Evaluation of Imputation of Covariates in an Impact Analysis With Regression Adjustment

Eric Grau[1], Susan Ahmed[2]
[1]Mathematica Policy Research, 600 Alexander Park, Princeton, NJ 08540-6346
[2]Mathematica Policy Research, 600 Maryland Avenue, SW, Suite 550, Washington, DC 20024-2512

Abstract
In an impact analysis using random assignment, researchers often deal with missing values in both the covariates and the outcome variables of regression models. Clearly rigorous methods are needed to impute missing values in the outcome variables to minimize the potential bias in impact assessments. When imputation is applied to covariates of the regression analyses, the effect of imputation is less clear on impact analyses. This paper assesses this effect, using a random assignment evaluation of the Growing America Through Entrepreneurship (GATE) program. Two outcome variables used in the original evaluation are modeled against a set of 10 covariates, a treatment indicator, and variables associated with the site of the evaluation. Impacts are assessed with different types of missingness in the covariates with values imputed using mean imputation and sequential hot deck.

Key Words: imputation, impact evaluation, sequential hot deck, regression adjustment, random assignment

1. Introduction

Considerable work has been done in the area of statistical inference with missing data. Special attention has been paid to the problem of missing values of the independent variables in regression. Little (1992) reviewed different methods of dealing with missing values in the independent variables, when the outcome variable is not missing. He determined that maximum likelihood methods that are based on both the independent variables and the dependent variable are preferred over other methods, with considerable promise shown for multiple imputation methods as well. Allison (2001) also recommends using likelihood-based methods, though he does indicate the list wise deletion is superior to all imputation methods (but particularly single imputation methods) in some instances. Little and Rubin's (2003) recommendation is to "impute draws (not means) from the conditional distribution of the missing covariates given the observed covariates and Y." They indicate that this will produce consistent estimates of the regression estimates.

The results discussed above were obtained by looking at estimates of the mean, variance, and coefficient parameters. In an impact evaluation, treatment status is included as one of the covariates in the regression. Analysts are interested in the difference between the treatment and control groups, often adjusted by controlling for other auxiliary covariates in the regression. The main objective of this study is evaluating what happens when there are missing data in these auxiliary covariates. There may be insufficient resources (or insufficient expertise) to implement sophisticated imputation procedures. The question therefore arises: is it necessary to impute missing values in the auxiliary covariates of a regression when the study is an impact evaluation, where the focus is on the comparison between treatment and control means? If it is necessary, does it matter which method is used?

In this paper, we look at the effect of missing data in independent variables in a regression, focusing on the estimated effects of the contrast which compares treatment to control outcomes. We compare list wise deletion with unconditional mean imputation (within treatment group), conditional mean imputation, and unweighted sequential hot deck imputation. For the remainder of this section, we discuss mechanisms that lead to missing data, and options for handling missing data. Section 2 describes the source data and the study from which the data came. The study methodology is described in Section 3. Results are presented in Section 4, and the paper concludes with conclusions and discussion in Section 5.

1.1 Missing Data Mechanisms

In other settings where it is necessary to accommodate missing data, the correct application of methods to analyze the data depends crucially on the missing data mechanism. It is no different with impact evaluations. As delineated by Rubin (1976) there are three missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR).

For data to be MCAR, the missing data mechanism is completely independent of the value of the response variable Y. In essence, this assumption is equivalent to saying that the set of nonrespondents could be considered a random sample from the full set of sample members and that respondents and nonrespondents are exactly alike in terms of Y. This is a stringent assumption, and is often violated for methods that require it.

A much weaker assumption is MAR, which states that the missing data mechanism is independent of the value of the response variable Y, but only after controlling for data that is observed. In practice, this means controlling for other covariates (X) in the analysis which have nonmissing data. This assumption only requires that respondents and nonrespondents are alike within the levels of the other covariates in the analysis. This is formally written as:

$$\Pr(Y_{missing} | Y, X) = P(Y_{missing} | X)$$

If the missing data mechanism is MAR, and the parameters that govern the missing data process are unrelated to the parameters being estimated, the missing data mechanism is ignorable.

If the missing data mechanism is neither MCAR nor MAR, then the missing data are not missing at random (NMAR). With this mechanism, the probability of missing data in Y is related to Y even after controlling for other variables in the analysis. This is a much more difficult situation, requiring explicit modeling of the missing data mechanism.

