WEB MATERIAL

"Inverse Probability Weighted Estimation for Monotone and Nonmonotone Missing Data"

BaoLuo Sun, Neil J. Perkins, Stephen R. Cole, Ofer Harel, Emily Mitchell, Enrique F. Schisterman and Eric J. Tchetgen Tchetgen

Web Appendix 1

Using the notation developed in the main manuscript, the missing data model corresponding to the 8 nonmonotone missing data patterns, $R = 1, 2, ..., 8$, for each CPP data set (as shown in Web Table 1) is

$$\begin{aligned}
\pi_2\,(age, sa, smk, blk, otr; \gamma_2) &= \left\{1 + \exp\left[-\left(\gamma_2(age, sa, smk, blk, otr)^T\right)\right]\right\}^{-1} \\
\pi_3\,(bmi, age, smk, blk, otr; \gamma_3) &= \left\{1 + \exp\left[-\left(\gamma_3(bmi, age, smk, blk, otr)^T\right)\right]\right\}^{-1} \\
\pi_4\,(age, smk, blk, otr; \gamma_4) &= \left\{1 + \exp\left[-\left(\gamma_4(age, smk, blk, otr)^T\right)\right]\right\}^{-1} \\
\pi_5\,(bmi, age, sa, blk, otr; \gamma_5) &= \left\{1 + \exp\left[-\left(\gamma_5(bmi, age, sa, blk, otr)^T\right)\right]\right\}^{-1} \\
\pi_6\,(age, sa, blk, otr; \gamma_6) &= \left\{1 + \exp\left[-\left(\gamma_6(age, sa, blk, otr)^T\right)\right]\right\}^{-1} \\
\pi_7\,(bmi, age, blk, otr; \gamma_7) &= \left\{1 + \exp\left[-\left(\gamma_7(bmi, age, blk, otr)^T\right)\right]\right\}^{-1} \\
\pi_8\,(age, blk, otr; \gamma_8) &= \left\{1 + \exp\left[-\left(\gamma_8(age, blk, otr)^T\right)\right]\right\}^{-1} \\
\pi_1 &= 1 - \sum_{m=2}^{8} \pi_m\,(\gamma_m)\,,
\end{aligned} \tag{1}$$

where *(bmi, age, sa, smk, blk, otr)* correspond to the variables body mass index, age, spontaneous abortion, smoking, black and other races respectively. v^T denotes the transpose of vector v and γ_j denotes the vector of parameters for the j^{th} missing data pattern. The observed log-likelihood corresponding to missing data model 1 is

$$\ell_N(\gamma) = \sum_{i=1}^{N} \left\{ \left[\sum_{m=2}^{8} I(R_i = m) \log \pi_m\,(\gamma_m) \right] + I(R_i = 1) \log \left[1 - \sum_{k=2}^{8} \pi_k\,(\gamma_k) \right] \right\}, \tag{2}$$

with the observable complete-case constraints given by

$$I(R_i = 1)\left\{\sum_{m=2}^{8} \pi_m\left(\gamma_m\right)\right\} < 1 - \sigma^* \text{ for } i = 1, 2, ..., N. \tag{3}$$

The posterior samples for γ are drawn from a distribution proportional to

$$f\left(\gamma\right) \exp\left\{\ell_N(\gamma)\right\} I(R_i = 1) I\left\{\sum_{m=2}^{8} \pi_m\left(\gamma_m\right) < 1 - \sigma^*\right\}, \tag{4}$$

where we specify independent diffuse prior distributions $N(0, 10^2)$ for $f\left(\gamma\right)$. The OpenBUGS code for sampling from the posterior distribution proportional to the expression given in 4 is shown below.

