République Algérienne Démocratique et Populaire
Ministére de l'enseignement supérieur et de la recherche scientifique

N° Attribué par la bibliothéque

Année univ.: 2023/2024

Testes non parametriques pour les données manquantes

Mémoire présenté en vue de l'obtention du diplôme de

Master Académique

Université de Saïda - Dr Moulay Tahar
Discipline : MATHÉMATIQUES
Spécialité : Analyse Stoquastiques Statistique des processus et Applications.

par

Kouidri **C**haimaa [1]

Sous la direction de

Dr. F. Benziadi

Soutenue le 12/06/2024 devant le jury composé de

Pr. S. Rahmani	Université Dr Tahar Moulay - Saïda	Présidente
Dr. F. Benziadi	Université Dr Tahar Moulay - Saïda	Encadreur
Dr. R. Rouane	Université Dr Tahar Moulay - Saïda	Examinatrice

[1]e-mail: chaimaakoui34@gmail.com

Dedication

All praise to Allah. Today, we wrap up the day's efforts and accomplishments within the pages of this humble work.

I dedicate my work to:

The one I prefer over myself, and why not; she has sacrificed for me and spared no effort to make me happy always ,my beloved mother.

The one who dominates our minds in every path we take remains the kind face and good deeds. He has never withheld from me throughout his life, and The person who sees me as the most beautiful and smartest girl ,my dear father.

My sisters: Khaoula, Asmaa, Aya and Houda, who have always stood by my side, ready to help me.
My brothers, Boualem and Mohamed.
My late niece Hiba , may God have mercy on her.
My nieces, Didaoui Hiba, Didaoui Djamila, Hakmi Khadidja and Hakmi Hiba.
My uncles, aunts and cousins.
My sister's husband, Didaoui Bouamama and Hakmi Oussama.
All those if my pen forget them, my heart will not forgotten them.
My friends and all who stood by my side and helped me in many ways, I present this research to you and hope it meets your approval.

Acknowledgments

In the Name of Allah, the Most Merciful, the Most Compassionate, all praise be to Allah, and prayers and peace be upon Mohammed His servant and messenger.

I extend my heartfelt gratitude to **ALLAH**, who granted me the strength, patience, and courage throughout my years of study.

I express my sincere appreciation to my supervisor, **Dr. Fatima Benziadi**, for her meticulous and effective guidance. I am profoundly grateful for her understanding, patience, assistance, and invaluable advice throughout the preparation of my dissertation. I would like to convey my heartfelt thanks for her unwavering support and guidance.

I am also grateful to the members of the jury for generously dedicating their valuable time to review and evaluate this research.

I am indebted to our university and the Laboratory of Stochastic Models, Statistics, and Applications, as well as its members, for providing me with the opportunity to undertake and complete this work.

I am deeply thankful to my parents for their unconditional love, prayers, care, and sacrifices in nurturing and preparing me for my future. I am also grateful to my brothers and my sisters for their love, understanding, prayers, and continuous support in completing this research work.

Contents

List of Figures

List of Tables

General introduction

Missing data are a common and challenging problem that complicates the statistical analysis of data collected in almost every discipline, including biology, psychology, sociology, and medicine. it arising when a sampled unit does not respond to the entire survey (unit nonresponse) or to a particular question (item nonresponse).No matter how carefully an investigator tries to have all questions fully responded to in a survey, or how well designed, any dataset frequently results in missing data.

This issue that dates back to the earliest known statistical operations, such as censuses in ancient empires. The first recorded census was conducted by the Babylonians around 3800 BC. Historical records indicate it was conducted every six or seven years and accounted for the number of people and livestock, as well as quantities of butter, milk, wool, and vegetables (see Kuhrt, 1995[18]). However, awareness of the problems caused by missing data has only emerged recently. Galton (1888)[11] was among the pioneers to study situations involving missing data. He encountered cases of incomplete measurements in his anthropometric work. His study was based on data (such as forearm length, hand length, palm, and palm width) from 350 men, but Galton noted that the exact number of 350 was not maintained throughout the study, as injuries to limbs reduced the number of individuals by 1, 2, or 3 in different cases. Later, Galton (1898)[12] considered a truncated distribution, specifically a right-truncated normal distribution, while analyzing data from Wallace's Year Book. This data included qualification times for runners who had to complete a mile in no more than 2 minutes and 30 seconds. Times for the slower runners were not recorded and thus excluded, with their number remaining unknown. He estimated the mean and identified quartiles to assess dispersion using the interquartile range. In 1931, British statistician R. A. Fisher revisited this problem using the maximum likelihood method, which was later used by Wilks (1932)[48] to estimate a covariance matrix in the presence of missing data for two variables. Twenty years later, Lord (1955)[28] extended this approach to three variables.

During the cholera outbreaks, while studying the distribution of households with 0, 1, 2, 3, or 4 cholera cases in an Indian village, McKendrick (1926)[30], a pioneer in mathematical epidemiology, found an unexpectedly high number of zeros for a Poisson distribution, while the true number of affected households and infected individuals was unknown.

Fisher (1934)[10] addressed the problem of albinism, noting the difficulty in distinguishing between families genetically capable of having albino children but who had none, and families incapable of having albino children.

During World War II, based on observations of bullet impacts on returning planes, Abraham Wald recommended reinforcing bombers everywhere, especially on the engines. The planes that returned had not been hit in the engine, while those that had been were missing. More details can be found in the works of Mangel and Samaniego (1984)[29], Wainer (2011)[47], and Ellenberg (2014)[6], with Ellenberg (2018)[7] being a French translation.

To analyze measurements of human skeletons from Jebel Moya in Sudan, Rao (1985)[38] used results from an archaeological study (Mukherjee and Rao, 1955[32]) on a sample of skulls. Some skulls were in good condition, described by four variables (capacity, length, width, and height), while other skulls were fractured, making certain measurements impossible.

"Missing data refers to a data value that should have been recorded but, for some reason, was not, Day (1999)[4]" .

When confronted with missing data, researchers employed conventional such as complete case analysis or available analysis, and noval methods such as multiple imputation or maximum likelyhood imputation, or Expectisation Maximization algorithm(EM algorithm). The majority of statistical research operates under the assumption that the data being analyzed are devoid of missing values. Often, the simplistic approach taken is to eliminate individuals with missing data.

"We find ourselves surrounded by missing data. The challenges they pose in statistical analysis have long been overlooked" (Van Buuren, 2018[45]).

However, this method of simply deleting missing data can significantly distort the outcomes of the statistical study.

"Missing data represent unobserved values that would hold significance in analysis if observed; essentially, a missing value conceals meaningful information" (Little and Rubin, 2020[26]).

Imputation involves filling in the dataset (i.e., predicting estimated values for the unobserved data). Numerous methods for imputing missing values have emerged, categorized into two branches: simple imputation and multiple imputation. The practical utilization of these techniques is increasingly widespread, with recent contributions from Van Buuren (2007)[44] and He et al. (2022)[16].

The essence of simple imputation lies in replacing missing values of survey variables with plausible substitutes. Following imputation, analyses proceed using these substituted values.

At times, considering multiple imputations of the same dataset proves beneficial. Known as multiple imputation (MI), pioneered by Rubin (1987)[41], as its name suggests, MI involves imputing missing values multiple times ($\mathbf{N}$ times with $\mathbf{N} > 1$), generating several complete datasets to amalgamate results and minimize imputation errors.

Another novel approach for approaching missing data was proposed by Orchard and Woodbury (1972)[5] using what is commonly referred to as an expectation maximization (EM) algorithm to produce unbiased estimates when the data are missing at random (MAR). ML and EM algorithms were also discussed in Dempster et al.'s (1977)[14] work.

Graham et al. (1997)[22] discussed using the EM algorithm to estimate means and covariance matrices from incomplete data. Papers from Little (1995)[23] and Little and Rubin (1989)[40] extended the concept of ML estimation in dealing data.

It is important to distinguish between two characteristics that describe the nature of missing data: one is called the "pattern," and the other relates to the process or mechanism behind the occurrence of missing data.

Rubin (1976)[24] was the first who study the mechanisms behind the presence of missing data. According to Rubin (1976)[24] and Little and Rubin (2002)[20], there are three major types of missing mechanisms that are This research has been supported in part by generally accepted and used in modern statistics: (a) missing completely at random (denoted as MCAR), if missingness does not depend on the data, missing or observed; (b) missing at random (denoted as MAR), if missingness depends only on the observed data, but not on the missing data, there another mechanism called covariate dependent missigness (denoted CDM), it's a secial case of MAR; (c) missing not at random (denoted as MNAR), if missingness depends on the missing data. Knowing the type of missing mechanisms is important for adopting the appropriate statistical procedure for the analysis of incomplete data. Many missing data methods, such as complete case analysis and available case analysis, as well as mean imputation methods generally require the MCAR assumption. If such procedures were used for the other two missing mechanisms, it would usually

cause biased inference. Therefore, it is necessary to test whether the MCAR assumption is satisfied before applying those procedures.

Little (1988)[19] first proposed a test of MCAR for incomplete multivariate data by testing the homogeneity of means across different missing pattern groups. The test is based on the likelihood ratio test assuming the normality for the data. Little (1988)[19] also mentioned a likelihood ratio test for testing homogeneity of both means and covariances across different missing pattern groups as another possible test of MCAR.

In this work, we present a non parametric test, a chis-quared test for testing MCAR mechanism for multivariate quantitative data proposed by Little (1988)[19], and a straightforward extension of Little's MCAR test for CDM assumption, highlighting their advantages, limitations.

This master memory falls into three chapters.

In chapter 1, we delve into the complex issue of missing data, aiming to provide a comprehensive understanding of its patterns, mechanisms, and methods for handling.

In chapter 2, we introduces the notation we use for Little's test for incomplete multivariate data and discusses the hypothesis testing problem corresponding to the test of MCAR. Also we present a straightforward extension of Little's MCAR test for CDM assumption.

In chapter 3, we report simulation studies for Little's test for MCAR and CDM mechanisms proposed by Little 1988[20] and Li 2014[19], respectivelly, to evaluate the performance of our proposed procedure, highliting the limitations of this test.

Chapter 1

Introduction to Missing Data

Missing data (or missing values) is defined as the data value that is not stored for a variable in the observation of interest. The problem of missing data is relatively common in almost all research and can have a significant effect on the conclusions that can be drawn from the data.

In this chapter, we delve into the complex issue of missing data, aiming to provide a comprehensive understanding of its patterns, mechanisms, and the methods for handling.

1.1 Missing Data: Reasons, Challenges, and Examples.

1.1.1 Reasons for Missing Data

The reason for the missing data is important to consider, because it helps you determine the type of missing data and what you need to do about it.
These three reasons tend to cover the largest areas of missing data in the data mining process:

- **Random errors :** such as equipment failures like malfunctioning sensors, instruments, or data collection devices, can result in missing data. Similarly, human errors, like mistakes during data entry or data processing, can lead to missing values.[3]

- **Refusal of response :** Some respondents may find certain questions in the survey offensive or be personally sensitive to certain questions. For example, some respondents may not have an opinion on certain issues, such as political or religious affiliation. In questions relating to level of education education, income, age or weight may be considered too private for some respondents to answer. In addition, respondents may simply not have sufficient knowledge to particular questions . Students or inexperienced people may have insufficient insufficient knowledge to answer certain questions. When they are asked about their future goals or career choices, they may not have time to study certain aspects of their of their career choice (such as salaries in different regions of the country, retirement retirement options, insurance choices,..., etc).[3]

- **Unworkable answers :** Sometimes questions are left blank simply because they apply apply to a more general population rather than to an individual respondent. If a subset of questions in a questionnaire questionnaire does not apply to the individual respondent, data may be be missing for a particular group within a data set. For example, many graduate students graduates may choose not to answer questions about social activities for which they simply for which they simply don't have the time. Similarly, adults who have never been married, or who are widowed or widowed or divorced are unlikely to answer a question about the number of years years of marriage.[3]

1.1.2 Problems with Missing Data

The existence of missing values may have significant influence on the analysis of the data and therefore on the conclusion of the data analysis. When missing data are present, we may have the following issues:

- **Power and variability:** With more missing data, we will have smaller sample size, which means we will have less statistical power for the analysis. And often since the extreme cases are more likely to be missing, we will have loss of data variability and the confidence interval will be forced to be narrower.

- **Bias:** For some circumstances, such as the situation where the participated interviewees in a survey are not a random sample of the population of interest, the bias issue exists. Bias is one of the worst effects that missingness brings. It also brings the issue of comparability of different groups and representativeness of the observed sample to the target population, as in some retrospective studies or observational studies.

1.1.3 Examples

Example 1: Income Nonresponse in the Current Population Survey

The Current Population Survey (CPS) is a crucial monthly survey conducted by the Census Bureau to gather diverse information from households. Specifically, in March of each year, the CPS includes a supplement to collect detailed income data. However, there's a challenge: not all individuals are willing to report their incomes. As a result, approximately 20% of surveyed individuals have missing data on one or more income items.

Moreover, the CPS encounters another issue: a small number of households fail to provide interviews at all. Income nonresponse not only reduces the efficiency of data analysis but also introduces biases because nonrespondents tend to differ from respondents. For instance, individuals with higher incomes are less likely to respond compared to those with middle incomes.

The consequence of these challenges is significant. Without adjusting for the differences between nonrespondents and respondents, analyzing CPS data can lead to biased conclusions. This is problematic given the importance of CPS data, as it serves as a key source for government figures on employment and income, and it's widely utilized by economists, social scientists, and various other professionals.

Therefore, despite the issue of income nonresponse, it's imperative for CPS databases to provide realistic answers for data analysts. This entails implementing methods to account for nonresponse biases and ensure the reliability and validity of the data for informed decision-making and policy formulation.[27]

Example 2: Nonresponse in the Fatal Accident Reporting System

The Fatal Accident Reporting System (FARS), administered by the National Highway Traffic Safety Administration, provides a publicly accessible database for analysis. This dataset contains comprehensive information on fatal accidents, including details such as accident location, vehicle involvement, and driver characteristics like age, sex, driving history, seatbelt usage, and blood alcohol content (BAC). However, critical variables like seatbelt use and BAC are often missing for many cases.

To address the issue of missing data, the National Highway Traffic Safety Administration has two primary objectives. Firstly, they aim to create a data file where missing values are filled in, enabling standard analytical methods designed for complete datasets to be applied. Secondly, they strive to equip analysts with tools to accurately estimate standard errors, which reflect the

information loss due to nonresponse. This approach ensures that analysts can conduct robust analyses and properly account for missing data in their interpretations.[27]

Example 3: The comparability of occupation codes across different time periods

In each decennial census, individuals provide details about their jobs through open-ended descriptions of occupations. These descriptions are then translated into standardized occupation codes by the Census Bureau. However, with every census, the system for classifying occupations is updated to reflect changes in job categories and economic trends.

A significant revision to this system occurred during the 1980 census, resulting in the introduction of new occupation codes. Consequently, the occupation codes in public-use datasets from the 1980 census cannot be directly compared to those from earlier censuses, such as the 1970 census. This lack of code comparability presents challenges for researchers interested in analyzing trends in occupation mobility and labor force composition over time by demographic characteristics.

The public-use datasets from the 1970 census are extensive, containing over one million records, making it impractical and costly to update them with the new 1980 codes. However, a subset of the 1970 census data, consisting of 120,000 units, has been coded with both 1980 and 1970 occupation codes.

This issue of occupation code comparability across different census years can be seen as a missing-data problem. While a small portion of the 1970 dataset contains both the 1970 and 1980 occupation codes, the majority of cases only have the old codes. Managing this discrepancy is essential for accurate longitudinal analysis and interpretation of trends over time.[27]

1.2 Basic concepts

1.2.1 Unit versus Item Nonresponse

In survey contexts, two types of missing data are commonly distinguished: **unit nonresponse**, where entire questionnaires are missing due to the inability to contact or interview a sampled individual, and **item nonresponse**, where specific questions are missing within an interview, either due to refusal to answer, interviewer errors, or deletion of inconsistent responses during the editing process.

Unit nonresponse often leads to a situation where survey variables are missing for nonrespondent units, while survey design variables (such as geographical information) are still available. This issue is often addressed through weighting adjustments, where nonrespondent units are excluded from the dataset, and weights are assigned to respondent units to correct for potential biases resulting from systematic differences between respondents and nonrespondents. On the other hand, **item nonresponse** is typically handled by imputing missing items or marking them with a code to indicate their absence.[27]

1.2.2 Notation

Suppose that if the data were complete, they could be arranged in an $(n \times p)$ data matrix $\mathbf{Y} = y_{ij}$, such that y_{ij} is the value of the jth variable for the ith unit, $i = 1, ..., n$; $j = 1, ..., p$. Let $\mathbf{M} = m_{ij}$ denote an $(n \times p)$ missing-data indicator matrix, such that :

$$\mathbf{M} = \begin{cases} 0 & \text{si} \quad y_{ij} \text{ missing} \\ 1 & \text{si} \quad y_{ij} \text{ observed} \end{cases}$$

We write $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$, where $\mathbf{Y}_{obs}$ represents the observed part of $\mathbf{Y}$ and $\mathbf{Y}_{mis}$ denotes the missing part. Let $\mathrm{P}(\mathbf{M} \mid \mathbf{Y}, \psi)$ denote the conditional distribution of $\mathbf{M}$ given $\mathbf{Y}$ and ψ, where ψ

is a set of unknown parameters (the parameter vector governing the model for the missingness mechanism).
Let $\mathbf{X} = x_{ij}$, such that x_{ij} is the value of the kth variable for the ith unit, be the $(n \times q)$ data matrix of covariate variables, $i = 1, ..., n$; $k = 1, ..., q$.

