Evaluating Online Ad Campaigns in a Pipeline: Causal Models At Scale

David Chan
Google
New York, NY
davidch@google.com

Rong Ge
Google
Kirkland, WA
rongge@google.com

Ori Gershony
Google
Kirkland, WA
orig@google.com

Tim Hesterberg
Google
Seattle, WA
rocket@google.com

Diane Lambert
Google
New York, NY
dlambert@google.com

ABSTRACT

Display ads proliferate on the web, but are they effective? Or are they irrelevant in light of all the other advertising that people see? We describe a way to answer these questions, quickly and accurately, without randomized experiments, surveys, focus groups or expert data analysts. Doubly robust estimation protects against the selection bias that is inherent in observational data, and a nonparametric test that is based on irrelevant outcomes provides further defense. Simulations based on realistic scenarios show that the resulting estimates are more robust to selection bias than traditional alternatives, such as regression modeling or propensity scoring. Moreover, computations are fast enough that all processing, from data retrieval through estimation, testing, validation and report generation, proceeds in an automated pipeline, without anyone needing to see the raw data.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications-Data Mining; G.3 [**Mathematics of Computing**]: Probability and Statistics

General Terms

Measurement

Keywords

Ad effectiveness, display ads, doubly robust estimation, irrelevant outcomes, observational data, propensity score, selection bias

KDD'10, July 25–28, 2010, Washington, DC, USA.

1. INTRODUCTION

Online display ads have many formats. They may be plain text, static images, video clips, or games that users can play. An ad campaign may use just one format or a mix of formats that changes over time. The ad creatives (ad content) may change over time even if the ad format does not change. Some campaigns may use only a few creatives for months, while others may use thousands of creatives, changing them by site and over time. The rate at which impressions are served may change abruptly with short bursts of intense activity interspersed among lulls in ad serving. Or, a campaign may move from narrowly targeted to broadly targeted, and the sites at which ads are shown or the geographical areas that are targeted may vary. In other words, a campaign is whatever the advertiser defines it to be. The advertiser may simultaneously run similar campaigns in other media such as print, radio or television and its competitors may simultaneously run their own campaigns, perhaps showing display ads on the same websites. The question is whether a display ad campaign, however complex or simple it and its environment may be, is effective.

Traditionally, online campaign effectiveness has been measured by "clicks" because someone who interacts with an ad was affected by it. Besides, counting clicks is nearly cost-free. However, many display ads are not click-able, or at least not obviously so, and some campaigns hope to build longer-term interest in the brand rather than drive an immediate response. Counting clicks alone then misses much of the value of a campaign. More subtle effects can be elicited from focus groups and panels, but these provide small samples and are too expensive for routine application. Delayed responses to a campaign can be measured by counting visitors to the advertiser's website or users searching for brand terms during the campaign, but these metrics overstate campaign effectiveness. Some people would have visited the advertiser's website or searched for brand terms even if the campaign had never run. Better measures of campaign effectiveness are based on the *change* in online brand interest that can be attributed to the display ad campaign alone.

We propose robust estimates of the change in the probability that a user searches for brand terms or navigates to brand sites that can be attributed to an online ad campaign. The estimates require only summary data from opt-in users and can be computed and validated in an automated pipeline so even the summary data are not viewed by anyone. Only

highly aggregated statistics need to be released to advertisers, and then only if validation tests are passed.

Our measures of campaign effectiveness are developed in the framework of causal models introduced by Rubin [20]. Randomized experiments that randomly assign a user to the "served an ad" group or to the "not served an ad" group are the gold standard for estimating treatment effects in causal models. However, true randomization for display ads requires an advertiser to forego showing ads to some users, and possibly to pay for public service announcements or blank ads that appear in its allotted space instead. Often advertisers are not keen to relinquish an opportunity to advertise or to pay for non-campaign ads. Randomization also requires the study to be set up before the campaign runs.

Estimation without randomization is more difficult but not always impossible. Section 4 describes estimation based on *propensity scoring*; a refinement known as *doubly robust estimation* is discussed in Section 5. To further protect against residual, hidden selection bias, we introduce a new nonparametric test for observational studies (Section 6). Namely, the test statistic for an outcome of interest to the advertiser is judged against a null distribution that is based on a set of outcomes that are irrelevant to the advertiser. Additional validation steps are discussed in Section 7. To make the ideas concrete, data from an actual campaign are introduced in Section 3 and discussed throughout. Experimental results are reported in Section 8.

We are not the first to use causal models to evaluate campaign effectiveness. In a careful analysis of one set of data, Rubin and Waterman [21] used stratified propensity scores to evaluate the effect of visits from pharmaceutical sales teams on the number of prescriptions that physicians write. Fulgoni and Morn [5] match exposed and control panelists on a set of characteristics $\mathbf{X}$ such as query volume and then compare the fractions of matched control and exposed samples that show interest in the advertiser after the campaign. But matching can fail when $\mathbf{X}$ has many dimensions, and it does not lead to an obvious test of residual, undetected selection bias. Our approach avoids matching on $\mathbf{X}$ and has built-in safeguards against hidden selection bias, and so is more suited for routine application. We believe that this is the first method that has enough built-in safeguards to run in an automated pipeline that retrieves data, computes estimates, and decides whether to release results, suppress results, or send them to an expert data analyst for review.

2. CAUSAL EFFECTS

Imagine a parallel universe exactly like ours, except for one small change: an advertiser never ran a particular display ad campaign. Competitors still advertised their products, and the advertiser still advertised in other media, but the campaign display ads did not run. What would have changed? Would fewer people have visited the advertiser's web site or searched for the advertiser's products and services, or would those counts not have changed? If nothing would have changed, then the campaign was ineffective. This notion of *counterfactuals*, or what would have happened had the campaign not been run, is fundamental to understanding the analysis of observational data found in corporate and government databases.

