Online Effects of Offline Ads

Diane Lambert
Google, Inc.
76 Ninth Ave
New York, NY 10011
dlambert@google.com

Daryl Pregibon
Google, Inc.
76 Ninth Ave
New York, NY 10011
daryl@google.com

ABSTRACT

We propose a methodology for assessing how ad campaigns in offline media such as print, audio and TV affect online interest in the advertiser's brand. Online interest can be measured by daily counts of the number of search queries that contain brand related keywords, by the number of visitors to the advertiser's web pages, by the number of pageviews at the advertiser's websites, or by the total duration of visits to the advertiser's website. An increase in outcomes like these in designated market areas (DMAs) where the offline ad appeared suggests heightened interest in the advertised product, as long as there would have been no such increase if the ad had not appeared. We propose a regression analysis to estimate the incremental value of the ad campaign beyond the baseline interest that would have been seen if the campaign had not been shown. A small print ad campaign illustrates the method.

General Terms

Statistical Inference, Difference-in-difference estimation, Lift, Causal Modeling, Bootstrapping

1. INTRODUCTION

Loosely speaking, an ad is effective if it generates more interest in the product advertised. But how should interest in a product be measured? Interest in low cost consumer goods that are often bought without much deliberation can often be measured by changes in sales. But sales may miss much of the effect of an ad when a product is bought infrequently and only after careful consideration. In that case, interest may mean a step towards purchase rather than a purchase itself. As people spend more time on the web, these steps toward purchase increasingly include searching for the advertiser's brand or visiting the advertiser's websites, even if the ad campaign was conducted in an offline medium such as print, radio or TV. That is, one measure of *offline* ad effectiveness is an increase in brand related *online* activity.

Still, measuring the effectiveness of offline ads is challenging. An obvious complication is that we would like to know how the ad affected the people who saw it, but it is usually impossible to know who saw an offline ad. Instead, we know only the designated market area (DMA) in which an ad appeared. A more subtle but equally thorny complication is that we would like to compare interest in a product after an ad campaign to the interest that would have been there without the ad campaign. It is only heightened interest above the baseline "that would have been there anyhow" that should be credited to the ad campaign. That is, an event that happened (*i.e.,* see the ad) has to be compared to an event that did not happen and for which there can be no data. Our goal is to provide a statistically sound method for estimating online effectiveness that circumvents both complications and can be routinely applied to data obtained by search engines or advertisers. Our approach uses notions from causal modeling, an approach which dates back at least to [5], and is sometimes dated back as far as [2].

Our approach to measuring the online effect of offline ads is based on daily online activity originating from the DMAs where the campaign appeared. Typical outcomes would include any of the following: the number of queries that contain brand related keywords, the number of visits or visitors to brand related websites, or the total duration of visits to brand related websites. To be specific, this paper takes the outcome to be number of daily visits to the advertiser's website. However outcomes are defined, though, it is convenient to consider the two potential outcomes for each DMA where the ad was shown during and after the campaign:

y_1, the daily outcome in the target DMA after the ad is shown

y_0, the daily outcome that would have been obtained in the target DMA had the ad not been shown.

The effect of the ad campaign in a DMA on a given day is then $y_1 - y_0$ since that is the difference in daily outcomes in the presence and absence of the ad campaign. But, we cannot observe these pairs because we cannot observe what would have happened if the ad had not been shown in a DMA where it was shown. Nevertheless, it is the unobservable $y_1 - y_0$ that is of interest, so we have to infer the unobservable daily outcomes y_0 from the available data.

There are two obvious strategies for estimating the unobservable y_0. First, we can assume that the past is like the present and use daily outcomes before the campaign ran. That is, we can obtain

y_b, the daily number of visits originating in the DMA *before* the ad ran.

The "before" number of visits y_b is not necessarily a good estimate of the "no ad" outcome, though, if interest in the product is expected to change over time even if no ad campaign is run. For example, if an ad is more likely to be run when interest in a product is high, then comparing counts-after to counts-before overstates the effect of the campaign. On the other hand, the advertiser may have run ads in other media during the pre-campaign period, so y_b may reflect those other ads and be higher than the "no ad" outcome y_0 would be. In that case, using y_b as a surrogate for y_0 understates the effect of the ad campaign. Any conclusion based only on the difference in means of before- and after-visits is thus suspect.

Alternatively, we could estimate the unobservable y_0 by the outcome in *control* DMAs, which are markets in which the ad did not appear. Comparing the data from control and targeted DMAs on the same days avoids the seasonality issue. Control DMAs also offer a way to adjust for other ad campaigns that ran during the one of interest. One problem, though, is that the advertiser may be more likely to advertise in DMAs where interest in the product is likely to be high. Then the level of interest in the control DMAs will be lower than y_0 and estimated effects based on the controls will be overstated. Another problem is that the increase in interest due to the ad campaign may depend on the current awareness of the product in the DMA. That is, the scale of the outcome in a control DMA may be different from that of the targeted DMA, so that the control DMA outcomes and targeted DMA outcomes have to be calibrated.

