DiD を理解する

北海道大学　経済学部　渡部元博

2023/02/13

目次

1 そもそも DiD とは

DiD(Difference in Difference) は大まかには、介入前後の差分を介入されたグループと介入されなかったグループでそれぞれ算出し、さらにグループ間でその差を取るという 2 回の差分を取る方法と言え、この差分の差分を取る手続きが DiD の名前の由来となっている。

地域 (企業と読み替えても可) ごとに複数の時期のデータを得られた際の DiD を適用しない単純な比較には二つの単純な比較が考えられる。一つは介入を受けた地域と介入を受けなかった地域で目的変数を比較する方法である。しかし、この場合地域と介入が完全に相関してしまう即ち、差を算出した場合に求められる効果には本来の介入の効果に加えて、地域固有の効果も含まれてしまう。

いまひとつは同一地域の介入前後の比較を行う方法である。この方法は前述した地域固有の効果が存在しないため、一見良さそうな方法のように感じられるが、この場合には、本来の効果に加えて時間を通した自然な変化 (トレンド) が算出されてしまうため、正確な介入効果を導き出せない。

このような単純な比較だけでは、推定量にバイアスが含まれてしまう。DiD はこのような問題に対処するために開発された手法であると言える。ここからは、DiD 分析で最も有名な John Snow のコレラの感染と水源の変化に関する分析を追いながら DiD がなんたるかを学習していく。

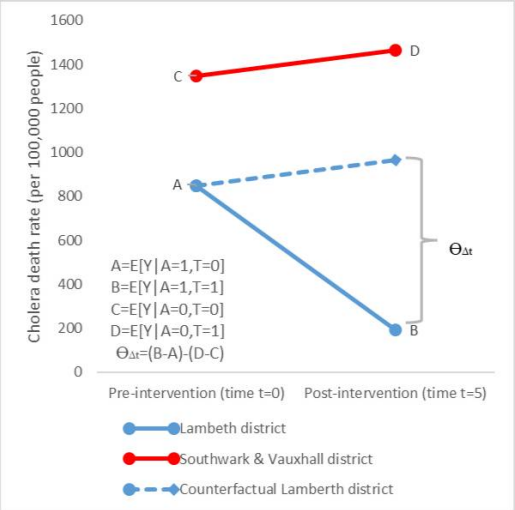
2 DiD の理論的背景

2.1 DiD の基礎

調査対象となった地域は Southwark and Vauxhall 社 (以下、SV 社) と Lambeth 社 (以下、L 社) という二つの会社によって水が提供されていた。1849 年においてはどちらの会社もテムズ川を水源としていたが、その後 L 社はより汚染が少ないと考えられる上流へと水源を移した。これにより SV 社によって水が提供されている地域では、1849 年も 1854 年も同じ水源が利用されているものの、L 社によって水が提供されている地域では 1854 年には別の水源へと変更されている状態になります。よって、もしコレラの感染源が水源にあるのだとすれば、L 社によって水が供給されている地域においてはこれらの死者数は減少するものと考えられます。以下の表で実際のデータを示す。

水道会社	1849 年のコレラ死亡者数	1854 年のコレラ死亡者数
SV 社	2,261	2,458
L 社	162	37
S 社 L 社混合	3,905	2,547

このデータからわかることは L 社が水を提供している地域では、コレラによる死亡者数は減少している一方で、SV 社のみが水を提供している地域では、逆にコレラ死亡者数は増加している。よって、もし L 社が水源変更しなかった場合に L 社が水を提供する地域で SV 社のみの地域と同様に死者数が増加するのであれば、水源の変化は死者の減少させたことを示しており、コレラの感染源は水にありそうだという説が有力である。



このように、非介入グループのデータの変化と、介入グループが仮に介入を受けなかった場合の変化が一致するだろうという仮定のことを**平行トレンド仮定 (Parallel Trend Assumption)** と呼ぶ。この仮定は DiD

