Homework 5

PubH 7406: Biostatistical Inference II – Jared D. Huling

Due: Tues, April 18, 2023 at the beginning of class

Missing Data HW - 40 points

Instructions

Turn in the homework in the form of a PDF. It is fine to use existing functions to answer questions. However, note that simply providing output from statistical software is not sufficient and will not receive full points. Any output/results must be interpreted in the context of the problem and accompanying explanations of the results are necessary. Use clearly-defined and explained statistical notation to accompany results. For example, any statistical tests should be accompanied by a formal statement of the null and alternative hypotheses with additional context explaining what the hypotheses mean in the context of the problem. Please follow the instructions on homeworks in the syllabus in order to receive full credit.

If you have any questions, please ask in the course Q&A on Canvas so that others can see any responses.

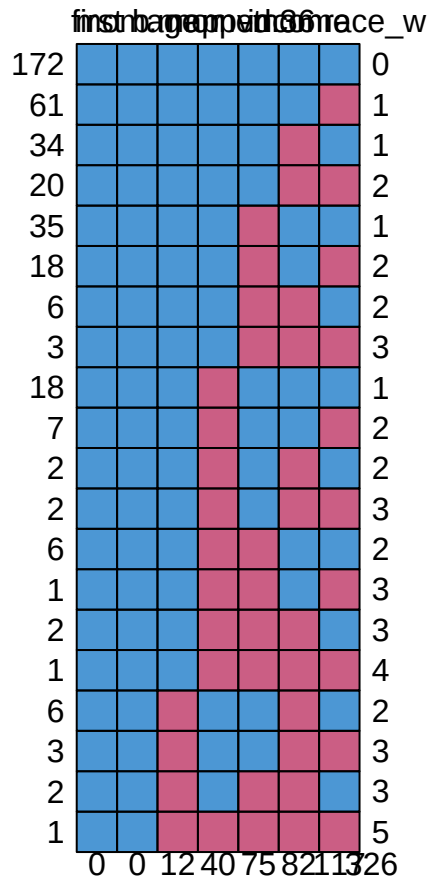
The Questions

1. Consider the data from the National Longitudinal Survey of Youth Extract that is part of the mi package in R. Note that you should use the mice package to analyze this dataset. A description of all the variables is given here (https://cran.r-project.org/web/packages/mi/mi.pdf).

```
library(mi)
library(mice)
data(nlsyV)
nlsyV$momrace_w <- as.factor(ifelse(nlsyV$momrace == 3, "white", "nonwhite"))
nlsyV$momed <- as.factor(nlsyV$momed)
nlsyV <- nlsyV[, - c(7)]
head(nlsyV)
```

```
##      ppvtr.36 first b.marr income momage momed momrace_w
## 535       105     1      1  21446     20     2     white
## 2932       91     1      1  12125     22     2     white
## 2906       89     0      1  13560     22     2  nonwhite
## 4510       85     0      1  24500     28     3      <NA>
## 3869       66     0      0   3304     20     1      <NA>
## 2952       NA     0      0   5832     27     2      <NA>
```

```
md.pattern(nlsyV)
```




```
##     first momage b.marr momed ppvtr.36 income momrace_w
## 172     1       1      1     1         1      1         1   0
## 61      1       1      1     1         1      1         0   1
## 34      1       1      1     1         1      0         1   1
## 20      1       1      1     1         1      0         0   2
## 35      1       1      1     1         0      1         1   1
## 18      1       1      1     1         0      1         0   2
## 6       1       1      1     1         0      0         1   2
## 3       1       1      1     1         0      0         0   3
## 18      1       1      1     0         1      1         1   1
## 7       1       1      1     0         1      1         0   2
## 2       1       1      1     0         1      0         1   2
## 2       1       1      1     0         1      0         0   3
## 6       1       1      1     0         0      1         1   2
## 1       1       1      1     0         0      1         0   3
## 2       1       1      1     0         0      0         1   3
## 1       1       1      1     0         0      0         0   4
## 6       1       1      0     1         1      0         1   2
## 3       1       1      0     1         1      0         0   3
## 2       1       1      0     1         0      0         1   3
## 1       1       1      0     0         0      0         0   5
##         0       0     12    40        75     82       117 326
```

a. (5 points) Use multiple imputation for the missing values and describe the imputation scheme you used.

b. (5 points) Justify that the imputation that you have chosen is reasonable

c. (5 points) Please evaluate the effect of being a first born child on the Peabody Picture Vocabulary Test after adjusting for the other covariates in the dataset. Be very explicit on the model that you fit and how you arrived at that model.

2. A simulation study is often an effective means of comparing the benefits and drawbacks of different statistical procedures. In a simulation study, the simulator (YOU) knows exactly how the data were generated so you know the true value of the parameters. However, within the simulation, you analyze the data using only information that is typically available. In other words, you can't cheat and use the true value of the parameters or the value of some missing data. Suppose that we use the following code to generate a response Y and 5 different predictor $X_1, \ldots X_5$ (note that X_1, X_2, X_3 are continuous and X_4, X_5 are binary). Unfortunately, we don't get to see some of the covariates on all subjects. So I've added some additional code to create some missing data. I've also obtained parameter estimates and standard errors using five methods. (A) Using complete cases only; (B) Imputing the mean value of the covariate; (C) Imputing the predicted value; (D) Single imputation with error; (E) Multiple imputation with error. Let's suppose that I am most interested in estimating the regression parameter associated with X_1.