1.2.3 Missing Data Patterns

A missing data pattern refers to the configuration of observed and missing values in a data set. This term should not be confused with a missing data mechanism, which describes possible relationships between the data and one's propensity for missing values. Roughly speaking, patterns describe where the holes are in the data, whereas mechanisms describe why the values are missing. Figure 1.1 shows four prototypical missing data patterns, with shaded areas representing the location of the observed values.
Here the matrix $\mathbf{M}$ describes the pattern of missing data. It is useful when discussing missing data analysis to treat $\mathbf{M}$ as a stochastic matrix.

the most commonly considered pattern is **univariate** nonresponse, where (possibly after rearrangement of the rows and columns), y_{ij} is observed for $i = 1, \ldots, n$ and $j = 1, \ldots, p-1$, and y_{ip} is observed for $i = 1, \ldots, n_o$ and missing for $i = n_o + 1, \ldots, n_0 + n_1 = n$. Thus, with univariate nonresponse, missing data are confined to variable p.

The **multivariate** pattern is obtained when the single incomplete variable Y_p in Figure 1.1(a) is replaced by a set of variables $Y_{j+1}, \ldots, Y_p$, all observed or missing on the same set of units (see Figure 1.1(b)).

Univariate nonresponse (Patterns (a)) and **multivariate** nonresponse (patterns (b)) is a special case of **monotone** missing data (see Figure 1.1(c)) , where (perhaps after rearranging columns), the variable Y_j is observed whenever Y_{j+1} is observed, for $j = 1, \ldots, p-1$. Thus, for any i, $m_{ij} = 1$ implies that $m_{ij'} = 1$ for all $j' < j$. In other words, the first variable in $\mathbf{Y}$ is at least as missing as the second variable, which is at least as missing as the third variable, and so on. Such a pattern of missingness, or a close approximation to it, is not uncommon in practice. Monotone patterns often arise in repeated-measures or longitudinal data sets, because if a unit drops out of the study in one time period, then the data will typically be missing in all subsequent time periods. Sometimes a nonmonotone missing-data pattern can be made monotone, or nearly so, by reordering the variables according to their missingness rates.

A **general** pattern is perhaps the most common configuration of missing values. As seen in Figure 1.1(d)), a general pattern has missing values scattered throughout the entire data matrix in a haphazard fashion.[8]

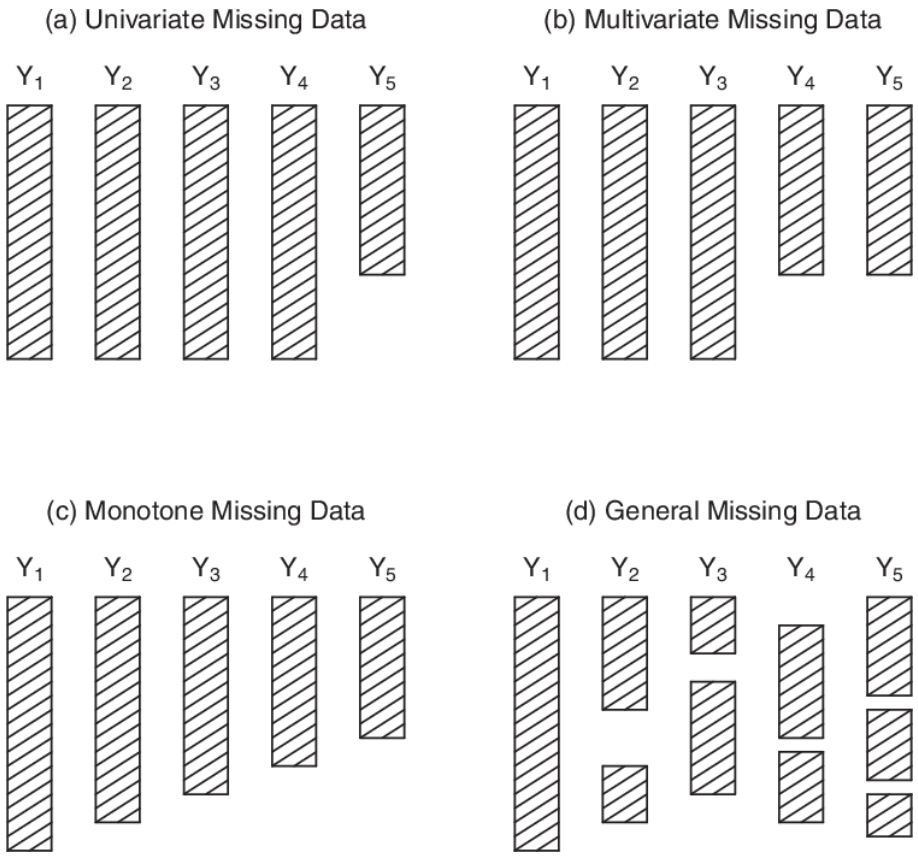


Fig. 1.1: Examples of missing-data patterns.

1.2.4 Missing Data Mechanisms

Rubin (1976)[40] first introduced a classification system for missing data problems that is widely used in the literature today. This work has generated three so-called missing data mechanisms that describe how the probability of a missing value relates to the data, if at all. In general, there are four types of missing data according to the mechanisms of missingness.

- Missing completely at random .
- Missing at random .
- Missing not at random .

Missing Completly At Random

A Missing Completely at Random (MCAR) mechanism states that the probability of missing values is unrelated to both the observed and missing parts of the data. This process is considered purely random missingness by researchers. Rubin's 1976[40] formal definitions involve the conditional distribution of the indicator variables in M given the observed data Y_{obs} and the missing data Y_{mis}. The distribution for an MCAR process is:

$$\mathrm{P}(\mathbf{M} \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \psi) = \mathrm{P}(\mathbf{M} \mid \psi) \tag{1.1}$$

The left side of the equation, which encompasses all possible associations between the indicators and the data, indicates that the probability of a missing value depends on both the observed and missing parts of the data, as well as some parameters governing missingness. The MCAR process on the right side of the equation (1.1) simplifies by removing all dependence on the realized data.

In other words, the equation (1.1) suggests that all participants have the same chance of having missing values.

For example, consider a scenario when a glass slide with biopsy material from a patient is accidentally broken such that pathology and histology tests cannot be performed, or when individuals had no blood pressure measured as the equipment was broken. Thus, under **MCAR**, missing data do not depend on either observed data or missing data. In this case, the glass slide of any patient can be broken.[9]

the statistical advantage of data that are MCAR is that the analysis remains unbiased. Power may be lost in the design, but the estimated parameters are not biased by the absence of the data.

Missing At Random

A Missing At Random (MAR) mechanism implies that the probability of missing values is related to the observed data but not to the missing data. The formal definition is given by:

$$\mathrm{P}(\mathbf{M} \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \psi) = \mathrm{P}(\mathbf{M} \mid \mathbf{Y}_{obs}, \psi) \tag{1.2}$$

This equation (1.2) means that the potential values in $\mathbf{Y}_{mis}$ do not provide any additional information about the missingness beyond what is already contained in the observed data. The term MAR is often misunderstood, as it suggests a random process rather than a systematic one. However, it actually means that missingness is random after accounting for the observed data. Data are considered MAR if they are missing because of some potentially observable, nonrandom, systematic process.[9]

Covariate Dependent Missingness

Note that here $\mathbf{M}$ is also independent of covariates $\mathbf{Y}$, as suggested by Little (1995)[22]. This means that under the MCAR assumption, the missingness should be totally independent of any observed variables. Instead, if $\mathbf{M}$ only depends on covariates $\mathbf{X}$

$$\mathrm{P}(\mathbf{M} \mid \mathbf{Y}_{obs}\mathbf{Y}_{mis}, \mathbf{X}, \psi) = \mathrm{P}(\mathbf{M} \mid \mathbf{X}, \psi). \tag{1.3}$$

Then Little (1995)[22] suggested that equation (1.3) be referred to as "covariate-dependent missingnes" (CDM). It is important to highlight that as per the definition, CDM is a special case of MAR since covariates $\mathbf{x}$ are always fully observed.

In longitudinal studies, researchers often include covariates in their analysis to help understand the relationships and interactions between different factors and the primary outcome, and to explain the missingness in the data, these covariates must be completely observed. For example, blood pressure outcome data could be CDM if missingness in blood pressure measurement depends on covariates (e.g. age, gender or weight), but given these, not on the blood pressure measurement itself. CDM is an example of a MAR mechanism when covariates are fully observed.

Missing Not At Random

A missing not at random mechanism (also referred to as a not missing at random process) states that the probability of missing values is related to the observed and missing parts of the data. The formal definition of this mechanism is as follows.

$$\mathrm{P}(\mathbf{M} \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \mathbf{X}, \psi). \tag{1.4}$$

Unlike the previous expressions, the conditional distribution of the missing data indicators doesn't simplify and features two distinct determinants of missingness.

For example, overweight or underweight individuals may be more likely to have their weight measured than individuals with normal weight, even after age is accounted for. Thus, the reason for missingness is related to unobserved characteristics of the individual, and thereby, data are MNAR. Another example is when data on income are missing, the probability of missingness may be related to the level of income; for instance, those with very low or high income might refuse to report their income.[9]

Example

To illustrate, let's examine the small dataset provided in Table 1.1. These data were crafted to simulate a scenario in employee selection by Enders (2010)[8], where candidates undergo an IQ (intelligence quotient) test during their job interviews, and later, their job performance is evaluated by a supervisor after a 6-month probationary period.

For the MCAR column, there is no relationship between IQ and the job performance ratings. A case with a lower IQ is just as likely to be missing as a case with a higher IQ.
Notably, in Table 1.1[8], the job performance ratings in the MAR column are absent for candidates with the lowest IQ scores. As a result, the likelihood of a missing job performance rating is solely determined by IQ scores and bears no relation to an individual's actual job performance.
The job performance ratings in the MNAR column are missing for the applicants with the lowest job performance ratings. Consequently, the probability of a missing job performance rating is dependent on one's job performance.[8]

Tab. 1.1: Job Performance Ratings with MCAR, MAR, and MNAR Missing Values

IQ	**Job Performance Ratings**			
	Complete	**MCAR**	**MAR**	**MNAR**
78	9	-	-	9
84	13	13	-	13
84	10	-	-	10
85	8	8	-	-
87	7	7	-	-
91	7	7	7	-
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	-	7	-
99	7	7	7	-
105	10	10	10	10
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	-	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	-	12	12

1.2.5 Ignorable and Nonignorable Missingness

The terms **ignorable** and **nonignorable** missingness are commonly used to refer to MAR (missing at random) and MNAR (missing not at random) processes, respectively. However, these terms have a broader definition, though the difference is often negligible in practice. Rubin's classification scheme involves two models: the focal analysis model, which you would estimate if the data were complete, and the missingness mechanism model. These models have parameters θ and ϕ, respectively. Usually interest is in the parameters θ, with ϕ are considered nuisances because they are not related to the main research goals. For example, the rows of $\mathbf{Y}$ may be assumed to have independent multivariate normal distributions with mean $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$, and $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, so here we interest in the parameters θ. The key question is when can we estimate θ from the observed data without needing to model the missingness mechanism or its parameters ϕ? This is the essence of ignorability.

Rubins 1976[40] shows that the missingness model is considered ignorable if:
(1) the missing values follow a MAR process, which is the important condition.
(2) the nuisance parameters in ϕ provide no information about the parameters in θ, the parameters θ and ψ are distinct, in the sense that the joint parameter space (θ, ψ), say $\Omega_{\theta,\psi}$, is the product of the parameter space Ω_θ of θ and the parameter space Ω_ψ of ψ, that is $\Omega_{\theta,\psi} = \Omega_\theta \times \Omega_\psi$.

The expression on the left of the equation represents a compact notation for the joint (multivariate) distribution of both the observed data and the indicators for missing data.

$$f(\mathbf{Y}_{obs}, \mathbf{M} \mid \theta, \phi) = f(\mathbf{M} \mid \mathbf{Y}_{obs}, \phi) \times f(\mathbf{Y}_{obs} \mid \theta) \tag{1.5}$$

Rubin's argument states that when θ and ϕ are independent, we can factorize the joint distribution as the right side of the equation (1.5). The missingness model is ignorable in this case because $f(\mathbf{M} \mid \mathbf{Y}_{obs}, \phi)$ acts as a constant, and estimating the focal model parameters from the observed data yields the same results with or without this term. Conversely, the missingness model is nonignorable if the missing values follow an MNAR process or if ϕ provides information about θ. In such cases, valid estimates of θ require pairing the focal analysis model with an additional model for missingness.[9]

1.3 Methods for Dealing with Missing Data

Several statistical approaches have been developed for dealing with missing data. The most common methods can be classified into one of the following groups:

- Traditionnal methods:
  - Deletion methods : complete-case analyses, available case analysis.
  - Single imputation methods: mean imputation, hot deck imutation, regression imputation, and stochastic regression imputation.
- Advanced methods:
  - Multiple Imputation (MI).
  - Maximum Likelyhood method (ML).
  - Expectation Maximization algorithm (EM).

1.3.1 Conventional Methods

Complete Case Analysis

The most common method for handling missing data is complete case analysis, also known as listwise deletion. This method simply deletes observations that have missing data on any variables in the model of interest. Only complete cases are used.

Advantages

Listwise deletion has two big and obvious attractions:

- It is easy and can be used with any statistical method. Furthermore, if the data are MCAR, listwise deletion will not introduce any bias into estimates. That is because, under MCAR, the subsample of complete cases is effectively a simple random sample from the original sample, and it is well known that simple random sampling does not introduce bias.
- Last, and quite important, listwise deletion produces estimated standard errors that consistently estimate the true standard errors. Thus, unlike conventional imputation methods, listwise deletion is "honest": it does not assume that one has more or better data than are actually available.

Disadvantages

Listwise deletion has two problems:

- The obvious downside of listwise deletion is that, quite often, it discards a great deal of potentially useful information. As a consequence, the true standard errors may be much higher than necessary, implying unnecessarily wide confidence intervals and high p-values.
- A second undesirable feature of listwise deletion is that parameter estimates may be biased if the data are MAR but not MCAR. For example, if men are less likely to report income than women, estimates of mean income for the whole population are likely to be biased downward. Violation of MCAR does not always result in biased estimates under listwise deletion, however. In fact, when predictor variables in regression analysis (either linear or logistic) have missing data, listwise deletion yields unbiased estimates of coefficients even when the data are not missing at random . Thus, even if high income people are less likely to report their income, coefficients for income as a predictor are not biased by listwise deletion. But deletion of what may be a large number of cases may still result in a loss of power.

Example

Diastolic blood pressure was measured for six patients; the results are presented in Table 1.2.

Tab. 1.2: Diastolic blood pressure (mm Hg) for six patients

Subject	Diastolic blood pressure (mm Hg)
1	75
2	?
3	90
4	92
5	?
6	80

In the complete case analysis, subjects with any missing observations are excluded. As a result, the final dataset for analysis would include only subjects 1, 3, 4, and 6, as shown in Table 1.3.

