Experiments as Instruments: Understanding Consumer Behavior in Sponsored Search*

Matthew Goldman†
UC San Diego

Justin M. Rao
Microsoft Research

Abstract

In this paper we show how one can re-purpose business experimentation to estimate causal effects unrelated to the intent of the original experiment. Using a large dataset from a major search engine, we apply this technique to examine the impact of ad position in sponsored search on click and purchasing patterns. Our findings reveal substantial heterogeneity in click position effects across advertisers. On average, moving up the page lowers conversion rate. Moving up the page also has a uniformly positive impact on click-through-rate, but the impact is far stronger for relatively medium-to-high quality advertisers. Both patterns are consistent with a satisficing model of search. The large heterogeneity in position impact has important consequences for the impact of the auction and poses new mechanism design challenges.

*We would like to thank the Microsoft Corporation for allowing us to publish this work and Susan Athey, Dean Foster and Sebsatien Lahaie for helpful discussion and comments.

†Much of this work was done when Goldman was an intern at Microsoft Research.

1 Introduction

Active experimentation is becoming increasingly common in business, especially online. Firms use field experiments to measure advertising effectiveness (Lewis and Reiley,), improve ranking and recommendation (Radlinski et al., 2008; Craswell et al., 2008) and optimize user interfaces (Kohavi et al., 2007; Kohavi et al., 2012). Experimentation has been evolved beyond simply a measurement tool to become an integral component of market design (Ostrovsky and Schwarz, 2011; Manzi, 2012).

A typical experiment measures the impact of a new policy as compared to the current best practice. For instance, a new auction mechanism might be used on a random subset of search traffic and a comparison of treatment and control search-instances reveals the impact on allocation and revenue. A company that relies heavily on such experiments can tally hundreds or thousands per year. In this paper, we show how a pool of past experiments can drive new academic and business insights unrelated to the intent of the original experiments by using each experiment as an instrument in an Instrumental Variable (IV) framework. Contrary to the classical IV case where one finds a few solid instruments through domain expertise, here we have a vast supply of ambiently available instruments, which affords causal inference freed of endogeneity that usually plagues historical data analysis.

We apply the "experiments as instruments" method in the largest, to the best of our knowledge, empirical study of the sponsored search marketplace. Sponsored search links—the paid advertisements on a search engine results page—are the monetization channel for search engines. Google and Bing/Yahoo![1] use a generalized second price auction (GSP) in which slots—up to four above the algorithmic search results—are allocated and priced based an advertiser's bid, quality score and relevance to the query (Edelman et al., 2007; Varian, 2007). This critical, and unusual, feature of the GSP means that ranking and pricing depend on parameters estimated by the auctioneer. Most important among them, perhaps, are "position effects," which capture differences in click-through-rates (CTR) based on *position alone*. They thus determine the differentiation between the slots—the goods being sold—and consequently fundamentally impact the degree to which one should shade their bid below their valuation, or put another way, how aggressively they should bid to get a higher slot. Moreover, since an advertiser's bid is weighted by the listing's "clickability," estimating the impact of position alone is necessary to normalize observed CTR.

A related determinant of return on investment is the conversion effect curve. If higher positions have a greater probability of sale conditional on click, then they are naturally more valuable to the advertiser. In the opposite case, the top slot would be very undesirable and equilibrium bids could be greatly depressed by this disincentive. All theoretical models of the GSP which we are aware of assume a multiplicatively separable impact of position on clicks (Edelman et al., 2007; Varian, 2007) and constant conversion effects across slots (Lahaie, 2006; Athey and Ellison, 2011). Yet

[1] Microsoft AdCenter operates the auction for both Yahoo! and Bing.

recent papers investigating the role of position on conversion rate, both working with a single firm, have equivocal findings. Agarwal et al. (2011) run a field experiment where the advertiser's bid is randomized in order to generate exogenous variation in rank. The authors find that conversion rate declined significantly moving towards the top of the page. Ghose and Yang (2009) use a structural approach and reach precisely the opposite conclusion. Recent work by (Narayanan and Kalyanam,) used a regression discontinuity design to show that impacts on conversion were statistically insignificant–unlike naive OLS estimators which revealed a strong positive impact of moving up the page. Further complicating matters is that the theories of search and the psychology of choice can produce arguments for both patterns (Lana, 1963; Brunel and Nelson, 2003; Joachims et al., 2005; Kempe and Mahdian, 2008), as does work on equilibrium allocation under different models of consumer search (Jerath et al., 2011; Athey and Ellison, 2011).

We use past experimentation in the ranking, relevance and click prediction algorithm by a major search engine to identify the role of the slot position on conversion and click rate. Our analysis includes hundreds of firms bidding on the top hundred queries revenue-wise. Using the exogenous variation generated by experiments is necessary because conversion rate and clickability are typically endogenous to position. For example, a firm running a sale might bid higher (or be rated as higher quality by the relevance algorithm), moving them up the page and generating a spurious correlation of position with conversion rate. Personalized ranking of ads (Agichtein et al., 2006; Dupret and Piwowarski, 2008) further exacerbates this issue.

We find strong evidence of heterogeneity in both click and conversion curves. We'll discuss the conversion results first. Since we are primarily interested signing the impact of moving up the page, we use a random coefficient linear IV model to maximize power. In aggregate, moving up the page causally reduces the probability of a conversion conditional on a click, but as evidence of the substantial heterogeneity, some advertisers benefit from a higher position. This likely explains the equivocal findings in the literature, since past work each used a single firm. Rational models of search in which users use position as a signal of quality would predict that moving up the page increases conversion rate, but we do not observe this in aggregate. Instead our results are consistent with behavioral models of search—such as satisficing—provided the threshold to buy exceeds the threshold to click.

For click curves, the specific shape fundamentally impacts the auction, so we fit a richer specification than the linear IV model. Since there are four ad slots, a fully flexible model would allow each position to have a different impact for each advertiser. This presents a methodological challenge, because the implicit requirement that our instrument set span R^4—four dimensional variation—is rarely satisfied because experiments often treat an advertiser more favorably (uniformly up the page) or less favorably (uniformly down). We solve this problem by adding some structure—but still far short of a fully linear model—to make estimation feasible for a large set of advertisers.

The IV click curves are significantly flatter click curves on average than those obtained through

OLS. Hausman-Wu tests confirm that these differences are indeed highly significant—strong evidence of the endogeneity of position. The IV estimates also reveal that position treatment effects seems to peak for ads with a baseline CTR of roughly 50% and to decay linearly in each direction. This means the highest quality advertisers don't benefit much from position—they get the clicks anyway. Similarly advertisers of low quality don't benefit much either, even if they move up the page users skip over them for more attractive alternatives. The advertisers that benefit the most are relatively high quality but not overwhelmingly relevant to the query. This is precisely the prediction from a satisficing model of search (Simon, 1967).

The heterogeneity across advertisers makes it difficult to establish general behavioral findings using experimentation at a single firm. Differing click position effects across ads has important implications for pricing and allocation in this market. Since moving up the page has a lower causal impact on clicks for low and very high quality advertisers if we use a uniform position effect curve—as implied by theoretical models—-we'll overestimate clickability for medium-high quality advertisers by attributing some of the position effect to the advertiser's quality. Due to the mechanics of the GSP, these advertisers will get higher positions at lower cost-per-click on average. It also raises mechanism design problem due to efficiency issues. Incorporating advertiser-specific position effects would enhance efficiency (on the margin higher slots given to those who benefit more from position), but would add substantial complexity to the GSP, in particular breaking the current model of pricing.

.

2 Data

Our data covers 4 months of keyword auctions for some of the highest revenue queries on two major search engines operated by Microsoft AdCenter. From this we draw 540 unique query-advertisement combinations that were served on at least 1,000 impressions[2]. For each query, we have the participating bidders, quality and relevance parameters attached to each bidder, the ranking and pricing parameters chosen by the auctioneer, serving logs, click logs and conversion tracking (for a subset of advertisers). Each query is assigned to an experimental traffic bucket using a process which we explain in detail below.

