MCAR Test

Professor: Gopinath Panda

Group No: 19

Vats Shah - 202201417
Aniruddha Shinde - 202411065
Monson Reji Verghese - 202411039

Date: September 22, 2024

Introduction

In many statistical studies, missing data presents a significant challenge. Understanding the mechanism behind missing data is essential for accurate analysis. One such mechanism is Missing Completely at Random (MCAR). One widely-used test to assess if data is Missing Completely at Random (MCAR) is Little's MCAR Test, developed by Roderick J.A. Little in 1988., where the missingness of data is entirely unrelated to the values of the data itself, either observed or unobserved. The MCAR test is used to verify this assumption and plays a critical role in determining how to handle missing data in a dataset.

The test checks whether the missing data mechanism is random, which is important for making unbiased inferences from incomplete datasets.

Missingpy Library Overview

missingpy is a Python library designed to handle missing data, especially in time series and other datasets where traditional machine learning methods struggle due to missing values. Unlike other popular libraries like pandas or sklearn, missingpy provides specialized imputation techniques, including ones based on nearest neighbors, iterative methods, and more sophisticated algorithms tailored for structured datasets. The main goal of missingpy is to offer efficient and flexible methods to impute missing data without losing too much information. Below are the key components and methods offered by this library.

KNNImputer

Purpose: Impute missing values using k-nearest neighbors.

How It Works: For each missing data point, the algorithm finds the k nearest neighbors (based on some distance metric) and imputes the missing value by averaging or taking the median of the neighboring points.

Key Parameters:

- `n_neighbors`: Number of neighboring samples to use for imputation.
- `weights`: Can be `"uniform"` or `"distance"`; determines whether all neighbors are weighted equally or closer neighbors have a higher influence.
- `metric`: The distance metric used (e.g., Euclidean, Manhattan, etc.).

MissForest

Purpose: Impute missing values using Random Forests, a non-parametric imputation method.

How It Works: The method builds multiple random forest models to predict the missing values in an iterative manner. For each iteration, the model predicts the missing values and updates the dataset, refining the predictions.

Key Parameters:

- `n_estimators`: Number of trees in the random forest.
- `max_iter`: Maximum number of iterations.
- `random_state`: Controls the randomness of the forest.
- `min_samples_leaf`: Minimum number of samples required at a leaf node in the trees.

IterativeImputer

Purpose: Impute missing data using multivariate feature modeling.

How It Works: The iterative method models each feature with missing values as a function of other features and iteratively updates the missing values. This method often leads to more accurate imputation since it leverages the relationship between variables.

Key Parameters:

- `max_iter`: Maximum number of iterations to impute missing values.
- `random_state`: For reproducibility of results.
- `initial_strategy`: Strategy for the initial imputation (mean, median, or most frequent).

Note:- We did not use missingpy library in this assignment because we encountered errors and also we were not able to run the missingpy library on popular environments like google colab or Jupyter notebook.

Background on MCAR Test

The MCAR test was developed to assess whether data missingness is random or related to the data itself. According to the MCAR assumption, the probability of any data point being missing is independent of both observed and unobserved data. This assumption simplifies the analysis, as data that is MCAR can be safely ignored or imputed without introducing bias.

If the data is MCAR, standard methods such as listwise deletion or mean imputation can be applied without significantly distorting the results. If, however, the data does not meet the MCAR assumption, more sophisticated techniques are required, such as multiple imputation or modeling the missing data mechanism.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Mathematical Formulation of the MCAR Test

The MCAR test is a statistical test that uses a chi-square distribution to determine whether the missing data follows the MCAR assumption. The test compares observed and expected patterns of missingness across different variables. A significant result (typically a p-value less than 0.05) suggests that the data is not missing completely at random.

Hypotheses of Little's MCAR Test

The test operates based on the following hypotheses:

- **Null Hypothesis (H_0):** The data is missing completely at random (MCAR).
- **Alternative Hypothesis (H_1):** The data is not missing completely at random (it could be Missing at Random (MAR) or Missing Not at Random (MNAR)).

The goal of Little's test is to determine if the observed patterns of missingness significantly deviate from what we would expect under the MCAR assumption.

Partitioning Data Based on Missingness Patterns

Let's assume a dataset X with n observations and p variables. Some data points in the dataset are missing. The test begins by partitioning the data into groups based on the *patterns of missingness*. Each distinct pattern of missing data is treated as a separate group. Let G represent the number of distinct missing data patterns.

Mean Differences Between Groups

For each group, the means of the observed variables are calculated and compared. Under the MCAR assumption, the missingness pattern should not systematically affect the means of the observed data. Thus, Little's test checks for statistically significant differences between these means across groups. If the data is MCAR, these differences should be small and attributable to random variation.

For each variable, the observed means are calculated for both the groups with missing data and the overall dataset, and these means are compared.

Chi-Square Test Statistic

The test statistic is then compared to a chi-square distribution to determine the likelihood that the data is MCAR. If the p-value from this test is greater than 0.05, we fail to reject the null hypothesis, indicating that the data may be missing completely at random. This is based on a *likelihood-ratio chi-square test*, which compares the observed and expected values of missingness for each group. The chi-square test statistic is calculated as:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where:

- O_i is the observed frequency,
- E_i is the expected frequency.

The degrees of freedom for the chi-square test are calculated as:

$$\text{df} = (r - 1)(c - 1)$$

Where:

- r is the number of rows (categories),
- c is the number of columns.

The degrees of freedom correspond to the number of independent comparisons being made between the groups for each variable.

P-Value Calculation

The chi-square test statistic is compared to a chi-square distribution with the calculated degrees of freedom. The *p-value* is derived from this distribution is interpreted as follows:

- **If the p-value $>$** 0.05: We fail to reject the null hypothesis, suggesting that the data is missing completely at random (MCAR). In this case, simpler techniques for handling missing data, such as listwise deletion or mean imputation, can be applied without introducing bias.
- **If the p-value $<$** 0.05: We reject the null hypothesis, indicating that the missing data mechanism is likely not MCAR (it may be MAR or MNAR). In this scenario, more sophisticated techniques such as multiple imputation or model-based approaches should be considered.

Results and Discussion

Based on the dataset used, we got the following metrics,

- Chi-square statistic: 363853
- Degrees of freedom: 1084460
- P-value: 1.0

Conclusion

The MCAR test is a valuable tool for assessing the randomness of missing data in a dataset. If the MCAR assumption holds, simple techniques can be employed to handle the missing data without introducing bias. If the assumption is violated, more advanced methods should be considered to properly address the missing data problem. The test plays an essential role in ensuring that analyses remain valid and reliable even in the presence of incomplete data.

In the 120 years Olympics dataset, since the p-value is 1.0, this suggests that the missing data is very likely to be missing completely at random (MCAR). In other words, the pattern of missingness does not appear to be related to any other variables in the dataset.

The chi-square statistic is high, but given the large degrees of freedom, this is not surprising. The test essentially compares observed vs. expected patterns of missingness, and the p-value indicates that any deviations from MCAR are statistically insignificant.

You can access our work here :GitHub Repository
Google Colab Link (please access using college ID): Colab Link