Tab. 1.3: Diastolic blood pressure (mm Hg) for six patients

Subject	Diastolic blood pressure (mm Hg)
1	75
3	90
4	92
6	80

Available Case Analysis

Available case analysis, or pairwise deletion, uses all available data to estimate parameters of the model. When a researcher looks at univariate descriptive statistics of a data set with missing observations, he or she is using available case analysis, examining the means and variances of the variables observed throughout the data set. When interest focuses on bivariate or multivariate relationships, the potential problems increase. Figure 1.2[36] illustrates a simple two-variable data matrix with only one variable subject to nonresponse. In pairwise deletion, all cases would be used to estimate the mean of Y_1, but only the complete cases would contribute to an estimate of Y_2, and the correlation between Y_1 and Y_2. Different sets of cases are used to estimate parameters of interest in the data:

$\bar{Y}_1 = \sum_{i=1}^{n} y_{i1}$

$\bar{Y}_2 = \sum_{i=1}^{m} y_{i2}$

$$s_1^2 = \frac{\sum_{i=1}^{n}(y_{i1} - \bar{Y}_1)^2}{n-1}$$

$$s_2^2 = \frac{\sum_{i=1}^{m}(y_{i2} - \bar{Y}_2)^2}{m-1}$$

$$r_{xy}^2 = \frac{\sum_{i=1}^{m}(y_{i1} - \bar{Y}_1)(y_{i2} - \bar{Y}_2)}{s_{1(m)}s_2}$$

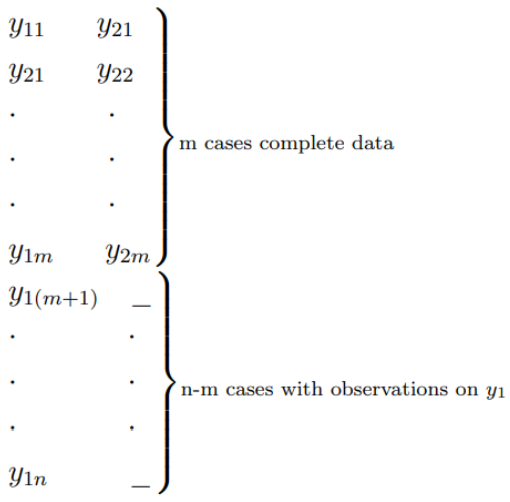


Fig. 1.2: Illustration of missing data restricted to one variable

where $\bar{Y}_1$, $\bar{Y}_2$, s_1^2, s_2^2, and r_{xy}^2, are, the mean of Y_1 , the mean of Y_2, the variance of Y_1, the variance of Y_2, correlation coefficient between variables Y_1 and Y_2.

The estimates can be improved by using available cases instead of complete cases, but there are problems with this procedure.[36]

Advantages

While this method has some advantages, such as using all available data and maintaining sample size, it also has several significant problems and limitations:

- Little (1992)[21] shows that available case analysis provides consistent estimates, when variables are moderately correlated in regression models. When variables are highly correlated, available case analysis provides estimates that are inferior to complete case; Another difficulty is that available case analysis can produce estimated covariance matrices that are implausible, such as estimating correlations outside of the range of -1.0 to 1.0.

- Errors in estimation occur because of the differing numbers of observations used to estimate components of the covariance matrix. The relative performance of complete-case analysis and available case analysis, with MCAR data, depends on the correlation between the variables; available case analysis will provide consistent estimates only when variables are weakly correlated. The major difficulty with available case analysis lies in the fact that one cannot predict when available case analysis will provide adequate results, and is thus not useful as a general method.[36]

Mean Imputation

In a mean substitution, the mean value of a variable is used in place of the missing data values for that same variable. Let y_{ij} be the jth value of Y_j for ith unit. The estimate of missing values y_{ij} is $\bar{Y}_j$, the mean of the recorded values of Y_j:

$$\bar{Y}_j = \frac{\sum_{i=1}^{r} y_{ij}}{r}$$

where r the number of the observed values for the jth variable.

Disadvantages

Little (1992)[21] points out that while mean imputation results in overall means that are equal to the complete case values, the variance of these same variables is underestimated. This underestimation derives from two sources:

- First, filling in the missing values with the same mean value does not account for the variation that would likely be present if the variables were observed. The true values probably vary from the mean.

- Second, the smaller standard errors due to the increased sample size do not adequately reect the uncertainty that does exist in the data. A researcher does not have the same amount of information present when some cases are missing important variables as he or she would have with completely observed data. Bias in the estimation of variances and standard errors are compounded when estimating multivariate parameters such as regression coeffcients. Under no circumstances does mean imputation produce unbiased results.

Example

Six patients were measured to find out their height and weight. The resulting measurements are presented in Table 1.4 below.

Tab. 1.4: Measured height and weight with missing values

Subject	Height (cm)	Weight (kg)
1	170	72
2	?	60
3	160	?
4	196	112
5	?	58
6	180	79

As can be seen from Table 1.4, one subject (subject number 3) is missing his/her weight and another one (subject number 5) his/her height. For the mean imputation method, the mean for height

$$\frac{170 + 160 + 196 + 180}{4} = 176.5 \text{ cm}$$

and the mean for weight

$$\frac{72 + 60 + 112 + 58 + 79}{5} = 76.2 \text{ kg}$$

are calculated based on the available data. Imputing the means for missing values leads to the following data set (Table 1.5).

Tab. 1.5: Measured height and weight with missing values

Subject	Height (cm)	Weight (kg)
1	170	72
2	176.5	60
3	160	76.2
4	196	112
5	176.5	58
6	180	79

The Hot Deck Imputation

The hot deck procedure involves imputing a participant's missing values using theoretically similar variables that are observed for another participant. A participant's missing values are imputed using the values from a participant with similar observed values.

Myers[31] suggested that the hot deck procedure can be performed under the following conditions: (a) up to 20% for data that is MCAR or MAR, and (b) up to 10% for data that is MNAR. Although, to address MNAR, counseling researchers will likely want to consider more robust methods for handling missing data, such as MI procedures , which are discussed in the following section 1.3.2.[31]

Advantages

The hot deck method of handling missing data offers several advantages over listwise and pairwise deletion. Primarily, hot deck procedures allow for retention of the complete sample of individuals, avoiding the loss of incomplete cases and the subsequent declines in statistical power

that are incurred as a result. Siddique and Belin (2008)[43] argue that the benefits of hot deck imputation include:
1) imputations tend to be realistic since they are based on values observed else where .
2) imputations will not be outside the range of possible values .

In the comparison of various techniques of handling missing data, the researchers found that hot deck imputation to be over 80 times more effective than listwise deletion and that hot deck imputation also outperformed pairwise deletion and mean substitution.

Furthermore, users of hot deck imputation are in good company, as many prominent large-scale surveys implement hot deck procedures to deal with missing data, including the U.S. and British censuses, the Current Population Survey, the Canadian Census of Construction, the U.S. Annual Survey of Manufacturers, and the U.S. National Medical Care Utilization and Expenditure Survey. Hot deck imputation is recommended by Roth (1994)[39] for all missing data scenarios, except those where the data are MNAR and constitute greater than 10% of the sample (in which case ML, MI, and EM techniques are recommended; see Figure 1.3[31]). Finally, the relative simplicity of the hot deck technique in comparison to model based techniques makes it an attractive alternative to listwise deletion and has the potential to facilitate wide use and application.[31]

Disadvantages

Despite the advantages of this method ,the use of the hot deck imputation does have several limitations. The first is that unique cases, that is, cases that are dissimilar to all others in the data set on the combination of sorting variables so that no "deck match" can be found, produce a Thus, there would be no "donor" available. This situation occurs more often in small data sets, when many sorting variables are used, when decks are defined by continuous variables, or when decks are defined by variables with many unique values. It is optimal to balance the size of the file with the number of sorting variables. A larger file can support the use of more sorting variables than a smaller file. Another problem noted by Siddique and Belin (2008)[**?**] is that single hot deck proceduresfail to account for the uncertainty due to the fact that the analyst does not know the values that might have been observed).[31]

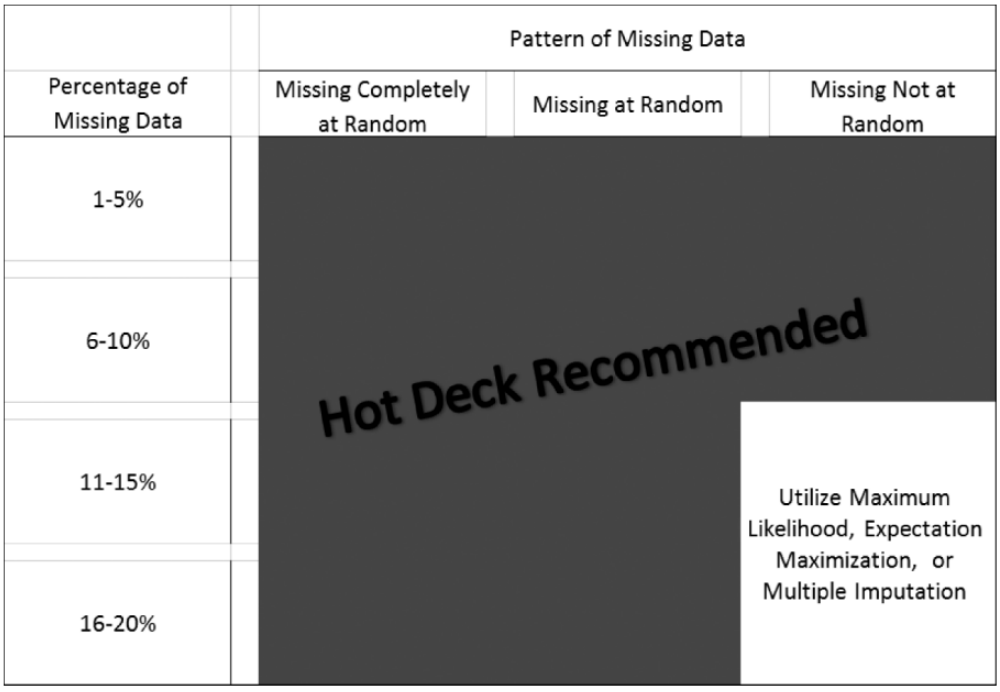


Fig. 1.3: Range of Hot Deck Applicability

Example

From the data set in Table1.6[3], it is noted that case three is missing data for item four. In this example, case one, two, and four are examined. Using hot deck imputation, each of the other cases with complete data is examined and the value for the most similar case is substituted for the missing data value. Case four is easily eliminated, as it has nothing in common with case three. Case one and two both have similarities with case three. Case one has one item in common whereas case two has two items in common. Therefore, case two is the most similar to case three.

Once the most similar case has been identified, hot deck imputation substitutes the most similar complete case's value for the missing value. Table1.7[3] provides the revised data set and displays the hot deck imputation results. Since case two contains the value of 13 for item four, a value of 13 replaces the missing data point for case three.[3]

Tab. 1.6: Illustration of hot deck imputation: incomplete data set

Case	Item 1	Item 2	Item 3	Item 4
1	10	2	3	5
2	13	10	3	13
3	11	10	3	???
4	2	5	10	2

Tab. 1.7: Illustration of hot deck imputation: imputed data set

Case	Item 1	Item 2	Item 3	Item 4
1	10	2	3	5
2	13	10	3	13
3	11	10	3	13
4	2	5	10	2

Regression Imputation:

This approche involves to replace the missing values for a unit by their predicted values from a regression of the missing variable on variables observed, usually calculated from units with both variables observed.

Disadvantages

Although regression imputation is useful for simple estimates, it has several inherent disadvantages:

- This method reinforces relationships that already exist within the data. As this method is utilized more often, the resulting data becomes more reflective of the sample and becomes less generalizable to the universe it represents.
- The variance of the distribution is understated.
- The assumption is implied that the variable being estimated has a substantial correlation to other attributes within the data set.
- The estimated value is not constrained and therefore may fall outside predetermined boundaries for the given variable. An additional adjustment may necessary.

Example

Consider univariate nonresponse, with Y_1, Y_2 fully observed and Y_3 observed for the first m units and missing for the last n-m units. Regression imputation computes the regression of Y_3 on Y_1, Y_2 based on the m complete units and then fills in the missing values as predictions from the regression. Table 1.8[3] displays an example of regression imputation. From the table, 20 cases with three variables (income, age, and years of college education) are listed. Income contains missing data and is identified as the dependent variable while age and years of college education are identified as the independent variables. The following regression equation is produced for the example :

$$Y_3^* = 33912.14 + 300.87(\text{age}) + 1554.25(\text{years of college education})$$

Predictions of income can be made using the regression equation and the right-most column

Tab. 1.8: Illustration of regression imputation

Case	Income($)	Age	Years of college education	Regression prediction($)
1	45,251.25	26	4	47,951.79
2	62,498.27	45	6	56,776.85
3	49,350.32	28	5	50,107.78
4	46,424.92	28	4	48,553.54
5	56,077.27	46	4	53,969.22
6	51,776.24	38	4	51,562.25
7	51,410.97	35	4	50,659.64
8	64,102.33	50	6	58,281.20
9	45,953.96	45	3	52,114.10
10	50,818.87	52	5	57,328.70
11	49,078.98	30	0	42,938.29
12	61,657.42	50	6	58,281.20
13	54,479.90	46	6	57,077.72
14	64,035.71	48	6	57,679.46
15	51,651.50	50	6	58,281.20
16	46,326.93	31	3	47,901.90
17	53,742.71	50	4	55,172.71
18	???	55	6	59,785.56
19	???	35	4	50,659.64
20	???	39	5	53,417.37

of the table displays these predictions. For cases 18, 19,and 20 ,income is predicted to be $59,785.56\$, 50,659.64\$$, and $53,417.37\$$, respectivelly.[3]

Stochastic Regression Imputation

Stochastic regression imputation also uses regression equations to predict incomplete variables from complete variables, but it takes the additional step of augmenting each predicted score with a random noise term from a normal distribution. Adding these residuals to the predicted values restores lost variability to the data and effectively eliminates the biases associated with standard regression imputation schemes. In fact, stochastic regression imputation is the only procedure in this section that is generally capable of producing unbiased parameter estimates when scores are MAR.

1.3.2 Novel Methods

Multiple Imputation

Multiple imputation is another useful strategy for handling the missing data. In a multiple imputation, instead of substituting a single value for each missing data, the missing values are replaced with **N** values for each missing item and creating **N** completed data sets, but the missing values are filled in with different imputations to reflect the natural variability and uncertainty of the right values.

This approach begin with a prediction of the missing data using the existing data from other variables . The missing values are then replaced with the predicted values, and a full data set called the imputed data set is created. This process iterates the repeatability and makes multiple imputed data sets (hence the term "multiple imputation"). Each multiple imputed data set produced is then analyzed using the standard statistical analysis procedures for complete data, and gives multiple analysis results. Subsequently, by combining these analysis results, a single overall analysis result is produced.

The benefit of the multiple imputation is that in addition to restoring the natural variability of the missing values, it incorporates the uncertainty due to the missing data, which results in a valid statistical inference. Restoring the natural variability of the missing data can be achieved by replacing the missing data with the imputed values which are predicted using the variables correlated with the missing data.

Incorporating uncertainty is made by producing different versions of the missing data and observing the variability between the imputed data sets. Multiple imputation has been shown to produce valid statistical inference that reflects the uncertainty associated with the estimation of the missing data. Furthermore, multiple imputation turns out to be robust to the violation of the normality assumptions and produces appropriate results even in the presence of a small sample size or a high number of missing data. With the development of novel statistical software, although the statistical principles of multiple imputation may be difficult to understand, the approach may be utilized easily.

Example

Suppose a data set has three variables, Y_1, Y_2 , and Y_3. Suppose Y_1 and Y_2 are fully observed, but Y_3 has missing data for, say, 20% of the cases. To impute the missing values for Y_3, a regression of Y_3 on Y_1 and Y_2 for the cases with no missing data yields the imputation equation(1.6)

$$\widehat{Y_3} = b_0 + b_1 Y_1 + b_2 Y_2 \tag{1.6}$$

Conventional imputation would simply plug in values of Y_1 and Y_2 for the cases with missing data and calculate predicted values of Y_3. But this imputed values have too small a variance. To correct this problem, we instead use the imputation equation(1.7)

$$\widehat{Y_3} = b_0 + b_1 Y_1 + b_2 Y_2 + sE \tag{1.7}$$

where E is a random draw from a standard normal distribution (with a mean of 0 and a standard deviation of 1) and s is the estimated standard deviation of the error term in the regression (the root mean squared error). Adding this random draw raises the variance of the imputed values to approximately what it should be and, hence, avoids the biases that usually occur with conventional imputation.

If parameter bias were the only issue, imputation of a single data set with random draws would be sufficient. Standard error estimates would still be too low, however, because conventional software cannot take account of the fact that some data are imputed. Moreover, the resulting

parameter estimates would not be fully efficient (in the statistical sense), because the added random variation introduces additional sampling variability.

The solution is to produce several data sets, each with different imputed values based on different random draws of E. The desired model is estimated on each dataset, and the parameter estimates are simply averaged across the multiple runs. This yields much more stable parameter estimates that approach full efficiency.

With multiple data sets we can also solve the standard error problem, by calculating the variance of each parameter estimate across the several data sets. This "between" variance is an estimate of the additional sampling variability produced by the imputation process. The "within" variance is the mean of the squared standard errors from the separate analyses of the several data sets. The standard error adjusted for imputation is the square root of the sum of the within and between variances (applying a small correction factor to the latter). The formula (Rubin, 1987[41]) is as follows:

$$\sqrt{\frac{1}{\mathbf{N}}\sum_{i=1}^{\mathbf{N}} s_i^2 + \left(1+\frac{1}{\mathbf{N}}\right)\left(\frac{1}{\mathbf{N}-1}\right)\sum_{i=1}^{\mathbf{N}}(a_i-\bar{a})^2} \tag{1.8}$$

In this formula, $\mathbf{N}$ is the number of data sets, s_i is the standard error in the ith data set, a_i is the parameter estimate in the ith data set, and $\bar{a}$ is the mean of the parameter estimates. The factor $(1 + (1/\mathbf{N}))$ corrects for the fact that the number of data sets is finite. How many data sets are needed? With moderate amounts of missing data, five are usually enough to produce parameter estimates that are more than 90 percent efficient. More may be needed for good estimates of standard errors and associated statistics, however, especially when the fraction of missing data is large.[2]

Maximum Likelihood for General Patterns of Missing Data: with Ignorable Nonresponse

In this section, we discuss the theory and implementation of maximum likelihood (ML) estimation for general patterns of missing data, under the assumption that the missing data mechanism is ignorable. The ML approach provides a unified and theoretically sound framework for handling missing data and is applicable to a wide range of models and data structures.

