

PLSC 503 – Spring 2024

Bootstrapping and Missing Data

March 11, 2024

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

Practical (Nonparametric) Bootstrapping

The General Idea:

- 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

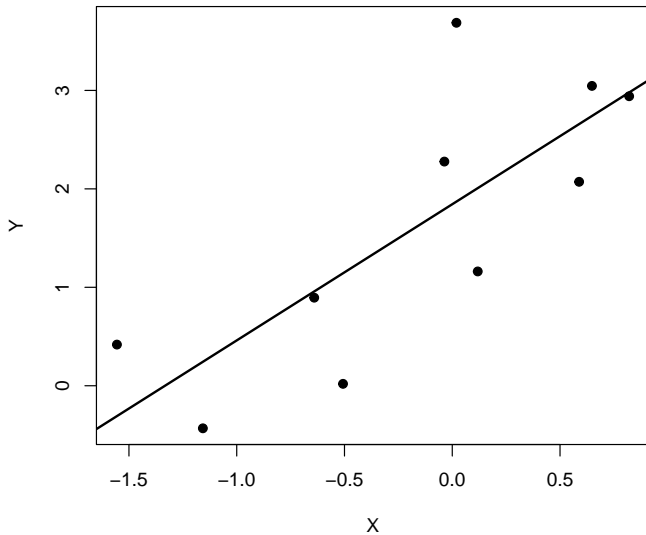
set.seed(1337)
X<-rnorm(N)
Y<-2+2*X+rnorm(N)
data<-data.frame(Y,X)
fitOLS<-lm(Y~X)
CI<-confint(fitOLS)

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)

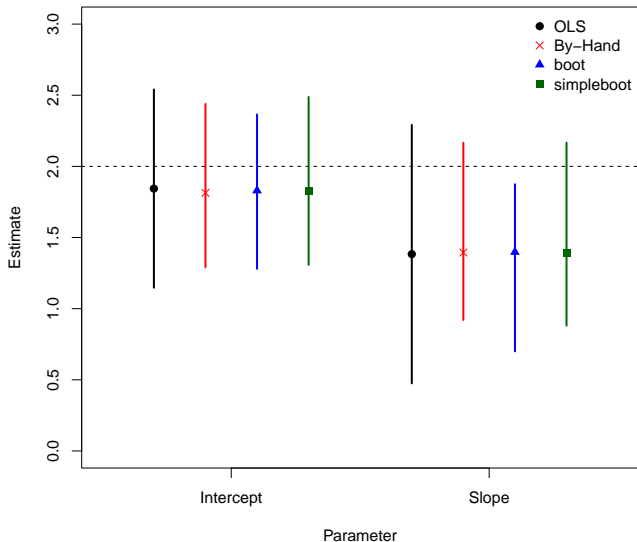
BootB0<-median(Boot.fit$t[,1])
BootB1<-median(Boot.fit$t[,2])
BootCI.B0<-boot.ci(Boot.fit,type="basic",index=1)
BootCI.B1<-boot.ci(Boot.fit,type="basic",index=2)
```

Bootstrapping Via simpleboot

```
library(simpleboot)

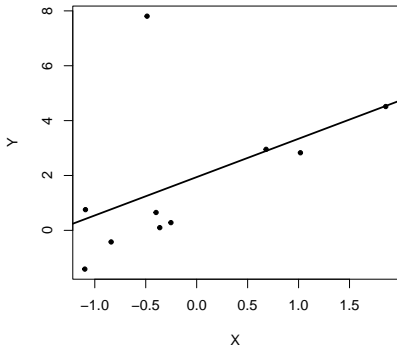
Simple<-lm.boot(fitOLS, reps)
SimpleB0<-perc(Simple, .50)[1]
SimpleB1<-perc(Simple, .50)[2]
Simple.CIs<-perc(Simple, p=c(0.025, 0.975))
```


Bootstrapping Results

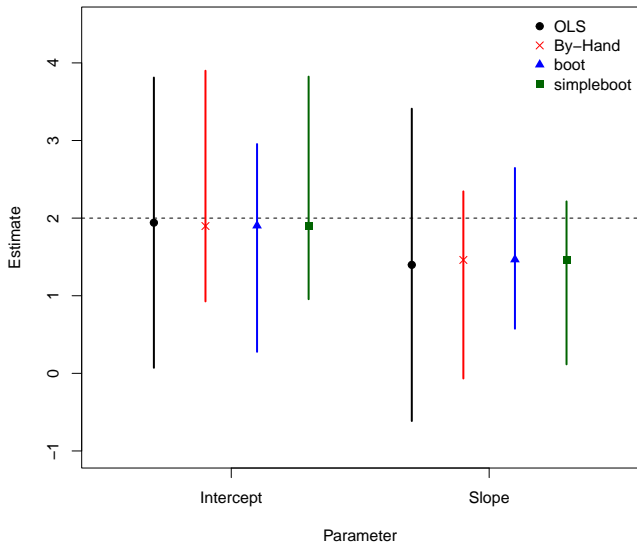


Bootstrapping: Skewed Residuals

