Homework: Missing Data

Due 10 April 2020

Katherine Wolf

Questions

1. Complete **Table 1**. **(10 points)**

Table 1: Descriptive statistics (means/proportions) for each fully observed covariate by indicator of whether BMI is observed or missing.

Covariate	BMI Missing	BMI Observed
N (%)	154 (31%)	346 (69%)
Hypertension, n (%)	99 (64.3)	251 (72.5)
Age, mean (sd)	46.3 (7.9)	51.0 (8.9)
Male, n (%)	3 (1.9%)	200 (57.8%)
Smoker, n (%)	83 (53.9%)	162 (46.8%)

2. Complete **Table 2**. **(40 points)**

Table 2: Parameter values ($\hat{\beta}$) and standard errors (in parentheses) from logistic regression model of hypertension on BMI, age, male, and smoking applying several missing data methods.

Model Coefficient	Full Analysis	Complete Case Analysis	IP Weighting	Multiple Imputation	Fully Bayesian
Intercept	-6.4149 (1.0694)	-6.1744 (1.3391)	-6.3174 (1.4915)	-6.3073 (1.0417)	-7.1258 (1.0694)
BMI	0.1895 (0.0332)	0.2053 (0.0412)	0.1834 (0.0461)	0.1843 (0.0314)	0.2107 (0.0332)
Age	0.0520 (0.0127)	0.0418 (0.0149)	0.0543 (0.0169)	0.0528 (0.0126)	0.0558 (0.0127)
Male	-0.1226 (0.2192)	-0.0935 (0.2773)	0.0985 (0.2894)	-0.0645 (0.2162)	-0.0772 (0.2192)
Smoker	0.1047 (0.2217)	-0.1231 (0.2726)	-0.2907 (0.3097)	0.0039 (0.2168)	0.0305 (0.2217)

3. State and describe the necessary assumption for the missing data mechanism (i.e. what does the probability of missing depend upon?) for the *complete case analysis* to be valid. Given the output in **Table 1** do you think a complete case analysis would be appropriate here? **(10 points)**

   For the complete case analysis to be valid, values for BMI must either (1) be missing completely at random (MCAR) or (2) missing at random (MAR) with the probability of missingness also independent of the outcome.

   Formally:

   - Let Y be the random variable representing the outcome, here, prevalent hypertension;
   - Let A be the random variable representing the exposure, here, BMI;
   - Let $\mathbf{W}$ be a vector of random variables representing the covariates (here, age, being male, and smoking);
   - Let $\mathbf{w^m}$ indicate missing values in the covariates $\mathbf{W}$;
   - Let $\mathbf{w^o}$ indicate observed values in the covariates $\mathbf{W}$; and
   - Let R_A be the binary indicator for missingness of the exposure A (here, missing values of BMI), with missingness indicated by $R_A = 1$ and non-missingness indicated by $R_A = 0$.

   MCAR then means that $P(R_A = 0|Y, A, \mathbf{W}) = P(R_A = 0)$, i.e., the probability of whether the value for A is missing does not depend on the values of any data, including the outcome Y, the exposure itself A, any of the covariates $\mathbf{W}$, or any unobserved variables.

   MAR then means that the probability of missingness of values of the variable of interest depends on the observed but only the observed data (and, notably, *not* on the value of the variable of interest itself), i.e., that $P(R_A = 0|Y, A, \mathbf{W}) = P(R_A = 0|Y, \mathbf{w^o})$.

   Finally, the probablity of missingness of A being independent of the outcome Y means that $R \perp Y$, i.e., $P(R_A = 0|Y, A, \mathbf{W}) = P(R_A = 0|A, \mathbf{W})$.

   Thus the assumptions of both MAR and $R \perp Y$ required for complete case analysis to be unbiased can be represented together as $P(R_A = 0|Y, A, \mathbf{W}) = P(R_A = 0|\mathbf{w^o})$.

