Observational Data for Heterogeneous Treatment Effects with Application to Recommender Systems

AKOS LADA, Facebook News Feed Ranking, USA
ALEXANDER PEYSAKHOVICH, Facebook Artificial Intelligence Research, USA
DIEGO APARICIO, IESE Business School, Spain; MIT Sloan, USA
MICHAEL BAILEY, Facebook Core Data Science, USA

Decision makers in health, public policy, technology, and social science are increasingly interested in going beyond 'one-size-fits-all' policies to personalized ones. Thus, they are faced with the problem of estimating heterogeneous causal effects. Unfortunately, estimating heterogeneous effects from randomized data requires large amounts of statistical power and while observational data is often available in much larger quantities the presence of unobserved confounders can make using estimates derived from it highly suspect. We show that under some assumptions estimated heterogeneous treatment effects from observational data can preserve the rank ordering of the true heterogeneous causal effects. Such an approach is useful when observational data is large, the set of features is high-dimensional, and our priors about feature importance are weak. We probe the effectiveness of our approach in simulations and show a real-world example in a large-scale recommendations problem.

CCS Concepts: • **Applied computing** → **Economics**; **Electronic commerce**; • **Computing methodologies** → Machine learning.

Additional Key Words and Phrases: Heterogeneous treatment effects; machine learning; observational data

ACM Reference Format:
Akos Lada, Alexander Peysakhovich, Diego Aparicio, and Michael Bailey. 2019. Observational Data for Heterogeneous Treatment Effects with Application to Recommender Systems. In *ACM EC '19: ACM Conference on Economics and Computation (EC '19), June 24–28, 2019, Phoenix, AZ, USA*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3328526.3329558

1 INTRODUCTION

Data-grounded approaches to decision making have become immensely popular in industry, health, and public policy [6, 35, 36, 41]. Making decisions using this data requires the ability to answer counterfactual questions, such as: 'What would happen if we implemented Z?' It is well known that because even the most comprehensive data sets will have unobserved confounders, models trained purely on observational data can give very misleading answers to questions about causal treatment effects [5, 35]. In this paper we show that, for certain counterfactual questions, observational data can be used to inform a prior on heterogeneity, which can improve the analysis design and the identification of treatment effects.

The traditional estimator in a randomized trial is the average treatment effect [5, 32, 33]. However, in practice most designs are likely to have different effects across different groups [24]. Understanding this heterogeneity is particularly important when a design is too expensive to implement

EC '19, June 24–28, 2019, Phoenix, AZ, USA

ACM ISBN 978-1-4503-6792-9/19/06…$15.00
https://doi.org/10.1145/3328526.3329558

on the entire population (and should therefore target those who will benefit the most) or when a design has positive effects for some, but may not be appropriate for others [1, 19, 33].

Unfortunately, in many domains average effects are small (in terms of signal-to-noise), the set of user-level covariates is high-dimensional, and we are a priori uncertain about which covariates are important predictors of heterogeneity. The combination of these conditions means that, in order to identify heterogeneous effects precisely, controlled analyses need to be very large and, in general, prohibitively costly [40].

Observational data is increasingly available in much larger quantities [35, 36, 49]. In online platforms, for example, these data can include precise measures on user activity, conversion rates, content order, and engagement over time. It is well known that using observational data as if it is randomized, in general, yields biased causal estimates due to the problem of confounding [2, 19]. However, we will show that this observational data can still be useful.

We suggest an approach that combines observational data with controlled experiments in order to learn heterogeneous effects. In essence, we propose using the observational data to estimate heterogeneous effects as if it were a randomized experiment. These estimates will be biased, however, we discuss the conditions on the structure of confounding that imply that rank-ordering of heterogeneity is preserved. We explore these conditions with a simple simulation study.

We then move to an example where observational data comes from a panel. Such data is common for many online applications. In the panel each unit has a covariate profile c_i, we observe some outcome y_i^t and variation in an endogenous variable of interest x_i^t at the individual level. We regress y on x at the individual level (alternatively: with fixed effects and interactions), take the individual level estimated coefficients, and train a classifier to predict each individual's implied regression coefficient (i.e. their personalized effect) from their covariate profile.

We apply this to predict ranks of treatment effects in two datasets. Our first is recommendations on a large social platform. We first use user-level observational panel data to estimate the relationship between page recommendations and user engagement with the platform. We then construct a classifier to construct a mapping from a large set of observable features c to the individual-level observationally estimated effect. We then use this classifier to predict treatment effects in a real randomized trial testing a new recommendation feature. Stratifying the trial by predicted treatment effects, we find an increasing relationship between the observationally estimated and the actual treatment effect. In addition to this large experiment, we also consider estimating heterogeneous effects in grants to microenterprises from recent work in development economics [17]. Here we use pre-experiment survey data as our observational component.

Of course, we do not claim that rank of heterogeneous effects will be preserved in all observational datasets. Further, we do not aim to replace randomization and standard heterogeneous effects analysis. Rather, we argue that analysts can simply add the observationally predicted heterogeneous effect into the feature space for whatever heterogenous effect model they plan to fit. This has a cost in terms of model complexity but if the covariate space is already high dimensional then this cost is negligible. However, if conditions are favorable, the benefits may be large.

1.1 Related literature

Our work relates to an extensive literature on the evaluation of heterogeneous treatment effects. Several methods traditionally allowed researchers to specify population subgroups and test for differences across them [24, 30]. These subgroups typically need to be specified before launching a controlled testing evaluation, e.g. time-stamping a pre-analysis plan, to avoid concerns that researchers might engage in p-hacking, or data snooping, that is to iteratively search across features or subgroups and present those that are significant [18, 23, 34, 42]. However, such restrictions can sometimes prevent researchers from finding unexpected true heterogeneous effects.