Let $\mathbf{Y} = (y_{ij}), i = 1, \ldots, n, j = 1, \ldots, p$ denote the data matrix if there were no missing values, with $y_{ij} \in \Omega_{ij}$, its sample space. We can model the density of the joint distribution of $\mathbf{Y}$ and $\mathbf{M}$ using the "selection model" factorization (defined by Little and Rubin 2002[24])

$$\mathrm{P}(\mathbf{Y}, \mathbf{M} \mid \theta, \psi) = f(\mathbf{Y} \mid \theta) f(\mathbf{M} \mid \mathbf{Y}, \psi), \tag{1.9}$$

where $f(\mathbf{Y} \mid \theta)$ represents the distribution of the data matrix $\mathbf{Y}$ assuming no missing values, $f(\mathbf{M} \mid \mathbf{Y}, \psi)$ represents the model for the missing-data mechanism, as we denote before θ is the parameter vector governing the data model, and ψ is the parameter vector governing the model for the missingness mechanism. The full likelihood based on the observed values (y_{obs}, m) and the assumed model (1.9) is defined to be

$$L_{\text{full}}(\theta, \psi \mid \mathbf{Y}_{obs}, m) = \int f(\mathbf{Y}_{obs}, \mathbf{Y}_{mis} \mid \theta) f(\mathbf{M} \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \psi)\, d\mathbf{Y}_{mis}, \tag{1.10}$$

considered as a function of the parameters (θ, ψ). The likelihood of θ ignoring the missingness mechanism is defined to be

$$L_{\text{ign}}(\theta \mid \mathbf{Y}_{obs}) = \int f(\mathbf{Y}_{obs}, \mathbf{Y}_{mis} \mid \theta) d\mathbf{Y}_{mis}, \tag{1.11}$$

which does not involve the model for $\mathbf{M}$. The term "ignorable likelihood" is sometimes used for equation (1.11), hence the notation L_{ign}.

Maximum-likelihood ignoring the missing-data mechanism means maximizing (1.11) rather than the full likelihood (1.10).

ML estimates can be found by solving the likelihood equation

$$D\ell(\theta \mid \mathbf{Y}_{obs}) \equiv \frac{\partial \ln L(\theta \mid \mathbf{Y}_{obs})}{\partial \theta} = 0. \tag{1.12}$$

When a closed-form solution of (1.12) cannot be found, iterative methods can be applied, like the EM algorithm.[26]

EM Algorithm

The EM algorithm (Expectation-Maximization) is a general iterative algorithm for ML estimation in incomplete-data problems. It provides a structured approach to dealing with missing data, a concept previously introduced in the precedent sections. This method involves the following steps:

i) substituting missing values with estimated values,

ii) estimating parameters,

iii) recalculating missing values based on the updated parameter estimates,

iv) reestimating parameters, and continuing this iterative process until convergence is reached.

Each iteration of EM consists of an expectation step (E step) and a maximization step (M step). The E step finds the conditional expectation of the "missing data" given the observed data and current estimated parameters, and then substitutes these expectations for the missing data.

Let $\theta^{(t)}$ be the current estimate of the parameter θ. The E step of EM finds the expected complete-data loglikelihood, evaluated at θ=$\theta^{(t)}$:

$$Q(\theta \mid \theta^{(t)}) = \int \ell(\theta \mid Y) f(\mathbf{Y}_{obs} \mid \mathbf{Y}_{mis}, \theta = \theta^{(t)})\, dY_{obs}.$$

The M step of EM determines $\theta^{(t+1)}$ by maximizing this expected complete-data loglikelihood:

$$Q(\theta^{(t+1)} \mid \theta^{(t)}) \geq Q(\theta \mid \theta^{(t)}), \text{ for all } \theta.$$

The new estimate $\theta^{(t+1)}$ then replaces $\theta^{(t)}$ in the next iteration. Also, under quite general conditions, EM converges to the maximum of this function. In particular, if a unique finite ML estimate of θ exists, EM will find it.[26]

Convergence Properties of EM

The distribution of the complete data Y can be factored as follows:

$$f(\mathbf{Y} \mid \theta) = f(\mathbf{Y}_{obs}, \mathbf{Y}_{mis} \mid \theta) = f(\mathbf{Y}_{obs} \mid \theta) f(\mathbf{Y}_{mis} \mid \mathbf{Y}_{obs}, \theta),$$

where $f(\mathbf{Y}_{obs} \mid \theta)$ is the density of the observed data $\mathbf{Y}_{obs}$ and $f(\mathbf{Y}_{mis} \mid \mathbf{Y}_{obs}, \theta)$ is the density of the missing data given the observed data. The corresponding decomposition of the log-likelihood is

$$\ell(\theta \mid \mathbf{Y}) = \ell(\theta \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}) = \ell(\theta \mid \mathbf{Y}_{obs}) + \ln f(\mathbf{Y}_{mis} \mid \mathbf{Y}_{obs}, \theta).$$

The objective is to estimate θ by maximizing the incomplete-data log-likelihood $\ell(\theta \mid \mathbf{Y}_{obs})$ with respect to θ for fixed observed $\mathbf{Y}_{obs}$; this task, however, can be difficult to accomplish directly.

First, write

$$\ell(\theta \mid \mathbf{Y}_{obs}) = \ell(\theta \mid \mathbf{Y}) - \ln f(\mathbf{Y}_{mis} \mid \mathbf{Y}_{mis}, \theta), \tag{1.13}$$

where $\ell(\theta \mid \mathbf{Y}_{mis})$ is the observed log-likelihood to be maximized, $\ell(\theta \mid \mathbf{Y})$ is the complete-data log-likelihood, which we assume is relatively easy to maximize, and $\ln f(\mathbf{Y}_{mis} \mid \mathbf{Y}_{obs}, \theta)$ can be viewed as the missing part of the complete-data log-likelihood.

The expectation of both sides of equation(2.14) over the distribution of the missing data $\mathbf{Y}_{mis}$, given the observed data $\mathbf{Y}_{obs}$ and a current estimate of θ, say $\theta^{(t)}$, is

$$\ell(\theta \mid \mathbf{Y}_{obs}) = Q(\theta \mid \theta^{(t)}) - H(\theta \mid \theta^{(t)}), \tag{1.14}$$

where

$$Q(\theta \mid \theta^{(t)}) = \int [\ell(\theta \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis})]\, f(\mathbf{Y}_{mis} \mid \mathbf{Y}_{obs}, \theta^{(t)})\, dY^{(1)}, \tag{1.15}$$

and

$$H(\theta \mid \theta^{(t)}) = \int [\ln f(\mathbf{Y}_{mis} \mid \mathbf{Y}_{obs}, \theta)]\, f(\mathbf{Y}_{mis} \mid \mathbf{Y}_{obs}, \theta^{(t)})\, d\mathbf{Y}_{mis}, \tag{1.16}$$

Note that

$$H(\theta \mid \theta^{(t)}) \leq H(\theta^{(t)} \mid \theta^{(t)}), \tag{1.17}$$

by Jensen's inequality (see Rao 1972[37]).

Let $\theta^{(0)}$ be an initial estimate of θ in its parameter space, may be based on a naive methods, for example, an estimate based on the completely case analysis, hot deck imutation, or mean imputation after the missing data $\mathbf{Y}_{mis}$ have been filled in by some approximations. Let $\theta(t)$ be the estimate at the tth iteration. Consider a sequence of iterates $(\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots)$. The difference in values of $\ell(\theta \mid \mathbf{Y}_{obs})$ at successive iterates is given by

$$\begin{aligned} \ell(\theta^{(t+1)} \mid \mathbf{Y}_{obs}) - \ell(\theta^{(t)} \mid \mathbf{Y}_{obs}) = &\left[Q(\theta^{(t+1)} \mid \theta^{(t)}) - Q(\theta^{(t)} \mid \theta^{(t)})\right] \\ &- \left[H(\theta^{(t+1)} \mid \theta^{(t)}) - H(\theta^{(t)} \mid \theta^{(t)})\right]. \end{aligned} \tag{1.18}$$

An EM algorithm chooses $\theta^{(t+1)}$ to maximize $Q(\theta \mid \theta^{(t)})$ with respect to θ. EM algorithm chooses $\theta^{(t+1)}$ so that $Q(\theta^{(t+1)} \mid \theta^{(t)})$ is greater than $Q(\theta^{(t)} \mid \theta^{(t)})$. Hence, the difference of Q functions in (1.18) is positive for EM algorithm. Furthermore, note that the difference in the H functions in (1.18) is negative by (1.17). Hence, for any EM algorithm, the change from $\theta^{(t)}$ to $\theta^{(t+1)}$ increases the log-likelihood.[26]

Example

The previous description of EM is conceptual in nature and omits many of the mathematical details of the procedure. This section expands the previous ideas and gives a more precise explanation of the E-step and the M-step. To illustrate the mechanics of EM, Enders 2010 [8] use a bivariate analysis example where one of the variables is incomplete. Throughout this section, He use Y_1 to denote the complete variable (e.g., IQ scores) and Y_2 to represent the incomplete variable (e.g., job performance ratings). This is a relatively simple estimation problem, but the basic ideas readily extend to multivariate analyses with general patterns of missing data.

With complete data, the following formulas generate the maximum likelihood estimates of the mean, the variance, and the covariance.

$$\widehat{\mu}_Y = \frac{1}{n}\sum Y \tag{1.19}$$

$$\widehat{\sigma}^2_Y = \frac{1}{n}\left(\sum Y^2 - \frac{(\sum Y)^2}{n}\right) \tag{1.20}$$

$$\widehat{\sigma}_{Y_1Y_2} = \frac{1}{n}\left(\sum Y_1Y_2 - \frac{\sum Y_1 \sum Y_2}{n}\right) \tag{1.21}$$

Notice that the sum of the scores (i.e., $\sum Y_1$ and $\sum Y_2$), the sum of the squared scores (i.e., $\sum Y_1^2$ and $\sum Y_2^2$), and the sum of the cross product terms (i.e., $\sum Y_1Y_2$) are the basic building blocks of the previous equations. Collectively, these quantities are known as sufficient statistics because they contain all of the necessary information to estimate the mean vector and the covariance matrix. As you will see, these sufficient statistics play an important role in the E-step.

The purpose of the E-step is to "fill in" the missing values so that the M-step can use Equations (1.19) through (1.21) to generate parameter estimates. More accurately, the E-step fills in each case's contribution to the sufficient statistics. The E-step uses the elements in the mean vector and the covariance matrix to build a set of regression equations that predict the incomplete variables from the observed variables. In a bivariate data set with missing value on Y_2, the necessary equations are:

$$\widehat{\beta}_{21} = \frac{\widehat{\sigma}_{12}}{\widehat{\sigma}_{11}} \tag{1.22}$$

$$\widehat{\beta}_0 = \widehat{\mu}_2 - \widehat{\beta}_{21}\widehat{\mu}_1 \tag{1.23}$$

$$\widehat{\sigma}_{22.1} = \widehat{\sigma}_{22} - \widehat{\beta}_{21}^2\widehat{\sigma}_{11} \tag{1.24}$$

$$\widehat{Y}_{i2} = \widehat{\beta}_0 + \widehat{\beta}_{21}Y_{i1} \tag{1.25}$$

where $\widehat{\beta}_0$ and $\widehat{\beta}_{21}$ are the intercept and slope coefficients, respectively, $\widehat{\sigma}_{22.1}$ is the residual variance from the regression of Y_2 on Y_1, and $\widehat{Y}_{i2}$ is the predicted Y_2 score for a given value of Y_1. The means, variances, and covariances that appear on the right side of the equations are elements from the mean vector and the covariance matrix.

The missing data complicate an otherwise straightforward analysis because the incomplete cases have nothing to contribute to $\sum Y_2$, $\sum Y_2^2$, and $\sum Y_1Y_2$. The E-step replaces the missing components of these sufficient statistics with their expected values (i.e., long-run averages). EM borrows information from other variables, so the algorithm actually uses so-called conditional expectations to replace the missing components of the formulas. To illustrate, consider the sum of the scores and the sum of the cross product terms (i.e., $\sum Y_2$ and $\sum Y_1Y_2$, respectively). The expected value of Y_2 is the predicted score from Equation (1.25), so the E-step replaces the missing components of $\sum Y$ and $\sum Y_1Y_2$ with $\widehat{Y}_{i2}$. Next, consider the sum of the squared scores, $\sum Y_2^2$. The expected value of a squared variable is $\widehat{Y}_{i2}^2 + \widehat{\sigma}_{22.1}$, where $\widehat{Y}_{i2}^2$ is the squared predicted score, and $\widehat{\sigma}_{22.1}$ is the residual variance from the regression of Y_2 on Y_1. The E-step replaces the missing components of $\sum Y_2^2$ with this expectation.

Notice that the E-step does not actually impute the raw data. Rather, it fills in the computational building blocks for the mean, the variance, and the covariance (i.e., the sufficient statistics). Once this process is complete, the M-step becomes a straightforward estimation problem that uses the filled-in sufficient statistics to compute Equations (1.19) through (1.21). The resulting parameter estimates carry forward to the next E-step, where the process begins anew.[8]

1.3.3 Comparison

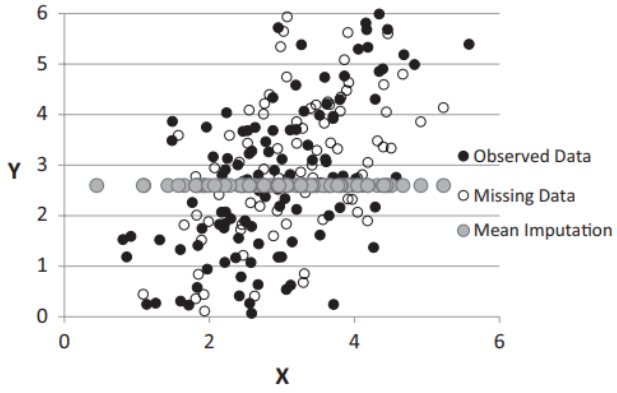
Single imputation techniques involve filling in each missing datum with a "good guess" as to what the missing datum should be. Fortunately, single imputation techniques are much less popular now than they once were. Common examples of single imputation are: (a) mean (across

persons) imputation: replacing each missing datum with the group mean for the corresponding variable, (b) hot deck imputation: replacing each missing datum with a value from a "donor" who has similar scores on other variables (which can be more error prone than listwise deletion , and (c) regression imputation: replacing each missing datum with a predicted value based on a multiple regression equation derived from observed cases.

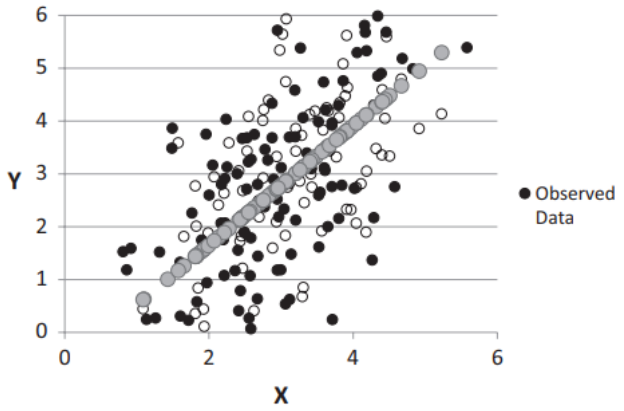
Single imputation suffers two major drawbacks. First, most single imputation techniques are biased under MCAR. For example, because mean imputation imputes a constant mean for each missing value (see Figure 1.4a[33]), the resulting sample estimates of the variance and the correlation will be downwardly biased—even if the missingness mechanism is completely random (MCAR). As another example, regression imputation leads to underestimation of the variance and overestimation of the correlation (because imputed values fall exactly on the regression line; see Figure 1.4b[33])—even if the missingness mechanism is MCAR! It is possible to improve regression imputation methods, however, by adding a random error term to the imputed values (i.e., stochastic regression imputation; see Figure 1.4c[33]). Stochastic regression imputation works to remove the missing data bias in regression imputation (described above) that previously underestimated the variance and overestimated the correlation (i.e., stochastic regression imputation is unbiased under both the MCAR and MAR missingness mechanisms). Nonetheless, even when considering stochastic regression imputation (which is unbiased under MAR), the researchers still do not recommend single imputation, for the following reason.

