Close Enough?
A Large-Scale Exploration of Non-Experimental Approaches to Advertising Measurement

Brett R. Gordon
Kellogg School of Management
Northwestern University

Robert Moakler
Ads Research
Meta

Florian Zettelmeyer
Kellogg School of Management
Northwestern University & NBER

Meta

Northwestern | Kellogg School of Management

Full paper (pdf) available on arXiv

Advertisers want to maximize incrementality, but experiments can be challenging

Advertisers want to maximize their return on advertising investments by maximizing the incrementality of their ads.

However, measuring incremental ad effects can be difficult.

Randomized Controlled Trials (RCTs) are often seen as the gold standard to measure incremental advertising effects

But, RCTs are not always available for ad measurement

Therefore many advertisers rely on non-experimental methods

- Propensity score matching has been widely used in many industries
- Double/debiased machine learning is becoming increasingly popular with people combining machine learning and casual inference.

What we tested: What happens when advertisers don’t or can’t run experiments?

We consider the hypothetical scenario

If an advertiser had not implemented a campaign as an RCT (i.e., without an experimental control group) what ad effect would they have estimated using a non-experimental method?

We explore a representative set of 1,673 RCTs and leverage thousands of detailed features for non-experimental modeling[1]

We highlight results from two research questions from this work

Do estimates from non-experimental methods like Double/Debiased Machine Learning (DML) and Stratified Propensity Score Matching (SPSM) come close to those from RCTs?

When do these methods tend to do better vs. worse and how close are they?

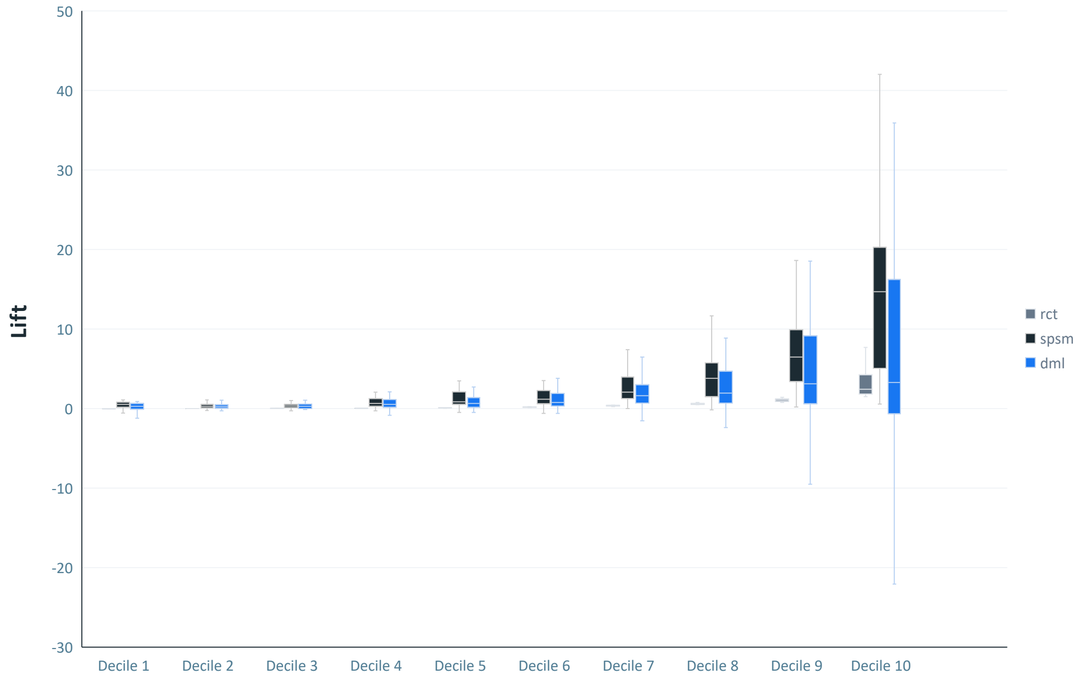
Source: This is a follow up from Gordon, B., F. Zettelmeyer, N. Bhargava, and D. Chapsky (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. Marketing Science 38 (2), 193–225.

Meta | Northwestern Kellogg School of Management

What we found: Non-experimental methods result in large measurement errors

For the best non-experimental method, DML, we found the median estimated lifts to be 143%, 126%, and 68% for upper, mid, and lower funnel outcomes, respectively. These estimates are very large given that the median RCT lifts are 28%, 19%, and 6% for the equivalent funnel outcomes.

