The case of the missing data

Received: 3rd January, 2002

Sam Koslowsky

is vice president of modelling solutions for Harte-Hanks, Inc. (NYSE:HHS), a worldwide direct and interactive marketing company that provides response management, marketing database and marketing service solutions. He is a regular speaker on data modelling at DMA conferences.

Abstract One of the most critical issues in model formulation and marketing analytics is how to handle missing data. If not handled correctly, even the best analysis efforts can fail, and even worse, an entire database marketing strategy can be seriously damaged.

This paper will deal with the typical data types that may be missing, both in marketing research and analysis settings. Several proven methods of negotiating the missing data dilemma will be discussed. Less desirable approaches will also be highlighted.

INTRODUCTION

Few, if any, analysts, have not encountered the plague of missing data. Whether they be market or survey researchers, database analysts, biometricians or a host of other professionals, all have confronted missing data, as they have sought knowledge within various data collections. Missing data have been described as questions without answers or variables without cases. Even a small percentage of missing data can be serious. Results can be misleading and conclusions difficult to make. The following case extracted from the pages of a market research study illustrates this.

A marketer is presented with the information shown in Table 1, it is about the participation of a surveyed population in life insurance products.

The information is designed to provide input as to whether or not the firm should proceed with offering a new

insurance service. The decision rule adopted by management states that if at least 33 per cent of the customer base maintain an insurance product, the firm will introduce the new service. Because of the large number of 'refused to answer' category, examination of the above analysis provides inconclusive results.

Several approaches are available in dealing with these data:

- the 194 'refused to answer' responses could be deleted. This leads to 136 responding affirmatively out of a possible 347, ie 39.1 per cent. This exceeds management's hurdle rate, and product development would continue
- as the analysis highlights only 25 per cent responding 'yes', which is the critical figure, any ideas of proceeding with the new service could be dropped since the 33 per cent threshold has not been reached

Sam Koslowsky

Vice President, Modeling Solutions, Harte Hanks CRM Analytics, 55 Fifth Ave, 14th Floor, New York, NY 10003-4301, USA.

Tel: +1 (212) 520 3259;
Fax: +1 (212) 696 9127;
e-mail: sam_koslowsky@harte-hanks.com

Table 1

Do you have a life insurance product?	Raw count	% of total
Yes	136	25.1
No	211	39.0
Refused to answer	194	35.9

- the 194 ‘refused to answer’ category could be recontacted. This scenario may or may not lead to the required 33 per cent. If, for example 31 additional individuals responded with ‘yes’, and the remaining 163 respond with ‘no’, then the true proportion would be $(136 + 31)/541$ or 30.8 per cent. Alternatively, if 31 respond ‘no’, and the 163 go for ‘yes,’ then the proportion is $(136 + 163)/541$ or 55.2 per cent
- it could be assumed that the missing data are random in nature and that the 194 will be distributed between the other two groups in the same proportion as they are now among those who did not respond. There are 347 ($136 + 211$) individuals that provided an answer. So $136/347$ or 39 per cent of the 194 could be taken. This results in 66 being added to the ‘yes’ category. And $211/347$ or 61 per cent of the 194 or 128 could be taken and assigned to the ‘no’ group.

What does management do? How can this sort of problem be handled? Missing data are a part of almost all research, and it is critical to be able to manage them intelligently.

MANAGING MISSING DATA

Missing data arise for a number of reasons. In survey research, respondents may be unwilling to answer all questions. Human error may be the cause. The interviewer may have forgotten to record

Table 2

Account number	Age of account
2001	34
2002	41
2003	Missing
2004	40
2005	37
2006	Missing
2007	43
2008	Missing

Table 3

Account number	Age of account
2001	34
2002	41
2003	69
2004	40
2005	37
2006	68
2007	43
2008	68

the responses. In database marketing, it may be impossible to secure a certain piece of data for some segments. It is common not to be able to enrich databases with data for younger consumers who are still in university. All these factors result in missing data.

With the proliferation of data mining technologies, managers and analysts still find themselves confronting serious problems when it comes to negotiating missing data. The issue, if not handled correctly, can damage the best modelling efforts and destroy well-conceived marketing programmes. The eight records in Table 2 offer an example.

Disregarding the missing data, an average age of 39 is calculated for the five ‘full’ records. With these results, a manager might consider framing a creative offer to fit this population. Investigating further, it can be determined what these missing points might have been (see Table 3). Assuming the ‘filled in’ ages are as shown in Table 3 a mean age of 50 is now calculated. A

Table 4

Account number	Married?	Gender	No. of cars owned	Age
1421	Yes	Male	0	Missing
1422	Yes	Missing	1	34
1423	Missing	Female	1	55
1424	No	Female	2	41
1425	Yes	Missing	Missing	26

new offer or creative might be appropriate.

