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

母集団の推論

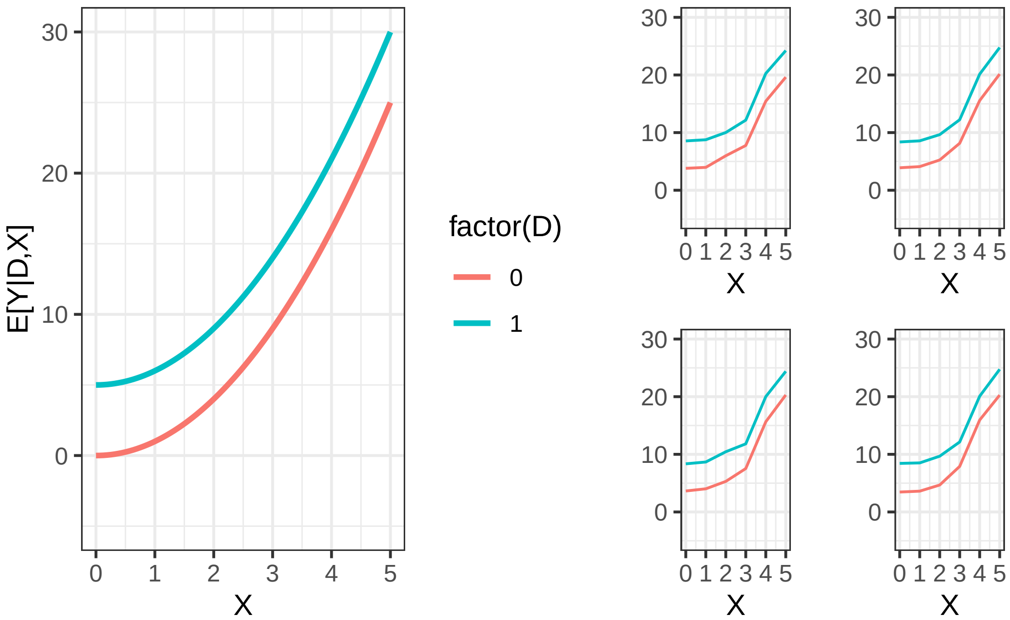
- 母集団の特徴を” 点 (値)“ではなく、” 棒 (区間)“として推定する
  - より信頼できる結果を得られる

データ分析の目的

- データの背後にある集団 (社会?) の理解とその応用
- 推定結果へ” 高い信頼性” が要求されるケースも多い
  - High-Stakes Decision making への応用: 政府の政策/企業の戦略/個人の人生設計の根拠などなど
- 教師付き学習は、その役割を果たせるか？