A search engine results page or "SERP" is divided into three key areas: "north" or "mainline" sponsored listings (ads), algorithmic results (the information retrieved for the query by the search engine) and "east" or "sidebar" ads. Mainline ads, if shown, are located directly above the algorithmic results and constitute the vast majority of revenue for the search engine. During the time period of study, the major U.S. search engines all capped the number of mainline ads at four per query. The number of ads shown (0-4) is determined by a reserve price set by the auctioneer

[2] A larger data pull is in the works.

(and other proprietary factors). For high value queries, it is very likely that 4 advertisers clear the reserve price. Queries with four ads thus tend to have relatively high cost-per-click and accordingly constitute a large share of publisher revenue. For this reason and to ensure maximal variation in position, we restrict our study to queries which consistently had four mainline ads. All other ads displayed on the impressions of these queries are grouped into the catch-all category of "sidebar".

Table 1: Average User Behavior by Ad Position

	ML 1	ML 2	ML 3	ML 4	Sidebar
Click Through Rate (Clicks per Impression)	1.000	0.423	0.358	0.279	0.127
Conversion Rate (Conversions per Click)	1.000	0.480	0.338	0.241	0.375
Average Dwell Time (Sec per Impression)	100.000	69.605	71.609	74.415	74.303

† Each row normalized separately to obscure scale.
Data has more than 1 million impressions and more than 100,000 clicks.

Regardless of position effects, these ads vary substantially in their innate clickability. This is demonstrated in the dispersion of the histograms in Figure 1. The x-scale is redacted, but is unchanged across the four panels.

2.1 Auction rules and experimentation

The generalized second price auction works as follows. For a given search query, each ad listing i is assigned a "rank score" s_i based on a quality score q_i, relevance score r_i, bid b_i and "squashing parameter" α:

$$s_i = f(r_i) q_i^{\alpha} b_i \tag{1}$$

where q_i is generally taken to be an advertiser's innate clickability. If α is set to 0, the auction is rank-by-bid. If α is set to 1, the ranking is effectively done by advertiser value; in general the revenue optimal α will lie somewhere in between (Lahaie, 2006; Athey and Nekipelov, 2010).

Ads are ranked by s_i and the top N are shown depending on reserve prices and auctioneer policies. In this paper, we study queries where four ads are consistently shown above the algorithmic search results. Define $S_i(b_i|\alpha, r_i)$ as rank score as a function of bid given the parameters of the action. The advertiser in slot k pays the lowest bid required to maintain the position given paramaters: $S_k > S_{k+1}$, or in terms of bid the advertiser pays a cost-per-click (CPC) of

$$c_k = S_k^{-1}(s_{k+1}).$$

The cost-per-click (CPC) for slot k is given by c_k. Let $p_{i,k}$ give the probability of click for advertiser

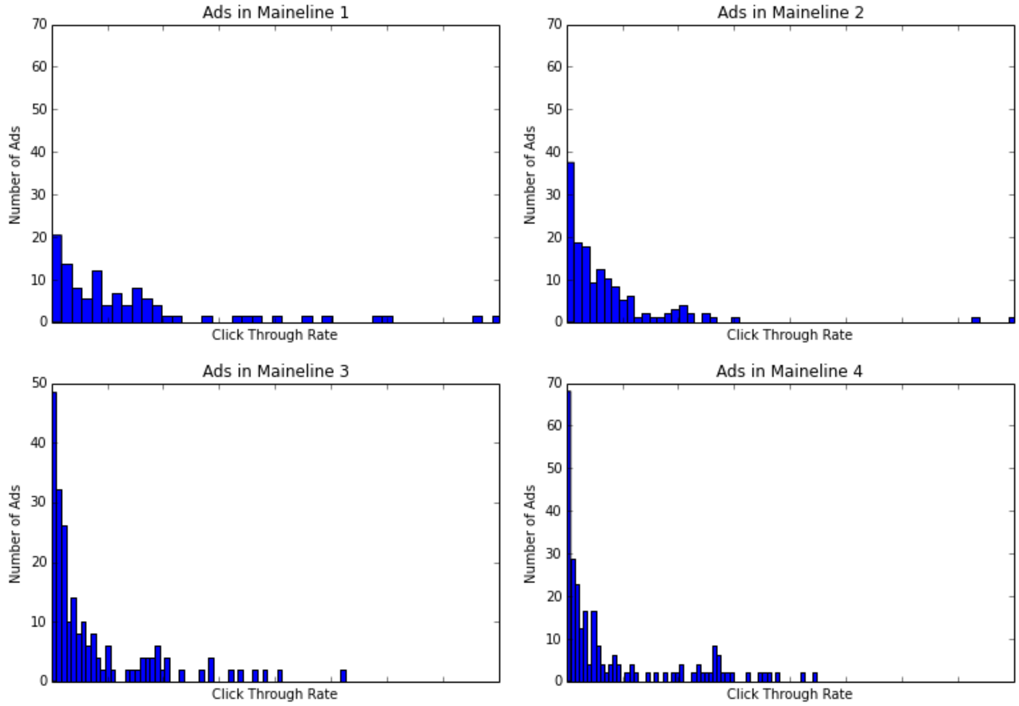


Figure 1: Average CTR at the advertisement level. All ads have atleast 400 impressions in each position for which they are displayed.

i in slot k. Expected revenue is thus:

$$\sum_{k=1}^{4} c_k * p_{i,k} \tag{2}$$

Typically $p_{i,k}$ is decomposed into a multiplicatively separable slot effect and the quality score indicated by q_i (Edelman et al., 2007), however we will not impose that assumption.

2.2 Experimentation

As is evident from (2), search engine revenue is impacted by the parameters of the rank score function (Lahaie, 2006; Athey and Ellison, 2011) thus motivating constant experimentation with new click prediction, relevance, and ranking algorithms. At any given time, hundreds of experiments are being run, each of which alters one or more parameters that govern ad ranking or display. Impressions are randomized into these experiments at either the user or impression level. Simply refreshing a Bing search several times will result in re-randomization into a different Bing experiment and may alter the displayed advertisements. Importantly, advertisers cannot bid differently by experimental condition and the probabilistic weights on each experiment are fixed at the monthly level. However, the probabilistic weights on each experiment are constantly altered as new experiments are introduced (see the right panel of Figure 2).

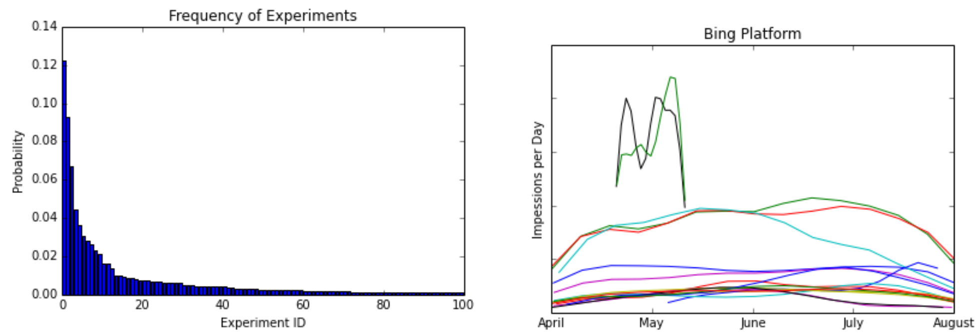


Figure 2: Histogram of experimental frequencies of the top 100 experiments on the Bing platform (left) and the time-varying frequency with which users were randomized into to the top 20 Bing experiments (right).

new experiments are introduced and certain queries are withheld from certain advertisements. Thus experimental assignment is exogenous to a host of potential advertiser related confounds.

These experiments have a wide range of goals, from improving user experience through more relevant ads, improving CTRs, and maximizing ad revenue. Although the experiments were designed for myriad purposes, many have the effect of shuffling the ranking of ads. This is intuitive, the ranking of ads is the primary impact the auctioneer has on the marketplace so most experiments impact it in some way. The most common such experiment is to tune the value of α in equation (1), thus altering the relative weight placed on ad quality and price in the ranking algorithm. Other experiments may impact ad ranking by re-parameterizing relevance estimation, changing minimum relevance thresholds for inclusion in the auction, or protecting a dominant advertiser against competitors.

