Robust likelihood-based analysis of multivariate data with missing values

Hyonggin An and Roderick Little
University of Michigan

Outline

- Develop robust missing data methods based on models, multiple imputation
- Univariate missing data:
  - Penalized Spline Prediction -- robust modeling based on penalized splines
  - Handling the curse of dimensionality -- propensity penalized spline prediction
  - Variance estimation, simulation studies
- Extensions of propensity spline prediction to general patterns

Missing data methods

- Likelihood-based
  - Maximum likelihood, Bayes
  - Multiple imputation (approximate Bayes) -- good for multiple analyses
- Other
  - “approximate” (pseudo) likelihood
  - Weighting approaches
  - Estimating equations other than likelihood

Why I like models and MI

- *Every* method effectively predicts the missing values (including GEE, methods that drop the incomplete cases).
- Most direct approach is to build a predictive distribution of the missing values, that incorporates plausible assumptions
- ML, Bayes, MI all involve prediction of missing data
- Models make assumptions, but every method makes assumptions; seek robust models (e.g. splines; see below)

Multiple imputation

- MI -- Unlinks imputation model from analysis model
  - Simple models suffice if missing information is small
  - Imputation model can condition on variables not included in the analysis model (e.g. Marie's intermediate variables in her talk)
  - Uniform treatment of missing data across analyses
  - Allows implicit as well as explicit imputation models (e.g. hot deck, predictive mean matching)
  - But, retains asymptotic optimality of ML if imputation and analysis model are same

Notation

- A random sample of missing data

$$\left(x_{i1},\ldots,x_{iK},\, y_{i1},\ldots,y_{ip},r_{i1},\ldots,r_{ip}\right),\quad i\in\{1,\ldots,n\}$$

$$r_{ij}=\begin{cases}0 & \text{, if } y_{ij} \text{ is missing}\\ 1 & \text{, if } y_{ij} \text{ is observed}\end{cases}$$

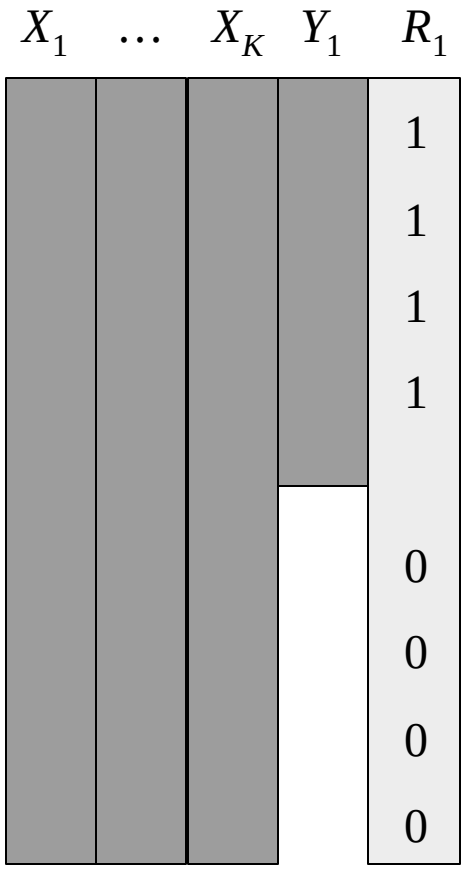
$$X_k=\left(x_{1k},\ldots,x_{nk}\right)^{\mathrm{T}} \text{ for } k=1,\ldots,K$$

$$Y_j=\left(y_{1j},\ldots,y_{nj}\right)^{\mathrm{T}} \text{ and}$$

$$R_j=\left(r_{1j},\ldots,r_{nj}\right)^{\mathrm{T}} \text{ for } j=1,\ldots,p$$

Introduction

- Assumptions
  - Missing at random (MAR)
  - Univariate pattern missing data
  - Later, I discuss extensions to a general pattern.
  - Independence over subjects
  - Y_1 continuous - focus on estimating its mean
- Data
  - $X_1, \ldots, X_K$ are fully observed
  - Y_1 has missing values.



Model-based prediction

$$m_1 = E\left(R_1 Y_1 + (1-R_1)E\left(Y_1 \mid X_1,\ldots,X_K\right)\right)$$

- Under MAR,

  $E(Y_1 \mid \mathbf{X}) = E(Y_1 \mid \mathbf{X}, R_1{=}1){=}E(Y_1 \mid \mathbf{X}, R_1{=}0)$

- One can estimate $E(Y_1 \mid X)$ using complete cases and predict the Y for each incomplete case by substituting the X for that case into the regression formula.

$$\hat{m}_1 = \frac{1}{n}\left(\sum_{i=1}^{r} y_{i1} + \sum_{i=r+1}^{n} \hat{y}_{i1}\right)$$

