統計的機械学習　（応用計量分析２）第４回

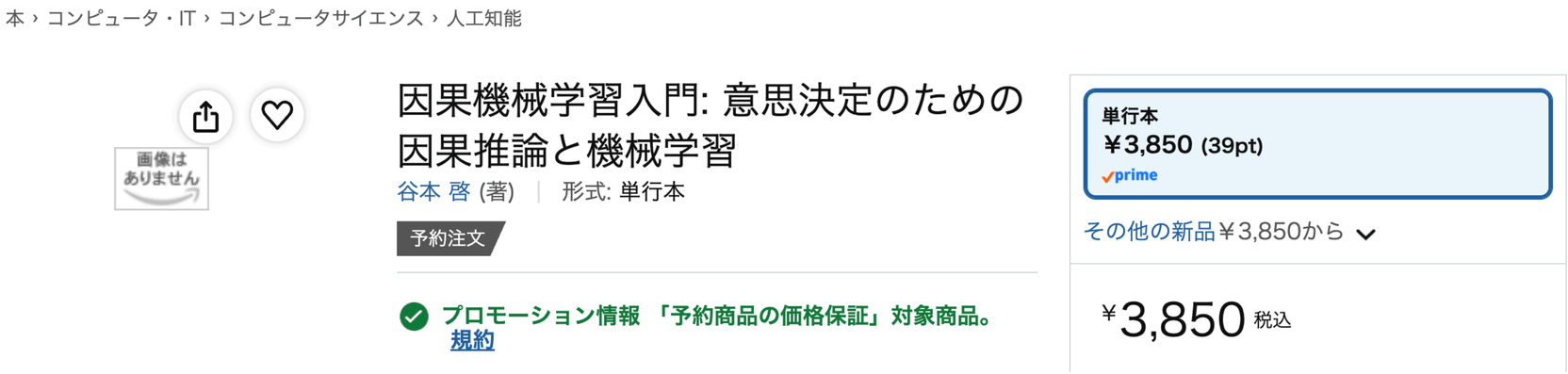
平均因果効果の推定法（参考pdf 4章）
条件付き平均因果効果の推定法（参考pdf 5章）にも少し触れる

参考書について

出欠コード
XXXX

『因果機械学習入門』11/30発売予定

- 参考pdfの内容が紙の本になります (Amazon)
- 購入は必須ではないが、
  - 今後もデータ分析をやっていきたい人
  - 機械学習等の研究がしたい人
  - せっかく時間を使って勉強したので振り返りたい人　などはぜひ。



出欠コード
XXXX

因果推論の意義と状況設定

- 因果推論の精度は意思決定の良さの保証を与える
- 潜在結果モデルはデータの表を拡張して反事実的結果を欠損とみなす
  - こうすることで因果効果を推定対象として表現でき、それが識別可能性を満たせばデータが無限にあれば（分布がわかれば）真値を特定可能
- （平均）因果効果の識別可能性を保証するには3つの仮定があれば十分
  - SUTVA、無視可能性、正値性
- IPW法は損失関数を傾向スコアの逆数で重み付ける

推定対象と評価指標

因果推論の典型的な精度は潜在結果や効果の精度

- 意思決定につながる推定対象（第2回内容）
  - **条件付き平均因果効果**（Conditional Average Treatment Effect; **CATE**） $\tau(x) = \mathbb{E}[y_1 - y_0 \mid x]$
    - ある程度個別化された効果。因果機械学習（機械学習分野における因果推論研究）でよく推定対象とされる
  - 条件付き平均潜在結果 $\mathbb{E}[y_a \mid x]$
- その他の典型的な推定対象
  - 平均因果効果（Average Treatment Effect; **ATE**） $\tau = \mathbb{E}[y_1 - y_0]$
  - 個別因果効果（Individual Treatment Effect; **ITE**） $\tau^i = y_1^i - y_0^i$
    - 究極的な目的だが、潜在結果が片方しか観測されないため識別のためには特別な追加仮定が必要（第7回以降）
- 評価指標
  - 因果効果の精度 $\mathrm{PEHE}(\hat{\tau}) = \mathbb{E}_x\left[\left(\tau(x) - \hat{\tau}(x)\right)^2\right]$
  - 潜在結果の精度 $\mathrm{MSE}^u(\hat{f}) := \mathbb{E}_x\left[\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}}\left(\mathbb{E}\left[y_a \mid x\right] - \hat{f}(x,a)\right)^2\right]$
  - 方策価値 $V(\pi) := \mathbb{E}_x\mathbb{E}_{a\sim\pi(a|x)}\mathbb{E}\left[y_a \mid x\right]$

