explicit Bayesian calculations, and some of these methods are described in the following sections. Any proper imputation procedure must condition on all observed data, and different sets of imputed values must be drawn independently so that they reflect all sources of uncertainty in the response process.

3.1 Linear Model

This section summarizes results from Brownstone (1991) on using multiple imputation techniques with the linear regression model. Suppose we are interested in estimating a K-vector, θ, in the standard linear model:

$$y = X\theta + \varepsilon, \tag{9}$$

where, conditional on X, the components of ε are independent and identically distributed random variables with mean 0 and variance σ^2. In the absence of missing data, θ would be estimated by the ordinary least squares estimator, $\hat{\theta}$, and inference would be based on:

$$(10) \qquad \left(\hat{\theta}-\theta\right) \approx \mathrm{N}\left(0, s^2(\mathrm{X}'\mathrm{X})^{-1}\right), \text{ where } s^2 = \mathrm{y}'\left(\mathrm{I} - \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\right)\mathrm{y}\Big/(N-K)$$

and I is the $K \times K$ identity matrix.

Suppose further that the first N_0 observations contain missing data in exogenous (X) and/or endogenous (y) variables, but that there are no missing data in the remaining $N_1 = N - N_0$ observations.

A proper multiple imputation procedure for the linear model above must produce imputed values for the missing observations, y_0^* and X_0^*, which at least match the first two asymptotic moments of the unobserved variables, y_0 and X_0. For any such set of proper imputations, the completed data least squares estimator is given by:

$$(11) \qquad \tilde{\theta} = \left[\left(\mathrm{X}_0^{*'}\mathrm{X}_0^*\right) + \left(\mathrm{X}_1^{*'}\mathrm{X}_1\right)\right]^{-1} \cdot \left[\mathrm{X}_0^{*'}\mathrm{y}_0^* + \mathrm{X}_1'\mathrm{y}_1\right]$$

and the associated covariance estimator:

$$(12) \qquad \tilde{\Omega} = s^{*2}\left[\left(\mathrm{X}_0^{*'}\mathrm{X}_0^*\right) + \left(\mathrm{X}_1^{*'}\mathrm{X}_1\right)\right]^{-1},$$

where s^{*2} is defined as in equation 10 with imputed values, y_0^* and X_0^*, replacing the unobserved variables. Since the first two asymptotic moments match, it follows that the probability limits of $\tilde{\theta}$ and $\tilde{\Omega}$ must equal the probability limits of the least squares estimator with no missing data. Note that $\tilde{\Omega}$ is a downward biased estimate of the variability of $\tilde{\theta}$ since it does not account for the variation induced by the non-response

process. Drawing repeated independent sets of proper imputations allows consistent estimation of this variance component (*B* in equation 5).

As discussed in Section 2, before we can construct a proper imputation procedure we need to make some assumptions about the process generating the missing data. To simplify the notation, we will assume that only the exogenous variables, X_0, are missing. If we further assume that the missing data process is ignorable, then the most straightforward way to generate proper imputations for the linear model is:

$$X_0^* = E(X_0 | y_0) + \eta_0^* \quad , \tag{13}$$

where η_0^* are independent draws from the distribution of the residuals, $X_0 - E(X_0 | y_0)$. The ignorability assumption implies that $E(X_0 | y_0)$ is identical to $E(X_1 | y_1)$, and can therefore be consistently estimated using parametric or nonparametric regression techniques (see Manski, 1991) on the observed data. Ignorability also implies that the distribution of the residuals can also be estimated from the distribution of residuals on the observed data.

If we further assume that the conditional distribution of X given y satisfies the standard linear model assumptions, then the imputation equation (13) can be implemented by first regressing the observed X_1 on y_1 to get the estimated normal sampling distributions of the slope parameter(s) and the residual variance. To draw one set of imputed values, first draw one set of slope and residual variance parameters and then draw the imputed residual vector, η_0^*, from independent normal distributions with mean zero and variance equal to the imputed residual variance parameter. The imputed values are then computed by adding this imputed residual to the predicted value from the regression using the imputed slope parameters. Additional sets of imputed values are drawn the same way beginning with independent draws of the slope and residual variance parameters. This imputation method, which Schenker and Welsh (1988) call the "normal imputation" procedure, is equivalent to

drawing from the Bayesian predictive posterior distribution when X and y follow a joint normal distribution with standard uninformative priors.

The assumption of joint normality in the previous paragraph is convenient but not necessary for producing proper imputations in this model. Nonparametric regression estimators can be used to estimate nonlinear conditional mean functions, $E(X_1 \mid y_1)$, and bootstrapping methods can be used to sample from residual distributions without assuming a normal distribution. However, the assumption of ignorability is crucial for the validity of these methods.

The simple model analyzed above is not very interesting from a practical perspective since the resulting multiple imputation estimator will have approximately the same distribution as the least squares estimator calculated from the observed data, X_1 and y_1. However, if there are additional fully observed variables, then these can be added to the conditioning set in equation (13) to yield improved estimators. The next section describes some circumstances where these additional variables may be readily available to the data collectors. Conditioning on additional variables, even if they are not directly related to the model in equation (9), can also make the crucial ignorability assumption more palatable.

3.2 Public Use Datasets

According to its developer (Rubin, 1996, page 473), "multiple imputation was designed to handle the problem of missing data in public-use databases where the database constructor and the ultimate user are distinct entities". The database constructor certainly has more information about the sampling design and surveying process than the ultimate user, and she may have access to confidential information (such as exact addresses) which cannot be released in a public-use file. The users of these data are assumed to have access to standard statistical packages which typically estimate a wide variety of models assuming random sampling and no missing data.

Multiple imputation methodology allows the database constructor to present her superior knowledge about the nonresponse process so that the end user can easily incorporate this

knowledge in his analysis. The database constructor produces 5 - 10 multiple imputed values for key variables (such as income) with substantial missing data. The end user only needs to repeat a standard complete-data analysis for each set of imputed values and then combine the results according to equations (3) - (6). These additional calculations can, and have, been incorporated as macros in standard statistical packages. The only other practical problem may be the extra storage required for the multiply imputed values, but this seems to be an increasingly small price to pay for statistically valid inference. Finally, Rubin (1996, section 4) points out that there are no general alternatives to multiple imputation for public-use files.

