Measuring the Return to Online Advertising: Estimation and Inference of Endogenous Treatment Effects

Shakeeb Khan[1], Denis Nekipelov[2], Justin Rao[3]

Abstract

In this paper we aim to conduct inference on the "lift" effect generated by an online advertisement display: specifically we want to analyze if the presence of the brand ad among the advertisements on the page increases the overall number of consumer clicks on that page. A distinctive feature of online advertising is that the ad displays are highly targeted- the advertising platform evaluates the (unconditional) probability of each consumer clicking on a given ad which leads to a higher probability of displaying the ads that have a higher *a priori* estimated probability of click. As a result, inferring the *causal* effect of the ad display on the page clicks by a given consumer from typical observational data is difficult. To address this we use the large scale of our dataset and propose a multi-step estimator that focuses on the tails of the consumer distribution to estimate the true causal effect of an ad display. This "identification at infinity " (Chamberlain (1986)) approach alleviates the need for independent experimental randomization but results in nonstandard asymptotics. To validate our estimates, we use a set of large scale randomized controlled experiments that Microsoft has run on its advertising platform. Our dataset has a large number of observations and a large number of variables and we employ LASSO to perform variable selection. Our non-experimental estimates turn out to be quite close to the results of the randomized controlled trials.

JEL Classification: C14, C31, C55, C90, M37.

Keywords: Endogenous treatment effects, randomized control trials, online advertising, lift effect.

[1] Department of Economics, Boston College.

[2] Departments of Economics and Computer Science, University of Virginia.

[3] Microsoft Research.

1 Introduction

In this paper we aim to determine the causal effect of online advertising on consumer behavior. We do so in the context of analyzing a dataset that characterizes consumer behavior (clicks) on the page of search results on Bing.com, Microsoft's search engine. The page of search results on Bing.com, similarly to other search engines contains the "organic" content, which is the list of links to the content that Bing's search algorithm identified as relevant to the search keywords. In addition, the page has clearly marked "paid" content which is the set of links that were placed on the page because the advertisers paid for them to be displayed. In our data we observe consumer search term, the content of the page with all organic and paid results, and the consumer clicks on the page within the same search session. Of interest in this study is to evaluate the effect of ads on the overall consumer clicks on the page of search results.

Evaluation of the effect of advertising is important for monetization in the industries whose revenues are exclusively or mainly rely on advertising revenues. At the same time, the increase in the targeting capabilities has led to an overall conflation of the effect of ad targeting from the effect of advertising per se. In this case it is difficult to distinguish whether the display of an ad has *caused* the consumer to act on that ad (such as click on it or purchase the product) or if the consumer with the targeted characteristics has already had an intent to perform an action that the ad was promoting. In typical marketing settings measuring the true causal effect of an ad requires some method to control for this endogeneity.

The particular form of advertising which we focus on is known as *search engine marketing* (SEM). Its causal effect on outcome variables of interest such as clicks or sales can be particularly difficult to recover due to endogeneity that can arise from one of two sources. First, unlike other advertising channels, the internet lets advertisers target their ads to the activity that users are engaged in. Second, the technology allows advertisers to track variables that should help measure the efficacy of ads. Failure to account for simultaneity (due to either or both reasons) will generally result in overestimating the treatment effect of SEM.

Consequently, recent work has estimated the causal effect of SEM using large scale field experiments to address endogenity. Lewis and Rao (2014) conclude that the SEM effect is not statistically significant and Blake, Nosko, and Tadelis (2015) find that the *return on investment* (ROI) for brand advertising can be negative in many experiments. These papers and approaches are welcome additions to literature, but the field experiment approach to controlling for endogeneity can be limited in scope. As discussed in Lewis and Rao (2014) such experiments can be extremely expensive when testing certain hypotheses.

Motivated by these conclusions, in this paper we propose new methods to conduct inference on the treatment effect of advertising using *observational data.* We propose an alternative inference procedure for the causal effect of the discrete action of placing an advertisement on clicks and/or sales. Generally speaking, discrete endogenous regressors (such as ad placement) are frequently encountered in econometric models, and failure to correct for endogeneity can result in incorrect inference. Furthermore, in nonlinear models estimation and inference on regression coefficients of these regressors can be especially difficult because "control function" approaches will not be valid (Blundell and Powell (2004)). Furthermore, identification can only be achieved at extreme values of observed variables, as shown in Blundell and Powell (2004), Lewbel (2000), Khan and Nekipelov (2010, 2016, 2017).

As we will explain in detail, our approach is based on limiting values of control variables to deal with endogeneity problem. Asymptotic properties of such an approach for models with discrete endogenous variables was recently studied in Khan and Nekipelov (2016) and were found to be nonstandard in two dimensions. For one the asymptotic approximation requires larger sample sizes as only data in the tails of the distribution are used in estimation. Second the limiting distribution was not normal and construction of confidence sets requires the use of subsampling methods. Fortunately we can deal with both these issues in this empirical example. This is because our data set is comprised of millions of observations, and for inference we can explore extreme behavior of explanatory variables as explored in Khan and Nekipelov (2010) and Khan and Nekipelov (2017).

The rest of the paper is organized as follows. The next section describes the econometric model we wish to estimate, and proposes an estimation procedure of the parameters of interest. Section 3 explains why inference for this parameter is nonstandard (regardless of the estimation procedure) and Section 4 proposes new estimation and inference procedures. Section 5 further explores the finite sample properties of our procedure by means of simulations, one of which is "empirical" in the sense that the designs used were based on features of the data used in the application. Section 6 discusses in detail the data made available from Microsoft that we will apply our estimation and inference procedures to. The results attained are then compared to those found using the field experiment approach. Finally, Section 7 concludes with a summary of results and suggest future areas of research including other models our procedure can apply to as well as other relevant data sets that can be used.

2 Semiparametric Model

We express our model as a separable semiparametric, or partially linear treatment effect model, in which we wish to evaluate the impact of a binary treatment D on the outcome

Y. The treatment assignment depends on the vector of observable consumer characteristics ξ and the unobservable disturbance V while the treatment outcome depends on the vector of observable characteristics ζ and the unobservable disturbance U. The full model can be written as

$$\begin{aligned}&Y = \alpha\, D + f(\xi) + U\\ &D = \mathbf{1}\{g(\zeta) - V \geq 0\},\end{aligned}$$

where the functions $f(\cdot)$ and $g(\cdot)$ are unknown. We assume that (U, V) has an arbitrary correlation structure while $E[U \,|\, \xi, \zeta] = 0$. The object of interest is the treatment effect α. We note that the standard treatment effect estimator in this case will suffer from the omitted variable bias due to correlation between U and V. In the controlled experiment settings in Blake, Nosko, and Tadelis (2015) and Lewis and Rao (2014) D can be viewed as exogenous, in which case estimation and inference for α can be performed using existing methods for the semi linear model- see, e.g. Robinson (1988), Newey and Donald (2014) when the dimension of ξ is fixed, and Cattaneo, Jansson, and Newey (2016) for the case when it increases with the sample size.

However, endogeneity of treatment complicates identification and estimation of α. To provide the identification argument for α denote $X = f(\xi)$ and $Z = g(\zeta)$. We fix those variables focusing on the identification of α. Note that

$$(Y - X)\, D = \alpha\, D + U\, D,$$

and thus

$$E[(Y - X)\, D \,|\, X, Z] = \alpha\, F_V(Z) + E[U\, \mathbf{1}\{V \leq Z\} \,|\, X, Z].$$

Next, assuming that sufficient measurability conditions are satisfied we notice that

$$\lim_{Z \to \infty} E[U\, \mathbf{1}\{V \leq Z\} \,|\, X, Z] = E[U \,|\, X] = 0$$

and

$$\lim_{Z \to -\infty} E[U\, \mathbf{1}\{V \leq Z\} \,|\, X, Z] = E[U \cdot 0 \,|\, X] = 0.$$

Also notice that $F_V(z)$, the conditional cdf of V, is the standard "propensity score" $P(D = 1 \,|\, Z = z)$ and we denote it by $P(z)$. As a result, we can write

$$\alpha = \lim_{t \to +\infty} \frac{E[(Y - X)\, D \,|\, X, Z = t] - E[(Y - X)\, D \,|\, X, Z = -t]}{P(t) - P(-t)}.$$

This will be the main object of interest that we recover from the data.

