PLAYING WITH HISTORY CAN AFFECT YOUR FUTURE: HOW HANDLING MISSING DATA CAN IMPACT PARAMATER ESTIMATION AND RISK MEASURE BY JONATHAN LEONARDELLI

March 1, 2012

ABSTRACT

Missing data is a common problem facing financial professionals. How to handle this problem not only depends on the objective at hand, but also on the type of missing data. To that end, the purpose of this paper is to provide an overview of the types of missing data and methods used to handle them. The types of missing data covered include: not missing at random (NMAR), missing at random (MAR), and missing completely at random (MCAR). The methods discussed to solve the missing data problem include: deletion, imputation, and interpolation. Results of a comparative analysis are included in this paper to show how the different missing data replacement methods performed when estimating model parameters and risk measures.

INTRODUCTION

The area of risk, like so many other areas, is reliant on data. For example, data is used to estimate the parameters of an equation, value instruments, and determine P&L. In a sense, data can be considered the life blood of a sound risk management practice.

Unfortunately, recipients of the data do not always get what they want. For a variety of reasons, time series and cross-sectional data can land in a table with missing values. When this happens, a question presents itself: What is the best way to handle this problem?

There are a variety of techniques available to handle missing data. However, not all techniques are appropriately suited for every task. The type of method you select to replace missing values can greatly affect your results.

This paper was written to provide an overview of the missing data problem. It discusses the types of missing data as well as some of the methods used to replace missing values. These two topics are covered in the first and second sections, respectively. The third section provides the results of an analysis that was performed to illustrate the effects the different missing data replacement methods have on parameter estimation and risk measures.

SECTION 1: TYPES OF MISSING DATA

Not all missing data is the same. There are three general classifications of missing data: Not Missing at Random (NMAR), Missing at Random (MAR), and Missing Completely At Random (MCAR). These classifications have distinct mechanisms which cause the data to be missing. If one can know the mechanism that causes the missing data, then one has insight into choosing an appropriate method to replace missing values. Unfortunately, this insight is rare.

Note: there are two examples in each of the following subsections. The first example is related to time series data; the second, cross-sectional data.

Not Missing at Random (NMAR)

This type of missingness is also known as Non-Ignorable Missingness (NIM). The mechanism causing this data to be missing is *related to the missing values*.

Examples of NMAR data:

- Upon reviewing a time series of a stock's price one notices that missing data tends to occur when the previous day's stock price is in the high $50s. Further investigation reveals that the missing values occur only when the stock price is above $60.

- A question on a survey of financial professionals asks how large their bonus is in relation to their base salary. In light of the recent financial turmoil, individuals who made large bonuses (e.g., > 50%) may not respond for fear of castigation by the survey-giver.

Missing At Random (MAR)

Data in this category is missing *due to the observed values of other variables*. While NMAR's missingness is related to the variable itself, this type of missing data is unrelated to a variable's missing values.

Examples of MAR data:

- Upon reviewing a time series of a stock's price it appears that the missing values randomly occur throughout the time period. Further investigation reveals that the missing values occur when two other stocks in the portfolio have an average value greater than $50.

- A question on a survey asks financial professionals the number of years they have been at their current job. Older people are proud of their commitment to the company, so they respond. Younger people may be embarrassed about their job hopping and decide not to respond. In this case, the missing values are related to age (another variable in the survey).

Missing Completely At Random (MCAR)

In this category, the missing data is *unrelated to values of it or other variables*. Data is missing due to an entirely random process.

Examples of MCAR data:

- Upon reviewing a time series of a stock's price it appears that the missing values randomly occur throughout the time period. Further investigation reveals…nothing. (The true reason may be related to magnetic storms from sunspots giving computers electronic spasms.)

- A survey-giver flips a coin to decide whether she will ask the first question on the survey.

One may be thinking that having MCAR data is not that bad. After all, if one were to drop the observations containing missing values, then one would be left with, essentially, a subset of the population. That thinking is correct. However, the truth is that data is rarely MCAR.