Neither before outcomes nor control outcomes alone suffice to estimate the unobservable outcomes y_0. This paper shows how they and the observable outcomes y_1 can be used together to estimate the online effect of an offline ad. The resulting estimator generalizes the difference-in-difference estimator [4] which is sometimes applied in similar settings, and it is straightforward to estimate its standard error and compute confidence intervals. The proposed estimator is simple enough to automate and apply routinely.

2. A PRINT AD CAMPAIGN

Our illustrative example relates to a print ad campaign that ran for one day in ten different newspapers in the U.S. The ad ran on one day in some newspapers and on another day in the others. For expository purposes the outcome of interest consists of daily counts of visits to the advertiser's website. These counts were obtained from Google Analytics which allows segregation by geo-region. Counts were obtained for either 28 or 35 days preceding the ad campaign and 28 days after the campaign. Target DMAs are defined by the audience of each newspaper. All other DMAs were collapsed into a single "control" DMA, and daily visits to the advertiser's website from the aggregate control DMA were obtained for the range of dates covered by the target DMA data. Here we suppress the newspaper names and simply refer to them as Paper A through Paper J.

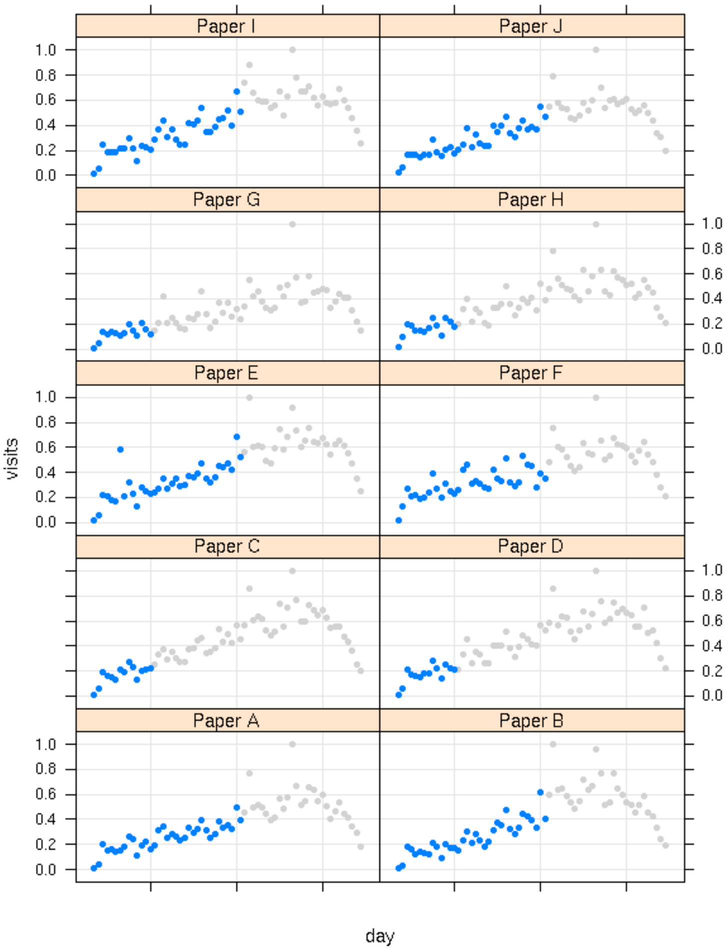

Figure 1: Daily visits originating from each target DMA before (blue points) and after (gray points) the print ad campaign. Daily visits in each target DMA are scaled by the maximum over the period so they range from 0 to 1.

Figure 1 conveys the challenge of assessing online effectiveness of print ad campaigns. Each panel of the figure shows the time series of the daily number of visits to the advertiser's web site originating from the DMAs associated with the newspapers that ran a campaign ad. The counts for each newspaper are scaled by the maximum number of visits originating from its DMA over the period. (The analysis and estimates provided use the raw counts; the data are scaled only to make the Figure panels more similar.) In each case the number of visits to the advertiser's website gradually increases over time, except for the last ten days when the visits uniformly fall. Drawing conclusions from these target DMAs alone is tricky because visits were trending up even before the print ad campaign ran and there is little to suggest from these data alone that any increase in visits to the advertiser's website is due to the ad campaign rather than seasonality. Similarly, the trend downwards may also reflect seasonality.

Figure 2 shows that the time series of daily visits in the aggregate control DMA has exactly the same seasonality as daily visits from the target DMAs, namely a steady increase in visits through time with a drop off in the final ten days. Thus, the control DMAs allow us to capture seasonality patterns that complicate the analysis of data from the target DMAs where the ad was shown.

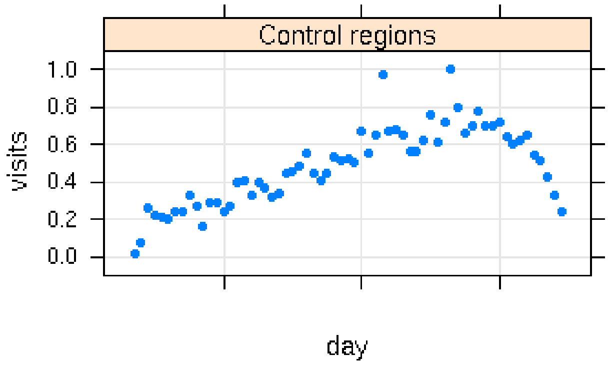

Figure 2: Daily visits from the aggregate control DMA, normalized by the maximum number of daily visits originating from the aggregate control DMA over the period.