Individuals are assigned to their respective missing data patterns, $R_k = 1, k = 1, ..., 8$, with $R_1 = 1$ denoting complete-cases. For a person with missing data pattern $R_h = 1$, the encoding follows that $R_j = 0$ for $j \neq h$. The first part of the code describes the contribution of each missing data probability of the N individuals to the likelihood function corresponding to the missing data model. Then the diffuse prior distributions for the parameters in the missing data model are specified as independent $N(0, 10^2)$. Finally, constraints are imposed on the N_C complete-cases with user-defined σ^*, where the input data set is ordered such that the first N_C individuals are complete-cases.

Web Table 1: Proportion of the 8 missing data patterns in the 3 CPP data sets.

Missing Data Pattern†							**Proportion (%) Data Set**		
R	**BMI**	**Age**	**SA**	**Smoke**	**Black**	**Other**	**I**	**II**	**III**
1	1	1	1	1	1	1	61.1	62.3	52.9
2	0	1	1	1	1	1	5.3	5.2	5.3
3	1	1	0	1	1	1	1.2	0.9	1.2
4	0	1	0	1	1	1	11.1	10.6	11.3
5	1	1	1	0	1	1	5.5	5.3	7.2
6	0	1	1	0	1	1	5.2	5.0	7.6
7	1	1	0	0	1	1	9.8	9.8	13.8
8	0	1	0	0	1	1	0.9	1.0	0.9

Abbreviations: BMI, body mass index; SA, spontaneous abortion.
† Observed variables are denoted by 1 and missing variables by 0.

OpenBUGS Code

```
Model <- function() {

for (i in 1:N){
z[i]   <- 1
z[i]    ~ dbern(p[i])
p[i]   <- L[i]
L[i]   <-  R1[i]*pi1[i]
          +R2[i]*pi2[i]
          +R3[i]*pi3[i]
          +R4[i]*pi4[i]
          +R5[i]*pi5[i]
          +R6[i]*pi6[i]
          +R7[i]*pi7[i]
          +R8[i]*pi8[i]

#Probability for each missing data pattern

logit(pi2[i])<-g[1]+g[2]*age[i]+g[3]*sa[i]+g[4]*smk[i]+g[5]*blk[i]+g[6]*otr[i]
logit(pi3[i])<-g[7]+g[8]*bmi[i]+g[9]*age[i]+g[10]*smk[i]+g[11]*blk[i]+g[12]*otr[i]
logit(pi4[i])<-g[13]+g[14]*age[i]+g[15]*smk[i]+g[16]*blk[i]+g[17]*otr[i]
logit(pi5[i])<-g[18]+g[19]*bmi[i]+g[20]*age[i]+g[21]*sa[i]+g[22]*blk[i]+g[23]*otr[i]
logit(pi6[i])<-g[24]+g[25]*age[i]+g[26]*sa[i]+g[27]*blk[i]+g[28]*otr[i]
logit(pi7[i])<-g[29]+g[30]*bmi[i]+g[31]*age[i]+g[32]*blk[i]+g[33]*otr[i]
logit(pi8[i])<-g[34]+g[35]*blk[i]+g[36]*otr[i]+g[37]*age[i]
pi1[i]       <-1-pi2[i]-pi3[i]-pi4[i]-pi5[i]-pi6[i]-pi7[i]-pi8[i]
}

#Priors for parameters in missing data model
for (j in 1:37) {
        g[j] ~ dnorm(0, 0.01)
}

# implementing the constraints for complete-cases
for (k in 1:N_C){
          ones[k] <- 1
          ones[k] ~ dbern(C[k])
          C[k] <- step(pi1[k]-sigma_star)
}
}
```