We will seek to use these experiments as IVs for causal inference in the next section. However, we must offer the caveat that some experiments directly manipulate the size or text of ad display; potentially altering consumer behavior directly. Such experiments produce endogenous IVs in our causal model and must be carefully pruned based on axillary data or carefully structured over-identification tests. Our current approach is use the subset of advertisements which are located in a fixed position (usually ML1) throughout the entirety of our sample period. All valid experiments should have identical click through rates *as a function of time* throughout the entirety of the sample. For example,

2.3 Measures of User Engagement

We use two measures as proxies for the value of a click to an advertiser: conversions and dwell time. Conversions are an advertiser-selected measure of user engagement after a click on a sponsored link.

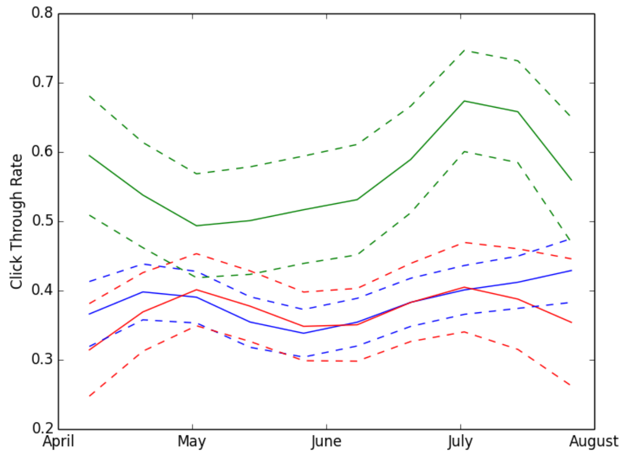


Figure 3: Local constant estimation of the average click through rate over time for the three most popular experiments on the Bing platform for one particular advertisement. The advertisement was exclusively listed in the first mainline position in all three experimental conditions. The green line appears represents an experiment in which the presentation of the bing ads is skewed to make the top ad more attractive - perhaps by the addition of site links.

For some advertisers a conversion may indicate only an expensive purchase, whereas for others it may indicate merely completing a free registration or signing up for an email list. As such, baseline conversion rates can vary dramatically across advertisers. Conversions are only tracked for a subset of advertisers and provide binary information about user engagement.

Dwell times are continuous and are provided for all advertisers. They track the amount of time between a user's initial click on a sponsored link and his eventual return to the bing search engine. In the even that a user never returns, the dwell time is labeled as missing. Figure 4 provides histograms summarizing the values of these variables. The left panel shows considerable heterogeneity in conversion rate at the advertiser level. The right panel shows the strongly skew (roughly log normal) distribution of observed dwell times.

We will use both conversions and larger dwell times as positive indicators of a valuable click. Obviously advertisers do not care directly about dwell time, but larger (or missing) dwell times seem like plausible indicators of an engaged consumer who was more likely to make a purchase. As demonstrated in the local linear regressions of Figure 5, all of the five most converting advertisers saw a positive relationship between dwell time and conversion probability.

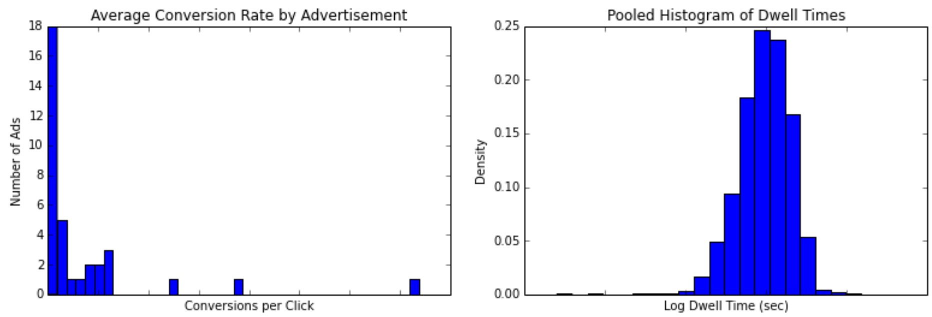


Figure 4: Histogram of logged dwell times (left panel) and a non-parametric smoothing of the relation between logged well time and conversion probability (right panel).

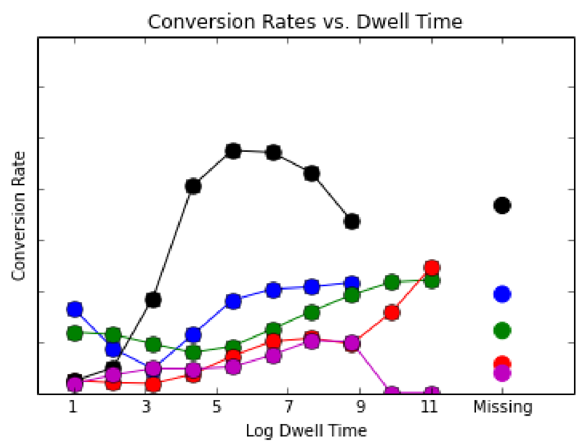


Figure 5: Histogram of logged dwell times (left panel) and a non-parametric smoothing of the relation between logged well time and conversion probability (right panel).

3 Estimation: Conversion and Dwell Time Effects of Ad Position

All analysis is performed at the level of a fixed query (e.g. "cheap cruises") and a fixed advertisement i which could be displayed in any of four mainline (or a sidebar) positions. We pose a random coefficient model of the probability that a given click on advertisement i leads to a conversion on search t. In order to maximize power, we consider a simple linear model (rather than allowing for a non-linear "conversion curve") of the impact of ad position on conversion probability. Different advertisers may have different standards for a conversion leading to different baseline conversion rates. Changes in advertiser popularity over time may also impact a given advertiser's baseline conversion rate. To control for these difficulties, all analysis is specific to a given advertisement and contains fixed effects at the month level. Our primary estimation equation is

$$P(\text{Conversion}_{i,t}|\text{Click}_{i,t}) = \alpha_i(t) + \beta_i(4 - \text{AdPos}_{i,t}) + U_{i,t}. \tag{3}$$

Or in the case of dwell times, we use the similar

$$E(\log[\text{Dwell}_{i,t}]|\text{Click}_{i,t}) = \alpha_i(t) + \beta_i(4 - \text{AdPos}_{i,t}) + U_{i,t}. \tag{4}$$

The interpretations of (3) and (4) are parallel. $\alpha_{i,\text{Month}_t}$ gives baseline for advertisement i in a given month and $\beta_{i,t}$ may be interpreted as the average change in conversion probability (or log dwell time) that occurs when the advertisement is moved up the page (AdPos is decreased). $U_{i,t}$ is a mean-zero shock representing the effects of time-varying, unobservable characteristics of advertiser i which are not captured by the month effects. These may include the beginning (or end) of a special promotion or a rise (or fall) in popularity. Recall that ad position is mechanically determined by an ads popularity (pClick) and the advertisers willingness to pay for a click (bid). Both of these seem likely to be positively correlated with $U_{i,t}$ and are thus likely channels for endogeneity. For this reason, we expect that OLS estimates of equation (3) will provide upward biased estimates of the *causal* impact of ad position on conversion rates.

Instead, we exploit exogenous variation generated by Bing's random experimentation (discussed in Section 2) in an instrumental variables framework. Let Z_t denote the designation of the experiment that search t is randomized into. We know that Z_t is randomized by design, with the randomization weights fixed at a monthly level. Thus we can be certain that

Assumption 1. Exogenous Instruments:

$$U_{i,t} \perp Z_t|\text{Month}_t$$

and IV estimates of equation (3) will be consistent.

Our IV estimates are formed by first constructing an instrument matrix of dummy variables for each possible experimental designation. There are over 400 active ranking experiments running at any given time. To avoid many instrument problems, we exclude data corresponding to experiments which yield less than 50 clicks on advertisement i[3].