Figure 3 shows the relationship between the daily visits to the advertiser's website from the aggregated control DMAs and those from the target DMAs. In each panel a point represents one day and shows the number of visits originating from the target DMA against the number of visits originating from the aggregated control DMAs on that day. The blue points correspond to days prior to the start of the ad campaign and the grey to days post-campaign. A strong linear relationship holds between the number of visits seen in each target DMA and the number of visits seen in the control markets before the campaign started, as shown by the least squares line fit to just the pre-campaign data in each panel. The R^2 statistics for the lines fit to the pre-campaign data range from 0.71 to 0.98, with a median of 0.94; these are exceptionally high R^2's for studies involving real data and evidence that the targeted DMAs do behave like the control DMAs in absence of the ad campaign.

Neither the target nor the control DMAs were exposed to the print ad before the campaign, so the least squares line, which fits the data well, summarizes the relationship between the target and control DMAs in the absence of a print ad campaign. This suggests that the line can be used to *predict* the daily number of post-campaign visits in each target DMA that would have occurred had there been no ad campaign (that is, the y_0 which we can never observe). In Section 3 we describe how the informal method suggested by these figures can be turned into formal estimates of offline ad effectiveness.

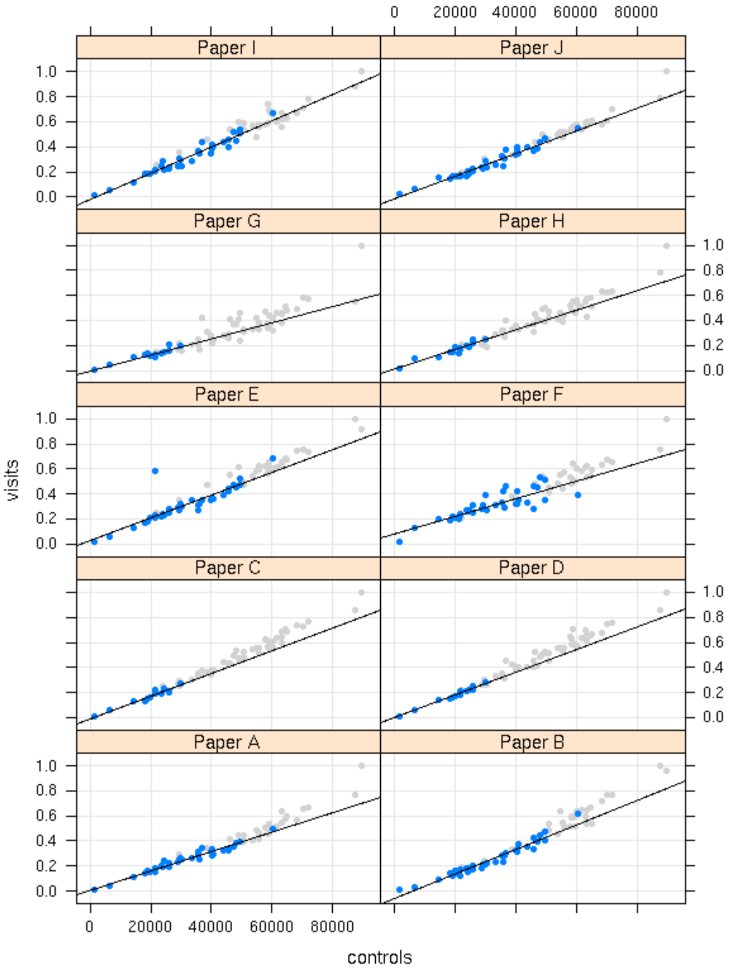

Figure 3: Daily visits for each newspaper's DMA against daily visits for the aggregated control DMA. Blue points denote pre-campaign days; gray points denote post-campaign days. A least squares line fit to only the pre-campaign days is superimposed on the data.

3. METHOD

We propose two measures of offline ad effectiveness. The first is the *incremental* effect δ or *average incremental visits* defined as the average daily difference between the number of post-campaign visits y_1 in target DMAs and the predicted number of visits y_0 that would have been observed post-campaign had the campaign not been run:

$$\delta = ave(y_1 - y_0). \qquad (1)$$

The second measure is the *relative* effect on the average number of visits over the period, sometimes called *lift* in the industry, which we denote by λ. More precisely, λ is the difference in the average number of visits with and without the campaign relative to the average number of visits without the campaign, or the average incremental effect relative to the baseline without the campaign:

$$\lambda = \frac{ave(y_1 - y_0)}{ave(y_0)} = \frac{\delta}{ave(y_0)} \qquad (2)$$

where the average is taken over the days in a fixed period after the campaign.