Re-stating in the language of statistics, an advertiser is interested in an outcome Y, such as "user visits a brand website" or "user searches for a brand term", where positive values of Y indicate interest in the advertiser's brand. Every user who could be exposed to a campaign ad has two potential values (Y_0, Y_1) of the outcome, where Y_0 will be the user's outcome if not shown a campaign ad and Y_1 will be the user's outcome if shown an ad. Both Y_0 and Y_1 cannot be observed for the same user because a user either is or is not shown a campaign ad. The unobservable outcome, regardless of whether it is Y_0 or Y_1, is called a *counterfactual*. The observed outcome will be denoted by Y.

The unobservable per-user difference $Y_1 - Y_0$ is the effect of the campaign on that user. If $Y_1 = Y_0$, then the campaign had no effect on that user, at least not as measured by the outcome Y. Only $Y_1 > Y_0$ is a favorable response. The average campaign effect on those served an ad is then

$$\Delta = E\left(Y_1 - Y_0\right) = E\left(Y_1\right) - E\left(Y_0\right),$$

where E denotes an average taken over all users who were served ads. Note that Δ does not include users who might have been served ads but were not. In other words, Δ measures the effect of the campaign as it was run, not what the effect would have been had the advertiser been able to serve everyone an ad. In the terminology of causal models, Δ is called the treatment effect on the treated [22].

The challenge is to estimate the average unobservable outcome $E\left(Y_0\right)$ for the users who were served ads. Randomly assigning users to a test group that is served campaign ads and a control group that is not served ads would make that easy. On average the randomized exposed and control groups will be similar except for their exposure status. We would expect large random samples of controls and exposed users to be equally active on the web, to have the same geographical distribution, and to be equally interested in the advertiser before the campaign, for example. Randomization alone then justifies using the average observed outcome $\bar{Y}_{control}$ of the controls as a proxy for the unobservable "without-campaign" average outcome $E\left(Y_0\right)$ for the exposed, so $\bar{Y}_{exposed} - \bar{Y}_{control}$ estimates Δ.

Unfortunately, the process of serving ads does not randomly assign users to exposed and control groups. The most glaring problem is that active web users are much more likely to fall in the exposed group than inactive users are. Anyone who arrives at a webpage when an advertiser is alloted space on the page and who satisfies the targeting conditions for the campaign is served an ad. The more a web user visits a web page that sometimes shows campaign ads and sometimes does not, the more likely the web user is to see a campaign ad. So, more active web users are more likely to be exposed and less active web users are more likely to be unexposed and hence potential controls, even if they satisfy the targeting conditions for the campaign. Unfortunately, more active web users are also more likely than inactive users to visit any site on the web, including the advertiser's site, even if they are not shown an ad. In other words, the without-campaign average outcome $E(Y_0)$ for the exposed is probably higher than $E(Y_0)$ for those not served ads, making the unexposed users poor surrogates for the exposed in the without-campaign state. In short, the ad serving process leads to *selection bias* or differences in the control and exposed that are related to the distribution of the outcomes (Y_0, Y_1). If ignored, this selection bias contaminates the estimated campaign effect. Our goal is to use the unexposed users to produce statistically sound estimates of $E\left(Y_0\right)$ for the exposed.

3. THE CONTROLS

Simply put, the controls were eligible to be served campaign ads but were not. More precisely, our controls are users that satisfy the following four conditions C1 - C4. The controls

C1 met targeting conditions, such as country and language,

C2 visited a website near a time that it served a campaign ad to an exposed user in the study,

C3 were served at least one non-campaign ad on that visit to the publisher site, so were not blocking ads served by google.com, and

C4 were not served campaign ads during the study period.

Websites that show campaign ads are called *publisher sites*.

The first ad served to an exposed user breaks its timeline into before and after exposure periods. Before and after periods are defined for each control by choosing one of its visits to a publisher site during the campaign to be a *pseudo-exposure*. The duration of the before period is the same for each user. The after period extends from the time of the first exposure or psuedo-exposure until the end of the campaign or until the end of a post-campaign follow-up period during which the advertiser may continue to accumulate responses to the campaign. An advertiser may also specify that the period to be analyzed end before the campaign ends, in which case there is no follow-up period.

Our estimates require summary (not personally identifiable) data on exposed and controls. The summary data are obtained from several sources, including the advertiser's own campaign information, ad serving logs, and sampled data from users who have installed Google toolbar and opted in to enhanced features. The summaries include features such as the country and language from which the user most often accessed the web and measures of online activity in the before period, such as total number of navigations, number of navigations to the sites that showed campaign ads, and number of navigations to the advertiser's site before exposure. Exposure data include time of first exposure and the number of campaign ads served to the user during the analysis period. If the analysis captures only a slice of an ongoing campaign, then campaign exposure before the analysis period is also measured. Finally, brand outcomes are actions that the advertiser believes will be affected by the campaign, such as whether the user navigated to the advertiser's website or searched for a brand term after exposure.

To protect privacy, estimates are not released for rare outcomes. All of the results reported to advertisers are aggregated over thousands of users. Moreover, all the data used in the analysis are anonymized, so it is not possible to match a summary record back to a small sample of records in the raw logs. For example, geographical areas are aggregated to meet privacy constraints. In addition, because the system is fully automated, no one sees even the summary records.

Although the controls met the targeting conditions C1 - C4, so could have been served ads, they are different from those who actually were served ads. In particular, as predicted in Section 2, the exposed are more active on the web than the controls are. Figure 1 compares the online activity of the controls and exposed before their first exposure for a campaign that lasted 42 days and had about 150 different ad creatives. The study includes about 15,000 exposed users and 70,000 controls. Each of the four panels in Figure 1 considers a different measure of online activity for these samples, and each point in a panel shows the ratio of a percentile of the activity for the exposed to the same percentile of the activity for the controls. If controls and exposed were equally active, their percentiles would be equal, and the ratios of their percentiles would be one. Clearly, that is not so; the exposed are much more active. The median exposed navigation count is 2.8 times the median control navigation count, the median exposed navigations to sites that published campaign ads is 4.1 times the median for the controls, and the median number of non-campaign display ads served to the exposed on the publisher sites is 8.7 times that of the controls. In other words, the controls did not act like the exposed before exposure, so the average outcome of the controls after exposure is likely to be a poor estimate of the average counterfactual $E(Y_0)$ for the exposed.

