

red

タグ Causal Inference | ML

発行年 2025

よさそう度 未入力

読む順 未入力

最終更新日時 2025年8月27日 8:00

URL arxiv.org/abs/...07462

作成日時 2025年8月26日 20:14

メモ 未入力

関連page 未入力

+ プロパティを追加する

コメント

コメントを追加...

ABSTRACT (要旨)

Double Machine Learning (DML) は、大規模観察データ集合における因果効果の推定に一般的に用いられる。とりわけ「residuals-on-residuals」回帰推定量 (RORR) は、その単純さと計算上の扱いやさから特に人気がある。しかし、処置効果が不均一である場合、RORRの適切な解釈は十分に理解されていない可能性がある。我々は、連続的なdose-response関数をもつ多値処置に対して、RORRが、観測データ集合には含まれない点で評価された導関数の条件付き分散重み付き平均に収束することを示す。これは一般にAverage Causal Derivative (ACD) とは異なる。したがって、すべてのユニットが同一のdose-response関数を共有する場合であっても、RORRは一般に、サンプルが代表する母集団における平均処置効果に収束しない。我々は大規模データ集合に適した代替推定量を提案する。我々は、例示的な数値例およびNetflixの実データへの応用の双方で、RORRの落とし穴と、提案推定量の望ましい性質を示す。

ACM Reference Format (ACM参考形式)

Apoorva Lal and Winston Chou. 2025. Does Residuals-on-Residuals Regression Produce Representative Estimates of Causal Effects?. In . ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

1 INTRODUCTION (序論)

Double Machine Learning (DML) [5, 8, 18]は、観測された共変量が与えられた条件で処置がランダム割付と同等であることを規定する条件付き無視可能 (ignorability) の仮定の下で、大規模・高次元データ集合における因果効果推定の標準になりつつある[13]。このような仮定を強化するため、さまざまな分野の研究者が、強い関数形の仮定にコミットすることなく多くの共変量に条件づけるためにDML手法を用いている [9, 12, 15]。DMLは大きな手法族を包含するが、本論文はPartially Linear Model (PLM) に焦点を当てる。PLMは、事前共変量 X_i に条件づけて、アウトカム Y_i を（連続値または離散値）処置 T_i に次のように関係づける。

$$Y_i = \theta T_i + g(X_i) + e_i$$

$$T_i = h(X_i) + u_i$$

PLMは、処置とアウトカムが共変量とどのように関係するかについて非常に弱い仮定しか課さない。また、 θ の直観的な二段階DML推定量であるresiduals-on-residuals回帰 (RORR) を動機づける。RORRはFrisch–Waugh–Lovell定理の自然な拡張であり、まず柔軟な機械学習法を用いて X_i の効果を「部分化」し、次に残差を形成する。

1 INTRODUCTION の「RORRはFrisch–Waugh–Lovell定理の自然な拡張で...残差を形成する」とは

Frisch–Waugh–Lovell (FWL) 定理は、線形回帰で「目的変数と説明変数を、制御変数に射影して取り除いた残差どうしで単回帰したときの係数」が、元の多変量回帰の対象係数と一致するという結果。本論文の部分線形モデルでは、共変量を機械学習で柔軟に「部分化 (partial out)」したうえで、残差回帰を行う。手順を式で示す ($\backslash\text{bm} X$ はベクトル) :

$$\begin{aligned} Y_i &= \theta T_i + g(X_i) + e_i, \quad T_i = h(X_i) + u_i \\ \tilde{Y}_i &:= Y_i - \hat{g}(X_i), \quad \tilde{T}_i := T_i - \hat{h}(X_i) \end{aligned}$$

$$\hat{\theta}_{\text{RORR}} := \arg \min_{\theta} \theta \sum_i (\tilde{Y}_i - \theta \tilde{T}_i)^2 = \frac{\sum_i \tilde{T}_i \tilde{Y}_i}{\sum_i \tilde{T}_i^2}.$$

直観：

- 最初に機械学習で「 X_i から予測できる部分」を取り去ることで、 \tilde{T}_i は X_i に起因しない処置の変動、 \tilde{Y}_i は X_i によらないアウトカムの変動だけを残す。
- 最後に \tilde{Y}_i を \tilde{T}_i に回帰して、 X_i による交絡を取り除いた効果を推定する。

この「予測→残差→単回帰」がFWLの拡張（機械学習による部分化）に当たる。

$$\tilde{Y}_i = Y_i - \hat{g}(X_i), \quad \tilde{T}_i = T_i - \hat{h}(X_i)$$

最後に、 \tilde{Y}_i を \tilde{T}_i に回帰して推定量 $\hat{\theta}$ を得る[19]。処置効果 θ が母集団のすべてのユニットで同一である場合、これは二値処置ではAverage Treatment Effect (ATE)、連続処置ではAverage Causal Derivative (ACD)、整数値処置ではAverage Incremental Effect (AIE) にも一致する。これに対し本論文は、処置効果が不均一な場合の $\hat{\theta}$ の確率極限 (plim) とその解釈を検討する。このような不均一性の下では、 $\hat{\theta}$ は因果効果の条件付き分散重み付き平均に収束し、処置値が予測しにくいユニットにより大きな重みを置く。例えば、 X が離散の場合、RORRの推定対象 (estimand) は、処置の分散が最も大きい層の処置効果に最も重みを置く[1, 3, 20]。処置が多値（例えは連続）であるとき、RORRは他の、より微妙なバイアスの影響を受け得る。本論文では、二値および多値の処置に対するRORRの一般的な分析を与える。我々は、これらのバイアスの経験的重要性を、スタイルイズドな数値例とNetflixの実データの双方で示す。最後に、Augmented Inverse Propensity Weighting (AIPW) を処置のビニング[7]と組み合わせた代替推定量を提案し、その有利な性質を理論的に、そして実証応用において示す。

2 RESIDUALS-ON-RESIDUALS REGRESSION WITH TREATMENT EFFECT HETEROGENEITY (2 処置効果の不均一性を伴うResiduals-on-Residuals回帰)



RORRには、実証応用においてそれを魅力的にする多くの有利な属性がある。前述のとおり、nuisance関数の推定を非常に柔軟に行える。さらに、RORRは二重にロバスト (doubly-robust) な性質をもち、処置モデルまたはアウトカムモデルのいずれかが正しく指定されれば真のθに収束する。両モデルが正しい場合には効率的推定量である[10]。RORRの近年の応用には、経済学[4]、生態学[11]、公衆衛生[21]の研究が含まれる。

商業的な現場——大規模データセット、短いタイムライン、技術的背景の多様なステークホルダーといった特徴をもつ——では、実務的理由からもRORRが用いられる。例えば、最終の回帰ステップの計算は効率的であり、の最終推定値を計算するのに必要な統計量の数は少數で済む。加えて、(1) 事前共変量で説明可能な変動の除去と (2) 残った外生的な T_i の変動が Y_i に与える効果の推定、という手順は直観的であり、非専門家にも説明しやすい。

1 INTRODUCTION の「商業的な現場でRORRが用いられる理由」とは

論文が挙げるポイントを実務目標で言い換えると次の通り。

- 計算効率が高い：最終ステップは単回帰なので、必要統計量は
 $\sum \tilde{T}_i, \sum \tilde{Y}_i, \sum \tilde{T}_i^2, \sum \tilde{T}_i \tilde{Y}_i$ 程度。ストリーミング処理や分散集計に向く。
- 実装が簡単：大規模データでも、nuisance推定 (\hat{g}, \hat{h}) を学習した後は、単純な回帰一発。締切が短い現場に適合。
- 伝えやすい手順：
 1. 事前共変量で説明できる変動を除去し、
 2. 残った外生的な T_i の変動が Y_i に与える影響を測る、
 というストーリーは、非専門家のステークホルダーにも直観的に説明しやすい。