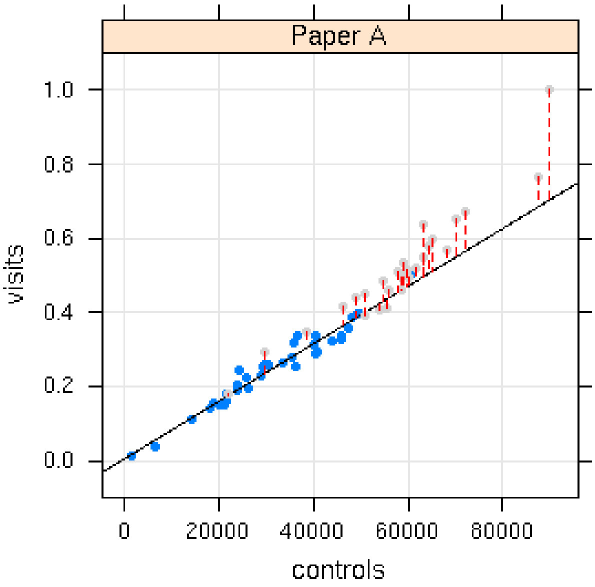

Figure 4: Illustration of the prediction formula for Paper A. The line is obtained by least squares regression using the pre-campaign data only (blue points). The dashed vertical lines connect the predictions of the post-campaign outcomes to their observed values (grey points) The average of the lengths of these vertical line segments provides the estimate of δ for Paper A.

Because y_0 is unobservable, it is estimated by $\hat{y_0}$, which is the prediction obtained from the least squares line fit to the daily pre-campaign visits of target versus aggregated control DMAs. This is illustrated in Figure 4 for a single target DMA. Substituting the prediction $\hat{y}_0$ into Equations (1) and (2) yields the estimated lift statistics:

$$\begin{aligned}\hat{\delta} &= ave(y_1 - \hat{y}_0) \\ \hat{\lambda} &= \frac{ave(y_1 - \hat{y}_0)}{ave(\hat{y}_0)}.\end{aligned}$$

Before presenting some technical details of the proposed method, we apply it to the data introduced in Section 2.

4. RESULTS

Table 1 provides the estimated incremental number of visits $\hat{\delta}$ and estimated lift in visits $\hat{\lambda}$ provided by the print ad campaign in each target market in the 28 days after the day the print ad appeared. (The standard error estimates are explained in Section 6.) For example, the DMA for Paper A enjoyed an additional 128 visits per day on average in the 28 days after the print ad appeared with an estimated standard deviation of 29 visits per day. The DMAs for the ten newspapers differ greatly in size, and the estimated incremental effects reflect these differences. The largest DMAs (that is, those with the most baseline visits) show the largest increase in number of visits. The total incremental visits per day over all ten target DMAs, which is obtained by summing across the ten target DMAs, is 925.4 with an estimated standard error of 245.9 visits.

Table 1: Estimates and bootstrapped standard errors of the incremental effect of the ad and lift.

Market	$\hat{\delta}$	stderr($\hat{\delta}$)	$\hat{\lambda}$	stderr($\hat{\lambda}$)
Paper A	128.4	29.2	0.122	0.029
Paper B	85.9	20.3	0.129	0.033
Paper C	53.9	25.2	0.100	0.051
Paper D	53.1	20.0	0.075	0.030
Paper E	459.4	215.8	0.106	0.054
Paper F	66.5	30.7	0.107	0.054
Paper G	11.0	25.3	0.032	0.077
Paper H	15.4	17.7	0.047	0.057
Paper I	19.0	13.3	0.037	0.027
Paper J	32.8	14.3	0.056	0.025

For Paper A, the additional 128 visits per day corresponds to an estimated 12% lift (relative increase over what we would have expected if the campaign had not been run). The estimated standard error of this value is 2.9% indicating that the estimated lift is statistically significant, using the usual rule of thumb that an estimate more than two standard errors from zero is statistically significant at the 5% level. The estimated lifts for the target DMAs range between 3% and 13%. They seem to fall into two groups, those slightly above 10% and those that are around 5%. The overall estimated lift, computed by dividing the total incremental visits by the total visits predicted by the model, is 9.6%.

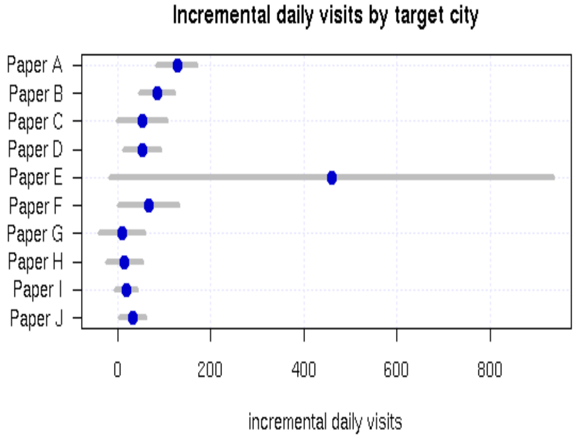

Figure 5: Estimated incremental number of visits by DMA summed over 28 days after the campaign with a two standard-error confidence interval.

