Dealing with Missing Data – The Art and Science of Imputation

Authors: Kimberly Roye, Dustin Hilton, Christian Smart

Galorath Federal

Abstract

Missing data is a common occurrence. Even when a data set includes many data points, many variables of interest will have omitted values. The most common way to deal with this situation is to exclude the data points from analysis. However, this is not ideal. We discuss a better way to deal with this issue, which is the use of imputation, a statistically rigorous method for filling in the holes.

Introduction

Missing data is a common phenomenon most analysts have experienced. Even when a dataset includes a significant number of data points, many of the variables of interest will have missing values. The most prevalent method for dealing with such data points is to leave them out of analysis. This method is not ideal for multiple reasons. One is that unless the data are missing completely at random, leaving out data points with missing values will bias the results of analysis (Enders, 2010, page 37). A second is that it leads to smaller data sets used for analysis. Deleting data points has been demonstrated through numerous empirical studies to be one of the worst methods for dealing with missing data (see for example Arbuckle 1996, Azen et al. 1989, Brown 1994, Ender 2001, Enders and Bandalos 2001, Haitovsky 1968, Kim and Curry 1977, Kromrey and Hines 1994, Wothke 2000).

Most of the time in defense and aerospace applications datasets are already small. Not using all the available data means that analyses are based on even smaller datasets, which reduces the power of the analysis and makes them more prone to overfitting. In this paper, we discuss the use of imputation to overcome the issue of missing data. Imputation is a proven statistical technique to fill in missing data points, allowing analysts to use all the available data. We discuss two current best practice techniques – Expectation Maximization (EM) and Multiple Imputation via Chained Equations (MICE) – and discuss the pros and cons of each. We discuss the limitations of imputation and the best situations for its application.

Understanding Missing Data

Data is the core of sound data analysis. Without data, we have nothing from which to draw conclusions; nothing from which to derive averages; nothing from

which we can make comparisons or forecasts. Indeed, data are the foundation of credible estimates. See Figure 1 for the conceptual data pyramid.

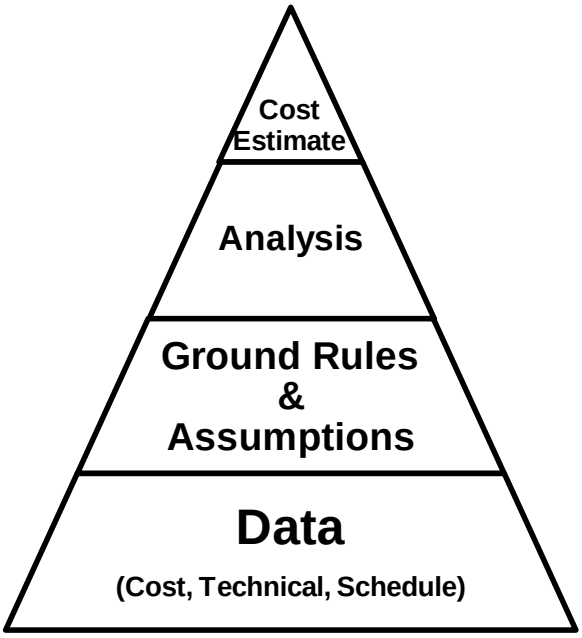

Figure 1. The Estimating Pyramid

Datasets often have missing values. When data is missing, we must first understand the reasons why. At times, the gaps mean more than the absence of a response; they can mean the answer does not exist or is not applicable. Think of the United States Census form. If no response is provided to the question of race for Person 1, it does not mean that the respondent belongs to no race. The respondent may have just declined selecting a race. But if the response for the name of Person 7 is left blank on the form from a household of six members, this blank is not an omission; the response is "Not Applicable".

Issues with Data Gaps

When data is missing at random, we often exclude the missing cases. This results in fewer degrees of freedom in the datasets used for analysis. When conducting analysis on typically smaller Department of Defense datasets, each data point is very valuable. If you have a large dataset with over 10,000 observations, the loss of 20 percent due to missingness will not reduce the predictive power nearly as much as it will on a dataset with 25 observations.

Besides reducing the degrees of freedom in a dataset, having gaps in your data will also prevent the use of certain machine learning techniques. Though algorithms like regression trees and random forests can be used on data with

missing values, other more advanced techniques like Support Vector Machines require data gaps to be filled. Imputation can be a useful technique that allows the use of all available data and facilitates the exploration of data using a variety of machine learning techniques.

Methods that Allow Missing Data

Missing data is a pervasive problem in virtually any discipline that employs quantitative research methods. However, there are multiple methods that can be incorporated in analysis when missing data is encountered.

Complete-case Analysis

This method is an approach in which you exclude any records with missing data. As mentioned, when dealing with Department of Defense data, datasets are often much smaller than commercial industry datasets. Each data point is like gold; we do not want to toss out a valuable source of information that could help to impact decisions supporting the warfighter. When performing regression in a program like R, data points are automatically ignored if missingness exists. A disadvantage with this method is that if there is a difference between the observations with the missing data and those that are more complete, bias becomes introduced into the analysis because throwing out data makes it likely that the model will be unrepresentative of the population.