```
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



When Should I Bootstrap?

A few canonical applications:

- When N is small, and the estimator is complex (/ consistent)
-

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)
- A good little [online intro](#), by James Scott
- 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

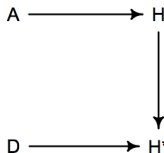
H: Homework

H*: Homework with missing values

A: Attribute of student

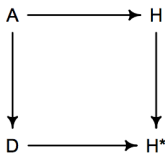
D: Dog (missingness mechanism)

**DOG EATS
ANY
HOMEWORK**



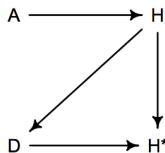
**MISSING COMPLETELY
AT RANDOM**

**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 ***

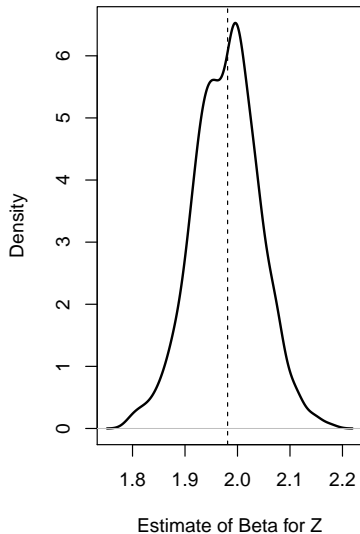
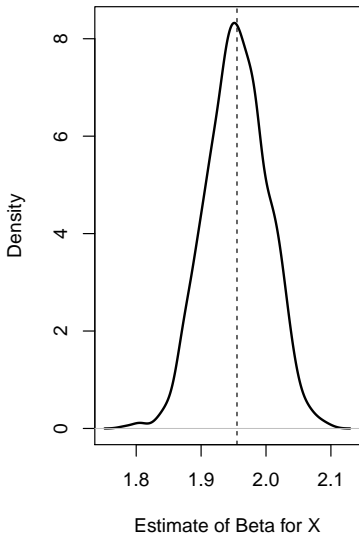
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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 ***

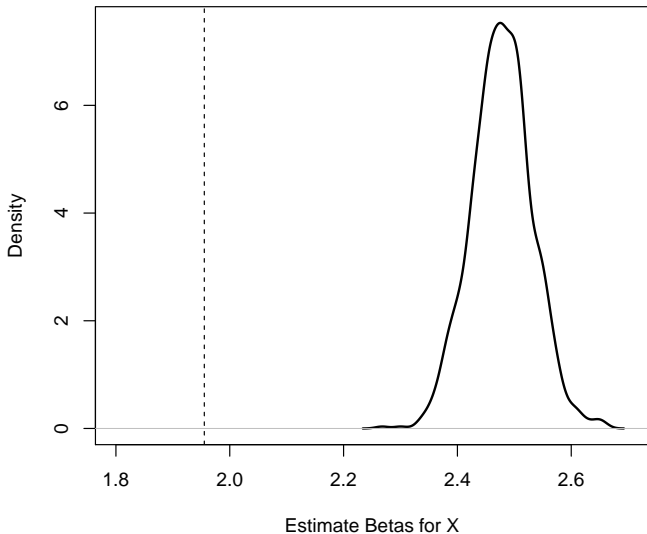
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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 ***