Figure 5 shows the number of incremental visits with a two standard error confidence interval, $\hat{\delta} \pm 2stderr(\hat{\delta})$, averaged

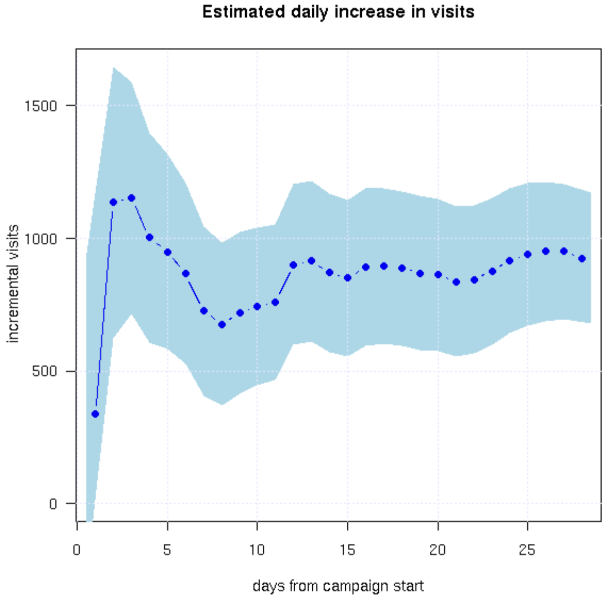

Figure 6: Estimated daily incremental visits summed over all target DMAs. The shaded region denotes a one standard error confidence band.

over the 28 days after the ad appeared for each targeted DMA. This plot highlights the differing market sizes and the uncertainty in the estimates.

Figure 6 shows the estimated incremental number of visits summed over all DMAs by day with a one standard error confidence interval.[We use a one standard error interval here to preserve the resolution that otherwise would be lost if we plotted the usual two standard error interval.] This plot shows the total effect of the ad campaign and allows us to assess how each successive day contributes to the overall estimate. There is only a modest lift on the day the ad appeared, with the biggest bumps occurring the next two days. The additional number of visitors gained on the first day an ad runs is often small because an ad cannot have an effect until it is seen, which means that additional visits are accrued over fewer than 24 hours on the first day it runs. Moreover, in this example the ad ran in the Sunday edition of each target DMA, and individuals exposed to the ad may have waited until a business day to explore the advertiser's web page. After approximately 10 days, the number of incremental daily visits stays between 900 - 950. This persistent lift is surprising and seems unlikely to hold in other campaigns.

5. THE PROPOSED ESTIMATES AND THEIR RELATIONSHIP TO OTHER METHODS

As suggested in Section 2, a least squares line fit to the pre-campaign visits for a targeted DMA and the control DMA can be used to estimate the number of visits y_0 that would have been observed in the post-campaign period had the ad not run in the targeted DMA. A little algebra shows that the least squares estimate of y_0 on a post-campaign day for which the control DMA saw x visits is

$$\hat{y}_0(x) = \bar{y}_b + \hat{\beta}(x - \bar{x}_b)$$

where $\bar{y}_b$ denotes the average daily number of visits in the targeted DMA prior to the campaign, $\bar{x}_b$ denotes the average daily number of visits in the control DMAs prior to the campaign, x is the number of daily visits to the control DMA on the day of interest after the campaign, and $\hat{\beta}$ is the least squares slope estimate.

The estimated effect $\hat{\delta}$ of the ad campaign in a DMA is then computed by averaging the daily differences $y_1 - \hat{y}_0$ of the observed-after-campaign and estimated-assuming-no-campaign number of visits in the targeted DMA. The algebraic form of $\hat{\delta}$ is

$$\begin{aligned} \hat{\delta} &= ave(y_1 - \hat{y}_0) \\ &= ave(y_1) - ave(\hat{y}_0) \\ &= \bar{y}_1 - (\bar{y}_b - \hat{\beta}(\bar{x}_a - \bar{x}_b)) \\ &= (\bar{y}_1 - \bar{y}_b) - \hat{\beta}(\bar{x}_a - \bar{x}_b) \end{aligned} \tag{3}$$

where $\bar{x}_a$ is the average number of visits per day in the control DMA in the period after the campaign ran in the targeted DMA.

Equation 3 for $\hat{\delta}$ is identical to the so-called difference-in-difference estimate [4] when $\hat{\beta} = 1$, namely

$$\widehat{DnD} = (\bar{y}_1 - \bar{y}_b) - (\bar{x}_a - \bar{x}_b).$$

The difference-in-difference estimate $\widehat{DnD}$ compares the post-campaign change in visits in the targeted DMAs to the post-campaign change in visits in the aggregated control DMAs. In our application, the least squares slope estimate $\hat{\beta}$ is necessary to scale the controls to the size of the targeted DMA. The fact that the proposed estimate reduces to the difference-in-difference estimate when $\hat{\beta} = 1$ shows that exploiting the relationship between targeted and non-targeted DMAs prior to the print ad campaign is consistent with existing practice. Not requiring that the targeted and control DMAs have matching number of visits before the campaign makes it easier to routinely apply the proposed estimator. It would be troublesome to find a control DMA with approximately the same number of pre-campaign visits for each target DMA; *i.e.*, to find controls for which $\beta = 1$. Equation (3) shows that this degree of matching is not necessary. Simply aggregate traffic from all non-targeted DMAs and use least squares to handle the attenuation factor. Note that not requiring $\beta = 1$ allows the control DMA to be much larger than the targeted DMA. The number of visits may be more stable in larger DMAs, providing more stable estimates of β.