Prediction Method

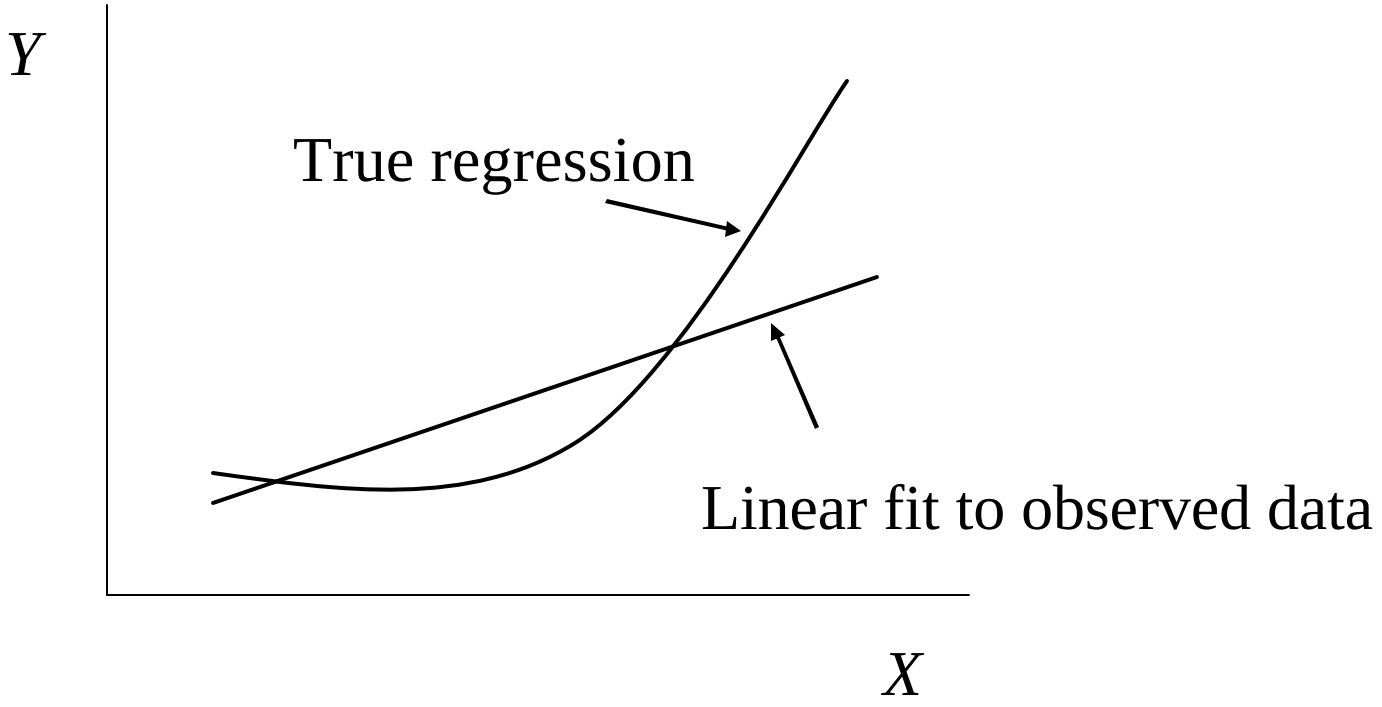
- If the prediction model is a linear regression, then

$$\hat{y}_{i1} = \hat{b}_0 + \sum_{k=1}^{K} \hat{b}_k \, x_{ik}, \; i \in \{i : r_{i1} = 0\}$$

- This linear regression prediction estimator is the maximum likelihood (ML) estimator of m_1 if we assume the data are from a multivariate normal distribution.
- Bayes/MI replaces *means* by *draws* from predictive distribution

Prediction Method

- The prediction method is sensitive to model misspecification, particularly if data are not MCAR



Relaxing Linearity: one X

- A simple way is to categorize X_1 and predict within classes -- link with *weighting* methods
- For continuous X_1 and sufficient sample size, a spline provides one useful alternative (Cheng 1994 JASA). We use a P-Spline approach:

$$(Y_1 \mid X_1, f) \sim N\left(s_1(X_1, f), s^2\right)$$

$$s_1\left(X_1, f\right) = f_0 + \sum_{j=1}^{q} f_j X_1^{\,j} + \sum_{k=1}^{K} f_{q+k}\left(X_1 - t_k\right)_+^q,$$