There are no tests to determine whether missing data are MCAR, MAR, or NMAR. Usually, however, ignorability is a reasonable assumption, particularly if the number of auxiliary covariates available for analysis is sufficient to account for the missing data mechanism.

1.2 Options for Handling Missing Data

With most available commercial software, the default method for handling missing data is list wise deletion. In particular, observations that have missing values for one or more variables used in the analysis are deleted. The obvious advantages of list wise deletion are its simplicity, in that it can be applied for any type of statistical analysis, and that no special computational methods are required. More importantly, the variance estimates are always appropriate. However, list wise deletion suffers from two major disadvantages, which make it unsuitable for use in most circumstances. First of all, list wise deletion requires the missing data mechanism to be MCAR, which is often an unrealistic assumption. In general, using list wise deletion when the data are not MCAR can lead to significant nonresponse bias in estimates. In this paper we will empirically determine whether this bias extends to impact estimates in an impact evaluation. In addition, the sample size can be severely reduced because, for each observation, data must be nonmissing for all variables.

Other methods that are based on completely recorded units include available case analysis (see Little and Rubin, 2003) and dummy variable adjustment (see Cohen and Cohen, 1985). Both of these methods have bias problems that make them less appealing than list wise deletion.

Another option is to use weighting methods. These methods assign factors to each observation that give greater relative importance to observations similar to those with missing data. It is the preferred method for unit nonresponse, where an entire unit from a survey is missing. It is impractical for item nonresponse, however, since different weights would have to be created for every analysis, depending upon the variables used.

For item nonresponse, the preferred option is imputation, or the replacement of missing values with imputed values using available information from the data set or from external sources. The advantages of imputation are that it will hopefully reduce nonresponse bias. It also mitigates the problem of lost information when many variables are used simultaneously. There are two major caveats, however: (1) the variance will be underestimated unless accounted for explicitly, and (2) the bias can be increased if the imputation is not done well.

Little and Rubin (2003) separate imputation methods into two types: those based on implicit models, and those based on explicit models. Imputation methods based on implicit models include the many hot deck methods, which use information from other similar records within the same data set, and replace missing values with the values from the similar records. Explicit models are used to define mean imputation, where the missing values are replaced by the mean of recorded values, and other methods with varying levels of sophistication. Unconditional mean imputation (also called marginal mean imputation), where the overall mean is used, will give biased estimates if the data are MAR. Conditional mean imputation, where the missing values are replaced by the means within adjustment cells of similar observations, can yield reasonable estimates, though the variance will be severely underestimated. Other explicit model methods include regression imputation, stochastic regression imputation, maximum likelihood, and multiple imputation. Although the latter methods have much to recommend them (particularly maximum likelihood and multiple imputation), it is not clear that the effort required to implement them is justified for the question posed in this paper.

In subsequent sections, we focus our attention on methods that are quick and easy to implement: in particular, we will be comparing list wise deletion, unconditional mean imputation (within treatment group), conditional mean imputation, and sequential hot deck, to determine if it really matters what method is used in an impact evaluation. This form of unconditional mean imputation was considered "unconditional" since treatment group was not related to the outcome variables.

When item nonresponse is very low, whether we impute and what method we use is less of an issue. When item nonresponse is high and the variable with missing data is important for analysis, whether we impute and what method is used is very important. This paper looks at a different question: what if item nonresponse is high and the variable with missing data is subsidiary to other variables in the analysis? Should we impute the missing data, and if so, does it matter what method we use?

2. Description of Source Data and the Study from Which It Came

2.1 Description of Study

The data for this paper was obtained from a study that Mathematica Policy Research, Inc. (MPR) performed under contract with the Employment and Training Administration of the Department of Labor. The project was called Growing America Through Entrepreneurship (GATE). This project was an evaluation of an effort to assist potential entrepreneurs in developing or growing their own business. Individuals in the study were randomly assigned to a treatment group, which involved training and assistance in entrepreneurship, and a control group, which did not receive any training or assistance. The study was conducted in 7 sites (Philadelphia, PA; Pittsburgh, PA; Minneapolis, MN; Duluth, MN; Bangor, ME; Portland, ME; and Lewiston, ME). For the purposes of analysis, the sites in Maine were combined. See Bellotti, et al (2006) and McConnell, et al (2004) for more details on this study.