More recent work in machine learning, statistics, and econometrics is focused on automating the heterogeneous effects estimation problem [5] to both streamline the detection of heterogeneous effects and attenuate some of the dangers described above. There are many methods for estimated such effects including non-parametric procedures [15, 48], regularized regression [31, 44], decision trees [26, 47], causal trees and random forests [4, 50], Bayesian additive regression trees [13], neural networks [46], double machine learning [12], 'virtual twin' analyses [22], custom generated 'types' [21], and mixtures of models [27]. We do not propose a new method, rather we propose a way to combine observational and randomized data to learn heterogeneous treatment effects. While in this paper we use a panel-based approach to actually estimate the mapping between covariates and treatment effects, any of these existing methods can be used in its place.

Our work combines observational and randomized data. Recent advances in machine learning have looked at combining collections of experiments [28, 43] to learn causal effects (rather than treatment effects). An interesting direction for future work is to combine these methods with observational data.

Our work is also related to research combining biased and unbiased survey samples using 'data enriched' linear regression [11]. More generally, it relates to a growing literature on semi-supervised learning [14, 51], and multitask learning [10]. The closest to our work looks at learning feature representations using observational data for analysis of experimental data [37]. The idea behind all of these approaches is that multiple tasks that a machine learning method can be asked to do are not independent. Therefore, learning to perform one task can make a model better at a correlated task. Exploring connections between our panel-style setup and these methods is an interesting direction for future work.

2 HETEROGENEOUS EFFECTS

2.1 The Model

Consider the standard panel data problem in causal inference. We define i as the units of observation. Each i has a covariate profile c_i which may be high dimensional. Time is indexed by t, y is the continuous potential outcome scalar x is the endogenous variable of interest with potential policies affecting values of x.

We assume each user has a linear response to the treatment: when x_i increments by one unit, y_i will increase by β_i. We assume linearity because in many cases of interest our treatment will have relatively small effects on x and thus we are interested in the locally linear approximation of the true response function.This covariate is assumed to be fixed per user and not affected by policy choices.

There is a mapping f from covariates to treatment effects, $\beta_i = f(c_i) + e_i$, and the key to design a good personalization is to learn this mapping. For simplicity, assume that this function is linear, such that $\beta_i = c_i \cdot \zeta$, where $\cdot$ represents the standard dot product.

One way to estimate $f(c_i)$ is to run a large scale experiment. We can raise x_i by one unit in the treatment group, while we leave x_i unchanged in the control group, and estimate the relationship:

$$\beta_i = \mathbb{E}(f(c_i)) = \mathbb{E}(y \mid T_i = 1, c_i) - \mathbb{E}(y \mid T_i = 0, c_i)$$

which can be estimated using any off-the-shelf methods in the literature [4, 22, 26, 27, 31, 46–48, 50].

However, when treatment effects are relatively small and the covariate space is large, these methods require that we run large and expensive experiments. Advertising campaigns, for example, often have very small conversion rate effects, and it is not rare for even large scale controlled experiments to find insignificant single homogeneous effects [3, 38]. And therefore, estimating heterogeneous effects across subgroups can be more challenging.

We consider the case where we have an observational panel generated by following process:

$$x_i^t = \theta_i + \epsilon_i^t + z_i^t \psi_i \tag{1}$$

and

$$y_i^t = \mu_i + x_i^t \beta_i + z_i^t \gamma_i + \eta_i^t \tag{2}$$

Here ϵ and η are *iid* white noise error terms, z is some time varying unobserved variable (which in general can be a vector but for ease of notation we write as a scalar quantity), μ and θ are unobserved variables that are fixed at the user level. For simplicity, all of these variables have mean 0 with finite variances. This is without loss of generality as we could include some observed variables O_i^t, and then use, not the original x and y, but x and y conditional on O (i.e. the residuals).

Suppose we had observational panel data of the form (x_i^t, y_i^t, z_i^t) with time $t \in \{1, \ldots, T\}$, where T is large and Z was observed. We could estimate individual effects running unit-level regressions of y_i^t on (x_i^t, z_i^t). This would yield a consistent estimate of β_i. We could then learn ζ by regressing these unit-level estimates on c_i.

Consider using the same strategy without observing z. Let $\hat{\beta}_i$ be the result of running a regression of y on x using only unit i's data without including z. The coefficient $\hat{\beta}_i$ is the solution to the least squares problem $\hat{\beta}_i = (x_i' x_i)^{-1}(x_i' y_i)$. Substituting the structural equation for y and some algebra gives usthe usual omitted variable bias (OVB) expression:

$$\mathbb{E}(\hat{\beta}_i) = \beta_i + \underbrace{\gamma_i \frac{Cov(x_i, z_i)}{Var(x_i)}}_{\text{OVB}} \tag{3}$$

This illustrates an important difference between prediction and evaluation problems [2, 5, 35, 49]: $x_i \hat{\beta}_i$ is the best unbiased linear predictor of y_i but is not an estimate of the causal effect. In fact, when the covariance between x and Z, and y and Z is high, $\hat{\beta}$ can be an extremely biased estimator.

2.2 Rank-ordering from observational data

Observational estimates (although potentially biased) can be used to inform a prior about heterogeneous treatment effects. In particular, suppose that there are two units i, j with $\hat{\beta}_i > \hat{\beta}_j$. When does does this mean that $\beta_i > \beta_j$? We refer to this as observational heterogeneous effects being *rank unbiased*.

When rank unbiasedness holds, $\hat{\beta}_i$ is a sufficient statistic for targeting constrained programs (e.g. when the analyst can only afford to implement certain design to a limited percentage of the population). This also means that if we can learn a function $g(\cdot)$ which maps covariates c_i to $\hat{\beta}_i$, then this function will be a monotonic transformation of the true heterogeneous effects function $f(\cdot)$.

