Statistical Analyses with Missing Data

Day 1

Cindy Chen and Bryan Shepherd

Vanderbilt-Nigeria Biostatistics Workshop

June 2–6, 2025

Overview

Missing data

Missing data

- Missing data are very common in biomedical research
- They can be a major setback to analyses
  (bias estimates, reduce sample size)
- It is important to understand approaches to address missing data

Course Topics

Introduction to Missing Data

- Potential problems with missing data
- Different types of missingness

Strategies for dealing with missing data

- Simple approaches
- Inverse probability weighted estimators
- Multiple imputation
- More advanced approaches

Our goal is for you to understand theory, methods, and application

- Lectures, simulation exercises, and real data analysis

Outline for Course

Day 1: Introduction to Missing Data

- Potential problems with missing data
- Different types of missingness
- Simple approaches for dealing with missing data

Day 2: Inverse probability weighted estimators

Days 3-4: Multiple imputation

Introduction to Missing Data

Some Reasons for Missing Data

- Refusal to respond to a survey
- Failure to take measurement (low funding, provider feels not necessary)
- Failure to record a measurement
- Drop out of a study (patient decision)
- Removal from a study (researcher / doctor decision)
- Death

Notation

Y = Outcome variable

R = Response indicator

- $R = 1$ if Y is observed
- $R = 0$ if Y is missing

X = Covariates of direct interest

V = Auxiliary covariates (available, not of direct interest)

Defining the estimation target with incomplete data

Possible targets of estimation

- **Full-data parameter:** Mean outcome among all individuals intended to be in the sample, whether or not they are observed

$$\mu = E(Y)$$

- **Observed-data parameter:** Mean response among all individuals whose outcome was observed

$$\mu_1 = E(Y|R = 1)$$

When does $\mu = \mu_1$?

Defining the estimation target with incomplete data

Possible targets of estimation

- **Full-data parameter:** Regression parameters among all individuals intended to be in the sample

$$E(Y|X) = X^T\beta$$

- **Observed-data parameter:** Regression parameters among individuals with observed outcome

$$E(Y|X, R = 1) = X^T\beta_1$$

When does $\beta = \beta_1$?

The need for assumptions to estimate full-data parameters

- Cannot estimate parameters for parts of data that are missing
- Hence need assumptions about the missing data
  - These are called missing data mechanisms
- Under most circumstances, these assumptions cannot be tested
- This motivates the need to:
  - State the assumptions unambiguously so others can critique them
  - Carry out sensitivity analysis when possible

Missing data mechanisms

Classification of association between R and Y

- MCAR: Missing completely at random
- MAR: Missing at random
- MNAR: Missing not at random

These are sometimes defined conditionally on covariates.

Statistical independence

The notation $R \perp Y$ means that the random variable R is independent of the random variable Y.

Implications: - Joint distribution can be factored

$$f(r, y) = f(r)f(y)$$

- Conditional distributions and expectations

$$f(y|r) = f(y)$$

$$E(Y|R) = E(Y)$$

Knowing R does not influence the distribution or expectation of Y

Missing data mechanism for univariate sampling

Missing values of Y are *missing completely at random (MCAR)* if $R \perp Y$, or equivalently, if

$$f(r|y) = f(r).$$

For univariate samples, this is also classified as *missing at random (MAR)*. More on this distinction later.

Missing data mechanism for univariate sampling

Missing values of Y are *missing not at random (MNAR)* if there exists at least one value of y such that

$$f(r|y) \neq f(r).$$

Or in words, the probability of response is systematically higher/lower for particular values of y.

