統計的機械学習 （応用計量分析２）第10回

代理変数法（参考pdf 12.3節）
回帰不連続デザイン（参考pdf 13章）

- CausalML2025_CATE_solution.ipynb
- 問１：R-Learner（DML）の実装間違い
  - $Y - \mathbb{E}[Y \mid X, W] = \theta(X) \cdot (a - \mathbb{E}[a \mid X, W]) + \varepsilon$
  - $\mathbb{E}[Y \mid X, W]$の推定はaを用いない
- 問２：X-Learnerの有利条件
  - 曝露・統制間のサンプル不均衡など
- 問３：任意の手法の有利・不利

```
def fit(self, X, T, Y):
    if self.cv > 1:
        m_hat, e_hat = self._oof(X, T, Y)
    else:
        XT = np.column_stack([X, T.reshape(-1, 1)])
        # self.m.fit(XT, Y)  # 間違い
        self.m.fit(X, Y)  # 正解
        self.g.fit(X, T)
        # m_hat = self.m.predict(XT)  # 間違い
        m_hat = self.m.predict(X)  # 正解
        e_hat = clip01(self.g.predict_proba(X)[:, 1])
    y_tilde = Y - m_hat
    t_tilde = T - e_hat
    eps = 1e-3
    z = y_tilde / (t_tilde + np.sign(t_tilde) * eps)
    w = t_tilde ** 2 + 1e-6
    self.tau.fit(X, z, sample_weight=w)
    return self
```

振り返り

未観測交絡因子によりバックドア基準を満たせない場合の手法を2つ紹介
フロントドア調整は媒介変数で分解して総合、操作変数法は外生変数をランダマイザとみなす

- 未観測交絡因子があってもバックドア基準を満たせる場合もある
- どうしても満たせない場合、または未観測交絡の疑いがいくらでも出てくる場合は別のアプローチも考えられる
- 2つのアプローチを紹介
  - フロントドア調整は未観測交絡因子から影響を受けていない、かつ$a \to y$の因果効果を完全に媒介する変数を用いて分解して総合
  - 操作変数法は未観測交絡因子と独立かつaを通してしかyに影響しない（かつaには影響する）変数をクジ（ランダマイザ）として、その影響によるaの変動のyへの伝播を分析

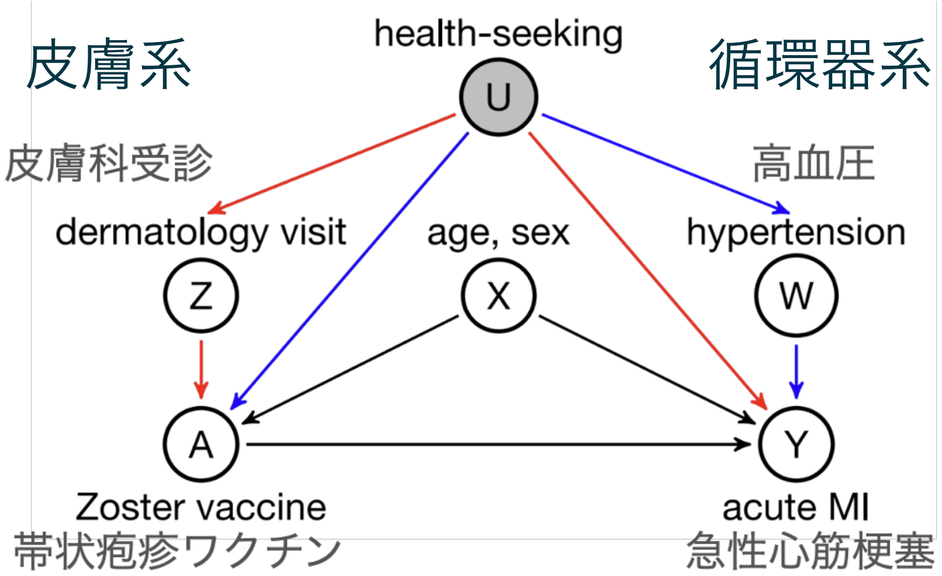
代理変数法・回帰不連続デザイン

- 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
  - CATEの推定法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. 構造方程式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、**回帰不連続デザイン、代理変数法**
- 10. 続き
- 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
- 15. まとめ