In spite of multiple imputation's obvious advantages for public-use files, there has been considerable controversy about this use of multiple imputation methodology. Fay (1991, 1992) provides several examples where the multiple imputation variance estimator, $\hat{\Sigma}$ in equation (4), is an inconsistent estimator of the sampling variance of the multiple imputation estimator, $\hat{\theta}$, even when the imputation model is correctly specified. Fay's examples do not violate Rubin's (1987a) results because the end-user's model conditions on an irrelevant variable not included in the imputer's model. Meng (1994) shows that this situation, which he calls "uncongeniality," typically leads to conservative inference (actual coverage probabilities higher than nominal coverage probabilities) from the standard multiple imputation formulas. In spite of this conservative bias, Meng shows that the multiple imputation intervals are still sharper than the "standard" inference from only non-missing observations. A simple partial solution to these problems is to make sure that the imputation models and assumptions used by the database constructor are fully documented.

Fay also criticizes practitioners of multiple imputation for not properly accounting for the highly stratified multistage sampling techniques used in many large-scale surveys. This is not a criticism of multiple imputation methodology, which assumes that the complete data covariance estimator, $\tilde{\Omega}$, is consistent. Rubin (1996) points out that multiple imputation can easily be combined with modern jackknife and bootstrap techniques for estimating variances in complex samples. Fay's criticism unfortunately applies to practically all end-

users of complex public-use databases, and the biases in estimated standard errors and inferences caused by ignoring complex sampling schemes may be larger than those caused by ignoring missing data.

The case for using multiple imputation in public-use databases is especially strong when the database creator has access to validation or non-response studies. These studies allow identification and imputation with nonignorable nonresponse. In these cases inference from non-missing observations will yield inconsistent point estimates in addition to inconsistent variance estimates. Unless these validation studies are also publicly available, the imputations will be conditioned on more information than can be made available to end users. Providing multiply imputed public-use databases is currently the only general method for giving the end-user the access to the database constructor's crucial confidential information about the nonresponse processes.

There are many opportunities for constructors of public-use transportation databases to obtain information about nonresponse (or erroneous response) mechanisms. Most large transportation surveys ask questions about household's vehicle holdings and vehicle miles traveled. Household reports of vehicle holdings can be checked against vehicle registration files. Household reports of vehicle miles traveled are known to be highly inaccurate (Lave, 1996), but in many states annual vehicle miles traveled can be checked against vehicle inspection records. Although there are substantial legal problems in many countries, household and personal income reports can sometimes be checked against tax records. Collecting this information can be difficult and time-consuming, but the examples described in the next section show large gains in statistical accuracy with relatively small validation study samples.

Modern activity analysis also depends on large-scale travel diary surveys. These surveys are difficult and expensive to collect, and even the best self-reported diaries contain substantial measurement error. New technology using on-vehicle computers equipped with global positioning system satellite receivers can almost totally automate collection of automobile trip details. While this technology is too expensive to deploy for a large-scale

activity survey, recent work with electric vehicle trials show that it is feasible to use these new accurate data collection techniques for a small validation sample (see Golob, Swertnik, *et. al.* 1996).

4. Missing or Erroneous Data

This section describes some studies where the multiple imputation methodology described in the previous section is used to compensate for missing and/or erroneous data. The studies are chosen to represent the wide range of applications that are possible, but this is not intended to be a review of all known applications of multiple imputation. Rubin (1996) and Meng (1994) provide more complete references to applications of multiple imputation. Heitjan and Little (1991) also give an interesting example applying multiple imputation to fill in missing blood alcohol content measures in the U.S. Fatal Accident Reporting System. Glynn *et. al.* (1993) give another application to missing data on alcohol drinking behavior in a study on the effects of drinking on retirement. They use a follow-up validation study to estimate a non-ignorable response model. Brownstone and Valletta (1996) use similar methodology in their study described in the next sub-section.

4.1 Linear Models

Wages (or labor earnings) are a key variable in labor economics, and they are also important in many transportation models. Wages and income are typically measured by directly asking survey respondents, but there is considerable evidence that these reports contain large errors. In particular, both Bound *et. al.* (1990, 1994) and Bound and Krueger (1991) find that measurement error in earnings is negatively correlated with true earnings and positively autocorrelated over time. These findings are based on two validation surveys: the Panel Study of Income Dynamics Validation Survey (PSIDVS) and the 1978 Current Population Survey-Social Security Earnings Records Exact Match File.

Brownstone and Valletta (1996) use the PSIDVS together with the main Panel Study of Income Dynamics (PSID) to examine the effects of wage measurement error on standard

linear models used to explain and predict wages. They use the PSIDVS to estimate an imputation model which is then used to multiply impute "true" wages in the main PSID. Accounting for measurement error in a PSID sample of approximately 2000 household heads for the survey years 1983 and 1987 increases the estimated return to general labor market experience, reduces the negative effect of blue-collar status, reduces the return to union status, and may affect the returns to tenure and other variables, in both a cross-section and longitudinal setting.

Since the PSID is a large panel study that has been running continuously since 1968, it was infeasible to carry out the validation study on a subsample of the PSID respondents. Therefore the PSIDVS consists of approximately 400 employees surveyed from a large Detroit, Michigan area manufacturing firm. An initial set of 534 interviews was attempted in 1983, of which 418 were completed. Reinterviews were successfully conducted with 341 individuals in 1987, of whom 275 were respondents in both 1983 and 1987. An additional sample of 151 hourly workers was interviewed in 1987. The resulting data set matches standard PSID survey responses with company personnel records on a variety of employment variables, including earnings, fringe benefits, hours, unemployment spells, and employment tenure. The company records are highly accurate and were treated as error-free.

The basic modeling assumption is that the true value of the logarithm of annual labor earnings, y^*, and the reported value, y, follow a bivariate normal distribution conditional on exogenous variables X and Z. Because y^* is unobserved in the main sample but is observed in the validation sample, while y is available in both the main and validation samples, it is convenient to write the model as:

(14a) $$y^* = X\beta + \varepsilon,\ \varepsilon \sim N(0,\ \sigma_\varepsilon^2 I)$$

(14b) $$y^* = y\gamma_0 + Z\gamma_1 + \eta,\ \eta \sim N(0,\ \sigma_\eta^2 I)\ ,$$

where X has N rows and K columns and Z also has N rows. The first equation (14a) represents the conditional distribution of y^* given X, which is the relationship we want to estimate. The second equation represents the conditional distribution of y^* given y and Z, and this is used to impute values for true earnings (y^*) in the main sample.

