HW 1

Question 1

Load the `Surrogate` package in R and load the dataset `Schizo_PANSS`:

```
library(Surrogate)
data("Schizo_PANSS")
```

The dataset combines five clinical trials aimed at determining if risperidone decreases the Positive and Negative Syndrome Score (PANSS) over time compared to a control treatment for patients with schizophrenia. These are longitudinal trials where patients are assessed at weeks 1, 2, 4, 6 and 8 after being assigned to a treatment arm.

Each row in the dataset is a different trial participant, and `Week1`, `Week2`, `Week4`, `Week6`, `Week8` records the change in PANSS from baseline. The variable `Treat` represents whether the patient was enrolled in the control (-1) or if the patient was in the risperidone arm (1).

Subset the data to include only `Week1`, `Week4` and `Week8`:

```
hw_data <- Schizo_PANSS[,c("Id","Treat","Week1","Week4","Week8")]
```

1. Summarize the missingness patterns in the `hw_data` dataset. How many missingness patterns are there? What are they? What proportion of patients are associated with each missingness pattern?
2. How would you assess whether there is evidence that treatment affects missingness? Is there evidence that treatment affects the missingness pattern?
3. How many people dropped out of the study after Week 1, or Week 4 vs. had intermittent missingness? For patients who dropped out, is there evidence that PANSS at the prior measurement predicted dropout? What about for the patients with intermittent missingness?
4. Is it reasonable to assume that missingness is MCAR for this dataset? Why or why not? What about MAR?
5. Subset the data to complete cases only and, using the algorithm we learned in class:

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

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

Fit the following model to the complete case data:

$$y_i \mid \text{Treat}_i \sim \text{Normal}(\mu + \beta_1 \text{Treat}_i + \beta_2 t + \beta_3 \text{Treat}_i t, \Sigma)$$

where t is the vector $1, 4, 8$ indicating at what time points the measurements were taken and μ is a scalar mean.

Include your MLEs for Σ and the vector $(\mu, \beta_1, \beta_2, \beta_3)$, and be sure to interpret your inferred coefficients in the context of the PANSS dataset.

It'll help to reshape the data into long format from wide format:

```
comp_case <- hw_data |>
  subset(
    !is.na(Week1) &
      !is.na(Week4) &
      !is.na(Week8)
  )
long_case <-
  stats::reshape(
    comp_case,
    direction = "long",
    varying = 3:5,
    sep = ""
  )[,-5]
names(long_case) <- c("Id","Treat","time","panss")
```

In order to make sure your algorithm is successful, include a test of your algorithm on on this simulated dataset, where you compare your algorithm's inferences to the true values of β and Σ:

```
set.seed(123)
n <- 10000
p <- 5
K <- 3
X <- list()
beta <- rnorm(p)
L <- matrix(rnorm(K * K),K,K)
Sigma <- L %*% t(L)
```

```r
y <- list()
for (i in 1:n) {
  X[[i]] <- matrix(rnorm(p * K),K,p)
  y[[i]] <- X[[i]] %*% beta + MASS::mvrnorm(mu = rep(0,K), Sigma = Sigma)
}
```

6. How might you expand this model based on your initial data analysis? There are no wrong answers.