『What if』の視点で読み解くPROC CAUSALTRT：IPWと標準化の理解と注意点

○峯脇 颯汰[1]、浦狩 保則[1]

（[1]シミック株式会社　データサイエンス事業本部）

Understanding PROC CAUSALTRT Through the “What If” Lens: IPW, Standardization, and Considerations

Sota Minewaki[1], Yasunori Uragari[1]

[1]Data Science Division, CMIC Co., Ltd.

要旨

観察研究では研究者は治療割付を制御できないため、治療群間の背景因子の不均衡が交絡を生じ、平均治療効果（ATE）の推定にバイアスをもたらす。交絡を調整しATEを推定する手法として逆確率重み付け（IPW）と標準化がある。これらの手法は因果推論の名著である What if で中心的に扱われており SAS の PROC CAUSALTRT で簡単に実装できる。CAUSALTRT では IPW を実装する際、method＝IPW/IPWR/IPWS の3つのオプションがある。IPW/IPWR はそれぞれ What if の Horvitz-Thompson 推定量、Hajek 推定量に対応する。これら 3 つのオプションの違いが ATE の推定精度に与える影響と実務における使い分けの指針について解説する。一方、CAUSALTRT で標準化を実装する際、治療群ごとにアウトカムモデルを想定するのに対し、What if は単一のアウトカムモデルを想定する。しかし、単一モデルと群別モデルの使い分けについて整理されていない。これらのモデルの違いや単一モデルにおける積項が ATE の推定精度に与える影響について解説し、実務における指針を示す。

キーワード：因果推論、IPW、標準化、CAUSALTRT

1. 序論

観察研究は治療や介入の「効果」を推定することを主な目的として行われる。しかし、治療割り付けを制御できないため、治療群間で背景因子の不均衡により交絡が生じる。そのため、群間のアウトカムを単純比較するだけでは治療の効果を正しく推定できない。Hernán and Robins の Causal inference: What if では「もし集団全体が治療群ならどのような結果になるか」「もし集団全体が対照群だったらどのような結果になるか」という反事実的な視点のもと、観察データから治療の因果効果を識別・推定する方法を体系的に説明している[1]。集団全体の治療の因果効果を推定する方法として IPW と標準化がある。これらの手法は What if で中心的に扱われており、SAS9.4 SAS/STAT14.2 から PROC CAUSALTRT で簡単に実装できるようになった[1,2]。本稿の目的は CAUSALTRT で扱える IPW と標準化を What if の視点から読み解くことで理論

と実装のつながりを明確にすることにある。また、CAUSALTRT で扱える IPW の手法の推定精度の違いや What if で説明されている標準化と CAUSALTRT の標準化（CAUSALTRT 上では正しくは回帰調整（Regression adjustment）と表現される）の推定精度の違いについても数値実験で検討し、これらの手法を扱う上での注意点を解説する。

2. 諸前提

本稿では、ベースライン時点で定義される二値治療A 、ベースライン共変量L 、およびアウトカムYからなる 1 時点の治療を考える。ここで$A=1$は治療あり、$A=0$は治療なしを表す。観察データでは、治療群と対照群で共変量分布が異なることが多く、観測されたアウトカムの差をそのまま治療の効果として解釈できない。因果効果を議論するため、潜在アウトカムを用いて治療の因果効果を定義し、そのうえで観測データからその量を識別するための条件を整理する必要がある。

2.1. 潜在アウトカムと ATE

因果推論では、各症例について「治療を受けた場合のアウトカム」と「治療を受けなかった場合のアウトカム」を考える。治療を受けた場合の潜在アウトカムをY^1、治療を受けなかった場合の潜在アウトカムをY^0とすると個体レベルの因果効果はY^1-Y^0として定義される[1]。しかし、同一症例のY^1とY^0は同時に観測できないため症例ごとの因果効果は求められない。そのため、関心の対象は母集団の平均的な因果効果となることが多い。母集団の平均的な因果効果である ATE は$E[Y^1]-E[Y^0]$と定義される。ATE は「対象集団全体が治療を受けた場合の平均アウトカム」と「対象集団全体が治療を受けなかった場合の平均アウトカム」の差として解釈できる。

2.2. 識別条件

実際はY^1とY^0は観測できないため、観測データから ATE を推定するために以下の 3 つの識別条件を仮定する必要がある[1]。これらの識別条件をすべて満たすことで観察研究を条件付きランダム化比較試験とみなせ、観測データYから ATE を推定することができる。

1. （条件付き）交換可能性

   観察研究では治療割り付けが無作為でないため、一般的に交換可能性$Y^a \perp A$は成立しない。そのため、共変量Lで条件づけた条件付き交換可能性$Y^a \perp A \mid L$を仮定する必要がある。

2. 一致性

   観察研究では、研究者が想定する治療の内容が曖昧だと対応する潜在アウトカムY^aも不明瞭になるため比較する治療aの定義を明確にする必要がある。そのうえで、実際に治療$A=a$を受けた症例は観測アウトカムYが潜在アウトカムY^aに一致するという一致性の仮定が必要である。一致性は$Y=AY^1+(1-A)Y^0$と表される。

3. 正値性

   条件付ける共変量Lの各水準において各治療水準が観測されることである。正値性は$P(A=a \mid L=l)>0$と表される。

2.3. IPW

2.3.1. What if で説明される IPW の役割

IPW の役割は観測集団を重みづけることで、あたかも治療がランダムに割り付けられ交換可能性が成り立つと考えられる疑似集団を作り出すことにある[1]。図 1 の右にある疑似集団を比較することで因果効果を推定できるようになる。つまり、疑似集団内では関連を因果とみなせる。

