Methods for Measuring Brand Lift of Online Ads

Rachel Fan Google
Tim Hesterberg Google
Ying Liu Google
Lu Zhang Google

Abstract
We describe ways to measure ad effectiveness for brand advertisements using online surveys. We estimate the causal effect of ads using randomized experiments. We focus on some technical issues that arise with imperfect A/B experiments–corrections for solicitation and response bias in surveys, discrepancies between intended and actual treatment, and comparing treatment group users who took an action with control users who might have acted. We discuss different methods for estimating lift for different slices of the population, to achieve different goals. We use regression, with a particular form of regularization that is particularly suited to this application. We bootstrap to obtain standard errors, and compare bootstrap methods.

Key Words: A/B experiment, brand lift, imperfect control, bootstrap

1. Introduction

In this paper we discuss issues that arise in causal modeling to estimate the effect of brand advertisements, using A/B experiments and surveys. Past research to measure survey based brand metrics includes (Chan et al., 2010; Hanna et al., 2011).

The population for a particular campaign consists of users who would normally see a campaign ad, if we were not running an experiment. Users are randomly split into two groups (*arms*), a treatment group who see the campaign ads as normal, and a control group for whom the campaign ad is held back; they may see a different ad, or no ad. Treatment users who see ads and controls who would have seen one are flagged as eligible to be surveyed. Some of them later visit sites where they can be surveyed, some are surveyed, and some of those respond.

We compare the survey responses from the two arms to measure the campaign's effect. Outcomes are binary, e.g. whether a user is familiar with a brand or not. We are interested in "brand lift", the difference in proportion of favorable outcomes between users who see or don't see ads, for the whole campaign and for gender, age, and other subpopulations.

The heart of this paper is a discussion of a number of practical issues that arise in this setting, including:

- Correcting for differences in covariates, either random differences (with good A/B experiments) or more systematic differences due to bias.
- Correcting for solicitation and response bias.
- For all corrections, we find regression to be more stable than propensity modeling.
- Regularization, using a particular form that combines advantages of L1 and L2 regularization.
- "Slice and dice", estimation for subpopulations. We discuss a number of methods, that are suitable for different purposes. Some methods provide estimates of incremental effect of covariates like age and gender that may be correlated, others provide non-incremental estimates. Some require that relevant covariates be included in the correction model, others do not.

- Discrepancies between intended and actual treatment. We use an intent-to-treat approach, but discuss issues that arise in this context. We note that it is easy to calculate standard errors incorrectly.
- Standard errors and confidence intervals. We use t intervals based on bootstrap standard errors. Not all bootstrap methods suitable in other contexts are suitable here.

Some of these are specific to brand lift, but most also occur in other causal modeling settings. We finish with a case study.

2. Estimation Methodology

Our goal is to estimate lift. In this section we discuss estimates for the whole population, beginning with estimates for respondents in Sections 2.1 and 2.2, then solicitation and response bias in Section 2.3. We defer estimation for subgroups to Section 2.3.1, and for mismatched comparisons (comparing a subset of the treatment group to all controls) to Section 4.

2.1 Difference in Proportions

Additive lift is $\Delta = p_{\text{treatment}} - p_{\text{baseline}}$, where $p_{\text{treatment}}$ is the fraction of people who see a relevant ad who would respond positively to a survey if they were asked, and p_{baseline} is the fraction of those same people who would respond positively, had they not seen the ad. Let

$$y = \begin{cases} 1 & \text{for a positive survey response (e.g. is aware of the brand)} \\ 0 & \text{otherwise} \end{cases}$$

A simple estimate of additive lift is the difference in positive response rates between treatment and control:

$$\hat{\Delta} = \bar{y}_{\text{treatment}} - \bar{y}_{\text{control}}.$$

We use $\bar{y}_{\text{treatment}}$ instead of $\hat{p}_{\text{treatment,respondents}}$ for simplicity and because variations of this notation are more useful later; similarly for controls.

This estimate may be biased due to response and solicitation bias (see Section 2.3), or if people aren't assigned to arms randomly.

