[R] CausalImpact でできること, できないこと

2019 年 10 月 9 日

概要

Brodersen et al. (2015) により提案され, R で実装された時系列因果推論フレームワーク, `CausalImpact` は, シンプルで分かりやすい difference in differences (DID) の因果推定理論に基づいており, マーケティングイベントがもたらすインパクトを計測するツールとして紹介されている. しかし, DID が非常にシンプルであれるのは, 厳格な仮定を置いているからであり, 利用する際には多くの注意が伴う. そこで今回は, より発展的な理論について考察したことを垂れ流してみる. あとついでに `tsibble` パッケージの使い方とかも少しだけ触れている.

この問題は CausalImpact の考案以前からある議論についても振り返る必要があるので, まず Rubin (1974), Rubin (1990) の因果推論フレームワークに沿って, 関連する話題, つまり **差の差推定** (DID), **シンセティック統制法** (synthetic control), **ベイズ構造時系列モデル** (bsts), そして causal impact フレームワークを解説した上で議論を進める.

対象読者層は, 企業などで causal impact を使って効果測定をしている人達だろうか. 大まかな要約と, 技術的に細かいパートをなるべく切り離し, 要点をよく強調して説明したつもりだが, まったくの初心者にとってはもしかすると難しいかも知れない.

1　イントロダクション

先日, 第 80 回, Tokyo.R で以下のような発表をした.

- 『計量経済学と機械学習の関係 –AI はさだめ, さだめは反事実的 –』

このスライドの内容に加筆修正してブログで公開する予定だったが, おそらくそれなりに時間がかかるため, 一旦本件と少しだけ関連のある話題について話す.

このスライドの中で, 近年の計量経済学と機械学習の融合の成果として `CausalImpact` パッケージを紹介した. `CausalImpact` パッケージは, 考案者の言葉を借りるなら, 「新プロダクトのリリース, 新機能の追加, そして広告キャンペーンの開始 or 終了といったマーケティング上のイベントのインパクト」の大きさを測るためのツールである.

- google/CausalImpact

CausalImpact のしくみを, 日本語で簡単に解説した資料はすでに存在する. 代表的なものとして, 以下 2 つ

がある. いわゆる因果推論の話を少し勉強した人ならば, これらを読んだだけでも何をやっているか察せると思う.

- 機械学習で広告の効果を推定したいお話。| 分析のおはなし。
- DID, Synthetic Control, CausalImpact

一方で, これらでは触れられていない DID の持つ強い仮定が存在する. **今回主に話したいのは, その仮定を無視するとどうなるかという話である.**

また, この記事によって, はるか昔, 2014 年に書いた記事での「今回は DID については書かないが, 暇があれば書く.」という伏線を回収した.

ところで, 下書きを途中まで書いて放置していたら, 以下のようにテーマが丸かぶりの投稿が発生した. しかし, よく読んだら私の主題とあまりかぶってなかったのでこのまま投稿することにする.

- {CausalImpact} を使う上での注意点を簡単にまとめてみた

タイミングが重なったのは対抗意識を燃やしているわけではなく, 本当に偶然である[*1].

本稿の以降の構成は, 次の通り: 第 2 節で, Causal Impact の下地となった方法論である, **差の差の推定法** (DID; Difference in Differences), **シンセティック統制法**[*2] (SC; Synthetic Control), そして Causal Impact の実装に用いられている**ベイズ構造時系列モデル** (BSTS; Bayesian Structural Time Series), 最後に Causal Impact の仕組みを再度, 定式化して説明する. この辺は適当にごまかして説明することもできるので, **Causal Impact の実用上の話を知りたいだけならば, 4 節まで飛ばしても問題ない.**

第 3 節では, Causal Impact の問題設定に由来する, フレームワークの限界について説明する. そして第 4 節では, 第 3 節での説明を踏まえた, 簡単なシミュレーションによる事例を紹介する. 第 5 節では, 以上の内容の要約と, それを踏まえてた発展的な議論をする.

2 方法論の解説

イントロダクションで紹介した資料の, 特に 2 番目のものでも DID から causal impact に至るまでの議論の解説はされているが, ここではより発展的な話題のため, 因果推論の基本的な考え方, DID, SC 法, BSTS について, 再度定式化して説明する.

2.1 Rubin 流の因果推論

今回紹介する Causal Impact は DID の方法論がベースであり, DID は Rubin (1974) の考案した因果推論のフレームワークに則したものである. この因果推論の基本的な考え方については, 黒澤 (2005), 星野 (2009), Angrist and Pischke (2009), 森田 (2014) が参考になるだろう[*3].

[*1] 計量経済学をかじった人間ならば「見せかけの相関」と理由付けるべきか

[*2] この訳語は私の考案したものであり, 少し調べた限りではまだ日本語訳は存在していない. control が統制実験, 対照実験などと訳される control experiment を意味すると思うので, 「シンセティック対照実験法」とか「合成的対照実験法」という訳でもよいかもしれない.

[*3] 日本語に限定すれば, (星野, 2009), 次いで (Angrist and Pischke, 2009) の邦訳, (黒澤, 2005) が最も詳しい. なお, 黒木学氏や宮川雅巳氏の「構造的因果モデル」や「統計的因果推論」を扱った教科書は, 今回説明する因果推論とはアプローチが異なるため, 今回のテーマに関しては参考書としてはおすすめできない. 今回紹介するフレームワークも「統計的」ではあるのだが, 和書で統計的因果推論というと, Judea Pearl 流のグラフ理論に基づく因果推論フレームワークを指すことが多いようだ. よって, Pearl の因果推論

患者への治療措置が健康状態をどの程度改善するかであったり, 自治体による政策によって市民の生活がどう向上したかであったり, 様々な事例に応用できる. そのため, これ以降は用語を一般化して, 施策, 治療, 政策といった行為を**処置** (tretment. 介入, intervention と呼ぶ文献も多い) と呼び, A/B テストの B に相当する, 処置を与えたグループを**処置群** (treatment group), A/B テストの A に相当する, 施策を与えていないグループを**対照群** (control group) と呼ぶ. KPI など結果を表す変数を, **結果変数** (outcome) と呼ぶ. 因果推論では, 処置をしたか, しないかによる違いによって発生した効果を**介入効果** (interventional effect) と呼び, どうやって介入効果の大きさを推定するかという研究がされている.

処置された場合, されなかった場合の結果変数をそれぞれ y_1, y_0 とする. y に対応するものは, 先ほど挙げたように好きな KPI を想像して良い. **割り当て変数** D は 1 またはゼロの 2 通りのみの取りうる変数で, 処置を与えた場合は 1, そうでなければゼロである. すると, ある処置による介入効果 (これを特に**平均処置効果**, ATE という) の大きさは両者の差の期待値,

$$ATE := \mathrm{E}\left[y_1 - y_0\right]$$

と表現できる. 単純に考えれば, 全ての個体に対して $y_1 - y_0$ の差を平均したものが ATE だが, **この方法では絶対に計算できない**. 実際に観測できる結果変数は, 割り当て変数 D に依存するので $y(D)$ として以下のように,

$$y(D) := Dy_1(D) + (1 - D)y_0(D)$$

と定義できる. $y(D)$ は観測できるが, $(D = 0)$ 個体が, **現実とは違って, もし仮に**, 処置群に割り当てられていた $(D = 1)$ 場合の結果である $y_1(0)$ と, 逆に処置群に割り当てられていた $(D = 1)$ が, 事実とは逆に対照群に割り当てられていた $(D = 0)$ 場合の結果 $y_0(1)$ は, **絶対に観測できない**ため, 上記は計算できない. つまり, 各個体ごとにある 2 通りの結果 $(y_{i,1}(D), y_{i,0}(D))$ について, 現実的にはどちらか片方が常に欠測している. そのため, $\mathrm{E}\left[y_1(1) - y_0(0)\right]$ と $\mathrm{E}\left[y(1) - y(0)\right]$ は一般に一致せず, このままでは ATE を計算できない. この ATE を中心としたフレームワークは, このような理由から**反事実的**または**反実仮想** (counterfactual) **モデル**と呼ばれる[*4].

しかし, A/B テストでは, 無作為に A/B 両グループへの割り当てを行うことが多い. このような場合は, 1 つ 1 つの個体の結果にばらつきがあっても, ランダム性により処置群・対照群の結果変数をそれぞれ平均したもので差を取れば個体ごとの効果の差が平均によって打ち消されるため, ATE を求めることができる. これは統計学では**ランダム化比較試験** (RCT) と呼ばれる. RCT で得た観測データがあれば, 以下のように単純に観測できた各群の平均の差をとることで ATE を推定できる. 以降, 簡単のために $y_1(D), y_0(D)$ を単に y_1, y_0 と表記する.

$$\begin{aligned}
\hat{ATE} :=& \frac{1}{card(T)} \sum_{i \in T} y_{1,i} - \frac{1}{card(C)} \sum_{i \in C} y_{i,0}, \\
T :=& \{i : y_{1,i} \text{ is observed}\}, \\
C :=& \{i : y_{0,i} \text{ is observed}\}
\end{aligned}$$

さらに, 処置群に対して, 現実どおり行った場合と現実に反して処置を行わなかった場合の差を見る**処置群での平均処置効果** (ATT; Average Treatment Effect on the Treated) という概念も存在する.

$$ATT := \mathrm{E}\left[y_1 - y_0 \mid D = 1\right]$$

を勉強したいのならばこれらの教科書も非常に参考になる.

[*4] このような考え方は Donald Rubin による反事実 (counterfactual) モデルあるいは**潜在効果** (potential outcome) モデルと呼ばれるが, counterfactual という語は因果推論の分野に限っても, 文脈によって意味が異なるので注意が必要である.

両者は名前も式も似ているが, はっきりと別物である. さらに, 対照群での平均処置効果 (ATU; Average Treatment Effect on the Untreated), $ATU := \mathrm{E}\left[y_1 - y_0 \mid D = 0\right]$ を定義すれば,

$$ATE := ATT \times p(D=1) + ATU \times p(D=0)$$

という関係が成り立つ. 実務上は ATE または ATT を求めたい場合が多く, ATU を意識することはあまりないが, この関係式が示唆することは重要である. 黒澤 (2005) にあるとおり, 両者が一致するには, 個体ごとの効果が平均的には処置群と対照群で同じであることが条件だが, ATE と ATT は必ずしも一致しないことがわかる. 制約の強さから, 実際には ATT を求めることが多いだろう. 逆に言えば RCT はそれだけ強力である[*5].

2.2 Difference in Differences (DID, DD) 推定法

ATT の式にも, 観測できない結果 y_0 が含まれている. これは RCT が前提ならば $y(0)$ で代用でき, よって ATE と同値になる. RCT が適用できない場合, さらに別の方法が必要であり, その 1 つが DID である.

DID は因果推論のフレームワークとして歴史が古く[*6], しかもシンプルな発想である. 前節で挙げた Rubin 的因果推論の文脈で書かれた教科書が多く, 例えば 星野 (2009), Angrist and Pischke (2009), 森田 (2014) や, 『岩波データサイエンス』シリーズでも 山口 (2016) で具体的な事例を交えて解説されている.

DID は **RCT よりも柔軟性をもたせた方法と位置づけられる**. RCT の場合は処置群と対照群がそれぞれ均質であると案に仮定しているが, DID ではその仮定が満たされなくとも ATT を推定できるという長所がある. DID は, ある処置を実行後だけでなく, **実行前でも処置群と対照群それぞれの結果変数を観測したデータが手元にある**という前提の方法である. そこで, 結果変数 $y(D)$ に時間を表す記号 t をつけて, $y_t(D)$ とする. とはいえ, 今考える必要があるのは, 処置前の時点 $t = b$ (before) と, 処置後の時点 $t = a$ (after) の 2 時点だけである[*7].