Next, we turn attention to X and Z and note that they are also identified. First of all, note that

$$P(Z) = E[D \,|\, Z] = E[D \,|\, \zeta].$$

Next, note that

$$E[Y \mid \zeta, \xi] = \alpha P(Z) + X + E[U \mid \xi, \zeta] = \alpha P(Z) + f(\xi).$$

Then for any ξ and $\xi' \neq \xi$ and any ζ we can write

$$E[Y \mid \xi', \zeta] - E[Y \mid \xi, \zeta] = f(\xi') - f(\xi),$$

which identifies $f(\cdot)$ up to scale. The scale then can be identified by noticing that

$$E[Y \mid \xi, \zeta] = E[Y \mid \xi, Z] = \alpha P(Z) + f(\xi).$$

And thus

$$f(\xi) = \lim_{Z \to -\infty} E[Y \mid \xi, Z].$$

Using this expression we can establish another expression relating α to observed variables as

$$\alpha = \frac{\lim\limits_{Z \to +\infty} E[Y \mid X, Z] - \lim\limits_{Z \to -\infty} E[Y \mid X, Z]}{\lim\limits_{Z \to +\infty} P(Z) - \lim\limits_{Z \to -\infty} P(Z)}. \tag{2.1}$$

3 Fisher information for causal effect α

The fact that the identification of the parameter of interest α requires the full exploration of the support of index variables X and Z has implications for the *quality* of its identification. More specifically, we are able to demonstrate that even in the model where function $f(\cdot)$ and $g(\cdot)$ are known, the Fisher information for α is equal to zero (in the terminology of Ibragimov and Has'minskii (1981) and Bickel, Klaassen, Ritov, and Wellner (1993)). To illustrate this property, we consider this simpler model by using our previous notation for the unobservable variables (U, V) and observable variables (X, Z) that we consider, we make the following assumption:

Assumption 1 *(i) The index variables X and Z have a joint distribution where Z has a full support on $\mathbb{R}$ with the joint support not contained in any proper one-dimensional subspace. The parameter of interest is in the interior of a convex compact set $\mathcal{A}$;*

(ii) (U, V) have an absolutely continuous density conditional on X and Z with full support on $\mathbb{R}^2$;

(iii) The conditional density of $U \mid V \leq v, X = x, Z = z$ is bounded, strictly positive for all v, x, z and has continuous derivative such that there exists function $q(\cdot, \cdot)$ with $E[q(X, Z)^2] < \infty$ which dominates this derivative.

We begin our analysis by noticing that we can construct examples of parametric distributions for the errors and covariates in the triangular model in which the variance of the score for parameter α is infinite. The simplest way to construct such examples is to consider cases of high correlation between errors U and V. This can reflect the situation where both equations determining consumer clicks and ad display are driven by common unobservable components.

The zero information result can be "repaired" in parametric models by assuming that covariates have bounded support with density bounded away from zero on that support. This assumption may not be suitable in semiparametric models: when the distributions of covariates and the unobserved shocks are unknown, the restriction on the covariate support often leads to a loss of point identification of the parameter of interest.

As a result, when we allow the model to be semiparametric with unknown distributions of errors and covariates, we can find parametric submodels that have zero information. It turns out that these submodels can be constructed for each smooth distribution of errors U and V. The structure of the least favorable submodel is such that the shocks U and V are highly correlated eveywhere on the support. The index variables X and Z, on the other hand are highly correlated at the tails. As a result, the parameter of interest is "nearly" not identified. We formally state the result regarding the Fisher information in the following theorem[4]:

Theorem 3.1 *Under Assumption 1, the Fisher information for the parameter α is zero.*

The result of Theorem 3.1 implies that there may not exist an estimator for α that converges at a parametric rate. The convergence rate of the estimators that explore the full support of the index variable Z and that converge in distribution comes from extreme value theory and can be significantly slower than parametric $\sqrt{n}$ convergence rate. If the estimators for the index functions $f(\cdot)$ and $g(\cdot)$ have convergence rates that are compatible with parametric, this implies that estimation error in these index function will be infinitesimal relative to the extreme value component in the estimator for α.

4 Empirical strategy

Equation (2.1) explicitly expresses the treatment effect of interest in terms of the propensity score and the conditional expectation of the outcome that can both be estimated from the data. As we will shortly see when we discuss the details of the data set we use, this is a

[4] The proof of this and all subsequent theorems is provided in the Appendix.

"large dimensional" problem with hundreds of explanatory variables. Consequently, from an implementation perspective, data driven dimension reducing methods will be needed and we use variations of LASSO (Tibshirani (1996)). One can use global and local versions of LASSO to estimate the components of the model. The local version can use pre-partitioning of the data (by dates, geography or other variables that may lead to "structural breaks" in the data). The global version uses the entire sample. Before running LASSO we need to transform all categorical variables intodummy variables. We also perform standard scaling of all continuous variables and interactions (Tibshirani (1996)).

The model is estimated in two steps. In the first step we estimate the propensity score and the conditional expectation of the outcome variable. We do so by, first, running LASSO with the logistic loss function (e.g. see (Negahban, Yu, Wainwright, and Ravikumar 2009)) to perform variable selection for both models. Then we estimate standard logisitc regressions with selected variables to reduce bias arising from regularization.

In the second step we use fitted values of the propensity score and the expected outcome to construct a sample analog of estimator (2.1). To do so, we use threshold δ to select the fitted values of the propensity score sufficiently close to upper and lower bounds of its support. Then we estimate the treatment effect using pairs of observations

$$\widehat{\alpha}_\delta = \frac{1}{n(n-1)} \sum_{i \neq j}^{n} (\widehat{Y}_i - \widehat{Y}_j) D_i (1 - D_j) \mathbf{1}\{|\widehat{P}_i - \widehat{P}_j| < \delta\}, \tag{4.2}$$

where D_i is the observed treatment status for observation i, $\widehat{Y}_i$ is the fitted value of the estimated conditional expectation of the outcome variable, and $\widehat{P}_i$ is the corresponding fitted value of the propensity score.

5 Simulation Study

In the this section we explore the finite sample performances of the newly proposed estimation procedure. We simulate from two designs of the following model:

$$\begin{aligned} d_i &= I[z_i'\delta_0 - \eta_i > 0] \\ y_i &= \alpha_0 d_i + x_i'\beta_0 + \epsilon_i \end{aligned}$$

Where in the simulated model, the variables observed to the econometrician are the scalars d_i, y_i and the vectors (of large dimension) z_i, x_i, whose values are used to estimate the parameter α_0, using the proposed three step procedure. In the above model, δ_0, β_0 are large dimensional vectors of constants unknown to the econometrician but only estimated for the purpose of estimating and conducting inference on the primary parameter of interest, α_0. Note that since we let both x_i, z_i be high dimensional, the linear index specification can be viewed a approximation of unknown functions. So in that sense $E[d_i|z_i]$ is approximated by a nonparametric propensity score $P(z_i)$ and the regression function for the second equation is approximately of a semi linear form $y_i \approx d_i\alpha_0+f(x_i)+\epsilon_i$. In our designs, the high dimensional vectors, x_i, z_i were distributed multivariate normal with varying covariance matrices. The disturbance terms η_i, ϵ_i were distributed bivariate normal, centered around 0 with varying degrees of correlation.

