3. 回帰分析の基礎

honocat

2025-12-28

```
HR <- read_csv('data/hr-data.csv')
glimpse(HR)
```

```
Rows: 8,803
Columns: 22
$ year       <dbl> 1996, 1996, 1996, 1996, 1996, 1996, 1996, 1996, 1996, 1996,~
$ ku         <chr> "aichi", "aichi", "aichi", "aichi", "aichi", "aichi", "aich~
$ kun        <dbl> 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3,~
$ status     <chr> "現 職", "元 職", "現 職", "新 人", "新 人", "新 人", "新 人", "現 職", "元
職", "新人",~
$ name       <chr> "KAWAMURA, TAKASHI", "IMAEDA, NORIO", "SATO, TAISUKE", "IWA~
$ party      <chr> "NFP", "LDP", "DPJ", "JCP", "others", "kokuminto", "indepen~
$ party_code <dbl> 8, 1, 3, 2, 100, 22, 99, 8, 1, 3, 2, 10, 100, 99, 22, 8, 1,~
$ previous   <dbl> 2, 3, 2, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 3, 1, 0, 0,~
$ wl         <chr> "当 選", "落 選", "落 選", "落 選", "落 選", "落 選", "落 選", "当 選", "落
選", "復活当選~
$ voteshare  <dbl> 40.0, 25.7, 20.1, 13.3, 0.4, 0.3, 0.2, 32.9, 26.4, 25.7, 12~
$ age        <dbl> 47, 72, 53, 43, 51, 51, 45, 51, 71, 30, 31, 44, 61, 47, 43,~
$ nocand     <dbl> 7, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8, 7, 7, 7, 7, 7,~
$ rank       <dbl> 1, 2, 3, 4, 5, 6, 7, 1, 2, 3, 4, 5, 6, 7, 8, 1, 2, 3, 4, 5,~
$ vote       <dbl> 66876, 42969, 33503, 22209, 616, 566, 312, 56101, 44938, 43~
$ eligible   <dbl> 346774, 346774, 346774, 346774, 346774, 346774, 346774, 338~
$ turnout    <dbl> 49.2, 49.2, 49.2, 49.2, 49.2, 49.2, 49.2, 51.8, 51.8, 51.8,~
$ exp        <dbl> 9828097, 9311555, 9231284, 2177203, NA, NA, NA, 12940178, 1~
$ expm       <dbl> 9.828097, 9.311555, 9.231284, 2.177203, NA, NA, NA, 12.9401~
$ vs         <dbl> 0.400, 0.257, 0.201, 0.133, 0.004, 0.003, 0.002, 0.329, 0.2~
$ exppv      <dbl> 28.341505, 26.851941, 26.620462, 6.278449, NA, NA, NA, 38.2~
$ smd        <chr> "当 選", "落 選", "落 選", "落 選", "落 選", "落 選", "落 選", "当 選", "落
選", "落選",~
$ party_jpn  <chr> "新進党", "自民党", "民主党", "共産党", "その他", "国民党", "無所属", "
```

新進党", "自民~

衆議院議員経験があることを表すダミー変数と、選挙費用を 100 万(1e6= 10^6)円単位で測定する変数を作る。

```
HR <- HR |>
  mutate(experience = as.numeric(status == '現職' | status == '元職'),
         expm       = exp / 1e6)
```

次に、データから 2009 年の結果だけ抜き出し、HR09 として保存する。

```
HR09 <- HR |>
  filter(year == 2009)
```

R で線形回帰分析を行う

説明変数が二値しか取らない(ダミー変数)とき(モデル 1)

得票率(結果変数)を議員経験(説明変数)で説明するモデルを考える。議員経験は、現職または元職の候補者なら 1、そうでなければ 0 を取る二値(binary)変数(ダミー変数)である。このモデルを式で表すと、

$$\text{得票率} = \beta_1 + \beta_2 \cdot \text{議員経験}_i + e_i$$

と、なる。

R では、lm() で回帰式を推定することができる。

```
fit_1 <- lm(voteshare ~ experience, data = HR09)
```

