Evaluation of Patterns of Missing Prices in CPI Data

November 2018

Harold Gomes
U.S. Bureau of Labor Statistics (BLS)
2 Massachusetts Ave NE, Room 3655, Washington, D.C. 20212
Gomes.Harold@bls.gov

Abstract
The validity of an imputation method to represent missing data depends on the assumptions made about the underlying data and the mechanism of missingness. In this empirical investigation, patterns of missing prices in the Consumer Price Index (CPI) microdata are evaluated, i.e., missing prices in relations to other covariates (auxiliary variables) are assessed. CPI is an official statistic that measures U.S. inflation and is estimated based on a multistage probability sample design. Price of a quote (item) is the variable of interest for the CPI target population, collected monthly from a representative market basket. CPI microdata are used to evaluate the missingness mechanism: Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR). Exploratory analysis, statistical tests, and data visualization are used in this study. A few important benefits of this research are: 1) to examine the validity of current imputation methods that use group means imputation with periodic updates to the group definitions; 2) to identify variables related to missingness in MAR situations; 3) to provide potential recommendations for future improvement.

Key Words: Consumer Price Index (CPI), missing data, imputation, Missing Completely at Random (MCAR), Missing at Random (MAR), Missing Not at Random (MNAR).

Any opinions expressed in this paper are those of the author and do not constitute policy of the Bureau of Labor Statistics.

1. Introduction

The Consumer Price Index for All Urban Consumers (CPI-U), often referred to as the U.S. "inflation rate", is a weighted-average of price change (in percent) over time in the prices of consumer items—goods and services—of the urban U.S. population. It is weighted by the expenditure of consumer items.

1.1 Consumer Price Index (CPI) Survey Design

CPI has been estimated through a scientific sampling method, a multistage probability sample design, since 1978. Price of a quote is the response variable, collected from an outlet within a Primary Sampling Unit (PSU), located in urban Areas representing the urban U.S. population (target population). CPI is an area-based sample with 211 Item Strata (Item) and 87 Primary Sampling Units (PSU) [2018 Area redesign has 75 PSU]. Quotes are nested into 211 Items, and the 211 Items are nested into 8 Major Groups. Similarly, Outlets (a store to attain a quote price) are nested into 87 PSUs, and 87 PSUs

are nested into 38 geographic areas (Index Areas or Areas) [2018 Area redesign has 32 Areas]. Many single PSUs also represent an Area (self-representing). CPI survey comprises two sampling components: 1) commodities and services sampling (C&S), about 70% of the CPI weight; 2) housing sampling, the remaining 30% of the CPI weight. In practice, C&S and Housing are two separate surveys combined into a final data repository before official index computation (BLS Handbook of Methods, Chapter 17).

1.2 Historical Perspective on Current CPI Imputation

While this study is not about the imputation methodology but on discovering the missing pattern, historical records and knowledge are incorporated to uncover a few perspectives on the current CPI imputation methodologies.

Reference to "Imputation Procedure" can be found as early as the 1966 edition of "BLS Handbook of Methods for Surveys and Studies", a decade prior to the implementation of the probability sampling methodology (1978). As with many surveys, a missing price is generated when a quote-price is not attained for the month from a specific outlet (or a city back then). Here is an excerpt from the 1966 edition:

> "Although prices are not obtained in all 56 cities every month,... it is necessary to represent all 56 cities in each monthly index computation.
> ...For new automobiles, a price change is imputed to the unpriced cities on the basis of changes in cities surveyed every month."

This confirms that the concept and practice of imputation is not new in the production of official statistics or in the discipline of statistics.

1.3 Motivation for this Study

(1) Imputation

Currently, CPI uses Group-Means imputation with periodic updates to the group definitions. In order to improve an imputation method, one must diagnose missing pattern as the first step and then treat (cure) the missing data—impute—as the second step. Although research on imputation has been conducted in the past, no studies have focused on *diagnosing* MCAR, MAR, and MNAR situations of missing pattern in the CPI micro data.

(2) Realized Sample Size Estimates

What proportion of prices are missing across *different* item strata, major groups, and PSUs? Inspection of this question of sample size, realized every month in the microdata, provided another motivation for this study. It enables an empirical estimate of the proportion of missing price (vs. collected price) across different item strata and major groups.

(3) Covariates

What are the potential covariates related to missingness? This question of what survey variables are associated with a price of a quote being missing provided another motivation for this study.

(4) Data Visual to Display Missing vs. Collected
How to display the distributions of missing prices for survey variables, such as Major Group, Item Strata, PSU?

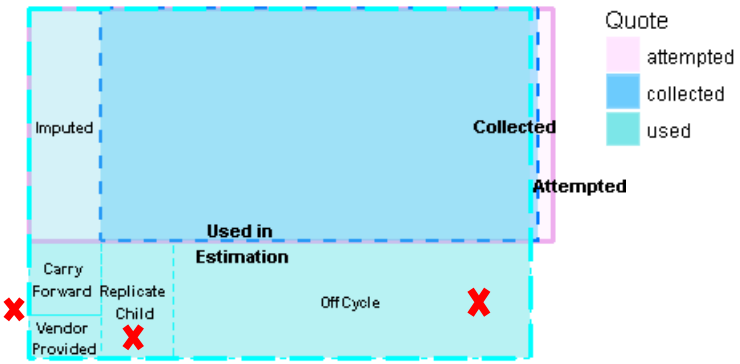
This study is **not about evaluating the imputation methodologies**, but on assessing the missing pattern in the CPI micro data and discovering the missingness mechanism.

2. Defining Missing Data and the Study Design

2.1 CPI Realized Sample: Price Attained from Multiple Pathways and Sources

The quote is the *Ultimate Sampling Unit* for which a price is attained from the target population (urban), such as a price of a banana from New York. CPI is not only a multistage sample design but also a multiple-data-frame sample design. Currently, monthly *realized* samples in CPI, known as, Used in Estimation, obtain prices from multiple pathways, from which the final official statistic is produced. The figure 1 shows a Venn diagram displaying the union, intersection and disjoint sets of different types of quotes. The attempted set of quotes are sent out for collection through field staff. From this set of quotes, the price is either successfully collected or not collected. Many quotes are *not* expected to change price frequently and, in order to optimize budget/resources, these quotes are priced *bimonthly* in most PSUs, and these quotes are known as *Offcycle* when not priced and the prices are forwarded from last month to the current month for index computation. Carry Forward quotes are also similar to Offcycle. Some vendors directly supply the data frames with quote-prices, hence identified as Vendor Provided data in this diagram. For a few items, quote-price is attained for a subset of PSUs (e.g., postage data, used car) and then the data is replicated to fill out other PSUs nested within an Index Area, and they are known as Replicate Child.

Figure 2.1: Venn Diagram of CPI Quote



2.2 Current Empirical Study Design

(1) Imputed vs. Collected

In this empirical study, quotes that are **Imputed** from the Attempted set and Used in Estimation (official statistic), are defined as "Imputed". Hence, missing data in this study refers to these Imputed quotes **excluding** the OffCycle, Carry Forward, Replicate Child, and Vendor Provided quotes that are due to sample design. We will refer to these quotes as *Design Quotes* throughout this paper for simplicity. Since the price of a quote is the

response variable of the target population, taking the intersection of Attempted and Used sets also ensures that there will be a price for certain, whether Imputed or Collected, enabling evaluation of the results. Full Sample Effective-Price-with-Tax is the response variable of the CPI target population.

(2) Dataset (Nov and Dec 2017)

In order to eliminate the *bimonthly* collection effect (even and odd) on the missing pattern, November and December 2017 data is used for Modeling and Evaluation.

(3) Cross-sectional Study

This is a *cross sectional* study and not a longitudinal study. One cross section or a complete set of quotes is composed of components from 2 months' sample sizes, i.e., monthly quotes (Dec'17) + even quotes (Dec'17) + odd quotes (Nov'17)

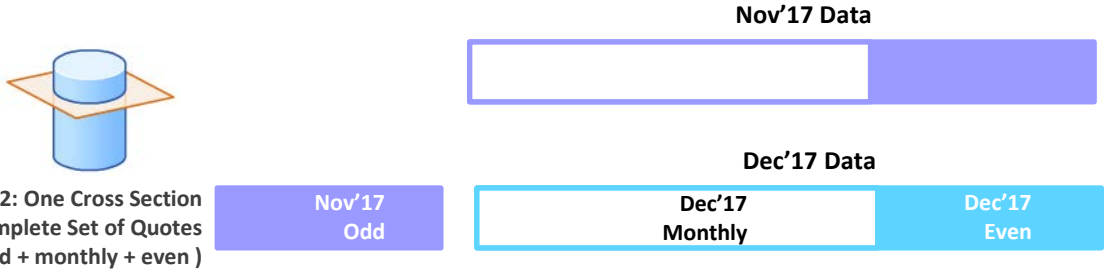


Figure 2.2: One Cross Section Complete Set of Quotes (odd + monthly + even)

(4) C&S only (excludes housing)

This study consists of the commodities and services component (C&S) only and excludes the housing component. There are 179 priced Item Strata, out of which 3 Item Strata samples are part of the Vendor Provided data. Hence, this study includes 176 Item Strata.

3. Exploratory Evaluation of Patterns of Missing Prices

Distributions are compared between Collected and Imputed quotes in various perspectives in order to evaluate the patterns of missing prices, presented with corresponding data visuals.

To evaluate price distributions of Collected and Imputed quotes, *aggregate* (Fig. 3.1.A) and *disaggregate* (by Major Group; Fig. 3.1.B) histograms and boxplots are produced with a single month data. December 2017 data is used representing monthly + even quotes.

To diagnose the pattern of missingness across a few important survey variables, such as Major Group, Item Strata, Primary Sampling Unit, data visuals are generated displaying the pattern in Imputed versus Collected quotes. This is one cross section so that the proportion displays a complete set of quotes, i.e., monthly quotes (Dec'17) + even quotes (Dec'17) + odd quotes (Nov'17).

3.1 Distribution of Missing Prices of CPI Quote

In this section, price-distributions are displayed—*aggregate* and *disaggregate* (by Major Group)—with corresponding median and mean (dot on the boxplot) for a single month (monthly + even quotes). Dashed lines correspond to the mean and median of the distributions (vertically coincides with boxplots).