$$\left(x\right)_+^q = x^q I\left(x \geq 0\right),$$

$t_1 < \cdots < t_K$ are selected fixed knots

$f_{q+1}, \ldots, f_{q+K}$ are random effects, shrink to zero

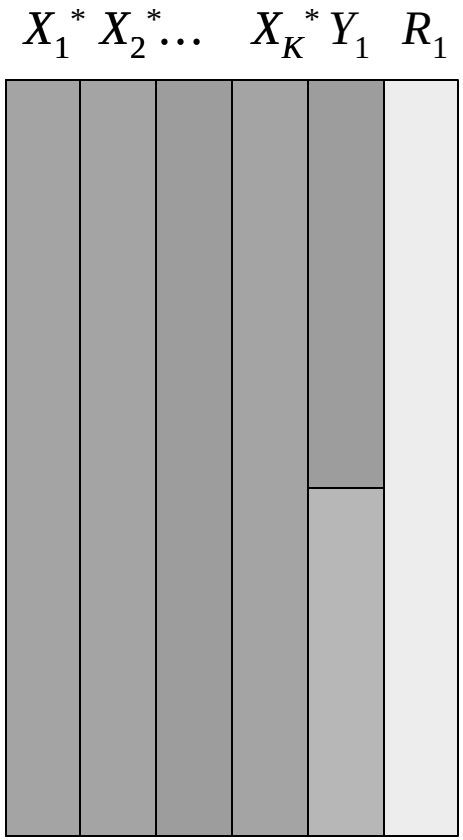
More than one covariate

- When we model the relationship with many covariates by smoothing, we have to deal with the "curse of dimensionality".
  - One approach is to "calibrate" the model by adding weighted residuals (e.g. Scharfstein and Izzarry 2004).
  - Our goal is to achieve both robustness and dimension reduction with many covariates, using the conceptually simple model-based approach.

Propensity P-spline Prediction

- Focus on a particular function of the covariates most sensitive to model misspecification, the response propensity score.
- Important to get relationship between *Y* and response propensity correct, since misspecification of this leads to bias
- Other *X*'s balanced over respondents and nonrespondents, conditional on propensity scores; so misspecification of regression of these is less important (loss of precision, not bias).

Propensity P-spline Prediction Model



1. $X_1^* = \text{logit}\left(\Pr\left(R_1 = 1 \mid X_1, \ldots, X_K\right)\right)$

2. For $k = 2, \ldots, K$,

   $X_k \mid X_1^* \sim N\left(s_k\left(X_1^*\right), s_{e_k}^2 \mathbf{I}_n\right)$,

   $X_k^* = X_k - \hat{s}_k\left(X_1^*\right)$.

3. $Y_1 \mid X_1^*, X_2^*, \ldots, X_K^*$

   $\sim N\left(s_1\left(X_1^*\right) + g\left(X_1^*, X_2^*, \ldots, X_K^*; x\right), s_e^2\right)$

   $\hat{y}_{i1} = \hat{s}_1\left(X_1^*\right) + g\left(X_1^*, X_2^*, \ldots, X_K^*; \hat{x}\right)$

   $g()$ parametric with $g(0,...,0;x) = 0$

Double Robustness Property

The proposed method yields consistency if

(a) Overall model relating Y_1 to $(X_1,\ldots,X_K)$ is correct,

or

(b1) X_1^* = logit(Pr(R_1=1|X_1, …, X_K)) is correctly specified and

(b2) Regressions of $Y_1,X_2,\ldots,X_K$ on X_1^* are correctly specified

Calibration approach yields consistency under (a) or (b1), without needing (b2)

But the additional condition (b2) is mild since these regressions are univariate and are modeled "nonparametrically" using splines.

Variance Estimation

- PPSP estimator is obtained based on predictions or imputations.

- Valid estimates of variance of the PPSP estimator need to incorporate
  - added uncertainty due to nonresponse,
  - and added variability due to propensity estimation

Asymptotic Variance

- Ignoring sampling error in estimating the propensity scores, and using the asymptotic variance for the PPSP estimator.

- Variance tends to be underestimated since the additional variability from estimating the propensity scores is ignored.

Bootstrap Variance Estimation

1. Generate a bootstrap sample by sampling with replacement from the complete and incomplete cases.
2. Apply the PPSP method to each bootstrap sample
3. Estimate variance from bootstrap distribution of parameter estimates.

Multiple Imputation

Bayesian Hierarchical Representation of PPSP

$$\left[Y_1 = (Y_{1\text{obs}}, Y_{1\text{mis}}) \mid X_1^*, X_2^*, \ldots, X_K^*, \boldsymbol{ß}, \mathbf{b}, s_{\mathbf{b}}^2, s_e^2\right] \sim N\left(\mathbf{X}^*\boldsymbol{ß} + \mathbf{Z}^*\mathbf{b},\ s_e^2\mathbf{I}_n\right)$$

$$X_1^* = \text{logit}\left(\Pr\left(R_1 = 1 \mid X_1, \ldots X_K; \mathbf{a}\right)\right) = \mathbf{X}\mathbf{a}, \text{ and}$$

$$\left[\mathbf{a} \mid X_1, \ldots, X_K\right] \sim N\left(\hat{\mathbf{a}}, I\left(\hat{\mathbf{a}}\right)^{-1}\right),$$