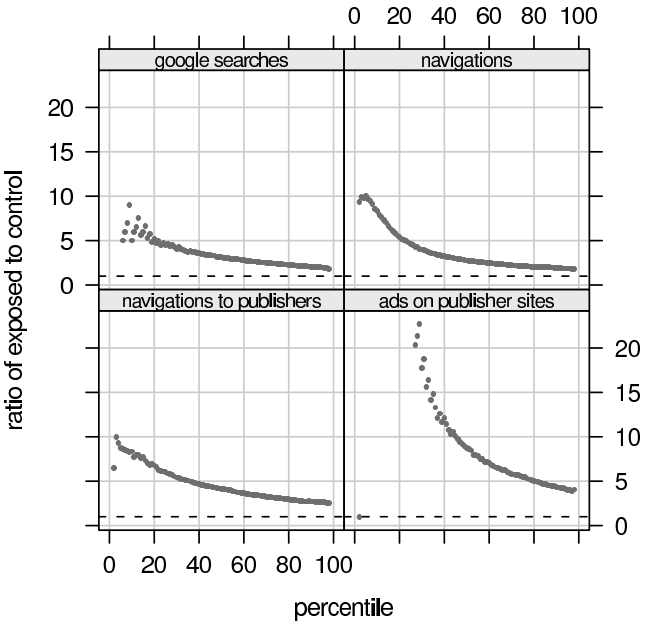

Figure 1: Ratios of the percentiles of the activity levels of exposed web users and controls satisfying conditions [C1] - [C4] before exposure. The dashed line corresponds to a ratio of one.

4. PROPENSITY SCORES

If a control and exposed user were identical before exposure, then it is reasonable to assume that they would have had the same probability of showing interest in the brand after exposure if the campaign had not been run. This suggests matching each exposed user to one or more controls pre-exposure and omitting the unmatched controls. More precisely, if an exposed and control user are matched on the K features $\mathbf{X} = (X_1, \ldots, X_K)$, then successful matching ensures that

$$P\left(Y_0 = 1 \mid \mathbf{X}, \textit{exposed}\right) = P\left(Y_0 = 1 \mid \mathbf{X}, \textit{control}\right)$$

when the outcome of interest is binary. All valid analyses of observational data either explicitly or implicitly assume that there is an $\mathbf{X}$ that breaks the dependence between exposure and the unobservable counterfactual Y_0 in this way.

4.1 Propensity Matching

Matching on $\mathbf{X}$ breaks down when $\mathbf{X}$ is high dimensional, especially if some dimensions are heavily skewed. Fortunately, a remarkable theorem in [18] states that matching on $\mathbf{X}$ is unnecessary. If

$$0 < p(\mathbf{X}) < 1,$$

so no user is certain to be a control or certain to be exposed, then matching on the one-dimensional *propensity score* $p(\mathbf{X})$ defined by

$$p(\mathbf{X}) = P(exposed \mid \mathbf{X}) \tag{1}$$

removes selection bias whenever matching on $\mathbf{X}$ itself removes selection bias. (Because our goal is to estimate the causal effect on those who were exposed rather on everyone who could have been exposed, we need only assume that every control could have been served; i.e., $p(\mathbf{X}) > 0$.) Surprisingly, matching on a consistent estimate $\hat{p}(\mathbf{X})$ of $p(\mathbf{X})$ can remove selection bias better than matching on $\mathbf{X}$ or $p(\mathbf{X})$ can [19].

After matching each exposed with a control, the campaign effect Δ can be estimated by the average within-pair difference of observed outcomes. Or, controls and exposed with similar propensities can be grouped together, and mean differences within groups averaged [19]. Propensity score grouping has been used extensively in clinical medicine, epidemiology and the social sciences. (See [12], [11] and the references therein, for example.) There are disadvantages, though. It may not be possible to pair each exposed to a different control or even to any control. Grouping may include too few controls at high propensities to give reliable group mean differences. Or, groups may be so coarse that the controls and exposed in a group are not well-matched. Stratifying is also inconsistent (asymptotically biased) when the mean outcome is not constant within groups [13].

Figure 2 shows the estimated propensities using logistic regression with variable selection for the study introduced in Section 3. (See Section 8 for a discussion of model selection in this context.) As expected, the controls have smaller $\hat{p}(\mathbf{X})$. The median exposed propensity score is 0.31 but the median control propensity score is only 0.10. The extreme exposed and control propensities are similar, though. The exposed range is $(.004, .94)$ and the control range is $(.004, .89)$. Pairing exposed and controls by $\hat{p}(\mathbf{X})$ is not likely to be successful here because many fewer controls than exposed have high propensities (e.g., above 0.5).

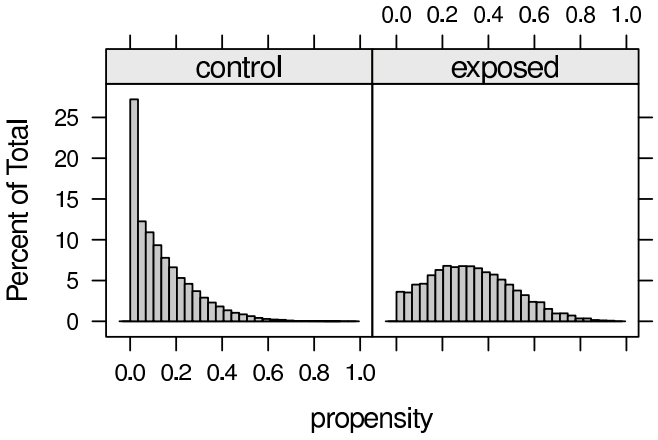

Figure 2: Estimated propensities for the controls and exposed in Figure 1.

