HW 03 - Missing Data

YOUR NAME

3/5/23

Setup

Import data & load libraries.

```
library(VIM)
library(tidyverse)
library(ggpubr)
#load("mi_example.Rdata") # fix your path
```

Part I. Exploring Imputation Techniques

Identify missing

Single Imputation

Use hotdeck imputation on `parent_overprotection` using the `hotdeck(dataset, variable = "var")` function in VIM. See `vignette("donorImp")` for more information.

Multiple Imputation

1. Create m imputed datasets
2. Calculate the point estimate Q and the variance U from each imputation.
3. Pool estimates

Comparison of Estimates

Calculate the estimate, SE and 95% CI for the average parental overprotection score under the following frameworks.

- Complete Case
- Single Imputation
- Multiple Imputation

Summary

Part II: Multiple Imputation using Chained Equations

Sticking with the Parental HIV data set (the one from the practice worksheet), build a better imputation model for `parental_overprotection`. Do this by imputing the `pb01-pb25`, then recreate the scale post-imputation. "Talk me" through your process.

1. Explore missing data patterns in other (non-scale) variables before you build your model. Not all variables should be considered in the imputation models. Use tables and plots. Discuss all output.

2. Multiply impute this data set between $m = 5$ and $m = 10$ times. Make sure the imputation models used for each variable are showing in your final output. Adjust any that may not make sense for their variable type.

3. Update the summary plot and compare how your new model did compared to the earlier ones from the worksheet.

4. After controlling for other measures, what is the effect of gender on the odds a student will skip school? Adjust the model for fit or stability as needed. Report your results in a nice table and interpret the effect of gender on skipping school.

4a. Fit this model on the complete cases (no imputation).

4b. Fit this model on the multiply imputed data sets and report the pooled estimates and intervals.

4c. Interpret the effect of gender on playing hookey. Did it change from the complete case model?

4d. Create a plot to compare the results for all coefficients in the model.

4e. What are the biggest differences you notice? Would the inference/interpretation of the effect of any covariate on the odds of a student skipping school change depending on what model you use?