Prediction Problem

川田恵介

keisukekawata@iss.u-tokyo.ac.jp

2025-05-05

1 社会に対する研究目標: 予測

1.1 予測問題

- 研究対象の一種
- 予測問題: 「Y が観察できない/Xは観察可能」な状況で、Y の値を予測する
  - 例: 建物の広さや立地から、市場における取引価格を予測する
- 推定"値": X から Y の値を自動計算する"式" $\hat{g}(X)$ (予測モデル/"AI")

1.2 例: OLS

```
data(CPS1985, package = "AER")

lm(wage ~ education + age, CPS1985)
```

```


Call:
lm(formula = wage ~ education + age, data = CPS1985)

Coefficients:
(Intercept)    education          age
    -5.5342       0.8211       0.1050
```

- 見慣れた数式に直すと、

$$\hat{g} = -5.5 + 0.8 \times education - 0.1 \times age$$

- education と age を代入すれば、価格の予測値を計算してくれるモデル

1.3 注意点

- 入門的講義では、推定値は β であることを前提とする場合が多い

- 予測問題では、β ではなく、 Y の計算式 $\hat{g}(X)$ を推定値であるとイメージする方が実践的
  - Nonparametric 推定 (含む、RandomForest/Boosting) などでは、β に相当する値が大量に存在し、人間には認知不可能であるため

1.4 予測性能

- ある事例についての予測誤差 = 実際のY − 予測値 $\hat{g}(X)$
  - ほとんどの応用で、まぐれあたりではなく、安定的な性能を目指す
- 予測対象も何らかの集団 (Target/Study Population)から抽出されたと想定する
  - よく用いられる性能指標は、Target Population 上で計算した平均二乗誤差

$$(Y-\hat{g}(X))^2\text{の}Target\text{上の平均}$$

1.5 識別: 最善の予測値

- 平均二乗誤差を性能指標とするのであれば
  - Target Population における平均値が最善の予測値
- データも、Target Populaion からランダムサンプリングされている (Source = Target) のであれば、
  - Source Population における母平均 $\mu(X)=E[Y \mid X]$ が最善の予測値
- 推定対象を**母平均**として、推定を行う必要がある

1.6 まとめ

- Target = Source Population であれば、母平均が最善の予測値であり、推定対象
  - 入門的な計量経済学と同じ推定対象!!!
    - 予測を超えた応用ができそう (後述)
- 課題: ここまでは全て、研究対象 ⇔ 推定対象 (識別)、であり、推定対象 ⇔ 推定値 (推定)、を論じないと実践できない
  - どのやって母平均を推定するか?
  - どうやって予測性能を実際に測定するのか?

2 予測性能の推定

2.1 研究計画: モデル評価

- **研究課題:** あるモデル $\hat{g}(X)$ の社会における予測誤差
  - Source = Target Population

- **推定課題:** $(Y - \hat{g}(X))^2$ の母平均
  - サンプル分割の活用
- **推定値:** 点推定値 (+ 信頼区間)

2.2 サンプル分割

- 代表的な方法

1. データをランダムに２分割する (Training/Test データ)
2. $\hat{g}(X)$ は、**Training** のみを使用して、推定
3. $(Y - \hat{g}(X))^2$ の **Test** 上での平均値を計算し、評価指標とする

2.3 Naive なアプローチ

- データ分割せずに、全データを用いて$\hat{g}(X)$ を推定し、全データについて $(Y - \hat{g}(X))^2$ を計算する
  - 多くの統計ソフトが、平均二乗誤差(ないし R^2)として出力する
- モデルの複雑性 (β の数) に比べて、事例数が十分に多ければ、母集団上での平均二乗誤差を近似する
  - 多くの機械学習の手法は、複雑なモデルを推定するため、予測性能を過大評価 (平均二乗誤差を過小に推定)しがち

2.4 例

- 食生活 ($= X$) から、運動能力 ($= Y$) を予測するモデルを推定したい
  - 大谷翔平選手の事例を収集 ($N = 1$)
  - “データ"に当てはめた結果、$\hat{g}$(オートミール) = 非常に高い能力
- **予測モデルを同じデータ(大谷選手)で評価**
  - 予測が完璧に当たっている…?
  - 大谷選手以外の事例で評価すべき

