Preparing for Basel II

Common Problems, Practical Solutions

Part 1: Missing Data

by Jeffrey S. Morrison

This article continues the series introduced in the May and June 2003 issues of *The RMA Journal* on modeling requirements for Basel II. Previous articles focused on the fundamentals associated with building PD and LGD models, as well as the steps necessary for model validation. This and forthcoming articles discuss common challenges in model development and offer practical statistical advice in overcoming them. To make this article as helpful as possible, examples use the SAS software language.

Missing or incomplete data is perhaps the most widespread challenge facing any model builder, regardless of the industry. Whether working with results from a survey, developing a telecom model to predict churn, predicting the response to a new credit card offer, or creating a model for probability of default, you must determine a strategy for handling missing information. You may, for example, have information on one loan applicant's income but not on others. The same may be true for the consumer credit score. If you feel that income and credit score are both very important in determining future default, what do you do with those accounts where the information simply is not available? Because your loan accounting system may be less than perfect, how do you handle an obvious input error, such as an annual income of -$100 or an LTV of 500%?

The truth is that *every* model builder will select some way to deal with the missing data problem before a model is estimated. As there are a variety of methods available, the task is to find the best. The decision will affect not only the model estimates or parameters, but also their statistical reliability. The remainder of this article presents several popular approaches for handling the missing data problem, beginning with the most simple. The last approach presented reflects the latest research on the subject and is made surprisingly easy to

 Jeff Morrison is vice president, Credit Metrics—PRISM Team, at SunTrust Banks, Inc., Atlanta, Georgia.

implement because of some recent improvements in certain statistical software packages.

Approach 1: Fix Those Errors

Many times, information is missing because of errors in the data collection process. If 40% of bureau-score data is missing, then you need to find out why. Was there a matching problem within the data warehouse? There will always be accounts that the credit bureaus are unable to score—but 40%? For income data, was there a systematic type of error made that could be corrected with some degree of certainty? Should blanks in your data really be interpreted as zeros? Most importantly, there is no substitute for careful examination of the dataset before doing any analyses.[1]

Approach 2: Delete Those Records

An obvious approach is to simply delete the records with missing information. However, doing so may create two problems.

First, if you have a number of predictor variables in your model with missing information, a deletion approach can significantly reduce the number of observations for modeling. For example, you may start off with 10,000 observations containing both defaulted and non-defaulted accounts. However, a deletion approach could result in a data set of 1,200. Sometime referred to as *case-wise deletion*, most regression routines automatically skip records where any predictor variable is missing.

Second, such an approach could bias your estimates. This happens because the population represented by your sample may have a different distribution of missing information. In other words, the sample needs to be as representative as possible of the overall characteristics of the general population. If the bias is severe, it could affect forecasting accuracy when you get to the validation stage of the process.

Approach 3: Substitute a Proxy for Missing Information

Since data is such a rare commodity in the model-building world, most modelers do not choose the second approach to handling missing information. Perhaps the most common approach is to substitute the mean, median, or mode of the variables for which you *do* have valid information. For example, if you are missing information on the credit bureau score, and the average bureau score in your sample is 689, then you would substitute that value each time you encounter missing data. The same procedure could be done using the median value of existing predictors. The decision to use the median rather than the mean value as a proxy could be empirically determined by looking at the validation results. In other words, try it both ways!

However, the proxy method, too, can cause a problem in the model-building process. Let's say that our credit bureau score is 20% missing. If we substitute 689 each time for the missing value, then we artificially reduce the variance of the predictor variable in the model. Since each missing value has been substituted with a single numeric value, the overall variability of the predictor has been made artificially low. Since this variance is crucial in determining whether a variable is statistically significant in the model-building process, you could be inadvertently adding statistical significance where none exists. Therefore, at the end of the model-building process, you may have predictors in the model that shouldn't be there.

Approach 4: Substitute a Proxy for Missing Information by Considering Default Rates

This approach is similar to the third approach, but here, we consider how to pick the proxy using information about the outcome event in question—in this case, payment default. Let's say the mean of available data for the bureau score is 689. If you are modeling default and using Approach 3, you implicitly assume that the average bureau score is the same between defaulters and non-defaulters. However, in real life, a missing bureau score may tend to be more prevalent for the defaulting population. Approach 4 allows you to pick a more realistic proxy for missing values. Although there are a number of ways to accomplish this, one procedure is illustrated by the following example, where you need only two pieces of data—a default indicator and your bureau score.

- Step 1: Divide the records for the variable in question (the bureau score, for example) into percentiles, leaving out those that have missing values.
- Step 2: Use these percentile values to create a new variable (bureau_score2) representing these percentile

ranges—a process referred to as *discretizing*. (See Figure 1.)

- Step 3: Put the accounts with a missing bureau score into a different group by assigning them an arbitrary value like −999999.
- Step 4: Compute the default rates for each discretized group. The default rate is simply the number of accounts that defaulted divided by the total number of accounts in the data.
- Step 5: To obtain your proxy for missing data, simply pick the bureau score value associated with the group that comes closest to the default rate of the missing group.

