Identification

労働経済学 1

川田恵介

1 記述 VS 構造研究

1.1 練習問題

- D が Y に与える因果効果を推論するために、以下の実証手順は一般に望ましくない。なぜか？
  - D/Y を設定し、それ以外の全変数をXに対して、Double-selection を適用
  - 選択された変数Zを使用して、$Y \sim D + Z$ を OLS 推定する

1.2 実証研究の工程

- (社会における)研究目標
  - $\underbrace{\rightarrow}_{\text{識別}}$ (母集団における)推定目標
  - $\underbrace{\rightarrow}_{\text{推定}}$ (データから計算する)推定値
- 本講義は、「推定」が主たる対象
  - 本スライドのみ、「識別」を議論

1.3 研究目標

- スライド: Models, Measurement, and the Language of Empirical Economics (Phil Haile) に従い、研究目標を大きく 2 種類に分類する
  - **Descriptive:** データから観察可能な変数の**母集団**における関係性を推論する
    - 例: 去年と今年で、賃金の平均値はどのくらい変化したのか?: Year $\sim$ Wage
  - **Structural:** 母集団の”背後”にある構造(仕組み)を推論する
    - 含む因果/経済/測定モデルの**構造推定**

1.4 Descriptive

- 観察可能な変数に関する議論なので

- 社会における関係性 $\underbrace{\simeq}_{Random\ Sampling}$ 母集団における関係性
  - $\underbrace{\simeq}_{\text{十分な事例数}}$ データ上の関係性

- 信頼区間などを応用した Valid inference が可能

1.5 構造 (Structure)

- **研究目標:** “社会の仕組み(構造)”の特徴把握
  - 母集団を生み出す構造があると想定
  - 構造は、データから観察できない要素も用いて定義されうる
- **推定目標:** 関心となる社会構造の特徴と母集団の特徴の関連性を示す必要がある
  - 識別の議論

1.6 構造モデル

- **構造モデル** (構造を議論する言語 + 構造についての仮定)を導入
  - {(未知の)研究関心となる構造、構造についての既知の知識(仮定)、観察可能な変数}を紐づける
- 例: 価格理論
  - データ: $\{P = Price, Q = Quantity\}$
  - 研究関心: 政策介入の効果など
  - 構造モデル: 完全競争/独占/寡占市場/サーチモデル…
    - 意思決定理論/生産関数/期待形成…

1.7 因果効果 $\in$ 構造モデル

- ある action(政策/介入等)が与える影響
  - 定義/識別を議論する有力な構造モデルが複数存在
  - 潜在結果や構造的因果モデル Chap 2, 4-7 in CausalML、“経済理論” (Heckman & Pinto, 2024) 、頑健性 (Peters et al., 2016) など
    - 不毛(?)な”学派”論争も散見されるが、共通点が多い (Goldberg, 2019; Imbens, 2020; Vytlacil, 2002)。
- 以下では、潜在結果 (Potential outcome)を採用
- 他の例: Section 4

2 潜在結果モデル

2.1 例: 留学の効果

- 例: 1 年次の短期留学は、語学力を向上するのか?
  - 短期留学"義務化"の是非についての、重要な指数(かも？)
- データ: $Y =$ TOEIC の点/$D =$ 短期の留学経験

2.2 例: データ: Summary statistics

$E[Y \mid D]$	D	N
400	0	7000
500	1	3000

- 平均差 $= 100$

2.3 推定

- 推定: 事例数は十分に大きいので、データ上の平均 $\simeq$ 母平均が期待できる
  - 単なる平均値なので、信頼区間も計算可能
- 識別: 母平均の差は因果効果と解釈できるか?

2.4 潜在結果モデル

- 以下の**観察できない潜在結果関数**を想定
- d : { 1 (プログラム参加), 0 (非参加) }
  - $y_i(d)$: 参加/非参加の場合の、i さんの点数
- 留学の個別効果 : $\tau_i = y_i(1) - y_i(0)$

2.5 観察可能な変数との関係性

- データ$=$ **現実の** $\{$**プログラム参加**(D_i), **点数**$(Y_i)\}$
- 観察できる/できない変数間に、以下の関係性を仮定
- $Y_i = D_i \times y_i(1) + (1 - D_i) \times y_i(0)$
  - $D_i = 1$ ならば $Y_i = y_i(1)$
  - $D_i = 0$ ならば $Y_i = y_i(0)$
- 注: 小文字はデータから観察できない変数/大文字は観察できる変数