Individuals applied for the program at these 7 sites, at which time background information was collected for all applicants. This information did not have missing data, and was used, in addition to the site indicators and treatment indicator, as covariates in the analysis. Surveys were conducted 6 months and 18 months after the start of the study. Two of the many questions in the survey will serve as outcome variables in this paper: monthly sales of the startup business and monthly expenses of the startup business. At baseline, data was obtained from 4,200 applicants, which were randomly allocated to a treatment group (2,096 applicants) and a control group (2,104 applicants). In some cases, multiple business partners from the same business applied to the program; the 4,200 applicants represented 4,071 businesses. Results were used from the 18-month survey, from which there were 2,946 respondents (1,516 in the treatment group and 1,430 in the control group). Analysis weights were computed to account for unit nonresponse and multiple respondents per business. The outcome data was perturbed for the results presented in this paper for illustrative purposes, though the covariances between key variables were maintained.

Two models were chosen from the original study. The outcome variables were monthly sales and monthly expenses, with missing values imputed using a sequential hot deck. For the illustrative purposes, these imputed values were considered real for the study described in this paper. Auxiliary covariates included gender, race (black or not black), household income (in three categories), existence of a formal business plan at baseline, whether the business grew out of a hobby, whether the applicant's family was very supportive, and whether the applicant had credit problems at baseline. Study site and treatment status were also included as covariates. Because the auxiliary covariates were obtained from the baseline application, there were no missing values in the original data. Treatment status and study

site were also nonmissing for all respondents. The treatment effect was measured using an estimated contrast. The effects of interest were regression-adjusted differences in monthly expenses and monthly sales between treatment and control respondents for each of the five sites and overall.

The covariates most highly correlated with the outcome variables were household income (r = 0.155 for monthly expenses, r = 0.143 for monthly sales), existence of a formal business plan at baseline (r = 0.105 for monthly expenses, r = 0.100 for monthly sales), and whether the applicant had credit problems at baseline (r = -0.111 for monthly expenses, r = -0.091 for monthly sales). Not surprisingly, monthly sales and monthly expenses were highly correlated with each other (r = 0.810). In the original data, limited impacts were detected for either monthly sales or monthly expenses.

3. Study Methodology

3.1 Models

For comparison purposes the regression-adjusted effects were calculated with no missing data. Outcome variables were perturbed by adding constants to the monthly sales and monthly expenses for treatment group members, where different amounts were added within various levels of the auxiliary covariates, with the intention of maintaining the original covariances as much as possible. This was done to manufacture a treatment effect, in order to make the comparison more informative. As stated earlier, imputed sales and imputed expenses were treated as real. Models were fitted with no missing data and with missingness induced using MCAR and MAR mechanisms for household income alone or household income and whether the applicant had a formal business plan. In addition, with monthly expenses as the outcome variable, monthly sales was added as a covariate. This was done to evaluate the effect of imputation in a covariate that was very highly correlated with the response variable. Regression-adjusted effects were estimated with no missing data, and with missingness induced in both household income and monthly sales. The models in which comparisons were made are shown in Table 1.

Table 1: Models Upon Which Comparisons Are Made

Response Variable	Covariates With Induced Missingness	Covariates With No Missing Data
Monthly sales	Household income	Business plan, Gender, Race, Hobby, Supportive family, Credit problems, Treatment/control, Site
Monthly sales	Household income, business plan	Gender, Race, Hobby, Supportive family, Credit problems, Treatment/control, Site
Monthly expenses	Household income	Business plan, Gender, Race, Hobby, Supportive family, Credit problems, Treatment/control, Site
Monthly expenses	Household income, business plan	Gender, Race, Hobby, Supportive family, Credit problems, Treatment/control, Site
Monthly expenses	Household income, monthly sales	Business plan, Gender, Race, Hobby, Supportive family, Credit problems, Treatment/control, Site

3.2 Induced Missingness

Results were compared for each model under the following scenarios:

1. No missing data
2. MCAR (30% missingness for each covariate with induced missingness)
3. MAR Setting 1 (30% missingness for each covariate with induced missingness)
4. MAR Setting 2 (30% for each covariate with induced missingness in the treatment group only and 15% for each covariate with induced missingness in the control group)

The covariates with induced missingness are given in Table 1. The original intention was to look at larger levels of missingness first, and if a clear difference between imputation methods was found, to look further at smaller missingness levels. However, the results at the 30% level were nuanced, and there was insufficient time and resources to investigate smaller missingness levels. This will be left for further research.