2.2 Adjusting for Differences in Covariates

We can improve on this estimate. We have additional signals — age, gender, etc., and can improve the estimates by adjusting for differences between these covariates in the treatment and control groups. For example, if age is positively correlated with y, and the treatment group has older users, the simple estimate would tend to overestimate Δ. By adjusting for such differences, we can mitigate assignment, solicitation and response bias, and reduce variability due to random imbalances.

We use logistic regression to correct for differences in covariates. We fit a logistic regression of the form

$$\text{logit}(E(Y)) = \beta_0 + \beta_1 Z + \beta_2 X + \beta_3 ZX \qquad (1)$$

where

- Z is a dummy variable (1 for treatment group, 0 for controls) and the β_1 term measures the overall effect of the ad

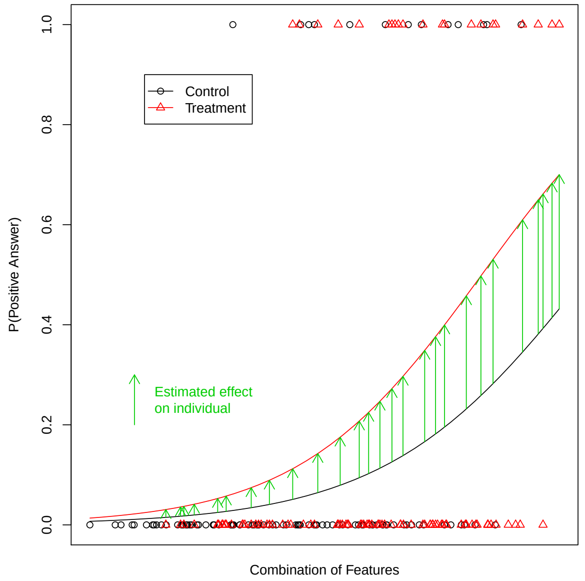

Figure 1: Estimate Δ as average of difference between treatment and control predictions. Arrows are shown for a subset of users, indicating the estimated effect of an ad on each user.

- X is a vector of additional signals such as age, gender, etc, and β_2 is a vector of corresponding regression coefficients that measure differences in baseline between groups, e.g. differences in how likely males and females are to answer positively without having seen an ad, and
- β_3 is a vector of coefficients corresponding to interactions between Z and X, that measures how different groups are affected by ads.

We use the estimated coefficients to produce two predictions for each person: (1) the treatment prediction $\hat{y}_1$, obtained by setting $Z = 1$, and (2) the control prediction $\hat{y}_0$, obtained by setting $Z = 0$. For any individual, the estimated effect of the ad for that person is the difference between these predictions. One estimate for each person is a counterfactual prediction — for controls $\hat{y}_1$ is the estimated probability of a positive response had they seen an ad, and for treated $\hat{y}_0$ is the estimated probability of a positive response had they not seen an ad. This is shown in Figure 1.

The estimated overall lift is the average of the individual differences,

$$\hat{\Delta} = \text{average}(\hat{y}_1 - \hat{y}_0). \tag{2}$$

In a non-randomized setting, this procedure generally reduces the bias of estimates. In a randomized setting where there is no bias, it reduces the asymptotic variance of estimates, compared to the simple estimate; the higher the multiple correlation between y and the covariates, the greater the improvement.

The model given in Equation 1 is roughly equivalent to fitting models of the form

$$\text{logit}(E(Y)) = \beta_0 + \beta_2 X \tag{3}$$

separately to treatment and control users. However, we use regularized regression (Section 2.4 below), that makes coefficients more similar between the two arms, so the combined model gives slightly different results than separate models would.

2.2.1 Average Over Both Arms, or Just Treatment Group

The average in Equation 2 can be taken over all respondents, or just over the treatment arm. For a good randomized experiment, averaging over all respondents is appropriate (Bloniarz et al., 2016), with lower variance. If there is no random assignment, and controls are chosen from some population to try to be like the exposed, then averaging over the treatment group (who saw ads for this campaign) is appropriate, to estimate the effect on those who actually saw ads.

