第 4 章: 予測
(4.3. 回帰分析と因果関係)

今井耕介 著

『社会科学のためのデータ分析入門 (QSS)』

2026-03-09

4.3 回帰分析と因果関係

予測と因果関係

- ここまで、回帰分析を「予測」のためのツールとして学んできた。
- しかし、社会科学では回帰分析を「**因果関係の推定**」に使うことが非常に多い。
- **ポイント**:
  1. RCT（無作為化比較試験）のデータを回帰分析にかけるとどうなるか？
  2. 複数の説明変数を用いた「重回帰分析」によって、交絡因子をどのように制御するか？
  3. 回帰不連続デザイン（RDD）による因果推論。

4.3.1 ランダム化比較試験と回帰分析

women データの読み込み (1)

- インドの村議会における女性のクオータ制（割当制）のデータを準備します（women.csv）。

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

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

RCT データへの回帰の適用 (2)

- 処置変数 reserved が 1 ならクオータあり、0 ならなし。

```
# 1. 飲用水設備 (water) の平均の差（因果効果）を手計算
mean(women$water[women$reserved == 1]) - mean(women$water[women$reserved == 0])
```

```
## [1] 9.252423
```

```
# 2. 単回帰分析による推定
# water を結果変数、reserved を説明変数とする
lm(water ~ reserved, data = women)
```

```
##
## Call:
## lm(formula = water ~ reserved, data = women)
##
## Coefficients:
## (Intercept)     reserved
##      14.738        9.252
```

- **結論**: RCT において、処置変数（0 or 1）を用いた単回帰分析の「傾き（係数）」は、**グループ間の平均の差（平均因果効果: SATE）と完全に一致する。**

4.3.2 複数の予測変数を用いた回帰分析

social データの読み込み (1)

- 複数の要因をコントロールするため、社会的圧力実験のデータを準備します（social.csv）。

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

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

重回帰分析 (Multiple Regression) (2)

- 第 2 章の社会的圧力実験 (social.csv) を使用。

```
# メッセージの種類（messages）による投票率への影響
# messages はカテゴリ変数のため、R が自動的にダミー変数に変換して計算する
fit <- lm(primary2006 ~ messages, data = social)
coef(fit)
```

```
##       (Intercept)   messagesControl messagesHawthorne messagesNeighbors
##       0.314537652      -0.017899344       0.007836968       0.063410569
```

- **解釈**: messagesNeighbors の係数 (約 0.081) は、ベースライン（Civic Duty）と比較して、Neighbors メッセージが投票率を約 8.1%上げることを意味する。

4.3.3 異質な処置効果

交差項 (Interaction Terms) の導入

- 処置の効果が、個人の属性（例：年齢）によって異なるかどうかを調べます（**異質な処置効果**）。

```
# 1. データの前処理
# Control と Neighbors グループだけを抽出
social.neighbor <- subset(social, messages %in% c("Control", "Neighbors"))
# 年齢を計算
social.neighbor$age <- 2006 - social.neighbor$yearofbirth

# 2. 交差項を含む回帰モデルの推定 (変数 A * 変数 B で交差項を指定)
fit.age <- lm(primary2006 ~ age * messages, data = social.neighbor)
coef(fit.age)
```

```
##             (Intercept)                      age     messagesNeighbors
##           0.0974732574           0.0039982107           0.0498294321
## age:messagesNeighbors
##           0.0006283079
```

- age:messagesNeighbors の係数が負であるため、年齢が上がるほど「近所に知らせる」メッセージの効果は小さくなることがわかる。

4.3.4 回帰不連続デザイン (RDD)

MPs データの読み込み (1)

- 当落の境界線付近を利用するため、イギリス下院議員のデータを準備します（MPs.csv）。

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

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

議会データの抽出と回帰 (2)

```r
# 1. 労働党 (labour) のデータを抽出
MPs.labour <- subset(MPs, subset = (party == "labour"))

# 2. 境界線（margin = 0）の左側（落選）と右側（当選）で別々に回帰分析
labour.fit1 <- lm(ln.net ~ margin, data = MPs.labour[MPs.labour$margin < 0, ])
labour.fit2 <- lm(ln.net ~ margin, data = MPs.labour[MPs.labour$margin > 0, ])
```