It is easy to see from equation 3 that a sufficient condition for rank unbiasedness is that

$$\frac{Cov(x_i, z_i)}{Var(x_i)} \gamma_i > \frac{Cov(x_j, z_j)}{Var(x_j)} \gamma_j.$$

A perhaps slightly more enlightening statement for the condition is to ask when the correlation between $\hat{\beta}$ and β is positive. This can be formally expressed as

$$Corr(\beta, \hat{\beta}) = Corr(\beta, \beta + \psi\gamma R) = \frac{Cov(\beta, \beta + \psi\gamma R)}{\sqrt{Var(\beta)Var(\beta + \psi\gamma R)}} =$$
$$= \frac{Var(\beta) + R * Cov(\beta, \psi\gamma)}{\sqrt{Var(\beta) * (Var(\beta) + R^2 * Var(\psi\gamma) + 2R * Cov(\beta, \psi\gamma))}} \quad (4)$$

Where $R \equiv \frac{\sigma_Z^2}{\psi^2\sigma_Z^2+\sigma_\epsilon^2}$ and for simplicity we denote as fixed. Although the expression depends on β, ψ, γ and their covariances, individualizing each effect is not straightforward.

2.3 An Example

The equations above are not straightforward to understand. We now give an economic model relevant to online behavior to help give intuition about the conditions. This simplified example will be the motivation for one of our later experiments. We first re-write our omitted variable bias equation substituting in the covariance and variance of various variables from our structural equations assuming unit variance for all variables to reduce notation

$$\hat{\beta}_i = \beta_i + \frac{\gamma_i\psi_i}{\psi_i^2 + 1}.$$

Consider an online social news aggregator with users i. Each time period t each user receives utility y_i^t from the platform as a function of many things, including how many articles are available, how much discussion their friends engage in, etc...

An analyst is interested in the effect of one particular variable: for which individuals does having more articles to read available on a given day (x_i^t) improve their experience most? We assume that the analyst observes a good proxy for y and x.

Users differ in their affinity to the platform and there is a daily shock z_i^t which is unobserved. The shock can be thought of as the 'newsworthiness' of a day. This shock affects baseline demand - this is γ_i in our structural equations above. In addition, this affinity affects the amount of articles a user has available to read on a given day, aka ψ_i in our equations.

We assume that higher affinity implies higher β_i. Now, our equations are simpler to understand. Consider fixing a user i and a counterfactual experiment of increasing their baseline affinity. This should increase β_i, γ_i, and ψ_i. If this increase in affinity increases γ_i by more than it increases ψ_i, then from our equation we see that omitted variable bias will increase. Thus, $\hat{\beta}_i$ will also increase whenever β_i increases.

In this example, such an effect seems reasonable: there are many features that affect how newsyness of a day affects demand for the news aggregator (e.g., how many friends will also discuss the news on the platform). A higher affinity affects all of them. By contrast, ψ only captures the impact of z through article supply, a much smaller channel.

3 EXPERIMENT 1: SIMULATION

The conditions discussed above are sufficient conditions for rank unbiasedness to hold. Necessary conditions are slightly harder to write down and make sense of. For this reason, we now investigate whether rank unbiasedness holds in a simulation study.

We generate unit data following structural equations (1) and (2). For each unit we draw $(\beta_i, \psi_i, \gamma_i)$ from a multivariate normal with expected value $(\mu_\beta, \mu_\psi, \mu_\gamma)$ and covariance matrix $(\sigma_{\beta,\psi}, \sigma_{\beta,\gamma}, \sigma_{\psi,\gamma})$. We set $T = 30$ days, $N = 50$ units, 100 replications. We ask whether $\hat{\beta}$ estimated from running unit level regressions correlates with the true β.

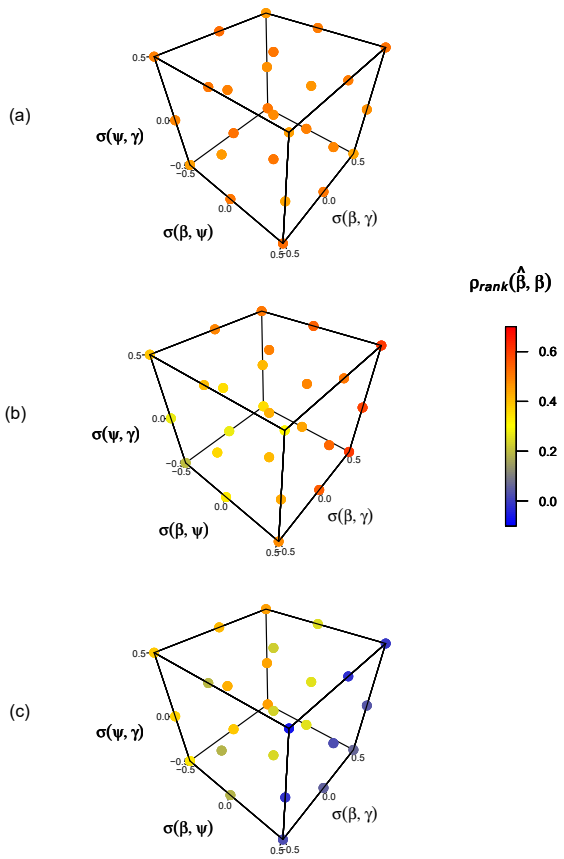

Fig. 1. Rank correlation

Notes: Results obtained from a Monte Carlo simulation where each user's $(\beta_i, \psi_i, \gamma_i)$ is drawn from a multivariate normal with expected value $(\mu_\beta, \mu_\psi, \mu_\gamma)$, variance 1, and covariance matrix $(\sigma_{\beta,\psi}, \sigma_{\beta,\gamma}, \sigma_{\psi,\gamma}$. Panel (a) shows the baseline case where $\mu_\beta, \mu_\psi, \mu_\gamma = 0$. Panel (b) shows a 'realistic' case where $(\mu_\beta, \mu_\psi, \mu_\gamma) = (0.5, 0.5, 0.75)$. Panel (c) shows an 'adversarial' case where $(\mu_\beta, \mu_\psi, \mu_\gamma) = (0.1, 0.1, -2)$, such that the correlation is usually poor. We note that these are selected examples for exposition purposes; results remain qualitatively the same under similar ranges of parameters.

