Part 1. R 프로그래밍 (데이터 분석 전문가 양성과정)

07

데이터 전처리 (1)

결측치와 이상치

경북대학교 배준현 교수
(joonion@knu.ac.kr)

07. 데이터 전처리 (1): 결측치와 이상치

- **데이터 전처리**: *Data Preprocessing*
  - **본격적인 통계분석을 시작하기 전에 필요한 데이터 정제 작업**
    - **결측치와 이상치**: *missing values* and *outliers*
    - **데이터의 변환**: *integration*, *filtering*, *sampling*, and so on.
    - **데이터의 표준화**: *standardization*

07. 데이터 전처리 (1): 결측치와 이상치

- **결측치**: missing values
  - 설문 결과, 실험 결과 등의 연구 데이터에서 누락된 관측값이 존재할 경우
    - 결측치가 포함된 관측값을 연구 데이터에서 제거하거나
    - 결측치를 적절한 다른 값으로 대체해야 함
  - NA: R에서 결측치를 나타내는 값

07. 데이터 전처리 (1): 결측치와 이상치

- **결측치가 포함된 벡터의 통계값 구하기**

```
> x <- c(45, NA, 87, 63, 55, NA, 72, 61, 59, 68)

> mean(x)
[1] NA
> mean(x, na.rm = T)
[1] 63.75

> var(x, na.rm = T)
[1] 155.0714
> sd(x, na.rm = T)
[1] 155.0714
```

07. 데이터 전처리 (1): 결측치와 이상치

- **벡터에 포함된 결측치를 다른 값으로 대체하기**

```
> x <- c(45, NA, 87, 63, 55, NA, 72, 61, 59, 68)
> is.na(x)
 [1] FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE

> x[is.na(x)]
[1] NA NA
> x[!is.na(x)]
[1] 45 87 63 55 72 61 59 68

> x[is.na(x)] <- mean(x, na.rm = T)
> x
 [1] 45.00 63.75 87.00 63.00 55.00 63.75 72.00 61.00 59.00 68.00
```

```
> ?airquality
> str(airquality)
'data.frame':  153 obs. of  6 variables:
 $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
 $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
 $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
 $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
 $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
 $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...
```

07. 데이터 전처리 (1): 결측치와 이상치

- complete.cases() 함수: 데이터 프레임에서 결측치가 포함된 관측값(행) 확인

```
> df <- airquality
> complete.cases(df)
  [1]  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE FALSE
 [11] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
.....(이하 생략)

> df[complete.cases(df), ]
    Ozone Solar.R Wind Temp Month Day
1      41     190  7.4   67     5   1
2      36     118  8.0   72     5   2
3      12     149 12.6   74     5   3
.....(이하 생략)

> df[!complete.cases(df), ]
 Ozone Solar.R Wind Temp Month Day
5      NA      NA 14.3   56     5   5
6      28      NA 14.9   66     5   6
10     NA     194  8.6   69     5  10
.....(이하 생략)
```