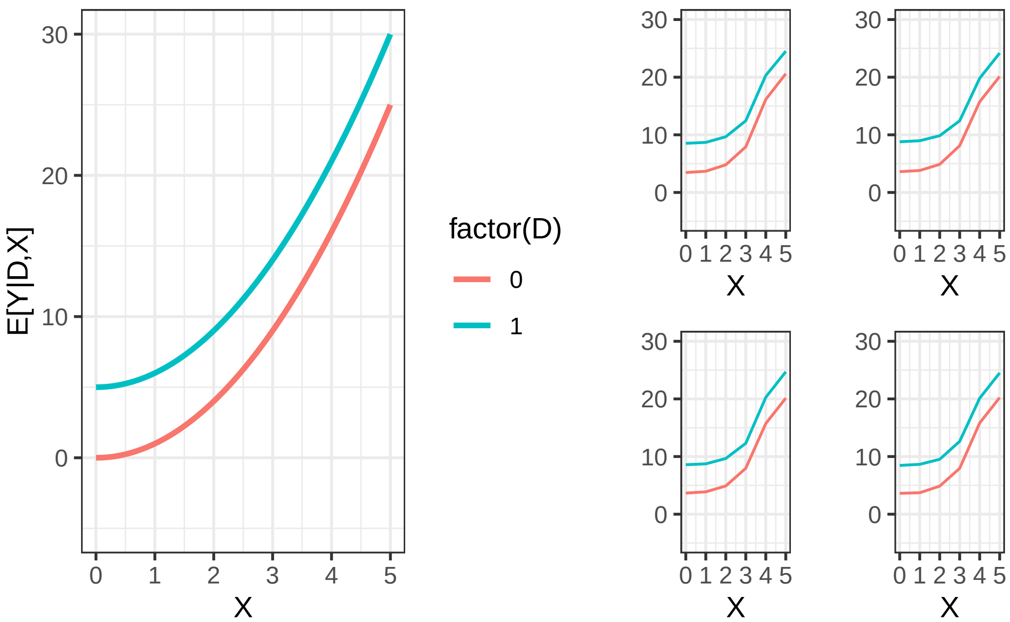
教師付き学習の問題点

- 5000 サンプル



教師付き学習の問題点

- 50000 サンプル



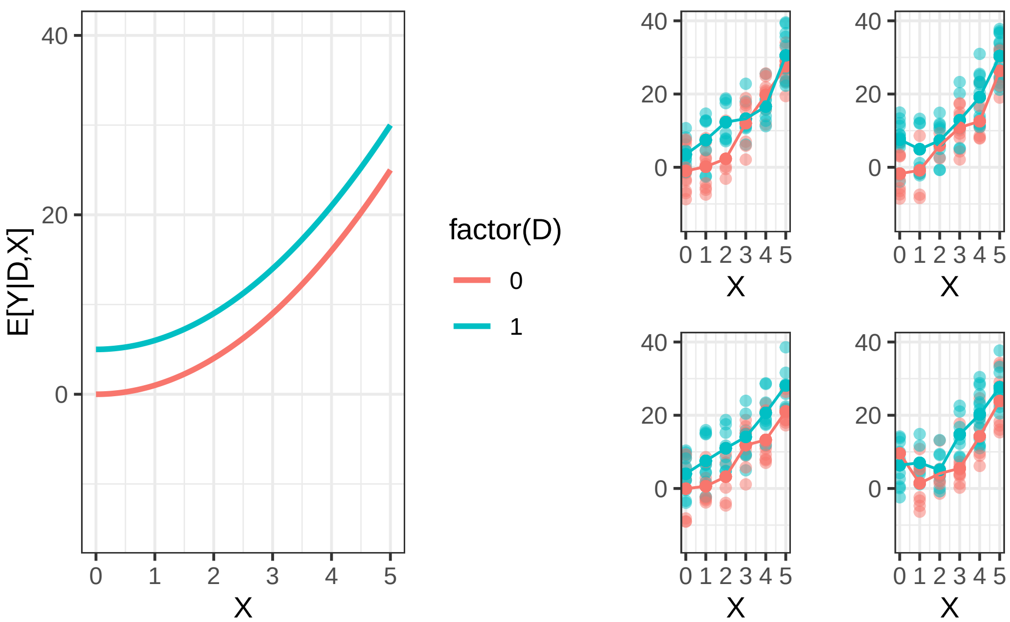
まとめ

- 代表的な教師付き学習による予測モデルは、母平均とは" 一致しない"
  - 事例数が" 無限大" になる必要がある
- 限られた事例数において推定されるモデルと、母平均関数はどのように乖離するのか？
  - 極めて不透明
- 推定されたモデルから、母平均の特徴を" 推論" することは極めて困難

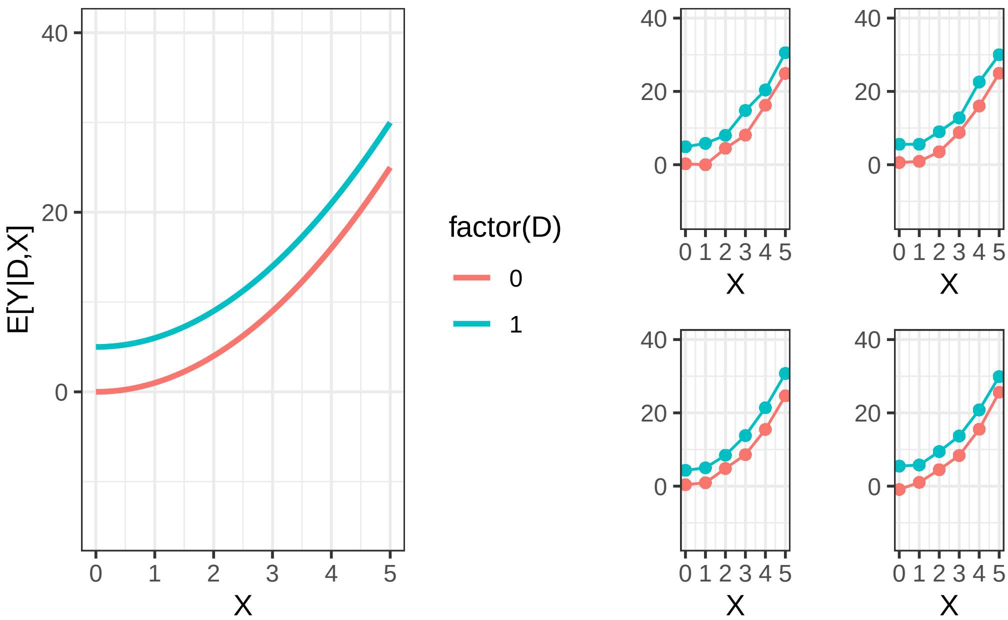
伝統的な計量経済学のアプローチ

- 母平均との関係性が" 明確" にできる、母集団の" 要約 (記述) 統計量" を推定
- より" 信頼性の高い" 推定が可能
  - 機械学習を応用すればさらに改善