Available-case Analysis

This approach allows the analysis of subsets of the complete dataset so that multiple aspects of a problem can be studied. For example, suppose you have a dataset with software data with over twenty factors available for your exploration. All the software systems provided total annual costs but only 55% of the systems reported the number of software changes implemented in the calendar year. You can summarize the distribution of total costs over all the systems and conduct another analysis on the distribution of the number of software changes for the records with responses. This will involve the creation of multiple datasets that are targeting the complete data and only include a subset of the available factors in the initial dataset. Again, with this method, bias is introduced if the systems with nonresponse differ significantly from those with responses.

Imputation Methods

To retain as much of the precious gold (data) as possible, we should consider using imputation methods. There are several methods you can choose to make a best statistical inference at a response that will close a data gap. This allows the entire dataset to be used in analysis; however, different types of biases can occur that must be considered. We'll begin the discussion by introducing some

commonly known methods of imputing data that are very simple to execute. Our core focus of the discussion with be on two methods: Expectation Maximization (EM) and Multiple Imputation by Chained Equations (MICE).

Mean Imputation

One of the easiest ways to calculate the missing value of a variable is to compute the mean of the observed values of the variable. This method is called mean imputation (also referred to as mean substitution and unconditional mean imputation). While this approach restricts the variability of the data, it weakens the magnitude of covariances and correlations. It also allows for values that are implausible or simply not possible for a system given other characteristics. An example of this implausibility is in an engine dataset, there are missing values for Number of Cylinders. The average of the existing data is seven cylinders, which is not possible.

Imputing using Related Observations

Data holes can be filled with responses from related observations. Suppose that we have a dataset, and we are missing labor rates for two software systems that are being maintained by the same company as six other systems within the dataset. We could fill in the labor rates for the two software systems with the same rates as the other representative data points. The issue with this method is that we are introducing measurement error.

Regression Imputation

This form of imputation replaces missing values with a predicted value based on the results of fitting a regression line to available data. This concept is the basis for one of the main methods we will discuss in this paper, MICE. The specific form of regression imputation is called stochastic regression. Rather than use the mean of the regression to fill in the missing data, stochastic regression imputation recognizes the uncertainty of the regression analysis and adds to the regression analysis a random draw from the residual distribution. The advantage of stochastic regression is that it is the only method that produces unbiased estimates under data that are Missing At Random (MAR) (Enders, page 46).

Multiple Imputation by Chained Equations

One approach to imputation is emerging as a preferred method due to its ability to create multiple imputations for a missing value that accounts for the statistical uncertainty in the imputation. This method is MICE.

MICE operates under the assumption that the missing data is MAR. MAR occurs when a data gap is fully accounted for by the variables where there is complete information. This means that there might be systematic differences between the

missing and observed values, but these can be entirely explained by another observed variables. An example of MAR is the following:

> Salary data is often left unanswered by respondents who have not completed a college degree. In this instance, a missing response to this question is deemed MAR, but there may be observable differences between salary data provided by respondents who graduated from college and those who did not.

MAR should not be confused with missing completely at random. This means that the missing data has nothing to do with any other factors in the data. A question could have been accidentally overlooked or the questionnaire could be lost in the mail.

With MICE, multiple regression models are conducted and each variable with missing data is modeled conditionally on the responses of the other variables within the dataset. With this method, each variable is modeled according to its own distribution.

How MICE Fills the Gaps

An analyst is provided a dataset with engine data. The dataset has seven features, with four of those seven features missing some data. Let Y_j with (j=1,…,p) be one of p incomplete variables, where $Y=(Y_1,\ldots,Y_p)$. Y_{obs} is the incomplete dataset. Q represents a model that is built on the engine dataset. To estimate Q without making unrealistic assumptions about the missing data, the gaps should be filled. Multiple imputation is a general framework that several imputed versions of the data by replacing the missing values by plausible data values. The values are drawn from a distribution that is modeled for each missing entry.

Multiple imputation can be viewed as a series of stochastic regression imputations. The additional conceptual complication is a Bayesian framework. It begins with an imputation step (I-step). This is conducted using stochastic regression. This provides one imputed data set that fills in all the holes. The goal, however, is to have several imputed data sets. This is where the second step, called the posterior step (P-step), comes into play. In the P-step, the mean and covariance distributions are calculated from the filled-in data that were obtained from the I-step. The P-step then proceeds by taking a random draw from the mean and covariance distributions, which are used to calculate regression coefficients in the succeeding I-step.

The number of iterations, m, must be specified. Users typically choose 3-5 full iterations of imputation. There are several theories on determining the sufficient number of imputations. A solid rule-of-thumb is to use the average percentage

rate of missingness as the m-number of imputations. For example, if there is 30% missing data on average in a dataset, use 30 imputations. (Bodner, T. E. (2008))

Because of the uncertainty in the missing data, it is useful to have multiple imputations to provide different views of the values that the missing data points are likely to be. Given the multiple imputations, the coefficients of the individual equation are averaged (using a simple, unweighted mean). The other parameters, including the degrees of freedom, standard errors, and R^2s are combined using what is known as *Rubin's Rules*, after the statistician who developed them. See Figure 2for an illustration of this process.

