

PLSC 503 – Spring 2022

Bootstrapping and Missing Data

March 16, 2022

The population is to the sample as the sample is to the bootstrap sample.

Practical (Nonparametric) Bootstrapping

- Draw one bootstrap sample of size N **with replacement** from the original data,
- Estimate the parameter(s) $\tilde{\theta}_{k \times 1}$,
- Repeat steps 1 and 2 R times, to get $\tilde{\theta}_r$, $r \in \{1, 2, \dots, R\}$, comprising elements $\tilde{\theta}_{rk}$,
- Examine the empirical characteristics of the resulting distribution(s) of $\tilde{\theta}_{rk}$.

Why Bootstrap?

- **It's intuitive.**
- **It's simple.**
- **It's robust.**

Bootstrapping: “By Hand”

```
N<-10 # small sample!
reps<-1001

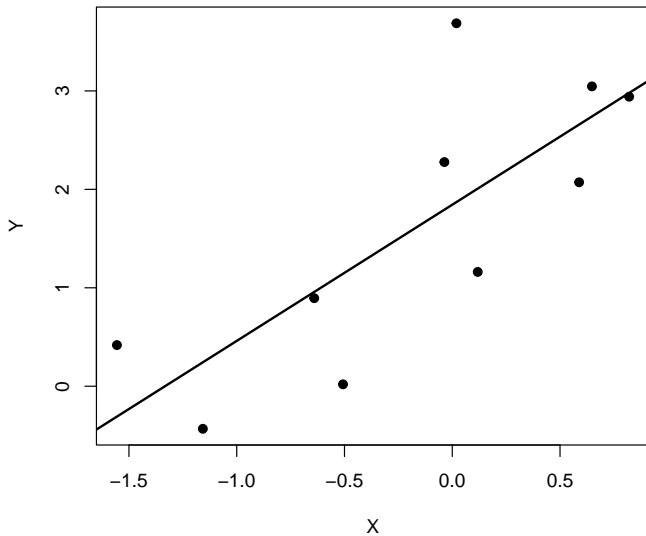
set.seed(1337)
X<-rnorm(N)
Y<-2+2*X+rnorm(N)
data<-data.frame(Y,X)
fitOLS<-lm(Y~X)
CI<-confint(fitOLS)

B0<-numeric(reps)
B1<-numeric(reps)

for (i in 1:reps) {
  temp<-data[sample(1:N,N,replace=TRUE),]
  temp.lm<-lm(Y~X,data=temp)
  B0[i]<-temp.lm$coefficients[1]
  B1[i]<-temp.lm$coefficients[2]
}

ByHandB0<-median(B0)
ByHandB1<-median(B1)
ByHandCI.B0<-quantile(B0,probs=c(0.025,0.975)) # <-- 95% c.i.s
ByHandCI.B1<-quantile(B1,probs=c(0.025,0.975))
```

Normal Residuals...



Bootstrapping Via boot

```
library(boot)

Bs<-function(formula, data, indices) { # <- regression function
  dat <- data[indices,]
  fit <- lm(formula, data=dat)
  return(coef(fit))
}

Boot.fit<-boot(data=data, statistic=Bs,
               R=reps, formula=Y~X)

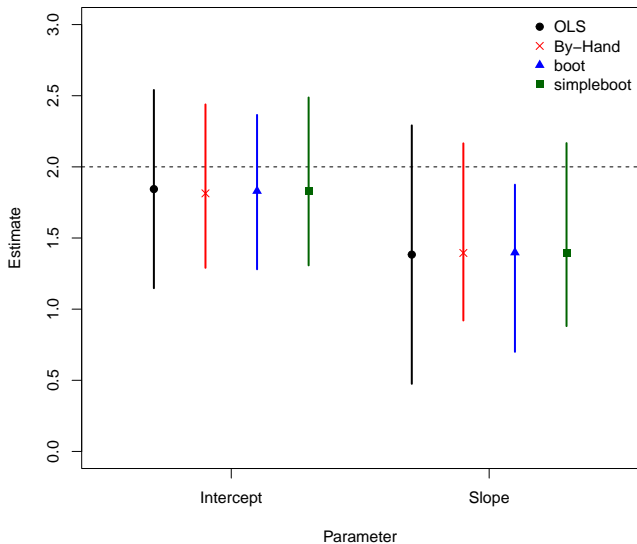
BootB0<-median(Boot.fit$t[,1])
BootB1<-median(Boot.fit$t[,2])
BootCI.B0<-boot.ci(Boot.fit,type="basic",index=1)
BootCI.B1<-boot.ci(Boot.fit,type="basic",index=2)
```