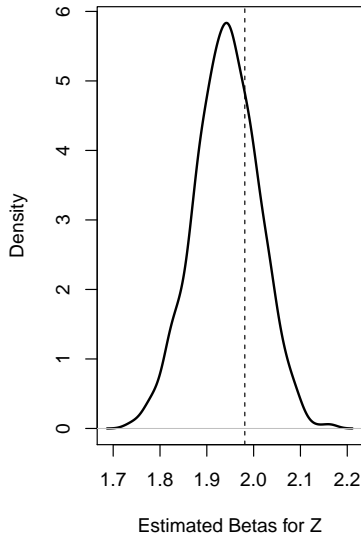
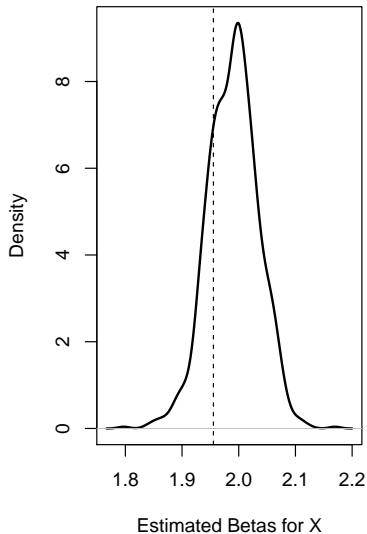
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

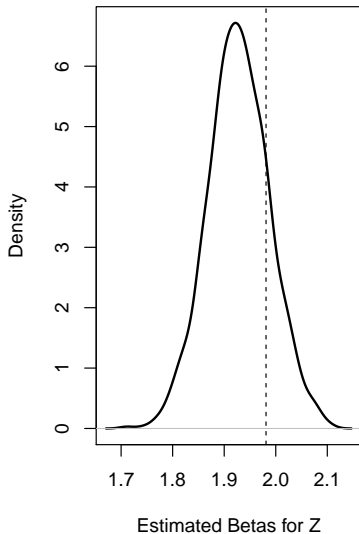
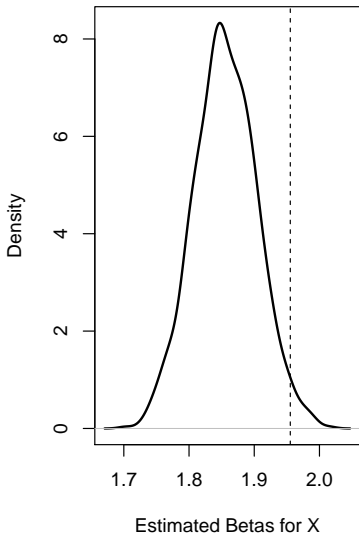
Residual standard error: 3.85 on 465 degrees of freedom

(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 ($N = 743$)

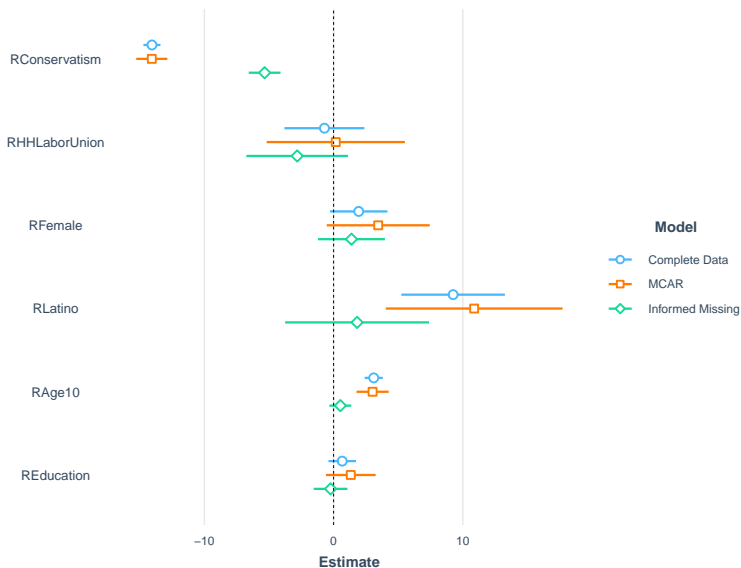
Biden Thermometer Models

	<i>Dependent variable:</i>		
	Biden Thermometer Score		
	Complete	2/3 MCAR data	MNAR (Trump Supporters)
R's Conservatism	−14.060*** (0.336)	−14.070*** (0.613)	−5.340*** (0.627)
R in Labor Union	−0.710 (1.578)	0.168 (2.723)	−2.817 (2.004)
R is Female	1.943* (1.135)	3.454* (2.029)	1.384 (1.317)
R is Latino	9.251*** (2.042)	10.880*** (3.483)	1.818 (2.835)
R's Age(/10)	3.106*** (0.357)	3.018*** (0.634)	0.524 (0.432)
R's Education	0.666 (0.542)	1.330 (0.978)	−0.234 (0.663)
Constant	86.870*** (3.306)	82.890*** (5.915)	40.440*** (4.750)
Observations	1,942	583	621
R ²	0.497	0.511	0.114
Adjusted R ²	0.495	0.506	0.105
Residual Std. Error	24.770 (df = 1935)	24.240 (df = 576)	16.350 (df = 614)
F Statistic	318.300*** (df = 6; 1935)	100.500*** (df = 6; 576)	13.150*** (df = 6; 614)

