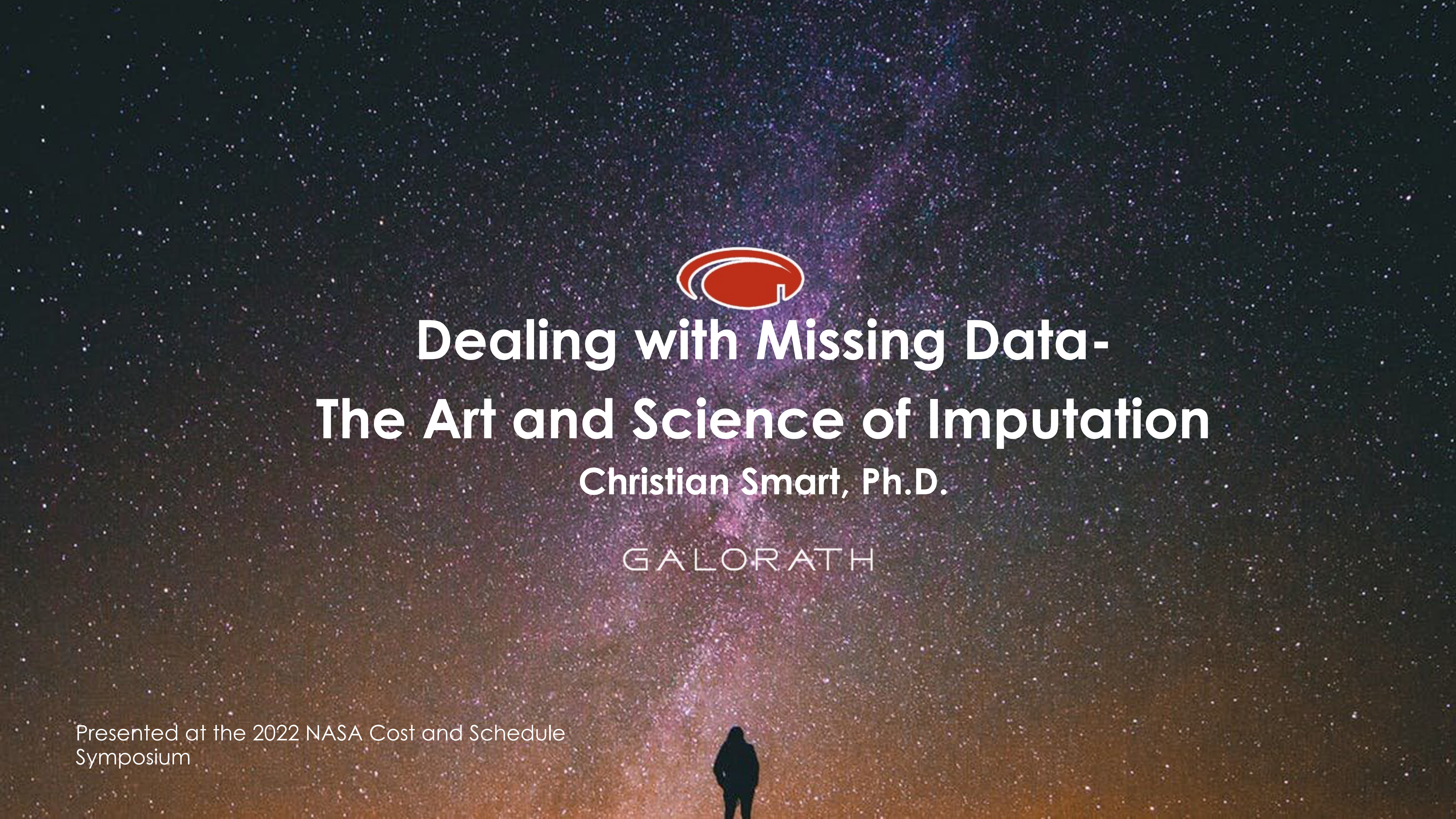


Dealing with Missing Data-
The Art and Science of Imputation

Christian Smart, Ph.D.

GALORATH

Presented at the 2022 NASA Cost and Schedule Symposium

IMPUTATION

FILLING IN HOLES IN DATASETS

THE PROBLEM OF MISSING DATA

A significant problem, especially for small datasets

Often dealt with by removing observations with missing data

FILLING IN HOLES WITH STATISTICS

A variety of techniques exist for filling in missing data, though some perform better than others

FILLING IN HOLES WITH STATISTICS

Recognizing the inherent uncertainty in missing data, we adopt and advocate the method of multiple imputation using Bayesian methods (“chained equations”)

DATA

Foundation of All Analyses

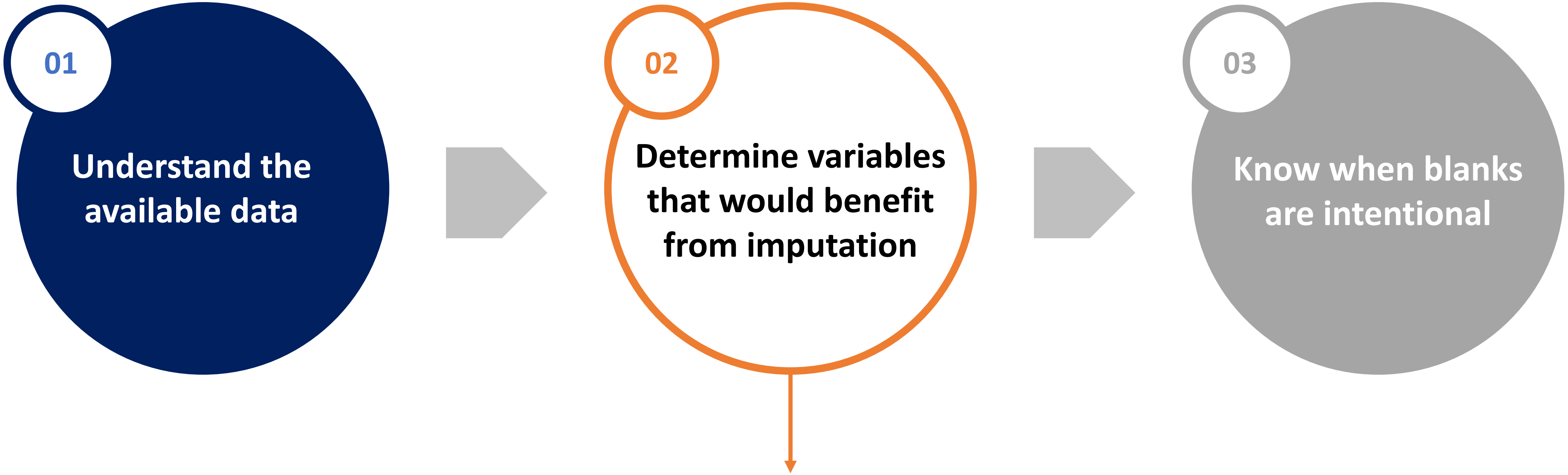
How Should We Handle It?