Figure 3.1.A: Monthly Sample Distribution of CPI Quote-Price [Dec 2017 Data]

Magnified distribution with limited *x*-axis ($0-$300) **Distribution with Mean displayed as a dot on the boxplot**

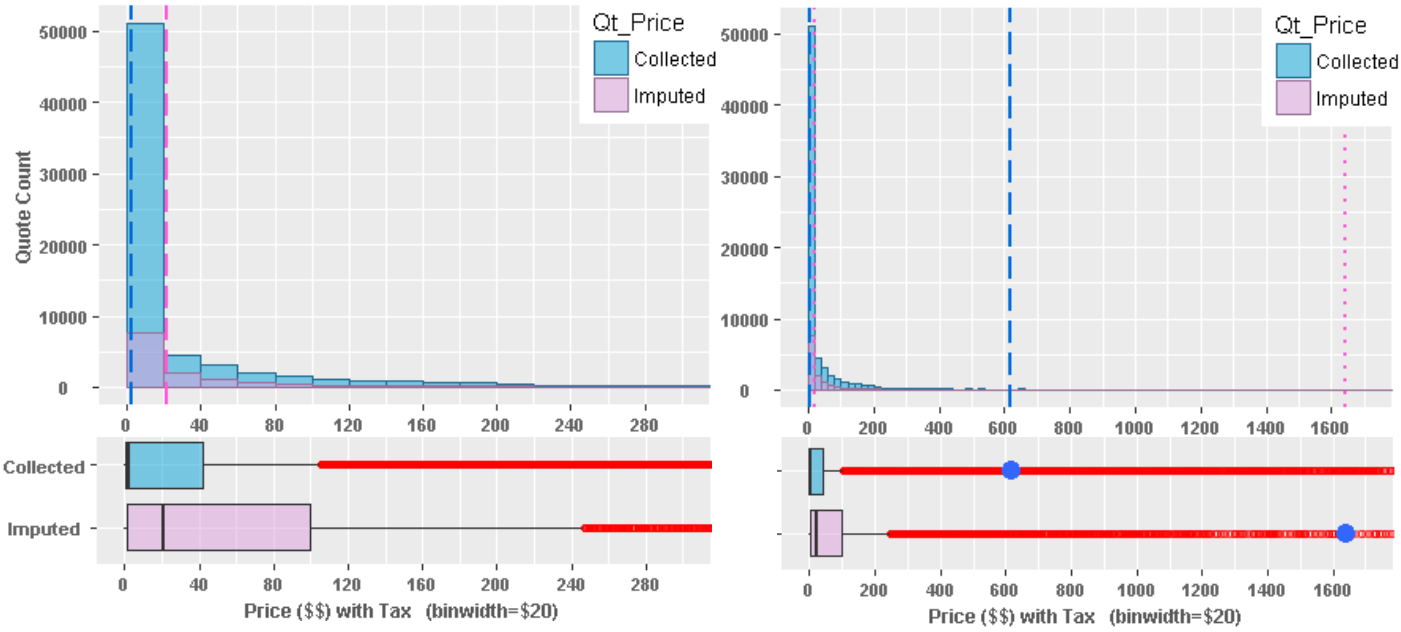
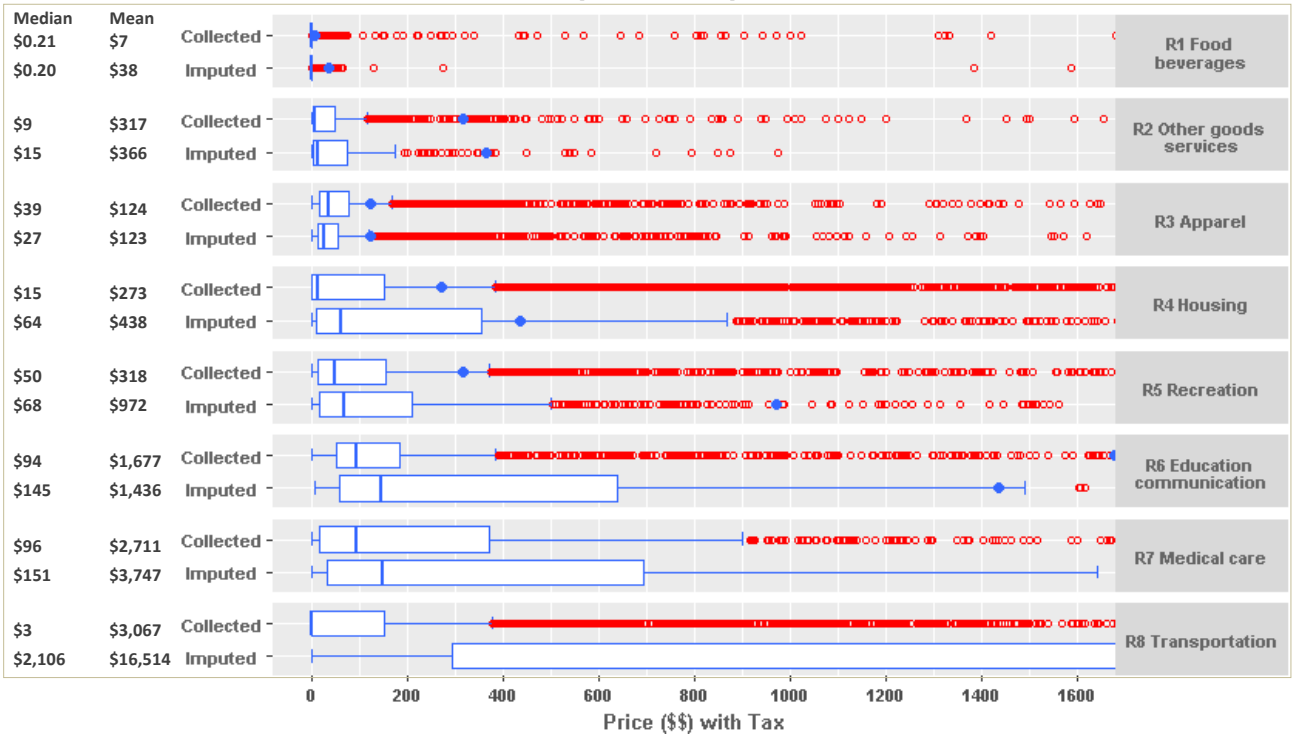


Figure 3.1.B: Distribution of 8 Major Groups, ranked by median-imputed (low to high) [Dec 2017 Data]



Observations from Price Distributions

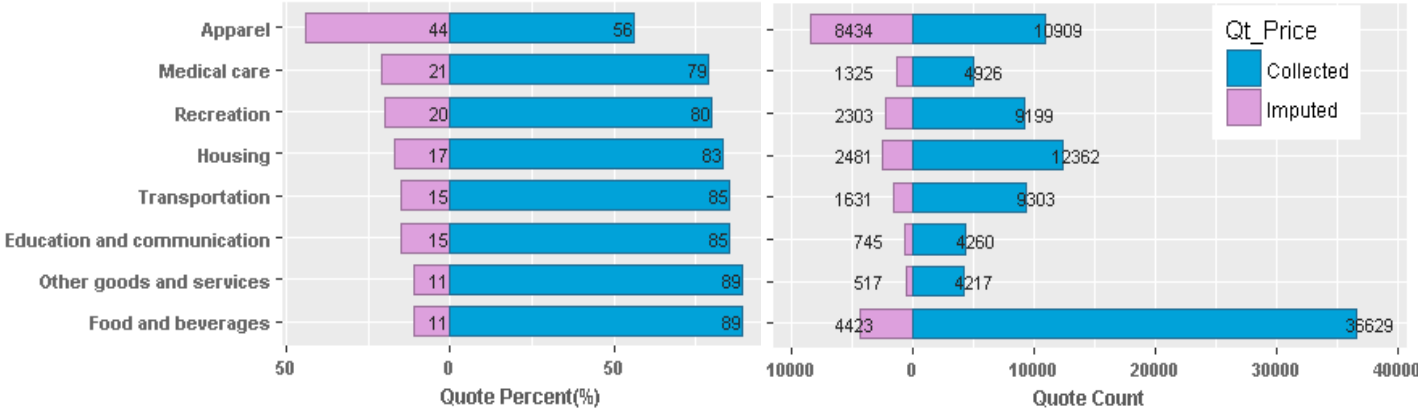
- ✓ Price displays a highly skewed distribution (positively skewed; Fig. 3.1.A) with a very long right tail. Price range goes up to ~$800,000 (e.g., Hospital Services). A few New Vehicles are also on the tail (~ $220,000).
- ✓ Sample average and median prices for imputed quotes are higher than collected for *All-items-All-area* 176 priced Item Strata.
- ✓ Each Major Group has a distinct sample average and median price (Fig. 3.1.B), varying from the overall sample mean and median. For example, the sample average price for collected quotes are higher than imputed quotes for Education and communication, and Apparel Major Groups, while the opposite is true for the rest.
- ✓ 83% (74,312) quotes are collected and 17% (15,418) quotes are imputed based on the December 2017 data. This sample size (n) *excludes* the Design Quotes and reflects only the monthly + even collection cycle.

3.2 Missing Prices as a function of *Major Group, Item Strata, and PSU*

In these sections, the proportion of quotes (percent) and number of quotes (count) are displayed as a function of a few important survey variables, in order to assess whether the distributions of imputed and collected are uniform across Major Groups, Items and PSU. Data represents one cross section, a complete set of quotes (i.e., monthly quotes (Dec'17) + even quotes (Dec'17) + odd quotes (Nov'17)).

