General Remedies to Local Problems: An Applied Researcher's Manual to Multiple Imputation

by

Gayaneh Kyureghian

Oral Capps, Jr.

Rodolfo M. Nayga, Jr.

Selected Paper prepared for presentation at the Agricultural & Applied Economics Association's 2011 AAEA & NAREA Joint Annual Meeting, Pittsburgh, Pennsylvania, July 24-26, 2011

Kyureghian is a Research Assistant Professor, University of Nebraska-Lincoln, email: gkyureghian2@unl.edu, Capps is a Professor, Texas A&M University, email: ocapps@tamu.edu , Nayga is a Professor, University of Arkansas, email: rnayga@uark.edu.

Abstract

Nonresponse is a pervasive and persistent problem in survey data. This research reviews several methods for imputing missing values. A special emphasis is placed on the multiple imputation methods as a more generalizable advanced remedy to missingness. An empirical application of these methods, along with a regression or conditional mean imputation, is provided. Contingent upon certain properties of data, inference based guidance to the choice and implementation of these methods is provided.

Introduction

Problems of survey nonresponse are not uncommon in social science. Censuses and surveys typically suffer from non-response to one or more questions due to sensitivity to disclosure of certain types of information, such as income; recall problems and bias; insufficient time to complete survey; technical limitations of measurement devices; attrition due to moving or death of some panel members in longitudinal surveys, etc. Any of these problems results in missing data points or even missing variables altogether (Rubin, 1987). If ignored the missingness creates considerable problems for field researchers as most data analysis procedures are designed for complete data matrices (Schafer & Graham, 2002). Inadequate treatment of missingness with *ad hoc* methods, such as unconditional mean or regression imputation, may deflate or inflate the correlations among variable, and in general increase the rates of Type I error over the nominal levels by ignoring the increased uncertainty due imputation (Schafer and Olsen, 1998).

Missingness due to factors beyond the researcher's control is inevitable in experimental and some survey data, where the experiment or survey designer has maximum control over data measurement and recording. Schenker *et al.* (2011) discuss the considerable proportion of missingness in dual-energy x-ray absorptiometry (DXA) data in the National Health and Nutrition Examination Survey (NHANES) data, even though the physical measurements are carried out by trained professionals in Mobile Examination Centers. Survey data are also subject to missingness due to recording mistakes.

Data missingness can be characterized by the pattern and the mechanism. The missingness patterns can be combined in uniform, monotone and random groups (Little & Rubin, 2002, Schafer & Graham, 2002). The mechanisms giving rise to missingness are commonly referred to as Missing Completely at Random (MCAR), Missing at Random (MAR), and Not Missing at Random (NMAR). We define and review these concepts in the Theoretical Setup section below. Subject to the patterns and mechanisms of missingness, different imputation methods were pioneered by a body of empirical literature (Schenker *et al.*, 2011, Rubin & Schenker, 1991, Castillo, *et al.*, 2010, Schafer & Graham, 2002). These studies demonstrate advantages and disadvantages of different remedies to missingness using simulated or observational data.

In the literature the performance of different imputation methods are demonstrated using either simulated or observational data. There are some characteristics of both simulated and observational data that are noteworthy. In the case of simulated data, the data generation

mechanism is known to the researcher by definition. The number of data sets, as the sample sizes and the missingness pattern are a part of the design as well. The main advantage of using simulated data is that the validation of imputed values is well defined and readily measured: actual parameter values, means, variances and covariances are known and can be compared against. The disadvantage is that simulated data are not necessarily representative of practical cases, where the missingness occurs and needs to be addressed and effectively remedied. Schafer & Graham (2002), for example, use 1000 simulated datasets of 50 observations each, to demonstrate the performance of the traditional (case deletion, conditional and unconditional mean and distribution) and ML estimation and multiple imputation (MI) methods in MCAR, MAR and NMAR scenarios. They compare the means, variances and covariances of variables in the imputed dataset against actual counterparts used to generate the data. Rubin & Schenker (1991) use MI to reconcile occupation category codes in 1970 and 1980 censuses. They demonstrate that the confidence intervals around single imputation are too narrow; therefore the confidence coverages are lower than their nominal level.

With observational data, on the other hand, the data generation mechanism is largely unknown. Therefore all the parametric inferences are subject to correct model specification (Schenker *et al.*, 2011). There is typically one dataset being used for imputing missing values. The missingness patterns can be haphazard, thereby reducing the generality and applicability of the conclusions to other data, and the assumptions about the missingness mechanism are assumptions at best. The sample size restrictions and availability of covariates may or may not permit the use of the recommended imputation methods, and the validation methods are limited as there is no way of knowing their actual values to compare the imputed values against. Castillo *et al.* (2010) impute missing price variable by merging country level data sets, and demonstrate the superiority of Country-Product-Dummy model and the Theil-Goldberger mixed estimator to marginal mean imputation. Schenker *et al.* (2011) impute missing data in 32 different DXA variables in NHANES survey dataset; they perform multiple and single imputation, and suggest that imputation helps to correct biases and increase the precision of estimates as well. Kyureghian *et al.* (2011) go a step further and use multiple observational data sets to create a missingness pattern and mechanism to conduct analysis similar to ones using simulated data. They impute prices to ACNielsen (now the Nielsen) survey data (2,101 datasets in total) using several single and multiple imputation methods, and report the averages of the bias and coverage (percentage of times the confidence interval around the imputed values contains the true value) of each imputation method over the 2,101 datasets.

This study seeks to build on the research by Kyureghian *et al.* (2011) and proceeds by (1) placing more emphasis on the multiple imputation methods as more efficient and generalizable (Schafer and Olsen, 1998); (2) interpreting the performance of the imputation methods from the points of view of bias and distribution rather than bias and coverage, which are more intuitively associated with the premises of underlying imputation methods; and (3) associate the performance of the methods to the properties of the data sets such as the price distribution (mean, skewness and kurtosis) and sample size. We discuss and implement several multiple imputation methods by drawing on the advantages of both methods with simulated and observational data sets as much as possible. Initiated as a funded research to impute price variable to NHANES dietary recall data, this research identified and utilized a comparable data set, the Nielsen HomeScan data sets, to model the price for ingredients that would eventually be used to make up recipes or USDA food codes that the NHANES participants reported consumed. The inherent

richness of the HomeScan and NHANES datasets made it possible to have 2,101 datasets (one per each food ingredient), with an average number of approximately 20,000 observations in each dataset, ranging from 75 to almost half a million. We control the missingness mechanism and pattern by randomly deleting 20% and 50% of the price variable, thereby making the mechanism MCAR and the pattern uniform. We use Markov Chain Monte Carlo (MCMC), regression, propensity score matching and expectation maximization (EM) multiple imputation methods to impute the missing prices. The single regression or conditional mean imputation results are provided as benchmark due to its performance demonstrated by Kyureghian *et al.*, (2011). The validation is performed by mean absolute percent errors (MAPE) and two-sample Kolmogorov-Smirnov asymptotic tests for differences in distributions, for each method in 2,101 datasets.