Some marketers see fewer missing data than others. Credit grantors keep transaction behaviour of their customers. Because billing the customer is a key function, data related to the 'billing' function are rarely, if ever, missing. In the same way, a retailer must secure all data related to a transaction. This includes price, discounts offered, time of day, number of items sold, etc. This is not a complex task. Retail point-of-sale systems capture this on a continuing basis.

These behavioural-associated measures are the primary forces in modelling assignments. Other types of data may, however, also be employed. A mutual fund organisation might require potential customers to complete an application. Frequently, the form is incomplete. On some lengthy application forms, it is not unusual to ask 15 or 20 questions. Not everyone will complete the form. Some applicants may apply by phone where, in order to streamline the application process, a select few questions are asked. All this results in missing data.

The entire problem is even more severe when external data are appended. Data appended on some geographic level will seldom result in missing data. Everyone lives somewhere. If a postcode can be identified, for example, it should be possible to append data. On the other hand, matching data by individual always results in missing data. Not all data are available for everyone.

Analysts are frequently asked to

identify data that best distinguish groups of people. For example, respondents *vs.* non-respondents, defectors *vs.* active account holders, or poor credit risks *vs.* satisfactory credit risks.

In most circumstances, the marketer confronts a situation where data may be missing for some individuals, and a different set of data is unavailable for yet another group of records. The records in Table 4, extracted from a marketer's database, have a variety of data issues. Only account number 1424 has a complete data record associated with it. There are no missing data. The other customers all have some piece of data missing. If, gender, for example, proved to be unimportant in the data mining exercise, then account numbers 1422 and 1425 would not create any negative impact. If, however, gender proved to be a predictive data element, the analyst could have trouble incorporating these two records because the gender data are missing.

No discussion of missing data would be complete without presenting an overall approach to determining the severity of the problem in a particular circumstance. A generally accepted procedure for understanding the problem involves building a table that summarises the various combinations of missing values. Table 5 is an illustration.

The analysis in Table 5 provides the analyst with some valuable insight as to how to proceed. The first step is to review what is available with the output. There are 11 patterns of data listed. Number 1 is complete. That is, there are

Table 5

Pattern no.	No. of observations	Occupation	Length of residence	Presence of children	Dwelling type	Number of children	Income	Occupation (2)			Gender (2)		Length of residence	Number of children
								Complete if (1)	Non-professional	Professional	Male	Female		
1	5,060							5060	2,730	2,330	1,630	3,430	4.30	2.10
2	230							5290	120	110	50	180	5.10	.
3	140							5200	80	60	60	80	7.30	.
4	310							5370	120	190	110	200	6.10	1.71
5	110							8050	30	80	30	80	3.50	0.82
6	2,570							7630	1,760	810	460	2,110	3.60	0.99
7	310							8570	0	0	80	230	4.60	1.60
8	630							5690	0	0	180	450	2.90	2.01
9	140							5200	80	60	99	41		2.21
10	120							7890	100	20	112	8		1.10
11	170							8030	130	40	70	100	12.10	

(1) No. of complete observations if variables missing in that pattern were excluded

(2) No. of observations with this combination of missing variables

5,060 records that have no data missing. Pattern 2 is missing 'number of children' data. The highlighted cell indicates 'missing values'. In this particular pattern, 230 observations are included. Pattern arrangements 5 and 11 are defined by having two data items missing. 'Dwelling type' and 'income' are missing for the former, while 'number of children' and 'income' are unavailable for the latter. An analyst interested in probing the data further by occupation and gender, for example, can split out the results by these demographics, as well. The centre part of the table highlights these results for each pattern. Finally, average values for critical variables can also be appended to provide additional guidance. The last section calculated these means. If, for example, income was a metric under study, pattern 6 shows that income is missing in about one-third of all records — a fairly substantial quantity.

Several methods have been suggested to deal with missing data. All the approaches have some merit and some

challenges. None of them are perfect.

These procedures include:

- converting all missing data to a zero value, simply placing a '0' wherever a missing data item is encountered
- computing the average of the data field, and inserting the average wherever the data are missing. In Table 2 under 'Age of account', five records contain a value. The average for these records is calculated and turns out to be '39'. This value is inserted for records '2003', '2006' and '2008'
- computing averages by some logical segment. 'Gender' may be used as a segment, for example. An average is calculated for MALE; another average is computed for FEMALE. These averages would then be substituted for missing data depending on the gender of the record
- calculating the median and inserting the median value for the missing data items

Table 6

Criteria for segmentation	Home value
Average for those under 42 years old	\$145,000
Average for those greater than or equal to 42	\$201,000
Average for all records	\$183,500

- deleting the records that have missing data items
- ignoring the data elements that appear to be missing for many individuals
- developing a separate model for the missing data items and using the result of this model as a substitute for the missing data
- developing separate models, one model where the data are present, and one where the data are missing
- utilising a flag to indicate when a piece of data is missing. This flag then becomes a candidate predictor in the model
- using software that ‘handles’ missing data.

