仮説検定と推定量分析

Naoko Ishibashi

2025-02-02

```
# install.packages("readxl")
# install.packages("pwr")
# install.packages("car")
library(readxl)  # Excelファイルを読み込むためのパッケージ
library(pwr)     # パワー分析を行うためのパッケージ
library(car)     # 回帰分析の診断などを行うためのパッケージ
```

```
## Loading required package: carData
```

```
# ------------------------------------------------------------------------------
# 1. ACSCountyData.Rdata データセットを読み込む。
#    回帰分析を行います。
#    独立変数 'percent.college'（各郡で大学を卒業した割合）と
#    従属変数 'median.income'（各郡の中央値の収入）の関係を分析します。

# ACSCountyData.Rdata データセットを読み込む
load("/Users/naoko/Dropbox/DATA310/Data/ACSCountyData.Rdata")

# データセットの変数名を表示
names(acs)
```

```
##  [1] "county.fips"             "county.name"
##  [3] "state.full"              "state.abbr"
##  [5] "state.alpha"             "state.icpsr"
##  [7] "census.region"           "population"
##  [9] "population.density"      "percent.senior"
## [11] "percent.white"           "percent.black"
## [13] "percent.asian"           "percent.amerindian"
## [15] "percent.less.hs"         "percent.college"
## [17] "unemployment.rate"       "median.income"
## [19] "gini"                    "median.rent"
## [21] "percent.child.poverty"   "percent.adult.poverty"
## [23] "percent.car.commute"     "percent.transit.commute"
## [25] "percent.bicycle.commute" "percent.walk.commute"
## [27] "average.commute.time"    "percent.no.insurance"
```

```
# 独立変数 (IV) = percent.college（大学卒業率）
# 従属変数 (DV) = median.income（中央値の収入）

# 回帰分析の実行
summary(lm(median.income ~ percent.college, data=acs))
```

```
##
## Call:
## lm(formula = median.income ~ percent.college, data = acs)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -37849  -6044   -623   5696  52223
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     29651.55     435.98   68.01   <2e-16 ***
## percent.college  1016.55      18.52   54.90   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9789 on 3139 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4899, Adjusted R-squared:  0.4897
## F-statistic:  3014 on 1 and 3139 DF,  p-value: < 2.2e-16
```

```
# ------------------------------------------------------------------------
```

```
# ------------------------------------------------------------------------
# 2. 回帰分析の出力を見て、
#    最初に 'percent.college' の推定係数が何を意味しているかを解釈します。
29651.55 + 1016.55 # 30668.1
```

```
## [1] 30668.1
```

分析　“percent.college” の推定係数は、大学を卒業した人々と郡の中央値収入との間の正の関係を示しています。 つまり、郡内で大学を卒業した人口の割合が1%増えるごとに、その郡の中央値収入は1016.55ドル増加することになります。

3. a. ‘percent.college’ のために自動的にテストされる帰無仮説は何ですか？
   b. 対立仮説は何ですか？
   c. その仮説検定の結果は何ですか？

解答　3a. ‘percent.college’ のために自動的にテストされる帰無仮説は？　ある郡内で大学を卒業した人口の割合が、その郡の中央値収入に影響を与えない、という仮説です。　なぜなら、帰無仮説は、大学を卒業した人口の割合と各郡の中央値収入との間に関係がないというものだからです。　例えば、percent.college に関連する 1016.55 の値が 0 である場合などです。

3b. 対立仮説は？　一方で、郡内で大学を卒業した人口の割合が、その郡の中央値収入に影響を与える、という仮説です。　対立仮説は、percent.college に関連する値が 0 以外の任意の値であることを意味します。

3c. その仮説検定の結果は？　大学を卒業した人口の割合が各郡の中央値収入に与える影響についての 1016.55 の値の確率は、　このサンプリング分布では 0 に近い <2e-16（0.0000000000000002）であり、ほぼゼロです。　したがって、帰無仮説が正しいというのは非常に稀であることを意味します。　これにより、大学を卒業した人口の割合が各郡の中央値収入に影響を与える可能性が高いことが示唆されます。

