記述統計量の推論

経済学のための機械学習入門

川田恵介

特定の変数間の関係性理解

- X が同じようなグループ内で、D と Yの関係性を推定する
  - 経済学における中心的な実証課題

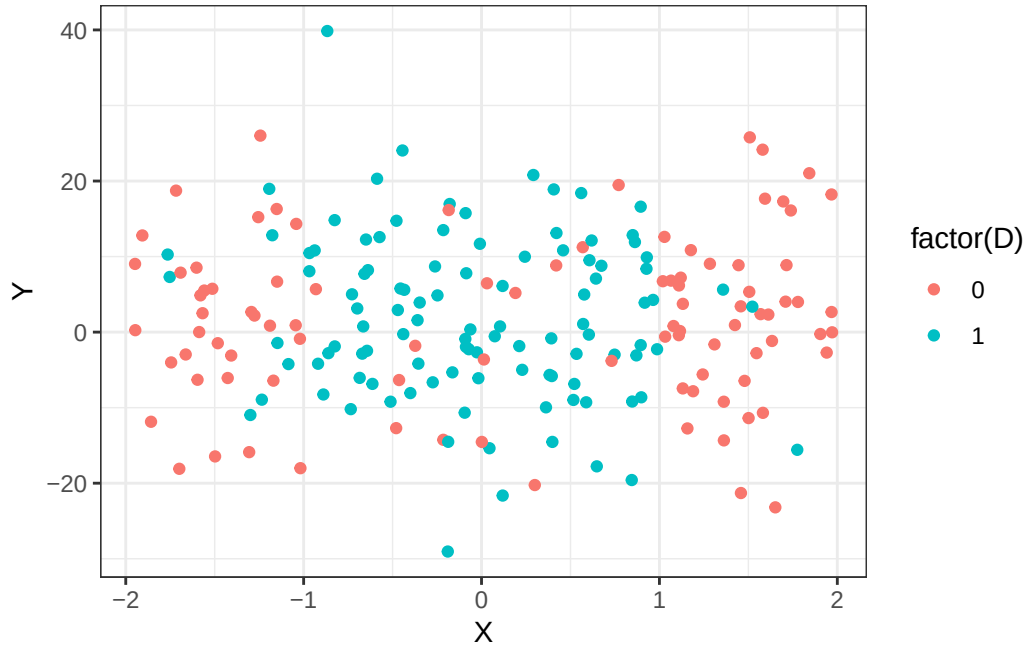
応用例

- 同一学歴 (X) 内男女間 (D) 賃金格差 (Y)
- 最低賃金 (D) が就業率 (Y) に与える因果効果の推定
  - $X =$ 地域の経済状態など
- キャッシュバックキャンペーン (D) が、新規携帯電話契約 (Y) に与える因果効果推定
  - $X =$ 個人の背景

数値例

- 「格闘ゲームをプレイした経験間で、主観的幸福度はどの程度異なるのか？」
  - 年齢と主観的幸福度、格闘ゲームのプレイ経験には強い相関がされるので、“コントロール”
- 母集団
  - 格闘ゲームのプレイ経験があるグループの方が、主観的幸福度は高い
  - 40 歳前後が最も格闘ゲームのプレイ経験は高い
  - 年齢と主観的幸福度の間には、U 字の関係がある

