第 4 章: 予測
(4.2. 線形回帰)

今井耕介 著

『社会科学のためのデータ分析入門 (QSS)』

2026-03-09

4.2 線形回帰 (Linear Regression)

線形回帰の基本概念

- **線形回帰**: 結果変数 Y と説明変数 X の関係を直線で近似する手法。
  - 数式: $\hat{Y} = \hat{\alpha} + \hat{\beta}X$
  - $\hat{Y}$: 予測値（Fitted value）
  - $\hat{\alpha}$: 切片（Intercept）
  - $\hat{\beta}$: 傾き（Slope / 係数）
- **最小二乗法 (OLS)**: 観測値 Y と予測値 $\hat{Y}$ の差（残差）の二乗和を最小にする $\hat{\alpha}$ と $\hat{\beta}$ を選ぶ。

4.2.1 外見と選挙結果

face データの読み込み (1)

- 候補者の顔の「有能さ」スコアと得票率のデータを準備します。

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

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

データの計算と確認 (2)

```
# 1. 民主党と共和党の得票率、およびその差（マージン）を計算
face$d.share <- face$d.votes / (face$d.votes + face$r.votes)
face$r.share <- face$r.votes / (face$d.votes + face$r.votes)
face$diff.share <- face$d.share - face$r.share

# 2. データの確認 (d.comp は有能さのスコア)
head(face[, c("d.comp", "diff.share")], n = 3)
```

```
##      d.comp diff.share
## 1 0.5645676 0.21012941
## 2 0.3419122 0.11943099
## 3 0.6123680 0.04990747
```

相関係数の計算

- 2 変数の線形な関係の強さを確認します。

```
# 有能さスコア（d.comp）と得票率の差（diff.share）の相関
cor(face$d.comp, face$diff.share)
```

```
## [1] 0.4327743
```

4.2.2 最小二乗法による推定

lm() 関数の使用: コード

- lm(結果変数 ~ 説明変数, data = データフレーム) という形式で回帰分析を行います。

```
# 1. 回帰モデルの推定
fit <- lm(diff.share ~ d.comp, data = face)

# 2. 推定された係数（切片と傾き）の表示
coef(fit)
```

```
## (Intercept)      d.comp
##  -0.3122259   0.6603815
```

```
# 3. 予測値（Fitted values）の最初の数件を確認
head(fitted(fit))
```

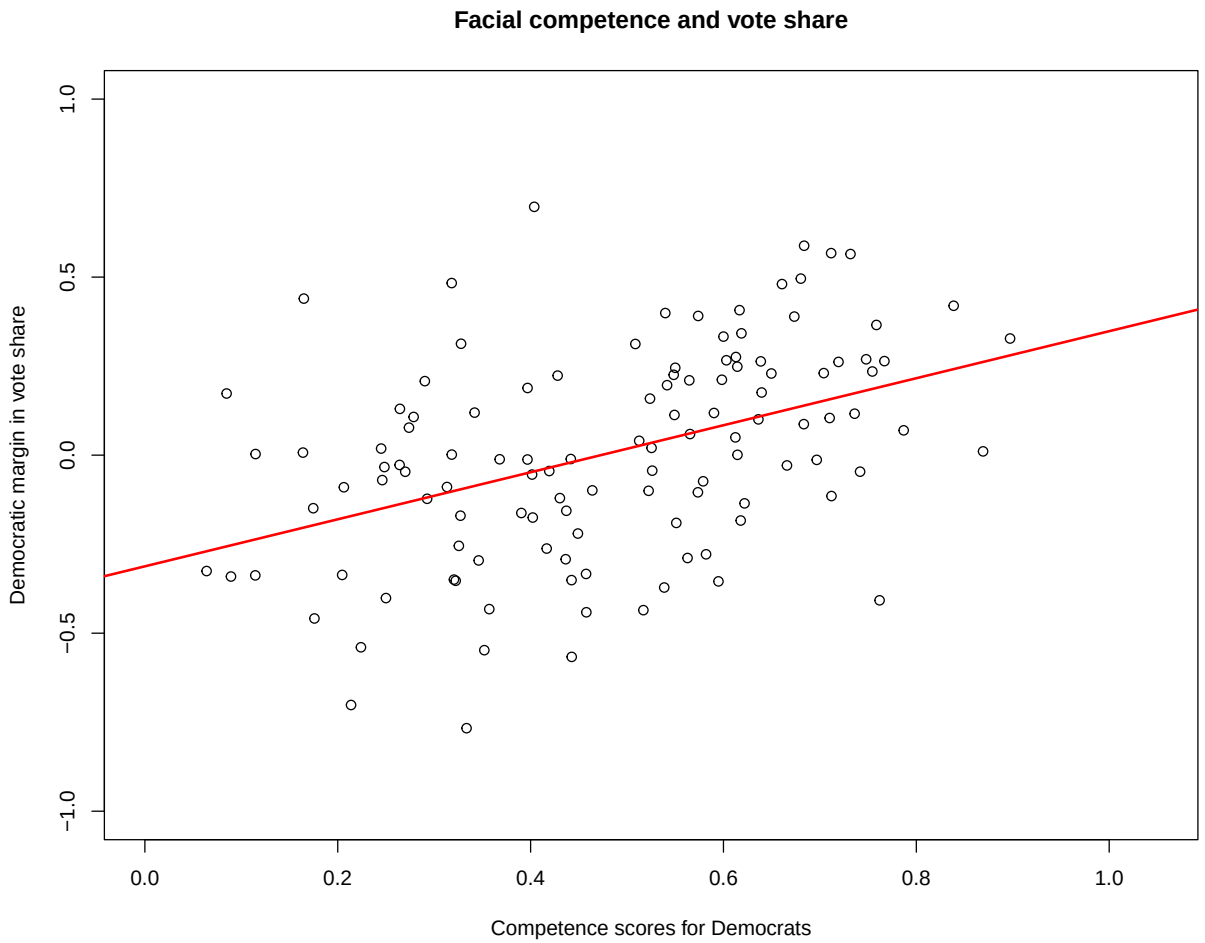
```
##           1           2           3           4           5           6
##  0.06060411 -0.08643340  0.09217061  0.04539236  0.13698690 -
0.10057206
```