Figure 1 shows the relationship between the parameters' covariance and the rank correlation between β and $\hat{\beta}$ (color scale). For instance, a top right edge shows the rank correlation, $\rho_{rank}(\hat{\beta}, \beta)$, for the case where β, ψ, γ are positively correlated. We discuss the simulation results from three scenarios. We start with a baseline case where (β, ψ, γ) are mean 0 and variance 1. Panel (a) in Figure 1 shows that the rank correlation is overall very stable across the covariance space, usually above 50%.

Panel (b) presents a realistic case, one in which β and ψ are positive but there is a higher expected value for γ, μ_γ, to illustrate a user demand that is mostly driven by some unobserved feature Z. A high rank correlation (in red) means that a high $\hat{\beta}_i$ predicts a high true β_i. Although $\hat{\beta}_i$ can be biased (in general overstating the effect from x_i), overall we find a reasonable rank correlation. And it is highest when the covariance is positive (top-right corner), because $\hat{\beta}$ also captures more

variation coming from Z and this is accentuated when β is also large. For instance, upward bias in nonexperimental studies is frequent when measuring advertising returns under activity bias, or when measuring schooling returns under ability bias [2, 9, 39]. This correlation breaks up when all covariances are negative (bottom-left corner), because when unobserved Z is large, users with large γ will have low β but high $\hat{\beta}$ (i.e. downward, negative bias).

So far we have seen that, across different covariances, the rank correlation is usually high, above 30%. However, this is not guaranteed to be the case for all types of data generating processes. If we continue to assume that user data is simulated through structural equations (1) and (2), what does it take for observational estimates to fail to predict true heterogeneity? Surprisingly, we find that it is difficult to construct covariance structures where observational estimates are 'inversely related' to causal estimates, and therefore rank correlation is low or incorrect.

We set $\mu_\beta, \mu_\psi, \mu_\gamma \in (-2, +2)$ and search the parameter space to yield low rank correlation. Panel (c) shows an adversarial case where this correlation is usually poor, and observational estimates fails to predict true heterogeneity. These correspond to $\mu_\beta = \mu_\psi = 0.25$ and $\mu_\gamma = -2$. However, it is worth discussing what these actually mean in practice. Such parameters imply that the variable of interest x (e.g. ads, online contents, ranking feature) has a low effect on demand, which is reasonable, but at the same that there is an unobserved factor, Z, that has a large negative effect. Under these conditions, for example when $\sigma_{\beta,\psi}$ is positive, then the rank correlation is close to 0. Why? Consider two users i, j such that $\beta_i < 0 < \beta_j$ but have otherwise similar parameters. Since i and j experience similar x but γ is negative and large, we will tend to estimate $\hat{\beta}_i > 0 > \hat{\beta}_j$, which is precisely the inverse rank order. This is because $\hat{\beta}$ corrects the sign to explain for the demand shock Z. Note that for this extreme choice of parameters $\sigma_{\beta,\psi}$ dominates the effect on the rank correlation.

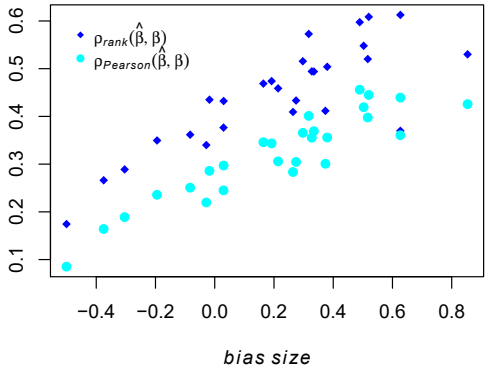

Fig. 2. Bias size

Notes: Size of the observational bias, and simple/rank correlation (between users' estimated $\hat{\beta}$ and true β). Same simulation as Panel (b) where $(\mu_\beta, \mu_\psi, \mu_\gamma) = (0.5, 0.5, 0.75)$

Finally, we also find that Spearman's rank correlation is less sensitive to outliers and higher than the simple correlation coefficient. See Figure 2. This suggests that, even though an observational bias (or simple correlation) can be large (small), the rank-ordering between observational and experimental estimates can remain very high. In other words, users' observational $\hat{\beta}$ can potentially be subject to a consistent (upwards or downward) bias, while it maintains the same experimental rank-ordering.

4 EXPERIMENT 2: FACEBOOK PAGE RECOMMENDATIONS

4.1 Experiment Details

We now discuss the use our approach to a large scale ranking recommendation problem at Facebook. We are interested in estimating users' heterogeneous response to page recommendations in order to better customize Facebook's News Feed. We first describe the problem.

The Facebook social network has about several billion monthly active users. Users each have content available to them for view. The News Feed is an individualized ranked list of content including status updates, photos, videos, links, app activity, and likes from people, pages and groups that users follow on Facebook.

Pages are non-user entities on Facebook than can post content (e.g. organizations, blogs, certain celebrities). Users can 'Like' a page to connect to it: liking a page makes posts created by this page eligible to appear in the user's News Feed. Page content corresponds to a portion of content individuals can view on their News Feed.

Helping users connect to the right pages can greatly improve their experience on the site. A personalized Facebook recommendation is used to suggest relevant pages to users. In practice, these recommendations show up in a user's News Feed in the form of 'Pages You May Like' units. See Figure 3 for a visual example.