Missing data mechanism for univariate sampling

- Under MCAR/MAR, methods applied to the observed data only will generally yield valid inferences about the population.
  - Estimates will be consistent
  - Standard errors will generally be larger than if you had the full data

- Under MNAR, methods applied to the observed data only generally will NOT yield valid inferences

Example of missing data mechanism for univariate sampling

Generating MCAR data

- $Y \perp R$

```r
set.seed(2)
y<-rnorm(25)
r<-rbinom(25,1,.8)
mean(y)
```

```
## [1] 0.3339737
```

```r
mean(y[r==1])
```

```
## [1] 0.3411507
```

In this simulation, by chance, $\hat{E}(Y|R=1) \approx \hat{E}(Y)$. But as the sample size increases, this is guaranteed to hold, i.e.,

$$\hat{E}(Y|R=1) \rightarrow E(Y).$$

```
##              y r
## 1  -2.21469989 1
## 2  -1.98935170 1
## 3  -0.83562861 1
## 4  -0.82046838 1
## 5  -0.62645381 1
## 6  -0.62124058 1
## 7  -0.30538839 1
## 8  -0.04493361 1
## 9  -0.01619026 1
## 10  0.07456498 1
## 11  0.18364332 0
## 12  0.32950777 1
## 13  0.38984324 1
## 14  0.48742905 1
## 15  0.57578135 1
## 16  0.59390132 0
## 17  0.61982575 1
## 18  0.73832471 1
## 19  0.78213630 0
## 20  0.82122120 1
## 21  0.91897737 1
## 22  0.94383621 1
## 23  1.12493092 1
## 24  1.51178117 0
## 25  1.59528080 1
```

Example of missing data mechanism for univariate sampling

Generating MNAR data

- R depends on Y

```r
set.seed(2)
y<-rnorm(25)
p<-ifelse(y<0.9,1,0.4)
r<-rbinom(25,1,p)
mean(y)
```

```
## [1] 0.3339737
```

```r
mean(y[r==1])
```

```
## [1] 0.1710997
```

$\hat{E}(Y|R=1) \neq \hat{E}(Y)$, and as the sample size increases, this inequality is guaranteed, i.e.,

$$\hat{E}(Y|R=1) \rightarrow \gamma \neq E(Y).$$

```
##              y r
## 1  -2.21469989 1
## 2  -1.98935170 1
## 3  -0.83562861 1
## 4  -0.82046838 1
## 5  -0.62645381 1
## 6  -0.62124058 1
## 7  -0.30538839 1
## 8  -0.04493361 1
## 9  -0.01619026 1
## 10  0.07456498 1
## 11  0.18364332 1
## 12  0.32950777 1
## 13  0.38984324 1
## 14  0.48742905 1
## 15  0.57578135 1
## 16  0.59390132 1
## 17  0.61982575 1
## 18  0.73832471 1
## 19  0.78213630 1
## 20  0.82122120 1
## 21  0.91897737 0
## 22  0.94383621 0
## 23  1.12493092 0
## 24  1.51178117 1
## 25  1.59528080 0
```

Missing data mechanism for multivariate sampling – regression

- Consider the setting where we are interested in the regression of Y on X. Let

$$\mu(X) = E(Y|X).$$

- Assume there are no other covariates available
- Our model is

$$g(\mu(X)) = \beta_0 + \beta_1 X.$$

  The function $g(\cdot)$ tells us the type of regression model we are fitting (linear, logistic, etc.)
- The full data are

$$(Y_1, X_1, R_1), (Y_2, X_2, R_2), \ldots, (Y_n, X_n, R_n).$$

Missing data mechanism for multivariate sampling – regression

The Y's are *missing completely at random (MCAR)* if

$$Y \perp R.$$

The Y's are *missing at random (MAR)* if

$$Y \perp R|X.$$

- MCAR implies that there is random missingness.
- MAR implies that there is random missingness within distinct levels of X.

Written alternatively:

- MCAR: $f(y|r) = f(y)$
- MAR: $f(y|r,x) = f(y|x)$

Example of missing data mechanism for multivariate sampling

Generating MCAR data

- $Y \perp R$

```r
set.seed(4)
x<-rnorm(25); y<-rnorm(25,x,1)
r<-rbinom(25,1,.8)
lm(y~x)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##     0.02285      0.89926
```

```r
lm(y~x,subset=r==1)
```

```
##
## Call:
## lm(formula = y ~ x, subset = r == 1)
##
## Coefficients:
## (Intercept)            x
##      0.1302       0.9033
```

