

Efficacy of ChatGPT as a missing data analysis problem-solving tool

Purpose

This study aims to evaluate the performance of ChatGPT (GPT-4.0) as a study tool for solving problems related to missing data analysis, specifically by applying it to a homework assignment on the topic.

1. Consider the following hypothetical data from a clinical trial for factors related to hypertension

ID:	1	2	3	4	5	6	7	8
Age:	23	35	31	35	28	37	51	41
SBP:	122	123	120	123	124	128	NA	NA
BMI:	21	19	17	22	22	25	36	41

2. Fit a linear regression of age and bmi on SBP with above data, what do you find?
3. Formulate the above data in the way of Y_{ij} , M_{ij} , and p_{ij} , and describe the role of each notation in the definition of MCAR.
4. Run Little's MCAR test on the missing data, can you claim MCAR at 0.05 significance level for variables underlying the data? Explain
5. Can you use the mean imputation method to impute the missing values? Explain
6. Print the data and examine the pattern of missing observation in the data, what do you find? (consider the rate of missing for people with age older than 40 and BMI higher than 30).
7. If you group the patients into two strata based on the observed data of BMI > 30 or not, within each stratum, what do you see for the missing rate?
8. Referring to Q7, if the stratification criterion is age > 40 instead of BMI > 30, what do you see?

The questions above are homework problems from my missing data analysis class. I am interested in evaluating how effectively ChatGPT (GPT-4.0) can assist in solving them.

I will present both ChatGPT's solutions and my own solutions to these problems for comparison and analysis.

The First Response from ChatGPT (GPT-4.0)

This refers to the initial output generated by ChatGPT (GPT-4.0) in response to my missing data analysis questions.

It seems that `test_mcar` from `statsmodels.imputation.mice` is not available in this environment. However, we can proceed by answering parts of your questions based on the available computations and data:

1. Data Overview

We constructed the dataset as follows:

ID	Age	SBP	BMI
1	23	122	21
2	35	123	19
3	31	120	17
4	35	123	22
5	28	124	22
6	37	128	25
7	51	NA	36
8	41	NA	41

2. Linear Regression (SBP ~ Age + BMI)

We fitted a linear regression model using complete cases (rows 1 to 6). The regression results (summary shown in the output) indicate how age and BMI relate to systolic blood pressure. Since the number of observations is small, estimates may not be highly reliable.

3. Notation for MCAR

We define the following notation for missing data mechanisms:

- Y_{ij} : The value of variable j for individual i (e.g., SBP)

- M_{ij} : A missingness indicator, where $M_{ij} = 1$ if Y_{ij} is missing, otherwise $M_{ij} = 0$
- $p_{ij} = P(M_{ij} = 1 \mid Y_{ij}, X_i)$: The probability that the data is missing, conditional on observed and unobserved data

MCAR (Missing Completely at Random) means that p_{ij} does **not** depend on either observed or unobserved data. That is, the probability of missingness is unrelated to the data values themselves.

4. Little's MCAR Test

We attempted to run Little's MCAR test, but the function `test_mcar` is unavailable in this environment. In a local R or Python environment with appropriate packages (e.g., `BaylорEdPsych` in R), you can perform this test.

If the p-value is greater than 0.05, we **fail to reject** the null hypothesis of MCAR, suggesting the data may be missing completely at random.

5. Mean Imputation?

Yes, mean imputation is technically possible, but **not recommended** unless missingness is MCAR. Otherwise, it may:

6. Missing Pattern

From the data:

- **Age > 40**: IDs 7 and 8 → Both have missing SBP → **100% missing**
- **BMI > 30**: IDs 7 and 8 → Again, both have missing SBP → **100% missing**

This suggests a **patterned missingness**, which is likely **not MCAR**.

