The challenge of missing data

- Let's assume that y_i is non-missing
- But the covariates, x_{ij}, are occasionally missing
- Missing Completely at Random (MCAR): $\mathrm{P}\left[x_{ij} \text{ is missing}\right] \perp y_i$
- If we can assume MCAR, a complete case analysis is unbiased yet, efficiency may suffer inviting missing imputation
- However, MCAR is only an assumption and a simplistic one
- Missing at Random (MAR) is more plausible $\mathrm{P}\left[x_{ij} \text{ is missing}\right]$ depends on y_i or covariates $x_{ij'}$
- MAR requires missing imputation to be unbiased

The challenge of missing data

	X_A	X_B	X_{AB}	X_O
A	1	0	0	0
B	0	1	0	0
AB	0	0	1	0
O	0	0	0	1
NA	0	0	0	0

- For categories, expand R factors into dummy variables
- Notice that this provides a natural missing category where $X_A = X_B = X_{AB} = X_O = 0$
- So imputation can be avoided when unnecessary
- Similarly, dichotomous covariates can be transformed into a category with three values
- For a continuous covariate like red blood hemoglobin, X_H
- You could add another variable as a missing indicator, X_M
- $X_M = 0$ when X_H is observed
- $X_M = 1$ when X_H is missing and X_H is set to a nominal value

The challenge of missing data: hot-decking

de Waal et al. 2011 *Handbook of Stat. Data Editing and Imputation*

- Legend has it that the US Census Bureau pioneered *hot-decking*
- There used to be a long-form household survey
- This asked a lot more questions such as your household income
- If a particular household declined to provide their income
- Then a nearby home was chosen randomly as a proxy
- Closeness was considered a good (or "hot") predictor
- So the corresponding income of this neighbor was imputed for the missing household

Adapting hot-decking to regression

- For y_i suppose that the one of the covariates, x_{ij}, is missing
- You can draw an $x_{i'j}$ from another record where $|y_i - y_{i'}| < \delta$
- However, challenges remain
- For example, how do you choose δ?
- And a related issue, what if y_i is not continuous?
- Nearness for dichotomous outcomes is not so simple
- Similarly, for time-to-event outcomes it is not well-defined

Cold-decking imputation for regression

- Cold-decking is the substitution of any neighbor's value to replace a missing value on the resident's form
- For one or more missing covariates, record-level cold-decking imputation can be employed
- It is biased towards the null, i.e., non-missing values from another record are randomly selected regardless of the outcome
- Sufficient for data sets with relatively few missing values

Cold-decking imputation for regression

- Suppose that we have the following 5 variables:
  household income, owned home vs. renting,
  age of home, number of rooms and number of occupants
- It is reasonable to assume that these variables are related
- Suppose record i has the observed/missingness pattern
  $\mathtt{A}_i$ $\mathtt{B}_i$ NA NA NA
- And we randomly draw record j to replace its values
  $\mathtt{C}_j$ $\mathtt{D}_j$ NA $\mathtt{E}_j$ $\mathtt{F}_j$
- Now, record i looks like this
  $\mathtt{A}_i$ $\mathtt{B}_i$ NA $\mathtt{E}_j$ $\mathtt{F}_j$
- So, we need to randomly draw again: record k
  $\mathtt{G}_k$ NA $\mathtt{H}_k$ $\mathtt{I}_k$ NA
- Now, record i looks like this
  $\mathtt{A}_i$ $\mathtt{B}_i$ $\mathtt{H}_k$ $\mathtt{E}_j$ $\mathtt{F}_j$