まず, 処置群の, 処置前後での結果の差,

$$\Delta y(1) := y_a(1) - y_b(1)$$

は, 処置の実行前後の結果の差をもって処置 D の「介入効果」と言えるだろうか? A/B テストは, A/B の 2 つのグループを比較するという方法論だった. しかしこれは処置群だけの差しか取っていない. 外部からの介入を完全に遮断する実験室での実験でもない限り, $t = b$ から $t = a$ の期間には, 処置以外にも結果に影響する出来事が多く発生しているはずである. 競合の参入, 市場構造の変化, 自社の営業の努力など, 色々な要因が考えられる. そこで, そこで, 処置を行わなかった対照群についても差,

$$\Delta y(0) := y_a(0) - y_b(0)$$

を考える. この差の大きさは, $\Delta y(1)$ と同じように, **期間中の様々な外部要因による結果変数の変化を表しているのではないだろうか?** しかし, $\Delta y(0)$ は, 前者と違って, 処置の影響がない. よって, $\Delta y(1)$ と $\Delta y(0)$ の差を取れば, **処置以外による期間中の変化を除去できる**のではないか, と考えられる. というわけで, この施策のある個体とそうでない個体の差による介入効果 τ_{DID} を, 以下のような 2 重の差として定義する.

[*5] ただし, それは理論上の話である. 実用上はさらにいろいろな問題を考える必要がある. こういう話について知りたい場合, 星野, 2016 が入門として優れている.

[*6] もともと DID は保健・公衆衛生の研究として生まれ, かなり古くからある概念である (https://www.mailman.columbia.edu/research/population-health-methods/difference-difference-estimation).

[*7] しかし, 後で確認したところ 星野 (2009) では a を処置前, b を処置後として記述していたので, かなり紛らわしいことになってしまった.

$$\tau_{DID} := \mathrm{E}\left[y_a(1) - y_b(1)\right] - \mathrm{E}\left[y_a(0) - y_b(0)\right]$$

実は, τ_{DID} は, ATT と同じである. そのため, もう一度厳密に定式化して説明する (が, 冗長なので退屈だったら節末のまとめまで読み飛ばしてもいい).

$$y_t(D) := Dy_{1,t}(D) + (1-D)y_{0,t}(D), t = b, a$$

と書ける. そこで, 処置群の, 処置前後での結果の差は, $\Delta y(1) := y_a(1) - y_b(1)$ と書け, 処置を行わなかった対照群についても差を, $\Delta y(0) := y_a(0) - y_b(0)$ と書ける. そして, この施策のある個体とそうでない個体の差による介入効果 τ_{DID} を, 以下のような 2 重の差として定義する.

$$\begin{aligned}
\tau_{DID} :=& \Delta y(1) - \Delta(0) \\
=& \mathrm{E}\left[y_a(D) - y_b(D) \mid D = 1\right] - \mathrm{E}\left[y_a(D) - y_b(D) \mid D = 0\right] \\
=& \mathrm{E}\left[y_{1,a}(1) - y_{1,b}(1)\right] - \mathrm{E}\left[y_{0,a}(1) - y_{0,b}(1)\right].
\end{aligned}$$

これをさらに変形すると,

$$\begin{aligned}
\tau_{DID} =& \mathrm{E}\left[y_{1,a}(D) - y_{1,b}(D) \mid D = 1\right] - \mathrm{E}\left[y_{0,a}(D) - y_{0,b}(D) \mid D = 1\right] \\
&+ \mathrm{E}\left[y_{0,a}(1-D) - y_{0,b}(1-D) \mid D = 1\right] - \mathrm{E}\left[y_{0,a}(D) - y_{0,b}(D) \mid D = 0\right]
\end{aligned}$$

となる. ここで, 1 段目の部分は, 2 つの項がいずれも $D = 1$ で条件づけられており, ATT と同じである. よって 2 段目の部分がゼロ, つまり,

$$\mathrm{E}\left[y_{0,a}(D) - y_{0,b}(D) \mid D = 1\right] = \mathrm{E}\left[y_{0,a}(1-D) - y_{0,b}(1-D) \mid D = 0\right]$$

ならば, τ_{DID} は ATT と同値になる. このように, DID もまた, Rubin 的な因果推論に沿ったフレームワークであることがわかる.

この条件は, 「処置群が処置の対象となった場合と, 仮にならなかった場合とで, 処置前後の変化が同じ」ということを意味しているため, **共通トレンド仮定**と呼ばれている. よって, 図 1 の $\hat{y}_b(1), \hat{y}_a(1)$ のように, 対照群と傾きが同じである場合のみ, DID による推定が ATT である.

最も単純な方法では, ATE のように各標本平均とその差を計算すれば DID 推定ができる. あるいは, 他に結果変数に影響することが分かっている共変量があるならば, 処置群, 対照群のデータをプールして次のような 1 つの重回帰モデルで推定することができる.

$$y_{i,t} = \alpha + \tau_{DID} D_i D_a + \beta D_i + \gamma D_a + \boldsymbol{x}_i' \boldsymbol{\theta} + \varepsilon_i$$

ここでは, 個体 i の処置前後の結果変数が $y_{i,t}$ であり, D_i が個体 i の処置の有無を表す指示関数, D_a が, $t = a$ であるときのダミー変数, $\boldsymbol{x}_i$ が結果に影響を及ぼす共変量で, ε_i が誤差項である.

2.2.1 DID のまとめ

DID の性質をまとめると, 以下のようになる.

- 差の差推定法 (DID) は, 施策の前後の差をとり, さらに施策の有無で差をとることで介入効果 (ATT) を推定する方法である.

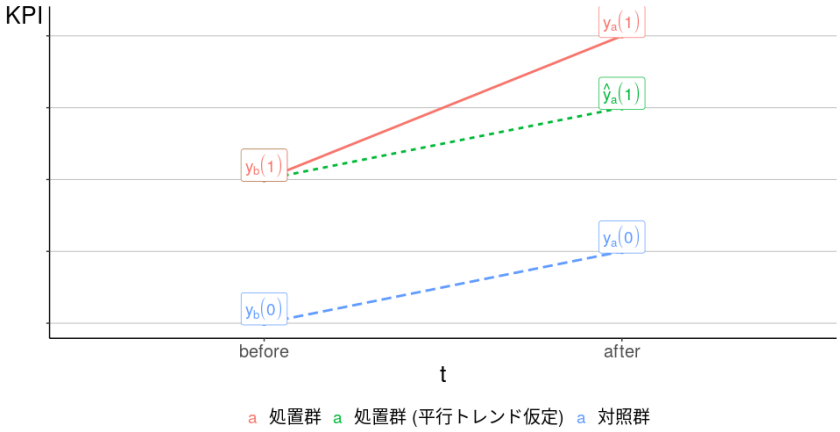

図1　DID と平行トレンド仮定のイメージ (山口, 2016 を参考に作成)

- 最もシンプルな DID は, 標本平均の差で求められる.
- DID は, 処置群・対照群は介入効果を除いて, 全く同じ外部要因から, 同じ大きさの影響を受けているという前提 (共通トレンド仮定) で ATT を推定しているのに等しい.

2.3　シンセティック統制法 (Synthetic Control Method)

SC 法は, DID の問題点を克服した方法として, Abadie and Gardeazabal (2003), Abadie et al. (2010) によって考案された. 彼らは DID の欠点を 2 つ挙げている.

1. DID のスキームでは, どのように比較対象を選ぶべきかのルールが曖昧である (いわゆる cherry-picking の危惧)
2. DID のスキームは従来的な統計推論が用いられており, これで定量化できる推定結果の不確実性は, 集計によるもの (例えば標本平均に対する標準誤差) に過ぎず, 反事実的な要因によるものを反映していない.

SC 法では, $t = T_0$ が介入直前時点で, それ以前は処置群・対照群いずれも処置がなされない. $t = T_0 + 1$ 時点以降, 処置群に対して処置がなされる. $J + 1$ のグループがあるとして, そのうちグループ 1 のみが処置群で, 残りの $2, \cdots, J + 1$ は対照群とする. DID と同様に, グループ j の t 時点の結果変数を $y_{j,t}(D)$ とする. DID と同様に, ATT を介入効果として推定する.

$$\tau_t := \mathrm{E}\left[y_{1,t}(D) - y_{j\neq 1,t}(D) \mid D = 1\right]$$

と書ける.

DID では単純に処置群・対照群の差分で s 求めた ATT を介入効果として推定する方法だったが,SC 法では先に挙げた批判を解消するため, 複数の対照群から擬似的に処置群の $y_{0,t}(1)$ を生成して介入効果を計算する. DID では単に 1 つだけの対照群を用意する (特定の州とか, 地方に限定して収集したデータを使うことが多い) ため, cherry-picking にもなりやすい, そこで, 複数の対照群 $\{y_{j,t}\}$ から $y_{1,t}$ を求める.

結果変数が, 以下のように, 個体要素と時系列要素で構成される簡単な線形回帰モデルであると仮定する.

$$y_{i,t} := \delta_t + \boldsymbol{z}_i^\top \boldsymbol{\theta}_t + \boldsymbol{\mu}_i^\top \boldsymbol{\lambda}_t + \varepsilon_{i,t} \tag{1}$$

なお, $\cdot^\top$ は転置記号である. この仮定は, この SC 法の持つ特徴のエッセンシャルな部分を抜き出す最も単純なケースである. 右辺の第 1 項 δ_t は未知の時間変化するトレンドで, 処置・対照群とで共通している. つまり DID でいう共通トレンド仮定に対応する. 第 2 項 $\boldsymbol{z}_i^\top \boldsymbol{\theta}_t$ は共変量 $\boldsymbol{z}_i$ と係数パラメータ $\boldsymbol{\theta}_t$ の内積である. 添字の通り, 共変量は個体ごとには異なるが, 時間に対して一定であり, 係数は時間に対して変化しうる.

第 3 項の $\boldsymbol{\mu}_i^\top \boldsymbol{\lambda}_t$ は, いわゆる, **観察できない個別効果** (unobserved indivisual effect) である. $\boldsymbol{\mu}_i$ もまた個体ごとの固有効果だが, データとしては取得できない. つまり, いわゆる**交絡効果** (confounder) を表している. 計量経済学に詳しくない人間に補足しておくと, 回帰モデルのパラメータの特定の際にはバイアスなく推定することが必要だが, このようなデータから観察できない交絡効果 (経済学では, **内生変数**と呼ぶことが多い) が潜んでいる場合はバイアスが発生することが知られている[*8]. 計量経済学の研究テーマは, 交絡効果の扱いがかなりの割合を占める. よって, 単純ではあるが (もちろん適切な仮定をおけばここから非線形モデルに拡張することもできるだろう), 観察データから介入効果を測定する際に重要な仮定が揃っている.

第 4 項はいわゆる誤差項で, ここでは i, t いずれに対しても独立であると仮定している. 特に DID との大きな違いは, 第 3 項の交絡効果項の存在である. 変数 $\boldsymbol{\mu}_i$ は観察できないため, 見かけ上は誤差項 $\varepsilon_{i,t}$ に含まれている. しかし, $\boldsymbol{z}_i$ と交絡しているため, 観察できる $\boldsymbol{z}_i$ だけで計算すれば推定結果にバイアスが発生する. DID では交絡効果を全て排除したという前提であるが, これはかなり強い仮定で, 観察データに含まれている情報だけであらゆる交絡効果を排除したと言い切ることは難しい.

SC 法では, J 個の対照群の加重平均から, 処置群の観察できない反事実的な結果 $y_{1,t}(0)$ を生成する. つまり, $\sum_{j=2}^{J+1} w_j = 1$ となるような非負の重み $w_j \geq 0$ で,

$$\begin{aligned} y_{1,t}(0) &\simeq \sum_{j=2}^{J+1} w_j^* y_{j,t}, \forall t = 1, \cdots, T_0, \\ \sum_j w_j^* \boldsymbol{z}_j &\simeq \boldsymbol{z}_1 \end{aligned} \tag{2}$$

を満たすような $\boldsymbol{w}^* := \begin{bmatrix} w_2 & \cdots & w_{J+1} \end{bmatrix}$ が存在すると一旦仮定する. このとき, (2) について, 対照群の結果変数 $y_{i,t}$ の重み平均が処置群の結果変数と一致すること, さらに共変量についても対照群の重み平均が処置群に一致することを条件としている. これは, 処置群の結果変数だけでなく共変量も対照群の合成で表現できる, という制約条件である. 単なるカーブフィッティングの発想ならば, 結果変数 $y_{1,t}$ のみに近似させればよいように見えるが, ここでは $\boldsymbol{\theta}_t$ が時間変化するため, 共変量もコントロールしなければ, $t = T_0 + 1$ 以降での処置群の推移を再現できないからである (ただし, 介入の前後で構造的な関係が大きく変化しないという制約は必要である.). つまりこの式は, 共変量を所与とした結果変数の条件分布 $p(y_{1,t} \mid \boldsymbol{z})$ だけでなく, 共変量と結果変数の同時分布も線形関数で近似できるというナイーブな仮定を意味している[*9]とみなせる.