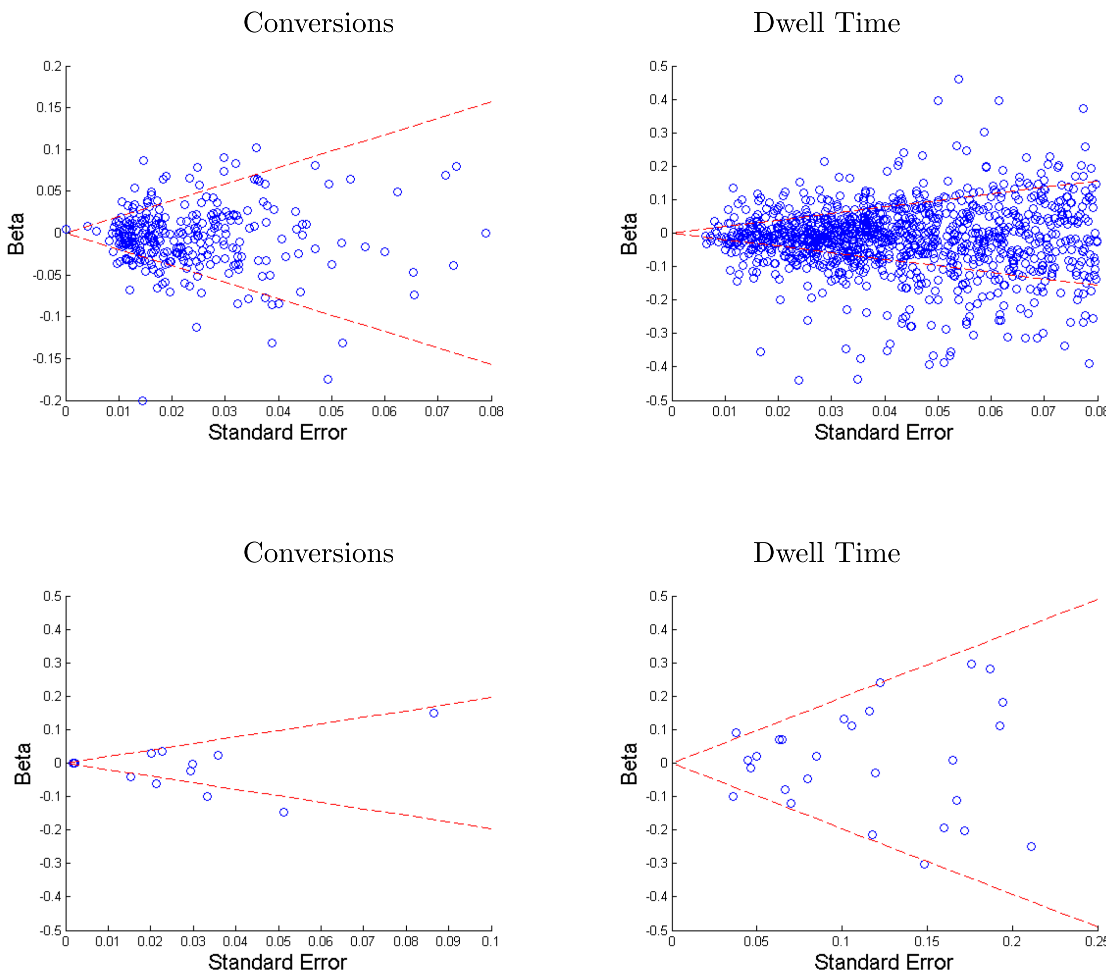
3.1 Main conversion rate results

Conversions are tracked for only a subset of advertisers who opt in with the search engine. Figure **??** demonstrates heterogeneous conversion effects for the opting in advertisers. Here the conversion effect is of moving one spot up the page without regard for which slot an advertiser is currently in.

[3]This always results in less than 10 columns in our instrument matrix.

Our IV model finds a statistically significant increase in conversion probability for some advertisers as we move up the page.

An additional measure the correlates strongly with conversion propensity that is available for all advertisers is dwell time on the advertiser's website following and ad click.



4 Estimation: Click Curves

Estimation of click curves poses additional challenges. Here interest attaches not just to the average impact of moving an ad up one slot, but rather to the exact shape of the click curve. Similar to the random coefficient specification we used in equation (3), we write

$$P(\text{Click}_{i,t}) = \sum_{s=1}^{4} \alpha_s \cdot 1\{\text{Ad Pos}_{i,t} = s\} + U_t \tag{5}$$

$$= \alpha_4 + \sum_{s=1}^{3} \beta_s \cdot 1\{\text{Ad Pos}_t \leq s\} + U_t, \tag{6}$$

where U_t is now a time varying, mean-zero shock to *click*-ability. Equations (5) and (6) define the same model in terms of different parameters. α_s can be understood as the average clickability when the ad is placed in position s and β_s is the number of marginal clicks generated by moving an ad from position $s+1$ up to s. Proportional increases in clickability are also of interest and are given by

$$\gamma_s = \frac{\beta_s}{\alpha_{s-1}}.$$

Once again, identification by the typical OLS moment conditions is very dubious and we will appeal to exogenous Microsoft experimentation for identification. However, now the problem is more difficult because we are seeking to estimate a richer model with four endogenous parameters instead of just two. As such, we need atleast four unique values of our instrument (order condition) and we need them to induce four-dimensional variation in the conditional distribution of ad position (rank condition).

4.1 Graphical depiction

The additional requirements on the data of modeling the entire click curve are illustrated in Figure 6. Each point represents an experimental condition or "flight" for a given advertiser on a given query. At the origin, all the advertisers impressions are at third position (ML3). This occurs for Flight 1 (which can be thought of as the control). Moving to the left increases the share of impressions at position 2, moving the right increases share at position 1. In Panel 1 we can see the Flight 2 is very favorable for this advertiser, pushing him to the top slot from ML3 half the time, whereas Flight 3 puts him position 2 half the time. As drawn, the flights provide variation to span the space necessary to infer causal effects of all three parameters (note that ML4 ads have been omitted from this illustration). In Panel 2 we have drawn a circumstance where this condition does not hold. In this case the three flights only span R^1, meaning we cannot separately identify the impact of moving from ML3 to ML2 from moving up the page 1 slot generally. Note that if we assumed linearity (as in Section 3), we would be fully identified.

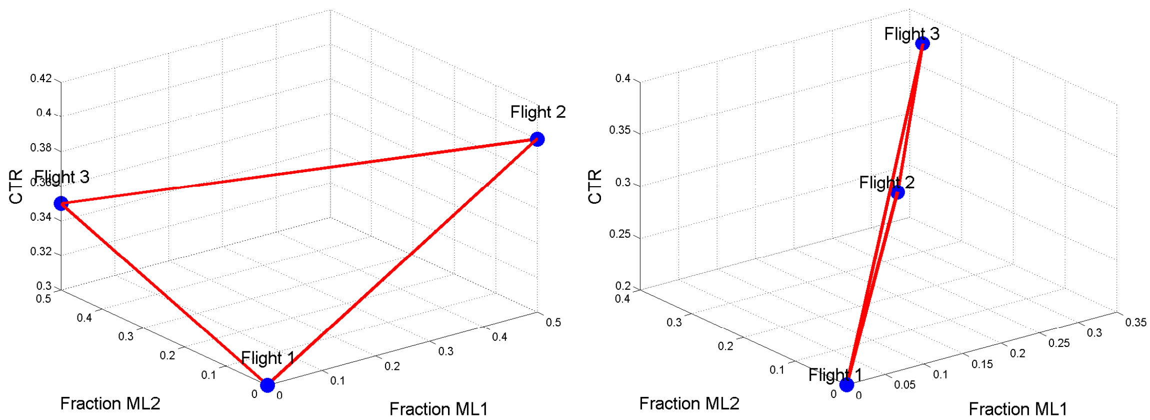


Figure 6: Panel 1 shows a situation where an instrument set spans R^2 for a given advertiser, allowing for identification. In Panel B only R^1 is effectively spanned because flights 2 and 3 push the advertiser up the page uniformly.

4.2 Partial Identification (technical and under development)

If we assume only the exogeneity of our instruments, then the full IV moments are the efficient (and only) basis of estimation. For the vast majority of advertisements in our sample, this poses too big a challenge to the data as our instruments fail multi-dimensional relevance tests. IV estimates then have very large variance and significant finite-sample bias.