For MCAR, we randomly set 30% of household income to missing in the one variable case. The random allocation to missing was done based on a random number from the Uniform(0,1) distribution. In the two variable case, we randomly set 30% of household income and 30% of the business plan variable to missing using different Uniform(0,1) random variables. As would be expected with such a random allocation, 9% of the observations were missing both (0.3 * 0.3), 42% were missing one or the other (0.3 * 0.7 * 2), and 49% were missing neither.

For MAR in the one variable case, we classified the 2,934 respondents into four clusters with varying income levels based on the values of marital status (married or not), education level (college graduate or not), and whether or not the applicant had credit problems. The first cluster, generally associated with the highest incomes, included respondents who were married college graduates with no credit problems. The fourth cluster, associated with the lowest incomes, included respondents who were unmarried, and they either were college graduates with credit problems, or they were not college graduates (with and without credit problems). We randomly set household income to missing for the first cluster (using a Uniform random number between 0 and 1) at twice the rate of the fourth cluster, with the second and third cluster in between, so that the overall level of missingness was 30% in Setting 1 (with equal missingness in the treatment and control groups). The overall level of missingness for Setting 2 was 15%, with 30% missing in the treatment group and no missing data in the control group.

In the MAR case for two variables, missingness in household income was induced in the same way as in the one variable case. For the business plan variable, the 2,934 respondents were classified into four clusters based upon different proportions of respondents who had a formal business plan at the time of application. The four clusters were based on the values of race (black or not black), level of family support for business venture (very supportive or not), and education level (college graduate or not). The first cluster, generally associated with the highest proportion who had a formal business plan, included respondents who were black college graduates with families very supportive of the business venture. The fourth cluster, associated with the lowest proportion who had a formal business plan, included respondents who were not black, not college graduates, and did not have families that were very supportive of the business venture. In the same manner as but independently of household income, we randomly set the business plan variable to missing for the first cluster (using a Uniform random number between 0 and 1) at twice the rate of the fourth cluster, with the second and third cluster in between, so that the overall level of missingness was 30% in Setting 1 (with equal missingness in the treatment and control groups). The overall level of missingness for Setting 2 was 15%, with 30% missing in the treatment group and no missing data in the control group.

When monthly sales of the business was included as a covariate, missingness was induced for the both settings of MAR (30% overall missingness and 30% missingness in the treatment group only), but only results from the second setting are presented here. For monthly sales, the 2,934 respondents were classified into four clusters based upon different levels of monthly sales. The four clusters were based on the values of household income (six categories), gender, and whether the respondent had a formal business plan. The first cluster, generally associated with the highest monthly sales, included male respondents who were in the second highest income group and had a formal business plan. The fourth cluster, associated with the lowest monthly sales, included all female respondents in the three lowest income groups, all male respondents in the three lowest income groups without a formal business plan, all female respondents in the fourth and fifth lowest income groups without a formal business plan, and all respondents in the four lowest income groups where information about a formal business plan was missing. In the same manner as but independently of household income, we randomly set monthly sales to missing for the first cluster (using a Uniform random number between 0 and 1) at twice the rate of the fourth cluster, with the second and third cluster in between, so that the overall level of missingness was 30% in Setting 1 (with equal missingness in the treatment and control groups).

3.3 Imputation Methods

Missing values for covariates were replaced by imputed values using three methods: unconditional mean imputation (within treatment group), conditional mean imputation, and unweighted sequential hot deck. For the mean imputation of binary variables, the missing value was replaced by the estimated proportion and not a 0 or 1.

3.3.1 Unconditional Mean Imputation

For the unconditional mean imputation, missing values are replaced by the overall mean of the recorded values of the outcome variable. In this application, missing values were replaced by the mean within each site and treatment group. The imputation was called "unconditional" because the treatment group is unrelated to the outcome variable for both monthly sales and monthly expenses.

3.3.2 Conditional mean imputation

For the conditional mean imputation, missing values are replaced by the mean within adjustment classes defined by variables closely related to the outcome variable. In this application, these variables included site, treatment group, marital status (married or not), education level (college graduate or not), and whether the respondent had credit problems.

3.3.3 Unweighted Sequential Hot Deck

For all hot deck imputation methods, missing values are replaced with values from a similar observation within the same sample data set. The donor is the observation that provides the data to the item with missing data, which is called the "donee" or "recipient." Imputation classes are formed from joint levels of nonmissing categorical variables highly correlated with the variable in question. In rare instances, an imputation class might contain recipients with no potential donors. In that case, it would be necessary to collapse imputation classes.

