6. 差分の差分法

honocat

2026-01-07

最低賃金の影響

まず、差分の差分(difference in differences; DID)法の仕組みを理解しよう。

Card and Krueger (1994) の研究

ここでは、Card and Krueger (1994) が分析した、最低賃金の上昇が失業に与える影響の研究を例に考えよう。最低賃金が引き上げられると、雇用は減るのだろうか。Card and Krueger (1994) は、1992 年にアメリカのニュージャージー州(NJ)で最低時給が 4.25 ドルから 5.05 ドルに上昇したのに対し、隣接するペンシルベニア州(PA)では最低時給の上昇がなかった事実を利用して、分析を行った。

データは David Card のウェブサイトで公開されている。

```
dir.create('tmp')
download.file(url = 'https://davidcard.berkeley.edu/data_sets/njmin.zip',
              destfile = 'tmp/njmin.zip')
unzip('tmp/njmin.zip', exdir = 'data')
```

```
myd <- read_table('data/public.dat', col_names = FALSE, na = '.')
```

`X47` がすべて欠測値なので削除する。

```
myd <- myd |>
  select(!X47)
names(myd)
```

```
 [1] "X1"  "X2"  "X3"  "X4"  "X5"  "X6"  "X7"  "X8"  "X9"  "X10" "X11" "X12"
[13] "X13" "X14" "X15" "X16" "X17" "X18" "X19" "X20" "X21" "X22" "X23" "X24"
[25] "X25" "X26" "X27" "X28" "X29" "X30" "X31" "X32" "X33" "X34" "X35" "X36"
[37] "X37" "X38" "X39" "X40" "X41" "X42" "X43" "X44" "X45" "X46"
```

このままだと変数の区別が難しいので、変数名をつける。変数は、先程ダウンロードした zip ファイルの中にある `codebook` にかかれている。コードブックから変数名だけを抜き出したファイルが矢内勇生先生のサイトに

ある。

```
var_names <- read_lines('https://yukiyanai.github.io/jp/classes/econometrics2/contents/data/card-kru
names(myd) <- var_names |>
  print()
```

```
 [1] "SHEET"    "CHAIN"    "CO_OWNED" "STATE"    "SOUTHJ"   "CENTRALJ"
 [7] "NORTHJ"   "PA1"      "PA2"      "SHORE"    "NCALLS"   "EMPFT"
[13] "EMPPT"    "NMGRS"    "WAGE_ST"  "INCTIME"  "FIRSTINC" "BONUS"
[19] "PCTAFF"   "MEALS"    "OPEN"     "HRSOPEN"  "PSODA"    "PFRY"
[25] "PENTREE"  "NREGS"    "NREGS11"  "TYPE2"    "STATUS2"  "DATE2"
[31] "NCALLS2"  "EMPFT2"   "EMPPT2"   "NMGRS2"   "WAGE_ST2" "INCTIME2"
[37] "FIRSTIN2" "SPECIAL2" "MEALS2"   "OPEN2R"   "HRSOPEN2" "PSODA2"
[43] "PFRY2"    "PENTREE2" "NREGS2"   "NREGS112"
```

コードブック(codebook)を読み、分析に使う変数をわかりやすい名前に変えて抜き出す。

```
minwage <- myd |>
  transmute(
    state            = ifelse(STATE == 1, 'NJ', 'PA'),
      # 州
    fulltime_before  = EMPFT,
      # 最低時給上昇前のフルタイム労働者の数
    parttime_before  = EMPPT,
      # 最低時給上昇前のパートタイム労働者の数
    wage_before      = WAGE_ST,
      # 最低時給前の賃金
    fulltime_after   = EMPFT2,
      # 最低時給上昇後のフルタイム労働者の数
    parttime_after   = EMPPT2,
      # 最低時給上昇後のパートタイム労働者の数
    wage_after       = WAGE_ST2,
      # 最低時給上昇後の賃金
    full_prop_before = fulltime_before
                       / (fulltime_before + parttime_before),
    full_prop_after  = fulltime_after
                       / (fulltime_after + parttime_after)
  ) |>
  na.omit() # 完全ケース分析にする
```