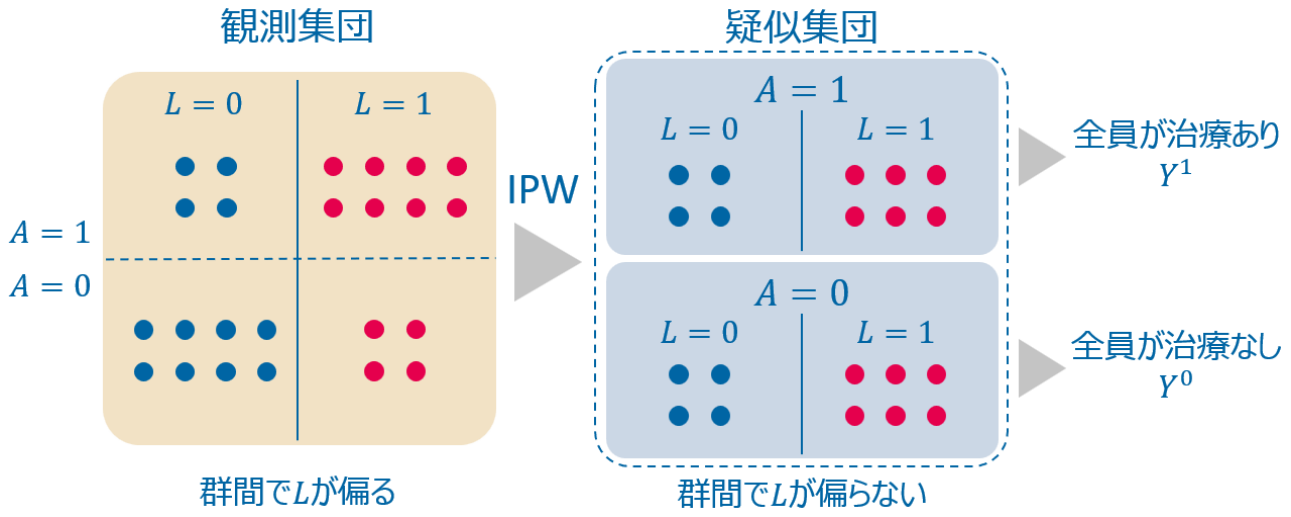


図 1：IPW の役割

各症例の重みは割付確率（傾向スコア）の逆数である$W^A = 1/P(A \mid L)$で定義され、傾向スコアを$e(L) = P(A = 1 \mid L)$とすると治療群では$1/e(L)$、対照群では$1/(1 - e(L))$が重みとなる。この重みを用いることで ATE は以下の式で求められる。

$$ATE = E[Y^1] - E[Y^0] = E\left[\frac{AY}{e(L)}\right] - E\left[\frac{(1-A)Y}{1-e(L)}\right]$$

What if ではW^Aにくわえて安定化重み$SW^A = P(A)/P(A \mid L)$も紹介されている。安定化重みを用いても交換可能性が成り立つと仮定できる疑似集団を作り出すという役割は変わらない。安定化重みを用いると極端な重みをもつことを防ぐため、有限標本における推定の不安定性を軽減できる。

2.3.2. Horvits-Thompson 型推定量、Hájek 型推定量

What if では ATE の推定において Horvits-Thompson（HT）型推定量と Hájek（H）型推定量を紹介しており、治療が二値の場合において$f(A|L)$をW^AまたはSW^Aとするとそれぞれ以下の式で表せる[1,3,4]。

$$\text{HT 型推定量：} \hat{E}[Y^a] = \hat{E}\left[\frac{\mathbb{I}(A=a)Y}{f(A|L)}\right] = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i=a)y_i}{\hat{f}(A|L_i)}$$

$$\text{H 型推定量：} \hat{E}[Y^a] = \left(\hat{E}\left[\frac{\mathbb{I}(A=a)}{f(A|L)}\right]\right)^{-1}\hat{E}\left[\frac{\mathbb{I}(A=a)Y}{f(A|L)}\right] = \left(\sum_{i=1}^{n}\frac{\mathbb{I}(A_i=a)}{\hat{f}(A|L_i)}\right)^{-1}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i=a)y_i}{\hat{f}(A|L_i)}$$

HT 型推定量は単純な重みづけ平均であるのに対し、H 型推定量は重みの総和で正規化した推定量であることが分かる。実務上は、H 型推定量のほうが有限標本で安定しやすく推定値が自然な範囲に収まりやすいという利点がある。HT 型推定量を用いる際はSW^Aを用いるとバイアスが生じるので注意が必要である。

2.3.3. CAUSALTRT における IPW 系の method と What if との対応

CAUSALTRT では、ATE を推定するための IPW 法として method=IPW / IPWR / IPWS が実装されている。重みはW^Aのみを想定しており、以下のように psmodel ステートメントで傾向スコアモデルを指定する。

```
proc causaltrt data = simdata method = IPW/IPWR/IPWS;
   psmodel A(event='1') = L;
   model Y / dist=normal;
   output out=result_IPW
run;
```

What if と CAUSALTRT の各 method は次のように対応付けることができる。

- method=IPW：What if の HT 型推定量に対応
- method=IPWR：What if の H 型推定量に対応
- method=IPWS：What if では紹介されていないが、ATE 推定量の大標本分散を最小化するために HT / H 型推定量を拡張したものであり、交換可能性が成り立つと仮定できる疑似母集団を作り出すという役割は変わらない[5]。また、H 型推定量と同様に重みの総和で正規化した推定量である。