Rubin (1987a, pp. 81-87) shows that the validation study data must be pooled with the main study data to estimate equation (14a), since otherwise the completed data estimators are not conditioned on all observed data. Similarly, Z must contain all variables in X for equation (14b) to generate proper multiple imputations. To help satisfy these constraints, Brownstone and Valletta used a subsample of the main PSID designed to more closely match the PSIDVS sample, and they also included additional variables in X and Z to facilitate pooling. The PSID subsample includes 2504 male and female household heads for 1983. Ideally, the validation sample is a probability sample from the main study, in which case pooling is not an issue.

The values of γ and σ_η^2 in equation (14b) can be estimated by $\hat{\gamma}$ and s_η^2, the least squares estimates from regressing y^* on y and Z using the validation study. Therefore, $\hat{\gamma}$ follows a $N(\gamma, \sigma_\eta^2\Delta^{-1})$ distribution (where $\Delta = [y\ Z]'[y\ Z]$) and ds_η^2/σ_η^2 follows an independent Chi-squared distribution with degrees of freedom d equal to the number of observations in the validation data minus the number of columns in [y Z]. Brownstone and Valletta use the following steps to create one set of valid imputations for the model in equations (14) according to equation (13):

(15)

a) set $\sigma_\eta^{2*} = s_\eta^2\chi^*/d$, where χ^* is drawn from a χ_d^2 distribution.

b) draw γ^* from a $N(\hat{\gamma}, \sigma_\eta^{2*}\Delta^{-1})$ distribution.

c) draw η^* independently from a $N(0, \sigma_\eta^{2*})$ distribution.

d) set $\hat{y}^* = y\gamma_0^* + Z\gamma_1^* + \eta^*$.

Table 1 reproduces a subset of Brownstone and Valletta's results. The first column gives the uncorrected results from regressing the reported values on the X variables. In the corrected regressions reported in the second column, the log of reported earnings for each observation in the main PSID sample is replaced by multiple imputed values of the log of true earnings, which are obtained through repeated application of algorithm (15) above.

The results in Table 1 show substantial changes from correcting for measurement error. Error correction increases the coefficient on potential experience and the absolute value of the coefficient on its square by about 33%. The negative effect of blue-collar status on earnings is almost halved due to error correction. The return to union status in 1983 is substantially reduced and becomes insignificant, the coefficients on tenure and tenure squared are reduced by 20% and 33%, and the negative effects of being black are substantially reduced. Finally, although the magnitude and statistical significance for most coefficients is reduced by error correction, this shrinkage is not uniform: the coefficients on blue-collar and black status remain relatively constant across the corrected and uncorrected specifications.

Table 1 - Partial Earnings Models Using Pooled 1983 PSID and PSIDVS Sample

dependent variable = ln(interview earnings)
(coefficients; standard errors in parentheses)

	Uncorrected	Corrected
Schooling (in years, topcoded at 17)	.0646*	.0584*
	(.0067)	(.0108)
Potential	.0172*	.0230*
Experience (= age-schooling-6)	(.0047)	(.0061)
(Potential	-.0320*	-.0413*
Experience)^2/100	(.0092)	(.0121)
Company tenure	.0379*	.0300*
	(.0043)	(.0071)

(Company tenure)^2/100	-.0666* (.0129)	-.0452* (.0167)
Black	-.114* (.052)	-.0707 (.0496)
Blue-collar	-.166* (.032)	-.0975* (.0366)
Union coverage	.184* (.029)	.122 (.090)
Number of Observations	2848	2848

* - indicates significance at 5% level, two-tailed test

Brownstone and Valletta also estimate a longitudinal fixed-effects wage equation using the 1983 and 1987 matched observations in the PSIDVS to correct for measurement error. As in the 1983 cross-section, error correction increases the size and significance of the coefficient on potential experience but decreases the negative effects of blue-collar and union status, including their changes. The measurement error process is highly autocorrelated, so the differencing required to remove the fixed effects does not significantly increase the relative measurement error variance from the cross-section case.

In addition to Brownstone and Valletta's multiple imputation methodology, there are a number of alternative approaches for estimating the model in equations (14). The simplest of these is complete case analysis, which is to estimate the earnings equation using the validation sample only. The simplicity gains of this alternative approach must be weighed against two main benefits of the multiple imputation approach: (1) combining information from both samples increases estimation efficiency; (2) the ability to account for possible differences across the two samples in the parameters of the earnings equation. In Brownstone and Valletta's longitudinal equation the standard errors of the multiple imputation parameter estimates are approximately 25%-50% lower than estimators just using the validation sample. Since validation samples are typically expensive to collect,

their sample sizes are typically an order of magnitude smaller than the surveys they are validating.

Another approach which yields asymptotically efficient estimates is proposed by Lee and Sepanski (1992). They propose (but do not implement) an instrumental variables method of moments estimators to combine validated and unvalidated data. Their techniques require more programming than multiple imputation and cannot easily be generalized to more complicated models with discrete and limited variables. The next subsection shows how multiple imputation can be used with these models.

4.2 Discrete Choice Models

Although there are a number of feasible methods for estimating the linear models discussed in the previous subsection, these more traditional methods become very difficult to implement with discrete choice models. Multiple imputation, on the other hand, can easily be used in these cases, as shown by Clogg *et. al.* (1991). They consider the problem of changing industry and occupation codes between the 1970 and 1980 U.S. Census. These coding changes were so dramatic that it is impossible even to compare major occupation groups across the coding schemes. As a result, it would be very difficult to use Census data to track changes in occupations across the 1970s. Of course, similar problems occur for transportation researchers when the census tract (or traffic analysis zone) boundaries are changed for each decennial census.

To alleviate problems caused by the changed coding system, the U.S. Census bureau randomly sampled about 125,000 individuals from the 1970 Census with known 1970 codes and recoded them using the new 1980 coding system. Clogg *et. al.* (1991) used this double-coded sample to create five multiply imputed 1980 occupation codes for each respondent in the 1970 Census Public Use Sample. Researchers studying occupational change could then estimate five models using these multiply imputed 1980 codes and then use the multiple imputation combining rules in equations 3 - 8 to carry out their analyses. If

the five multiply imputed codes are all different for most observations, then it will obviously be very difficult to estimate any reasonable change models. Conversely, if the five multiply imputed codes are identical, then the coding system change will not affect the analyses.