When averaging only over the treatment group, and when the Z term is included in the model without regularization, we have

$$\hat{\Delta} = \text{average}_{\text{treatment}}(\hat{y}_1 - \hat{y}_0) = \bar{y}_{\text{treatment}} - \text{average}_{\text{treatment}}(\hat{y}_0). \tag{4}$$

In other words, the estimate is the average actual response for the treatment group, minus an estimated *baseline* — what the average response would have been for that group, had they not seen ads.

2.2.2 Linear or Poisson Regression

For non-binary response variable, linear or Poisson regression may be used in place of logistic regression. The approach is similar, except that the left side of equations 1 and 3 would be $E(Y)$ or $\log(E(Y))$, respectively.

2.2.3 Regression vs. Propensity Methods

In work with observational data (non-randomized) with substantial differences between the treatment and control groups, we tried propensity methods (Chan et al., 2010), but ultimately abandoned them; they were too unstable. Propensity methods result in estimates of the form

$$\hat{\Delta} = \bar{y}_{\text{treatment}} - \sum_{\text{control}} w_i y_i, \tag{5}$$

a weighted average with weights that depend on the propensities, and high variation in the weights results in unstable estimates. A common workaround is to truncate the weights, but that results in bias.

For comparison, regression estimates can also be written as weighted averages of the y values, with weights depending on the x values (this is true exactly for linear regression without regularization, and approximately for logistic regression), and these weights give smaller variance for Δ estimates.

2.3 Solicitation and Response Bias

The people who respond to surveys are a subset of the people who see campaign ads. Some people who see ads never visit sites where we can solicit them for surveys, and some people surveyed do not respond. When estimating Δ for the campaign audience we encounter solicitation bias (differences between those we can solicit, and the whole campaign audience) and response bias (differences between those who respond, and those who do not).

We use a similar logistic regression to reduce solicitation and response biases. We use signals such as age, gender, number of ad impressions for this campaign, and the number of ads people see for other campaigns (this serves as a proxy for how active people are on the internet — more active people are more likely to be aware of a typical brand, even without seeing an ad for the brand).

When extrapolating, we can only correct for differences in signals that we observe. Hence, Δ estimates for the solicited non-respondents and the larger population may have some bias. Still, the regression does allow us to correct for some important differences between the respondents and the campaign audience.

Some signals are available for everyone, like demographics and ad impressions. Other signals are only available for some groups: solicitation gap (delay between ad impression and survey solicitation) is only available for those solicited, and response speed (how quickly someone answers a survey) is only available for those who respond. For later use, let X_A correspond to variables that are available for All, X_S correspond to the (larger) set of variables that are available for Solicited (those solicited for a survey), and X_R correspond to the (largest) set of variables that are available for Respondents.

To make best use of available information, we use a two-step process for extrapolating from respondents to the whole population.

2.3.1 Extrapolation Step 1 to Solicited: Correct for Response Bias

Respondents may differ from solicited non-respondents in two ways:

- Differences that affect both arms, e.g. some age groups are more likely to respond to a survey;
- Differences specific to the treatment group, e.g. users who see the campaign ad may become interested in the sponsor's product, and hence more likely to take the time to respond to the survey. (There do not appear to be many such users; response rates are actually slightly higher for controls than for treatment in recent data.)

We focus on the first category of differences, including differences in demographics, internet activity, or other signals between respondents and non-respondents. These differences matter if they relate to differences in lift, e.g. if one age group is a higher proportion of respondents and that age group also has more or less lift than other ages.

We use a logistic regression as before, but now using only X_S — covariates that are present for both the respondents "R" and solicited non-respondents "SNR" groups. We fit the model to the R data (y is present only for this data), then produce treatment and control predictions for the treatment group, for both R and SNR. Then, for each person, we use the best prediction — for respondents using the earlier model based on X_R, and for SNR using the second model based on X_S:

$$\hat{y}_k = \begin{cases} \hat{y}_{k,X_R} & \text{Respondents} \\ \hat{y}_{k,X_S} & \text{SNR} \end{cases} \tag{6}$$

for $k = 0, 1$ (control, treatment).

The lift estimate is the average over both R and SNR of those predictions

$$\hat{\Delta} = \text{average}(\hat{y}_1 - \hat{y}_0). \tag{7}$$

As noted in Section 2.2.1, the average may be over both arms, or just the treatment group.