The second major drawback is that single imputation suffers the inability to calculate accurate SEs for hypothesis testing (i.e., there is usually no single value of n that corresponds well to all the parameter estimates). This problem is coupled with the real and common danger that many researchers tend to use the maximum n (treating the partially imputed data set as though it were a complete data set), which naturally leads to deflated SEs and creates Type I errors of inference (a.k.a., mirages, where incorrect hypotheses are falsely supported). As described in the following, multiple imputation solves this problem.

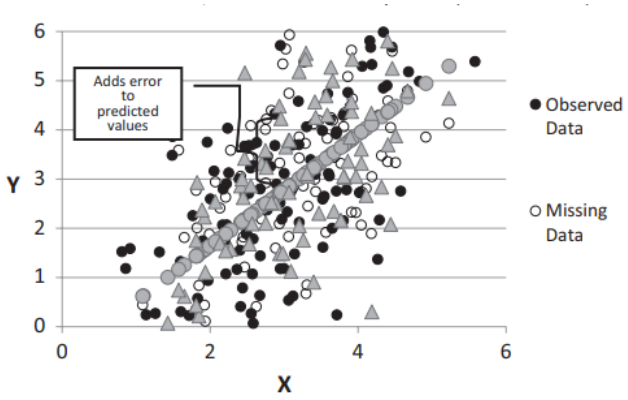
Overall, the main reason to place a moratorium on single imputation is because multiple imputation has all of the advantages of single imputation, but none of its major drawbacks. Thus, for typical data-analytic applications (e.g., involving correlation, regression), single imputation should be forbidden.[33]



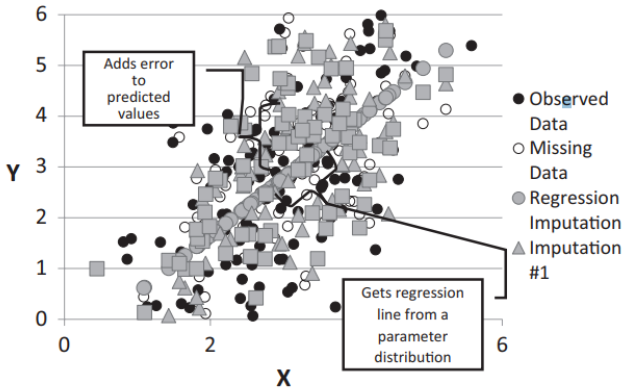
(a) Mean Imputaon:
Underesmated Variance and Correlation,Overesmated Sample Size (Inaccurate SEs)



(b) Regression Imputaon:
Underesmated Variance,Overesmated Correlaon and Sample Size (Inaccurate SEs)



(c) Stochasc Regression Imputaon:
Unbiased Variance and Correlation, Overesmated Sample Size (Inaccurate SEs)



(d) Mulple Imputaon:
Unbiased Variance and Correlation, Accurate SEs

Fig. 1.4: Missing Data Bias

Chapter 2

Non Parametric Test for Missing Completely At Random

In this chapter, I will introduce Little's test as a key tool for assessing MCAR assumption and CDM assumption, describe the methodology of Little's test and its interpretation.

2.1 statistical Methods

In this section we introduce some basic statistical methods, ANOVA test and likelyhood ratio statistic, which will be useful later.

2.1.1 Analyse of Variance (ANOVA)

Analysis of variance (ANOVA) is a statistical test for detecting differences in group means when there is one parametric dependent variable and one or more independent variables. The independent variables are called factors and the measured quantities are the dependent variables. For example, consider a clinical trial in which three different diagnostic imaging modalities are used on both men and women in different centers. The three elements used for classifications (center, sex, and treatment) identify the source of variation of each datum and are called factors. The individual classes of the classifications are the levels of the factor (e.g., the three different treatments T1, T2, and T3 are the three levels of the factor treatment). Male and female are the two levels of the factor sex, and center1, center2, center3 are the three levels of the factor center. A subset of the data present for a "combination" of one level of each factor under investigation is considered a cell of the data. In this example we have, three factors: center (3 levels), sex (2 levels), and treatment (3 levels). The specific type of ANOVA used is determined by the number of independent variables (factors) in the study, there are three types : one way ANOVA, two way ANOVA, and factorial ANOVA.

In this chapter we interest about One way ANOVA, which is the most simple form testing differences between three or more groups based on one independent variable. For example, comparing the sales performance of different stores in a retail chain.

ANOVA test relies on three main assumptions that must be met for the test results to be valid:

- Normally distributed data within each group: it follows that a fundamental assumption of parametric ANOVA is that each group of data (each level) be normally distributed.
- Homogeneity of variance within each group: referring again to the notion that ANOVA compares normal distribution curves of data sets, these curves need to be similar to each other in shape and width for the comparison to be valid. In other words, the amount of data dispersion (variance) needs to be similar between groups.

➤ Independent Observations: a general assumption of parametric analysis is that the value of each observation for each subject is independent of (i.e., not related to or influenced by) the value of any other observation.

One Way ANOVA

Suppose that y_{ij}, $i = 1, \ldots, n_i$, $j = 1, \ldots, p$ represents a random sample of i normal populations with means μ_1 , μ_2 ,$\mu_3, \ldots, \mu_p$ and common variance σ^2. The data point y_{ij} denotes the ith observation on the jth population of size n_j and its value assumed to follow a normal distribution: $y_{ij} \sim N(\mu_i, \sigma^2)$.

In ANOVA (Analysis Of Variance), the null hypothesis (H_0) states that there is no significant difference among the means of the groups being compared. In other words:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_p$$

The alternative hypothesis (H_1) in ANOVA (Analysis of Variance) states that there is a significant difference among the means of the groups being compared. In other words: at least one of the population means differs from the others.Symbolically, this can be represented as:

$$H_1 : \mu_1 \neq \mu_2 \neq \mu_3 \neq ... \neq \mu_p$$

The Mechanics of Calculating a One Way ANOVA

Let us suppose that there are p groups of scores where first group has n_1 scores, second has n_2 scores, and so on, and pth group has n_p scores. If y_{ij} represents the ith score in the jth group ($i = 1, 2, \ldots, n_i; j = 1, 2, \ldots, p$), then these scores can be shown as follows:

Here, $\mathbf{n} = n_1 + n_2 + \ldots n_p$, the total of all the scores

The total variability among the above-mentioned n scores can be attributed due to the variability between groups and variability within groups. Thus, the total variability can be broken into the following two components:

$$\text{Total variability } = \text{Variability between groups} + \text{Variability within} Groups$$

or

$$\text{TSS } = \text{SSB } + \text{ SSW}$$

This is known as one-way ANOVA model where it is assumed that the variability among the scores may be due to the groups. After developing the model, the significance of the group variability is tested by comparing the variability between groups with that of variability within groups by using the F-test. The null hypothesis which is being tested in this case is that whether variability between groups (SSB) and variability within the groups (SSW) are the same or not. If the null hypothesis is rejected, it is concluded that the variability due to groups is significant, and it is inferred that means of all the groups are not same. On the other hand, if the null hypothesis is not rejected, one may draw the inference that group means do not differ significantly. Thus, if **p** groups are required to be compared on some criterion variable, then the null hypothesis can be tested by following the below mentioned steps:

a) Hypothesis construction: The following null hypothesis is tested H_0 against the alternative hypothesis that at least one mean differs.

b) Level of significance: The level of significance may be chosen beforehand. Usually it is taken as 0.05 or 0.01.

c) Statistical test: The F-test is used to test the above mentioned hypothesis. If F-value is significant, it indicates that the variability between groups is significantly higher than the variability within groups; in that case, the null hypothesis of no difference between the group

means is rejected.
F-value is obtained by computing the sum of squares between groups (SSB), and sum of squares within groups (SSW):

➤ **Sum of squares between groups (SSB):** The sum of squares between groups can be defined as the variation of group around the grand mean of the data set. In other words, it is the measure of variation between the group means and is usually denoted by SSB. This is also known as the variation due to assignable causes. The sum of squares between groups is computed as

$$\text{SSB} = \sum_{j=1}^{p} n_j(\bar{y}_j - \bar{y})^2$$

Since **p** samples are involved in one way ANOVA, the degrees of freedom for between groups is $p-1$.
Thus, mean sum of squares for between groups MSB is obtained by dividing SSB by its degrees of freedom $k-1$, where :

$$\text{MSB} = \frac{\text{SSB}}{df_B} = \frac{\sum_{j=1}^{p} n_j(\bar{y}_j - \bar{y})^2}{p-1}$$

➤ **Sum of squares within groups (SSW):** The sum of squares within groups is the residual variation and is referred as variation due to non-assignable causes. It is the average variation within the groups and is usually denoted by SSW:

$$\text{SSW} = \sum_{j=1}^{p}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2 = \sum_{j=1}^{p}(n_j - 1)s_j^2$$

The degrees of freedom for the sum of squares within groups is given by $n-p$, and, therefore, mean sum of squares for within groups MSW is obtained by dividing SSW by $n-p$.

$$\text{MSW} = \frac{\text{SSW}}{df_W} = \frac{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2}{n-p}$$

➤ **ANOVA table:** After computing all sum of squares, these values are used in the analysis of variance(ANOVA) table 2.1 [46] for computing F-value as shown below.

➤ **F-statistic:** Under the normality assumptions, the F-value obtained in the above table, that is,

$$\text{F} = \frac{\text{MSB}}{\text{MSW}}$$

follows an F-distribution with $(p-1, n-p)$ degrees of freedom.This test statistic F is used to test the null hypothesis of no difference among the group means.

d) Decision criteria: The tabulated value of F at 0.05 and 0.01 level of significance with $(p-1, n-p)$ degrees of freedom may be obtained from F-distribution Tables. If calculated value of F is greater than tabulated F, the null hypothesis is rejected. And in that case it is concluded that at least one of the means will be different.

Tab. 2.1: ANOVA table

Sources of variation	SS	df	MS	F-value
Between groups	SSB	$p-1$	MSB=$\frac{\text{SSB}}{p-1}$	F= $\frac{\text{MSB}}{\text{MSW}}$
Within groups	SSW	$n-p$	MSW = $\frac{\text{SSW}}{n-p}$	
Total	TSS	$n-1$	$\frac{\text{TSS}}{n-1} = s^2$	

Application of One Way ANOVA

An audio company predicts that students learn more effectively with a constant low-tune melodious music in background, as opposed to an irregular loud orchestra or no music at all. To verify this hypothesis, a study was planned by dividing 30 students into three groups of ten each. Students were assigned to these three groups in a random fashion, and all of them were given a comprehension to read for 20 min. Students in group 1 were asked to study the comprehension with low-tune melodious music at a constant volume in the background. Whereas the students in group 2 were exposed to loud orchestra and group 3 to no music at all while reading the comprehension. After reading the comprehension, they were asked to solve few questions. The marks obtained are shown in the Table 2.2[46]. Do these data confirm that learning is more effective in particular background music Test your hypothesis at 5% level.
Solution Following steps shall be taken to test the required hypothesis:
a) Hypotheses construction: The researcher is interested in testing the following null hypothesis: $H_0 : \mu_{\text{Music}} = \mu_{\text{Orchestra}} = \mu_{\text{Without Music}}$ against the alternative hypothesis that at least one mean is different.
b) Level of significance: 0.05.
c) Statistical test: One-way ANOVA shall be used to test the null hypothesis.
In order to complete the ANOVA table, first, all the sum of squares are computed. We have:
Number of groups = $p = 3$
Sample size in each group$n_i = 10$
Total number of scores n=30
The computation of group total, group means, and grand total has to be computed first which is shown in Table 2.3[46]

➤ **Sum of squares between groups:**
SSB = $\sum_{j=1}^{p} n_j(\bar{y}_j - \bar{y})^2 = 10((6.4-4.5)+(4-4.5)+(3.1-4.5)) = 58.2$

➤ **Sum of square within groups:**
SSW = $\sum_{j=1}^{3}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_j)^2 = (8-6.4)^2+(4-6.4)^2+(8-6.4)^2+(6-6.4)^2+(6-6.4)^2+(7-6.4)^2+(3-6.4)^2+(7-6.4)^2+(9-6.4)^2+(6-6.4)^2+(4-4)^2+(6-4)^2+(3-4)^2+(4-4)^2+(3-4)^2+(8-4)^2+(3-4)^2+(2-4)^2+(4-4)^2+(3-4)^2+(3-3.1)^2+(4-3.1)^2+(6-3.1)^2+(2-3.1)^2+(1-3.1)^2+(2-3.1)^2+(6-3.1)^2+(4-3.1)^2+(1-3.1)^2+(2-3.1)^2 = 89.3$

d) Decision criteria: From F-distribution Tables, $\text{F}_{0.05}(2,27) = 4.22$.
Since calculated F(=8.79) $>$ $\text{F}_{0.05}(2,27)$, the null hypothesis may be rejected. It is therefore concluded that learning efficiency in all the three experimental groups is not same.

Tab. 2.2: Comprehension scores of the students under three different environments

Music	Orchestra	Without music
8	4	3
4	6	4
8	3	6
6	4	2
6	3	1
7	8	2
3	3	6
7	2	4
9	4	1
6	3	2

Tab. 2.3: ANOVA table for the data on comprehension test

Sources of variation	SS	df	MS	F-value
Between groups	58.2	$n-1=2$	29.5	8.79
Within groups	89.3	$n-p=27$	3.31	
Total	147.5	29		

2.1.2 Complete Data LRTs

In complete data sets, an LRT(Likelyhood Ratio Test) is performed by comparing two models with respect to the log-likelihood of the data, $l(\theta) = \log f(Y|\theta)$, given a set of model parameters θ. Specifically, the LRT is performed by calculating

$$d = 2[l(\widehat{\theta}_0 - l(\widehat{\theta})], \tag{2.1}$$

where $\widehat{\theta}$ denotes the estimated parameters under the full model, and $\widehat{\theta}_0$ denotes the estimated parameters under a restricted (or null) model. The LRT statistic (also known as the **deviance**) is typically compared with a distribution with k degrees of freedom, where k is the number of parameters being tested, that is, the number of restricted elements in θ_0.

2.2 Little's Test

In the first chapter, Rubin's (1976)[40] definition of MCAR requires that the probability of a value from data being missing is the same for all the observations. In orher words, the observed data are a simple random sample of the hypothetically complete data set. This implies that the cases with missing data belong to the same population (and thus share the same mean vector and covariance matrix) as the cases with complete data. One way to test for homogeneity of means is to separate the missing and the complete cases on a particular variable and examine group mean differences on other variables in the data set. Finding that the missing data patterns share a common mean vector and a common covariance matrix provides evidence that the data are MCAR, whereas group differences in the means or the covariances provide evidence that the data are not MCAR.

For example, suppose that a psychologist is studying quality of life in a group of cancer patients and finds that patients who refused the quality of life questionnaire have a higher average age and

a lower average education than the patients who completed the survey. These mean differences provide compelling evidence that the data are not MCAR and suggest a possible relationship between the demographic variables and the probability of missing data.[9]

Little (1988)[20] first proposed a test of MCAR for incomplete multivariate data by testing the homogeneity of means across different missing-pattern groups.

2.2.1 Notations

Introduce an index of notations to be used throughout this chapter and in subsequent sections. Let:"

- $\mathbf{Y} = y_{ij}$ be a $(n \times p)$ dimensional multivariate normal data with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, with part of the components in $\boldsymbol{y}_i$s missing.
- $\boldsymbol{m}_i$ be a $(1 \times p)$ vector of missing data indicators for case i.
- $J =$ number of distinct missing data patterns $\mathbf{m}_i$ in the data set. Fully observed cases, if present, count as a pattern.
- $\mathbf{S}_j$ be a set of cases with missing-data pattern $j(j = 1, ..., J)$.
- $\mathbf{r}_j$ number of cases in $\mathbf{S}_j$; $\sum \boldsymbol{r}_j = n$.
- p_j be the number of observed variables for cases in $\mathbf{S}_j$.
- $\mathbf{R}_j$ be a $(p \times p_j)$ matrix indicating which variables are observed for pattern j. The matrix has one column for each variable present, consisting of $p - 1$ Os and one 1 corresponding to the variable identified.
- $\mathbf{y}_{obs.j}$ be a $(1 \times p_j)$ vector of values of observed variables in case i.
- $\bar{\mathbf{y}}_{obs.j} = \boldsymbol{m}_j^{-1} \sum_{i \in \mathbf{S}_j} \mathbf{y}_{obs.i}$ is $(1 \times p_j)$ is the observed sample average for the jth missing pattern.
- $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}$, be the ML estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, assuming the $\mathbf{y}_i$ are iid normal and the missing-data mechanism is ignorable.
- $\tilde{\boldsymbol{\Sigma}} = n\hat{\boldsymbol{\Sigma}}/(n-1)$, be the ML estimate of $\boldsymbol{\Sigma}$ with a correction for degrees of freedom.
- Let $\boldsymbol{\mu}_{obs.j} = \boldsymbol{\mu}\mathbf{R}_j$ be a $(1 \times p_j)$-dimensional mean vector of only the observed components for jth missing pattern.
- $\boldsymbol{\Sigma}_{\boldsymbol{obs.j}} = \boldsymbol{R}_j^T \boldsymbol{\Sigma} \mathbf{R}_j$ be a $(p_j \times p_j)$ covariance matrix of only the observed components for jth missing pattern.

