第 2 章: 因果関係 (2.2. R でデータを部分集合化する)

今井耕介 著

『社会科学のためのデータ分析入門 (QSS)』

2026-03-09

2.2 R でデータを部分集合化する

データの抽出（部分集合化）の目的

- 大規模なデータセット全体を分析するだけでなく、特定の条件を満たす「部分集合」に注目したい場合が多い。
  - 例：男性だけのデータ、特定の州のデータ、ある年齢層のデータなど。
- R では、**論理演算子**と**インデックス（`[]` や `subset()`）**を組み合わせて、柔軟にデータを抽出できる。

2.2.1 論理演算子 (Logical Operators)

基本的な論理演算子

- 2 つの値を比較し、TRUE（真）か FALSE（偽）を返す。

```
# 1. 等しい (==)
5 == 5
```

```
## [1] TRUE
```

```
5 == 3
```

```
## [1] FALSE
```

```
# 2. 等しくない (!=)
5 != 3
```

```
## [1] TRUE
```

```
# 3. 大なり、小なり (>, <, >=, <=)
5 > 3
```

```
## [1] TRUE
```

複数の条件の組み合わせ

- 「かつ（AND）」や「または（OR）」を使って条件を繋げる。

```
# 1. かつ (&): 両方の条件が真なら TRUE
(5 > 3) & (2 < 4)
```

```
## [1] TRUE
```

```
# 2. または (|): どちらか一方が真なら TRUE
(5 > 3) | (2 > 4)
```

```
## [1] TRUE
```

2.2.2 データの読み込みと論理演算

resume データの読み込み (1)

- 実験で収集されたデータを R に読み込みます。

```
# 1. ローカルに保存した resume データの読み込み（推奨）
resume <- read.csv("resume.csv")

#（参考）URL から直接読み込むことも可能
# resume <- read.csv("https://ayumu-tanaka.github.io/QSS/QSS_Data/resume.csv")

# 2. データの次元（行数と列数）を確認
dim(resume)
```

```
## [1] 4870    4
```

ベクトルの要素ごとに比較する (2)

- データの最初の数行を表示して確認し、論理演算を行います。

```
# 最初の 3 行を表示
head(resume, n = 3)
```

```
##   firstname    sex  race call
## 1   Allison female white    0
## 2   Kristen female white    0
## 3   Lakisha female black    0
```

```
# race 変数の最初の 5 つの要素が "black" かどうかを確認
head(resume$race == "black", n = 5)
```

```
## [1] FALSE FALSE  TRUE  TRUE FALSE
```

```
# コールバック (call) が 1 であった行の TRUE/FALSE
head(resume$call == 1, n = 5)
```

```
## [1] FALSE FALSE FALSE FALSE FALSE
```

2.2.3 データの抽出 (Indexing)

角括弧 [] を使った抽出 (1)

- データ [行の条件，列の指定] という形式でデータを抽出する。

```
# 1. 黒人 (race == "black") の行だけを抽出
# カンマの後は空にすると「全ての列」を意味する
resumeB <- resume[resume$race == "black", ]

# 2. 白人 (race == "white") の行だけを抽出
resumeW <- resume[resume$race == "white", ]

# 抽出したデータの行数を確認
nrow(resumeB)
```

```
## [1] 2435
```

```
nrow(resumeW)
```

```
## [1] 2435
```

角括弧 [] を使った抽出 (2)

- 抽出したサブセットに対して、統計量を計算する。

```
# 黒人のコールバック率（平均）
mean(resumeB$call)
```

```
## [1] 0.06447639
```

```
# 白人のコールバック率（平均）
mean(resumeW$call)
```

```
## [1] 0.09650924
```

2.2.4 subset() 関数の活用

subset() による直感的な抽出

- subset(データ, subset = 条件) という書き方で、より読みやすく抽出できる。

```
# 黒人かつ男性 (sex == "male") のデータを抽出
resumeBm <- subset(resume, subset = (race == "black" & sex == "male"))

# データの最初の数行を表示
head(resumeBm, n = 3)
```

```
##    firstname  sex  race call
## 10    Tyrone male black    0
## 21     Leroy male black    0
## 42    Tyrone male black    0
```

```
# 黒人男性のコールバック率
mean(resumeBm$call)
```

```
## [1] 0.0582878
```

2.2.5 まとめ

このセクションのまとめ

- **論理演算子**: `==`, `!=`, `>`, `<`, `&`, `|` を使って、データが条件を満たすかを判定する。
- **データの抽出**:
  - **角括弧 `[]`**: `resume[resume$race == "black", ]` のように記述。
  - **subset() 関数**: `subset(resume, race == "black")` のように、よりシンプルに記述可能。
- **因果推論への応用**: グループごとにデータを切り分けることで、処置群と対照群の平均を簡単に比較できる。