SECTION 2: REPLACING MISSING VALUES

Before the methodologies for replacing missing values are presented two topics need to be covered.

Imputation versus Interpolation

The *Oxford Dictionary of Statistical Terms* defines imputation as "the process of replacing missing data in a large-scale survey" (Dodge, 2006, p. 194). Interpolation, on the other hand, is defined as "the use of a formula to estimate an intermediate data value" (Dodge, 2006, p. 204). The key distinction here is the use of the words "intermediate data value". In general, imputation is the act of *replacing missing values with an estimated or observed value*[1]. Interpolation is the act of *creating new data points between other points*.

On the surface these two concepts may appear the same. After all, both slot in a value for a missing data point. But there is a subtlety that might often be overlooked. For interpolation, an element of *continuity* is required.

For example, one may use interpolation to connect the 3 month and 6 month term on a yield curve. In this case, the continuity component is time. Contrast this with cross-sectional data. Suppose one reviews a data set and notices missing observations within, for example, the bonus percent variable. This individual may be lulled into a sense of false continuity from the position of the observations in the data set[2]. Interpolation in this case would be wrong.

Deleting Data

One of the most common ways of handling missing data is to delete the observation. There are two types of methods for deleting the data: listwise deletion and pairwise deletion.

Listwise Deletion

This is the process of deleting an observation when *one or more* variables in that observation contain a missing value.

Advantages:

- Simple
- Can compare univariate statistics (e.g., mean) across variables because the number of observations is constant from variable to variable

Disadvantages:

- (Possibly huge) loss of information
- Produces biased estimates due to lack of variability if data is not MCAR

[1] It should be noted that while the definition for imputation mentions survey data, imputation can also be used for time series data.

[2] In this far-fetched example, the person may be tempted to interpolate between the last observation with data (e.g., observation 4) and the next observation with data (e.g., observation 11).

Example:

Obs	Stock_ENR	Stock_SAM	Stock_DAL
1	.	92.47	11.76
2	66.18	92.51	11.35
3	.	.	11.53
4	67.38	93.83	11.4
5	67.59	93.16	.

In the above table, three observations would be deleted. PROC CORR performs listwise deletion when the NOMISS option is specified.

Pairwise Deletion
This is the process of *keeping* an observation when *two or more variables contain* values. This allows, for example, correlation calculations to use as many of the non-missing pairs as possible.

Advantages:

- Simple
- Uses all available data

Disadvantages:

- The number of observations varies from one variable to the next.
- Biased estimates are produced if the data is not MCAR.

Example:

Obs	STOCK_ENR	STOCK_SAM	STOCK_DAL	Correlations
1	.	92.47	11.79	SAM and DAL
2	66.18	92.51	11.35	ENR and SAM, ENR and DAL, SAM and DAL
3	.	.	11.53	None
4	67.38	93.83	11.4	ENR and SAM, ENR and DAL, SAM and DAL
5	67.59	93.16	.	ENR and SAM

In the above table, only one observation is omitted from analysis. This is because no correlation can be computed. In all other cases, at least one correlation can be performed. Therefore that observation is "preserved". PROC CORR performs pairwise deletion when the NOMISS option is not specified.

Imputation Methods

Imputation can be broken into two groups: single and multiple. Single imputation replaces the missing data with one value. The PROC EXPAND procedure in SAS can perform most of these types of imputations. Multiple imputation, on the other hand, replaces the missing data with *multiple* values. The PROC MI and PROC MIANALYZE procedures in SAS perform this type of imputation.

Mean Imputation (Single)

Uses the mean of a variable's observed observations for all missing values.

Advantage:

- Simple.

Disadvantages

- Distorts the empirical distribution of the sample data
- Provides poor estimates of quantities such as variance and covariance.
- Does not use all available information in the data set (i.e., the other variables).

Hot Deck Imputation (Single)

There are multiple ways to accomplish this, but the general procedure replaces the missing values with those from similar observations *in sample*. Cold deck imputation uses similar observations *out of sample*.

Advantages

- Conceptually straightforward.
- Uses relationships that exist in the data.