What distinguishes the different hot deck methods is the method of choosing a donor for a particular recipient within an imputation class. Hot deck methods include sequential hot deck, nearest neighbor hot deck, and random hot deck. With sequential hot deck, donors and recipients are sorted together according to a continuous variable or variables that is (are) highly correlated with the outcome variable. The order of the sorting variables matters, with the primary sorting variable having the most influence on the selection of the donor. Often the sort used is a serpentine sort, where the secondary variable sort alternates between ascending and descending whenever the level of the primary variable changes. (Subsequent variables alternate in a similar manner.) This ensures that the sequential hot deck is not so dependent on the values of the primary sorting variable. Nearest neighbor hot deck donors are selected according to a well-defined distance metric. Donors in a random hot deck are selected randomly within an imputation class. In some cases, a random component is added to the sequential hot deck and nearest neighbor hot deck procedures. In general, hot decks can be weighted or unweighted.

In this application, we applied an unweighted sequential hot deck, where site, marital status (married or not), education level (college graduate or not), and whether the respondent had credit problems were used to define imputation classes for the household income imputation. Three indicator variables were used as sorting variables, in the following order: whether the respondent had 5 or more years of managerial experience, whether the respondent had collected unemployment insurance, and whether the respondent had a family member in the household drawing a regular salary. For the imputation of the business plan variable, the imputation classes were defined by race (black or not), level of family support for the new business (very supportive or not), and education level (college graduate or not). Three indicator variables were used as sorting variables, in the following order: gender, whether the respondent was self-employed at the time of the application, and whether the respondent had a family member in the household drawing a regular salary. Finally, for the imputation of monthly sales when that variable was used as a covariate for the outcome variable monthly expenses, the imputation classes were defined by an 8-level categorical variable corresponding to different levels of monthly expenses and a 5-level household income variable. Three indicator variables were used as sorting variables, in the following order: whether the respondent had a formal business plan at the time of application, gender, and whether the respondent had credit problems. All of the hot decks included a random number as a final variable in the list of sorting variables. The unweighted sequential hot deck was not applied to the MCAR situation, since it was clear that even unconditional mean imputation of missing values in the covariates reproduced test statistics and p-values that were close to the original (nonmissing) data.

3.4 Method of Evaluation

Means and standard errors of household income, the business plan variable, and monthly sales were produced before missingness was induced, for each missingness mechanism, and for each imputation method within each missingness mechanism (although the hot deck was not applied when the data were MCAR). In addition, estimated effects were compared within sites and overall for each of these situations. In this paper, we evaluated the impact of inducing

different types of missingness, and determined how well each of the imputation methods were able to reproduce the estimated effects, test statistics, and their p-values obtained prior to inducing missingness.

4. Results

4.1 Results, MCAR

When MCAR was the induced missingness mechanism, the mean value for the covariate being imputed was unaffected by the missingness, but the standard errors increased (as expected). The regression-adjusted estimated effects gave somewhat different conclusions than was found with the original data, though in the case where covariates were not highly correlated with the outcome variable, an unconditional mean imputation was all that was needed to reproduce estimated effects, test statistics, and p-values that were close to those of the original data. With a covariate highly correlated with the outcome variable, the unconditional mean imputation did not fare as well, and the hot deck was not tested with MCAR data.

4.2 Results, MAR Setting 1

The first setting of MAR data involved a constant level of nonresponse across the two treatment groups. Estimated means and standard errors for the covariates themselves, before and after missingness was induced, are presented in Table 2. As is apparent, list wise deletion results in biased estimates, where all the parameter estimates are underestimates, though the standard errors are appropriately larger due to the smaller sample size. Bias is still a problem with unconditional mean imputation, with standard errors severely underestimated. Conditional mean imputation improves the bias and standard errors, though the best method for providing estimates close to the original data is the hot deck. It should be noted that even though the standard error is close to that of the original data, it is still an underestimate of the true standard error (which would exceed that of the original data), since we have not accounted for the imputation uncertainty.