For $k = 2, \ldots, K$,

$$\left[X_k \mid X_1^*, \boldsymbol{ß}_k, \mathbf{b}_k, s_{e_j}^2\right] \sim N\left(\mathbf{X}^*\boldsymbol{ß}_k + \mathbf{Z}^*\mathbf{b}_k,\ s_{e_j}^2\mathbf{I}_n\right),$$

$$X_k^* = X_k - \left(\mathbf{X}^*\boldsymbol{ß}_k + \mathbf{Z}^*\mathbf{b}_k\right),$$

$$\left[\mathbf{b}_k \mid s_{\mathbf{b}_j}^2\right] \sim N\left(0,\ s_{\mathbf{b}_j}^2\mathbf{I}_S\right),$$

$$\left[\mathbf{b} \mid s_{\mathbf{b}}^2\right] \sim N\left(0,\ s_{\mathbf{b}}^2\mathbf{I}_S\right)$$

$$\left[\boldsymbol{ß}, s_{\mathbf{b}}^2, s_e^2, \boldsymbol{ß}_2, s_{\mathbf{b}_2}^2, s_{e_2}^2, \ldots, \boldsymbol{ß}_K, s_{\mathbf{b}_K}^2, s_{e_K}^2\right] \sim \text{Prior distribution}$$

Multiple Imputation

- Impute *M* sets of imputed data using the Gibbs' sampler for Bayes version of model.
- Apply Rubin's (1987) rule to the multiply imputed data sets to estimate the mean of Y_1 and its associated variance.
- Not quite fully Bayes, but close

Multiple Imputation by Bootstrap

1. Bootstrap the samples *M* times
2. For each bootstrapped samples,
    1. Estimate relevant parameters for PPSP
    2. Predict the missing values using the estimated parameters from the bootstrap data and fully observed variables for incomplete cases.
3. Using the multiply imputed data sets, estimate the mean of *Y* and its associated standard error by the MI combining rules.

Simulation

- Objective: To evaluate and compare the proposed variance estimation methods.
- X_1, X_2 ~ind. $N(0, 1)$
- $Y_1|X_1,X_2$ ~
  - Linear mean: $N(10+3X_1+3X_2, 3)$
  - Additive mean: $N(1+(X_1+2)^2+(X_2+1)^2, 3)$
  - Non-additive mean: $N(10+5X_1+5X_2+5X_1X_2, 5)$
- $E(Y_1) = 10$ for all three mean structures.
- Response propensity: $E(R_1)=0.5$
  - $\text{logit}(\Pr(R_1=1|X_1,X_2)) = X_1+X_2$
- 500 data sets with sample size 100 and 1000 are generated for each mean structure.

Simulation

- **Prediction methods**
  - **BD**: Before deletion estimator
  - **CC**: Complete case estimator
  - **LP**: Linear prediction estimator
  - **APSP**: Additive P-spline prediction estimator
  - **PPSP**: Propensity P-spline prediction estimator
- **Variance estimation methods**
  - **SI**: Variance estimate based on single imputation
  - **AV**: Asymptotic variance estimate ignoring sampling error in the estimated propensity scores
  - **BOOT**: Bootstrap variance estimate
  - **MI-Boot**: Multiple imputation by bootstrap
  - **MI-Bayes**: Multiple imputation estimate from the Bayesian joint posterior distribution using the Gibbs' sampler.

Simulation

- For P-splines,
  - 20 equal percentile knots
  - linear truncated spline basis
- For multiple imputation
  - Flat priors on parameters
  - 10 sets of imputed data
- For each method over 500 simulated data sets,
  - BIAS: empirical bias
  - Emp.SE: empirical standard error
  - Est.SE: mean of the estimated standard errors
  - Ave.CI: average length of confidence intervals (95%)
  - COV: coverage rate (95%)

Simulation: Linear mean structure (n = 1000)