処置の確認

まず、最低時給の上昇が実際に時給を引き上げたかどうかを確認しよう。そのために、賃金(時給)が 5.05 ドル未満のファーストフード店の割合を求める。

```
minwage |>
  group_by(state) |>
  summarize(before  = mean(wage_before < 5.05),
            after   = mean(wage_after < 5.05),
            .groups = 'drop')
```

```
# A tibble: 2 x 3
  state before   after
  <chr>  <dbl>   <dbl>
1 NJ     0.911 0.00344
2 PA     0.940 0.955
```

処置(NJ での最低時給引き上げ)前には、どちらの州でも大半(9 割以上)の労働者の時給が 5.05 ドル未満である。それに対して処置後は、処置がなかった PA での割合に大きな変化はない一方で、処置を受けた NJ では時給が 5.05 ドル未満なのは 0.3% のみであり、基本的には最低時給が守られていることがわかる。つまり、法律上の最低時給の引き上げは、実際に時給を引き上げたことが確認できる。

単純比較 I: 個体間比較

最低時給の引き上げが雇用にどのような影響を与えたか、単純比較による推定を試みよう。まず、処置具の NJ と PA のフルタイム労働者の割合を比較する。

```
minwage |>
  group_by(state) |>
  summarize(fulltime = mean(full_prop_after),
            .groups  = 'drop')
```

```
# A tibble: 2 x 2
  state fulltime
  <chr>    <dbl>
1 NJ       0.320
2 PA       0.272
```

単純比較を信じるなら(もちろん信じてはいけない)、最低賃金の上昇は、フルタイム労働者を 4.8 ポイント増やしたということになる。言い換えると、最低賃金の上昇は雇用を増やす。

単純比較 II: 前後比較

次に、処置の前後でフルタイム労働者の割合を単純に比較してみる。

```
minwage |>
  filter(state == 'NJ') |>
  summarize(across(.cols = starts_with('full_'), mean))
```

```
# A tibble: 1 x 2
  full_prop_before full_prop_after
             <dbl>           <dbl>
1            0.297           0.320
```

単純比較を信じるなら(もちろん信じてはいけない)、最低賃金の上昇は、フルタイム労働者を 2.4 ポイント増やしたということになる。言い換えると、最低賃金の上昇は雇用を増やす。

DID

差分の差分によって、因果効果を推定する。まず、個体間と処置前後のフルタイム労働者の割合を表にしてみよう。

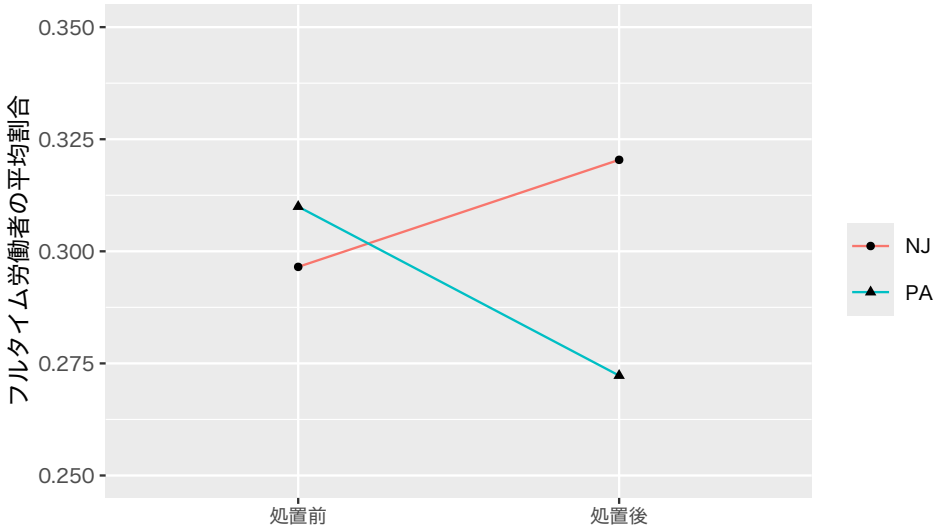
```
d_full <- minwage |>
  group_by(state) |>
  summarize(across(.cols = starts_with('full_'), mean),
            .groups = 'drop') |>
  print()
```

```
# A tibble: 2 x 3
  state full_prop_before full_prop_after
  <chr>            <dbl>           <dbl>
1 NJ               0.297           0.320
2 PA               0.310           0.272
```