Bootstrapping Via simpleboot

```
library(simpleboot)

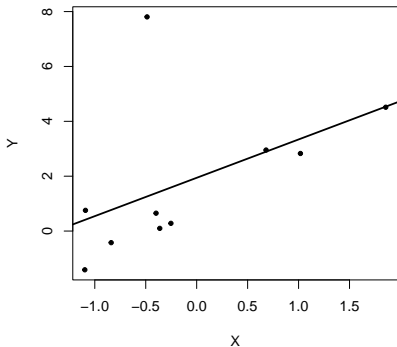
Simple<-lm.boot(fit0LS, reps)
SimpleB0<-perc(Simple, .50)[1]
SimpleB1<-perc(Simple, .50)[2]
Simple.CIs<-perc(Simple, p=c(0.025, 0.975))
```


Bootstrapping Results

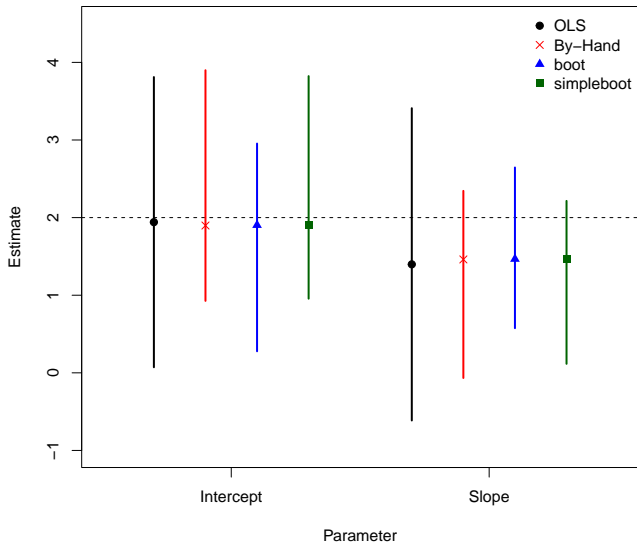


Bootstrapping: Skewed Residuals

```
N<-10  
reps<-1001  
  
set.seed(1337)  
X<-rnorm(N)  
ustar<-rgamma(N,shape=0.2,scale=1)*6 # <- skewed u.s  
Y<-2+2*X+(ustar-mean(ustar))  
data<-data.frame(Y,X)  
fitOLS<-lm(Y~X)  
CI<-confint(fitOLS)
```



Skewed Residuals: Results



R things:

- A [simple introduction](#) at StatMethods
- [Bootstrap in R](#) (at DataCamp)
- Packages: [boot](#), [bootstrap](#), [simpleboot](#), [car::Boot](#), [broom](#) (tidy), many more

Other Resources:

- Efron's [original \(1979\) paper](#)
- [Chernick and Labudde \(2011\)](#) (a solid R-based bootstrapping book)
- Many other books, etc.

Missing Data

Why are data missing?

- The observation itself does not exist
- Data don't exist for that observation
- Data exist, but are *impossible* to measure
- Data exist, but were not measured

Missing Data, Part II: Flavors

Notation:

$$\mathbf{X}_i \in \{\mathbf{W}_i, \mathbf{Z}_i\}$$

\mathbf{W}_i have some missing values,
 \mathbf{Z}_i are “complete”

$$R_{ik} = \begin{cases} 1 & \text{if } W_{ik} \text{ is missing,} \\ 0 & \text{otherwise.} \end{cases}$$

$$\pi_{ik} = \Pr(R_{ik} = 1)$$

Missing Data, Part II: Rubin's Flavors

Missing completely at random ("MCAR"):

$$\mathbf{R} \perp \{\mathbf{Z}, \mathbf{W}\}$$

Missing at random ("MAR"):

$$\mathbf{R} \perp \mathbf{W} | \mathbf{Z}$$

Anything else is "informatively" (or "non-ignorably," or sometimes "MNAR") missing.