4.2 Inverse Propensity Weighting

Propensity score weighting is an alternative to propensity score matching. It can be motivated by a simple analogy. Suppose that 35% of a target population is from Canada and 65% from the U.S., and the outcome Y is smaller in Canada. If 60% of a random sample is from Canada, then the average over the random sample underestimates the mean in the target population. A better estimate weights everyone in the sample from Canada by .35/.6 and everyone in the sample from the U.S. by .65/.4 before averaging. That is, there is selection bias because a feature (country) is correlated with both sample selection and the outcome, but re-weighting the data with weights inversely proportional to the probability of selection removes the bias. Averaging with known inverse sampling weights was introduced by Horvitz and Thompson in 1952 [9] and has been further studied in recent KDD papers such as [10].

In nonrandomized studies, the controls can be weighted to resemble the sample of exposed before exposure by giving each control i a weight proportional to

$$w_i = \hat{p}(\mathbf{X}_i) / (1 - \hat{p}(\mathbf{X}_i)) \tag{2}$$

where the sum of the weights over the controls equals one. This is called *inverse propensity weighting*. The exposed users are not weighted. Weighting the exposed by $1/\hat{p}(\mathbf{X})$ and the controls by $1/(1-\hat{p}(\mathbf{X}))$ is appropriate when the goal is to estimate the potential campaign effect on all users.

Figure 3 shows that inverse propensity weighting effectively matches the controls to the exposed for the campaign in Section 3. It reduces the ratios of the exposed medians to the control medians for the four activities shown from $(3.1, 2.8, 4.1, 8.7)$ to $(1.04, 1.06, 1.07, 1.14)$. The ratios for all but the smallest percentiles now lie between 0.8 and 1.20. Moreover, search activity on google.com is now balanced, even though variable selection deleted that term from the estimated propensity model.

Because inverse propensity weighting matches the controls to the sample of exposed, the *inversely propensity weighted (IPW) estimate* of Δ is defined by

$$\hat{\Delta}_{IPW} = \bar{Y}_{exposed} - \left(\sum_{controls} w_i Y_i / \sum_{controls} w_i \right) \tag{3}$$

where Y_i is the observed outcome for a control and w_i is defined by equation (2).

Although $\hat{\Delta}_{IPW}$ is asymptotically unbiased, it has high variance if $p(\mathbf{X})$ is close to one for some controls. Propensities close to one arise if $\mathbf{X}$ nearly separates the controls and exposed. In that case, estimation by any method may be unwise because too few controls resemble the exposed. Estimated propensities close to one may also occur because the algorithm used to estimate the propensities is unstable. In a criminal justice application, McCaffrey, Ridgeway and Morral [14] were able to match propensity weighted controls to exposed using boosted stumps but not logistic regression, for example.

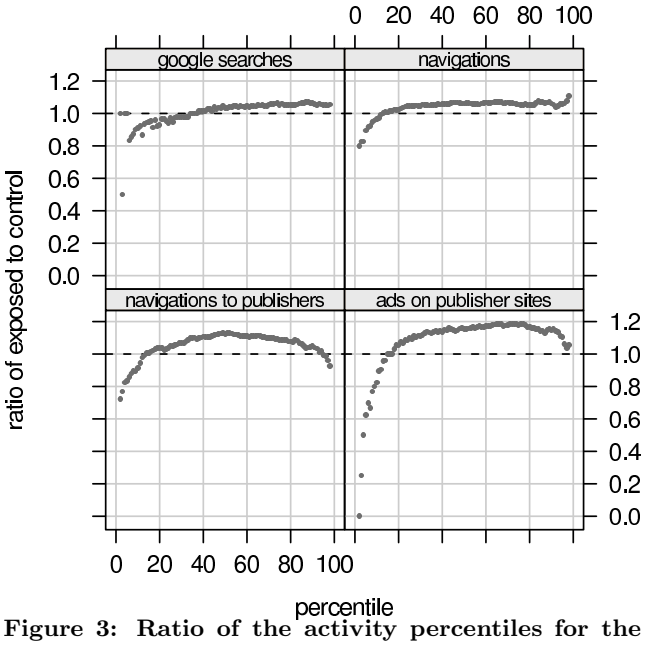

Figure 3: Ratio of the activity percentiles for the exposed and propensity weighted controls.

5. DOUBLY ROBUST ESTIMATION

Inverse propensity weighting is fundamental to causal estimation because any estimator of Δ that is asymptotically unbiased is equivalent to an estimate based on inverse propensity weighting [16]. The question is which estimate with inverse propensity weighting is best in the sense of smallest asymptotic variance among all consistent estimates of Δ?

The variability of $\hat{\Delta}_{IPW}$ depends on random noise in the observed outcomes, which affects any estimate, and errors in $\hat{p}(\mathbf{X})$. To adjust for those errors, first define the exposure indicator $Z_i = 1$ if user i is exposed and $Z_i = 0$ if not for the $i = 1, \ldots, n$ users in the study and let $\hat{p}_i = \hat{p}(\mathbf{X}_i)$. Then consider the estimate

$$\hat{\Delta}_* = \sum_{i=1}^{n} \hat{p}_i \left\{ \frac{Z_i Y_i}{\hat{p}_i} - \frac{(1 - Z_i) Y_i}{1 - \hat{p}_i} \right\} / \sum_{i=1}^{n} \hat{p}_i,$$

which is identical to $\hat{\Delta}_{IPW}$ if $\hat{p}(\mathbf{X}_i)$ is constant but not in general. The estimate $\hat{\Delta}_*$ is consistent when $\hat{p}(\mathbf{X})$ is consistent. If we were interested in the causal effect on everyone who might have been served a campaign ad rather than the effect on those who were served a campaign ad, the leading multiplier $\hat{p}_i$ outside the braces would be omitted and the denominator would be n rather than $\sum \hat{p}_i$.