Table 2: Means and standard errors of covariates with 30% MAR (Setting 1) missing data: Household Income, Business Plan, and Monthly Sales

Site	No Missing	30% MAR, No Imputation	30% MAR, Unconditional Mean Imputation	30% MAR, Conditional Mean Imputation	30% MAR, Hot Deck
Household income	$42,159 ($552)	$40,531 ($639)	$41,121 ($457)	$42,633 ($494)	$41,840 ($544)
Business plan	0.2291 (0.0081)	0.2209 (0.0095)	0.2193 (0.0067)	0.2241 (0.0068)	0.2239 (0.0080)
Monthly sales	$3,159 ($160)	$2,891 ($172)	$2,946 ($124)	$3,045 ($142)	$3,088 ($163)

The question arises, however, about whether these biases and underestimated standard errors have a deleterious effect on the estimated effects. We do not present the situations where only one covariate in the regression, household income, was set to missing, as the conclusions drawn for the two-variable case can be extrapolated to the one-variable case. Estimated effects and p-values are shown for the regression involving monthly sales as the outcome variable in Tables 3 and 4, respectively, where missingness was induced in the covariates household income and the business plan variables. (Conclusions for monthly expenses were similar to those of monthly sales.) The only clear result is that with 30% missingness, list wise deletion is not recommended. There is no clear directional bias across sites, but the differences between the estimated effects for the original data effects and the list wise-deletion estimates are dramatic. Conclusions would differ as well for Philadelphia, Pittsburgh, and Maine. Imputation is therefore recommended, but the choice of imputation method is less clear. For estimates of means, bias was almost as large with unconditional mean imputation as it was for the list wise deletion case. However, this bias did not translate in most cases to the estimated effects. In fact, the estimated effects for no-missing-data case were closer in value to the estimated effects in the unconditional mean imputation case than with other methods of imputation. On the whole, however, there wasn't an appreciable difference between any of the imputation methods, as is especially apparent when comparing P-values in Table 4. Only in the Duluth site does hot deck marginally outperform the other methods.

Table 3: Estimated Effects for 30% MAR (Setting 1), Monthly Sales as Outcome Variable, Missingness Induced in Covariates Household Income and Business Plan

	Estimated Effect	Relative Difference from No Missing Data Case (30% MAR)			
Site	No Missing	30% MAR No Imputation	30% MAR Unconditional Mean Imputation	30% MAR Conditional Mean Imputation	30% MAR Hot Deck
Philadelphia	$1,772	-31.35%	-1.28%	-4.53%	-2.15%
Pittsburgh	$1,318	90.76%	1.16%	3.25%	3.34%
Minneapolis	$2,125	21.28%	-1.22%	3.87%	-1.65%
Duluth	$1,998	-19.58%	7.51%	8.00%	2.67%
Maine	$2,569	-3.72%	-1.40%	-1.70%	-3.34%
Total	**$1,956**	**6.20%**	**0.83%**	**1.65%**	**1.72%**

Table 4: P-Values for 30% MAR (Setting 1), Monthly Sales as Outcome Variable, Missingness Induced in Covariates Household Income and Business Plan

Site	No Missing	30% MAR No Imputation	30% MAR Unconditional Mean Imputation	30% MAR Conditional Mean Imputation	30% MAR Hot Deck
Philadelphia	<0.0001	0.0316	<0.0001	<0.0001	<0.0001
Pittsburgh	0.1272	0.0001	0.1364	0.1301	0.1301
Minneapolis	0.0003	0.0095	0.0004	0.0002	0.0005
Duluth	0.1264	0.3838	0.0971	0.0979	0.1388
Maine	0.0011	0.0558	0.0013	0.0014	0.0019
Total	**<0.0001**	**0.0001**	**<0.0001**	**<0.0001**	**<0.0001**

The biases in the estimate of the mean of monthly sales follow the same pattern as household income and the business plan variable, as is apparent in Table 2. We now consider the regression with monthly expenses as the outcome variable, and missingness induced in the covariates household income and monthly sales. Estimated effects and p-values are shown for this situation in Tables 5 and 6, respectively. As with the situation where missingness was induced in the household income and the business plan variables, list wise deletion is generally not recommended. For the P-values in particular, list wise deletion results in effects deemed significant when they were not significant in the original data. No imputation method is clearly superior, however; in fact, although all imputation methods are better than list wise deletion, none does a consistently good job of reproducing the results in the original data.