注意点として、処置効果が不均一・非線形だと、RORRの推定対象は平均効果 (ACD/AIE/ATE) と一致しないことがある、というのが本論文全体の主張。

「残った外生的な T_i の変動」とは、共変量 X_i で説明できる (=交絡になりうる) 部分を取り除いたあとに残る、 T_i の X_i と無関係な「ゆらぎ」のことを指す。

$$T_i = h(\mathbf{X}_i) + u_i, \quad E[u_i | \mathbf{X}_i] = 0 \\ Y_i = \theta T_i + g(\mathbf{X}_i) + e_i, \quad E[e_i | \mathbf{X}_i] = 0, \quad E[e_i u_i | \mathbf{X}_i] = 0$$

- 第1段階で T_i を \mathbf{X}_i で予測し、残差をつくる：
 $\tilde{T}_i := T_i - \hat{h}(\mathbf{X}_i) \approx u_i$

この \tilde{T}_i が「残った外生的な T_i の変動」。定義的に \mathbf{X}_i と直交し (平均的に独立)、上式の仮定のもとではアウトカム誤差 e_i とも条件付きで無相関。

- 直観：
 1. \mathbf{X}_i が決めている“系統的な”処置のばらつき (交絡の源) を除く。
 2. 残った \tilde{T}_i は、 \mathbf{X}_i を固定したときに「ランダムに近い」処置のゆらぎ (=外生的変動) とみなせる。
 3. だから第2段階で $\tilde{Y}_i := Y_i - \hat{g}(\mathbf{X}_i)$ を \tilde{T}_i に回帰すると、 \mathbf{X}_i による交絡の影響を取り除いた“因果的な”関係だけを拾いやすい。
- 何ではないか：
 - 物理的にランダム化した処置でも、道具変数 (IV) で誘発した変動でもない。あくまで「 \mathbf{X}_i で説明できない」「 \mathbf{X}_i に条件づけて外生」と仮定される統計的残差。
 - したがって、条件付き無視可能性 (θ 同質なら $T_i \perp Y_i | \mathbf{X}_i$ に相当、PLMでは $E[e_i u_i | \mathbf{X}_i] = 0$) が破れれば、 \tilde{T}_i は外生とは言えず、推定はバイアスする。

要するに、「残った外生的な T_i の変動」とは

$$\tilde{T}_i = T_i - \hat{h}(\mathbf{X}_i)$$

という“ \mathbf{X}_i で説明できない処置の残差”的ことで、 \mathbf{X}_i を所与としたとき Y_i の誤差と相関しない (=外生) と仮定して、その部分が Y_i に与える影響を推定する、という意味だ。

しかし、この単純さは代償を伴い得る。すなわち、 θ を「平均的」処置効果 (ATE、ACD、AIEのいずれであれ) として解釈することは、PLMに内在する同質な処置効果の仮定に依存しており、この仮定は応用で成り立たない場合がある。本節では、この仮定の二つの一般的な違反、すなわち、個体間で効果が不均一な二値処置と、連続的なdose-response関数をもつ多値処置の下での θ の解釈を論じる。

2.1 Binary Treatments with Heterogeneous Treatment Effects (二値処置における不均一処置効果)

まず、処置効果が不均一な二値処置に対するRORR推定量のplimを検討する。 $T_i \in \{0, 1\}$ を二値処置指標とし、次のモデルを考える：

$$Y_i = \theta_i T_i + g(X_i) + e_i \quad (1) \\ T_i = h(X_i) + u_i$$

ここで、 θ_i は個別効果である。さらに、誤差は条件付き外生かつ無相関であると仮定する： $E[e_i | X_i] = 0$ 、 $E[u_i | X_i] = 0$ 、 $E[e_i u_i | X_i] = 0$ 。これは条件付き無視可能性の線形化である。この仮定から、 $g(X_i) = E[Y_i - \theta_i T_i | X_i]$ および $h(X_i) = E[T_i | X_i]$ が含意される。また、 θ_i は X_i を与えた条件で T_i と独立、すなわち $\theta_i \perp T_i | X_i$ と仮定する。

観測 $O_i = (Y_i, T_i, X_i)$, $i = 1, \dots, N$ に対し、 $Y_i - g(X_i)$ を $T_i - h(X_i)$ に回帰するOLSのplimを考える。観測 $O_1, \dots, O_N \sim O$ は独立かつ同一分布と仮定する。実務では g と h を推定する必要があるが、記述を簡潔にするため、ここではこれらが一貫推定されると仮定し、(真の)極限の g と h に焦点を当てる。1

まず次を観察する：

$$\hat{\theta} \xrightarrow{p} \frac{E[(T_i - h(X_i))(Y_i - g(X_i))]}{E[(T_i - h(X_i))^2]} \quad (2) \\ = \frac{E[\theta_i, (T_i^2 - T_i h(X_i))]}{E[(T_i - h(X_i))^2]}$$

T_i が二値であることと反復期待の法則を用いると、これは次のように書き換えられる：

$$\frac{E[\theta_i, (T_i^2 - T_i h(X_i))]}{E[(T_i - h(X_i))^2]} = \frac{E[\theta_i, E[T_i(1 - h(X_i)) | X_i]]}{E[(T_i - h(X_i))^2]} \quad (3) \\ = \frac{E[\theta_i, (T_i - h(X_i))^2]}{E[(T_i - h(X_i))^2]} \quad (4)$$



ここで(3)は、 θ_i が T_i と条件付きで独立であることから従う。これは既知の結果、すなわち二値処置においては、線形回帰は個体別処置効果の条件付き分散重み付き平均に収束することを示している[1, 3]。

直観的な言い換えのため、条件付き分散重みを

$$\omega_i := \frac{(T_i - h(\mathbf{X}_i))^2}{E[(T_i - h(\mathbf{X}_i))^2]}$$

とおく（構成上 $E[\omega_i] = 1$ である）。このとき、ATEに対するRORRのplim ($\tilde{\theta}$ と表す) のバイアスは次のように書ける：

$$\begin{aligned} \tilde{\theta} - E[\theta_i] &= E[\omega_i \theta_i] - E[\theta_i] \\ &= \text{Cov}(\omega_i, \theta_i). \end{aligned} \quad (5)(6)$$

言い換えると、処置効果が不均一な二値処置におけるRORRのバイアスは、個体別処置効果と、処置残差の条件付き分散との共分散に等しい。この共分散は特殊な場合（例えば処置が一様に無作為に割り当てられる場合）を除きゼロにはならないため、一般に $\tilde{\theta} \neq E[\theta_i]$ である。2

□ 前提（この節で使う仮定）

$$Y_i = \theta_i T_i + g(\mathbf{X}_i) + e_i$$

$$T_i = h(\mathbf{X}_i) + u_i$$

$$\begin{aligned} E[e_i | \mathbf{X}_i] &= 0, \quad E[u_i | \mathbf{X}_i] = 0, \quad E[e_i u_i | \mathbf{X}_i] = 0, \quad \theta_i \perp T_i | \mathbf{X}_i, \\ T_i &\in 0, 1 \quad (\text{二値処置}; \text{ゆえに } T_i^2 = T_i), \quad h(\mathbf{X}_i) = E[T_i | \mathbf{X}_i]. \end{aligned}$$

□ (2) 式の導出

まず、分子を展開しているだけ。

$$\begin{aligned} \hat{\theta} &\xrightarrow{p} \frac{E[(T_i - h(\mathbf{X}_i))(Y_i - g(\mathbf{X}_i))]}{E[(T_i - h(\mathbf{X}_i))^2]} = \frac{E[(T_i - h(\mathbf{X}_i))(\theta_i T_i + e_i)]}{E[(T_i - h(\mathbf{X}_i))^2]} \\ &= \frac{E[\theta_i(T_i - h(\mathbf{X}_i))T_i]}{E[(T_i - h(\mathbf{X}_i))^2]} + \frac{E[(T_i - h(\mathbf{X}_i))e_i]}{E[(T_i - h(\mathbf{X}_i))^2]}. \end{aligned}$$

ここで第2項は

$$E[(T_i - h(\mathbf{X}_i))e_i] = E[E(T_i - h(\mathbf{X}_i))e_i | \mathbf{X}_i] = E[(T_i - h(\mathbf{X}_i)), E(e_i | \mathbf{X}_i)] = 0$$

