IMPUTATION OF MISSING DATA: TESTING A MODEL-BASED APPROACH

Eric Melton, Millward Brown IntelliQuest
Valentine Appel, Consultant

Introduction

One of the most pressing issues surrounding media and marketing research is the decline in respondent cooperation. Despite years of experimentation and implementation of unique techniques intended to cajole prospective respondents to participate in surveys, and maximize their level of completion, as an industry we are at best holding our own. With increased respondent confusion in the U.S. over the differences between legitimate survey research versus telemarketing, direct marketing, and fund-raising, researchers are encountering a suspicious, uncooperative, and often angry population.

As an industry, we have compounded the issue by continuing to increase the amount of information and the level of detail we ask respondents to reveal. Readership researchers learned long ago that we needed to separate the critical element of print measurement from the onerous task of completing detailed product purchase behavior and other measures in order to achieve the highest achievable levels of cooperation. Even with this separation, we clearly have seen evidence of respondent bias or laziness when faced with the burden of overly complicated or lengthy readership questions. In 1999, we presented a paper at the Florence Symposium which showed that a readership screening process which hid the follow-up questions regarding frequency of readership and reader quality measures resulted in significantly higher screen-in rates and cooperation compared to questionnaires that revealed upfront to the respondent the "penalty" of responding affirmatively to the screen-in question (Appel, Elder, and Incalcatera, 1999).

We continue our efforts to simplify and improve the respondent experience to reduce the number of non-responders, and minimize missing data within the "completed" surveys, while devoting a great deal of resources in trying to improve the accuracy of imputing unknown data. The experiments conducted by Millward Brown IntelliQuest, using data from CIMS^{TM} syndicated studies, serve potentially to bring a better approach to the handling of missing product data, but also explore the appropriate usage of the imputation approach. As this paper demonstrates, while model-based imputation may be a better surrogate for product information than what is commonly used by syndicated researchers today, it is *not a reasonable replacement* for the readership habits of respondents. We believe the results of this experiment, while improving upon the process, also point out the need to refocus our efforts as an industry on “the accurate collection of data” rather than ameliorating “missingness” in syndicated studies.

Background

Millward Brown IntelliQuest has been actively pursuing more advanced methods of imputing missing data (i.e., replacing missing data with plausible values for use in complete cases analyses) since early 2002 culminating in the research by Vriens and Melton (2002) which attempted to show that model-based multiple imputation (MBMI) outperformed simpler procedures for imputing missing data. Model-based multiple imputation is a technique published by Rubin (1987) where missing data is replaced with a number of different plausible values. Each imputed version of the data set may be analyzed using complete case methods, and the different versions may be synthesized into a single data set with imputed values. Multiply imputing missing values increases the precision of the prediction and allows analysts a way to calculate confidence intervals that incorporate the uncertainty surrounding missing data prediction. Schafer developed a specialized Windows-based software application called NORM that produces model-based multiple imputations. It is freely available on the World Wide Web.

As we continue to investigate MBMI as a replacement imputation technique, we found that we needed to expand on the research conducted in 2002. There were three reasons for this:

First, the basis for judging the effectiveness of MBMI and other imputation techniques involved the comparison of aggregate predicted means versus actual means. Aggregate means can be easy to reproduce with any number of techniques especially when data is missing completely at random. The key is looking to see if the predictions hold among sub-groups of the sample.

Second, the original research may be biased in favor of MBMI over the random hot deck (RHD) approach. The original research was an investigation of new techniques versus the random hot deck approach as currently implemented in the CIMS Business Influencer Study, Millward Brown IntelliQuest's annual, syndicated media study. The CIMS implementation of RHD incorporates four variables for the imputation of missing data due to item non-response: Job role (MIS, senior manager, departmental), Number employed in worksite (1 to 9, 10 to 49, 50 to 249, 250 to 999, 1,000+), Type of technology influencer (primary, authorize, informal, evaluate) and Gender. The choice of these four variables originated with their use for imputing a

large number of variables for whole book (unit) non-response of the product portion of the CIMS survey. As the MBMI procedure inherently selects variables known to be correlated with the variables to be imputed, this difference may be more important than the difference in imputation technique employed. Had both model-based and random hot deck techniques used the same variable set as predictors the difference may not have been as pronounced.