More Advanced Missing Imputation Methodology

- Let's assume that y_i is non-missing
- But the covariates, $\boldsymbol{X}_i$, are occasionally missing
- Order the covariates by increasing level of missingness: 1 to p

$$
\begin{aligned}
\boldsymbol{X}_i &= (\boldsymbol{X}_{i0}, X_{i1}, \ldots, X_{ip}) && \boldsymbol{X}_{i0} \text{ are fully observed} \\
&= (\boldsymbol{X}_i^{\text{obs}}, \boldsymbol{X}_i^{\text{mis}}) && \text{alternate construction} \\
\boldsymbol{r}_i &= (r_{i1}, \ldots, r_{ip}) && \text{response indicators} \\
\Pr\left(r_{ij} = 1 | \boldsymbol{X}_i, y_i, \boldsymbol{\theta}_r\right) &= \Pr\left(r_{ij} = 1 | \boldsymbol{\theta}_r\right) && \text{MCAR assumption} \\
\Pr\left(r_{ij} = 1 | \boldsymbol{X}_i, y_i, \boldsymbol{\theta}_r\right) &= \Pr\left(r_{ij} = 1 | \boldsymbol{X}_i^{\text{obs}}, y_i, \boldsymbol{\theta}_r\right) && \text{MAR assumption}
\end{aligned}
$$

Joint Sequential Missing Imputation

- Chained equations (CE) are a popular approach for imputation $\left[X_{ij}|\boldsymbol{X}_{i,-j},\boldsymbol{\theta}_j\right]$
- However, CE are NOT guaranteed to provide valid inference since they do NOT correspond to a joint distribution
- As an alternative, let's construct a joint distribution sequentially

$$\begin{aligned}\left[X_{i1},\ldots,X_{ip},y_i|\boldsymbol{X}_{i0},\boldsymbol{\theta}\right] = {} & [X_{i1}|\boldsymbol{X}_{i0},\boldsymbol{\theta}_1]\,[X_{i2}|\boldsymbol{X}_{i0},X_{i1},\boldsymbol{\theta}_2]\cdots \\ & \left[X_{ip}|\boldsymbol{X}_{i0},X_{i1},\ldots,X_{i,p-1},\boldsymbol{\theta}_p\right]\left[y_i|\boldsymbol{X}_i,\boldsymbol{\theta}_y\right]\end{aligned}$$

Sequential BART

Xu, Daniels and Winterstein 2016 *Biostatistics*
`https://cran.r-project.org/src/contrib/Archive/sbart`

1. BART to impute continuous and dichotomous covariates
2. For categories, Bayesian CART (a single tree) where all categorical covariates are combined into a single vast categorical variable
3. And the outcome, y_i, is modeled by linear regression

- Yet Sequential BART has several challenges
- For example, choices 2. and 3. are not based on BART
- The **sbart** R package is archived on CRAN
- However in my experience, **sbart** is unstable
- **BART3** has a working variant based on MPI BART
- See the sub-directory `seqBART` after install

Sequential BART2: research in progress

1. BART to impute continuous and dichotomous covariates
2. ~~2.~~ For categories, Bayesian CART (a single tree) where all categorical covariates are combined into a single vast categorical variable
3. And the outcome, y_i, is modeled by ~~linear regression~~ BART
4. It is with propositions 2. and 3. that we part ways
   we want to replace them with more natural BART alternatives

$$\boldsymbol{X}_{i,j-1} = (\boldsymbol{X}_{i0}, X_{i1}, \ldots, X_{i,j-1})$$

$$X_{ij}|(\boldsymbol{X}_{i,j-1}, \boldsymbol{\theta}_j) \sim \begin{cases} \mathrm{N}\left(f_j(\boldsymbol{X}_{i,j-1}),\ \sigma_j^2\right) & \text{Continuous} \\ \mathrm{B}\left(\Phi(f_j(\boldsymbol{X}_{i,j-1}))\right) & \text{Dichotomous} \\ \text{more to come} & \text{Categorical} \end{cases}$$