The hard part is modeling the variability in the double-coded sample so that proper multiple imputations can be produced. Clogg *et. al.* (1991) use separate Bayesian multinomial logit models for each of the approximately 500 three-digit 1970 occupational codes. The dependent variables in these models are the 1980 codes assigned to each respondent in the double-coded sample with the same 1970 occupational code. The independent variables include sex, race, age, education, employment status, and region. Due to the small sample sizes and skewness of the dependent variables for many 1970 occupational codes, maximum likelihood estimation of these models was infeasible. Instead Clogg *et. al.* (1991) use an empirical Bayes procedure that shrinks the coefficients of the covariates towards a model with just alternative-specific constants.

Since they use a full Bayesian model, Clogg *et. al.* (1991) sample from the posterior predictive distribution to draw the multiple imputations for the 800,000 Public Use Sample respondents. Recent advances in Bayesian computations have made direct sampling from these posterior distributions feasible for very complex models. Clogg *et. al.* (1991) use Rubin's sampling-importance resampling algorithm (Rubin, 1987b). Clogg *et. al.* (1991) report extensive Monte Carlo and other evaluations of this methodology which show that the imputation method and the resulting multiple imputation inferences are valid for this application.

This application highlights many of the strengths of the multiple imputation methodology. Even if end-users had access to the double-coded sample, it is very unlikely that they would have the statistical and computational resources to carry out the modeling and imputations performed by Clogg *et. al.* (1991). The multiple imputation combining equations (3 - 6) only require the end-user to perform some simple matrix computations to take advantage of all of this modeling and data collection effort. Also, the presence of five

different 1980 occupation codes makes it very difficult for even the most casual end-user to ignore the errors in the recoding process. The current practice of filling in missing data in public use files by a single imputation makes it easy for end-users to ignore any potential problems with the missing data process. In addition, standard end-use analysis of singly-imputed data always underestimates parameter standard errors.

Brownstone and Golob (1992) use multiple imputation to compensate for missing data in a key variable explaining commute mode choice in Southern California. The main commute modes in this region are drive alone (about 80%) and carpool. The number of employees at the respondents work site is an important mode-choice predictor. It is easier to find suitable carpool partners from a larger set of employees, so larger worksites should have more car-poolers. Unfortunately Brownstone and Golob's unusual sampling design caused employer size information to be missing for 30% of the sample, and the remainder of the sample were employed in a small geographic area near the city of Irvine. This design makes complete case analysis using only the respondents with employer-size information very inefficient.

Since Brownstone and Golob's mode choice model use only a dummy variable for whether the employer had more than 200 employees at the respondent's worksite, they use a binomial probit model calibrated on the subsample where employer size is observed. The dependent variable for this imputation model is equal to one if the respondent's employer size is greater than 200, and zero otherwise. Important explanatory variables include respondent age, carpool status, household size and income, commuting distance, and the presence of various employer-provided carpooling incentives. These covariates improved the prediction success rate of the model from 56% to 78%. Brownstone and Golob's procedure assumes that the non-response process is ignorable, which is justified since the non-response is due to sample design features.

Drawing proper multiple imputations from this probit model requires accounting for two sources of errors: the error in the parameter estimates and the error in predicting the actual value of the dummy variable from the estimated probit probability. Brownstone and Golob

created their imputations by first sampling from the asymptotic multivariate normal approximation to the sampling distribution of the maximum likelihood probit estimators. Conditional on a single draw from this distribution, they compute the probit probabilities for each respondent with missing data. The imputed values were then drawn from a binomial distribution with probability of success given by the probit probabilities. The next set of multiple imputations then begin with a new independent draw from the multivariate normal sampling distribution.

The specification of Brownstone and Golob's imputation equation might be criticized by econometricians unfamiliar with multiple imputation methodology since the dependent variable from main model, mode choice, is being used to help impute an "exogenous" variable in the main model. In fact, the imputation model must condition on all observed data, including the "endogenous" ones, for the multiple imputation method to yield consistent estimates. Inference for the main model is conditioned on the true values of the missing data, so the completed data estimators need to be consistent for this conditional model. The econometrician's "simultaneous equations bias" occurs when inference is not conditioned on the true values of the missing data.

4.3 Panel Data

Panel studies, where a sample is followed over time with repeated interviews, are becoming increasingly important in transportation demand analysis (Golob and Kitamura, 1996). Even if the missing data process is completely random, the compounding of a 20% nonresponse rate (which is common for income measures) over only three interviews can leave very few respondents with complete data. In addition, small amounts of completely random measurement error can cause serious problems when using common panel data models which are based on differences between panel waves.

These problems are balanced by the fact that since transportation decisions and most human behavior exhibit inertia, observed data for the panel respondents provide excellent information about the missing data. For example, attempts to impute missing income data

in cross-section studies using either standard methods or multiple imputation usually do not produce substantial efficiency gains over complete case analysis. The reason for this is that models for predicting income in cross-sections typically have very poor fit, with R^2 measures below .2. However, if previous or future income values are available from other waves of a panel study, then the fit of the imputation models greatly improves. In an unpublished study using the Swedish HUS panel, Brownstone's (1990) models for imputing missing income values had R^2s around .7. The key to this improved fit is that previous and/or future income values for the same household are typically available in panel studies. This suggests that multiple imputation of missing values may yield substantial efficiency gains in panel studies.

The same inertia governing observed human behavior might also characterize measurement error processes in panel data. Brownstone and Valletta (1996) found that measurement error in annual earnings reports were highly autocorrelated. They found that this autocorrelation was high enough so that fixed effects models were no more affected by measurement error than cross section models. Of course, the methods used by Brownstone and Valletta crucially depend on the existence of a validation study, and these studies are currently very rare.

Multiple imputation techniques are well-suited to panel data since they require very little additional computation and programming beyond that required for a complete data analysis. Transportation models using panel data, such as dynamic models of mode choice and dynamic vehicle type choice models, are based on complex dynamic discrete econometric models. Even without considering missing data and attrition problems, these models are very difficult to estimate - frequently requiring simulation estimators (McFadden and Ruud, 1994). Although joint full-information maximum likelihood estimation of linear models with missing data is feasible (see Fuller, 1987), these methods are not computationally feasible with many panel data applications.