The theory in [16], [13], and [8] shows that the minimum asymptotic variance is achieved by modifying $\hat{\Delta}_*$ by the errors $Z_i - \hat{p}_i$ as follows:

$$\hat{\Delta}_{DR} = \sum_{i=1}^{n} \hat{p}_i \hat{\delta}_i / \sum_{i=1}^{n} \hat{p}_i, \tag{4}$$

where

$$\hat{\delta}_i = \frac{Z_i Y_i - \hat{m}_{1i}(Z_i - \hat{p}_i)}{\hat{p}_i} - \frac{(1 - Z_i) Y_i + \hat{m}_{0i}(Z_i - \hat{p}_i)}{1 - \hat{p}_i}$$

and $\hat{m}_{zi} = \hat{m}_z(\mathbf{X}_i)$ is consistent for $E(Y \mid \mathbf{X}_i, Z_i = z)$. Usually, $\hat{m}_1(\mathbf{X})$ is obtained by regressing the observed Y on $\mathbf{X}$ for the exposed, and $\hat{m}_0(\mathbf{X})$ is obtained by regressing Y on $\mathbf{X}$ for the controls. In other words, computing $\hat{\Delta}_{DR}$ is not much more difficult than computing $\hat{\Delta}_{IPW}$. (Here we have used the fact that the standard theory remains valid if the term in braces is multiplied by a weight like $\hat{p}_i$ that does not depend on the outcome or Z.)

The estimate $\hat{\Delta}_{DR}$ is *doubly robust* because it remains consistent if the outcome models are wrong but the propensity model is right or if the propensity model is wrong but the outcome models are right, although in those cases there is no guarantee that $\hat{\Delta}_{DR}$ has minimum asymptotic variance. Extensive simulations (e.g., [13]) show that the asymptotic claims hold in standard settings and that the standard error of $\hat{\Delta}_{DR}$ has a simple estimate $\hat{s}_{DR}$ defined by

$$\hat{s}_{DR}^2 = n^{-2} \sum_{i=1}^{n} q_i^2 \left(\hat{\delta}_i - \hat{\Delta}_{DR} \right)^2, \quad \text{where} \tag{5}$$

$$q_i = \frac{\hat{p}_i}{n^{-1} \sum_{i=1}^{n} \hat{p}_i}$$

Doubly robust estimation is used in applications ranging from medicine [2] to criminal justice [15]. The statistical analysis systems SAS [1] and STATA [3] include procedures for computing doubly robust estimates, standard errors, and model diagnostics.

Table 1 gives $\hat{\Delta}_{IPW}$, $\hat{\Delta}_{DR}$ and a regression estimate

$$\hat{\Delta}_{reg} = n_1^{-1} \sum_{exposed} \left(\hat{m}_1(\mathbf{X}_i) - \hat{m}_0(\mathbf{X}_i) \right), \quad n_1 = \sum_{i=1}^{n} Z_i$$

for the campaign introduced in Section 3. Brand navigation refers to navigation to the advertiser's website; brand search refers to searches for the advertiser or its advertised product. Lift is the ratio of $\hat{\Delta}_{DR}$ to the average prediction of the counterfactual Y_0 for the exposed. The campaign increased interest in the brand and, to a lesser extent, its competition. Increased interest in the competition is unavoidable when exposed users comparison shop.

Table 1: Estimates of the causal effect for brand navigations, brand searches, and competitor navigations for the campaign in Section 3.

Outcome	$\hat{\Delta}_{IPW}$	$\hat{\Delta}_{reg}$	$\hat{\Delta}_{DR}$	$\hat{se}_{DR}$	lift
brand nav	.025	.027	.009	.002	.37
brand srch	.048	.037	.016	.002	.34
compet nav	.294	.091	.022	.004	.08

6. HYPOTHESIS TESTING

The obvious test of the null hypothesis $H_0 : \Delta \leq 0$ against $H_1 : \Delta > 0$ with size α (false alarm probability α) compares $T_{DR} = \hat{\Delta}_{DR} / \hat{s}_{DR}$ to a standard normal$(0, 1)$ percentile q_α. This test should perform well when the propensity model and outcome regression models are adequate. But if a feature that is correlated with Z and Y but not the features in $\mathbf{X}$ has been missed, then differences in the missing feature for the controls and the exposed that are not germane to the campaign may cause T_{DR} to wrongly reject H_0. This is the problem of hidden, residual selection bias.

Rosenbaum [17] suggests detecting hidden bias by testing H_0 for an outcome that should be unaffected by the treatment (the campaign, in our context). For example, a campaign that advertises Google's web browser Chrome may

have recipes as an irrelevant search term. If H_0 is rejected for the irrelevant outcome then there may be residual selection bias and a test of H_0 based on T_{DR} for an outcome of interest may have more than probability α of a false rejection.

We take the idea of irrelevant outcomes a step further, and compute T^* for many irrelevant outcomes. The T^*'s for the irrelevant outcomes should act like a random sample from the null distribution of the brand test statistic. When H_0 is true, the brand outcome should act like an irrelevant outcome too. In other words, instead of comparing T_{DR} to a percentile of a normal distribution, which assumes that selection bias has been controlled, compare T_{DR} to a null distribution that allows for partially removed selection bias.

If there are K irrelevant outcomes, then there are K test statistics $T_1^*, \ldots, T_K^*$ and so under H_0 there are $K+1$ observations in all, including T_{DR} for the brand outcome, from the null distribution. In keeping with standard testing practice, H_0 is rejected only if T_{DR} is larger than a tail percentile of this null distribution. A fully nonparametric test is obtained by rejecting H_0 only if T_{DR} is larger than the $(k_\alpha+1)/(K+1)$ quantile of the empirical null distribution where $k_\alpha = min\{k : (k+1)/(K+1) \geq \alpha\}$. It is possible to fit a parametric distribution to the $K+1$ test statistics and use a percentile of the parametric distribution to test H_0, but we prefer not to make assumptions about the shape of the null distribution.