Estimation based only on traditional IV moments tends to provide little information about our parameters. Instead we impose a number of additional assumptions.

Assumption 2. *Monotone Treatment Response:* $\beta_1, \beta_2, \beta_3 \geq 0$.

Assumption 3. *Monotone Selection into Treatment:*

$$(j < k) \implies E(U|\text{AdPos} = j, Z = z_1) \geq E(U|\text{AdPos} = k, Z = z_2).$$

Assumption 4. *Commensurate Selection Strength Across Instruments:*

$$\Big[P(\text{AdPos} > s|Z = z_1) \geq P(\text{AdPos} > s|Z = z_2)\Big] \implies \Big[E(U|\text{AdPos} > s, Z = z_1) \leq E(U|\text{AdPos} > s, Z = z_2)\Big].$$

Assumption 2 merely states that moving up the page as a weakly positive impact on click-ability. Assumption 3 assumes that advertisements are more likely to be higher on the page when their time-varying click-ability is higher. Conditional on the instrument value, ad position is mediated by an ads pClick and bid. Since pClick is mechanically correlated with ad click-ability, this assumption requires only that advertisers tend to increase their bid when their click-ability is high. The final assumption requires that the degree of selection is not dramatically different across experiments.

For example, I could compare data from two experiments and find that a given advertisement is in ML1 10% of the time in experiment 1 and only 5% of the time in experiment. Assumption 3 would tell me that the ML1 bucket of experiment 2 contains better residuals on average.

We can use these three assumptions to generate additional information about the parameters of the click curve. These can be framed as estimation models yielding bounds on parameters or as moment inequalities. Below we introduce four estimation models and their corresponding moment inequalities. At the end a theorem collects these models and signs the biases associated with naive estimation of each one.

For purposes of notation let R be a square matrix representing the distribution of Z and P denote the $(J+1)\times 4$ matrix giving the distribution of position conditional on the experiment. That is,

$$R \equiv \text{diag}\{P(Z=j)\}_{j=0}^{J} \tag{7}$$

$$P \equiv \begin{bmatrix} P(\text{Pos}=1|Z=0) & P(\text{Pos}\leq 2|Z=0) & P(\text{Pos}\leq 3|Z=0) & P(\text{Pos}\leq 4|Z=0) \\ P(\text{Pos}=1|Z=1) & P(\text{Pos}\leq 2|Z=1) & P(\text{Pos}\leq 3|Z=1) & P(\text{Pos}\leq 4|Z=1) \\ . & & & \\ . & & & \\ . & & & \\ P(\text{Pos}=1|Z=J) & P(\text{Pos}\leq 2|Z=J) & P(\text{Pos}\leq 3|Z=J) & P(\text{Pos}\leq 4|Z=J) \end{bmatrix}$$

$$\equiv \begin{bmatrix} P_1 & P_2 & P_3 & P_4 \end{bmatrix}.$$

Condition(s, W)**.** For a positive definite $(J+1)\times(J+1)$ matrix, W, and $s \in \{1,2,3,4\}$, a given advertisement's selection into position is said to satisfy this condition if for all $r \in \{1,2,3,4\}\setminus s$,

$$\frac{P_s' W P_r}{P_s' W P_s} \geq \frac{\mathbf{1}_{J+1}' W P_r}{\mathbf{1}_{J+1}' W P_s}.$$

Intuitively, this condition says that (mediated by some positive definite W) there is a positive correlation between conditional (on possible instrument values) CDFs of Position at values at s and r. If we choose weighing matrix $W = I_{J+1}$, then the condition reduces to the requirement of a positive correlation between P_s and P_r. If the conditional distributions of position can be ranked by stochastic dominance (instrument values that induce more time in Position$=1$ also induce more time in Position≥ 2), then this condition will hold for any positive definite W. For any given utilized weighting matrix, one can compute the sample analog to this constraint and see if it is satisfied.

4.2.1 Full IV Model

Assumption 1 is the standard exogeneity of insturmental variables and gives the $(J+1)$-vector of moment equalities,

$$E[Z(Y-\alpha-\sum_{s=1}^{3}\beta_s X_s)] = E[ZU] = \mathbf{0}. \tag{8}$$

We refer to the equations in 8 as *full IV moments*. Without further restrictions, identification of the entire model requires that $\text{rank}(P) \geq 4$ which may fail even if we have access to $J \geq 3$ different experiments[4]. Further if any of the eigenvalues of $E(X'P_Z X)$ are close to zero, then the IV estimator will be biased and preform poorly in finite samples (Stock and Yogo, 2002)[5].

4.2.2 OLS Model

Naive estimation of an OLS model in (5) yields the four moments ($j \in \{1,2,3,4\}$),

$$E(1\{\text{Pos}=j\}(Y-\alpha_j)) = E(U|\text{Pos=j}),$$

each of which is of ambiguous sign. But by appeal to Assumptions 1 and 3 we have 3 unconditional moments. For $j \in \{1,2,3\}$,

$$E\Big(1\{\text{Pos}=j\}(Y-\alpha_j) - 1\{\text{Pos}=j+1\}(Y-\alpha_{j+1})\Big) \geq 0, \tag{9}$$

or 3(J+1) conditional moments,

$$E\Big(1\{\text{Pos}=j\}(Y-\alpha_j) - 1\{\text{Pos}=j+1\}(Y-\alpha_{j+1})\Big|Z=z\Big) \geq 0.$$

4.2.3 Upper IV Model

We consider IV estimation of α_{s+1}, β_s in the *misspecified* model

$$\begin{aligned} P(Y=1) &= \alpha_{s+1} + \beta_s 1\{Pos \leq s\} + V^U \\ &\equiv X \begin{bmatrix} \alpha_{s+1} \\ \beta_s \end{bmatrix} + V^U, \end{aligned}$$

[4]See Figure 6 for an example of the failed identification of a two-dimensional model by two non-control experiments.

[5]The statistic in Cragg and Donald (1993) provides a convenient multi-dimensional analog to the commonly cited "rule of ten" for F-statistics with a single endogenous variable.

where the treatment effects in β_{-s} are left unmodeled leaving us with residual

$$V^U \equiv \sum_{r=1}^{s-1} \beta_r 1\{\text{Pos} \le r\} - \sum_{r=s+1}^{3} \beta_r 1\{\text{Pos} > s\} + U.$$

Intuitively this is a correct model if there are no treatment effects beyond the one of interest (moving for $s+1$ to s). The model is more dramatically misspecified if β_{-s} contains large elements. Define the naive moments for this model as

$$\begin{aligned} Q_j^s &\equiv E\left[1\{Z=j\}(Y - \alpha_{s+1} - \beta_s 1\{\text{Pos} \le s\})\right] \\ &= E\left[1\{Z=j\}V^U\right] \\ &= P(Z=j)\left[\sum_{r=1}^{s-1} \beta_r P(\text{Pos} \le r | Z=j) - \sum_{r=s+1}^{3} \beta_r P(\text{Pos} > r | Z=j)\right] \end{aligned}$$

and the stacked column vector is given by

$$Q^s = R\left[\sum_{r=1}^{s-1} \beta_r P_r - \sum_{r=s+1}^{3} \beta_r (1 - P_r)\right]$$

If we set $s = 1$ (or $s = 3$) then by Assumption 2 the error in our moment conditions is all negative (positive) and we may write the moment inequalities,

$$Q^1 \le \mathbf{0}, \qquad Q^3 \ge \mathbf{0}. \tag{10}$$

In the event that the elements of β_{-s} are all zero, these moment conditions could be replaced by equalities. For the case of $s = 2$, the moment conditions can not be signed and it is not obvious how to write down a useful moment inequality. But in Theorem 1(ii), we demonstrate how this model can still be applied to partially identify of β_2 and α_3.

4.2.4 Lower IV Model

Now consider the exact same model but estimated only on data from advertisements that fall in slots s or $s+1$. Inference in this model is no longer confounded by unmodeled treatment effects but instead by the selection of a biased sample. This model is more dramatically misspecified if there are strong forces for selection and not misspecified at all allocation of position is also exogenous.