   Complete case analysis is not appropriate here. Table 1 shows that the data are likely not MCAR, as only 2% of those with missing BMI but 58% of those with observed BMI are male. Thus the missingness seems at least to depend on sex, violating the MCAR assumption. Whether the data are MAR and depend only on observed variables, or are NMAR and depend on either the outcome or unobserved values of other variables (here, they would have to be variables not included in the analysis at all, since the values for all the variables except BMI are completely observed), is hard to tell just from looking at a table of only the observed data. That said, though, 31% of the BMI values (and none of the other values for any of the other variables!) are missing, so even if the data are MAR and the missingness is independent of the outcome and using complete case analysis would thus be unbiased, the precision of the estimate of the outcome would suffer from throwing out 31% of the values of all the other variables in the analysis. Thus another method that accounts for the missingness but allows the use of the observed values of the other variables for the participants with missing BMI will ultimately produce at least more precise coefficient estimates with smaller standard errors (and probably also a less biased one, closer to the truth).

4. State and describe the necessary assumption (again, regarding the probability of missing) for the missing data mechanism for the *IPW*, *multiple imputation* and *fully Bayesian* analyses to be valid. Do you think these are appropriate in this case? **(10 points)**

   For IPW, multiple imputation, and fully Bayesian analyses to be valid, the data must be MAR, i.e., the probability of missingness can depend on the observed but only the observed data (and, notably, *not* on the value of the variable with missing values itself). Formally, using the abbrevations above, we assume that $P(R_A = 0|Y, A, \mathbf{W}) = P(R_A = 0|Y, \mathbf{w^o})$. Whether the data are MAR is hard to tell from looking at the data itself. In particular, whether the missingness of the BMI value is associated with that value is impossible to check using only the observed BMI values! That said, however, these data are more likely to meet the necessary assumptions for these approaches to result in unbiased estimates than the more stringent assumptions required for complete case analysis to result in unbiased estimates, and with 31% of the BMI values missing, we have to account for the missingness somehow. So to get the best possible estimates after the data have already gone missing, these approaches are the least likely to result in biased estimates. (Moreover, the Bayesian approach does not always require the MAR assumption and can sometimes also accomodate situations where the data are not missing at random, i.e., where the missingness depends on unobserved values of variables or entirely unobserved variables.)

5. How do the *IPW* and *imputation/Bayesian* methods vary in terms of what is being modeled? When might you prefer to use one method over another? **(10 points)**

   The IPW method models the predicted probability of observing the value of the variable with missing values for each participant (i.e., $P(R_A = 1)$) using the observed values of the outcome and other covariates. (The method then models the outcome using only the data from the complete cases, but weighted by the inverse of the predicted probability calculated in the first step.) In contrast, the imputation and Bayesian methods directly model the missing values of the covariates with missing values using the observed values of the outcome and other covariates. (Both the imputation and Bayesian methods model the outcome using the imputed missing values; multiple imputation does this in a separate step and then repeats both steps multiple times to produce multiple models of the outcome and pools the results for the covariates, while Bayesian methods model both the missing values and the outcome simultaneously.)

   IPW relies on accurately predicting the probability of missingness, but does not require any modeling of what those missing values are, whereas multiple imputation and, to a lesser extent, Bayesian methods rely on specifying accurate parametric models for those values.

   IPW is based on theories in which sample sizes approach infinity and thus likely produces models with more biased and less precise estimates in small samples than models generated using multiple imputation or Bayesian methods. (Multiple imputation and Bayesian methods are more statistically efficient in general and thus produce more precise parameter estimates than IPW.)

   I would use IPW if I was confident that the missingness was MAR, had a large sample size, had either only one variable with missing values or missing values across variables always occuring simultaneously within participants, AND had little confidence about specifying a parametric model for the missing values.

   I would use either multiple imputation or Bayesian methods if I was confident that the missingness was MAR and if either I had a small sample or complicated patterns of missingness across variables.

   I would use Bayesian or more sophisticated methods if I suspected that my data were NMAR.