Table 1 reports results from design 1 where we assume to covariates x_i, z_i are mutually independent of each other as well as of the unobserved terms η_i, ϵ_i. Each of the two vectors were assumed to be distributed standard multivariate normal. The disturbance terms η_i, ϵ_i were drawn from bivariate normal distributions with marginal distributions that were standard normal, and we considered varying correlations of the two terms. For generating observed dependent variables we assumed that both β_0 and δ_0 were one hundred dimensional vectors of parameters which took evenly spaced values between -5 and 5. The parameters of interest, α_0 was set to 1.

Implementation of our three step procedure involved using lasso methods twice- once to estimate the propensity score function, by regressing d_i on z_i, and the other to estimate the "reduced form" function by regressing y_i on x_i, z_i. In each case we implemented this using the lasso command on Matlab, and constructed the lasso fit using ten-fold cross-validation.

Recall that our estimator of α_0 involved trimming the data, effectively only using observations where the estimated propensity scores were near the extremes of 0 and 1. Letting $\widehat{P}(z_i)$ denote the first stage estimated propensity score values, the trimming procedure we adopted only uses observations where $\widehat{P}(z_i)$ exceeds (1-δ_n) or is less than δ_n where $\delta_n = 1/n$. α_0 then was then estimated via (4.2).

Table 1 reports mean bias, median bias, RMSE, Median Absolute Deviation (MAD) of this estimator for sample sizes ranging from 500 to 8000, with 10000 replications. To allow for varying degrees of endogeneity of treatment, we let ρ, the correlation between ϵ_i and η_i vary between 0 and 0.75.

The performance of the estimator agrees well with our asymptotic theory. Vital statistics degrease with the sample size but not at the parametric rate. Furthermore, finite sample performance deteriorates the higher the degree of endogeneity. But quite encouragingly, even in the case when $\rho = 0.75$, the estimator performs quite well once we consider sample sizes of 4000 or larger. We consider this encouraging because the data set in our application has many more observations.

Design 2 complicates the design by allowing for correlation among the regressors. Specifically, each pair of regressors were correlated at level proportional to how far the regressors were apart. So, for example the first and second regressor were much more correlated than the first and the one hundredth were. Precisely, we used the function $\exp(-2 \cdot |ii - jj|)$ to denote the pairwise correlation matrix, as ii, jj each went from one to 100, denoting which component of x_i or z_i we were referring to. Results for this design are reported in table 2. We see that finite sample biases are noticeably larger in this designs as is to be expected. But for all levels of endogeneity, acceptable levels are still achieved as the sample size reaches 4000. Again we deem tis more than satisfactory as the sample size in our application is much larger than this.

Finally we consider a design that is motivated by our application. Here we simulate data that is based on the data used in our application. In that sense this part of the section can be referred to as an “empirical monte carlo study”. To generate data for this design we worked with the 36 remaining ”post-lasso” regressors discussed in Section 4. We took summary statistics of these variables from our observational data sample, including means, variances and pairwise covariances of all selected 36 variables. Using these summary statistics, we generate draws of regressor values by drawing from a 36 dimensional multivariate normal distribution with the tabulated mean vector and covariance matrix. Once we have these regressor draws we generated values of treatment indicators and outcome variables using regression coefficients in each equation that were of small order, specifically 10^{-4} which was chosen to reflect the order of magnitude of the indexes if each coefficient were 1. The treatment effect coefficient was set to 1. For the disturbance terms we drew from a bivariate normal distribution with correlation 0.5, mean of 0 and variances of 1.

To implement our estimator we used a larger number of regressors by including all interaction terms and second moments of the the 36 regressors as explanatory variables. Then, to use the three stage estimator, as we just did, we used the lasso command in matlab, again attaining lasso fits using 10-fold cross validation to select the regularization parameter. Trimming was performed as it was for Designs 1,2. Table 3 reports results from 10000 replications using sample sizes of 1,2,4, and 8 thousand observations. The results are for the same summary statistics we considered in tables 1 and 2. As results indicate, our proposed estimation procedure performs quite well in this empirical design. All statistics appear to decline with the sample size though again, and in accordance with the theory, not at the parametric rate.

TABLE 1
Design 1

	$n = 500$	$n = 1000$	$n = 2000$	$n = 4000$	$n = 8000$
$\rho = 0$					
Mean	-0.0200	-0.0442	-0.0111	-0.0221	0.0080
Median	-0.0101	-0.0706	-0.0354	-0.0394	-0.0290
RMSE	0.3311	0.3139	0.2141	0.1629	0.1443
MAD	0.2142	0.2096	0.1452	0.1034	0.0916
$\rho = 0.25$					
Mean	-0.0014	-0.0332	-0.0394	-0.0167	-0.0331
Median	-0.0022	-0.0472	0.0590	-0.0239	-0.0396
RMSE	0.3233	0.2889	0.1981	0.1657	0.1434
MAD	0.2067	0.1855	0.1267	0.1078	0.0947
$\rho = 0.5$					
Mean	-0.2534	-0.0574	-0.0112	-0.0222	-0.0512
Median	-0.2505	-0.0474	-0.0206	-0.0214	-0.0494
RMSE	0.3235	0.1490	0.0787	0.0409	0.0587
MAD	0.2725	0.0983	0.0570	0.0376	0.0494
$\rho = 0.75$					
Mean	-0.0432	-0.0806	-0.0879	-0.0802	-0.0968
Median	-0.0502	-0.0978	-0.0991	-0.0964	-0.1084
RMSE	0.3256	0.3170	0.2059	0.1779	0.1523
MAD	0.2055	0.2297	0.1387	0.1286	0.1191

TABLE 2
Design 2

	$n = 500$	$n = 1000$	$n = 2000$	$n = 4000$	$n = 8000$
$\rho = 0$					
Mean	-0.2654	-0.0246	-0.0097	-0.0051	-0.0007
Median	-0.2861	-0.0394	-0.0306	-0.0246	0.0100
RMSE	0.4001	0.2985	0.2123	0.1723	0.1451
MAD	0.3149	0.2154	0.1452	0.1151	0.0948
$\rho = 0.25$					
Mean	-0.3010	-0.0485	-0.0542	-0.0287	-0.0405
Median	-0.3221	-0.0815	-0.0721	-0.0481	-0.0513
RMSE	0.4065	0.2961	0.2106	0.1634	0.1414
MAD	0.3295	0.2135	0.1469	0.1070	0.0973
$\rho = 0.5$					
Mean	-0.2944	-0.0454	-0.0561	-0.0628	-0.0597
Median	-0.3162	-0.0654	-0.0716	-0.0716	-0.0660
RMSE	0.4122	0.3035	0.1955	0.1654	0.1455
MAD	0.3361	0.2135	0.1352	0.1197	0.1000
$\rho = 0.75$					
Mean	-0.2963	-0.0958	-0.0908	-0.0680	-0.0937
Median	-0.3032	-0.1017	-0.1061	-0.0078	-0.1042
RMSE	0.4218	0.3004	0.2120	0.1707	0.1537
MAD	0.3279	0.2127	0.1480	0.1196	0.1166

TABLE 3
Design 3

	$n = 500$	$n = 1000$	$n = 2000$	$n = 4000$	$n = 8000$
$\rho = 0$					
Mean	-0.0959	-0.0450	-0.0205	-0.0096	-0.0039
Median	-0.0937	-0.0454	-0.0206	-0.0102	-0.0044
RMSE	0.1367	0.0815	0.0491	0.0341	0.0229
MAD	0.0991	0.0568	0.0326	0.0231	0.0147
$\rho = 0.25$					
Mean	-0.4655	-0.4292	-0.4118	-0.4042	-0.4005
Median	-0.4633	-0.4348	-0.4128	-0.4043	-0.4021
RMSE	0.4743	0.4335	0.4146	0.4055	0.4012
MAD	0.4633	0.4348	0.4128	0.4043	0.4021
$\rho = 0.5$					
Mean	-0.8198	-0.8053	-0.8026	-0.7991	-0.7990
Median	-0.8198	-0.8051	-0.8019	-0.8001	-0.7986
RMSE	0.8239	0.8075	0.8037	0.7997	0.7993
MAD	0.8198	0.8051	0.8019	0.8001	0.7986
$\rho = 0.75$					
Mean	-1.1646	-1.1829	-1.1917	-1.1928	-1.1970
Median	-1.1641	-1.1813	-1.1925	-1.1934	-1.1969
RMSE	1.1667	1.1840	1.1922	1.1931	1.1972
MAD	1.1641	1.1813	1.1925	1.1934	1.1969