このような $\boldsymbol{w}^*$ が分かれば, $(y_{j,t}, \boldsymbol{z}_j)$ の外挿によって $t = T_0 + 1$ 以降も $y_{1,t}(0)$ を擬似的に生成することができ, $t \leq T_0 + 1$ でも介入効果 $\tau = y_{1,t}(1) - y_{0,t}(1)$ を計算できる.

[*8] 交絡効果 (内生変数) の問題はほとんどの計量経済学の教科書で言及されているが, 特に今回のような設定のモデルに関する説明は, Cameron and Trivedi (2005), Angrist and Pischke (2009), Wooldridge (2010) あたりが参考になるだろう. 日本語文献ならば, 北村 (2009) や Hsiao (2014) の邦訳がある.

[*9] つまり, 標準的な機械学習の想定とは違い, out-of-sample で共変量分布が変化することをモデルに織り込んでいる.

2.3.1　シンセティック統制法の計算方法

もちろん, 加重平均だけで $y_{1,t}(0)$ が計算できるというのはかなり都合の良い話で, これが成立するには一定の条件がある. どのようにしてこの重み係数 $\boldsymbol{w}^*$ 求めるかを, 説明を簡単にするために最も前提を簡略化した場合で解説する. 処置群・対照群それぞれについて, 共変量 $\boldsymbol{z}_i$ と $y_{i,t}$ を重ねた行列をそれぞれ $\boldsymbol{x}_1, \mathbf{X}_1$ とする.

$$\mathbf{x}_1 := \begin{bmatrix} \boldsymbol{z}_0 \\ y_{1,1} \\ \vdots \\ y_{1,T_0} \end{bmatrix},$$

$$\mathbf{X}_0 := \begin{bmatrix} \boldsymbol{z}_2 & \boldsymbol{z}_3 & \cdots & \boldsymbol{z}_{J+1} \\ y_{2,1} & y_{3,1} & \cdots & y_{J+1,1} \\ \vdots & \vdots & & \vdots \\ y_{2,T_0} & y_{3,T_0} & \cdots & y_{J+1,T_0} \end{bmatrix}$$

ここで, (2) を満たすようにするには, $\boldsymbol{x}_1$, $\mathbf{X}_0\boldsymbol{w}$ の誤差を最小化することになる. よって, 以下のような二乗誤差最小化問題の解が, 最適な $\boldsymbol{w}^*$ となる.

$$\boldsymbol{w}^* := \arg\min_{\boldsymbol{w}} \|\boldsymbol{x}_1 - \mathbf{X}_0\boldsymbol{w}\|$$

つまり, 最小二乗法の計算と同様である. このようにして求めた $\boldsymbol{w}^*$ によって, $y_{1,t}(0)$ の不偏推定 (つまり期待値が一致すること) が可能であることまで Abadie et al. (2010) では示されている. 標準誤差についても, 強い非線形性がある場合どうするかとか, よりモデルの仮定を緩和した場合についても議論されている.

2.3.2　シンセティック統制法のまとめ

DID から SC 法へ発展したことでの変化をまとめる.

- DID では処置群と対照群の選び方について, 諸々の仮定を満足しているかについて定量的な議論が難しかった. うがった見方をするなら, 都合のいい結果を得られるように対照群を選ぶ, つまり cherry-picking もできてしまうところ, SC 法では対照群の選び方を, 介入前の結果変数が両群で同等か, といった定量的な議論に落とし込むことができるようになった.
- 交絡因子の係数 λ_t の時間変化を許容している: DID では共通トレンドを前提としていたが, 個体・時間によって変化するケースも許容されるようになった.

一方で, 新たに増えた制約としては, 以下 2 点がある.

1. $\boldsymbol{\lambda}_t$ の非特異性.
2. モデル構造の不変性. DID では単なる差分だったが, (1) というモデルに特定しているため, 期間中はこの構造が変化しない必要がある. それでも共通トレンド仮定よりは緩和されている.

2.4 ベイズ構造時系列モデル (BSTS)

ここまで, DID と SC 法とで, 2 時点だけのデータを扱う方法を紹介した. しかし, この後に出てくる causal impact は, 2 時点よりも多くの期間での介入効果を見ることができる[*10]. そのために, Causal Impact の実装ではベイズ構造時系列 (BSTS) モデルを利用している.

BSTS は, Google の研究チームに所属する Scott and Varian (2014) で提案され, R をインターフェースとした実装が用意されている.

- cran/bsts

詳しくは上記の論文か, 私が過去に書いたものを参考にしてほしい.

- http://ill-identified.hatenablog.com/entry/2017/09/08/001002

BSTS についてここでは causal impact の説明の文脈に最低限必要なぶんだけの説明だけをしておく. BSTS はいわゆる状態空間モデルである[*11]. DID フレームワークが 2 時点間の, 静的な回帰モデルであったのに対し, Causal Impact では, 自己回帰・ラグ回帰といった時間変化を考慮した「動的」なモデルを構築できるように状態空間モデルを用いている. そして, 時系列として扱うため, 3 時点以上の期間についての結果変数の推移を見ることもできるのが強みとなる[*12].

加えて, モデルの予測精度向上の為, BSTS にベイズモデル平均化 (BMA) 法という, 一種のアンサンブル学習を用意している. これは, 対照群を 1 つだけ用意するのではなく, 複数の対照群を用いた SC 法での工夫に対応していると言える[*13].

なお, 上記で挙げた解説ではモデルが数式で書かれているのでわかりにくい, という場合は, Brodersen et al. (2015) にグラフィカルモデルでの表記があるので, そちらを確認するのもいいかもしれない.

2.5 Causal Impact

Brodersen et al. (2015) が考案した causal impact フレームワークは, ここまでで紹介した DID, SC 法の考え方を踏まえて, ベイズ構造時系列モデルの特性を取り入れている. 冒頭に挙げたサイトでも causal impact を解説しているが, ここでは因果推論の文脈に沿って改めて説明する.

処置の有無について, 処置したものにもし処置しなかった場合/処置しなかったものにもし処置した場合, という反事実的な結果は観察できないという根本的な問題について, DID 法では単純に処置群と対照群の差で比較し, SC 法では, 介入前までの対照群の結果変数の加重平均で処置群の反事実的な結果を近似していた. Causal impact では, $t = 1, \cdots, T_0$ までが施策前の時系列データであり, 施策後のデータが $t = T_0 + 1, \cdots, T$ までであるとする. この場合, 施策前のデータ $\{y_1, \cdots, y_T\}$ で学習し, 事後予測分布を求める. 事後予測分布から, 施策後の期間の予測値 $\{\tilde{y}_{T_0+1}, \cdots \tilde{y}_T\}$ を求める[*14]. そして, $t = T_0 + 1, \cdots, T$ の期間について, 実績値と予測

[*10] DID や SC でも, 同じような拡張にすること自体は可能だと思う.

[*11] 時系列モデルの用語としては, 現在の状態変数が現在の他の変数と相関しうる「同時決定モデル」を構造時系列モデルと呼ぶという認識だったが, 一般には違うのだろうか?

[*12] とはいえ, DID のフレームワークでも複数時点に拡張すること自体は可能だと思う.

[*13] これは比喩的すぎてあまり正確な解説とは言えないかもしれない.

[*14] bsts の出力する予測値は確定的な値ではなく, 予測分布に基づく乱数であることに注意.

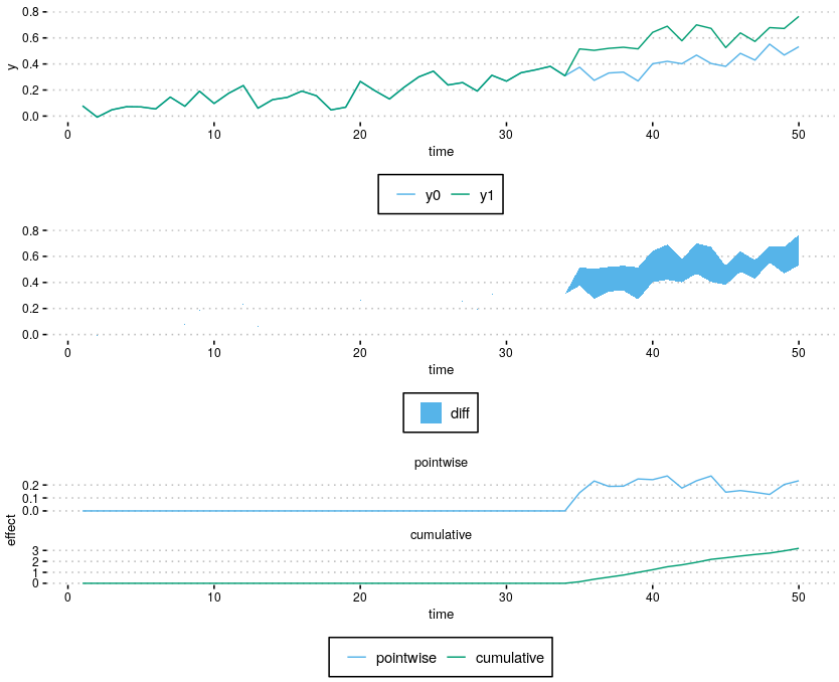

図2　causal impact フレームワーク. y_0, y_1 のプロット (上段), その差 (中段), 各時点の介入効果 ϕ_t とその累積 (下段)

値の差

$$\phi_t := y_t(1) - \tilde{y}_t(0)$$

を得る. これが DID/SC フレームワークでの介入効果に対応する. あとは期中の ϕ_t の累積や平均を計算するなどして, いろいろな方法で期間中の「施策によるインパクトの大きさ」を評価することができる (図 2).

この $\tilde{y}_t(0)$ を推定する部分は, すでに Scott and Varian (2014) で提案されているベイズ構造時系列モデルの一部でもあり, 実装上は `bsts` パッケージを利用している. 対照群ではなく処置群のデータから, $y_1(0)$ に対応するデータを生成している. つまり, 処置群の,

仮に処置のなかった場合という反事実的な結果を $y_{0,t}(0)$ をモデルに基づいて生成するという方法は SC 法に似ているが, causal impact ではもはや**対照群のデータすら不要になった**, というのが大きな違いである. A/B テストの B が必要なくなったのである.

3　DID と Causal Impact の限界

ここまでが, causal impact についての教科書的な解説であり, すぐにアクセスできる情報から読み取れる内容だ.

しかし, よく理解している人間ならば, この説明を聞いただけでは予測残差を使った古典的な異常検知とどう違うのかわからない, と考えるだろう. **ありていに言えば, 同じである.** 正常系であるデータで回帰し, 予測残差の大きさを異常値として扱うというのは, すでに様々な異常検知の教科書で紹介されている. 誰が提案したかすら書かれていないので, 相当古くからある, 素朴な発想ということなのだろう. 以前 Tokyo.R の応用セッション『再考: お買い得物件を機械学習で見つける方法 』で私が指摘したように, この方法は当てはめるデータ

が正常である (あるいはパラメータを学習するまでもなく正常系のモデルが自明である) という前提でのみ成り立つし, さらに前回の発表でも言ったように, 「標準的な機械学習」は独立変数 (特徴量) と従属変数 (目的変数) との相関関係だけを見て条件分布または単に条件期待値を構成しているだけなので, 共変量 (特徴量) x の分布が変化 (共変量シフト) したときにはうまく当てはまらない可能性がある.