Table 5: Estimated Effects for 30% MAR (Setting 1), Monthly Expenses as Outcome Variable, Missingness Induced in Covariates Household Income and Monthly Sales

	Estimated Effect	Relative Difference from No Missing Data Case (30% MAR)			
Site	No Missing	30% MAR No Imputation	30% MAR Unconditional Mean Imputation	30% MAR Conditional Mean Imputation	30% MAR Hot Deck
Philadelphia	$743	-37.18%	6.78%	-35.92%	24.56%
Pittsburgh	$402	113.06%	-53.34%	-22.16%	-2.40%
Minneapolis	$391	68.05%	10.47%	-18.31%	18.99%
Duluth	$59	244.92%	-846.88%	455.32%	-1054.27%
Maine	$695	8.62%	56.19%	25.41%	84.85%
Total	**$458**	**28.34%**	**-10.20%**	**0.80%**	**-6.64%**

Table 6: P-Values for 30% MAR (Setting 1), Monthly Expenses as Outcome Variable, Missingness Induced in Covariates Household Income and Monthly Sales

Site	No Missing	30% MAR No Imputation	30% MAR Unconditional Mean Imputation	30% MAR Conditional Mean Imputation	30% MAR Hot Deck
Philadelphia	0.0002	0.0001	<0.0001	0.0058	0.0007
Pittsburgh	0.2356	<0.0001	0.7365	0.4215	0.3463
Minneapolis	0.0712	0.0026	0.1452	0.2181	0.0731
Duluth	0.9255	0.4220	0.6872	0.4414	0.5813
Maine	0.0702	<0.0001	0.0117	0.0137	0.0013
Total	**0.0369**	**<0.0001**	**0.1795**	**0.0381**	**0.1330**

4.3 Results, MAR Setting 2

Heretofore, all of the comparisons that have been made assume that the level nonresponse is the same between the treatment group and the control group. What happens if the level of nonresponse differs between the treatment and control group, also known as differential attrition? The second setting of MAR data involved 30% missingness in the treatment group and no missingness in the control group, which would be considered an extreme (and unrealistic) level of differential attrition. We only present results for the regression with missingness induced in household income and the business plan variable. Estimated effects and P-values are presented in Tables 7 and 8, respectively. Here we see a very different result than was found with constant nonresponse. In this case, it appears that list wise deletion is actually superior to mean imputation, whether we are referring to unconditional means or conditional means. The P-values for list wise deletion are actually closer to the original data than the mean imputation P-values. The best method in this instance, however, appears to be hot deck. Not only are the P-values closer to those of the original data, but the estimated effects appear to have less bias.

Table 7: Estimated Effects for 30% MAR (Setting 2), Monthly Sales as Outcome Variable, Missingness Induced in Covariates Household Income and Business Plan

	Estimated Effect	Relative Difference from No Missing Data Case (30% MAR)			
Site	No Missing	30% MAR No Imputation	30% MAR Unconditional Mean Imputation	30% MAR Conditional Mean Imputation	30% MAR Hot Deck
Philadelphia	$1,772	-24.20%	7.70%	-8.25%	-3.35%
Pittsburgh	$1,318	5.88%	15.22%	-17.55%	1.18%
Minneapolis	$2,125	11.19%	14.58%	-31.37%	0.35%
Duluth	$1,998	-4.03%	23.50%	-40.52%	2.15%
Maine	$2,569	6.46%	2.67%	-27.08%	-5.87%
Total	**$1,956**	**-0.29%**	**12.11%**	**-26.06%**	**-1.48%**

Table 8: P-Values for 30% MAR (Setting 2), Monthly Sales as Outcome Variable, Missingness Induced in Covariates Household Income and Business Plan

Site	No Missing	30% MAR No Imputation	30% MAR Unconditional Mean Imputation	30% MAR Conditional Mean Imputation	30% MAR Hot Deck
Philadelphia	<0.0001	0.0002	<0.0001	<0.0001	<0.0001
Pittsburgh	0.1272	0.1333	0.0797	0.0853	0.1264
Minneapolis	0.0003	0.0056	<0.0001	<0.0001	0.0004
Duluth	0.1264	0.1516	0.0539	0.2875	0.1173
Maine	0.0011	0.0078	0.0008	<0.0001	0.0027
Total	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**

We now consider the regression with monthly expenses as the outcome variable, and missingness induced in the covariates household income and monthly sales. Estimated effects and p-values are shown for this situation in Tables 9 and 10, respectively. In these tables, it is apparent that none of the methods show a clear advantage. The advantages that were apparent with the hot deck with the induced missingness in covariates that were not highly correlated with the outcome variable are not apparent here.