TABLE 4
Design 4

	$n = 500$	$n = 1000$	$n = 2000$	$n = 4000$	$n = 8000$
$\rho = 0$					
Mean	-0.0973	-0.0473	-0.0193	-0.0074	-0.0013
Median	-0.0995	-0.0497	-0.0171	-0.0095	-0.0007
RMSE	0.1383	0.0847	0.0533	0.0347	0.0243
MAD	0.1035	0.0598	0.0360	0.0235	0.0168
$\rho = 0.25$					
Mean	-0.3516	-0.2977	-0.2656	-0.2440	-0.2305
Median	-0.3503	-0.3035	-0.2645	-0.2445	-0.2311
RMSE	0.3640	0.3059	0.2708	0.2463	0.2318
MAD	0.3503	0.3035	0.2645	0.2445	0.2311
$\rho = 0.5$					
Mean	-0.5882	-0.5378	-0.5083	-0.4873	-0.4612
Median	-0.5909	-0.5396	-0.5078	-0.4882	-0.4605
RMSE	0.5952	0.5419	0.5107	0.4886	0.4619
MAD	0.5909	0.5396	0.5078	0.4882	0.4605
$\rho = 0.75$					
Mean	-0.8217	-0.7878	-0.7506	-0.7188	-0.6870
Median	-0.8229	-0.7883	-0.7502	-0.7183	-0.6872
RMSE	0.8258	0.7902	0.7519	0.7194	0.6874
MAD	0.8229	0.7883	0.7502	0.7183	0.6872

TABLE 5
Design 3

	$n = 1000$	$n = 2000$	$n = 4000$	$n = 8000$
$\rho = 0.5$				
Mean	0.0210	0.0108	-0.0037	-0.0166
Median	0.0265	0.0155	-0.0031	-0.0158
RMSE	0.0486	0.0322	0.0197	0.0211
MAD	0.0305	0.0238	0.0117	0.0160

6 Empirical Context and Results

Our empirical context is “sponsored search.” Search engines return two classes of results, “algorithmic” or “organic” results which are based on webpage relevance to the intent of the search query and sponsored results that are based on relevance and an advertiser submitted bid. Sponsored results, if there are any, appear above the organic results in the most visually prominent portion of the page. Advertisers pay for “consideration,” as measured by clicks, and clicks are thus the central unit of analysis in sponsored search. However, estimating the causal impact of an ad is not as simple as computing the average profit made from a click and comparing this to the “cost per click” paid. The reason is that many advertisers appear in both the sponsored and organic results and past work has shown that some of the clicks on sponsored links would have gone to the organic link in the absence of the sponsored results (Reiley, Li, and Lewis (2010)). When the a search query contains a trademarked brand name, this “cannabilization” of organic clicks for the “focal brand” can be very large, meaning that naive estimates of ad effectiveness vastly overstate the true effect as shown in Blake, Nosko, and Tadelis (2015). Furthermore, advertisers tend to bid higher for geolocations and time periods in which they are more inherently clickable, introducing a form of omitted variable bias.

Given these biases, most practitioners have viewed experiments as the only way to get reliable estimates of ad effectiveness in this setting (Lewis, Rao, and Reiley (2014)). Indeed, the data in our study come from a series of randomized experiments on the Bing search engine. The experiments were conducted on a small fraction of U.S.-located users over nine days in January of 2014 with randomization at the user level. Four experiments took place, in which the maximum number of mainline ads was limited to 0, 1, 2, and 3. Each experiment had a balanced control group, which corresponded to the maximum of 4 mainline ads, the typical production setting. This is standard practice in online experimentation, as it provides a check that each experimental “line” was executed correctly.

The treatment limited the number of ads that could be shown, but often this cap did not bind. For instance, in the treatment group that limited mainline ads to a maximum of 3 (“Cap 3” to employ the terminology we will use throughout), if there were not enough bidders that met the reserve price to fill the 3 slots, then fewer than 3 ads were shown. We carefully control for this issue by selecting only queries that matched into bidding data where an ad would have been shown in the absence of the experiment.

To identify brands, we extracted 87,000 retailer and brand names from the Open Directory Project.[5] A search is characterized as a brand query if and only if (1) the query is on this list, meaning it is a verified firm brand, and (2) the query matches the domain name in the first organic position. We focus only on brands that are in the first organic link because this selects true brand queries. Queries that for brands that are not in the first organic position might be searches of a different nature, perhaps not meant to get directly to the brand page, but to a broader set of sites. Figure 2 provides an example of a brand query. Queries are simplified using standard techniques, e.g. we treat “Macy’s,” “macys.com,” “macys,” and “macy’s” as the same query. We focus on searches with 0 or 1 clicks on the page, ignoring rare instances of 2 or more clicks.[6]

Table 1 gives the number of brands binned by the number of observations for those brands in all control conditions combined. 64.7% of all brands in the control group have less than 10 exposures but represent only 0.19% of all traffic, whereas 96% of traffic comes from the 1045 brands that have 1000 or more exposures. We keep the 2517 companies with over 350 exposures, which cover 98.7% of market activity.[7] Out of the selected 2517 companies, 824 of companies advertise on their own brand keywords more than 90% of the time. In estimating the direct returns to brand search advertising we focus on these 824 brands.

6.1 Experimental Estimation

For each brand j, we observe a number of brand searchers in each experimental condition c, N_{jc}. For each search, among other things we observe the URLs of organic links shown on the page, URLs of paid links shown, and click decisions of consumers. We classify the URLs as belonging to the focal brand if it matches the brand name and belonging to competitors otherwise. We estimate the probability of clicking on a focal brand’s link across experimental conditions using a simple frequency estimator:

$$\hat{Pr}(\text{click } j \text{ in } c) = \frac{1}{N_{jc}} \sum_{i}^{N_{jc}} I(i \text{ clicks } j \text{ in } c)$$

[5] dmoz.org, the project uses volunteer annotators to “classify the web.”

[6] In these occurrences, the searcher often visits all advertisers, making it less interesting to study. Further, search engines often refund clicks from such patterns.

[7] With this selection rule we are balancing the number of firms against the inclusion of brands that don’t provide meaningful information because they are so small. We have done substantial robustness around this threshold and the results are not materially impacted.

Table 1: Ad coverage in the control condition

Number of exposures in Control	Number of brands in Control	Percentage of brands (%)	Percentage of traffic (%)	Percentage of own ads in ML1 (%)	Percentage of competitor's ads in ML1 (%)
1	4869	23.1	0.02	3	30.6
2	2773	13.1	0.02	4.1	32.4
3	1686	8	0.02	6.3	30.8
4 - 10	4315	20.5	0.12	10.2	34.5
11 - 100	4200	19.9	0.64	19.8	34.6
101 - 1000	2202	10.4	3.6	42.64	28.5
> 1000	1045	5	95.6	43.8	13.6
Total	21090	100	100	14.4	31.4

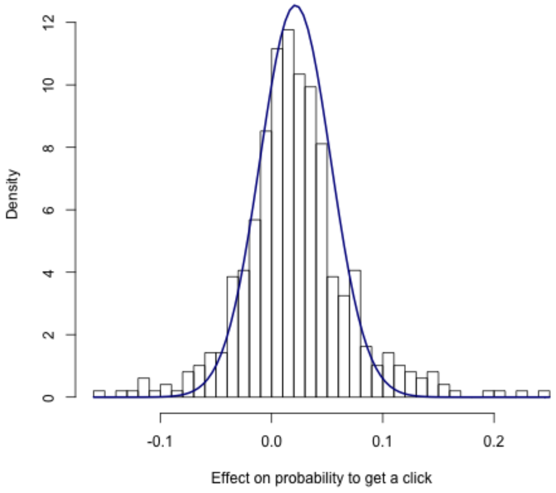
Percentage of ads is computed across companies. For example, companies with 4 exposures and companies with 10 exposures are given the same weight in group 4-10. Total frequency is also computed across companies, unweighted.