However, facilitating these connections is costly. There is an opportunity cost, because each of the 'Recommended Pages' takes up some space on the News Feed, and there is also a user experience cost, because users who do not want more connections can be inconvenienced with unnecessary recommendations. Therefore, our question of interest is **not** *which page* should we suggest as in standard recommender systems [8]. Rather, we ask: *which users'* experience will benefit the most with additional page recommendations [16, 45]?

Here we interpret structural equations (1) and (2) as follows. We define the outcome variable or user demand, y_i^t, as user's i measure of engagement with the platform on day t, on either a desktop or mobile device.

As measures we simply use time spent on the platform. We note this is a crude measure of user demand but is sufficient for the purposes of this paper and that other measures of engagement give qualitatively similar results. We let x_i^t as user i's page inventory supply at time t. The page inventory of user i is the number of posts made on a given day by all pages that she or he is connected to (i.e. has selected to be a fan of). Adding more page connections will (stochastically) increase page inventory levels. Recall that x_i^t varies from day to day because some days experience more events than others, and therefore pages produce more content. Importantly, a user may not necessarily view, or engage with, all of her or his page inventory.

Our observational analysis will be confounded by many unobserved confounders including time varying affinity for the platform from each user as well as unobserved shocks that jointly increase activity in individuals and pages (creating more page inventory).

During Fall 2015 Facebook tested a new type of recommendation system which could show a 'representative' post from a page in a user's News Feed (with the header 'Recommended For You'). The eligible 'Recommended For You' recommendations were obtained from a pool of user-page pairs which had very high similarity scores according to Facebook's standard recommender system.

To evaluate whether these units improved user experience, Facebook performed a controlled version testing on randomly chosen users. From the approximately 8 million people who were eligible for this test (eligibility required that the underlying recommender system be confident in their suggestions of potential new pages), close to 400,000 were randomly chosen to see these new recommendation units. We use a control group of 8 million individuals who were eligible but did not see the new unit.

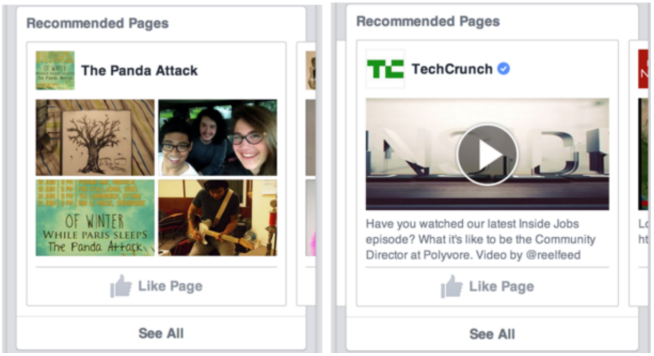

Fig. 3. An example page recommendation unit on Facebook from 2014.

To define the unit-level covariates, c, we use a large set of variables that are generally useful in the platform's machine learning systems [20, 25, 29]. For instance, user time-constant covariates (geography, age, operating system), network-related metrics (number of friends, URLs), and past engagement measures.

4.2 Estimating Observational Heterogeneity

We estimate observational treatment effects using a large user-level dataset. We take a random, deidentified panel of 120 million Facebook users. We consider 60 days of data per individual. We then compute $\hat{\beta}_i$ for each of the users by running user-level panel regressions on 60 days of (y_i^t, x_i^t) pairs. Due to the fat-tailed nature of y we use a log transformation. This yields our set of user-level coefficients $\hat{\beta}_i$.

To be able to generalize our estimates to individuals outside the sample we train a model predict an individual's $\hat{\beta}$ from c_i. We refer to this as $\hat{g}(c_i)$. Due to outliers, skewness, and noise in the estimates of $\hat{\beta}$ that often occur with sparse data, we instead use a version of quantile regression [2]. Here we report the results of a classifier which assigns a label of 1 to those users whose estimated $\hat{\beta}_i$ is in the top 20% of all estimated $\hat{\beta}_i$, and 0 otherwise.

We found this to be useful for both reducing noise and increasing interpretability of how well the procedure is performing since a classification problem also allows us to evaluate the performance of the machine learning in terms of AUC (which for ranking is more interpretable than MSE). We used what was, at the time, the default Facebook machine learning system, based on gradient boosted decision trees as feature transformers followed by a final linear layer to train a classifier using these features. See [29] for implementation details. We achieve an AUC of approximately $\sim .78$.

This trained classifier estimates our function $\hat{g}(\cdot)$. Note that the output of the classifier is the probability that a user with covariate profile c_i is in the top quantile of the treatment effects, this gives us an ordering of individuals that we can compare to randomized estimates.

4.3 Estimating Experimental Heterogeneity

We now evaluate the extent to which these observational estimates predict treatment effects in randomized data. We take y, is defined as overall time spent during 1 week of the experiment. We transform the variable taking natural logs and reduce variance by difference-out 1 week of pre-analysis engagement per user. That is, the outcome becomes Δy instead of y. Because y is highly autocorrelated and right-skewed, these two transformations increase statistical power.

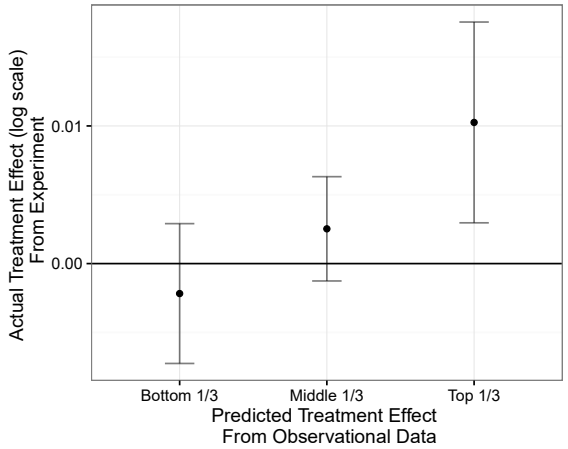

Fig. 4. Rank-ordered relationship between the actual treatment effect and the predicted treatment effects (probability of being in the top 20%).