3.2.1 Missing Prices as a function of Major Group

Figure 3.2.A: Distribution of 8 Major Groups, ranked by percent-imputed
[One Cross Section: odd (Nov'17) + monthly (Dec'17) + even (Dec'17)]

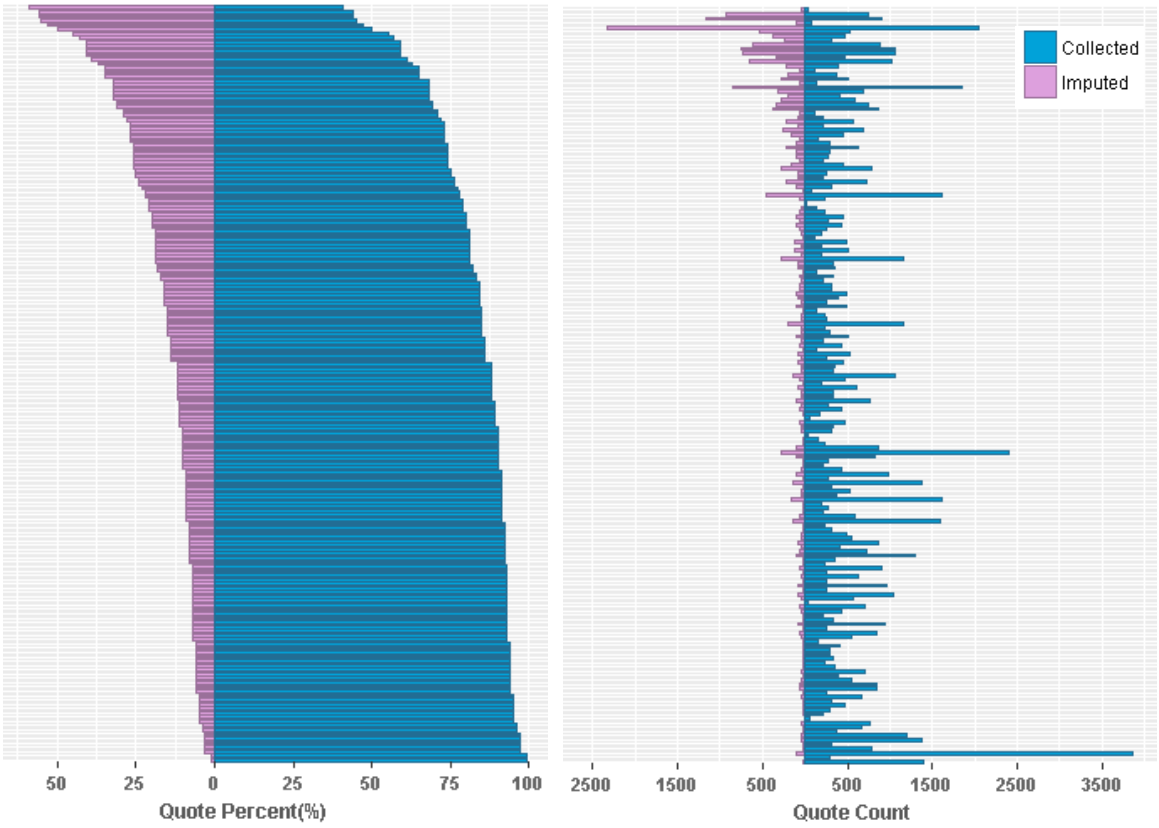


Observations from Major Group (3.2.1)

- ✓ Based on one cross section, a full dataset (odd [Nov'17] + monthly [Dec'17] + even [Dec'17]), the proportion of imputed quotes ranges from 11% to 44% depending on the Major Group.
- ✓ Apparel displays the highest proportion (and count) of imputed quotes. Food and beverages displays a low proportion but second highest of imputed quotes based on count.
- ✓ The cross section indicates 81% (91,766) collected and 19% (21,898) imputed quotes, a total sample size of 113,664 quotes.

3.2.2 Missing Prices as a function of Item Strata

Figure 3.2.B: Distribution of 176 Item Strata, ranked by percent-imputed
[One Cross Section: odd (Nov'17) + monthly (Dec'17) + even (Dec'17)]



Observations from Item Strata (3.2.2)

- ✓ Highly sparse (unbalanced) *realized sample sizes* across Item Strata, ranges from 20 to 4,000 quotes in an Item Stratum.
- ✓ Highly sparse (unbalanced) *imputed* quotes within an Item Stratum, ranges from 1% to 60% quotes being imputed depending on the Item Strata.
- ✓ The next sets of graphs display the tail of the distribution (magnified) in order to examine the item strata with lowest and highest imputed quotes (Figs. 3.2.C-3.2.F).

Figure 3.2.C: Distribution of Top 10 Items, ranked by *percent-imputed*
[One Cross Section: odd (Nov'17) + monthly (Dec'17) + even (Dec'17)]

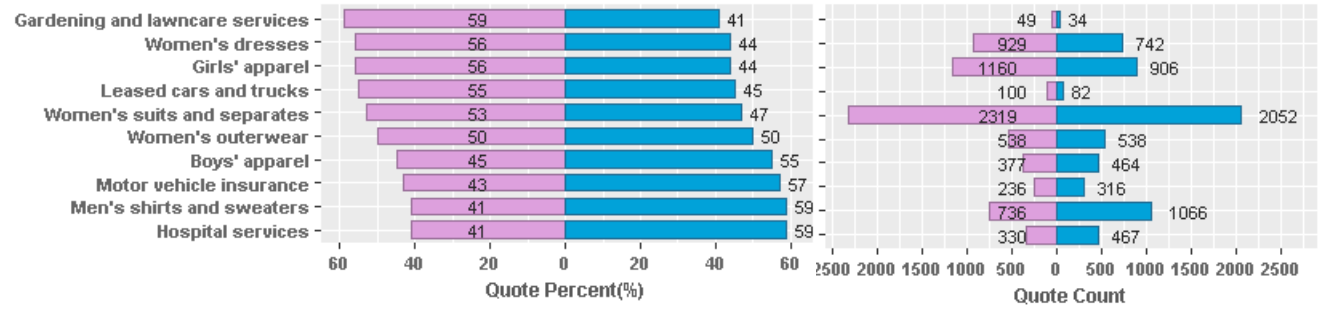


Figure 3.2.D: Distribution of Top 10 Items, ranked by *count-imputed*
[One Cross Section: odd (Nov'17) + monthly (Dec'17) + even (Dec'17)]

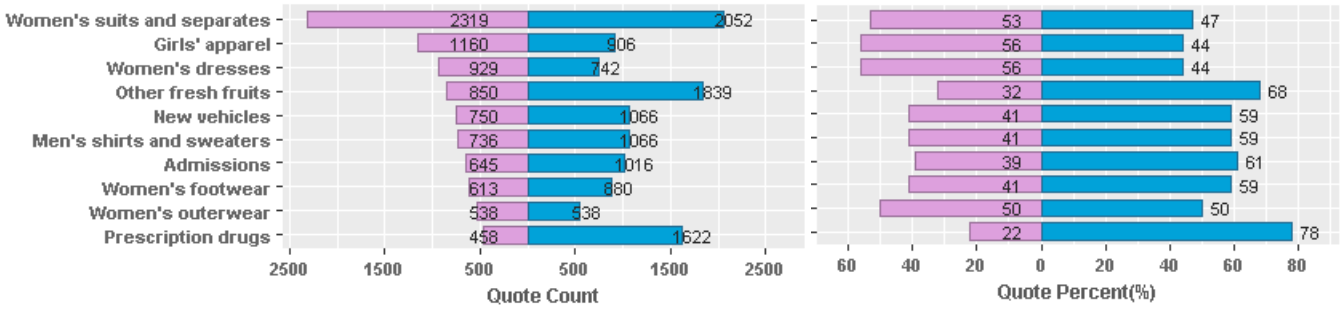


Figure 3.2.E: Distribution of Bottom 10 Items, ranked by *percent-imputed*
[One Cross Section: odd (Nov'17) + monthly (Dec'17) + even (Dec'17)]

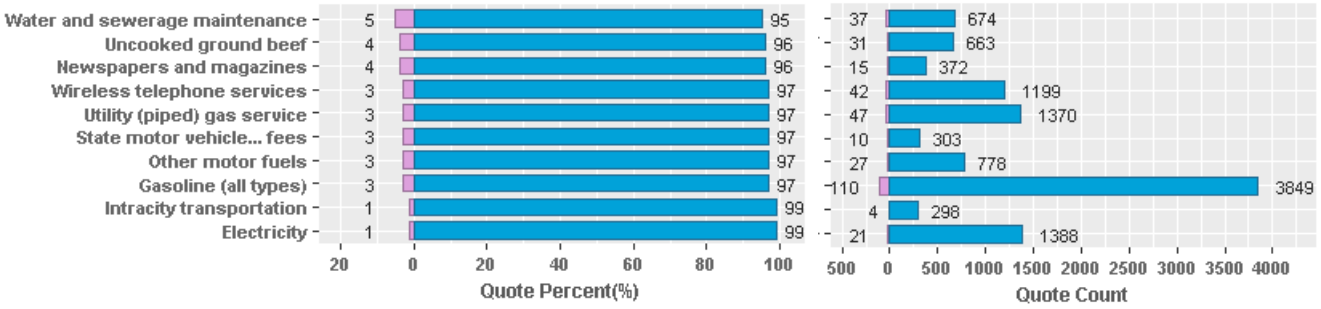
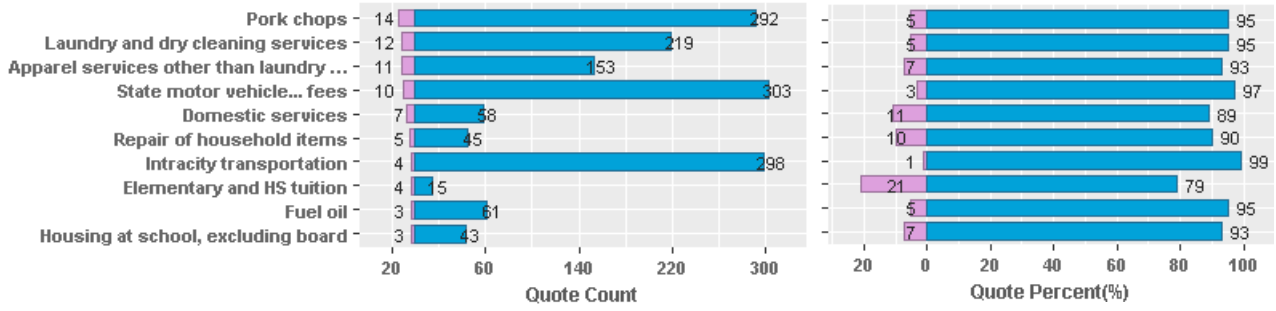
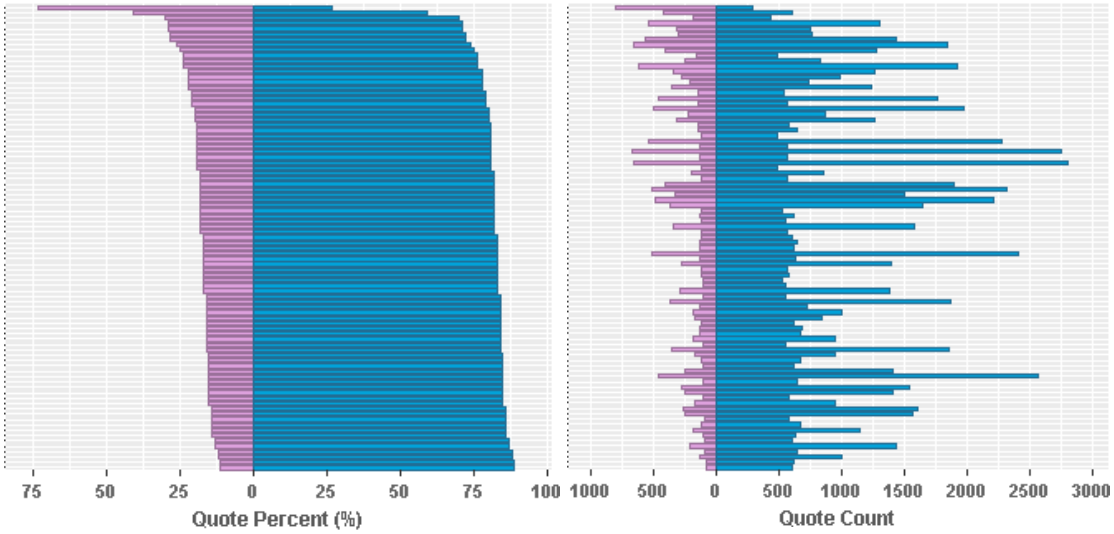