2.2.2 Little's MCAR Test

The Test Statistic, When $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are Known

To motivate the test statistic, Little first consider the (unrealistic) case When $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are Known. Assuming the missing data are ignorable. The Little's $\boldsymbol{\chi^2}$ test statistic for MCAR takes the following form:

$$d_0^2 = \sum_{j=1}^{J} \boldsymbol{r}_j (\bar{\boldsymbol{y}}_{obs.j} - \boldsymbol{\mu}_{obs.j}) \boldsymbol{\Sigma}_{obs.j}^{-1} (\bar{\boldsymbol{y}}_{obs.j} - \boldsymbol{\mu}_{obs.j})^T \tag{2.2}$$

if the data are MCAR, then conditional on the missing indicator $\boldsymbol{m}_i$, the following null hypothesis holds

$$H_0 : (\boldsymbol{y}_{obs.j}|\boldsymbol{m}_i) \sim N(\boldsymbol{\mu}_{obs.j}, \boldsymbol{\Sigma}_{obs.j}) \quad \text{if} \quad i \in \boldsymbol{S}_j, 1 \leq j \leq J \tag{2.3}$$

Instead, if (2.3) is not true, then conditional on the missing indicator $\boldsymbol{m}_i$, the means of the observed $\boldsymbol{y}$'s are expected to vary across different patterns, which implies

$$H_1 : (\boldsymbol{y}_{obs.j}|\boldsymbol{m}_i) \sim N(\boldsymbol{\nu}_{obs.j}, \boldsymbol{\Sigma}_{obs.j}) \quad \text{if} \quad i \in \boldsymbol{S}_j, 1 \leq j \leq J \tag{2.4}$$

where $\boldsymbol{\nu}_{obs.j}$, $j = 1, 2..., J$ are mean vectors of each pattern j and can be distinct. Rejecting (2.3) is sufficient for rejecting the MCAR assumption (1.1), but not necessary.

Little (1988)[20] proved that the statistic (2.2) is the likelihood ratio statistic for testing (2.3) against (2.4). If the normality assumption holds, then d_0^2 follows a chis-quared distribution with $f = \sum_{j=1}^{J} p_j - p$ df. If $\boldsymbol{y}_i$'s are not multivariate normal but has the same mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, then by the multivariate central limit theorem (see part (c) of the lemma in Little (1988)[20], under the null assumption of MCAR, d_0^2 follows a $\boldsymbol{\chi}^2$ distribution asymptotically.

The Test Statistic, When $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are Unknown

In practice, since $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are usually unknown, Little (1988)[20] proposed to replace $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in (2.2) with the unbiased estimators with the $\widehat{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$, $\tilde{\boldsymbol{\Sigma}}_{obs}$ it's a submatrix of $\widehat{\boldsymbol{\Sigma}}$, Which give

$$d_0^2 = \sum_{j=1}^{J} \boldsymbol{r}_j (\bar{\boldsymbol{y}}_{obs.j} - \widehat{\boldsymbol{\mu}}_{obs.j}) \tilde{\boldsymbol{\Sigma}}_{obs.j}^{-1} (\bar{\boldsymbol{y}}_{obs.j} - \widehat{\boldsymbol{\mu}}_{obs.j})^T \tag{2.5}$$

Asymptotically, d_0^2 follows a chis-quared distribution with $f = \sum_{j=1}^{\boldsymbol{J}} p_j - p$ df, and (2.3) is rejected if $d_0^2 > \chi^2_{\text{df}}(1-\alpha)$ where α is the significance level.$\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ can be obtained from EM algorithm using the observed data $\boldsymbol{y}_{obs}$ (Little and Rubin 1987 [25], for more details see section1.3.2).

The Test Statistic for Monotone Missing Data

The small sample null distribution of d_0^2 is extremely complex for a general pattern of missing data, but simplifies for particular missing-data patterns. Consider first the special case of $p = 2$ variables Y_1, and Y_2, where Y_1, is observed for all n cases and Y_2, is observed for $n_2 < n$ cases, say $i = 1, ..., n_2$. There are $J = 2$ patterns:
Pattern 1 denotes cases with Y_1, and Y_2, present and pattern 2 denotes cases with only Y_1 present. Then, $\mathbf{y}_{obs.i} = (y_{i1}, y_{i2})$ for $\mathbf{y}_{obs.i} = (y_{i1}, y_{i2})$ $i = 1, ..., n_2$ and $\mathbf{y}_{obs.i} = Y_{i1}$, for $i = n_2 + 1, ..., n$; $\bar{\mathbf{y}}_{obs.i} = (\bar{y}_1, \bar{y}_2)$, the sample means of Y_1, and Y_2, based on the first n_2 cases; $\bar{\mathbf{y}}_{obs.2} = \bar{y}_1^*$, the sample mean of Y_1, based on the last $n - n_2$ cases; $\tilde{\boldsymbol{\Sigma}}_{obs.1} = \tilde{\boldsymbol{\Sigma}}$; and $\tilde{\boldsymbol{\Sigma}}_{obs.2} = \tilde{\sigma}_1$. The covariance matrix for (Y_1, Y_2) is given by:

$$\tilde{\boldsymbol{\Sigma}} = \begin{pmatrix} \tilde{\sigma}_{11} & \tilde{\sigma}_{12} \\ \tilde{\sigma}_{21} & \tilde{\sigma}_{22} \end{pmatrix}$$

The inverse of the covariance matrix is:

$$\tilde{\boldsymbol{\Sigma}}^{-1} = \frac{1}{\det(\tilde{\boldsymbol{\Sigma}})} \begin{pmatrix} \tilde{\sigma}_{22} & -\tilde{\sigma}_{12} \\ -\tilde{\sigma}_{21} & \tilde{\sigma}_{11} \end{pmatrix}$$

where

$$\det(\tilde{\mathbf{\Sigma}}) = \tilde{\sigma}_{11}\tilde{\sigma}_{22} - \tilde{\sigma}_{12}\tilde{\sigma}_{21}$$

Thus (2.5) becomes

$$d_0^2 = n_2 \begin{pmatrix} \bar{y}_1 - \mu_1 \\ \bar{y}_2 - \mu_2 \end{pmatrix}^T \tilde{\mathbf{\Sigma}}^{-1} \begin{pmatrix} \bar{y}_1 - \mu_1 \\ \bar{y}_2 - \mu_2 \end{pmatrix} + (n - n_2)(\bar{y}_1^* - \hat{\mu}_1)^2/\tilde{\sigma}_{11}, \tag{2.6}$$

This simplifies to:

$$d_0^2 = \frac{n_2}{\det(\tilde{\mathbf{\Sigma}})}\left[\tilde{\sigma}_{22}(\bar{y}_1 - \hat{\mu}_1)^2 - 2\tilde{\sigma}_{12}(\bar{y}_1 - \hat{\mu}_1)(\bar{y}_2 - \hat{\mu}_2) + \tilde{\sigma}_{11}(\bar{y}_2 - \hat{\mu}_2)^2\right],$$

We have this from section 1.3.2:

$$\tilde{\sigma}_{22.1} = \tilde{\sigma}_{22} - \frac{\tilde{\sigma}_{12}\tilde{\sigma}_{21}}{\tilde{\sigma}_{11}}$$

and

$$\tilde{\beta}_{21.1} = \frac{\tilde{\sigma}_{21}}{\tilde{\sigma}_{11}}$$

Substitute $\tilde{\sigma}_{22.1}$ into the test statistic:

$$d_0^2 = \frac{n_2}{\tilde{\sigma}_{11}\tilde{\sigma}_{22.1}}\left[\tilde{\sigma}_{22}(\bar{y}_1 - \hat{\mu}_1)^2 - 2\tilde{\sigma}_{12}(\bar{y}_1 - \hat{\mu}_1)(\bar{y}_2 - \hat{\mu}_2) + \tilde{\sigma}_{11}(\bar{y}_2 - \hat{\mu}_2)^2\right] + \frac{(n - n_2)(\bar{y}_1^* - \hat{\mu}_1)^2}{\tilde{\sigma}_{11}},$$

which can written as:

$$d_0^2 = \frac{n_2}{\tilde{\sigma}_{11}\tilde{\sigma}_{22.1}}\left[(\tilde{\sigma}_{22.1} + \frac{\tilde{\sigma}_{12}^2}{\tilde{\sigma}_{11}})(\bar{y}_1 - \hat{\mu}_1)^2 - 2\tilde{\sigma}_{12}(\bar{y}_1 - \hat{\mu}_1)(\bar{y}_2 - \hat{\mu}_2) + \tilde{\sigma}_{11}(\bar{y}_2 - \hat{\mu}_2)^2\right] + \frac{(n - n_2)(\bar{y}_1^* - \hat{\mu}_1)^2}{\tilde{\sigma}_{11}},$$

$$d_0^2 = n_2\left[(\frac{(\bar{y}_1 - \hat{\mu}_1)^2}{\tilde{\sigma}_{11}}) + \left(\frac{\tilde{\sigma}_{12}^2}{\tilde{\sigma}_{11}^2\tilde{\sigma}_{22.1}}(\bar{y}_1 - \hat{\mu}_1)^2 - 2\tilde{\sigma}_{12}(\bar{y}_1 - \hat{\mu}_1)(\bar{y}_2 - \hat{\mu}_2) + \tilde{\sigma}_{11}(\bar{y}_2 - \hat{\mu}_2)^2\right)\right] + \frac{(n - n_2)(\bar{y}_1^* - \hat{\mu}_1)^2}{\tilde{\sigma}_{11}},$$

This can be rearranged to:

$$d_0^2 = n_2\left[\frac{(\bar{y}_1 - \hat{\mu}_1)^2}{\tilde{\sigma}_{11}} + \frac{(\bar{y}_2 - \hat{\mu}_2 - \tilde{\beta}_{21.1}(\bar{y}_1 - \hat{\mu}_1))^2}{\tilde{\sigma}_{22.1}}\right] + \frac{(n - n_2)(\bar{y}_1^* - \hat{\mu}_1)^2}{\tilde{\sigma}_{11}},$$

which can written as

$$\frac{n_2(\bar{y}_1 - \hat{\mu}_1)^2}{\tilde{\sigma}_{11}} + \frac{n_2[\bar{y}_2 - \hat{\mu}_2 - \widehat{\beta}_{21.1}(\bar{y}_1 - \hat{\mu}_1)]^2}{\tilde{\sigma}_{22.1}} + \frac{(n - n_2)(\bar{y}_1^* - \hat{\mu}_1)^2}{\tilde{\sigma}_{11}}, \tag{2.7}$$

Explicit ML estimates of the parameters are available for this problem (for more details see Anderson 1957 [1]):$\hat{\mu}_2 = \bar{y}_2 + \widehat{\beta}_{21.1}(\hat{\mu}_1 - \bar{y}_1$. Substituting these in (1.16) yields, after a little algebra,

$$d_0^2 = [n_2(\bar{y}_1 - \hat{\mu}_1)^2 + (n - n_2)(\bar{y}_1^* - \hat{\mu}_1)^2]/\tilde{\sigma}_{11}, \tag{2.8}$$

$$= \mathbf{SSB}_1/\mathbf{MST}_1 = (n - 1)F/(n - 2 + F). \tag{2.9}$$

where $\mathbf{SSB}_1$,$\mathbf{MST}_1$, and F are, respectively, the between groups sum of squares, the total mean square, and the F statistic from the analysis of variance (**ANOVA**) of Y_1 on the missing-data

pattern. Since there are two patterns here, $F = t^2$, where t is t statistic for comparing pattern. Hence the test based on d_0^2 is equivalent to the t test. Under the null hypothesis of MCAR and assuming that the values of Y_1 are normal, F has an F distribution with 1 and $n-2$ df. More generally, suppose that the data can be arranged in a monotone pattern, where variable Y_q, is more observed than Y_{q-1}, for $q = 1, ..., p-1$. Then, if n_q, is the number of cases for which Y_q is observed, $n = n_1 \geq n_2 \geq ... \geq n_p$. A generalization of the previous analysis yields

$$\begin{aligned} d^2 &= \mathbf{SSB}_1/\mathbf{MST}_1 + \mathbf{SSB}_{2.1}/\mathbf{MST}_{2.1} + ... \\ &+\mathbf{SSB}_{p-1.12...p-2}/\mathbf{MST}_{p-1.12...p-2} \\ &= \textstyle\sum_{q=1}^{p-1}(n_q - 1)(k_q - 1)F_{q.12...q-1} \\ &\div n_a - k_a + (k_a - 1)F_{a.12...a-1}, \end{aligned} \tag{2.10}$$

where n_q, is the number of cases with Y_q observed; k_q is the number of patterns with Y_q observed; $\mathbf{SSB}_1$,$\mathbf{MST}_1$, and F_1 are, respectively, the between-groups sum of squares, total mean square, and F statistic from the **ANOVA** of Y_1 on all k_2 patterns (see section 2.1.1).
$\mathbf{SSB}_{2.1}$,$\mathbf{MST}_{2.1}$, and $F_{2.1}$ are the between- groups sum of squares, total mean square, and F statistic from the analysis of covariance of Y_2 on all k_2 patterns with Y_2 observed, adjusting for Y_1 and the remaining terms are defined similarly. Under normality and MCAR, each of the contributions in (2.10) is independent, so the small sample null distribution of d^2 is a sum of functions of independent F statistics. In large samples, these functions become chi-squared distributed, and d^2 has the asymptotic chi-squared distribution discussed in section2.2.2 (d^2 like d_0^2 is chi-squared distribution with f df).[20]

2.2.3 Little's CDM Test

A natural extension of Little's test of MCAR is to test the CDM assumption (1.3) of $\boldsymbol{y}_i$ conditional on $\boldsymbol{x}_i$ when covariates $\boldsymbol{x}_i$'s are present. For simplicity, we assume that $\boldsymbol{x}_i$ contains the constant term 1 as one of its components. If $\boldsymbol{y}_i$ depends linearly on $\boldsymbol{x}_i$, then the model becomes

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{x} + \epsilon$$

where $\boldsymbol{B}$ is a $p \times q$ matrix of coefficients and $\epsilon \sim N(0, \boldsymbol{\Sigma})$. Under homoscedasticity assumption(homogeneity of covariances), $\boldsymbol{\Sigma}$ does not depend on $\boldsymbol{x}$. Compared with the model without covariates, we need to replace every unconditional mean of $\boldsymbol{y}$ with the conditional mean of $\boldsymbol{y}$ given $\boldsymbol{x}$, and test whether the coefficient matrix $\mathbf{B}$ varies among different missing patterns. The χ^2 test statistic (2.2) now becomes

$$\begin{aligned} d_0^2 &= \sum_{j=1}^{J}\sum_{i\in\mathbf{S}_j}(\tilde{\boldsymbol{B}}_{obs.j}\boldsymbol{x}_i - \boldsymbol{B}_{obs.j}\boldsymbol{x}_i)\boldsymbol{\Sigma}_{obs.j}^{-1}(\tilde{\boldsymbol{B}}_{obs.j}x_i - \boldsymbol{B}_{obs.j}x_i)^T \\ &= \sum_{j=1}^{J}\sum_{i\in\mathbf{S}_j}\boldsymbol{x}_i(\tilde{\boldsymbol{B}}_{obs.j} - \boldsymbol{B}_{obs.j})\boldsymbol{\Sigma}_{obs.j}^{-1}(\tilde{\boldsymbol{B}}_{obs.j} - \boldsymbol{B}_{obs.j})\boldsymbol{x}_i^T \end{aligned} \tag{2.11}$$

where $\boldsymbol{B}_{obs.j}$ is a $p_j \times q$ submatrix of $\boldsymbol{B}$, whose rows correspond to the jth missing pattern, and $\tilde{\boldsymbol{B}}_{obs.j}$ is the OLS (Ordinary Least Squares) estimator of $\boldsymbol{B}_{obs.j}$ using the observed data from pattern j. It is straightforward to see that d_0^2 in (2.2) is a special case of d_0^2 in (2.11) when $\boldsymbol{x}$ only contains the constant component 1. Accordingly, we are now testing the null hypothesis

$$H_0 : (\boldsymbol{y}_{obs.i}|\boldsymbol{m}_i, \boldsymbol{x}_i) \sim N(\boldsymbol{B}_{obs.j}\boldsymbol{x}_i, \boldsymbol{\Sigma}_{obs.j}) \quad \text{if} \quad i \in \boldsymbol{S}_j, 1 \leq j \leq J \tag{2.12}$$

versus

$$H_1 : (\boldsymbol{y}_{obs.i}|\boldsymbol{m}_i, \boldsymbol{x}_i) \sim N(\boldsymbol{D}_{obs.j}\boldsymbol{x}_i, \boldsymbol{\Sigma}_{obs.j}) \quad if \quad i \in \boldsymbol{S}_j, 1 \leq j \leq J \tag{2.13}$$

where under H_1, the CDM assumption does not hold and $\boldsymbol{y}_{obs.j} = \boldsymbol{D}_{obs.j}\boldsymbol{x} + \epsilon$ for pattern j, with $\boldsymbol{D}_{obs.j}$ potentially different among all patterns, but the error terms still sharing the same multivariate distribution $N(0, \boldsymbol{\Sigma})$.