Disadvantages

- Can be difficult to define the characteristics of a similar observation.
- Continuous data can be problematic.
- Assumes data is MAR.

Regression Imputation (Single)

Uses univariate or multivariate regression models to impute missing values. For imputation purposes, the dependent variable is the variable with missing values and the independent variables are one or more of the other variables in the data set.

Advantages

- Fairly easy to implement.
- Uses existing relationships among the variables.

Disadvantages

- Reduces the variance of the sample distribution.
- Can lead to overstating relationships among variables.
- Estimated values may fall outside of accepted ranges (e.g., negative stock prices).
- Assumes data is MAR.

Multiple Imputation (Multiple)

As stated above, multiple imputation produces multiple values for every missing value. This process involves three steps. The first step creates multiple *complete* datasets (i.e., data sets with no missing values)[3]. The second step involves the user performing the desired statistical analysis on each of these data sets. The third step takes the parameter estimates from each of these analyses and combines them into one, using what is known as Rubin's Rules[4].

[3] Typically, the number of data sets created is between 3 and 10.

[4] These rules are named after the person who created them.

For example, suppose a person wants to create a regression model from a table with missing data. The first step would involve using PROC MI to create complete tables. For the sake of this example, suppose four tables have been created. Next the person might use PROC REG to estimate the betas of the model. This would be performed four times – once for each table. Once the beta estimates are created, the person would then use PROC MIANALYZE to aggregate the four estimated sets of betas in to one set of betas.

Advantages

- Accounts for the variableness in missing data.
- Produces unbiased estimates.

Disadvantages

- Complex calculations are required to provide estimates.
- Need to combine multiple results for a single inference.
- Assumes data is MAR.

Interpolation Methods

As mentioned earlier, interpolation methods are used to insert values between two points. Basically, these methods connect the dots. Interpolation techniques can be broken into two groups: deterministic and stochastic.

Linear and Cubic Spline Interpolation (Deterministic)

These two methods are deterministic (i.e., not random) interpolation methods. The simpler method is linear interpolation. This process uses a line to connect two points. Cubic spline interpolation is a bit more complex. It uses a spline to connect the points[5]. Regardless of the method, values that fall on the curve are substituted for missing data points.

Advantages

- Simple.
- Connects data points, therefore providing continuity of data.

Disadvantages

- Does not take into consideration other variables when inserting values for missing data.

Brownian Bridge (Stochastic)

A Brownian Bridge is considered a stochastic (i.e., influenced by randomness) interpolation method. Like other interpolation methods, a Brownian Bridge produces points between a beginning and ending value. The difference, though, is how it goes about doing this.

Basically, a Brownian Bridge allows one to simulate Brownian motion between two points. However, the Brownian motion process is constrained by the initial and final values. This allows, then, randomness to occur while preserving the starting and ending values.

Advantages

- Connects data points, therefore providing continuity of data.
- Allows one to incorporate randomness into a process' path.

[5] A spline is defined as a piecewise-polynomial function.

Disadvantages

- Does not take into consideration other variables when inserting values for missing data.
- Can only be used for processes following Brownian motion.

SECTION 3: COMPARING THE EFFECTS OF MISSING DATA STRATEGIES

The purpose of this investigation was to determine the effects that the different missing data replacement strategies have on parameter estimates and risk measures. Of interest was the question: Of all the available methods, which one performs best in mimicking the results obtained from a full history of data?

Setup

A three stock portfolio was created which consisted of: Energizer (ENR), The Boston Beer Company (SAM), and Delta Airlines (DAL). It was determined that about 15% of the ENR historical data would be missing. To accomplish this, values of the SAM and DAL stocks were used to set ENR to missing (which made the missing data MAR). A total of 35 different missing data scenarios were created.

It should be noted that stocks were chosen because their price can be modeled using a process known as geometric Brownian motion (GBM). There are two parameters in the GBM model: mu and sigma. Mu is defined as the expected return of the stock; sigma, volatility of the returns. These parameters are estimated from historical data.

After the missing data scenarios were created, all 35 scenarios were passed through seven methods for replacing missing values. The four imputation methods discussed in this paper were used in addition to the three interpolation methods.

