回帰分析２：Rによる実践

講師：遠山祐太

最終更新：2024-11-16

はじめに

実証分析例：カリフォルニア州の学校データ

- 問：生徒と教師の比率が試験の成績に与える影響について分析しよう。
- AER パッケージに入っているカリフォルニア州の学校データを用いる。
  - 詳細はこのページを参照のこと。
- 以下、２ステップで分析を進めていこう。
  - Step 1：下準備から記述的分析
  - Step 2：回帰分析
- 謝辞：このスライドは Introduction to Econometrics with R に基づく。

Step 1：下準備から記述的分析

パッケージの準備

- 以下のパッケージを用いる。
  - `AER`：データソース
  - `dplyr`：データの処理
  - `summarytools`：記述統計の表示
  - `stargazer`：結果表示全般
  - `fixest` ：計量経済学的な推定

```r
# clear workspace
rm(list =ls())

# load packages
library(AER)
library(tidyverse)
library(summarytools)
library(stargazer)
library(fixest)
```

注：プログラミングのTips

- Rで外部ライブラリを利用する際には、以下の二通りがある。
  - `library()`で事前に読み込み、パッケージ内の関数を呼び出す。例：require(dplyr)の後にmutate().
  - パッケージ全体を事前に読み込みはせず、特定のパッケージの関数を使う際に[package_name]::[function_name]として呼び出す。例：dplyr::mutate().

- 関数の使う際に、`function(XX, YY)` と `function(input1 = XX, input2 = YY)`という二通りの書き方がある。
  - 前者の場合は、関数で指定されている順に引数を書く必要がある。
  - 後者の場合はその必要はない。例えば、`function(input2 = YY, input1 = XX)` としてもOK.

データの読み込み

- AER パッケージに入っているデータCASchoolsを読み込む。

```
# ワークスペースにデータを読み込む。
data(CASchools)
```

- データフレームCASchoolsがEnviornmentの中に登場する。

データの概観

- `head()` 関数でデータの概観を見る。

```
head(CASchools)
```

```
##   district                          school  county grades students teachers
## 1    75119              Sunol Glen Unified Alameda  KK-08      195    10.90
## 2    61499            Manzanita Elementary   Butte  KK-08      240    11.15
## 3    61549     Thermalito Union Elementary   Butte  KK-08     1550    82.90
## 4    61457 Golden Feather Union Elementary   Butte  KK-08      243    14.00
## 5    61523        Palermo Union Elementary   Butte  KK-08     1335    71.50
## 6    62042         Burrel Union Elementary  Fresno  KK-08      137     6.40
##   calworks   lunch computer expenditure    income   english  read  math
## 1   0.5102  2.0408       67    6384.911 22.690001  0.000000 691.6 690.0
## 2  15.4167 47.9167      101    5099.381  9.824000  4.583333 660.5 661.9
## 3  55.0323 76.3226      169    5501.955  8.978000 30.000002 636.3 650.9
## 4  36.4754 77.0492       85    7101.831  8.978000  0.000000 651.9 643.5
## 5  33.1086 78.4270      171    5235.988  9.080333 13.857677 641.8 639.9
## 6  12.3188 86.9565       25    5580.147 10.415000 12.408759 605.7 605.4
```

- `View()` 関数を使って、データセット全体をウィンドウで見ることもできる。

変数の作成

- dplyr パッケージを用いて、必要な変数を作成する。
  - STR：生徒と学生の比率
  - score：試験点数の平均

```
CASchools <- CASchools %>%
    mutate(STR = students / teachers) %>%
    mutate(score = (read + math) / 2)
```

- コメント：パイプ演算子 %>% と 代入演算子 -> に慣れよう。(詳しくは口頭で)

記述統計

- 記述統計（descriptive statistics）を作る方法はいろいろ。
- `summary()` 関数:Base Rのもの。見た目がイマイチ。
- `stargazer` パッケージ：個人的好み。しかし開発停止によりいろいろ不具合。
- `summarytools` パッケージ：比較的新しいパッケージ