In practice, we replace $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ in (2.11) with unbiased estimators $\widehat{\boldsymbol{B}}$ and $\tilde{\boldsymbol{\Sigma}} = n\hat{\boldsymbol{\Sigma}}/(n-q)$ where $\widehat{\boldsymbol{B}}$ and $\hat{\boldsymbol{\Sigma}}$ are the maximum likelihood estimators using all data under H_0, and calculate

$$d^2 = \sum_{j=1}^{J} \sum_{i \in \mathbf{S}_j} \boldsymbol{x}_i (\tilde{\boldsymbol{B}}_{obs.j} - \widehat{\boldsymbol{B}}_{obs.j}) \tilde{\boldsymbol{\Sigma}}_{obs.j}^{-1} (\tilde{\boldsymbol{B}}_{obs.j} - \widehat{\boldsymbol{B}}_{obs.j}) \boldsymbol{x}_i^T \tag{2.14}$$

which asymptotically follows $\boldsymbol{\chi}^2$ distribution with degrees of freedom $df = q(\sum_{j=1}^{J} p_j - p)$, and (2.12) is rejected if $d^2 > \boldsymbol{\chi}^2_{df}(1-\alpha)$ where α is the significance level. Again, when there are no covariates, and $\boldsymbol{x}$ only contains the constant component 1 with $q = 1$, then $df = \sum_{j=1}^{J} p_j - p$, which coincides with the degrees of freedom in the test of MCAR.[19]

2.2.4 Adjustment for Unequal Variances

A important limitation of d^2 in (2.5) and (2.14) is that the covariance matrix of observed $\boldsymbol{y}$ (or observed $\boldsymbol{y}$ conditional on $\boldsymbol{x}$) is still the same for all missing-value patterns even in the alternative hypotheses (2.4) and (2.13). This assumption may not be satisfied in general, especially when the number of missing patterns is large. Therefore, we can relax this limitation on covariance matrices and replace the alternative hypothesis (2.4) and (2.13) with (2.15) (2.16) respectively

$$H_1 : (\boldsymbol{y}_{obs.i}|\boldsymbol{m}_i) \sim N(\boldsymbol{\nu}_{obs.j}, \boldsymbol{\Gamma}_{obs.j}) \quad \text{if} \quad i \in \boldsymbol{S}_j, 1 \leq j \leq J \tag{2.15}$$

$$H_1 : (\boldsymbol{y}_{obs.i}|\boldsymbol{m}_i, \boldsymbol{x}_i) \sim N(\boldsymbol{D}_{obs.j}\boldsymbol{x}_i, \boldsymbol{\Gamma}_{obs.j}) \quad \text{if} \quad i \in \boldsymbol{S}_j, 1 \leq j \leq J \tag{2.16}$$

where the covariance matrices $\boldsymbol{\Gamma}_{obs.j}$, like the means, $\boldsymbol{\nu}_{obs.j}$, contain distinct parameters for each pattern j. To test (2.12) against (2.13) (or to test (2.3) against (2.4), we can derive the following likelihood ratio statistic as in Little (1988)[20]

$$d^2_{aug} = d^2 + \sum_{j=1}^{J} \boldsymbol{r}_j \{(\boldsymbol{S}_{obs.j} \hat{\boldsymbol{\Sigma}}_{obs.j}^{-1}) - p_j - \log|\boldsymbol{S}_{obs.j}| + \log|\hat{\boldsymbol{\Sigma}}_{obs.j}|\} \tag{2.17}$$

where d^2 is the same as in (2.5) without covariates or (2.14) with covariates, $\boldsymbol{S}_{obs.j}$ is the sample covariance matrix of the observations with pattern j in (2.15), and $\boldsymbol{S}_{obs.j}$ is the estimated covariance matrix of residuals from the regression of observed $\boldsymbol{y}_{obs.j}$ on $\boldsymbol{x}$ in pattern j in (2.16), and $\hat{\boldsymbol{\Sigma}}_{obs.j}$ is the same as in (2.5) and (2.14). aug stands for "augmented" since more parameters need to be estimated for covariance matrices in the new test. Asymptotically, d^2_{aug} follows χ^2 distribution with degrees of freedom

$$df = q(\textstyle\sum_{j=1}^{J} p_j - p) + \sum_{j=1}^{J} \dfrac{p_j(p_j - p)}{2} - \dfrac{p(p+1)}{2},$$

when there are no covariates, and $\boldsymbol{x}$ only contains the constant component 1 with $q = 1$, then,

$$df = (\textstyle\sum_{j=1}^{J} \dfrac{p_j(p_j + 3)}{2} - \dfrac{p(p+3)}{2},$$

and (2.3) or (2.12) is rejected if $d^2 > \boldsymbol{\chi}^2_{df(1-\alpha)}$ where α is the significance level. This augmented test using d^2_{aug} tends to have higher power than the test using d^2 for large sample sizes. On the other hand, d^2_{aug} may not be applicable if some patterns have too small sample sizes such that $\boldsymbol{r}_j < p_j + q$, since $\boldsymbol{S}_{obs.j}$ will then be singular; thus $\log|\boldsymbol{S}_{obs.j}|$ in the expression of d^2_{aug} cannot be computed.

Chapter 3

Simulation study

In this chapter, we delve into a comprehensive simulation study to evaluate the performance of Little's chi-square test for the Missing Completely at Random (MCAR) assumption and the Conditional Missing Data Mechanism (CDM) assumption using the `mcartest` command (it's available in SATA software).

3.1 The mcartest Command

3.1.1 Description

`mcartest` performs Little's chi-square test for the MCAR assumption, and accommodates arbitrary missing-value patterns. depvars contains a list of variables with missing values to be tested. depvars requires at least two variables. indepvars contains a list of covariates. When indepvars are specified, mcartest tests the CDM assumption for *depvars* conditional on *indepvars*. The test statistic uses multivariate normal estimates from the EM algorithm . The `unequal` option performs Little's augmented chi-square test which allows unequal variances between missing-value patterns. To install the Stata user-written program, mcartest (Li 2013), in Stata, type "search mcartest, all", click on "st0318", and then install or type:
net install st0318.pkg, replace.[19]

3.1.2 Syntax

Test for MCAR
`mcartest` *depvars* [*if*] [*in*] [, `noconstant unequal emoutput` *em_options*]
Test for CDM
`mcartest` *depvars*= *indepvars* [*if*] [*in*] [, `noconstant unequal emoutput` *em_options*]

3.1.3 Options

`noconstant` suppresses constant term.
`unequal` specifies to allow unequal variances between missing-value patterns.
By default, the test assumes equal variances between different missing value patterns.
emoutput specifies to display intermediate output from EM estimation.
em_options specifies the options in EM algorithm.

3.2 Application

Cheng Li(2013)[19] demonstrate the use of the mcartest command with an example. The fictional dataset in question comprises blood test results from a study on obesity, featuring 371 observations and 11 variables: cholesterol level, triglycerides level, diastolic blood pressure, systolic blood pressure, age, gender, height, weight, exercise time per week, alcohol consumption, and smoking habits. Let's focus on the first four variables, labeled as `chol`, `trig`, `diasbp`, and `sysbp`, while using the remaining seven as auxiliary variables, labeled as `age`, `female`, `height`, `weight`, `exercise`, `alcohol`, and `smoking`.The descriptions of these variables are presented in Table 3.1.

Tab. 3.1: Descriptions of the variables

Variable	Type	Description
chol	Continuous	**Cholesterol level**
trig	Continuous	**Triglycerides level**
diasbp	Continuous	**Diastolic blood pressure**
sysbp	Continuous	**Systolic blood pressure**
age	Categorical	**1 if 21-30, 2 if 31-40, 3 if 41-50, 4 if above 50**
female	Categorical	**1 if female, 0 if male**
height	Continuous	**Height in inches**
weight	Continuous	**Weight in lbs**
exercise	Discrete	**Exercise in hours per week**
alcohol	Categorical	**1 if drinking alcohol, 0 if not**
smoking	Categorical	**1 if smoking, 0 if not**

After loading the data, we can examine the patterns of missing values using the misstable command.

```
. use bloodtest
(fictional blood test data)
. misstable summarize
```

				Obs<.		
Variable	Obs=.	Obs>.	Obs<.	Unique values	Min	Max
chol	90		281	265	187.73	224.57
trig	70		301	280	103.22	136.21
diasbp	34		337	24	66	90
sysbp	73		298	32	106	138

```
. misstable pattern, freq
  Missing-value patterns
    (1 means complete)
```

	Pattern			
Frequency	1	2	3	4
122	1	1	1	1
72	1	1	1	0
70	1	0	1	1
55	1	1	0	1
34	0	1	1	1
18	1	1	0	0
371				

```
Variables are  (1) diasbp  (2) trig  (3) sysbp  (4) chol
```

The results suggest that the dataset contains missing values in the first four variables, but all the other variables are completely observed. 122 observations out of the 371 in total are complete, while over 2/3 of the observations contain missing values, with six missing-value patterns in total that are not monotone.

3.2.1 Little's MCAR Test for Bloodtest Data:

Now, let's determine if the data are missing completely at random (MCAR) using the mcartest command. Initially, we will not include any auxiliary variables in the analysis. Instead, we will apply Little's MCAR test to the variables `chol`, `trig`, `diasbp`, and `sysbp`. We will perform both the regular MCAR test and the test that accounts for unequal variances.

```
. mcartest chol trig diasbp sysbp, emoutput nolog
Expectation-maximization estimation      Number obs           =       371
                                         Number missing       =       267
                                         Number patterns      =         6
Prior: uniform                           Obs per pattern: min =        18
                                                          avg =  61.83333
                                                          max =       122
Observed log likelihood = -2623.2645 at iteration 17
```

	chol	trig	diasbp	sysbp
Coef				
_cons	206.2264	120.5829	78.8161	121.196
Sigma				
chol	41.91012	22.33289	3.762825	3.48862
trig	22.33289	42.08035	6.622086	10.69249
diasbp	3.762825	6.622086	18.45518	14.37273
sysbp	3.48862	10.69249	14.37273	35.92427

```
Little´s MCAR test
Number of obs       = 371
Chi-square distance = 25.7412
Degrees of freedom  = 14
Prob > chi-square   = 0.0279
```

We specified the `emoutput` option to display the EM estimates and also suppressed the log using the `nolog` option within the *em options*. If the EM algorithm does not converge, mcartest will generate a warning message in blue, similar to what mi impute mvn does. In this case, the EM algorithm has converged. The regular Little's MCAR test gives a χ^2 distance of 25.74 with 14 degrees of freedom and a p-value of 0.0279. This result provides evidence that the missing data in the four variables of interest are not MCAR at the 0.05 significance level.

We can also specify the `unequal` option to run the test with unequal variances.

```
. mcartest chol trig diasbp sysbp, unequal
Little´s MCAR test with unequal variances
Number of obs       = 371
Chi-square distance = 56.7101
Degrees of freedom  = 41
Prob > chi-square   = 0.0522
```

This test gives a χ^2 distance of 56.71 with 41 degrees of freedom and a p-value of 0.0522. The p-value is only slightly larger than 0.05, suggesting that although the evidence against MCAR is not strong, the power of the test could be relatively low. Both tests raise doubts about the MCAR assumption.

3.2.2 Little's CDM Test for Bloodtest Data:

To examine the CDM assumption, we incorporate auxiliary variables as covariates in the test. `age` is categorized into four brackets, and `female` has two categories, so we use the factor variables `i.age` and `i.female` in the test. Additionally, we specify the emoutput option to display the EM estimates of the linear regression coefficients.

```
. mcartest chol trig diasbp sysbp = weight i.age i.female, emoutput nolog
Expectation-maximization estimation      Number obs           =       371
                                         Number missing       =       267
                                         Number patterns      =         6
Prior: uniform                           Obs per pattern: min =        18
                                                          avg =  61.83333
                                                          max =       122

Observed log likelihood = -2477.8319 at iteration 24
```

	chol	trig	diasbp	sysbp
Coef				
weight	.0898433	.1155952	.0035606	.0315919
1b.age	0	0	0	0
2.age	-.0790635	-.598354	.0120911	-.6006885
3.age	-.3147961	-.6971391	-.4392923	-1.07614
4.age	-2.220313	-2.172395	.4254206	-.582046
0b.female	0	0	0	0
1.female	2.10565	-4.386112	-4.315367	-2.971464
_cons	191.5976	103.5614	79.32499	117.3274
Sigma				
chol	38.04902	15.04927	2.537881	1.435059
trig	15.04927	21.60197	-.5490975	1.695223
diasbp	2.537881	-.5490975	14.83308	10.89443
sysbp	1.435059	1.695223	10.89443	32.07185

```
Little´s CDM test

Number of obs       = 371
Chi-square distance = 89.4992
Degrees of freedom  = 84
Prob > chi-square   = 0.3204
```

This CDM test results in a χ^2 distance of 89.50 with 84 degrees of freedom and a p-value of 0.3204. For this dataset, incorporating age, female, and weight as covariates successfully passes the CDM test. The EM outputs in the table provide the EM estimates for the multivariate linear regression of `chol`, `trig`, `diasbp`, and `sysbp` on `weight`, age, and `female`, including the regression coefficients (Coef) and the covariance matrix of the errors (Sigma). For comparison, we also conduct the test using all seven auxiliary variables as covariates.

```
. mcartest chol trig diasbp sysbp = weight height exercise i.age i.female i.alcohol i.smoking

Little´s CDM test

Number of obs       = 371
Chi-square distance = 141.1465
Degrees of freedom  = 140
Prob > chi-square   = 0.4569
```

This CDM test results in a χ^2 distance of 141.15 with 140 degrees of freedom and a p-value of 0.4569. Both CDM tests are highly nonsignificant, suggesting that even though `chol`, `trig`, `diasbp`, and `sysbp` are not MCAR, the missing data mechanism can be reasonably considered CDM given the auxiliary variables age, female, and weight. Consequently, for this dataset, any analysis of `chol`, `trig`, `diasbp`, and `sysbp` using only the 122 completely observed samples without adjusting for the effect of the auxiliary variables is invalid because the MCAR assumption is violated. The means of these four variables differ significantly between the 122 completely observed samples and the samples with missing values. However, the plausible CDM assumption implies

that the means of these four variables change linearly with the auxiliary variables. For instance, the mean cholesterol level varies linearly with the subject's `weight`, `age`, and `gender`, as shown by the linear regression coefficients in the EM estimates. Since CDM is a special case of MAR, as mentioned in Section 2.1, this example also indicates that simple methods like complete case analysis may not be effective under the broader MAR assumption.[19]

3.3 Simulation Study

Cheng Li(2013)[19] evaluate the performance of Little's chi-square test of MCAR and CDM through simulation studies. In general, when the true missing data mechanism is MCAR, the empirical rejection probability of Little's test of MCAR fits well with the nominal significance level, with a stable performance even for small samples, different proportions of missing values, and different numbers of variables with missing values, as was found in Little (1988).

In this simulation, in order to evaluate the relative performance of the various test statistics under null conditions, a series of simulated experiments was conducted. In each of those experiments, 10000 samples were generated from a specific population under known conditions: (i) number of covariates for Little's CDM test, (ii) sample size, and iii) the mechanism that we have MCAR or MAR or MNAR . The various proposed Little's tests were then computed in each sample. The performance of these tests across all samples was then evaluated empirically and calculte the empirical rejection rate[1] for each scenario. This process was then repeated 10000 times under different conditions.

3.3.1 Little's MCAR Test Simulation

In this section we interest to compare the performance of Little's MCAR test statistic d^2 with that of the augmented test statistic d^2_{aug} when the covariance matrices vary among different missing-value patterns. Li (2013)[19] simulated the following simple model without covariates

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right\} \tag{3.1}$$

where y_2 remains complete across all observations, and y_1 is missing with a probability of 0.5 based on the mechanisms described below. We can compare both the rejection probabilities when the null hypothesis (2.3) or (2.12) is satisfied by the true model and the power of these tests when the null hypothesis is violated. The alternative hypothesis could be either (2.4) or (2.13), and will be examined in the 5 cases below. In the following, $\Phi(.)$ denotes the cumulative distribution function of the standard normal distribution, and $\Phi^{-1}(.)$ is its inverse.