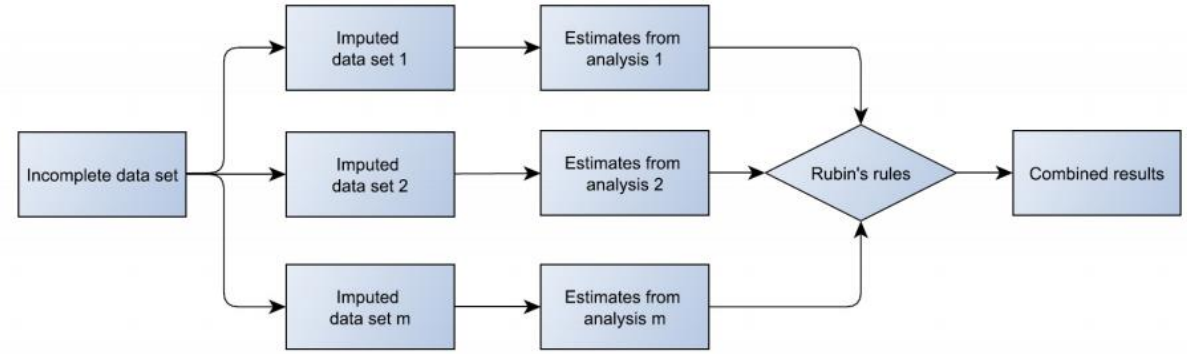

Figure 2. Illustrating the Multiple Imputation Process

Implementing MICE

We will be using R to demonstrate an example on how to implement MICE. A first step when you are working with new data with missing values is to determine how prevalent the gaps are within the dataset. Table 1displays the variables included within the data used for analysis.

Engine Data – Included Features	
Brake Horsepower (bHP)	Engine Speed (EngSP)
Cylinders (CYL)	Dry Weight (DryWGT)
Peak Torque (PkTor)	Displacement (DISP)
Unit Cost in Dollars (UC)	

Table 1: Data Features

Using the ***mice*** package in R, the ***md.pattern()*** function can provide an output of the number of complete and incomplete records across all variables.

```
install.packages('mice')

library(mice)
```

```
data<-read.csv("Example.csv")

md.pattern(data)
```

Using the dataset, which includes engine data from Army programs, and the md.pattern function, we produce Figure 8.

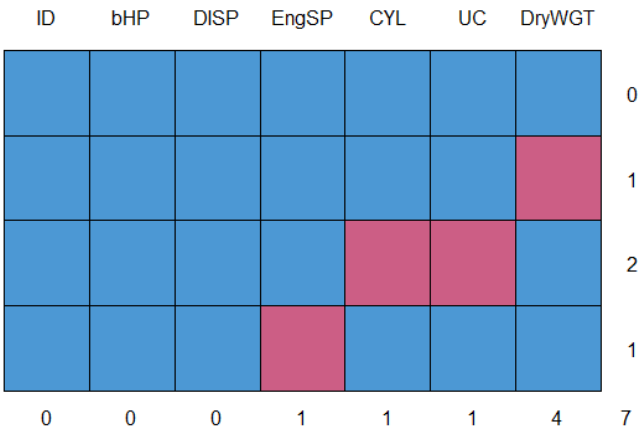

Figure 3: Missing Pattern Count Chart

This graph provides counts of the missing values in the dataset. The first three variables have no missing values, but the remaining variables do have gaps. Dry Weight has the most missing values at four.

Next, we will use the ***mice()*** function to conduct imputation. We will set out imputation iterations to five.

```
imputdata <- mice(data, m=5,maxit=50,meth='pmm',seed=500)

summary(imputData)
```

Setting *m=5* will produce five iterations. The function *meth* refers to the imputation method. The chosen method is the predictive mean matching (pmm) method. This method is a semi-parametric imputation approach. It fills in missing values with an observed value from a data point whose predicted values are closest to the predicted value of the missing data. One of the advantages of using pmm is that it considers the empirical characteristics of the underlying data set, such as skew. There are multiple methods that can be selected that are detailed in the R documentation for the MICE package. The *pmm* method will be the chosen method for this example. (The pmm method is appropriate for numerical variables that take on a continuous range of values (non-categorial data). For categorical data, appropriate methods are logistic regression ('logreg') for two-level factors

and multinomial logistic regression ('polyreg') for categorical variables that can take on more than two values.)

The output provides information on the number of imputations, the methods used to fill gaps, the number of logged events (iterations), and other information on the imputation run.

To see the completed dataset, use the ***complete()*** function.

```
completedData <- complete(imputdata,1)
```

The missing values that were present initially have been filled. The number one in the ***complete*** function indicates that you want to see the first iteration. To see the other 2-5 datasets, you will need to write functions to create and view those datasets.

To identify any relationships between variables in our data, we should fit models to each of the imputed datasets and then we can pool the results together and determine how this helps to improve the results.

To fit a linear model to a dataset, use the ***lm()*** function. Then, pool the *m* estimates $\hat{Q}^{(1)}, \ldots, \hat{Q}^{(m)}$ into one model $\bar{Q}$. Fit1 <- with(imputedata, lm(Y~ X+U+Z))

```
summary(pool(Fit1))
```

This produces one model using all the results of the imputation iterations. For this example, the model Unit Cost (UC) = Maximum Brake Horsepower (bHP) is used to explore a linear relationship.