回帰直線のプロット: コード

```
# 1. 散布図の作成
plot(face$d.comp, face$diff.share, xlim = c(0, 1.05), ylim = c(-1, 1),
     xlab = "Competence scores for Democrats",
     ylab = "Democratic margin in vote share",
     main = "Facial competence and vote share")

# 2. 推定された回帰直線を重ねる
abline(fit, col = "red", lwd = 2)
```

回帰直線のプロット: 描画結果



4.2.3 データの結合と平均への回帰

選挙データの読み込みと結合 (1)

- 異なる年のデータを州名略称 (state) をキーにして統合します。

```
# 1. ローカルに保存したデータの読み込み（推奨）
pres08 <- read.csv("pres08.csv")
pres12 <- read.csv("pres12.csv")

#（参考）URL から直接読み込むことも可能
# pres08 <- read.csv("https://ayumu-tanaka.github.io/QSS/QSS_Data/pres08.csv")
# pres12 <- read.csv("https://ayumu-tanaka.github.io/QSS/QSS_Data/pres12.csv")

# 2. データの結合 (merge)
pres <- merge(pres08, pres12, by = "state")
```

結合データの確認 (2)

```
# 結合後のデータを確認 (Obama.x は 2008 年、Obama.y は 2012 年)
head(pres[, c("state", "Obama.x", "Obama.y")], n = 3)
```

```
##   state Obama.x Obama.y
## 1    AK      38      41
## 2    AL      39      38
## 3    AR      39      37
```

4.2.4 モデルの適合度

florida データの読み込み (1)

- モデルの当てはまりを確認するため、フロリダ州のデータを使用します。

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

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

決定係数 (R^2) の計算 (2)

- R^2: モデルがデータの変動をどれだけ説明できているかを示す指標（0〜1）。

```
# 1. 回帰分析の実行
fit2 <- lm(Buchanan00 ~ Perot96, data = florida)

# 2. summary() 関数から決定係数を取り出す
summary(fit2)$r.squared
```

