MM6761: Take-home Assignment 4

Dai Yao (dai@{yaod.ai, mle.bi})

April 11, 2024

1 Zhang and Zhu (AER 2011), Wu and Zhu (MNSC 2022)

In ZZ (2011), they simply call the number of contributors as group size. In WZ (2022), they call the number of contributors as competition (or competition intensity). Why?

Answer:

In ZZ (2011), users of Wikipedia are content contributor and consumers simultaneously, and they may write in the same article when contributing content to the platform. In WZ (2022), there is a clear distinction between content contributors (i.e., the novelists) and content consumers on the platform, and each novelist writes his or her own novels. Hence, there is a clear competition between novelists on the supply side for content consumers on the demand side.

What are the major differences (2 to 3 points) between the two studies? (please avoid talking about trivial things such as business context, user sample, etc.).

Answer:

Some noticeable major differences:

- The core model in ZZ (2011) is a regression discontinuity model (Model 1 on page 1608), while that in WZ (2022) is a difference-in-differences model (Model 1 on page 8621).
- ZZ (2011) focuses on the impact on the extensive margins only (i.e., addition, deletion), while WZ (2022) focuses on the impact on both the extensive margins as well as the intensive margins (i.e., novelty).
- ZZ (2011) conducts some analysis on the underlying mechanism (Model 2 on page 1609) of the effects of group size, while WZ (2022) does not have any mechanism analysis.

2 Goli et al (MKSC 2024)

Intuitive speaking (i.e., by words), what is the main idea in the paper, and what is the proposed method?

Answer:

Marketplace experiments involving ranking algorithms can suffer from significant interference bias because they are unable to simulate what the counterfactual market equilibrium would be if the test algorithm were to be scaled up to the entire platform. The authors in this paper develop a procedure that takes data from a series of existing experiments and identifies microslices of data that are close to the counterfactual market equilibrium under a particular algorithm a in each individual experiment.

Formally (i.e., by mathematical equations or formulas), what is the proposed method?

Answer:

The true total average treatment effect (TATE), when algorithm a is scaled up to the total population, is:

$$\text{True TATE} \quad = \quad \sum_{j \in J} q_a^{(d)}(\mathbf{x}_j; P_j^a(\cdot), P_j^a(\cdot)). \tag{1}$$

Here, $P_j^a(\cdot)$ is the probability for item j to be place at each r of the potential ranks ($r \in \{1, 2, ..., \infty\}$. Actually, the maximum of r is J). The summation is done over all the items.

There are however multiple ranking algorithms (A is the set of all the ranking algorithms) deployed on the platform simultaneously, and m_a is the mass of customers assigned to algorithm a. Given that the platform has a unit mass of users, $\sum_{a \in A} m_a = 1$. The TATE observed across all the experiments is as follows:

$$\text{Observed TATE} \quad = \quad \sum_{j \in J} q_a^{(d)}(\mathbf{x}_j; P_j^a(\cdot), P_j^{\mathbf{m}}(\cdot)), \tag{2}$$

where $P_j^{\mathbf{m}}(\cdot) = \sum_{a' \in A} m_a P_j^{a'}(\cdot)$.

Thus, on the one hand, unless $P_j^{\mathbf{m}}(\cdot) = P_j^a(\cdot)$, we expect a discrepancy between the observed and true TATE. On the other hand, the gap between the two determines the discrepancy between the observed and true TATE. Proposition 1-3 (page 8-9) are developed based on this intuition.

Is there any limitation in the method? Can you generalize the setting from ranking experiments to others?

Answer:

The method relies on a few assumptions. Assumption 1 (page 6) and 3 (page 8) seem to be fine, as Assumption 1 is a standard assumption and has been tested (in Web Appendix B), and Assumption 3 just ensures continuity in the function. Personally, I feel that Assumption 2 (page 7) is too strong. It basically assumes that the state variables associated with j are fully capture by its own characteristics ($\mathbf{x}_j$) and the overall ranking distribution ($P_j^{\mathbf{m}}(\cdot)$).