Table 9: Estimated Effects for 30% MAR (Setting 2), Monthly Expenses as Outcome Variable, Missingness Induced in Covariates Household Income and Monthly Sales

	Estimated Effect	Relative Difference from No Missing Data Case (30% MAR)			
Site	No Missing	30% MAR No Imputation	30% MAR Unconditional Mean Imputation	30% MAR Conditional Mean Imputation	30% MAR Hot Deck
Philadelphia	$743	-4.89%	-11.51%	-14.53%	-3.80%
Pittsburgh	$402	-18.63%	-2.17%	-9.88%	40.18%
Minneapolis	$391	-22.00%	31.08%	-10.20%	-15.63%
Duluth	$59	-941.64%	-65.76%	-258.66%	8.62%
Maine	$695	-10.46%	35.21%	-14.29%	30.68%
Total	**$458**	**-27.96%**	**10.17%**	**-15.74%**	**12.68%**

Table 10: P-Values for 30% MAR (Setting 2), Monthly Expenses as Outcome Variable, Missingness Induced in Covariates Household Income and Monthly Sales

Site	No Missing	30% MAR No Imputation	30% MAR Unconditional Mean Imputation	30% MAR Conditional Mean Imputation	30% MAR Hot Deck
Philadelphia	0.0002	0.0052	0.0020	0.0018	0.0003
Pittsburgh	0.2356	0.2503	0.2417	0.1622	0.0923
Minneapolis	0.0712	0.2295	0.0260	0.1056	0.1523
Duluth	0.9255	0.4039	0.9759	0.8822	0.9230
Maine	0.0702	0.0184	0.0276	0.0888	0.0273
Total	**0.0369**	**0.1305**	**0.0304**	**0.0766**	**0.0217**

5. Discussion

If we have regression-adjusted impact estimates with a high level of MAR missingness in some of the covariates, we have shown that it is almost always advisable to impute the missing values in the covariates. With equal levels of nonresponse between the treatment and control groups, there does not appear to be an advantage to using hot deck against using a simple mean imputation. We did not consider regression imputation, though it could be argued that the results for regression imputation would be similar to hot deck and conditional mean imputation, since it accounts for the same set of covariates.

With differential nonresponse, the situation was less clear. If the covariates had a low correlation with the outcome variable, hot deck was clearly better; however, with missingness induced in a covariate highly correlated with the outcome variable, none of the methods were close to what was observed with the original data.

When we looked at differential nonresponse, we were looking at an extreme case which would, in all likelihood, never occur. This needs to be explored further, to determine at what level of differential nonresponse does the imputation method matter. Further work also needs to be done to determine if there is a mathematical reason why hot deck outperforms mean imputation in the case we tested. In addition, we have only looked at imputations that condition on other covariates X. Little and Rubin (2003) have indicated that imputing draws (not means) from the conditional distribution of the missing covariates given the observed covariates and Y will yield consistent estimates of the regression coefficients. We could test this to see if that also applies to estimated effects. Finally, we have not looked at methods involving explicit models, including regression imputation, stochastic regression imputation, maximum likelihood, or multiple imputation. Further work needs to be done incorporating these other methods.

A key point that has not been addressed in this paper is the amount of missingness. It may be that with small amounts of missingness, list wise deletion will perform adequately. Certainly, with low levels of item nonresponse, the choice of imputation methods when imputing the covariates will have less of an impact on the results of the analysis.

References

Allison, Paul D. (2001). *Missing Data.* Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136. Thousand Oaks, CA: Sage.

Bellotti, Jeanne, Sheena McConnell, and Jacob Benus. "Growing America Through Entrepreneurship: Interim Report." Submitted to U.S. Department of Labor, Employment and Training Administration. Princeton, NJ: Mathematica Policy Research, Inc., August 2006.

Cohen, J. and P. Cohen (1985). *Applied Multiple Regression and Correlation Analysis for the Behavioral Sciences*, 2nd edition, Hillsdale, NJ: Erlbaum.

Little, Roderick J.A. (1976) "Inference about means from incomplete multivariate data," *Biometrika* **63**, 593-604.

Little, Roderick J.A. (1992) "Regression with missing X's: a review," *J. Am. Statist. Assoc.* **87**, 1227-1237.

Little, Roderick J.A., and Donald B. Rubin (2003). *Statistical Analysis with Missing Data*, 2nd edition, New York: Wiley

McConnell, Sheena, Irma Perez-Johnson, Jeanne Bellotti, Nuria Rodriguez-Planas, and Walter Corson. "Project GATE: Evaluation Design." Final report submitted to the U.S. Department of Labor, Employment and Training Administration. Washington, DC: Mathematica Policy Research, Inc., August 2004.