2.6 因果推論の根本問題

- 個人効果 τ_i の定義: $\tau_i = y_i(1) - y_i(0)$

- $y_i(1)$ と $y_i(0)$ を同時に観察できないため、個別効果 τ_i は観察できない

2.7 集計

- 観察できる変数の集計値

$$E[Y_i \mid D_i = 1, X] - E[Y_i \mid D_i = 0, X]$$

- 観察できない構造の集計値 (Conditional treatment effect)

$$E[\tau_i \mid D_i = 0, X]$$

  - $\{X, D = 0\}$ 中での平均効果

2.8 観察可能な分布との関係性

- $E[Y_i \mid D_i = 1, X] - E[Y_i \mid D_i = 0, X]$
- $= E[y_i(1) \mid D_i = 1, X] - E[y_i(0) \mid D_i = 0, X]$
- $$= \underbrace{E[y_i(1) \mid D_i = 0, X] - E[y_i(0) \mid D_i = 0, X]}_{=E[\tau_i \mid D_i = 0, X]}$$
  $$+\underbrace{E[y_i(1) \mid D_i = 1, X] - E[y_i(1) \mid D_i = 0, X]}_{=Selection}$$
  - Selection = **仮に**全員留学したとしても残る格差

2.9 識別問題

- 平均差を”説明する”構造は無数に存在
- $$\underbrace{E[Y_i \mid D_i = 1, X] - E[Y_i \mid D_i = 0, X]}_{=0.2} =$$
  $$\underbrace{E[\tau_i \mid D_i = 0, X]}_{-0.6} + \underbrace{Selection}_{=0.8}$$
- 短期留学の平均効果は負だが、強力な Selection によって正の効果がもたらされている

2.10 Observational equivalent

- 異なる構造
- $$\underbrace{E[Y_i \mid D_i = 1, X] - E[Y_i \mid D_i = 0, X]}_{=0.2} =$$
  $$\underbrace{E[\tau_i \mid D_i = 0, X]}_{0.1} + \underbrace{Selection}_{=0.1}$$
- 短期留学の効果は、Selection により過大評価

2.11 Identification by mean independence

- $$\underbrace{E[Y_i \mid D_i = 1, X] - E[Y_i \mid D_i = 0, X]}_{=0.2} =$$

  $$\underbrace{E[\tau_i \mid D_i = 0, X]}_{0.2} + \underbrace{\underbrace{Selection}_{=0 \ (Mean\ independence)}}$$

- Mean independence の正当化: X内で D は
  - 完全ランダムに決まっている

2.12 Identification by mean independence

- Mean independence のもとで、$Y \sim D + X$ の Population OLS の結果として定義される β_D は、Xについての balanced comparison
- $$\beta_D = \sum_X \omega(X) \times [E[Y \mid D = 1, X] - E[Y \mid D = 0, X]]$$

  $$\underbrace{=}_{Mean\ independence} \sum_X \omega(X) \times \tau(X)$$

2.13 識別における変数選択

- 推定対象: X をバランスさせた balanced comparison
  - 例えば、Xについて十分に複雑にした Population OLS: $Y \sim D + X$
- 識別における変数選択: Population OLS において、どの変数をXに加えるべきか?
  - 例: $D = 1$ 年次の留学、$Y =$ 在学中の TOEIC の最高点、であったときに"2 年次の上級英会話コースへの参加"をXに加えるべきか?

2.14 識別における変数選択

- 理想: 現実社会におけるD の決まり方についての知識から、Dがランダムに決まっているサブグループを同定する X を選ぶ
- 妥協: 理想的な RCT を模倣するために、RCT を実施しても差が生じないと「想像できる」Xを選ぶ
  - 例: 1 年時に留学しても、**年齢**は不変なのでXに含めるが、**上級英会話の受講**は変化しうるので含めない

2.15 識別における変数選択

- 実験の結果、D間で差異が生じる変数は、Bad control, Post treatment (文脈によっては mediator)と呼ばれ、バランスの対象とすべきではない
- 詳細/発展は、 VanderWeele (2019) ; Cinelli et al. (2024) などを参照

2.16 補論: 仮定の緩和

- $$\underbrace{E[Y_i \mid D_i = 1, X] - E[Y_i \mid D_i = 0, X]}_{=0.2} = \underbrace{E[\tau_i \mid D_i = 0, X]}_{\leq 0.2} + \underbrace{Selection}_{\geq 0}$$
- 効果の上限が識別される
- 一般には、Partial identification/Moment inequality (Canay et al., 2023; Kline & Tamer, 2023) と呼ばれる文脈

2.17 まとめ

- Population の推定はデータ分析の古典的な関心だが、経済学の主たる関心は構造であることが昔から多かった
  - “Population” versus “Structure” (Koopmans & Reiersol, 1950)
    - 識別の議論が必須
- 伝統的には経済モデルを用いてきたが、他の枠組み(潜在結果等)の利用も増えている
  - 研究課題に応じて、柔軟に活用する必要がある

3 推定(復習)