5. Panel Attrition and Choice-Based Sampling

Panel studies are often plagued by the attrition of survey respondents. Attrition can bias the sample and limit the usefulness of the panel for long-term dynamic analysis. If the attrition process is correlated with the endogenous variables in the model (called *non-ignorable attrition*), then standard estimation techniques ignoring attrition will yield inconsistent inferences and estimators. Even if the attrition process is independent of the endogenous variables (called *ignorable attrition*), uncorrected attrition may bias forecasts and policy simulations based on the remaining sample. Both of these problems occur in transportation panels. If survey questions are concentrated on the adoption of a new mode or technology, then users of this new mode or technology may be less likely to attrite and the attrition process will be correlated with the endogenous choice variable. Since the main purpose of many analyses of transportation panels is to produce forecasts of the effects of proposed policy changes, it is important to account for the effects of attrition on model forecasts and policy simulations.

This section, which is closely based on Brownstone and Chu (1996), describes a multiple imputation methodology for obtaining consistent estimates and forecasts from panel models where non-ignorable attrition is present. If the non-attrition probabilities are known, then their inverses can be used as weights in Manski and Lerman's (1977) Weighted Exogenous Maximum Likelihood Estimator (WESMLE). These weights make the weighted sample look like the initial panel wave.

Manski and Lerman (1977) show that a simple modification of the standard Maximum Likelihood estimator for discrete-choice models yields consistent and asymptotically normal parameter estimates in the presence of choice-based sampling when the proportion of the population choosing each discrete alternative is known. If $L_i(\theta, x_i)$ is the log likelihood function for the i^{th} observation, then Manski and Lerman's WESMLE maximizes:

$$(16) \qquad \sum_i \omega_i L_i(\theta, x_i),$$

where θ is a vector of parameters to be estimated, x_i is the vector of observed characteristics for the i^{th} observation, and the sampling weight, ω_i, is the inverse of the probability that the i^{th} observation (individual) would be chosen from a completely random sample of the population. Of course, if the sampling scheme were completely random, then all of the sampling weights would be equal and the WESMLE would simply be the usual maximum likelihood estimator. Manski and Lerman (1977) show that the WESMLE is consistent and asymptotically normal, but not fully efficient (see Imbens, 1992 for fully efficient alternative estimators). Manski and Lerman's proof actually shows that the WESMLE's properties hold for any regular maximum likelihood estimator as long as the sampling weights are known with certainty.

A major advantage of the WESMLE is that it can be computed very easily by modifying existing maximum likelihood programs. The WESMLE for both the linear regression model and the multinomial logit model can be computed by appropriately weighting the variables and applying standard maximum likelihood programs. Unfortunately, this procedure yields downward biased standard error estimates, but the consistent estimates given by Manski and Lerman are easy to compute (see DuMouchel and Duncan, 1983, for a similar analysis of the linear model). This downward bias can be substantial in common applications. Brownstone and Valletta (1996) find a 30% downward bias in their weighted regressions from using the incorrect weighted regression covariance estimator.

A panel survey can always be viewed as the result of the original sampling process and the attrition process. Although in a well-designed panel study the properties of the sampling process are known with certainty, the properties of the attrition process are typically unknown. If they *were* known, then the sampling weights could be easily computed as the inverse of the product of the sampling and attrition probabilities and the WESMLE could be applied to get consistent parameter estimates. Fortunately, there is at least one wave of information about panel attriters, and with some modeling assumptions this information can be used to estimate a model of the attrition process. Unfortunately, the resulting predicted attrition probabilities cannot be used to generate weights for the WESMLE, since this would violate the assumption that the weights are known with certainty.

Suppose we have a procedure for making independent simulated draws from the sampling distribution of the attrition probabilities (which are given from our estimated attrition model). Conditional on this set of simulated attrition probabilities, we can compute a vector of sampling weights (as the inverse of the product of the attrition probabilities and the sampling probabilities for the first wave of the panel). This weight vector can in turn be used to get a consistent (conditional on that particular set of weights) estimate of θ and its covariance using the WESMLE. After drawing a number of independent attrition probabilities, equations (3-6) can be used to combine the resulting WESMLE estimators for final inference. This procedure appears to have been first proposed in Brownstone (1991), but it is a simple modification of Rubin (1986).

If the attrition model is correctly specified, then the resulting multiple imputation estimators, $\hat{\theta}$ and $\hat{\Sigma}$, are consistent whether the attrition process is ignorable or not. The standard unweighted maximum likelihood estimators, $\overline{\theta}$ and $\overline{\Sigma}$, which ignore the sampling and attrition weights, are efficient if both the sampling and attrition processes are ignorable, but inconsistent otherwise. Therefore the statistic:

$$(17) \qquad T = \left(\hat{\theta} - \overline{\theta}\right)' \left(\hat{\Sigma} - \overline{\Sigma}\right)^{-1} \left(\hat{\theta} - \overline{\theta}\right),$$

is a valid Hausman (1978) test statistic for the null hypothesis that both the sampling and attrition processes are ignorable. Under the null hypothesis, T has a chi-squared distribution with degrees of freedom equal to the rank of $\left(\hat{\Sigma} - \overline{\Sigma}\right)$.

Relative to joint maximum likelihood estimation of the attrition and choice model, the methodology described above is inefficient. However, this methodology is much easier to calculate than joint maximum likelihood, which is frequently intractable in complex models. Simple Hausman (1978) tests can be applied to test for the non-ignorability of the attrition (or missing data) process. Since the WESMLE was originally designed to

provide consistent estimates with choice (or response)-based sampling designs, the methodology proposed here can be trivially modified to yield consistent estimates and forecasts for choice-based panels with non-ignorable attrition.

Brownstone and Chu (1996) apply the above multiple imputation methodology using a dynamic commute mode choice model calibrated from the University of California Transportation Center's Southern California Transportation Panel. The study region and survey methodology are more fully described in Uhlaner and Kim (1993). The panel was selected from respondents to a mail survey, and was initiated in February 1990. The first wave of data were drawn from the original sample and from a refreshment sample introduced three months later. The overall response rate for the first-wave mail survey was approximately 50%. The total sample size for the first wave was 2,189 commuters (approximately 1,850 had complete data). Almost all respondents were employed full-time. The fifth wave of the panel was collected beginning in July 1991. The attrition rate (from Wave 1) was 40%, leaving 1,107 respondents whose data were suitable for dynamic analysis.