where $I(i \text{ clicks } j \text{ in } c)$ follows a Bernoulli distribution. The estimator has expectation of p_{jc} and the variance of $\frac{p_{ic}(1-p_{jc})}{N_{jc}}$, where p_{jc} is the true probability of a click. Similarly, we estimate the probability of clicking on competing firms, j', after searching for brand j in the experimental condition c as $\hat{Pr}(\text{click } j' \text{ in } c)$. We compute $\hat{Pr}(\text{click } j \text{ in } c)$ for each combination of brand j and experimental condition c. Experimental conditions were balanced to compare treatment and control groups. Comparing Cap 0 to Cap 1 allows to isolate the effect of own brand advertising in the absence of competitors in positions 2, 3 and 4. To make sure different conditions can be compared without bias, we ensure that the associated control conditions do not differ from each other. The mean effect is a 2.27 percentage points increase in the probability of visit to the advertising brand's website. That is, the ad does causally increase visits, but the effect is small relative to a baseline visit rate of 78% in the absence of the ad. Figure 5 plots the distribution of firm level ad effects for this case.

6.2 Estimation from observational data

We deployed our empirical strategy outlined in Section 4 to produce non-experimental evaluation of the causal effect of the ad display on clicks. In order to be able to make direct comparison, we produce estimation for the same list of brands that were included in the randomized experiments. However, unlike the experimental evaluation, we include the entire data sample for estimation. To ensure sufficient "effective data size" for each brand, we

Figure 1: The Distribution of Brand-Specific Heterogeneity



Source: Simonov, Nosko and Rao, 2016

chose to focus only on top 350 brands by advertising spent on Bing.com for our estimation and comparison. That generates the sample of 9.3 million observations across all considered brands with over 8.4 thousand raw explanatory variables, their polinomials and interactions up to degree 3 (before variable selection). We provide the description of the variables that we used as well as the sample in Appendix C.1. Using these variables and the observed instances of ad displays, we estimate the first stage model for the probability of ad display as well as the model for the conditional expectation of the number of consumer clicks as a function of transformed observable variables and their interactions using the standard logit-LASSO procedure with the regularization parameter chosen based on cross-validation.

Given that in the second step we use only the observations with “extreme” values of the propensity score, we need to make sure that our sample is sufficiently large to warrant reliable second step inference. On Figure 2 we display the empirical distribution of the fitted probability of ad display for the top 5% quantiles of that distribution and for the bottom 10%. While this figure shows that the empirical density of the probability of ad display approaches zero towards the top and the bottom of its support (i.e. there are virtually no users who are always shown an ad or never shown an ad), the density in the neighborhood of the top and the bottom of the support sharply increases that allows us to use the data for users for whom the fitted probability of ad display is nearly zero or nearly one.

As specified in our empirical strategy, to estimate the causal effect we choose the threshold that determines the neighborhood of the upper and lower bound of the support of the fitted

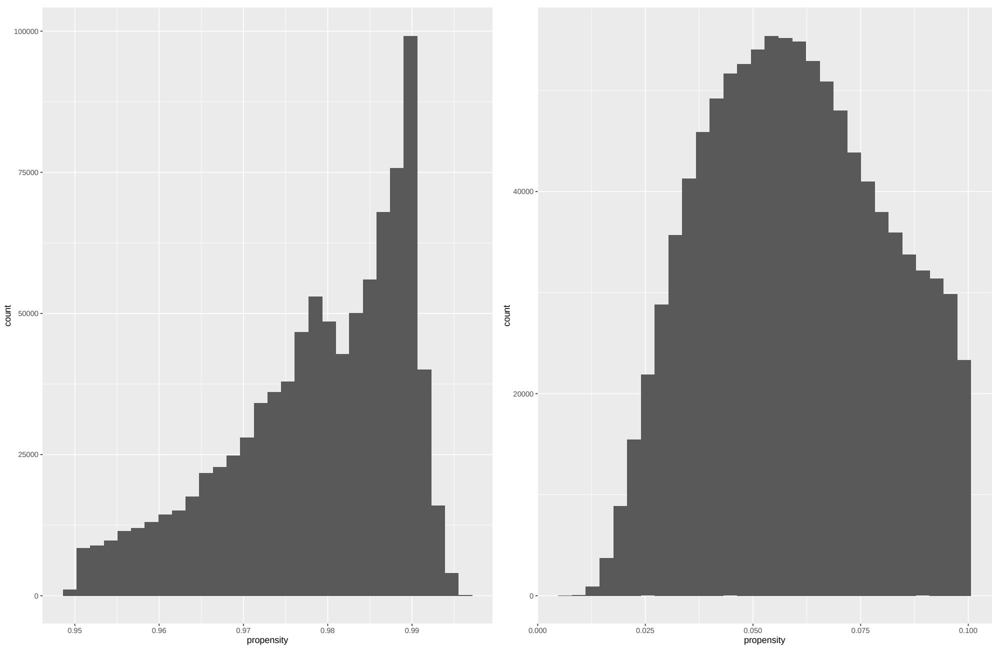


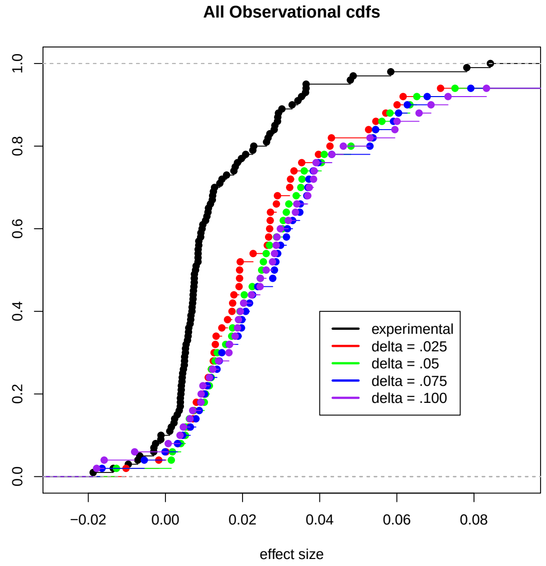
Figure 2: Empirical distribution of the fitted values of the probability of ad display for top and bottom distribution quantiles

probability of an ad display to include the data whose fitted probability is within those two neighborhoods. Our results are presented on Figure 3 for the top 350 brands that were chosen for the randomized experiment. Figure 3 plots the empirical cdf of the causal effect of ad display across the chosen 350 advertisers obtained from the randomized experiment and the cdf of the causal effect obtained using our empirical strategy for different sizes of the neighborhoods of the top and the bottom of the support of the fitted probability of ad display. The figure shows that the inclusion of more observations that are further from the top and the bottom of the support leads to an increased positive bias of the estimated causal effect of the ad display. This is clearly visible from the shift of the cdf to the right as the threshold parameter δ increases.