より0。よって

$$\hat{\theta} \xrightarrow{p} \frac{E[\theta_i(T_i^2 - T_i h(\mathbf{X}_i))]}{E[(T_i - h(\mathbf{X}_i))^2]}.$$

あなたの式の $E[\theta_i, (T_i^2 - T_i h(\mathbf{X}_i))]$ のカンマは掛け算を意味しており、上と同じことである。

□ (3)への変形（反復期待と条件付き独立）

$$\begin{aligned} E[\theta_i(T_i^2 - T_i h(\mathbf{X}_i))] &= E[\theta_i, T_i(1 - h(\mathbf{X}_i))] \quad (\text{二値なので } T_i^2 = T_i) \\ &= E\left[\theta_i, E\{T_i(1 - h(\mathbf{X}_i)) | \mathbf{X}_i\}\right] \\ &= E\left[\theta_i, E\{T_i(1 - h(\mathbf{X}_i)) | \mathbf{X}_i\}\right] \end{aligned}$$

最後の行では、 $\theta_i \perp T_i | \mathbf{X}_i$ により θ_i を内側の条件付き期待の外に出している。これが(3)の式

$$\frac{E[\theta_i, E(T_i(1 - h(\mathbf{X}_i)) | \mathbf{X}_i)]}{E[(T_i - h(\mathbf{X}_i))^2]} \text{に相当する。}$$

□ (4)への変形（Bernoulliの分散恒等式）

二値 T_i について

$$\begin{aligned} E[T_i(1 - h(\mathbf{X}_i)) | \mathbf{X}_i] &= (1 - h(\mathbf{X}_i)), E[T_i | \mathbf{X}_i] \\ &= h(\mathbf{X}_i), (1 - h(\mathbf{X}_i)) \\ &= E[(T_i - h(\mathbf{X}_i))^2 | \mathbf{X}_i]. \end{aligned}$$

最後の等号は

$$\begin{aligned} E[(T_i - h)^2 | \mathbf{X}_i] &= E[T_i^2 - 2T_i h + h^2 | \mathbf{X}_i] \\ &= E[T_i | \mathbf{X}_i] - 2h, E[T_i | \mathbf{X}_i] + h^2 \\ &= h - 2h^2 + h^2 = h(1 - h) \end{aligned}$$

（ここで $h = h(\mathbf{X}_i)$ と略記）を使ったもの。よって

$$\frac{E[\theta_i, E(T_i(1 - h) | \mathbf{X}_i)]}{E[(T_i - h)^2]} = \frac{E[\theta_i, E(T_i - h)^2 | \mathbf{X}_i]}{E[(T_i - h)^2]} = \frac{E[\theta_i, (T_i - h)^2]}{E[(T_i - h)^2]},$$

が(4)に一致する。

□ 重み付き平均としての解釈と共分散表示

分子・分母が同じ「重み」($T_i - h(\lfloor \log X_i \rfloor)$)²を持つので、正規化重み

$$\omega_i := \frac{(T_i - h(\mathbf{X}_i))^2}{E[(T_i - h(\mathbf{X}_i))^2]}$$

を置けば $E[\omega_i] = 1$ で、

$$\tilde{\theta} = \frac{E[\theta_i, (T_i - h)^2]}{E[(T_i - h)^2]} = E[\omega_i, \theta_i].$$

すると

$$\tilde{\theta} - E[\theta_i] = E[\omega_i \theta_i] - E[\theta_i] = E[\omega_i \theta_i] - E[\omega_i]E[\theta_i] = \text{Cov}(\omega_i, \theta_i),$$

となる($E[\omega_i] = 1$ を使用)。これが最後の等式の意味。

共分散の公式

$$\text{Cov}[X, Y] = E[XY] - \mu_X \mu_Y$$

$\text{Cov}[X, Y]$	共分散
$E[XY]$	X と Y の積の期待値
$\mu_X = E[X]$	X の期待値
$\mu_Y = E[Y]$	Y の期待値

要するに、

- (2) : モデルを代入して直項を消し、分子を $E[\theta_i(T_i^2 - T_i h(\mathbf{X}_i))]$ にする

- (3) : 二値性と反復期待・条件付き独立で $E[\theta_i, E\{T_i(1 - h) | \mathbf{X}_i\}]$ にする

- (4) : Bernoulli の恒等式 $E[T(1-h) | X] = E[(T-h)^2 | X]$ を使う
 - 最後 : 重み ω_i を定義して共分散表示に落とす
- という流れである。

2.2 Many-Valued Treatments (多値処置)

ここでは、連続値または整数値の多値処置で、連続的なdose-response関数をもつ場合に注目する。先行研究は、線形処置効果推定量の解釈に対する処置効果の不均一性の影響を検討してきたが、その多くは二値処置および／または線形の処置効果という文脈で行われてきた[3]。しかし多くの応用では、処置は連続であるか、あるいはアウトカムに対して非線形の効果（例えば限界効果の遮断）をもつ。このような非線形性は処置効果の不均一性の重要な形態であるが、いくつかの例外を除き[2, 22]、先行研究では十分な注目を受けてこなかった。

具体的には、次のモデルを考える：

$$\begin{aligned} Y_i &= f(T_i) + g(\mathbf{X}_i) + e_i \\ T_i &= h(\mathbf{X}_i) + u_i \end{aligned} \quad (7)$$

ここで、 f は多値処置 T_i の連続微分可能な関数である。

- 2.2 Many-Valued Treatments の「 f は多値処置 T_i の連続微分可能な関数」とは

$$\text{モデル } Y_i = f(T_i) + g(\mathbf{X}_i) + e_i, \quad T_i = h(\mathbf{X}_i) + u_i$$

で、処置 T_i のアウトカムへの影響を、未知の関数 $f(\cdot)$ によって表しているという意味。

- 「多値処置」なので T_i は連続値や整数値を取り得る。
- 「連続微分可能」は、全域で微分 $f'(t)$ が存在し、その変化が連続であること。これにより平均値の定理を使って

$$\hat{\theta} = \frac{E[(T_i - h(\mathbf{X}_i))^2 f'(T_i^*)]}{E[(T_i - h(\mathbf{X}_i))^2]}$$
- のような導出が可能になる。
- また、 T_i が離散でも平均値の定理を適用できるよう、 f は T_i の凸包上で定義されると仮定する（本文の技術的仮定）。

解釈： $f(t)$ は「dose-response曲線」、その導関数 $f'(t)$ は「その処置水準での限界因果効果」。本論文では、RORRが平均するのは $t = T_i$ ではなく $t = T_i^*$ （受用量と条件付き平均の凸結合点）で評価された $f'(t)$ で、しかも分散重み付きになる点が重要となる。

平均値の定理

関数 $f(x)$ が $a \leq x \leq b$ で連続で、
 $a < c < b$ で微分可能であるとき

$$\frac{f(b) - f(a)}{b - a} = f'(c) \quad \text{かつ} \quad a < c < b$$

式1

式2

を満たす実数 c が少なくとも 1 つ存在する

もし $f'(x)$ が単調増加なら
式より $f'(a) < f'(c) < f'(b)$ となり、
式1を代入して $f'(c)$ を消去する

$(b - a) f'(a) < f(b) - f(a) < (b - a) f'(b)$

a=n, b=n+1 とする

$f'(n) < f(n+1) - f(n) < f'(n+1)$

この式は、不等式の証明はもちろん、 $f(n+1) - f(n)$ の極限を問われたときの「はさみうちの原理」にも活用可能！

後に明らかになる理由により、たとえ T_i 自体が連続でなくとも、 f は T_i の凸包上で良好に定義されていると仮定する。先と同様に、条件付き外生性、 g と h の一貫推定量の存在、ならびに独立同分布 (iid) の観測を仮定する。

これらの仮定の下で、RORRの推定量は確率収束により次に収束する：

$$\hat{\theta} \xrightarrow{P} \frac{E[(T_i - h(\mathbf{X}_i)), f(T_i)]}{E[(T_i - h(\mathbf{X}_i))^2]}. \quad (8)$$