性能評価自体も因果推論問題

- 評価自体も難しい
  - 評価が正しくできるならその評価指標を損失とする学習器が作れるはず
    - 損失は微分可能であることが通常なので若干異なるが、計算量の問題が異なるだけで統計的な問題は同じ
- いくつかの典型的な検証方法があり、一長一短ある
  - 擬似的に潜在結果が揃っているデータを使う（例：Twinsデータ）
    - 双子のうち体重の大小を行動とみなし、それぞれの死亡率への影響を調べる
      双子は非常によく似た背景をもつため、片方の結果をもう片方の反事実の代理として用いる
    - そのようなデータが取れる状況は限定的
  - 潜在結果に人工データを用いる（例：乳幼児健康発達プログラム（IHDP）データセット）
    - 人工データであるため、実データの複雑さ（例えばノイズ分散のx依存性など）を再現できない可能性
  - RCTされたデータを用いる（例：職業訓練（Jobs）データセット）
    - PEHE等そのものは測れず、ノイズありの精度で測るか、意思決定性能で測る

ノイズあり/なしの推定精度

ノイズありの精度でもモデル間の比較には十分

- 実際に観測される（潜在）結果は一般にノイズつき
  - $y_a = \mathbb{E}[y_a \mid x] + \varepsilon \quad (\mathbb{E}[\varepsilon \mid x] = 0)$
- y_aに対するMSEと潜在結果の条件付き期待値に対するMSEは期待値においてモデルによらない定数を除いて同等

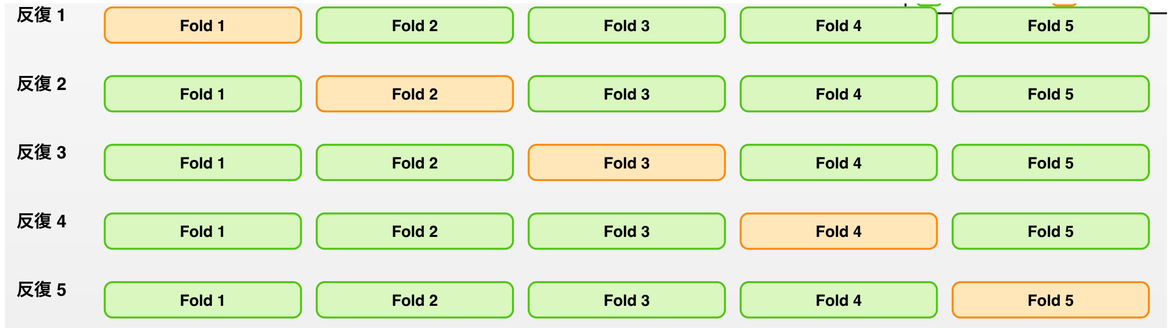
$$\mathbb{E}\left[\left(\underbrace{\mathbb{E}[y_a \mid x] + \varepsilon}_{y_a} - \hat{f}(x,a)\right)^2 \mid x\right] =$$

$$\left(\mathbb{E}[y_a \mid x] - \hat{f}(x,a)\right)^2 + 2\underbrace{\mathbb{E}[\varepsilon \mid x]}_{=0}\left(\mathbb{E}[y_a \mid x] - \hat{f}(x,a)\right) + \underbrace{\mathbb{E}[\varepsilon^2 \mid x]}_{\text{モデルによらない}}$$