Which procedure is the one to use? Each one is discussed briefly below.

Converting all missing data to a zero value

This approach should be avoided. It is simple to do, but most analysts would agree that there is a significant analytic problem. How is it possible to distinguish between a real value of ‘0’ and a substituted value? It can be argued that replacing the missing value with any number would accomplish the same thing.

Computing the average of the data field and inserting the average wherever the data are missing

Known by analysts as mean substitution, this is probably the most

often used solution. If for example, the mean income is \$53,500, then records where income is missing would be assigned a value of \$53,500. The thinking behind this approach is that, without any additional knowledge about a person, given a normally distributed variable, the best guess is probably the mean.

If, however, many of the records have a missing value for the same data field, this approach would force all these missing data items to have the same value. This could complicate profiling. If fewer records have missing values, then this option is more viable.

Computing averages by some logical segment

Known more familiarly as ‘group mean substitution’, this is certainly a step up from direct mean substitution. The idea is to locate groups that are known to be reasonably homogenous but may have logical variations with the missing data. The assumption is that a mean closer to the real missing data value can be computed when splitting up averages by segments. If, for example, age is the criterion used for segmentation and ‘Home value’ is missing on many records the scenario in Table 6 may occur. Clearly, the substituted missing value would be different if the overall average or the segmented ones were used.

This routine will produce better results than mean substitution. It is important to be sure that a meaningful

number of records are available for each segment. Otherwise an average that is not the true value may be computed.

Calculating the median and inserting the median value for the missing data items

A frequently used approach, this practice moderates the problem that often occurs when averages are used on skewed populations. For example, assume the income of five individuals is:

- 1 – \$18,000
- 2 – \$27,000
- 3 – \$29,000
- 4 – \$36,000
- 5 – \$1,500,000

A simple average of this data set results in \$322,000. The median value is \$29,000. Of course, an argument can be made that the average is not consistent with any of the individual values. The median, on the other hand, appears to be more appropriate.

Deleting the records that have missing data items

‘Listwise deletion’, as it is popularly called, is a conservative and usually reliable solution. It is most often the default option available on many statistical analysis software packages.

If only a few of the records are deleted, this works fine. If, however, all the deleted records represent a distinct profile or mind-set, the marketer may be missing out on leveraging this ‘unique’ population. For example, it is typically more difficult appending wealth indicators to higher income individuals. The data arrive missing. Many marketers do not want to delete this segment from consideration.

Ignoring the data elements that appear to be missing for many individuals

‘Pairwise deletion’, as it commonly called, is used by many researchers. This approach keeps the set of data elements that is ‘complete’ and disregards the set of data elements that are ‘incomplete’. So if ‘age’ were missing for many data records, the entire data element would be fully ignored, and not considered in any subsequent analysis.

Many managers are, of course, interested in incorporating critical indicators, such as age, into their analysis. Deleting these data elements would, of course, make ‘age’ difficult or impossible to use.

Developing a separate model for the missing data items

A creative way of dealing with missing data, this can work well. Since the data are missing, it is necessary to arrive at a best estimate of what the missing values are. So a model is developed to predict what these missing values may be. These estimates are then used as potential predictors for the analysis.

While requiring more effort on the part of the analyst in terms of performing multiple analyses, it does permit the researcher to use a ‘complete’ set of data.

Developing separate models

This is a viable but infrequently used method. Records with non-missing data are subjected to one model, and the other records are input into another model. Perhaps the incremental resources needed to develop a second model make this approach a seldom-used one. The author finds it to be an excellent solution.

Utilising a flag to indicate when a piece of data is missing

Missing data may themselves be meaningful. The author creates an indicator signifying the presence of missing data for a particular record. The indicator is turned off when the data are present. This flag becomes a potential predictor. This is particularly useful when dealing with wealth-related marketing segmentation. Frequently, it is more difficult securing data for wealthier individuals.

Using software that 'handles' missing data

Software exists that will, 'behind the scenes', fill in the blanks for missing data. These programs are available either as part of a data-mining suite or as stand-alone programs. The author's general rule of

thumb is to be sure to know what is going on and how the missing data are being managed. These programs are designed for analytics professionals.

CONCLUSION

Successful data mining and analysis requires completion of many tasks. Some of these take time and consistent effort. Indeed, some problems cannot be easily solved, if at all. Missing data, however, are not one of these. There are viable approaches to handling the missing data predicament, depending on the firm's resources, needs and circumstances. One thing is certain, however. Ignoring the problem with the expectation that results will not be affected is incorrect. The analysis will be lacking, and worse, the resulting programmes may very well be ineffective.