HT 型推定量と H 型推定量の正規化の有無と使用できる重みの対応を表 1 に示す。

表 1：HT 型推定量と H 型推定量の正規化の有無と使用できる重みの対応

	$W^A = 1/P(A \mid L)$	$SW^A = P(A)/P(A \mid L)$
Horvits-Thompson 型（正規化なし）	What if のW^A CAUSALTRT の method=IPW	
Hájek 型（正規化あり）	CAUSALTRT の method=IPWR	What if のSW^A

CAUSALTRT で実装できる IPW 、IPWR、IPWS の ATE 推定精度を数値実験にて様々な条件下で比較する。

2.4. 標準化

2.4.1. What if で説明される標準化の役割

標準化の役割は各共変量組合せの層ごとの条件付き平均を集団全体に拡張して ATE を求めることにある[1]。離散的な共変量Lを考えると潜在アウトカムの平均は以下のように表せる。

$$E[Y^a] = \sum_l E[Y \mid A = a, L = l]\, P(L = l)$$

$$= E_L[E[Y \mid A = a, L]]$$

つまり、各層$L = l$における条件付き平均$E[Y \mid A = a, L = l]$を求め、各層の出現率$P(L = l)$で重みづけ平均することで集団全体が治療aを受けた場合の潜在アウトカムの平均を得る。

共変量が高次元または連続変数を含む場合、各層における条件付き平均をノンパラメトリックに推定するのは非現実的である。そのため、条件付き平均に対して回帰モデルを仮定し、その推定値を用いて標準化する。これが parametric g-formula である[1]。手順としては、まず観測データセットを用いてアウトカム回帰モデル$E[Y \mid A = a, L = l]$を推定する。次に観測データセットを共変量はそのままにして、「すべての症例が治療群」、「すべての症例が対照群」であるデータセットをそれぞれ用意する。これらのデータセットに対してアウトカム回帰モデルの推定式$\hat{E}[Y \mid A = a, L = l]$を用いて、各症例の潜在アウトカムの予測値$\hat{Y}^1$、$\hat{Y}^0$を得る。最後に$\hat{\mu}_1 = (1/n)\sum_{i=1}^n \hat{Y}_i^1$、$\hat{\mu}_0 = (1/n)\sum_{i=1}^n \hat{Y}_i^0$を算出し、$\widehat{ATE} = \hat{\mu}_1 - \hat{\mu}_0$を得る。これは条件付き平均アウトカムを plug-in して g-formula を評価する方法であるため、plug-in g-formula とよばれる[1]。

2.4.2. CAUSALTRT における REGADJ と What if との対応

CAUSALTRT では ATE を推定するための標準化に対応する方法として method=REGADJ が実装されている。以下のように model ステートメントで推定するアウトカム回帰モデルを指定する。

```
proc causaltrt data = simdata method = REGADJ;
   psmodel A(event='1');
   model Y = L1-L10 / dist=normal;
   output out=result_IPW
run;
```

CAUSALTRT では群ごとにアウトカム回帰モデル$\hat{Y}^1 = g^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}}_{\boldsymbol{t}})$、$\hat{Y}^0 = g^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}}_{\boldsymbol{c}})$ を推定する[2]。ここでgはリンク関数、$\hat{\beta}_t$と$\hat{\beta}_c$は治療群と対照群で推定されたパラメータである。観測データからアウトカム回帰モデルを推定し、推定したモデルから「全員が対照群」「全員が治療群」という 2 つの反事実アウトカムを予測し、平均するという点で、What if の標準化と同じ手順を踏む。ところがアウトカム回帰モデルの推定方法が少し異なる。What if では$E[Y \mid A = a, L = l] = A\beta_A + \boldsymbol{L\beta} + A\boldsymbol{L\gamma}$というかたちで 1 つのモデルでアウトカム回帰モデルを推定するのに対し、CAUSALTRT では群ごとにアウトカム回帰モデルを推定する。この違いが ATE 推定にどのような影響を与えるのか数値実験にて様々な条件下で検討する。

3. 方法

3.1. IPW のシミュレーションの設定

3.1.1. シミュレーション手順

以下の手順でシミュレーションデータを生成する。

1. 共変量$L_1, \ldots, L_{10}$の生成

   平均$\boldsymbol{0}$、分散共分散行列が$\boldsymbol{\Sigma}$である 10 変量正規分布から中間変数$X_1, \ldots, X_{10}$をそれぞれn個生成する。nはサンプルサイズであり、nと$\boldsymbol{\Sigma}$の設定は 3.1.2 で述べる。そのあと、以下の式で表すように$L_1, \ldots, L_5$は中央化して二値化し、$L_6, \ldots, L_{10}$はそのまま連続量として扱う。

$$L_i = \begin{cases} \mathbb{I}\{X_i > 0\} - 0.5 & (i = 1, \ldots, 5) \\ X_i & (i = 6, \ldots, 10) \end{cases}$$