6.3 Comparison of Results

With results from two distinct approaches to estimating the treatment effects of ad placement on outcome variables, such as, for example, the probability of visiting a brand's website, we can compare findings. If the results found are similar, this can serve to validate our proposed approach to conduct inference that is based on observational data. Few doubt the benefits of controlling for endogeniety by using experimental data- see Lewis, Rao, and Reiley (2014). However, the costs of collecting data from experiments can also be be enormous- see Blake, Nosko, and Tadelis (2015) and Lewis and Rao (2014). Thus if results based on observational

Figure 3: CDFs of the Experimental and Observational Estimates of the Causal Effect of an Ad Displayfor Selected 350 Brands

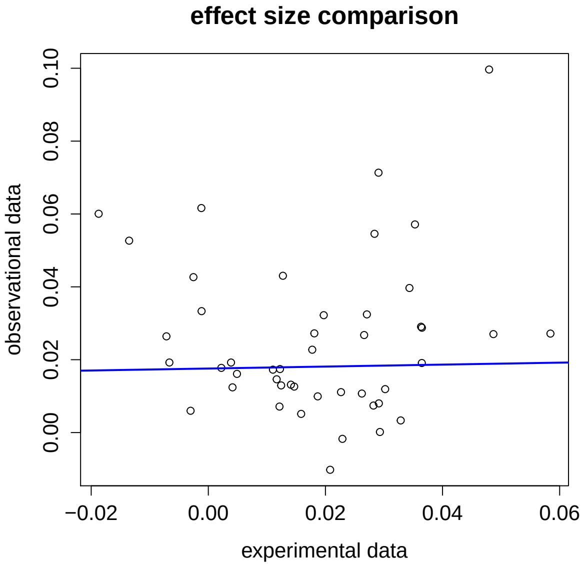


data can be validated, it can be regarded as a cost effective alternative.

Comparisons are illustrated in Figures 3 that we presented before and Figure 4 displayed below.

Figure 3 plots the cdf's of estimated causal effects from both experimental and observational approaches. For observational estimates we plot cdf's for different choices of the trimming parameter determining the size of the neighborhood of upper and lower bounds of the support of probability of an ad display. The median is positive in all approaches, though slightly smaller in the experimental approach. The estimated causal effect is statistically significantly positive in all cases. Figure 4 displays the scatterplot of the estimated causal effect of ad displays obtained from the experimental data against our estimates obtained from the observational data with the smallest chosen threshold parameter ($\delta = .025$). The scatter plot demonstrates an upward sloping pattern, especially in the region where the results attained from experimental data are positive, which was found in most of the brands considered. However, es expected, the observational data lead to relatively small "effective sample" for each brand (given that we only focus on the observations with extreme values of the fitted probability of the ad display). This generates the noise clearly observable on the scatterplot. This effect indicates that further improvement of the accuracy of observational results may require even larger samples.

Figure 4: Scatterplot of Experimental and Observational Estimates of the Causal Effect of an Ad Display for Selected 350 Brands*



* Regression line is displayed in blue

7 Conclusions

In his paper we considered estimation of and inference on the treatment effect of the displays of targeted online advertising on consumer behavior. The endogeneity of treatment was addressed in two ways. One was based on an identification argument for a partially linear model with a binary endogenous variable that is based on using the extreme values of observables for which probabilities of treatment are close to zero or 1. Inference for such a model is nonstandard as observed in Khan and Nekipelov (2017). The other approach to addressing endogeneity was to use experimental data, as practitioners often view randomized controlled experiments as the best way to get reliable estimates of ad effectiveness in this setting. The data in our study came from a series of randomized experiments on the Bing search engine and were conducted on a small fraction of U.S.-located users with randomization at the user level. The finite sample validity of our new estimation and inference procedures was demonstrated by the similarity in findings of the two approaches, and furthermore, by a small scale simulation study using designs that were based on features of the data attained from the experiments. We thus conclude that our new inference procedures can be regarded as a viable alternative method to attaining reliable estimates of treatment effects, especially when data sets are sufficiently large as in our empirical example. This should be a welcome alternative to working with experimental data when the costs of runnings such experiments is very large, such as examples detailed in recent work by Blake, Nosko, and Tadelis (2015) and Lewis and Rao (2014).

References

Bickel, P., C. Klaassen, Y. Ritov, and J. Wellner (1993): *Efficient and adaptive estimation for semiparametric models.* Johns Hopkins University Press Baltimore.

Blake, T., C. Nosko, and S. Tadelis (2015): "Consumer Heterogeneity and Paid Search Effectivness: A Large Scale Field Experiment," *Econometrica*, 83(1), 155–174.

Blundell, R., and J. Powell (2004): "Endogeneity in Binary Response Models," *Review of Economic Studies*, 73.

Cattaneo, M., M. Jansson, and W. Newey (2016): "Alternative Asymptotics and the Partially Linear Model with Many Regressors," *Econometric Theory*, 1, 1–25.

Chamberlain, G. (1986): "Asymptotic efficiency in semi-parametric models with censoring," *Journal of Econometrics*, 32(2), 189–218.

Ibragimov, I., and R. Has'minskii (1981): *Statistical Estimation Asymptotic Theory.* Springer-Verlag.

Khan, S., and D. Nekipelov (2010): "Information Bounds and Impossibility Theorems for Simultaneous Discrete Response Models," Duke University Working Paper.

——— (2016): "On Uniform Inference in Nonlinear Models with Discrete Endogenous Variables," Working Paper.

——— (2017): "Information Structure and Statistical Information in Discrete Response Models," *Quantitative Economics*, forthcoming.

Lewbel, A. (2000): "Semiparametric Qualitative Response Model Estimation with Unknown Heteroscedasticity or Instrumental Variables," *Journal of Econometrics*, 97, 145–177.

Lewis, R. A., and J. M. Rao (2014): "The Unfavorable Economics of Measuring the Returns to Advertising," *Quarterly Journal of Economics*, 130, 1941–1973.

Lewis, R. A., J. M. Rao, and D. H. Reiley (2014): "Measuring the Effects of Advertising: The Digital Frontier," in *The Economics of Digitization*, ed. by A. Goldfarb, S. Greenstein, and C. Tucker. NBER Press.

Negahban, S., B. Yu, M. J. Wainwright, and P. K. Ravikumar (2009): “A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers,” in *Advances in Neural Information Processing Systems*, pp. 1348–1356.

Newey, W., and S. Donald (2014): “Series Estimation of Semilinear Models,” *Journal of Multivariate Analysis*, 130, 1941–1973.

Reiley, D. H., S.-M. Li, and R. A. Lewis (2010): “Northern exposure: A field experiment measuring externalities between search advertisements,” in *Proceedings of the 11th ACM conference on Electronic commerce*, pp. 297–304. ACM.

Robinson, P. (1988): “Root-n-Consistent Semiparametric Regression,” *Econometrica*, 56, 931–954.

Tibshirani, R. (1996): “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

A Proof of Theorem 3.1

To derive the information of the model we follow the approach in Chamberlain (1986) by demonstrating that for each triangular model generated by a distribution satisfying the conditions of Theorem 3.1 we can construct a parametric submodel passing through that model for which the information for the parameter α is equal to zero. Suppose that Γ contains all distributions of errors that satisfy the conditions of Theorem 3.1 along with all distributions of indices X and Z for which $E[q(X,Z)^2] < \infty$ for $q(\cdot,\cdot)$ defined in the statement of the theorem such that X and Z have a continuous joint distribution with a full support for Z.

We now construct the likelihood function. Let $\varphi_{xz}(u)$ be conditional density of $U \,|\, V \leq z, X = x, Z = z$ and $\mathcal{P}_{xz}(v)$ be the conditional cdf of $V \,|\, X = x, Z = z$. We use notation $P = (\varphi_{xz}, \mathcal{P}_{xz})$ for the entire set of nuissance parameters in the model.

The conditional likelihood function of the model can be then written as

$$r(y,d,x,z;\alpha,P) = \varphi_{xz}(y-\alpha-x)^d \mathcal{P}_{xz}^d(z)\varphi_{xz}(y-x)^{1-d}(1-\mathcal{P}_{xz}(z))^{1-d}.$$

Consider measurable continuously differentiable function $\eta_{\cdot,\cdot}(\cdot)$ such that for each x and z $\int \varphi_{xz}(u)\eta'_{xz}(u)\,du = 0$, $\int \eta^2_{xz}(u)\,du < \infty$ and $\int (\eta'_{xz}(u))^2\,du < \infty$. We use this function to construct local parametrization of the original model. We denote the likelihood function corresponding to the perturbed model $l_\lambda(y,d,x,z;\alpha,\delta)$. We define $\tilde{\Lambda}$ as the collection of paths through the original model such that for each path $\lambda \in \tilde{\Lambda}$ corresponding to a specific choice of $\eta_{xz}(\cdot)$ and parameter δ

$$\begin{aligned} l_\lambda(y,d,x,z;\alpha,P) = {} & \varphi_{xz}\left(y-\alpha-x+\delta(\eta_{xz}(y-\alpha-x)-1)\right)^d \mathcal{P}_{xz}^d(z) \\ & \times \varphi_{xz}(y-x)^{1-d}(1-\mathcal{P}_{xz}(z))^{1-d}. \end{aligned}$$

where we note that for sufficiently small δ these paths maintain the properties of the joint probability distribution.