2.5 まとめ

- モデルの推定に用いたデータで、モデルを評価すると、性能が過大評価される
  - “２度漬け (Double Dipping)” と呼ばれる問題
- 最もシンプルな解決策は、データの推定と評価を、ランダム分割で生成した異なるデータで行う
  - 事例数に比べて単純なモデルを推定する場合のみ、R^2 などの伝統的な評価指標は有効

3 実践への含意

3.1 復習

- 分析フロー全体に注意を払う必要がある
- 実務/社会/政策課題
  - $\rightarrow$ 研究対象
  - $\rightarrow$ 推定対象
  - $\rightarrow$ 推定値
  - $\rightarrow$ 計算、発信、、、

3.2 $\rightarrow$ 研究対象

- そもそも何をY/Xとするかが極めて重要
  - 学術研究: 研究動機(研究の重要性)をしっかり説明できるか?
    - Einav et al. (2018) の Motivation などは好例
  - 実務: 実務上に役が立つ/弊害がない(少ない)か?
    - Algorithm Fairness (Mitchell et al., 2021; Berk, Kuchibhotla and Tchetgen Tchetgen, 2023)

3.3 研究対象

- Y の値を完璧に予測するための、**社会に対する**前提条件は、**最善の予測値**$\mu(X) = Y$
  - 社会において、X が同じであれば、Y についての個人差がない
- 人間行動についてあり得えない場合が高い
  - 例: 一卵性の双子でも、人生は異なる

3.4 研究対象 $\rightarrow$ 推定対象

- 予測対象と母集団がずれていると、予測が難しい
  - 伝統的なサンプリングバイアスに注意 (生存バイアス, 回答バイアス, 選択バイアス)
  - Concept drift: モデル推定から時間が経つと、社会 (Target Populaiton)が変化し、予測性能は悪化する
    - 予測性能の監視と必要に応じた再推定が必要
- 詳細な解説とチャレンジの紹介

3.5 推定対象 $\rightarrow$ 推定値

- **予測誤差** $= Y -$ **予測値**

- $= \underbrace{Y - Target\text{上での平均値}}_{\text{Irreducible Error}}$
- $+Target\text{上での平均値} - Source\text{上での平均値}$
- $\underbrace{+Source\text{上での平均値} - \text{推定されたモデル}}_{\text{推定の問題}}$

3.6 OLS の問題

- 大きな問題は、

$$\underbrace{\mu(X)}_{\text{予測問題が求める推定対象: 母平均}} \neq \underbrace{g^{Pop}(X)}_{\text{OLS の実質的な推定対象: Population OLS}}$$

- OLS を十分に複雑化すれば、$\mu(X) \simeq g^{Pop}(X)$ が期待できる
  - 川田が知る限り、「十分な複雑性」は、現状経験則以上のものはない
  - 例えば、Duflo によるチュートリアルセッション では、連続変数については二乗項、および 2 変数間の交差項を導入

3.7 OLS の問題

- モデルを複雑にしすぎると、推定精度が悪化
  - 元々のXの数が多いと、モデルは容易に複雑化
  - 例: CPS1985 (AER package) を用いて、賃金を予測する
    - $X =$ [教育年数、経験年数、年齢、人種、地域、性別、職種、産業、組合、結婚]
      - 2 次項 + 交差項を加えると、βの数: 16 → 107
- 次回以降、データ主導の解決策を議論

3.8 推定値 →

- “機械学習による予測性能の改善"を研究対象とするのであれば、ベンチマークとなる予測モデルと比較する必要がある
  - 単純な OLS や単純平均値など

3.9 Reference

Bibliography

Berk, R. A., Kuchibhotla, A. K. and Tchetgen Tchetgen, E. (2023) “Fair risk algorithms,” Annual Review of Statistics and Its Application, 10(1), pp. 165–187

Einav, L. et al. (2018) “Predictive modeling of US health care spending in late life,” Science, 360(6396), pp. 1462–1465

Mitchell, S. et al. (2021) “Algorithmic fairness: Choices, assumptions, and definitions,” Annual review of statistics and its application, 8(1), pp. 141–163