Note, I generate the data and then analyzed 100 different times. It is possible that on one particular dataset, a particular estimation approach works well and on the next dataset works much more poorly.

```
set.seed(8172013)
library(mvtnorm)
expit <- function(x) {
  expit <- exp(x)/(1+exp(x))
  return(expit)}

# Number of Datasets
S <- 100

# Sample Size
n <- 200

# Parameter and Standard Error Estimates
parm.est <- matrix(0, nrow = S, ncol = 5)
se.est <- matrix(0, nrow = S, ncol = 5)
colnames(parm.est) <- colnames(se.est) <- paste("Method", letters[1:5], sep = ".")
```

```r
for (q in 1:S) {
  X_123 <- rmvnorm(n, mean = rep(0, 3),
                   sigma = matrix(c(1, 0.6, 0.36, 0.6, 1, 0.6, 0.36, 0.6, 1),
  nrow = 3, ncol = 3))
  X1 <- X_123[, 1]
  X2 <- X_123[, 2]
  X3 <- X_123[, 3]
  X4 <- rbinom(n, 1, expit(0.5*X1 + 0.5*X2 - 0.5*X3))
  X5 <- rbinom(n, 1, expit(0.5*X1 - 0.5*X2 + 0.5*X3 - 0.5*X4))
  Y <- 2 + X1 + X2 + X3 + 2*X4 + 2*X5 + rnorm(n, mean = 0, sd = 3)

  # Add in some missing data
  X1_obs <- rbinom(n, 1, expit(1 + 0.25*X2 + 0.25*X3 + 0.125*X4 + 0.125*X5))
  X4_obs <- rbinom(n, 1, expit(1 + 0.25*X1 + 0.25*X2 + 0.25*X3 +  0.125*X5))
  X1 <- ifelse(X1_obs == 1, X1, NA)
  X4 <- ifelse(X4_obs == 1, X4, NA)
  final_data <- data.frame(Y, X1, X2, X3, X4, X5)

  # Method A
  method.A <- summary(lm( Y ~ X1 + X2 + X3 + X4 + X5,
                          data = final_data))$coefficients
  method.A.coef <- method.A[2, 1]
  method.A.se <- method.A[2, 2]

  # Method B
  X1_impB <- ifelse(is.na(X1) == TRUE, mean(X1, na.rm = TRUE), X1)
  X4_impB <- ifelse(is.na(X4) == TRUE, mean(X4, na.rm = TRUE), X4)
  final_data <- data.frame(final_data, X1_impB, X4_impB)
  method.B <- summary(lm( Y ~ X1_impB + X2 + X3 + X4_impB + X5,
                          data = final_data))$coefficients
  method.B.coef <- method.B[2, 1]
  method.B.se <- method.B[2, 2]

  # Method C
  X1_mod <- lm(X1 ~ X2 + X3 + X5, data = final_data)
  X4_mod <- glm(X4 ~ X2 + X3 + X5, data = final_data, family = "binomial")
  X1_impC <- ifelse(is.na(X1) == TRUE, predict(X1_mod), X1)
  X4_impC <- ifelse(is.na(X4) == TRUE, predict(X4_mod, type = "response"), X4)
  final_data <- data.frame(final_data, X1_impC, X4_impC)
  method.C <- summary(lm( Y ~ X1_impC + X2 + X3 + X4_impC + X5,
                          data = final_data))$coefficients
  method.C.coef <- method.C[2, 1]
  method.C.se <- method.C[2, 2]

  # Method D
  impD <- mice(final_data[, 1:6], maxit = 20, m = 1, print = FALSE)
  method.D <- summary(lm( Y ~ X1 + X2 + X3 + X4 + X5,
                          data = complete(impD)))$coefficients
  method.D.coef <- method.D[2, 1]
  method.D.se <- method.D[2, 2]
```

```
  # Method E
  impE <- mice(final_data[, 1:6], maxit = 20, m = 20, print = FALSE)
  method.E <- summary(pool(with(impE, lm( Y ~ X1 + X2 + X3 + X4 + X5))))
  method.E.coef <- method.E[2, 2]
  method.E.se <- method.E[2, 3]

  # All methods
  parm.est[q, ] <- c(method.A.coef, method.B.coef,
                     method.C.coef, method.D.coef, method.E.coef)
  se.est[q, ] <- c(method.A.se, method.B.se,
                   method.C.se, method.D.se, method.E.se)

  print(q)
}
```

a. (5 points) Remember bias is the average difference between $\hat{\beta}_1$ and β_1. What is the average bias of each of these methods?

b. (5 points) What is the standard deviation in the estimates of $\hat{\beta}_1$ for each of these 5 methods?

c. (5 points) Is the average standard deviation for $\hat{\beta}_1$ close to the value in part (b)? Why is this important?

d. (5 points) What is the average squared error ($mean\left\{(\hat{\beta}_1 - \text{truth})^2\right\}$) in the estimates of $\hat{\beta}_1$ for each of these 5 methods?

e. (5 points) What are your overall conclusions?

f. (**Extra Credit**: 12 points) Do your conclusions differ if the sample size (n) changes? Or the amount of missingness (e.g., changing the intercept in X1_obs and X4_obs models)? Or the *degree to which* the missingness depends on the other covariates? Write up your conclusions in a paragraph with supporting table.