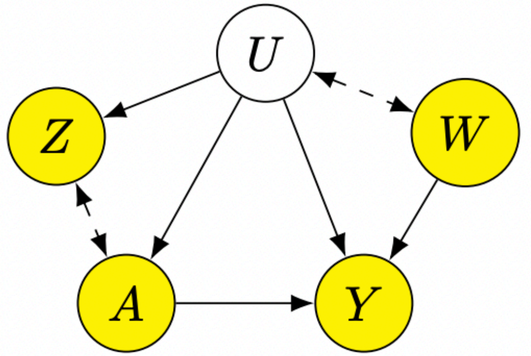
代理変数法とは

未観測交絡因子によるバックドアパスの効果を代理変数を通して推定して調整

- 例：帯状疱疹ワクチンと心筋梗塞
  - 元々高血圧の人が健康に対する意識が高く、ワクチンを打ちやすく、ワクチンとは無関係に高血圧のため心筋梗塞になりやすいかもしれない
- 未観測交絡因子uから(a, y)と同様に影響を受けていて、行動aと結果yにそれぞれ関連する2つの因子を未観測交絡因子uの代理として用いる
  - 行動側のz（陰性対照曝露）はyと独立 $y \perp\!\!\!\perp z \mid a, u$ と仮定
  - 結果側のw（陰性対照結果）はa, zと独立 $w \perp\!\!\!\perp (a, z) \mid u$ と仮定
  - → 帯状疱疹は皮膚の病気なので皮膚科受診zと関係しそう、心筋梗塞と高血圧は循環器系なので相互にはuを通してしか関連しない
- 陰性対照z, wはuの情報を十分に受け継いでいると仮定
- このときATE（xがあればCATE）が識別可能



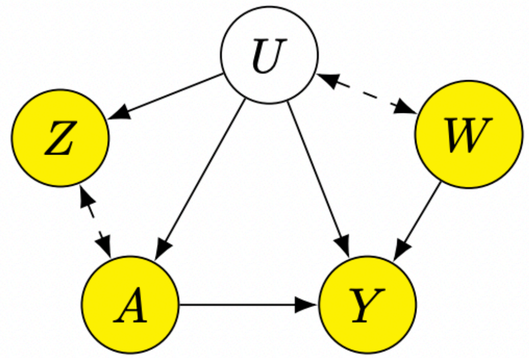
因果関係は一部逆でもよい



代理変数法（線形モデルの場合）

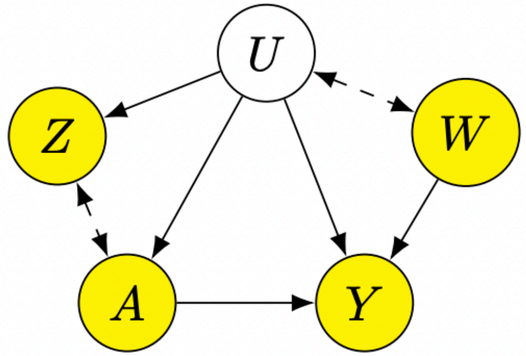
(z, a)から予測したwをuの代わりに説明変数として用いる

- 推定対象：下記のβ_{ay}
  - $\mathbb{E}[y \mid a, z, u] = \beta_{y0} + \beta_{ay} a + \beta_{uy} u$　（独立性$y \perp\!\!\!\perp z \mid a, u$から$z$の項は無し）
  - uがあればバックドア基準を満たすため、β_{ay}は因果効果
- 同様にwについては、以下で書ける
  - $\mathbb{E}[w \mid a, z, u] = \beta_{w0} + \beta_{uw} u$　（独立性$w \perp\!\!\!\perp (a, z) \mid u$から$a, z$の項は無し）
- 上2式を $p(u \mid a, z)$ について期待値を取り$\mathbb{E}[u \mid a, z]$を消去すると
  - $\mathbb{E}[y \mid a, z] = \beta_{y0} + \beta_{ay} a + \dfrac{\beta_{uy}}{\beta_{uw}} \left(\mathbb{E}[w \mid a, z] - \beta_{w0} \right)$
  - 係数を整理して $\mathbb{E}[y \mid a, z] = \beta'_{y0} + \beta_{ay} a + \beta_{wy} \mathbb{E}[w \mid a, z]$
- 結局、wをa, zに回帰して得た予測値$\hat{w}$をuの代わりに用いて回帰する