The above missing data model can be run directly in OpenBUGs, which also provides summary statistics for each parameter. Alternatively, the posterior mean, median and 95% credible intervals can also be obtained directly from Markov-chain Monte-Carlo sampling in R through BRugs, an interface to the OpenBUGS software. The above missing data model can be specified directly in R as a function. A sample R code to estimate the missing data model parameters is as follows:

```
library("BRugs")

writeModel(Model, model_file) #write missing data model to user-specified file
```

```
bugsData(Data, data_file)     #write input data set to user-specified file

modelCheck(model_file)     # check model file
modelData(data_file)       # read data file
modelCompile(numChains=k)  # compile model with k chains
modelInits(inits_file)     # read initialization parameter values

modelUpdate(20000)         # burn in of 20000 iterations
samplesSet("g")            # monitor parameters (array ‘‘g’’ in model).
modelUpdate(20000)         # 20000 more iterations, sampled values are recorded.

samplesStats("g")$mean     # posterior mean and median values
samplesStats("g")$median
samplesStats("g")$val2.5pc # Quantiles for constructing 95% credible intervals
samplesStats("g")$val97.5pc
```

Web Appendix 2

Let *(bmi, age, sa, smk, blk, otr)* denote the variables body mass index, age, spontaneous abortion, smoking, black and other races respectively. Then the substantive model of interest is given by

$$\Pr\left(sa = 1|bmi, age, smk, blk, otr; \beta\right) = \left\{1 + \exp\left[-\left(\beta(1, bmi, age, smk, blk, otr)^T\right)\right]\right\}^{-1} = \mu(\beta), \tag{5}$$

where $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ and our interest lies primarily in estimating β_3, the association between smoking and spontaneous abortion while adjusting for other potential confounders. Let $g(\cdot;\beta)$ be an arbitrary 6×1 dimensional function of the variables $(bmi, age, smk, blk, otr)$, for instance $g(\cdot;\beta) = (1, bmi, age, smk, blk, otr)^T$. In addition let $U_g(\beta) = g(\cdot;\beta)\mu(\beta)$. The class of AIPW estimators are given by solutions to the estimating equations

$$N^{-1}\sum_i^N \left\{\frac{I(R_i = 1)}{\pi_1(\hat{\gamma})}U_g(\beta) + \sum_{r=2}^{8}\left[\frac{I(R_i = 1)}{\pi_1(\hat{\gamma})} - \frac{I(R_i = r)}{\pi_r(\hat{\gamma})}\right]\phi_r(L_{(r),i})\right\} = 0, \tag{6}$$

where $\phi_r(L_{(r)})$ is an arbitrary 6×1 dimensional function of the observed variables $L_{(r)}$ corresponding to missing data pattern $R = r$ (1). The observed variables for each missing data pattern are tabulated in Web Table 1; the estimates $\hat{\gamma}$ are obtained as posterior medians of samples from distribution 4 described in the previous section. Note that there is an optimal choice of $g(\cdot;\beta)$, denoted as $g^{\text{opt}}(\cdot;\beta)$, as well as an optimal choice for each of $\phi_r(L_{(r)})$, $r = 2, 3, ..., 8$, denoted as $\phi_r^{\text{opt}}(L_{(r)})$, such that the solution to estimating equation 6 is the most efficient asymptotically. For nonmonotone missing data such as the CPP data sets, these optimal functions are generally not available in closed forms and involve complicated integrals that are computationally challenging to solve. Instead, as described in the main manuscript, the proposed approach entails approximating the optimal functions by finite sums of basis functions. To approximate $g^{\text{opt}}(\cdot;\beta)$, we construct the

6×1 dimensional vector

$$P_{6\times 9}(1, smk, bmi, age, blk, otr, bmi^2, age^2, bmi \times age)^T = P_{6\times 9}\mathbb{G}^*_{9\times 1}(\beta), \tag{7}$$

where $P_{6\times 9}$ is a 6×9 constant matrix. To approximate the augmentation term in 6, we construct the following $\phi^*_r(L_{(r)})$ for each $r = 2, 3, ..., 8$:

$$\begin{aligned}
\phi^*_2 =& (1, age, sa, smk, blk, otr, age^2, age \times sa, age \times smk, age \times blk, age \times otr, age \times smk, \\
& sa \times blk, smk \times blk, smk \times otr) \\
\phi^*_3 =& (1, bmi, age, smk, blk, otr, bmi^2, age^2, bmi \times age, bmi \times smk, bmi \times blk, bmi \times otr, \\
& age \times smk, age \times blk, age \times otr, smk \times blk, smk \times otr) \\
\phi^*_4 =& (1, age, smk, blk, otr, age^2, age \times smk, age \times blk, age \times otr, blk \times smk, otr \times smk) \\
\phi^*_5 =& (1, bmi, age, sa, blk, otr, bmi^2, age^2, bmi \times age, bmi \times sa, bmi \times blk, bmi \times otr, \\
& age \times sa, age \times blk, age \times otr) \\
\phi^*_6 =& (1, age, sa, blk, otr, age^2, age \times sa, age \times blk, age \times otr) \\
\phi^*_7 =& (1, bmi, age, blk, otr, bmi^2, age^2, bmi \times age, bmi \times blk, bmi \times otr, age \times blk, age \times otr) \\
\phi^*_8 =& (1, blk, otr, age, age^2, age \times blk, age \times otr).
\end{aligned} \tag{8}$$

Let $\mathbb{A}^*$ be the 86×1 vector

$$\mathbb{A}^*_{86\times 1} = \left\{ \left[\frac{I(R_i = 1)}{\pi_1(\hat{\gamma})} - \frac{I(R_i = 2)}{\pi_2(\hat{\gamma})} \right] \phi^*_2, ..., \left[\frac{I(R_i = 1)}{\pi_1(\hat{\gamma})} - \frac{I(R_i = 8)}{\pi_8(\hat{\gamma})} \right] \phi^*_8 \right\}^T. \tag{9}$$

Then the class of AIPW estimators "approximated" by basis functions is given by the solution to the estimating equation.

$$N^{-1} \sum_i^N \left\{ \frac{I(R_i = 1)}{\pi_1(\hat{\gamma})} P_{6\times 9}\mathbb{G}^*_{9\times 1,i}(\beta)\mu_i(\beta) + Q_{6\times 86}\mathbb{A}^*_{86\times 1,i} \right\} = 0, \tag{10}$$

where $Q_{6\times 86}$ is a constant matrix of dimensions 6×86. The optimal choices of the constant matrices $P_{6\times 9}$ and $Q_{6\times 86}$ in 10 remains to be estimated. Using a result due to Tsiatis (2), the optimal choices $(P^{\text{opt}}_{6\times 9}, Q^{\text{opt}}_{6\times 86})$ within the class 10 are given by the solution to

$$\left[P^{\text{opt}}_{6\times 9}, Q^{\text{opt}}_{6\times 86}\right] \begin{bmatrix} U_{11} & U_{12} \\ U_{12}^T & U_{22} \end{bmatrix} = [H_1, H_2]\,, \tag{11}$$

where

$$U_{11} = E\left\{\frac{\mathbb{G}^*_{9\times 1}(\beta)\mu(\beta)\mu(\beta)^T\mathbb{G}^{*T}_{9\times 1}(\beta)}{\pi_1(L)}\right\}^{9\times 9}$$

$$U_{12} = E\left\{\frac{I(R=1)}{\pi_1(L)}\mathbb{G}^*_{9\times 1}(\beta)\mu(\beta)\mathbb{A}^{*T}_{86\times 1}\right\}^{9\times 86}$$

$$U_{22} = E\left\{\mathbb{A}^*_{86\times 1}\mathbb{A}^{*T}_{86\times 1}\right\}^{86\times 86}$$

$$H_1 = \left(-E\left\{\frac{\partial\left[\mathbb{G}^*_{9\times 1}(\beta)\mu(\beta)\right]}{\partial\beta}\right\}^T\right)^{6\times 9}$$

$$H_2 = 0^{6\times 86}.$$

The superscripts denote dimensions of the matrices. The matrices (U_{11}, H_1) that involve full data L can be estimated from the complete cases only by standard inverse probability weighted empirical averages and the matrices $(U_{22}, , U_{12})$ by an empirical average of the observed data, i.e.