の分析結果を正しいものと考えるために必要となる重要な仮定であり、この仮定が成立しない場合、推定された効果は本来の効果にトレンドの乖離分を加えたものになる。ただ、「介入グループが介入されなかった場合」というのは反実仮想であり、実際には観測されないため、データからこの仮定を確認することは基本的にできない。介入前のデータを十分に入手できる場合には、介入までのデータがトレンドが似ているかを確認することで、この仮定が満たされているかを確認できる。他の対策方法としてはトレンドが同一となるサンプルを自動的に検出する**合成コントロール (Synthetic Control)** の適用や、トレンドの乖離をもたらす共変量を投入することが考えられる。

今回の分析の状況を数式で記述する。

$$
\begin{aligned}
Y_{1854,treat} &= Time_{1854} + Area_{treat} + \tau \\
Y_{1849,treat} &= Time_{1849} + Area_{treat} \\
Y_{1854,conrol} &= Time_{1854} + Area_{control} \\
Y_{1849,control} &= Time_{1849} + Area_{control}
\end{aligned}
$$

添字の *treat,control* はそのデータが介入されたか以下を示している。*Area* は介入が行われる地域とそうでない地域における時間で変化しない地域特有の効果を表しており、*Time* は時間の固定効果を表すものである。そして、τ は介入の効果を表す。水源は 1854 年の介入が行われた地域のみで変化しているので、$Y_{1854,treat}$ のみに含まれている。

介入効果 τ は以下のような分解によって求められる。

$$
\tau = (Y_{1854,treat} - Y_{1849,treat}) - (Y_{1854,cotrol} - Y_{1849,control})
$$

この計算は二段階になっており、まずそれぞれの地域において前後比較を行い、次に前後比較の結果を地域間で比較する。一段階目の前後比較によって地域の固定効果が取り除かれる。

$$
\begin{aligned}
Y_{1854,treat} - Y_{1849,treat} &= Time_{1854} - Time_{1849} + \tau \\
Y_{1854,control} - Y_{1849,control} &= Time_{1854} - Time_{1849}
\end{aligned}
$$

この時、介入を受けた地域の前後比較に着目すると、時間による効果を水源の効果が混ざった状態となっている。この状態を改善するためにさらに差分を取る。

$$
(Time_{1854} - Time_{1849} + \tau) - (Time_{1854} - Time_{1849}) = \tau
$$

この結果、2 回の差分を取ることで、介入の効果のみを推定できることがわかった。

2.2 DiD と回帰分析

DiD は回帰分析のフレームワークに当てはめられる。先ほどの John Snow のデータを例として取り上げる。最も単純なモデルとして目的変数 Y である死者数が地域の固定効果と時間の固定効果によって、完全に説明されるモデルを考える。モデル式は以下の通りである。

$$
Y_i = \beta_0 + \beta_1 LSV_i + \beta_2 D53_i + \beta_3 LSV_i \times D53_i + u_i
$$

このモデルは両方の会社が水を供給している地域であることを表す変数 LSV と、1853 年のデータであることを表す変数 $D53$ とそれらの交差項が含まれている。交差項=1 となるときに介入が発生しているので、今回の分析で興味があるパラメータは β_3 である。

通常の回帰分析では、一つの観測対象から一つのデータが得られる一方で、DiD 分析では、同一対象に対して幾つかの期間において取得したデータが利用される。このような場合、**自己相関**と呼ばれる状態を持つデータを得る可能性がある。自己相関とは、ある時点で取得された変数の値がその近辺の時間で取得される同じ変数の値と相関するような状態を示す。この場合、誤差項の値は同一店舗で似通った値を持つことになり、誤差項の分散が小さくなってしまう。回帰分析のパラメータの標準誤差は誤差項の分散を利用するため、結果として標準誤差が過小に算出されてしまい、有意差検定の観点では、統計的に有意な結果が過剰に得られてしまう事態に陥る。このような場合には、**クラスター標準誤差**を利用してパラメータの標準誤差を算出しなければならない。

3 CausalImpact

3.1 DiD の欠点

DiD には二つの欠点がある。一つは、効果の影響を調べたい変数が複数の場所や磁気で得られている必要があることだ。スーパーの売上に対する価格変更の効果を考えると、売上データは実際に介入が行われた地域のみではなく、介入のなかった別の地域においても手に入れる必要がある。しかし、介入を行なった対象のデータしか所持していないことがよくある。今一つは、どのデータを分析に用いるのかが分析者の仮説に依存している点である。DiD 分析は、平行トレンド仮定により介入グループと非介入グループの時間変化が本来は同質であるという仮定が必要、即ち分析者はこの過程が満たされるように非介入グループのデータを形成するか共変量を調整する必要がある。しかし、平行トレンドは実際に満たされているかどうか基本的には確認できない。