1. (MCAR) y_1 is missing completely at random with probability 0.5.
2. (MAR) y_1 is missing if and only if $\Phi^{-1}(0.1) \leq y_2 \leq 0$ or $y_2 \geq \Phi^{-1}(0.9)$.
3. (MAR) y_1 is missing if and only if $|y_2| \geq \Phi^{-1}(0.75)$.
4. (MNAR) y_1 is missing if and only if $\Phi^{-1}(0.2) \leq y_1 \leq 0$ or $y_1 \geq \Phi^{-1}(0.8)$.
5. (MNAR) y_1 is missing if and only if $|y_1| \geq \Phi^{-1}(0.75)$.

Note that y_1 is missing with a probability of 0.5 in all 5 cases, resulting in 2 missing-value patterns. We always test the full vector of $y = (y_1, y_2)^T$. Consequently, the true missing data mechanism for Case 1 corresponds to MCAR. Cases 2 and 3 are MAR, while Cases 4 and 5 are

[1]The proportion of times that the null hypothesis is rejected when it is true.

MNAR. The covariance structures of the 2 missing-value patterns are the same in Cases 1, 2, and 4 due to symmetry, but differ in Cases 3 and 5. Under the null hypothesis (2.3), d^2 in (2.5) asymptotically follows a χ^2 distribution with 1 degree of freedom, and d^2_{aug} in (2.17) asymptotically follows a χ^2 distribution with 2 degrees of freedom. The empirical rejection rates of both tests at a significance level of $\alpha = 0.05$ are reported using sample sizes of 100, 250, 500, and 1000, based on 10,000 Monte Carlo replications for each of the 5 missing data mechanisms. The results are summarized in Table 3.2[19].

Tab. 3.2: Empirical rejection rates when $\alpha = 0.05$ for d^2 and d^2_{aug}

Missingness of y_1	**Test stat**	**Sample size**			
		100	250	500	1000
Case 1 (MCAR)	d^2	0.051	0.043	0.050	0.048
	d^2_{aug}	0.053	0.048	0.050	0.050
Case 2 (MAR)	d^2	0.182	0.346	0.566	0.851
	d^2_{aug}	0.184	0.303	0.490	0.780
Case 3 (MAR)	d^2	0.052	0.051	0.051	0.050
	d^2_{aug}	1.000	1.000	1.000	1.000
Case 4 (MNAR)	d^2	0.363	0.728	0.953	0.999
	d^2_{aug}	0.292	0.626	0.916	0.998
Case 5 (MNAR)	d^2	0.050	0.053	0.048	0.052
	d^2_{aug}	0.261	0.572	0.882	0.996

We can compare the results of d^2 and d^2_{aug} in Table 3. In Case 1, where y_1 is MCAR, the empirical rejection rates for both d^2 and d^2_{aug} are close to the nominal level. In Case 2 (MAR) and Case 4 (MNAR), both tests also perform similarly, although d^2 seems to have slightly higher power than d^2_{aug}. This is expected because, in the true model, the covariance matrices of the two missing patterns are exactly the same, and d^2_{aug} is less efficient as it estimates two covariance matrices separately.

However, in Case 3 (MAR), where y_1 is missing if $|y_2| \geq \Phi^{-1}(0.75)$, or in Case 5 (MNAR), where y_1 is missing if $|y_1| \geq \Phi^{-1}(0.75)$, the missing data and the observed data have the same mean (zero) but different variances. As a result, the empirical rejection rates for d^2 are very low, indicating weak power of Little's test in these situations. The power of d^2 does not improve significantly even with a sample size of 1000. Conversely, after adjusting for unequal variances, d^2_{aug} shows much higher power, increasing to 1 as the sample size grows from 100 to 1000. This implies that d^2 may not be reliable when differences between missing-value patterns do not lie in their means, whereas d^2_{aug} can overcome this weakness when the covariance structure varies significantly across different missing-value patterns.

Although the augmented test for unequal variances demonstrates better power in some situations, such as Case 3 and Case 5, it may be too conservative with small sample sizes and complex missing-value patterns. In extreme cases, according to formula (2.17), d^2_{aug} cannot be computed when some missing-value patterns contain too few observations.

In the following, Cheng Li(2013)[19] simulate the same sample from Little (1988)[20] and compare the finite sample performance of d^2 and d^2_{aug} with more complicated missing-value patterns. Little (1988)[20] considered a multivariate normal model with 4 variables $y = (y_1, y_2, y_3, y_4)^T$,

generated by

$$
\begin{aligned}
y_1 &= z_1 \\
y_2 &= z_1\sqrt{0.9} + z_2\sqrt{0.1} \\
y_3 &= z_1\sqrt{0.2} + z_2\sqrt{0.1} + z_3\sqrt{0.7} \\
y_4 &= z_1\sqrt{0.6} + z_2\sqrt{0.25} + z_3\sqrt{0.1} + z_4\sqrt{0.05}
\end{aligned}
$$

where z_1, z_2, z_3, z_4 are independent standard normal random variables. We only observe y_1, y_2, y_3, y_4 but not z_1, z_2, z_3, z_4, and the missing data mechanism of y_1, y_2, y_3, y_4 is MCAR. For $y = (y1, y2, y3, y4)^T$, Little (1988)[**?**] considered 7 missing-value patterns in total, which can be represented by the missing indicator vector $\boldsymbol{r} = 1111, 1110, 1100, 1101, 1001, 1011, 1010$. For example, $\boldsymbol{r} = 1110$ means that y_1, y_2, y_3 are observed and y_4 is missing. The proportions of the 7 missing-value patterns in the sample are 0.4, 0.1, 0.1, 0.1, 0.1, 0.1 and 0.1 respectively. We examine the empirical rejection rates of d^2 and d_a^2ug with the sample size ranging from 100, 250, 500, 1000 to 2000, based on 10,000 Monte Carlo replications. The results are summarized in Table 3.3[19] and the Monte Carlo standard errors are displayed in the parentheses. The results are summarized in Table 3.3: Given these 7 missing-value patterns, the chi-square degrees of freedom

Tab. 3.3: Empirical rejection rates when $\alpha = 0.05$ for d^2 and d^2_{aug}

Test stat	**Sample Size**				
	100	**250**	**500**	**1000**	**2000**
d^2	0.043	0.047	0.054	0.051	0.049
d^2_{aug}	0.213	0.096	0.070	0.060	0.053

for d^2 and d^2_{aug} are 15 and 42 respectively.

The results in Table 4 suggest that with too many parameters in the covariance matrices to estimate, the empirical rejection rates for d^2_{aug} are too conservative and only get close to the nominal level 0.05 when the sample size is 2000. In comparison, d^2 has already achieved acceptable accuracy when the sample size is 250. This implies that d^2_{aug} not perform as well as d^2 in small samples when the missing-value patterns become more complicated. Moreover, as pointed out in Little (1988)[20], d^2_{aug} may be sensitive to departure from the normality assumption as d^2_{aug} involves the comparison of variances, while simulation results in Little (1988)[20] suggest that d^2 is relatively robust to non-normality of the data. Therefore the augmented test works best for nearly multivariate normal data when the covariance structure differs significantly among missing-value patterns and a sufficient number of observations are available in each pattern.

3.3.2 Little's CDM Test Simulation

For Little's test of CDM, the natural extension of MCAR test, it remains unclear whether increasing the number of covariates has an impact on its finite sample performance. Cheng Li(2013)[19] explored this by simulating the following model

$$
\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = B \begin{pmatrix} x_1 \\ x_2 \\ . \\ . \\ x_q \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \tag{3.2}
$$

where B is a $p \times (q)$ matrix of all 1's, $x_1, x_2, ..., x_{q-1}$ are independent N(0, 1) variables, and the error terms follow

$$\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \sim N \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right\} \tag{3.3}$$

y_1 is missing completely at random with probability 0.5 and y_2 is always completely observed, yielding 2 missing value patterns. $y = (y_1, y_2)^T$ is tested for CDM with auxiliary variables (covariates) $x = (x_1, ..., x_{q-1})^T$. The number of covariates $q-1$ (constant term not included) varies among 0, 1, 2, 5, 10, and 20, and the sample size increases from 100, 250, 500 to 1000. For each scenario, 10,000 Monte Carlo replications are used. Under the null hypothesis (2.12), d^2 in (2.14) asymptotically follows χ^2 distribution with df $= q$. At significance level $\alpha = 0.05$, Li(2013)[19] report the empirical rejection probability of the CDM test in Table 3.4[19].

Tab. 3.4: Empirical rejection rates of the CDM test with $\alpha = 0.05$

Covariates	$\boldsymbol{\chi^2}$	**Sample Size**			
		100	**250**	**500**	**1000**
0	**1**	0.051	0.043	0.050	0.048
1	**2**	0.051	0.052	0.050	0.052
2	**3**	0.044	0.049	0.049	0.048
5	**6**	0.045	0.049	0.050	0.051
10	**11**	0.036	0.045	0.046	0.047
20	**21**	0.023	0.039	0.045	0.046

Table 3.4 illustrates that in this model, with a small number of covariates, the empirical rejection rate of Little's CDM test is close to the nominal level of 0.05 with sample sizes of 100 or 250. However, as the number of covariates increases to 10 and 20, the empirical rejection rate falls significantly below the nominal level of 0.05 for these smaller sample sizes. Therefore, in small samples, the CDM test becomes more conservative as the number of covariates increases.

Conclusion

In this work, firstly, we present a comprehensive introduction to the problem of missing data, I outlined the primary mechanisms through which data can be missing and the different patterns, I explore both conventional and novel methods for handling missing data, examining their strengths, limitations, and practical implications in statistical analysis. By understanding the characteristics and applications of these methods.

We have discussed Little's test for MCAR for multivariate quantitative data proposed by Little (1988), which tests whether there exists significant difference between the means of different missing-value patterns. The test statistic takes a similar form to the likelihood ratio statistic for multivariate normal data and is asymptotically chi-square distributed under the null hypothesis that there are no differences between the means of different missing-value patterns. Rejection of the null provides sufficient evidence to indicate that the data are not MCAR. This indication is vital for analysts because it guides them in choosing the right methods for dealing with missing data. If data are not MCAR, methods like multiple imputation or ML approaches may be more appropriate than simple techniques like Complete Case Analysis.

Also, We present Li's mcartest command 2014 [19] that implements Little's chisquare test of the MCAR assumption or the CDM assumption. The methodology is mainly based on Little (1988) and can be extended to testing the CDM assumption when covariates are included in the test. The command also allows adjustment for unequal variances via the unequal option. We demonstrated how to use this command and the caveats of choosing covariates through an example. Finally I examined the performance of the MCAR/CDM test, compared the strengths and weaknesses of the regular test and the test with unequal variances by simulation and provided some suggestions for how to use them in practice.

Bibliography

[1] Anderson, T. W. (1957). Maximum Likelihood Estimates for a Multivariate Normal Distribution when Some Observations are Missing. *Journal of the American Statistical Association*, 52(278), 200-203.

[2] Allison, P. D. (2010). Missing Data. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research* (2nd ed). Emerald Group Publishing Limited.

[3] Brown, M. L., & Kros, J. F. (2003). Data mining and the impact of missing data. *Industrial Management & Data Systems*, 103(8), 611–621. doi:10.1108/02635570310497657.

[4] Day, S. (1999). *Dictionary for clinical trials*. Hoboken: John Wiley & Sons.

[5] Dempster, A., Laird, N. & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, Vol. B39, pp. 1-38.

[6] Ellenberg, J. (2014). *How Not to Be Wrong: The Power of Mathematical Thinking*. Penguin UK: Westminster.

[7] Ellenberg, J. (2018). *L'art de ne pas dire n'importe quoi : ce que le bon sens doit aux mathématiques*. Gillingham: Cassini.

[8] Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

[9] Enders, C. K. (2022). *Applied missing data analysis*. Guilford Publications.

[10] Fisher, R. A. (1934). Effect of methods of ascertainment upon the estimation of frequencies. *Annals of Human Genetics*, 6, 13–25.

[11] Galton, F. (1888). Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society*, 45, 135–145.

[12] Galton, F. (1898). An examination into the registered speeds of American trotting horses, with remarks on their value as hereditary data. *Proceedings of the Royal Society of London*, 62, 310–315.

[13] Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology*, 60(1), 549–576. doi:10.1146/annurev.psych.58.110405.085530.

[14] Graham, J., Hofer, S., Donaldson, S., MacKinnon, D. and Schafer, J. (1997), "Analysis with missing data in prevention research", in Bryant, K., Windle, W. and West, S. (Eds), New Methodological Approaches to Alcohol Prevention Research, American Psychological Association, Washington, DC.

[15] Hansen, M. H., Hurwitz, W. N. & Madow, W. G. (1953). *Sample Survey Methods and Theory, Volume 1*. New York: Wiley.

[16] He, Y., Zhang, G. & Hsu, C. (2022). *Multiple imputation of missing data in practice: Basic theory and analysis strategies*. New York: John Wiley and Sons.

[17] Kenett, R. S., & Salini, S. (2011). *Modern analysis of customer surveys: With applications using R*. Wiley. DOI:10.1002/9781119961154.

[18] Kuhrt, A. (1995). *The Ancient Near East (Routledge History of the Ancient World) c. 3000–330 B.C.E.* London: Routledge.

[19] Li, C. (2013). Little's Test of Missing Completely at Random. Northwestern University, Evanston, IL. Retrieved from .

[20] Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83, 1198–1202.

[21] Little, R. J. A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*, 87, 1227-1237.

[22] Little, R. J. A. (1995). Modeling the Drop-Out Mechanism in Repeated-Measures Studies. *Journal of the American Statistical Association*, 90(431), 1112–1121. https://doi.org/10.1080/01621459.1995.10476615.

[23] Little, R. J. A. & Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods and Research*, Vol. 18, pp. 292-326.

[24] Little, R. J. A. & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd edn.). New York: Wiley.

[25] Little, R. J. A., & Rubin, D. B. (1987). *Statistical Analysis With Missing Data*. New York: John Wiley.

[26] Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

[27] Little, R. J. A., & Schenker, N. (1995). Missing Data. In: Arminger, G., Clogg, C. C., & Sobel, M. E. *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. https://doi.org/10.1007/978-1-4899-1292-3.

[28] Lord, F. M. (1955). Estimation of parameters from incomplete data. *Journal of the American Statistical Association*, 271, 870–876.

[29] Mangel, M. & Samaniego, F. J. (1984). Abraham wald's work on aircraft survivability. *Journal of the American Statistical Association*, 386, 259–267.

[30] McKendrick, A. (1926). Applications of mathematics to medical problems. *Proceeding of the Edinburgh Mathematical Society*, 44, 98–130.

[31] Myers, T. A. (2011). Goodbye, Listwise Deletion: Presenting Hot Deck Imputation as an Easy and Effective Tool for Handling Missing Data, *Communication Methods and Measures*, 5(4), 297-310. DOI: 10.1080/19312458.2011.624490.

[32] Mukherjee, R. & Rao, C. R. (1955). The ancient inhabitants of Jebel Moya, Sudan, Volume 123. England: Cambridge University Press.

[33] Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods*, 17(4), 372-411. https://doi.org/10.1177/1094428114548590.

[34] Orchard, T. & Woodbury, M. (1972). A missing information principle: theory and applications. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 697-715.

[35] Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., & Petersen, I. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical Epidemiology*, 9, 157–166. https://doi.org/10.2147/CLEP.S129785.

[36] Pigott, T. D. (2001). A Review of Methods for Missing Data, *Educational Research and Evaluation: An International Journal on Theory and Practice*, 7(4), 353-383. http://dx.doi.org/10.1076/edre.7.4.353.8937.

[37] Rao, C. R. (1972). *Linear Statistical Inference and Its Applications*. New York: Wiley.

[38] Rao, C. R. (1985). Weighted Distributions Arising Out of Methods of Ascertainment: What Population Does a Sample Represent? In: Atkinson A.C., Fienberg S.E. (eds). New York: A Celebration of Statistics. Springer.

[39] Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, 47(3), 537–570.

[40] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592.

[41] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York, NY.

[42] Rubin, D. B. (1996). Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, 91(434), 473–489. https://doi.org/10.2307/2291635.

[43] Siddique, J., & Belin, T. R. (2008). Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in Medicine*, 27(1), 83–102. https://doi.org/10.1002/sim.2999.

[44] Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16, 219–242.

[45] Van Buuren, S. (2018). *Flexible imputation of missing data* (Second edition). New York: Chapman and Hall.

[46] Verma, J.P. (2013). One-Way ANOVA: Comparing Means of More than Two Samples. In: Data Analysis in Management with SPSS Software. Springer, India.

[47] Wainer, H. (2011). *Uneducated Guesses: Using Evidence to Uncover Misguided Education Policies*. Princeton: Princeton University Press.

[48] Wilks, S. S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. *The Annals of Mathematical Statistics*, 3, 163–195.