6. Answer the following using the results in Table 2:

   i. Describe the differences in the inference on the BMI-hypertension relationship across the missing data methods in terms of the magnitude of the association, and its precision, compared to the full data model (you may need to consider 4-5 decimal places). **(10 points)**

Inverse probability weighting produces the BMI estimate closest to the true estimate, whereas multiple imputation offers the smallest precision (standard error). The fully Bayesian analysis produces the least accurate estimate, followed by the complete case analysis and then multiple imputation. IPW has the worst precision (largest standard error), followed by the complete case method, the fully Bayesian method, and then the multiple imputation method.

   ii. In practice you will not have the full data model to compare to–if you only had the results from the missing data models, which would you present? Why? **(10 points)**

Honestly, I would probably run and present all four in the supplemental materials of a paper since the mechanism of missingness is impossible to verify and to be transparent about what I tried. For the main analysis I would probably have chosen multiple imputation to present because of its statistical efficiency compared the other methods, i.e., because its standard error (precision) on the outcome of interest, the BMI variable, was the smallest, and because I would have thought it likely that I met its assumptions behind that efficiency, that the missingness was at random and that the multiple imputation algorithm could specify an accurate parametric model for the missing values.

R code

```
knitr::opts_chunk$set(echo = FALSE,
                      eval = TRUE,
                      results = 'hide',
                      warning = FALSE,
                      message = FALSE,
                      global.par=TRUE)


library("tidyverse")
library("blm")
library("geepack")
library("mi")
library("R2jags")
library("coda")
library("doBy")
library("tableone")


load("BMI_HTNData.Rdata")


##### Full analysis (this is already completed in Table 2)
logistic.full <- glm(hyperten~bmi + age + male + cursmoke, data=comp_data,
            family = binomial(link="logit"))
summary(logistic.full)
########## Question 1: Complete Case Analysis
# Descriptive statistics
CreateTableOne(vars = c("hyperten",
                        "age",
                        "male",
                        "cursmoke"),
               data = comp_data,
               strata = "r",
               factorVars = c("male","cursmoke"),
               test=FALSE)
##### Complete-case analysis
completecase.htn <- glm(hyperten~bmi_missing + age + male + cursmoke,
                        data=comp_data, family = binomial(link="logit"))
summary(completecase.htn)
##### Inverse probability weighting
model.r <- glm(r~age + male + cursmoke + hyperten, family=binomial,
               data=comp_data) # Model for observed/missing
summary(model.r)
phat.r <- predict(model.r, type="response") # Predicted probability of observed
```

```r
w <- 1/phat.r # Weight according to probability of being observed

data.cc <- na.omit(as.data.frame(cbind(comp_data,w)))
# IPW for missing data:
logistic.ipw <- geeglm(hyperten~bmi_missing + age + male + cursmoke,