We now ask ourselves the following. Is there an increasing relationship between our $\hat{g}(c)$ and the actual treatment effect? Figure 4 shows that this is indeed the case. Stratifying the testing population by predicted treatment effects, we find an increasing and monotonic relationship between $\hat{g}(c)$ and the actual controlled treatment effect. As usual, the latter is computed as the average difference between treatment and control groups. Note that the error bars are quite large because even though there are approximately 8 million people in the control group, there are close to 400,000 people in the treatment group and the analysis involves a very small change in the user experience. This is consistent with the challenges of estimating economically meaningful magnitudes even in large scale experiments [3, 38]. Moreover, the overall effect is relatively small as the treatment involves inserting a single extra recommendation unit into users' News Feed.

Variable	**Coefficient**	**Standard Error**
Intercept	-.0316	.001***
Treatment Dummy	-.0064	.0033*
Predicted Effect	.2045	.002***
Predicted * Treatment	.035	.01**

Regression of pre-post treatment change of log engagement on treatment, predicted effect and the interaction. *** p<0.001, ** p<0.01, *p<0.1.

Table 1. Regression table

Our results are reinforced in a linear regression. We regress experimental y on an intercept, treatment dummy, predicted treatment effect, and the interaction of treatment dummy with observationally predicted treatment effect. We also control for the pre-analysis user-level y to reduce

variance. We find a statistically significant interaction effect ($p < 0.01$) which confirms the visual impression in Figure 4.

5 EXPERIMENT 3: RETURNS TO CAPITAL IN MICROENTERPRISES

Predicting the ranking of unit heterogeneity is best suited to estimation problems where the researcher has experimental data together with observational data, possibly with high dimensional covariates. This is often the case in online platforms where rich browsing user-level data can be collected at any point in time.

We now consider a slightly different applications: development economics. Researchers in this field often run comprehensive baseline surveys prior to launching a complete large scale, expensive experiment [7, 19, 49]. These surveys help the researcher to gain preliminary insight into potential pitfalls of the experiment, power calculations, logistics issues related to data collection, attrition and non-compliance rates, useful covariates and outcome variables, all of which are better to address before rolling out an experiment to the entire sample. In fact, experiments in development can cost millions of dollars and often take months, or even years, to complete [7, 19, 42].

We consider the dataset of De Mel et al.. This work studies microenterprises in Sri Lanka and tries to estimate the returns to capital. In the language of our structural equations, the outcome variable y are enterprise profits, x is the enterprise's capital stock. There is relatively comprehensive covariate data on each of the microenterprises including (e.g. education, gender, age, family size, assets, loans, ability measure, source of credit, industry, size).

Like many development economics experiments, this dataset includes a randomized trial where extra capital is randomly assigned to microenterprises as well as an observational component of baseline surveys conducted on 408 enterprises meeting the program eligibility criteria, including regional areas, firms with invested capital below US$1,000, self-employed workers, and between the ages of 20 and 65.

Following equations (1) and (2), the basic regression to be estimated is:

$$y_i^t = \alpha + \mu_i + \delta_t + x_i^t \beta_i + O_i^t \kappa_i + \eta_i^t \tag{5}$$

where y_i^t is profits and x_i^t is capital stock for enterprise i at time t; μ_i and δ_t are enterprise- and time- fixed effects; and O_i^t are a set of observable covariates.

However, a grant randomization allows to break up this OVB because enterprises in the treatment group who receive a cash grant are otherwise similar to those with no cash grant in the control group. Thus this exogenous grant can be used as an 'instrument' to estimate the returns to capital. In practice, this is estimated using two-stage least squares (2SLS). 2SLS is the ratio of the coefficients from the 'reduced form' (regression of profits on the instrument) and the 'first stage' (regression of capital stock on the instrument). See [2, 32] for additional details on instrumental variables.

We will use the baseline surveys along with these covariates to estimate the observational model ignoring Z. Since we have a cross-sectional dataset, not a panel as in our discussion above we need to perform a slightly different analysis. First, we discretize the continuous covariates into indicator variables, which take the value of 1 if feature s for household i is above the median. For example, $c_{i,\text{assets}} = 1$ if household i has above median durable assets. Second, we interact capital stock, x_i, with these household-level features. That is, we use the baseline survey to estimate a regression of the form:

$$y_i = \alpha + x_i \beta + \sum_{s=1}^{S} \phi_s x_i * c_{i,s} + \sum_{s=1}^{O} c_{i,s} \kappa_s + \eta_i \tag{6}$$

where $c_{i,s}$ is observable covariate s for enterprise i; and $s \in S, S \in O$ because we do not test all possible interactions. To reduce over-fitting and multicollinearity, we interact the capital stock

with four features: ability, english speaking, education, and assets. The parameter ϕ_s captures how the effect varies with feature s. Note that there are no subindices t because the baseline survey is a cross-section.