The bulk of the time in analytics should be spent on collecting, normalizing and verifying data. In defense and aerospace applications, datasets are small. Data should be preserved when possible

> The goal is to turn data into information, and information into insight.
>
> -Carly Fiorina

IMPUTATION

To impute or not to impute, that is the question



01

Understand the available data

02

Determine variables that would benefit from imputation

03

Know when blanks are intentional

Imputation is a powerful method that is useful for filling blanks when they are missing within a dataset

An analyst must understand the data intimately to know if a blank means that the factor is not applicable for that data point

Sometimes a blank does not reflect a nonresponse and should be observed “as is”

Is the response missing at random?

The US Census Bureau deals with missing data all the time. If no response is provided for the name of Person 7 on the Census form from the household of six members, this missing value is not an omission; the response is "Not Applicable"

ISSUES WITH DATA GAPS

What can go wrong?

Fewer Degrees of Freedom

Removing observations with missing values results in fewer degrees of freedom in models

Reduction of Predictive Power

Predictive power is diminished when degrees of freedom are small

Inability to Use Advanced Methods

Certain Machine Learning methods cannot be applied when missing values are prevalent

METHODS ALLOWING MISSING DATA

Complete-Case Analysis

Approach that excludes any records with missing data. Disadvantage – bias becomes introduced into the analysis due to the removal of data that may provide insight into the population

Available-Case Analysis

Approach allows the analysis of subsets of the complete dataset so that multiple aspects of a problem can be studied. Disadvantage – bias is again introduced if data are missing in a pattern

Alternative to Allowing Missingness

Though methods exist to continue with analysis upon removal of missing data, better alternatives exist for filling data gaps

IMPUTATION METHODS

Mean Imputation

Filling missing values with the mean of the observed values

Imputing using Related Observations

Filling missing values with responses from related observations

Regression Imputation

Replacing missing values with a predicted value based on the results of fitting a regression line to the available data

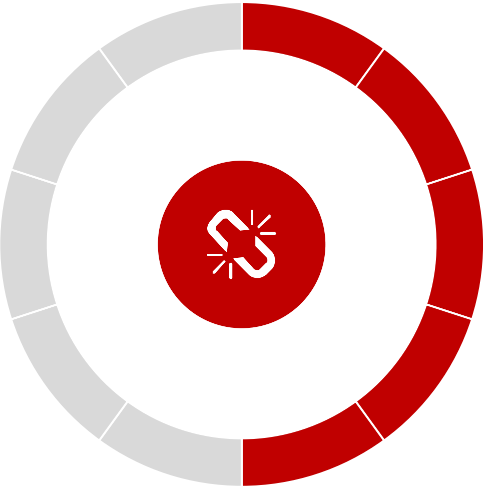
Expectation Maximization

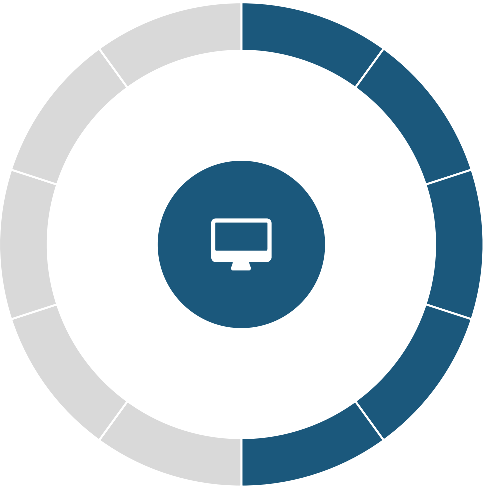
Replacing missing values by exploring the covariation among variables in order to infer values for the missing data

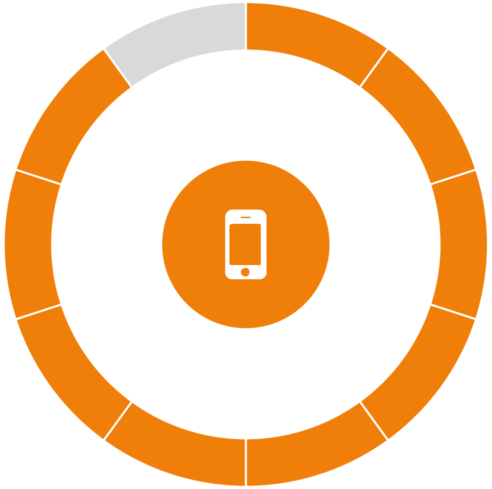
To retain as much of the precious gold (data) as possible, we should consider using imputation methods. There are several methods you can choose to make a best statistical inference at a response that will close a data gap

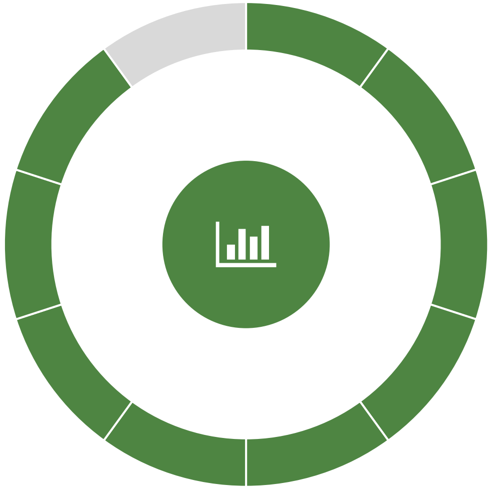
IMPUTATION METHODS

How do they compare?