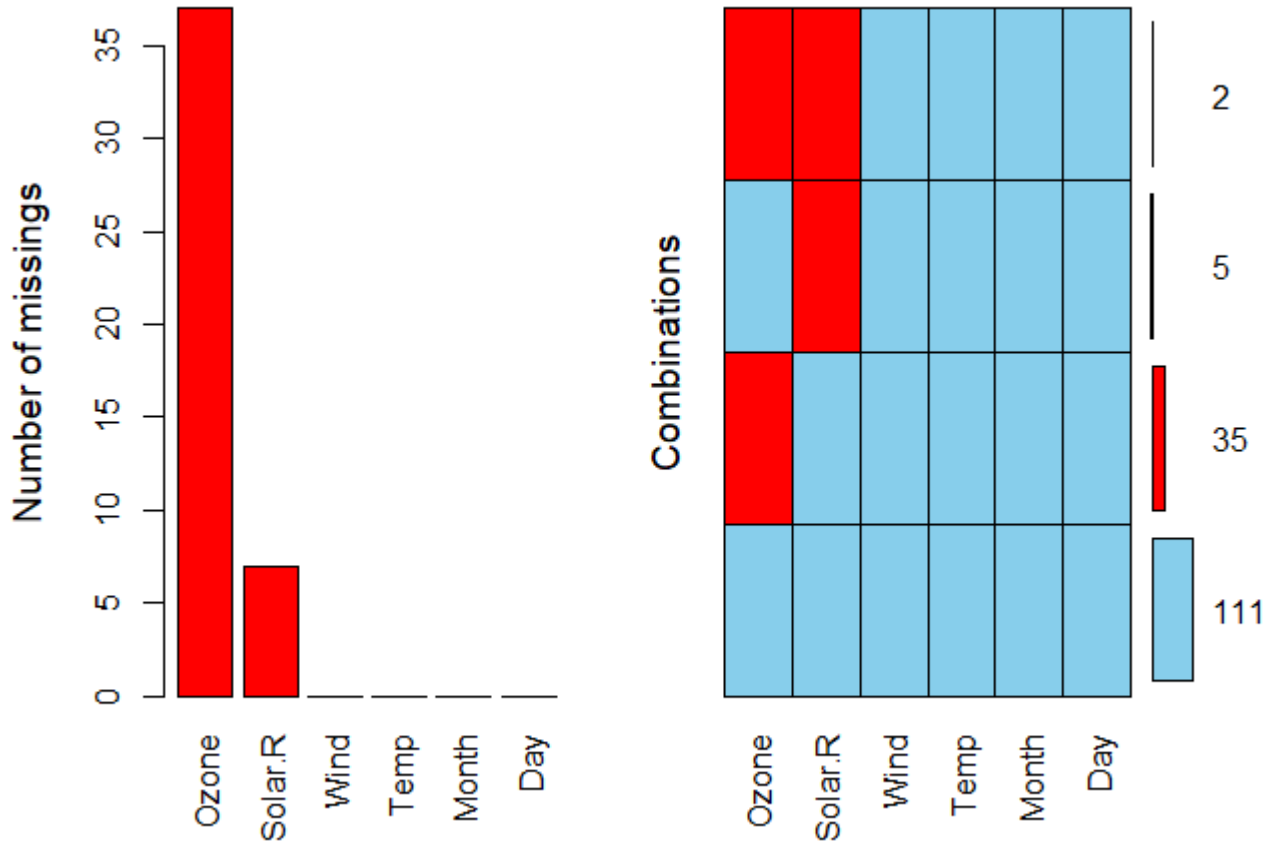
- 결측치에 관련된 정보 확인: 결측치의 개수와 비율

```
> sum(is.na(df$Ozone))
[1] 37
> sum(is.na(df$Solar.R))
[1] 7
> sum(is.na(df$Solar.R) & is.na(df$Ozone))
[1] 2

> sum(!complete.cases(df))
[1] 42
> mean(!complete.cases(df))
[1] 0.2745098
```

07. 데이터 전처리 (1): 결측치와 이상치

- VIM 패키지의 aggr() 함수: 변수별로 결측치의 분포와 발생 패턴을 시각화

```
> library(VIM)
> ?aggr
> aggr(airquality, prop = F, numbers = T, sortVar = T)
```



07. 데이터 전처리 (1): 결측치와 이상치

- na.omit() 함수: 데이터 프레임에서 결측치를 제거

```
> airquality[complete.cases(airquality), ]
> nrow(airquality[complete.cases(airquality), ])
[1] 111

> df <- na.omit(airquality)
> str(df)
'data.frame':  111 obs. of  6 variables:
 $ Ozone  : int  41 36 12 18 23 19 8 16 11 14 ...
 $ Solar.R: int  190 118 149 313 299 99 19 256 290 274 ...
 $ Wind   : num  7.4 8 12.6 11.5 8.6 13.8 20.1 9.7 9.2 10.9 ...
 $ Temp   : int  67 72 74 62 65 59 61 69 66 68 ...
 $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
 $ Day    : int  1 2 3 4 7 8 9 12 13 14 ...
 - attr(*, "na.action")= 'omit' Named int [1:42] 5 6 10 11 25 26 27 32 33 34 ...
  ..- attr(*, "names")= chr [1:42] "5" "6" "10" "11" ...
```

