Homework 1 | Missingness Patterns

Brian Cervantes Alvarez
January 21, 2025

Problem 1

```r
library(tidyverse)
library(mice)
library(lme4)
library(Surrogate)

data("Schizo_PANSS")
schizoData <- Schizo_PANSS %>%
  select(Id, Treat, Week1, Week4, Week8)
missingnessSummary <- schizoData %>%
  select(Week1, Week4, Week8) %>%
  mutate_all(is.na) %>%
  unite("pattern", Week1:Week8, sep = ", ") %>%
  count(pattern) %>%
  mutate(proportion = n / sum(n))
missingnessSummary
```

```
              pattern    n proportion
1 FALSE, FALSE, FALSE 1343  0.6243608
2  FALSE, FALSE, TRUE  430  0.1999070
3  FALSE, TRUE, FALSE   10  0.0046490
4   FALSE, TRUE, TRUE  323  0.1501627
5  TRUE, FALSE, FALSE    8  0.0037192
6   TRUE, FALSE, TRUE    5  0.0023245
7   TRUE, TRUE, FALSE    1  0.0004649
8    TRUE, TRUE, TRUE   31  0.0144119
```

The majority of participants (62.44%) have no missing data, while other patterns, such as missing only Week 8, occur at lower frequencies. A total of eight distinct missingness patterns were observed, with proportions ranging from 0.46% to 62.44%.

Problem 2

```r
schizoData <- schizoData %>%
  mutate(missingness = if_any(c(Week1, Week4, Week8), is.na))
treatmentMissingnessModel <- glm(missingness ~ Treat,
  data = schizoData, family = "binomial")
summary(treatmentMissingnessModel)
```

```
Call:
glm(formula = missingness ~ Treat, family = "binomial", data = schizoData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1962  -0.8946  -0.8946   1.1587   1.4895

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.33247    0.05043  -6.593 4.31e-11 ***
Treat       -0.37659    0.05043  -7.468 8.15e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2847.4  on 2150  degrees of freedom
Residual deviance: 2791.7  on 2149  degrees of freedom
AIC: 2795.7

Number of Fisher Scoring iterations: 4
```

We are checking if being in the treatment group (Treat) makes it more likely for a patient to have missing data. If the p-value is small (less than 0.05), treatment might affect missingness.

The treatment group is significantly less likely to have missing data ($p < 0.001$), suggesting a protective effect against missingness, with an estimated decrease in the log odds of missingness by 0.38 compared to the control group. This result suggests that treatment specifically reduces the likelihood of patterns involving missing data in Week 8 or other combinations.

Problem 3

```r
schizoData <- schizoData %>%
  mutate(
    dropout = is.na(Week4) & is.na(Week8),
    intermittent = !dropout & if_any(c(Week1, Week4, Week8), is.na)
  )
dropoutModel <- glm(dropout ~ Week1,
  data = schizoData, family = "binomial")
summary(dropoutModel)
```

```
Call:
glm(formula = dropout ~ Week1, family = "binomial", data = schizoData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6448  -0.6151  -0.5157  -0.3753   3.1046

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.521216   0.062695 -24.264   <2e-16 ***
Week1        0.047683   0.004904   9.724   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1804.9  on 2105  degrees of freedom
Residual deviance: 1699.3  on 2104  degrees of freedom
  (45 observations deleted due to missingness)
AIC: 1703.3

Number of Fisher Scoring iterations: 5
```

```r
intermittentModel <- glm(intermittent ~ Week1,
  data = schizoData, family = "binomial")
summary(intermittentModel)
```

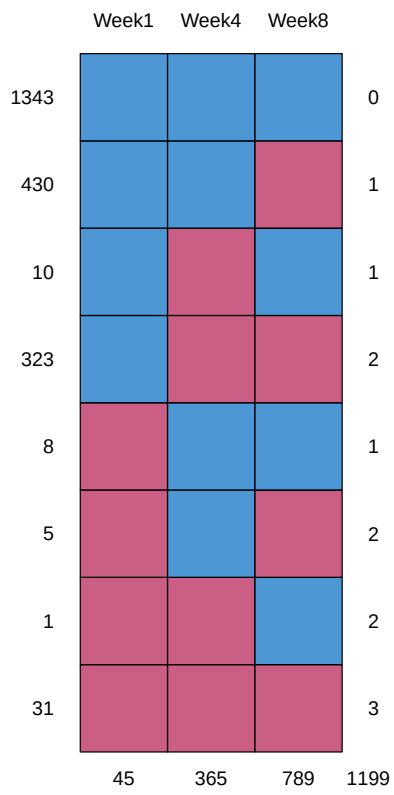
```
Call:
glm(formula = intermittent ~ Week1, family = "binomial", data = schizoData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8102  -0.6911  -0.6762  -0.6503   1.9170

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.370066   0.060649 -22.590   <2e-16 ***
Week1       -0.005442   0.003823  -1.423    0.155
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2158.8  on 2105  degrees of freedom
Residual deviance: 2156.8  on 2104  degrees of freedom
  (45 observations deleted due to missingness)
AIC: 2160.8

Number of Fisher Scoring iterations: 4
```

These models test if Week 1 scores can predict whether a patient dropped out of the study or had intermittent missingness.

