MLM Nested Simulation Project (Spring 2020)

1. You will generate simulated data for a single school with 100 classrooms, each of which has 200 students.
    a. Outcome for student i in classroom j: Y_{ij}.
    b. There is a single predictor, $X_{ij} \sim U(0,1)$ (uniform on [0,1]).
    c. There is a classroom random effect, $\eta_j \sim N(0,\sigma^2_\eta)$, where $\sigma^2_\eta = 2$.
    d. Subject level error, $\varepsilon_{ij} \sim N(0,\sigma^2_\varepsilon)$, where $\sigma^2_\varepsilon = 2$.
    e. set.seed(2042001) once at the beginning of your code.

    f. Generate the random quantities in this order to ensure the same solution for everyone: X, η_j, ε_{ij}
    g. The outcome has the following form (DGP, given the modeling parameters above):

    $$Y_{ij} = 0 + 1X_{ij} + \eta_j + \varepsilon_{ij}; \quad \eta_j \sim N(0,\sigma^2_\eta), \varepsilon_{ij} \sim N(0,\sigma^2_\varepsilon), \text{ indep.}$$

    h. Generate a single simulated dataset (you will need an "classid" variable to track classrooms; you can optionally assign a "studentid")
        i. **Important**: construct classid such that classrooms appear consecutively within the dataframe. As per: `rep(1:J,each=n_j)`
2. Fit the model corresponding to the DGP on your simulated data.
    a. Report coefficient estimate for slope on X.
    b. Does a 95% confidence band for this coefficient estimate cover the "truth" that you used to generate the data?
3. Next, we simulate missing data in several ways. This is the first:
    a. Make a copy of the data, then modify the copy following these instructions:
    b. Generate $Z_{ij} \sim \text{Bernoulli}(p)$, with $p = 0.5$.
    c. Set Y to NA when $Z_{ij} == 1$. This should look a lot like "MCAR" missingness.
    d. Refit the model on the new data and report the coefficient estimate for slope on X. Look at the other parameter estimates as well.
    e. Do you see any real change in the β_X estimate?
        i. Does a 95% confidence band for this coefficient estimate cover the "truth" that you used to generate the data?
    f. What is the total sample N size used in the model fit?
4. Missing Data II: Make another copy of the *original* data, then modify the copy as follows:
    a. Generate $Z_{ij} \sim \text{Bernoulli}(X_{ij})$, with X_{ij} your predictor generated previously.
    b. Set Y to NA when $Z_{ij} == 1$. This should look a lot like "MAR" missingness.
    c. Refit the model on the new data and report the coefficient estimate for slope on X. Look at the other parameter estimates as well.
    d. Do you see any real change in the β_X estimate?
        i. Does a 95% confidence band for this coefficient estimate cover the "truth" that you used to generate the data?
    e. What is the total sample N size used in the model fit?
5. Missing Data III: Make another copy of the *original* data, then modify the copy as follows:
    a. First, define the expit function: expit <- function(x) exp(x)/(1+exp(x))
    b. Generate $Z_{ij} \sim \text{Bernoulli}(expit(Y_{ij}))$, with Y_{ij} your *outcome* generated previously.
    c. Set Y to NA when $Z_{ij} == 1$. This should look like a violation of "MAR" missingness (missingness depends on outcome and cannot be *simply* predicted with the predictor set – Y should be correlated with X, though, so it might not be too bad a violation).
    d. Refit the model on the new data and report the coefficient estimate for slope on X. Look at the other parameter estimates as well.
    e. Do you see any real change in the β_X estimate?
        i. Does a 95% confidence band for this coefficient estimate cover the "truth" that you used to generate the data?
    f. What is the total sample N size used in the model fit?