07. 데이터 전처리 (1): 결측치와 이상치

- mice 패키지의 mice() 함수: 결측치를 여러 가지 통계적 방법으로 대체(*imputation*)

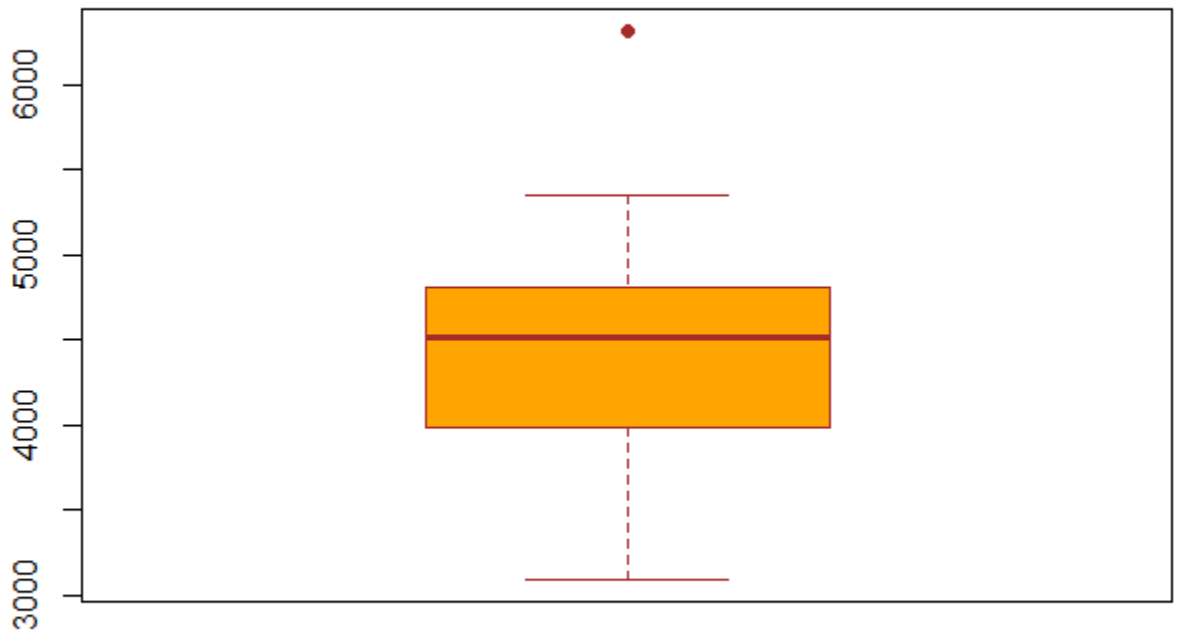
```
> result <- mice(airquality, method="mean", m = 2, maxit = 2)
 iter imp variable
  1   1  Ozone  Solar.R
  1   2  Ozone  Solar.R
  2   1  Ozone  Solar.R
  2   2  Ozone  Solar.R

> result$imp$Ozone
           1        2
5   42.12931 42.12931
10  42.12931 42.12931
.....(이하 생략)

> result$imp$Solar.R
          1        2
5  185.9315 185.9315
6  185.9315 185.9315
.....(이하 생략)
```

- **이상치:** *outliers* or *anomalies*
  - 특이값: 정상적인 데이터의 분포 범위 밖에 위치하는 관측값
    - 입력 오류에 의해 발생한 이상치:
      - 키의 데이터에서 단위가 다른 경우 등
    - 실제로 특이한 값을 가진 이상치:
      - 부모의 월 소득 정보에서 재벌 2세가 포함된 경우 등

07. 데이터 전처리 (1): 결측치와 이상치

- boxplot() 함수: 데이터셋에 이상치가 존재하는 지를 시각화

```
> df <- data.frame(state.x77)
> boxplot(df$Income, pch = 19, col = "orange", border = "brown")
```



- boxplot.stats() 함수를 이용한 이상치에 대한 상세 확인

```
> boxplot.stats(df$Income)
$stats
[1] 3098 3983 4519 4815 5348

$n
[1] 50

$conf
[1] 4333.093 4704.907

$out
[1] 6315

> outlier <- boxplot.stats(df$Income)
> df[df$Income == outlier$out, ]
       Population Income Illiteracy Life.Exp Murder HS.Grad Frost   Area
Alaska        365   6315        1.5    69.31   11.3    66.7   152 566432
```