ここで示した 4 つの数字の関係を可視化してみよう。

```
p_did <- d_full |>
  pivot_longer(cols         = starts_with('full'),
               names_to     = 'time',
               names_prefix = 'full_prop_',
               values_to    = 'prop') |>
  mutate(time = factor(time,
                       levels = c('before', 'after'),
                       labels = c('処置前', '処置後'))) |>
```

```
  ggplot(aes(x = time, y = prop, group = state)) +
  geom_line(aes(color = state)) +
  geom_point(aes(shape = state)) +
  ylim(0.25, 0.35) +
  labs(x = '', y = 'フルタイム労働者の平均割合') +
  scale_color_discrete(name = '') +
  scale_shape_discrete(name = '')
plot(p_did)
```



NJ と PA のフルタイム労働者の割合に**平行トレンドが仮定できるなら**、差分の差分のよって、最低時給上昇の処置効果を推定することができる。平行トレンドがあると仮定して、差分の差分を計算してみる。

```
minwage |>
  group_by(state) |>
  summarize(across(.cols = starts_with('full_'), mean),
            .groups = 'drop') |>
  mutate(dif_ba = full_prop_after - full_prop_before) |>
  with(dif_ba[state == 'NJ'] - dif_ba[state == 'PA'])
```

```
[1] 0.06155831
```

差分の差分を使うと、最低賃金の引き上げは、フルタイム労働者の割合を 6.2 ポイント上昇させると推定される。この推定値は、個体間の単純比較や前後比較による推定値よりも大きい。

回帰分析による DID の推定

DID による推定値を、回帰分析によって得る方法を考えよう。DID 回帰のために必要なのは、処置群を表すダミー変数 D、処置後を表すダミー変数 P とそれらの交差項である。ここまで使ってきたデータは横長（wide: 処置前と処置後の結果変数の値が異なる列にある）なので、縦長に変換する。

```
minwage_long <- minwage |>
  select(state, starts_with('full_')) |>
  pivot_longer(cols         = starts_with('full'),
               names_to     = 'time',
               names_prefix = 'full_prop_',
               values_to    = 'prop') |>
  mutate(D = ifelse(state == 'NJ', 1, 0),
         P = ifelse(time == 'after', 1, 0))
```

このデータフレームを使って回帰分析を行う。

```
did_fit00 <- lm(prop ~ D * P, data = minwage_long)
tidy(did_fit00) |>
  select(term, estimate)
```

```
# A tibble: 4 x 2
  term        estimate
  <chr>          <dbl>
1 (Intercept)   0.310
2 D            -0.0134
3 P            -0.0377
4 D:P           0.0616
```

得られた推定値のうち、`D:P`（D と P の交差項）の係数が、DID による推定値（先程計算した差の差の値と同じ）であることがわかる。

このように、DID 推定値は回帰分析によって得ることができる。交絡因子が想定される場合には、交絡を回帰式に含めた重回帰分析を行う。

法定飲酒年齢の影響

Angrist and Pischke (2015) の例

Angrist and Pischke (2015) の 5.2 節（pp.191-203）にある、法定飲酒年齢（minimum legal drinking age; MLDA）の変更が 18 歳から 20 歳までの若者の死に与える影響を分析してみる。

データは、Mastering 'Metrics から入手できる。

```
download.file(
  url = 'https://masteringmetrics.com/wp-content/uploads/2015/01/deaths.dta',
  destfile = 'data/deaths.dta'
)
```

```
MLDA <- haven::read_dta('data/deaths.dta')
```

分析に必要なデータは一部なので、その部分を抜き出す。Angrist and Pischke (2015) は、(1) すべての死; All deaths (dtype = 1)、(2) 自動車事故による死; Motor vehicle accidents (dtype = 2)、(3) 自殺; Suicide (dtype = 3)、(4) 内臓疾患による死; All internal causes (dtype = 6) の 4 種類の死を結果変数としているが、ここでは (1) のみを扱う。

```
myd <- MLDA |>
  filter(year  <= 1983,
         agegr == 2,    # 18-21 years old
         dtype == 1) |> # all deaths
  mutate(state    = factor(state),
         year_fct = factor(year))
```

このデータは、州(state)と年(year)の組み合わせが 1 つひとつの行を構成する州-年パネル(state-year panel)である。州の数と観測期間の数を確認しよう。