Method		BIAS	Emp. SE	Est. SE	Ave. CI	Coverage
BD		**0.000**	**0.151**	**0.145**	**0.567**	**94.8**
CC		**2.173**	**0.185**	**0.180**	**0.706**	**0.0**
LP	**(SI)**	**-0.001**	**0.168**	**0.139**	**0.547**	**90.0**
	(AV)	**-0.001**	**0.168**	**0.152**	**0.596**	**92.6**
	(BOOT)	**-0.001**	**0.168**	**0.160**	**0.629**	**93.0**
	(MI-BOOT)	**-0.001**	**0.169**	**0.163**	**0.642**	**93.8**
	(MI-Bayes)	**-0.001**	**0.168**	**0.171**	**0.676**	**96.4**
APSP	**(SI)**	**-0.002**	**0.169**	**0.140**	**0.547**	**88.4**
	(AV)	**-0.002**	**0.169**	**0.152**	**0.597**	**91.2**
	(BOOT)	**-0.002**	**0.169**	**0.163**	**0.640**	**92.2**
	(MI-BOOT)	**-0.004**	**0.170**	**0.165**	**0.653**	**93.8**
	(MI-Bayes)	**-0.004**	**0.176**	**0.167**	**0.661**	**94.4**
PPSP	**(SI)**	**-0.002**	**0.170**	**0.140**	**0.547**	**89.8**
	(AV)	**-0.002**	**0.170**	**0.153**	**0.599**	**92.0**
	(BOOT)	**-0.002**	**0.170**	**0.167**	**0.653**	**94.2**
	(MI-BOOT)	**0.000**	**0.171**	**0.169**	**0.669**	**95.0**
	(MI-Bayes)	**-0.009**	**0.176**	**0.168**	**0.667**	**94.0**

Simulation: Additive mean structure (n=1000)

Method		BIAS	Emp. SE	Est. SE	Ave. CI	Coverage
BD		**-0.022**	**0.280**	**0.285**	**1.119**	**95.0**
CC		**3.539**	**0.457**	**0.458**	**1.794**	**0.0**
LP	**(SI)**	**-1.246**	**0.361**	**0.331**	**1.298**	**4.8**
	(AV)	**-1.246**	**0.361**	**0.328**	**1.285**	**4.6**
	(BOOT)	**-1.246**	**0.361**	**0.358**	**1.404**	**6.8**
	(MI-BOOT)	**-1.254**	**0.379**	**0.456**	**1.826**	**17.2**
	(MI-Bayes)	**-1.248**	**0.368**	**0.435**	**1.734**	**13.8**
APSP	**(SI)**	**-0.022**	**0.293**	**0.272**	**1.105**	**92.4**
	(AV)	**-0.022**	**0.293**	**0.293**	**1.147**	**94.4**
	(BOOT)	**-0.022**	**0.293**	**0.310**	**1.180**	**96.0**
	(MI-BOOT)	**-0.015**	**0.303**	**0.357**	**1.427**	**97.6**
	(MI-Bayes)	**-0.032**	**0.296**	**0.306**	**1.202**	**95.8**
PPSP	**(SI)**	**-0.041**	**0.332**	**0.281**	**1.100**	**89.8**
	(AV)	**-0.041**	**0.332**	**0.360**	**1.409**	**95.8**
	(BOOT)	**-0.041**	**0.332**	**0.338**	**1.324**	**93.2**
	(MI-BOOT)	**-0.041**	**0.337**	**0.378**	**1.508**	**96.4**
	(MI-Bayes)	**-0.026**	**0.367**	**0.409**	**1.650**	**96.8**

Simulation: Non-additive mean structure (n=1000)

Method		BIAS	Emp. SE	Est. SE	Ave. CI	Coverage
BD		**0.007**	**0.264**	**0.283**	**1.108**	**95.8**
CC		**3.636**	**0.433**	**0.436**	**1.711**	**0.0**
LP	**(SI)**	**-1.370**	**0.362**	**0.328**	**1.286**	**2.2**
	(AV)	**-1.370**	**0.362**	**0.316**	**1.237**	**1.6**
	(BOOT)	**-1.370**	**0.362**	**0.376**	**1.474**	**3.6**
	(MI-BOOT)	**-1.370**	**0.369**	**0.430**	**1.718**	**9.0**
	(MI-Bayes)	**-1.372**	**0.366**	**0.431**	**1.719**	**6.0**
APSP	**(SI)**	**-1.731**	**0.471**	**0.352**	**1.381**	**1.0**
	(AV)	**-1.731**	**0.471**	**0.338**	**1.324**	**0.8**
	(BOOT)	**-1.731**	**0.471**	**0.471**	**1.497**	**5.2**
	(MI-BOOT)	**-1.735**	**0.514**	**0.600**	**2.985**	**27.6**
	(MI-Bayes)	**-1.785**	**0.500**	**0.436**	**1.728**	**2.8**
PPSP	**(SI)**	**-0.063**	**0.351**	**0.269**	**1.055**	**86.8**
	(AV)	**-0.063**	**0.351**	**0.328**	**1.284**	**93.4**
	(BOOT)	**-0.063**	**0.351**	**0.366**	**1.435**	**94.6**
	(MI-BOOT)	**-0.058**	**0.359**	**0.404**	**1.644**	**96.4**
	(MI-Bayes)	**-0.075**	**0.356**	**0.354**	**1.412**	**94.6**

General Pattern Missing Data

- Bayes/Gibbs based on model for joint distribution
  - Principled, but requirement of a coherent joint distribution imposes limitations.
- Raghunathan et al. (2001) proposed a sequential regression multivariate imputation.
  - Approximate relevant conditional distributions for Gibbs' sampler by a sequence of regressions of one variable on all the others.
  - Another advantage is the SRMI can create multiply imputed data sets.
  - Inference can be made by Rubin's (1987) rule.