$h(\mathbf{X}_i)$ は \mathbf{X}_i を与えた条件で定数であることと反復期待の法則を用いれば、上式は次のように書き換える：

$$\frac{E[(T_i - h(\mathbf{X}_i)), f(T_i^*)]}{E[(T_i - h(\mathbf{X}_i))^2]} = \frac{E[E(T_i - h(\mathbf{X}_i)), (f(T_i) - f(h(\mathbf{X}_i))) | \mathbf{X}_i]}{E[(T_i - h(\mathbf{X}_i))^2]}. \quad (9)$$

平均値の定理により、各 \mathbf{X}_i について T_i と $h(\mathbf{X}_i)$ の間に T_i^* が存在して、

$$\begin{aligned} \frac{E[(T_i - h(\mathbf{X}_i)), (f(T_i) - f(h(\mathbf{X}_i))) | \mathbf{X}_i]}{E[(T_i - h(\mathbf{X}_i))^2]} &= \frac{E[(T_i - h(\mathbf{X}_i))^2, f'(T_i^*)]}{E[(T_i - h(\mathbf{X}_i))^2]} \\ &= \frac{E[\omega_i, f'(T_i^*)]}{E[\omega_i]}, \end{aligned} \quad (10)$$

が成り立つ。これは、二値の場合と同様に、 $\hat{\theta}$ が因果効果の条件付き分散重み付き平均に収束することを示している。3 しかしこれの場合と異なり、平均される量は、サンプルが代表する母集団において「処置を増加させること」に対応する因果効果として解釈できない。というのも、点 T_i^* は個体 i が実際に受け取った処置量ではなく、受用量 T_i とその条件付き平均 $h(\mathbf{X}_i)$ の凸結合だからである。（凸結合上の点 $= T_i$ と $h(\mathbf{X}_i)$ を結ぶ線分上の任意の点）したがって、 T_i^* は観測された処置レベルとは限らない。 T_i が連続でない場合には、 T_i^* はそもそも実現可能な処置値ではない可能性がある。

- 2.2 Many-Valued Treatments の「平均値の定理により、各 \mathbf{X}_i について T_i と $h(\mathbf{X}_i)$ の間に T_i^* が存在して」とは

関数 f が区間 $[\min\{T_i, h(\mathbf{X}_i)\}, \max\{T_i, h(\mathbf{X}_i)\}]$ 上で連続、かつその内部で微分可能なら、平均値の定理により各 i について次が成り立つ。

$$f(T_i) - f(h(\mathbf{X}_i)) = f'(T_i^*) \cdot (T_i - h(\mathbf{X}_i))$$

ここで T_i^* は T_i と $h(\mathbf{X}_i)$ の間のどこかの点（個体*i*に依存するランダム量）。これを式(9)の内側に代入すると

$$E[(T_i - h(\mathbf{X}_i)) f(T_i) - f(h(\mathbf{X}_i)) | \mathbf{X}_i] = E[(T_i - h(\mathbf{X}_i))^2 f'(T_i^*) | \mathbf{X}_i],$$

したがって式(10)の形

$$\frac{E[(T_i - h(\mathbf{X}_i))^2 f'(T_i)]}{E[(T_i - h(\mathbf{X}_i))^2]}$$

が得られる。注意点は、 T_i^* は実際に観測された処置水準とは限らず、 T_i と $h(\mathbf{X}_i)$ の凸結合上の点であること（特に T_i が離散のとき）。

Proposition 1 (命題1)

$O_i = (Y_i, T_i, \mathbf{X}_i)$, $i = 1, \dots, N$ が (7) 式のモデルに従う分布からのiid標本であるとする。 f と h は一貫推定され、かつ $E[e_i | \mathbf{X}_i] = 0$, $E[u_i | \mathbf{X}_i] = 0$, $E[e_i u_i | \mathbf{X}_i] = 0$ と仮定する。さらに f は至る所で微分可能で、その導関数 f' は Lipschitz 連続で定数 L をもつと仮定する。最終的に $E[(T_i - h(\mathbf{X}_i))^2] > 0$ を仮定する。このとき residuals-on-residuals 回帰 (DODD) の $\text{nlm}(\tilde{A} + f)$ が affine ($I - \alpha$) のときの限

Proof (証明)

iid観測、条件付き無視可能性、 f と h の一貫推定の下で、 $\tilde{\theta}$ は (10) 式で与えられる。ACD に対する $\tilde{\theta}$ のバイアスは、次のように足し引きして二つの項に分解できる：

$$\begin{aligned} \frac{E[(T_i - h(\mathbf{X}_i))^2 f'(T_i^*)]}{E[(T_i - h(\mathbf{X}_i))^2]} - E[f'(T_i)] &= \underbrace{\frac{E[(T_i - h(\mathbf{X}_i))^2 f'(T_i^*)]}{E[(T_i - h(\mathbf{X}_i))^2]} - \frac{E[(T_i - h(\mathbf{X}_i))^2 f'(T_i)]}{E[(T_i - h(\mathbf{X}_i))^2]}}_{:=A} \\ &\quad + \underbrace{\frac{E[(T_i - h(\mathbf{X}_i))^2 f'(T_i)]}{E[(T_i - h(\mathbf{X}_i))^2]} - E[f'(T_i)]}_{:=B}. \end{aligned} \quad (11)$$

第一項 A は、RORR の plim と、サンプルで実際に観測された処置分布における分散重み付き平均導関数との差である。A の絶対値は L と T_i の分布の関数として上から抑えられる。まず A を書き換える：

$$A = \frac{E[(T_i - h(\mathbf{X}_i))^2, (f'(T_i^*) - f'(T_i))]}{E[(T_i - h(\mathbf{X}_i))^2]}. \quad (12)$$

f' が Lipschitz であるという仮定から、任意の u, v についてある定数 $L \geq 0$ が存在して

$$|f'(u) - f'(v)| \leq L|u - v|. \quad (13)$$

が成り立つ。 T_i^* は T_i と $h(\mathbf{X}_i)$ の間にあるので、

$$|f'(T_i^*) - f'(T_i)| \leq L|T_i^* - T_i| \leq L|T_i - h(\mathbf{X}_i)|. \quad (14)$$

が成り立つ。両辺に $(T_i - h(\mathbf{X}_i))^2$ を掛けて期待値をとると、

$$\begin{aligned} E[(T_i - h(\mathbf{X}_i))^2, |f'(T_i^*) - f'(T_i)|] \\ \leq L, E[(T_i - h(\mathbf{X}_i))^2, |T_i - h(\mathbf{X}_i)|] = L, E[|T_i - h(\mathbf{X}_i)|^3] \end{aligned} \quad (15)$$

を得る。両辺を $E[(T_i - h(\mathbf{X}_i))^2]$ で割れば、次の評価が得られる：

$$|A| \leq L, \underbrace{\frac{E[|T_i - h(\mathbf{X}_i)|^3]}{E[(T_i - h(\mathbf{X}_i))^2]}}_{:=\kappa}. \quad (16)$$

ここで $|T_i - h(\mathbf{X}_i)|^3$ は常に 0 以上である。したがって、 κ が 0 となるのは $T_i = h(\mathbf{X}_i)$ がほとんど確実に成り立つ場合に限られる。しかし仮定より $E[(T_i - h(\mathbf{X}_i))^2] > 0$ であるため、 $T_i = h(\mathbf{X}_i)$ はほとんど確実には成り立たない。ゆえに、(16) 式右辺が 0 になるのは $L = 0$ の場合に限られる。第二項 B は、分散重み付き平均導関数と ACD の間の既知のバイアスであり、その解釈は二値処置の場合と同様である。すなわち、再び条件付き分散重み ω_i を用いると、

$$E[\omega_i f'(T_i)] - E[f'(T_i)] = \text{Cov}(\omega_i, f'(T_i)), \quad (17)$$

となる。これは (6) 式の連続版に相当する。 f が affine であれば $f'(T_i)$ は定数となり、 $\text{Cov}(\omega_i, f'(T_i)) = 0$ である。ゆえに、 $\tilde{\theta}$ の絶対バイアスは

$$|A + \text{Cov}(\omega_i, f'(T_i))| \leq |A| + |\text{Cov}(\omega_i, f'(T_i))| \leq L\kappa + |\text{Cov}(\omega_i, f'(T_i))| = 0$$