要約: 100 事例



要約: 1000 事例



Best Linear Projection

- 最善の線形近似: $E_P[Y|X]$ を可能な限り再現した**仮想的な**一直線

$$BLP(X) = \beta_0 + \beta_1 X_1 + .. + \beta_L X_L$$

`- $\beta:$` **パラメータ**

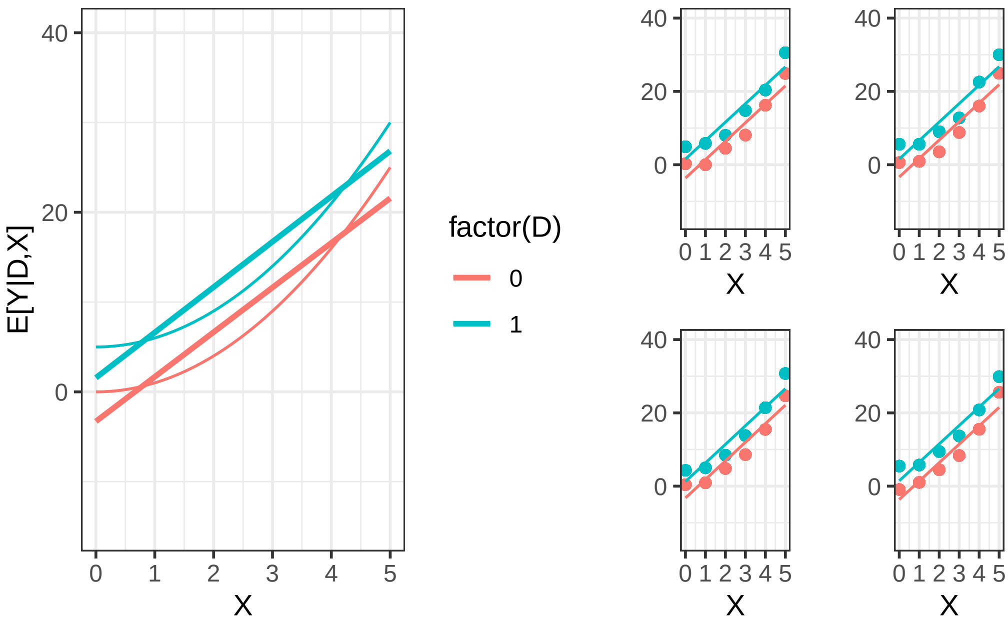
- 母集団において、以下を最小化するように設定

$$E_P[(Y - BLP(X))^2]$$

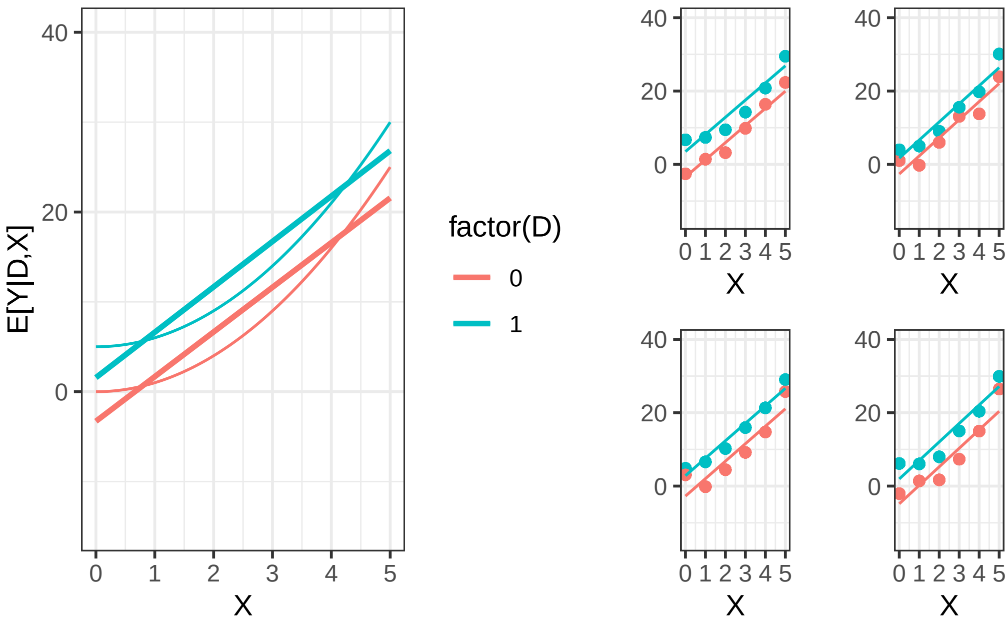
BLP の推定

- データに合うように $g(X) = \beta_0 + .. + \beta_L X_L$ を推定する
  - 二乗誤差 $E[(Y - g(X))^2]$ の最小化
- 事例数に比べて、十分に単純なモデル (β の数が少ない) であれば、非常に優れた方法
  - 計量経済学や統計学の講義で確実に学ぶ

BLP の推定: 1000 事例



BLP の推定: 200 事例

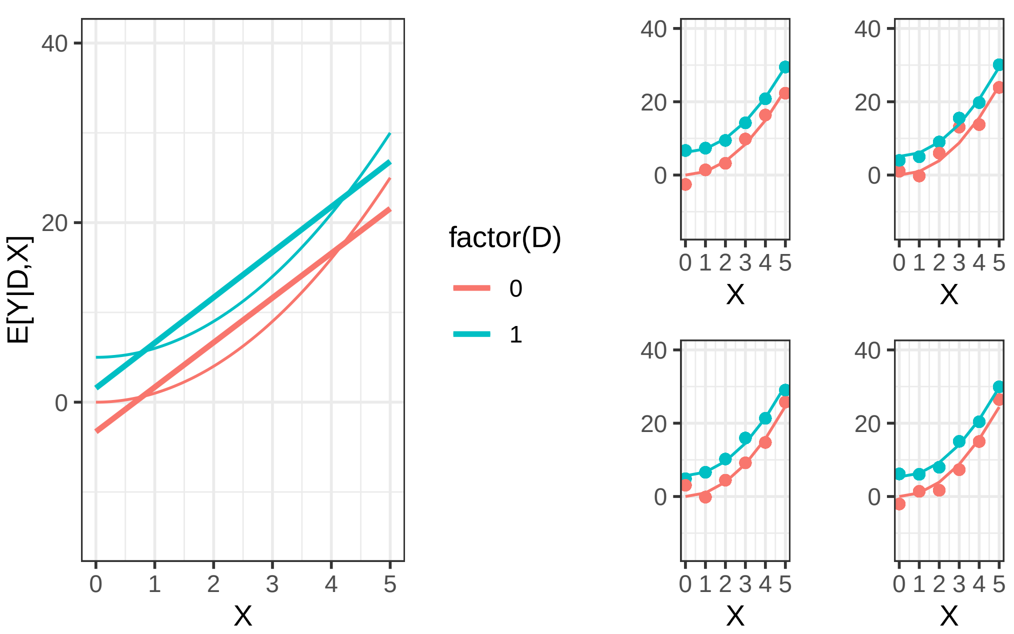


Well-specified model

- **特殊**なケース: $E_P[Y|X] = BLP(X)$
  - 入門的な教科書が想定
- 例: $E_P[Y|X] = X^2$ ならば

$$BLP(X) = \underbrace{\beta_0}_{=0} + \underbrace{\beta_1}_{=0} X + \underbrace{\beta_2}_{=1} \underbrace{X^2}_{:=X_2}$$