2.3.2 Extrapolation Step 2 to All: Correct for Solicitation Bias

The ultimate goal is to estimate lift for the campaign audience, including respondents "R", solicited non-respondents "SNR", and non-solicited "NS"; let "All" denote the union. Here we have two choices. One approach is to extrapolate directly from R to All, using the same procedure described above for extrapolating from R to Solicited, except using X_A in place of X_S.

Instead, we extrapolate from Solicited to All. This presents an obvious difficulty — y is unknown for the SNR. Our solution is to use predictions from step 1. We build a logistic regression model, on all solicited users, similar to step 1, but

- We only use X_A, covariates that are available for All — excluding terms like solicitation gap that are nonexistent for those who were never solicited.
- The response variable y' is a composite. For R, we use the actual y values. For SNR, we use the predictions from step 1.

$$y' = \begin{cases} y & \text{Respondents} \\ \hat{y}_{k,X_S} & \text{SNR} \end{cases} \tag{8}$$

After fitting that model, we use the best prediction for each person:

$$\hat{y}_k = \begin{cases} \hat{y}_{k,X_R} & \text{Respondents} \\ \hat{y}_{k,X_S} & \text{SNR} \\ \hat{y}_{k,X_A} & \text{NS.} \end{cases} \tag{9}$$

As in Equations 2 and 7, the lift estimate is

$$\hat{\Delta} = \text{average}(\hat{y}_1 - \hat{y}_0)$$

where now the average is over All users (either both arms, or just treatment, see Section 2.2.1).

The three models are summarized in this table:

	Training set	Variables	Prediction	Predict for
Respondents	R	X_R	$\hat{y}_{k,X_R}$	R
Extrapolation 1	R	X_S	$\hat{y}_{k,X_S}$	SNR
Extrapolation 2	S	X_A	$\hat{y}_{k,X_A}$	NS

This two-step extrapolation is motivated by multiple imputation for missing data (Schafer, 1997). When one or more variables are missing, their estimated conditional distribution given other variables is used to generate random values for the variables. This is true regardless of whether the missing variables include y or x's. We take two shortcuts. First, rather than multiple imputation, we do single imputation using the predicted value of y rather than its distribution. We do not need multiple imputation because (1) the missingness pattern is monotone (all variables are known for R; a subset are known for SNR; and a subset of those are known for NS) and (2) we estimate variability another way, by bootstrapping. Second, we only impute y, not missing x's, and do later regressions using a smaller set of x's. We would gain no information by imputing missing x's from present x's and then using both the present and imputed x's in a regression. Note that if seeing ads affects whether people respond, e.g. if people who saw an ad and therefore recognize a brand are more likely to respond, that makes the data MNAR (missing not at random) and causes bias we cannot address.

Here is a toy example that illustrates the difference between direct and two-step extrapolation. Suppose there is a covariate X with two values, and $E(\hat{\Delta}|X = 0) = 0.04$ and $E(\hat{\Delta}|X = 1) = 0.06$. Suppose that only 10% of the respondents have X=1; then

$$E_R(\hat{\Delta}) = 0.9 \cdot 0.04 + 0.1 \cdot 0.06 = 0.042.$$

Suppose that 40% of the SNR have X=1; then the model in step 1 gives

$$E_{SNR}(\hat{\Delta}) = 0.6 \cdot 0.04 + 0.4 \cdot 0.06 = 0.048.$$

Suppose that X is not present for NS. Then, extrapolating directly from R to All uses no variables, and the prediction for all non-respondents would be 0.042; it ignores the better prediction (0.048) that is available for the non-respondents who were solicited. Direct extrapolation combines 0.042 for R, 0.042 for SNR, 0.042 for NS, overall $\hat{\Delta} = 0.042$. In contrast, the two-step procedure uses 0.048 for the SNR, and the average of R and SNR when extrapolating to NS, e.g 0.045 if there are equally many R and SNR. The resulting estimates are 0.042 for R, 0.048 for SNR, 0.045 for NS, and overall $\hat{\Delta} = 0.045$.