Figure 3.2.F: Distribution of Bottom 10 Items, ranked by *count-imputed*
[One Cross Section: odd (Nov'17) + monthly (Dec'17) + even (Dec'17)]



3.2.3 Missing Prices as a function of Primary Sampling Unit

Figure 3.2.G: Distribution of 87 PSUs, ranked by percent-imputed
[One Cross Section: odd (Nov'17) + monthly (Dec'17) + even (Dec'17)]



Observations from Primary Sampling Unit (3.2.3)

- ✓ Highly sparse (unbalanced) *realized sample sizes* across PSU, ranges from 80 to 3,400 quotes in a PSU.
- ✓ Sparse (unbalanced) *imputed* quotes within a PSU, most PSUs ranges from 11% to 30% quotes imputed (one PSU with 73% imputed). This is not as highly sparse as the Item Strata distribution.

Since the sample sizes are unbalanced across item strata, PSU, and Major Group, looking at the percent-imputed measures may be misleading. For example, "Food and beverages" and "Other goods and services" show the same proportion of imputed quotes, 11%. However, looking at the count-imputed distribution, we see that the Food and beverages display 9 times more missing quotes (~ 4500 quotes imputed) than Other goods and services (~500 quotes imputed). Hence, count-imputed distributions are also produced to generate an equitable context.

3.3 Remarks from Exploratory Analysis

Missing price as a function of Major Group, Item Strata and PSU does not display uniform distributions (imputed versus collected), suggesting a potential relationship may exist between these survey variables and the underlying missingness mechanism. That is, propensity for a missing price *varies within* item strata, PSU and Major Group. Formal evaluation (modeling) needs to be employed to examine this relationship while controlling for other covariates.

4. Missing Data Generating Process: MCAR, MAR, and MNAR

There are 3 processes that could generate missing values in data (Rubin, 1976; Schafer, 1997; Little and Rubin, 2002). They are: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). Each process has its own assumptions. The three charts below summarize each situation (Molenberghs et at, 2015, Grace-Martin, K.)

Missing Completely at Random (MCAR)

- missing data is *not related* to values of any survey variables, whether missing or observed.
- propensity for a data point to be missing is completely random. There's NO relationship between *whether* a data point is missing and any values in the data set, missing or observed.
- e.g., A quote-price of a random item was not collected because the outlet was closed on a collection day.
- Evaluate and diagnose the following Probability Model:

$$P(y_i = \text{missing} \mid X, X^*) = P(y_i = \text{missing})$$

X = Observed Covariates
X = Unobserved Covariates*

Missing at Random (MAR)

[Missing conditionally at Random]

- missing data is conditionally related to the values of another survey variable (covariates, auxiliary var.)
- propensity for a data point to be missing is not related to missing data, but related to some of the observed data of other variables.
- e.g., A quote-price of a specific type of item is mostly missing over other items.
- Evaluate and diagnose the following Probability Model:

$$P(y_i = \text{missing} \mid X, X^*) = P(y_i = \text{missing} \mid X)$$

X = Observed Covariates
X = Unobserved Covariates*

Missing Not at Random (MNAR)

- missing data is related to its values.
- there is a relationship between the propensity of a data point to be missing and its values (same variable).
- e.g., A quote-price of a specific item is missing for cases when price is expensive.
- True way to evaluate is to conduct a follow-up survey to non-respondents, and if non-respondents answer very differently than respondents, that's a good evidence for MNAR.
- Evaluate and diagnose the following Probability Model:

$$P(y_i = \text{missing} \mid X, X^*) = P(y_i = \text{missing} \mid X^*) \text{ OR}$$
$$P(y_i = \text{missing} \mid X, X^*) = P(y_i = \text{missing} \mid y_i = c)$$

X = Observed Covariates
X = Unobserved Covariates*

4.1 Diagnosing Missing Data Mechanism

There is no single *litmus* test to diagnose the missing data mechanism when dealing with real data. Instead, there are a variety of techniques, statistical methodologies, employed to diagnose MCAR, MAR, and MNAR situations. Diagnosing missing data mechanisms is more like gathering a body of evidence to determine the underlying process, especially when dealing with limited scope and real survey data (Grace-Martin, K; Molenberghs et at, 2015).

Also, null-hypothesis based testing (such as, Little's test), may provide significant results (p-value) for studies with large sample sizes (such as this study; n=113,664) when distributions are compared between missing and collected data, even if there is no practical significance.

If diagnostic results indicate MNAR situation, it eliminates the possibility for MCAR and MAR situations. If diagnostic results indicate MAR situation, it eliminates the possibility for MCAR but not for MNAR. If diagnostic results indicate neither MNAR nor MAR, only then the mechanism could be concluded as MCAR (Molenberghs et at, 2015, Grace-Martin, K). Using this premise and the hierarchy of logic, formal evaluation methods are developed to examine the MCAR, MAR and MNAR situations.

5. Formal Evaluation of Missing Price Mechanism

One way to investigate missingness mechanisms is to find potential covariates (auxiliary information) in the CPI Survey that must be common to both (common support region), imputed and collected quotes. Also, it is important to note that response (price) could be missing in the dataset but the covariates should not have any missing entry. Otherwise these cases need to be deleted (case-deletion) or some type of proxy measure has to be used in missing observations.

A probability model (propensity) is developed in the context of missing price for formal evaluation of missingness mechanisms.

5.1 Probability of a Missing Quote-Price (Observed or Empirical)

Probability of an event MP, missing price, estimated from observed data is defined as:

$$P(MP) = \frac{n_{event\ MP}}{N_{event\ total}}$$

$$Probability(MissingPrice) = \frac{Number\ of\ Quote_{missing}}{Number\ of\ Quote_{missing} + Number\ of\ Quote_{not-missing}}$$

$$Probability(MissingPrice) = \frac{Number\ of\ Quote_{imputed}}{Number\ of\ Quote_{imputed} + Number\ of\ Quote_{collected}}$$

$$\boldsymbol{Probability(MissingPrice) = Proportion_{imputed}} \quad \textbf{...Eqs (1)}$$

Odds (Ω) of an event MP, missing price, is defined as:

$$\boldsymbol{\Omega}(MP) = \frac{P(MP)}{1 - P(MP)} = \frac{\pi}{1 - \pi}$$

$$Odds(MissingPrice) = \frac{P(MissingPrice)}{1 - P(MissingPrice)} = \frac{Proportion_{imputed}}{1 - Proportion_{imputed}} = \frac{Proportion_{imputed}}{Proportion_{collected}}$$

5.2 Probability of a Missing Quote-Price as a Function of Covariates (Predicted)
Probability of an event, missing price, estimated from covariates is defined as:

$$\boldsymbol{Probability(MissingPrice) = f(Commom\ Covariates) + Error} \quad \textbf{...Eqs (2)}$$

This could be deployed using binomial regression model (Eqs. 3).

5.2.1 Multiple Logistic Regression: Diagnose Missing at Random (MAR)

In binomial regression, commonly known as logistic regression or logit model, the probability of a success (or failure) is related to some explanatory variables ($x_1,\ldots,x_p$) and the response (missing price) is a categorical variable (yes | no). *Missing* implies *imputed* price and *not-missing* implies *collected* price.

Multiple logistic regression equation for a missing price:

$$Log(Odds\ of\ MissingPrice) = ln\left[\frac{Proportion_{imputed}}{1 - Proportion_{imputed}}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

$$\frac{Proportion_{imputed}}{1 - Proportion_{imputed}} = exp\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\right)$$

$$Proportion_{imputed} = \frac{exp\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\right)}{1 + exp\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\right)}$$

$$\boldsymbol{Probability(MissingPrice) = \frac{exp\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\right)}{1 + exp\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\right)}} \quad \textbf{...Eqs (3)}$$

Multiple logistic regression is deployed to diagnose *Missing at Random* i.e., missing price is conditionally related to the values of another survey variable (covariates, auxiliary var.) (see section 4).

5.3 Potential Survey Covariates (Auxiliary Variables x_p)
Eleven (11) variables were selected as potential candidates of covariates/auxiliary information after inspecting the variables in CPI microdata repositories. During this process of initial variable selection, prior knowledge of CPI survey design and data collection process were considered. For example, if a quote is seasonal, it might have a higher probability of being missing than others; or if a PSU has a lot of vacancies of data collectors, this PSU may generate a lot of missing quote-price due to overload of work; or if some item strata in apparel is sold out before re-pricing again, this item strata may exhibit a higher probability of missing price than other item strata. This is often the reality of data collection in the survey paradigm. Here is a short explanation of each of these variables. All of them were indicator or categorical variables.

Item Strata: There are 176 priced item strata, part of attempted quote (Fig 2.1).

Major Group: All the item strata are combined into 8 major group.

Index PSU: 87 Primary Sampling Units in the United States (prior to 2018 Area Design).

Collection PSU: These are administrative units to manage the data collection process. These are not the Survey Design PSUs.

PSUsize: A, B, and C size of PSU based on the city size.

Mode: The process a data is collected. Personal Visit (P), Telephone (T), Web (W), Blank (few imputed quotes had no indicator, so assigned blank).