2. 治療Aの生成

   以下の式で表される処理生成モデルによって治療Aを生成する。

$$P(A = 1) = \exp\{\alpha_0 + \eta_A(\boldsymbol{L})\}/[1 + \exp\{\alpha_0 + \eta_A(\boldsymbol{L})\}]$$

$$A \sim \mathrm{Ber}\{P(A = 1)\}$$

   ここで、α_0は$P(A = 1)$の割合を調整する定数であり、$\eta_A(\boldsymbol{L})$は治療割付に対する共変量$\boldsymbol{L}$の影響を表す線形予測子である。$E[L_i] = 0$より$E[\eta_A(\boldsymbol{L})] = 0$であるので、$\alpha_0$を調整することで治療割合$P(A = 1)$の期待値を調整することができる。一方、共変量の影響$\eta_A(\boldsymbol{L})$により個々の症例割付に偏りが生じる。α_0と$\eta_A(\boldsymbol{L})$の設定は 3.1.2 で述べる。

3. アウトカムYの生成

   以下の式で表されるアウトカム生成モデルによってアウトカムYを生成する。

$$Y = \tau A + \boldsymbol{L\beta} + \varepsilon, \qquad \varepsilon \sim N(0,1),$$

$$\boldsymbol{\beta} = (0.6,\ -0.5,\ 0,\ 0.4,\ -0.3,\ 0.4,\ 0,\ 0.2,\ 0.1,\ 0)^{\mathrm{T}},$$

   Aの回帰係数τの設定は 3.1.2 で述べる。

4. 各 IPW 法の実行

   手順 1.~3.で生成したデータセットに対して CAUSALTRT を用いて IPW、IPWR、IPWS を実行

しATEの推定値$\hat{\theta}_r$ $(r = 1, \ldots, R,\ R$はシミュレーション回数)を得る。各methodに対してpsmodelステートメントはpsmodel A(event='1') = L1-L10;と指定する。

3.1.2. シナリオの設定

本シミュレーションでは、IPW、IPWR、IPWS の ATE の推定精度がサンプルサイズ、共変量間の共分散、治療割り付けの期待値、共変量による治療割り付けの偏りの強さ、アウトカム生成モデルにおける A の係数τによってどう変わるかを系統的に比較するため、以下の表 2 に示した項目を組み合わせた 120 通りのシナリオを設定する。各シナリオに対してシミュレーション回数$R = 3000$とする。

表 2：IPW のシミュレーションのシナリオ設定

項目	設定
サンプルサイズn（2 通り）	$n = 250, 1000$
共変量の分散共分散行列$\boldsymbol{\Sigma}$（2 通り）	$V[L_i] = 1, Cov[L_i, L_j] = 0$ $V[L_i] = 1, Cov[L_i, L_j] = 0.25$
$P(A = 1)$の期待値を調整する定数α_0（5 通り）	$P(A = 1) = 0.1:\ \alpha_0 = -2.197$ $P(A = 1) = 0.3:\ \alpha_0 = -0.847$ $P(A = 1) = 0.5: \alpha_0 = 0$ $P(A = 1) = 0.7:\ \alpha_0 = 0.847$ $P(A = 1) = 0.9:\ \alpha_0 = 2.197$ $\eta_A(\boldsymbol{L}) = 0$のときに上記の$P(A = 1)$を満たす$\alpha_0$を設定する
処理生成モデルにおける治療割付の偏りの強さ$\eta_A(\boldsymbol{L})$（3 通り）	偏りなし：$\eta_A(\boldsymbol{L}) = 0$ 偏りあり：$\eta_A(\boldsymbol{L}) = 0.5L_1 - 0.5L_2 + 0.4L_3 + 0.3L_6 - 0.3L_7$ 偏り強い：$\eta_A(\boldsymbol{L}) = 1.5L_1 - 1.5L_2 + 1.2L_3 + 1.0L_6 - 1.0L_7$
Aの回帰係数τ（2 通り）	$\tau = 0, 0.5$

3.1.1 節の手順 2.3.における処理生成モデル、アウトカム生成モデルに含まれる共変量の対応を表 3 に示す。

表 3：処理生成モデル、アウトカム生成モデルに含まれる共変量の対応

処理生成モデル	アウトカム生成モデル	共変量
含む	含む	L_1, L_2, L_6
含む	含まない	L_3, L_7
含まない	含む	L_4, L_5, L_8, L_9
含まない	含まない	L_{10}

表 3 において処理生成モデルとアウトカム生成モデルの両方に影響を与えるL_1, L_2, L_6が交絡因子である。

3.1.3. 評価指標

各シナリオに対して、3.1.1 節の手順 3.のアウトカム生成モデルにおける群間差τ を真の ATE とする。各 IPW 手法の ATE の推定精度を比較するため表 4 に示す評価指標を用いる。

表 4：各 IPW 手法の ATE の推定精度を比較するための評価指標一覧

評価指標	式
バイアス	$\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_r - \tau\right)$
RMSE	$\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_r - \tau\right)^2}$
95% CI Coverage	$\frac{1}{R}\sum_{r=1}^{R}\mathbb{I}\{L_r \le \tau \le U_r\}$ L_rは$\hat{\theta}_r$の 95%信頼下限、U_rは$\hat{\theta}_r$の 95%信頼上限
Empirical Power Rate	$\frac{1}{R}\sum_{r=1}^{R}\mathbb{I}\{L_r > 0 \text{ or } U_r < 0\}$

Empirical Power Rate は$\tau = 0$のとき第一種の過誤確率、$\tau = 0.5$のとき検出力として解釈できる。

3.2. 標準化のシミュレーションの設定

3.2.1. シミュレーション手順

基本的に 3.1.1 と同じ手順で以下のようにシミュレーションデータを生成する。

1. 共変量$L_1, \dots, L_{10}$の生成

   3.1.1 の手順 1.と同様の方法で共変量$L_1, \dots, L_{10}$を生成する。nと$\boldsymbol{\Sigma}$の設定は 3.2.2 で述べる。