Analyzing Imputed Results - MICE

We now have our imputed dataset free of holes and it is ready for analysis. First, it is often helpful to create scatter plots to compare the original and imputed data. An indicator that the imputed datasets are on the right track is if the shape of the imputed values matches the shape of the observed values.

From the scatterplot in Figure 4, we can see there is somewhat of a linear relationship between UC and bHP. The pattern of the relationship seems plausible for the imputed values (pink) as compared to the observed values (blue).

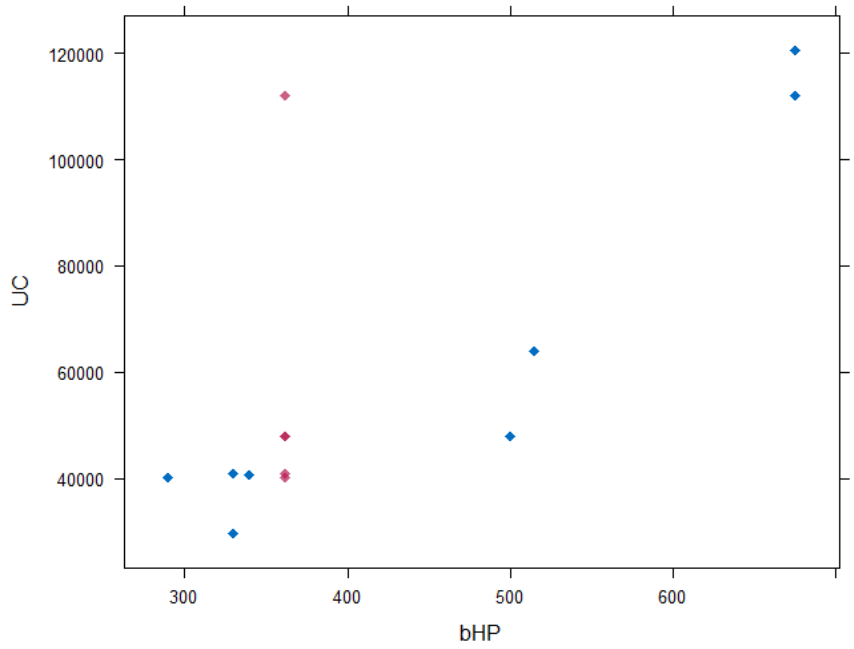

Figure 4: Scatterplot for UC and bHP - MICE

To continue further, viewing the density plot will also allow analysts to determine how the shape of the feature compares for the observed and imputed values. We can see by the density plot in Figure 5 for Dry Weight that the imputations follow a similar shape, but a couple of the imputation iterations have higher peaks than the observed values. This function works on all variables with two or more missing values.

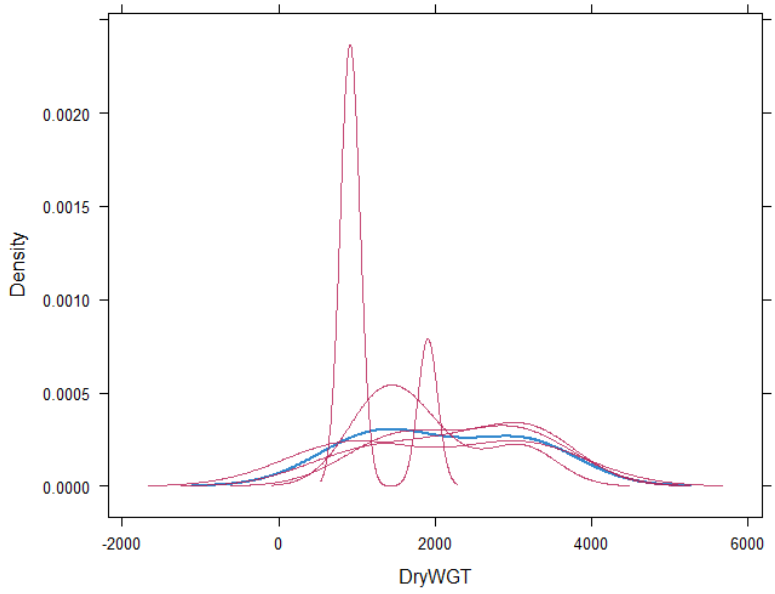

Figure 5: Density Plot for Dry Weight - MICE

To determine how the linear relationship between the original dataset and the imputed dataset, let us first look at the results from a linear model being fit on the original dataset. Figure 6 presents the results on the dataset without imputation.

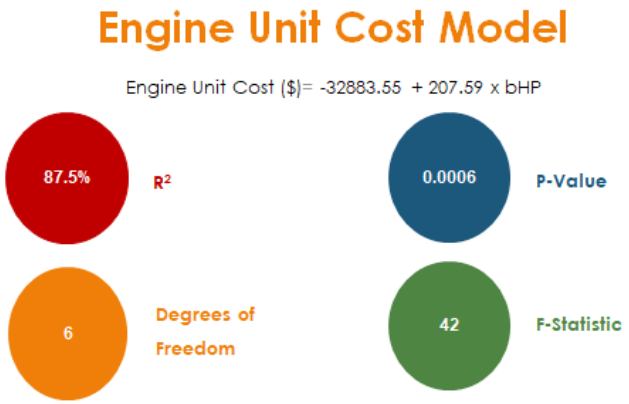

Figure 6. Output of Engine Unit Cost Model - No Imputation

The model is a solid one with a statistically significant p-value (less than alpha = 0.05) and an R^2 equal to 87.5%. One data point was removed due to missingness present for unit cost. The goal is to determine if the model can be improved using imputation.

