Applied Metrics Papers

Author: Wenxiao Yang

Institute: Haas School of Business, University of California Berkeley

Date: 2024

All models are wrong, but some are useful.

Contents

Chapter 1 Identification of Prediction Errors

1.1 [Rambachan(2024)]: Identifying Prediction Mistakes in Observational Data

Uncovering systematic prediction mistakes in empirical settings is challenging because

1. the decision maker's preferences and
2. the information set

are unknown to us.

1.1.1 Expected Utility Maximization at Accurate Beliefs

A decision maker (DM) makes a binary choice $c \in \{0, 1\}$ for each individual, which is summarized by characteristics $x \in \mathcal{X}$ and an unknown outcome $y^* \in \mathcal{Y}$ (observable when $c = 1$).

Example 1.1 (Pretrial Release)

A judge decides whether to detain ore release defendants $C \in \{0, 1\}$. The outcome $Y^* \in \{0, 1\}$ is whether a defendant would fail to appear in court if released. X is the recorded information of the defendant.

Example 1.2 (Medical Testing and Diagnosis)

$C \in \{0, 1\}$ is whether to conduct a test. $Y^* \in \{0, 1\}$ is whether the patient had a heart attack. X is the recorded information of the patient.

Example 1.3 (Hiring)

$C \in \{0, 1\}$ is whether to hire a candidate. Y^* is a vector of on-the-job productivity measures. X is the recorded information of the candidate.

These three variables are summarized by a joint distribution, $(X, C, Y^*) \sim P(\cdot)$. We assume finite full support of x, i.e. there is a $\delta > 0$ such that $P(x) := P(X = x) \geq \delta, \forall x \in \mathcal{X}$. As the Y^* is only observable when $C = 1$. We define

$$Y := C \cdot Y^*$$

The observable data is the joint distribution $(X, C, Y) \sim P(\cdot)$. The DM's conditional choice probabilities are

$$\pi_c(x) := P(C = c | X = x), c \in \{0, 1\}, x \in \mathcal{X}$$

The observable conditional outcome probabilities are

$$P_1(y^* \mid x) := P(Y^* = y^* \mid C = 1, X = x), y^* \in \mathcal{Y}, x \in \mathcal{X}$$

The $P_0(y^* \mid x)$ and the true outcome probabilities $P(y^* \mid x)$ are not identified due to the missing-data problem.

Assumptions

Note *In the main context of paper: (i). The decision maker makes a binary choice $c \in \{0, 1\}$ for each individual; (ii). The decision maker's choice does not have a direct causal effect on the outcome.*

Assumption 1.1 (Bounds on the Unobserved Conditional Outcome Probabilities)

For each $x \in \mathcal{X}$, there exists a known subset $\mathcal{B} \subseteq \Delta\mathcal{Y}$ such that $P_0(\cdot \mid x) \in \mathcal{B}_x$.

Given this assumption, the identified set for the true outcome probabilities given $x \in \mathcal{X}$, denoted by

$$\begin{aligned}\mathcal{H}(P(\cdot \mid x); \mathcal{B}_x) := \big\{\tilde{P}(\cdot \mid x) \in \Delta\mathcal{Y} : \tilde{P}(y^* \mid x) = \tilde{P}_0(y^* \mid x)\pi_0(x) + P_1(y^* \mid x)\pi_1(x), \\ \forall y^* \in \mathcal{Y} \text{ and for some } \tilde{P}_0(\cdot \mid x) \in \mathcal{B}_x\big\}\end{aligned}$$

Bibliography

[Rambachan(2024)] Rambachan, A. (2024). Identifying prediction mistakes in observational data. *The Quarterly Journal of Economics*, page qjae013.