が f が affine のときに成り立ち、主張が得られる。□

□ Proposition 1 の affine ($L=0$) とは

命題では f' が Lipschitz で

$$|f'(u) - f'(v)| \leq L \cdot |u - v|$$

を仮定している。 $L = 0$ であることは $f'(u) = f'(v)$ が任意の u, v に対して成立することを意味し、すなわち f' が定数。ゆえに f は一次関数 (切片つきの線形関数、すなわち affine)

$$f(t) = a + bt$$

となる。言い換えると、 f が affine $\Leftrightarrow f'$ が定数 $\Leftrightarrow f'$ の Lipschitz 定数 $L = 0$ 。

Corollary 1 (系1)

命題1は、 f が affine であることが $\tilde{\theta} = E[f'(T_i)]$ の十分条件であることを述べる。さらに $L > 0$ が成り立ち、 $A \neq 0$ かつ $A \neq -B$ (A と B は (11) で定義) を仮定すれば、 $\tilde{\theta} = E[f'(T_i)]$ となるのは f が affine の場合に限る。

言い換えれば、 f が affine に近いほど、 $\tilde{\theta}$ は f' の条件付き分散重み付き平均に近づく。逆に、その条件付き分散重み付き平均が $E[f'(T_i)]$ に等しくなるのは $\text{Cov}(\omega_i, f'(T_i)) = 0$ のときであり、これは f が affine なら自明に成り立つ。したがって、 f が affine (すなわち PLM が正しく指定) であれば、二つのバイアスはいずれも消える。しかし f が affine でない場合、作戦的な場合 (例えば二つのバイアスがちょうど相殺される) を除き、 $\tilde{\theta} \neq E[f'(T_i)]$ である。

□ 要点を一気にまとめる。2.2章が主張しているのは「多値（連続・整数）処置で dose-response が非線形のとき、RORR は“平均的な限界因果効果 (ACD)”には一般にならない」ということだ。理由は、RORR が (i) 導関数を観測された処置値ではなく擬似点で評価し、かつ (ii) 処置残差の条件付き分散で重み付けして平均するからだ。以下、骨子。

1. 前提と目標

モデルは

$$\begin{aligned} Y_i &= f(T_i) + g(\mathbf{X}_i) + e_i \\ T_i &= h(\mathbf{X}_i) + u_i \end{aligned}$$

で、 f は連続微分可能 (さらに f' は Lipschitz)。条件付き外生などの標準仮定の下で、RORR が収束する量 (plim) を、平均的な限界因果効果 ACD

$$\text{ACD} := E[f'(T_i)]$$

とどう関係づけられるかを明らかにするのが目的。

2. RORR の plim は「擬似点」での導関数の分散重み付き平均。ます

$$\hat{\theta} \xrightarrow{p} \frac{E[(T_i - h(\mathbf{X}_i)), f(T_i)]}{E[(T_i - h(\mathbf{X}_i))^2]}.$$

ここで平均値の定理を使うと、各 i について T_i と $h(\mathbf{X}_i)$ の間に T_i^* が存在して

$$\hat{\theta} \xrightarrow{p} \frac{E[(T_i - h(\mathbf{X}_i))^2, f'(T_i^*)]}{E[(T_i - h(\mathbf{X}_i))^2]} = \frac{E[\omega_i, f'(T_i^*)]}{E[\omega_i]},$$

となる。重みは

$$\omega_i := \frac{(T_i - h(\mathbf{X}_i))^2}{E[(T_i - h(\mathbf{X}_i))^2]} \quad (E[\omega_i] = 1).$$

重要なのは、導関数の評価点が T_i ではなく T_i^* (T_i と $h(\mathbf{X}_i)$ の凸結合上の点) だということ。 T_i が離散だと T_i^* は実現不可能な値になりうる。(RORRの極限で現れる導関数が観測処置 T_i ではなく、その“途中点” T_i^* で評価されること)

3. なぜACDと一致しないのか (2つのバイアス)

ACDとの差は

$$\frac{E[(T_i - h)^2 f'(T_i^*)]}{E[(T_i - h)^2]} - E[f'(T_i)] = A + B$$

に分解できる (h は $h(\mathbf{X}_i)$ の略記)。

- A : 評価点のズレによるバイアス

$$A := \frac{E[(T_i - h)^2 f'(T_i^*)] - E[f'(T_i)]}{E[(T_i - h)^2]}$$

f' がLipschitz (定数L) なら

$$|A| \leq L, \kappa := \frac{E[|T_i - h(\mathbf{X}_i)|^3]}{E[(T_i - h(\mathbf{X}_i))^2]}.$$

- B : 条件付き分散重み付けによるバイアス

$$B := \frac{E[(T_i - h)^2 f'(T_i)]}{E[(T_i - h)^2]} - E[f'(T_i)] = \text{Cov}(\omega_i, f'(T_i)).$$

1. いつACDに一致するか (命题1の結論)

f がaffine (一次関数) すなわち $f'(t)$ が定数なら、 $L = 0$ で $A = 0$ 、かつ $f'(T_i)$ が定数なので $B = 0$ 。したがって

$$\tilde{\theta} = \text{ACD} = E[f'(T_i)]$$

が成り立つ。逆に、 f がaffineでない限り一般には $\tilde{\theta} \neq E[f'(T_i)]$ 。

2. 直観 (バイアスの向き)

多くの応用で $f''(t) < 0$ (遞減限界効果)。RORRは

- ω_i により「処置が予測しにくい (残差分散が大きい)」領域を上方重み付けし、
- さらに T_i^* による評価点のズレで「より大きい処置側」で f' を見る傾向になる。

遞減なら高いほど $f'(t)$ は小さいため、全体として $\tilde{\theta}$ は ACD より小さくなりやすい (数値例・実データで観察された“負方向の乖離”)。

1. 実務上の含意

- RORRは「平均的限界効果」を測っているつもりでも、実は上の2つのバイアスを含む“別の平均”に収束しうる。
- f がほぼ線形 (affineに近い)、かつ $\text{Cov}(\omega_i, f'(T_i))$ が小さい状況でのみ、ACDに近い解釈が可能。
- そうでなければ、4章のcoarsened AIPWのように、処置分布を明示的に尊重する推定法が望ましい。

ひとと言で言えば：

「連続・多値処置かつ非線形なdose-responseでは、RORRはACDではなく、観測外の擬似点 T_i^* で評価した導関数の“条件付き分散重み付け平均”に収束する。ACDと一致するのは f がaffineのときだけ」というのが2.2章の主張だ。

3 NUMERICAL EXAMPLE (3 数値例)

直観の助けするために、本節では様式化した数値例を提示する。4閉形式の解析を容易にするため、我々はいくつか単純化した仮定を置くが、同時にこれらの選択は実データの質的侧面を反映することも意図している。特に、次を仮定する。

(1) $E[Y_i | T_i, \mathbf{X}_i]$ は T_i に関して単調増加である一方で、限界効果の遞減も示す。すなわち、 $f(T_i)$ を (7) 式で定義するとき、 $f'(T) > 0$ かつ $f''(T) < 0$ 。

(2) T_i は過分散をもつ計数変数であり、そのため $E[T_i | \mathbf{X}_i]$ の正しいモデルであっても不均一分散誤差 (heteroskedastic errors) をもつ。

△ 奥山怜央 10時間
 $Y_i = f(T_i) + g(\mathbf{X}_i) + e_i \quad (7)$
 $T_i = h(\mathbf{X}_i) + u_i,$