Figure 1 shows that those accounts with missing bureau scores had a default rate of .55%. Group 2 (accounts with bureau scores from 581 to 628) was found to have a default rate of .58%, a value closest to that of the missing group. Therefore, we would select a value of 628 as the new proxy for a missing bureau score rather than a value of 689 as might have been suggested using Approach 3.

Approach 5: Dummy Variables

Another popular approach is to create categorical, or dummy, variables. This method lets your model directly estimate the default risk associated with accounts that do not have, for example, a valid bureau score. An example of this coding in SAS might be as follows:

```
If Bureau Score = .  Then DUMMY=1;
Else DUMMY=0;
```

Missing values in SAS are often coded as a dot (.), as in the above example. If the variable DUMMY were included in the regression model, then that parameter estimate would reflect the risk associated with missing information for the bureau score. Although this method has its advantages, there are two disadvantages. First, if you have to create these types of variables for all the predictors in your model, then these series of 1s and 0s can create multicollinearity problems. *Multicollinearity* is a statistical problem where two or more variables are so highly correlated that the regression procedure is unable to isolate their unique contribution in explaining the dependent variable. This can happen when you have missing information occurring for many of the same observations across a set of predictors. Second, the parameter estimates from such an approach can be biased and impact the accuracy of the model.[2]

Approach 6: Single Imputation

Some schools of thought argue that a better way of handling missing values is through single imputation. This method uses regression to predict the values of missing data. For example, suppose we are missing information on the bureau score. If we had other attributes correlated with the bureau score apart from our default condition, then we could use them for prediction. However, a couple drawbacks of this approach are 1) finding correlated predic-

Figure 1

Discretizing the Bureau Score into Groupings by Percentile (SAS code)

```
Group 1:  if Bureau_score >=300  and Bureau_score<=581  then Bureau_score2=581
Group 2:  if Bureau_score >581   and Bureau_score<=628  then Bureau_score2=628
Group 3:  if Bureau_score >628   and Bureau_score<=656  then Bureau_score2=656
Group 4:  if Bureau_score >656   and Bureau_score<=678  then Bureau_score2=678
Group 5:  if Bureau_score >678   and Bureau_score<=699  then Bureau_score2=699
Group 6:  if Bureau_score >699   and Bureau_score<=720  then Bureau_score2=720
Group 7:  if Bureau_score >720   and Bureau_score<=740  then Bureau_score2=740
Group 8:  if Bureau_score >740   and Bureau_score<=759  then Bureau_score2=759
Group 9:  if Bureau_score >759   and Bureau_score<=778  then Bureau_score2=778
Group 10: if Bureau_score >778   and Bureau_score<=850  then Bureau_score2=850
```

Description	Bureau Score Lower Range	Bureau Score Upper Range	Bureau_ Score2 Value	Default Rate	Missing Group Default Rate	Missing Group Default Rate Difference
Missing	-	-	-999999	0.55	0.55	0.00
Group 2	**581**	**628**	**628**	**0.58**	**0.55**	**0.03**
Group 3	628	656	656	0.34	0.55	0.21
Group 4	656	678	678	0.24	0.55	0.31
Group 5	678	699	699	0.14	0.55	0.41
Group 6	699	720	720	0.12	0.55	0.43
Group 7	720	740	740	0.05	0.55	0.50
Group 8	740	759	759	0.04	0.55	0.51
Group 9	759	778	778	0.03	0.55	0.52
Group 10	778	850	850	0.02	0.55	0.53
Group 1	300	581	581	1.67	0.55	1.12

tors may be difficult and 2) if you have a number of predictor variables with missing information, then this procedure can be cumbersome. The next approach takes this idea and builds on it significantly.

Approach 7: Multiple Imputation

Multiple imputation is at the forefront of research today in dealing with the missing-data problem. Although the computations behind the scenes are quite complex, statistical software packages make this approach not only practical, but also easy to implement. As stated earlier, most of the other approaches to handling missing data have at least one major drawback: They end up understating the variance of the predictor variables because they use a single proxy for missing. This is important in the modeling process because variable selection procedures such as stepwise regression are affected by the variance of the predictor variables. Even if no variable selection procedure is used, the modeler still will not know how reliable the variances are. Under these conditions, you could end up choosing a model containing variables that should really be discarded because they were statistically insignificant.

Multiple imputation works by creating new data sets containing no missing values. They are based on correlation information from other variables plus a random draw component. These complete data sets represent random samples of the missing values but with the same statistical properties as your original data. Since this method depends on having as much correlation information as possible about your predictor variables, all relevant data should be included in the process. The good news is that the modeler does not have to do any special pre-analysis for multiple imputation—the software performs this automatically. However, multiple imputation carries with it certain assumptions that should be met before the full benefits of the procedure can be realized. Typically, the most important of these are easily met, but the reader is encouraged to review them in the references listed at the end of this article.

Figure 2 shows an example of data with missing observations for X2 and X3. Variables Y and X1 have no missing data.