In this simulation, by chance, $\hat{\beta}_1 \approx \hat{\beta}$. But as the sample size increases, this is guaranteed to hold, i.e., $\hat{\beta}_1 \to \beta$.

```
##               x           y r
## 1  -1.28124663 -0.22931405 1
## 2  -0.54249257  0.71339145 1
## 3  -0.28344457 -1.10443815 1
## 4  -0.21314452 -0.96735573 1
## 5  -0.10036844 -1.78841702 0
## 6  -0.04513712 -0.51103300 1
## 7  -0.04420400  1.24830834 1
## 8   0.01571945 -0.21168596 1
## 9   0.03435191 -0.60319159 1
## 10  0.16516902  0.26401271 1
## 11  0.16902677  1.51273540 1
## 12  0.21675486 -0.06618882 1
## 13  0.38305734  1.31715351 1
## 14  0.56660450  0.16208467 1
## 15  0.59289694 -1.20448508 1
## 16  0.59598058 -0.33204753 1
## 17  0.68927544  0.84273962 1
## 18  0.89114465  1.80098380 1
## 19  1.16502684  1.34656222 1
## 20  1.28825688  2.01216103 1
## 21  1.30762236  0.93196722 0
## 22  1.54081498  0.67866884 1
## 23  1.63561800  2.87579884 1
## 24  1.77686321  2.63799509 0
## 25  1.89653987  0.41435075 0
```

Example of missing data mechanism for multivariate sampling

Generating MNAR data

- R depends on Y

```r
set.seed(4)
x<-rnorm(25); y<-rnorm(25,x,1)
p<-ifelse(y<1.5,1,0.4)
r<-rbinom(25,1,p)
lm(y~x)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##     0.02285      0.89926
```

```r
lm(y~x,subset=r==1)
```

```
##
## Call:
## lm(formula = y ~ x, subset = r == 1)
##
## Coefficients:
## (Intercept)            x
##     -0.1002       0.7089
```

```
##              x           y r
## 1  -1.28124663 -0.22931405 1
## 2  -0.54249257  0.71339145 1
## 3  -0.28344457 -1.10443815 1
## 4  -0.21314452 -0.96735573 1
## 5  -0.10036844 -1.78841702 1
## 6  -0.04513712 -0.51103300 1
## 7  -0.04420400  1.24830834 1
## 8   0.01571945 -0.21168596 1
## 9   0.03435191 -0.60319159 1
## 10  0.16516902  0.26401271 1
## 11  0.16902677  1.51273540 0
## 12  0.21675486 -0.06618882 1
## 13  0.38305734  1.31715351 1
## 14  0.56660450  0.16208467 1
## 15  0.59289694 -1.20448508 1
## 16  0.59598058 -0.33204753 1
## 17  0.68927544  0.84273962 1
## 18  0.89114465  1.80098380 0
## 19  1.16502684  1.34656222 1
## 20  1.28825688  2.01216103 0
## 21  1.30762236  0.93196722 1
## 22  1.54081498  0.67866884 1
## 23  1.63561800  2.87579884 1
## 24  1.77686321  2.63799509 0
## 25  1.89653987  0.41435075 1
```

In this simulation, $\hat{\beta}_1 \neq \hat{\beta}$. As the sample size increases, this inequality is guaranteed, i.e., $\hat{\beta}_1 \to \gamma \neq \beta$.