Theoretical Setup

When faced with a missing value problem, practitioners need to address the missingness to force the dataset into a rectangular shape in order to use regular analytical methods. The choice between case deletion and imputation depends on the pattern and the mechanism of the missingness in the data (Little & Rubin, 2002). The former essentially describes which data are available and which are missing and group them by the patterns made by the missing data. The latter describes the relationship between the values in the data set and the event of missingness.

To illustrate these concepts, we define an indicator matrix with the same dimensions as the data set the elements of which take a value of 1 if the particular cell is observed, and 0 if the observation is missing. The matrix R that describes the probabilities of missingness, is a set of random variables with a joint probability distribution$P(R|Y,\varepsilon)$, where ε is an unknown parameter (Schafer, 1997). We also denote the data set as $Y_{com} = (Y_{obs}, Y_{mis}) = y_{ij}$, where $i = 1,\ldots,n, j = 1,\ldots,k$. By *iid* assumption, the probability density function of the complete data set is

$$P(Y|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

Using the notation above the missingness pattern can be described as univariate, monotone or arbitrary. Univariate missingness occurs when missingness is confined to a single variable such that one of the k variables was observed for only $1 < n_1 < n$ rows. Multivariate or monotone missingness occurs when a set of variables has missing values on a monotonically increasing number of rows. That is, the k variables in Y can ordered in such a way that if an observation is missing in Y_j, then it is missing in the subsequent $Y_{j+1}, \ldots, Y_k$ variables as well (Schafer & Graham, 2002). A more general pattern of missingness is when data are missing in an arbitrary pattern, e.g. survey non-response.

The mechanisms that give rise to missingness are missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). The conditional distribution of R given Y describes the underlying mechanism that gives rise to the missing data. If

$$P(R|Y_{com}, \varepsilon) = P(R|Y_{obs}, \varepsilon)$$

then the data are said to be MAR. This means that the probability that an observation is missing depend on other variables' values, but not on the missing datum value. The assumption that the missingness is not conditioned by the data, both missing and observed, is a strong assumption. In this case the data are said to be MCAR.

$$P(R|Y_{com}, \varepsilon) = P(R|\varepsilon)$$

In the case of NMAR, the probability of missingness depends on both observed and missing values. To demonstrate the concept of the missing data mechanisms, consider a dataset that has variables of age and gender. Suppose age has some missing observations. If the probability of the outcome of 'missing' of a certain age observation is independent of the values of age and gender of that observation, then the missingness mechanism is MCAR. If the probability of the outcome of 'missing' of a certain age observation is independent of the values of age, but is not independent of the values of gender for that observation, then the missingness mechanism is MAR. That is, the probability of not revealing age is higher if the respondent is a female, for example, but the age of that particular female does not condition the probability of not revealing age.

A last assumption we need to make is the distinctness of parameters θ and ε. This means that the joint parameter space of (θ, ε) is the product of parameter space of θ and parameter space of ε (Little & Rubin, 2002; Schafer, 1997). If MAR and distinctness hold, then missing data mechanism can be ignored when making likelihood based inferences about θ (Little & Rubin, 2002).

Missing Data Imputation Methods

Historically the missingness problems in incomplete data sets were addressed using *ad hoc* methods, such as case deletion or mean imputation. Case deletion, which implies that the observations with missing data points be removed from analysis altogether, has been very popular for its simplicity and lack of methods and computing power to handle missingness otherwise. This is the default option on many statistical programs (Schafer & Graham, 2002). The loss of information due to discarding observations has two aspects: loss of precision and bias. Although this method can be appropriate when the loss is minimal, usually under very stringent conditions (small proportion of missingness, MCAR and uniform pattern of missingness), Schafer & Graham (2002) argue that it is still inefficient.

Unconditional or conditional mean imputations imply replacing the missing values of a variable by the unconditional or conditional (regression) mean value of observed or non-missing observations for that variable. Once very popular because of computational simplicity, it is not a

desirable course of action (Kyureghian *et al.*, 2011). Even if it yields unbiased estimates (under MCAR), the other distributional properties are distorted: sample variance of the filled in data set underestimates the variance as a result of imputing missing values at the mean of the distribution. The conditional mean imputation method fits a regression model for the cases with non-missing observations. Then the estimated equation with the parameter estimates and known or non-missing covariates is used to predict the missing values of the dependent variable. Schafer & Graham (2002) do not recommend this method for analyses of covariances and correlations, because it overstates the relationship between the dependent variable and independent variables by imputing values on the regression line.

Multiple Imputation Methods

The above mentioned methods fill in the missing data gaps with single values. Multiple imputation methods, on the other hand, produce several values for each missing datum and imply uncertainty associated with filling in a value that was previously missing. Therefore, incorporating a random component in each imputed value rules out the uniqueness of the complete data set. Hence a vector of imputed values, rather than a single imputed value, is generated for each missing datum.

To generate multiple imputations by a parametric Bayesian model, following the notation in Schafer (1999), suppose $Y = (Y_{obs}, Y_{mis})$ with n_1 and n_0 number of observed and missing observations, respectively. Y follows a parametric model $P(Y|\theta)$, where θ has a prior distribution and the ignorability is satisfied. Then

$$P(Y_{mis}|Y_{obs}) = \int P(Y_{mis}|Y_{obs}, \theta)P(\theta|Y_{obs})d\theta$$

An imputation for Y_{mis} can be created by first simulating a random draw of the unknown parameters from their observed-data posterior

$$\theta^* \sim P(\theta|Y_{obs}),$$

followed by a random draw of the missing values from their conditional predictive distribution

$$Y_{mis}^* \sim P(Y_{mis}|Y_{obs}, \theta^*)$$

Since the observed data posterior is not typically easily simulated (Schafer, 1999), techniques such as Markov Chain Monte Carlo (MCMC) are utilized to create pseudorandom draws from probability distributions. The limiting marginal distribution of this sequence of random draws is the target distribution.

Multiple imputation method makes a use of the posterior distribution of the parameters to construct new parameters to calculate new fitted dependent variables. While it is desirable for the

single imputation to impute from conditional distribution $P(Y_{mis}|Y_{obs};\hat{\theta})$ where $\hat{\theta}$ is an estimate derived from observed data, multiple estimation first simulates m independent $\theta_1, \ldots, \theta_m$, then generating a m-vector of Y_{mis}^t from $P(Y_{mis}|Y_{obs};\theta_t)$, for *t = 1, …, m* (Schafer & Graham, 2002).

Let $\hat{Q}_j$ and U_j represent the estimated parameter and variance in j^{th} imputation. Rubin (1987) suggests combining estimates by averaging the vector of *m* estimates created by multiple imputation:

$$\bar{Q} = m^{-1}\sum_{j=1}^{m}\hat{Q}_j$$

The variance for $\bar{\beta}$ has two parts: within-imputation variance, denoted as $\bar{\sigma}$:

$$\bar{U} = m^{-1}\sum_{j=1}^{m}U_j$$

and between-imputation variance:

$$B = (m-1)^{-1}\sum_{j=1}^{m}\left(\hat{Q}_j - \bar{Q}\right)^2$$

The total variance is calculated as $T = \bar{U} + (1+m^{-1})B$. This estimator is distributed approximately as

$$\frac{\bar{Q}-Q}{T^{\frac{1}{2}}} \sim t_{\nu}, \qquad \text{where } \nu = (m-1)\left[1+\frac{\bar{U}}{(1+m^{-1})B}\right]^2$$

MI Methods

The Expectation Maximization (EM) algorithm capitalizes on the relationship between the missing values and parameters of a data model (Schafer, 1997; Schafer and Olsen, 1998). This two stem algorithm first predicts the missing values by maximizing the log likelihood function using initial assumed parameter values. In the second step it calculates the parameter estimates using both observed and imputed data from the first step, which are subsequently used in the first step to get new predictions for the missing values, and so forth. The resulting sequence of parameters converges to the maximum likelihood estimates.