2.4 Regularization

We use regularized logistic regression that shrinks regression coefficients (other than the intercept) toward zero. For each column in the design matrix we add a dummy observation with $y = 0.5$, weight 4, and an x value that we discuss below. Greenland and Mansournia (2015) suggest a similar approach.

The dummy observation approach is equivalent to a penalty that is roughly quadratic for coefficients near zero, and approaches linear for larger coefficients. In general, adding p dummy observations with $y = 0.5$, $x_j = v_j$, $x_k = 0 \forall k \neq j$ and weight λ_j for $j = 1, \ldots p$ corresponds to maximizing

$$\log(\text{likelihood}) - \sum_{j=1}^{p} \lambda_j \log(2 * \cosh(v_j \beta_j / 2)). \tag{10}$$

We prefer this to both L1 and L2 regularization. L1 regularization is not smooth, sampling distributions are a mix of continuous and discrete distributions (positive probability of being exactly zero) and standard errors are unusable for coefficients at or near zero, so confidence intervals are difficult. Slice estimates (see below) may not be reasonable. L2 regularization gives poor predictions due to large penalties on the coefficients for important variables (Hesterberg et al., 2008). Elastic net, a sum of L1 and L2 penalties, combines the worst features of each — non-smooth, and quadratically increasing penalties for large coefficients.

There are two additional differences between regularization commonly found in software and the literature, and our procedures. First (and somewhat trivial), we scale penalties rather than variables. That is, the common procedure is to divide each predictor by its standard deviation, use the modified predictors $x'_j = x_j / s_j$ in the model to estimate coefficients $\hat{\beta}'_j$, and regularize using those scaled coefficients. Instead, we rewrite the linear combination in the model as $\sum \beta'_j x'_j = \sum \beta_j x_j$ with $\beta_j = \beta'_j / s_j$, and rewrite the regularization penalty, e.g. $\lambda \sum |\hat{\beta}'_j| = \lambda \sum s_j |\hat{\beta}_j|$ for an L1 penalty.

Second, we do not base the scaling terms s_j on standard deviation. For typical dummy variables, binary variables that take on values 0 and 1, we do not scale at all. Other variables are penalized as if they were scaled to have the same standard deviation as a dummy variable with the same skewness. Compared to the common approach of standardizing variables by dividing by standard deviations, this procedure causes more shrinkage for coefficients for variables with less information — those corresponding to rare factor levels or interactions, and skewed variables where more of the variation is contained in a small number of observations.

For a binary variable x_j, our default parameters correspond to adding four dummy observations with $x_j = 1$, $x_k = 0$ for $k \neq j$, and $y = 0.5$. This is motivated by the $(X+2)/(n+4)$ approximation to the Wilson confidence interval for binomial proportions.

3. Slice and Dice

It is often important to estimate lift for different subgroups, such as age, gender, and groups based on the number of campaign impressions. This is useful for advertisers planning future campaigns. We call this slice and dice.

We describe four approaches to calculating slice and dice lift results. The first method is a raw comparison — Δ for a particular slice is estimated by the difference in y between treatment and control users in that slice, e.g. treatment males to control males:

$$\hat{\Delta}_{\text{raw,male}} = \bar{y}_{T,\mathrm{male}} - \bar{y}_{C,\mathrm{male}}. \tag{11}$$

This does not adjust for any differences in control variates. We do not use this estimate.

3.1 Subset Method for Slice and Dice

The subset method is similar, but compares predictions:

$$\hat{\Delta}_{\text{subset,male}} = \mathrm{average}_{\mathrm{male}}(\hat{y}_1 - \hat{y}_0). \tag{12}$$

The average can be taken across both arms, or just the treatment group. In the latter case, a minor variation is

$$\hat{\Delta}_{\text{subset,male}} = \bar{y}_{T,\mathrm{male}} - \mathrm{average}_{\mathrm{male}}(\hat{y}_0). \tag{13}$$

The two estimates (12) and (13) are equivalent if the $Z \cdot \mathrm{gender}$ interaction is not regularized.

For continuous covariates, we use the variable or a transformation in the model as normal, then produce slice estimate by bucketing the variable, then averaging across observations within a bucket.