Recently Lee [3] proposed a pair of generalizations of the difference-in-difference estimate:

$$\begin{aligned} GD_\gamma &= (\bar{y}_1 - \bar{y}_b) - \gamma(\bar{x}_a - \bar{x}_b) \\ DG_\eta &= (\bar{y}_1 - \eta\bar{y}_b) - (\bar{x}_a - \eta\bar{x}_b). \end{aligned}$$

We see that the first of these has exactly the same form as our estimator defined in Equation 3. However Lee formulated the problem differently and did not see a way to use data to estimate γ. Instead, Lee focused attention on DG_η

and proposed an estimate of η based on an analysis of panel data.

Equation (3) also shows what happens if there is no statistically significant relationship between the outcomes in the targeted DMAs and the control DMA. In this case, the slope estimate is 0 and $\hat{\delta}$ is the standard difference of the before and after averages. That is, if there is no predictive power in the non-target markets, then the best estimate of post-campaign average daily traffic is the pre-campaign daily average. More formally, a statistical hypothesis test of $H_0 : \beta = 0$ can be employed to determine when the control market does not provide enough information about the targeted DMAs pre-campaign to be used to estimate the 'without campaign' outcome y_0.

6. STANDARD ERROR ESTIMATES

Point estimates such as $\hat{\delta}$ and $\hat{\lambda}$ are only meaningful if presented with estimates of their sampling variability. In this section we derive an explicit formulas for the variance of $\hat{\delta}$ and $\hat{\lambda}$. As a check on the accuracy of these formulas, especially for $\hat{\lambda}$, we propose an alternative approach based on re-sampling methods.

6.1 Standard Error of Incremental Visits

Equation (3) leads to a simple estimate of the variance of $\hat{\delta}$. From the theory of least squares regression, the variance of the estimated slope $\hat{\beta}$ in the linear regression of pre-campaign DMA daily visits against pre-campaign control daily visits is

$$var(\hat{\beta}) = \frac{\sigma^2}{(n_b - 1)s_b^2},$$

where σ^2 is the residual variance of the pre-campaign visits y_b in a targeted DMA given that there were x visits in the control DMA that day, n_b is the number of days included in the pre-campaign period, and s_b is the standard deviation across days of the daily number of visits in the control market in the pre-campaign period. By the assumptions that underlie linear regression modeling, the residual variance σ^2 is the same for all levels x_b. Moreover, $\hat{\beta}$ is independent of $\bar{y}_b$, and the post-campaign data are independent of the pre-campaign data. Putting all this together, conditional on the outcomes in the control DMA,

$$var(\hat{\delta}) = \sigma^2 \left(\frac{1}{n_a} + \frac{1}{n_b} + \frac{(\bar{x}_a - \bar{x}_b)^2}{(n_b - 1)s_b^2} \right) \qquad (4)$$

where n_a is the number of days in the study after the ad campaign ran. Note that all quantities in this equation can be computed from the data with the exception of σ^2. However the residual mean squared error from the linear regression model provides a convenient and accurate estimate of σ^2.

Table 2 gives the estimated standard errors of $\hat{\delta}$ for each target market using equation (4). We see good agreement with the bootstrap standard error estimates for $\hat{\delta}$ (which we explain in Section 6.3) that are reported in Table 1.

6.2 Standard Error of Lift

Lift is the ratio of the estimate of incremental visits and the average daily visits that we predict we would have observed had the campaign not been run. Thus it is the ratio of two random variables and as such, a simple and accurate variance estimate is problematic. One standard approach is to use a "first order" approximation based on a Taylor series expansion of the ratio of random variables around their means.

Table 2: Estimates of the incremental number of visits δ and their standard errors obtained from equation (4).

Market	$\hat{\delta}$	stderr($\hat{\delta}$)	$\hat{\lambda}$	stderr($\hat{\lambda}$)
Paper A	128.4	21.3	0.122	0.029
Paper B	85.9	18.0	0.129	0.039
Paper C	53.9	26.3	0.100	0.062
Paper D	53.1	19.1	0.075	0.033
Paper E	459.4	237.5	0.106	0.079
Paper F	66.5	31.9	0.107	0.074
Paper G	11.0	24.0	0.032	0.084
Paper H	15.4	18.8	0.047	0.070
Paper I	19.0	11.5	0.037	0.031
Paper J	32.8	13.6	0.056	0.032

To proceed, note that we can write $\hat{\lambda} = ave(y_1)/ave(\hat{y}_0) - 1$ and that $ave(y_1)$ and $ave(\hat{y}_0)$ are statistically independent. Using this fact, the approximate variance of $\hat{\lambda}$ as

$$var(\hat{\lambda}) \approx \left(\frac{ave(y_1)}{ave(\hat{y}_0)} \right)^2 \left[\frac{var(ave(y_1))}{ave^2(y_1)} + \frac{var(ave(\hat{y}_0))}{ave^2(\hat{y}_0)} \right] \qquad (5)$$

where the variance terms in the brackets are

$$\begin{aligned} var(ave(y_1)) &= \frac{\sigma^2}{n_a} \\ var(ave(\hat{y}_0)) &= \sigma^2 \left(\frac{1}{n_b} + \frac{(\bar{x}_a - \bar{x}_b)^2}{(n_b - 1)s_b^2} \right) \end{aligned}$$

Table 2 gives the estimated standard errors of $\hat{\lambda}$ for each target market using equation (5). In contrast to those for $\hat{\delta}$ we see that these are much larger than the bootstrap standard error estimates for $\hat{\lambda}$ that are reported in Table 1.