Another iterative process, data augmentation (DA), that is used in Markov Chain Monte Carlo multiple imputation holds strong affinity to EM in that it too starts with an assumed parameter value to produce predictions for the missing data points. In the second step it uses both observed and imputed data to randomly draw new parameters from Bayesian posterior distribution which are fed into the first step of missing data imputation. This process creates a Markov chain of simulated data and parameters that eventually converge in distribution to the target distribution. The important difference between these two algorithms is that the in the convergence under EM parameters no longer change, but in the convergence under DA the distribution of parameters no longer changes (Schafer and Olsen, 1998).

Multiple regression imputation for a univariate normal linear regression model is performed first generating new stochastic parameters, β_* and σ_*, from the posterior predictive distribution of the parameter. By Bayesian calculations,

$$\sigma_*^2 = \frac{\hat{\sigma}_1^2(n_1-q)}{g} \text{ and } \beta_* = \hat{\beta}_1 + \sigma_*[V]^{1/2}Z,$$

where $g \sim \chi^2_{n_1-q}$, $n_1$is the number of non-missing values, q is the number of parameters to be estimated, $\hat{\beta}_1$ and $\hat{\sigma}_1^2$ are the parameter estimates based on the non-missing values, $V = [X'X]^{-1}$, and Z is a standard normal random variable. The missing values are subsequently filled in as $Y_{miss} = X\beta_* + Z\sigma_*$ (Rubin, 1997).

The propensity score method is another multiple imputation technique for monotone missing data by assigning each missing value to a particular group conditional to observed covariates. This method creates an indicator variable R_j that takes values of 0 if the j^{th} observation is missing, and 1 if it is observed. Then it fits a logistic regression where the dependent variable is the probability that the j^{th} observation is missing, conditional on a set of covariates with all observed values. Based on the predicted probabilities, the observations are grouped into groups. Approximate Bayesian Bootstrap is used to first draw, with replacement, a new set of $n_1$values from Y_{obs}, and then randomly draw $n_0$values from this new set to fill in the missing values in the group. This is repeated multiple times and the statistical averages are reported as the value of the missing datum.

Empirical Application

National Health and Nutrition Examination Survey (NHANES) is a database designed to assess the health and nutritional status of adults and children, and associate it with risk factors and prevalence of major diseases related to dietary intake. The dietary intake data are based on 24-hour recall of foods and quantities consumed. The dietary intake in NHANES 2001-2002 is recorded in *Food and Nutrient Database for Dietary Studies* (FNDDS), USDA, food codes. The Center for Nutrition Promotion and Policy (CNPP) broke down these food codes into ingredients by recipes provided by FNDDS. CNPP matched the ingredients with the foods commercially purchased, and therefore with observed prices, from a reference dataset. The reference dataset was identified to be The Nielsen Company HomeScan data. For example, a respondent in NHANES declares consuming beef stew at home. Since one cannot buy beef stew from a store, therefore there is no commercial price for beef stew. But one can buy the ingredients of beef stew and cook at home. Suppose FNDDS recipe calls for beef, carrots and onions for preparing beef stew. CNPP matched the ingredients in the recipes: beef, carrots and onions, to UPC's in the Nielsen, converted different measurements to a uniform measurement (grams), and provided weights for each ingredient needed to make up 100 grams of each recipe.

The Nielsen Company (formerly ACNielsen) recruits a representative panel of households from 48 contiguous states based on demographic characteristics. 8216 and 8685 households participated in the panel in 2001 and 2002, respectively. Each participating household is asked to scan/record each purchase made throughout each week. The households record the quantities purchased and amount paid for the purchase, and whether it was purchased at a regular or

discount price or a coupon was used. Observations from this data set for the UPC's identified by CNPP are used to model price as a function of quantity, poverty-income ratio (PIR, a variation of per capita income), region of residence, year and quarter of purchase, outlet status, and if the price was promotional or regular. Imputation techniques will be used to eventually merge the two datasets and impute the missing price variable for NHANES by combining the estimated ingredient prices by recipe weights. This study is a natural extension to the groundwork laid by Kyureghian *et al.* (2011), for analyzing the choice of the adequate imputation method to impute prices to NHANES dietary intake data.

Variables

The choice of variables was guided first and foremost the desire to obtain maximum possible explanatory power (Schafer, 1997, pp. 138-143; Rubin & Schenker, 1991). Although possible collinearity issues may arise for the analyst models using NHANES imputed price variable in the case of single conditional mean imputation, we suppressed this issue here and do not perform sensitivity analysis for different model specifications.

The price variable is expressed in cents per 100g. Since this variable takes only strictly positive values, it is typically right skewed. To conform to normality a log transformation is used in the models, and transformed back after the imputation (Schafer & Graham, 2002). Incidentally, this will also guarantee that the imputed values are strictly positive. Quantities are expressed in 100g units. PIR's are calculated as the midpoint of categorical Income variable divided by government issued poverty levels by household size:

$$PIR = \frac{\frac{Upper\ Bound + Lower\ Bound}{2}}{\$8{,}590 + \$3{,}020 \times (Household\ Size - 1)}$$

The region of residence is defined by the four census regions: East, Central, South, and West. The fourth quarter of the year and the West region are left out of the models as the base quarter and region. Store types provided in the data are grouped as (i) grocery, (ii) drug, (iii) mass merchandiser, (iv) supercenters, (v) club, (vi) convenience and (vii) all other. We combined them in Grocery (base), Club (includes (iii), (iv), and (v)), and Other (includes the rest).

The ingredients with sample size of less than 75[1] observations were not used in this study. There were several individual transactions where the price for an item was recorded as zero. As mentioned above, in these cases the items were purchased as a part of a deal. We did not attempt to examine these instances on a case-by-case basis due to overwhelming sample size, the insignificance of the proportions of zero price transactions and the potential complexity of promotional deals. Therefore individual transactions with zero prices were discarded. There were 42,114,592 observations for 2,101 ingredients in the Nielsen dataset. The sample sizes vary from 75 to 472,204 with an average sample size of 20,045.

[1] The Nielsen Company recommends this threshold for calculating average prices. We take the number 75 at face value. In the initial analysis these ingredients were assigned the overall mean price value for the particular ingredient.