7. Stratification by BMI > 30

- **BMI > 30 (TRUE)**: 2 observations → Both missing SBP → **100% missing**
- **BMI ≤ 30 (FALSE)**: 6 observations → None missing → **0% missing**

⇒ Strong evidence that SBP missingness is related to **BMI**.

8. Stratification by Age > 40

- **Age > 40 (TRUE)**: 2 observations → Both missing SBP → **100% missing**
- **Age ≤ 40 (FALSE)**: 6 observations → None missing → **0% missing**

⇒ Again, strong indication of **non-random missingness** related to **Age**.

MY comments

Missing Data Test Not Actually Done. ChatGPT stated that it couldn't run the MCAR (Missing Completely at Random) test due to the unavailability of the required function. Although it attempted to explain each question, the explanations lacked sufficient depth for me to consider the task fully completed

Wrong R packages suggested. ChatGPT suggested some R packages for my analysis, but these packages do not exist in the current R environment. Upon further investigation, I found that they were removed from **CRAN** as far back as 2012.

No Justification for Excluding Missing Data. In Question 2, ChatGPT did not clearly explain why the missing data should be excluded. It suggested running the linear regression on the complete cases (rows 1 to 6), which implied excluding rows 7 and 8. Although this approach yields a valid result, it caused me to overlook an important feature of the `lm()` function in R—namely, that it automatically handles missing values by excluding them. ChatGPT did not mention this behavior or justify the exclusion, which could lead to misunderstandings about how missing data is treated in regression analysis.

Contradictory Conclusions About Missingness. ChatGPT initially stated that MCAR could not be tested, yet later concluded that the missingness is likely not MCAR based on observed patterns. This is a form of circular reasoning—you cannot determine whether data is not MCAR without conducting a proper statistical test. Why it matters: If the data is truly MCAR, simple methods like mean imputation may be acceptable. However, if the data is not MCAR—as ChatGPT suspects—more robust imputation methods would be required. Without a valid test, the analysis remains inconclusive and potentially misleading.

Mean Imputation Suggestion is Problematic. ChatGPT mentions that mean imputation is “possible but not recommended,” but fails to offer any viable alternatives. In reality, mean imputation is generally a poor choice because it distorts relationships between variables and reduces variability in the data. Why it matters: For instance, if systolic blood pressure (SBP) is missing primarily for older individuals, imputing with the overall mean would likely underestimate their true SBP, introducing bias and compromising the validity of the analysis.

No Practical Alternatives Provided. ChatGPT’s analysis highlights some issues with handling missing data but fails to offer practical solutions. It does not suggest more appropriate methods such as multiple imputation or explicitly modeling the missingness mechanism, both of which are essential for rigorous missing data analysis.

Next I asked ChatGPT to help me implement this suggestions in R and the following were the results

1. Enter the data

```
# Load necessary packages
library(dplyr)
#library(BayLorEdPsych) # For LittleMCAR test
library(mice)           # For pattern analysis
```

```
## Warning: package 'mice' was built under R version 4.4.3
```

```
##
## Attaching package: 'mice'
```

```
## The following object is masked from 'package:stats':
##
## filter
```

```
## The following objects are masked from 'package:base':  
##  
##      cbind, rbind
```

```
# Create the data frame  
df <- data.frame(  
  ID = 1:8,  
  Age = c(23, 35, 31, 35, 28, 37, 51, 41),  
  SBP = c(122, 123, 120, 123, 124, 128, NA, NA),  
  BMI = c(21, 19, 17, 22, 22, 25, 36, 41)  
)
```

Comment

This part was correctly addressed in concept; however, ChatGPT referenced a non-existent **R package named BaylorEdPsych**. Since this package is not available, the code could not be executed as written, which prevented this section from running correctly.