summary() 関数

```
summary(CASchools)
```

```
##    district           school                   county      grades
##  Length:420         Length:420         Sonoma     : 29   KK-06: 61
##  Class :character   Class :character   Kern       : 27   KK-08:359
##  Mode  :character   Mode  :character   Los Angeles: 27
##                                        Tulare     : 24
##                                        San Diego  : 21
##                                        Santa Clara: 20
##                                        (Other)    :272
##     students          teachers         calworks          lunch
##  Min.   :   81.0   Min.   :   4.85   Min.   : 0.000   Min.   :  0.00
##  1st Qu.:  379.0   1st Qu.:  19.66   1st Qu.: 4.395   1st Qu.: 23.28
##  Median :  950.5   Median :  48.56   Median :10.520   Median : 41.75
##  Mean   : 2628.8   Mean   : 129.07   Mean   :13.246   Mean   : 44.71
##  3rd Qu.: 3008.0   3rd Qu.: 146.35   3rd Qu.:18.981   3rd Qu.: 66.86
##  Max.   :27176.0   Max.   :1429.00   Max.   :78.994   Max.   :100.00
##
##     computer       expenditure       income          english
##  Min.   :   0.0   Min.   :3926   Min.   : 5.335   Min.   : 0.000
##  1st Qu.:  46.0   1st Qu.:4906   1st Qu.:10.639   1st Qu.: 1.941
##  Median : 117.5   Median :5215   Median :13.728   Median : 8.778
```

(参考) 手書きバージョン

```
# STR と score の標本平均の計算
avg_STR <- mean(CASchools$STR)
avg_score <- mean(CASchools$score)

# STR と score の標本標準偏差の計算
sd_STR <- sd(CASchools$STR)
sd_score <- sd(CASchools$score)

# 分位点の設定と計算
quantiles <- c(0.10, 0.25, 0.4, 0.5, 0.6, 0.75, 0.9)
quant_STR <- quantile(CASchools$STR, quantiles)
quant_score <- quantile(CASchools$score, quantiles)

# データフレームに集計
DistributionSummary <- data.frame(Average = c(avg_STR, avg_score),
                                  StandardDeviation = c(sd_STR, sd_score),
                                  quantile = rbind(quant_STR, quant_score))
```

```
head(DistributionSummary)
```

```
##                Average StandardDeviation quantile.10. quantile.25. quantile.40.
## quant_STR    19.64043          1.891812      17.3486     18.58236     19.26618
## quant_score 654.15655         19.053347     630.3950    640.05000    649.06999
##             quantile.50. quantile.60. quantile.75. quantile.90.
## quant_STR        19.72321      20.0783     20.87181     21.86741
## quant_score     654.45000     659.4000    666.66249    678.85999
```

stargazer パッケージ

- 個人的なおすすめは `stargazer::stargazer()` 関数を使うこと。

```
stargazer(CASchools,
          type = "text")
```

```
##
## ===================================================
## Statistic    N    Mean    St. Dev.     Min       Max
## ---------------------------------------------------
## students    420 2,628.793 3,913.105    81      27,176
## teachers    420  129.067   187.913    4.850   1,429.000
## calworks    420  13.246    11.455     0.000    78.994
## lunch       420  44.705    27.123     0.000    100.000
## computer    420  303.383   441.341      0       3,324
## expenditure 420 5,312.408  633.937  3,926.070 7,711.507
## income      420  15.317     7.226     5.335    55.328
## english     420  15.768    18.286     0.000    85.540
## read        420  654.970   20.108    604.500   704.000
## math        420  653.343   18.754    605.400   709.500
## STR         420  19.640     1.892    14.000    25.800
## score       420  654.157   19.053    605.550   706.750
```

- 報告したい要約統計を選択できる。

```
stargazer(CASchools,
          type = "text",
          summary.stat = c("n", "p75", "sd"))
```

```
##
## ===================================
## Statistic    N  Pctl(75)  St. Dev.
## -----------------------------------
## students    420   3,008   3,913.105
## teachers    420  146.350   187.913
## calworks    420  18.981    11.455
## lunch       420  66.865    27.123
## computer    420   375.2    441.341
## expenditure 420 5,601.401  633.937
## income      420  17.629     7.226
## english     420  22.970    18.286
## read        420  668.725   20.108
## math        420  665.850   18.754
## STR         420  20.872     1.892
## score       420  666.662   19.053
## -----------------------------------
```