母平均の推定: 200 **事例**



まとめ

- データではなく、研究者が推定するモデルを設定する
- 母平均関数そのものではなく、単純化した BLP を推定している
  - ハードルの低いゴール:「現実が一直線だ」と仮定しているわけではない
  - “安定” する
  - データへの依存度が減り、異なる分析者間で推定結果が似てくる

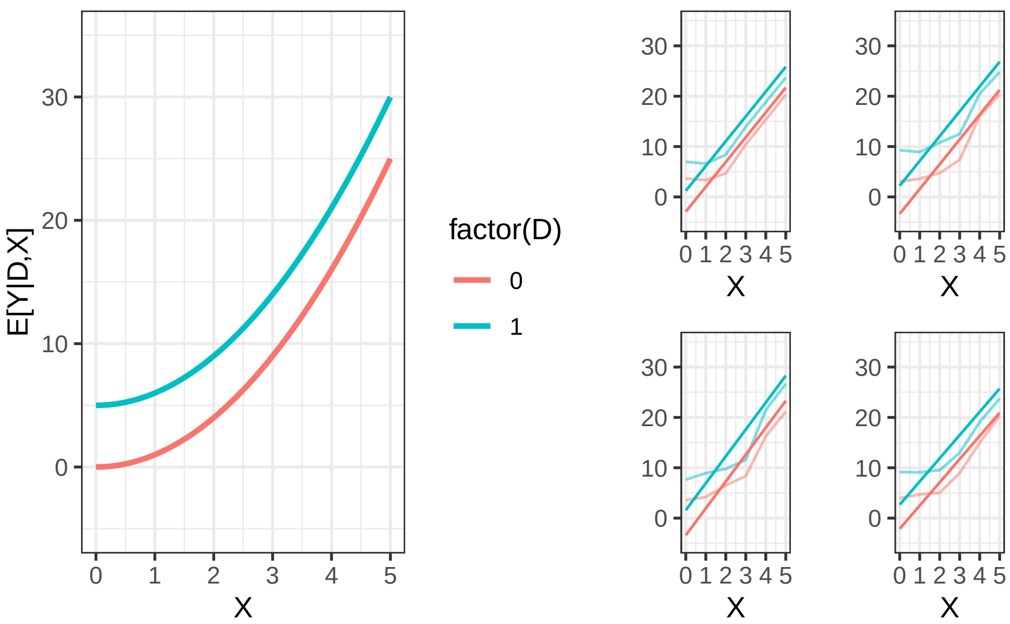
- 適切かつ単純な BLP を設定できれば、母平均を優れた推定値
  - 非現実的!?

予測研究への応用

- 教師付き学習の一つの手法
  - スムーズな母平均関数に対する、有力な手法

例: RandomForest VS OLS

- 500 サンプル



性質の比較

- $X=4$ を予想

$$Y-g(X=4)=\underbrace{Y-E_P[Y|X=4]}_{\text{どうしようもない個人差}}$$

$$\underbrace{+E_P[Y|X=4]-\underbrace{g_{\infty}(X=4)}_{\text{事例数が無限大ある場合の予測値}}}_{\text{(母集団における)近似誤差}}$$

$$+\underbrace{g_\infty(X=4)-g(X=4)}_{\text{推定誤差}}$$

RandomForest/決定木

$$Y-g(X=4)=Y-E_P[Y|X=4]$$

$$+\underbrace{E_P[Y|X=4]-g_\infty(X=4)}_{\simeq 0}$$

$$+\underbrace{g_\infty(X=4)-g(X=4)}_{?}$$

シンプルな OLS

$$Y-g(X=4)=Y-E_P[Y|X=4]$$

$$+\underbrace{E_P[Y|X=4]-g_\infty(X=4)}_{?}$$

$$+\underbrace{g_\infty(X=4)-g(X=4)}_{\sim\text{正規分布}}$$

まとめ

- データ主導でモデルを設定する: RandomForest/Tree
  - 事例数が増えれば、勝手に複雑なモデルが推定されるので、誤差は一般に少ない
    * 予測研究における明確な利点
  - 推定誤差がどうなるかわからない