Results for Upper Funnel Outcomes



RTC Lift Deciles

(Figures exclude the top 1% lifts for each position in the purchase funnel)

Meta | Northwestern Kellogg School of Management

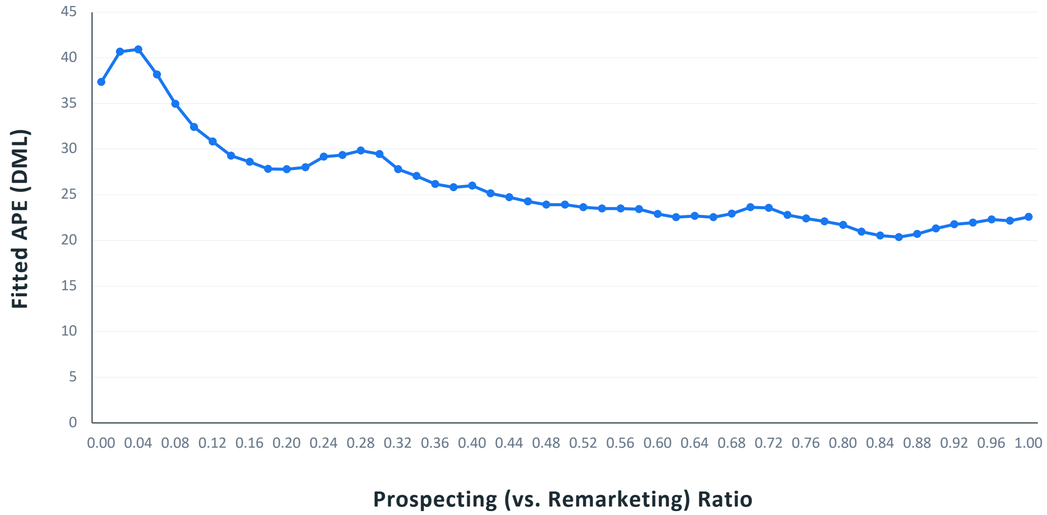
What we found: Even in the best case, non-experimental methods have large errors

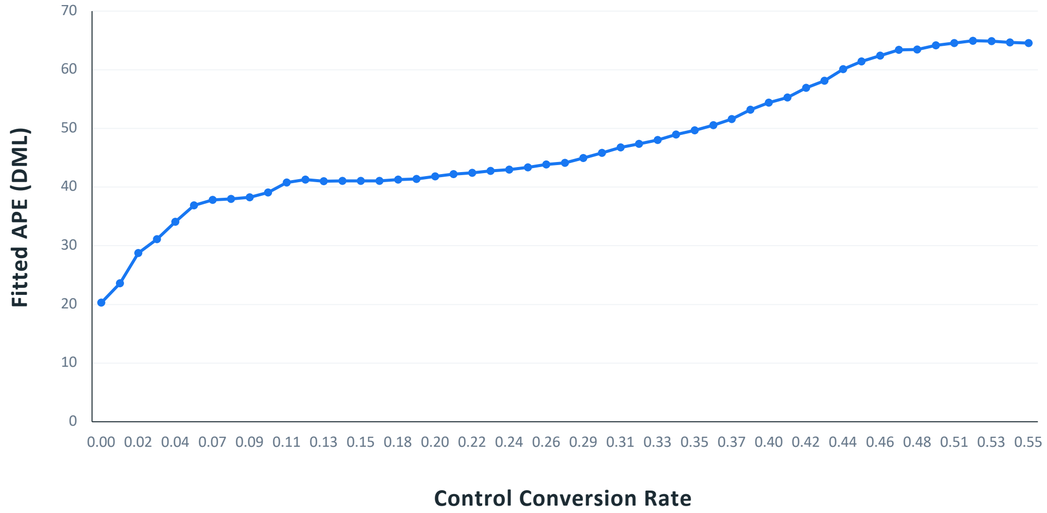
While we see overall that DML doesn't estimate RCT effects well, are there cases where it performs better?

By building and analyzing partial dependence plots, we see some moderate increases in performance for campaigns

- Using more prospecting ads vs. remarketing
- With lower baseline conversion rates
- Measuring upper funnel conversion events

However, even in these cases, we still see large error rates





Key Takeaways

Non-experimental methods are unlikely to succeed unless advertising platform pursue one of two paths.

01

Fundamentally change what data advertising platforms log and how they implement non-experimental methods.

02

Reframe using non-experimental data to estimate ad campaign effects as a prediction problem.

- Advertising platforms may run a large set of RCTs and therefore have a collection of “ground truths” of advertising effects.
- The unit of observation is now an RCT itself.
- Model the relationship between RCT lift and a set of easily observed non-causal “proxy” metrics such as simple last click counts.
- A companion paper to this one will pursue this approach.[1]

Source: (1) Gordon, B., R. Moakler, and F. Zettelmeyer (2022). Predicting Advertising Incrementality. Working Paper.

Place end of document link here