Provided the assumed dominance condition, it will be mean-square differentiable at $(\alpha, 0)$. In other words, we can find functions $\psi_\alpha(x,z)$ and $\psi_\delta(x,z)$ such that $l_\lambda^{1/2}(\cdot;\alpha,\delta) = \psi_\alpha(x,z)(a-\alpha) + \psi_\delta(x,z)\delta + R_{a,\delta}$, with $E\left[R^2_{a,\delta}\right]/\left(|a-\alpha|+|\delta|\right)^2 \to 0$ as $a \to \alpha$, $\delta \to 0$. We can explicitly derive the mean-square derivatives. In particular, the derivative with respect to the finite-dimensional parameter can be expressed as

$$\psi_\alpha(x,z) = -\tfrac{1}{2}d\varphi_{xz}(y-\alpha-x)^{-1/2}\mathcal{P}_{xz}(z)^{1/2}\varphi'_{xz}(y-\alpha-x),$$

and the derivative with respect to λ can be expressed as

$$\psi_\delta(x,z) = \tfrac{1}{2}d\varphi_{xz}(y-\alpha-x)^{-1/2}\mathcal{P}_{xz}(z)^{1/2}\varphi'_{xz}(y-\alpha-x)(\eta_{xz}(y-\alpha-x)-1).$$

Suppose that μ is the Lebesgue measure over y, x and z. Then we use the fact that the Fisher information can be bounded as

$$I_{\lambda,\alpha} \leq 4 \int (\psi_\alpha - \psi_\lambda)^2 \, d\mu \leq \int \tfrac{\mathcal{P}_{xz}(z)}{\varphi_{xz}(y-\alpha-x)} \left(\varphi'_{xz}(y-\alpha-x)\right)^2 \eta^2_{xz}(y-\alpha-x) \, d\mu(y,x,z)$$

We can define the measure on Borel sets in $\mathbb{R}^3$ as

$$\pi(A) = \int_A \tfrac{\mathcal{P}_{xz}(z)}{\varphi_{xz}(u)} \left(\varphi'_{xz}(u)\right)^2 \, d\mu(u+\alpha+x, x, z),$$

Following Chamberlain (1986), we let $L_2(\pi)$ denote the space of measurable functions $q : R^3 \to R$ such $\int q^2 d\pi < \infty$, allowing us to conclude that $I_{\lambda,\alpha} \leq \|\eta_{xz}\|^2_{L_2(\pi)}$.

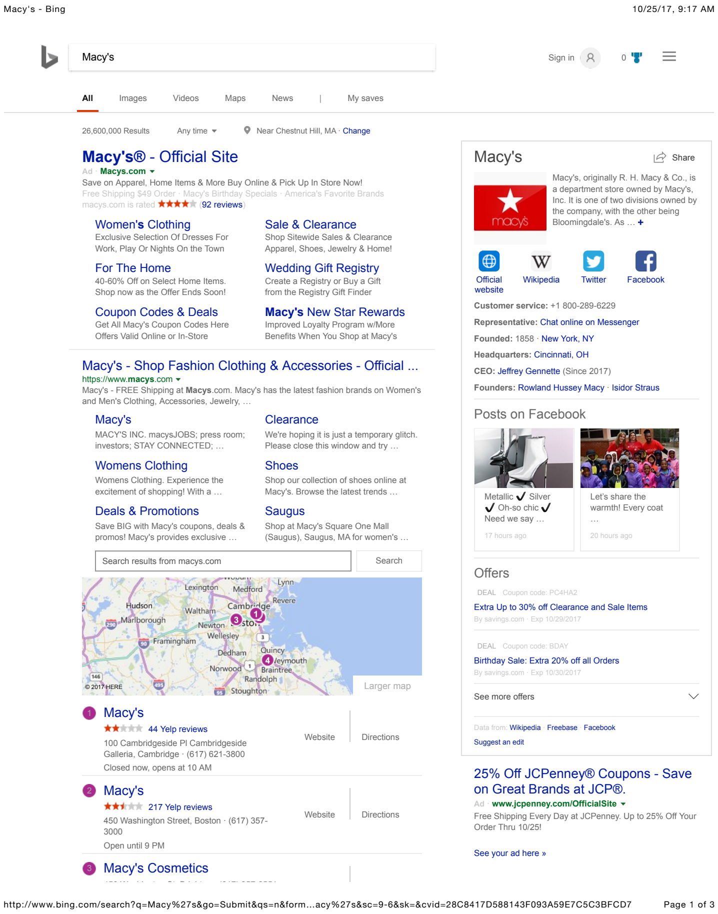
Chamberlain (1986) demonstrates that the space of differentiable functions with compact support is dense in $L_2(\pi)$. Moreover, we require the derivative of h to be continuous in the interior of its support. Let S be the support of h. We take $\epsilon^* > 0$ and construct the set S_{ϵ^*} to be a compact subset of S such that the Euclidean distance of the boundary of S from the boundary of S_{ϵ^*} is at least ϵ^*, where ϵ^* is selected such that $\pi(S \setminus S_{\epsilon^*}) < \sqrt{\epsilon}$. Since the set of differentiable functions is dense in $L_2(\pi)$, for any $\epsilon > 0$ we can find $a_{xz} \in C^2_c(\mathbb{R}^3)$ (where $C^2_c(\mathbb{R}^3)$ denotes the set of real-valued functions on $\mathbb{R}^3$ that have compact support and continuous partial derivatives of order 2) such that $\|a_{xz}\|_{L_2(\pi)} < \sqrt{\epsilon}$. The derivative $a'_{xz}(\cdot)$ is continuous in the interior of S. Provided that $S_{\epsilon^*} \subset S$, this derivative is continuous on the entire set S_{ϵ^*} and, due to its compactness it is uniformly continuous there. As a result, there exists $M = \sup\limits_{S_{\epsilon^*}} |a'_{xz}(u)|$. There also exists $M' = \sup\limits_{S} |a_{xz}|$. Then we pick the direction η^*_{xz} as function with support on S such that $\eta^*_{xz} = B(a_{xz}/M)$ in S_{ϵ^*}. Then we note that

$$\|\eta^*_{xz}\|_{L_2(\pi)} \leq \frac{B}{M}\|a_{xz}\|_{L_2(\pi)} + \frac{BM'}{M}\|\mathbf{1}_{S\setminus S_{\epsilon^*}}\|_{L_2(\pi)} < \frac{B(M'+1)}{M}\sqrt{\epsilon}.$$

As a result, $I_{\lambda,\alpha} \leq \frac{B^2(M'+1)^2}{M^2}\epsilon$. As the choice of ϵ was arbitrary, this proves that $\inf\limits_{\lambda \in \tilde{\Lambda}} I_{\lambda,\alpha} = 0$.

B Results

Figure 5: Example of Brand Query



Macy's

26,600,000 Results · Any time · Near Chestnut Hill, MA · Change

Macy's® - Official Site
Ad · Macys.com
Save on Apparel, Home Items & More Buy Online & Pick Up In Store Now!
Free Shipping $49 Order · Macy's Birthday Specials · America's Favorite Brands
macys.com is rated ★★★★☆ (92 reviews)

Women's Clothing
Exclusive Selection Of Dresses For Work, Play Or Nights On the Town

Sale & Clearance
Shop Sitewide Sales & Clearance Apparel, Shoes, Jewelry & Home!