Table 1. Summary Statistics of the Variables in the Model

Variable	Description	Mean	Standard Deviation	Median	Minimum	Maximum
Price	Cents per 100g	49.49	48.48	41.62	0.03	43,638.20
Quantity	Quantity in 100g	8.52	13.57	4.80	0.01	7,838.21
PIR	Poverty Income Ratio	4.08	2.47	3.66	0.09	14.55
Promo Deal	Equals to 1 if deal, and 0 o/w	0.40	0.49	0	0	1
Club	Equals to 1 if purchased from Club, and 0 o/w	0.11	0.31	0	0	1
Other	Equals to 1 if purchased from Other, and 0 o/w	0.05	0.22	0	0	1
Central	Equals to 1 if purchase is in the Central region, and 0 o/w	0.20	0.40	0	0	1
East	Equals to 1 if purchase is in the East region, and 0 o/w	0.22	0.41	0	0	1
South	Equals to 1 if purchase is in the South region, and 0 o/w	0.37	0.48	0	0	1
Year2002	Equals to 1 if purchase is in 2002, and 0 o/w	0.51	0.50	1	0	1
Quarter1	Equals to 1 if purchase is in quarter 1, and 0 o/w	0.26	0.44	0	0	1
Quarter2	Equals to 1 if purchase is in quarter 2, and 0 o/w	0.25	0.43	0	0	1
Quarter3	Equals to 1 if purchase is in quarter 3, and 0 o/w	0.24	0.43	0	0	1

Imputation

To create controlled missingness in our data, subsamples of 20% and 50% for each ingredient were selected and the values of the price variable were removed from those observations. Since these subsamples were selected purely randomly and the selection did not depend either on the values of other variables or on the values of the removed prices for corresponding observations, the missingness mechanism of MCAR was literally imposed on the data. Imposing missingness on one variable – price, resulted in the Univariate pattern of missingness.

We illustrate the performance of Marcov Chain Monte Carlo (MC), Regression (Reg), Propensity Score (Prop), Expectation Maximization (EM) MI methods. Conditional mean (CM) imputation results are used as benchmarks (Kyureghian et al., 2011). The appropriate numbers of imputations for each missingness scenario were determined by the efficiency of m imputations relative to one based on infinite number of imputations, expressed as

$$\left(1+\frac{\lambda}{m}\right)^{-1}$$

where λ is the rate of missingness (Rubin, 1987). We committed our models to 95% efficiency, rendering $m = 4$ and $m = 10$ imputations for our 20% and 50% missingness scenarios,

respectively. Since the choice of the default imputation number in the software used is 5 imputations, we settled with 5 and 10 imputations for the 20% and 50% missingness scenarios, respectively.

The price variable is naturally skewed to the right, which is at odds with the normality assumption for the multiple imputations. To make this normality assumption plausible, we used the natural logarithmic transformation of the price at the imputation stage, which was transformed back to the original scale afterwards. Schafer and Olsen (1998) maintain the robustness of MI to mild departures from normality.

Results

For each imputation method-missingness proportion combination two types of validation methods were used. The first method deals with measuring the goodness or distance of each imputed datum from the true or observed value of that datum. To measure this we used the mean percent absolute error (MAPE). The second method deals with the goodness of imputation measured by matching the distributions of the missing and imputed values. Summary statistics, graphical and regression analyses for the price, MAPE and distribution match are used in each case.

The summary statistics of the price variables of the full sample without missing observations as well as those imputed by different methods are presented in Table 2. The price imputations by multiple imputation exhibited erratic behavior at the extreme values of the price distribution. To rule out unrealistic or nonsensical imputed values we imposed a price maximum value of twice the observed maximum price. Since the maximum of the observed price was $436.38 per 100 g, the imputed prices above $872.76 per 100g were truncated at that value. As can be seen from the results in Table 2, MC and EM were the only two methods affected by this restriction.

With the exception of CM, the imputation methods do a remarkably good job at preserving the mean of the price distribution. CM systematically undershoots the mean value, although it demonstrates the lowest standard deviation among the methods. While very well behaved at the center of the distribution, MC and EM have the highest standard deviations even after truncating at the maximum value.

Although all the methods seem to undershoot the means and overshoot the medians in both missingness scenarios, the means and medians migrated closer to each other in 50% missingness scenario. The standard deviations for MC and EM have increased, while standard deviations for other methods have decreased in 50% missingness scenario. CM, by definition overstates the correlations between variables by imputing values on the regression line. Hence the unreasonably low standard deviation in CM imputed prices.

In general, the distributions of prices in two missingness scenarios across all the imputation methods appear to be remarkably similar, which, in part, might be attributed to MCAR mechanism of missingness. The graphical display of the distribution of imputed prices vs. actual price for the subsamples with missing prices were not revealing and were therefore omitted.

Table 2. Summary Statistics of Full and Imputed Price Variables

	Mean	Median	STD	Min	Max
Price Full	**49.49**	**41.62**	**48.48**	**0.03**	**43638.20**
20% Missingness					
MC	49.46	42.99	51.54	0.00	84,616.20
Reg	49.45	42.99	46.58	0.00	43,638.20
Prop	49.50	43.21	46.91	0.04	43,638.20
EM	49.48	42.99	47.95	0.00	70,253.10
CM	48.71	42.77	45.78	0.00	43,638.20
50% Missingness					
MC	49.42	44.91	52.20	0.00	87,276.40
Reg	49.40	44.91	41.61	0.00	43,638.20
Prop	49.49	45.99	40.56	0.04	43,638.20
EM	49.42	44.92	54.17	0.00	81,289.20
CM	47.54	44.09	39.16	0.00	43,638.20

Defined as the absolute difference between the imputed and observed values as percentage of the observed value, the MAPE's are a useful method to measure the distance between the imputed value and the true or observed value. Figures 1 and 2 depict the kernel distributions of MAPE's for 20% and 50% missingness scenarios, respectively[2].

The comparison across the graphs, once again, reveals that at higher missingness rates all five distributions roughly preserve the shape. From the graphs it is clear that the CM imputation clearly outperforms other imputation methods. MC, EM and MR have similar performance and although they have more mass to the right from the CM distribution, they clearly outperform propensity score method.

The distributional characteristics of MAPE in Table 3 demonstrate what was visually apparent in Figures 1 and 2 above. The CM imputation results the lowest mean MAPE in 20% and 50% missingness scenarios, demonstrating precision of being off the true value by approximately 28% only, or 28 cents in each dollar. Although the comparison of the sensitivity of imputation methods to model specification was not a part of the objective of this research, we can draw from and compare to other research in this area. For example, Kyureghian (2009), compares conditional and unconditional mean imputation precision, using the same data sets from the Nielsen data as this research, but uses a much more parsimonious model specification that accounts for regional and seasonal variations in price only. The average MAPE reported in that research is 34-35% compared to 28% found in this study. Given the *ceteris paribus* condition, we attribute this decline to the choice of variables in the model.

[2] Again, for the convenience of visual display we restricted the MAPEs to values less than or equal to 80. As a result the graphs do not show the performance of the methods at the extreme values. The remaining samples represent the 1996 (95.00%) and 1988 (94.62%) of 2101 food codes for the 20% and 50% missing subsamples, respectively.