Viewing the output from the ***summary(pool(Fit1))*** function, we can see the new function created by using imputation on the data. The code also produces the R^2 statistic.

```
summary(pool(Fit1))

pool.r.squared(Fit1, adjusted = FALSE)
```

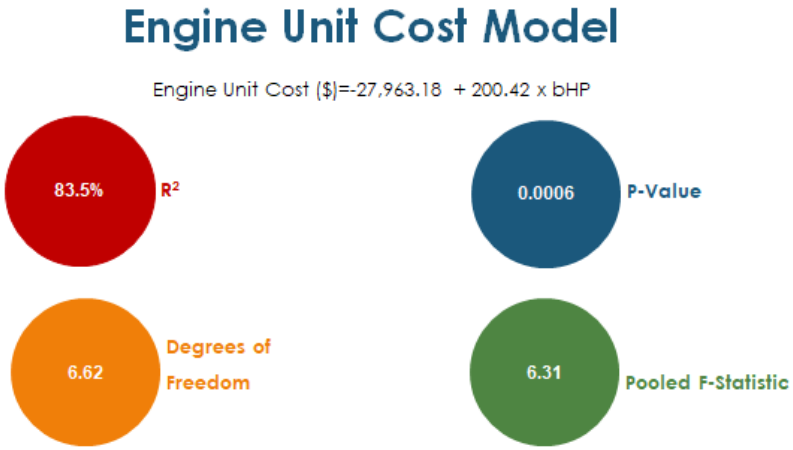

Figure 7. Output of Engine Unit Cost Model – Imputed using MICE

Though the R^2 statistic is lower than the original dataset, we gained some degrees of freedom with the use of imputation with the creation of this statistically significant model. The missing value in bHP was filled. The model does not gain a full degree of freedom since the iterations are pooled. We must also consider that with the removal of data points, we may be losing understanding in the true structure of the data; therefore, we are limiting the understanding of the relationships of features on the entire population.

Expectation Maximization

A method that is not much more complicated than the mean imputation is Expectation Maximization (EM). This maximum likelihood method can also be thought of as an optimization problem. The algorithm uses available data to impute a value for the missing variable, then it checks to determine if this value is most likely. If not, the algorithm imputes a more likely value. This iterative process continues until the most likely value is determined. It differs from other optimization algorithms in that it does not require computing first or second derivatives.

The way this method estimates missing values is that the covariation among variables is used to infer probable values for the missing data. Even partially complete cases are used in the estimation of missing values. Using these records helps to increase the precision of parameter estimates and decrease parameter estimate bias.

Expectation Maximation uses a two-step process to fill data gaps:

Step 1: Data gaps are filled using listwise deletion, pairwise deletion or some other method of imputation. Listwise deletion is the same as complete-case analysis, in that you remove any records with missing values. Likewise, pairwise deletion mimics available-case analysis. With this method, analyses are conducted on all records, but only for the variables with complete data. For example, if height is missing from observation two, this record will be excluded in analyses involving height but will be included in analyses involving the remaining features. Once the missing values are replaced with the conditional mean of the missing values, the initial covariance matrix is estimated.

To clarify further, consider an example in that we want to conduct a simple univariate linear regression on a data set where the *X* values are complete but some of the *Y* values are missing.

We can estimate the regression equation using all the data with expectation maximization.

In Step 1, the missing components of sums are replaced with their averages. In this step we have estimates for $\beta_0$and β_1 so in the sums $\sum Y$and $\sum XY$ we replace the missing values with $\hat{Y}_i$. We have that

$$Y_i = \hat{Y}_i + \varepsilon$$

The residual ε has mean 0 and variance equal to $\sigma^2_{Y|X}$, so

$$E(\varepsilon^2) = \sigma^2_{Y|X} + E(\varepsilon)^2 = \sigma^2_{Y|X}, \text{ so}$$

$$E(Y_i^2) = E(\hat{Y}_i^2) + \sigma^2_{Y|X}$$

So, we replace the missing components of Y^2 with this expected value.

$$\hat{\mu}_Y = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

$$\hat{\sigma}_Y^2 = \frac{1}{N}\left(\sum_{i=1}^{N} Y_i^2 - \frac{(\sum_{i=1}^{N} Y_i)^2}{N}\right)$$

$$\hat{\sigma}_{X,Y} = \frac{1}{N}\left(\sum_{i=1}^{N} X_i Y_i - \frac{\sum_{i=1}^{N} X_i \sum_{i=1}^{N} Y_i}{N}\right)$$

Step 2: The maximum likelihood estimates of the mean vector and covariance matrix are calculated. The covariance matrix is then used to

derive regression equations for the next iteration and the cycle begins again. This iterative process repeats until the difference between the covariance matrices in subsequent runs falls below some convergence criteria that is specified by the analyst.