2. 治療Aの生成

   以下の式で表される処理生成モデルによって治療Aを生成する。

$$P(A = 1) = \exp\{\alpha_0 + \eta_A(\boldsymbol{L})\}/[1 + \exp\{\alpha_0 + \eta_A(\boldsymbol{L})\}]$$
$$A \sim \text{Ber}\{P(A = 1)\}$$

   ここで、$\eta_A(\boldsymbol{L}) = 0.5L_1 - 0.5L_2 + 0.4L_3 + 0.3L_6 - 0.3L_7$である。$\alpha_0$の設定は 3.2.2 で述べる。

3. アウトカムYの生成

   以下の式で表されるアウトカム生成モデルによってアウトカム$\boldsymbol{Y}$を生成する。

$$Y = \tau A + \boldsymbol{L\beta} + A\boldsymbol{L\gamma} + \varepsilon, \qquad \varepsilon \sim N(0,1),$$
$$\boldsymbol{\beta} = (0.6,\ -0.5,\ 0,\ 0.4,\ -0.3,\ 0.4,\ 0,\ 0.2,\ 0.1,\ 0)^{\mathrm{T}}$$

   τと$\boldsymbol{\gamma}$の設定は 3.2.2 で述べる

4. 標準化の実行

   手順 1.~3.で生成したデータセットに対して以下に示す解析モデル M1~M3 を用いて標準化を実行し ATE の推定値$\hat{\theta}_r$を得る。

$$\text{M1: } Y = \beta_0 + A\beta_A + \boldsymbol{L\beta}$$
$$\text{M2: } Y = \beta_0 + A\beta_A + \boldsymbol{L\beta} + A\boldsymbol{L\gamma}$$
$$\text{M3: } Y = \beta_{a0} + \boldsymbol{L\beta}_a$$

   M1 と M2 は PROC GENMOD を用いて標準化し、ATE の推定値は ESTIMATE ステートメントにおいて、M1 では estimate “ATE” A 1、M2 では estimate “ATE” A 1 A*L_1 $\bar{L}_{r1}$ … A ∗ L_{10} $\bar{L}_{r10}$ とすることで得る。一方、M3 は CAUSALTRT の model ステートメントを model Y = L1-L10 / dist=normal;として標準化する。M1 と M2 は 1 つのモデルで推定するのに対し M3 は CAUSALTRT の仕様上、群ごとにモデルを推定する。

3.2.2. シナリオの設定

本シミュレーションでは、各解析モデルの ATE の推定精度がサンプルサイズ、共変量間の共分散、治療割り付けの期待値、アウトカム生成モデルにおけるAと$\boldsymbol{L}$の積項の係数$\boldsymbol{\gamma}$ と A の係数τによってどう変わるかを系統的に比較するため、以下の表 5 に示した項目を組み合わせた 120 通りのシナリオを設定する。各シナリオに対してシミュレーション回数$R = 3000$とする。

表 5：IPW のシミュレーションのシナリオ設定

項目	設定
サンプルサイズn（2 通り）	$n = 250, 1000$
共変量の分散共分散行列$\boldsymbol{\Sigma}$（2 通り）	$V[L_i] = 1, Cov[L_i, L_j] = 0$ $V[L_i] = 1, Cov[L_i, L_j] = 0.25$
$P(A = 1)$の期待値を調整する定数α_0（5 通り）	$P(A = 1) = 0.1：\ \alpha_0 = -2.197$ $P(A = 1) = 0.3：\ \alpha_0 = -0.847$ $P(A = 1) = 0.5：\alpha_0 = 0$ $P(A = 1) = 0.7：\ \alpha_0 = 0.847$ $P(A = 1) = 0.9：\ \alpha_0 = 2.197$ $\eta_A(\boldsymbol{L}) = 0$のときに上記の$P(A = 1)$を満たす$\alpha_0$を設定する
アウトカム生成モデルにおける Aと$\boldsymbol{L}$の積項の係数$\boldsymbol{\gamma}$（3 通り）	積項なし：$\boldsymbol{\gamma} = \boldsymbol{0}$ 積項あり：$\boldsymbol{\gamma} = (0.4,\ 0,\ 0,\ 0,\ 0,\ -0.5,\ 0,\ 0.3,\ 0,\ 0)^T$ 積項強い：$\boldsymbol{\gamma} = (2.0,\ 0,\ 0,\ 0,\ 0,\ -2.5,\ 0,\ 1.5,\ 0,\ 0)^T$
Aの回帰係数τ（2 通り）	$\tau = 0, 0.5$

また、3.2.1 節の手順 2.3.における処理生成モデル、アウトカム生成モデルに含まれる共変量の対応は 3.1.2 節の表 3 と同じである。

3.2.3. 評価指標

各シナリオに対して手順 4.で実行した各方法の ATE の推定精度を比較するため表 4 の評価指標を用いる。Empirical Power Rate は$\tau = 0$のときは第一種の過誤確率、$\tau = 0.5$のときは検出力として解釈できる。