- → 学習やモデル選択（モデル f 間の相対評価）には使える
  - 意思決定性能の評価（第2回講義参照）に使うと過剰に悪い評価になるので注意
- 欠損データからノイズ付き因果効果を推定できる
  - $\tau(x) = \mathbb{E}[y_1 \mid x] - \mathbb{E}[y_0 \mid x] = \dfrac{\mathbb{E}[ay \mid x]}{\mu(1 \mid x)} - \dfrac{\mathbb{E}[(1-a)y \mid x]}{\mu(0 \mid x)}$ ・・・観測不可能な反事実的結果にはゼロが掛かるので平均は計算可能

検証用データでハイパーパラメータを調整する

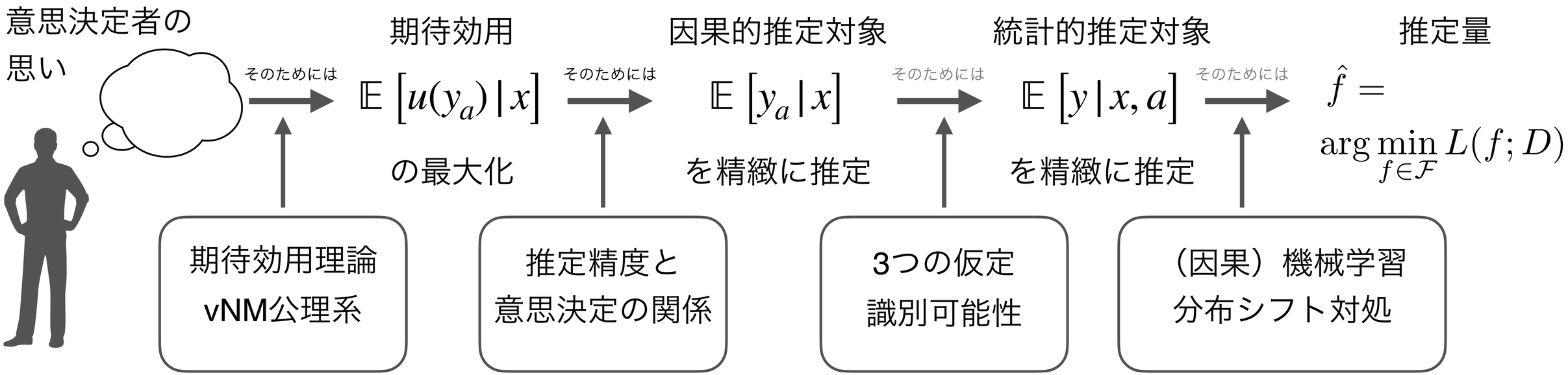
- 多くの機械学習手法はハイパーパラメータが存在
  - 仮説集合の広さや正則化の強さなど、学習途中ではなく事前に設定するパラメータ
  - 複数のモデルを学習し、学習に用いずに取っておいた検証用データの精度に従い選択するのが標準的
    - 因果推論では、RCT（ランダム化実験）データを検証用に用いる場合も
  - データが少なくできるだけ全て使いたい場合は交差検証などの方法も



まとめ

意思決定とデータ分析をつなぐ各ステップで仮定や近似（上界）がある

- 統計的推定には誤差がつきもの
- 因果推論の評価指標の意味で高精度なモデルは意思決定に資する
- そのようなモデルを得るための手続きが因果推論



出欠コード
XXXX

小テスト・前回の内容で質問等あれば

- 1. ガイダンス・因果推論と機械学習の概論
- 2. 意思決定理論（期待効用理論）の復習、因果推論との関係
- 3. 潜在結果モデルに基づく因果推論の枠組み
- **4. 平均因果効果の推定法**
- 5. 条件付き平均因果効果（CATE）の推定法 １：メタ学習器、２：二重機械学習
- 6. CATEの推定法３：決定木と決定森、深層学習に基づく方法
- 7. 構造的因果モデルとバックドア基準
- 8. 因果探索
- 9. 発展的な因果推論手法：フロントドア調整、操作変数法
- 10. 発展的な因果推論手法：代理変数法、回帰不連続デザイン
- 11. 発展的な意思決定理論
- 12. 強化学習
- 13. オフライン強化学習
- 14. バンディット