For The Home
40-60% Off on Select Home Items. Shop now as the Offer Ends Soon!

Wedding Gift Registry
Create a Registry or Buy a Gift from the Registry Gift Finder

Coupon Codes & Deals
Get All Macy's Coupon Codes Here Offers Valid Online or In-Store

Macy's New Star Rewards
Improved Loyalty Program w/More Benefits When You Shop at Macy's

Macy's - Shop Fashion Clothing & Accessories - Official ...
https://www.**macys**.com
Macy's - FREE Shipping at **Macys**.com. Macy's has the latest fashion brands on Women's and Men's Clothing, Accessories, Jewelry, …

Macy's
MACY'S INC. macysJOBS; press room; investors; STAY CONNECTED; …

Clearance
We're hoping it is just a temporary glitch. Please close this window and try …

Womens Clothing
Womens Clothing. Experience the excitement of shopping! With a …

Shoes
Shop our collection of shoes online at Macy's. Browse the latest trends …

Deals & Promotions
Save BIG with Macy's coupons, deals & promos! Macy's provides exclusive …

Saugus
Shop at Macy's Square One Mall (Saugus), Saugus, MA for women's …

Search results from macys.com

1 Macy's
★★☆☆☆ 44 Yelp reviews
100 Cambridgeside Pl Cambridgeside Galleria, Cambridge · (617) 621-3800
Closed now, opens at 10 AM

2 Macy's
★★½☆☆ 217 Yelp reviews
450 Washington Street, Boston · (617) 357-3000
Open until 9 PM

3 Macy's Cosmetics

Macy's

Macy's, originally R. H. Macy & Co., is a department store owned by Macy's, Inc. It is one of two divisions owned by the company, with the other being Bloomingdale's. As … +

Customer service: +1 800-289-6229
Representative: Chat online on Messenger
Founded: 1858 · New York, NY
Headquarters: Cincinnati, OH
CEO: Jeffrey Gennette (Since 2017)
Founders: Rowland Hussey Macy · Isidor Straus

Posts on Facebook

Metallic ✓ Silver ✓ Oh-so chic ✓ Need we say …
17 hours ago

Let's share the warmth! Every coat …
20 hours ago

Offers

DEAL Coupon code: PC4HA2
Extra Up to 30% off Clearance and Sale Items
By savings.com · Exp 10/29/2017

DEAL Coupon code: BDAY
Birthday Sale: Extra 20% off all Orders
By savings.com · Exp 10/30/2017

See more offers

Data from: Wikipedia · Freebase · Facebook
Suggest an edit

25% Off JCPenney® Coupons - Save on Great Brands at JCP®.
Ad · www.jcpenney.com/OfficialSite
Free Shipping Every Day at JCPenney. Up to 25% Off Your Order Thru 10/25!

See your ad here »

http://www.bing.com/search?q=Macy%27s&go=Submit&qs=n&form...acy%27s&sc=9-6&sk=&cvid=28C8417D588143F093A59E7C5C3BFCD7 Page 1 of 3

C Summary statics

C.1 Summary statistics of raw variables

Below we report raw values of the variables that were used to form explanatory variables for the models of the propensity score and the model of the conditional expectation of the outcome variable. Our entire sample contains 9,318,608 observations. To generate variables for LASSO we used interactions and polynomials of variables up to degree 3. That creates an overall of 8,436 variables that were used in the LASSO step. The "Request time" variable is collected but not reported to avoid revealing the exact timing of the experiment.

Feature	Average	Max	Min	First Quartile	Median	Third Quartile	St Dev
User_SessionSequenceNumber	1.748	34.0	0.0	1.0	1.0	3.0	1.657
User_PageViewSequenceNumber	72.601		0.0	10.0	34.0	88.0	109.886
User_IsWindowsLiveUser	0.22	1.0	0.0	0.0	0.0	0.0	0.414
User_IsNew	0.028	1.0	0.0	0.0	0.0	0.0	0.165
User_IsFacebookUser	0.227	1.0	0.0	0.0	0.0	0.0	0.419
User_Duration	24167	86398	0	5842	22124	37856	19772
User_ClickCount	19.413	4359.0	0.0	4.0	10.0	23.0	32.524
Session_PageViewSequenceNumber	28.917	1412.0	0.0	0.0	7.0	32.0	58.629
Session_PageViewCount	61.412	1473.0	1.0	8.0	27.0	74.0	95.023
Session_Duration	3051	14399	0	356	1720	4484	3507
Request_HasRMSCookie	0.211	1.0	0.0	0.0	0.0	0.0	0.408
Query_ResultsAdultScore	0.003	1.0	-1.0	0.0	0.0	0.0	0.044
Query_IsSpellSuggestionCorrection	0.062	1.0	0.0	0.0	0.0	0.0	0.24
Query_IsQueryAlteration	0.376	1.0	0.0	0.0	0.0	1.0	0.484
Query_IsAutoSuggest	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Query_AdultScore	-2.429	100.0	0.0	0.0	0.0	0.0	155.273
Page_PageNumber	1.042	86.0	1.0	1.0	1.0	1.0	0.573
Page_DistinctQueryCountVerticalChangeWithinVisit	0.794	1.0	0.0	1.0	1.0	1.0	0.404
Metrics_WebIndexItemCount	24.274	201.0	1.0	18.0	22.0	29.0	8.874
Metrics_WebIndexClickCount	0.598	34.0	0.0	0.0	1.0	1.0	0.707
Metrics_TotalTimeToFirstClick	14.104	9316.055	-0.594	1.782	4.491	10.95	93.14
Metrics_RightRailAdItemCount	1.285	11.0	0.0	0.0	0.0	2.0	1.896
Metrics_RelatedItemCount	11.833	33.0	0.0	8.0	14.0	16.0	5.918
Metrics_RelatedClickCount	0.036	16.0	0.0	0.0	0.0	0.0	0.204
Metrics_QueryViewCountWithRightRailAdPresent	0.437	1.0	0.0	0.0	0.0	1.0	0.496
Metrics_QueryViewCountWithResultsSuccessClicks	0.479	1.0	0.0	0.0	0.0	1.0	0.5
Metrics_QueryViewCountWithQuickBackClicks	0.189	1.0	0.0	0.0	0.0	0.0	0.391
Metrics_QueryViewCountWithDCardPresent	0.337	1.0	0.0	0.0	0.0	1.0	0.473
Metrics_QueryViewCountWithCoreTopAdPresent	0.469	1.0	0.0	0.0	0.0	1.0	0.499
Metrics_QueryDwellTime	37	16036	0	2	6	18	174
Metrics_PaginationClickCount	0.021	8.0	0.0	0.0	0.0	0.0	0.145
Metrics_Overall_PLT	1184.4	59994.0	-2.0	750.0	910.0	1175.0	1538.2
Metrics_DominantResultWebIndexItemCount	3.07	19.0	0.0	0.0	0.0	9.0	4.4
Metrics_DCardAnswerItemCount	0.752	44.0	0.0	0.0	0.0	0.0	2.849
Metrics_CoreTopAdItemCount	0.469	1.0	0.0	0.0	0.0	1.0	0.499
Metrics_Browser_TimeToPageLoadComplete	777	119557	0	442	619	803	1392
Metrics_AnswerItemCount	31.9	386.0	0.0	20.0	30.0	42.0	17.3

C.2 Summary of the LASSO step

Our online appendix lists our generated variables and shows which variables were selected at the LASSO step for the propensity score equation and for the outcome equation. The demonstrated results are provided for our entire dataset that contains an overall of 350 top brands (by search volume) and a subset of the top 50 brands.

For the entire dataset LASSO selected 2,166 variables for the propensity score and 2,114 variables for the outcome equation (out of 8,436). For the subset of top 50 brands LASSO selected 1,907 variables for the propensity score equation and 1,954 for the outcome equation.