PH 718 Data Management and Visualization in `R`

Part 5: Missiningness Imputation

Zhiyang Zhou (zhou67@uwm.edu, zhiyanggeezhou.github.io)

2026/02/28 17:07:16

Missing data frequently occurs in real-world datasets, leading to biased estimates and reduced statistical power.

Primary types of missingness

- Missing completely at random (MCAR): The missingness is unrelated to any observed or unobserved data. Examples:
  - A participant accidentally skips a survey question due to a technical glitch.
  - A random subset of data is lost due to equipment malfunction during data collection.
- Missing at random (MAR): The missingness is related to observed data but NOT to the missing values themselves. Examples:
  - Participants with lower educational backgrounds might not report their income due to misunderstanding of the survey question.
  - Older respondents might skip answering online survey questions due to difficulty using the technology.
- Missing not at random (MNAR): The missingness is related directly to the missing data itself. Examples:
  - Individuals with high incomes may choose not to disclose their exact income in a survey.
  - Patients experiencing severe health issues may systematically refuse to answer questions about their health condition.

Ex. 5.1. Identify the type of missingness

- For each scenario below, classify the missingness as: MCAR, MAR, or MNAR, and provide a brief justification for your classification.
  - In a hospital survey, a random server outage causes 5% of patient responses to be lost across all departments.
  - In a mental health questionnaire, older patients are less likely to complete the online form due to difficulty navigating the system.
  - Patients with very high alcohol consumption are more likely to skip the question asking about alcohol intake.
  - A lab machine malfunctions for two hours and loses all cholesterol measurements during that time period.
  - People with lower BMI are less likely to respond to a question about exercise habits, but BMI itself is fully observed.

Commonly used methods

Complete-case analysis

- Remove records with missing values.

- Easy but may result in substantial data loss and biased results if data aren't MCAR.

Simple imputation (with mean/median/mode)

- Replace missing numeric values with the mean or median of the observed values.
- Replace missing categorical values with the mode (i.e., the most frequent value).
- Easy but can lead to biased results and underestimated variance.
- May produce unbiased point estimates under MCAR

Iterative regression imputation (single imputation)

- Predict missing values using regression, iteratively updating until stable.
- Better than the simple imputation, but doesn't fully account for uncertainty.
- Can work for MCAR and MAR if the regression model is correctly specified

```r
# Illustrative example: airquality
## Preprocessing
old_df <- airquality
old_df$lnOzone = log(old_df$Ozone)
colSums(is.na(old_df))
## Initial simple imputation
missing_ozone <- which(is.na(old_df$Ozone))
missing_solar <- which(is.na(old_df$Solar.R))
old_df$Ozone[missing_ozone] <- mean(old_df$Ozone, na.rm = TRUE)
old_df$Solar.R[missing_solar] <- mean(old_df$Solar.R, na.rm = TRUE)
## Set convergence criteria
tolerance <- 1e-4
max_iterations <- 20

## Iterative process with convergence check
for (i in 1:max_iterations) {
  if (i==1) {
    curr_df <- old_df
  }else{
    curr_df <- next_df
  }
  #### Regressions
  model_ozone <- lm(
    lnOzone ~ Solar.R + Wind + Temp + as.factor(Month) + as.factor(Day),
    data = curr_df
  )
  model_solar <- lm(
    Solar.R ~ Ozone + Wind + Temp + as.factor(Month) + as.factor(Day),
    data = curr_df
  )
  #### Updating data
  next_df = curr_df
  next_df$Ozone[missing_ozone] <- exp(predict(model_ozone, newdata = curr_df[missing_ozone, ]))
  next_df$lnOzone = log(next_df$Ozone)
  next_df$Solar.R[missing_solar] <- predict(model_solar, newdata = curr_df[missing_solar, ])

  #### Check the mean squared difference
  diff <- mean((curr_df$Ozone - next_df$Ozone)^2 + (curr_df$Solar.R - next_df$Solar.R)^2)
  cat("Iteration:", i, "Difference:", diff, "\n")
  if (diff < tolerance) {
    cat("Convergence achieved after", i, "iterations.\n")
    break
  }
}
View(next_df)
```

Multiple imputation (MI)

- Designed to produce unbiased and efficient estimates under MCAR and MAR, provided the imputation model is correctly specified.
- Procedure:
  1. Create multiple imputed datasets.
  2. Analyze each separately.
  3. Pool the results using the Rubin's rule (Rubin 1987).
- Rubin's rule

$$\text{Pooled point estimate: } \bar{Q} = m^{-1}\sum_{j=1}^{m} Q_j, \quad \text{where } Q_j \text{ is a point estimate}$$

$$\text{Average within-imputation variance: } V_w = m^{-1}\sum_{j=1}^{m} \text{var}(Q_j)$$

$$\text{Between-imputation variance: } V_b = \frac{1}{m-1}\sum_{j=1}^{m}(Q_j - \bar{Q})^2$$

$$\text{Total variance: } V_t = V_w + (1+m^{-1})V_b$$

$$\text{Degrees of freedom: } \nu = (m-1)[1 + V_w/\{(1+m^{-1})V_b\}]^2$$

$$95\% \text{ confidence interval: } \bar{Q} \pm t_{\nu,0.975}\sqrt{V_t}$$

```
# Illustrative example: airquality
## Using mice::mice() with default Bayesian regression models:
### predictive mean matching (pmm, for numeric data)
### logistic regression (logreg, for factors with 2 levels)
### polytomous regression (polyreg, for unordered factor > 2 levels)
### proportional odds model for (polr, ordered factors > 2 levels)
library(mice)
old_df <- airquality
colSums(is.na(old_df))
imp_obj <- mice(
  old_df,
  m = 2, # number of imputed datasets
  maxit = 200,
  method = c("pmm","pmm","", "", "", ""),
  seed = 123
)
## Check convergence visually
### Each trajectory should stabilize and mix well over iterations
### No upward/downward trends = good convergence.
### Wiggly lines or divergence = possible convergence issues.
```

```r
plot(imp_obj)
## Extracting imputed data
for (i in 1:imp_obj$m) {
  data_name <- paste0("data_imp_", i)
  assign(data_name, complete(imp_obj, i))
}
## Model fitting with imputed data
fit <- with(data = imp_obj, exp = lm(Ozone ~ Solar.R + Wind + Temp))
pooled <- pool(fit, rule = "rubin1987") # Pool the results using Rubin's rules
summary(pooled, conf.int = T) # Summarize the pooled result
```

```r
# Illustrative example: mice::nhanes2
library(mice)
colSums(is.na(nhanes2))
imp_obj <- mice(
  nhanes2,
  m = 2,
  maxit = 100,
  method = c("", "pmm", "logreg", "pmm"),
  seed = 123
)
plot(imp_obj)
for (i in 1:imp_obj$m) {
  data_name <- paste0("data_imp_", i)
  assign(data_name, complete(imp_obj, i))
}
fit <- with(data = imp_obj, exp = lm(chl ~ age + bmi + hyp))
pooled <- pool(fit, rule = "rubin1987")
summary(pooled, conf.int = T)
```

Maximum likelihood estimation

- Incorporates the density of missingness into the likelihood when estimating unknown parameters. Sophistication is needed in constructing and optimizing complex likelihood functions.

- Required for MNAR.

Bibliography

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.