      family=binomial, weights = w, id=randid, data=data.cc,
      std.err='san.se', corstr="independence", scale.fix=T)
summary(logistic.ipw)

##### Multiple Imputation
# Make a data frame with all variables for missing data analysis

to.drop <- names(comp_data) %in%
     c("educ",
       "randid",
       "bmi",
       "prevhyp",
       "timehyp",
       "r")
mdf <- missing_data.frame(comp_data[!to.drop])
show(mdf)


mdf <- change(mdf,
              y = c("bmi_missing",
                    "age"),
              what = "transformation",
              to = c("identity",
                     "identity"))
show(mdf)
summary(mdf)

imputations <- mi(mdf, n.iter = 200, n.chains = 4)
round(mipply(imputations, mean, to.matrix = TRUE), 3)
Rhats(imputations)

logistic.mi <- pool(hyperten~bmi_missing + age + male + cursmoke,
                 data=imputations, family=binomial, m=10)
summary(logistic.mi)
##### Bayesian Modeling of Missing Data

logistic.model <- function() {
  # SAMPLING DISTRIBUTION
  for (i in 1:N) {
    logit(p[i]) <- b[1] + b[2]*bmi_missing[i] + b[3]*age[i] +
         b[4]*male[i] + b[5]*cursmoke[i];
    hyperten[i] ~ dbin(p[i],1);
    # DISTRIBUTION ON COVARIATE WITH MISSING DATA:
    mu.bmi[i] <- a[1]+ a[2]*age[i] + a[3]*male[i] + a[4]*cursmoke[i];
    bmi_missing[i] ~ dnorm(mu.bmi[i], tau.bmi);
  }

  # PRIOR ON TAU FROM LINEAR MODEL FOR BMI (vague uniform prior)
  tau.bmi ~ dunif(0.001,10);

  # MULTIVARIATE NORMAL PRIORS ON BETAS
  b[1:N.y] ~ dmnorm(mu.b[1:N.y],tau.b[1:N.y,1:N.y]);

  # MULTIVARIATE NORMAL PRIORS ON ALPHAS
  a[1:N.x] ~ dmnorm(mu.a[1:N.x],tau.a[1:N.x,1:N.x]);
}
N <- length(comp_data$hyperten) # Number of observations
N.y <- 5 # Number of slope parameters in model for hypertension
N.x <- 4 # Number of slope parameters in model for BMI (variable w/ missingness)

# Data, parameter list and starting values
mu.b <- rep(0,N.y)
tau.b <- diag(0.001,N.y)

mu.a <- rep(0,N.x)
tau.a <- diag(0.001, N.x)
attach(comp_data)
data.logistic <- list(N=N, N.y=N.y, N.x=N.x, hyperten=hyperten,
                      bmi_missing = bmi_missing, age=age,
                      male=male, cursmoke=cursmoke,
                      mu.b=mu.b, tau.b=tau.b, mu.a=mu.a, tau.a=tau.a)
detach(comp_data)
parameters.logistic <-c("b","a","tau.bmi") # Parameters to keep track of

# Use ML estimates for starting values for beta, alpha, and
# 1/var(residual) for tau

lm.bmi <- lm(bmi_missing ~ age + male + cursmoke, data=comp_data)
cc.bmi <- glm(hyperten ~ bmi_missing + age + male + cursmoke, data=comp_data,
                family=binomial)

inits.logistic <- function() {list (b= coefficients(cc.bmi),
                                    a=coefficients(lm.bmi),
                                    tau.bmi=1/var(lm.bmi$residuals))}
## THIS WILL TAKE A WHILE TO RUN.
set.seed(114011)
logistic.sim<-jags(data=data.logistic,inits=inits.logistic,
                   parameters.logistic,n.iter=50000,
                   model.file=logistic.model,
                   n.thin=10, n.chains = 1)
logistic.mcmc <- as.mcmc(logistic.sim)
pdf("TraceplotBayes.pdf")
plot(logistic.mcmc)
dev.off()

pdf("AutoCorrelation.pdf")
autocorr.plot(logistic.mcmc)
dev.off()
print(logistic.sim,2)
# Coefficient estimates:
cbind(coef(logistic.full),
      coef(completecase.htn),
      coef(logistic.ipw),
