**Case Study 10 Write Up**

**Abstract:**

In this case study, we will be working with the Boston Housing data and simulating missing values. Missing values is a phenomenon, or a reality, that can impact results of analytical studies by delivering inaccurate outcomes that are biased. Missing values plays a major part in cleaning up the data set, a step in the statistical & analytical process that makes up a disproportionate amount of the project, but is a much-needed step in order to have reliable outcomes. The cause of missing values can be human based or machine based or both.

**Introduction:**

Quality of data being used for statistical analysis, has a major bearing on the outcome of the results. In this case study we are tasked with four steps for conducting the study:

In step 1 we’ll be obtaining the Boston Housing dataset using sklearn and fitting a linear regressor to the data as a baseline and establish a loss and goodness of fit. This will be our baseline for comparison

In step 2, we will select between 1, 5, 10, 20, 33, and 50% of the data from a single column, completely at random, and replace the vale with NaN and perform imputation of that value. A goodness of fit will be performed with the imputed data and the loss and goodness of fit will be compared to the baseline established in step 1.

In step 3, we will be taking 2 different columns to create data “Missing at Random” when controlled for a third variable. We will have to impute 10, 20, and 30% of the missing data using our best guess. Again we’ll be comparing our results to the baseline.

In the last step, we’ll be creating a Missing not at Random pattern in which 25% of the data is missing for a single column. We’ll impute the data, fit the results and compare to the loss and goodness of fit baseline.

Before we start with our case study, we need to understand what different forms the missing data can take. Below is a brief explanation of the three major forms of missing data:

***Missing Completely at Random (MCAR)*** - Missing completely at random (MCAR) is the only missing data mechanism that can actually be verified. Missing data are MCAR when the probability of missing data on a variable is unrelated to any other measured variable and is unrelated to the variable with missing values itself [1] There’s no relationship between whether a data point is missing and any values in the data set, missing or observed [2] Examples include census questionnaires being lost in the mail by chance. Thus, missing data is completely random.

***Missing Not at Random (MNAR) -*** when there is a relationship between the propensity of a value to be missing and its values [2] In other words, when the missing data is related to the outcome of the experiment or the response it is considered MNAR. For example, when data are missing on IQ and only the people with low IQ values have missing observations for this variable [4]

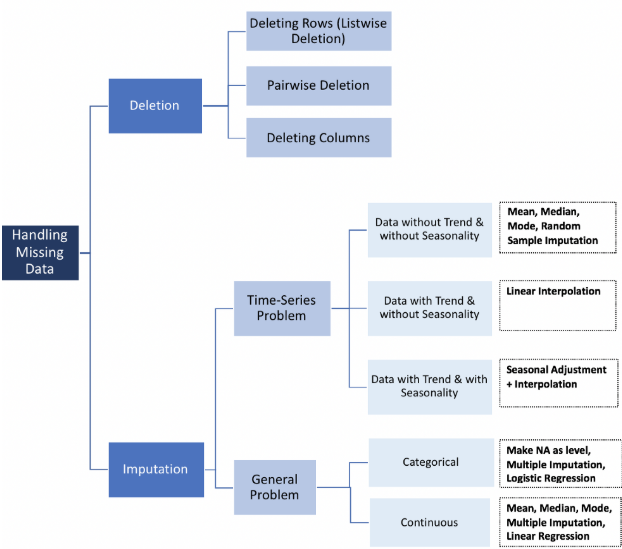
***Missing at Random (MAR) -*** Missing data are missing at random (MAR) when the probability of missing data on a variable is related to some other measured variable in the model, but not to the value of the variable with missing values itself [5]. Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data.

MNAR is called “non-ignorable” because the missing data mechanism itself has to be modeled as you deal with the missing data. You have to include some model for why the data are missing and what the likely values are.

“Missing Completely at Random” and “Missing at Random” are both considered ‘ignorable’ because we don’t have to include any information about the missing data itself when we deal with the missing data [2] But with MNAR, removing observations with missing values can produce a bias in the model.

Now that we know the different categories of missing data, we also need to understand the different methods available to us for missing data and which one would be most suitable. Below graph summarizes the appropriate method to be used for dealing with missing data [3]

**Graph 1: Methods of Handling Missing Data**



(Towards data science a better reference than <https://www.kaggle.com/questions-and-answers/88476#latest=512775>, keep that)

**Deletion:**

We have three deletion methods available that can be used for missing data

* *Listwise Deletion* - With this method all rows with any missing data on one or more variables are excluded from the analysis. The advantage of this method is that the remaining dataset is complete. However, this complete dataset has a reduced sample size and power, caused by the loss of the incomplete cases. [1] It is applicable with MCAR data
* *Pairwise Deletion* – Also called available case analysis, uses datasets that may have some missing data. It does not include cases where a particular variable has missing values, but that case will be included for analyzing other variables with non-missing values. Thus, it analyses all cases in which the variables of interest are present. It makes the assumption that missing data are MCAR. With this method means and standard deviations computed are based on all available data for each variable which can result in misleading numbers.
* *Deleting Columns* – With this a variable is dropped if it has missing values. Some argue that a column can be deleted if more than 60% of its observations are missing and if it is also insignificant [3]