Example of missing data mechanism for multivariate sampling

Generating MAR data

- $Y \perp R | X$

```r
set.seed(4)
x<-rnorm(25); y<-rnorm(25,x,1)
p<-ifelse(x<1.5,1,0.4)
r<-rbinom(25,1,p)
lm(y~x)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##     0.02285      0.89926
```

```r
lm(y~x,subset=r==1)
```

```
##
## Call:
## lm(formula = y ~ x, subset = r == 1)
##
## Coefficients:
## (Intercept)            x
##     0.02445      0.67098
```

```
##              x           y r
## 1  -1.28124663 -0.22931405 1
## 2  -0.54249257  0.71339145 1
## 3  -0.28344457 -1.10443815 1
## 4  -0.21314452 -0.96735573 1
## 5  -0.10036844 -1.78841702 1
## 6  -0.04513712 -0.51103300 1
## 7  -0.04420400  1.24830834 1
## 8   0.01571945 -0.21168596 1
## 9   0.03435191 -0.60319159 1
## 10  0.16516902  0.26401271 1
## 11  0.16902677  1.51273540 1
## 12  0.21675486 -0.06618882 1
## 13  0.38305734  1.31715351 1
## 14  0.56660450  0.16208467 1
## 15  0.59289694 -1.20448508 1
## 16  0.59598058 -0.33204753 1
## 17  0.68927544  0.84273962 1
## 18  0.89114465  1.80098380 1
## 19  1.16502684  1.34656222 1
## 20  1.28825688  2.01216103 1
## 21  1.30762236  0.93196722 1
## 22  1.54081498  0.67866884 0
## 23  1.63561800  2.87579884 0
## 24  1.77686321  2.63799509 0
## 25  1.89653987  0.41435075 1
```

In this simulation, $\hat{\beta}_1 \neq \hat{\beta}$. However, as the sample size increases, $\hat{\beta}_1 \to \beta$.

Example of missing data mechanism for multivariate sampling

Generating MAR data

- $Y \perp R | X$

```
set.seed(4)
x<-rnorm(2500); y<-rnorm(2500,x,1)
p<-ifelse(x<1.5,1,0.4)
r<-rbinom(2500,1,p)
lm(y~x)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##   -0.003017     1.022946
```

```
lm(y~x,subset=r==1)
```

```
##
## Call:
## lm(formula = y ~ x, subset = r == 1)
##
## Coefficients:
## (Intercept)            x
##   -0.003317     1.021907
```

```
##              x           y r
## 1  -1.28124663 -0.22931405 1
## 2  -0.54249257  0.71339145 1
## 3  -0.28344457 -1.10443815 1
## 4  -0.21314452 -0.96735573 1
## 5  -0.10036844 -1.78841702 1
## 6  -0.04513712 -0.51103300 1
## 7  -0.04420400  1.24830834 1
## 8   0.01571945 -0.21168596 1
## 9   0.03435191 -0.60319159 1
## 10  0.16516902  0.26401271 1
## 11  0.16902677  1.51273540 1
## 12  0.21675486 -0.06618882 1
## 13  0.38305734  1.31715351 1
## 14  0.56660450  0.16208467 1
## 15  0.59289694 -1.20448508 1
## 16  0.59598058 -0.33204753 1
## 17  0.68927544  0.84273962 1
## 18  0.89114465  1.80098380 1
## 19  1.16502684  1.34656222 1
## 20  1.28825688  2.01216103 1
## 21  1.30762236  0.93196722 1
## 22  1.54081498  0.67866884 0
## 23  1.63561800  2.87579884 0
## 24  1.77686321  2.63799509 0
## 25  1.89653987  0.41435075 1
```

With a large sample size, $\hat{\beta}_1 \approx \hat{\beta}$.

MAR in regression – some practical issues

- MAR: $Y \perp R | X$

$$\Rightarrow E(Y|X, R = 1) = E(Y|X)$$

- Under MAR, inferences can still be valid even if
  - The X distribution is different between those with missing and observed Y's
- Questions to subjectively assess MAR:
  - Is the missing data mechanism a random deletion of Y's among people who have the same X values?
  - Is the relationship between X and Y the same among those with missing and observed Y values?

More Generally

In the above examples, we only considered missingness in Y.

Of course,

- X could be missing
- either X or Y missing
- both X and Y missing

There may also be many variables

- $Y, X_1, X_2, X_3, \ldots, X_p$

There could be many missing data indicators

- $R_0, R_1, R_2, R_3, \ldots, R_p$

More Generally

Let Y^m denote missing Y, Y^o denote observed Y.

Similarly, let X_1^m denote missing X_1, X_1^o denote observed X_1, $\ldots$.