もう一度 BSTS の仕組みについて考えてみる. `bsts` パッケージでできるのはデータ内での当てはまりだけを考慮する, 「従来的な機械学習」である. よって, `bsts` によって得られる予測分布関数は, t 時点までの y_t と $t+1$ 時点までの x_t の情報をもとに, y_{t+1} を予測する関数

$$p(y_{t+1} \mid \{y_s\}_{s=1}^{t}, \{x_s\}_{s=1}^{t+1})$$

である. そして causal impact ではさらに x_t で周辺化しているので, 介入後の $\tilde{y}_t(0)$ を求めるのに x_t の情報は必要ではなくなっている. これは, causal impact フレームワークでは, 対照群ではなく処置前の処置群のデータから SC 法に基づいて仮想的なシンセティック対照群を生成しているという見方もできる. また, BSTS では BMA を採用している. 一般に BMA というかアンサンブル学習は, モデルのバリアンスを低減することに関して有効だが, これもまた「標準的な機械学習」の範疇であり, 共変量シフトに対してロバストなモデルになるという保証はない[*15]. さらに, `CausalImpact` の ソースコードによれば, モデルの自動選択をしてくれるわけではない. 特に指定をしない場合, 時系列モデルには季節周期成分すらなく, **最も単純な種類のモデルの 1 つである, ローカルレベルモデルが採用される**. よって, ある程度複雑な分布をによるデータでは, causal impact 誤った結果を出してしまう可能性がある.

3.1 部分識別 (?) で考える

奥村 (2018) によれば, **部分識別** (partial identification) やバウンド識別というのは, 推定したいパラメータの理論上の上下界 (バウンド) を特定するというアプローチである. 多くの統計的/機械学習的モデルでは様々な仮定を課して計算しているが, 部分識別は逆にまったく仮定を課さない状態から仮定を追加すると, 推定値のバウンドがどう変化するかを検証するという方法論であり, 定性的な仮定と, 定量的なバウンドの識別との対応関係を知ることができる. つまり, 従来の統計学や機械学習でなされる点推定とは全く違うアプローチである. 部分識別は区間推定や信用区間とも全く異なる. これらは従来の, 強い仮定に基づく点推定まわり確率密度を与えているだけである. 一方部分識別は, ある仮定のもとでのバウンドのみを答えるもので, 確率密度に対して何か答えることを目的としたものではない. 奥村 (2018) ではこれを, データの多さに対して変わるかどうかという観点で特徴づけている. つまり, 従来の推定論に基づく区間推定は, データが増えるほど狭まっている. これはデータを仮定を課しているからであるが, 部分識別はデータの数とは関係ない (部分識別のバウンド自体推定する際に, 従来的な統計推論を利用することはある). causal impact の目的はビジネスインパクトの評価だから, 形式的に点推定することにこだわる必要はなく, 予測結果の不確実さを別の何らかの方法で評価できさえすればよいのではないかと思う[*16].

ただし, 今回のような時系列モデルに関する部分識別の研究は見つからない[*17]. なのでシャープバウンドかどうかをいちいち証明するなど, ゼロから理論的な開拓はをするのはだいぶ面倒だ, 部分識別の考え方をヒントに, 適当な仮定下で causal impact (以下, CI) がどの程度のことまで言えるのかを考える. つまり, 以降の話は部分識別に着想を受けてはいるが, 部分識別と言うほど厳密でテクニカルではない.

[*15] もしなったとしても, ごく限られた条件下, あるいは「偶然」だろう. ごく限られた条件下については, この後実例を示す.
[*16] 新規に予測モデルの開発をするかどうかの判断材料として, 最悪のケースと最良のケースを調べる, というのは私はよくやる.
[*17] 私は部分識別を専門的に研究していたわけではないので見落としている可能性もある

Causal impact は, DID が課す仮定である共通トレンド仮定が, モデルの構造パラメータの時間不変性に転嫁されたに過ぎない. ベイズ構造時系列モデルの趣旨から,「短期的トレンド不変仮定」とでも呼ぶべきだろうか. つまり, $T_0 \leq t$ 以降は $y_t(1)$ を観察することができるが, $y_t(0)$ はできない. Causal impact は, 介入直前のトレンドが, 介入直後の少なくとも一定の短い期間だけは変わらず続くという仮定に依存することで, この問題に対処している (そしてこの仮定が正しいのか検証するのは現実には困難である.). そこで, CI で要求される仮定を, 以下の 4 つだと考えてみた[*18].

I. (加法性) 介入効果は介入時の結果変数と, 反事実的な非介入時の結果変数の差である.

$$\phi_t := y_{1,t} - y_{0,t}$$

II. (BSTS の自由度) Causal Impact で表現できるのは線形状態空間モデル + ベイズモデル平均法 (BMA) の範囲である.

III. (無作為性) 共変量 X_t に対して割り当て変数 $D(t)$ は独立している.

$$D(t) \perp X_t$$

IV. (構造の定常性) 介入の前後で真のモデルの構造パラメータが変化しない.

仮定 I は, 元論文でも明示的な仮定である. 処置群と対照群の差分を介入効果とみなしているのは, DID のフレームワークと対応している.

仮定 II は, BSTS モデルによる予測値 $\tilde{y}_t(0)$ が $y_t(0)$ の良い推定値になっていることを担保するためである. ただし, このままでは介入前の $t < T_0$ に対してしか意味がない. そこで, 仮定 III, IV が必要になる.

仮定 III は, 因果推論の文脈で言えば, 強く無視できる割り当て (SIA; strongly ignorable assignment) の仮定と似ているが, これより強い仮定である. SIA の仮定は共変量を調整することで割り当て変数と結果変数の条件独立が成り立つと家庭するものであるのに対し, 共変量が割り当て変数と独立ということから, いわゆる共変量シフトが介入の前後で発生しない, という仮定である. CI は BSTS モデルの周辺事後予測分布を特徴量 X_t に対する y_t の条件分布として使用しているから, いわゆる相関モデルであり, 共変量分布の変化は考慮されていない.

仮定 IV は, 因果推論に限らずほとんどの統計モデルや機械学習アルゴリズムで暗黙に仮定されているものである. 共変量シフトは条件分布の構造はそのままで, 共変量の分布が変化することだが, 仮定 IV に反するのは条件分布そのものが変化する場合である. BSTS モデルで言うなら, 時変パラメータ以外のパラメータが変化してしまうことを意味する. ただし, 時間経過に伴う当てはまりの悪化の原因が, 観測されていない変数の変化によるもの (仮定 II に対する違反) なのか, 構造パラメータの変化によるものなのか, さらには多くの場合時間とともに減衰しうる ϕ_t によるものなのか (たとえば, 一般均衡効果とか, 同時決定効果とかいうべき現象によって起こる) の識別は, 答え合わせのできるシミュレーションでのみ検証可能なことが多く, 正解のわからない実データに当てはめる場合は困難であることが多い (例えばスイッチモデルとみなせば, パラメトリックなモデルでも表現できるため).

CI が適正に使われているかを確認する方法として, 公式ドキュメントでは, "How can I check whether the model assumptions are fulfilled?" というセクションで, 結果の sanity check について言及している.

[*18] Brodersen et al. (2015) では仮定 I, III 以外は明記されていない. これは私が考えたものであるが, 標準的な介入効果の推定で課される仮定と大きな違いはない. 厳密に考えれば, おそらく互いに排他的でかつより緩和された条件に置き換えることができそうだが, 今回説明する限りではこの程度の大雑把な設定でも問題ないと思われる.

First of all, it is critical to reason why the covariates that are included in the model (this was x1 in the example) were not themselves affected by the intervention. Sometimes it helps to plot all covariates and do a visual sanity check. Next, it is a good idea to examine how well the outcome data y can be predicted before the beginning of the intervention. This can be done by running CausalImpact() on an imaginary intervention. Then check how well the model predicted the data following this imaginary intervention. We would expect not to find a significant effect, i.e., counterfactual estimates and actual data should agree reasonably closely.

以上で言及されるチェック項目は以下の 3 点に要約される.

1. 共変量をプロットし, 介入前後で共変量が変化していないことを確認する.
2. 介入前のデータにモデルが当てはまっていることを確認する.
3. 介入後のデータ ($y_t(1)$) とモデルの予測値 $\hat{y}_t(0)$ が大きくかけ離れていないことを確認する.

(1) は明らかに, 共変量シフトが発生していないか, つまり仮定 I を確認するものである. (2) は, 仮定 III の確認である. (3) は, 介入による差を見たいのに両者に差がないことを確認する, というのは一見すると矛盾しているように聞こえるかもしれないが, 仮定 II の加法性のもとでは, 両者は平行に推移するはずで, その傾きが大きく変化することはないであろう, という前提による.

また, Brodersen et al. (2015) も指摘しているように, 施策によるインパクトがすぐに収縮する可能性もあり, その期間がどれくらいかは `CausalImpact` は答えてくれない. しかし, 多くの場合は収縮するので, 問題としてはあまり大きなものではないから, 今回はこの話は省略する.

4 実演

CI がうまく機能しない事例について, シミュレーションによる例を紹介する. シミュレーションであればデータの生成過程を我々は知っているので, CI の推定結果が正確に評価することができる.

4.1 CASE 1: チュートリアルの改変

まずは基本モデルとして, チュートリアルに対して少しだけ改変を加えて生成したデータを使う. $t = 1$ から $t = 100$ までの 100 期間として, 共変量である x_t は 1 階の自己回帰成分 (AR) と非確率的な線形トレンドの和で, 結果変数 y_t は x_t の 1 次式とする.

$$\begin{aligned} y_{0,t} =& 1.20 x_t + \eta_t, \\ x_t =& 0.90 x_{t-1} + 0.50 t + \varepsilon_t \end{aligned}$$

$T_0 = 70$ で介入し, 介入効果の大きさは一律で 10 とする. 共変量 x_t の情報を与えて介入効果 ϕ_t を推定し, 予測値, pointwise effect (ϕ_t), ϕ_t の累積値 (cumulative effect) について, 真の値と CI による推定結果の比較を 図 3 に掲載した. 破線は現実には観測できないはずのデータである. 今回はシミュレーションによる確認のため見ることができるが, 実際の分析作業ではこの値は知ることが出来ない.

CASE 1 は, ほとんどチュートリアルと同じであり, 推定もうまくいっている. 以降, このデータを基本として, どのように変更を加えた場合にが変わるのかを確認していく.

なお, `print(, "report")` や `summary()` で数値を表示することもできる. 前者は「レポート」を出力してくれるとあるが, 英語の定型文を出すだけなのであまり役に立たない.