Exploring the mathematics calculation for Step 2, consider:

$$\hat{\beta}_1 = \hat{\sigma}_{X,Y}$$

$$\hat{\beta}_0 = \hat{\mu}_Y - \hat{\beta}_1 \hat{\mu}_X$$

Noting that

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \varepsilon$$

it follows that

$$\hat{\sigma}_Y^2 = \hat{\beta}_1^2 \hat{\sigma}_X^2 + \hat{\sigma}_\varepsilon^2$$

The variance of the residuals is the conditional variance of Y | X, so we write

$$\hat{\sigma}_{Y|X}^2 = \hat{\sigma}_Y^2 - \hat{\beta}_1^2 \hat{\sigma}_X^2$$

Another basic fact that we use in Step 2 is

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \text{ for all } i = 1, \ldots N.$$

We iterate between Steps 1 and 2 until the regression coefficients converge (i.e., stop changing within a specified tolerance such as 1x10^-5).

An advantage of this method over mean imputation is EM preserves the relationship with other variables, which is imperative when using linear regression to analyze your data. A disadvantage of this method is that it underestimates the standard error.

Implementing EM

EM can be implemented in R. The function **prelim.norm** if used on a matrix of the x (bHP) and y (cost) variables to sort rows according to the missingness patterns.

```
a<-prelim.norm(cbind(y,x)
```

Next, the function **em.norm** is applied to the matrix and performs maximimum-likelihood estimation using the EM algorithm. This function produces a vector which can then be used to return a list of parameters.

```
b<-em.norm(a)

c1<-getparam.norm(a ,b)
```

Next, the average of the imputations is calculated for the variable with missing values.

```
c1$mu[1]
```

The estimates for the coefficients of the model are then estimated.

```
b.est<-c(c1$mu[1]-
(c1$sigma[1,2]/c1$sigma[2,2])*c1$mu[2],c1$sigma[1,2]/c1$sigma[2,2])
```

Analyzing Imputed Results – EM

Now that the coefficients have been calculated, the models can be compared. The results produced by the EM algorithm are different from those obtained from MICE. EM produces one single model as opposed to multiple imputed datasets. Using the EM model, analysts can then calculate goodness-of-fit metrics.

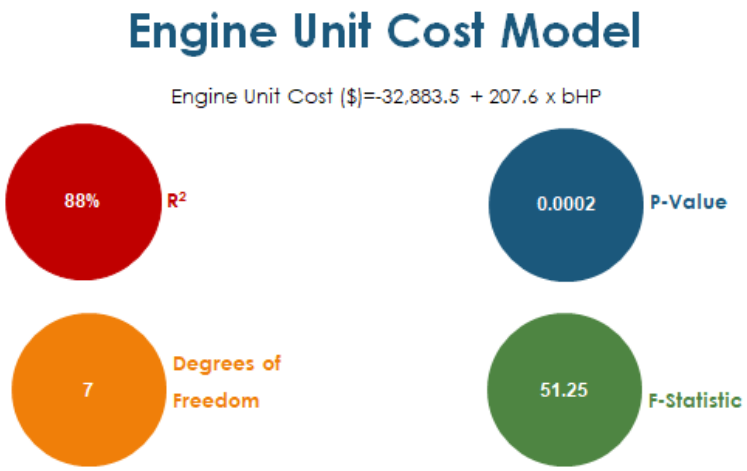

Figure 8. Output of Engine Unit Cost

The results presented in Figure 8 are based on using EM to fill in the missingness for Unit Cost. Compared to the results produced from removing the data points with missing values, this is a better performing model. A degree of freedom was gained and the R^2 metric increased while the model retained statistical significance.

MICE vs. EM

MICE and EM are based on similar assumptions and in practice they often produce similar results. The Bayesian estimation in MICE is asymptotically equivalent to the maximum likelihood estimates in EM (Enders 2010), so for large data sets the two methods should provide similar results. For small data sets, it is wise to run both and compare the results, as small differences in the methods

could have an outsized impact when the number of data points is limited. Many of the differences between the two come into play when looking at advanced applications such as structural equation modeling and interaction effects, which is beyond the scope of this paper. The interested reader should consult *Applied Missing Data Analysis* (Enders 2010) for more information.

Figure 9 plots the iterations together to provide a visual of how the algorithms performed in relation to each other. The fifth MICE imputation is an outlier compared to the other Unit Cost estimates, which causes the pooled model to be a little less strong than the EM imputation.

