第 4 章: 予測

Elena Llaudet & Kosuke Imai.
Data Analysis for Social Science: A Friendly and Practical Introduction.

2026-03-09

4.1 予測とは何か？

予測の重要性

- 過去のデータを用いて、未知の状況や未来の結果を推測する。
- 社会科学における例：
  - 選挙結果の予測
  - 経済指標（GDP 成長率など）の予測
  - 犯罪率や教育効果の予測
- 予測の精度を上げるために、**線形回帰** (Linear Regression) という手法を学ぶ。

4.2 散布図と線形関係

precincts データの読み込み (1)

- 2019 年ウクライナ大統領選挙の投票区レベルデータを読み込みます。

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

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

データの確認 (2)

```
# 最初の数行を表示
head(precincts)
```

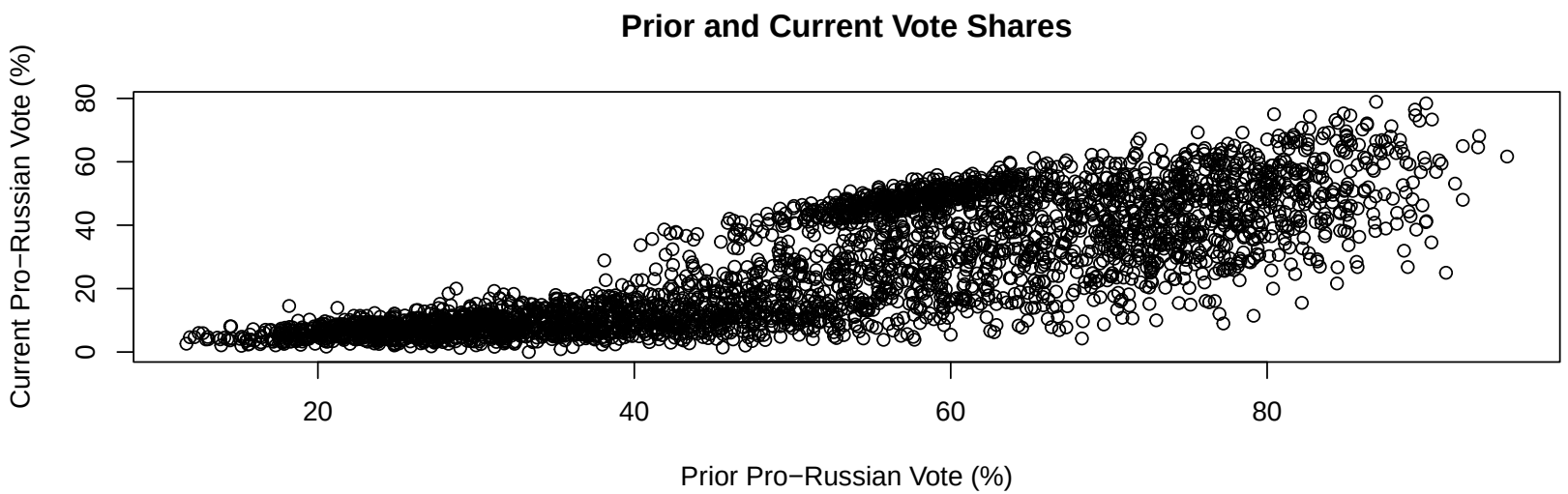
```
##   russian_tv pro_russian prior_pro_russian within_25km
## 1          0   2.7210884           25.14286           1
## 2          0   0.8928571           35.34483           0
## 3          1   1.6949153           20.53232           1
## 4          0  72.2689076           84.47761           1
## 5          0   1.2820513           28.99408           0
## 6          1   1.4285714           45.58824           0
```

散布図の作成

- 2 つの数値変数の関係を視覚化する。
- 例：過去の親ロシア派への投票率（`prior_pro_russian`）と、今回の親ロシア派への投票率（`pro_russian`）。

散布図のコード

```
# 散布図のプロット
plot(precincts$prior_pro_russian, precincts$pro_russian,
     xlab = "Prior Pro-Russian Vote (%)",
     ylab = "Current Pro-Russian Vote (%)",
     main = "Prior and Current Vote Shares")
```



4.3 線形回帰モデル

回帰直線

- 散布図の点の間を通る「最もよく当てはまる」直線を引く。
- 数式：$\hat{Y} = \hat{\alpha} + \hat{\beta}X$
  - $\hat{Y}$: 予測値
  - X: 説明変数（独立変数）
  - $\hat{\alpha}$: 切片 (Intercept)
  - $\hat{\beta}$: 傾き (Slope)

最小二乗法 (OLS)

- **最小二乗法** (Ordinary Least Squares: OLS)
- 観測値 Y_i と予測値 $\hat{Y}_i$ の差（**残差**, Residual）の二乗和を最小にする $\hat{\alpha}$ と $\hat{\beta}$ を求める。

4.4 Rでの回帰分析

lm() 関数の使用

- `lm(従属変数 ~ 説明変数, data = データフレーム)`

```
# 回帰モデルの推定
model <- lm(pro_russian ~ prior_pro_russian, data = precincts)
```

モデルの結果

```
summary(model)
```

```
##
## Call:
## lm(formula = pro_russian ~ prior_pro_russian, data = precincts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.202  -7.524  -0.107   7.119  25.247
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -14.321933   0.530968  -26.97   <2e-16 ***
## prior_pro_russian   0.796611   0.009651   82.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.17 on 3587 degrees of freedom
## Multiple R-squared:  0.6551, Adjusted R-squared:  0.655
## F-statistic:  6814 on 1 and 3587 DF,  p-value: < 2.2e-16
```

推定値の解釈

- `(Intercept)` ($\hat{\alpha}$): 説明変数が 0 のときの期待値（理論値）。
- `prior_pro_russian` ($\hat{\beta}$): 過去の投票率が 1%上がると、今回の投票率が何%変化するか。

4.5 予測値の算出

予測の実行

- 推定された係数を用いて、特定の過去の投票率（例：60%）のときの予測値を算出する。

```
# 係数の取り出し
coef(model)
```

```
##         (Intercept) prior_pro_russian
##          -14.3219329           0.7966111
```

```
# 過去の投票率が 60%の場合の予測値
# pro_russian = alpha + beta * 60
-14.32 + (0.797) * 60
```

```
## [1] 33.5
```

predict() 関数

```
# predict 関数を用いた予測
new_data <- data.frame(prior_pro_russian = 60)
predict(model, newdata = new_data)
```

```
##        1
## 33.47473
```

4.6 モデルの適合度

決定係数 (R-squared)

- R^2 は、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。
- 0 から 1 の間の値を取り、1 に近いほどモデルの当てはまりが良い。

```
# summary の結果から R-squared を確認
summary(model)$r.squared
```

```
## [1] 0.6551177
```

4.7 まとめ

第 4 章のまとめ

- 線形回帰を用いて、ある変数から別の変数を予測する方法を学んだ。
- 最小二乗法 (OLS) の仕組みを理解した。
- R の `lm()` 関数を用いて回帰分析を実行し、その結果（切片、傾き、決定係数）を解釈した。
- `predict()` 関数を用いて未知のデータの予測値を算出した。