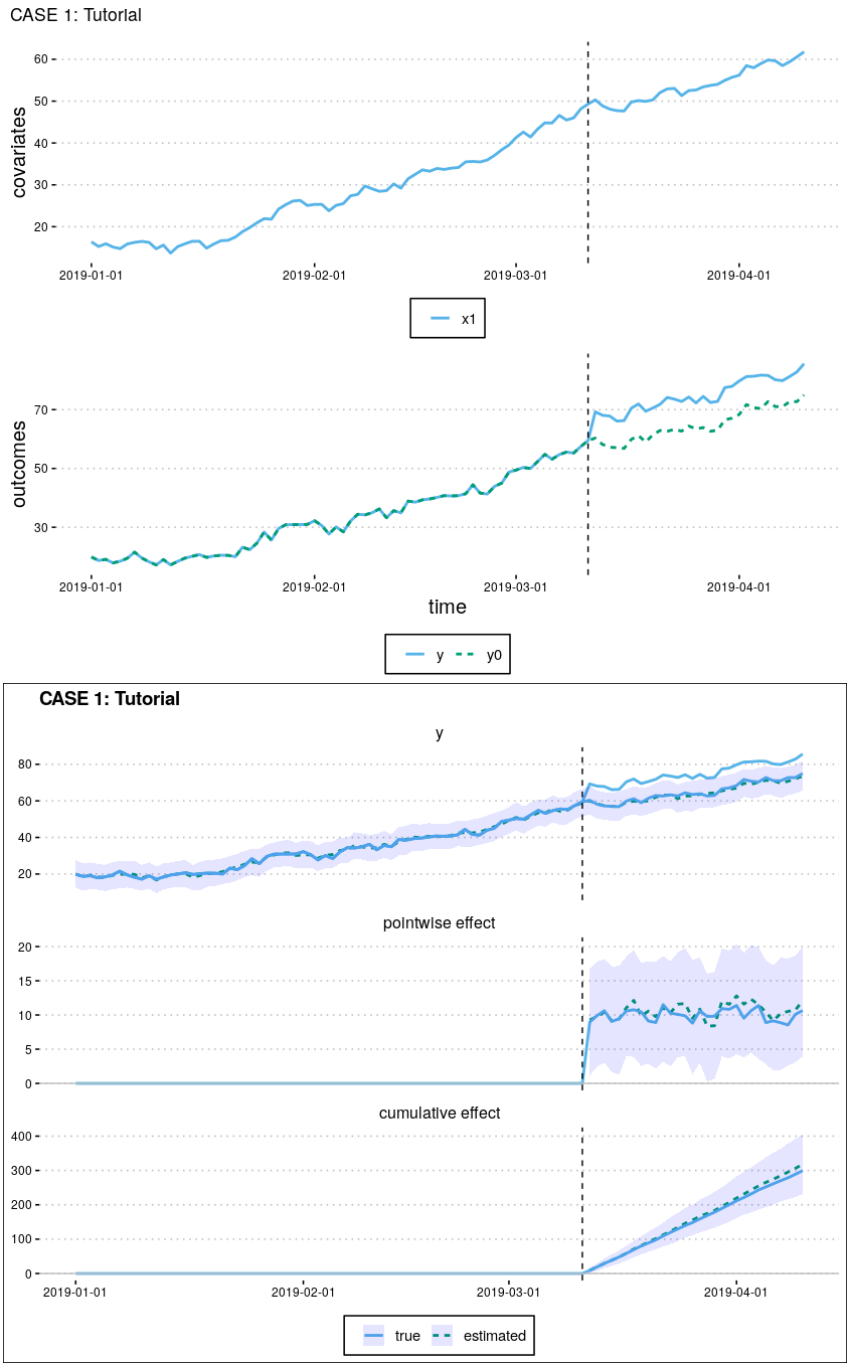

図3　CASE 1 の生データ (上段), 推定結果 (下段)

4.2 CASE 2: 介入直後の共変量シフトの発生

次に, 仮定 III の違反として, 介入の直後から共変量シフトを発生させる. 具体的には, 以下のように自己回帰成分をなくす. 介入効果の大きさはそのままである.

$$
\begin{aligned}
y_{0,t} =& 1.2x_t + \eta_t \\
x_t =& \begin{cases} 0.90x_{t-1} + 0.50t + \varepsilon_t & \text{if } t \leq T_0 \\ 0.50t + \varepsilon_t + x_{70} & \text{if } t \geq T_0 \end{cases}
\end{aligned}
$$

これは, 割り当て変数を決める t に応じて x が変化していることから, 仮定 III を違反したという想定である. 現実にありうるこのタイプの現象は, 施策を受ける側ユーザーが, 施策に反応して KPI に間接的に影響する x_t を変えたということだろうか. 例えば PV を KPI として, PV を上げようとサイトデザインを変更したところ, ユーザー側は不便に感じ, サイトの利用時間 x_t を減らした, その結果, 間接的に PV が減少した, という例が考えられる.

シミュレーションの結果は 図 4 のようになった.

理想的な状況である CASE 1 と同様, 推定結果に大きな誤差がみられない. これは, 共変量シフトに結果が左右されない特殊なケースである. つまり, シフト後の x_t は単に自己回帰成分が消えただけであり, トレンド定常まわりの平均は変わっておらず, また結果変数と共変量は単純な 1 次式の関係であり, 結局のところ期待値に関して言えば変化の影響を受けていないからである (x_t の自己回帰成分ではなく, トレンドの傾きのほうを変えれば, 結果は変わってくると予想される. 果たして本当にそうなるかは, 「読者への宿題」とする.). Case 2 は共変量の分布の変化が結果に影響しなかったが, 一般に共変量シフトによってどの程度推定が狂うかはケースバイケースである. 少なくともチェック項目 (2) の, x のプロットによる確認は必要である. 次以降でそのような状況を見てみる.

4.3 CASE 3: 非線形性

観測方程式を非線形形式に置き換える. つまり, 仮定 II を違反した想定である. x_t の状態方程式が非線形というのもありうるが, 今回は省略する.

$$
\begin{aligned}
y_{0,t} =& 0.01x_t^2 + \eta_t \\
x_t =& 0.90x_{t-1} + 0.50t + \varepsilon_t
\end{aligned}
$$

ベイズ構造時系列モデルは線形状態空間であるので, 非線形な関係から生成されるデータを当てはめると, 平均から離れるほど誤差が大きくなるはずである (テイラー展開). 現実のデータでは, CASE 2 と違い, 一般にこのような単純な一次関係だけで表現できるとは限らない. 例えば, 当初はサービスの利用者数が順調に増えていたが, 人間は有限なので, そのうち伸び悩む (いわゆるサチる). つまり, KPI の変動も単純な直線の式では表現できず, 曲線でなければ適切に表現できないかもしれない.

結果は予想通りで, 図 5 のように, pointwise effect の推定結果は時間経過に伴って 10 から離れ, 過大推定されていく.

なお, さらに介入以降の共変量シフトを加えれば, 図 6 のようにこの差はより顕著になる.

チェック項目 (3) では, 介入後の y_t と予測値が大きくかけ離れていないかを確認する, とあるが, 実際にはどれくらい離れていれば問題かと言うのは難しい. CASE 3 では, cumulative effect が加速的に増加しているよ

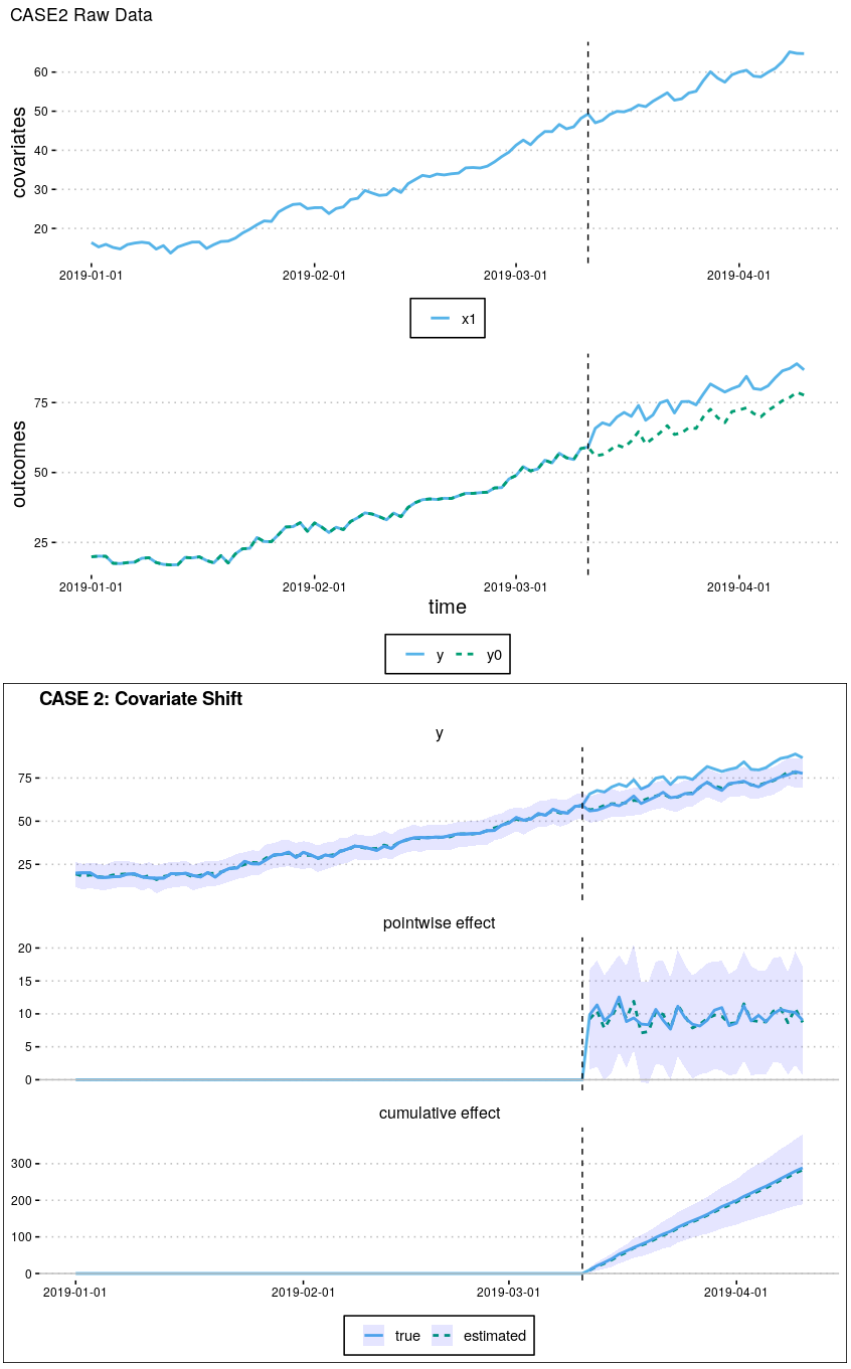

図4　CASE 2 の生データ (上段), 推定結果 (下段)

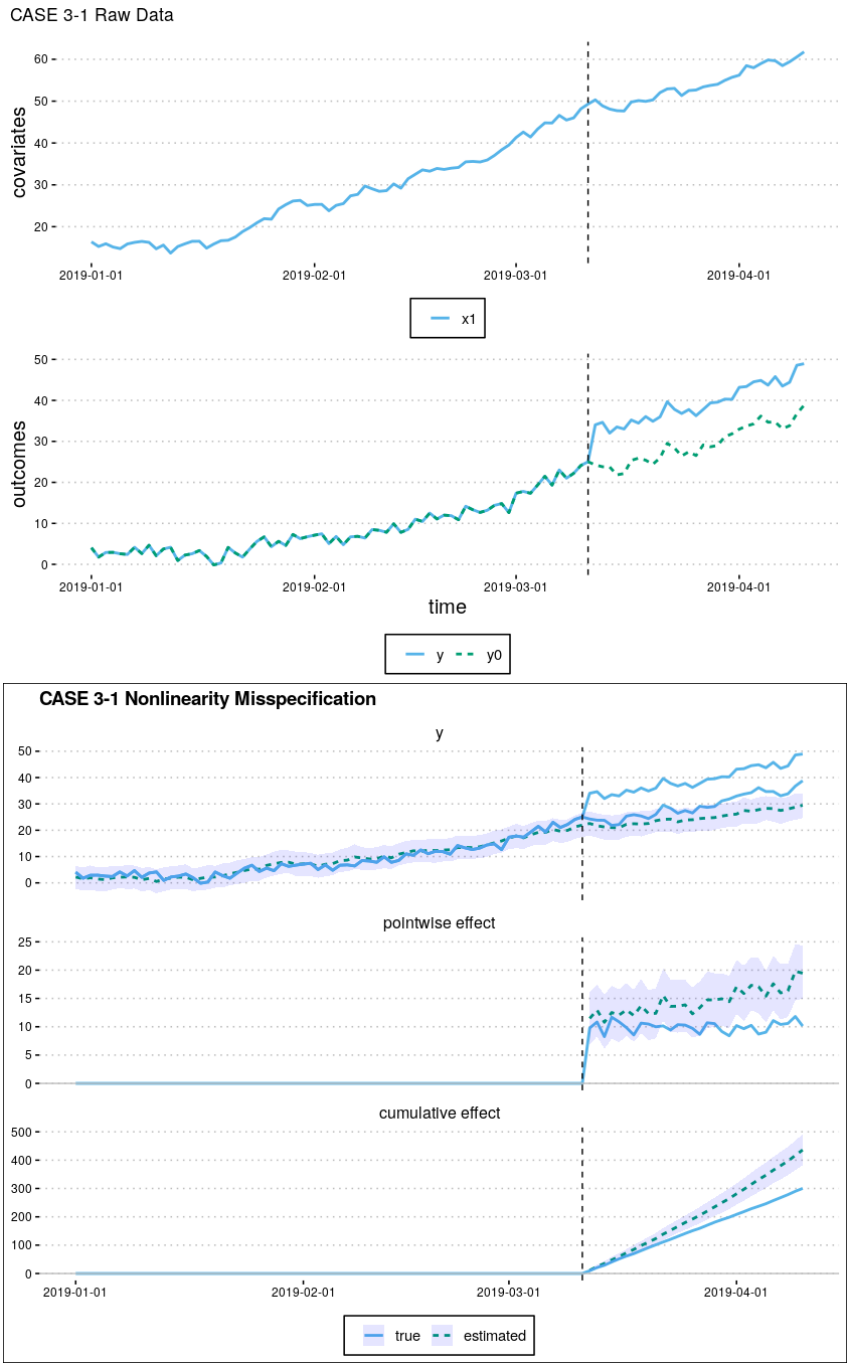

図5　CASE 3 の生データ (上段) と, 推定結果 (下段)