出欠コード
XXXX

基本的な推定対象：（曝露群/対照群における）平均因果効果

- 推定対象
  - 平均因果効果（Average Treatment Effect; **ATE**）
    - $\tau = \mathbb{E}[y_1 - y_0]$
  - **曝露群**（サンプルのうち$a = 1$の群）における因果効果
    （Average Treatment Effect on the Treated; **ATT**）
    - $\tau_1 = \mathbb{E}[y_1 - y_0 | a = 1]$
  - 対照群（同$a = 0$の群）における因果効果
    （Average Treatment Effect on the Untreated; ATU）
    - $\tau_0 = \mathbb{E}[y_1 - y_0 | a = 0]$
- 評価指標
  - 推定対象が1次元なので単に数値どうしを比較する

←政策などを適用した効果を定量化したい場合

ATE推定の手法（1/4）

マッチング法：バランスされた標本を作って統計をとる

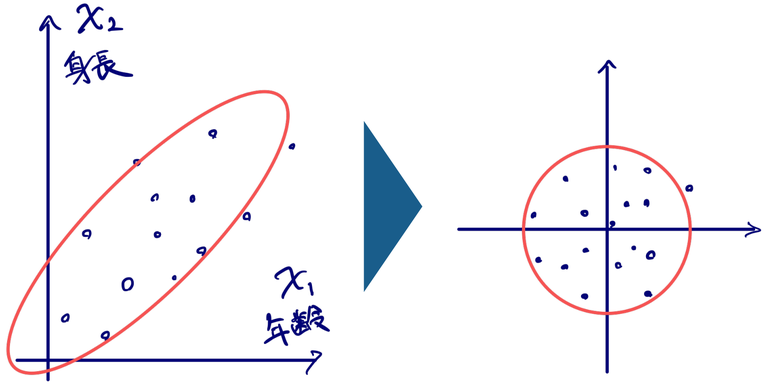
- 行動以外が似ているデータ点のペアを作る
  - 同じような背景 x をもつ人（データ点）を比較すると行動 a の効果だけに
  - 無視可能性 $(y_{a'})_{a'} \perp\!\!\!\perp a \mid x$ の仮定を満たす x が同じ

    $\Rightarrow$ 行動と潜在結果は独立（=RCTと同じ状況を再現）
- 曝露群のデータ点 i に対して距離 $d(x^i, x^j)$ が最小の対照群のデータ点を取得し、ペアとする
  - 曝露群の方がサンプルサイズが小さいことが多いため
  - 十分類似したペアが得られれば（曝露群の分布$p(x \mid a=1)$が対照群の分布$p(x \mid a=0)$に覆われていれば）

    曝露群における平均因果効果（ATT）の推定量となる
- そのペア間の結果の差を平均
  - $\hat{\tau} = \dfrac{1}{|M|} \displaystyle\sum_{(i,j)\in M} y_1^i - y_0^j$ ， Mはペアの集合

ユークリッド距離・マハラノビス距離・傾向スコア差

- マッチングに使用する距離 d はいくつかの選択肢がある
  - ユークリッド距離
    - 標準化（xの各変数平均0分散1にする）しておく
  - マハラノビス距離
    - 説明変数の分散共分散行列 $\hat{\Sigma}$ を用いて
    - $d_M(x, x') = \sqrt{(x - x')^\top \hat{\Sigma}^{-1} (x - x')}$
  - 傾向スコア差・傾向スコアの対数オッズ差（なぜこれで良いかは後述）
    - $d_p(x, x') = \left| \hat{\mu}(a = 1 \mid x) - \hat{\mu}(a = 1 \mid x') \right|$
    - $d_\phi(x, x') = \left| \log \frac{\hat{\mu}(a = 1 \mid x)}{\hat{\mu}(a = 0 \mid x)} - \log \frac{\hat{\mu}(a = 1 \mid x')}{\hat{\mu}(a = 0 \mid x')} \right|$
    - 1次元なので「近傍」を見つけやすい
- 傾向スコアを推定して用いる傾向スコアマッチングは実用面で不安定という指摘もある（King & Nielsen 2019）
  - 曝露群/対照群に分ける際にも使うことになる（同じデータを二度使う）
  - 次元が高すぎなければマハラノビス距離を使っておくのが安全