**Imputation:**

Approaches to imputation can vary based on if the problem being analyzed is time-series based or a general problem

**Time Series specific Approaches:**

* *Data Without Trend and Seasonality* – When trending and seasonality is absent from the dataset, calculating mean median, or mode from all non-missing values for imputing can be used. It needs to be kept in mind that it only works with numeric data and it reduces variance in the dataset [3]
* ***Data with Trend and Without Seasonality* – Linear interpolation is recommended for time series that have some trending without any seasonality. It provides an estimated value from other observations for the variable and assumes a linear relationship between the missing and non-missing values [6]**
* ***Data with Trend and Seasonality –* Using linear interpolation with seasonal adjustment is the recommended method for imputing data with a time series component that has trending and seasonality**
* *Last Observation Carried Forward (LOCF)* – It is a single imputation method and is specific to longitudinal designs. This technique imputes the missing value with the last/most recent value of the sample as it tracks the same sample at varying points in time and the assumption is that there is no trending present in the data and hence no changes. It uses mean/median/mode. It introduces bias and will not perform well with trending as a main assumption is that the samples have no change over time [3]
* *Next Observation Carried Backward* – This is the opposite of LOCF as explained above

**Non Time Series Specific Approaches:**

* *Categorical Data* - With categorical data multiple imputation, logistic regression and making NA as a level are all available.

In Logistic Regression, a correlation matrix is used to identify predictors of variable with missing values. The best predictors are selected and used as independent variables in a regression equation and the variable with the missing data is used as the dependent variable. Instances with complete data for the predictor variables are used in the equation to predict and impute missing values for incomplete data [3]

In single regression imputation the imputed value is predicted from a regression equation. For this method the information in the complete observations is used to predict the values of the missing observations. [5]

The multiple imputation process contains three phases: A) the imputation phase, B) the analysis phase and C) the pooling phase. A) Imputed values are estimated using mean and covariance of the data and regression equation is used to predict the incomplete values. B) Statistical analysis is performed on each imputed data sets. C) the multiple sets of results are combined into a single set of results using the average of all imputed data sets [7]

Missing values can be treated as a separate category by itself. We can create another category for the missing values and use them as a different level. This is the simplest method [3]

* *Continuous Data* – The below may be used to impute continuous data:

Imputation using mean

Imputation using mode

Imputation using multiple imputation and

Imputation using linear regression

Besides the one listed in the graph other techniques available to impute data include:

* *KNN (K Nearest Neighbors*) – It is a widely used imputation technique. In this method, k neighbors are chosen based on some distance measure and their average is used as an imputation estimate. It can predict both discrete and continuous attributes [3]
* *Imputation by Most Frequent Values* – It imputes missing data with the most frequent values within each column [8]
* Imputation by *Zero/Constant Values* – In this technique missing values are replaced by either a 0 or user defined constant [8]
* *Stochastic Regression Imputation –* It adds a random residual value to the predicted value from a regression [9]
* *Hot Deck Imputation* - It makes use of a randomly chosen value to impute missing values from a case in the dataset who has similar values on other variables.
* *Cold Deck Imputation –* Similar to Hot Deck imputation but without the value being randomly chosen to impute missing values. Instead it selects values systematically from a case/instance with similar values on other variables

**Background:**

The Boston housing dataset being used has 506 rows, 14 variables and no missing data points. The variables included in the dataset are:

Table 1: Dictionary of the Variables in the Boston Housing Dataset

|  |  |
| --- | --- |
| **Variable** | **Description** |
| CRIM | per capita crime rate by town |
| ZN | proportion of residential land zoned for lots over 25,000 sq.ft. |
| INDUS | proportion of non-retail business acres per town |
| CHAS | Charles River dummy variable (= 1 if tract bounds river; 0 otherwise) |
| NOX | nitric oxides concentration (parts per 10 million) |
| RM | average number of rooms per dwelling |
| AGE | proportion of owner-occupied units built prior to 1940 |
| DIS | weighted distances to five Boston employment centers |
| RAD | index of accessibility to radial highways |
| TAX | full-value property-tax rate per $10,000 |
| PTRATIO | pupil-teacher ratio by town |
| B | 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town |
| LSTAT | % lower status of the population |
| MEDV | Median value of owner-occupied homes in $1000's |

References:

[1] <https://www.iriseekhout.com/missing-data/missing-data-mechanisms/mcar/>

[2] <https://www.theanalysisfactor.com/mar-and-mcar-missing-data/>

[3] <https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4>

[4] <https://www.iriseekhout.com/missing-data/missing-data-mechanisms/mnar/>

[5] <https://www.iriseekhout.com/missing-data/missing-data-mechanisms/mar/>

[6] Julia Zhang and David Chen, INTERPOLATION CALCULATION MADE EZ, 2001 (<https://www.lexjansen.com/nesug/nesug01/ps/ps8026.pdf>)

[7] <https://www.iriseekhout.com/missing-data/missing-data-methods/multiple-imputation/>

[8] <https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779>

[9] <https://www.theanalysisfactor.com/seven-ways-to-make-up-data-common-methods-to-imputing-missing-data/>