Homework 4: Missing Data

Due Thursday 18 April 2024 at 11:59pm via Gradescope

Read all questions carefully before answering. You may work in small groups of no more than 3 individuals and turn in a single assignment (and everyone in the group will receive the same grade). Work through the entire assignment individually first, then come together to discuss and collaborate. Please maintain numbering on sub-questions, type your responses, and **please keep answers brief.**

Load required packages and read data:

```
library("geepack")
library("mi")
library("R2jags")
library("coda")
library("doBy")
library("tableone")

load("CVD_dataHW4.Rdata")
```

We will estimate the relationship between smoking (`cursmoke`) and incident CVD with a modification of the dataset from the previous homework assignment. We will explore several approaches to dealing with missing data with a version of the `cursmoke` variable that has some values missing.

Full Model

Estimate the model for the full analysis (if you did not have any missing data), adjusting for age, sex, and education level:

```
##### Full analysis (this is already completed in Table 2)
logistic.full <- glm(cvd ~ cursmoke + age + factor(sex) + factor(educ),
                     data=CVD.data.miss,
                     family = binomial(link="logit"))
summary(logistic.full)
```

Missing Data Models

Describe the pattern of missingness

The dataset contains an indicator (`r`) of if smoking status is observed (`r=1`) or missing (`r=0`), and the smoking variable with missing values is called `cursmoke.miss`. Calculate descriptive statistics for the fully-observed covariates, by missingness of `cursmoke`:

```
########## Question 1: Complete Case Analysis
# Descriptive statistics
CreateTableOne(vars=c("cvd","age", "sex", "cursmoke", "educ"),
               data=CVD.data.miss, strata="r",
               factorVars =c("cvd","sex","cursmoke", "educ"), test=FALSE)
```

Complete-case analysis

Next, fit a logistic regression model for the CVD-smoking relationship using the `cursmoke` variable **with missing values**. (Note the message at the bottom of the summary that indicates how many observations were deleted due to missing values.)

```
##### Complete-case analysis
logistic.cc <- glm(cvd ~ cursmoke.miss + age + factor(sex) + factor(educ),
                     data=CVD.data.miss,
                     family = binomial(link="logit"))
summary(logistic.cc)
```

Inverse probability weighting

Next, use inverse probability weighting (IPW) to account for the probability smoking status is missing:

1. Fit a logistic regression model for the probability smoking status is observed (indicated by `r`) as a function of the fully observed covariates and the outcome (`cvd`):

```
##### Inverse probability weighting
model.r <- glm(r ~ age + factor(sex) + factor(educ) + cvd, family=binomial,
               data=CVD.data.miss) # Model for observed/missing
summary(model.r)
```

2. Estimate predicted probabilities of observed:

```
phat.r <- predict(model.r, type="response") # Predicted probability of observed
w <- 1/phat.r # Weight according to probability of being observed

summary(w) # Weights seem well-behaved (nothing too large/too small)
data.cc <- na.omit(as.data.frame(cbind(CVD.data.miss,w)))
```

3. Fit a logistic regression for CVD on the observed smoking values, and other covariates, weighted by the inverse of the probability of being observed:

```
# IPW for missing data:
data.cc$id <- seq(1:nrow(data.cc)) # Create ID variable for GEE function

logistic.ipw <- geeglm(cvd ~ cursmoke.miss + age + factor(sex) + factor(educ),
                       family=binomial, weights = w, id=id, data=data.cc,
                       std.err='san.se', corstr="independence", scale.fix=T)
summary(logistic.ipw)
```

Multiple Imputation

Use multiple imputation to account for missing smoking values.