07. 데이터 전처리 (1): 결측치와 이상치

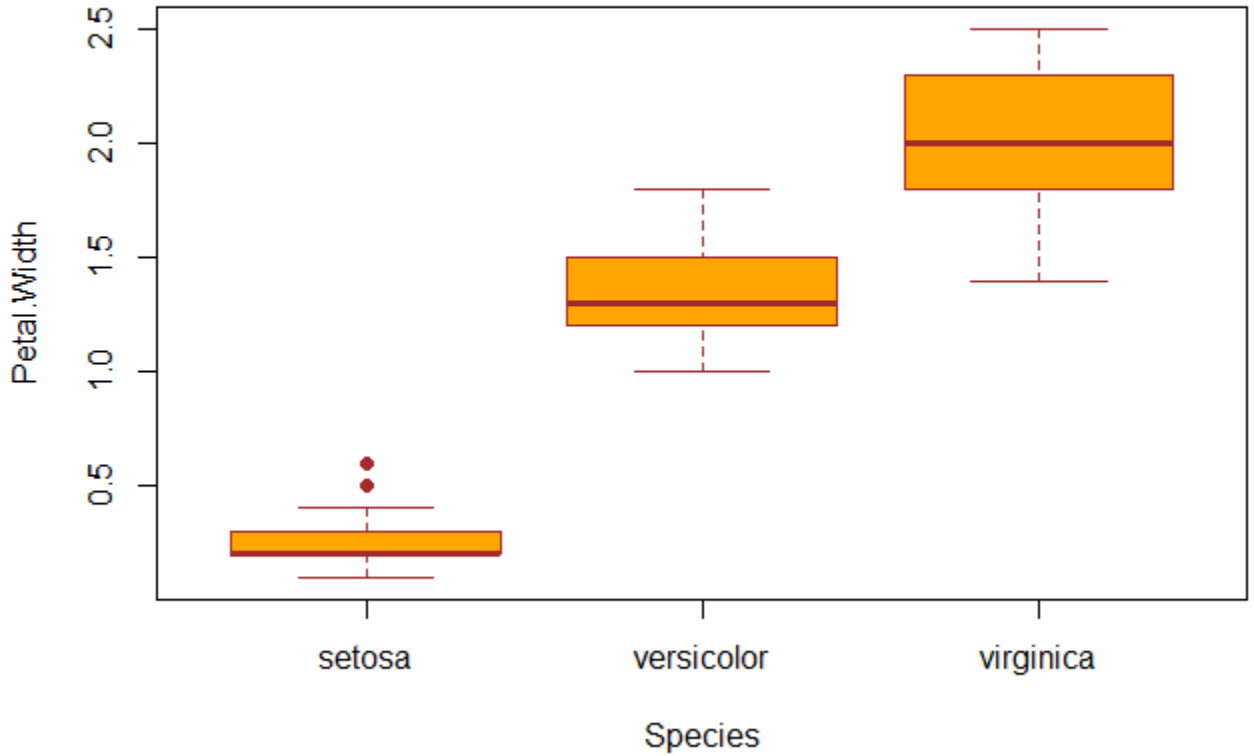
- 이상치가 통계량을 왜곡할 때는 결측치로 변환하여 통계분석 대상에서 제외

```
> df <- data.frame(state.x77)
> nrow(df)
[1] 50

> df[df$Income == outlier$out, ] <- NA
> df[!complete.cases(df), ]
Population Income Illiteracy Life.Exp Murder HS.Grad Frost Area
Alaska         NA     NA          NA       NA     NA       NA    NA   NA

> df.no.outlier <- na.omit(df)
> nrow(df.no.outlier)
[1] 49
```

07. 데이터 전처리 (1): 결측치와 이상치

- 범주로 구분할 수 있는 데이터는 범주별로 이상치를 확인할 수 있음

```
> boxplot(Petal.Width ~ Species, data = iris,
          pch = 19, col = "orange", border = "brown")
```