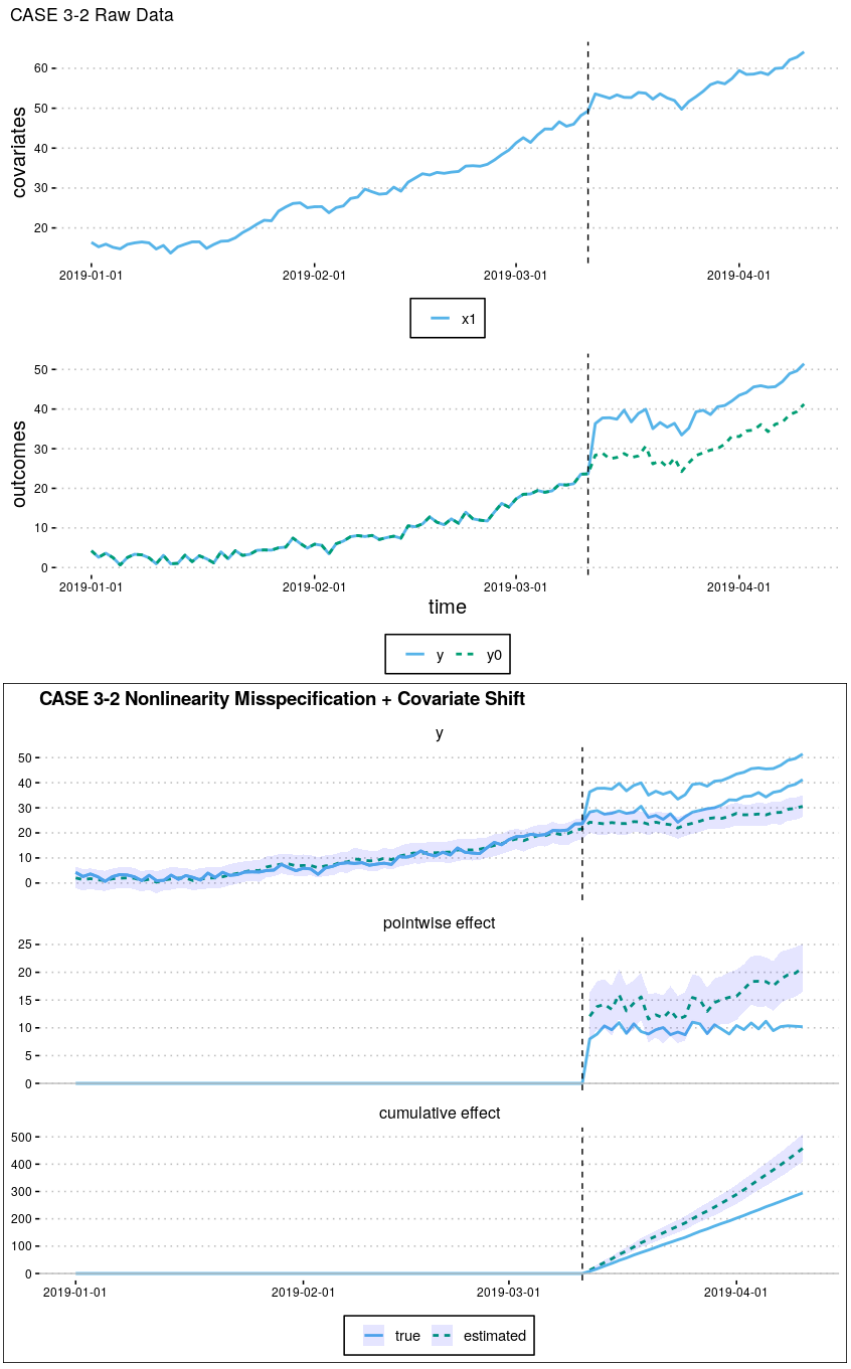

図6　CASE 3-2 の生データ (上段), 推定結果 (下段)

うに見えるので, かろうじてここから推定結果に問題があることがわかるヒントになるだろう. このような誤りを避けるには, 手動で `bsts` モデルの構造を指定したり, 残差プロットで残差に偏りがないかを確認するなど, 地道な試行錯誤が必要になる.

もう 1 つ, モデルの特定の誤りの例を提示する. y_t が以下のようにして決まる場合を考える.

$$y_{0,t} = 1.2x_t + 0.5y_{0,t-1} + 0.3y_{0,t-2} + 5\sin(2\pi t/10) + \eta$$

これはつまり, CASE 1 に 2 階の自己回帰項と, 10 期間で 1 周期になる周期成分を追加したということである. このデータと推定結果をプロットすると 図 7 のようになる. デフォルトの設定では周期成分のあるモデルは使われないため, pointwise effect の推定結果にも周期的なバイアスが発生している.

図 7 かあるいは残差をプロットすれば (図 8), データと予測値の間に周期的な誤差が発生しているのが分かる. このように, 周期性があるのが疑わしい, あるいは最初から周期性があると分かっている場合, 手動でモデルを設定することもできる. そこで, 周期成分をモデルに追加したのが 図 9 上段である. さらに自己回帰項も追加した結果が下段だが, かえって過剰適合しているようだ (この解決は読者の課題とする. 時系列モデルをどうするかという話は, 北川, 2005, 萩原他, 2018 などを参考に.).

4.4 CASE 4: 構造パラメータの変化

仮定 IV が成り立たない場合である. 答えそのものを変えているのだから, 結果は自明なのだが一応掲載しておく. CASE 4 が難しいのは, 表面的には同じような変化に見えても, その原因の識別が難しいことである.

反事実的な結果 $y_{0,t}$ の構造だけが変化するケース, $y_{1,t}$ の構造だけが変化するケース, 両方の構造がそれぞれ同様に変化するケースの 3 通りがある.

CASE 3 とは対照的に, CASE 4 ではある施策によってユーザーの行動パターンの KPI に直結する部分が変化するような場合を意味する. 例えば, 自社が取り扱っているある商品 (あるいはサービス) の 1 つを割引した場合. 割引されたものの需要は増えるかも知れないが, 消費者の財布の中身は有限なので, 代わりに他の割引しなかった他の商品を買い控えるかもしれない. すると, 売上が単純に増加するとも言えなくなる. あるいは,「増税の駆け込み需要」も, 増税という施策の直後ではなく直前に行動が変化するという違いはあるが, これと同類の問題と言える. このような場合は, どう介入効果を測定するかというよりも, 施策の効果をどう最大化するかという問題のほうが重要だろう.

このケースで結果が歪むのは自明ではあるが, 一応数値実験の例を掲載する. 以下のように, 介入後に構造が変化するケースである.

$$y_{1,t} = \begin{cases} 1.2x_t + \eta_t & \text{if } t \leq T_0 \\ 1.0x_t + \phi_t + \eta_t & \text{if } t > T_0 \end{cases}$$

結果は以下の 図 10 のようになる. 介入効果を「真の値」よりも過小評価してしまっているが, 一方で介入と同時に発生した y_t の構造変化は, 介入効果の一部ではないかとも考えられる. その点から言えば, これは仮定 I の違反でもある.

4.5 CASE 5: 非定常分布

時系列データはしばしば非定常である. 反例はなるべくシンプルなものを示すほうが良く, モデルの特定を誤った場合はすでに CASE 3 で示しており蛇足感はある. しかし時系列データというと早押しクイズのように

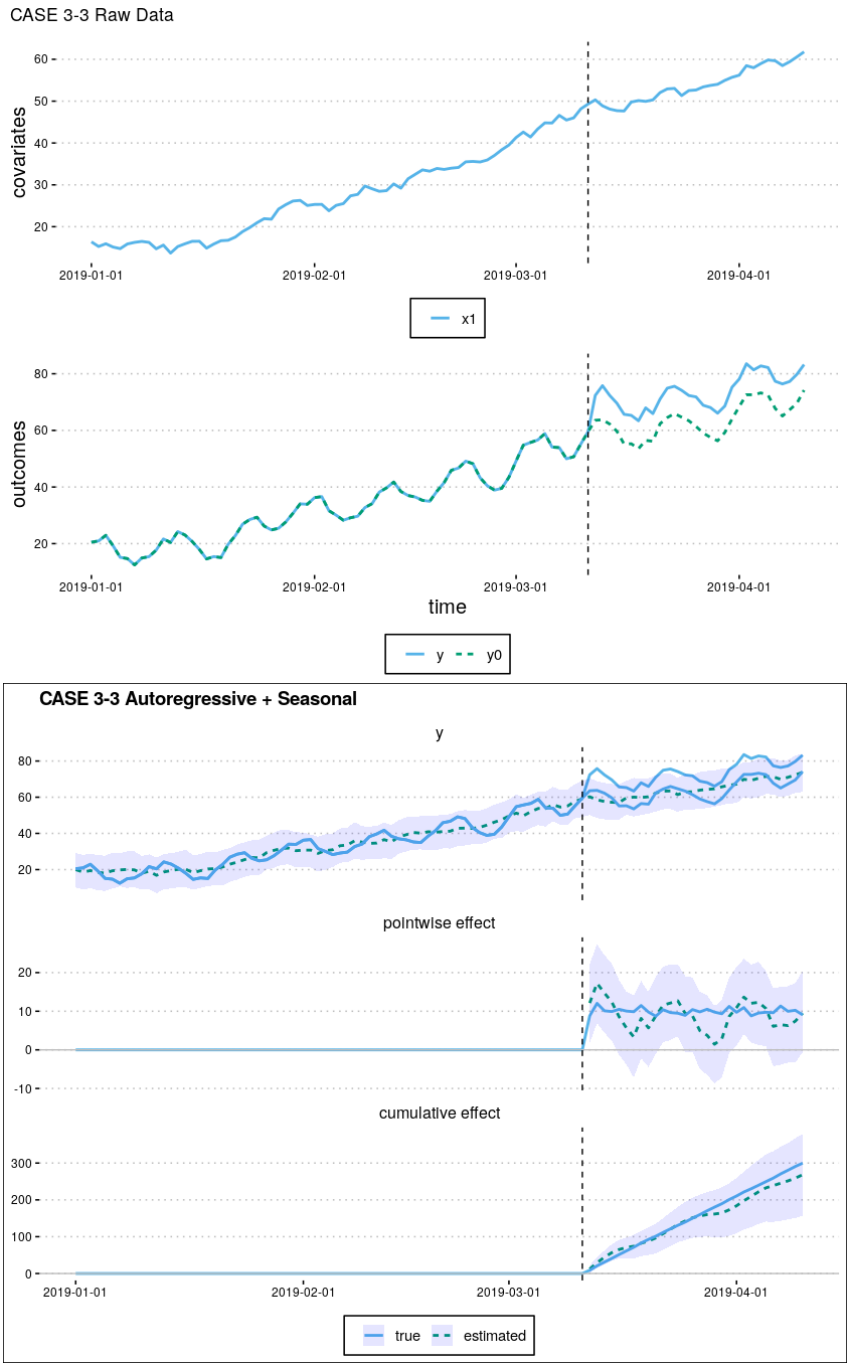

図7　CASE 3-3 の生データ (上段), 推定結果 (下段)