stargazer パッケージの補足

- 詳細はこのページ を参照。
- stargazer パッケージは回帰分析の結果を報告するのにも使われる。

summarytools::descr() 関数

```
descr(CASchools,
      style = "simple",
       transpose = TRUE)
```

```
## Non-numerical variable(s) ignored: district, school, county, grades

## Descriptive Statistics
## CASchools
## N: 420
##
##                       Mean    Std.Dev       Min         Q1     Median         Q3        Max        MAD
## ----------------- --------- --------- --------- --------- ---------- ---------- ---------- ----------
##          calworks     13.25     11.45      0.00      4.38      10.52      19.03      78.99      10.19
##          computer    303.38    441.34      0.00     46.00     117.50     376.50    3324.00     134.18
##           english     15.77     18.29      0.00      1.94       8.78      23.00      85.54      11.76
##       expenditure   5312.41    633.94   3926.07   4906.13    5214.52    5603.19    7711.51     487.17
##            income     15.32      7.23      5.34     10.64      13.73      17.64      55.33       4.79
##             lunch     44.71     27.12      0.00     23.26      41.75      66.88     100.00      32.20
##              math    653.34     18.75    605.40    639.35     652.45     665.90     709.50      19.79
##              read    654.97     20.11    604.50    640.40     655.75     668.75     704.00      20.83
##             score    654.16     19.05    605.55    640.00     654.45     666.67     706.75      19.39
##               STR     19.64      1.89     14.00     18.58      19.72      20.87      25.80       1.70
##          students   2628.79   3913.10     81.00    379.00     950.50    3011.00   27176.00    1158.65
##          teachers    129.07    187.91      4.85     19.52      48.56     146.40    1429.00      58.66
##
## Table: Table continues below
##
##
##
##                       IQR      CV   Skewness   SE.Skewness   Kurtosis   N.Valid   Pct.Valid
## ----------------- --------- ------ ---------- ------------- ---------- --------- -----------
##          calworks     14.59   0.86       1.68          0.12       4.55    420.00      100.00
```

summarytools::descr() 関数

- こちらも報告したい要約統計を選択できる。

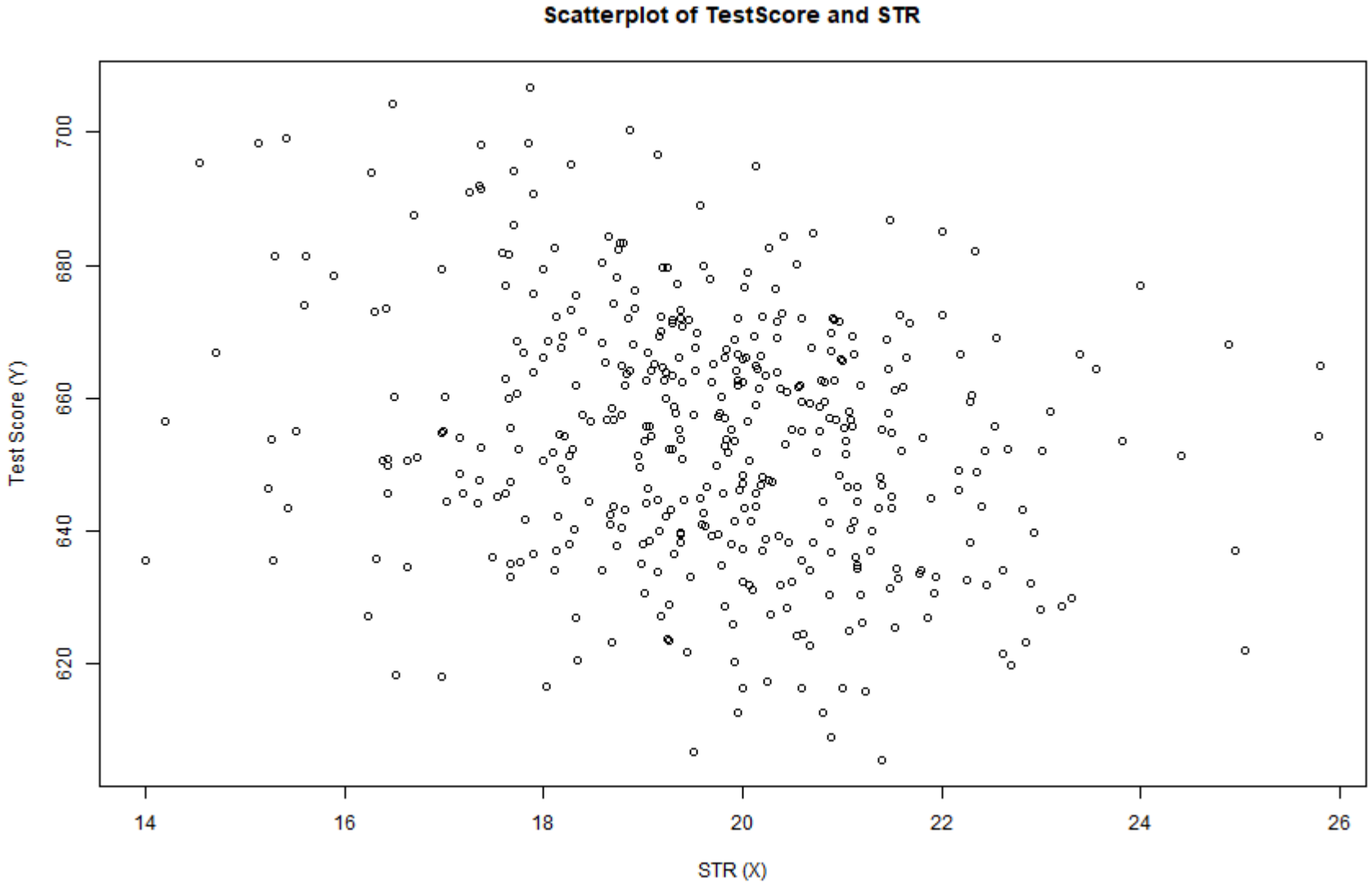
```
descr(CASchools,
      style = "simple",
      stats = c("n.valid", "q3", "sd"),
      transpose = TRUE)
```

```
## Descriptive Statistics
## CASchools
## N: 420
##
##                     N.Valid         Q3   Std.Dev
## ----------------- --------- --------- ---------
##           calworks    420.00     19.03     11.45
##           computer    420.00    376.50    441.34
##            english    420.00     23.00     18.29
##        expenditure    420.00   5603.19    633.94
##             income    420.00     17.64      7.23
##              lunch    420.00     66.88     27.12
##               math    420.00    665.90     18.75
##               read    420.00    668.75     20.11
```