Sequential regression MI

- Fill in each missing values of each variable by draws from predictive distribution given observed or imputed values of other variables (as in a Gibbs' sampler)
- Prediction by drawing from posterior distribution of parameters of regression, and then missing values given drawn values of parameters
- Cycles through variables one at a time, conditioning on latest draws of missing values for other variables

Sequential Regression MI

- When the regression model is linear additive, the draws of missing values from SRMI procedures are equivalent to the draws from joint predictive distribution under a multivariate normal distribution with an improper prior on the mean and the covariance.
- For more robust prediction, we incorporate PPSP regressions into the SRMI approach.

PPSP in sequential MI method

- For the missing values of each variable:
- Condition on imputed values of other variables – reduces problem to univariate missing data
- Compute propensity to respond for that variable given the other variables
- Apply the PPSP method to create draws of missing values for that variable
- Cycle through all the variable until “convergence”
- Details in Hyonggin An’s thesis

Simulation1

- logit(Pr(R_1=0|X_1,X_2,Y_2) and logit(Pr(R_2=0|X_1,X_2,Y_1) are not exactly linear but approximately linear.
- The propensity model specified in the PPSP method is relatively correct.
- Linear Additive, Non-linear additive and nonlinear nonadditive mean structures are simulated
- 200 simulated data sets with sample size 500 for each mean model.

Simulation1

- X_1, X_2 ~ind. $N(0, 1)$
- Linear mean model:
  - $Y_1|X_1,X_2 \sim N(X_1+X_2, 5)$
  - $Y_2|X_1,X_2,Y_1 \sim \mathrm{N}(\mathrm{X1+X2+2Y1}, 5)$
- Additive mean model
  - $Y_1|X_1,X_2 \sim N(-4 +X_1+X_2+2X_1^2+2X_2^2, 5)$
  - $Y_2|X_1,X_2,Y_1 \sim N(-25+X_1+X_2+Y_1+X_1^2+X_2^2+Y_1^2, 5)$
- Non-additive mean model
  - $Y_1|X_1,X_2 \sim N(X_1+X_2+5X_1X_2, 5)$
  - $Y_2|X_1,X_2,Y_1 \sim \mathrm{N}(-10+X_1+X_2+Y_1+5X_1Y_1+5X_2Y_1, 5)$
- $\mathrm{E}(Y_1)=\mathrm{E}(Y_2)=0$

Simulation1

- Generating missing data

$$
\begin{aligned}
P_{10} &= \Pr\left(R_1 = 1, R_2 = 0 \mid X_1, X_2, Y_1\right) \\
&= \Pr\left(R_1 = 1 \mid R_2 = 0, X_1, X_2\right) \Pr\left(R_2 = 0 \mid X_1, X_2, Y_1\right) \\
&= \left(\frac{\exp\left(0.1\left(1 + X_1 + X_2\right)\right)}{1 + \exp\left(0.1\left(1 + X_1 + X_2\right)\right)} \right) \left(\frac{1}{1 + \exp\left(0.1\left(1 + X_1 + X_2 + Y_1\right)\right)} \right)
\end{aligned}
$$

$$
\begin{aligned}
P_{01} &= \Pr\left(R_1 = 0, R_2 = 1 \mid X_1, X_2, Y_2\right) \\
&= \Pr\left(R_1 = 1 \mid R_2 = 1, X_1, X_2, Y_2\right) \Pr\left(R_2 = 1 \mid X_1, X_2\right) \\
&= \left(\frac{1}{1 + \exp\left(0.01\left(1 + X_1 + X_2 + Y_2\right)\right)} \right) \left(\frac{\exp\left(X_1 + X_2\right)}{1 + \exp\left(X_1 + X_2\right)} \right)
\end{aligned}
$$

Simulation1

$$P_{00} = \Pr\left(R_1 = 0, R_2 = 0 \mid X_1, X_2\right)$$
$$= \Pr\left(R_1 = 0 \mid R_2 = 0, X_1, X_2\right) \Pr\left(R_2 = 0 \mid X_1, X_2\right)$$
$$= \left(\frac{\exp\left(0.1\left(1 + X_1 + X_2\right)\right)}{1 + \exp\left(0.1\left(1 + X_1 + X_2\right)\right)} \right) \left(\frac{1}{1 + \exp\left(X_1 + X_2\right)} \right)$$

$$P_{11} = 1 - P_{10} - P_{01} - P_{00}$$

Simulation1

% of observed values

Mean Model	$R_1 = 1$	$R_2=1$	$R_1=1, R_2=1$	$R_1=1, R_2=0$	$R_1=0, R_2=1$	$R_1=0, R_2=0$
Linear	**62**	**61**	**48**	**14**	**13**	**25**
Additive	**64**	**60**	**49**	**15**	**11**	**25**
Non-additive	**62**	**60**	**47**	**15**	**13**	**25**