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:

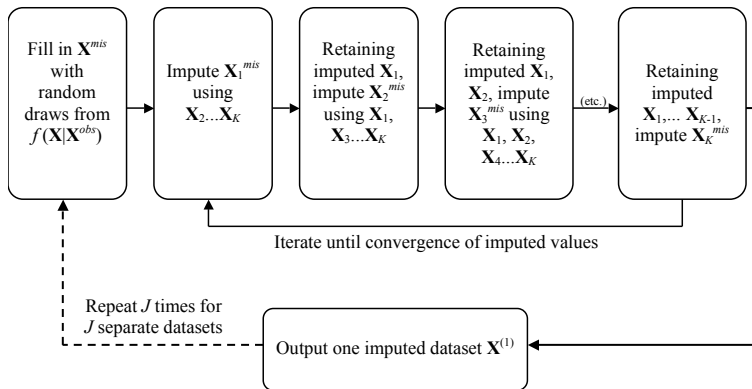
$$\mathbf{R} \perp \mathbf{W} | \mathbf{Z}$$

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	$\hat{X}_{12}^{(1)}$	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	$\hat{X}_{15}^{(1)}$	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

Basically:

- 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: “Set $M \geq$ 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)]

Multiple Imputation Example: ANES

Earlier, we created a data frame with $\approx 75\%$ MCAR missingness on the `BidenThermometer` variable:

```
> describe(MCAR.ANES)
      vars    n  mean    sd median trimmed   mad min max range  skew kurtosis   se
MCARBidenTherm 1  583 47.85 34.50   50.0   47.66 51.89 0.0 100 100.0 -0.12   -1.43 1.43
RConservatism  2 1942 4.11  1.75    4.0    4.11  2.97 1.0  7  6.0 -0.05   -1.14 0.04
RHHLaborUnion  3 1942 0.15  0.36    0.0    0.06  0.00 0.0  1  1.0  1.95    1.80 0.01
RFemale        4 1942 0.52  0.50    1.0    0.53  0.00 0.0  1  1.0 -0.10   -1.99 0.01
RLatino        5 1942 0.09  0.28    0.0    0.00  0.00 0.0  1  1.0  2.95    6.71 0.01
RAge10         6 1942 5.27  1.62    5.4    5.29  2.08 1.9  8  6.1 -0.08   -1.13 0.04
REducation     7 1942 3.57  1.07    4.0    3.62  1.48 1.0  5  4.0 -0.26   -0.67 0.02
```

We can multiply impute values for `MCARBidenTherm` using (e.g.) `mice`:

```
> mice.mcar<-mice(MCAR.ANES,m=75,seed=7222009) # MICE object

iter imp variable
1    1  MCARBidenTherm
1    2  MCARBidenTherm
1    3  MCARBidenTherm
.
.
.
5   74  MCARBidenTherm
5   75  MCARBidenTherm
```

Multiple Imputation Example: ANES (continued)

Re-run the regression on the multiply-imputed data:

```
> fit.imputed.mcar<-with(mice.mcar,lm(MCARBidenTherm~RConservatism+
+                                     RHHLaborUnion+RFemale+RLatino+RAge10+
+                                     REducation))

> summary(pool(fit.imputed.mcar))
```

	term	estimate	std.error	statistic	df	p.value
1	(Intercept)	84.5889	5.1475	16.4330	152.9	1.336e-35
2	RConservatism	-14.0159	0.5088	-27.5478	163.4	2.676e-63
3	RHHLaborUnion	0.4795	2.3107	0.2075	179.2	8.358e-01
4	RFemale	3.5353	1.9616	1.8022	123.1	7.396e-02
5	RLatino	12.0907	3.2318	3.7412	147.3	2.617e-04
6	RAge10	2.8790	0.5532	5.2045	154.3	6.114e-07
7	REducation	0.8884	0.9066	0.9800	131.2	3.289e-01