3.2 CausalImpact のアイデア

DiD が「介入が行われたサンプルがもし介入が行われなかったら...」という反実仮想について、非介入データで補完する分析手法であった一方で、**CausalImapact** はさまざまな種類の変数 X **を**利用して、目的変数 Y をうまく予測できるようなモデルを、介入が行われる前の期間のみで作成する。例えば、ある店舗で価格を変えた効果を推定する際には、Y として売上を用いる一方で、X としては価格を変えた商品名の検索回数などを利用する。検索回数はその商品に興味のあるユーザー数やどのくらい認知されているかによって変動するため、商品の潜在的な売上などと強い相関を持つ。よって、価格を変えなかった場合の売上とも強い相関を持つ変数となる。(もちろん同時に他の変数も X として投入して良い)

さらに、CausalImapct では、介入前のデータを利用してどの変数のデータが Y の予測に役立つのかを判別し、自動的に利用するデータを決定しつつモデルを学習する。学習されたモデルは、X を入力すると介入前の状態の Y を予測するため、介入後の X のデータを入力することで介入が行われなかった場合の Y の値を予測として出力する。よって、この予測値と本来の Y の差が効果として得られる。

CausalImapct を利用する場合においても平行トレンド仮定は重要な役割を果たす。X も介入の影響を受けるような変数の場合にはやはり問題が生じてしまい、予測値も介入の効果を受けて変動し、最終的に推定され

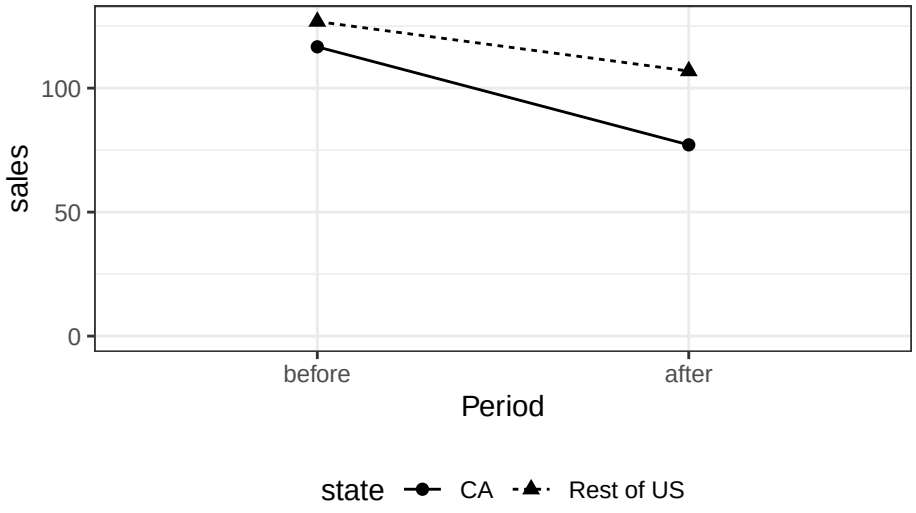
る効果の値にバイアスが生じる。また、Y から影響を受けるような変数も介入が Y に影響を与え、連鎖的に X も影響を受けるため、これも効果にバイアスを生じさせる。そのため、DiD と同様にどのデータを使うのかという点に関しては分析者に大きく依存することになる。

4 実装例: 禁煙キャンペーンの効果

DiD と CausalImpact の具体的な例として、カリフォルニア州で行われた大規模な禁煙キャンペーンの Proposition99 がどの程度タバコの消費に影響を与えたのか推定する。