2. Fit linear regression: $SBP \sim Age + BMI$

GhatGPT Approach

```
model <- lm(SBP ~ Age + BMI, data = df, na.action = na.omit)  
summary(model)
```

```
##
## Call:
## lm(formula = SBP ~ Age + BMI, data = df, na.action = na.omit)
##
## Residuals:
##      1      2      3      4      5      6
## -0.09361  0.76254 -0.04772 -1.64698  0.37397  0.65180
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 101.87244    4.46807   22.800 0.000185 ***
## Age          0.14585    0.09889    1.475 0.236715
## BMI          0.80317    0.18948    4.239 0.024038 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.136 on 3 degrees of freedom
## (2 observations deleted due to missingness)
## Multiple R-squared:  0.8905, Adjusted R-squared:  0.8175
## F-statistic: 12.2 on 2 and 3 DF, p-value: 0.03624
```

Comment

ChatGPT provided the correct R code but did not offer any explanation of how the model works, why `na.action = na.omit` was used, or how to interpret the output.

Innocent Abaa's Approach

```
# Create data frame
data <- data.frame(
  Age = c(23, 35, 31, 35, 28, 37, 51, 41),
  SBP = c(122, 123, 120, 123, 124, 128, NA, NA),
  BMI = c(21, 19, 17, 22, 22, 25, 36, 41)
)

# Fit linear regression
model <- lm(SBP ~ Age + BMI, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = SBP ~ Age + BMI, data = data)
##
## Residuals:
##      1      2      3      4      5      6
## -0.09361  0.76254 -0.04772 -1.64698  0.37397  0.65180
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 101.87244    4.46807   22.800 0.000185 ***
## Age          0.14585     0.09889    1.475 0.236715
## BMI          0.80317     0.18948    4.239 0.024038 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.136 on 3 degrees of freedom
## (2 observations deleted due to missingness)
## Multiple R-squared:  0.8905, Adjusted R-squared:  0.8175
## F-statistic: 12.2 on 2 and 3 DF, p-value: 0.03624
```

In this small sample, we find that **BMI is a statistically significant predictor** of systolic blood pressure (SBP), with an estimated increase of **0.80 mmHg per unit increase in BMI** ($p = 0.024$). Age has a positive but **not statistically significant** relationship with SBP ($p = 0.24$). The model explains a large portion of SBP variation ($R^2 = 0.89$), though results should be interpreted cautiously due to the small sample size.

3. Describe MCAR Notation

GhatGPT Approach


```
#  $Y_{i,j}$ : Observed value of variable  $j$  for individual  $i$ 
#  $M_{i,j}$ : Indicator variable: 1 if  $Y_{i,j}$  is missing, 0 otherwise
#  $p_{ij}$ : Probability that  $M_{i,j} = 1$ 
# Under MCAR,  $p_{ij}$  is unrelated to both observed and unobserved data.
```

Innocent Abaa's Approach

Formulating Data as Y_{ij} , M_{ij} , and p_{ij}

For MCAR (Missing Completely At Random) analysis, we can define:

- Y_{ij} : Complete data matrix (Age, SBP, BMI for all patients)
- M_{ij} : Missingness indicator matrix (1 if missing, 0 if observed)
- p_{ij} : Probability that Y_{ij} is missing

In MCAR, the probability that a value is missing (p_{ij}) is independent of both the observed and unobserved data.

observed M_{ij} table

```
data2 <- data.frame(
  ID = 1:8,
  Age = c(23, 35, 31, 35, 28, 37, 51, 41),
  SBP = c(122, 123, 120, 123, 124, 128, NA, NA),
  BMI = c(21, 19, 17, 22, 22, 25, 36, 41))

# Create missingness indicators  $M_{ij}$ 
M <- as.data.frame(is.na(data2[,c("Age", "SBP", "BMI")])) * 1)
colnames(M) <- c("M_Age", "M_SBP", "M_BMI")
data_with_M <- cbind(data2, M)

# Print the data in desired format
print(data_with_M)
```

##	ID	Age	SBP	BMI	M_Age	M_SBP	M_BMI
## 1	1	23	122	21	0	0	0
## 2	2	35	123	19	0	0	0
## 3	3	31	120	17	0	0	0
## 4	4	35	123	22	0	0	0
## 5	5	28	124	22	0	0	0
## 6	6	37	128	25	0	0	0
## 7	7	51	NA	36	0	1	0
## 8	8	41	NA	41	0	1	0

Q4 Little's MCAR test

Interpretation:

If $p > 0.05$, fail to reject MCAR (i.e., data could be MCAR).