Therefore, instead of predicting an individual-level $\hat{\beta}_i$, we use the coefficients trained in the observational data, to inform a prior of heterogenous treatment effects for those subgroups with covariate profile c in the experimental dataset. That is,

$$\hat{\beta}^{\text{pred.}}(c) = \hat{\beta}^{obs} + \sum_{s=1}^{S} \hat{\beta}_s^{obs} * c_i.$$

Once we obtain the predictions $\hat{\beta}^{\text{pred.}}(c)$ in the experimental sample, we stratify enterprises into two groups: enterprises with predicted treatment effects above or below the median. We then compare these predicted treatment effects with the actual 2SLS experimental estimates in each group.[1]

	(1)	(2)	(3)
	Full sample	Low $\hat{\beta}^{\text{pred.}}$	High $\hat{\beta}^{\text{pred.}}$
2SLS			
Capital stock	0.36	0.25	0.66
	$(0.11)^{***}$	$(0.12)^{**}$	$(0.24)^{***}$
First stage			
Treatment	0.33	0.41	0.18
	$(0.04)^{***}$	$(0.04)^{***}$	$(0.04)^{***}$
Adjusted R^2	0.69	0.70	0.65
Observations	2,620	1,652	968

Notes: Log real monthly profits as outcome, log capital stock as endogenous variable, and grant amount as instrumental variable. Capital stock and profits are measured in Sri Lankan rupees. All regressions include time and enterprise fixed effects. Standard errors clustered at the enterprise-level shown in parenthesis. *** p<0.01, ** p<0.05, * p<0.1.

Table 2. Heterogenous returns to capital

We find that enterprises with larger predicted treatment effects correlate with larger causal treatment effects. The magnitudes, shown in Table 2, are qualitatively similar to [17]'s estimates of returns to capital (see also their Online Appendix). The coefficients on the first stage show that the grant is highly significant in instrumenting the capital stock. Column (1) shows that if we run 2SLS on the full experimental sample, we obtain a coefficient of 0.36 which implies a rate of return of 5.25% per month. This point estimate is very close to that in Table IV in [17]. The estimates in columns (2) and (3), 0.25 and 0.66, respectively, suggest large and significant heterogenous treatment effects across households. If we examine the covariates that account for this heterogeneity we find

[1]Leave-one-out or repeated split estimators can also be used to stratify the sample to reduce in-sample overfitting [1]. This is beyond the scope of this paper. The predicted outcomes are not based on in-sample experimental controls; and observational over-fitting tends to produce upward (downward) bias for low (high) predicted units, which would therefore make the magnitudes in Table 2 conservative.

that those in the group with higher predicted treatment effects have statistically larger values of education, english speakers, ability, and assets.

Finally, these results complement [17]'s comparison of nonexperimental and experimental estimates. Although in this study nonexperimental estimates of returns to capital can suffer from attenuation bias (downward bias due to measurement error), our results suggest that observational data can be used to find a prior on treatment effects heterogeneity across groups. Therefore, rich observational data combined with experimental data can be used to increase statistical power or to personalize programs when treatment is limited.

6 CONCLUSION

Estimating heterogeneous treatment effects is a particularly challenging problem when the set of individual-level covariates is large and our priors about the important ones are weak. Moreover, the magnitudes of treatment effects are often very small, and therefore even large scale controlled tests can fail to find statistically significant results. Although controlled experiments will remain the gold standard for causal inference, there are increasingly large observational datasets in online platforms, policy, and health, which can be used to improve the personalization problem.

We suggest combining time series observational estimates with controlled testing to learn the mapping from unit-level features to the size of the unit-level causal treatment effects. We show that, when observational estimates preserve the rank ordering of individuals' true heterogeneous causal effects, these data can be used to construct a useful prior on heterogeneous treatment effects. These estimates are informative to design a costly treatment when controlled analyses are unfeasible or to facilitate the identification of heterogeneity.

In the main empirical application, Section 4, we use this approach to estimate heterogeneity in users' engagement with page recommendations at Facebook's News Feed. We believe this is particularly relevant to a broader class of problems in online platforms, where unprecedented amounts of content need to be ranked and delivered to each user's limited time and space interface. And therefore improving the interaction with these contents can substantially improve the user experience.

ACKNOWLEDGMENTS

The authors would like to thank Mert Demirer, Dean Eckles, Glenn Ellison, Mateo Montenegro, and Martin Savransky for helpful comments. The views expressed in this paper are those of the authors, and do not necessarily reflect those of Facebook. All errors are our own. An earlier version of this paper with a smaller coauthor list, fewer experiments, and analysis was circulated on arxiv.org and presented in non-archival venues.

REFERENCES

[1] Alberto Abadie, Matthew M Chingos, and Martin R West. 2013. Endogenous stratification in randomized experiments. *NBER WP 19742* (2013).

[2] Joshua D Angrist and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion.* Princeton university press.

[3] Diego Aparicio and Drazen Prelec. 2018. Choice overload in online platforms. *Working Paper* (2018).

[4] Susan Athey and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113, 27 (2016), 7353–7360.

[5] Susan Athey and Guido W Imbens. 2017. The state of applied econometrics: Causality and policy evaluation. *The Journal of Economic Perspectives* 31, 2 (2017), 3–32.

[6] E. Bakshy, D. Eckles, and M. S. Bernstein. 2014. Designing and Deploying Online Field Experiments. In *Proceedings of the 23rd ACM conference on the World Wide Web*. ACM.

[7] Abhijit Banerjee and Esther Duflo. 2012. *Poor economics: A radical rethinking of the way to fight global poverty.* PublicAffairs.

[8] James Bennett and Stan Lanning. 2007. The netflix prize. In *Proceedings of KDD cup and workshop*, Vol. 2007. 35.

[9] Thomas Blake, Chris Nosko, and Steven Tadelis. 2015. Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. *Econometrica* 83, 1 (2015), 155–174.

[10] Rich Caruana. 1998. Multitask learning. In *Learning to learn*. Springer, 95–133.

[11] Aiyou Chen, Art B Owen, Minghui Shi, et al. 2015. Data enriched linear regression. *Electronic Journal of Statistics* 9, 1 (2015), 1078–1112.

[12] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, et al. 2016. Double machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060* (2016).

[13] Hugh A Chipman, Edward I George, Robert E McCulloch, et al. 2010. BART: Bayesian additive regression trees. *The Annals of Applied Statistics* 4, 1 (2010), 266–298.