6.3 Bootstrap Confidence Intervals

The estimated standard error for $\hat{\delta}$ is easy to compute and accurate, while the estimated standard error for the lift estimate $\hat{\lambda}$ is problematic. To obtain a reliable and honest estimate of variability for $\hat{\lambda}$, we recommend a re-sampling technique known as the bootstrap [1]. The *bootstrap* method of estimating the sampling variability of estimates has a rich history in statistics and is often used in cases such as ours where a reliable and honest method of estimating a standard error is required but there is no simple formula for the standard error. The idea is to generate data sets that resemble the one under study by randomly sampling the original data, or some combination of the original data and a model for the data, with replacement. The lift estimate $\hat{\lambda}$ is then computed for each generated data set, and confidence intervals can be formed by

- computing the sample standard deviation of these values (s_λ) and use this to define the limits of a 95% confidence interval as $(\hat{\lambda} - 2s_\lambda, \hat{\lambda} + 2s_\lambda)$;
- computing the quantiles of the bootstrap estimates and define the limits of a 95% confidence interval as

$$(\hat{\lambda}_{.025}, \hat{\lambda}_{.975}).$$

In cases where the bootstrap distribution is unimodal and symmetric, these two procedures lead to approximately the same confidence intervals. We later show that this is the case for our estimation problem.

In simple one-sample problems, bootstrapping proceeds by taking random samples of size n with replacement from the original n observations. In linear regression problems with a model that fits the data well, it is common to re-sample residuals from the fitted model instead of the raw observations themselves. The re-sampled residuals are then added to the n fitted values from the original data to generate new data, the model is fit to the newly generated data, and then $\hat{\lambda}$ is estimated from the re-fit model.

In the context of lift, we find the following formulation convenient. First, write

$$y = (1 - z)(\alpha_b + \beta_b x) + z(\alpha_a + \beta_a x) + e \qquad (6)$$

where z is a binary indicator denoting whether the outcome was recorded in the before ($z = 0$) or after ($z = 1$) period, and e denotes Gaussian noise with mean 0 and residual variance σ^2. This model specifies a different intercept and slope for the two periods and can be fit with any standard statistical analysis system. If we fit this model to our data, we derive fitted values $\hat{y}$ and residuals r, which can be split into two groups:

before: $\{\hat{y}_i : z_i = 0\}$ and $\{r_i : z_i = 0\}$

after: $\{\hat{y}_i : z_i = 1\}$ and $\{r_i : z_i = 1\}$.

Bootstrapping involves sampling from the before- and after- distributions of the residuals, and adding these to their respective fitted values. Thus, bootstrap values for the before- and after- periods can be written as

before: $y_b^* = \hat{y}_b +$ random draw from $\{r_i : z_i = 0\}$

after: $y_a^* = \hat{y}_a +$ random draw from $\{r_i : z_i = 1\}$

Drawing separately from the before- and after- periods allows the residual variance of the outcome to be different in the two periods. When there is limited data, we would sacrifice this generality and randomly draw from the combined residuals and fitted values to avoid sampling from a small set. For the example data in this paper, we have either 14 or 35 observations in the before period and 28 observations in the after- period, so that we feel comfortable with separate draws.

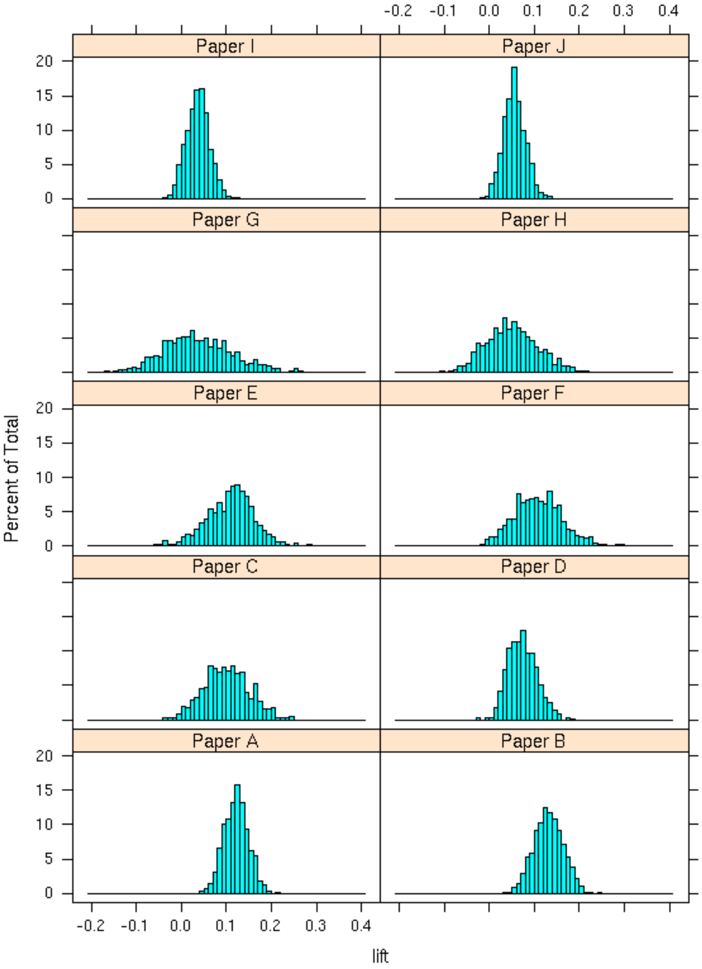

Figure 7: Histograms of bootstrap estimates of λ. 1000 samples (with replacement) were drawn from the distribution of before- and after- residuals.