PriorityQuote: There are 10 ranks, low priority (1) to high priority (10), that a data collector may follow when assigned a workload. A quote with rank 10 gets high priority over a quote with rank 1 in the collection process. This priority rank is defined as: Median SE*Relative Importance of an item-area combination.

SeasonQuote: Some quotes are only available during specific season. So it has 3 indicators: Not Seasonal, Seasonal Non-Food, Seasonal Food (based on WO_DESIGNATED_SEAS_TYPE variable).

OutletStatus: If an outlet does not respond or coordinate, all (or most) quotes from that outlet may not be collected. Hence, all data from this specific outlet may exhibit imputed price. Two indicator levels were created (1=responded; 0=not-responded) from the disposition code of DER_INTERVIEW_CD variable.

1=responded (disposition code 11)
0=not-responded (disposition code not 11)

DER_INTERVIEW_CD
11 Available-at least one usable quote, etc.
19 Temp unavail-no quotes can be priced in outlet-due to temporary reason
23 Out of season-outlet is a seasonal outlet-and it is out of season
97 Deletion pending
98 Delete
99 Unknown-outlet is still in survey but status is unknown

Estimator Type: CPI uses 2 types of estimators to calculate the basic index: Geometric mean, Laspeyres. Geometric mean estimator could be seen as a proxy that accounts for the consumer substitution behavior within strata.

Bimonthly: Some quotes are collected every month, some are collected every other month. 1=monthly 0=bimonthly (even or odd). It is based on COMPUTATION_CYCLE variable.

MissingPrice (Response variable): MissingPrice (1 or 0) is the response variable of this Model. 1=Missing or Imputed data; 0=Not-Missing or Collected data.

With these potential covariates, the Logit model from section 5.2.1 could be written as follows:

$$Log(Odds\ of\ MissingPrice) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p =$$

$$\beta_0 + \beta_{is} ItemStrata_{(1\ldots176)} + \beta_{mg} MajorGroup_{(1\ldots8)} + \beta_{psu} PSU_{(1\ldots87)} + \beta_{ps} PSUsize_{(1\ldots3)} + \beta_{cpsu} CollectionPSU_{(1\ldots95)} + \beta_{mode} Mode_{(1\ldots4)} + \beta_{pq} PriorityQuote_{(1\ldots10)} + \beta_{os} OutletStatus_{(1,0)} + \beta_{sq} SeasonalQuote_{(1,2,3)} + \beta_{cs} ConsumerSubstitution_{(1,2)} + \beta_{bm} Bimonthly_{(1,2)} + Error$$

This equation will be the basis for model building, variable selection, and evaluation process in next sections.

6. Model Building, Implementation and Evaluation

6.1 Logistic Regression Accounting for Survey Design

The CPI Survey is a multistage sample design (not simple random sampling); hence, linear models need to account for *strata* and *cluster* structures for valid inference. Accounting for the survey design enables estimation of corrected standard errors, p-values, and confidence intervals. The Wald test uses these standard errors for assessing significance (Valliant R. et al., 2013; Lewis, T., 2010, 2012).

6.2 Model Deployment

SAS *PROC surveylogistic* is deployed to account for the survey design (stratum=Item; cluster =PSU). The Taylor Series option is used for variance estimation. The p-values and confidence intervals are calculated from this Taylor-Series variance. This p-value is used to make determination for variable selection from the set of covariates (SAS Institute Inc., "The SURVEYLOGISTIC Procedure", 2013).

On the other hand, *PROC logistic* is deployed to output the Goodness-Of-Fit Statistics, such as Hosmer and Lemeshow Test for Lack-of-fit, and Deviance Test for overdispersion. *PROC surveylogistic* does not produce these outputs as there is no consensus on how these tests are adjusted due to the presence of strata and cluster. *PROC surveylogistic* and *PROC logistic* generate same coefficients (beta) but different standard errors, p-values and confidence intervals (SAS Institute Inc., "The LOGISTIC Procedure", 2011; Allison P. & SAS Institute, 2012). Additional Goodness-of-Fit statistics are generated to assess model fit; however, each of these measures has limitations and trade-offs (see Allison, P., "Measures of Fit for Logistic Regression", 2014).

Table 1 summarizes these measures for each candidate model being compared to other, and an approximate guideline for each measure.

6.3 Variable Selection and Model Building Process

SAS *PROC logistic* has an automated option (selection=) to select variables using stepwise, backward, or forward method. However, this selection method uses p-values without accounting for the design. *PROC surveylogistic* does not have an automated variable selection option but generates robust standard errors, corrected p-values for the survey design. Hence, variable selection process started with the deployment of the full model (Model 1) with all potential covariates and an interaction term (ItemStrata*PSUsize), and reduced into a *parsimonious* model (Model 4a).

Based on the exploratory analysis (3.2), it is observed that some item strata have higher propensity for missing price than others. Similarly, some PSUs have higher propensity for missing price than others. As a result, interaction between these two variables (cross product) is a reasonable assumption to make, i.e., an item with high propensity of missing-price, being collected from a PSU with high (versus low) propensity of missing-price, may have different conditional probabilities for being missing. However, many cells will have 0 counts in these cases, causing a singularity in maximum likelihood estimation. Interaction between item strata with PSUsize is a potential cure, since 3 PSUsize (A, B, C) is a linear combination of 87 PSUs. It would reduce the chance of 0 counts in cell. Based on this premise, Model 1 was deployed, however, quasi-complete separation was still detected (SAS output), making the result not valid.

Model 2 was deployed excluding the interaction term from Model 1 but including the following 3 interactions: PSUsize*SeasonQuote; PSUsize*bimonthly; PSUsize*Estimator Type. Results indicated that interactions, bimonthly, and mode effects were insignificant ($p > 0.05$). Type 3 Analysis Table displayed significance for mode variable based on the reference group "Blank", but after inspecting the coefficients, it was evident that no significant difference exist between Personal Visit (P), Telephone (T), Web (W) collection; the coefficients were almost identical. The final model was deployed after eliminating these interactions and mode variable.

Any variable, a linear combination of another variable, was either eliminated before model deployment (e.g., Major Group), or SAS set the parameters as 0 (e.g., Estimator Type and PSUsize). Index PSU (87 parameters) and Collection PSU (95 parameters) were very similar. As a result, both variables were not deployed in the same model as covariates; instead, separate model was deployed for each case.

A total of 9 models (1, 2, 3, 4a, 4b, 5, 6, 7, 8) were deployed using *PROC surveylogistic* and *PROC logistic* and an intercept only model (null, saturated) before the parsimonious model was selected as the winning model.

6.4 Final Model (Parsimonious): Explaining Missingness Mechanism

Model fit statistics were compared across all the 8 models (except Model 1 because of invalid results, quasi-complete separation) before selecting the final model.

Hosmer and Lemeshow Test indicated no evidence of lack-of-fit (p-value > 0.05) for all 8 models. Deviance goodness-of-fit Test indicated no evidence of overdispersion based on Chi-square/DF close to 1 for Model 2 (0.9563), Model 3 (0.9485), Model 4a (0.9678), and Model 4b (0.9842). These 4 models were subsequently considered as candidate models. AIC was very similar for these 4 models (range: 68762 to 69060), although Model 2 (68762) had the lowest out of all 8 models. BIC was very similar for these 4 models (range: 71557 to 71642), although Model 5 had the lowest (71408) out of all 8 models.

Model 2 used Collection PSU, while Model 3 used Index PSU as covariate—the only difference. Similarly, Model 4a used Index PSU, while Model 4b used Collection PSU as covariate. Model 4a is derived from Model 3 excluding non-significant terms (interactions, Mode, Estimator Type, Bimonthly, Priority Quote, and PSUsize). Similarly, Model 4b is derived from Model 2 excluding non-significant terms. Because of the exclusion of non-significant terms, AIC and BIC slightly increased. At this stage, Model 4a and Model 4b were the only models under consideration and deemed as *parsimonious*.

Table 1: Model Comparison and Evaluation || ***Model 4a is Parsimonious and Interpretable***

Type3 Analysis Table	Model Null	Model 1	Model 2	Model 3	***Model 4a***	Model 4b	Model 5	Model 8	Sample Size (n)
Variables (SurveyLogistic)	**Intercept only**	**p Value**	**p Value**	**p Value**	**p Value**	**p Value**	**p Value**	**p Value**	Sample Size (n) = 113,664
Item Strata		**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	MissPrice = 0 (91,766 81%)
Index PSU		**<.0001**		**<.0001**	**<.0001**			**<.0001**	MissPrice = 1 (21,898 19%)
Collection PSU			**<.0001**			**<.0001**			
PSU size		**.**	**0.341**	**.**					
Mode		**<.0001**	**<.0001**	**<.0001**				**<.0001**	
Outlet Status		**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**		
Seasonal Quote		**<.0001**	**<.0001**	**<.0001**	**<.0001**	**<.0001**		**<.0001**	
Estimator Type		**.**	**.**	**.**					
Bimonthly		**0.3126**	**0.5174**	**0.6903**					
Priority Quote			**0.2878**	**0.2971**				**0.5457**	
PSU size*Season Quote			**0.2355**	**0.2364**				**0.1657**	
PSU size*Bimonthly			**0.3604**	**0.5914**					
PSU size*Estimator Type			**0.2114**	**0.2261**					
Item Strata*PSU size		**<.0001**							
Model Assessment (Logistic)									Approximate Guideline
Lack-of-fit Test: *Hosmer and Lemeshow* *(Chi-Square / DF)*			**7.5808 / 8**	**5.6331 / 8**	**5.5939 / 8**	**8.8701 / 8**	**0**	**12.5295 / 8**	
p-value (Hosmer)		***Quasi-complete separation***	**0.4755**	**0.6882**	**0.6926**	**0.3534**	**1**	**0.1291**	P > 0.05 => no evidence of lack of fit
Overdispersion: *Deviance Test* *(Chi-Square / DF)*			**0.9563**	**0.9485**	**0.9678**	**0.9842**	**0**	**1.243**	Chi/DF ~ 1 => no evidence of overdispersion
p-value (Deviance)			**1**	**1**	**0.9981**	**0.9136**	**1**	**<.0001**	
Overdispersion: Pearson Test* *(Chi-Square / DF)			**1.3149**	**1.3031**	**1.2372**	**1.3064**	**0**	**1.5684**	Chi/DF ~ 1 => no evidence of overdispersion
p-value (Pearson)			**<.0001**	**<.0001**	**<.0001**	**<.0001**	**1**	**<.0001**	
Number of Parameters (*P*)			**293**	**286**	**265**	**272**	**177**	**280**	
Log Likelihood	**-55701**		**-34086**	**-34114**	**-34265**	**-34238**	**-34674**	**-38568**	
-2 Log L	**111402**		**68172**	**68228**	**68530**	**68475**	**69348**	**77136**	smaller the better
AIC	**111404**		**68762**	**68800**	**69060**	**69019**	**69702**	**77696**	smaller the better
BIC	**111414**		**71606**	**71557**	**71615**	**71642**	**71408**	**80395**	smaller the better
Cox and Snell R^2	**0**		**0.316**	**0.316**	**0.314**	**0.315**	**0.309**	**0.260**	Larger the better
Max-Rescaled R^2	**0**		**0.506**	**0.506**	**0.503**	**0.503**	**0.495**	**0.417**	Larger the better
McFadden's $R^2 = 1-(L_M/L_0)$	**0**		**0.388**	**0.388**	**0.385**	**0.385**	**0.378**	**0.308**	Larger the better
Tjur (2009) *Coefficient of Discrimination*	**0**		**0.408**	**0.408**	**0.405**	**0.406**	**0.397**	**0.325**	Larger the better
Kappa	**0**		**0.512**	**0.505**	**0.502**	**0.507**	**0.514**	**0.421**	Larger the better
Area Under Curve (from ROC)	**0.5**		**0.878**	**0.878**	**0.876**	**0.877**	**0.872**	**0.844**	Larger the better
Error Rate	**0.1927**		**0.128**	**0.128**	**0.129**	**0.129**	**0.131**	**0.146**	smaller the better