Since Brownstone and Golob (1992) found that the Wave 1 nonresponse process is ignorable, Brownstone and Chu just model the attrition process. Table 2 reproduces Brownstone and Chu's binomial logit attrition model results for attrition between Waves 1 and 5 of the panel. The model is specified so that positive coefficients favor attrition. Since at least some of the coefficients on the mode choice variables and their interactions are significantly different from zero, the attrition process is not ignorable. The many interactions between mode choice and the demographic variables show the complexity of the process. These results imply that white, middle-aged homeowners with an annual household income of less than $75,000, more education, and more than three vehicles are less likely to attrite from the panel. Those respondents who receive the survey at their work sites (and presumably fill it out during their normal working hours) are also less likely to attrite.

The significant coefficients on the mode choice variables in Table 2 suggest that non-ignorable attrition is a problem for this application. However, the Hausman test given in equation (16) does not reject the null hypothesis that the attrition process is ignorable for Brownstone and Chu's dynamic mode choice model. Their model is quite large, so it is easier to examine the effects of correcting for non-ignorable attrition in a policy experiment simulated from their model. Table 3 shows the results of giving all commuters

Table 2. Binomial logit attrition model

Dependent Variable	Count	Percent
In Both Waves	1107	59.97
Attrited (dropped out before Wave 5)	739	40.03
Independent Variables*	**Estimated Coefficient**	**t-Statistic**
Annual household income<=$75,000	-0.20233	-1.81388
High school graduate	-0.90640	-2.08486
Some college, but no degree	-1.03234	-2.48198
College degree, including graduate	-0.96502	-2.30214
Older than 24 and younger than 35	-0.40301	-1.95426
Older than 34 and younger than 45	-0.31492	-1.52823
Older than 44 and younger than 55	-0.46445	-2.08844
Older than 54 and younger than 65	-0.47694	-1.80652
Production/manufacturing	0.85561	3.86404
Sales	0.61101	2.83996
Other occupation	0.60538	2.30767
Survey received at work site	-0.25986	-2.37420
Always lived in Southern Ca.	0.29458	2.77893
Considered moving next year	0.29522	2.52530
Non-white	0.47706	3.42080
Arrived at work between 7:00 and 9:00	-0.13672	-1.17095
Years lived at present address (years)	-.01998	-2.20902
Reserved parking for rideshare	0.30232	2.54872
Household owned vehicles<=3	0.33727	2.00087
Home owner	-0.17925	-1.50927
Always rideshare in last two weeks	-0.69192	-1.71807
Always rideshare and household income<=$75,000	0.57605	1.44266
Always rideshare and moving next year	0.75718	1.80477
Always rideshare and having kids under 16	0.43372	1.20571
Sometime rideshare in last two weeks	1.12516	2.52191
Sometime rideshare and college degree	-0.73210	-2.57659
Sometime rideshare and age>24and<35	-0.74236	-2.53687
Sometime rideshare and household vehicles<=3	-0.73703	-1.74108
Sometime rideshare and having kids under 16	0.49162	1.83682
Constant	0.65782	1.35897

Auxiliary statistics	At Convergence	Initial
Log likelihood	-1164	-1279.5
Number of observations	1846	
Percent correctly predicted	64.6	

*All variables defined as dummies except for years lived at present address.

in the sample an employer-paid guaranteed ride home in emergencies. As expected, there is an increase in the number of commuters remaining or switching to ridesharing and a corresponding decrease in drive-alone commuting. However, these results are not significantly affected by correcting for non-ignorable attrition.

Table 3. Estimated effects from giving everyone access to a guaranteed ride home

	Ignoring Attrition		Multiply Imputed WESMLE	
	% Change	Std. Error	% Change	Std. Error
RS[1] → RS	21.43712	8.08602	18.21683	8.79526
DA[2] → RS	68.47014	12.81228	70.49838	14.19468
RS → DA	-37.61644	6.00235	-44.80742	6.53906
DA → DA	-16.15972	5.80547	-14.19958	6.56034

[1] "RS" means 'rideshare at least once in last 2 weeks'.
[2] "DA" means 'always drive alone'.

Non-ignorable attrition did not turn out to cause serious biases in Brownstone and Chu's application, but there is no reason to believe that this will be true in other transportation applications. The multiply imputed WESMLE estimator described in this section provides a simple way of testing and correcting for biases caused by non-ignorable attrition.

6. Non-random Survey Response and Synthetic Sampling

Transportation analysts use surveys for two purposes: calibrating behavioral models and to provide a representative sample for microsimulation forecasts. Modern transportation surveys place heavy burdens on respondents. They frequently require multiple telephone contacts as well as detailed travel and activity diaries. Unless the survey response mechanism is completely random, the resulting sample will not be representative. If the response mechanism is independent of the endogenous variables, then the response process is *ignorable* and behavioral models can be efficiently estimated by standard unweighted maximum likelihood techniques. However, even if the response process is ignorable for

estimation purposes, the sample will probably still need to be "adjusted" to make it representative for forecasting purposes.

This section briefly reviews current methods for reweighting survey samples for forecasting and estimation. All of these methods ignore any estimation error in creating these weights, and I will indicate how this can be remedied using multiple imputation techniques similar to those discussed in Section 5. A more important problem is that it is impossible to make any progress on these problems without external data and/or untestable assumptions (see Section 2). Fortunately, in many cases external data from censuses and administrative records such as vehicle registration files and tax records are useful for the methods described here. No method can test or correct for non-ignorable response caused by variables that aren't measured in external data.

There are frequently good external data for commute mode choice available from transit meter counts and highway traffic counters. These data have been used by transportation demand analysts to implement choice-based sampling where the survey sample is stratified by the mode choice variable. Choice-based sampling is clearly a non-ignorable sampling scheme for calibrating mode choice models. However, choice-based samples are much cheaper to collect when some modes have very small shares. Manski and Lerman (1977) show that the WESMLE yields consistent estimators when the sample strata are weighted to match the known mode shares.