Formally, we have

$$\begin{aligned}P(Y=1) &= \alpha_{s+1} + \beta_s 1\{Pos \leq s\} + V^L \\ &\equiv X_t \begin{bmatrix}\alpha_{s+1} \\ \beta_s\end{bmatrix} + V^L.\end{aligned} \tag{11}$$

By restricting the sample such that $\text{Pos} \in \{s, s+1\}$ we leave no treatment effects unmodeled but observe a biased sample of residuals with moments

$$E(V^L|Z) \equiv E(U|Z, \text{Pos} \in \{s, s+1\})$$

Define the naive moments for this model as

$$\begin{aligned}Q_j^s &\equiv E\left[1\{Z=j, \text{Pos} \in \{s, s+1\}\}(Y - \alpha_{s+1} - \beta_s 1\{\text{Pos} \leq s\})\right] \\ &= E\left[1\{Z=j, \text{Pos} \in \{s, s+1\}\} V^L\right] \\ &= P(Z=j)\left[\sum_{r=1}^{s-1} P(\text{Pos} \leq r|Z=j) \cdot E(U|\text{Pos} \leq r, Z=j)\right. \\ &\quad \left. - \sum_{r=s+1}^{3} P(\text{Pos} > r|Z=j) \cdot E(U|\text{Pos} > r, Z=j)\right]\end{aligned}$$

and Q^s as the stacked column vector. For $s=1$ (or $s=3$) there is only selection on one side of the distribution of U and we have the inequalities

$$Q^1 \leq \mathbf{0}, \qquad Q^3 \geq \mathbf{0}. \tag{12}$$

Once again, it is less obvious how to write down moments for the case of $s=2$, but Theorem 1(iii) shows how information from this model can still be used to partially identify our parameter of interest.

Results on all four of these estimation procedures are collected in Theorem 1 and proved in the appendix.

Theorem 1. *Given the data generating process in* (6) *and the estimation routines described in subsections 4.2.1-4.2.4, we have the following*

i. If Assumption 1 is true and $\text{rank}(P) = 4$, *then estimation based on a full IV model as in 4.2.1 is consistent for all parameters.*

ii. If Assumption 3 is true then OLS estimates of $\beta_1, \beta_2, \beta_3$ *are biased up and OLS estimates of* α_4 *are biased down.*

iii. If Assumptions 1 and 2 are true and Condition(s, W) holds then Upper IV model (with moment weighting matrix W) of 4.2.3 yields an upward biased estimate of β_s and a downward biased estimate of α_s.

iv. If Assumptions 1, 3, and 4 are true and Condition(s, W) holds then Lower IV model (with moment weighting matrix W) of 4.2.3 yields an downward biased estimate of β_s and an upward biased estimate of α_s.

4.3 Main click-through-rate results

Combining these moment-inequalities will yield efficient inference on the click (conversion curve). We ultimately plan to follow the grid search methodology, similar to as outlined in (Rosen, 2008). For now we present estimates based on the OLS, upper IV, and lower IV models for selected advertisements. Recall that we have proved under our stated conditions that the OLS and upper IV estimates should provide upper bounds on the true causal effect while the lower IV estimates should provide a lower bound.

Our results are presented graphically below in Figures 7-8. They demonstrate:

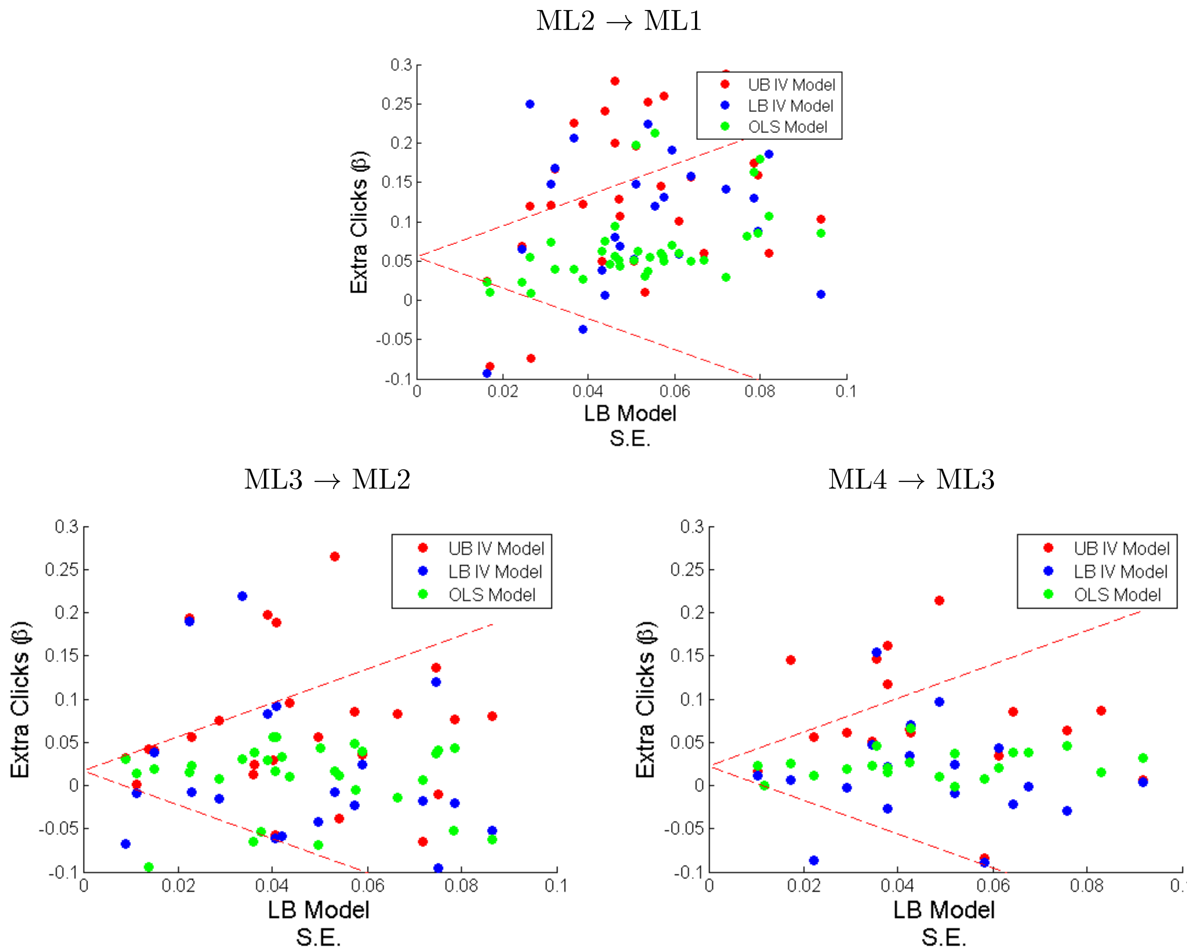
- Substantial heterogeneity in click curves.
- Flatter click curves for IV than OLS estimates.
- Steeper click curves for Ads with baseline CTR closest to 50%.

5 Discussion

5.1 Models of search

Eye-tracking evidence indicates that most people start at the top of webpage and scroll down (Granka et al., 2004; Joachims et al., 2005). This pattern, combined with a incomplete search or a threshold rule of some sort is the most logical explanation for increased CTR as an ad moves up the page. By studying conversions we address a related question: as you get more clicks, what happens to the probability of sale conditional on a click. If, as we observe in aggregage, it drops, it means these marginal clicks are of lower value.

The signalling model of search (in which occupying the top slot is a costly signal of quality) predicts the precise opposite. Of course for a few reasons this does not mean signalling is not going on. First, we see significant heterogeneity, so signalling might benefit some advertisers, such as smaller players. Second, signalling could still be going on and without it the effect would be even more negative in aggregate. However, we can conclude that empirically the *net* effect in the market is not consistent with the signalling model.