MCAR vs. MAR vs. MNAR, Explained

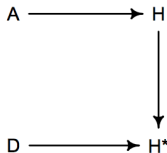
H: Homework

H*: Homework with missing values

A: Attribute of student

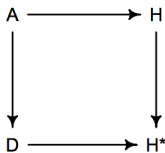
D: Dog (missingness mechanism)

**DOG EATS
ANY
HOMEWORK**



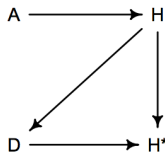
**MISSING COMPLETELY
AT RANDOM**

**DOG EATS
STUDENTS'
HOMEWORK**



**MISSING
AT RANDOM**

**DOG EATS
BAD
HOMEWORK**



**MISSING NOT
AT RANDOM**

(Source)

Missing Data: Consequences

In general:

- MCAR:
 - Missing data are a fully random sample of all the data
 - \rightarrow Missingness does not bias $\hat{\theta}$, *but*
 - There is some loss of information (and therefore efficiency)
- MAR
 - Missing data are a nonrandom sample of all the data
 - Ignoring missingness can lead to bias in $\hat{\theta}$, *but*
 - Conditioning on the variable(s) that drive the missingness can eliminate the bias
- Informative Missingness / MNAR
 - Missing data are a nonrandom sample of all the data
 - Ignoring missingness can lead to bias in $\hat{\theta}$
 - In general, conditioning cannot eliminate the bias

Example, Simulated

```
> set.seed(7222009)
> Npop <- 1000
> X<-runif(Npop,0,10)    # NOTE: X, Z are correlated a bit...
> Z<-(0.3*X)+(0.7*runif(Npop,0,10))
> Y<-0+(2*X)+(2*Z)+rnorm(Npop,mean=0,sd=4)
> DF<-data.frame(X=X,Z=Z,Y=Y)
> fit.pop<-lm(Y~X+Z,DF)
> summary(fit.pop)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.4051	0.3260	1.24	0.21
X	1.9553	0.0466	41.97	<2e-16 ***
Z	1.9812	0.0617	32.09	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.98 on 997 degrees of freedom

Multiple R-squared: 0.823, Adjusted R-squared: 0.823

F-statistic: 2.32e+03 on 2 and 997 DF, p-value: <2e-16

Simulated MCAR

```
> pmis<-0.50
> DF$Ymcar<-rbinom(Npop,1,pmis)
> DF$Ymcar<-ifelse(DF$Ymcar==1,NA,DF$Y)
>
> # Regression w/listwise deletion:
>
> fit.s<-lm(Ymcar~X+Z,DF) # <-- looks fine
> summary(fit.s)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.4442	0.4653	0.95	0.34
X	1.9661	0.0658	29.87	<2e-16 ***
Z	1.9763	0.0862	22.92	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

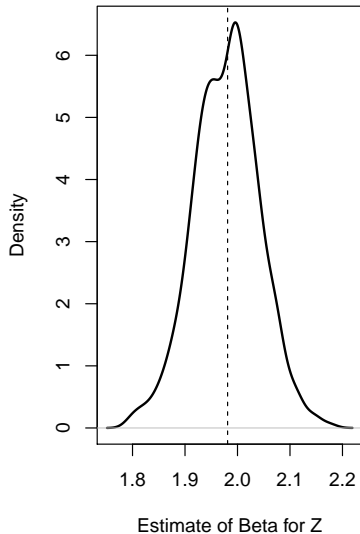
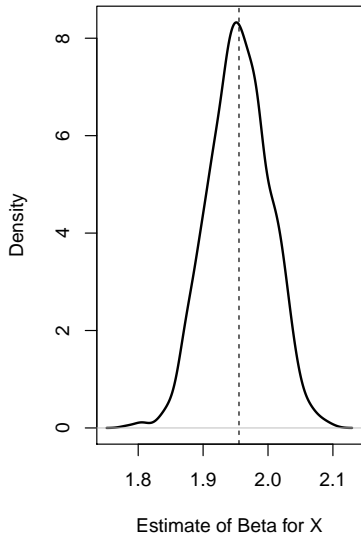
Residual standard error: 4 on 507 degrees of freedom

(490 observations deleted due to missingness)