The subset method requires that both the variable (e.g. age) and its interaction with Z be included in the logistic regression model. Recall that main effects in the model measure baseline characteristics of the person and interactions measure how the users are affected by the ad. If the interaction is omitted from the model, then differences between slice estimates for different groups will be practically zero.

3.2 Residual Method for Slice and Dice

The residual method is similar, but is used for a variable when the interaction of the variable with Z is not in the model; the main effect may be included or not. In this case, we do not expect the model to give good predictions for a subgroup, so we use residuals from the model to make up for the model's shortcomings. The estimate is

$$\hat{\Delta}_{\text{residual,male}} = \mathrm{average}_{\mathrm{male}}(\hat{y}_1 - \hat{y}_0) + \mathrm{average}_{T.\mathrm{male}}(y - \hat{y}_1) - \mathrm{average}_{C.\mathrm{male}}(y - \hat{y}_0). \tag{14}$$

The first average can be over both arms, or just over the treatment group; in the latter case this reduces to

$$\hat{\Delta}_{\text{residual,male}} = \mathrm{average}_{T.\mathrm{male}}(y - \hat{y}_0) - \mathrm{average}_{C.\mathrm{male}}(y - \hat{y}_0). \tag{15}$$

which matches the subset method, plus an adjustment based on control residuals.

3.3 Mutation Method for Slice and Dice

The previous methods do not control for correlations with other variables. Suppose, for example, that younger users have greater lift than other groups and that gender is irrelevant — within any age group the lift for the two sexes is equal — but that males in the campaign tend to be younger than females. Then those methods will show greater lift for males than females. In contrast, coefficients in the logistic regression measure the incremental effect of each variable — e.g. the effect of gender with age held constant. We use that property of the coefficients in the next method.

Recall the technique described above for correcting for covariates. We fit the model, then produce both model predictions, by changing the Z variable to either treatment or control (or to 1 or 0) and feeding the modified data into the prediction model. We use a similar approach for the mutation method. To obtain a slice estimate, say for age 65+, we fit models, then mutate all users by changing their age to 65+, and compare mutated treatment and control predictions.

$$\hat{\Delta}_{\text{mutation,65+}} = \text{average}(\hat{y}_{1,65+} - \hat{y}_{0,65+}). \tag{16}$$

The average is over all users (or just treatment users). Similarly for other age groups. Hence the slice estimates compare the effect of different ages, with all other variables held constant (because we're averaging across the same values for other variables).

While this method has the attractive property of canceling out correlations with other covariates, it is disquieting. To estimate the slice for male, for example, we change all females to males for predictions. This could result in impossible combinations — say the campaign targeting criteria excluded males 18-24 but allowed females in that range, the mutations would result in males in that range. More importantly, it is not clear that advertisers want and know how to interpret estimates that correct for correlations.

4. Failure to Treat

There could be some treatment users who do not actually receive real ad impressions, for technical reasons. In the statistical literature this is called noncompliance. Our goal is to estimate lift for users who actually see ads, so the ideal comparison would be treatment users who see ads vs. control users who would have seen ads. Unfortunately, we do not know who the latter group are. We use an intent-to-treat analysis, comparing treatment group users whom we intended to show ads to similar controls, to estimate the total effect for the campaign, then divide by the number who actually saw ads to estimate the per-person effect (Bloom, 1984). Equivalently, we divide the (estimated Δ based on comparing treatment to control) by the (fraction of treatment users who actually saw ads). This is known as the method of moments estimator of the complier average causal effect (Imbens and Rubin, 2015).

To be more definite, consider four groups of users:

- RT: "real treatment", users who are exposed to real ads
- PT: "pseudo-treatment", users who are not exposed to real ads
- RC: "real controls", users who would have been exposed to real ads if assigned to treatment
- PC: "pseudo-controls", users who would NOT have seen real ads if assigned to treatment

with corresponding means y and counts n. We observe n_{RT}, n_{PT}, n_{*C}, $\bar{y}_{RT}$, $\bar{y}_{PT}$,and $\bar{y}_{*C}$, but not n_{RC}, n_{PC}, $\bar{y}_{RC}$, or $\bar{y}_{PC}$. The unadjusted Δ estimate (ignoring covariate adjustment for now, for simplicity) would be $\bar{y}_{*T} - \bar{y}_{*C}$, and an adjusted estimate is

$$\hat{\Delta}_{\text{adjusted}} = (\bar{y}_{*T} - \bar{y}_{*C})/(n_{RT}/n_{*T}). \tag{17}$$