散布図

- 生徒と教師の比率と試験の点数の相関を見よう。

```
plot(score ~ STR,
     data = CASchools,
     main = "Scatterplot of TestScore and STR",
     xlab = "STR (X)",
     ylab = "Test Score (Y)")
```


Scatterplot of TestScore and STR
Test Score (Y)
620
640
660
680
700
14
16
18
20
22
24
26
STR (X)

相関係数

- 数値ベクトルの相関を計算するには、 `cor()` 関数を用いる。

```
cor(CASchools$STR, CASchools$score)
```

```
## [1] -0.2263627
```

Step 2：回帰分析

回帰分析のパッケージ

- パッケージの例（その他にもいろいろ）
  - `lm()` 関数: Base R 組み込み。標準誤差の計算が均一分散を仮定。
  - `estimatr` パッケージ：不均一分散における標準誤差を考慮 `lm_robust()`など。
  - `lfe` パッケージ： `felm`関数。大量の固定効果の処理が得意。。
  - `fixest` パッケージ：`feols`関数。 `lfe`よりも計算が早い。操作変数法もOK。

- 本講義では最初から`fixest` パッケージを利用する。

回帰モデル

- 以下の回帰式を考えよう。

$$score_i = \beta_0 + \beta_1 size_i + \epsilon_i$$

ただし、 $size_i$ はクラスサイズ（生徒と教師の比率）である。
  - 以下、変数 STR をクラスサイズとよぶ。

feolsによる推定

```
# 変数`STR`の名前付け直し
CASchools <- CASchools %>%
    dplyr::rename(size = STR)

# モデルの推定と、変数`model1_summary`への結果の格納
model1_summary <- feols(fml = score ~ size,
                        data = CASchools,
                        vcov = "hetero")
```

- fml: 回帰式の指定。lmと同じ。なお、固定効果や操作変数を考えると複雑になる（来週以降）
- data: 利用するデータフレームの指定。
- vcov: 標準誤差の計算方法の指定。クラスターに頑健な標準誤差でも利用（来週以降）

推定結果

- summaryで結果を表示する。

```
summary(model1_summary)
```

```
## OLS estimation, Dep. Var.: score
## Observations: 420
## Standard-errors: Heteroskedasticity-robust
##               Estimate Std. Error  t value   Pr(>|t|)
## (Intercept) 698.93295  10.364362 67.43618  < 2.2e-16 ***
## size         -2.27981   0.519489 -4.38856 1.4467e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 18.5   Adj. R2: 0.04897
```