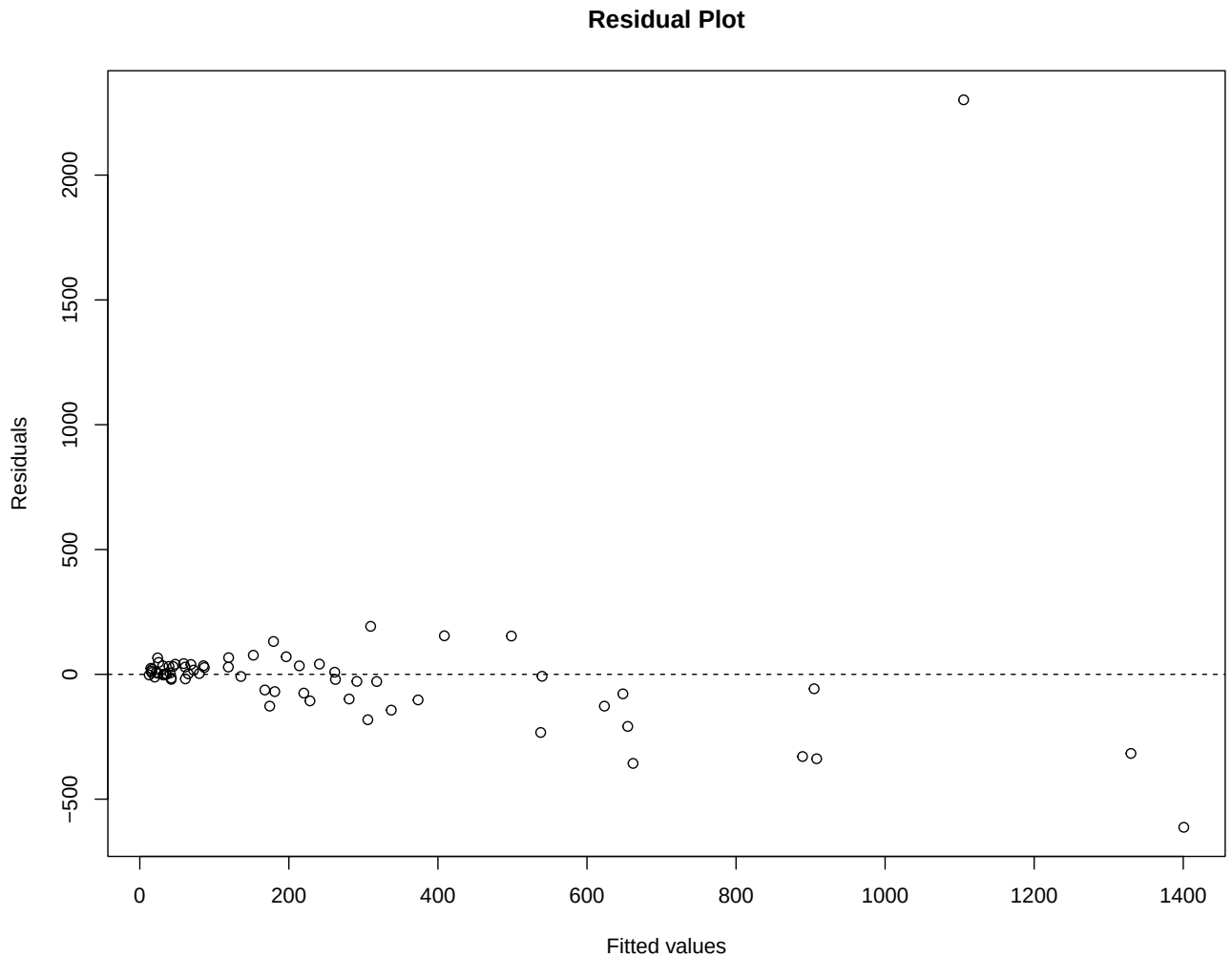
```
## [1] 0.5130333
```

残差プロット (Residual Plot): コード

- 予測が系統的に外れていないか、外れ値がないかを確認します。

```
# 横軸に予測値、縦軸に残差をプロット
plot(fitted(fit2), resid(fit2),
     xlab = "Fitted values", ylab = "Residuals",
     main = "Residual Plot")
# 残差 0 の水平線を追加
abline(h = 0, lty = "dashed")
```

残差プロット: 描画結果


Residual Plot
Residuals
2000
1500
1000
500
0
−500
0
200
400
600
800
1000
1200
1400
Fitted values

4.2.5 まとめ

このセクションのまとめ

- **線形回帰**: Y と X の関係を直線 $\hat{Y} = \hat{\alpha} + \hat{\beta}X$ でモデル化する。
- **最小二乗法**: 残差の二乗和を最小にすることで、最適な係数を推定する。
- **データの統合**: `merge()` を使って、複数のソースからのデータを適切に結合する。
- **適合度の評価**: 決定係数 (R^2) や残差プロットを用いて、モデルの良し悪しを判断する。
- **R の操作**: `lm()`, `coef()`, `fitted()`, `resid()`, `merge()`, `summary()$r.squared`。