This can be derived another way. Assume that the same proportion of controls would have seen the ad as for exposed $n_{RC}/n_{*C} = n_{RT}/n_{*T}$, and that the mean for the pseudo-controls matches the mean for pseudo-treatment $\bar{y}_{PC} = \bar{y}_{PT}$ (since neither of these groups saw the ad). Then some algebra yields the adjusted estimate, as well as an adjusted baseline estimate

$$\bar{y}_{RC} = \bar{y}_{RT} - \hat{\Delta}_{\text{adjusted}}. \tag{18}$$

If this results in $\bar{y}_{RC}$ outside the range (0, 1) we truncate and adjust $\hat{\Delta}_{\text{adjusted}}$ accordingly.

In practice, we use this method in combination with covariate adjustment, and slicing estimates.

Note that it would be easy to calculate standard errors (SE) incorrectly. The estimate appears to be a difference in proportions, divided by a constant; in which case the SE would be easy. However, (n_{RT}/n_{*T}) is not a constant, and affects $\bar{y}_{*T}$; when the fraction is high (low) then $\bar{y}_{*T}$ contains more (fewer) real treatment users and is closer (farther) to $\bar{y}_{RT}$. The multivariate delta method (first-order Taylor series approximation) could be used to calculate a standard error; we bootstrap instead.

5. Confidence Intervals

We have derived analytical standard errors for some Δ estimates, including (4) and (7), based on influence functions. The influence of any observation may be a sum of two terms. First, observations in the training set influence $\hat{\beta}$, which in turn influences all predictions. Second, observations included in the average have additional influence proportional to $\hat{y}_1 - \hat{y}_0$ for the observation.

However, we are not using these analytical SEs, because there are additional complicating factors: two-step extrapolation from respondents to the campaign audience, multiple slicing methods, failure-to-treat adjustments, regularization, and additional issues that are beyond the scope of this article (stratification, oversampling certain subpopulations and correcting for that bias by weighting, and estimates for the contrasts between slices), that combine to make analytical SEs complicated.

Instead, we bootstrap to calculate SEs, then calculate confidence intervals using the usual form of a t interval

$$\hat{\Delta} \pm t_{\alpha/2,df}\text{SE}[.] \tag{19}$$

The degrees of freedom should reflect both the uncertainty due to sample size in the original data, and uncertainty due to random bootstrapping, $\frac{1}{df} = \frac{1}{n-1} + \frac{1}{r-1}$, where n is the sample size and r the number of bootstrap replications. In our case r is much smaller than n (100 bootstrap replications, due to computational expense of bootstrapping for thousands of campaigns running simultaneously, analyzing multiple questions per campaign and many slices per question), so we use the approximation $df = r - 1$.

We use these t intervals rather than other bootstrap intervals, such as bootstrap percentile intervals (Efron and Tibshirani, 1993), because 100 bootstrap samples is not enough for percentile intervals. Even for t intervals, the extra uncertainty in standard errors due to using 100 bootstrap replications instead of infinite results to a loss of power equivalent to discarding about 2% of the observations.

There are two common ways to bootstrap in regression applications — to resample observations, or resample Y conditional on x (keeping all x values fixed). The latter corresponds to resampling residuals in linear regression; for logistic regression it means to fit the initial model, calculate $\hat{y}$ for each observation, then generate Y values independently using $P(Y = 1 \,|\, X) = \hat{y}$. The two approaches are roughly comparable for estimating the variability of regression coefficients, or variability of predictions at fixed values of x. However, here we average across predictions at random x values; the latter approach misses variation due to the observations being averaged over being a random sample; this especially matters when extrapolating.

5.1 Relative Lift

In addition to additive lift

$$\Delta = p_{\text{treatment}} - p_{\text{baseline}},$$

we can also calculate relative lift:

$$\text{lift} = \Delta / p_{\text{baseline}}.$$

We use t intervals using bootstrap standard errors for Δ, and Fieller intervals (Fieller, 1954) (using t quantiles) for lift, using the bootstrap estimate of the covariance matrix for Δ and baseline.