MCAR:

- $(Y, X_1, \ldots, X_p) \perp (R_0, R_1, \ldots, R_p)$
- $(Y^m, X_1^m, \ldots, X_p^m) \perp (R_0, R_1, \ldots, R_p)$

MAR:

- $Y \perp R_0 | (X_1, \ldots, X_p); X_1 \perp R_1 | (Y, X_2, \ldots, X_p); \ldots; X_p \perp R_p | (Y, X_1, \ldots, X_{p-1})$
- $(Y^m, X_1^m, \ldots, X_p^m) \perp (R_0, R_1, \ldots, R_p) | (Y^o, X_1^o, \ldots, X_p^o)$
- The probability of being missing is the same within groups defined by the observed data.

MNAR:

- Missing data that are not MCAR or MAR.

Simulations Exercise 1A

1. Generate $n = 1000$ i.i.d. realizations of (Y, X, V) from a known joint distribution:
   - $V \sim N(0, 1)$, then $X|V \sim N(V, 1)$, and then $Y|X, V \sim N(X - V, 1)$
2. Estimate the following:
   - Regression coefficients for the model $E(Y|X) = \beta_0 + \beta_1 X + \beta_2 V$
   - Regression coefficients for the model $E(Y|X) = \gamma_0 + \gamma_1 X$
   - The mean of X

Estimate $(\beta_1, \beta_2), \gamma_1$, and $E(X)$ among those without missing data in the following scenarios:

3. Create MCAR data in X with approximately 50% missing (i.e., $P(R|X, Y, V) = P(R)$).
4. Create $\sim$ 50% missing data in X such that $P(R|X, Y, V) = P(R|V)$.
5. Create $\sim$ 50% missing data in X such that $P(R|X, Y, V) = P(R|X)$.
6. Create $\sim$ 50% missing data in X such that $P(R|X, Y, V) = P(R|Y)$.

Compare estimates in each setting with those under the full data model (question 2). Are they close? How do their standard errors compare?

Some simple (naive) solutions with missing data

- Complete case analysis
- Single imputation
- Missing data indicators

Complete case analysis

- This approach performs analyses only on those with complete data.
- This is probably the most common approach with missing data.
- But it is not a good approach; it tries to ignore the problem.
- We have seen that this leads to bias unless the data are MCAR (or in certain settings MAR).
- We have also seen that this leads to lower power and wider confidence intervals.
  - Especially the case when there are lots of variables subject to missingness.

MCAR with many variables

Generating MCAR data

- $X_1, X_2, X_3 \perp R_1, R_2, R_3$

```r
set.seed(5)
 n<-25
    x1<-rnorm(n); x2<-rnorm(n,x,1); x3<-rnorm(n,x,1)
    r1<-rbinom(n,1,3/4); r2<-rbinom(n,1,0.5); r3<-rbinom(n,1,0.5)
    r<-r1*r2*r3
    lm(y~x)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##     0.02285      0.89926
```

```r
    lm(y~x,subset=r==1)
```

```
##
## Call:
## lm(formula = y ~ x, subset = r == 1)
##
## Coefficients:
## (Intercept)            x
##    -0.20096      0.06505
```

```
##             x1          x2           x3           y r1 r2 r3 r
## 1  -2.18396676 -0.11878292 -0.345906279  1.24830834  1  0  0 0
## 2  -1.25549186  2.38991847  1.913167506  1.80098380  0  1  1 0
## 3  -1.08039260 -1.61741540 -0.195313080  1.31715351  1  0  0 0
## 4  -1.07176004 -0.10825622 -0.725706023 -0.60319159  0  0  0 0
## 5  -0.84085548 -0.07672699  1.680003426 -0.06618882  1  1  0 0
## 6  -0.80177945 -0.99381319  0.870549877 -0.21168596  1  1  1 1
## 7  -0.63537131  2.00231605 -0.325753589 -0.96735573  1  0  0 0
## 8  -0.60290798  1.00519048 -0.235677644  0.84273962  1  1  1 1
## 9  -0.59731309  0.36260366 -0.937302281  1.34656222  1  1  0 0
## 10 -0.47216639 -0.17155246 -0.527941832 -0.22931405  1  1  1 1
## 11 -0.28577363  3.11364351  1.832448944  0.41435075  1  0  0 0
## 12 -0.25935541 -0.74001351 -0.563110679 -1.10443815  0  0  0 0
## 13 -0.15753436 -1.80732299  0.451224423 -0.51103300  0  1  0 0
## 14 -0.13898614  1.71908714 -0.172359496  1.51273540  1  1  0 0
## 15  0.07014277 -0.06110152  0.004145744 -0.33204753  1  0  1 0
## 16  0.13810822  3.25608500  2.010138507  2.63799509  0  1  0 0
## 17  0.24081726  1.79529951 -1.372751885 -1.78841702  0  1  1 0
## 18  0.70676109  0.56392839  1.320624721  2.01216103  1  0  0 0
## 19  0.81900893  0.52368578  1.006428230 -1.20448508  1  1  1 1
## 20  0.90051195  2.10303834  1.336717660  0.67866884  1  0  0 0
## 21  0.94186939 -0.72183949 -0.060445166  0.26401271  0  0  0 0
## 22  1.22763034  1.51817833 -0.569978305  0.16208467  1  0  1 0
## 23  1.38435934  0.87609650 -0.354766475  0.71339145  1  0  0 0
## 24  1.46796190  0.84737778  1.654650812  0.93196722  0  1  1 0
## 25  1.71144087  0.78282256  1.523417346  2.87579884  1  0  1 0
```