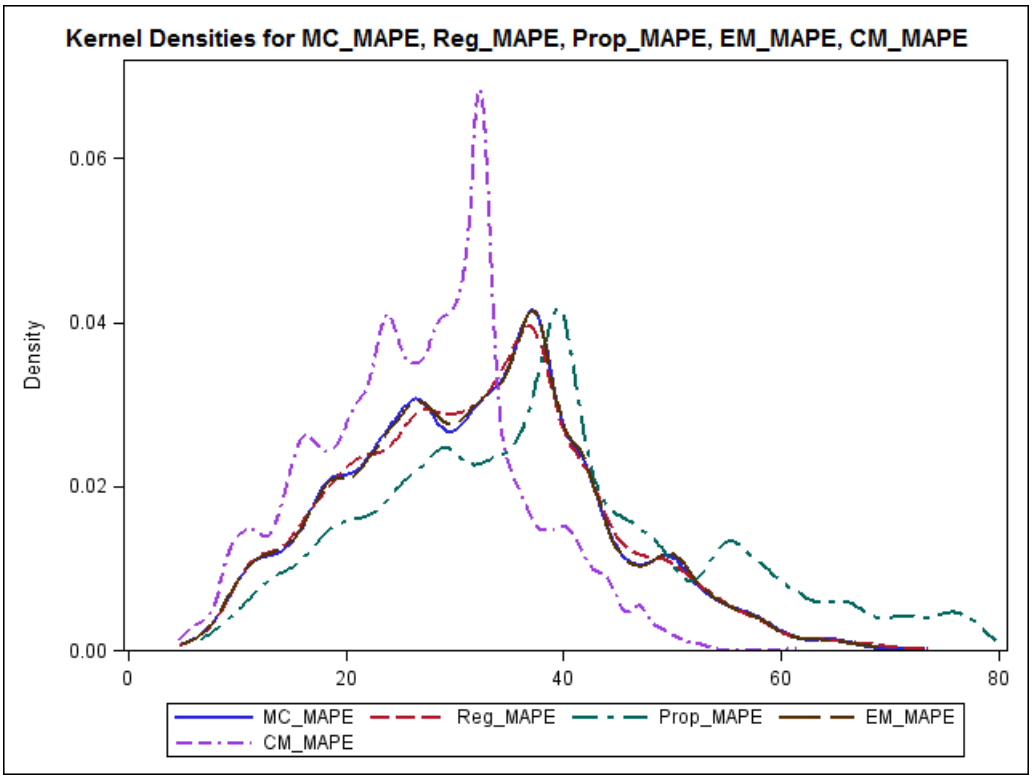

Figure 1. Kernel Densities the Mean Absolute Percent Errors (MAPE) for Imputation Methods for 20% Missingness.

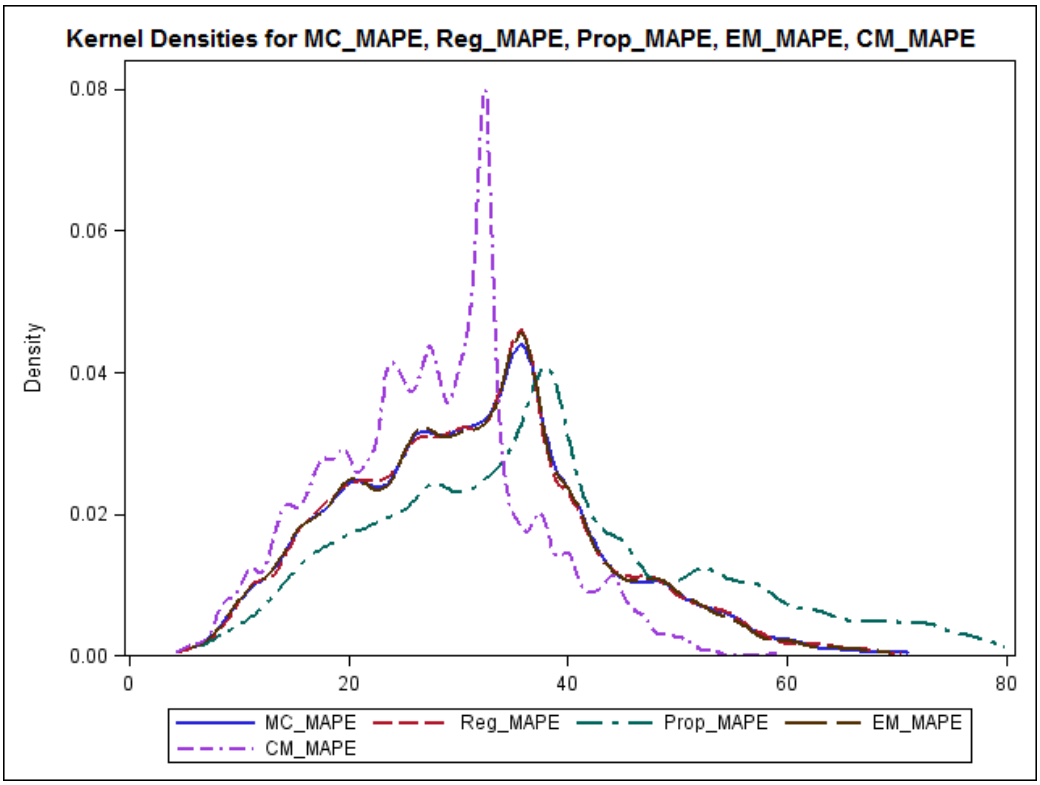

Figure 2. Kernel Densities the Mean Absolute Percent Errors (MAPE) for Imputation Methods for 50% Missingness.

The MAPEs for the multiple regression method – Reg, are very close to CM results. MC and EM have outlier problems again, giving rise to very large standard deviations and means. The graphical display of these methods, which ignores the behavior at the tails of distributions, reveals that although very erratic at the tails, these methods can be reasonably good. This is also evidenced by the proximity of MC and EM distributions to Reg distribution. Provided the generality of these methods, they appear to provide evidence of superiority over the CM method.

Table 3. Mean Absolute Percent Errors (MAPE) of Imputation Methods.

Statistic Name	Mean	Median	STD	Min	Max
20% Missingness					
MC_MAPE	86.25	33.15	1,234.00	4.93	42,047.02
Reg_MAPE	34.49	33.10	23.00	4.74	497.17
Prop_MAPE	42.49	38.26	40.01	6.57	1,260.87
EM_MAPE	51.98	33.11	761.67	5.05	34,915.78
CM_MAPE	28.20	28.02	12.39	4.65	203.07
50% Missingness					
MC_MAPE	59.25	32.21	461.53	5.48	14,299.31
Reg_MAPE	33.48	31.84	22.05	4.47	457.28
Prop_MAPE	41.42	37.01	43.96	6.33	1,512.75
EM_MAPE	61.20	32.19	505.49	5.65	17,574.67
CM_MAPE	28.50	28.09	12.25	4.35	202.08

The second line of the comparison of imputation methods is based on the distribution match between the imputed and missing values for the five imputation methods for both missingness scenarios. We calculated two-sample asymptotic Kolmogorov-Smirnov tests statistic D as

$$D = \max_j \left| F_1(x_j) - F_2(x_j) \right|,$$

where $j = 1,2,\ldots,2101$ and $F(x)$ is the empirical distribution function.