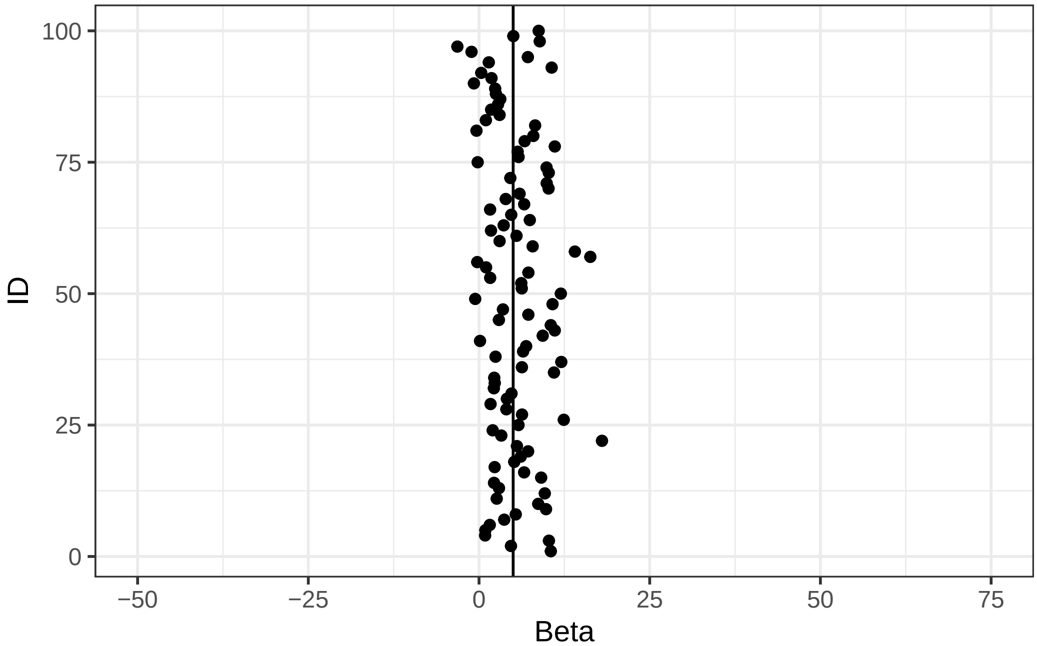
大標本理論に基づく推論

- ランダムサンプリングの仮定が持つ含意は？
- ある程度サンプルサイズが大きければ (典型的には 200 事例)、母集団への詳細な仮定なしで、BLP を推論できる

Well-specified model

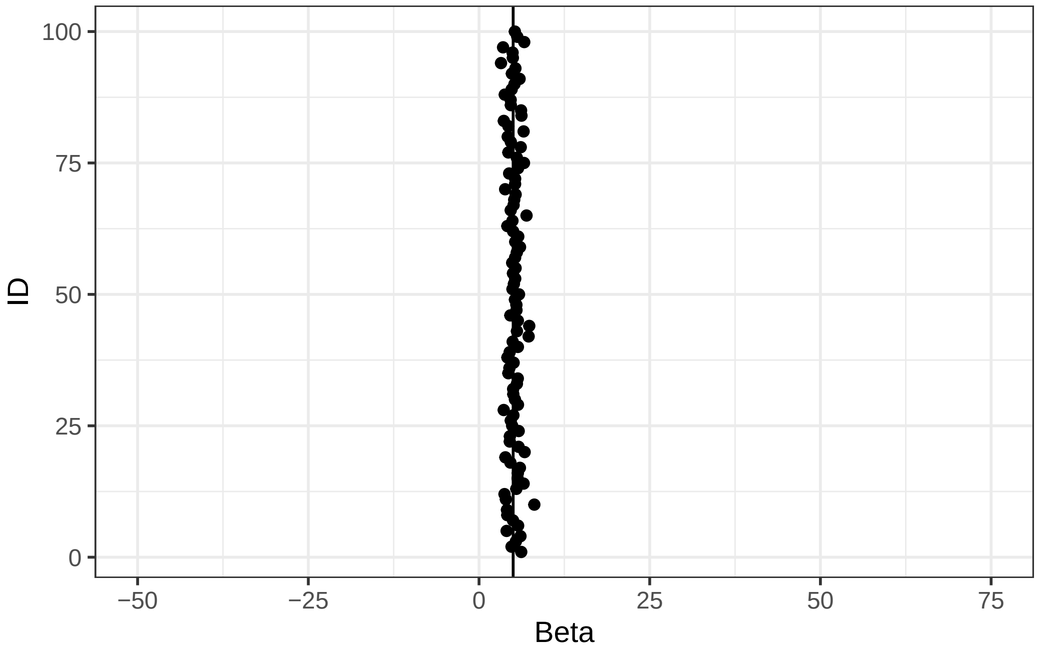
- “入門教科書” 的な問題設定
- $g(D, X) = \beta_0 + \beta_D D + \beta_1 X_1 + .. + \beta_L X_L$
- $E_P[Y|D, X] = g(D, X)$ を達成する β^P が存在