The revised historical tables from these seven methods were then passed into PROC MODEL to estimate GBM's mu and sigma[6]. The resulting models were then loaded into SAS Risk Dimensions® and, using Monte Carlo simulation, were used to calculate two common risk measures: Value at Risk (VaR)[7] and Expected Shortfall (ES)[8]

Along with these seven methods of data replacement, an additional condition that preserved the missing values was included in the analysis. These eight conditions were compared to each other as well as values from the Ideal scenario (i.e., no missing data).

Results

For comparative purposes, it was determined that Analysis of Variance (ANOVA) would be used. The purpose of using this statistical method was to determine whether there were differences in the estimates' means between the replacement, and missing data, conditions. While ANOVA provides an overall test of difference, what was of particular interest was how these conditions relate to the Ideal case and each other. As a result, confidence intervals were used to show the potential relationships.

[6] Parameters of the GBM models for all three stocks were estimated.

[7] VaR is defined as the maximum amount of money that can be lost over a period of time at a given confidence. E.g., a 10-day VaR of $1,000 at 95th percentile can be interpreted as: I am 95% confident I will not lose more than $1,000 over the next 10 days.

[8] ES is the average of values in a loss distribution that exceed the VaR value. E.g., a 10-day ES of $2,500 at 95th percentile can be interpreted as: Given that I'm in the 5% worst case scenario, the average amount I can lose over 10 days is $2,500.

Figure I shows the confidence intervals for the mu estimate. Welch's ANOVA was used due to the non-homogeneity of variances. As can be seen from the p-value, the null hypothesis can be rejected. Therefore, it can be concluded that the various methods for missing data replacement are different from each other. The value of mu for the Ideal case is 0.00064. Some comments:

- Confidence intervals for the interpolation methods tend to be smaller than those of the imputation methods. (This is common in all other parameter estimates).
- Most of the interpolation methods provide estimates for mu *below* the Ideal case; imputation methods produce estimates *above* the Ideal case.
- Not only does Hot Decking provide the largest estimates of mu, it also has the largest confidence interval (excluding the Missing condition)
- Two conditions have confidence intervals that contain the mu value from the Ideal case – Missing and Brownian Bridge.

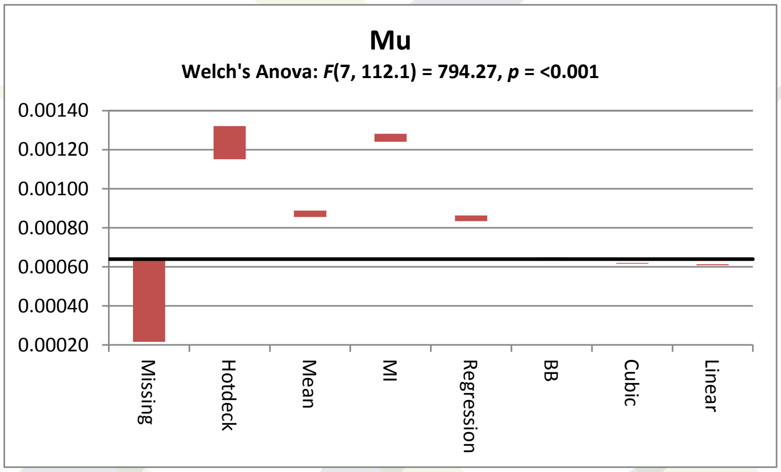

Figure 1 Confidence level comparisons for mu

Figure 2 shows the confidence intervals for the sigma estimate. As can be seen by the p-value, the null hypothesis can be rejected. Again, it can be concluded that the various methods for missing data replacement are different from each other. The value of sigma for the Ideal case is 0.01744. Some comments:

- The confidence interval for the Missing condition became smaller.
- While it is difficult to tell from the figure, none of the confidence intervals include the Ideal value. However, Missing is slightly below; Brownian Bridge, slightly above.
- Interpolation methods still tend to provide estimates below the Ideal value; imputation methods, above.