If $p \leq 0.05$, data is not MCAR.

Comment

ChatGPT provided an incorrect R function—`LittleMCAR` does not exist in the current R environment, making the code unusable. While the interpretation of the p-value was reasonable, ChatGPT failed to clearly state the null and alternative hypotheses being tested.

Innocent Abaa's Approach

We test whether the missing data mechanism is Missing Completely At Random (MCAR).

$$H_0 : P(M_{ij} = 1 \mid Y_{obs}, Y_{mis}) = P(M_{ij} = 1), \quad \forall i, j$$

(Missingness is completely random, independent of observed and unobserved data)

$$H_1 : P(M_{ij} = 1 \mid Y_{obs}, Y_{mis}) \neq P(M_{ij} = 1), \quad \text{for some } i, j$$

(Missingness depends on observed or unobserved data)

```
mcar_test <- mcar_test(data)
print(mcar_test)
```

```
## # A tibble: 1 × 4
##   statistic    df p.value missing.patterns
##   <dbl> <dbl>   <dbl>         <int>
## 1     7.36     2  0.0252           2
```

At a significance level of $\alpha = 0.05$, we reject the null hypothesis because:

$$p\text{-value} = 0.0252 < 0.05$$

This provides **statistically significant evidence that the missingness mechanism is *not* MCAR** (Missing Completely At Random). Therefore, the probability of missingness may depend on observed or unobserved values in the data.

Q5 Mean Imputation?

ChatGPT Approach

```
mean_SBP <- mean(df$SBP, na.rm = TRUE)
df_mean_imp <- df
df_mean_imp$SBP[is.na(df_mean_imp$SBP)] <- mean_SBP
```

You can do this, but only safe if MCAR. Otherwise, it biases variance and relationships.

Comment

ChatGPT correctly showed how to perform mean imputation and mentioned that it is only appropriate if the data is MCAR. While this guidance is helpful, the reasoning was not clearly articulated. It did not explain why mean imputation is problematic under MAR or MNAR, particularly in terms of how it can distort variance and bias relationships between variables.

Innocent Abaa's Approach

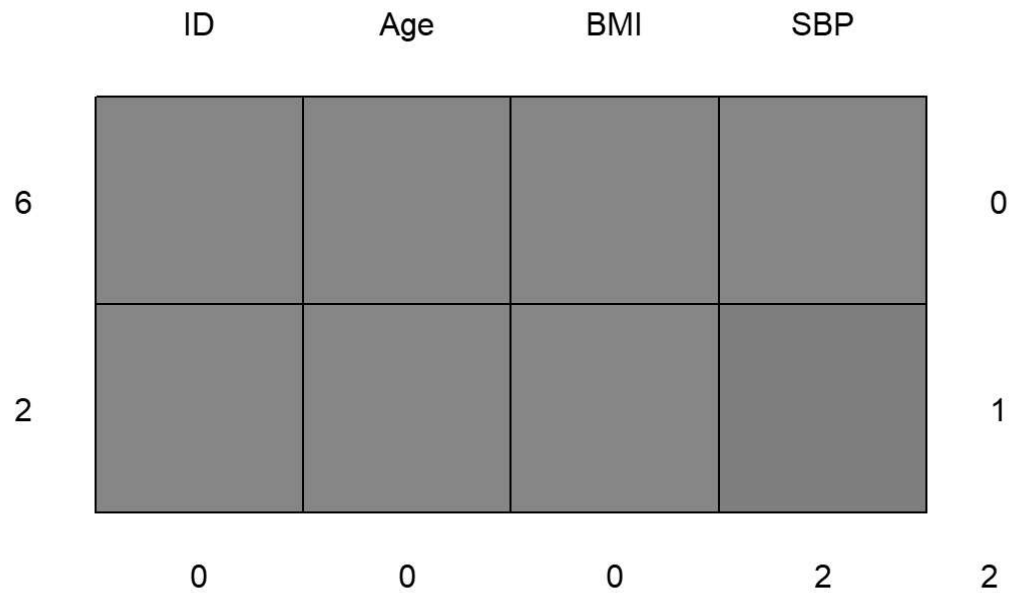
No, mean imputation is not appropriate here.