Figure 7 is based on 1000 bootstrap samples derived from the data from Section 2 and the generative model described by Equation (6). The bootstrap distributions of $\hat{\lambda}$ are unimodal and reasonably symmetric. In this case, confidence intervals can be constructed using the standard deviation of the bootstrap samples of $\hat{\lambda}$.

Since bootstrapping applies to any statistic computed from the data, we can compute bootstrap confidence intervals for both $\hat{\delta}$ and $\hat{\lambda}$. We reported these values in Table 1. We prefer these, especially for $\hat{\lambda}$, since those based on the first order approximation are often overly conservative. Indeed in the present study, several DMAs change from having significant lift (using the bootstrap method) to insignificant lift (using the first order approximation).

7. SUMMARY

This paper proposes two metrics for evaluating the effectiveness of ad campaigns in traditional media like print, radio, and TV:

δ: the incremental online activity that can be attributed to the ad campaign, and

λ: the increase in online activity relative to the baseline activity that would be expected without the campaign.

Both these parameters require estimating what would have happened if the ad campaign had not run in the targeted markets. We have provided simple estimates of the online effectiveness metrics that allow us to estimate these counterfactual outcomes, along with ways to compute confidence intervals for the estimated effectiveness metrics. These estimates accommodate strong seasonal effects, like those seen in the illustrative example in this paper. We propose that the generalized difference-in-difference method be used to quantify effects in offline ad campaigns in general, and print ad campaigns in particular. It is intuitive and relatively easy to compute using off-the-shelf software.

Our method is premised on certain data requirements and underlying assumptions. We require that online activity data be available 2 to 4 weeks *before* the start of an offline ad campaign. We also require that data be collected outside the campaign markets, (i.e., not just in the targeted DMAs). Both of these data sources are necessary in order to predict what we would observe had the campaign not been run.

Since the slope estimate accounts for any differences in the units between the outcome and predictor variable, we do not require that the "out of market" control match the targeted DMA according to demographics or online interest in the product before the campaign. Moreover, the predictor variable (*.e.g,* number of visits to the advertiser's website) measured in the out-of-market control areas need not be the same as the outcome used to evaluate the effectiveness of the campaign. It is convenient to do so, but the only requirement is that there is a strong linear relationship between the in-market outcome and the out-of-market predictor variable. In some cases, it might be necessary to build more complicated prediction models to get a model that fits the pre-campaign data well. In that case, standard errors and confidence intervals can still be obtained by bootstrapping, although the closed form formula for the standard error of $\hat{\delta}$ given in Section 6 may no longer apply.

Fitting a separate linear regression model in each targeted DMA allows for an interaction between seasonal effects and in-market and out-of-market outcomes, which means that it is not necessary to assume that the same seasonality pattern applies in all targeted DMAs. For example, suppose an ad campaign for "Tom's Tanning Booth" runs in both Seattle and Miami and the weather changes over the study period in Seattle but not in Miami. Then the seasonality will be expected to be different in the two markets. Fitting a separate regression model in each targeted DMA also allows the product awareness before the ad campaign to vary substantially across DMAs. Otherwise, the residual variance σ^2 around the regression line may not be the same in all target DMAs. However if all the target DMAs have the same seasonality patterns and market awareness, it is reasonable to aggregate data across all targeted DMAs and construct a single least squares model for predicting what would have occurred had the campaign not been run. Our preference is to build separate models in each targeted DMA, not only for robustness to this strong assumption of no interaction, but also to simplify computations in situations where the campaigns do not all start and stop on the same day.

Finally, ad campaigns are often more complicated than the one presented here. There may be two different versions of an ad to compare. In a cross-over design, each targeted DMA may see both ads, but half the targeted DMAs will see ad A first and the remaining will see ad B first. Or, the advertiser may want to decide when to stop running an ad based on an ongoing analysis of its effectiveness. The simple estimation method proposed here would need to be extended to handle these situations, but the twin principles of comparing what happened after a campaign to an estimate of what would have happened had the campaign not run and basing the estimate on a control DMA consisting of markets where the campaign did not run should still apply.

8. REFERENCES

[1] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap.* CRC Press, 1993.

[2] R. Fisher. *The Design of Experiments.* Hafner Publishing Company, 1935.

[3] M. jae Lee. Difference in generalized-differences with panel data: Effects of moving from private to public school on test scores. *Discussion Paper Series: The Institute of Economic Research, Korea University*, (07-21):1–30, 2007.

[4] B. Meyer. Natural and quasi-natural experiments in economics. *Journal of Business and Economic Statistics*, XII:151–162, 1995.

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