1. Create a missing data frame:

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

# Create data frame omitting the true smoking variable and missingness indicator:
comp_data_subset <- subset(CVD.data.miss,
                           select= -c(cursmoke,r))
mdf <- missing_data.frame(comp_data_subset)
show(mdf)
```

The `mi` package automatically standardizes continuous variables–let's override that (`identity` option means no transformation):

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

2. Impute 3 chains from the posterior predictive distribution, and examine convergence of the chains:

```
imputations <- mi(mdf, seed=123, n.chains=3, n.iter=100, parallel=FALSE)
round(mipply(imputations, mean, to.matrix = TRUE), 3)
Rhats(imputations)
```

3. Fit a logistic regression for incident CVD on 20 imputed datasets and pool the results:

```
logistic.mi <- pool(cvd ~ cursmoke.miss + age + factor(sex) + factor(educ),
                    data=imputations, family=binomial, m=20)
summary(logistic.mi)
```

Fully Bayesian Model

Finally, we will use a Fully Bayesian approach to missing data by specifying priors on the regression parameters (as in our previous Bayesian analyses), but also an additional model on the variable with missing values.

1. First we specify the outcome model of interest–a **logistic regression** model for CVD status as a function of smoking status, age, sex, and education:

$$\begin{aligned} \text{logit}(\pi_{cvd,i}) &= \beta_1 + \beta_2\, cursmoke_i + \beta_3\, age_i + \beta_4\, male_i + \beta_5\, educ_{2,i} + \beta_6\, educ_{3,i} + \beta_7\, educ_{4,i} \\ y_i &\sim \text{Binomial}(\pi_i, 1) \end{aligned}$$

   where y_i is the binary indicator of incident CVD.

2. Second, we specify the **logistic regression model** for the distribution on the binary covariate with missing values (`cursmoke`) as a function of age, sex, and education, along with the

age-by-sex interaction:

$$\begin{aligned} \text{logit}\pi_{cursmoke,i} &= \alpha_1 + \alpha_2\, age_i + \alpha_3\, male_i + \alpha_4\, educ_{2,i} + \alpha_5\, educ_{3,i} + \alpha_6\, educ_{4,i} + \\ &\quad \alpha_7\, age_i \times male_i \\ cursmoke_i &\sim \text{Binomial}(\pi_{cursmoke,i}, 1) \end{aligned}$$

3. Finally, **vague priors for the parameters** in these models:

$$\begin{aligned} \beta_i &\sim \text{N}(0, \tau_\beta = 0.001) \text{ for } i = 1, \ldots, 7 \\ \alpha_i &\sim \text{N}(0, \tau_\alpha = 0.001) \text{ for } i = 1, \ldots 6 \end{aligned}$$

Complete the following tasks:

1. Specify the JAGS model:

```
##### Bayesian Modeling of Missing Data

logistic.model <- function() {
  # SAMPLING DISTRIBUTION
  for (i in 1:N) {
    logit(p[i]) <- b[1] + b[2]*cursmoke.miss[i] + b[3]*age[i] +
                   b[4]*male[i] + b[5]*educ.2[i] + b[6]*educ.3[i] + b[7]*educ.4[i];
    cvd[i] ~ dbin(p[i],1);

    # DISTRIBUTION ON COVARIATE WITH MISSING DATA:
    logit(p.cursmoke[i]) <- a[1]+ a[2]*age[i] + a[3]*male[i] +
                            a[4]*educ.2[i] + a[5]*educ.3[i] + a[6]*educ.4[i] +
                            a[7]*age[i]*male[i];
    cursmoke.miss[i] ~ dbin(p.cursmoke[i], 1);
  }

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

  # VAGUE NORMAL PRIORS ON ALPHAS
  a[1:N.x] ~ dmnorm(mu.a[1:N.x], tau.a[1:N.x,1:N.x]);
}
```

2. Specify the parameters needed for the model (number of observations, number of parameters in each regression model, values for the hyperparameters of the priors):

```
N <- nrow(CVD.data.miss) # Number of observations
N.y <- 7 # Number of parameters in model for cvd
N.x <- 7 # Number of parameters in model for cursmoke.miss (variable w/ missingness)