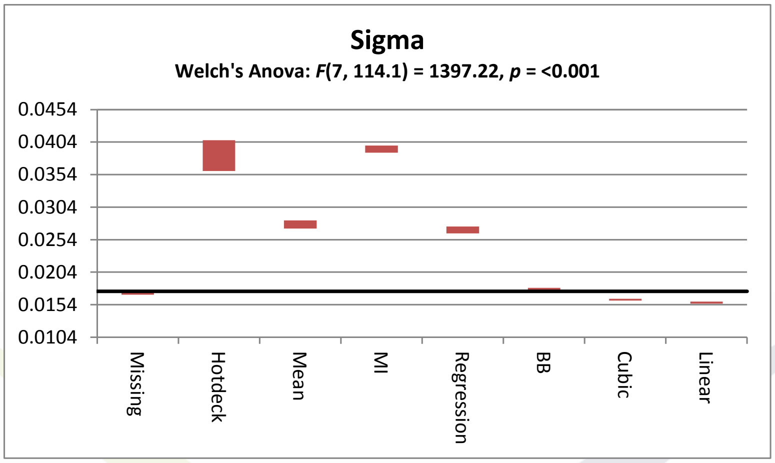

Figure 2 Confidence level comparisons for sigma

Figure 3 shows the confidence intervals for the VaR estimate. As can be seen by the p-value, the null hypothesis can be rejected. Again, it can be concluded that the various methods for missing data replacement are different from each other. The value of VaR for the Ideal case is 2.10. Some comments:

- For this metric, the Missing condition is the only confidence interval that contains the Ideal value. The confidence interval for Brownian Bridge is slightly above the Ideal value.
- Interpolation methods still tend to provide estimates below the Ideal value; imputation methods, above.

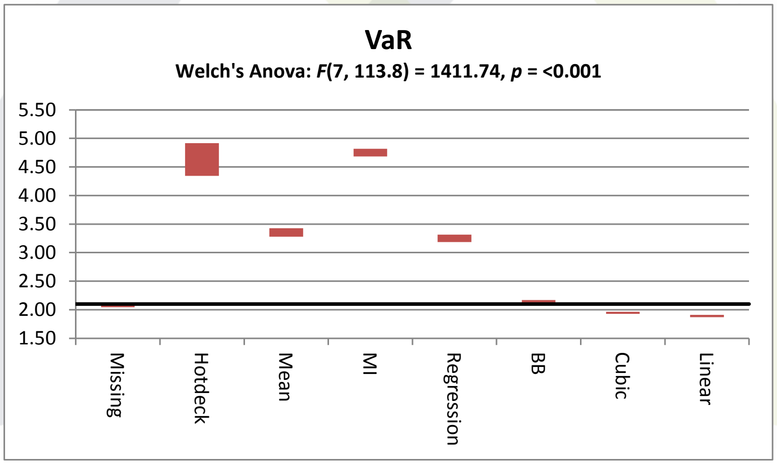

Figure 3 Confidence level comparisons for VaR

Figure 4 shows the confidence intervals for the ES estimate. As can be seen by the p-value, the null hypothesis can be rejected. Again, it can be concluded that the various methods for missing data replacement are different from each other. The value of ES for the Ideal case is 2.64. Some comments:

- Like VaR, the Missing condition is the only confidence interval that contains the Ideal value. The confidence interval for Brownian Bridge is, again, slightly above the Ideal value.
- Interpolation methods still tend to provide estimates below the Ideal value; imputation methods, above.