Single Imputation

- Fill in the missing values with some other value.
- Lots of different choices for the imputation:
  - Mean / median / mode
  - Conditional mean
  - Conditional mean plus noise / random draw from fitted distribution
  - The same value as a similar observation in the dataset with complete data
- All of these approaches have limitations

Example Dataset

```r
#install.packages("mice")

library("mice")
summary(airquality)  # 153 observations
```

```
##      Ozone           Solar.R           Wind             Temp      
##  Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00  
##  1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00  
##  Median : 31.50   Median :205.0   Median : 9.700   Median :79.00  
##  Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88  
##  3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00  
##  Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00  
##  NA's   :37       NA's   :7                                       
##      Month            Day      
##  Min.   :5.000   Min.   : 1.0  
##  1st Qu.:6.000   1st Qu.: 8.0  
##  Median :7.000   Median :16.0  
##  Mean   :6.993   Mean   :15.8  
##  3rd Qu.:8.000   3rd Qu.:23.0  
##  Max.   :9.000   Max.   :31.0  
## 
```

Mean imputation

Replace missing values with their means

```
imp <- mice(airquality, method = "mean", m = 1, maxit = 1)
```

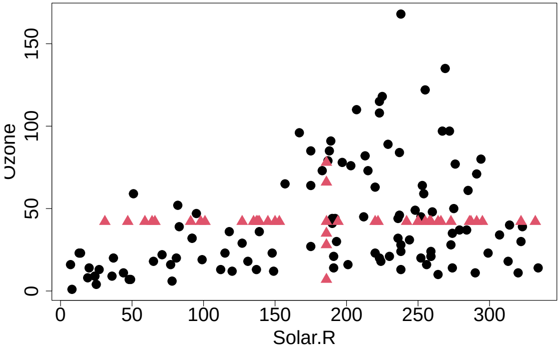
```
##
##  iter imp variable
##   1   1  Ozone  Solar.R
```

Complete case analysis:

- Standard deviation of $Y = 33.0$
- $E(Y|X) = 18.6 + 0.127X$

Analysis with mean imputation:

- Standard deviation of $Y = 28.7$
- $E(Y|X) = 23.7 + 0.099X$



Problems with Mean Imputation

- Distorts the distribution
- Under-estimates variation
- Results in bias in almost all situations
  - Unbiased if estimating mean and data are MCAR
- Changes (typically weakens) associations
- Although it's quick and easy, mean imputation should be avoided unless only a handful of values are missing.