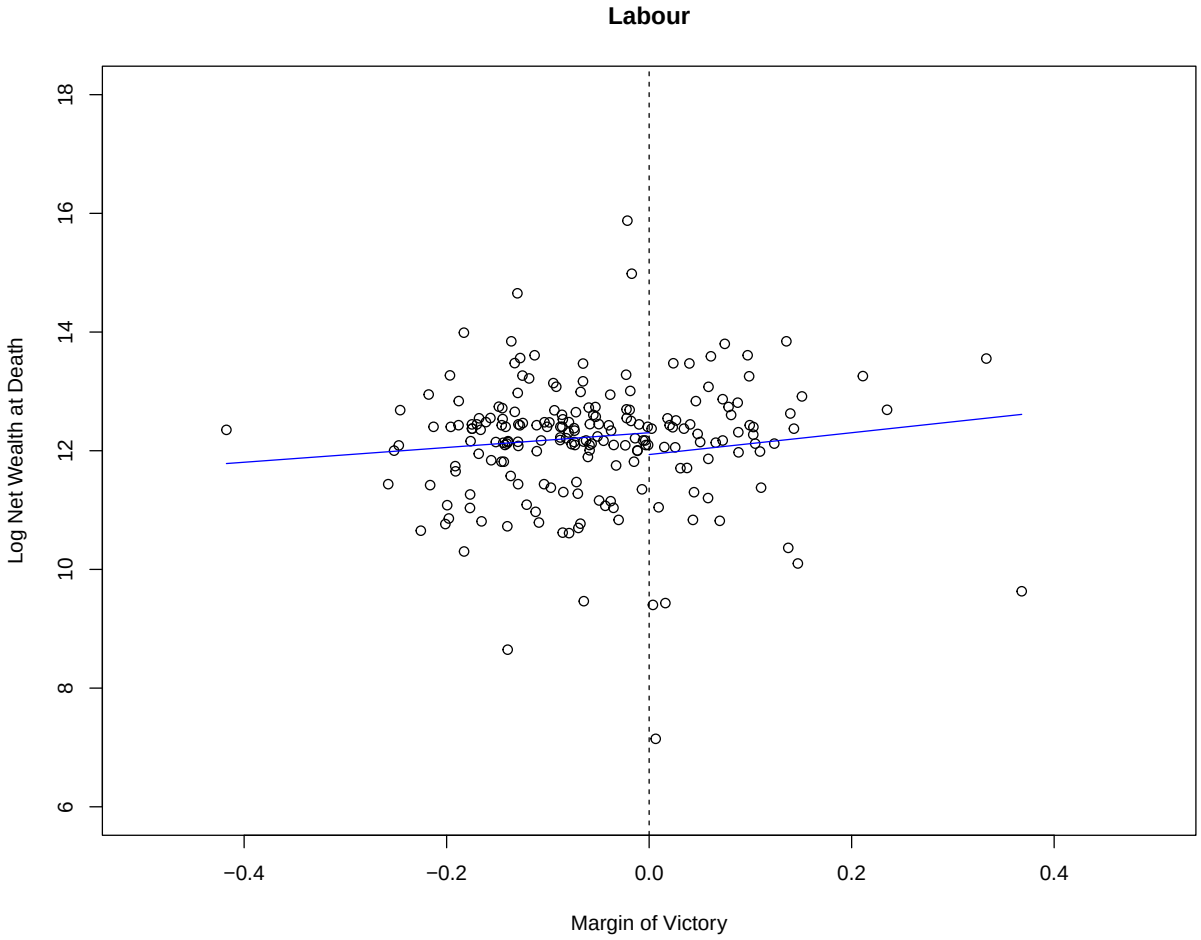
RDD の可視化: コード

```
# 1. 散布図のプロット
plot(MPs.labour$margin, MPs.labour$ln.net, main = "Labour",
     xlim = c(-0.5, 0.5), ylim = c(6, 18),
     xlab = "Margin of Victory", ylab = "Log Net Wealth at Death")

# 2. 境界線 (margin = 0) を追加
abline(v = 0, lty = "dashed")

# 3. 予測値の計算と回帰直線の追加
y1l.range <- c(min(MPs.labour$margin), 0) # 左側
y2l.range <- c(0, max(MPs.labour$margin)) # 右側
lines(y1l.range, predict(labour.fit1, newdata = data.frame(margin = y1l.range))
lines(y2l.range, predict(labour.fit2, newdata = data.frame(margin = y2l.range))
```

RDD の可視化: 描画結果



- **結論**: 境界線 (`margin = 0`) において、線に大きな「ジャンプ」が見られないため、労働党において議員になることが死後の資産を増やしたとは言えない[16]

4.3.5 まとめ

このセクションのまとめ

- **RCT と回帰**: ダミー変数（0, 1）を用いた回帰係数は、グループ間の平均の差（因果効果）と一致する。
- **重回帰と交絡**: 複数の変数をモデルに組み込むことで、観察研究でも交絡因子の影響を一定程度コントロールできる。
- **異質な処置効果**: 交差項 (* または :) を用いることで、グループごとの効果の違いを検証できる。
- **RDD**: 恣意的な境界線の前後を比較することで、実験ができない状況でも因果関係を推定する強力な手法。
- **R の操作**: `lm()` におけるカテゴリ変数や交差項の扱い。