The p-value for this test is the probability that D is greater the observed d under the null hypothesis of no difference between the distributions of missing and imputed. An auxiliary statistic – Diff, which is a binary variable that measures if D is significant at 1% level, is constructed to help interpret D across 2101 tests. Diff equals to 1 if the null is rejected and 0 otherwise. The summary statistics are reported in Tables 4 and 5 below.

The results indicate overall low match between missing and imputed distributions. EM, MC and Reg have the highest probabilities of failing to reject the no difference in both missingness scenarios. Propensity score method has the lowest average probability of failing to reject the null.

Table 4. Summary Statistics of P-values of the Asymptotic Kolmogorov-Smirnov Statistic

Statistic Name	Mean	Median	STD	Min	Max
20% Missingness					
MC_KS	0.0573	0	0.1585	0	0.9971
Reg_ KS	0.0566	0	0.1559	0	0.9718
Prop_ KS	0.0249	0	0.0896	0	0.9251
EM_ KS	0.0579	0	0.1613	0	0.9898
CM_ KS	0.0375	0	0.1261	0	0.9639
50% Missingness					
MC_ KS	0.0154	0	0.0713	0	0.8386
Reg_ KS	0.0139	0	0.0612	0	0.6744
Prop_ KS	0.0029	0	0.0221	0	0.5038
EM_ KS	0.0154	0	0.0737	0	0.9272
CM_ KS	0.0129	0	0.0677	0	0.9272

Table 5. Summary Statistics of the Auxiliary Statistic Diff

Statistic Name	Mean	Median	STD	Min	Max
20% Missingness					
MC_KS	0.7587	1	0.4280	0	1
Reg_ KS	0.7625	1	0.4257	0	1
Prop_ KS	0.8391	1	0.3675	0	1
EM_ KS	0.7611	1	0.4265	0	1
CM_ KS	0.8296	1	0.3761	0	1
50% Missingness					
MC_ KS	0.8948	1	0.3069	0	1
Reg_ KS	0.8967	1	0.3044	0	1
Prop_ KS	0.9676	1	0.1770	0	1
EM_ KS	0.8948	1	0.3069	0	1
CM_ KS	0.9148	1	0.2792	0	1

This is not surprising as this imputation method uses the covariate information only to associate it with whether the variable is missing or not. CM does a poor job at retaining the distributional properties which is closely related to the nature of this method. The results in Table 5 are similar to those in Table 4. CM and Prop have the highest proportion of instances that the null is rejected at 1% confidence level in both missingness scenarios. The other three methods performance is roughly similar. As expected, in the 50% missingness scenario the performance of all methods is strictly worse. The kernel densities of the p-values of this statistic are presented in Figures 3 and 4 below. To maintain the null of no difference between the distributions, we want these density functions away from 0. The graphical results support the summary evidence from Tables 5 and 6.

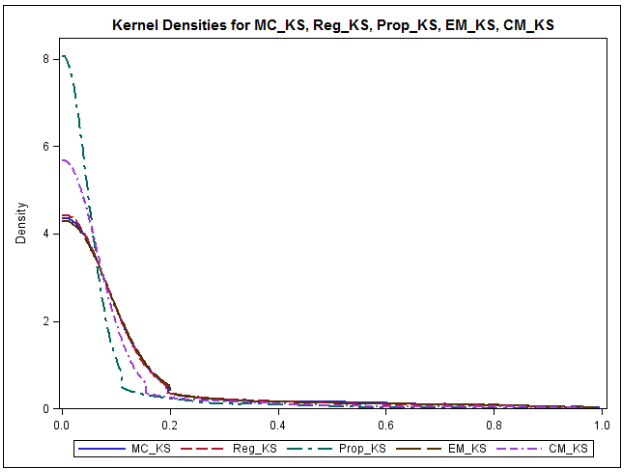

Figure 3. Kernel Densities the Kolmogorov-Smirnov (KS) Statistic for Imputation Methods for 20% Missingness.

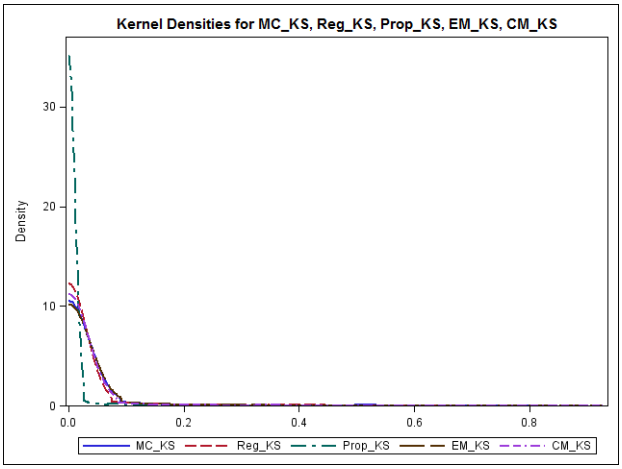

Figure 4. Kernel Densities the Kolmogorov-Smirnov (KS) Statistic for Imputation Methods for 50% Missingness.

Finally we would like to be able to generalize our findings on the statistics created for performance check. That is we would like to associate the properties of our data sets to these statistics and draw inference rather than report averages and other summary statistics. For this reason we regress certain data properties against the validation statistics: MAPE, the p-value of the statistic Kolmogorov-Smirnov statistic - P_KSA, and Diff. The advantage of this approach is rendering a richer framework to interpret partial correlations conditional upon data properties, such as the average price level, skewness, kurtosis and sample size. This meta-analytic approach will produce empirical evidence to be generalizable to the extent as to help other researchers that face the problems of missingness to associate their method of choice with the properties of their data sets.

For this purpose we created a data set where each observation corresponds to one food code or original dataset. The entries are the MAPE, KS and Diff for that particular dataset (these are the dependent variables), average price level, skewness, kurtosis and the sample size (covariates). Therefore, the resulting final data set has 2,101 observations. The regression estimates, along with the fit statistics are reported in Tables 6 and 7.

Three models were estimated for each imputation model:

$$MAPE_{ji} = \beta_{j0} + \beta_{j1} \times Price_{ji} + \beta_{j2} \times SK_{ji} + \beta_{j3} \times KT_{ji} + \beta_{j4} \times Size_{ji} + \epsilon_{ji} \quad (1)$$

$$P\_KSA_{ji} = \beta_0 + \beta_1 \times Price_i + \beta_2 \times SK_i + \beta_3 \times KT_i + \beta_4 \times Size_i + \epsilon_i \quad (2)$$

$$Diff_{ji} = \beta_0 + \beta_1 \times Price_i + \beta_2 \times SK_i + \beta_3 \times KT_i + \beta_4 \times Size_i + \epsilon_i \quad (3)$$

where $i = 1,2, \ldots, 2101, j \in \{MC, Reg, Prop, EM, CM\}$. Therefore 15 equations were estimated altogether. Models in (1) and (2) were estimated by the least square and the model in (3) was estimated by logit estimation methods.

Table 6. MAPE Regression Parameter Estimates (P-values) for the Imputation Methods in 20% and 50% Missingness Scenarios.