基本的は結果は、summary() で。

```
summary(fit_1)
```

```
Call:
lm(formula = voteshare ~ experience, data = HR09)

Residuals:
    Min      1Q  Median      3Q     Max
-43.867 -12.072  -5.567   8.583  52.123
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.8772     0.6203   22.37   <2e-16 ***
experience   30.9898     0.9783   31.68   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.19 on 1137 degrees of freedom
Multiple R-squared:  0.4688,    Adjusted R-squared:  0.4684
F-statistic:  1004 on 1 and 1137 DF,  p-value: < 2.2e-16
```

broom::tidy() も使える。

```
broom::tidy(fit_1)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic   p.value
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)     13.9     0.620      22.4 3.78e- 92
2 experience      31.0     0.978      31.7 2.18e-158
```

この出力の **estimate** の列の係数の推定値(coefficient estimates)が示される。これにより、$\hat{\beta}_1 = 13.88$, $\hat{\beta}_2 = 30.99$ が得られた。したがって、

$$\widehat{\text{得票率}} = 13.88 + 30.99 \cdot \text{議員経験}$$

と、なる。

傾きの値を、分散と共分散を利用して求めてみる。分散は var()、共分散は cov() で計算できる。

```
with(HR09, cov(voteshare, experience) / var(experience))
```

```
[1] 30.98979
```

lm() で求めた傾きと一致する。

次に、行列計算で回帰係数を求める。結果変数の N 次元列ベクトルを $N \times 1$ 行列として用意する。

```
y <- matrix(HR09$voteshare, ncol = 1)
```

行列計算は、第 1 列がすべて 1、第 2 列が議員経験なので、

```
N <- length(y)
X <- matrix(c(rep(1, N), HR09$experience), ncol = 2)
```

行列の掛け算は %*%、転置は t()、逆行列は slove() で求められるので、回帰係数 b_hat は、

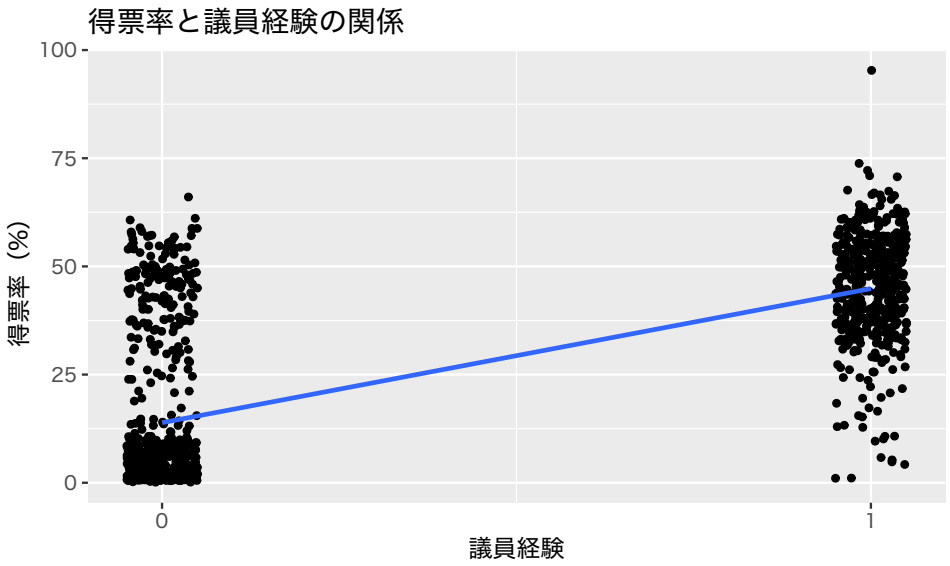
```
b_hat <- t(X) %*% X |>
  solve() %*% t(X) %*% y
b_hat
```

```
         [,1]
[1,] 13.87724
[2,] 30.98979
```

lm() を使った場合と同じ結果が得られる。

この結果を図示する。

```
p1 <- ggplot(HR09, aes(x = experience, y = voteshare)) +
  scale_x_continuous(breaks = c(0, 1)) +
  geom_jitter(position = position_jitter(width = 0.05), size = 1) +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(x = '議員経験', y = '得票率(%)')
plot(p1 + ggtitle('得票率と議員経験の関係'))
```