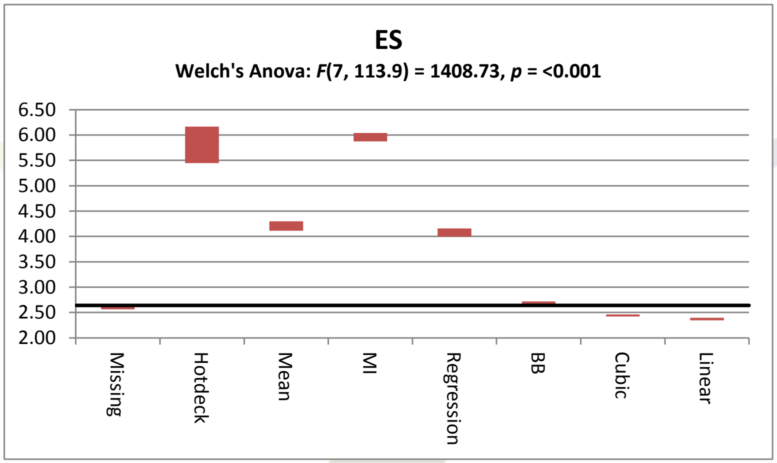

Figure 4 Confidence level comparisons for ES

General Discussion

After reviewing all of the results, there are additional comments that can be made:

- While the Missing condition appears to have performed best, caution should be taken with that conclusion. The fact that the Missing condition contains the Ideal value for three out of the four estimates suggests that the data might be MCAR instead of MAR. As stated previously in the paper, if the data truly is MCAR then one would expect the sample to be a good estimate of the population.
- The Brownian Bridge technique, though an interpolation method, performed the best out of all data replacement strategies for this example. This is likely due to the fact that when building a bridge, one uses the process' volatility as well as random shocks.
- In this example, imputation methods result in higher risk measures. This implies that these methods produce more financially conservative estimates. However, these estimates are on the order of 1.25 to 2.25 times the Ideal case.
- Cubic and Linear interpolation, on the other hand, produce risk measures that were lower than the Ideal value. It could be argued, after looking at the tightness of their confidence intervals, that these methods reduce the volatility of the data. This is not necessarily a good thing when one is trying to assess the risk in a portfolio.

Future Directions

The purpose of this analysis was to act as a "cautionary tale". Given the results obtained from the analysis, though, there are some areas for future research:

1. Adjust the number of missing values in a scenario. For the purpose of this study about 15% of the data was set to missing. How do the replacement methods perform if there was only 5% missing data? Thirty-five percent?
2. Create a different mechanism to make the data MAR.
3. Develop a different technique to implement Hot Decking for continuous data.
4. Create a portfolio of financial instruments that are more correlated in order to see if there are improvements with the imputation methods.

CONCLUSION

Missing data happens. Therefore, in order for individuals to best handle the situation when it arises, they need to ask themselves two questions: *Do I have an idea as to what type of missing data this is? What tools are available to me to replace the values?* The answers to these questions should help guide individuals to a solution.

To provide an illustration as to the impacts of missing data replacement, an analysis comparing the results from the different strategies was performed. From this analysis, it was shown that the replacement methods have a drastic effect on parameter estimates and risk measures. In addition, the analysis showed that not one replacement method performed best (i.e., always contained the Ideal value), although some appeared to perform better than others.

REFERENCES

Allison, P. D. (2002). *Missing data.* Thousand Oaks, CA: Sage Publications, Inc.

Dodge, Y. (2006). *The oxford dictionary of statistical terms.* New York, NY: Oxford University Press.

Enders, C. K. (2010). *Applied missing data analysis.* New York, NY: The Guilford Press.

London, J. (2005). *Modeling derivatives in C++.* Hoboken, NJ: John Wiley & Sons, Inc.

Wang, J. (2003). *Data mining: Opportunities and challenges.* Hershey, PA: Idea Group Publishing.

Wayman, J. C. (2003). *Multiple imputation for missing data: what is it and how can I use it?* Paper presented at 2003 Annual Meeting of the American Educations Research Association, Chicago, IL. Retrieved from http://www.csos.jhu.edu/contact/staff/jwayman_pub/wayman_multimp_aera2003.pdf

ACKNOWLEDGEMENT

I would like to thank Henry Skiba, an employee at the Financial Risk Group, for helping set up the Risk Dimensions® environment and for his willingness to do any ad hoc work I threw his way.

CONTACT INFORMATION

For comments and questions, please contact the author at:

Jonathan Leonardelli, FRM
The Financial Risk Group
Email: jonathan.leonardelli@frgrisk.com