$$\hat{U}_{11}(\beta) = N^{-1}\sum_{i=1}^{N}\frac{I(R_i=1)}{\pi_1^2(\hat{\gamma})}\left\{\mathbb{G}^*_{9\times 1,i}(\beta)\mu_i(\beta)\mu_i(\beta)^T\mathbb{G}^{*T}_{9\times 1,i}(\beta)\right\}$$

$$\hat{U}_{12}(\beta) = N^{-1}\sum_{i=1}^{N}\frac{I(R_i=1)}{\pi_1(\hat{\gamma})}\left\{\mathbb{G}^*_{9\times 1,i}(\beta)\mu_i(\beta)\mathbb{A}^{*T}_{86\times 1,i}\right\}$$

$$\hat{U}_{22} = N^{-1}\sum_{i=1}^{N}\left\{\mathbb{A}^*_{86\times 1,i}\mathbb{A}^{*T}_{86\times 1,i}\right\}$$

$$\hat{H}_1(\beta) = -\left[N^{-1}\sum_{i=1}^{N}\frac{I(R_i=1)}{\pi_1(\hat{\gamma})}\left\{\frac{\partial\left[\mathbb{G}^*_{9\times 1,i}(\beta)\mu_i(\beta)\right]}{\partial\beta}\right\}\right]^T.$$

Then we can estimate $(P^{\text{opt}}_{6\times 9}, Q^{\text{opt}}_{6\times 86})$ by

$$\hat{P}^{\text{opt}}_{6\times 9}(\beta) = \hat{H}_1(\beta)\left\{\hat{U}_{11}(\beta) - \hat{U}_{12}(\beta)\hat{U}_{22}^{-1}\hat{U}_{12}^T(\beta)\right\}^{-1}$$

$$\hat{Q}^{\text{opt}}_{6\times 86}(\beta) = -\hat{H}_1(\beta)\left\{\hat{U}_{11}(\beta) - \hat{U}_{12}(\beta)\hat{U}_{22}^{-1}\hat{U}_{12}^T(\beta)\right\}^{-1}\hat{U}_{12}(\beta)\hat{U}_{22}^{-1},$$

and the AIPW estimator, denoted by $\hat{\beta}_{\text{aipw}}$ is given by the solution to the estimating equation

$$N^{-1}\sum_{i}^{N}\left\{\frac{I(R_i = 1)}{\pi_1(\hat{\gamma})}\hat{P}^{\text{opt}}_{6\times 9}(\beta)\mathbb{G}^*_{9\times 1,i}(\beta)\mu_i(\beta) + \hat{Q}^{\text{opt}}_{6\times 86}(\beta)\mathbb{A}^*_{86\times 1,i}\right\} = 0. \quad (12)$$

A consistent estimator for the asymptotic variance of $\hat{\beta}_{\text{aipw}}$ is given by

$$\left\{\hat{H}_1(\hat{\beta}_{\text{aipw}})\left\{\hat{U}_{11}(\hat{\beta}_{\text{aipw}}) - \hat{U}_{12}(\hat{\beta}_{\text{aipw}})\hat{U}_{22}^{-1}\hat{U}_{12}^T(\hat{\beta}_{\text{aipw}})\right\}^{-1}\hat{H}_1^T(\hat{\beta}_{\text{aipw}})\right\}^{-1}. \quad (13)$$

Note that $\hat{\beta}_{\text{aipw}}$ is guaranteed to be asymptotically more efficient than the simple IPW estimator, since the latter belongs to the class of estimators given in 10 by setting

$$P_{6\times 9} = [I_{6\times 6}, 0_{6\times 3}]$$

$$Q_{6\times 86} = 0_{6\times 86}.$$

Web Table 2: Estimates along with 95% confidence intervals from unconstrained maximum likelihood estimation (UMLE) of missing data model parameters in data set II. Asterisk denotes $p < 0.05$.