# Create indicator variable for education variable:
X <- model.matrix(~ factor(educ)-1, data=CVD.data.miss)
educ.1 <- X[,1] # Unused in model
```

```
educ.2 <- X[,2]
educ.3 <- X[,3]
educ.4 <- X[,4]

# Data, parameter list and starting values
mu.b <- rep(0,N.y) # Vector of 0's for means
tau.b <- diag(0.01,N.y) # Diagonal matrix for variance-covariances

mu.a <- rep(0,N.x)
tau.a <- diag(0.01, N.x)
```

3. Construct the lists required by JAGS:

```
data.logistic <- list(N=N, N.y=N.y, N.x=N.x,
                      cvd=CVD.data.miss$cvd,
                      cursmoke.miss = CVD.data.miss$cursmoke.miss,
                      age=CVD.data.miss$age,
                      male=as.integer(CVD.data.miss$sex=="male"),
                      educ.2=educ.2,
                      educ.3=educ.3,
                      educ.4=educ.4,
                      mu.b=mu.b, tau.b=tau.b,
                      mu.a=mu.a, tau.a=tau.a)
parameters.logistic <-c("b","a") # Parameters to keep track of
```

4. Sample from the posterior using JAGS, with 20000 total samples, a burn-in of 10000 samples and thinning every 5th iteration. Here we will use 3 chains.

```
## THIS WILL TAKE A WHILE TO RUN.
logistic.sim<-jags.parallel(data=data.logistic,
                    parameters.to.save=parameters.logistic,
                    n.iter=20000,
                    model.file=logistic.model,
                    n.thin=5, n.chains = 3,
                    jags.seed=114011)

# Convert results to MCMC object:
logistic.mcmc <- as.mcmc(logistic.sim)
```

5. Visually assess convergence:

```
pdf("TraceplotBayes.pdf")
plot(logistic.mcmc)
dev.off()

pdf("AutoCorrelation.pdf")
autocorr.plot(logistic.mcmc)
dev.off()
```

6. Print results:

```
print(logistic.sim,2)
```

Bringing the results together:

```
# Coefficient estimates:
beta.full <- coef(logistic.full)
beta.cc <- coef(logistic.cc)
beta.ipw <- coef(logistic.ipw)
beta.mi <- coef(logistic.mi)
beta.bayes <- summary(logistic.mcmc)$quantile[,3][paste("b[",1:N.y,"]",sep="")]

cbind(beta.full, beta.cc, beta.ipw, beta.mi, beta.bayes)

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

cbind(se.full, se.cc, se.ipw, se.mi, se.bayes)
```

Questions

1. Complete **Table 1** (next page). **(5 points)**
2. Complete **Table 2** (next page). **(20 points)**
3. State and describe the necessary assumption for the missing data mechanism for the *complete case analysis* to be unbiased/valid (*i.e.* what does the probability of missing depend upon?). Given the output in **Table 1** do you think a complete case analysis would be appropriate here? **(5 points)**
4. State and describe the necessary assumption for the missing data mechanism for the *IPW*, *multiple imputation* and *Fully Bayesian* analyses to be unbiased/valid Do you think these are appropriate in this case? **(5 points)**
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? **(5 points)**
6. Answer the following using the results in Table 2:
    i. Describe the differences in the inference on the CVD-smoking 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)**
    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)**

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

Covariate	Smoking Missing	Smoking Observed
N		
CVD		
Age		
Male		
HS grad		
Some college		
College grad		

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

Model Coefficient	Full Analysis	Complete Case Analysis	IP Weighting	Multiple Imputation	Fully Bayesian
Intercept	-3.7616				
	(0.28234)				
Smoking	0.2546				
	(0.08442)				
Age	0.0458				
	(0.00492)				
Male	0.3042				
	(0.08289)				
HS grad	-0.2813				
	(0.09745)				
Some college	-0.4917				
	(0.12398)				
College grad	-0.5185				
	(0.14267)				