Mean Imputation

This method helps to restrict the variablility of the data

Disadvantage: it weakens covariances and correlations amount features

Related Observations

This method also helps to restrict variability in the data

Disadvantage: Introduces measurement error

Regression Imputation

This method uses regression to predict missing values

Advantage: Produces unbiased estimates with data that are Missing At Random (MAR)

Expectation Maximization

This method uses maximum likelihood method to estimate missing values

Advantage: Increases precision and decreases parameter bias

MICE

MULTIPLE IMPUTATION BY CHAINED EQUATIONS

MICE

Method

This method creates multiple imputations for a missing value that accounts for the statistical uncertainty in the imputation

Assumptions

This method operates under the assumption that the missing data is MAR. MAR occurs when a data gap is full accounted for by variables where there is complete information

Iterations

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 GAPS

Several imputed versions of the data are created using plausible data values

01

NUMBER #01

Multiple imputation is a series of stochastic regression imputations

02

NUMBER #02

The first step is an imputation step (I-step) that fills data gaps using stochastic regression

03

NUMBER #03

The number of iterations, m, are specified for the number of imputations that are conducted in the I-step

04

NUMBER #04

In posterior step (P-step), the mean and covariance distributions are calculated from the filled-in data

05

NUMBER #05

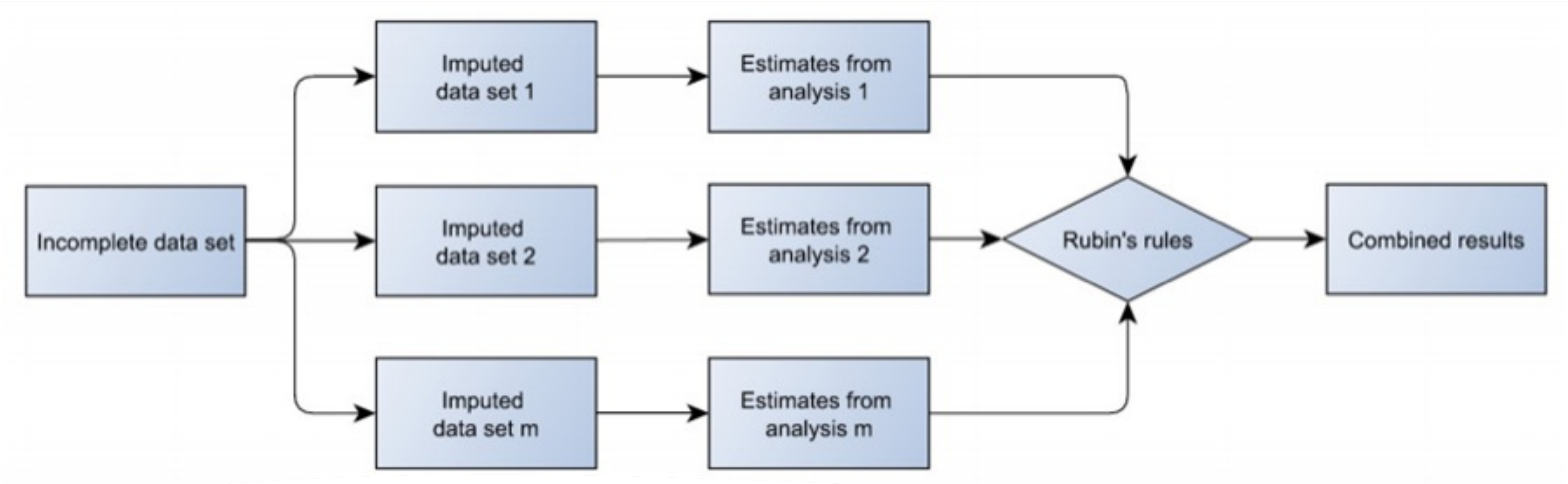
The P-step proceeds by taking a random draw from the mean and covariance distributions, which are used to calculate regression coefficients

06

NUMBER #06

The coefficients of the individual equation are averaged using a simple, unweighted mean. Goodness-of-fit measures are calculated using the pooled results

THE MICE PROCESS



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

UNDERSTANDING THE DATA

Exploring engine data

Dataset

The data used for analysis is a Wheeled and Tracked Vehicle Engine dataset. The dataset is small, which makes the use of imputation very important