Finally, as a consequence of how the specialized imputation software can be used, the models produced by MBMI were tested on the same sample upon which they were developed. This would give a competitive advantage to MBMI over RHD in determining the accuracy of predicted responses.

The present research serves to address these issues using data from the v9.0 CIMS Business Influencer Study and will incorporate questions from the product behavior portion of the study as well as readership screen-in data from the media portion of the study. To our knowledge, an in-depth comparison using MBMI on readership data has not yet been presented; hence our results will be directly relevant for all researchers currently in the business of measuring audiences. In order to produce a valid test against RHD we must dispense with MBMI for the moment. In practice, with the specialized model-based imputation software coming up with the model itself is not the goal. The sole purpose of the software is to impute missing values based on whatever relationships are at hand; the modeling itself takes place behind the scenes. Consequently, there is no interaction with the software where regression coefficients can be input from another data set to impute a test data set. Also, the software does not output model equations for use elsewhere. In order to make the comparison against RHD, a model-based single imputation (MBSI) approach using multiple linear regression will allow for the cross-validation of regression models and a fairer comparison against the random hot deck approach. A key difference between the MBSI approach taken in this research and what specialized imputation software provides is that the imputation software introduces random error or noise in the data to better simulate real data. A consequence of our MBSI approach is that all predicted values fall on the regression line defined by the model. Variance in the imputed data from MBSI will be underestimated. Still, we should expect that if MBSI does not do better than RHD that the introduction of random error will not suddenly make model-based approaches materially better.

Methodology (Readership)

The v9.0 CIMS Business Influencer Study data set contained 8,781 respondents with all variables fully edited and missing values imputed. Within each of seven general-interest groups presented in the survey, a single title was randomly selected. The following titles are the subjects of this research (group lettering corresponds to the lettering in the CIMS questionnaire):

Group	Title
G.	Fast Company
H.	Worth
I.	Time
J.	Outside Magazine
K.	National Geographic
L.	Rolling Stone
M.	Spirit (Southwest Airlines)

No imputation takes place on media screen-in questions for CIMS, so all screen-in questions for the seven test titles were complete. The data set was then randomized and split into two approximately equal subsets. The first half is Group A which was used to build the predictive models. The second half is Group B which acted as the test set where values were imputed.

Multiple linear regressions employing stepwise procedures were performed on each of the seven test titles using a series of demographic, firmographic and broad purchase behavior questions as predictors (see Appendix A). Like the media screen-in questions, the predictor variables are asked of all respondents in the study and are primarily complete in the CIMS data set.

Random hot deck imputations were produced on Group B respondents for each test title as follows:

- Group A and Group B were combined back together. Group A acted as the pool of donator respondents.
- Respondents were sorted so that respondents with similar attributes were grouped together. Categorizing was done on the basis of the predictor variables identified in the regression analysis on Group A. Within each category respondents were randomly sorted to disperse respondents with missing data among respondents with complete data.
- The recipient respondent (from Group B) with a missing value for the test title received its imputed value from the nearest prior donor respondent (from Group A) with a good value for that title. The process continued with subsequent recipient respondents.

A model-based single imputation for each test title was made by applying the regression equation developed from Group A to the data in Group B to yield probabilities. Cut points for classifying each probability were determined for each title by multiplying the predicted mean rating by the sample size of Group B. For example, for *Fast Company* the predicted mean rating was about 4.58% and the sample size 4,443. The cut point for *Fast Company* was 203 meaning that the 203 respondents with the highest probabilities were given a value of '1' or 'Yes' for the screen-in. All other respondents were given a value of '0' or 'No' for the screen-in.

Hence, each respondent in Group B had three sets of values for each test title: an imputed value from RHD, an imputed value from MBSI, and the actual value.

Results (Readership)

Correlations (phi coefficients) were calculated between each of the two imputed values and the actual value on a respondent by respondent basis for each test title within Group B (See Exhibit A).

Exhibit - (Readership – Correlation of Group B imputed values versus actual) *

Publication	THD	MSI
Fast Company	0.012	0.042
Worth	0.011	0.018
TIME	0.003	0.093
Outside Magazine	0.026	0.083
National Geographic	0.056	0.143
Rolling Stone	0.028	0.122
Spirit (Southwest Airlines)	0.035	0.013

* n=4,443

Correlation coefficients approaching a value of 1 would show that either imputation approach is both successfully predicting a screen-fail and predicting a screen-in. Here, our results show that MBSI performed slightly better than RHD (the correlation coefficients of MBSI were closer to 1 than RHD) on all test titles except for one. However, MBSI does a poor job of recovering the actual value of the screen-in.