Multiple R-squared: 0.822, Adjusted R-squared: 0.821

F-statistic: 1.17e+03 on 2 and 507 DF, p-value: <2e-16

Do That A Bunch Of Times...



```
> set.seed(7222009)
> DF$Ymar<-rbinom(Npop,1,(DF$Z/10))
> DF$Ymar<-ifelse(DF$Ymar==1,NA,DF$Y)
>
> summary(lm(Ymar~X,DF))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.6600	0.3610	10.1	<2e-16 ***
X	2.9923	0.0648	46.2	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

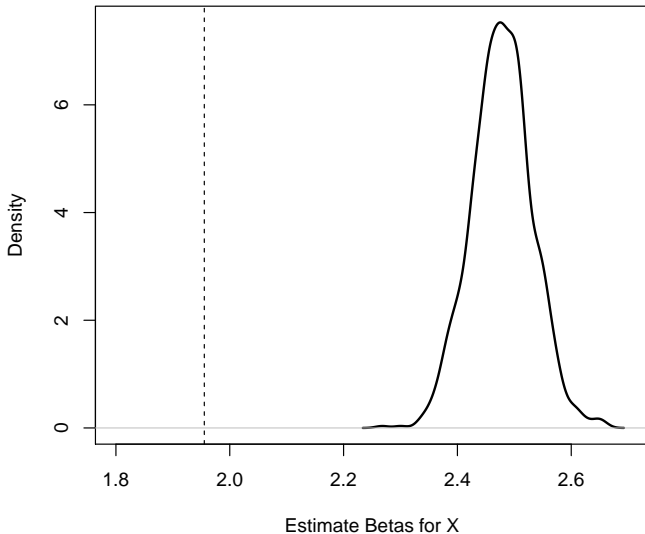
Residual standard error: 4.75 on 547 degrees of freedom

(451 observations deleted due to missingness)

Multiple R-squared: 0.796, Adjusted R-squared: 0.795

F-statistic: 2.13e+03 on 1 and 547 DF, p-value: <2e-16

Do That A Bunch Of Times...



More MAR: Add Z...

```
> summary(lm(Ymar~X+Z,DF))
```

Call:

```
lm(formula = Ymar ~ X + Z, data = DF)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.2529	0.4367	0.58	0.56
X	2.0200	0.0663	30.49	<2e-16 ***
Z	1.9499	0.0979	19.91	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

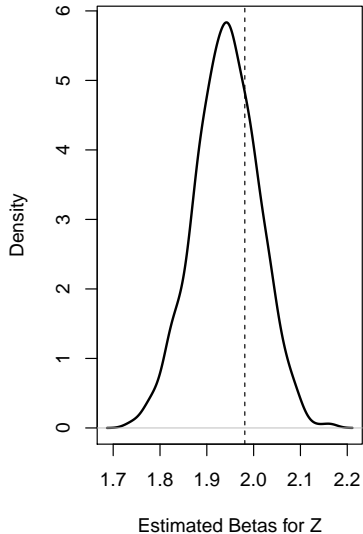
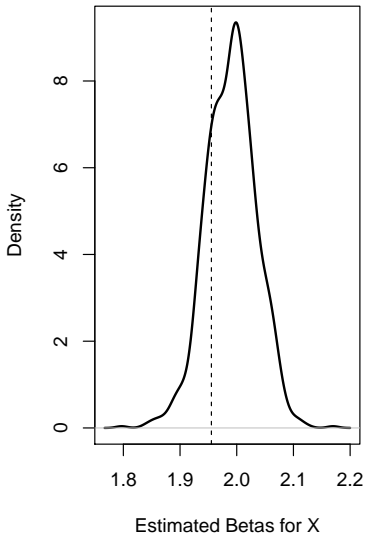
Residual standard error: 4.02 on 499 degrees of freedom

(498 observations deleted due to missingness)

Multiple R-squared: 0.801, Adjusted R-squared: 0.8

F-statistic: 1e+03 on 2 and 499 DF, p-value: <2e-16

Do That A Bunch Of Times...