Imputation with conditional expectation

Replace missing values with their conditional expectation

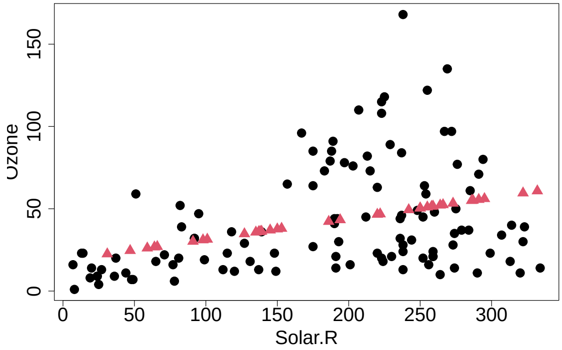
Complete case analysis:

- Standard deviation of $Y = 33.0$
- $E(Y|X) = 18.6 + 0.127X$
- Standard error of $\hat{\beta}_x = 0.033$

Analysis with mean imputation:

- Standard deviation of $Y = 29.4$
- $E(Y|X) = 18.6 + 0.127X$
- Standard error of $\hat{\beta}_x = 0.025$

(This analysis only imputed Y)



Problems with Conditional Mean Imputation

- Distorts the distribution
- Under-estimates variation
- Results in bias in most situations
  - Unbiased if estimating conditional expectation and data are MAR or MCAR
- Unrealistically strengthens associations
- Imputations are too good to be true
- Although it's quick and easy, conditional mean imputation should be avoided unless only a handful of values are missing.

Single imputation from the fitted distribution

Replace missing values with their conditional expectation plus some residual

- The residual can be estimated from the non-missing data

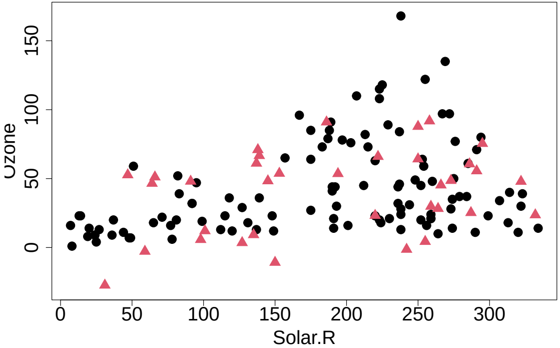
Complete case analysis:

- Standard deviation of $Y = 33.0$
- $E(Y|X) = 18.6 + 0.127X$
- Standard error of $\hat{\beta}_x = 0.033$

Analysis with imputation from fitted distribution:

- Standard deviation of $Y = 32.3$
- $E(Y|X) = 19.2 + 0.121X$
- Standard error of $\hat{\beta}_x = 0.028$

(This analysis only imputed Y)



Single imputation from the fitted distribution (a new random draw)

Replace missing values with their conditional expectation plus some residual

- The residual can be estimated from the non-missing data

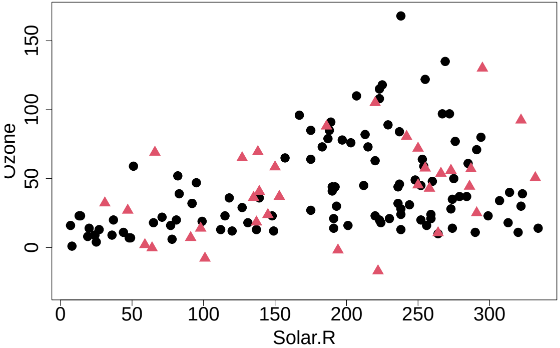
Complete case analysis:

- Standard deviation of $Y = 33.0$
- $E(Y|X) = 18.6 + 0.127X$
- Standard error of $\hat{\beta}_x = 0.033$

Analysis with imputation from fitted distribution:

- Standard deviation of $Y = 34.0$
- $E(Y|X) = 17.9 + 0.129X$
- Standard error of $\hat{\beta}_x = 0.030$

(This analysis only imputed Y)



Single imputation from the fitted distribution

- Not too bad
- Shape of distribution is generally similar
  - Although need to be sure imputations are within range of data
- Variation is about right
- Unbiased estimation if data are MAR and the fitted distribution is properly specified
  - In this example, the fitted distribution is properly specified if the conditional mean model is correct and the residuals are normally distributed with constant variance
  - Not exactly correct in our analysis, but not a bad approximation
- Problem is that we will get different estimates depending on the chosen sample
- Maybe we could repeat this multiple times and take averages?
  - This is the basic idea behind multiple imputation (lectures 3 and 4)