余計な相関を代理変数で除去

- 用いている式
  - $\mathbb{E}[y \mid a, z] = \beta_{y0} + \beta_{ay} a + \frac{\beta_{uy}}{\beta_{uw}} \left(\mathbb{E}[w \mid a, z] - \beta_{w0} \right)$
- 知りたい因果効果β_{ay}とバックドアの相関AU×UYを分離したい
- AU×UWは$\mathbb{E}[w \mid a, z]$のaの係数として推定可能
  - しかしUY/UWだけスケールが間違っている
- UY,UWそれぞれは推定不可能だが、その比は推定可能
  - UY/UW = (ZU+AU)×UY / (ZU+AU)×UW
  - = $\mathbb{E}[y \mid a, z] \big/ \mathbb{E}[w \mid a, z]$



回帰不連続デザイン

正値性 $\mu(a \mid x) > 0$ が満たされない場合

決定論的にaが変化する点の周辺を比較する

回帰不連続デザイン（Regression Discontinuity Design; RDD）とは

割り当てカットオフ点周辺はランダムとみなせることを利用

- 正値性 $\mu(a\,|\,x) > 0 \quad \forall(x, a)$ が満たされない場合

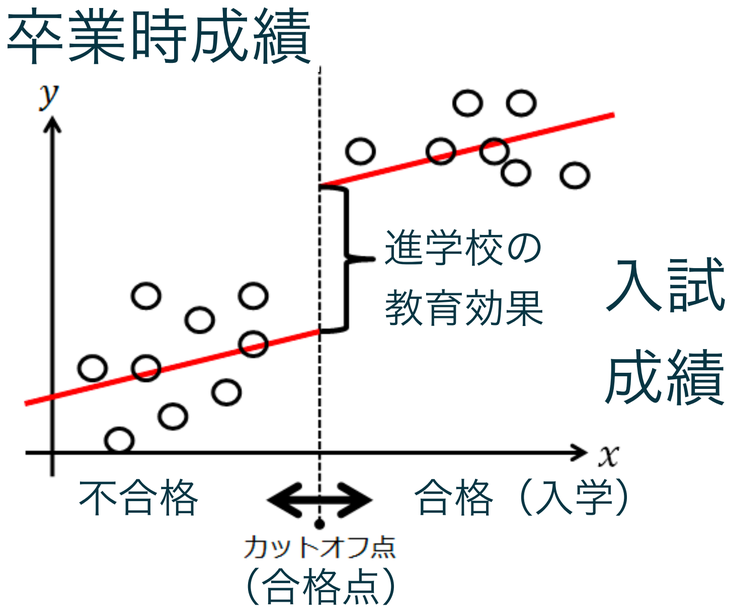
  → x 全体では推定のしようがない

  - 中でも決定論的に割り当てが決まる場合を考える
  - $a = \begin{cases} 1 & g(x) \geq c \\ 0 & g(x) < c \end{cases}$
- カットオフ点 $z := g(x) = c$ の周りではほとんどランダムにaが決まる