4. 結果

4.1. IPW のシミュレーションの結果

紙面の都合上、方法間の相違が比較的大きかった結果のみ図示し、残りの結果は補足資料に記載した。

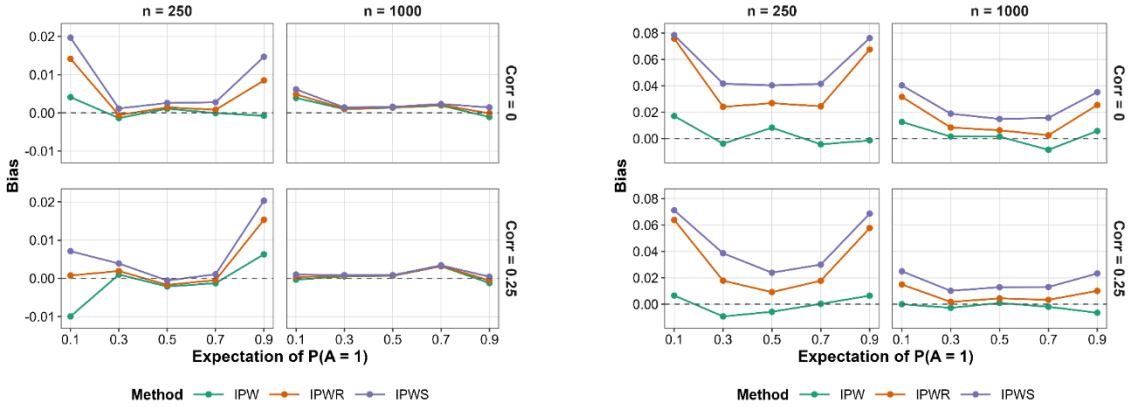


図 2：$\tau = 0$で処理生成モデルにおける治療割付の偏りがある（左）と偏りが強い（右）ときのバイアス

図 2 に示すように処理生成モデルにおける治療割付の偏りが強くなるほど、IPWR、IPWS のバイアスが大きくなった。このバイアスはサンプルサイズが大きくなるほど、治療割付の期待値が均等になるほど小さくなった。RMSE は処理生成モデルにおける治療割付の偏りが強いときのみ IPW の RMSE が大きくなったが、そうでないときは方法間で差はなかった。95% CI Coverage は処理生成モデルにおける治療割付の偏りが強いときのみ IPWR、IPWS の Coverage が 95%を大きく下回った。そうでないときは治療割付の期待値が 0.1, 0.9 のとき以外は方法間で差はなかった。

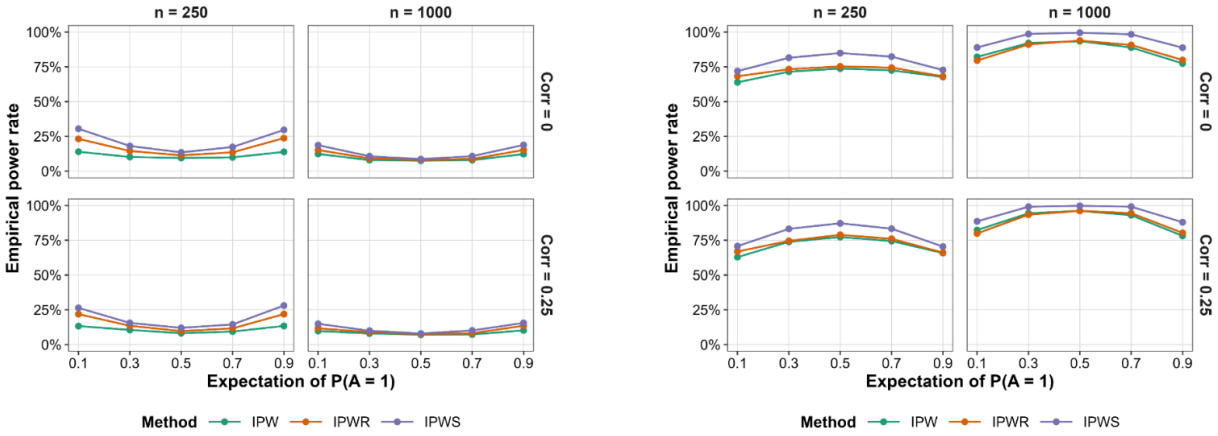


図 3：処理生成モデルにおける治療割付の偏りが強いときにおける$\tau = 0$（左）と$\tau = 0.5$（右）の Empirical Power Rate

図 3 に示すように治療割付の偏りが強いとき、$\tau = 0$（図 3 の左図）から分かるように IPWR、IPWS の第一種の過誤確率が大きくなった。サンプルサイズが大きいと IPW の方法間の差は小さくなった。$\tau = 0.5$（図 3 の右図）から分かるように IPWS の検出力が大きくなった。治療割付の偏りがない場合とある場合で方法間の差はほとんどなかった。

4.2. 標準化のシミュレーションの結果

紙面の都合上、解析モデル間の相違が比較的大きかった結果のみ図示し、残りの結果は補足資料に記載した。アウトカム生成モデルに$\boldsymbol{A}$と$\boldsymbol{L}$の積項がある場合、解析モデル M1 のバイアスは大きくなった。また、治療割合の期待値が 0.5 から乖離するほど M1 のバイアスが大きくなった。M2 と M3 はいずれの設定においてもバイアスがほとんどなかった。積項がない場合はどの解析モデルもバイアスはほとんどなかった。RMSE はいずれの設定でも M2 と M3 の値がほとんど同値だった。