Informative Missingness / “MNAR”

```
> set.seed(7222009)
> DF$Yim<-rbinom(Npop,1,rescale(DF$Z-(4*DF$Y)))
> DF$Yim<-ifelse(DF$Yim==1,NA,DF$Y)
>
> summary(lm(Yim~X+Z,DF))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.0518	0.5463	3.76	0.00019 ***
X	1.8420	0.0671	27.45	< 2e-16 ***
Z	1.9171	0.0859	22.32	< 2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

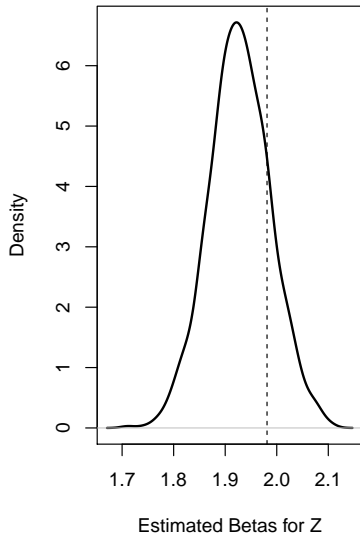
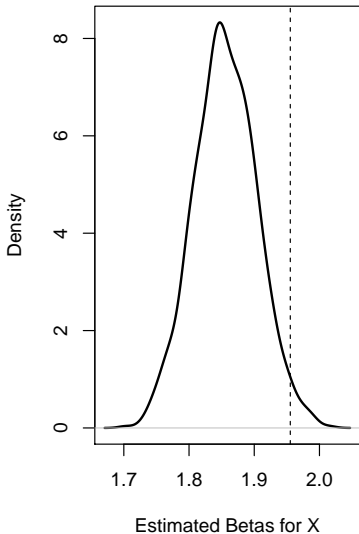
Residual standard error: 3.85 on 465 degrees of freedom

(532 observations deleted due to missingness)

Multiple R-squared: 0.797, Adjusted R-squared: 0.796

F-statistic: 911 on 2 and 465 DF, p-value: <2e-16

Do That A Bunch Of Times...



A Real-Data Examples: 2020 ANES

Model is:

$$\begin{aligned}\text{Biden Thermometer}_i &= \beta_0 + \beta_1 \text{R's Conservatism}_i + \\ &= \beta_2 \text{R Labor Union}_i + \beta_3 \text{Female}_i + \\ &= \beta_4 \text{Latino}_i + \beta_5 \text{Age} / 10_i + \\ &= \beta_6 \text{Education}_i + u_i\end{aligned}$$

Data: ANES 2016-2020 Panel data, 2020 pre-election survey ($N = 2839$).

Three models:

- All data ($N = 2291$)
- 67% MCAR (via simulation) ($N = 709$)
- (MNAR) Data *only* on individuals who stated that they “strongly approved” of how then-President Trump was doing his job ($N = 743$)

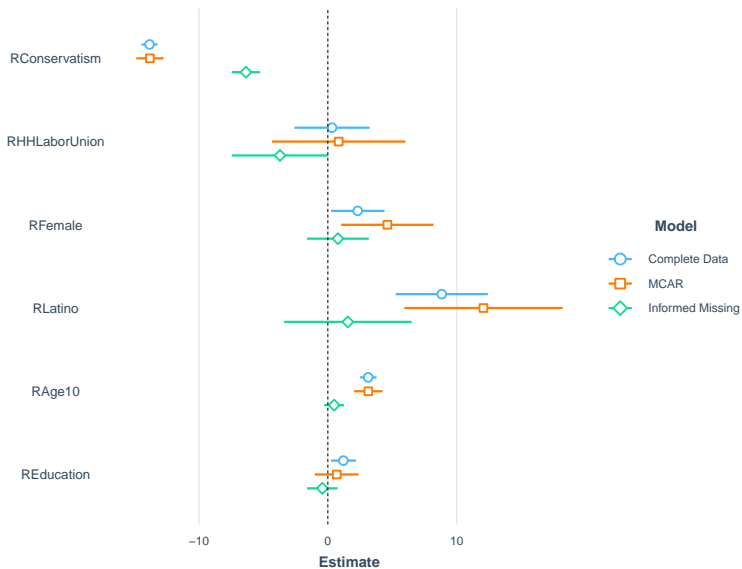
Biden Thermometer Models

	<i>Dependent variable:</i>		
	Biden Thermometer Rating		
	All Data	MCAR	MNAR
R's Conservatism	-13.841*** (0.319)	-13.817*** (0.538)	-6.354*** (0.561)
R Labor Union	0.325 (1.485)	0.844 (2.640)	-3.714* (1.906)
Female	2.317** (1.058)	4.621** (1.828)	0.783 (1.224)
Latino	8.842*** (1.824)	12.077*** (3.129)	1.550 (2.524)
Age / 10	3.131*** (0.328)	3.142*** (0.563)	0.490 (0.394)
Education	1.204** (0.498)	0.690 (0.864)	-0.427 (0.603)
Constant	83.222*** (3.039)	83.938*** (5.198)	47.563*** (4.165)
Observations	2,291	709	743
R ²	0.478	0.512	0.159
Adjusted R ²	0.477	0.508	0.152
Residual Std. Error	25.104 (df = 2284)	24.066 (df = 702)	16.616 (df = 736)

Note:

* p<0.1; ** p<0.05; *** p<0.01

Biden Thermometer Models (II)



How Much Missing Data Is A Problem?

"It is often supposed that there exists something like a critical missing rate up to which missing values are not too dangerous. The belief in such a global missing rate is rather stupid."

– Vach (1994, 113)