It may be challenging to obtain data on many irrelevant outcomes for medical studies, but that is not the case for online campaigns. There are many websites and search terms that are irrelevant to an advertiser. These can be found by looking at advertisers in other segments or clusters, for example. An automated system may wrongly declare such a term or site to be irrelevant, but that should happen so infrequently that the nonparametric test is hardly affected. (See, for example, Section 8.) Finally, all results and summary data on irrelevant outcomes can be withheld because by definition the irrelevant outcomes are unimportant to the advertiser. Also note that a test that compares a brand outcome to irrelevant outcomes may be easier for advertisers to appreciate than a standard statistical hypothesis test is.

The three outcomes in Table 1 are statistically significant at the $\alpha = .05$ level of significance when T_{DR} is compared to either a percentile of a normal distribution or a nonparametric null distribution defined by a set of irrelevant outcomes. However, the experiments in Section 8 suggest that the test based on normal distributions is too liberal even when the propensity and outcome models are correct. The nonparametric test is more trustworthy.

7. MODEL SELECTION AND VALIDATION

The estimated campaign effect $\hat{\Delta}_{DR}$ depends on two outcome models and a propensity model. In this paper the outcomes Y and exposure status Z are binary, and any method for estimating probabilities that includes feature selection can potentially be used to compute $\hat{\Delta}_{DR}$, including regression trees, boosted stumps, leaps and bounds [6], or L_1 and L_2 penalized regression [4].

It is not enough for a propensity model to fit the data well, though. Its primary goal is to balance $\mathbf{X}$ across controls and exposed, in the sense that the distribution of $\mathbf{X}$ conditional on $(\hat{p}(\mathbf{X}), Z)$ should be independent of Z even for features X that are not included in the fitted propensity model. In small studies, this assumption can be tested by stratifying users into several groups (a common recommendation is seven groups [19]) according to their $\hat{p}(\mathbf{X}_i)$'s, computing the difference $D_g = \bar{X}_{g,exposed} - \bar{X}_{g,control}$ in each group g, and then using an analysis of variance (anova) F-test to test whether the within-group mean differences are zero. An alternative in large studies is to compute D_g corresponding to a fine grid of $\hat{p}(\mathbf{X}_i)$ rather than a coarse grid, fit a smooth function to D_g as a function of the group mean $\hat{p}(\mathbf{X})$, and then test whether the smooth curve is different from a horizontal line at zero. This can be done in the open source software environment for statistical computing and graphics R (`www.r-project.org`) by applying the anova function in R to the output from the gam (generalized additive model) function, for example.

The propensity weights $w_i = \hat{p}(\mathbf{X}_i)/(1-\hat{p}(\mathbf{X}_i))$ themselves are useful for testing the validity of the results of a study. For example, an analysis may be declared invalid if a small subset of controls accounts for too high a fraction of the total weight on the controls because in that case $\hat{\Delta}_{DR}$ and $\hat{\Delta}_{IPW}$ may be determined by only a small subset of the data and $\hat{\Delta}_{reg}$ may be based on extrapolation. For the campaign introduced in Section 3, only 2% of the controls have $w_i > 1$ and only 0.09% have a weight larger than 3. The largest weight is 9.4, so the most extreme control has as much weight as 9.4 exposed users.

As another check of the validity of a study, note that if n independent observations $\{A_1, \ldots, A_n\}$ have the same mean and same variance and the n weights $\{W_i, \ldots, W_n\}$ are fixed, then the weighted mean $\sum W_i A_i / \sum W_i$ has variance proportional to $\sum W_i^2 / (\sum W_i)^2$. The same result is approximately true if the weights are random and independent of the $A_i's$. Because the sample mean $\bar{A}$ has variance proportional to $1/n$, the effective sample size for a weighted mean in either case is $(\sum W_i)^2 / \sum W_i^2$. This suggests failing a study when the effective sample size for the controls is too small. For the campaign introduced in Section 3, the effective number of controls is about 30% of the total number of controls, giving a ratio of 1.4 effective controls per exposed user in the study, which also suggests that the study is valid.

Finally, Hainmueller [7] estimates the propensity function by minimizing the deviance under an assumed model, like logistic regression, subject to the constraint that the means of $\mathbf{X}$ for the weighted control sample equal the means of $\mathbf{X}$ for the exposed. That is an interesting proposal, but not one that we have tried to automate. It also raises the question of what to match, e.g. means of the features themselves, means and second moments of the features, or means of transformations of the features?

8. EXPERIMENTS

Often less than 5% of the exposed searched for the advertiser's brand or navigated to the advertiser's website before the campaign. If a campaign increases that rate by 50%, then the effect to be estimated is less than 0.025. Here we provide experimental evidence that it is possible to estimate such small effects for online campaigns.

The experiments simulate users like those in the campaign introduced in Section 3, except that exposed outcomes correspond to a specified lift. First, a model to generate users

was built in stages using the fact that

$$P(X_1, \ldots, X_K) = \prod_{k=1}^{K} P(X_k | X_1, \ldots, X_{k-1}), \ k = 2, \ldots, K.$$

At the first stage, the probability that a user in the study is exposed (and not a control) is set to the fraction of exposed users in the study. At the second stage, the distribution of geography conditional on exposure status is taken to be the empirical distribution of geography for the exposed and the empirical distribution of geography for the controls. At later stages, the log of an activity metric is taken to be normally distributed with a conditional mean and variance that depend on the features previously modeled. The conditional mean and variance are estimated from a linear regression with leaps and bounds variable selection. Although the conditional models are normal, the unconditional distributions for the controls and exposed are as long-tailed as those in the original data due to mixing over the conditioning variables. The final multivariate distribution is then a mixture of binary, categorical and continuous variables that is much more complex but also much more realistic than one based on standard parametric distributions. It is, however, not pathological so we do not claim that the simulations represent a worst possible scenario.

Models were produced for 31 outcomes irrelevant to the advertiser and the three brand outcomes in Table 1. The models for (Y_0, Y_1) for each irrelevant outcome were obtained from separate logistic regressions for the controls and exposed. A model of the "without-campaign" brand outcomes Y_0 for both the control and exposed users was obtained by fitting a logistic regression to the brand outcomes for the controls. The model of Y_1 is then chosen to give either $\Delta = 0$ or a Δ corresponding to 50% lift according to the following three scenarios.