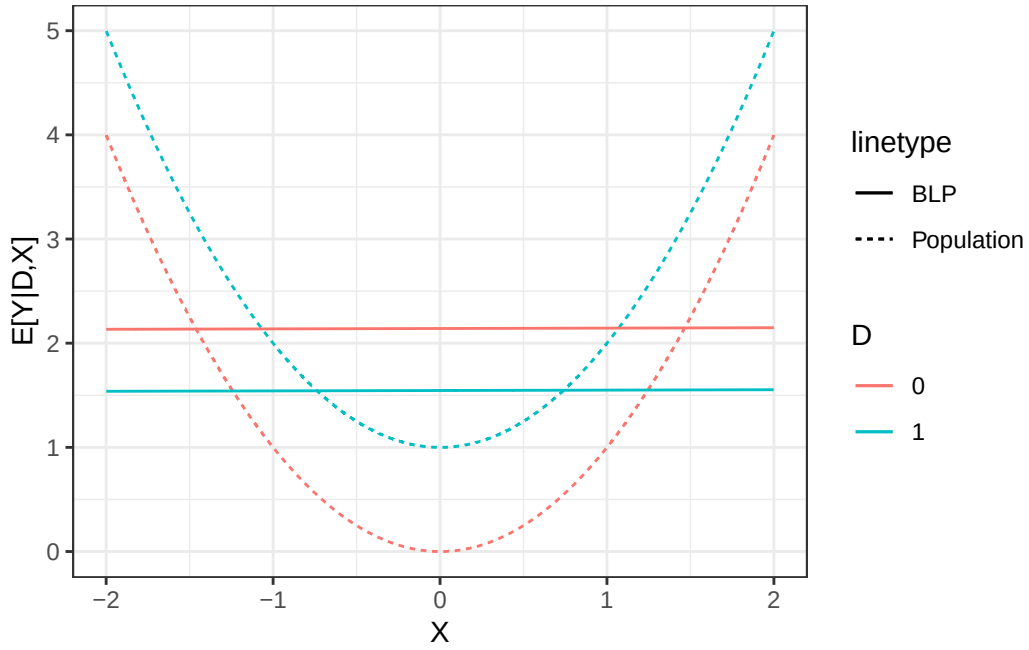
数値例



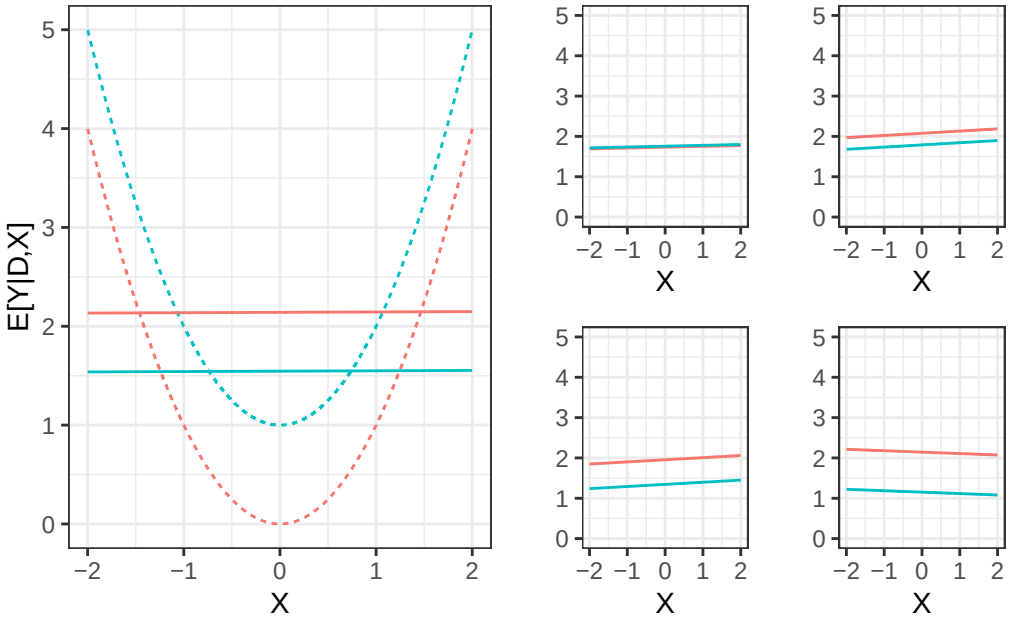
OLS 推定

- 真の関係性は、$Y \sim D + X^2$
- $Y \sim \beta_0 + \beta_1 D + \beta_2 X$ を回帰
  - どのような BLP を推定することになるのか?