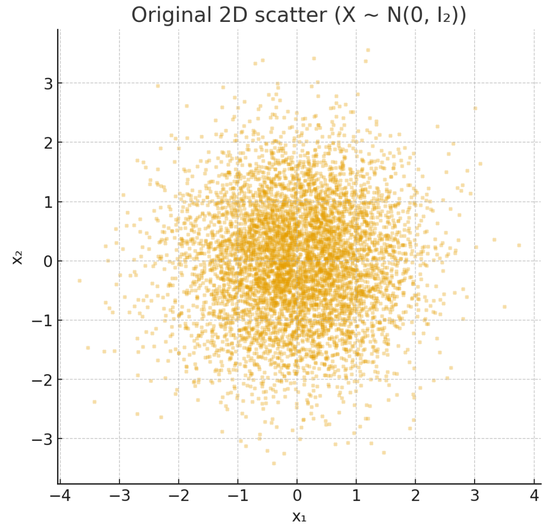


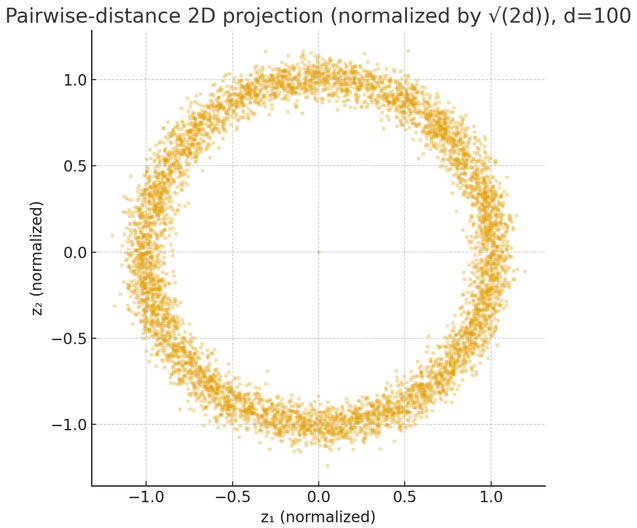
マハラノビス距離は説明変数間の相関を除去した空間でのユークリッド距離

出欠コード
XXXX

次元の呪い：100次元のジャガイモはほぼ皮

- 高次元では体積が距離に対して爆発的に増える
  - 次元 p において距離 r での微小体積は r^{p-1} に比例
- よって多次元正規分布でも原点からの距離が一定の層に集中
- 原点を多少取り替えても同じ
  →データ点間の距離も同じ
- 高次元では「全ての人どうしがほとんど同程度に違って見える」



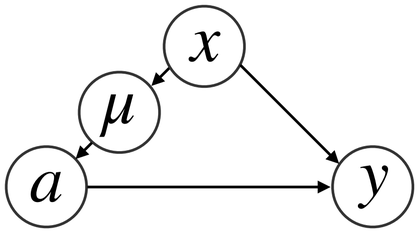


傾向スコアのバランシング特性

出欠コード
XXXX

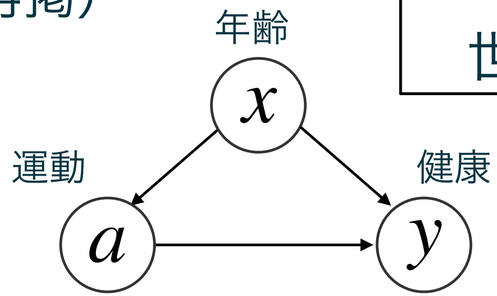
バランシングスコアが同じペアで十分、傾向スコアはその１つ

- バランシングスコアbは、xの関数であって、以下を満たすもの
  - $a \perp\!\!\!\perp x \mid b(x)$
- バランシングスコアをxの代わりに用いても無視可能性を満たせる
  - $(y_{a'})_{a'} \perp\!\!\!\perp a \mid b(x)$
- 定理：傾向スコア$\mu(x) = \mu(a = 1 \mid x)$ はバランシングスコアである
  - 証明略（pdf参照）