この図に推定の不確実性を示すには、geom_smooth(method = 'lm', se = TRUE) とすればよい。

この直線の切片である、13.88 は、議員経験がない候補者の平均得票率(予測得票率)である。予測値の式の「議

員経験」に 0 を代入すれば、これは明らかである。議員経験がある候補者の平均得票率（予測得票率）は「議員経験」に 1 を代入することで得られる。代入してみると、$13.88 + 30.99 \cdot 1 = 44.87$ となる。

議員経験ごとに平均得票率を求め、予測値と一致するか確かめる。

```
HR09 |>
  group_by(experience) |>
  summarize(voteshare = mean(voteshare),
            .groups   = 'drop')
```

```
# A tibble: 2 x 2
  experience voteshare
       <dbl>     <dbl>
1          0      13.9
2          1      44.9
```

このように、予測値は説明変数の値を与えられたときの、結果変数の平均値であることがわかる。

説明変数が連続値をとるとき（モデル 2）

同様に、得票率を選挙費用（測定単位：100 万円）で説明するモデルは、次のように推定できる。

```
fit_2 <- lm(voteshare ~ expm, data = HR09)
broom::tidy(fit_2)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic   p.value
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)     7.74    0.757       10.2 1.61e- 23
2 expm            3.07    0.0958      32.1 1.14e-160
```

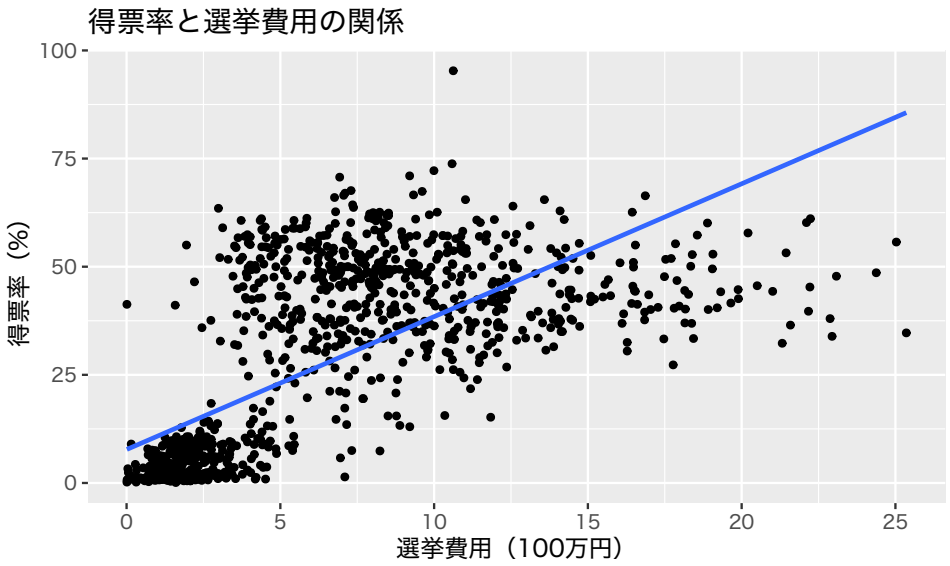
傾きの値を、分散と共分散を利用して求める。expm には欠測値があるので、除外する。

```
HR09 |>
  filter(!is.na(expm)) |>
  with(cov(voteshare, expm) / var(expm))
```

```
[1] 3.071721
```

回帰直線を図示する。

```
p2 <- ggplot(HR09, aes(x = expm, y = voteshare)) +
  geom_point(size = 1) +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(x = '選挙費用(100万円)', y = '得票率(%)')
plot(p2 + ggtitle('得票率と選挙費用の関係'))
```



95% 信頼区間を加える。

```
p2_ci95 <- p2 + geom_smooth(method = 'lm')
plot(p2_ci95 + ggtitle('得票率と選挙費用の関係'))
```