The results of the test prompted a re-examination of the imputation of missing data within the product behavior/profiling portion of the CIMS study.

Methodology (Online Behavior and Attitudinals)

Again, the test data was derived from v9.0 CIMS Business Influencer Study. A post-imputation version of the data was filtered by those who had reached the end of both the media portion of the survey and the product portion of the survey. In other words, none of the respondents in this test were subject to imputation due to whole-book (unit) or sectional (respondent drop-out) non-response. The data set contained a total of 8,113 respondents. Four variables were randomly selected from the last section of the questionnaire dealing primarily with online behavior. No effort was made to use variables that might have produced good results. The following variables were selected:

Q11B – Year first accessed the Internet
Q11GH – In minutes, avg. length of online session at home
Q11V.04 – attitudinal - agree or disagree that low cost is most important factor in tech decision
Q11V.10 – attitudinal - agree or disagree that keeping up with tech changes is a hassle

The following actions were taken:

1. All cases where the test variable or variables were missing or contained imputed data were excluded from the file. Only cases where the respondent actually gave a good response were used in the test.
2. The file contained all CIMS variables including media data. A selection of predictor variables had to be made from all the variables in the file. First, a listing was created of all variables where the number of missing values was no more than 5% of the remaining test cases. From this listing a large number of variables were selected that did not further erode too much the number of cases available for regression analysis. (i.e., we were trying to find the happy medium between not discarding too many variables that may have a relationship with the test variable and not having to sacrifice a lot of cases to retain a large number of variables.)
3. After the selection of predictor variables, the test data set had no missing values. The data set was then randomized and split into two approximately equal subsets. The first half of the file was designated Group A, and the second half of the file was designated Group B.
4. A multiple, linear regression analysis on Group A was run on the test variable using the remaining variables as independent variables. Stepwise procedures were used.
5. The resulting regression model was limited to at maximum five independent variables. For MBSI, we applied the resulting regression equation to predict the values of the test variable for Group B.
6. RHD was performed as described in the readership test. The result was that each respondent in Group B had three sets of values for each test variable: an imputed value from RHD, an imputed value from MBSI, and the actual value.

Results (Product Behavior Test)

Correlations (Pearson coefficients) were calculated between each of the two imputed values and the actual value on a respondent by respondent basis for each test variable within Group B (See Exhibit B).

Exhibit B – (online Behavior and Attitudinals – Correlation of Group B imputer values versus actual)

Variable	RHD	MBSI	n
Q11B	0.159	0.313	2,293
Q11GH	0.180	0.351	2,140
Q11V.04	-0.005	0.068	2,905
Q11V.10	0.038	0.237	2,905

Again, the results show that MBSI performed somewhat better than RHD (the correlation coefficients of MBSI are closer to 1 than RHD), but MBSI itself did an unexceptional job of recovering the actual value of the test variable.

Conclusion

The tests reveal that model-based approaches may be better than random hot deck, but it is not the miracle way to get unbiased predictions for CIMS data especially where the underlying relationships between variables are not strong. Still, the model-based approach demonstrates an incremental step in quality over the traditional method as currently implemented in CIMS. Remember that random hot deck as executed in CIMS incorporates just four variables with at best one additional variable (Job role, Number employed at work site, Type of technology influencer, and Gender). These four variables may have or may not have a predictive relationship with any particular variable that is subject to item non-response imputation. A model-based approach does better than the random hot deck approach, and a model-based approach should do even better than a random hot deck approach limited to four variables.

What is apparent, however, is that, under the conditions within this study, neither imputation technique can reliably predict readership habits. And, points out that readership is a highly unique experience which does not correlate to the current set of data in the CIMS study, making it imperative that studies strive for less “missingness” overall. This will have a positive impact on the collection of critical readership information.

