Chapter 12: Missing Data

Gunwoong Park

Lecture Note

University of Seoul

Outline

- Type of Missing Data
- Remedy

Reason for Missing Data

Some values of some cases are missing. Why?

- The reason may be non-informative (equivalent to random deletions)
- The reason may be linked to the values of predictors and/or response (e.g., people may drop out of a drug trial if they feel the treatment is not working)
- The amount of bias introduced by missing values depends on the reason

In other words,

- Fail to observe a complete case (x_i, y_i) at random.
- Incomplete cases because of a special reason; censored data.

Type of Missing Data

- Missing Completely at Random (MCAR): The probability that a value is missing is the same for all cases.
- Missing at Random (MAR): The probability of a value being missing depends on a known mechanism. For example, in surveys, certain groups are less likely to provide information than others.
- Missing not at Random (MNAR): The probability that a value is missing depends on some unobserved variable or, more seriously, on what value would have been observed. For example, people who have something to hide are typically less likely to provide information that might reveal something embarrassing or illegal.

Problems:

- It is very hard to determine the type of missing data.
- No diagnostic methods

Some Solutions

Work only when the reason is non-informative:

1. Delete the case with missing values – easiest
2. Impute the missing values – two options
   - ▷ Fill in each missing value with the mean of that predictor
   - ▷ Regress each predictor on other predictors to fill in missing values
3. EM algorithm – model missing values as parameters (complicated)

Remedy: Deletion

- Simple and unbiased method
- A lot of samples are necessary

Chicago Insurance Example

Insurance redlining: practice of refusing to issue insurance to certain types of people or within certain geographical areas.

Question of interest: Which variables influence denial of insurance? E.g., using fire rates is fine, but using race is illegal.

- Data: Chicago, 1977–1978, $n = 47$, $p = 6$.
- FAIR: offered as a default policy to homeowners who were rejected by the voluntary market.
- Do not have information about individuals. All variables are measured at zip code level.

Example: Chicago insurance

```
> data(chmiss, package="faraway")
> head(chmiss)
      race fire theft  age involact income
60626 10.0  6.2    29 60.4       NA  11.74
60640 22.2  9.5    44 76.5      0.1   9.32
60613 19.6 10.5    36   NA      1.2   9.95
60657 17.3  7.7    37   NA      0.5  10.66
60614 24.5  8.6    53 81.4      0.7   9.73
60610 54.0 34.1    68 52.6      0.3   8.23
```

Chicago Insurance Data: Variables

- Response: `involact` – new FAIR plan policies and renewals per 100 housing units
- `race`: minority percentage
- `fire`: fires per 100 housing units
- `theft`: theft per 1000 population
- `age`: percent of housing units build before 1939
- `income`: median family income in 1000 dollars
- `side`: North or South side of Chicago – won't use here

Chicago insurance with missing data

```
> data(chmiss)
> chmiss
race fire theft  age involact income
60626 10.0  6.2    29 60.4       NA 11.744
60640 22.2  9.5    44 76.5      0.1  9.323
60613 19.6 10.5    36   NA      1.2  9.948
- - - - - -
> dim(chmiss)
[1] 47  6
##Number of missing values
> sum(is.na(chmiss))
[1] 20
```

Example: Chicago insurance

```
> summary(chmiss)
race            fire           theft
Min.   : 1.0   Min.   : 2.0   Min.   :  3.0
1st Qu.: 3.8   1st Qu.: 5.6   1st Qu.: 22.0
Median :24.5   Median : 9.5   Median : 29.0
Mean   :35.6   Mean   :11.4   Mean   : 32.7
3rd Qu.:57.6   3rd Qu.:15.1   3rd Qu.: 38.0
Max.   :99.7   Max.   :36.2   Max.   :147.0
NA's   :4      NA's   :2      NA's   :4

age           involact         income
Min.   : 2.0   Min.   :0.000   Min.   : 5.58
1st Qu.:48.3   1st Qu.:0.000   1st Qu.: 8.56
Median :64.4   Median :0.500   Median :10.69
Mean   :60.0   Mean   :0.648   Mean   :10.74
3rd Qu.:78.2   3rd Qu.:0.925   3rd Qu.:12.10
Max.   :90.1   Max.   :2.200   Max.   :21.48
NA's   :5      NA's   :3       NA's   :2
```