□ 3 NUMERICAL EXAMPLE の「 T_i は過分散をもつ計数変数」とは

計数変数 T_i は $0, 1, 2, \dots$ の整数值をとる。過分散とは平均に対して分散が大きい (典型的には $\text{Var}(T_i | \mathbf{X}_i) > E[T_i | \mathbf{X}_i]$) という性質を指す。ポイントは、 $E[T_i | \mathbf{X}_i] = h(\mathbf{X}_i)$ を正しくモデル化しても $\text{Var}(T_i | \mathbf{X}_i)$ が \mathbf{X}_i に依存して変わるために、残差 $u_i = T_i - h(\mathbf{X}_i)$ が不均一分散 (heteroskedastic) になること。

数値例では $T_i | X_i = j \sim \text{Poisson}(\lambda_j)$ を仮定しており、この場合でも $\text{Var}(T_i | X_i = j) = \lambda_j$ は \mathbf{X}_i によって異なるため、OLS的には不均一分散で、線形ガウス誤差の仮定とは異なる分散構造になる。過分散の語は一般に「ポアソンよりも分散が大きい」状況を指すことが多いが、ここで著者が強調しているのは「分散が一定でない (X により変わる)」という実務的含意 (=ヘテロスケ) である。

\mathbf{X}_i を、値 $j = 1, \dots, J$ をとるCategorical変数で、その確率を π_1, \dots, π_J とし、 T_i を \mathbf{X}_i に条件づけたPoisson分布で、そのパラメータを λ_j とする。さらに $f(T_i) = \log(T_i + 1)$ と仮定する。これにより、 \mathbf{X}_i が与えられたときの T_i に関する Y_i の条件付き期待導関数の式として、次を導ける (付録A.1参照) :

$$E[f'(T_i) | X = j] = \frac{1 - \exp(-\lambda_j)}{\lambda_j}. \quad (18)$$

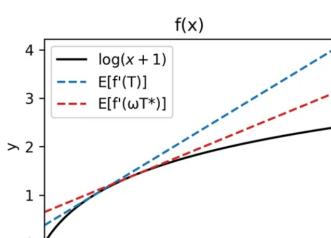
このとき ACD は単に $\sum_j \pi_j \frac{1 - \exp(-\lambda_j)}{\lambda_j}$ である。

また、RORRのplimも解析的に次のように導ける:

$$\hat{\theta} \xrightarrow{p} \hat{\theta} = \frac{\sum_j^J \pi_j E[(T_i - \lambda_j)^2 f'(T_i^*) | X_i = j]}{\sum_j^J \pi_j \lambda_j}, \quad (19)$$

ここで、先と同様に T_i^* は T_i と λ_j の間の点である。ACDに対する二つのバイアスに注意せよ。第一に、 f' を T_i ではなく T_i^* で評価している。第二に、各 $f'(T_i^*)$ を平均からの二乗偏差でさらに重みづけしている。

Figure 1は、このデータ生成過程に基づくシミュレーションにより生じるバイアスを示している。



1段目 (f(x))

- 黒実線は $f(x) = \log(x+1)$ (上に凹: $f''(x) < 0$)。
- 青破線の傾きは $E[f'(T)]$ (ACDの勾配) を表す。
- 赤破線の傾きは $E[f'(\omega T^*)]$ (RORRが暗黙に評価している勾配) を表す。
- 上に凹なので、大きい x ほど $f'(x)$ は小さくなる。

2段目 (Observed Treatment Distribution)

- 観測された処置 T の分布。平均は約 3.00。

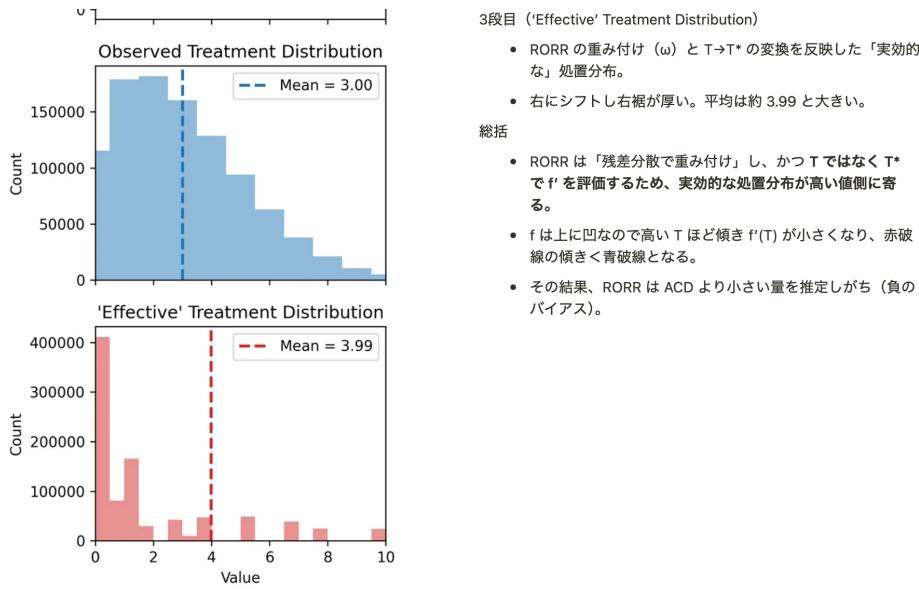


Figure 1: Bias of Residuals-on-Residuals Regression

まず、Figure 1 の上段パネルに $f(T_{-i}) = \log(T_{-i} + 1)$ を描き、さらに傾きが $E[f'(T_i)]$ (青) と $E[f'(\omega_i T_i^*)]$ (赤) に等しい接線を描く。ここで $\omega_i T_i^*$ は、RORR が分析する「effective」な処置である。重要なポイントは、RORR は ACD とは異なる（しかもより小さい）量を推定対象としている、ということである。5 級くパネルはこの結果に対する直観を与える。すなわち、 ω_i による重み付けと T_i を T_i^* に変換した後の「effective」な処置分布は、観測された処置分布よりもはるかに右に歪んでいる。これは、 f の傾きをより大きな T_i の値で評価しがちであることを意味する。 $f''(T_i) < 0$ なので、これが負のバイアスにつながる。

Table 1 には、種々のサンプルサイズでのシミュレーションから得られた推定的経験的 RORR を報告する。比較のため、経験的 ACD ($\frac{1}{T_i+1}$ の標本平均として計算) と、(18) 式を用いて計算した真の ACD も報告する。

Table 1: Simulation Results for RORR and ACD

Sample Size	Empirical RORR	RORR 95% CI	RORR Plim
10,000	0.250	(0.237, 0.262)	0.247
100,000	0.248	(0.244, 0.252)	0.247
1,000,000	0.248	(0.247, 0.249)	0.247
Sample Size	Empirical ACD	ACD 95% CI	True ACD
10,000	0.362	(0.357, 0.367)	0.365
100,000	0.364	(0.362, 0.366)	0.365
1,000,000	0.365	(0.364, 0.365)	0.365

Table 1 (RORRとACDのシミュレーション)

- 真の ACD は 0.365。経験的 ACD はサンプルサイズが増えるほど 0.365 に収束しており、一貫性を確認できる。
- 経験的 RORR は約 0.248、理論的 plim も 0.247 で一致しており、ACD より明確に小さい（バイアスは約 -0.118、相対で約 -32%）。
- サンプルサイズ拡大で信頼区間が急速に狭まり、RORR が「ACD とは別の量」に安定収束していることを示す。

この例では T_i が整数値であることに注意されたい。したがって、より妥当な因果推定対象は、おそらく Average Incremental Effect (AIE) であり、次式で定義される：

$$E[Y_i(T_i + 1) - Y_i(T_i)] = \sum_{t=0}^{\infty} (f(t+1) - f(t)) p(t), \quad (20)$$

ここで p は T_i の確率質量関数である。しかし、RORR の plim は導関数の重み付き平均であるため、Table 1 では ACD に焦点を当て、AIE については第4節で一貫推定量を提案する。Table 1 が示すように、ACD に対して RORR の plim は負にバイアスしている。これは、処置の大きい値における dose-response 曲線の導関数に、より大きな重みを置くためである。次節では、整数値の処置に対する AIE、および連続処置に対する ACD の一貫推定量を提案する。