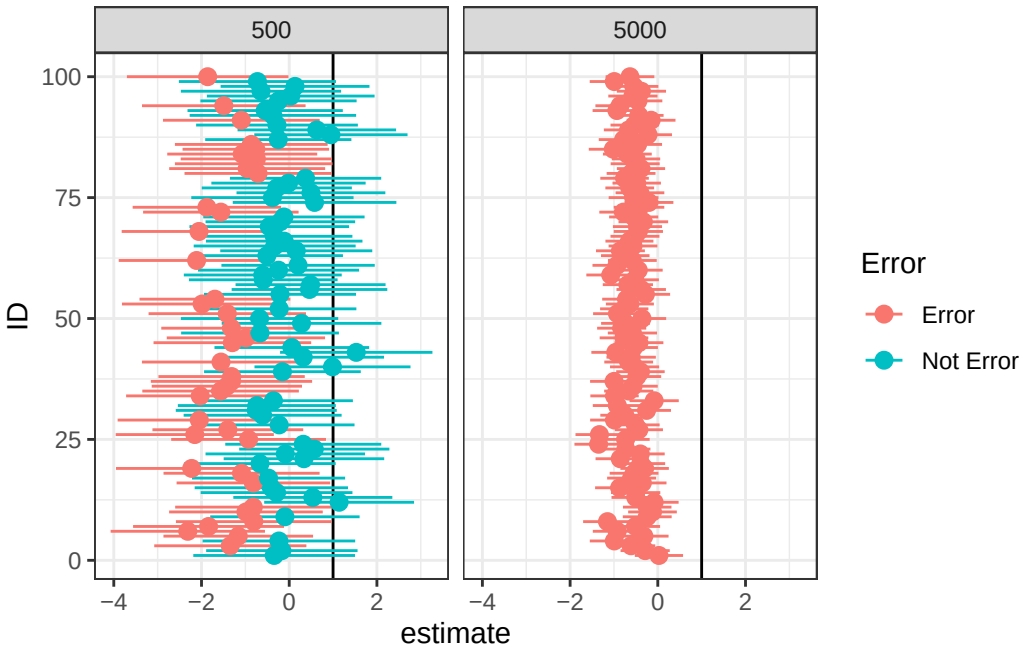
数値例



数値例



数値例: 信頼区間



集計した条件付き平均差

条件付き平均差

- $$\tau_P(X) = E_P[Y|D=1,X] - E_P[Y|D=0,X]$$
- 事例数が十分あり、かつ X の取りうる値が限られていれば、サブサンプル平均差として推定できる
  - 多くの応用例で、難しい

集計した条件付き平均差

- $$E_P[\tau_P(X)]$$
- 事例数が限られていたとしても、推定できる可能性は高い

線形モデル

- $E_P[Y|D,X] \simeq \beta_0 + \tau_P \times D + \beta_1 \times X_1 + .. + \beta_L \times X_L$ を推定する

- 以下のどちらかの条件が成り立てば OK
  - $E_P[Y|D, X] = \beta_0 + \tau_P \times D + \beta_1 \times X_1 + .. + \beta_L \times X_L$ となるような τ_P, β が存在する
    - モデルが正しく定式化されている
  - D がランダムに決定されている

FWL 定理

- **F**risch–**W**augh–**L**ovell 定理
- OLS の推定結果は以下の推定結果と一致する

1. $Y \sim X_1, .., X_L$, $D \sim X_1, .., X_L$ を OLS で推定し、“予測モデル” $g_Y(X), g_D(X)$ を獲得
2. 予測誤差同士を OLS で推定する $Y - g_Y(X) \sim D - g_D(X)$
3. 係数値 $= D$ の係数値

Well specified model

- g_Y または g_D が well-specified であれば、D は条件付き平均差の優れた推定値
  - 信頼区間を計算可能
- Well-specified なモデルを設定するのは、非常に難しい
  - 機械学習の活用

Partialling-out 推定

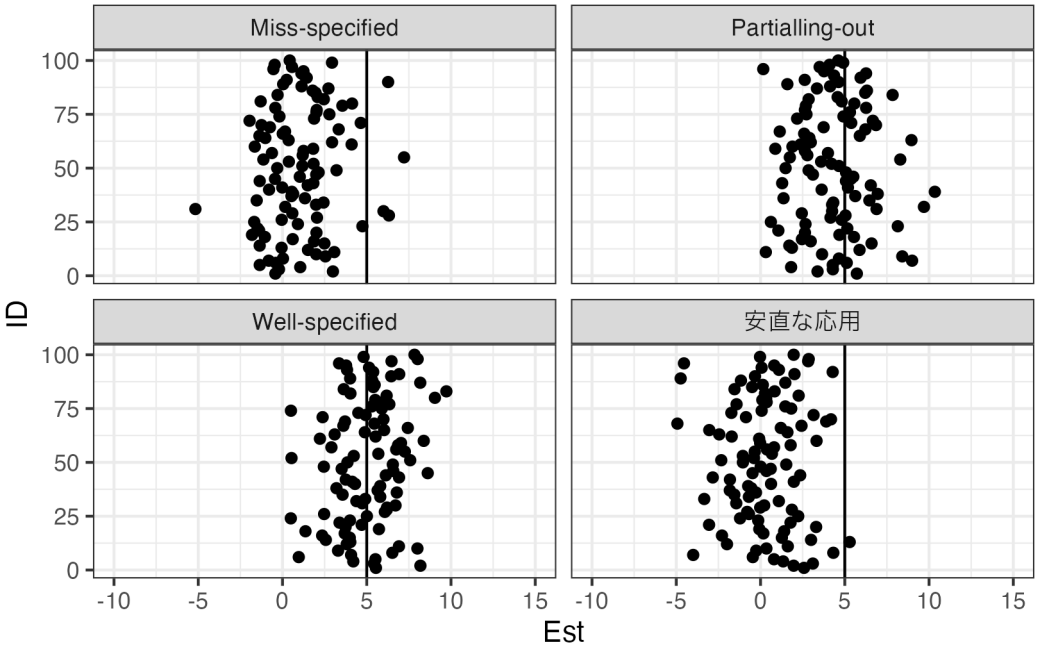
1. Y, D の予測モデル $g_Y(X), g_D(X)$ を、**何らかの方法**で交差推定する
2. 予測誤差 $Y - g_Y(X), D - g_d(X)$ を単回帰する $(Y - g_Y(X) \sim D - g_D(X))$
3. 単回帰の係数 $=$ 条件付き平均差の集計値
   - “高性能” なアルゴリズムを使用できていれば、OK