When external data are available, then some procedure is needed for estimating the sampling weight for each sample respondent. This sampling weight model can then be used to multiply impute sampling weights, and the WESMLE can be used as in Section 5 to get consistent inferences. The Hasuman test in equation (17) can also be used to test the null hypothesis that the response process is ignorable. Of course, there is no point implementing any of this unless the external data contain information on endogenous variables for the models under consideration.

There are a variety of statistical matching (or poststratification) methods available for estimating sampling weights to match an external reference sample. The simplest is to discretize all variables common to both the survey and external datasets and use the discrete values to define "bins." Respondents assigned to the same bin are assumed to be identical, and the population in the bin is estimated by the sum of the weights of all of the external sample respondents assigned to the bin. The estimated weight for the survey sample is estimated by the ratio of this estimated population and the number of survey respondents assigned to the particular bin. Brownstone and Golob (1992) used this method to match their commute-choice sample to the U.S. Census Bureau's Current Population Survey (CPS). Although the CPS contains excellent demographic and labor market data, it has no information on vehicle holdings or travel time. Therefore Brownstone and Golob were only able to test for non-ignorability to the extent that the response process was correlated with demographic and labor data.

As long as the response process only depends on exogenous variables which are included in the behavioral model, standard unweighted maximum likelihood techniques will yield efficient estimators and hypothesis tests. Of course, any stratification variables included in the sample design should normally be included in the model to control for design effects. Even if the sample doesn't need to be weighted for estimation purposes, consistent forecasts require weights unless the response and sampling processes are totally random (independent of all endogenous and exogenous variables). Reweighting is also needed when forecasts are needed for populations (or geographic regions) which are different from the sampled populations. For example, behavioral models may be estimated based on a regional or national sample, but separate forecasts are required for smaller urban regions.

As long as the survey sample has at least one respondent in each bin, the simple matching procedure used in Brownstone and Golob (1992) will produce weights which make the weighted survey sample exactly match the joint distribution of the discrete match variables from the external sample. The main source of error in this procedure is the estimation error in the target joint distribution. Using a standard multinomial model for this joint

distribution it is straightforward to multiply impute estimates of these joint distributions, and these imputed distributions can be used to generate multiply imputed weight vectors for the survey sample. In many cases the external sample is very large, in which case the estimation error in the weights is likely to be small enough to ignore.

The simple bin matching procedure has a number of drawbacks in some applications. If there are many matching variables and/or many discrete values, then the size of the joint distribution can become very large and some of the estimated cells (or bins) will be empty. These empty cells cause no problems if the true population count for the cell is zero, but this is rarely the case in practice. Rubin (1986) proposes a predictive mean matching algorithm which is also used by Heitjan and Little (1991). This uses multivariate regression models to predict match variables for both the main and external sample. These predicted variables are then combined to a univariate distance measure, and matches are randomly chosen from the closest respondents using this metric. Additional matching and weighting methods are discussed in Little (1988) and Imbens and Hellerstein (1995).

If the target population is a small geographic area such as a traffic analysis zone or census tract, then data privacy restrictions preclude direct access to census observations. This problem arises in generating synthetic samples for microsimulation models of detailed travel behavior (Kitamura, 1996). In this situation the external data typically consist of univariate (occasionally bivariate) marginal distributions of census variables for the small region as well as a public use sample for a larger region with a population of at least a few million and containing the smaller region. Beckman, Baggerly and McKay (1996) give a procedure to use iterative proportional fitting to estimate the joint distributions for the small regions given this external data. If their method is applied to all subregions comprising the larger region, then the estimated joint distributions for the subregions will aggregate up to exactly match the joint distribution for the larger region.

Little and Wu (1991) give the sampling distribution of the iterative proportional fitting estimators used by Beckman, *et. al.* (1996), and Gange (1995) shows how to draw values from these sampling distributions. These results can be used to implement a multiple

imputation method to account for the estimation errors in these matching processes by drawing multiple estimated joint distributions. Little (1988) shows how other matching algorithms can be modified to yield approximately proper multiple weights.

While the methods discussed in this section allow researchers to make the most use from expensive transportation survey data, they are no substitute for obtaining a well-designed survey sample to begin with. For example, if the sample only contains a small number (relative to the population proportion) of respondents corresponding to a particular minority group, then the weighting schemes discussed here will yield very high weight values for these minority respondents. This means that weighted estimates and forecasts will depend very heavily on these few respondents, especially if the results are stratified by minority group status. Researchers should therefore always examine the weights resulting from these reweighting procedures. Having a few sample members in a particular "post-strata" is only slightly better than having no observations. In either case, the only practical solution is to further aggregate the match variables.

7. Conclusion

This chapter has shown how multiple imputation methods can be used to help alleviate problems caused by survey non-response and missing data. Multiple imputation is like an adjustable wrench - it is rarely the ideal tool for any particular job, but it works well for a wide variety of problems. The examples given in this chapter show that multiple imputation can be successfully implemented for real applied problems using existing software packages. Furthermore, Brownstone and Valletta's (1996) application shows that using this methodology can make a substantial difference in the qualitative conclusions.

Manski's work reviewed in Section 2 shows that missing data causes serious problems with identification and inference from even simple models. The best way to circumvent these problems is to put more resources into reducing response biases during survey administration. The next best solution is to collect external validation data which allow

identification of the non-response process. If these validation data become more widely available, then the multiple imputation methods presented in this chapter provide an easy and consistent way for transportation researchers to incorporate this information into their modeling and forecasting efforts.

8. References

Beckman, Richard J., Keith A. Baggerly and Michael D. McKay (1996), "Creating Synthetic Baseline Populations," *Transportation Research Part A-Policy And Practice*, 30:6, 415-429.

Bound, John, Charles Brown, Gregory J. Duncan, and Willard L. Rodgers, "Measurement Error in Cross-Sectional and Longitudinal Labor Market Surveys: Validation Study Evidence," in J. Hartog, G. Ridder, and J. Theeuwes (eds.), *Panel Data and Labor Market Studies*, (Amsterdam: North Holland, 1990).

Bound, John, Charles Brown, Gregory J. Duncan, and Willard L. Rodgers, "Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data," *Journal of Labor Economics*, 12 (July 1994), 345-368.

Bound, John and Alan B. Krueger, "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?," *Journal of Labor Economics* 9 (Jan. 1991), 1-24.

Brownstone, D. (1990), "Multiple Imputations for Panel Studies," Department of Economics Working Paper, UCI, April 1990. (revised version of paper presented at HUS conference, Gothenberg, Sweden, August 1989).