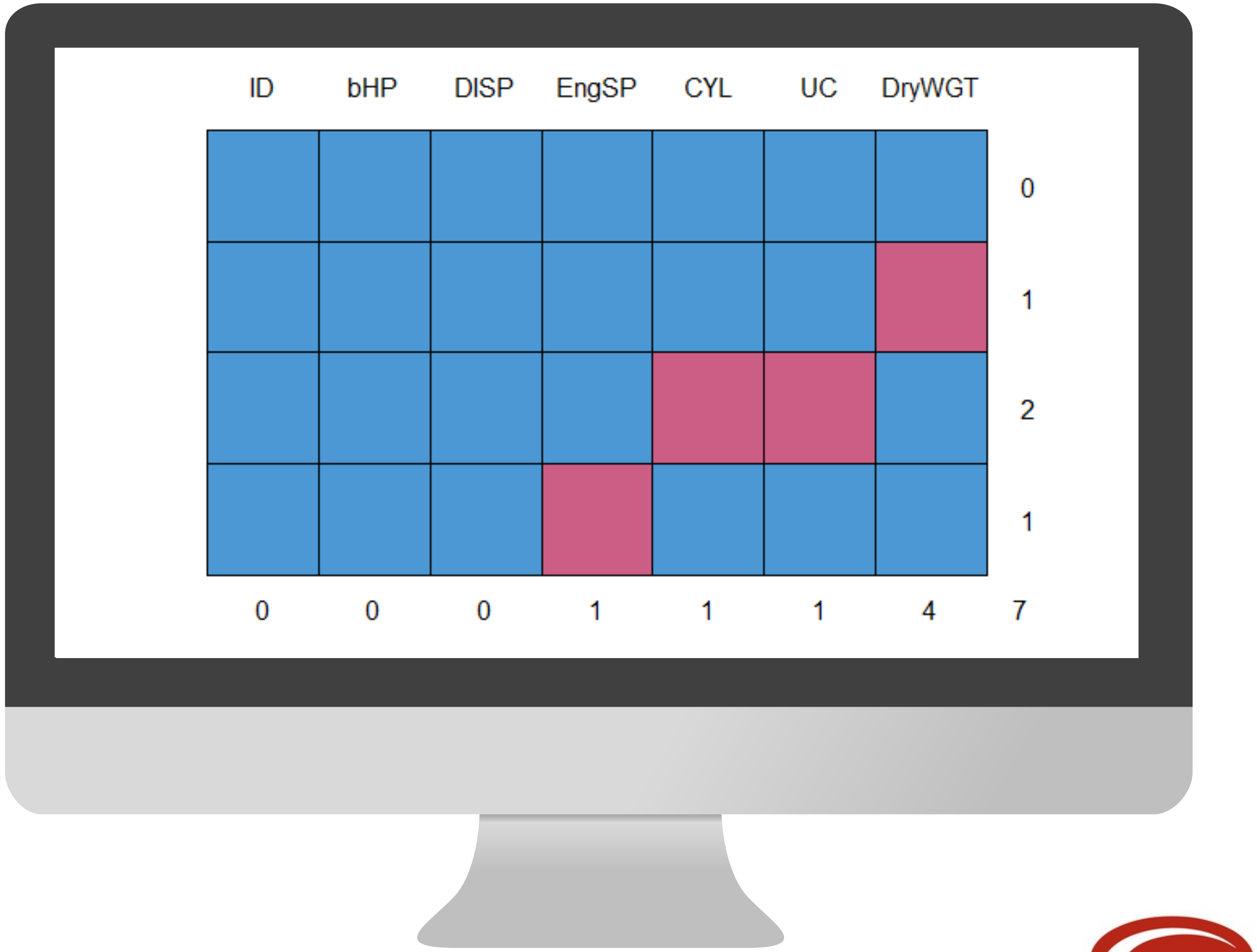
Included Features

Identification (ID), Brake Horsepower (bHP), Displacement (DISP), Engine Speed (EngSP), Cylinders (CYL), Unit Cost in Dollars (UC), Dry Weight (DryWGT)

TITLE GOES HERE

Of the seven features included in the dataset, four of those seven have missing values.

N=9



ID	bHP	DISP	EngSP	CYL	UC	DryWGT	
🟦	🟦	🟦	🟦	🟦	🟦	🟦	0
🟦	🟦	🟦	🟦	🟦	🟦	🟥	1
🟦	🟦	🟦	🟦	🟥	🟥	🟦	2
🟦	🟦	🟦	🟥	🟦	🟦	🟦	1
0	0	0	1	1	1	4	7

IMPLEMENTING MICE

01

We used the statistical programming platform R and the ‘mice’ package to calculate imputed data

R code:
install.packages('mice’)
library(mice)
data<-read(“Example.csv”)
imputdata<-mice(data, m=5, meth=‘pmm’, seed=23109)

Fixed seed to ensure the analysis is repeatable

The default in mice is m=5. This parameter will need to be included if another value of imputations is desired

02

Conduct linear regression on each of the five imputed datasets

To view each of the imputed datasets, we use the complete() function:

completedData<-complete(imputedata,1)

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

03

Pooling Results

Combining the results of these separate analyses is referred to as pooling

The pooled regression equation has coefficients that are the arithmetic means of the coefficients for the five individual regressions

Let *m* denote the number of imputed datasets, β_i denote the *i*th coefficient, and β_{ij} denote the *i*th coefficient for the *j*th imputed dataset; then:

$$\beta_i = \frac{\sum_{j=1}^{m} \beta_{ij}}{m}$$

IMPLEMENTING MICE

04

Pooling Results - 2

To fit a linear model to a dataset, use the ***lm()*** function. Then, pool the *m* estimates $\hat{Q}^{(1)}, \dots, \hat{Q}^{(m)}$ into one model $\bar{Q}$.

R code:

```
Fit1<-with(imputedata,lm(UC~bHP))
Summary(pool(Fit1))
```

05

Goodness-of-Fit Statistics

Unlike the coefficients, you cannot simply average the R^2 values, standard errors, the F-stats, etc., in order to calculate the goodness-of-fit statistics

```
pool.r.squared(fit4, adjusted = FALSE)

poolF<-mi.anova(mi.res=imputedata,
formula="UC~bHP")
```