What to Do About Missing Data?

- Listwise Deletion...
- Mean Substitution / Imputation
- “Nearest Neighbor” methods
- “Hot Deck” Imputation
- Multiple Imputation
- Model-Based Solutions

For MAR data:

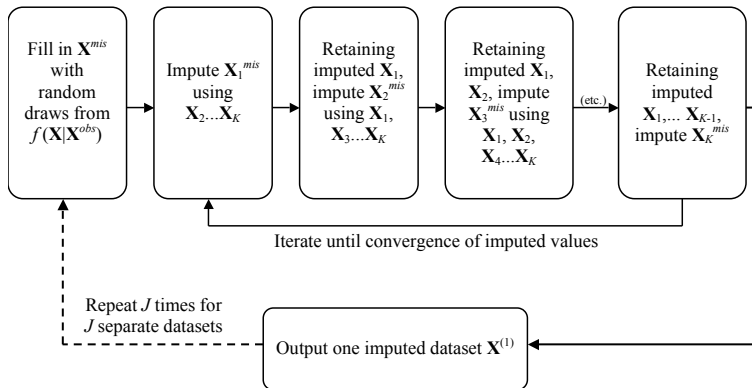
$$\mathbf{R} \perp \mathbf{W} | \mathbf{Z}$$

so \mathbf{W} and \mathbf{Z} factorize independently.

Sources of variation we need to consider:

1. Prediction
2. Predictive variation
3. Parameter variation / uncertainty

MAR: Multiple Imputation



Multiple Imputation (continued)

Original Data \mathbf{X} With Missing Data

i	X_1	X_2	X_3	X_4	...	X_K
1	X_{11}	X_{21}	X_{31}	X_{41}	...	X_{K1}
2	•	X_{22}	X_{32}	•	...	X_{K2}
3	X_{13}	X_{23}	•	X_{43}	...	X_{K3}
4	X_{14}	•	X_{34}	X_{44}	...	X_{K4}
5	•	X_{25}	X_{35}	•	...	•
6	X_{16}	X_{26}	X_{36}	X_{46}	...	X_{K6}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
N	X_{1N}	X_{2N}	X_{3N}	X_{4N}	...	X_{KN}

Multiple Imputation (continued)

Iteration One:

Step One: "Fill In" Missing Values of \mathbf{X}

i	X_1	X_2	X_3	X_4	...	X_K
1	X_{11}	X_{21}	X_{31}	X_{41}	...	X_{K1}
2	R_{12}	X_{22}	X_{32}	R_{42}	...	X_{K2}
3	X_{13}	X_{23}	R_{33}	X_{43}	...	X_{K3}
4	X_{14}	R_{24}	X_{34}	X_{44}	...	X_{K4}
5	R_{15}	X_{25}	X_{35}	R_{45}	...	R_{K5}
6	X_{16}	X_{26}	X_{36}	X_{46}	...	X_{K6}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
N	X_{1N}	X_{2N}	X_{3N}	X_{4N}	...	X_{KN}

Multiple Imputation (continued)

Step Two: Use $\{X_2, X_3, \dots, X_K\}$ To Impute X_1^{mis}

i	X_1	X_2	X_3	X_4	...	X_K
1	X_{11}	X_{21}	X_{31}	X_{41}	...	X_{K1}
2	$X_{12}^{(1)}$	X_{22}	X_{32}	R_{42}	...	X_{K2}
3	X_{13}	X_{23}	R_{33}	X_{43}	...	X_{K3}
4	X_{14}	R_{24}	X_{34}	X_{44}	...	X_{K4}
5	$X_{15}^{(1)}$	X_{25}	X_{35}	R_{45}	...	R_{K5}
6	X_{16}	X_{26}	X_{36}	X_{46}	...	X_{K6}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
N	X_{1N}	X_{2N}	X_{3N}	X_{4N}	...	X_{KN}

Multiple Imputation (continued)

Step Three: Use The Imputed X_1 , Along With $\{X_3, X_4, \dots, X_K\}$ To Impute

X_2^{mis}

i	X_1	X_2	X_3	X_4	...	X_K
1	X_{11}	X_{21}	X_{31}	X_{41}	...	X_{K1}
2	$I_{12}^{(1)}$	X_{22}	X_{32}	R_{42}	...	X_{K2}
3	X_{13}	X_{23}	R_{33}	X_{43}	...	X_{K3}
4	X_{14}	$I_{24}^{(1)}$	X_{34}	X_{44}	...	X_{K4}
5	$I_{15}^{(1)}$	X_{25}	X_{35}	R_{45}	...	R_{K5}
6	X_{16}	X_{26}	X_{36}	X_{46}	...	X_{K6}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