No Effect, No Hidden Bias The brand outcomes Y_1 for the exposed are simulated from the brand model for Y_0. All features $\mathbf{X}$ of the simulated users are available for model fitting, but some may be dropped during model selection. There is selection bias in this scenario, but there no hidden bias unless model fitting deletes an important feature from the propensity and outcome models.

Positive Shift The brand outcomes Y_1 for the exposed are generated by adding a shift θ to the intercept in the model for Y_0, where θ gives a lift of 50%. The θ for brand navigation, brand search, and competitor navigation are 0.51, 0.58 and 0.84 respectively, which correspond to campaign effects of $\Delta = .013, .024, .147$. The mean no-campaign outcome $E(Y_0)$ for the exposed are .026, .050 and .295 respectively. There is no hidden selection bias in this scenario, except that due to variable selection.

No Effect, Hidden Bias The brand outcomes (Y_0, Y_1) are generated from the, same model, so there is no campaign effect, but number of navigations to the websites that served ads and number of display ads served by Google on those sites were not used to fit the outcome or propensity models.

Each scenario was simulated 500 times, and 80,000 users were generated in each simulation trial.

Table 2 summarizes the results for the no effect, no hidden bias scenario. Because there is no hidden bias, all three estimates $\hat{\Delta}_{DR}$, $\hat{\Delta}_{reg}$ and $\hat{\Delta}_{IPW}$ should behave well and they do, although, as the theory predicts, $\hat{\Delta}_{DR}$ behaves slightly better. However, the large sample standard error estimate $\hat{s}_{DR}$ (6) is much too optimistic. As a result, the parametric test that compares T_{DR} to the .10 quantile of a normal(0,1) distribution has a false alarm rate of about 20% for the two advertiser related outcomes instead of 10%. (The test for the effect on navigations to competitor sites appears to be unbiased). The nonparametric test against irrelevant outcomes is a better choice. Under the null hypothesis the mean p-value should be .50 and 10% of the p-values should be below .10 (except for small differences due to discreteness). While the mean nonparametric p-value is less than .50 for the two advertiser related outcomes, the simulated false alarm rate is also smaller than 10% for all three outcomes so the nonparametric test is conservative.

Table 2: Results for the no effect, no hidden bias scenario. $rmse$ is root mean squared error, $(bias^2 + sd^2)^{1/2}$. $\bar{P}_{DR}$ is the mean simulated p-value under the null distribution defined by the irrelevant outcomes. $\bar{P}(T_{DR} > z_{.10})$ is the fraction of (false) rejections of a one-sided $\alpha = .10$ test in the simulation using the large sample normal theory test. $\bar{P}(P_{DR}) < .10)$ is the fraction of false alarms for the nonparametric test based on irrelevant outcomes.

	brand nav	brand search	comp
rmse($\hat{\Delta}_{DR}$)	.0051	.0053	.0168
rmse($\hat{\Delta}_{reg}$)	.0056	.0061	.0243
rmse($\hat{\Delta}_{IPW}$)	.0059	.0060	.0253
mean($\hat{s}_{DR}$)/sd($\hat{\Delta}_{DR}$)	.83	.82	.80
$\bar{P}(T_{DR} > z_{.10})$	.19	.21	.09
mean(P_{DR})	.40	.38	.50
mean(P_{DR}) < .10)	.05	.06	.03

Table 3 gives the results for the scenario with positive shift and no feature withheld from model fitting. Surprisingly, $\hat{\Delta}_{DR}$ is less biased than either $\hat{\Delta}_{reg}$ or $\hat{\Delta}_{IPW}$ for the advertiser related outcomes, and only slightly more biased than $\hat{\Delta}_{reg}$ for the competitor outcome. The relative biases $bias/\Delta$ for $\hat{\Delta}_{DR}$, $\hat{\Delta}_{reg}$ and $\hat{\Delta}_{IPW}$ for brand navigations are 2%, 5% and 10% respectively, while those for brand search are -1%, 4% and 9% respectively. As in the null case, the large sample estimate $\hat{s}_{DR}$ is smaller than it should be, but only by about 10%. The nonparametric test for positive shift rejects for most simulation trials, as it should.

Table 4 shows the results of omitting two of the features that were important to the models for the study data and were used to generate the simulated outcomes. Both features increase the chance that $Y_1 = 1$, so omitting both should overstate the effect of the campaign. The theory suggests that $\hat{\Delta}_{DR}$ should behave well if the omitted features are included in either the propensity model or the outcome models, but here they are excluded from both models. Although the theory is silent on this case, omitting features from both models seems more realistic than omitting them from only the propensity model or only the outcome models. This experiment shows that the three estimates respond differently to hidden bias. The regression and propensity weighted es-

Table 3: Results for the positive shift scenario.

	brand nav	brand search	comp
bias($\hat{\Delta}_{DR}$)	.00030	.00034	-.00037
bias($\hat{\Delta}_{reg}$)	.00066	.00105	-.00021
bias($\hat{\Delta}_{IPW}$)	.00130	.00220	.00650
sd($\hat{\Delta}_{DR}$)	.0023	.0033	.0060
sd($\hat{\Delta}_{reg}$)	.0021	.0031	.0048
sd($\hat{\Delta}_{IPW}$)	.0028	.0036	.0062
mean($\hat{s}_{DR}$)/sd($\hat{\Delta}_{DR}$)	.89	.91	.87
mean(P_{DR}) < .10)	.98	1.00	1.00

timates are more biased but less variable than $\hat{\Delta}_{DR}$. The ratios of the rmse of $\hat{\Delta}_{reg}$ to $\hat{\Delta}_{DR}$ fall between 1.10 and 1.44, while the ratios for $\hat{\Delta}_{IPW}$ to $\hat{\Delta}_{DR}$ fall between 1.14 and 1.56, so $\hat{\Delta}_{DR}$ has the better overall behavior. Note that although $\hat{\Delta}_{DR}$ is biased high and its estimated standard error is too small, the nonparametric test of the null hypothesis of no campaign effect does not break down for brand navigation and brand search and is not too far off for competitor navigation.