      coef(logistic.mi),
      summary(logistic.mcmc)$quantile[,3][paste("b[",1:N.y,"]",sep="")])

# Standard error estimates:
cbind(sqrt(diag(vcov(logistic.full))),
      sqrt(diag(vcov(completecase.htn))),
      summary(logistic.ipw)$coefficients$Std.err,
      sqrt(diag(summary(logistic.mi)$cov.scaled)),
      summary(logistic.mcmc)$statistics[,2][paste("b[",1:N.y,"]",sep="")])


table_one <- CreateTableOne(vars = c("hyperten",
                        "age",
                        "male",
                        "cursmoke"),
               data = comp_data,
               strata = "r",
               factorVars = c("male","cursmoke"),
               test=FALSE)

comp_data_tibble <-
  tibble(comp_data)

n_total <- comp_data %>%
  nrow()

n_miss <- comp_data %>%
  filter(r == 0) %>%
  nrow()

n_obs <- comp_data %>%
  filter(r == 1) %>%
  nrow()
```

```
n_p_miss <- round(n_miss / n_total, 2)

n_p_obs <- round(n_obs / n_total, 2)

htn_miss_n <- comp_data %>%
  filter(r == 0) %>%
  select(hyperten) %>%
  filter(hyperten == 1) %>%
  nrow

htn_miss_perc <- comp_data %>%
  filter(r == 0) %>%
  pull(hyperten) %>%
  mean() %>%
  `*`(100) %>%
  round(1) %>%
  sprintf("%.1f", .)

htn_obs_n <- comp_data %>%
  filter(r == 1) %>%
  select(hyperten) %>%
  filter(hyperten == 1) %>%
  nrow

htn_obs_perc <- comp_data %>%
  filter(r == 1) %>%
  pull(hyperten) %>%
  mean() %>%
  `*`(100) %>%
  round(1) %>%
  sprintf("%.1f", .)

age_miss_mean <- comp_data %>%
  filter(r == 0) %>%
  pull(age) %>%
  mean() %>%
  round(1) %>%
  sprintf("%.1f", .)

age_miss_sd <- comp_data %>%
  filter(r == 0) %>%
  pull(age) %>%
  sd() %>%
  round(1) %>%
  sprintf("%.1f", .)

age_obs_mean <- comp_data %>%
  filter(r == 1) %>%
  pull(age) %>%
  mean() %>%
  round(1) %>%
  sprintf("%.1f", .)

age_obs_sd <- comp_data %>%
  filter(r == 1) %>%
  pull(age) %>%
  sd() %>%
  round(1) %>%
  sprintf("%.1f", .)

male_miss_count <- comp_data %>%
  filter(r == 0) %>%
  select(male) %>%
  filter(male == 1) %>%
  nrow()

male_miss_perc <- comp_data %>%
  filter(r == 0) %>%
  pull(male) %>%
  mean() %>%
  `*`(100) %>%
  round(1) %>%
  sprintf("%.1f", .)

male_obs_count <- comp_data %>%
  filter(r == 1) %>%
  select(male) %>%
  filter(male == 1) %>%
  nrow()

male_obs_perc <- comp_data %>%
  filter(r == 1) %>%
  pull(male) %>%
  mean() %>%
  `*`(100) %>%
  round(1) %>%
  sprintf("%.1f", .)

smoke_miss_count <- comp_data %>%
  filter(r == 0) %>%
  select(cursmoke) %>%
  filter(cursmoke == 1) %>%
  nrow()
```

```
smoke_miss_perc <- comp_data %>%
  filter(r == 0) %>%
  pull(cursmoke) %>%
  mean() %>%
  `*`(100) %>%
  round(1) %>%
  sprintf("%.1f", .)

smoke_obs_count <- comp_data %>%
  filter(r == 1) %>%
  select(cursmoke) %>%
  filter(cursmoke == 1) %>%
  nrow()

smoke_obs_perc <- comp_data %>%
  filter(r == 1) %>%
  pull(cursmoke) %>%
  mean() %>%
  `*`(100) %>%
  round(1) %>%
  sprintf("%.1f", .)


# Coefficient estimates:
coefficients <-
  cbind(coef(logistic.full),
        coef(completecase.htn),
        coef(logistic.ipw),
        coef(logistic.mi),
        summary(logistic.mcmc)$quantile[,3][paste("b[",1:N.y,"]",sep="")])

# Standard error estimates:
standard_errors <-
  cbind(sqrt(diag(vcov(logistic.full))),
        sqrt(diag(vcov(completecase.htn))),
        summary(logistic.ipw)$coefficients$Std.err,