N	X_{1N}	X_{2N}	X_{3N}	X_{4N}	...	X_{KN}

Multiple Imputation (continued)

Step Four: Use The Imputed X_1 and X_2 , Along With $\{X_4, \dots, X_K\}$ To Impute X_3^{mis}

i	X_1	X_2	X_3	X_4	...	X_K
1	X_{11}	X_{21}	X_{31}	X_{41}	...	X_{K1}
2	$I_{12}^{(1)}$	X_{22}	X_{32}	R_{42}	...	X_{K2}
3	X_{13}	X_{23}	$I_{33}^{(1)}$	X_{43}	...	X_{K3}
4	X_{14}	$I_{24}^{(1)}$	X_{34}	X_{44}	...	X_{K4}
5	$I_{15}^{(1)}$	X_{25}	X_{35}	R_{45}	...	R_{K5}
6	X_{16}	X_{26}	X_{36}	X_{46}	...	X_{K6}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
N	X_{1N}	X_{2N}	X_{3N}	X_{4N}	...	X_{KN}

(etc.)

Multiple Imputation (continued)

Step $K + 1$: Use The Imputed X_1, X_2, \dots, X_{K-1} To Impute X_K^{mis}

i	X_1	X_2	X_3	X_4	...	X_K
1	X_{11}	X_{21}	X_{31}	X_{41}	...	X_{K1}
2	$I_{12}^{(1)}$	X_{22}	X_{32}	$I_{42}^{(1)}$...	X_{K2}
3	X_{13}	X_{23}	$I_{33}^{(1)}$	X_{43}	...	X_{K3}
4	X_{14}	$I_{24}^{(1)}$	X_{34}	X_{44}	...	X_{K4}
5	$I_{15}^{(1)}$	X_{25}	X_{35}	$I_{45}^{(1)}$...	$I_{K5}^{(1)}$
6	X_{16}	X_{26}	X_{36}	X_{46}	...	X_{K6}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
N	X_{1N}	X_{2N}	X_{3N}	X_{4N}	...	X_{KN}

Multiple Imputation (continued)

Iteration Two:

Step One: Use The Imputed X_2, X_3, \dots, X_K To Impute X_1^{mis}

i	X_1	X_2	X_3	X_4	...	X_K
1	X_{11}	X_{21}	X_{31}	X_{41}	...	X_{K1}
2	$I_{12}^{(2)}$	X_{22}	X_{32}	$I_{42}^{(1)}$...	X_{K2}
3	X_{13}	X_{23}	$I_{33}^{(1)}$	X_{43}	...	X_{K3}
4	X_{14}	$I_{24}^{(1)}$	X_{34}	X_{44}	...	X_{K4}
5	$I_{15}^{(2)}$	X_{25}	X_{35}	$I_{45}^{(1)}$...	$I_{K5}^{(1)}$
6	X_{16}	X_{26}	X_{36}	X_{46}	...	X_{K6}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
N	X_{1N}	X_{2N}	X_{3N}	X_{4N}	...	X_{KN}

Multiple Imputation (continued)

Step Two: Use The Imputed X_1, X_3, \dots, X_K To Impute X_2^{mis}

i	X_1	X_2	X_3	X_4	...	X_K
1	X_{11}	X_{21}	X_{31}	X_{41}	...	X_{K1}
2	$I_{12}^{(2)}$	X_{22}	X_{32}	$I_{42}^{(1)}$...	X_{K2}
3	X_{13}	X_{23}	$I_{33}^{(1)}$	X_{43}	...	X_{K3}
4	X_{14}	$I_{24}^{(2)}$	X_{34}	X_{44}	...	X_{K4}
5	$I_{15}^{(2)}$	X_{25}	X_{35}	$I_{45}^{(1)}$...	$I_{K5}^{(1)}$
6	X_{16}	X_{26}	X_{36}	X_{46}	...	X_{K6}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
N	X_{1N}	X_{2N}	X_{3N}	X_{4N}	...	X_{KN}

(etc.)

Multiple Imputation (continued)

Step K : Use The Imputed X_1, X_2, \dots, X_{K-1} To Impute X_K^{mis}

i	X_1	X_2	X_3	X_4	...	X_K
1	X_{11}	X_{21}	X_{31}	X_{41}	...	X_{K1}