Example: Chicago insurance

```
> rowSums(is.na(chmiss))
60626 60640 60613 60657 60614 60610 60611 60625
1     0     1     1     0     0     0     0
60618 60647 60622 60631 60646 60656 60630 60634
1     1     0     0     1     0     0     1
60641 60635 60639 60651 60644 60624 60612 60607
0     0     0     1     1     0     0     1
60623 60608 60616 60632 60609 60653 60615 60638
0     1     1     0     1     0     0     0
60629 60636 60621 60637 60652 60620 60619 60649
1     0     1     0     0     1     0     1
60617 60655 60643 60628 60627 60633 60645
1     0     0     1     0     0     1
```

Example: Chicago insurance

```
> modmiss = lm(involact ~., chmiss)
> sumary(modmiss)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.11648    0.60576   -1.84  0.07947
race         0.01049    0.00313    3.35  0.00302
fire         0.04388    0.01032    4.25  0.00036
theft       -0.01722    0.00590   -2.92  0.00822
age          0.00938    0.00349    2.68  0.01390
income       0.06870    0.04216    1.63  0.11808

n = 27, p = 6, Residual SE = 0.3, R-Squared = 0.79
```

Example: Chicago insurance without missing values

```
> data(chredlin, package="faraway")
> modfull = lm(involact ~race + fire + theft + age
  + income, chredlin)
> sumary(modfull)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.60898    0.49526   -1.23  0.22585
race         0.00913    0.00232    3.94  0.00031
fire         0.03882    0.00844    4.60    4e-05
theft       -0.01030    0.00285   -3.61  0.00083
age          0.00827    0.00278    2.97  0.00491
income       0.02450    0.03170    0.77  0.44398

n = 47, p = 6, Residual SE = 0.3, R-Squared = 0.75
```

Remedy: Deletion

- Small sample size $n = 27$.
- Large standard of error
- May lose some information

Remedy: Single Imputation

- Average
- Regression

Example: Chicago insurance

```
## Average
> cmeans = colMeans(chmiss, na.rm = T)
> cmeans
race     fire     theft      age  involact   income
35.609   11.424   32.651   59.969    0.648   10.736
> mchm = chmiss
> for(i in c(1:4, 6)) mchm[is.na(chmiss[,i]),i] = cmeans[i]
```

Example: Chicago insurance

```
> imod = lm(involact ~.,mchm)
> sumary(imod)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.07080    0.50945    0.14    0.890
race         0.00712    0.00271    2.63    0.012
fire         0.02874    0.00939    3.06    0.004
theft       -0.00306    0.00275   -1.11    0.272
age          0.00608    0.00321    1.90    0.066
income      -0.02709    0.03168   -0.86    0.398

n = 44, p = 6, Residual SE = 0.4, R-Squared = 0.68
```

Example: Chicago insurance

```
## Regression
> lmodr = lm(age ~ fire+theft+race+income,chmiss)
> chmiss[is.na(chmiss$age),]
      race fire theft age involact income
60613 19.6 10.5    36  NA      1.2   9.95
60657 17.3  7.7    37  NA      0.5  10.66
60644 59.8 16.5    40  NA      0.8   9.78
60620 71.2 11.9    46  NA      0.9  11.04
60645  3.1  4.9    27  NA      0.0  13.73

> predict(lmodr,chmiss[is.na(chmiss$age),])
60613 60657 60644 60620 60645
74.1  71.8  64.3  59.1  45.8
```

Summary

- Missing data may invalidate all analysis if the values are missing not at random
- If the values are missing at random, deleting these observations will only increase variance (but not bias)
- Imputation introduces bias which may or may not be offset by reduction in variance
- Regression imputation works better on highly correlated predictors