Proposition99 は近代において初めての大規模な禁煙キャンペーンで、1988 年から実施された。まずタバコ 1 箱に対して 25 セントの税金をかけ、そこから得られる税収は健康やタバコの健康被害に対する教育やメディア上の禁煙キャンペーンの予算に利用された。そのほかにも室内での空気環境を改善するための条例を設立するなど多岐にわたる活動が行われ、一年あたり一億ドルの予算が使われた。このキャンペーンは当時では類を見ないほどに大規模で、かつ初の試みであったため、同様の試みは 1993 年になるまで他の週では実施されなかった。カリフォルニア州全体でこの介入が行われているため、非介入グループをカリフォルニア州の中から用意することはできないので、DiD のような分析を実施し、介入が行われなかった場合のカリフォルニア州の状態を容易する必要がある。以下はカリフォルニア州とその他の州の売上の前後比較を示したものである。この図から、効果があったと予想できる。

```
`summarise()` has grouped output by 'period'. You can override using the
`.groups` argument.
```



実際に DiD を行うために以下のような回帰式を推定する。

$$Sales_i = \beta_0 + \beta_1 ca_i + \beta_2 post_i + \beta_3 ca_i \times post_i + \sum_{t=1970}^{T} \gamma_t year_{t,i} + u_i$$

ca はデータがカリフォルニア州のものであれば 1、そうでなければ 0 となるダミー変数で、$post$ は介入が行

われた期間であることを示す変数である。$year$ は各年固有の効果を示す変数で、例えば、$year_{1970,i}$ であればサンプル i が 1970 年のデータであった場合には 1、そうでない場合に 0 を取る変数になる。分析結果を以下に示す。

介入の効果を示す交差項の係数の推定値は-20.5 となっているため、Proposition99 は一人当たりのタバコの売上を 20 箱程度減少させたことがわかった

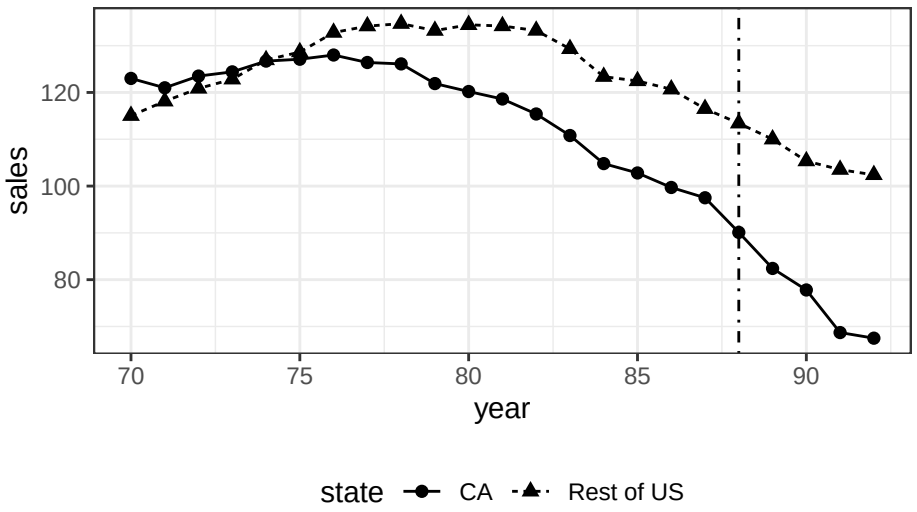
```
library(stringr)
```

```
Warning: package 'stringr' was built under R version 4.1.1
```