In terms of interpretability, Model 4a (with Index PSU; 86 parameter estimates) is more interpretable than Model 4b (Collection PSU; 93 parameter estimates). Index PSU is the *Primary Sampling Unit* of the CPI survey design and had fewer parameters to estimate. Furthermore, while inspecting the 93 estimates (beta) for each Collection PSU, one beta was extremely large compare to others due to presence of 1 sample unit attempted from it (100% not-missing). Hence, **Model 4a** was deemed as the **final model** that is most *parsimonious* and *interpretable*.

All other measures of model assessment—Cox and Snell R^2, Max-Rescaled R^2, McFadden's $R^2 = 1-(L_M/L_0)$, Tjur *Coefficient of Discrimination* (2009), Kappa, Area Under Curve (AUC), Error Rate —were very similar for all these 8 Models. Pearson chi-square Test could be sensitive to large sample size (n=113,664), hence Deviance goodness-of-fit Test was chosen for assessing overdispersion over the Pearson chi-square Test.

Table 1 summarizes the measures for selective few candidate models including intercept only model (null / saturated model).

The final model that explains the missingness mechanism in CPI quote collection could be written as:

$$Log\left(Odds\ of\ MissingPrice_{(1\ or\ 0)}\right) = ln\left[\frac{P(MissingPrice)}{1-P(MissingPrice)}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

$$= \beta_0 + \beta_{is} ItemStrata_{(1,\ldots,176)} + \beta_{psu} IndexPSU_{(1,\ldots,87)} + \beta_{isos} OutletStatus_{(1,0)} + \beta_{sq} SeasonalQuote_{(1,2,3)} \quad \text{..Eq(4)}$$

$$Probability(MissingPrice) = \frac{exp\left(\beta_0 + \beta_{is} ItemStrata_{(1,\ldots,176)} + \beta_{psu} IndexPSU_{(1,\ldots,87)} + \beta_{isos} OutletStatus_{(1,0)} + \beta_{sq} SeasonalQuote_{(1,2,3)}\right)}{1 + exp\left(\beta_0 + \beta_{is} ItemStrata_{(1,\ldots,176)} + \beta_{psu} IndexPSU_{(1,\ldots,87)} + \beta_{isos} OutletStatus_{(1,0)} + \beta_{sq} SeasonalQuote_{(1,2,3)}\right)}$$

For any significant Item or PSU ($p < 0.05$), probability of price being missing could be explained by the following estimated model (survey variables).

$$Log\left(Odds\ of\ MissingPrice_{(1)}\right) = -2.4005 + (-14.568, \ldots, 1.9787) ItemStrata_{(1,\ldots,119)} + (0.4156, \ldots, 1.0084)\ IndexPSU_{(1,\ldots,11)} + 30.2357\ OutletStatus_{(non-response)} + 0.9941\ SeasonalQuote_{Non-Food\ SeasonalQ}$$

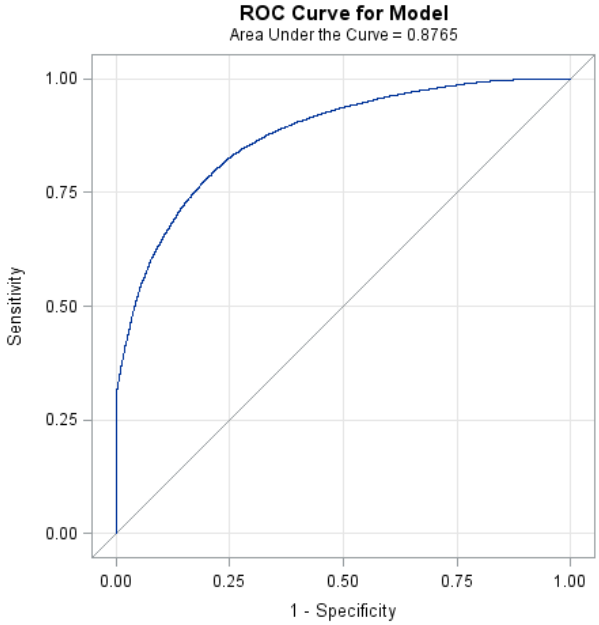
$$Odds\ of\ MissingPrice_{(1)} = exp(Log\left(Odds\ of\ MissingPrice_{(1)}\right) = = \frac{P(MissingPrice = 1)}{1 - P(MissingPrice = 1)} = \frac{P(MissingPrice = 1)}{P(MissingPrice = 0)}$$

$$= 0.091 \times (0.006\ to\ 7.233) ItemStrata_{(1,\ldots,119)} \times (1.515\ to\ 2.741)\ IndexPSU_{(1,\ldots,11)} \times (1.353 \times 10^{13}) OutletStatus_{Non-respond} \times 2.702\ SeasonalQuote_{Non-Food\ SeasonalQ}$$

If a linear coefficient $\beta > 0$ for an item stratum or PSU, then $exp(\beta)>1$, suggesting it favors the outcome of success (probability of missing; 1); if a linear coefficient $\beta < 0$ for an item stratum or PSU, then $exp(\beta) < 1$, suggesting it favors the outcome of failure

(probability of not-missing; 0); if a linear coefficient $\beta = 0$ for an item stratum or PSU, then $exp(\beta) = 1$, suggesting it favors the outcome of success or failure equally, or about the same.

The figure displays the ROC curve for the final model (Model 4a). The ROC curve examines the trade-off of true positive rate (sensitivity) against the false positive rate (1-



specificity) at various decision rules (cut-off points that convert probability-outputs of logit model into 0/1 classes). The true positive rate (sensitivity) is the probability of predicting (proportion) a missing-price when the price is indeed missing in the data (observed). The false positive rate (1-specificity) is the probability of predicting (proportion) missing-price although price is not-missing (collected) in the data (observed). The closer the ROC curve is to the upper left corner (AUC 1.0), the higher the overall accuracy of the Model; while the diagonal line (AUC 0.5) is a non-informative model with no predictive power. The AUC measures discrimination—the ability of the model to correctly classify quote as observed from the data. The probabilistic interpretation is that if a quote is randomly chosen with true missing value classified as missing, and with not-missing value (collected) falsely classified as missing, the probability that the true missing price outranks the falsely classified missing quote is 87%.

6.5 Reference Group (effect or dummy coding) for Interpretability

In regards to dummy coding (effect coding), logistic regression requires a reference group—1 for 176 levels of item strata and 1 for 87 PSUs—for interpretability of the model. Highly sparse (unbalanced) missing proportions (1% to 60%) and sample size (counts) across item strata, and presence of so many levels of class (176; 87) for a single variable create a challenge in finding this "ideal" case (that fits all). Additionally, p-values could slightly change depending on the reference group it is compared to, and due to sparsity in missing proportion distribution (Fig. 3.2.B; 3.2.G). Hence, the ideal case is constructed based on the marginal mean of a group. The average proportion of missing price (marginal mean) is 16.5% for item strata and 19% for PSU. An item with a marginal mean proportion of missing price is assigned as the reference group. The same holds for PSU. It enables better interpretability of the model.

7. Remarks from Logit Results

7.1 Missing Price and Covariates

Based on the diagnosis, the probability of a missing price seems to be *conditionally* related to a few survey variables ($P<0.05$), and could be explained by Item Strata, PSU, Outlet Status, and Seasonal Quote. That is, the propensity for a quote price to be missing is *conditionally related* to the observed data of other variables, suggesting *Missing at Random* (MAR) mechanism (if no evidence found for *Missing Not at Random*, MNAR).

7.2 Significant Item Strata or PSU

An item stratum that shows significance ($p <0.05$) implies, propensity for a quote price to be missing is partly explained by that item, partly explained by its geographic location (when PSU $p <0.05$), even after controlling for outlet status (whether responded or not; $p <0.0001$) and whether the quote is seasonal non-food ($p <0.0001$).

7.3 Non-Significant Item Strata or PSU

Any item stratum or PSU with no significance ($p \geq 0.05$) implies that the item stratum or PSU does not have much impact on the propensity of a price being missing compared to an average proportion missing (marginal mean proportion of item and of PSU).

7.4 Outlet Status

Outlet status is significant ($p <0.0001$) for all cases, indicating it is a certainty explanatory variable for missingness. Outlet Status is an indicator of the following attributes: outlet may be a high-end (or some other type) and does not have time to coordinate with the data collectors, outlet may be effected by a disaster (like hurricane), or data collector could not collect all quotes due to vacancies not filled (lack of data collector). More auxiliary variables about the outlet could explain why some outlets do not respond over other outlets. Outlet status not-responded spans all 176 items and all 87 PSUs, suggesting it is not an isolated situation for a few selective items or PSU.

7.5 PSU Attributes

PSU seems to be an indicator or proxy measure of the following attributes: geographic location, vacancies of data collectors at a given time, interaction between outlet, geographic location and propensity of coordination with government collection.