結果の解釈

```
summary(model1_summary)
```

```
## OLS estimation, Dep. Var.: score
## Observations: 420
## Standard-errors: Heteroskedasticity-robust
##               Estimate Std. Error  t value   Pr(>|t|)
## (Intercept) 698.93295  10.364362 67.43618  < 2.2e-16 ***
## size         -2.27981   0.519489 -4.38856 1.4467e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 18.5   Adj. R2: 0.04897
```

- 解釈
  - 教師一人あたりの生徒が1人増えると、試験の点数が2.2点下がる。
  - 標準誤差はとても小さい。クラスサイズが試験の点数に与える影響は統計的に有意である。
  - (自由度を調整していない)$R^2 = 0.051$ であり、被説明変数の分散のうち5.1%がモデルによって説明されると示唆される。
  - 問：影響の経済的な大きさについて考えてみよう。

標準誤差の比較

- summaryのvcovに"iid"を指定すると、均一分散における標準誤差を報告する。

```
summary(model1_summary, vcov = "iid")
```

```
## OLS estimation, Dep. Var.: score
## Observations: 420
## Standard-errors: IID
##              Estimate Std. Error  t value   Pr(>|t|)
## (Intercept) 698.93295   9.467491 73.82452  < 2.2e-16 ***
## size         -2.27981   0.479826 -4.75133 2.7833e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 18.5   Adj. R2: 0.04897
```

```
summary(model1_summary, vcov = "hetero")
```

```
## OLS estimation, Dep. Var.: score
## Observations: 420
## Standard-errors: Heteroskedasticity-robust
##              Estimate Std. Error  t value   Pr(>|t|)
```

説明変数を追加

詳しくは こちら を参照。

```
# 異なるモデルの推定
spec1 <- feols(score ~ size, data = CASchools, vcov = "hetero")
spec2 <- feols(score ~ size + english, data = CASchools, vcov = "hetero")
spec3 <- feols(score ~ size + english + lunch, data = CASchools, vcov = "hetero")
spec4 <- feols(score ~ size + english + calworks, data = CASchools, vcov = "hetero")
spec5 <- feols(score ~ size + english + lunch + calworks, data = CASchools, vcov = "hetero")
```

推定結果の表の作成

- **【重要】** summary() で出した結果は自分で見る分にはOKだが、レポートするものとしては不十分。
- 宿題やレポート、論文において、決してsummary()の結果を貼り付けないこと！
- 推定結果は表としてレポートするのが良い。
- 推定結果を表でまとめる方法もいろいろある。
  - `stargazer`: 最近のパッケージ(estimatr::lm_robustやfixest:feols)を受け付けない。
  - `modelsummary`: フレキシブル。その分慣れが必要
  - `fixest::etable`: fixestパッケージに付属するもの。
- 以下では`fixest::etable`を用いる。

推定結果をまとめた表を作成

```
result <- etable( list(spec1, spec2, spec3, spec4, spec5),
                  se = "hetero",
                  fitstat = c("r2", "n" ) ,
                  signif.code = NA,
                  digits = 2, digits.stats = 2,
                  depvar = FALSE)
```

- 注：`dict = c(size = "教員・学生比率",`(Intercept)`= "定数項")`オプションで変数名を日本語に変更可能。
- `signif.code =NA` で統計的有意性の記号を外している。昨今の論文は記号を載せない傾向。

推定結果

```
print(result)
```

```
##                    model 1      model 2      model 3      model 4      model 5
## Constant     698.9 (10.4)  686.0 (8.7)  700.1 (5.6)  698.0 (6.9)  700.4 (5.5)
## size          -2.3 (0.52)  -1.1 (0.43)  -1.0 (0.27)  -1.3 (0.34)  -1.0 (0.27)
## english                   -0.65 (0.03) -0.12 (0.03) -0.49 (0.03) -0.13 (0.04)
## lunch                                  -0.55 (0.02)               -0.53 (0.04)
## calworks                                                -0.79 (0.07) -0.05 (0.06)
## ____________ ____________ ____________ ____________ ____________ ____________
## S.E. type    Hetero.-rob. Hetero.-rob. Hetero.-rob. Hetero.-rob. Hetero.-rob.
## R2                   0.05         0.43         0.77         0.63         0.77
## Observations          420          420          420          420          420
```

練習問題

- 問い：説明変数を増やすとクラスサイズの係数が減少するのはなぜだろうか？（ヒント：欠落変数バイアス）