	ML2 → ML1			ML3 → ML2			ML4 → ML3		
	Mean	Median	Var-Weighted Mean	Mean	Median	Var-Weighted Mean	Mean	Median	Var-Weighted Mean
OLS	0.166 (0.0013)	0.125 (0.0050)	0.156 (0.0005)	0.0565 (0.0011)	0.0523 (0.0022)	0.0529 (0.0005)	0.0464 (0.0011)	0.0377 (0.0021)	0.0427 (0.0005)
Lower IV	0.136 (0.0077)	0.146 (0.0273)	0.12 (0.0057)	0.0741 (0.0058)	0.0581 (0.0113)	0.0463 (0.0042)	0.0319 (0.0070)	0.0351 (0.0178)	0.0126 (0.0049)
Upper IV	0.197 (0.0076)	0.175 (0.0243)	0.183 (0.0051)	0.0863 (0.0053)	0.0765 (0.0103)	0.0565 (0.0039)	0.0423 (0.0061)	0.0441 (0.0135)	0.0257 (0.0044)

Figure 7: Estimates of the impact of moving up one slot: ML2 to ML1 (Top), ML3 to ML2 (Left) and ML4 to ML3 (Right) by Advertiser. Each advertiser receives a colored dot corresponding to each of our three estimation models. OLS models tend to have much lower standard errors, but for purposes of comparisons, estimates from all three dots for a given advertiser are shown at the advertisers standard error on the lower IV model.

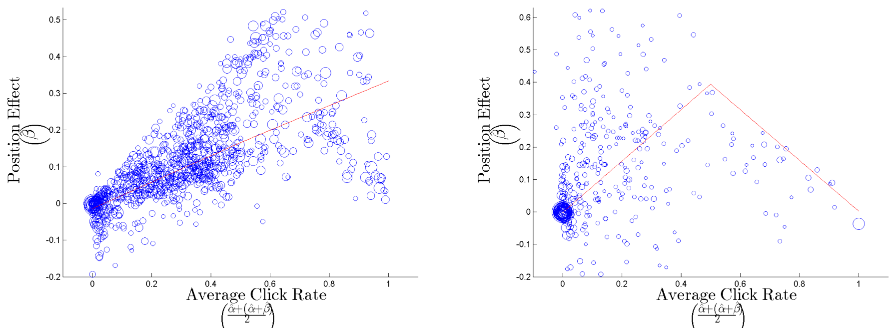


Figure 8: Position effect of moving from the second slot (ML2) to the first slot (ML1) by advertiser. x-axis gives the average CTR over ML1 and ML2. In Panel 1 we show OLS estimates, which are biased. Panel 2 gives lower IV estimates which uncovers a diminishing return of position for dominant ads.

Satisficing models of search in which a users have heterogeneous quality thresholds to click which is positively related to purchase probability would predict the behavior we observe. To see how this works, imagine a user with low threshold for clicking. The hypothesized association says this user is less likely to buy conditional on click than someone with a high threshold. Perhaps this is because she is the type with low search costs in the second stage or because purchasing thresholds are uniformly high across the population. Since low threshold types are more likely to click a higher slot and, conversely, high threshold types are more likely to search down the page to find the right fit, conversion rate drops as one moves up the page (but total conversions still go up).

5.2 Heterogeneity, efficiency and mechanism design

If advertisers have different benefits from higher slots, a mechanism that ignores this will naturally sacrifice efficiency. In the extreme, suppose there two advertisers, one that benefits mightily from being in the top slot, the other who gets a constant fraction of clicks wherever the ad appears. Since the current GSP ignores this possibility, we would often wastefully allocate the top position.

There are two hurdles to incorporating heterogeneity of this sort into the mechanism. The first is an empirical challenge. But of course this precisely what the methodology developed in this paper was designed to address. The second is a theoretical one. All current equilibrium analysis we are aware relies both on multiplicatively separable impact of position (advertiser quality times position effect) and a constant position effects across advertisers. Even if we are not concerned with equilibrium analysis, there are philosophical questions surrounding pricing, given that historically an advertiser is compensated for their clickability, which has been viewed as “quality”.

6 Concluding Remarks

In this paper we develop and apply an econometric method for re-purposing platform-level business experiments to the estimation of firm-level structural parameters. Borrowing from the econometric literature on instrumental variables and moment inequalities we are able to test many of the underlying assumptions of the auction literature on sponsored search. Our results demonstrate the importance of firm heterogeneity and the endogeneity of advertiser position and thus the profitability of this approach to generate unbiased causal inference from historical "big data."

References

Agarwal, A., Hosanagar, K., and Smith, M. (2011). Location, location, location: An analysis of profitability of position in online advertising markets. *Journal of Marketing Research*.

Agichtein, E., Brill, E., and Dumais, S. (2006). Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–26. ACM.

Athey, S. and Ellison, G. (2011). Position auctions with consumer search. *The Quarterly Journal of Economics*, 126(3):1213–1270.

Athey, S. and Nekipelov, D. (2010). A structural model of sponsored search advertising auctions. In *Sixth Ad Auctions Workshop*.

Brunel, F. F. and Nelson, M. R. (2003). Message order effects and gender differences in advertising persuasion. *Journal of Advertising Research*, 43(03):330–341.

Brynjolfsson, E., Dick, A. A., and Smith, M. D. (2010). A nearly perfect market?: Differentiation vs. price in consumer choice. *Quantitative Marketing and Economics*, 8(1):1–33.

Chakravarti, A., Janiszewski, C., and Ülkümen, G. (2006). The neglect of prescreening information. *Journal of Marketing Research*, pages 642–653.

Cragg, J. G. and Donald, S. G. (1993). Testing identifiability and specification in instrumental variable models. *Econometric Theory*, 9:222–222.

Craswell, N., Zoeter, O., Taylor, M., and Ramsey, B. (2008). An experimental comparison of click position-bias models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 87–94. ACM.

Dupret, G. E. and Piwowarski, B. (2008). A user browsing model to predict search engine click data from past observations. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 331–338. ACM.

Edelman, B., Ostrovsky, M., and Schwarz, M. (2007). Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American Economic Review*, 97(1):242–259.

Ghose, A. and Yang, S. (2009). An empirical analysis of search engine advertising: Sponsored search in electronic markets. *Management Science*, 55(10):1605–1622.

Granka, L. A., Joachims, T., and Gay, G. (2004). Eye-tracking analysis of user behavior in www search. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 478–479. ACM.

Jerath, K., Ma, L., Park, Y.-H., and Srinivasan, K. (2011). A position paradox in sponsored search auctions. *Marketing Science*, 30(4):612–627.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161. ACM.

Kempe, D. and Mahdian, M. (2008). A cascade model for externalities in sponsored search. In *Internet and Network Economics*, pages 585–596. Springer.

Kohavi, R., Deng, A., Frasca, B., Longbotham, R., Walker, T., and Xu, Y. (2012). Trustworthy online controlled experiments: Five puzzling outcomes explained. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 786–794. ACM.

Kohavi, R., Henne, R. M., and Sommerfield, D. (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*, pages 959–967. ACM.

Lahaie, S. (2006). An analysis of alternative slot auction designs for sponsored search. In *Proceedings of the 7th ACM conference on Electronic commerce*, pages 218–227. ACM.

Lana, R. E. (1963). Interest, media, and order effects in persuasive communications. *The Journal of Psychology*, 56(1):9–13.

Lewis, R. and Reiley, D. Online ads and offline sales: Measuring the effects of retail advertising via a controlled experiment on yahoo! In *Working paper*.

Manzi, J. (2012). Uncontrolled: The suprising power of trial and error for business, politics and society.

Narayanan, S. and Kalyanam, K. Position effects in search advertising: A regression discontinuity approach.

Ostrovsky, M. and Schwarz, M. (2011). Reserve prices in internet advertising auctions: A field experiment. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 59–60. ACM.

Radlinski, F., Kurup, M., and Joachims, T. (2008). How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 43–52. ACM.

Rosen, A. M. (2008). Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities. *Journal of Econometrics*, 146(1):107–117.

Stock, J. H. and Yogo, M. (2002). Testing for weak instruments in linear iv regression.

Varian, H. R. (2007). Position auctions. *International Journal of Industrial Organization*, 25(6):1163–1178.