Because the data are not MCAR, mean imputation is not appropriate. It would introduce bias by ignoring the dependence of missingness on Age and BMI, leading to distorted relationships in any subsequent analysis.

Q6 Examine missingness pattern

ChatGTP's Approach

```
md.pattern(df)
```



```
##   ID Age BMI SBP
## 6  1  1  1  1 0
## 2  1  1  1  0 1
##   0  0  0  2 2
```

```
# Missing rate for Age > 40
mean(is.na(df$SBP[df$Age > 40]))
```

```
## [1] 1
```

```
# Missing rate for BMI > 30
mean(is.na(df$SBP[df$BMI > 30]))
```

```
## [1] 1
```

Comment

Maybe correctly implemented

Innocent Abaa's Approach

```
data3 <- data.frame(
  ID = 1:8,
  Age = c(23, 35, 31, 35, 28, 37, 51, 41),
  SBP = c(122, 123, 120, 123, 124, 128, NA, NA),
  BMI = c(21, 19, 17, 22, 22, 25, 36, 41))
print(data3)
```

```
##   ID Age SBP BMI
## 1  1  23 122  21
## 2  2  35 123  19
## 3  3  31 120  17
## 4  4  35 123  22
## 5  5  28 124  22
## 6  6  37 128  25
## 7  7  51  NA  36
## 8  8  41  NA  41
```

1. Missingness in Patients with Age > 40 and BMI > 30

- Patients meeting criteria (Age > 40 & BMI > 30):
 - ID 7: Age = 51, BMI = 36 → SBP missing
 - ID 8: Age = 41, BMI = 41 → SBP missing
- Missing rate in this group: **100% (2 out of 2)**.

2. Missingness in Patients with Age ≤ 40 or BMI ≤ 30

- All other patients (IDs 1–6) have:
 - Age ≤ 40 and BMI ≤ 30
- SBP is **fully observed** in this group (**0% missing, 0 out of 6**).

The missingness in SBP is **not random** but follows a **clear pattern**:

- Missing **only in older patients (Age > 40) with obesity (BMI > 30)**.
- SBP is **fully observed in all younger and normal-weight patients (Age ≤ 40 and BMI ≤ 30)**.

Q7. Stratify by BMI > 30

GhatGPT's Approach

```
df <- df %>% mutate(BMI_group = BMI > 30)
df %>% group_by(BMI_group) %>%
  summarise(missing_rate = mean(is.na(SBP)))
```

```
## # A tibble: 2 × 2
##   BMI_group missing_rate
##   <lgl>          <dbl>
## 1 FALSE          0
## 2 TRUE           1
```

Comment

Correctly implemented

Innocent Abaa's Approach

```
data4<- data.frame(
  ID = 1:8,
  Age = c(23, 35, 31, 35, 28, 37, 51, 41),
  SBP = c(122, 123, 120, 123, 124, 128, NA, NA),
  BMI = c(21, 19, 17, 22, 22, 25, 36, 41))
data4$BMI_group <- ifelse(data4$BMI > 30, ">30", "<=30")
aggregate(is.na(data4$SBP), by=list(BMI_Group=data4$BMI_group), mean)
```

```
## BMI_Group x
## 1      <=30 0
## 2      >30 1
```