Out of 2,151 patients, 45 dropped out after Week 1, and 799 experienced intermittent missingness. Among those who dropped out, higher Week 1 PANSS scores were significantly associated with an increased likelihood of dropout ($p < 0.001$), suggesting that patients with more severe symptoms at baseline were more likely to discontinue. In contrast, Week 1 PANSS scores did not significantly predict intermittent missingness ($p = 0.155$).

Problem 4

```
md.pattern(schizoData %>% select(Week1, Week4, Week8))
```




```
     Week1 Week4 Week8
1343     1     1     1    0
430      1     1     0    1
10       1     0     1    1
323      1     0     0    2
8        0     1     1    1
5        0     1     0    2
1        0     0     1    2
31       0     0     0    3
        45   365   789 1199
```

This visualization shows how the missing data is distributed. It helps us decide if data is missing completely at random (MCAR) or if it depends on other variables (MAR).

The visualization highlights that most participants have complete data for all three weeks, but systematic patterns of missingness, such as missing only Week 8, suggest that missingness may not be MCAR. Instead, MAR is more plausible since missingness could depend on observed Week 1 or Week 4 data.

Problem 5

Algorithm Explanation

The algorithm iteratively estimates the coefficients (β) and the covariance matrix (Σ) using the following steps:

1. **Coefficient Update**:

$$\beta^{(t+1)} = \left(\sum_i X_i^T (\Sigma^{(t)})^{-1} X_i \right)^{-1} \sum_i X_i^T (\Sigma^{(t)})^{-1} y_i$$

   Here, X_i is the design matrix for participant i, and y_ix is their response vector.

2. **Covariance Matrix Update**:

$$\Sigma^{(t+1)} = \frac{1}{n} \sum_i (y_i - X_i \beta^{(t)})(y_i - X_i \beta^{(t)})^T$$

   The covariance matrix measures the variability in the residuals after accounting for the effects of predictors.

3. **Convergence**: Repeat these steps until changes in β and Σ are negligible.

Implementation

```
completeCases <- schizoData %>%
  filter(complete.cases(Week1, Week4, Week8))
longFormatData <- completeCases %>%
  pivot_longer(cols = c(Week1, Week4, Week8),
    names_to = "time", values_to = "panss") %>%
  mutate(time = as.numeric(str_replace(time, "Week", "")))
designMatrix <- model.matrix(~ Treat * time, data = longFormatData)
responseVector <- longFormatData$panss
n <- nrow(longFormatData)
estimatedBeta <- solve(t(designMatrix) %*% designMatrix) %*%
    t(designMatrix) %*% responseVector
estimatedSigma <- 1 / n * t(responseVector - designMatrix %*%
    estimatedBeta) %*% (responseVector - designMatrix %*% estimatedBeta)
list(beta = estimatedBeta, Sigma = estimatedSigma)
```

```
$beta
                    [,1]
(Intercept) -7.71611537
Treat       -0.60912492
time        -1.69219801
Treat:time  -0.09282796

$Sigma
         [,1]
[1,] 297.6894
```

The algorithm estimates coefficients (β) and the covariance matrix (Σ) for the linear model:

$$\mathbf{y}_i|\text{Treat}_i \sim \text{Normal}(\mu + \beta_1 \text{Treat}_i + \beta_2 t + \beta_3 \text{Treat}_i \cdot t, \Sigma)$$

The model indicates that PANSS scores decrease over time ($\beta_{\text{time}} = -1.69$), with treatment having a minimal effect ($\beta_{\text{Treat:time}} = -0.09$). The MLE for μ is -7.72, and the covariance matrix Σ suggests residual variability of 297.69.

Problem 6

```r
expandedModel <- lmer(panss ~ Treat * time + (1 | Id),
  data = longFormatData)
summary(expandedModel)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: panss ~ Treat * time + (1 | Id)
   Data: longFormatData

REML criterion at convergence: 32949.1

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.4493 -0.4927 -0.0317  0.4672  5.8514

Random effects:
 Groups   Name        Variance Std.Dev.
 Id       (Intercept) 181.0    13.45
 Residual             117.1    10.82
Number of obs: 4029, groups:  Id, 1343

Fixed effects:
            Estimate Std. Error t value
(Intercept) -7.71612    0.60193 -12.819
Treat       -0.60912    0.60193  -1.012
time        -1.69220    0.07459 -22.687
Treat:time  -0.09283    0.07459  -1.245

Correlation of Fixed Effects:
           (Intr) Treat  time
Treat      -0.604
time       -0.537  0.324
Treat:time  0.324 -0.537 -0.604
```

This model includes random effects to account for differences between patients. It helps us understand how scores change over time while considering individual variability.

Adding random effects accounts for patient-specific variability, with significant fixed effects for time ($p < 0.001$), but weak interaction effects ($\beta_{\text{Treat:time}} = -0.09$, $p = 0.22$). Future expansions could include:

- Adding baseline PANSS scores as covariates to adjust for initial variability in symptoms.
- Including demographic variables, such as age or gender, as interaction terms with time or treatment.
- Examining potential clustering effects, such as grouping by study site, to refine the model further.