8. MNAR Diagnosis and Results

8.1 Diagnosis Method without a Follow-up Survey to Non-Respondents

The best way to diagnose the MNAR situation is to conduct a follow-up survey to non-respondents, and if answers of non-respondents *differ* substantially from respondents, that is evidence a MNAR process is generating missing data (Grace-Martin, K). However, this is out of scope for us because there was no budget for a follow-up survey. Additionally, once a quote is sold out to a customer before re-pricing by the data collector, that price is no longer available (missing).

With this challenge and limited scope, an innovative method is implemented to diagnose the MNAR process. Here is the intuition behind the analysis. Each quote (observation) in our dataset must have had a collected price at some point of time since CPI is a longitudinal survey, regardless of prices being imputed currently for some quotes. A distribution of *Last Observed Price* could be constructed for each item stratum. If this *Last Observed Price* distribution is controlled as a covariate in the model, and the

variable displays significant association with the response variable (1=missing, 0=not missing), it would indicate that missing quotes in this item strata seem to be missing in certain values of price, i.e., missing prices are related to price itself (its values). It enables one to make a potential case for MNAR situation without conducting a follow-up survey when budget constraints exist. Each item stratum must be analyzed separately so that the distribution is specific to the values of itself (same item stratum) and not confounded by the values of other item strata.

8.2 Model Implementation and Evaluation

Last Observed Price for each quote, collected within 12 months, was matched with the current dataset. This process retained a sample size of $n_{matched}$=96,952 out of n=113,664. An exploratory examination was conducted before selecting a few item strata to conduct this part of analysis since sample sizes across item strata were highly unbalanced and proportions of imputed versus collected were highly sparse (Fig 3.2.B).

MNAR diagnosis for a *single* item strata (from section 4):
[P(y_i = missing | X, X*) = P(y_i = missing | y_i=c)]

$$Log\left(Odds\ of\ MissingPrice_{(1\ or\ 0)\boldsymbol{Item\ A}}\right) = \beta_0 + \beta_{LP}LastObservedPrice + \beta_{PS}PSUsize_{(1,\dots,3)} + \beta_{md}Mode_{(1,\dots,4)} + \beta_{pq}PriorityQuote_{(1,\dots,10)} + \beta_{os}OutletStatus_{(1,0)} + \beta_{sq}SeasonalQuote_{(1,2,3)} + \beta_{et}Estimator\ Type_{(1,2)} + \beta_{bm}Bimonthly_{(1,2)} \quad \dots \text{Eq.5}$$

This full model (Eq.5) could generate quasi-complete separation in MLE, depending on a specific item stratum.

> "…most common cause of quasi-complete separation is a dummy predictor variable that has the following property: at one level of the dummy variable either every case has a 1 on the dependent variable or every case had a 0" (Allison, P. & SAS Institute, 2012; pg. 50).

As a mitigation to this challenge, two models were developed:

1) *Reduced Sample (Full Model):* Exclude **cases** (delete observations) that generate the quasi-complete separation in the model. This enables one to deploy the full model with all dependent variables of Eq. 5.
2) *Full Sample (Reduced Model):* Exclude **variables** (dependent variable) that generate quasi-complete separation in the model. This enables one to use the entire sample size of an item strata while reducing the number of dependent variables of Eq. 5.

Furthermore, a PSU may contain zero (n=0) sample units for some item strata since the basic CPI index is computed for an area-item level and not a PSU-item level. As a mitigation to this challenge, the model was deployed in both forms—Proc Logistic (not accounting for PSU as Cluster) and Proc Surveylogistic for assessment. This generated 4 conditions for evaluation: Reduced Sample (Logistic), Reduced Sample (Surveylogistic), Full Sample (Logistic), Full Sample (Surveylogistic).

The p-value for *Last Observed Price* is assessed for all 4 models—whether or not this variable has a significant relationship (p-value<0.05) with price being missing, while controlling for other covariates.

If the logistic procedure detected overdispersion based on the Deviance Goodness-of-Fit Test, the covariance matrix was multiplied by the heterogeneity factor (Deviance / DF) to adjust for overdispersion [scale=Deviance] (Allison, P. & SAS Institute, 2012), and then output the adjusted standard errors and p-values. *Last Observed Price* was standardized [standard normal, N(0, 1)] before model deployment. Table 2 summarizes the assessment results for selected few item strata.

Table 2: MNAR Diagnosis Results					
Item Strata	**Boys' and girls' footwear**				
	Full Sample Model		**Reduced Sample Model**		**Conclusion**
Last Observed Price N(0,1)	Coefficient	*p*-value	Coefficient	*p*-value	Significant (Potentially)
SurveyLogistic	-0.252	0.0617	-0.4109	0.0039	
Logistic	-0.252	0.0148	-0.4109	0.0005	
Sample Size (n)	687		635		
Collected	465		465		
Missing	222		170		
Item Strata	**Girls' apparel**				
	Full Sample Model		**Reduced Sample Model**		**Conclusion**
Last Observed Price N(0,1)	Coefficient	*p*-value	Coefficient	*p*-value	Not significant
SurveyLogistic	-0.0156	0.8201	-0.00844	0.9007	
Logistic	-0.0103	0.8499	-0.00844	0.9066	
Sample Size (n)	1415		1396		
Collected	696		696		
Missing	719		700		
Item Strata	**Physicians' services**				
	Full Sample Model		**Reduced Sample Model**		**Conclusion**
Last Observed Price N(0,1)	Coefficient	*p*-value	Coefficient	*p*-value	Not significant
SurveyLogistic	0.0298	0.8059	0.1453	0.2743	
Logistic	0.0298	0.774	0.1453	0.1922	
Sample Size (n)	625		519		
Collected	475		444		
Missing	150		75		
Item Strata	**Other fresh fruits**				
	Full Sample Model		**Reduced Sample Model**		**Conclusion**
Last Observed Price N(0,1)	Coefficient	*p*-value	Coefficient	*p*-value	Not significant
SurveyLogistic	-0.0355	0.5161	Same as Full Sample Model.		
Logistic	-0.0355	0.5774	No Case generated Quasi-		
Sample Size (n)	2365		Complete-separation in MLE		
Collected	1673				
Missing	692				

9. Final Study Conclusions

9.1 Explaining Missingness Mechanism

Propensity for a quote-price being missing could be explained by four (4) survey variables: *Outlet Status, Item Strata, PSU* and *Seasonal Quote Status*.

Up to 6% (6,632) quotes being missing could be attributed to outlet not responding (including lack of data collectors or vacancies). When outlets respond, a propensity of up to 13% (15,266) quote being missing still exist and could be explained due to item strata, PSU and Seasonal Quote status. Non-food Seasonal Quotes are 3 times more likely to be missing than Non-seasonal quote in average (OR= 2.702 $p<0.0001$ OR Lower= 2.348 OR Upper= 3.109). Some 119 Item Strata out of 176 have a significant ($p < 0.05$) relationship with the propensity of a quote being missing even after controlling for PSU, Seasonal Quote and Outlet Status; different coefficient size or effect size (Odds: 0.006 to 7.233) suggests certain items have higher propensity of being missing than other items. Similarly, 11 Index PSU out of 87 have a significant ($p < 0.05$) relationship with the propensity of a quote being missing even after controlling for Item Strata, Seasonal Quote and Outlet Status; different coefficient size or effect size (Odds: 1.515 to 2.741) suggests in certain PSU a quote has a higher propensity of being missing than other PSUs (even for a same item).

9.2 Diagnosis of MNAR, MAR, MCAR in Item Strata

If *Last Observed Price* is significant ($p < 0.05$) for an item stratum (Table 2), it provides some evidence for a potential MNAR situation for that item strata (e.g., Boys' and girls' footwear), superseding the results of MAR case. If the MNAR case is not significant ($p \geq 0.05$), the MAR result is valid for an item stratum (e.g., Girls' apparel. Fig 9.2.A; 9.2.B). If an item stratum is not significant at all (neither MNAR nor MAR), it could be concluded as a weaker MAR situation or a MCAR situation (e.g., Sports equipment).

Figure 9.2.A

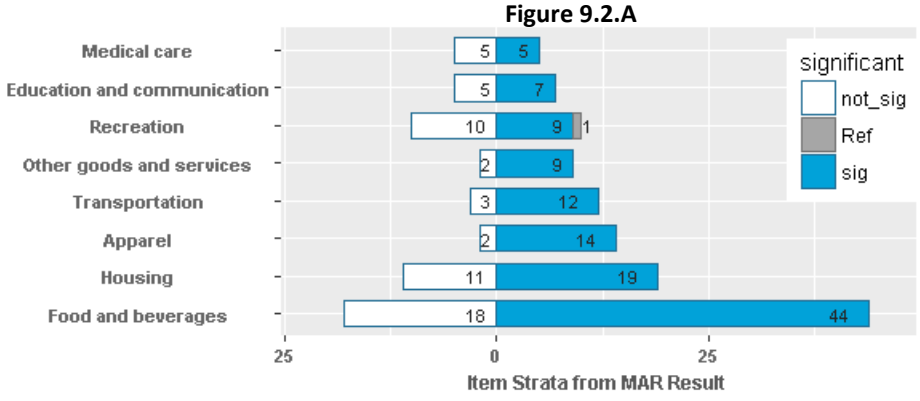


Figure 9.2.A displays the results from MAR Logit Model 4a. 119 Item Strata out of 176 are significant ($p <0.05$). These 176 Item Strata (significant, not-significant and 1 reference group) are shown in the data visual *disaggregated* by Major Group.