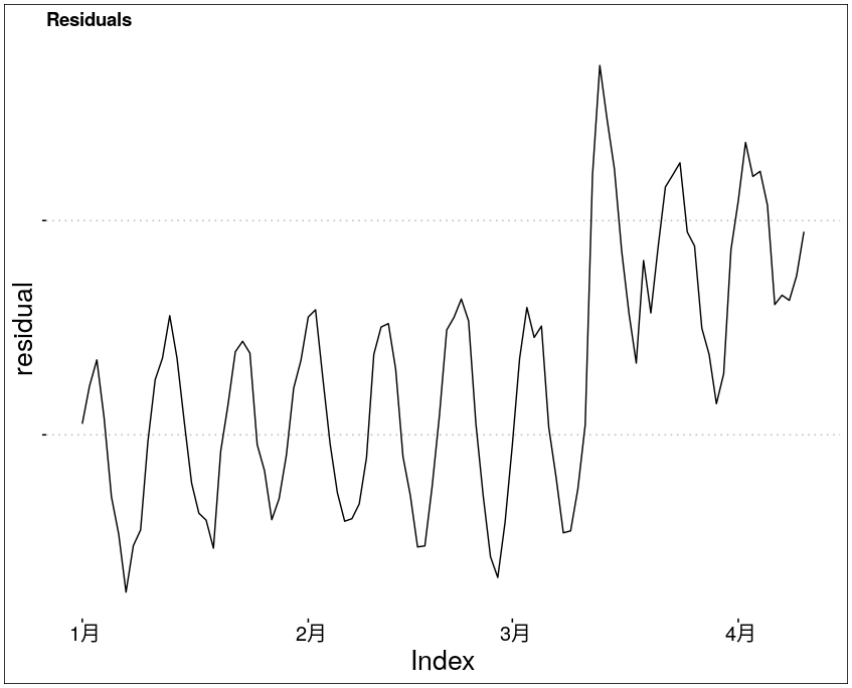

図8　残差の時系列プロット

非定常の話をしたがる人は一定数いるし, 非定常な分布の生データは時系列でははよくあるので実用例として書いておく.

既に書いたように, CI はデフォルトでローカルレベルモデルを使用するので, 単純なランダムウォーク成分がある程度では大きな誤差にならない. そこで, CASE 1 の y_t にランダムウォーク成分を加えた. 標準的な状態空間表現で書けば,

$$
\begin{aligned}
y_{0,t} =& 1.20x_t + \alpha_t + \eta_t, \\
x_t =& 0.90x_{t-1} + 0.50t + \varepsilon_t, \\
\alpha_t =& \alpha_{t-1} + \zeta_t
\end{aligned}
$$

となるデータを生成した. ζ_t はホワイトノイズである. CI での推定結果は図 11 の通りである. 残念ながら推定と実際の値にはかなり差がある. 非定常な分布は一般に定常な分布に比べ予測誤差が非常に大きくなるので, 分布を正しく特定できたとしても誤差が大きくなることが多い. CI のチェック項目 (1) や (3) に引っかかることが多いので問題は発見しやすいが, 効果的な対処法がないこともある.

5　結論

今回は, 因果推論の基本的な考え方を復習した上で, 因果推論に理論的根拠を置く causal impact フレームワークの特性について考えた.

Causal impact はユーザーにとって非常に簡単に操作できるフレームワークである. しかし, 今回の複数のケースで挙げたように, 必ずしも結果が信頼できるとは限らない. 推定・推論とは仮説と結果の検証の繰り返しであり, causal impact だろうが何であろうが, スタートボタンを押して後はぼんやり待っているだけで絶対うまくいくフレームワークなどというものは存在しない. 常に自分自身に対して批判的な姿勢で推論の反証可能