Brownstone, D. (1991) Multiple Imputations for Linear Regression Models. Technical Report MBS 91-37, Research Unit in Mathematical Behavioral Sciences, University of California, Irvine, California.

Brownstone, D. and X. Chu (1996), "Multiply imputed sampling weights for consistent inference with panel attrition," Chapter 10 in T. Golob and R. Kitamura ,eds., *Panel Data for Transportation Planning*, Kluwer Academic Publishers, Boston, in press.

Brownstone, D. and Golob, T.F. (1992) The effectiveness of ridesharing incentives: Discrete-choice models of commuting in Southern California. *Regional Science and Urban Economics*, **22**, 5-24.

Brownstone, D. and R.G. Valletta (1996), "Modeling earnings measurement error: a multiple imputation approach," *Review of Economics and Statistics*, 78:4: 705-717.

Clogg, Clifford C., Donald B. Rubin, Nathaniel Schenker, Bradley Schultz, and Lynn Weidman, "Multiple Imputation of Industry and Occupation Codes in Census Public-use Samples Using Bayesian Logistic Regression," *Journal of the American Statistical Association* 86 (March 1991), 68-78.

DuMouchel, William H. and Gregory J. Duncan, "Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples," *Journal of the American Statistical Association* 78 (June 1983), 535-543.

Fay, R.E. (1991), "A design-based perspective on missing data variance," in *Proceedings of the 1991 Annual Research Conference, U.S. Bureau of the Census*, 429-440.

Fay, R.E. (1992), "When are inferences from multiple imputation valid?," in *Proceedings of the Survey Research Methods Section, American Statistical Association,* 227-232.

Fuller, Wayne A., *Measurement Error Models* (New York: John Wiley, 1987).

Gange, Stephen J. (1995), "Generating Multivariate Categorical Variates Using the Iterative Proportional Fitting Algorithm," *The American Statistician*, Vol. 49, No. 2, pp. 134-38.

Glynn, R.J., N.M. Laird, and D.B. Rubin (1993), "Multiple imputation in mixture models for nonignorable nonresponse with follow-ups," *Journal of the American Statistical Association*, 88, 423, 984-993.

Golob, T. and R. Kitamura ,eds., (1996) *Panel Data for Transportation Planning*, Kluwer Academic Publishers, Boston, in press.

Golob T.F. and K. Swertnik (1996), Report on Southern California Edison Co.'s Electric Vehicle Trials, delivered to Southern California Edison Co.

Hausman, Jerry A., "Specification Tests in Econometrics," *Econometrica* 46 (Nov. 1978), 1251-1271.

Heckman, J.J. (1976), "The common structure of statistical models of truncation, sample selection, and limited dependent variables, and a simple estimator for such models," *Annals of Economics and Social Measurement*, 5, 475-592.

Heitjan, D. F. and R.J.A. Little (1991), "Multiple imputation for the fatal accident reporting system," *Applied Statistics*, 40, 1, 13-29.

Horowitz, J.L. and C.F. Manski (1995), "Censoring of outcomes and regressors due to survey nonresponse: identification and estimation using weights and imputations," working paper, Department of Economics, University of Iowa, October, 1995.

Imbens, G. (1992) An efficient method of moments estimator for discrete choice models with choice-based sampling. *Econometrica*, **60**, 1187-1214.

Imbens, G. and J.K. Hellerstein (1995), "Imposing moment restrictions by weighting," working paper, Department of Economics, Harvard University, May, 1995.

Kitamura, R., (1996), "Two computational-process models of daily activity-travel behavior," paper presented at Theoretical Foundations of Travel Choice Modeling Conference, Stockholm, Sweden, August 7 - 11, 1996.

Lave, C. (1996), "Are Americans really driving so much more?," *Access*, 8 (Spring 1996), 14-18.

Leamer, E.E. (1983), "Lets take the con out of econometrics," *American Economic Review,* 73 (1), 31-43.

Lee, Lung-fei and Jungsywan H. Sepanski, "Estimation of Linear and Nonlinear Errors-in-Variables Models Using Validation Data," *Journal of the American Statistical Association* 90 (March 1995), 130-140.

Little, R.J.A. (1988), "Missing data adjustments in large surveys (with discussion)," *Journal of Business and Economic Statistics,* 6, 287-297.

Little, Roderick J. A. and Mei-Miau Wu (1991), "Models for Contingency Tables with Known Margins when Target and Sampled Populations Differ," *Journal of the American Statistical Association*, Vol. 86, No. 413, pp. 87-95.

Manski, C.F. (1991), "Regression," *Journal of Economic Literature*, 29, 34-50.

Manski, C.F. (1995), *Identification problems in the social sciences*, Harvard University Press, Cambridge, Massachusetts.

Manski, C.F. and Lerman. S. (1977) The estimation of choice probabilities from choice-based samples. *Econometrica*, **45**, 1977-1988.

McFadden, D., and P. Ruud, 1994, "Estimation by Simulation," *Review of Economics and Statistics*, Vol. 76, No. 4, pp. 591-608.

Meng, Xiao-li, "Multiple Imputation with Uncongenial Sources of Input" (with discussion), *Statistical Science* 9, (Spring 1994) 583-574.

Meng, Xiao-li and Donald B. Rubin , "Performing Likelihood-Ratio Tests with Multiply-Imputed Data Sets," *Biometrika* 79 (Jan. 1992), 103-111.

Rubin, D.B. (1986) Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, **4**, 87-94.

Rubin, Donald B. (1987a), *Multiple Imputation for Nonresponse in Surveys* (New York: John Wiley, 1987).

Rubin, Donald B. (1987b), "Discussion of Tanner and Wong," *Journal of the American Statistical Association* 82, 543-546.

Rubin, Donald B., "Multiple Imputation After 18+ Years," *Journal of the American Statistical Association* 91 (June 1996), 473-489.

Schenker, Nathaniel and A.H. Welsh, "Asymptotic Results for Multiple Imputation," *Annals of Statistics* 16 (Dec. 1988), 1550-1566.

Uhlaner, C.J. and Kim. S. (1993) Designing and Implementing a Panel Study of Commuter Behavior: Lessons for Future Research. Working Paper 93-2, Institute of Transportation Studies, University of California, Irvine, California.