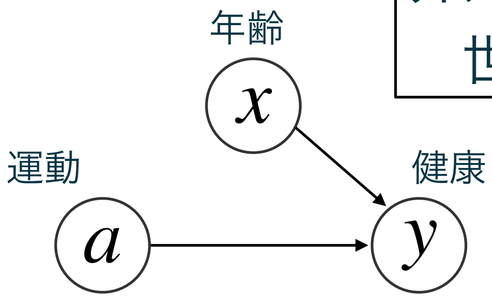
（再掲）

介入前の世界



年齢が高い人が運動する

介入後の世界



年齢が若くても運動する

出欠コード
XXXX

プロパーな損失を用いて学習・クリッピングも

- 傾向スコアの学習は漸近的に真値に一致することが推定バイアスの観点では望ましい
  - $\hat{\mu} = \arg\min_{\mu \in \mathscr{F}} \frac{1}{N} \sum_i \ell(a^i, \mu(a \mid x^i))$
  - 関数クラス $\mathscr{F}$ が真の傾向スコアを含む
  - 損失 ℓ として、期待値が真の傾向スコアにおいて唯一最小値をとる損失（**プロパーな損失**）を用いる
    - 真の傾向スコア値$\bar{a} = \mu(a = 1 \mid x)$に対して以下となる
    - $\bar{a} = \arg\min_{a' \in [0,1]} \mathbb{E}_{a \sim \mu(a|x)} \ell(a, a')$
    - 交差エントロピーや二乗損失はプロパーである
- ただし傾向スコアは逆数で使うことが多いため、小さい値は後段の推定分散の観点から避けたい
  - データ点の重み付け $1/\hat{\mu}(a^i \mid x^i)$ が1点だけ100などになると$N = 100$でもほぼその1点だけを重視した学習になる
  - 20以上は20に切り捨てるクリッピングなどを行うことが多い

ATE推定の手法（2/4）

IPW推定の応用：ホーヴィッツ＝トンプソン推定量

- 行動と潜在結果の積は観測できることを使う
  - $\tau = \mathbb{E}_x \mathbb{E}_{a\sim\mu(a|x)} \mathbb{E}_y \left[\left(\frac{a}{\mu(x)} - \frac{1-a}{1-\mu(x)} \right) y \right]$
    - 第3回でも既出の変換。最後の期待値は訓練データの分布 $p(y,a,x) = p(y|a,x)\mu(a|x)p(x)$ になっていることに注意

$$\leftarrow \mathbb{E}_{y_1,y_0,x}\left[y_1 - y_0\right] = \mathbb{E}_x \mathbb{E}_{a\sim\mu(a|x)} \mathbb{E}_{y_1,y_0} \left[\frac{a}{\mu(x)} y_1 - \frac{1-a}{1-\mu(x)} y_0 \right]$$
$$= \mathbb{E}_x \mathbb{E}_{a\sim\mu(a|x)} \mathbb{E}_y \left[\left(\frac{a}{\mu(x)} - \frac{1-a}{1-\mu(x)} \right) y \right]$$