A Mathematical Proofs

Proof of Theorem 1

Proof. i. This is the standard consistency of IV estimators.

ii. If the OLS model is estimated with naive moment equalities, the estimator is

$$\begin{aligned}
\hat{\alpha}_{OLS} &= (T^{-1}1\sum_{t=1}^{T} X_t X_t')^{-1}(T^{-1}1\sum_{t=1}^{T} X_t Y_t) \\
&\to \alpha + E(X_t X_t')^{-1} E(X_t U_t) \\
&= \alpha + \begin{bmatrix} P(\text{Pos}=1) & 0 & 0 & 0 \\ 0 & P(\text{Pos}=2) & 0 & 0 \\ 0 & 0 & P(\text{Pos}=3) & 0 \\ 0 & 0 & 0 & P(\text{Pos}=4) \end{bmatrix}^{-1} \begin{bmatrix} E(U \cdot 1\{\text{Pos}=1\}) \\ E(U \cdot 1\{\text{Pos}=2\}) \\ E(U \cdot 1\{\text{Pos}=3\}) \\ E(U \cdot 1\{\text{Pos}=4\}) \end{bmatrix} \\
&= \alpha + \begin{bmatrix} E(U|\text{Pos}=1) \\ E(U|\text{Pos}=2) \\ E(U|\text{Pos}=3) \\ E(U|\text{Pos}=4) \end{bmatrix}. \qquad (13)
\end{aligned}$$

Recalling Assumption 3 it is easy to see, from 13, the prediction that OLS will yield upward biased estimates of α_1 and downward biased estimates of α_4. Then, since

$$\begin{aligned}\beta^j_{OLS} &= \alpha^j_{OLS} - \alpha^{j+1}_{OLS}\\ &= \beta_j + \Big[E(U|\text{Pos}=j) - E(U|\text{Pos}=j+1)\Big]\\ &\geq \beta_j,\end{aligned}$$

we see that estimates of all β_j are biased up.

iii. The standard IV estimator based on this model has probability limit

$$\beta^U_{IV} \xrightarrow{p} \begin{bmatrix}\alpha_{s+1}\\ \beta_s\end{bmatrix} + [P'WP]^{-1}[P'WQ], \tag{14}$$

where W is a positive definite weighting matrix, $Q = \sum_{r=1}^{s-1}\beta_r Q_r - \sum_{r=s+1}^{3}\beta_r(1-Q_r)$, and

$$Q_r = \Big[P(\text{Pos}\leq r) \quad P(Z_1=1,\text{Pos}\leq r) \quad P(Z_2=1,\text{Pos}\leq r) \quad ... \quad P(Z_J=1,\text{Pos}_r\leq s)\Big]'$$

$$P = \begin{bmatrix}1 & P(Z_1=1) & P(Z_2=1) & ... & P(Z_J=1)\\ P(\text{Pos}_t\leq s) & P(Z_1=1,\text{Pos}_t\leq s) & P(Z_2=1,\text{Pos}_t\leq s) & ... & P(Z_J=1,\text{Pos}_t\leq s)\end{bmatrix} = \begin{bmatrix}P_1\\ P_2\end{bmatrix}.$$

Bias of this estimator is given by calculating the second term in (16) as

$$\begin{aligned}[P'WP]^{-1}[P'WQ] = \frac{1}{|P'WP|}\Bigg(&\sum_{r=1}^{s-1}\beta_r \overbrace{\begin{bmatrix}P_2WP_2' & -P_2WP_1'\\ -P_1WP_2' & P_1WP_1'\end{bmatrix}\begin{bmatrix}P_1WQ_r\\ P_2WQ_r\end{bmatrix}}^{\Omega_{r,+}}\\ &+ \sum_{r=s+1}^{3}\beta_r \underbrace{\begin{bmatrix}P_2WP_2' & -P_2WP_1'\\ -P_1WP_2' & P_1WP_1'\end{bmatrix}\begin{bmatrix}P_1W(Q_r-1)\\ P_2W(Q_r-1)\end{bmatrix}}_{\Omega_{r,-}}\Bigg).\end{aligned}$$

With individual terms

$$\begin{aligned}\Omega_{r,+} &\equiv \begin{bmatrix}(P_2WP_2')(P_1WQ_r) - (P_2WP_1')(P_2WQ_r)\\ -(P_1WP_2')(P_1WQ_r) + (P_1WP_1')(P_2WQ_r)\end{bmatrix}\\ \Omega_{r,-} &\equiv \begin{bmatrix}(P_2WP_2')(P_1W[Q_r-1]) - (P_2WP_1')(P_2W[Q_r-1])\\ -(P_1WP_2')(P_1W[Q_r-1]) + (P_1WP_1')(P_2W[Q_r-1]).\end{bmatrix}\end{aligned}$$

The elements of $\Omega_{r,+}$ and $\Omega_{r,-}$ can both be calculated directly from the data. However, it is plausible to believe that each will contain a negative first element and positive second. This by Assumption 2 leads to a downward biased estimate of α_{s+1} and upward biased estimate of β_s in this model.

iv. The standard IV estimator based on this model is

$$\beta_{IV}^{L} = \left[\left(\sum_{t=1}^{T} X_t' Z_t\right) W \left(\sum_{t=1}^{T} Z_t' X_t\right)\right]^{-1} \left[\left(\sum_{t=1}^{T} X_t' Z_t\right) W \left(\sum_{t=1}^{T} Z_t' Y_t\right)\right] \tag{15}$$

$$\xrightarrow{p} \begin{bmatrix} \alpha_{s+1} \\ \beta_s \end{bmatrix} + [P'WP]^{-1}[P'WQ], \tag{16}$$

where W is a positive definite weighting matrix, $Q = \sum_{r=1}^{s-1} \beta_r Q_r - \sum_{r=s+1}^{3} \beta_r (1 - Q_r)$, and

$$Q_r = \begin{bmatrix} P(\text{Pos} \leq r) & P(Z_1 = 1, \text{Pos} \leq r) & P(Z_2 = 1, \text{Pos} \leq r) & ... & P(Z_J = 1, \text{Pos}_r \leq s) \end{bmatrix}'$$

$$P = \begin{bmatrix} 1 & P(Z_1 = 1) & P(Z_2 = 1) & ... & P(Z_J = 1) \\ P(\text{Pos}_t \leq s) & P(Z_1 = 1, \text{Pos}_t \leq s) & P(Z_2 = 1, \text{Pos}_t \leq s) & ... & P(Z_J = 1, \text{Pos}_t \leq s) \end{bmatrix} = \begin{bmatrix} P_1 \\ P_2 \end{bmatrix}.$$

Bias of this estimator is given by calculating the second term in (16) as

$$[P'WP]^{-1}[P'WQ] = \frac{1}{|P'WP|}\left(\sum_{r=1}^{s-1} \beta_r \overbrace{\begin{bmatrix} P_2WP_2' & -P_2WP_1' \\ -P_1WP_2' & P_1WP_1' \end{bmatrix} \begin{bmatrix} P_1WQ_r \\ P_2WQ_r \end{bmatrix}}^{\Omega_{r,+}} + \sum_{r=s+1}^{3} \beta_r \underbrace{\begin{bmatrix} P_2WP_2' & -P_2WP_1' \\ -P_1WP_2' & P_1WP_1' \end{bmatrix} \begin{bmatrix} P_1W(Q_r - 1) \\ P_2W(Q_r - 1) \end{bmatrix}}_{\Omega_{r,-}}\right).$$

With individual terms

$$\Omega_{r,+} \equiv \begin{bmatrix} (P_2WP_2')(P_1WQ_r) - (P_2WP_1')(P_2WQ_r) \\ -(P_1WP_2')(P_1WQ_r) + (P_1WP_1')(P_2WQ_r) \end{bmatrix}$$

$$\Omega_{r,-} \equiv \begin{bmatrix} (P_2WP_2')(P_1W[Q_r - 1]) - (P_2WP_1')(P_2W[Q_r - 1]) \\ -(P_1WP_2')(P_1W[Q_r - 1]) + (P_1WP_1')(P_2W[Q_r - 1]). \end{bmatrix}$$

The elements of $\Omega_{r,+}$ and $\Omega_{r,-}$ can both be calculated directly from the data. However, it is plausible to believe that each will contain a negative first element and positive second. This by Assumption 2 leads to a downward biased estimate of α_{s+1} and upward biased estimate of β_s in this model.

□