Hot deck imputation

Replace missing values with the observed value from a similar unit

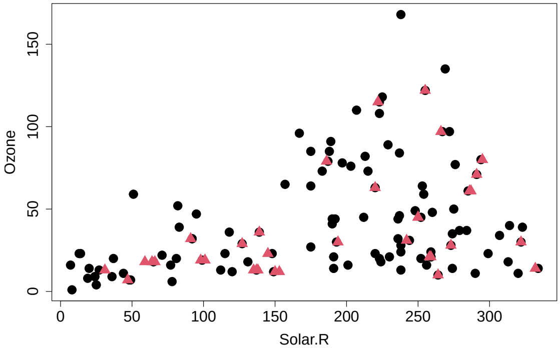
Complete case analysis:

- Standard deviation of $Y = 33.0$
- $E(Y|X) = 18.6 + 0.127X$
- Standard error of $\hat{\beta}_x = 0.033$

Analysis with hot deck imputation:

- Standard deviation of $Y = 32.6$
- $E(Y|X) = 15.9 + 0.135X$
- Standard error of $\hat{\beta}_x = 0.028$

(This analysis only imputed Y)



Hot deck imputation

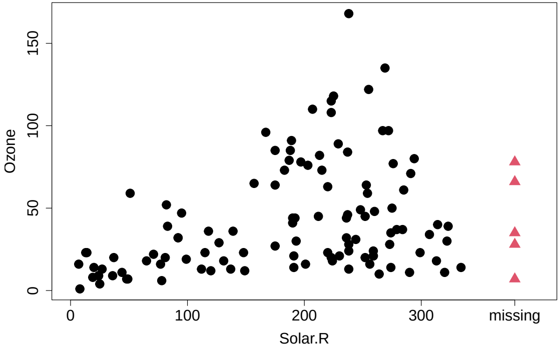
- Not too bad
- Shape of distribution is generally similar
- Variation appears about right
  - Although unclear how to incorporate uncertainty into inference
- Avoids parametric assumptions
- Requires good donor matches
- Unbiased estimation if data are MCAR
  - Unclear for MAR
- Interesting idea that is popular among survey statisticians, but less well understood than some other approaches

Missing Data Indicators

In regression model, include R_x.

- Include interaction between R_x and X.
- Include missing data indicator, $(1 - R_x)$.
- This only accounts for missing X.

```
##
## Call:
## lm(formula = y5 ~ I(x5 * r) + I(1 - r))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.292 -21.113  -8.332  18.062 119.136
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.59873    6.72998   2.764  0.00668 **
## I(x5 * r)    0.12717    0.03269   3.890  0.00017 ***
## I(1 - r)    24.20127   15.51199   1.560  0.12152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.25 on 113 degrees of freedom
## Multiple R-squared:  0.1181, Adjusted R-squared:  0.1025
## F-statistic: 7.568 on 2 and 113 DF,  p-value: 0.0008234
```



Missing Data Indicators

- Simple and retains all those with non-missing outcomes
  - Addresses missing covariates
- Cannot be used for missing outcomes
- Changes interpretation
  - Slope among those with complete data
  - Change in mean among those with missing X
  - Not clear that these interpretations are of interest
- Estimation can be biased even with MCAR data
- Generally not recommended except in special cases

Simulations Exercise 1B

Using the 4 simulated datasets that you generated earlier today, obtain regression coefficient estimates for the model $E(Y|X) = \beta_0 + \beta_1 X + \beta_2 V$ using the following methods:

1. Mean imputation
2. Conditional expectation imputation
3. Single imputation from the fitted distribution
4. Missing indicator approach
5. Bonus: Hot deck imputation

Compare and contrast estimates with estimates based on the original data and based on complete case analyses.

Summary

Today we have

- Defined missing data mechanisms
  - MCAR, MAR, NMAR
- Discussed simple (typically naive) analysis approaches
  - Complete case analysis
  - Single imputation
    - Mean imputation, conditional expectation imputation, single imputation from fitted distribution, hot deck imputation
  - Missing data indicators