要約　帰無仮説 (Ho)　対立仮説 (Ha)　Ho = 大学を卒業した人口の割合は各郡の中央値収入に影響を与えない　Ha = 大学を卒業した人口の割合は各郡の中央値収入に影響を与える　‘percent.college’ の P値 = 2e-16。これは、P値が非常にゼロに近いことを意味します。　したがって、帰無仮説を棄却した場合、タイ

プ1エラーを犯す確率はほぼありません。　結論として、対立仮説が示すように、大学を卒業した人口の割合は各郡の中央値収入に影響を与える可能性が高いです。

帰無仮説 (Ho): 大学を卒業した人口の割合がその郡の中央値収入に影響を与えない。 対立仮説 (Ha): 大学を卒業した人口の割合がその郡の中央値収入に影響を与える。 P値: 非常に小さい値（2e-16）から、帰無仮説が正しい確率が非常に低いことを示しており、対立仮説が支持されることがわかります。

```
# ------------------------------------------------------------------------------
# 問題 2
# 1. 同じacsデータを使用し、複数回帰分析を実行します。
#    'percent.walk.commute'（徒歩通勤の割合）を従属変数、
#    'census.region'（地域）を独立変数として使用します。

table(acs$census.region) # midwest northeast south west
```

```
##
##   midwest northeast     south      west
##      1055       217      1422       448
```

```
# IV = census.region（独立変数）
# DV = percent.walk.commute（従属変数）
summary(lm(percent.walk.commute ~ census.region, data=acs))
```

```
##
## Call:
## lm(formula = percent.walk.commute ~ census.region, data = acs)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.858 -1.372 -0.577  0.538 47.403
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               3.3944     0.1105  30.710   <2e-16 ***
## census.regionnortheast    0.5072     0.2676   1.895   0.0582 .
## census.regionsouth       -1.5012     0.1459 -10.291   <2e-16 ***
## census.regionwest         2.9867     0.2026  14.741   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.59 on 3137 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.1493, Adjusted R-squared:  0.1485
## F-statistic: 183.5 on 3 and 3137 DF,  p-value: < 2.2e-16
```

```
# ------------------------------------------------------------------------------
```

2. 回帰分析の出力を見て、推定された係数が何を意味するか解釈してください（インターセプトも解釈するのを忘れずに！）。

回答 ミッドウェストはインターセプト（またはベースライン）であり、徒歩通勤の割合の平均は3.3944です。

ノースイースト地域の平均は、徒歩通勤の割合のベースラインに修正を加えます。
つまり、徒歩通勤の平均割合は3.9016です。これは3.3944に0.5072が加わった結果です。

サウス地域の平均は、徒歩通勤の割合のベースラインに修正を加えます。
つまり、徒歩通勤の平均割合は1.8932です。これは3.3944から-1.5012が引かれた結果です。