補論: 機械学習の安易な応用の問題点

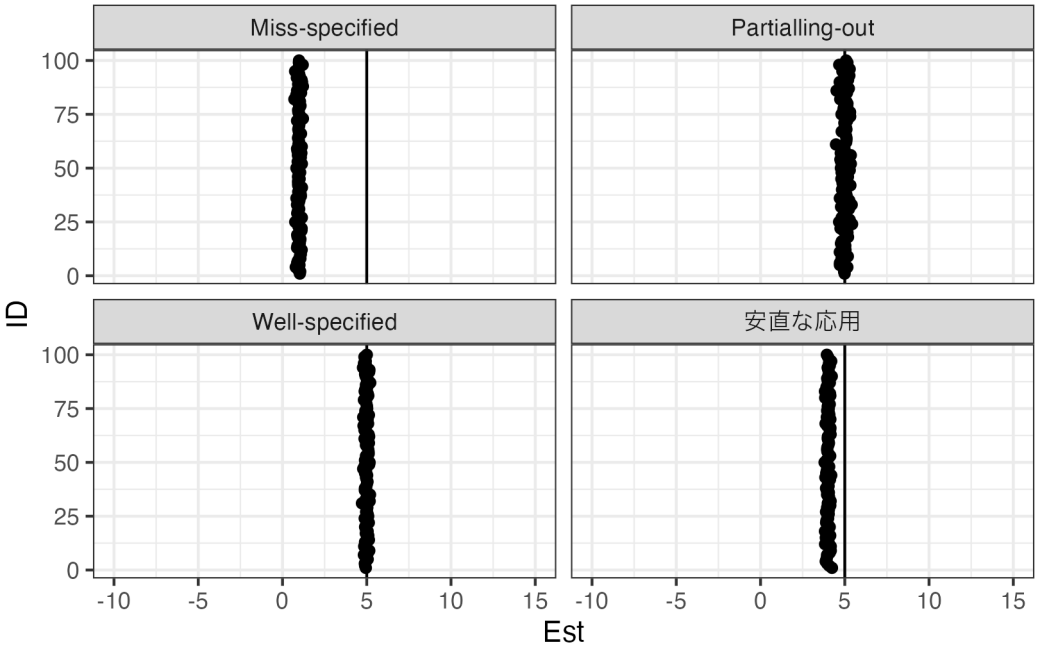
- 古典的 (かつ批判の多い) アイディア

1. D, X から Y の予測モデル $(g_Y(D, X))$ を作る
2. 予測値の差 $g_Y(1, X) - g_Y(0, X)$ の平均値を推定値とする

性質: 小規模サンプル



性質: 大規模サンプル



収束

- 現実的な事例数の元で、真の値 (平均差) を中心とした正規分布で近似できることを保証したい
  - OLS: Well-specified でないかぎり、中心が常にズレる
  - 安易な応用: 極めて大きいサンプルサイズがないと、中心がズレる
- このため Partialling-out が推奨される

まとめ

- BLP $\neq$ 条件付き平均差
  - 一般に OLS を条件付き平均差の推定に使うことは不適切
  - 実験データに近いのであれば、大きな問題はない
- 教師付き学習も、収束の遅さが問題
- 推奨は Partialling-out 推定

補論: OLS

- OLS は引き続き用いられる
  - ランダム化実験データであれば、D は X と独立して決定されているので、大きな問題は起きない
- 明らかに相関しているデータで同じ処理を行うと問題