Figure 3 shows the results of a multiple imputation procedure using PROC MI in SAS. Note that no changes are made to the original data if the values are not missing. On the other hand, if the data was missing, each imputation (_Imputation_) results in a different value for the missing information. For example, the second observation for variable X3 was

Figure 2

Example of Data with Missing Information

OBS	Y	X1	X2	X3
1	1	22	7	22000
2	0	32	.	.
3	0	65	15	23500
4	1	67	.	55000
5	0	54	7	21000
6	0	44	6	.
7	0	32	5	.
8	0	55	4	78000
9	1	22	7	22000
10	0	32	.	.

Figure 3

Multiple Imputation Results
Three imputations for 10 observations
(Note: all imputations' observations are shown in bold print)

Obs	_Imputation_	Y	X1	X2	X3
1	1	1	22	7	22000
2	1	0	32	2.582024	**59104.43**
3	1	0	65	15	23500
4	1	1	67	9.719319	55000
5	1	0	54	7	21000
6	1	0	44	6	47099.09
7	1	0	32	5	34756.91
8	1	0	55	4	78000
9	1	1	22	7	22000
10	1	0	32	4.320581	39477.55
1	2	1	22	7	22000
2	2	0	32	7.117512	**23099.72**
3	2	0	65	15	23500
4	2	1	67	9.959226	55000
5	2	0	54	7	21000
6	2	0	44	6	46834.58
7	2	0	32	5	34075.01
8	2	0	55	4	78000
9	2	1	22	7	22000
10	2	0	32	7.593331	33365.2
1	3	1	22	7	22000
2	3	0	32	8.802116	**14794.59**
3	3	0	65	15	23500
4	3	1	67	9.279476	55000
5	3	0	54	7	21000
6	3	0	44	6	47439.81
7	3	0	32	5	38025.51
8	3	0	55	4	78000
9	3	1	22	7	22000
10	3	0	32	5.925172	31309.48

originally missing. In Figure 3, we now have three different values for this observation: 59104.43, 23099.72, and 14794.59.

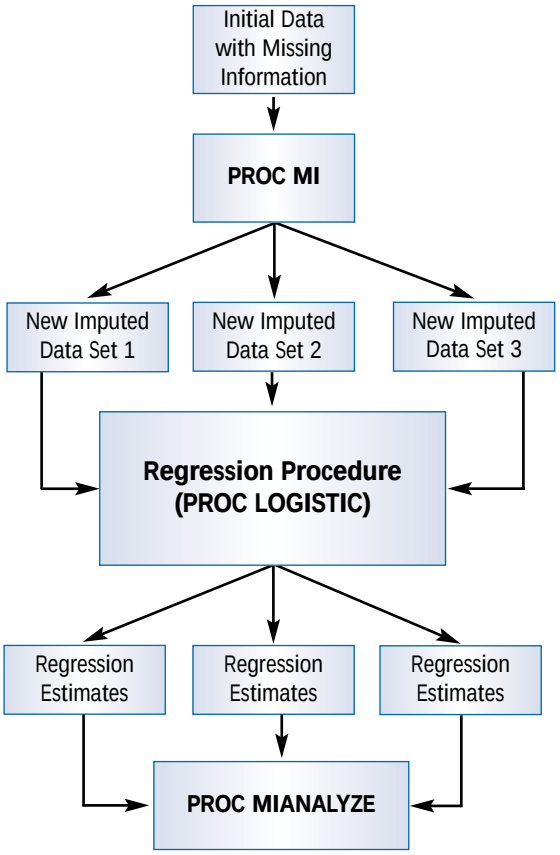
Once the imputation process is complete, the next step is to estimate your model by running each complete data set (three in all) against a standard regression procedure like logistic regression. Here, you can specify a variable selection method such as stepwise, or simply force the regression to use all predictor variables. Afterward, the final step is to simply average the results of the regressions that compute the correct standard errors and parameter estimates. In SAS, PROC MIANALYZE can be used for this exact purpose. Figure 4 shows the imputation, model development, and summarization process as implemented in the SAS software solution.

Summary

The unfortunate truth is that there is no perfect solution to the problem of missing data. Since there is no perfect solution, then data collection efforts should be geared to produce and maintain the highest level of data quality possible. However, when data does appear incomplete, there exist a number of options available to the modeler, ranging from the simple to the advanced. Regardless of which method is selected, it is important to understand why the data is missing and its implications for the modeling process. As we have seen, each approach has its advantages as well as its disadvantages. My recommendation is to try a number of techniques and perform an empirical test or validation to see which method works best in that particular situation. If time permits, it may be advantageous to seriously consider the multiple imputation method, which many view as the best approach to the problem of missing data. ❐

Figure 4

Multiple Imputation Process in SAS



Contact Morrison by e-mail at Jeff.Morrison@suntrust.com.

Notes

1 Harrell, Frank E., Jr., *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis*, Springer-Verlag New York, Inc., 2001, pp. 41-52.

2 Allison, Paul D., *Missing Data*, Sage Publications, Inc., 2001.