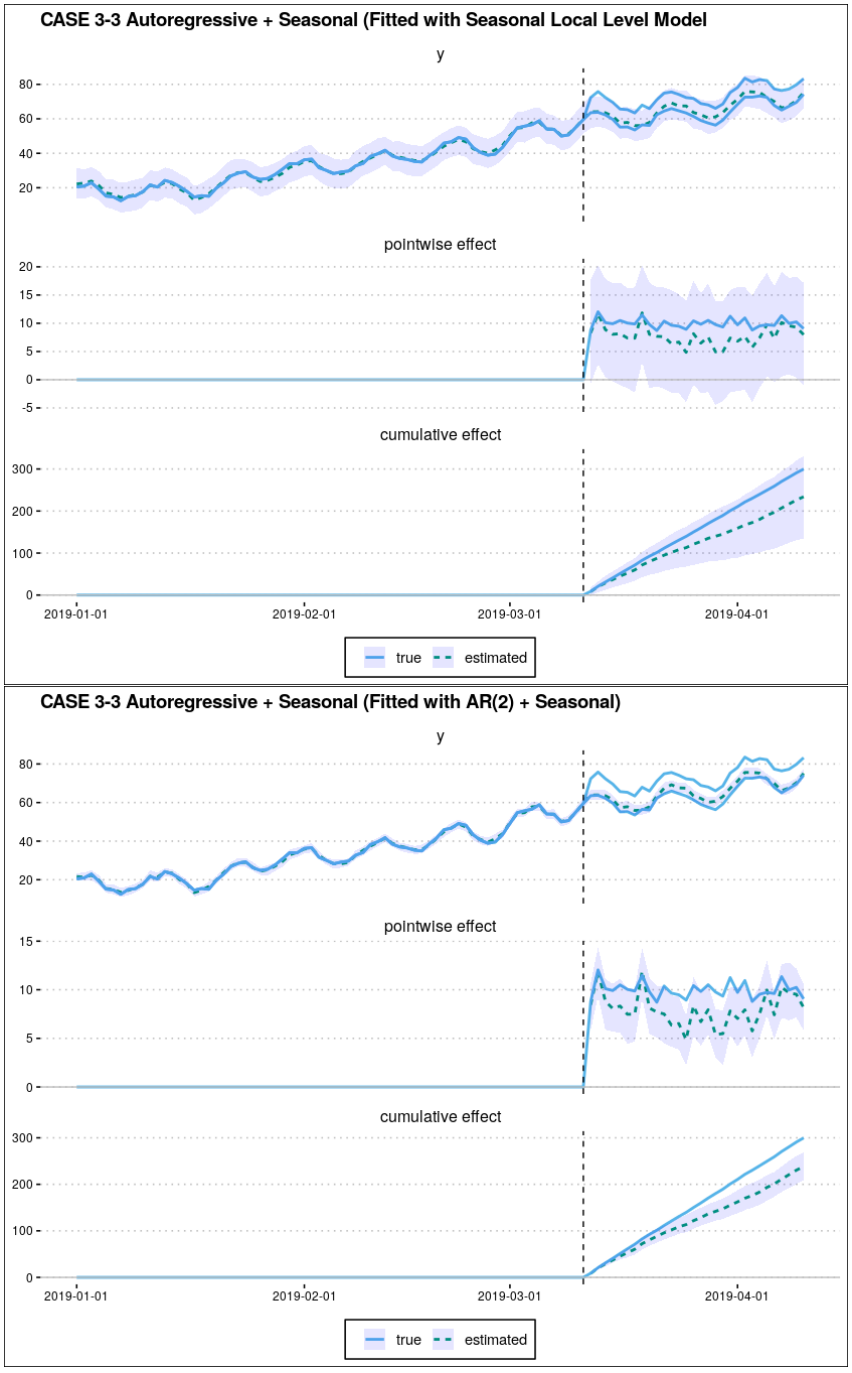

図9　モデルを手動設定した場合の CASE 3-3 の推定結果

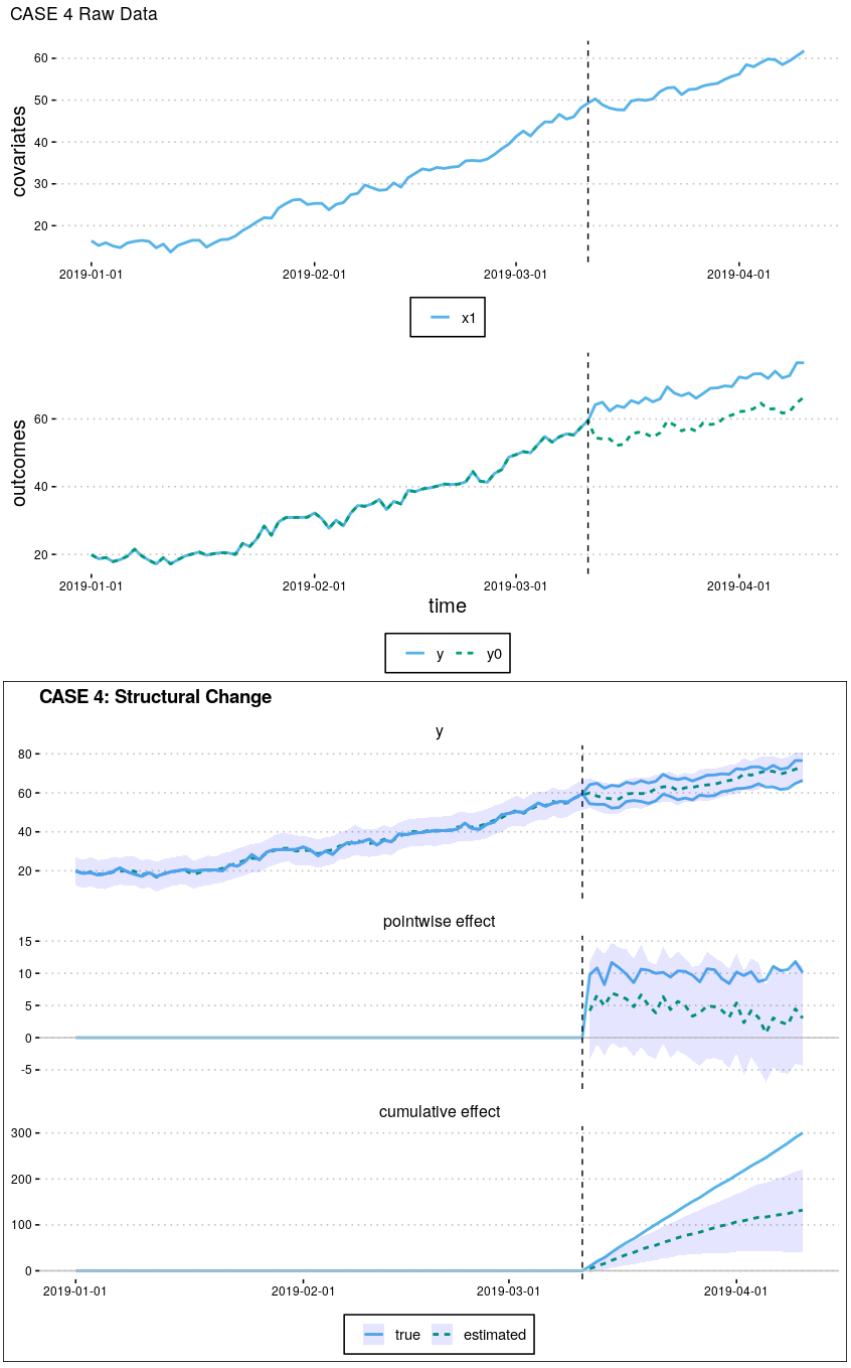

図10　CASE 4 の結果

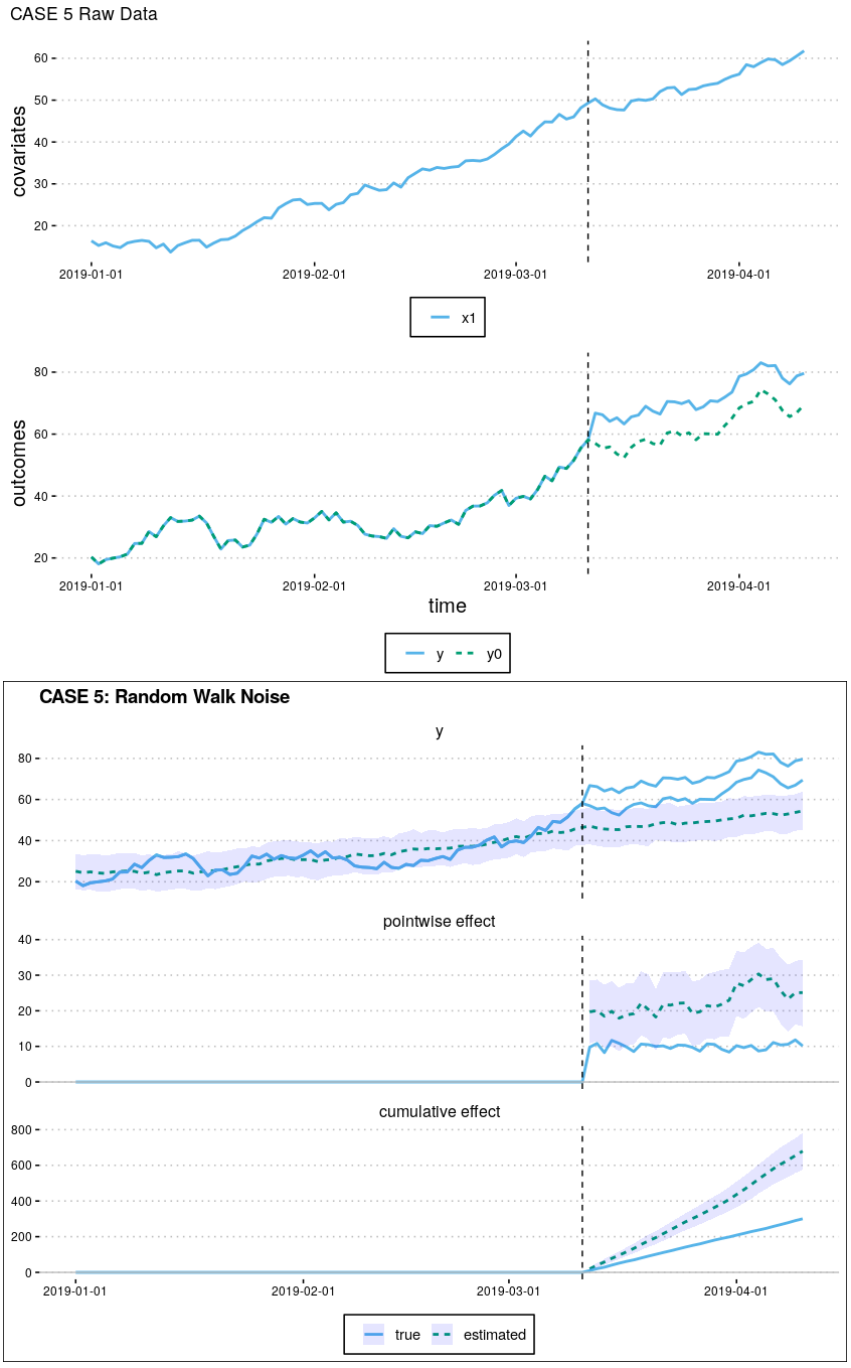

図11　CASE 5 の結果

性について考える必要がある. 学術研究とビジネスでは目的は違えど, 意味のある結果が要求されるのは同じだろう.

(もう少しおもしろい発展的な話題を考えられないかと前々回の Tokyo.R 後 2 週間ほどあれこれいじってみたが, わりとありきたりで教科書的な内容になってしまったので 2 ヶ月も放置していた. 散発的に思いつくことを書いたせいで, むしろ内容を他人が読める文にまとめるのに時間がかかった. あとはてなブログの構文に書き換えるのがとてもしんどい. 今度から pdf だけ貼り付けてしまいたい.)

補足: プログラムについて

Causal impact は R の `CausalImpact` パッケージで実装されている[*19]. ヘルプではデータの入力を `data.frame` でも `xts` 等の時系列データオブジェクトでも受け付けているのだが, 時間インデックスも同時に扱える時系列オブジェクトのほうが結果の加工が楽なので, そちらを使うことにした. 近年は tidyverse により配列構造のデータ (`data.frame`) の扱いが非常に簡単になったが, `zoo`, `xts` といった時系列データを扱うオブジェクトの操作は昔から変わっていない. 今となっては R の古い構文はとても煩雑で面倒に感じる. 時系列データオブジェクトを tidy に扱うパッケージには, `tsibble`, `tidytime` という 2 つがあるが, 今回は `tsibble` を使った. r-wakalang で教えてもらったところによれば, `tsibble` は `tidyverts` プロジェクトの一環である.

- `https://github.com/tidyverts`
- `https://tsibble.tidyverts.org`

使い方は簡単で, `data.frame` を `as_tsibble()` で変換するだけでよい. 時系列データのため, 時間インデックスが必須で, 任意でグループ変数も指定する必要があるが, `index_by()` と `summarise()` を併用することでサンプリングレートを変更 (時間単位から日単位, 週単位, など) できたり, `slide()`, `tile()`, `stretch()` などで rolling 集計ができるなど, 時系列データの特有の処理も簡単である. `filter`, `select()`, `mutate()`, `gather()`[*20] などの `dplyr`/`tidyr` でおなじみの関数もほぼ同じ感覚で使うことができるので, 時刻処理機能の充実した `lubridate` などと併用すればだいぶ簡単になるだろう. また, `forecast` パッケージの `acf` 関数などにもそのまま入力として与えられる. ただし, **現状 `ts` には対応しているものの, `zoo`/`xts` との相互変換はサポートしていない**. また, 関数がコンフリクトするので, `tsibble` と従来の `zoo` などを使った処理を併用するようなプログラムを書く場合は注意が必要である.

今回は CI にデータを入力するため, `tsibble` と `zoo` を相互に変換する簡易的な関数を用意してみた (プログラム 1).

また, モデルを手動で指定して CI に適用する方法についても補足しておく. モデルを変更したい場合, `CausalImpact()` の `model.args` である程度指定できる. 例えば, CASE 3-3 のように周期成分を追加したい場合, `nseasons=` を与えることができる. 詳しい説明はヘルプにあるのでそちらを参照されたい.

model.args で可能な範囲を超える場合, 自分で状態空間モデルを指定し, `bsts()` で計算した結果を `CausalImpact()` に与える必要がある. まず, `bsts` をでモデルにフィットさせる際は, 介入以降の結果変数

[*19] 非公式だが, Python でも causal impact が作られている (`https://github.com/google/CausalImpact/issues/14`). ただし, フレームワークを真似ているだけで, 時系列モデルの部分は古典的な statsmodels のものを流用しているだけなので, BSTS は利用できない.

[*20] 時間 index を - しなくとも残してくれるので便利

プログラム 1 tsibble/zoo の相互変換

```
tsibble2zoo␣<-␣function(x){
␣␣stopifnot(inherits(x,␣"tbl_ts"))
␣␣zoo::read.zoo(dplyr::select(x,␣index_var(x),␣everything()))
}

zoo2tsibble␣<-␣function(x){
␣␣stopifnot(inherits(x,␣"zoo"))
␣␣as_tsibble(fortify.zoo(x),␣index=Index)
}
```

プログラム 2 CausalImpact とのモデル変更の例

```
data3_3_input␣<-␣data3_3␣%>%␣mutate(y=if_else(Index␣%within%␣pre_span,␣y,␣NA_
    real_))
ss3_3␣<-␣list()␣%>%␣AddAr(y=data3_3_input$y,␣lags=2)␣%>%
␣␣AddSeasonal(y=data3_3_input$y,␣nseasons=10)
model3_3␣<-␣bsts(formula=y~x1,␣state.specification=ss3_3,␣timestamps=data3_3$
    Index,
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣data=data3_3_input␣%>%␣as_tibble,␣niter=1000,␣seed=42)

impact3_3_mod␣<-␣CausalImpact(bsts.model=model3_3,␣post.period.response=filter(
    data3_3,␣Index␣%within%␣post_span)$y,␣seed=42)
```

を全て NA に置き換える必要がある. そして CausalImpact に bsts オブエジェクトと, 介入後の結果のベクトルを与える. 例えば CASE 3-3 では, プログラム 2のように書いている.

今回使ったプログラム全文は以下で公開している.

- https://github.com/Gedevan-Aleksizde/20190728_DID

参考文献

Abadie, Alberto and Javier Gardeazabal (2003) "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, Vol. 93, No. 1, pp. 113–132, February, DOI: 10.1257/000282803321455188.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010) "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the*

American Statistical Association, Vol. 105, No. 490, pp. 493–505, June, DOI: 10.1198/jasa.2009.ap08746.

Angrist, Joshua D and Jörn-Steffen Pischke (2009) *Mostly Harmless Econometrics: An Empiricist's Companion*: Princeton University Press, retrieved from *here*,（大森義明・小原美紀・田中隆一・野口晴子訳,『ほとんど無害な計量経済学 –応用経済学のための実証分析ガイド–』, NTT 出版, 2013 年）.

Brodersen, Kay H., Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott (2015) "Inferring Causal Impact Using Bayesian Structural Time-Series Models," *The Annals of Applied Statistics*, Vol. 9, No. 1, pp. 247–274, March, DOI: 10.1214/14-AOAS788.

Cameron, AC and PK Trivedi (2005) *Microeconometrics: Methods and Applications*, Cambridge: Cambridge University Press, DOI: 10.1017/CBO9781107415324.004.

Hsiao, Cheng (2014) *Analysis of Panel Data*, Cambridge: Cambridge University Press, 3rd edition, DOI: 10.1017/CBO9781139839327,（国友直人訳,『ミクロ計量経済学の方法 –パネルデータ分析』, 東洋経済新報社, 2007 年）, 邦訳 ISBN: 978-4-492-31384-8.

Rubin, Donald B. (1974) "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.," *Journal of Educational Psychology*, Vol. 66, No. 5, pp. 688–701, DOI: 10.1037/h0037350.

——— (1990) "Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies," *Statistical Science*, Vol. 5, No. 4, pp. 472–480, November, DOI: 10.1214/ss/1177012032, On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.

Scott, Steven L. and Hal R. Varian (2014) "Predicting the Present with Bayesian Structural Time Series," *International Journal of Mathematical Modelling and Numerical Optimisation*, Vol. 5, No. 1/2, p. 4, DOI: 10.1504/IJMMNO.2014.059942.

Wooldridge, Jeffrey M (2010) *Econometric Analysis of Cross Section and Panel Data, Second Edition*: The MIT Press, 2nd edition, retrieved from *here*.

奥村綱雄 (2018)『部分識別入門: 計量経済学の革新的アプローチ』, 日本評論社, 東京, retrieved from *here*, OCLC: 1057483434.

北川源四郎 (2005)『時系列解析入門』, 岩波書店.

北村行伸 (2009)『ミクロ計量経済学入門』, 日本評論社, 東京, retrieved from *here*, OCLC: 308167431.

星野崇宏 (2009)『調査観察データの統計科学 –因果推論・選択バイアス・データ融合』, 岩波書店.

——— (2016)「統計的因果効果の基礎–特に傾向スコアと操作変数を用いて」,『岩波データサイエンス Vol. 3』, 岩波書店, 62–90 頁.

森田果 (2014)『実証分析入門: データから「因果関係」を読み解く作法』, 日本評論社, 東京, OCLC: 881836881.

山口慎太郎 (2016)「差の差法で検証する「保育所整備」の効果」,『岩波データサイエンス Vol. 3』, 岩波書店, 112–128 頁.

萩原淳一郎・瓜生真也・牧山幸史 (2018)『基礎からわかる時系列分析 = Understanding Time Series Analysis with R: R で実践するカルマンフィルタ・MCMC・粒子フィルタ』, 技術評論社, 東京, OCLC: 1035562613.

黒澤昌子 (2005)「積極労働政策の評価–レビュー」,『フィナンシャル・レビュー』, 第 77 巻, 197–220 頁, 7 月, retrieved from *here*.