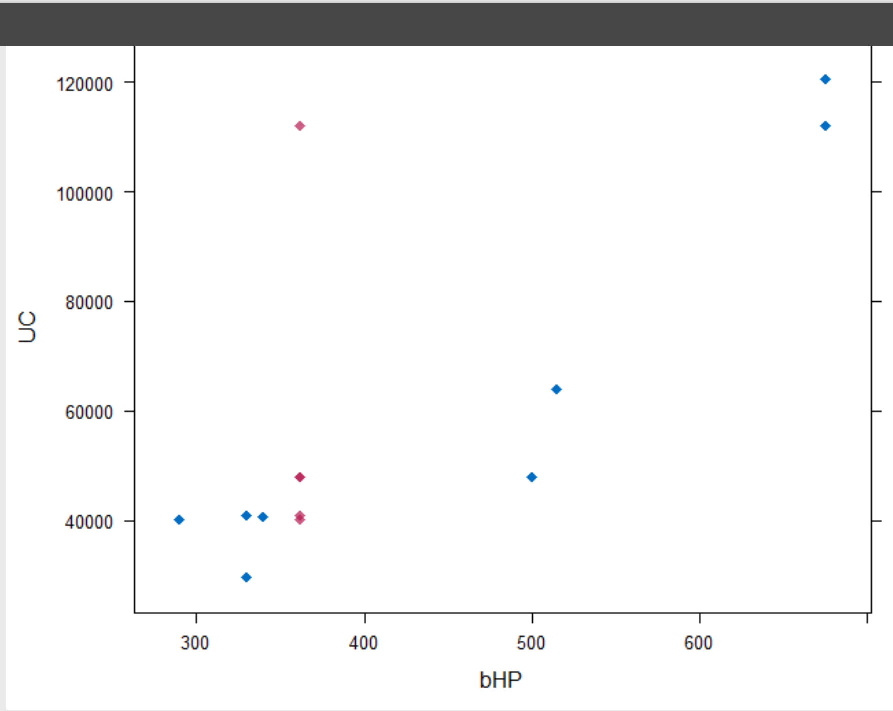
06

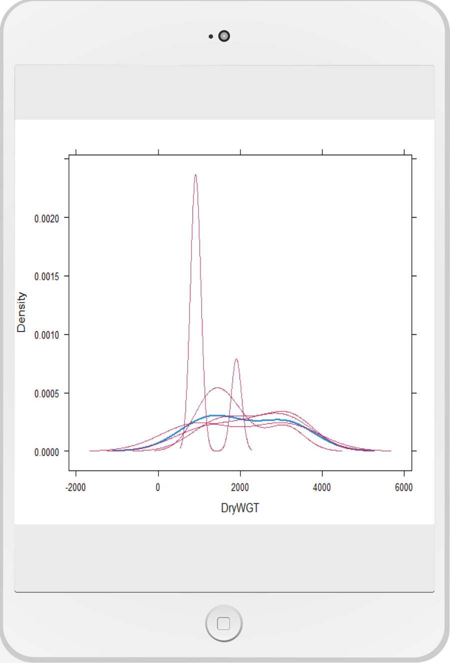
Compare Results

Compare the results from the imputed dataset to the original dataset with missing values removed

ANALYZING RESULTS

Creating plots to determine reasonableness of imputations





Scatterplot Analysis

There is 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)

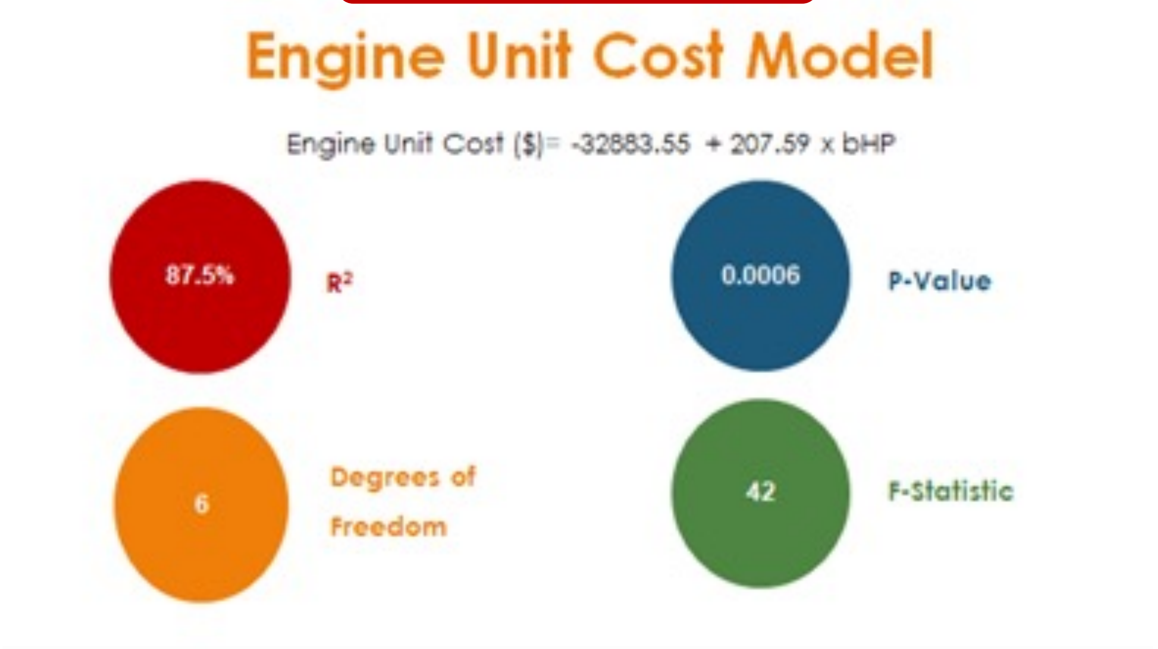
Density Plot Analysis

Density plots provide a visual into the shapes of each imputation. The plot is useful to determine outlier imputations and works for variables with two or more missing values