Simulation1 Results

Linear mean model when propensities are correctly specified

Method	m_1					m_2				
	BIAS	**RMSE**	**Emp. SE**	**Est. SE**	**Cov.**	**BIAS**	**RMSE**	**Emp. SE**	**Est. SE**	**Cov.**
BD	**0.001**	**0.119**	**0.119**	**0.119**	**96.0**	**0.009**	**0.293**	**0.293**	**0.294**	**97.0**
CC	**0.135**	**0.200**	**0.148**	**0.152**	**85.5**	**1.115**	**1.174**	**0.365**	**0.373**	**15.5**
LP	**-0.001**	**0.145**	**0.145**	**0.194**	**97.5**	**0.007**	**0.332**	**0.332**	**0.456**	**97.0**
APSP	**-0.055**	**0.187**	**0.179**	**0.234**	**98.5**	**-0.122**	**0.431**	**0.413**	**0.545**	**98.5**
PPSP	**-0.021**	**0.147**	**0.146**	**0.149**	**94.0**	**0.039**	**0.340**	**0.338**	**0.362**	**96.0**

Simulation1 Results

Additive mean model when propensities are correctly specified

Method	m_1					m_2				
	BIAS	**RMSE**	**Emp. SE**	**Est. SE**	**Cov.**	**BIAS**	**RMSE**	**Emp. SE**	**Est. SE**	**Cov.**
BD	**-0.007**	**0.215**	**0.215**	**0.214**	**95.0**	**-0.121**	**2.541**	**2.538**	**2.555**	**90.5**
CC	**0.336**	**0.452**	**0.302**	**0.286**	**79.0**	**4.556**	**6.001**	**3.906**	**3.870**	**88.5**
LP	**-0.033**	**0.285**	**0.283**	**0.415**	**98.5**	**-3.560**	**4.604**	**2.920**	**6.022**	**97.5**
APSP	**-0.025**	**0.227**	**0.226**	**0.233**	**95.0**	**-0.028**	**2.616**	**2.616**	**2.760**	**94.0**
PPSP	**0.012**	**0.253**	**0.253**	**0.249**	**97.0**	**-0.238**	**2.855**	**2.855**	**3.108**	**93.0**

Simulation1 Results

Non-additive mean model when propensities are correctly specified

Method	m_1					m_2				
	BIAS	**RMSE**	**Emp. SE**	**Est. SE**	**Cov.**	**BIAS**	**RMSE**	**Emp. SE**	**Est. SE**	**Cov.**
BD	**0.037**	**0.255**	**0.253**	**0.251**	**93.0**	**0.096**	**3.036**	**3.035**	**2.960**	**94.5**
CC	**0.535**	**0.625**	**0.323**	**0.337**	**66.0**	**7.834**	**9.007**	**4.444**	**4.440**	**55.5**
LP	**0.108**	**0.371**	**0.355**	**0.521**	**99.0**	**-6.405**	**7.430**	**3.765**	**6.543**	**90.0**
APSP	**1.889**	**2.156**	**1.040**	**0.631**	**25.5**	**21.921**	**28.925**	**18.871**	**9.007**	**51.0**
PPSP	**0.084**	**0.331**	**0.320**	**0.299**	**91.0**	**-0.392**	**3.653**	**3.632**	**3.299**	**91.0**

Simulation2

- logit(Pr(R_1=0|X_1,X_2,Y_2)=logit(P_{01}+P_{00}) and logit(Pr(R_2=0|X_1,X_2,Y_1)=logit(P_{10}+P_{00}) are clearly not linear.
- The propensity model is not correctly specified the PPSP method.
- Linear Additive, Non-linear additive and nonlinear nonadditive mean structures are simulated
- 200 simulated data sets with sample size 500 for each mean model.