N = 10

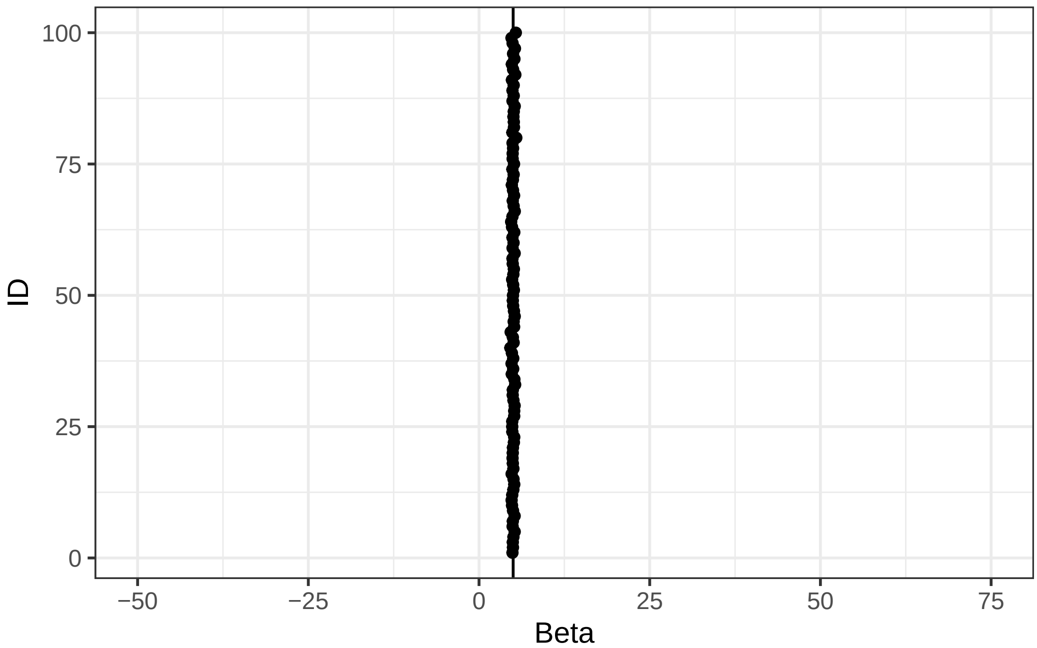


- 100 名研究者が独立して研究 (事例数 = 10)

N = 200


100
75
50
25
0
ID
−50
−25
0
25
50
75
Beta

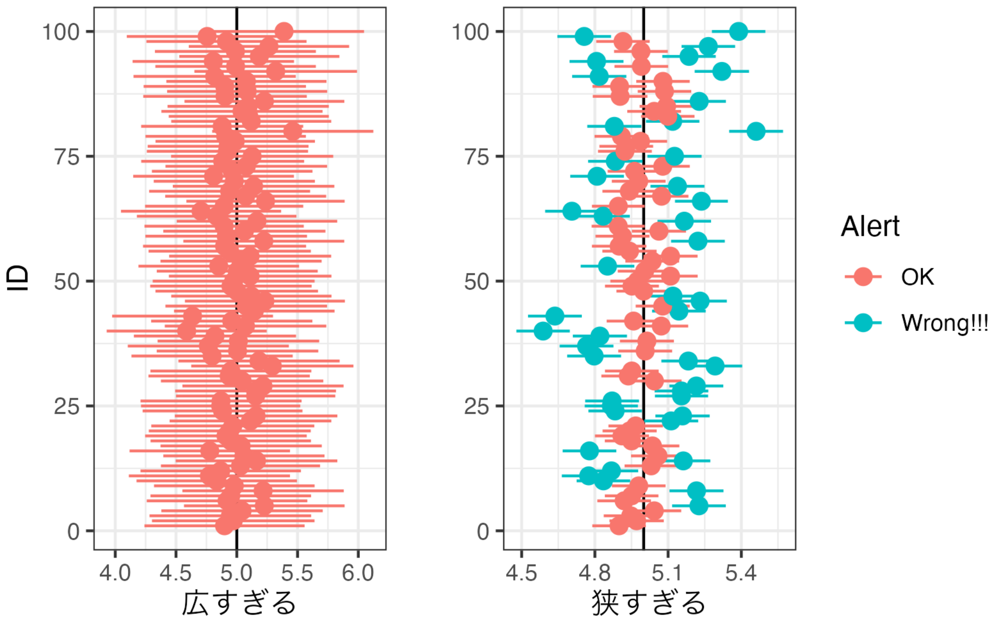
N = 5000


100
75
50
25
0
ID
−50
−25
0
25
50
75
Beta

信頼区間

- 「推定値 = 真の値」を前提に議論を始めると、“100%” 間違う
  - 事例数が無限でない限り、絶対に乖離
  - 独立した研究者間での合意も不可能
- ハードルを下げる
  - 大多数 (典型的には 95%) の研究者について、真の値を含む区間 (信頼区間) を計算する