FIT RESULTS

Comparing results from the original dataset to the imputed (pooled) dataset

Linear Model



Engine Unit Cost Model

Engine Unit Cost ($)= -32883.55 + 207.59 x bHP

87.5% R²

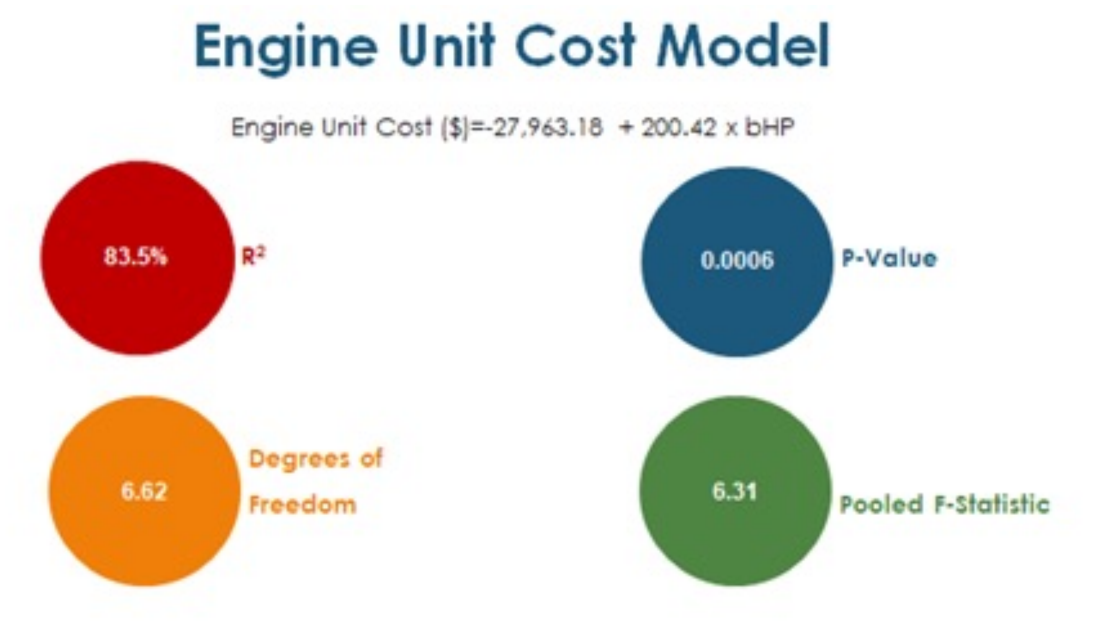
0.0006 P-Value

6 Degrees of Freedom

42 F-Statistic

The model is a solid one with a statistically significant p-value less than alpha = 0.05 and an R2 equal to 87.5%. One data point was removed due to missing a unit cost value

MICE Imputed Model



Engine Unit Cost Model

Engine Unit Cost ($)=-27,963.18 + 200.42 x bHP

83.5% R²

0.0006 P-Value

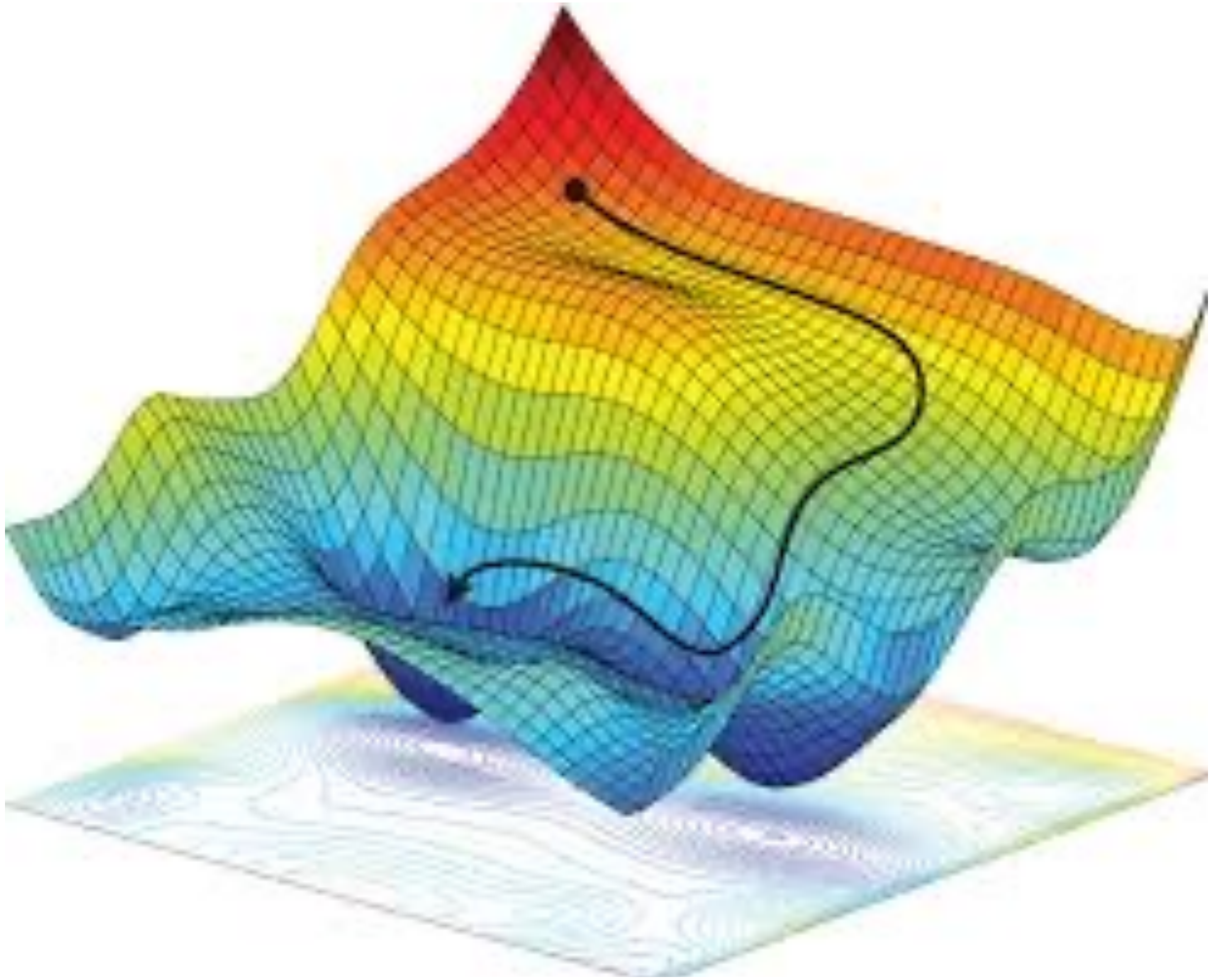
6.62 Degrees of Freedom

6.31 Pooled F-Statistic

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 model does not gain a full degree of freedom since the iterations are pooled