  → それらの差は（$z = c$での）因果効果

  - カットオフ点における平均因果効果（ATE at the cutoff）

    $\tau_c = \mathbb{E}[y_1\,|\,z = c] - \mathbb{E}[y_0\,|\,z = c] = \lim_{z \downarrow c} \mathbb{E}[y\,|\,z] - \lim_{z \uparrow c} \mathbb{E}[y\,|\,z]$
  - 例：進学校の教育効果を、入試成績がギリギリ合格・不合格の群を比較



https://ja.wikipedia.org/wiki/回帰不連続デザイン

RDDの詳細

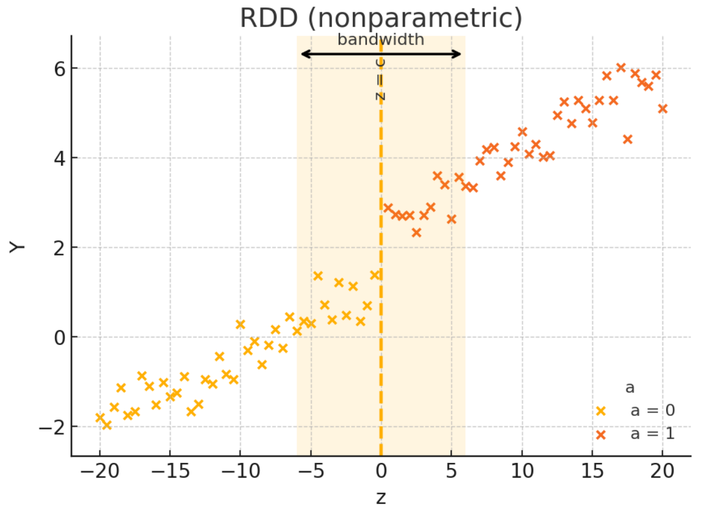
出欠コード
XXXX

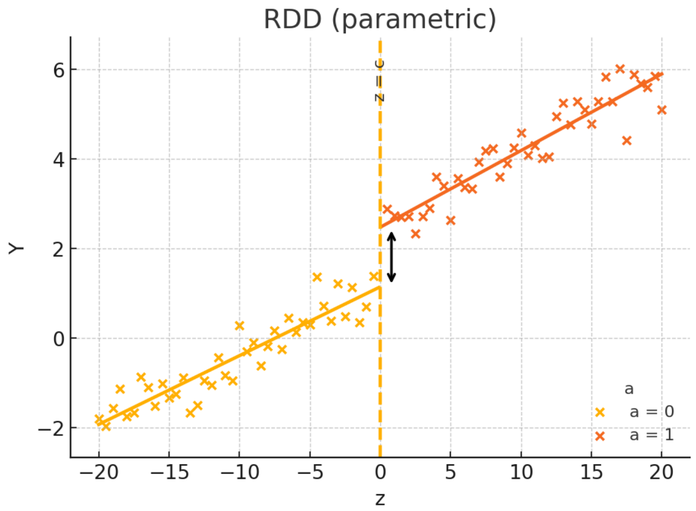
真の関数が連続的＋カットオフ点周りの正値性を仮定
関数をフィットするパラメトリック法とカットオフ付近を平均するノンパラメトリック法

- 仮定
  - 合格点付近が「割り当て以外同様の人」と見なせる必要あり

    $$\lim_{z \downarrow c} \mathbb{E}[y_a \mid z] = \lim_{z \uparrow c} \mathbb{E}[y_a \mid z] \quad \forall a \in \{0,1\}$$

    - 18歳から大学に入るしパチンコもできる場合、
      大学入学の効果かパチンコの効果かわからなくなる
  - カットオフ点周りにサンプルが存在： $p(z = c) > 0$
- 手法２：ノンパラメトリックな方法
  - 一定のバンド幅 δ をとり、その領域の平均間を比較
- 手法１：パラメトリックな方法
  - カットオフ点以後かどうか（$a = I(z \geq c)$）で分割
  - a ごとに重回帰等をフィット





まとめ

未観測交絡因子への対処法（その３）代理変数法
正値性を満たさない決定論的行動割り当て状況の手法RDD

- 代理変数法は未観測交絡を通したバックドアパスの相関を代理変数を用いて推定して差し引く
- RDDはカットオフ点周りで決定論的に行動の割り当てが決まる場合に、その周辺を比較することでカットオフ点周りの因果効果を推定する
  - パラメトリックな手法とノンパラメトリックな手法