	MC	Reg	Prop	EM	CM
20% Missingness					
Intercept	113.4159	29.8222	31.3107	58.7528	25.2895
	(0.0015)	(<.0001)	(<.0001)	(0.0078)	(<.0001)
Mean Price	-0.1665	0.0358	0.1316	-0.0451	**0.0152**
	(0.4702)	(<.0001)	(<.0001)	(0.7510)	**(<.0001)**
SK	-0.8756	0.2799	0.4973	-0.0676	**0.1804**
	(0.8127)	(<.0001)	(<.0001)	(0.9764)	**(<.0001)**
KT	0.0025	-0.0008	**-0.0013**	0.0002	-0.0005
	(0.8239)	(0.0002)	**(0.0001)**	(0.9738)	(<.0001)
Sample Size	-0.0035	0.0004	0.0004	-0.0009	**0.0003**
	(0.3493)	(<.0001)	(0.0013)	(0.6828)	**(<.0001)**
F Value	0.37	31.34	100.20	0.07	42.36
	(0.8327)	(<.0001)	(<.0001)	(0.9920)	(<.0001)
R^2	0.0007	0.0564	0.1605	0.0001	0.0748
Adj R^2	-0.0012	0.0546	0.1589	-0.0018	0.0730
50% Missingness					
Intercept	69.2058	28.5486	28.7971	72.3879	25.4222
	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)
Mean Price	-0.0344	0.0433	0.1595	-0.0423	**0.0195**
	(0.6896)	(<.0001)	(<.0001)	(0.6542)	**(<.0001)**
SK	-0.4922	0.2629	0.4325	-0.5391	**0.1752**
	(0.7216)	(<.0001)	(0.0003)	(0.7216)	**(<.0001)**
KT	0.0013	-0.0007	**-0.0011**	0.0015	-0.0005
	(0.7478)	(0.0002)	**(0.0018)**	(0.7475)	(<.0001)
Sample Size	-0.0006	0.0001	0.0001	-0.0007	0.0001
	(0.2704)	(<.0001)	(0.0037)	(0.2702)	(<.0001)
F Value	0.40	41.70	121.54	0.41	48.20
	(0.8101)	(<.0001)	(<.0001)	(0.8036)	(<.0001)
R^2	0.0008	0.0737	0.1883	0.0008	0.0842
Adj R^2	-0.0011	0.0719	0.1867	-0.0011	0.0825

Given the fact that, in general, we are interested in reducing (or at least not increasing) MAPE and Diff, and increasing (or at least not decreasing) P_KSA, we would favor methods that have the largest negative (or smallest positive) coefficient in the first two models and the largest positive (or smallest negative) coefficients in P_KSA model across each covariate.

In the MAPE model (Table 6) the results indicate that the MAPEs under MC and EM methods are not associated with data properties in any meaningful way, in both missingness scenarios. The signs and significance of the parameters in other methods are remarkably similar. In all of these methods higher level of average prices are associated with higher MAPEs. In other words, the percentage error of prediction increases for higher-priced foods. Similarly the error of prediction increases if the data are skewed, which is quite normal as these methods impute at the conditional mean, therefore skewness would distort any measure of the proportional departure from the mean. Logically, as data become more peaked and thinner tailed or kurtosis increases, the MAPEs decrease. Interestingly, the prediction error increases with the sample size. These results are robust to the change in the level of missingness. The best method choices across the relevant covariates are boldfaced in Table 6.

Table 7. Kolmogorov-Smirnov P-value Regression Parameter Estimates (P-values) for the Imputation Methods in 20% and 50% Missingness Scenarios.

Variables	MC		Reg		Prop		EM		CM	
	P_KSA	Diff	P_KSA	Diff	P_KSA	Diff	P_KSA	Diff	P_KSA	Diff
20% Missingness										
Intercept	0.0758	0.6790	0.0748	0.6816	0.0334	0.7879	0.0771	0.6817	0.0523	0.7763
	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)
Mean Price	**0.0001**	**-0.0003**	0.0001	-0.0002	0.0000	-0.0002	0.0001	-0.0002	0.0000	-0.0002
	(0.0087)	**(0.0009)**	(0.0103)	(0.0048)	(0.1211)	(0.0019)	(0.0155)	(0.0013)	(0.9812)	(0.0007)
SK	-0.0022	0.0085	-0.0021	0.0084	**-0.0010**	**0.0057**	-0.0022	0.0085	**-0.0013**	**0.0060**
	(<.0001)	(<.0001)	(<.0001)	(<.0001)	**(0.0003)**	**(<.0001)**	(<.0001)	(<.0001)	**(0.0007)**	**(<.0001)**
KT	0.0000	-0.0000	0.0000	-0.0000	0.0000	-0.0000	0.0000	-0.0000	0.0000	-0.0000
	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.0013)	(<.0001)	(<.0001)	(<.0001)	(0.0025)	(<.0001)
Sample Size	-0.0000	0.0000	-0.0000	0.0000	-0.0000	0.0000	-0.0000	0.0000	-0.0000	0.0000
	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)
F Value	28.30		28.16		16.11		27.85		16.97	
	(<.0001)		(<.0001)		(<.0001)		(<.0001)		(<.0001)	
R^2	0.0512	0.1170*	0.0510	0.1138*	0.0298	0.0731*	0.0505	0.1155*	0.0314	0.0784*
Adj R^2	0.0494		0.0492		0.0280		0.0487		0.0295	
Mean		0.7587		0.7625		0.8391		0.7611		0.8296
		(0.4280)		(0.4257)		(0.3675)		(0.4265)		(0.3761)
Sigma		0.4021		0.4006		0.3537		0.4011		0.3609
		(<.0001)		(<.0001)		(<.0001)		(<.0001)		(<.0001)
50% Missingness										
Intercept	0.0177	0.8579	0.0189	0.8680	0.0040	0.9555	0.0188	0.8585	0.0163	0.8955
	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(<.0001)
Mean Price	**0.0001**	-0.0001	0.0000	-0.0002	0.0000	-0.0000	0.0000	-0.0001	0.0000	**-0.0003**
	(<.0001)	(0.0433)	(0.2291)	(<.0001)	(0.8995)	(0.5865)	(0.0005)	(0.0334)	(0.0171)	**(<.0001)**
SK	-0.0007	0.0045	-0.0006	0.0045	-0.0001	0.0013	-0.0006	0.0044	**-0.0005**	**0.0035**
	(0.0017)	(<.0001)	(0.0009)	(<.0001)	(0.1008)	(0.0177)	(0.0031)	(<.0001)	**(0.0108)**	**(<.0001)**
KT	0.0000	-0.0000	0.0000	-0.0000	0.0000	-0.0000	0.0000	-0.0000	0.0000	-0.0000
	(0.0062)	(<.0001)	(0.0037)	(<.0001)	(0.1469)	(0.0361)	(0.0096)	(<.0001)	(0.0251)	(0.0002)
Sample Size	-0.0000	0.0000	-0.0000	0.0000	-0.0000	0.0000	-0.0000	0.0000	-0.0000	0.0000
	(<.0001)	(.)	(<.0001)	(.)	(0.0027)	(.)	(<.0001)	(.)	(<.0001)	(.)
F Value	15.70		11.20		3.31		11.97		8.73	
	(<.0001)		(<.0001)		(0.0103)		(<.0001)		(<.0001)	
R^2	0.0291	0.0474*	0.0209	0.0527*	0.0063	0.0127*	0.0223	0.0472*	0.0164	0.0464*
Adj R^2	0.0272		0.0191		0.0044		0.0205		0.0145	
Mean		0.8977		0.8967		0.9676		0.8948		0.9148
		(0.3032)		(0.3044)		(0.1770)		(0.3069)		(0.2792)
Sigma		0.2949		0.2962		0.1758		0.2945		0.2726
		(<.0001)		(<.0001)		(<.0001)		(<.0001)		(<.0001)

* The R^2 values are calculated as the squared Pearson correlation between the observed and the predicted dependent variable values.