EXPECTATION MAXIMIZATION

Expectation Maximization



Imputing by optimizing

Maximum Likelihood

The maximum likelihood method is used to impute missing values. This method uses available data to impute a value and then checks to determine the reasonableness of the guess

Covariance

The covariation among variables is used to infer probable values for the missing data

Two-Step Process

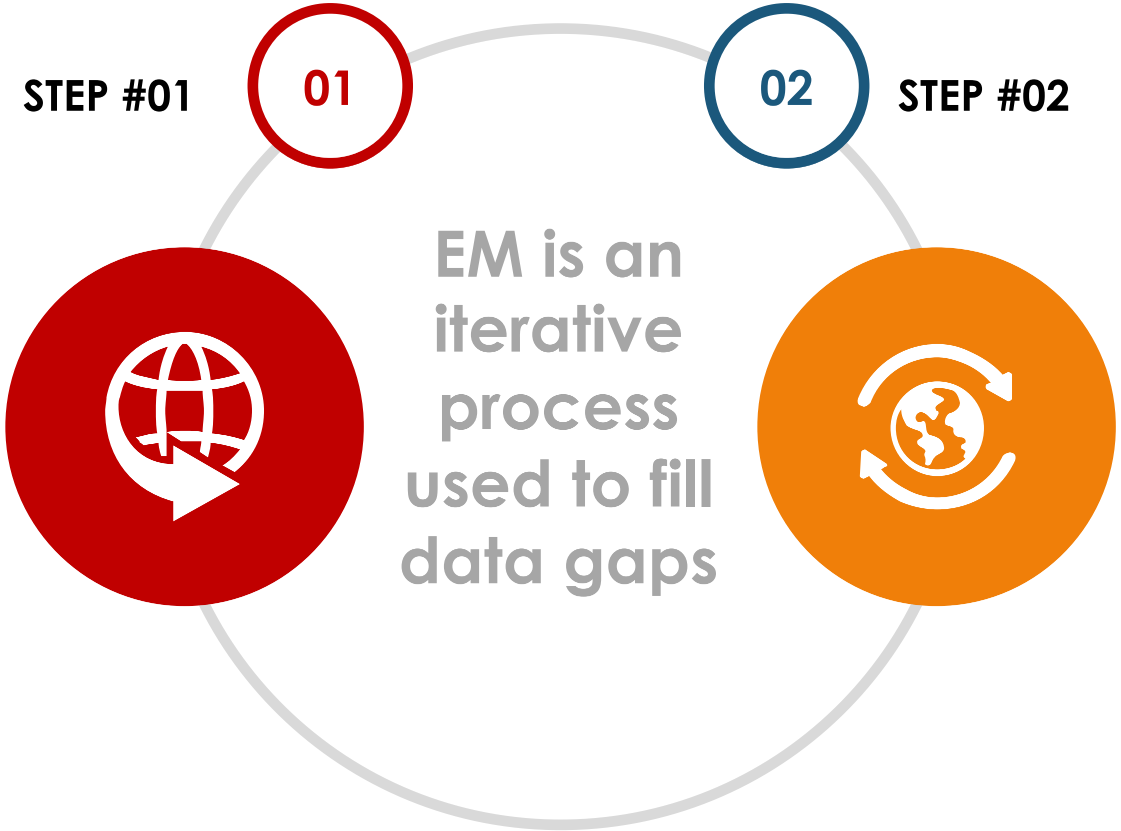
The method follows a two-step process to fill in missing data

EM TWO-STEP PROCESS

How EM fills data gaps

First Pass at Filling Gaps

The algorithm begins by filling the gaps with the conditional mean of the missing values.



STEP #01

01

02

STEP #02

EM is an iterative process used to fill data gaps

Iterative Process

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 continues until the difference between the covariance matrices in subsequent runs falls below the convergence criteria

IMPLEMENTING EM

01

Show missingness patterns

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 Fixed seed to ensure the analysis is repeatable

R code:

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

02

Performing maximum likelihood estimation using EM algorithm

R code:

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

This function produces a vector which can then be used to return a list of parameters

03

Pooling Results

The average of the imputations is calculated for the variable with missing values

R code:

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

The estimates for the coefficients of the model are then estimated

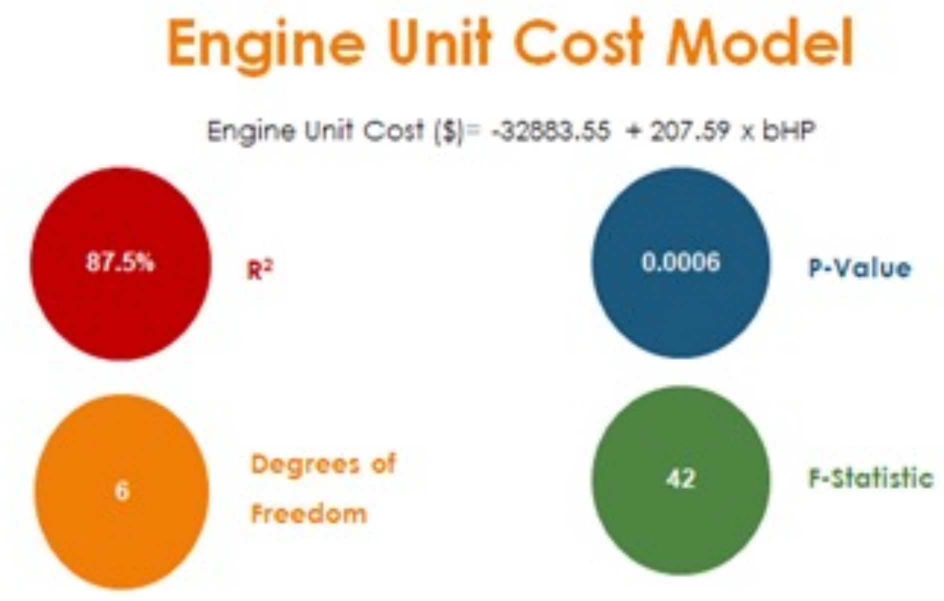
```
b.est<-c(c1$mu[1]-(c1$sigma[1,2]/c1$sigma[2,2])*c1$mu[2],c1$sigma[1,2]/c1$sigma[2,2])
```