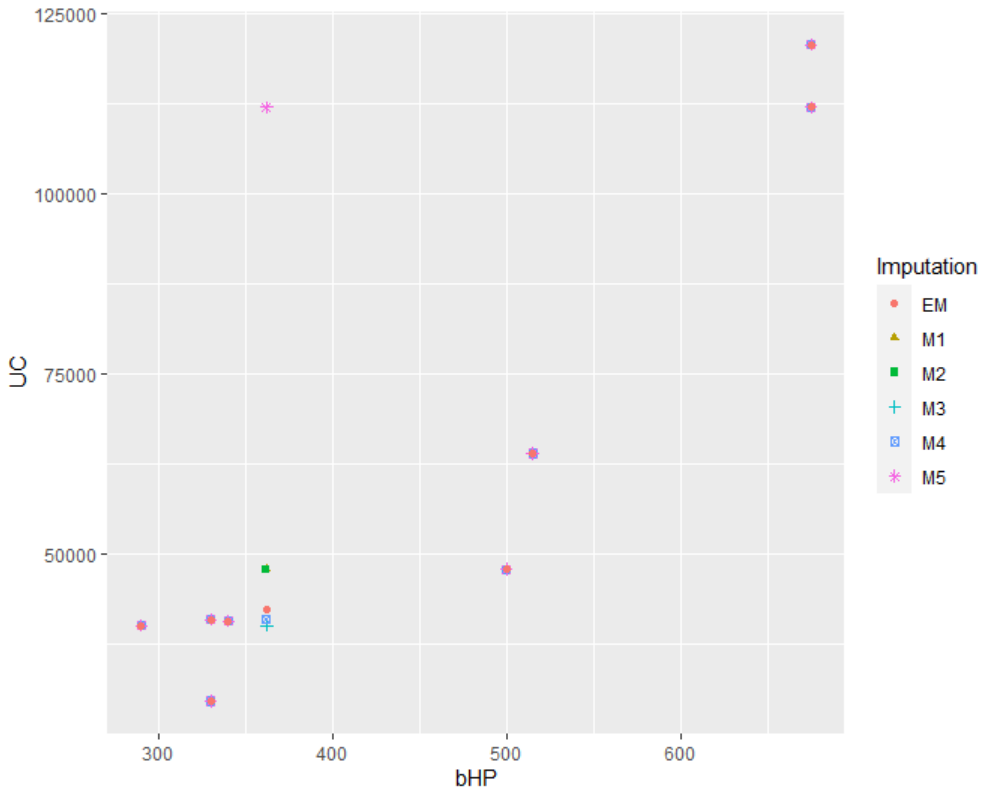

Figure 9. Imputation Iteration Plot

Conclusions

Imputation is a useful method allowing for the replacement of missing values within datasets instead of the removal of incomplete observations. This method helps to protect the degrees of freedom of the dataset and preserve valuable data points that can provide informative insight into future costs.

There are multiple methods which can be used to impute data. Two of the strongest techniques, MICE and EM, should be considered first as they preserve relationships between independent and dependent variables and estimate error more accurately.

The MICE method for imputation has an edge over EM since MICE calculates multiple imputations for the missing values instead of one single estimate. Having these multiple estimations for the missing values considers the uncertainty in the imputations, which increases the accuracy of standard errors.

Technical Note: Linear Vs. Nonlinear Regression

As we have discussed, the imputation methods for numerical variables relies heavily on linear relationships between the variables and the predictors. If you are applying imputation to a dataset in which you believe the relationships are nonlinear, the data should be transformed before applying imputation. For example, the power equation $Y = aX^b$ can be linearized by applying a log transformation.

Technical Appendix: Pooling Goodness-of-Fit Statistics from Multiple Imputations

As mentioned earlier, to pool the results of regression analysis across multiple imputations, the regression analysis is applied to each of the imputed data sets. The coefficients are then combined by taking a simple, unweighted mean. Pooling the goodness-of-fit statistics is more complicated. The general approach is referred to as *Rubin's Rules*, after one of the pioneers who developed much of the statistical theory of imputation, including how to pool goodness-of-fit statistics from multiple imputation. Different methods are required for each metric.

1. R^2 and Correlation Coefficients

   To pool correlations, a Fisher's z-transformation is required. Let r_j denote the correlation from a single imputed data set. Fisher's z-transformation is

   $$z_j = \frac{1}{2} ln\left(\frac{1 + r_j}{1 - r_j}\right)$$

   The arithmetic mean of the individual z_j values is then calculated, denote this by $\bar{z}$. This is then transformed back (inverse z transformation) to yield the pooled correlation:

   $$\bar{r} = \frac{e^{2\bar{z}} - 1}{e^{2\bar{z}} + 1}$$

   To apply to R^2, first take the square root to obtain a correlation.

2. Standard Errors

   To pool standard errors, first calculate the average variance across the imputations (referred to as *within variance*):

   $$V_1 = \frac{\sum_{i=1}^{m} SE_i^2}{m}$$

However, there is another source of uncertainty – the imputed values; this needs to be accounted for as well (referred to as *between variance*):

$$V_2 = \frac{1}{m-1}\sum_{i=1}^{m}(SE_i - \overline{SE})^2$$

This uncertainty is compounded by the fact that there is only a finite number of imputations (*m*), so the total variance is:

$$V_T = V_1 + \left(1 + \frac{1}{m}\right)V_2$$

3. Degrees of Freedom

   The pooled degrees of freedom are based on the within (V1) and between (V2) variances we provided in the discussion on pooling standard errors

$$\lambda = \frac{\left(1 + \frac{1}{m}\right)V_2}{V_1 + \left(1 + \frac{1}{m}\right)V_2}$$

$$df = \frac{n-k+1}{n-k+3}(n-k)(1-\lambda)$$

   where *n* is the number of data points (included imputed data points) and *k* is the number of parameters/independent variables in the regression.