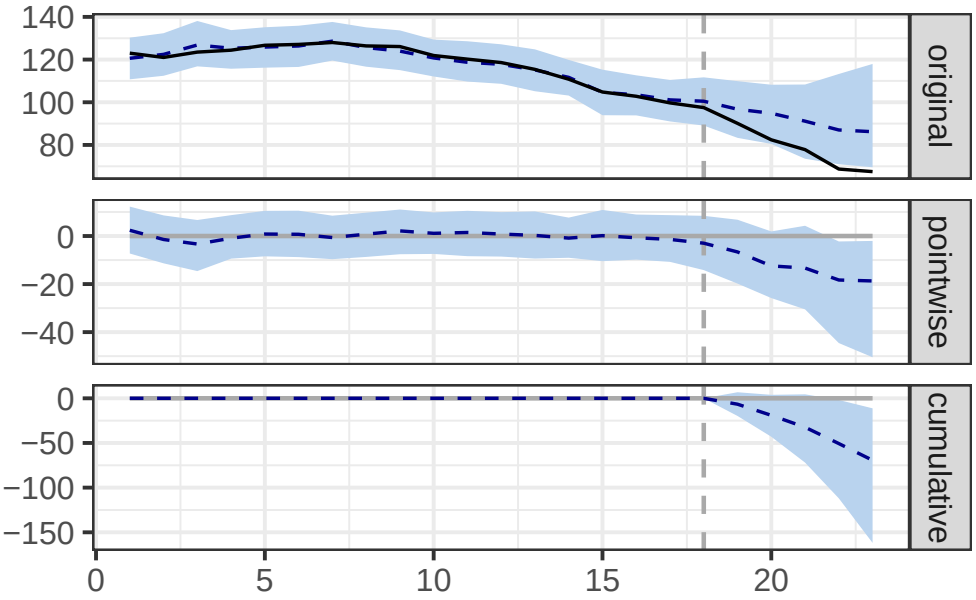
```
Cigar_did_sum_reg <- Cigar_did_sum %>%
  lm(data = ., sales ~ ca + post + ca:post + year_dummy) %>%
  tidy() %>%
  filter(!str_detect(term, "state"),
         !str_detect(term, "year"))
Cigar_did_sum_reg
```

```
# A tibble: 4 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   124.       4.52     27.3  6.69e-18
2 ca             -9.09     2.07     -4.38 2.60e- 4
3 post          -23.8      6.61     -3.60 1.67e- 3
4 ca:post       -20.5      4.45     -4.62 1.48e- 4
```

```
`summarise()` has grouped output by 'year'. You can override using the
`.groups` argument.
```



しかしながら、平行トレンド仮定を満たしているか確認してみると、介入前の売上のトレンドが、カリフォルニア州とその他の州で平行に動いていることが確認できなかったため、トレンドに差異があり、DiD 分析に適していないデータであると考えられる。そこで、CausalImpact を実施する。CausalImapact は目的変数 Y と予測に利用する X、そして介入期間を表す変数をそれぞれ入力する必要がある。この時変数 X は今まで登場してきたような共変量ではなく、平行トレンドの仮定を満たすような変数であると注意する。目的変数はカリフォルニアの売上高である。続いて予測に利用する変数を用意する。ここでは、カリフォルニア州以外のタバコの売上を利用する。



original とタイトルのついたグラフは、カリフォルニアにおけるタバコの売り上げとそれに対する予測値を横軸に時間をとって表している。実践が実際の売上で点線が予測値となっており、介入が始まったタイミングを示す縦の点線の前後で売上高の実測値と予測値に乖離が生まれていることがわかる。

pointwise と名前のついたグラフは、original のグラフで示す実測値と予測値の乖離をプロットしたもので、各時点における介入の効果を示しているグラフである。

cumulative と名前のついたグラフは、pointwise のグラフで示す値を介入が始まったタイミング以降で積み上げ上げたものであり、proposition99 はタバコの売上を低下させる施策であるので、今回は下に増加していっている。pointwise のグラフを見て分かる通り、CausalImapact の特徴に各年で効果を推定できる点がある。DiD においては介入開始後の平均効果しかわからない。このことから段階的に強化されるような介入であっても、その効果が反映されるかをみることができる。またモデルの結果を保存した変数を呼び出すことで、推定された効果を実数と比率で確認することができる。

```
impact
```

```
Posterior inference {CausalImpact}

                         Average          Cumulative
```

```
Actual                    77              386
Prediction (s.d.)         91 (6.8)        456 (34.0)
95% CI                    [80, 110]       [398, 548]

Absolute effect (s.d.)    -14 (6.8)       -69 (34.0)
95% CI                    [-32, -2.2]     [-162, -11.2]

Relative effect (s.d.)    -15% (6.2%)     -15% (6.2%)
95% CI                    [-30%, -2.8%]   [-30%, -2.8%]

Posterior tail-area probability p:   0.0152
Posterior prob. of a causal effect:  98.48%

For more details, type: summary(impact, "report")
```

Absolute effect は効果を実際の売上の箱数で表しており、左側 (Average) に平均的な効果、右側 (Cumlative) には積み上げの効果を示している。それぞれの値に続く括弧の中にはその標準誤差が報告されている。よって、ここでの効果は平均的には-14 箱でありその標準誤差は 7.5 になる。この時の 95% 信用区間はさらにその下に表示されている。