        sqrt(diag(summary(logistic.mi)$cov.scaled)),
        summary(logistic.mcmc)$statistics[,2][paste("b[",1:N.y,"]",sep="")])


fa_int_cf <- sprintf("%.4f", round(coefficients[[1,1]], 4))
fa_int_se <- sprintf("%.4f", round(standard_errors[[1,1]], 4))

fa_bmi_cf <- sprintf("%.4f", round(coefficients[[2,1]], 4))
fa_bmi_se <- sprintf("%.4f", round(standard_errors[[2,1]], 4))

fa_age_cf <- sprintf("%.4f", round(coefficients[[3,1]], 4))
```

```
fa_age_se <- sprintf("%.4f", round(standard_errors[[3,1]], 4))

fa_male_cf <- sprintf("%.4f", round(coefficients[[4,1]], 4))
fa_male_se <- sprintf("%.4f", round(standard_errors[[4,1]], 4))

fa_smoke_cf <- sprintf("%.4f", round(coefficients[[5,1]], 4))
fa_smoke_se <- sprintf("%.4f", round(standard_errors[[5,1]], 4))


cc_int_cf <- sprintf("%.4f", round(coefficients[[1,2]], 4))
cc_int_se <- sprintf("%.4f", round(standard_errors[[1,2]], 4))

cc_bmi_cf <- sprintf("%.4f", round(coefficients[[2,2]], 4))
cc_bmi_se <- sprintf("%.4f", round(standard_errors[[2,2]], 4))

cc_age_cf <- sprintf("%.4f", round(coefficients[[3,2]], 4))
cc_age_se <- sprintf("%.4f", round(standard_errors[[3,2]], 4))

cc_male_cf <- sprintf("%.4f", round(coefficients[[4,2]], 4))
cc_male_se <- sprintf("%.4f", round(standard_errors[[4,2]], 4))

cc_smoke_cf <- sprintf("%.4f", round(coefficients[[5,2]], 4))
cc_smoke_se <- sprintf("%.4f", round(standard_errors[[5,2]], 4))


ipw_int_cf <- sprintf("%.4f", round(coefficients[[1,3]], 4))
ipw_int_se <- sprintf("%.4f", round(standard_errors[[1,3]], 4))

ipw_bmi_cf <- sprintf("%.4f", round(coefficients[[2,3]], 4))
ipw_bmi_se <- sprintf("%.4f", round(standard_errors[[2,3]], 4))

ipw_age_cf <- sprintf("%.4f", round(coefficients[[3,3]], 4))
ipw_age_se <- sprintf("%.4f", round(standard_errors[[3,3]], 4))

ipw_male_cf <- sprintf("%.4f", round(coefficients[[4,3]], 4))
ipw_male_se <- sprintf("%.4f", round(standard_errors[[4,3]], 4))

ipw_smoke_cf <- sprintf("%.4f", round(coefficients[[5,3]], 4))
ipw_smoke_se <- sprintf("%.4f", round(standard_errors[[5,3]], 4))


mi_int_cf <- sprintf("%.4f", round(coefficients[[1,4]], 4))
mi_int_se <- sprintf("%.4f", round(standard_errors[[1,4]], 4))

mi_bmi_cf <- sprintf("%.4f", round(coefficients[[2,4]], 4))
mi_bmi_se <- sprintf("%.4f", round(standard_errors[[2,4]], 4))
```

```
mi_age_cf <- sprintf("%.4f", round(coefficients[[3,4]], 4))
mi_age_se <- sprintf("%.4f", round(standard_errors[[3,4]], 4))

mi_male_cf <- sprintf("%.4f", round(coefficients[[4,4]], 4))
mi_male_se <- sprintf("%.4f", round(standard_errors[[4,4]], 4))

mi_smoke_cf <- sprintf("%.4f", round(coefficients[[5,4]], 4))
mi_smoke_se <- sprintf("%.4f", round(standard_errors[[5,4]], 4))


fb_int_cf <- sprintf("%.4f", round(coefficients[[1,5]], 4))
fb_int_se <- sprintf("%.4f", round(standard_errors[[1,5]], 4))

fb_bmi_cf <- sprintf("%.4f", round(coefficients[[2,5]], 4))
fb_bmi_se <- sprintf("%.4f", round(standard_errors[[2,5]], 4))

fb_age_cf <- sprintf("%.4f", round(coefficients[[3,5]], 4))
fb_age_se <- sprintf("%.4f", round(standard_errors[[3,5]], 4))

fb_male_cf <- sprintf("%.4f", round(coefficients[[4,5]], 4))
fb_male_se <- sprintf("%.4f", round(standard_errors[[4,5]], 4))

fb_smoke_cf <- sprintf("%.4f", round(coefficients[[5,5]], 4))
fb_smoke_se <- sprintf("%.4f", round(standard_errors[[5,5]], 4))
```