Ceteris paribus, with the increase of the price level and skewness, the increase in MAPEs is much milder in models imputed under the CM method compared to Reg and especially to Prop. On the other hand if the data happen to have higher kurtosis, Prop appears to be a better choice for reducing the prediction error, ceteris paribus. The effect of the sample size is rather uniform across these methods.

In 20% missingness scenario the results from the P_KSA models reveal that, with the exception of MC, the average price level has no effect on the probability of the missing and

imputed distributions being not different at 1% confidence level. MC seems to do a better job of mimicking the distribution for higher-priced foods. As expected, the more skewed the data, the lower the probability that any imputation method will preserve the distributional likeness in predictions. Nevertheless, skewness under the Prop and CM methods has the minimal negative effects on P_KSA. Increase in kurtosis will on the other hand improve the probability of distributional match. Just as the increased sample size increased the error in prediction in the MAPE models it decreases the probability of distribution match in this case. Although the coefficients of these two covariates are significant in majority of the estimations, the parameter estimates are virtually identical in all cases; therefore these variables do not increase our ability as to the choice of the method. Consequently, we will not consider them any further.

In conclusion, for both missingness levels MC outperforms the other methods for higher-priced foods, and Prop does a better job in skewed data. In the 20% missingness scenario CM does a comparably good job in imputing values to skewed data. These models have reasonably good fits witnessed by the F-values.

In 20% missingness scenario the kurtosis and sample size variables had identical parameter estimates for all of the methods, and therefore were excluded from consideration as in P_KSA models above. The results from the Diff models favor choice of methods identified in P_KSA models: in higher-priced food data MC does a better job of decreasing the probability of the distributions being different. Prop and CM do comparably good job when the data are skewed. Interestingly, in 50% missingness scenario, the Prop models exhibit a rather poor fit. In this missingness scenario the CM truly stands out and outperforms the other methods where both the average price level and skewness are concerned.

The auxiliary variable, sigma, reported in the outcomes indicates the improvement of the standard error of the mean dependent variable, and is the equivalent of the root mean square error in the models with continuous dependent variable. For comparison purposes with calculated the R^2 values for these models as the squared correlations between the fitted value and the actual value of the dependent variable. Prop and CM consistently have the lowest fit levels in both scenarios.

The estimation results for P_KSA and Diff models are reported in Table 8. The best method choices across the relevant covariates are boldfaced.

A note of warning is in place before we conclude. The definition of the higher-level dependent variable may seem not well defined and subject or data specific, and therefore not very well applicable to other researches or data in general. Nonetheless it is hard to deny its usefulness and generalizability for food price data, and the Nielsen HomeScan data in particular. Therefore, other researchers using consumer level food consumption data may still find the results insightful. Also a reminder is in place that the results from imputation methods based on explicit models are subject to model specification.

Discussion and Conclusion

The focus of this research is to examine, validate and recommend techniques for handling the problem of missingness in observational data. The widespread nature of the problem of missingness and the wide variety of idiosyncrasies of different data sets make such recommendations incredibly hard and of little practical value. With this in mind, we tackle the problem using a rich observational data set, the Nielsen data, which effectively combines elements from simulated data sets: large number of observations and variables allowing elements of 'design' that typically come with simulated data, and is observational in nature.

We created 20% and 50% missingness in our datasets and employed several widely used multiple imputation methods to fill in the data gaps. We then compared these methods by utilizing the mean absolute percent error of prediction and Kolmogorov-Smirnov two-sample test statistic for testing the null hypothesis of no difference in distributions of missing and imputed values. The summary statistics of imputed prices reveal that the prices imputed under the MI methods consistently perform better at preserving the full sample mean price level. But these methods perform rather erratically at the tails of the distribution and have the higher standard deviations compared to the prices imputed under CM. When comparing the percentage difference between the true value and imputed value for each data point, we clearly arrive at CM as the method that on average outperforms the others. It comes as no surprise that on average MC, Reg and Prop perform better in mimicking the distributions, than CM and Prop.

We extend our results from the summary statistics to regression inference by estimating the association between MAPE and distributional match and data properties. Based on our results, if the researchers interested in increased prediction precision and facing missingness in data with higher level dependent variable and increased skewness would be advised to opt for CM imputation method in both 20% and 50% missingness scenarios. Data exhibiting increased level of kurtosis would be better predicted by Prop. On the other hand, if the researchers are interested in preserving the distributions of the missing data rather than improving the precision of the prediction, MC is the method to adopt for data with higher-level dependent variable and Prop is the method if the data are skewed, in 20% missingness scenario. In data sets with larger proportion of the data missing, CM appears to perform better from both high-level dependent variable and skewness points of view.

References

Castillo, M.J., Useche, P.P., & Moss, C.B. (2010). Missing agricultural price data: an application of mixed estimation. *Applied Economics Letters*, 17, 537-541.

Durrant G.B. (2009). Imputation methods for handling item-nonresponse in practice: methodological issues and recent debates. *International Journal of Social Research Methodology*,12(4), 293-304.

Kyureghian, G. (2009). Methodology and applications in imputation, food consumption and obesity research. Doctoral dissertation. Texas A&M University.

Kyureghian, G., Capps, O. & Nayga, R.M. (2011). A missing variable imputation methodology with an empirical application. *Advances in Econometrics: Missing Data Methods*, Volume 27, 2011: Forthcoming.

Little, R.J.A. & Rubin, D.B. (2002). Statistical analysis with missing data. New York: John Wiley & Sons, Inc.

Rubin, D.B. (1987). Multiple imputation for nonresponse in surveys. New York: John Wiley & Sons, Inc.

Rubin, D.B. & Schenker, N. (1991). Multiple imputation in health-care databases: an overview and some applications. *Statistics in Medicine*, 10, 585-598.

Schafer, J.L. (1997). Analysis of incomplete multivariate data. London: Chapman & Hall.

Schafer, J.L. (1999). Multiple imputation: a primer. *Statistical Measures in Medical Research,* 8, 3-15.

Schafer, J.L. & Graham J.W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7, 147-177.

Schafer, J.L. & Olsen M.K. (1998). Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivariate Behavioral Research*, 33(4), 545-571.

Schenker, N., Borrud, L.G., Burt, V.L., Curtin, L.R., Flegal, K.M., Hughes, J., Johnson, C.L., Looker, A.C., & Mirel, L. (2011). Multiple imputation of missing dual-energy X-ray absorptiometry data in the National Health and Nutrition Examination Survey. *Statistics in Medicine*, 30, 260-276.