The model can then be used to calculate the missing values for the dataset

FIT RESULTS - 2

Comparing results from the original dataset to the EM imputed dataset

Linear Model



Engine Unit Cost Model

Engine Unit Cost ($)= -32883.55 + 207.59 x bHP

87.5% R²

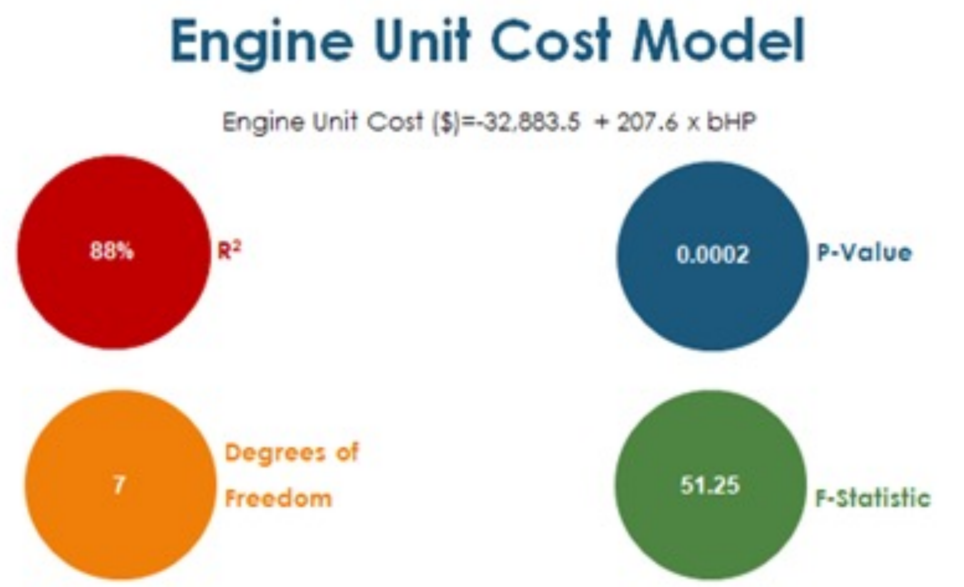
0.0006 P-Value

6 Degrees of Freedom

42 F-Statistic

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 missing a unit cost value

EM Imputed Model



Engine Unit Cost Model

Engine Unit Cost ($)=-32,883.5 + 207.6 x bHP

88% R²

0.0002 P-Value

7 Degrees of Freedom

51.25 F-Statistic

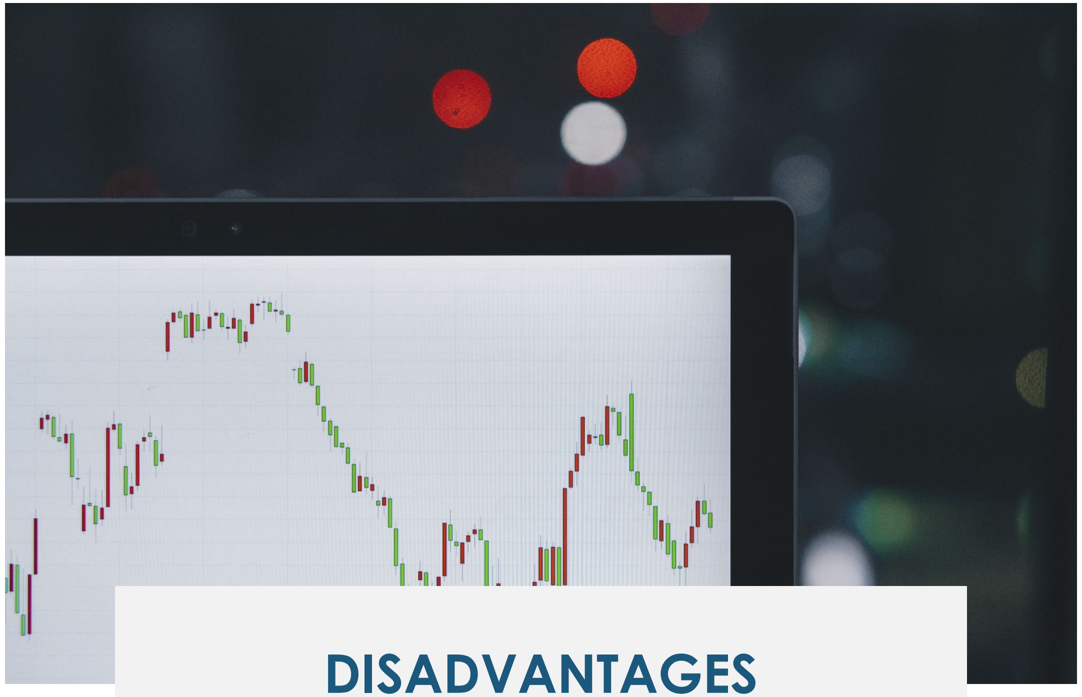
Compared to the results produced from removing the data points with missing values, this is a better performing model. A degree of freedom was gained and the R^2 metric increased while the model retained statistical significance

EXPECTATION MAXIMIZATION

Why choose EM?

ADVANTAGES

EM preserves the relationship with other variables, unlike mean imputation



DISADVANTAGES

EM can sometime underestimate standard error

COMPARING METHODS

MICE VERSUS 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, 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

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.

Questions?

GALORATH