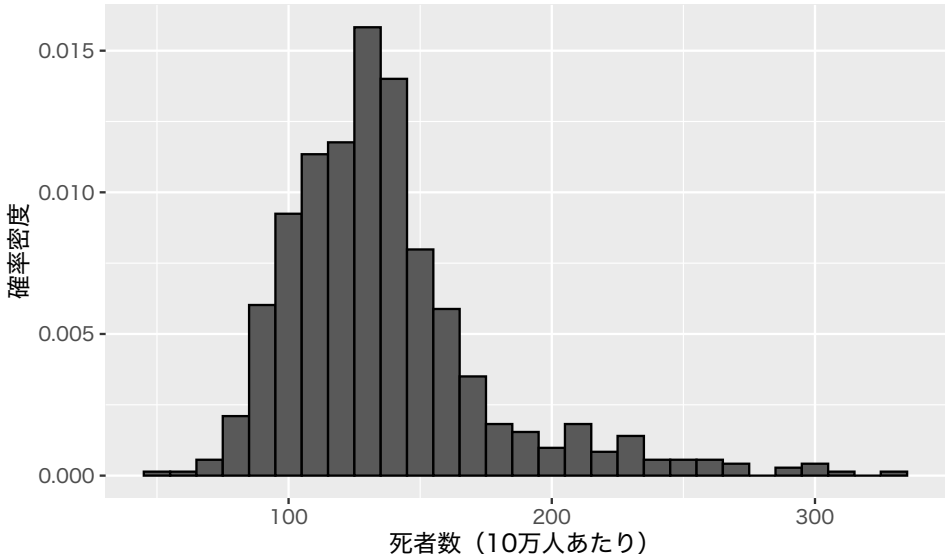
```
myd |>
  summarize(across(.cols = c(state, year),
                   .fns  = n_distinct))
```

```
# A tibble: 1 x 2
  state  year
  <int> <int>
1    51    14
```

対象となる州は 51（50 州とワシントン D.C.)、年は 14 期(14 年)あることがわかる。

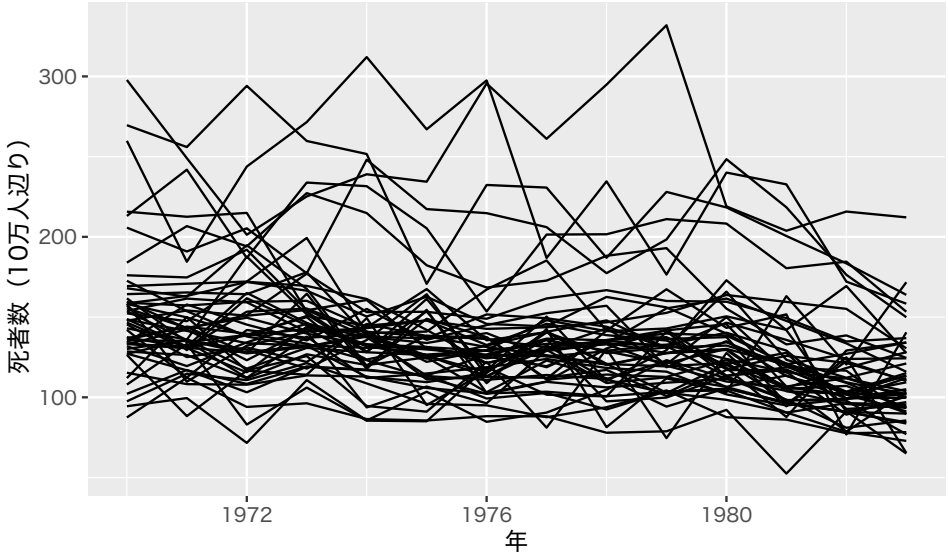
分析に使う結果変数は `mrate`（死亡率; mortality rate)である。10 万人あたりの死者数が記録されている。分布を確認しておこう。

```
hist_mrate <- ggplot(myd,
                     aes(x = mrate,
                         y = after_stat(density))) +
  geom_histogram(color = 'black', binwidth = 10) +
  labs(x = '死者数(10 万人あたり)', y = '確率密度')
plot(hist_mrate)
```



死亡率の時系列変化を可視化する。

```
ts_mrate <- ggplot(myd,
                   aes(x     = year,
                       y     = mrate,
                       group = state)) +
  geom_line() +
  labs(x = '年', y = '死者数(10万人辺り)')
plot(ts_mrate)
```