不適切な区間



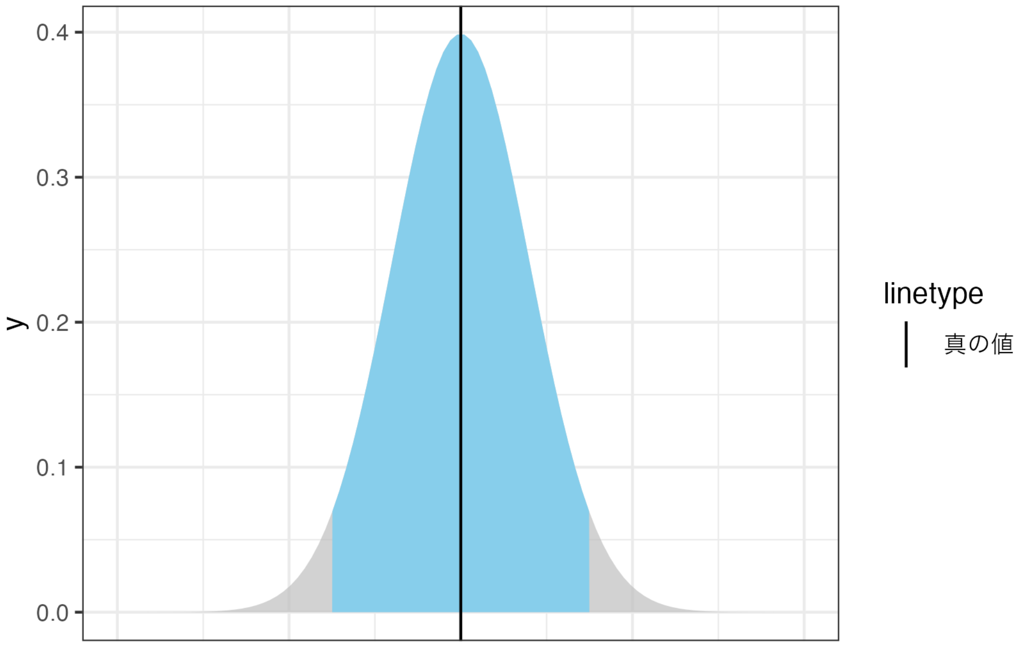
漸近性質の活用

- 信頼区間を計算するには、推定値の分布 (研究者間の散らばり具合) への仮定が必要
  - 本当の分布は、母分布に依存
- 代表的なアプローチは、サンプリング方法への仮定 (ランダムサンプリング)“のみ” に基づいて導出される、漸近性質 (サンプルサイズがある程度大きければ、近似的になりたつ性質) を活用

漸近正規性

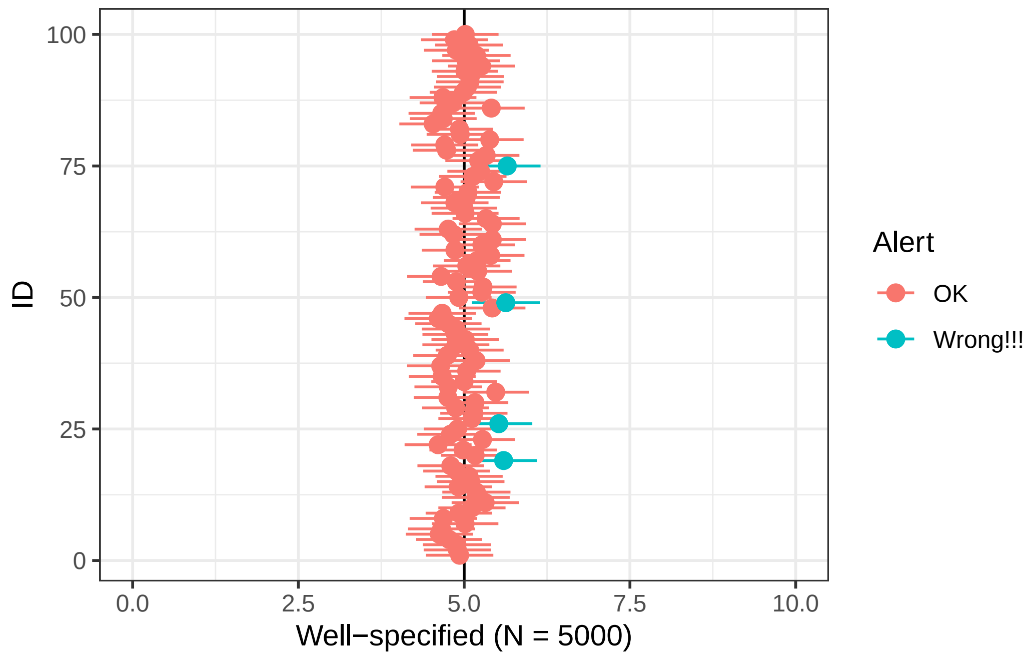
- サンプルサイズがある程度大きければ、正規分布で近似できる
    - 真の値よりも、“早め” に収束する