4 COARSENED AUGMENTED IPW ESTIMATOR (4 粗化Augmented IPW推定量)

連続処置に対する Average Causal Derivative (ACD) を推定する一般的なベンチマークは、Generalized Propensity Score (GPS) [14] である。しかし GPS は処置の条件付き密度を推定する必要があり、単純化するパラメトリック仮定がなければ、収束が遅い・不安定といった問題に直面しうる [16]。

RORR と GPS の双方に対する代替として、我々は単純な coarsened Augmented Inverse Propensity Weighting (AIPW) 推定量を提案する。これは、反事実平均の AIPW 推定量を構成要素として用いる [7]。本推定量はまず、 T_i の台を互いに素な K 個の区間（あるいは bin） $\{S_1, S_2, \dots, S_K\}$ に分割するところから始める。 $S_k = [t_{k-1}, t_k]$ とする。各 bin S_k について、bin 水準の propensity score $p_k(\mathbf{X}_i) := \Pr(T_i \in S_k | \mathbf{X}_i)$ を、多クラス分類モデルを適合するなどして推定する。この推定量を $\hat{p}_k(\mathbf{X}_i)$ と表す。あわせて、 $m_k(\mathbf{X}_i) := E[Y_i | T_i \in S_k, \mathbf{X}_i]$ の柔軟なアウトカム回帰も推定し、 $\hat{m}_k(\mathbf{X}_i)$ と表す。

次に、bin S_k における周辺平均の通常の AIPW 推定量を

$$\hat{\psi}_k := \frac{1}{N} \sum_{i=1}^N \left(\frac{1(T_i \in S_k)}{\hat{p}_k(\mathbf{X}_i)}, (Y_i - \hat{m}_k(\mathbf{X}_i)) + \hat{m}_k(\mathbf{X}_i) \right)$$

RORR と同様に、この推定量は二重にロバストである。すなわち、 \hat{p}_k または \hat{m}_k のいずれかが一貫推定されれば、 $\hat{\psi}_k$ も一貫である。 $\hat{\psi}_k$ の plim は、bin 内平均化のためやや繊細な解釈をもつ。標準的仮定の下では、「 \mathbf{X}_i と $T_i \in S_k$ が与えられた条件付き分布に従って処置された場合の、母集団平均の潜在アウトカム」と解釈できる。

- 4 COARSENED AUGMENTED IPW ESTIMATOR の $\hat{\psi}_k$ の plim の解釈

bin $S_k = [t_{k-1}, t_k]$ を固定し、 $m(t, \mathbf{x}) := E[Y | T = t, \mathbf{X} = \mathbf{x}]$ 、 $F_{T|\mathbf{X}}(\cdot | \mathbf{X}, T \in S_k)$ を用いると、AIPW が標的一致で収束する量は

$$\begin{aligned} \psi_k &:= E_{\mathbf{X}} \left[\int_{S_k} m(t, \mathbf{X}) dF_{T|\mathbf{X}}(t | \mathbf{X}, T \in S_k) \right], \\ &= E_{\mathbf{X}} [E(Y | \mathbf{X}, T \sim \mathcal{L}(T | \mathbf{X}, T \in S_k))]. \end{aligned}$$

ここで $\mathcal{L}(T | \mathbf{X}, T \in S_k)$ は \mathbf{X} を固定し、 T を条件付き分布 $F_{T|\mathbf{X}}$ を S_k に切り詰めて再正規化した分布から無作為に割り付ける」という意味。現在アウトカム未記 $Y(t)$ を用いれば、条件付き無差可能性的下で $\psi_k = E_{\mathbf{X}} \left[\int E(Y(t) | \mathbf{X}) dF_{T|\mathbf{X}}(t | \mathbf{X}, T \in S_k) \right]$ 。

精英 奥山怜央 10時間
比較手法になりそう

と解釈できる。したがって

- $\hat{\psi}_k$ は「単一の処置水準 t での平均潜在アウトカム」ではなく、
- 「BIN S_k 内の条件付き処置分布に従って処置が割り付けられたときの母集団平均潜在アウトカム」

を推定している。これが「BIN内平均化に伴うやや繊細な解釈」という意味であり、その後に隣接差分

$$\hat{\Delta}_k = \hat{\psi}_{k+1} - \hat{\psi}_k, \quad \hat{\delta}_k = \frac{\hat{\psi}_{k+1} - \hat{\psi}_k}{\bar{t}_{k+1} - \bar{t}_k}$$

を用いて、BIN間の離散的な処置効果や離散導関数を可視化・集約している。

さらに、各BINの重み w_k を、その経験的比率に比例するように定義する：

$$w_k = \begin{cases} \frac{\Pr(T_i \in S_k)}{\sum_{\ell=1}^{K-1} \Pr(T_i \in S_\ell)} & \text{for } k < K, \\ 0 & \text{for } k = K. \end{cases}$$

最後に、全体の ACD 推定量は次で与えられる：

$$\hat{\psi} := \sum_{k=1}^{K-1} w_k \left(\frac{\hat{\psi}_{k+1} - \hat{\psi}_k}{\bar{t}_{k+1} - \bar{t}_k} \right),$$

ここで $\bar{t}_k = \frac{t_{k+1} + t_k}{2}$ は S_k の中点である。ヒューリスティックには、 $\hat{\psi}$ は区分的線形関数で f^* を近似し、その下側セグメントの経験分布を用いて加重平均を計算している。付録Bでは、N と K が無限大に発散する極限で、この推定量が ACD の一貫推定量となることを証明する。

Table 2 は、第3節のシミュレーションデータに本推定量を適用した結果を示す。処置が整数値であるため、各観測された処置値をそのまま BIN として設定すればよい。表が示すように、これは AIE に対する一貫推定量を与える。6

Table 2: Simulation Results for Coarsened AIPW and AIE

Sample Size	Empirical AIE	AIE 95% CI	True AIE
10,000	0.277	(0.247, 0.308)	0.295
100,000	0.291	(0.282, 0.300)	0.295
1,000,000	0.295	(0.290, 0.300)	0.295

Table 2 (粗化AIPWとAIEのシミュレーション)

- 真のAIEは0.295。提案する粗化AIPWによる経験的AIEは0.277→0.291→0.295と収束しており、一貫推定になっている。
- 信頼区間もサンプルサイズとともに急速に縮小。整数値処置ではAIE推定に有効であることを示す。

5 EMPIRICAL APPLICATION (5 実証応用)

本節では、Netflixの実データを用いて、我々の理論的分析の経験的な妥当性を示す。機密保持上、共有できる情報には制約があるものの、本節の主眼は、上で論じた理論的バイアスが実データにおいても（そして我々の経験ではしばしば）現れ得ることを示す点にある。

本アプリケーションでは、我々がFeature Aと呼ぶある機能の使用が、将来のNetflixへの訪問にどのような影響を与えるかを理解することを目的とした。この問い合わせるために、2,971,128名のメンバーから無作為標本を抽出し、28日間のウィンドウにおいてFeature Aを使用した回数を数えた。次に、この回数を同期間における当該メンバーのNetflixへの訪問回数で割ることで、連続処置であるFeature A Usage Rateを定義した。続いて、各メンバーの次の28日間ウィンドウにおけるNetflix訪問回数をアウトカムとして定義した。共変量としては、各メンバーのFeature A使用回数、および処置期間に先行する7日・14日・28日におけるNetflix訪問回数を含めた。7

データ集合は、おおよそ等サイズの学習・検証・テストの各データセット（各≈980,000ユニット）に分割した。PLMにおけるnuisanceパラメータを推定するため、学習データで観測された処置変数とアウトカム変数に対して、勾配ブースティング回帰木を適合し、検証データを用いてブースティングランダム数を調整した。最後に、テストデータにおいてアウトカム残差を処置残差に回帰して、RORRの処置効果推定量を得た。Table 3に示すとおり、Feature Aがその後の訪問に与えるRORR推定値は小さく、負で、統計的に有意であった。8 この所見は、Feature AがNetflix訪問を増やすだろうという我々の事前の信念と矛盾していた。