[14] Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*. ACM, 160–167.

[15] Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. 2008. Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics* 90, 3 (2008), 389–405.

[16] Abhinandan S Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. 2007. Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th international conference on World Wide Web*. ACM, 271–280.

[17] Suresh De Mel, David McKenzie, and Christopher Woodruff. 2008. Returns to capital in microenterprises: evidence from a field experiment. *The quarterly journal of Economics* 123, 4 (2008), 1329–1372.

[18] Angus Deaton. 2010. Instruments, randomization, and learning about development. *Journal of economic literature* 48, 2 (2010), 424–455.

[19] Esther Duflo, Rachel Glennerster, and Michael Kremer. 2007. Using randomization in development economics research: A toolkit. *Handbook of development economics* 4 (2007), 3895–3962.

[20] Dean Eckles and Eytan Bakshy. 2017. Bias and high-dimensional adjustment in observational studies of peer effects. *Working Paper* (2017).

[21] Ziv Epstein, Alexander Peysakhovich, and David G Rand. 2016. The good, the bad, and the unflinchingly selfish: Cooperative decision-making can be predicted with high accuracy when using only three behavioral types. In *Proceedings of the 2016 ACM Conference on Economics and Computation*. ACM, 547–559.

[22] Jared C Foster, Jeremy MG Taylor, and Stephen J Ruberg. 2011. Subgroup identification from randomized clinical trial data. *Statistics in medicine* 30, 24 (2011), 2867–2880.

[23] Annie Franco, Neil Malhotra, and Gabor Simonovits. 2014. Publication bias in the social sciences: Unlocking the file drawer. *Science* 345, 6203 (2014), 1502–1505.

[24] M Gail and R Simon. 1985. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* (1985), 361–372.

[25] Brett R Gordon, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. 2019. A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Marketing Science* (2019).

[26] Donald P Green and Holger L Kern. 2012. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opinion Quarterly* (2012), nfs036.

[27] Justin Grimmer, Solomon Messing, and Sean J Westwood. 2014. Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Unpublished manuscript, Stanford University, Stanford, CA* (2014).

[28] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. 2017. Deep IV: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 1414–1423.

[29] Xinran He, Junfeng Pan, Ou Jin, Tianbing Xu, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers, et al. 2014. Practical lessons from predicting clicks on ads at facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*. ACM, 1–9.

[30] James J Heckman and Edward Vytlacil. 2005. Structural equations, treatment effects, and econometric policy evaluation. *Econometrica* 73, 3 (2005), 669–738.

[31] Kosuke Imai, Marc Ratkovic, et al. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics* 7, 1 (2013), 443–470.

[32] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

[33] Guido W Imbens and Jeffrey M Wooldridge. 2009. Recent developments in the econometrics of program evaluation. *Journal of economic literature* 47, 1 (2009), 5–86.

[34] John PA Ioannidis. 2005. Why most published research findings are false. *PLoS medicine* 2, 8 (2005), e124.

[35] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. Prediction policy problems. *The American economic review* 105, 5 (2015), 491–495.

[36] Ron Kohavi, Randal M Henne, and Dan Sommerfield. 2007. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 959–967.
[37] Soren R Kunzel, Bradly C Stadie, Nikita Vemuri, Varsha Ramakrishnan, Jasjeet S Sekhon, and Pieter Abbeel. 2018. Transfer Learning for Estimating Causal Effects using Neural Networks. *Working paper* (2018).
[38] Randall A Lewis and Justin M Rao. 2015. The unfavorable economics of measuring the returns to advertising. *The Quarterly Journal of Economics* 130, 4 (2015), 1941–1973.
[39] Randall A Lewis, Justin M Rao, and David H Reiley. 2011. Here, there, and everywhere: correlated online behaviors can lead to overestimates of the effects of advertising. In *Proceedings of the 20th international conference on World wide web*. ACM, 157–166.
[40] Mark W Lipsey. 1990. *Design sensitivity: Statistical power for experimental research*. Vol. 19. Sage.
[41] Michelle N Meyer. 2015. Two Cheers for Corporate Experimentation: The A/B Illusion and the Virtues of Data-Driven Innovation. *J. on Telecomm. & High Tech. L.* 13 (2015), 273.
[42] Benjamin A Olken. 2015. Promises and perils of pre-analysis plans. *The Journal of Economic Perspectives* 29, 3 (2015), 61–80.
[43] Alexander Peysakhovich and Dean Eckles. 2018. Learning Causal Effects From Many Randomized Experiments Using Regularized Instrumental Variables. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 699–707.
[44] Alexander Peysakhovich and Jeffrey Naecker. 2017. Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity. *Journal of Economic Behavior & Organization* 133 (2017), 373–384.
[45] Paul Resnick and Hal R Varian. 1997. Recommender systems. *Commun. ACM* 40, 3 (1997), 56–58.
[46] Uri Shalit, Fredrik Johansson, and David Sontag. 2016. Bounding and Minimizing Counterfactual Error. *Working Paper* (2016).
[47] Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David M Nickerson, and Bogong Li. 2009. Subgroup analysis via recursive partitioning. *The Journal of Machine Learning Research* 10 (2009), 141–158.
[48] Matt Taddy, Matt Gardner, Liyun Chen, and David Draper. 2014. A nonparametric Bayesian analysis of heterogeneous treatment effects in digital experimentation. *arXiv preprint arXiv:1412.8563* (2014).
[49] Hal R Varian. 2014. Big data: New tricks for econometrics. *The Journal of Economic Perspectives* 28, 2 (2014), 3–27.
[50] Stefan Wager and Susan Athey. 2015. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Working Paper* (2015).
[51] Xiaojin Zhu. 2006. Semi-supervised learning literature survey. *Computer Science, University of Wisconsin-Madison* 2, 3 (2006), 4.