2	$I_{12}^{(2)}$	X_{22}	X_{32}	$I_{42}^{(2)}$...	X_{K2}
3	X_{13}	X_{23}	$I_{33}^{(2)}$	X_{43}	...	X_{K3}
4	X_{14}	$I_{24}^{(2)}$	X_{34}	X_{44}	...	X_{K4}
5	$I_{15}^{(2)}$	X_{25}	X_{35}	$I_{45}^{(2)}$...	$I_{K5}^{(2)}$
6	X_{16}	X_{26}	X_{36}	X_{46}	...	X_{K6}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
N	X_{1N}	X_{2N}	X_{3N}	X_{4N}	...	X_{KN}

Multiple Imputation: Summary

- Repeat this process for $J \approx 10$ iterations until convergence of the $I_{ki}^{(j)}$ s.
- Output the resulting imputed data $\mathbf{X}^{(1)}$.
- Repeat this entire process M times to create M imputed datasets $\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \dots, \mathbf{X}^{(M)}\}$.
- Rule of thumb: $M \geq$ the percentage of cases in your data with missingness.
- Estimate models and conduct inference using multiple analysis and model averaging (see e.g. Schafer 1997, Ch. 4).

For MNAR data:

$$\Pr(\mathbf{R}) = f(\mathbf{W}, \mathbf{Z})$$

i.e., missingness is *nonignorable*.

Common causes / situations:

- Omitted variables (\rightarrow can't condition on all elements of \mathbf{Z})
- Differential response due to unmeasured factors
- Truncation / censoring

MNAR and Model-Based Solutions

For MNAR data, we must model the joint distribution $\Pr(\mathbf{X}, \mathbf{R})$...

Approaches:

- *Selection* model: $\Pr(\mathbf{X}, \mathbf{R}) = \Pr(\mathbf{X}) \Pr(\mathbf{R}|\mathbf{X})$
 - E.g., Heckman (1976, 1979, etc.)
 - Specifies a (usually, regression) model for $\Pr(\mathbf{R} | \mathbf{X})$
- *Pattern-Mixture* model: $\Pr(\mathbf{X}, \mathbf{R}) = \Pr(\mathbf{X}|\mathbf{R}) \Pr(\mathbf{R})$
$$= \Pr(\mathbf{X}|\mathbf{R} = 0) \Pr(\mathbf{R} = 0) + \Pr(\mathbf{X}|\mathbf{R} = 1) \Pr(\mathbf{R} = 1)$$
 - E.g., Glynn, Laird, and Rubin (1986)
 - Mixture-type model across “responders” and “non-responders”
- Others... [see, e.g., Little and Rubin (2002)]

Missing Data Resources, R and Otherwise

- The [Missing Data CRAN Task View](#)
- Packages:
 - [Amelia](#)
 - [mi](#), [mice](#), and [miceFast](#)
 - [miceMNAR](#) (MNAR imputation using a Heckman-style selection model)
 - [naniar](#) (tidy-cult, but enables [cool visualizations](#))
 - [VIM](#) (joint visualization and imputation of missing data; also used to have a GUI)
 - Many others...
- van Buuren's [Flexible Imputation of Missing Data](#) e-book