It is clear from these tests that the type of imputation technique to be used is not alone a sufficient answer to addressing the issue of item non-response. Front-end strategies must be employed along with imputation to manage missing data. For example, with the design of the next version of CIMS, a great deal of care was taken to reduce respondent confusion and improve the respondent experience with the study; response lists have been reduced, language has been made more plain, direct and specific. And, perhaps most important, as CIMS is conducted online programming logic can dictate that respondents are only presented questions for which they are suited. All these are efforts that can reduce respondent fatigue that lead to item non-response.

Appendix A - Variables used in regression analysis (Readership Test):

employed in worksite (weighting variable – 1 to 9, 10 to 49, 50 to 249, 250 to 999, 1,000+)
Gender (weighting variable)
Job role (weighting variable – MIS, senior manager, departmental)
Type of influencer (weighting variable – primary, authorize, informal, evaluate)
Products involved (Q3A – desktops, notebooks, handheld devices, servers, software, peripherals, networking, telecommunications, services)
Scope of influence (Q3C – highest among all categories involved)
Home technology influencer (Q3E)
Total expenditures planned next 12 months (sum of all categories)
Sports/activities (Q10B)
Age (Q10C)
Children in household (Q10D)
Education level (Q10E)
Formal influencer (Q10G)
Primary influencer (Q10H)
Full-time/part-time (Q10I)
Work location (Q10J)
Job title (Q10K)
Manager (Q10L)
Primary job function (Q10M2)
E-commerce/e-business (Q10Q)
Household income (Q10S)
Business activity (Q10T)
employed in workgroup (Q10W)
employed in department (Q10Y)
employed in organization (Q10CC)
PCs in department (Q10EE)
PCs in organization (Q10HH)
Company revenue (Q10II)
Location of PC use (Q11A)
Year began using PC (Q11B)
Computer ability (Q11C)
Access the Internet (Q11D)
Location of Internet use (Q11E)
times access the Internet at work in avg month (Q11F(w))
times access the Internet at home in avg month (Q11F(h))
Total spent online last 3 months for work (Q11L(w))
Total spent online last 3 months for personal (Q11L(p))
Year first accessed the Internet (Q11O)
Provide advice in organization (Q11U)

Appendix B - Variables used for missing data imputation (Product Behavior Test):

(Q11B – Year first accessed the Internet)

Computer ability (Q11C)
Age (Q10C)
Education level (Q10E)
Uses Manufacturer/Reseller Sites for vendor selection (Q11N)
% of pages read – *The Atlantic Monthly* (Q1D)

(Q11GH – In minutes, avg. length of online session at home)

Time spent with media yesterday - Internet (Q2G)
% of time spent online for business purposes (Q11H)
Uses Internet for forums, chat, USENET groups for personal reasons (Q11I)
times in avg. month access the Internet at work (Q11F1a)
Access the Internet at work (Q11E)

(Q11V.04 – attitudinal - agree or disagree that low cost is most important factor in tech decision)

Attitudinal - agree or disagree that it is better to use proven technology rather than the newest technology (Q11V)
Purchased music (CDs, tapes) online in past 3 months (Q11J2)
Purchased software direct from a manufacturer's in-person sales force (Q5N)
Program frequency – Becker (Q2D)
employed in workgroup (Q10W)

(Q11V.10 – attitudinal - agree or disagree that keeping up with tech changes is a hassle)

Attitudinal - agree or disagree that I worry about using my credit card for an online purchase (Q11W)
products intend to purchase – remote access routers or concentrators with VPN or firewall capabilities (Q7H)
Attitudinal - agree or disagree that it is better to use proven technology rather than the newest technology (Q11V)
Publication screen-in – *Entertainment Weekly* (Q1A)
Attitudinal - agree or disagree that it's hard to get enough information on things sold over the Internet (Q11W)

References

Appel, V., A. Elder, T. Incalcatera (1999), "Measuring Print Audiences Via the Internet" ***Proceedings of the Worldwide Readership Research Symposium*****,** Florence, 509-516.

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*. J. Wiley & Sons, New York.

Schafer, J.L. (1999) NORM: Multiple imputation of incomplete multivariate data under a normal model, version 2. Software for Windows 95/98/NT, available from http://www.stat.psu.edu/~jls/misoftwa.html.

Vriens, M. and E. Melton (2002), "Managing Missing Data: Improving Data Quality with Multiple Imputation", *Marketing Research*, Fall, 12-17.