3.1 推定問題

- 識別(Section 2)によって定義された推定目標を、限られたデータから推論する
  - 先の例では、$Y \sim D + X$ の Population OLS の結果を推論する
- X の数が十分に少なければ、データ上で OLS

3.2 推定における変数選択

- Double-selection: X の一部をZとして選ぶ
- ただし、データ上での$Y \sim D + Z$ で得られる β_D が、母集団上での$Y \sim D + X$ に極力近づくように選ぶ
  - **推定目標を変化させているわけではない**

3.3 まとめ

- 識別における変数選択: 推定目標の設定
  - 研究課題および識別の議論から、変数選択は行われる
    - 一般にデータ**ではなく**、構造についての背景知識を用いる

        - 因果推論については、CausalML (Chap. 8)や Maiti et al. (2025) などさまざまな”補助ツール”が活用できる
- 推定における変数選択: 推定目標と推定値を近づけるために行われる
  - データに基づいて決めることが可能

3.4 まとめ

- 一般に、社会構造の推論は、以下を要求する
  - データ + 推定についての仮定 + **識別に関する仮定**
    - **識別に関する仮定**の一例 = Mean independence
      - Mean independence を保証する一例 = RCT の実施
    - **推定についての仮定**の一例 = Approximately Sparsity

3.5 実践への含意

- 前提を明示した上で、結果の解釈を示す
- 知りたい構造は、極力厳密に定義
- “ドメイン知識”に基づいた、妥当な仮定を構造に導入
  - 経済理論等に基づいた明確なシナリオ
- 機械学習等を活用し、少ない仮定で推定

4 補論: Item response theory

4.1 例: Item response theory

- **データ:** 以下の回答結果
  - Q1. $Y = 2 \times X + 5, Y = -4 \times X + 6$ の解
  - Q2. $Q_D = -3 \times P_D + 6$, $Q_S = 2 \times P_S + 10$ における完全競争市場均衡
- **研究関心:** 完全正答できる”知識” を持つ回答者の割合

4.2 Structural model

- Knowledge space：ある問に正答 = 知識を持つかどうか × ケアレスミスする確率 × 紛れ当たりする確率
- 仮定
  - Q2 だけ回答できる知識は存在しない
- 識別: Noventa et al. (2024)

4.3 Identification

- Q1 と Q2 に正答 = Q1,Q2 に正答できる知識を持つ × (1 - “Q1 をミスる”) × (1 - “Q2 をミスる”)
- Q2 のみ正答 = Q1,Q2 に正答できる知識を持つ × “Q1 をミスる” × (1 - “Q2 をミスる”)
- $$\frac{Q1, Q2\text{の正答率}}{Q2\text{のみの正答率}} = \frac{1 - Q1\text{をミスる}}{Q1\text{をミスる}}$$
  - Q1 をミスる確率を識別できる
    - “Q1 をミスる確率” = “Q2 をミスる確率”であれば、Q1,Q2 に正答できる知識を持つ回答者割合を識別できる

5 Reference

Bibliography

Canay, I. A., Illanes, G., & Velez, A. (2023). A User’s guide for inference in models defined by moment inequalities. Journal of Econometrics, 105558. https://doi.org/https://doi.org/10.1016/j.jeconom.2023.105558

Cinelli, C., Forney, A., & Pearl, J. (2024). A crash course in good and bad controls. Sociological Methods & Research, 53(3), 1071–1104.

Goldberg, L. R. (2019,). The Book of Why: The New Science of Cause and Effect: by Judea Pearl and Dana Mackenzie, Basic Books (2018). ISBN: 978-0465097609. Taylor & Francis.

Heckman, J., & Pinto, R. (2024). Econometric causality: The central role of thought experiments. Journal of Econometrics, 105719.

Imbens, G. W. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. Journal of Economic Literature, 58(4), 1129–1179.

Kline, B., & Tamer, E. (2023). Recent developments in partial identification. Annual Review of Economics, 15(1), 125–150.

Koopmans, T. C., & Reiersol, O. (1950). The identification of structural characteristics. The Annals of Mathematical Statistics, 21(2), 165–181.

Maiti, A., Plecko, D., & Bareinboim, E. (2025). Counterfactual Identification Under Monotonicity Constraints.

Noventa, S., Ye, S., Kelava, A., & Spoto, A. (2024). On the identifiability of 3-and 4-parameter item response theory models from the perspective of knowledge space theory. Psychometrika, 89(2), 486–516.

Peters, J., Bühlmann, P., & Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. Journal of the Royal Statistical Society Series B: Statistical Methodology, 78(5), 947–1012.

VanderWeele, T. J. (2019). Principles of confounder selection. European Journal of Epidemiology, 34, 211–219.

Vytlacil, E. (2002). Independence, monotonicity, and latent index models: An equivalence result. Econometrica, 70(1), 331–341.