ウェスト地域の平均は、徒歩通勤の割合のベースラインに修正を加えます。
つまり、徒歩通勤の平均割合は6.3811です。これは3.3944に2.9867が加わった結果です。

```
# ------------------------------------------------------------------------------
# 3. 回帰分析の出力を使って回帰式を構築してください。
#    Y = B0 + B1*X1 + B2*X2 + B3*X3 + e
#    推定値  = ベースライン + 3.3944*(もしノースイーストの場合) + 3.3944(もしサウスの場合) + 3.3944(もしウェストの場合)
lm(formula = percent.walk.commute ~ census.region, data = acs)
```

```
##
## Call:
## lm(formula = percent.walk.commute ~ census.region, data = acs)
##
## Coefficients:
##            (Intercept)  census.regionnortheast      census.regionsouth
##                 3.3944                  0.5072                 -1.5012
##      census.regionwest
##                 2.9867
```

```
# ミッドウェスト
3.3944 + (0.5072 * 0) + (-1.5012 * 0) + (2.9867 * 0) # 3.3944
```

```
## [1] 3.3944
```

```
# ノースイースト
3.3944 + (0.5072 * 1) + (-1.5012 * 0) + (2.9867 * 0) # 3.9016
```

```
## [1] 3.9016
```

```
# サウス
3.3944 + (0.5072 * 0) + (-1.5012 * 1) + (2.9867 * 0) # 1.8932
```

```
## [1] 1.8932
```

```
# ウェスト
3.3944 + (0.5072 * 0) + (-1.5012 * 0) + (2.9867 * 1) # 6.3811
```

```
## [1] 6.3811
```

```
# ------------------------------------------------------------------------------
```

```
# ------------------------------------------------------------------------------
# 4. 上記の式を使って、ノースイースト地域で徒歩通勤する人の割合を推定してください。

# ノースイースト
3.3944 + (0.5072 * 1) + (-1.5012 * 0) + (2.9867 * 0) # 3.9016
```

```
## [1] 3.9016
```

```
# 答え
# ノースイースト地域で徒歩通勤する人の割合は3.9%です。
# ------------------------------------------------------------------------------
```

5. この回帰分析で自動的に実行される仮説検定は何ですか？

答え この回帰分析では帰無仮説検定が自動的に実行されます。

理由は、帰無仮説が「徒歩通勤の割合が地域別の国勢調査に影響を与えない」と述べているからです。
ミッドウェスト（ベースライン）を基準に、ノースイースト、サウス、西部と比較します。そのため、各地域ごとにベースラインを修正する必要があります。

例えば：

ミッドウェスト地域では徒歩通勤の割合が国勢調査に影響を与えない。
これは帰無仮説が「徒歩通勤の割合とミッドウェスト地域の国勢調査に関係がない」と述べているからです。
例えば、徒歩通勤の割合に関連する`3.3944`の値が`0`である場合です。

ノースイースト地域では徒歩通勤の割合が国勢調査に影響を与えない。
これは帰無仮説が「徒歩通勤の割合とノースイースト地域の国勢調査に関係がない」と述べているからです。
例えば、徒歩通勤の割合に関連する`3.9016`の値が`0`である場合です。

サウス地域では徒歩通勤の割合が国勢調査に影響を与えない。
これは帰無仮説が「徒歩通勤の割合とサウス地域の国勢調査に関係がない」と述べているからです。
例えば、徒歩通勤の割合に関連する`1.8932`の値が`0`である場合です。

西部地域では徒歩通勤の割合が国勢調査に影響を与えない。
これは帰無仮説が「徒歩通勤の割合と西部地域の国勢調査に関係がない」と述べているからです。
例えば、徒歩通勤の割合に関連する`6.3811`の値が`0`である場合です。

6. これらの仮説検定の結果はどうでしたか？

答え

ミッドウェスト（ベースライン）とオプション1ノースイーストの差
ミッドウェスト（ベースライン）とオプション2サウスの差
ミッドウェスト（ベースライン）とオプション3ウェストの差

テスト1 = Ho = ミッドウェストとノースイースト間の徒歩通勤の割合に統計的に有意な差がある、P値は0.0582です。
したがって、帰無仮説を棄却して徒歩通勤がミッドウェストの国勢調査に与える影響がノースイーストの人々と同じであるとすると、58.2%の確率でタイプIエラーを犯す可能性があると結論します。

テスト2 = Ho = ミッドウェストとサウス地域の徒歩通勤の割合に差はない。

テスト3 = Ho = ミッドウェストとウェスト地域の徒歩通勤の割合に差はない。

したがって、ミッドウェストとノースイーストの徒歩通勤割合には差があると結論できます。P値は0.05より大きいです。