				UMLE			
Pattern	**Intercept**	**BMI**	**Age**	**Abort**	**Smoke**	**Black**	**Other**
1	-2.84(-3.22,-2.46)∗		-0.44(-1.88, 1.00)	-0.37(-0.89, 0.15)	0.10(-0.09, 0.29)	0.09(-0.10, 0.27)	-0.00(-0.36, 0.36)
2	-5.36(-6.54,-4.19)∗	1.45(-3.06, 5.96)	1.67(-1.59, 4.92)		-0.43(-0.94, 0.08)	0.07(-0.36, 0.49)	0.05(-0.74, 0.84)
3	-2.29(-2.56,-2.02)∗		0.63(-0.38, 1.64)		-0.06(-0.20, 0.08)	0.02(-0.12, 0.15)	0.06(-0.19, 0.30)
4	-2.52(-3.06,-1.98)∗	-0.94(-3.08, 1.20)	-0.52(-2.00, 0.96)	-0.24(-0.73, 0.25)		-0.04(-0.23, 0.14)	-0.15(-0.51, 0.22)
5	-3.09(-3.48,-2.71)∗		0.40(-1.04, 1.85)	0.08(-0.36, 0.51)		0.08(-0.11, 0.26)	0.12(-0.22, 0.47)
6	-1.80(-2.20,-1.40)∗	-0.08(-1.64, 1.48)	-1.53(-2.65,-0.42)∗			-0.06(-0.20, 0.08)	0.03(-0.23, 0.28)
7	-4.88(-5.72,-4.04)∗		0.52(-2.60, 3.64)			0.32(-0.08, 0.72)	0.40(-0.30, 1.10)

Web Table 3: Estimates along with 95% confidence intervals from unconstrained maximum likelihood estimation (UMLE) of missing data model parameters in data set III. Asterisk denotes $p < 0.05$.

				UMLE			
Pattern	**Intercept**	**BMI**	**Age**	**Abort**	**Smoke**	**Black**	**Other**
1	-3.25(-3.63,-2.86)∗		-0.35(-1.75, 1.06)	0.87(0.57, 1.17)∗	1.06(0.89, 1.23)∗	0.14(-0.04, 0.33)	-0.17(-0.59, 0.25)
2	-6.29(-7.36,-5.23)∗	3.79(-0.01, 7.59)	2.62(-0.27, 5.52)		0.92(0.56, 1.28)∗	-0.19(-0.61, 0.24)	0.39(-0.28, 1.06)
3	-2.59(-2.86,-2.31)∗		0.38(-0.63, 1.39)		1.16(1.04, 1.28)∗	0.06(-0.07, 0.20)	0.01(-0.26, 0.28)
4	-2.85(-3.29,-2.40)∗	1.91(0.19, 3.63)∗	-0.52(-1.77, 0.74)	0.52(0.19, 0.84)∗		-0.07(-0.23, 0.09)	-0.30(-0.63, 0.04)
5	-2.35(-2.67,-2.04)∗		-0.70(-1.90, 0.50)	0.96(0.69, 1.23)∗		-0.06(-0.22, 0.09)	0.05(-0.24, 0.33)
6	-1.78(-2.12,-1.44)∗	0.25(-1.09, 1.58)	-0.46(-1.40, 0.48)			0.05(-0.07, 0.16)	-0.10(-0.33, 0.13)
7	-4.40(-5.31,-3.49)∗		-1.18(-4.66, 2.31)			-0.02(-0.46, 0.41)	-1.18(-2.60, 0.23)

References

1. Robins JM, Rotnitzky A, Zhao LP. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*. 1994;89(427):846-866.

2. Tsiatis A. *Semiparametric Theory and Missing Data*. New York, NY: Springer; 2006.