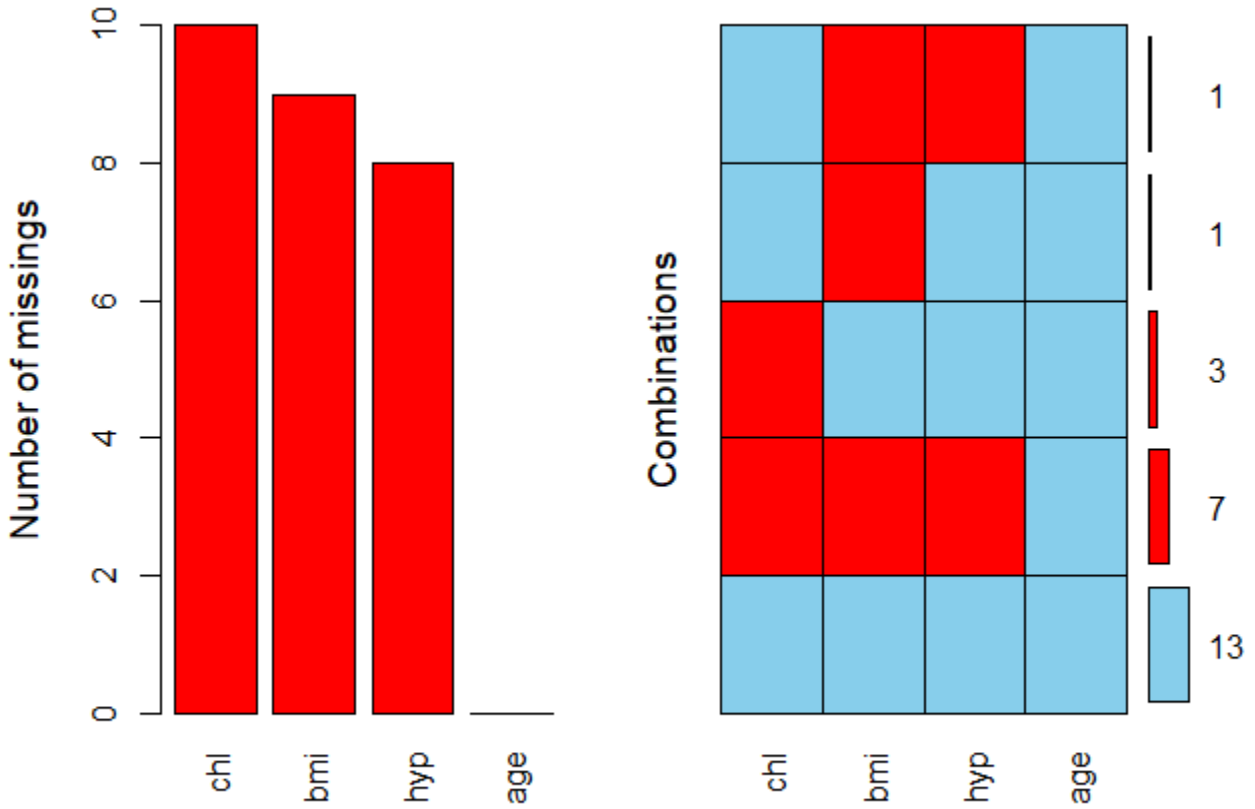


- 이상치가 여러 개인 경우에는 %in% 연산자를 활용하여 결측치를 제거

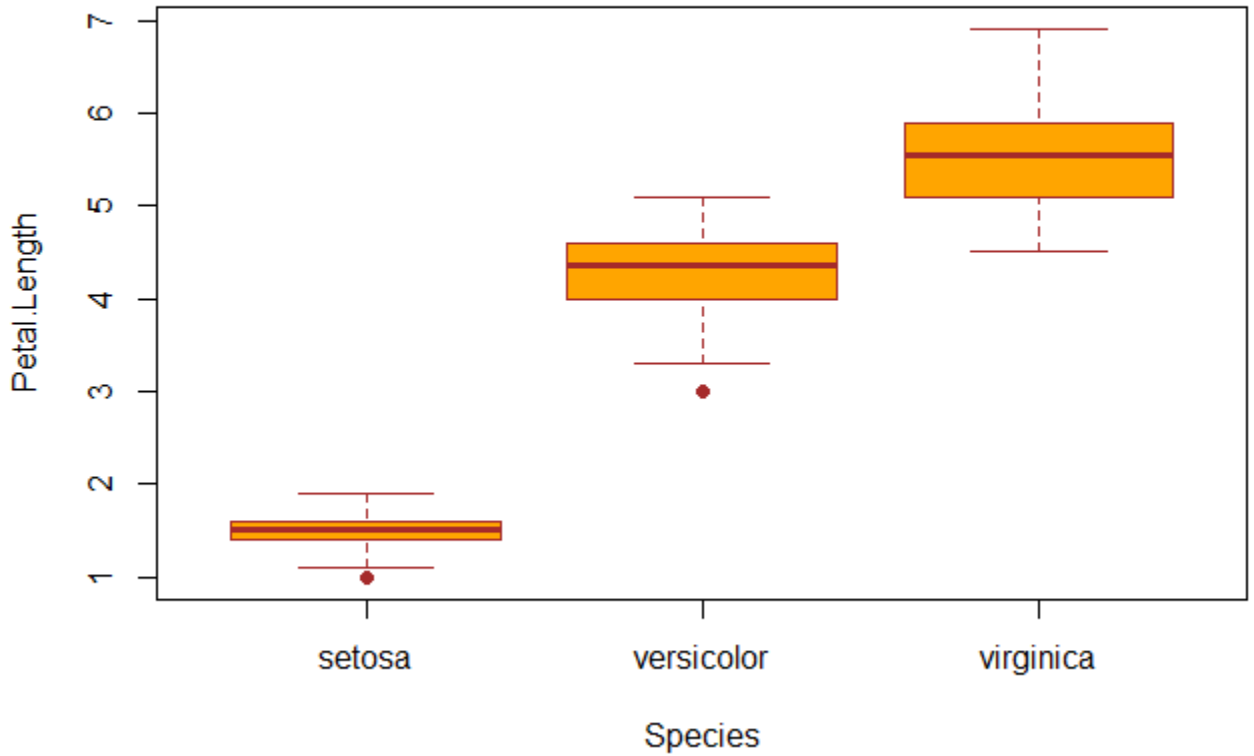
```
> df <- with(iris, iris[Species == "setosa", ])
> boxplot.stats(df$Petal.Width)
$stats
[1] 0.1 0.2 0.2 0.3 0.4

$n
[1] 50

$conf
[1] 0.1776554 0.2223446

$out
[1] 0.5 0.6
> outlier <- boxplot.stats(df$Petal.Width)$out
> df[df$Petal.Width %in% outlier, ] <- NA
> df.no.outlier <- na.omit(df)
> nrow(df.no.outlier)
[1] 48
```

- **연습문제 7.1:**
  - mice 패키지의 nhanes 데이터셋을 로드하고, 다음 R 코드를 작성하시오.
    - 변수와 관측값의 개수는 각각 얼마인가?
    - NA가 포함되지 않은 관측값들을 모두 출력하시오.
    - NA가 포함된 관측값들을 모두 출력하시오.
    - NA가 포함된 관측값들의 개수는 몇 개인가?
    - 각각의 변수별로 NA의 개수는 각각 몇 개인가?
    - VIM 패키지의 aggr() 함수로 결측치의 분포를 확인하시오.

07. 데이터 전처리 (1): 결측치와 이상치



- **연습문제 7.2:**
  - iris 데이터셋에 대해서, 다음 R 코드를 작성하시오.
    - Petal.Length에 대해서 박스플롯을 그려보시오.



- setosa 품종에서 Petal.Length의 이상치를 out.set 변수에 저장하시오.
- versicolor 품종에서 Petal.Length의 이상치를 out.ver 변수에 저장하시오.
- iris 데이터셋을 df 라는 이름의 변수에 따로 저장하시오.
- df 데이터셋에서 out.set, out.ver에 저장된 이상치를 가진 관측값에 대해
  - 관측값을 NA로 변경한 후에
  - NA값을 가진 관측값을 제거하시오.

Any Questions?