上の最低賃金の例で見たように、DID に必要な説明変数は、個体の差（処置群と統制群の区別）を表す D、時間（処置・施策が実施される前後の区別）を表す P、処置を受けたあとの観測値であることを示す $D \times P$ の 3 つだった。

この分析では、個体の差には州を表す `state` が、時間には年を表す `year_fct` が使える（`year` をそのまま使うと数値として扱われてしまうので、factor 型の `year_fct` を使う）。しかし、処置後の観測値であるかどうかは、これら 2 つの交差項では表現できない。

この分析で考える処置（介入）は MLDA の変更であるが、MLDA は 18 歳、19 歳、20 歳 21 歳のいずれかである。また、MLDA が変更されるタイミングは州によって異なる。さらに、（少なくとも理論的には）複数回の MLDA の変更も可能である。

そこで、このデータセットには処置変数として `legal` が用意されている。この変数は、特定の年の特定の州で、18 歳から 20 歳までの人口のうち何割が合法的に飲酒できるかを表す。例えば、t 年の s 州で MLDA が 21 歳だとすると、20 歳以下で合法的に飲酒できるものはいないので、この変数の値は 0 となる。MLDA が 18 歳なら、18 歳以上の全員が合法的に飲酒できるので、この変数は 1 となる（さらに、年の途中で MLDA が変更された場合には、その期間に応じて値が調整される）。この値が大きいほど、若者（18-20 歳）がアルコールにアクセスしやすいということを意味する。MLDA を引き下げるという処置（施策）を実行すると、`legal` の値が大きくなるということである。

DID 回帰

分析に使う変数が揃ったので、差分の差分を利用した回帰分析によって、MLDA の平均処置効果を推定する。`state` は 51 州を表すカテゴリ変数、`year_fct` は 14 年のうちいずれかの年を表すカテゴリ変数であるが、参照カテゴリを用意する代わりに回帰式の **formula に `0` を書くことで切片なしのモデルを推定**する。

```
fit_ap0 <- lm(mrate ~ 0 + legal + state + year_fct,
              data = myd)
tidy(fit_ap0) |>
  select(term, estimate, std.error) |>
  filter(term == 'legal')
```

```
# A tibble: 1 x 3
  term  estimate std.error
  <chr>    <dbl>     <dbl>
1 legal     10.8      3.14
```

変数 `legal` の効果が 10.8 と推定された。これが平均処置効果である。これは、18-20 歳の人々がアルコールを摂取できない状況から摂取できる状況に変化すると、10 万人あたりの死者数が約 11 人増えるということを意味する。言い換えると、MLDA を下げると、酒を飲めるようになった年齢の人たちのなかで死者が増える。

これで Angrist and Pischke (2015: 196) の表 5.2 と同じ推定値が得られたが、標準誤差が異なる。ここで得た標準誤差のほうが小さい。それは、この分析では同一州内の観測値が似ているという問題を考慮していないため、不確実性を低く見積もってしまっているからである。

そこで、州というクラスタを考慮に入れた標準誤差（cluster-robust standard error）を求めよう。そのために、`estimatr::lm_robust()` を使う。`lm_robust()` でクラスタ標準誤差を得るためには、`lm()` を使うときに指

定する無いように加え、cluster と se_type を指定する。cluster には、クラスタを表す変数を指定する。ここでは、cluster = state とする。se_type には、標準誤差を計算する方法を指定する。既定値は CR2 だが、ここでは Angrist and Pischke (2015) と同じ結果を得るために、Stata と同じ計算方法である stata を指定する。

```
fit_ap1 <- estimatr::lm_robust(
  mrate    ~ 0 + legal + state + year_fct,
  data     = myd,
  clusters = state,
  se_type  = 'stata'
)
tidy(fit_ap1) |>
  select(term, estimate, std.error) |>
  filter(term == 'legal')
```

```
   term estimate std.error
1 legal 10.80414  4.592205
```

legal の係数の推定値は先程とまったく同じであるが、標準誤差が先程より大きくなった。クラスタ化を考慮しない分析では、標準誤差が過小評価されていたことがうかがえる。この係数の推定値と標準誤差の値は、Angrist and Pischke (2015: 196) の表 5.2 の”All deaths” の行の (1) の列にある数値に一致する。

平行トレンドの仮定

- MLDA の DID 回帰は、交差項がない。なんであの回帰式で分析できるのか
- クラスタとは