Simulation2

- Generating missing data

$$P_{00} = \Pr(R_1 = 0, R_2 = 0 \mid X_1, X_2)$$
$$= \left(\frac{\exp(-2 + X_1 + X_2)}{1 + \exp(-2 + X_1 + X_2)} \right),$$
$$P_{10} = \Pr(R_1 = 1, R_2 = 0 \mid X_1, X_2, Y_1)$$
$$= \left(\frac{1}{2} \left(\frac{1}{1 + \exp(0.5 Y_1)} \right) \right) \times (1 - P_{00}),$$
$$P_{01} = \Pr(R_1 = 0, R_2 = 1 \mid X_1, X_2, Y_2)$$
$$= \left(\frac{1}{2} \left(\frac{1}{1 + \exp(0.1 Y_2)} \right) \right) \times (1 - P_{00}),$$
$$P_{11} = 1 - P_{10} - P_{01} - P_{00}$$

Simulation2

% of observed values

Mean Model	$R_1 = 1$	$R_2=1$	$R_1=1$, $R_2=1$	$R_1=1$, $R_2=0$	$R_1=0$, $R_2=1$	$R_1=0$, $R_2=0$
Linear	**62**	**62**	**43**	**19**	**20**	**18**
Additive	**17**	**58**	**47**	**24**	**11**	**18**
Non-additive	**67**	**63**	**48**	**19**	**15**	**18**

Simulation2 Results

Linear mean model when propensities are incorrectly specified

Method	m_1					m_2				
	BIAS	**RMSE**	**Emp. SE**	**Est. SE**	**Cov.**	**BIAS**	**RMSE**	**Emp. SE**	**Est. SE**	**Cov.**
BD	**-0.003**	**0.111**	**0.111**	**0.118**	**97.5**	**-0.007**	**0.288**	**0.288**	**0.293**	**95.0**
CC	**-0.511**	**0.530**	**0.143**	**0.146**	**6.5**	**-1.787**	**1.821**	**0.349**	**0.355**	**0.0**
LP	**0.002**	**0.141**	**0.141**	**0.188**	**98.0**	**0.014**	**0.340**	**0.340**	**0.468**	**98.0**
APSP	**0.077**	**0.215**	**0.201**	**0.239**	**98.0**	**0.167**	**0.500**	**0.472**	**0.554**	**96.0**
PPSP	**0.001**	**0.133**	**0.133**	**0.132**	**97.0**	**0.000**	**0.322**	**0.322**	**0.324**	**97.0**

Simulation2 Results

Additive mean model when propensities are incorrectly specified

Method	m_1					m_2				
	BIAS	**RMSE**	**Emp. SE**	**Est. SE**	**Cov.**	**BIAS**	**RMSE**	**Emp. SE**	**Est. SE**	**Cov.**
BD	**0.004**	**0.219**	**0.219**	**0.214**	**95.0**	**-0.120**	**2.461**	**2.458**	**9.958**	**96.5**
CC	**-0.451**	**0.517**	**0.252**	**0.242**	**54.5**	**-1.960**	**3.512**	**2.915**	**11.025**	**79.0**
LP	**-0.490**	**0.557**	**0.263**	**0.340**	**76.0**	**-8.040**	**8.371**	**2.332**	**18.283**	**54.0**
APSP	**0.031**	**0.257**	**0.256**	**0.244**	**94.5**	**0.573**	**3.260**	**3.209**	**11.462**	**93.5**
PPSP	**-0.335**	**0.416**	**0.246**	**0.249**	**73.0**	**-4.396**	**5.007**	**2.396**	**10.963**	**60.0**

Simulation2 Results

Non-additive mean model when propensities are incorrectly specified

Method	m_1					m_2				
	BIAS	**RMSE**	**Emp. SE**	**Est. SE**	**Cov.**	**BIAS**	**RMSE**	**Emp. SE**	**Est. SE**	**Cov.**
BD	**-0.005**	**0.267**	**0.267**	**0.253**	**92.5**	**-0.078**	**3.159**	**3.158**	**2.965**	**92.0**
CC	**-0.623**	**0.682**	**0.278**	**0.270**	**38.5**	**-9.614**	**9.409**	**2.135**	**2.356**	**2.5**
LP	**-0.770**	**0.842**	**0.339**	**0.380**	**56.0**	**-4.647**	**5.697**	**3.296**	**3.453**	**80.0**
APSP	**-0.739**	**0.902**	**0.516**	**0.397**	**50.0**	**-10.159**	**12.450**	**7.196**	**3.982**	**35.5**
PPSP	**0.125**	**0.392**	**0.371**	**0.314**	**89.5**	**-3.158**	**4.661**	**3.427**	**2.714**	**74.0**

Conclusions

- Models, multiple imputation: direct, flexible
- Make model more flexible and robust to avoid model misspecification
- A key idea is to single out the response propensity scores for the robust form of modeling to achieve dimension reduction and robustness.
- Simulation studies show that the PPSP works well over wide range of population with different mean structures when the propensity is correctly specified under the MAR assumption.

Future work

- Comparing the PPSP method with the calibration method when missing data pattern is general. – comparing double robustness.
- Extend the PPSP method to non-continuous missing variables such as binary, categorical, and counting data.
- Generalize the method to estimate other functions rather than means, such as covariance of missing variables and density estimation.
- Extend the method to non-ignorable missing data. Bayesian sensitivity analysis is my prefered approach