- μを$\hat{\mu}$で、期待値を平均で置き換えたものがホーヴィッツ＝トンプソン推定量
  - $\hat{\tau}_{\mathrm{HT}} := \frac{1}{N} \sum_i \left(\frac{a^i}{\hat{\mu}(x^i)} - \frac{1-a^i}{1-\hat{\mu}(x^i)} \right) y^i$
- 上記は、結果の期待値 $\bar{y} = \mathbb{E}[y]$ が非ゼロであった場合、 ()内の重み部分のばらつきにより$\bar{y}$の影響が出て推定精度が悪化する
  - 因果推論では通常τよりも$\bar{y}$の方がスケールが大きいので重みの合計が曝露群と対照群で同じでないと$\bar{y}$の影響が大きく出る
    - yが一定値で$\tau = 0$でも重みの合計が違うと$\hat{\tau}$は非ゼロ
  - よって重みを正規化した修正版を用いることが多い（ハーイェク推定量ともいう）
  - $\hat{\tau}_{\mathrm{Hájek}} := \frac{\sum_i \frac{a^i y^i}{\hat{\mu}(x^i)}}{\sum_i \frac{a^i}{\hat{\mu}(x^i)}} - \frac{\sum_i \frac{(1-a^i) y^i}{1-\hat{\mu}(x^i)}}{\sum_i \frac{(1-a^i)}{1-\hat{\mu}(x^i)}}$

G計算：CATE推定量を利用する

- ATEはCATEの平均
  - $\tau = \mathbb{E}\left[y_1 - y_0\right]$
  - $= \mathbb{E}_x[\mathbb{E}[y \mid x, a = 1] - \mathbb{E}[y \mid x, a = 0]]$
- CATE推定量 $\hat{f}(x, a)$ を（なんらか）学習して代入
  - $\hat{\tau}_{\mathrm{gc}} = \frac{1}{N} \sum_i \hat{f}(x^i, a = 1) - \hat{f}(x^i, a = 0)$
    - $f(x, a) = af_1(x) + (1 - a)f_0(x)$ として曝露群の関数f_1と対照群の関数f_0に分ける場合も
      - T(wo)-Learnerという（次回内容）

ATE推定の手法（4/4）

IPWとG計算のハイブリッド：二重頑健推定量（Doubly robust estimator、Augmented IPW; AIPW）

- 以下の２つはそれぞれ$\mathbb{E}[a\,|\,x]$、$\mathbb{E}[y\,|\,a,x]$の推定精度に依存
  - IPW $\hat{\tau}_{\mathrm{HT}} := \frac{1}{N}\sum_i \left(\frac{a^i}{\hat{\mu}(x^i)} - \frac{1-a^i}{1-\hat{\mu}(x^i)}\right) y^i$
  - G計算 $\hat{\tau}_{\mathrm{gc}} = \frac{1}{N}\sum_i \hat{f}(x^i, a=1) - \hat{f}(x^i, a=0)$
- これら2つの推定量を組み合わせる
  - $\hat{\tau}_{\mathrm{dr}} = \frac{1}{N}\sum_i \hat{f}(x^i, a=1) - \hat{f}(x^i, a=0) + \left(\frac{a^i}{\hat{\mu}(x^i)} - \frac{1-a^i}{1-\hat{\mu}(x^i)}\right)(y^i - \hat{f}(x^i, a^i))$
- $\hat{f}$ **または** $\hat{\mu}$ が正確であれば $\hat{\tau}_{\mathrm{dr}}$ も正確（二重頑健性）
  - 実務的には、両方がそこそこ正確であれば $\hat{\tau}_{\mathrm{dr}}$ はかなり正確
  - 理論的には、それぞれの収束率の積が効率的であれば $\hat{\tau}_{\mathrm{dr}}$ の推定も効率的

ATE（ATT）推定法

- 4つの推定法
  - マッチング法
    - 距離の定義によってマッチの取り方がいくつかある
    - 高次元では距離は難しいので1次元の傾向スコアを使う方法も
      - 傾向スコアはバランシングスコアなのでそれが類似ならペアとしてよい
    - 傾向スコアの推定にはプロパーな損失を用いる
  - IPW法（ホーヴィッツ＝トンプソン推定量）
    - 行動と結果の積は観測可能、行動をかけて歪んだ分布を傾向スコアで重み付けて戻す
  - G計算（CATE推定量を用いる）
  - 二重頑健推定量（AIPW）
    - $\hat{\mu}$ と $\hat{f}$ のどちらかが正しければ正しい推定