Some additional work needs to be done to generalize the method to other settings than ranking experiments. A potentially fruitful direction is to evaluate the true TATE of certain firm intervention (e.g., price promotion) on socially connected consumers, based on the observed TATEs from a set of field experiments.

3 Huang et al (WP 2024)

Intuitive speaking (i.e., by words), what is the main idea in the paper, and what is the proposed method?

Answer:

In practise, we may want to estimate the effects of long-term treatments (e.g., updates to product functions, user interface designs, and recommendation algorithms). However, there are a lot of constraints of conducting long-term experiments. Hence, there needs to be an approach to infer the effects of long-term treatments based on short-term experimental results.

The authors propose a framework to use longitudinal surrogates, defined as the intermediate outcomes of subjects during the experiment that saturate the causal links between historical treatments and future outcomes. They iteratively make use of these longitudinal surrogates, and establish "longitudinal surrogate index" and "pivot index functions." These index functions enable them to extrapolate the longitudinal surrogates from the short-term experimental periods to the long-term future periods, thus providing estimations of effects in the long term.

Formally (i.e., by mathematical equations or formulas), what is the proposed method?

Answer:

The surrogate index, which is the conditional expectation of the primary outcome at time t, given the surrogate outcomes at time 0, the pre-treatment variables, and the treatment assignments, is denoted as:

$$h_t(s, x, w_{1:t}) \quad = \quad \mathbb{E}_{\mathcal{F}}[Y_{it}|S_{i0} = s, X_i = x, W_{i,1:t} = w_{1:t}]. \tag{3}$$

The pivot index, the conditional expectations of the surrogate outcomes at time t, given the surrogate outcomes at time 0, the pre-treatment variables, and the treatment assignments, is denoted as:

$$g_t(s, x, w_{1:t}) \quad = \quad \mathbb{E}_{\mathcal{F}}[S_{it}|S_{i0} = s, X_i = x, W_{i,1:t} = w_{1:t}]. \tag{4}$$

In addition, denote the conditional surrogate outcomes at time t, given the surrogate outcomes at time 0, the pre-treatment variables, and the treatment assignments, as:

$$G_t(s, x, w_{1:t}) \quad = \quad S_{it}|S_{i0} = s, X_i = x, W_{i,1:t} = w_{1:t} \tag{5}$$

With Assignment 1 and 2, the average effect of long-term treatments on the primary outcome is said to be the following:

$$\begin{aligned} \pi_T \quad = \quad & \mathbb{E}_{\mathcal{F}}[h_{\Delta t_{K+1}}(G_{\Delta t_K}(\cdots G_{\Delta t_1}(S_{i0}, X_i, 1_{\Delta t_1}) \cdots, X_i, 1_{\Delta t_K}), X_i, 1_{\Delta t_{K+1}})] - \\ & \mathbb{E}_{\mathcal{F}}[h_{\Delta t_{K+1}}(G_{\Delta t_K}(\cdots G_{\Delta t_1}(S_{i0}, X_i, 0_{\Delta t_1}) \cdots, X_i, 1_{\Delta t_K}), X_i, 0_{\Delta t_{K+1}})] \end{aligned} \tag{6}$$

The model can be converted to a linear model if Assumption 3 is applied.

Do you think this model works? If so, why? If not, what's the main issue?

Answer:

I feel the inference is artificial. The main reason is that the assumption about the intermediate outcomes (i.e., surrogate outcomes) might not hold, that they only affect the primary outcome one period afterwards rather than the primary outcome in the same period (See Figure 2 on page 9). In the empirical studies, there is no reason to believe that the four surrogate outcomes (see Table 2 on page 19, $search_{qv}, recall_{qv}, expose_{qv}, click_{qv}$) occur prior to the primary outcome, i.e., $search_{uv}$.