4. F-statistics

   To pool F-statistics, first calculate the average F-statistic across the imputations:

$$\bar{F} = \frac{\sum_{i=1}^{m} F_i}{m}$$

   Next, calculate the average relative increase in variance:

$$ARIV = (1 + \frac{1}{m})\left[\frac{1}{m-1}\sum_{i=1}^{m}\left(\sqrt{F_i} - \overline{\sqrt{F}}\right)^2\right]$$

where $\overline{\sqrt{F}}$ is the average of the square roots of the F-statistics for each imputation.

The Pooled F-statistic is then calculated as

$$F = \frac{\frac{\bar{F}}{k} - \frac{m+1}{m-1}ARIV}{1 + ARIV}$$

The numerator degrees of freedom is k (number of parameters/independent variables) and the denominator degrees of freedom is

$$v = k^{-\frac{3}{m}}(m-1)\left(1 + \frac{1}{ARIV}\right)^2$$

5. Pooling Likelihood Ratios

F-statistics apply to linear regression; for nonlinear regression need another measure, such as the likelihood ratio:

$$LR = -2(logL_{Null} - logL_{Model})$$

where $logL^{Null}$ is the log likelihood of the model with no parameters except a constant (mean) and $logL_{Model}$ is the regression model.

First, the average of the likelihood ratios is calculated across the imputed data sets:

$$\overline{LR} = \frac{\sum_{i=1}^{m} LR_i}{m}$$

Second, the log likelihood ratio of the pooled model is recalculated across all *m* imputed data sets and then average to yield

$$\overline{LR}_{Constrained}$$

Next, calculate the average relative increase in variance

$$ARIV = (\frac{m+1}{k(m-1)})[\overline{LR} - \overline{LR}_{Constrained}]$$

where *k* is the number of parameters/independent variables.

A pooled F statistic is then calculated as

$$F = \frac{\overline{LR}_{Constrained}}{k(1 + ARIV)}$$

where the numerator degrees of freedom is k (number of parameters/independent variables) and the denominator degrees of freedom is

$$v = 4 + (k(m-1) - 4)\left[1 + \left(1 - \frac{2}{k(m-1)}\right)\frac{1}{ARIV}\right]^2 \text{ if } k(m-1) \geq 5$$

and

$$v = \frac{k(m-1)(1+1/k)\left(1+\frac{1}{ARIV}\right)^2}{2} \text{ if } k(m-1) \leq 4$$

References

Arbuckle, J. L. (1996). "Full information estimation in the presence of incomplete data." In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 243–277). Mahwah, NJ: Erlbaum.

Azen, S. P., M. Van Guilder, and M.A. Hill (1989). "Estimation of parameters and missing values under a regression model with non-normally distribution and non-randomly incomplete data." Statistics in Medicine, 8, 217–228.

Azur, M., E. Stuart, C. Frangakis, and P. Leaf (2011). "Multiple Imputation by Chained Equations: What is it and how does it work?", *International Journal of Methods in Psychiatric Research*, 20(1): 40–49.

Bodner, T. E. (2008). "What Improves with Increased Missing Data Imputations?" *Structural Equation Modeling*, Volume 15, Issue 4, 651-675.

Brown, R. L. (1994). "Efficacy of the indirect approach for estimating structural equation models with missing data: A comparison of five methods." Structural Equation Modeling: A Multidisciplinary Journal, 1, 287–316.

Enders, C. K. (2001). "The impact of nonnormality on full information maximum likelihood estimation for structural equation models with missing data." Psychological Methods, 6, 352–370.

Enders, C. K., & Bandalos, D. L. (2001). "The relative performance of full information maximum likelihood estimation for missing data in structural equation models." Structural Equation Modeling: A Multidisciplinary Journal, 8, 430–457.

Enders, C. K. (2010) *Applied Missing Data Analysis*. New York, New York: Guilford Press.

Haitovsky, Y. (1968). "Missing data in regression analysis." Journal of the Royal Statistical Society, Series B, 30, 67–82.

Kim, J., and J. Curry (1977). "The treatment of missing data in multivariate analyses." Sociological Methods and Research, 6, 215–240.

Kromrey, J. D., and C.V. Hines (1994). "Nonrandomly missing data in multiple regression: An empirical comparison of common missing-data treatments." Educational and Psychological Measurement, 54, 573–593.

Peters, C., L. Okleshen, and C. Enders, (2002). "A primer for the estimation of structural equation models in the presence of missing data: Maximum likelihood

algorithms," *Journal of Targeting, Measurement and Analysis for Marketing*, 11, 81–95.

Van Buuren,S. and K. Groothuis-Oudshoorn, (2011). "mice: Multivariate Imputation by Chained Equations in R", *Journal of Statistical Software*, 45:3, 1-66.

Van Buuren, S., MICE: Multiple Imputation via Chained Equations, RDocumentation, https://www.rdocumentation.org/packages/mice/versions/3.6.0/topics/mice, retrieved February 2, 2021.

Wothke, W. (2000). Longitudinal and multigroup modeling with missing data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), "Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples" (pp. 219–240). Mahwah, NJ: Erlbaum.