$$\text{where } f_j \overset{\text{prior}}{\sim} \text{BART}\ (\mu_j)$$

$$y_i \sim \mathrm{N}\left(f_y(\boldsymbol{X}_{ip}),\ \sigma_y^2\right)$$

$$\text{where } f_y \overset{\text{prior}}{\sim} \text{BART}\ (\mu_y)$$

Sequential BART2: imputation for continuous variables

- after $\boldsymbol{\theta}^m$ have been generated from the posterior
- draw X_{ij}^{mis} with Metropolis-Hastings at the mth MCMC sample
- the design matrix up to covariate $j-1$
  $\boldsymbol{X}_{i,j-1}^m = (\boldsymbol{X}_{i0}, X_{i1}^m, \ldots, X_{i,j-1}^m)$
- the design matrix up to covariate $k \geq j$
  $\boldsymbol{X}_{ik}(x) = (\boldsymbol{X}_{i,j-1}^m, x, X_{i,j+1}^{m-1}, \ldots, X_{ik}^{m-1})$
- draw proposal: $X_{ij}^* | (\boldsymbol{X}_{i,j-1}^m, \boldsymbol{\theta}_j) \sim \mathrm{N}\left(f_{jm}(\boldsymbol{X}_{i,j-1}^m),\ \sigma_{jm}^2\right)$
- $\eta^*(x) = \phi(y_i | f_y(\boldsymbol{X}_{ip}(x)), \sigma_y^2) \prod_{k=j+1}^{p} \left[X_{ik}^{m-1} | \boldsymbol{X}_{i,k-1}(x), \boldsymbol{\theta}_k\right]$
- accept proposal X_{ij}^* with probability $\min(1, \eta^*(X_{ij}^*)/\eta^*(X_{ij}^{m-1}))$

Sequential BART2: imputation for dichotomous variables

- draw X_{ij}^{mis} with Bayes' rule at the mth MCMC sample
- the design matrix up to covariate $j-1$
  $\boldsymbol{X}_{i,j-1}^{m} = (\boldsymbol{X}_{i0}, X_{i1}^{m}, \ldots, X_{i,j-1}^{m})$
- the design matrix up to covariate $k \geq j$
  $\boldsymbol{X}_{ik}(x) = (\boldsymbol{X}_{i,j-1}^{m}, x, X_{i,j+1}^{m-1}, \ldots, X_{ik}^{m-1})$

$$\text{draw } X_{ij}^{\text{mis}} \sim \mathrm{B}\left(\frac{\eta(1)}{\eta(0)+\eta(1)}\right)$$

$$\eta(x) = \Pr\left(X_{ij} = x | \boldsymbol{X}_{i,j-1}^{m}, \boldsymbol{\theta}_j\right) \phi(y_i | f_y(\boldsymbol{X}_{ip}(x)), \sigma_y^2) \times \prod_{k=j+1}^{p} \left[X_{ik}^{m-1} | \boldsymbol{X}_{i,k-1}(x), \boldsymbol{\theta}_k\right]$$

$$\text{draw } Z_{ij}^{m} \sim \mathrm{N}\left(f_{jm}(\boldsymbol{X}_{i,j-1}^{m}),\ \sigma_{jm}^2\right) \begin{cases} \mathrm{I}(-\infty, 0) & \text{if } X_{ij} = 0 \\ \mathrm{I}(0, \infty) & \text{if } X_{ij} = 1 \end{cases}$$

Sequential BART2: imputation for categorical variables

- a simple extension of dichotomous covariates
- replace ~~Bayesian CART~~ with Categorical BART
- each category has more than 2 values: $1, \ldots, L_j$
- draw X_{ij}^{mis} with Bayes' rule at the mth MCMC sample

$$\text{draw } X_{ij}^{\text{mis}} \sim \text{Multinomial}\left(1, \begin{bmatrix} \pi(1) \\ \vdots \\ \pi(L_j) \end{bmatrix}\right)$$

$$\text{where } \pi(x) = \frac{\eta(x)}{\sum_{x'=1}^{L_j} \eta(x')}$$