漸近正規性

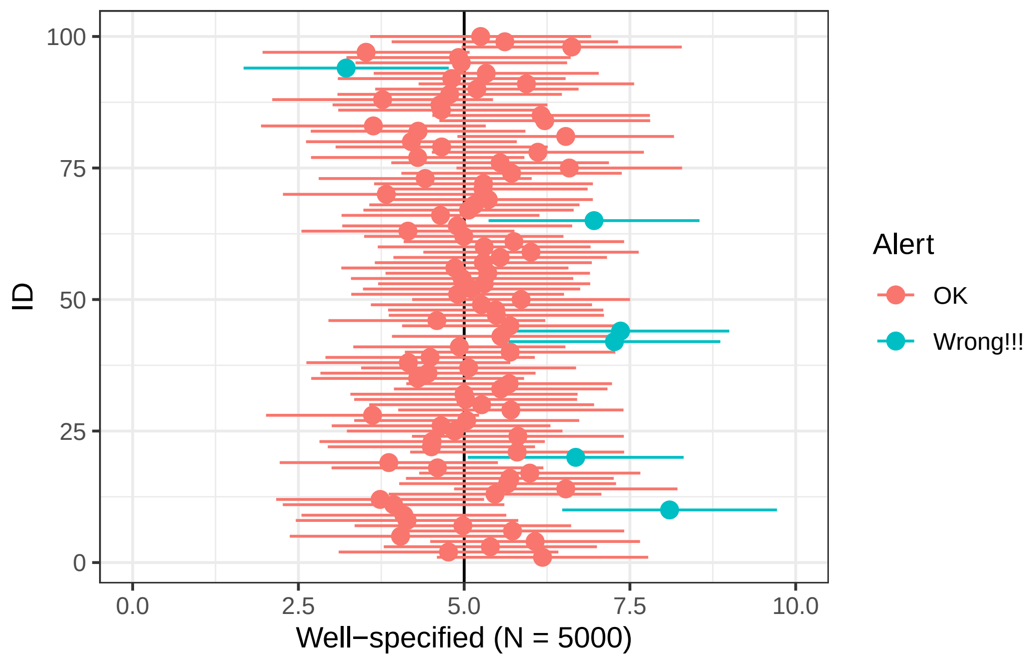


- 注意: 真の値からの” 距離” だけわかる

95% 信頼区間: 2000 事例



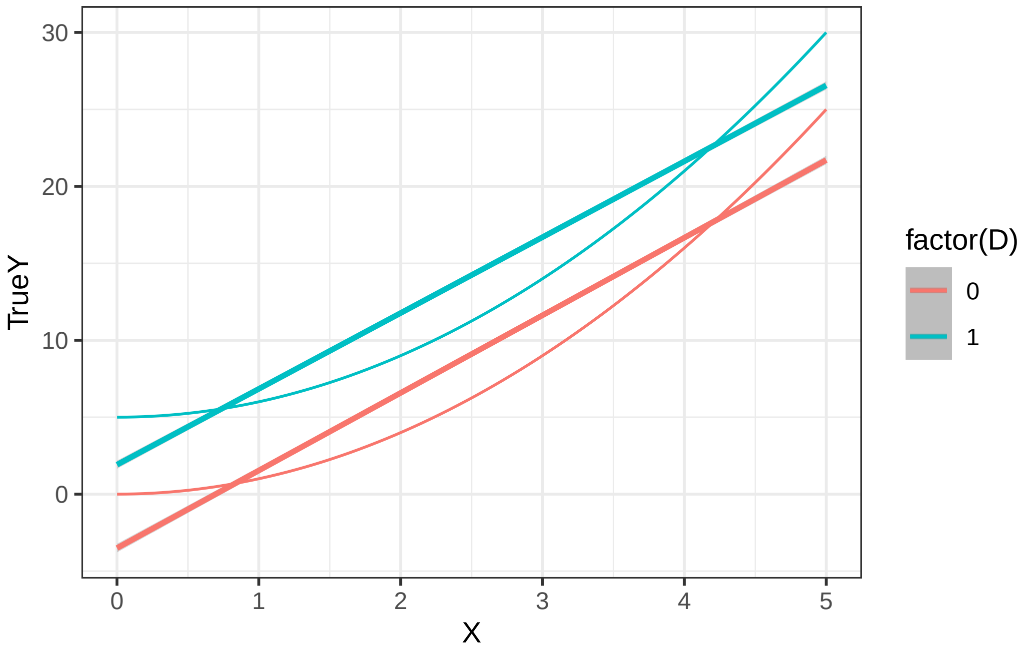
サンプルサイズの影響: 200 事例



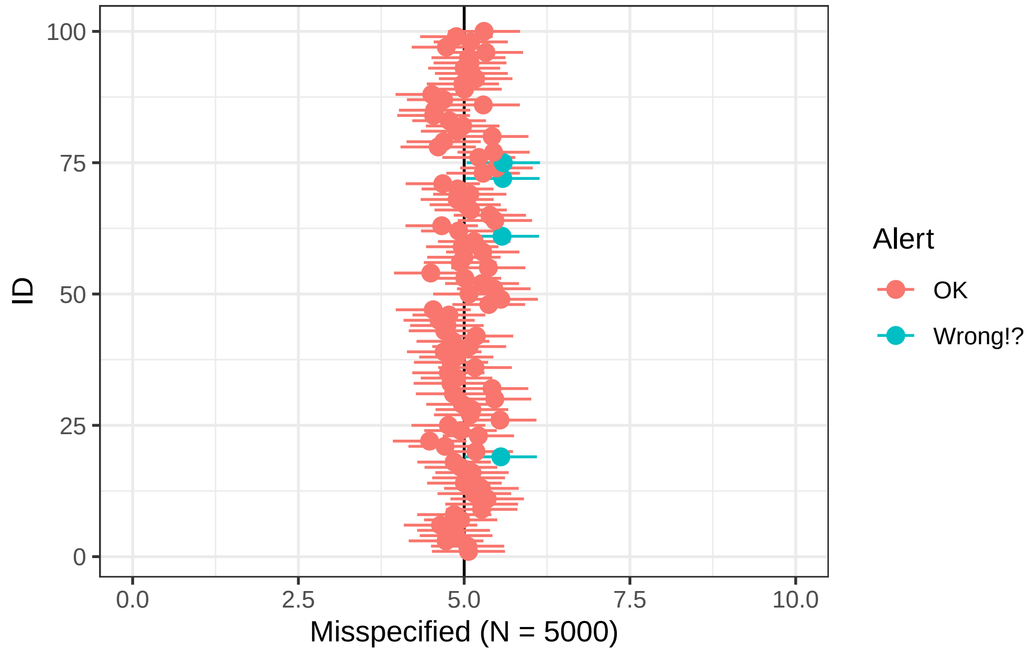
Misspecified model

- $g(D, X) = \beta_0 + \beta_D D + \beta_1 X_1 + .. + \beta_L X_L$
  - β を推定
- β をどう選んでも、$E_P[Y|D, X] \neq g(D, X)$ (近似誤差)
  - BLP について信頼区間を提供

例: BLP



95% 信頼区間: 2000 事例



まとめ

- ランダムサンプリングの仮定のみで、BLP についての信頼区間を導出できる
  - 大部分の研究者が真の値を含んだ区間を得られる
- BLP が" 研究関心" となる母集団の特徴を捉えているのであれば、有益な方法
  - 予測の手法としては問題があっても**関係ない**
- BLP 以外の記述統計量を推定したい場合は?