Here Missing rate in BMI > 30 group: 100%

This tells us that missing probability for BMI > 30 group is very high

Also, Missing rate in BMI ≤ 30 group: 0%

This tells us that missing probability for BMI ≤ 30 group is very low

```
dat <- data.frame(
  age = c(23, 35, 31, 35, 28, 37),
  sbp = c(122, 123, 120, 123, 124, 128),
  bmi = c(21, 19, 17, 22, 22, 25)
)

mcar_test2 <- mcar_test(dat)
print(mcar_test2)
```

```
## # A tibble: 1 × 4
## statistic    df p.value missing.patterns
##      <dbl> <dbl>   <dbl>          <int>
## 1         0     0     1             1
```

We define the probability that observation M_{ij} is missing as a function of BMI using indicator functions:

$$P(M_{ij} = 1) = -0.3 \cdot I(\text{BMI} \leq 30) + 0.5$$

$$P(M_{ij} = 1) = 0.4 \cdot I(\text{BMI} > 30) + 0.5$$

where:

- $I(\cdot)$ is an indicator function that equals 1 if the condition inside is true, and 0 otherwise.

8. Stratify by Age > 40

ChatGPT's Approach

```
df <- df %>% mutate(Age_group = Age > 40)
df %>% group_by(Age_group) %>%
  summarise(missing_rate = mean(is.na(SBP)))
```

```
## # A tibble: 2 × 2
##   Age_group missing_rate
##   <lgl>          <dbl>
## 1 FALSE          0
## 2 TRUE           1
```

Innocent Abaa's Approach

```
data$age_group <- ifelse(data$Age > 40, ">40", "<=40")
aggregate(is.na(data$SBP), by=list(Age_Group=data$age_group), mean)
```

```
##   Age_Group x
## 1      <=40 0
## 2      >40 1
```

Here Missing rate in Age > 40 group: 100%

This tells us that missing probability for Age > 40 group is very high

Also, Missing rate in Age ≤ 40 group: 0%

This tells us that missing probability for Age ≤ 40 group is very low

$$P(M_{ij} = 1) = -0.3 \cdot I(\text{Age} \leq 40) + 0.5$$

$$P(M_{ij} = 1) = 0.4 \cdot I(\text{Age} > 40) + 0.5$$

Conclusion

ChatGPT was effective in addressing some aspects of the assignment, particularly in Questions 6, 7, and 8, and its initial explanations were helpful in framing the problems. However, it did not fully or correctly answer most of the questions, even after multiple attempts. In several cases, it suggested R packages that do not exist or have been removed from **CRAN**, resulting in code that would produce errors without human correction.

Furthermore, where ChatGPT did attempt to provide solutions, the responses were often incomplete or lacked the depth needed to fully answer the questions. In some cases, the answers were even contradictory, which required me to apply my own understanding of missing data mechanisms to identify and correct the issues. Overall, while ChatGPT can serve as a useful supplementary tool, it is not a substitute for a solid grasp of the subject matter and critical thinking.

Additional Conclusion Options

Usefulness as a Learning Aid: While ChatGPT did not always provide correct or complete answers, its explanations helped clarify concepts and guided my thinking. This suggests that ChatGPT can be a valuable learning aid, especially when used to supplement—not replace—traditional instruction(Dr. John Chen) or my critical thinking.

Limitations in Handling Statistical Inference: ChatGPT struggled with tasks requiring formal statistical inference, such as performing hypothesis tests like Little’s MCAR test. Its inability to execute code or access current R packages in real-time limits its usefulness in applied data analysis without human oversight.

Strength in Repetition, Not Critical Thinking: Even when prompted multiple times, ChatGPT didn't always refine or correct its errors. This suggests that it lacks the critical thinking and iterative refinement that a human tutor or expert would provide during problem-solving.