Table 3: RORR and AIPW Estimates of the Effect of Feature A Usage on Netflix Visits (N = 980,139)

	RORR	Std. Err.	95% CI
Feature A Usage Rate	-0.0038	0.001	(-0.005, -0.002)
	AIPW	Std. Err.	95% CI
Feature A Usage Rate	5.343	0.010	(5.324, 5.362)

Table 3 (Netflix実データ：Feature A使用率→訪問回数)

- RORRは-0.0038 (SE=0.001, 95%CI: -0.005, -0.002) で有意な負。著者らの事前期待（正の効果）と逆。
- 粗化AIPWは5.343 (SE=0.010, 95%CI: 5.324, 5.362) で大きく正。
- 両者が食い違う理由は、RORRが「処置残差の条件付き分散で重み付け」し、さらに観測外の点 T_i^* で導関数を評価するため、実効的な処置分布が高い値側に偏ることに起因する（非線形・適応効果の下で負方向のバイアス）。一方、粗化AIPWはビニング×AIPWでACD/AIEに整合的な量を推定する設計で、実データでもより妥当な符号・大きさを示している。

table1~3総括

- シミュレーションでは理論どおり、RORRはACDより小さく収束（負のバイアス）。
- 粗化AIPWはAIE（整数処置）・ACD（連続処置）の推定に一致。
- 実データでも、RORRはミスリーディングになり得る一方、粗化AIPWは直觀と整合的な結果を与える。

我々のcoarsened AIPW推定量は、この不可解な結果に対する直觀を与える。この推定量を適合させるため、まず処置を5つのBINにcoarsen（粗化）し、その結果生じたBINに対して勾配ブースティングを用いた多クラス分類器を適合した。0値（すなわち処置期間中にFeature Aの使用なし）は第1BINに割り当て、残る非ゼロ値を四分位点で分割した。アウトカム回帰については、RORRの適合に用いたものと同じ関数を再利用した。

Figure 2は標準的な診断図であり、各BINにおける前処置期アウトカムの標準化値とBIN1との差を、逆傾向スコア重み付けの前後で描いている。図が示すように、IPWはアウトカム変数の前処置差を有意に縮小し、BIN間の比較可能性を高め、条件付き無視可能性（ignorability）の信頼性を強化する。

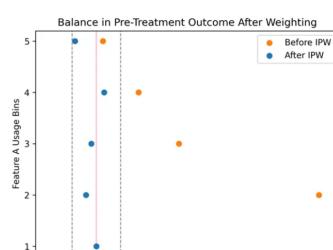


Figure 2 (IPW後の前処置アウトカムのバランス)

- 横軸は「BIN1（ゼロ使用）との差の標準化平均差」。点が0に近いほど、前処置のアウトカムがBIN間で似ていることを意味する。
- オレンジ（IPW前）は大きくズれているが、青（IPW後）はすべて0付近に集まり、BIN間の比較可能性が大きく改善している。



Figure 2: Balance in Pre-Treatment Outcomes After Inverse Propensity Score Weighting

主要結果はFigure 3に描かれており、上から順に、各処置ビンにおける処置後アウトカムの反事実平均、各ビンを1段階増やすことに対応する推定処置効果、そして各ビンに占めるデータセットの割合（ゼロ使用のビンに明確に集中）を示している。

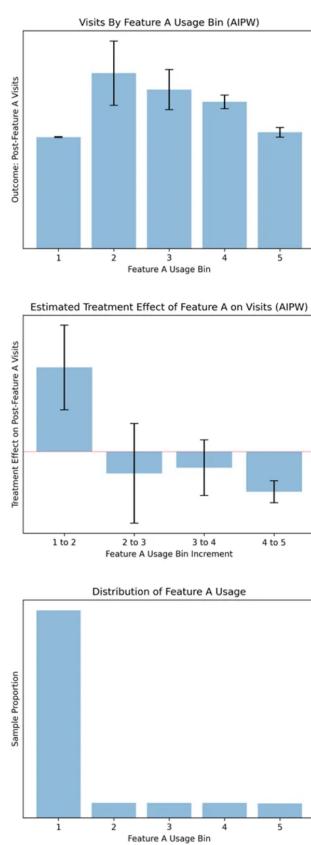


Figure 3 (AIPWでの主結果)

- 上段：各使用ビンの処置後訪問回数の反事実平均。ビン1が最も低く、ビン2で大きく上がり、その後はやや低下傾向。「各ビンに入ったときの反事実的な平均アウトカム（訪問回数）」の水準

$$\hat{\psi}_k := \frac{1}{N} \sum_{i=1}^N \left(\frac{1}{\hat{p}_k(\mathbf{X}_i)} (Y_i - \hat{m}_k(\mathbf{X}_i)) \hat{m}_k(\mathbf{X}_i) \right)$$

- 中段：ビンを1段階上げたときの推定処置効果。1→2は大きな正、2→3・3→4・4→5は小さな負（過減・右下がりの領域）。「中段は「ビンを1段階上げたときのアウトカムの差（離散的な限界効果）」

$$\hat{\Delta}_k := \hat{\psi}_{k+1} - \hat{\psi}_k \quad (k = 1, \dots, K-1)$$

- 下段：使用度分布。データの大半がビン1（ゼロ使用）に集中。

まとめ

- IPWで前処置差が解消され、因果比較の信頼性が向上。
- AIPWでは「ゼロ使用→少し使用」の効果が大きく正で、母集団の多くが該当するため、代表的効果は正になる。
- 高使用域では限界効果が小さくなる（あるいは負）ため、RORRが強く重み付けすると全体を過小評価しうるが、AIPWは分布を代表する効果を可視化している。

Figure 3: Treatment Effects of Feature A Usage on Netflix Visits After AIPW Weighting. Note that we remove y-axis labels to preserve confidentiality.

Figure 3が示すように、AIPWは、ゼロ使用ビン（ビン1）から次のビン（ビン2）へ移行する際の処置効果を大きく正と推定する。さらに、Feature Aの使用はゼロ過剰（zero-inflated）であるため、ビン1が最も代表的なビンである。したがって、Table 3に示すとおり、coarsened AIPWの推定値は正で統計的に有意であり、その大きさはRORR推定値よりもはるかに大きい。この乖離は、coarsened AIPW推定量が、処置分布を示すように処置効果に明示的な重み付けを行なうのに対し、RORRは処置の高い値をもつユニット（dose-response曲線が右下がりとなる領域）に重みを上げるために生じる。意思決定にとってこの乖離はきわめて重要である。すなわち、RORRはFeature Aがアウトカムに対して負の処置効果をもつことを示唆する一方で、AIPWの結果は、Feature Aの使用を増やすことが大多数のメンバーにとって正の効果をもつことを示す。実際、ゼロでないすべてのFeature A使用ビンは、ゼロ使用ビンよりも高い条件付き平均を示しており、何らかのFeature A使用は使用なしより望ましいことを示している。

6 CONCLUSION (6 結論)

DML推定量は学術研究・商業研究の双方でますます一般的になっているが、研究者は——従来と同様に——特定の応用に対する適合性を慎重に評価しなければならない。本論文は、residuals-on-residuals回帰（RORR）という特別なケースに焦点を当て、処置効果が不均一である場合や、処置が多値でdose-response関数が非線形である場合に、RORRをどのように正しく解釈すべきかを検討した。我々は、後のケースにおいて、RORRが因果導関数の条件付き分散重み付き平均に収束すること、しかもそれらの導関数がデータに観測される処置分布とは異なる「疑似処置（pseudo-treatment）」分布上で評価されるという追加の複雑性を伴うことを示した。

我々の実証アプリケーションが示すとおり、いかなる「平均的」処置効果に対しても、RORRが有する微妙なバイアスは意思決定に重大な影響を及ぼし得る。これらのバイアスに対処するため、我々はcoarsened AIPW推定量を提案し、これが因果効果のより代表性のある推定値をもたらすことを示した。