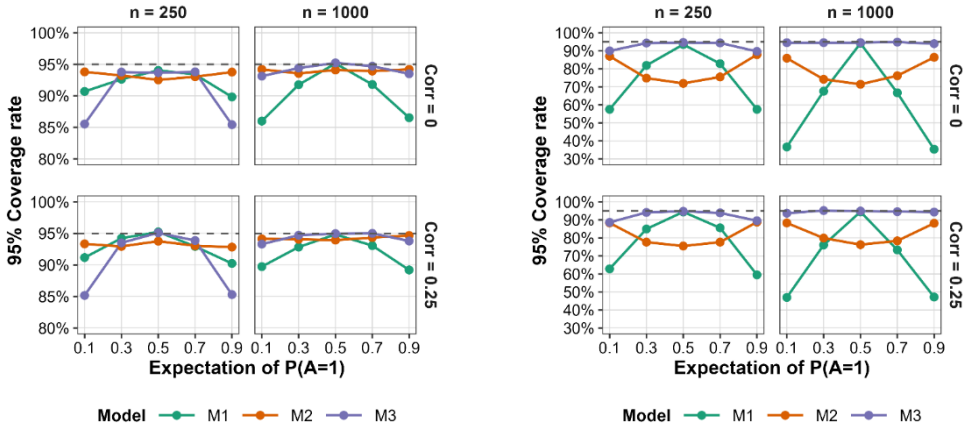


図 4：$\tau = 0$でアウトカム生成モデルにおける積項がある（左）と積項が強い（右）ときの 95% CI Coverage

95% CI Coverage は図 4 に示したように M2 は積項の影響が強い場合は名目水準から乖離した。M3 は治療割合の期待値が 0.1, 0.9 のときは名目水準から乖離するものの、積項の影響が強い場合でも名目水準に近い値を示した。M1 は治療割合の期待値が 0.5 から乖離するほど名目水準から乖離した。

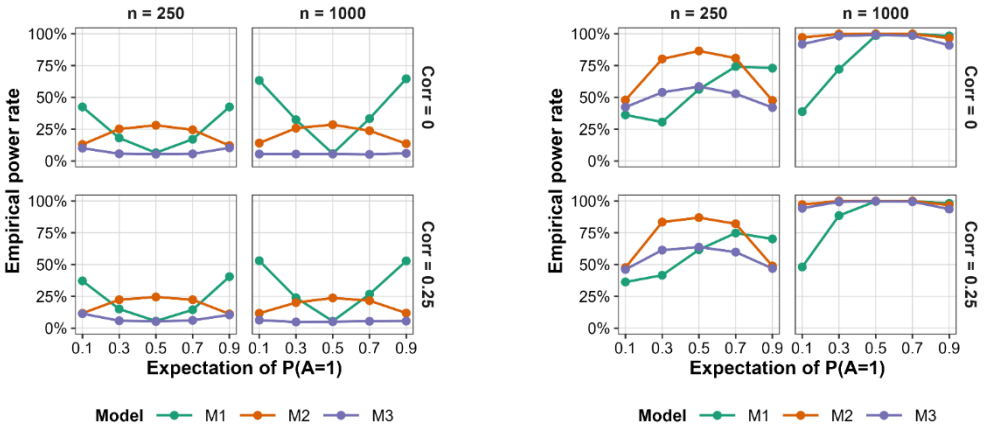


図 5：アウトカム生成モデルにおけるの積項が強いときにおける$\tau = 0$（左）と$\tau = 0.5$（右）の Empirical Power Rate

図 5 に示すようにアウトカム生成モデルの積項が強いとき、M2 の第一種の過誤確率と検出力が大きくなった。$\tau = 0.5$（図 5 の右図）でサンプルサイズが大きいとき、M2、M3 の検出力はほとんど同じだった。アウトカム生成モデルの積項が強くないとき、解析モデル間で第一種の過誤確率と検出力に差はほとんどなかった。

5. 考察

5.1. IPW のシミュレーション

処理生成モデルにおける治療割付の偏りが強い場合、IPWR と IPWS は正規化しているので RMSE は小さくなったがバイアスが生じた。これは IPWR と IPWS が比率推定量であることに起因すると考えられる。2.3.2 節において$A = \sum \frac{\mathbb{I}(A_i=a)y_i}{\hat{f}(A|L_i)}$、$B = \sum \frac{\mathbb{I}(A_i=a)}{\hat{f}(A|L_i)}$とすると H 推定量は$E\left[\frac{A}{B}\right] \approx \frac{E[A]}{E[B]} - \frac{Cov[A,B]}{E[B]^2} + \frac{E[A]Var[B]}{E[B]^3}$と一次近似できる。右辺第 2 項以降が有限標本バイアスとして現れるため、正規化を用いる IPWR と IPWS でバイアスが生じたと考えた。

処理生成モデルにおける治療割付の偏りが強い場合、IPWR と IPWS の 95% CI Coverage は名目水準である 95%より小さくなり、第一種の過誤確率と検出力が大きくなった。これは上記と同様の理由で標本バイアスが生じたことと、重みづけにより標準誤差が小さくなることに起因すると考えた。

5.2. 標準化のシミュレーション

標準化のシミュレーションでは、M1 と M2 が What if に由来する 2 群を併合した単一の解析モデル、M3 が CAUSALTRT の群ごとに推定する解析モデルであった。アウトカム生成モデルに積項が含まれる場合、積項を考慮していない解析モデル M1 は治療割合の期待値が 0.5 から乖離するほどバイアスが大きくなった。治療割合の期待値が小さいあるいは大きい条件では、片方の群のデータが少なくなるため、モデル誤特定に基づく予測誤差がより強く現れたと考えた。M2 は積項を考慮しており、M3 は群別にモデルを推定しているので、いずれの設定でもバイアスが小さくなったと考えた。それに伴って M2 と M3 の RMSE はほとんど同じ値を示したと考えた。