Table 4: Results when there is hidden bias, no shift.

	brand nav	brand search	comp
bias($\hat{\Delta}_{DR}$)	.0037	.0062	.011
bias($\hat{\Delta}_{reg}$)	.0041	.0074	.018
bias($\hat{\Delta}_{IPW}$)	.0044	.0080	.020
sd($\hat{\Delta}_{DR}$)	.0020	.0028	.0068
sd($\hat{\Delta}_{reg}$)	.0017	.0022	.0041
sd($\hat{\Delta}_{IPW}$)	.0019	.0025	.0044
rmse($\hat{\Delta}_{DR}$)	.0042	.0068	.0128
rmse($\hat{\Delta}_{reg}$)	.0046	.0077	.0184
rmse($\hat{\Delta}_{IPW}$)	.0048	.0084	.0200
mean($\hat{s}_{DR}$)/sd($\hat{\Delta}_{DR}$)	.86	.78	.61
$\bar{P}(T_{DR} > z_{.10})$	.78	.86	.78
mean(P_{DR})	.28	.22	.25
mean(P_{DR}) < .10)	.08	.09	.14

9. DISCUSSION

This paper has focused on estimating the effect of a campaign on everyone who was served a campaign ad. The expression (4) can also be used to estimate effects on subsets of users, as long as the subset can be identified without knowing the outcomes or exposure status of the users. For example, an advertiser may want to estimate the effectiveness of a campaign by region. A propensity model and outcome models can then be fit to the data from all regions together, and the effect within region j estimated by

$$\hat{\Delta}_{DR}(j) = \sum_{i=1}^{n} W_{ij}\hat{\delta}_i / \sum_{i=1}^{n} W_{ij},$$

where $W_{ij} = \hat{p}_i U_{ij}$ and $U_{ij} = 1$ if user i is from region j and $U_{ij} = 0$ otherwise. The standard theory holds because the region of a user can be identified without knowing if it is exposed or its outcome. Note that fitting propensity and outcomes models to all users together should give more reliable estimates than fitting propensity and outcome models within each region separately if the effect of the other features in the model is not highly correlated with region. That is, it is not necessary to estimate a separate propensity and outcome models for each region. Of course, region may also be included in the propensity and outcome models, and there should be enough users in each region to detect an effect of interest if present.

It is only slightly more difficult to estimate how Δ varies with a continuous feature of the users that can be computed without knowing the exposure status or outcome of the user. For example, suppose an advertiser wants to understand how Δ varies with the time t that a user first sees a campaign ad, and that Δ is thought to be a smooth function of t. Then define

$$\hat{\Delta}_{DR}(t) = \sum_{i=1}^{n} s_i(t)\hat{p}_i\hat{\delta}_i / \sum_{i=1}^{n} \hat{p}_i,$$

where $s_i(t_i)$ is a smoothing weight that depends on how close the first exposure time t_i for user i is to t. Because the weights no longer sum to one, the large sample standard error estimate $\hat{s}_{DR}$ has to be multiplied by $\sum_i s_i(t)\hat{p}_i / \sum_i \hat{p}_i$, but otherwise the standard theory holds.

Observational data are ubiquitous in KDD applications, and so is the need to identify important but numerically small effects for many outcomes. Examples other than measuring campaign effectiveness for advertisers are as diverse as mining health records to evaluate the effectiveness of myriad medical treatments without randomized trials and mining cellular network data to evaluate the quality delivered to subsets of users, such as those with particular mobile devices in particular regions. Each of these applications may need to run routinely, without review by a skilled data analyst who understands the dangers inherent in reasoning from observational data. This paper has shown how that goal can be realized, and suggests safeguards against hidden selection bias

10. REFERENCES

[1] *SAS Online Documentation 9.1.3*. SAS Institute, 2006.

[2] H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–972, 2005.

[3] R. Emsley, M. Lunt, A. Pickles, and G. Dunn. Implementing double-robust estimators of causal effects. *The Stata Journal*, (3):343–353, 2008.

[4] J. H. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalize models using co-ordinate descent. *Stanford University Technical Report*, 2009.

[5] G. M. Fulgoni and M. P. Morn. *How Online Advertising Works: Whither the Click?* http://www.comscore.com, 2008.

[6] G. Furnival and R. Wilson. Regression by leaps and bounds. *Technometrics*, 16:499–511, 1974.

[7] J. Hainmueller. Synthetic matching for causal effects: a multivariate reweighting method to produce balanced samples in observational studies. *Technical Report, Harvard University*, 2009.

[8] K. Hirano, G. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the

estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

[9] D. Horvitz and D. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685, 1952.

[10] J. A. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Schlkopf. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Prcessing Systems 19: Proceedings of the 2006 Conference*, pages 601–608, 2007.

[11] G. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

[12] R. J. Little and D. B. Rubin. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual Review of Public Health*, 21:121–145, 2000.

[13] J. K. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, 23:2937–2960, 2007.

[14] D. McCaffrey, G. Ridgeway, and A. Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4):403–425, 2004.

[15] G. Ridgeway and J. M. MacDonald. Doubly robust internal benchmarking and false discovery rates for detecting racial bias in police stops. *Journal of the American Statistical Association*, 104:661–668, 2009.

[16] J. Robins, A. Rotnitzky, and L. Zhao. Analysis of semiparametric regression models with missing data. *Journal of the American Statistical Association*, 90:106–121, 1994.

[17] P. Rosenbaum. *Observational Studies*. Springer, 1995.

[18] P. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[19] P. Rosenbaum and D. B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79:516–524, 1984.

[20] D. B. Rubin. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

[21] D. B. Rubin and R. Waterman. Estimating the causal effects of marketing interventions using propensity score methodology. *Statistical Science*, 21:206–222, 2006.

[22] J. Wooldridge. *Econometric Analysis of Cross-Section and Panel Data*. MIT Press, 2001.