Figure 9.2.B

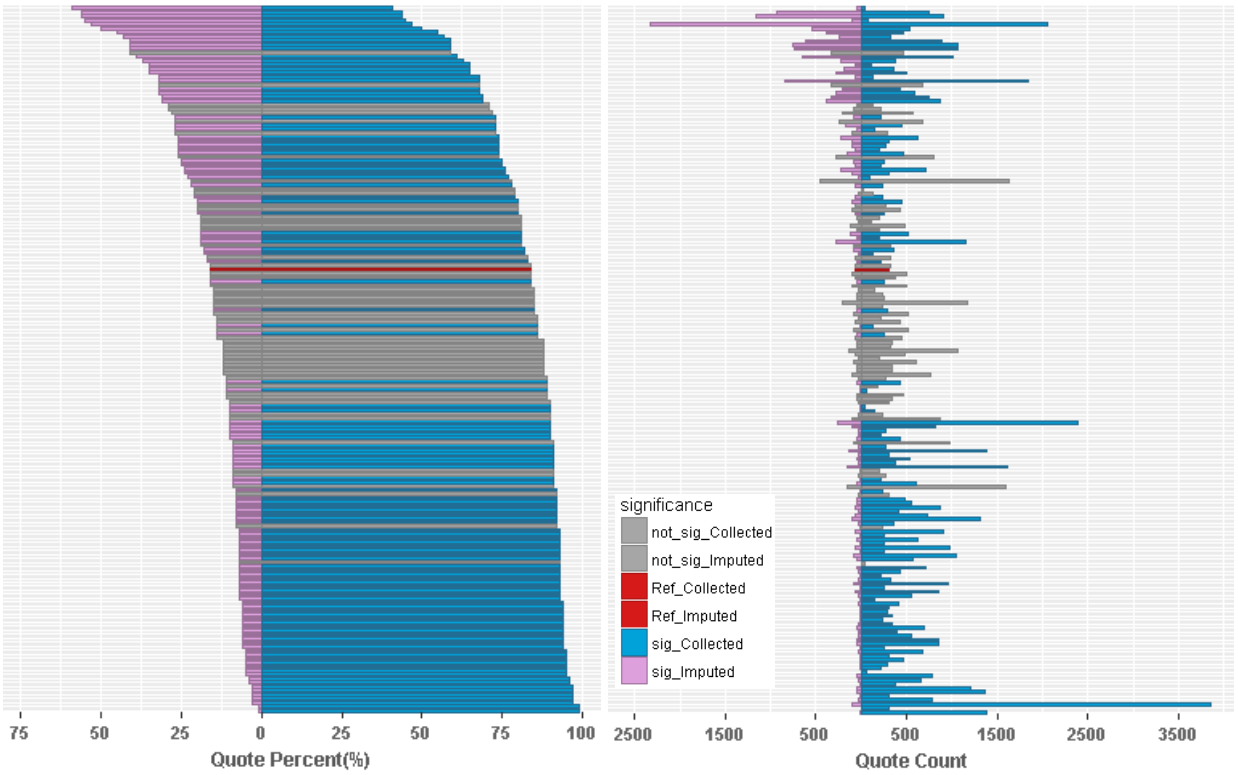


Figure 9.2.B incorporates the results of MAR Logit Model 4a. and redisplays the Item Strata distribution from exploratory analysis (Fig 3.2.B), *differentiating* significant versus non-significant (grey) Item Strata, and the reference group (red). This highlights the Item Strata that exhibit significant relationships ($p < 0.05$) with the propensity of missing price compare to the reference group, *after* adjusting for covariates (PSU, Outlet Status, Seasonal Quote Status). When a linear coefficient is negative ($\beta < 0$; $exp(\beta) < 1$) for an item stratum and significant ($p < 0.05$), it favors the probability of a not-missing (0; collected quote) outcome. In other words, these Item Strata do not have much missing data compare to the reference group after adjustment for covariates. Out of 119 significant item strata, 81 favor the outcome of collected (not-missing) propensity ($\beta < 0$), 38 favor the outcome of missing propensity ($\beta > 0$). Hence, these 38 Item Strata are the primary concern for this study. "Collected" data (not-missing) is the preferred outcome.

10. Benefits of this Study

10.1 First Step before Imputation Evaluation

The first step before evaluating imputation methodologies should be to diagnose the missing data mechanism. This study provides the first step—evaluates patterns of missing price in the micro data.

First Step: Diagnose → Missing Pattern
Second Step: Cure → Imputation

Currently CPI deploys the Group-Mean imputation with periodic updates to the group definitions. This is the first comprehensive study to diagnose MCAR, MAR and MNAR situations in the CPI micro data, albeit at a higher level than the CPI performs imputation

which is usually at the PSU – Entry Level Item level. An ELI is a subset of an item stratum.

10.2 Targeted Intervention to Imputation (Return on Investment)
This study provides a guide, or blueprint, to direct resources for an item stratum (within limited budget and staff) considering multiple factors:

- ✓ Missing proportion of price data, specific to an item stratum.
- ✓ Sample size of an item stratum.
- ✓ Missing data mechanism, specific to an item stratum.
- ✓ Relative Importance of an item stratum to CPI weight.

Although price is the target variable (collected or imputed), the CPI survey does not produce an average or sum of the target variable, but percent-change in price as the ultimate survey output. Percent-change is a non-symmetric measure, e.g., change from \$1 to \$2 in current month implies 100% increase $\left(\frac{\$2-\$1}{\$1} \times 100\right)$; however, change from \$2 to \$1 implies 50% decrease $\left(\frac{\$1-\$2}{\$2} \times 100\right)$.

Additionally, any method of imputation may work for an item stratum displaying MCAR situation. A missing quote that does not change its price frequently (somewhat stable) may have negligible impact in percent-change calculation that feeds into price index.

Hence, these factors could be valuable information for decision making processes, before investing resources for improvement to specific item strata.

10.3 Targeted Intervention for Collection
Based on this study, up to 6% (6,632) missing quotes could be attributed to the outlet not responding (including lack of data collectors). This might be an area for opportunity to explore options for targeted intervention in order to increase response rates. When an outlet status is not-responded, all quotes from that specific outlet may be missing.

10.4 Data Visualization
Generating data visualization with proportion imputed versus collected price (distributions) is an effective way to construct a narrative about the missing-data and to share with the stakeholders.

10.5 MNAR Diagnosis Method for Other Longitudinal Surveys
In many federal surveys, budget is often a constraint to conduct another follow-up survey for non-respondents or it is simply infeasible. This MNAR diagnosis method provides an innovative approach that may be useful to other longitudinal surveys.

11. Limitations of this Study

11.1 Longitudinal Study for Robust Conclusions
This is a cross-sectional study and does not inform whether the same item stratum or PSU has a propensity of being missing over time. A longitudinal study may need to be conducted (e.g., Generalized Linear Mixed Model; Proc GLIMMIX) to investigate the

stability of the missingness mechanism over time for correlated observations (dependent data).

11.2 Include More Auxiliary Variables

The current study selects common covariates (common support region between imputed and collected) as the first step to assess missing pattern for a recent dataset. More auxiliary variables need to be included as covariates—specific to item strata or Outlet characteristics—in a future extension of the study. These auxiliary variables must have complete case observations (without missing value), which may be a challenge. Another challenge may be to find covariates without too many levels (class) in evaluating small sample size item strata. Otherwise, too many parameters are estimated for a small sample size item stratum (large p, small n). Potential confounding is always a challenge to rule out completely in any observational studies.

Acknowledgements

Special thanks to Bill Johnson—Supervisory Mathematical Statistician; Chief of the CPI Survey Research and Analysis Branch—for providing guidance, historical knowledge, and helpful feedback throughout this study.

References

Allison, P. & SAS Institute. (2012). *Logistic Regression Using SAS: Theory and Applicatio*n; Second Edition. SAS Institute.

Allison, P. (2014). "Measures of Fit for Logistic Regression." *Proceedings of the SAS® Global Forum 2014 Conference*, paper 1485-2014. Cary, NC: SAS Institute Inc. https://support.sas.com/resources/papers/proceedings14/1485-2014.pdf.

Bureau of Labor Statistics. (1966). "Chapter 10: Consumer Prices". *The BLS Handbook of Methods for Surveys and Studies.* Retrieved from https://fraser.stlouisfed.org/files/docs/publications/bls/bls_1458_1966.pdf.

Bureau of Labor Statistics. (2018). "Chapter 17: The Consumer Price Index". *BLS Handbook of Methods.* Retrieved from http://www.bls.gov/cpi/cpi_methods.htm.

Grace-Martin, K. (n.d.). *Missing Data Mechanisms: A Primer*. The Analysis Factor©. Retrieved February 14, 2018, from https://www.theanalysisfactor.com/causes-of-missing-data/.

Grace-Martin, K. (n.d.). *What is the difference between MAR and MCAR missing data?* The Analysis Factor©. Retrieved February 14, 2018, from https://www.theanalysisfactor.com/causes-of-missing-data/.

Grace-Martin, K. (n.d.). *How to Diagnose the Missing Data Mechanism*. The Analysis Factor©. Retrieved February 14, 2018, from https://www.theanalysisfactor.com/causes-of-missing-data/.

Lewis, T. (2010). "Principles of Proper Inferences from Complex Survey Data." *Proceedings of the SAS® Global Forum 2010 Conference*, paper 266-2010. Cary, NC: SAS Institute Inc. http://support.sas.com/resources/papers/proceedings10/266-2010.pdf.

Lewis, T. (2012). “Modeling Complex Survey Data.” *Proceedings of the 2012 MidWest SAS Users Group Conference*, paper SA-07-2012. Cary, NC: SAS Institute Inc. https://www.mwsug.org/proceedings/2012/SA/MWSUG-2012-SA07.pdf.

Little, R., Rubin, D. (2002). *Statistical analysis with missing data*. New York: Wiley.

Molenberghs, G., Fitzmaurice, G., Kenward, M., Tsiatis, A., Verbeke, G. (Eds). (2015). *Handbook of missing data methodology*. Boca Raton: Chapman & Hall/CRC.

Rubin, D.B. (1976). Inference and Missing Data. Biometrika, 63, 581-592.

SAS Institute Inc. (2011). “The LOGISTIC Procedure”. *SAS/STAT® 9.3 User’s Guide*. Cary, NC: SAS Institute Inc. https://support.sas.com/documentation/onlinedoc/stat/930/logistic.pdf.

SAS Institute Inc. (2013). “The SURVEYLOGISTIC Procedure”. *SAS/STAT® 13.1 User’s Guide*. Cary, NC: SAS Institute Inc. https://support.sas.com/documentation/onlinedoc/stat/131/surveylogistic.pdf.

SAS Institute Inc. “Fit Statistics for Scored Data Sets, The LOGISTIC Procedure”. *SAS/STAT® 9.3 User’s Guide*. Cary, NC: SAS Institute Inc. https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_logistic_sect050.htm.

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

Valliant R, Dever J, Kreuter F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer.

Wickham, Hadley. (2010). "A layered grammar of graphics". *Journal of Computational and Graphical Statistics* 19 (1): 3–28. http://byrneslab.net/classes/biol607/readings/wickham_layered-grammar.pdf.

Wickham, Hadley. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag: New York.