95% CI Coverage と Empirical Power Rate について、アウトカム生成モデルの積項の影響が強い場合、95% CI Coverage と Empirical Power Rate のトレードオフの関係が顕著であることが分かる。M3 の 95% CI Coverage は比較的、名目水準に近い値を示した。これは CAUSALTRT で用いられるロバスト・サンドイッチ分散に起因すると考えられる。CAUSALTRT の REGADJ では、群別の解析モデルのスコア関数と、潜在アウトカムの平均の推定方程式を積み上げたロバスト・サンドイッチ分散が用いられる。ロバスト分散では回帰係数の不確実性だけでなく、母集団から抽出された標本における共分散の不確実性も考慮する。つまり、母集団から抽出された各標本の患者背景の違いによる ATE の不確実性も考慮できる。そのため、母集団の ATE 推定の不確実性をより正確に評価でき、M3 では名目水準に近い 95% CI Coverage が得られたと考えた。一方、M2 の 95% CI Coverage が名目水準から乖離したのは ATE 推定の際に GENMOD の ESTIMATE ステートメントを用いたことに起因すると考えられる。ESTIMATE ステートメントで ATE を推定する際、3.2.1 節の手順 4.で示したように各共変量の標本平均を用いる必要がある。これにより母集団から抽出された標本における共変量の不確実性を考慮できなくなる。そのため、GENMOD で推定された ATE 推定値の分散は過小評価され、95% CI Coverage が名目水準より小さくなったと考えた。M2 の検出力は大きいが、ATE 推定値の分散は過小評価されているため解釈には注意が必要である。M1 はモデル誤特定による影響で 95% CI Coverage が名目水準より小さくなり、第一種の過誤が大きくなったと考えた。

6. 総括

本稿では、What if の視点で IPW および標準化の考え方を読み解くとともに、CAUSALTRT で扱うことのできる各手法とそれらの What if との対応を整理し、シミュレーションにより各手法の ATE 推定精度を比較した。

IPW のシミュレーションの結果から、共変量による治療割付の偏りが弱いと仮定できる状況では、IPW、IPWR および IPWS の間で結果が大きく異なる可能性は低いと考えられる。一方、共変量による治療割付の偏りが強く、かつサンプルサイズが小さい場合、正規化を伴う IPWR および IPWS で有限標本バイアスが大きくなった。このような場合、IPWR、IPWS を用いた ATE の推定はバイアスの観点から妥当でない可能性が高く、推定値の扱いは慎重に行うべきである。

What if で説明される標準化では、2 群のデータを併合した単一の解析モデルを用いる。共変量の値によって治療のアウトカムへの効果が異なるという効果修飾が想定される共変量について、治療とその共変量の積項を含めるといったモデル化が可能である。一方、CAUSALTRT は群ごとに解析モデルを推定する。治療と共変量の積項を事前に特定しなくてもよい利点があるが、群ごとにモデル化するため各解析モデルを推定するための群ごとのサンプルサイズが小さくなり、推定が不安定になる場合がある。標準化のシミュレーションの結果から、単一の解析モデル、群ごとの解析モデルのどちらを用いてもバイアスなく推定できることが分かった。一方、ATE 推定値の信頼区間や検出力を算出する際は ATE 推定値の分散推定の方法に注意が必要である。CAUSALTRT では共分散の不確実性も考慮して母集団の ATE を推定できるため妥当な分散を推定できるが、GENMOD の場合は共分散の不確実性を考慮できないため ATE 推定値の分散が過小評価される恐れがある。単一のモデルで ATE 推定とその信頼区間を求めたい場合は積み上げ式ロバスト分散を考慮すべきである。

本シミュレーションでは、実務において複数の交絡因子候補を同時に考慮しながら因果効果を推定する状況を模擬するため 10 個の共変量が存在する状況を設定した。解析に含める共変量について What if では、交絡調整に必要な共変量がすべて測定され解析に含まれていること、すなわち未測定交絡がないことを仮定す

る。ところが、この仮定が成立するかを観察データのみから確認することは困難である。したがって、実務では、測定された共変量のうち、治療選択およびアウトカムに共通して関係する可能性のある因子を当該領域の専門知識や因果ダイアグラム等を用いて検討し適切に解析へ含めることが重要となる。

本稿では、IPW と標準化それぞれについて整理検討したが、両手法の使い分けについては議論しなかった。モデル誤特定への対策として、What if は IPW と標準化がどちらとも可能である場合は両方実行し、さらに可能なら二重ロバスト推定量を用いるという立場をとっている。モデル誤特定という観点から IPW、標準化、二重ロバスト推定量の推定精度を比較することを今後の検討課題としたい。

参考文献

[1] Hernán MA, Robins JM. *Causal Inference: What If*. Chapman & Hall/CRC. 2020.

[2] SAS Institute Inc. *SAS/STAT® 14.2 User's Guide*. SAS Institute Inc. 2016.

[3] Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*. 1952; **47**: 663-685. doi:10.1080/01621459.1952.10483446.

[4] Hájek J. Comment on a paper by D. Basu. In: Godambe VP, Sprott DA (eds). *Foundations of Statistical Inference*. 1971: 236.

[5] Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*. 2004; **23**: 2937-2960. doi:10.1002/sim.1903.