Compare to the “complete” data::

```
> summary(fit.all)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	86.868	3.306	26.28	< 2e-16 ***
RConservatism	-14.060	0.336	-41.88	< 2e-16 ***
RHHLaborUnion	-0.710	1.578	-0.45	0.653
RFemale	1.943	1.135	1.71	0.087 .
RLatino	9.251	2.042	4.53	0.0000063 ***
RAge10	3.106	0.357	8.71	< 2e-16 ***
REducation	0.666	0.542	1.23	0.219

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

Residual standard error: 24.8 on 1935 degrees of freedom
Multiple R-squared: 0.497, Adjusted R-squared: 0.495
F-statistic: 318 on 6 and 1935 DF, p-value: <2e-16

Does this work for MNAR data?

MNAR ANES data:

```
> describe(MNAR.ANES)
      vars   n  mean   sd median trimmed  mad min max range  skew kurtosis   se
MNARBidenTherm  1  621 12.40 17.28   0.0   9.14 0.00 0.0  90  90.0  1.53    2.24 0.69
RConservatism   2 1942  4.11  1.75   4.0   4.11 2.97 1.0   7   6.0 -0.05   -1.14 0.04
RHHLaborUnion   3 1942  0.15  0.36   0.0   0.06 0.00 0.0   1   1.0  1.95    1.80 0.01
RFemale         4 1942  0.52  0.50   1.0   0.53 0.00 0.0   1   1.0 -0.10   -1.99 0.01
RLatino         5 1942  0.09  0.28   0.0   0.00 0.00 0.0   1   1.0  2.95    6.71 0.01
RAge10         6 1942  5.27  1.62   5.4   5.29 2.08 1.9   8   6.1 -0.08   -1.13 0.04
REducation      7 1942  3.57  1.07   4.0   3.62 1.48 1.0   5   4.0 -0.26   -0.67 0.02

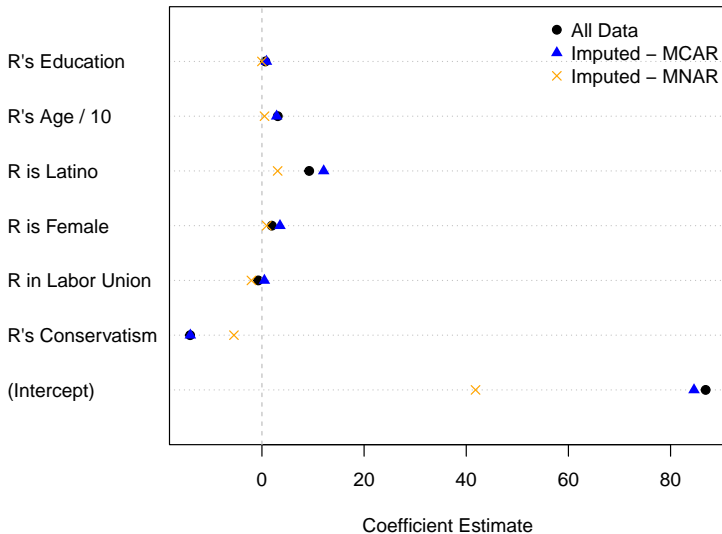
> mice.mnar<-mice(MNAR.ANES,m=75,seed=7222009) # MICE object

iter imp variable
  1   1 MNARBidenTherm
  1   2 MNARBidenTherm
.
.
.

> fit.imputed.mnar<-with(mice.mnar,lm(MNARBidenTherm~RConservatism+RHHLaborUnion+RFemale+RLatino+
+                                     RAge10+REducation))

> summary(pool(fit.imputed.mnar))
      term estimate std.error statistic    df p.value
1 (Intercept) 41.816443   4.9672   8.418599 153.9 2.472e-14
2 RConservatism -5.478921   0.5402  -10.141586 132.6 3.051e-18
3 RHHLaborUnion -2.058441   2.5739   -0.799749 129.1 4.253e-01
4 RFemale      0.936825   1.8626   0.502967 127.6 6.159e-01
5 RLatino      3.073983   3.5359   0.869358 115.8 3.864e-01
6 RAge10      0.508512   0.5681   0.895163 135.3 3.723e-01
7 REducation  -0.004455   0.7982   -0.005581 162.0 9.956e-01
```

Imputed Thermometer Model Estimated $\hat{\beta}$ s



Missing Data Resources, R and Otherwise

Check out:

- 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 2e](#) e-book