6. Case Study

This article is based on the application of Google Brand Lift to estimating lift from TrueView advertising campaigns, with surveys on YouTube. In this case we use a randomized experiment. For some other applications we use observational controls. We begin with further details on Google Brand Lift, then look at one example campaign.

6.1 Experiment Framework

Users are assigned to treatment or control using random number generation based on a hashed version of their cookies. 70% are assigned to the treatment group and 30% to the control group.

When users visit YouTube and are to be shown an ad, we normally use an auction to determine which ad to show. For campaigns to be analyzed, we instead begin with a "simulated auction" that mimics the real auction, but is not used to place ads. For treatment users, ads that win the simulated auction are sent to the real auction, where they normally win and are shown to users. There are rare cases where they do not win the real auction (failure to treat, see Section 4). For controls, ads that win are sent to a real auction that excludes the campaign ads, and users are shown the ad that wins that auction; we call this the replacement ad. Controls do not see campaign ads.

The 70-30 split is based on trade-off between cost and accuracy. Increasing the fraction of control users would reduce the random error in lift estimates, but increase cost and reduce system effectiveness, because replacement ads usually have a lower bidding price than real ads, and are less effective for their sponsors.

The simulated auction is run for a specific impression and a specific user. If the campaign ad wins the simulated auction, the corresponding user's cookie is eligible for a campaign impression, i.e., it is the cookie which would have been shown the campaign ad if no experiment is running. We call those campaign impressions predicted by the simulated auction as "virtual impressions", and cookies that are eligible for the virtual impressions as “eligible cookies”. The eligible cookies which are in the treatment group are served a campaign ad. Those in the control group are served a replacement ad.

6.2 Surveys

Questions are designed to address different stages of the advertising funnel: ad recall, brand awareness, consideration, etc. For example, for ad recall a typical multiple choice question is: “Which of the following online ads have you seen recently?”, where one of the choices is the target brand. Similarly, a brand awareness question is “Which of the following brands have you heard of?”

Treatment users who see ads, and control users who would have, are flagged for surveys (Sostek and Slatkin, 2017). However, they are not surveyed immediately, but rather only after a minimum 1 hour delay, in order to estimate the persistent effect of the ads, rather than the very short-term effect right after an ad. They are not surveyed if their last virtual impression happened too many days in the past.

6.2.1 Solicitation and Response Biases

We set out to measure the ad effect for the “campaign audience”, i.e., users who see (or controls who would see) the campaign ad(s). As in any survey, solicitation and response biases are inevitable (Callegaro et al., 2014). Solicitation bias occurs when users who are solicited differ in meaningful ways from who are not. We solicit treatment and controls the same way, so this bias should be the same for both arms. To be solicited, a user must visit a page where they can be surveyed within a specified time period after their last virtual impression. Hence more active cookies have a better chance to be solicited, and we do observe more internet activity and greater campaign ad frequency for solicited cookies than for non-solicited cookies. While both arms are affected the same way, it could be that high-frequency (and more-solicited) users have greater or lesser lift than low-frequency users.

Response bias is due to audience choices, but influenced by the survey system. A survey is presented in place of a YouTube ad preceding a video. If a user does not want to answer, she can close or refresh the tab, or skip/wait for Y seconds for the content video to start. She is more likely to do so if not that interested in the following video. She might try to answer, but be timed out if she does not submit the answer within 30 seconds. She may be more likely to answer if interested in the sponsor or product, and this interest might be affected by having seen the ad or not. We remove users who answer too quickly, because they are more likely to be answering randomly, just to get to the following video. These factors may make the answers we receive from respondents different than answers for the whole population exposed to the ads, and answers we receive from the treatment group different than answers from the control group.

6.2.2 Cookie vs. User

We set up experiments based on cookies and user accounts, rather than real users. Users may have multiple cookies — because they refresh cookies, use multiple devices, multiple browsers, or both apps and browsers on mobile. We estimate lift per cookie rather than per user. This underestimates per-person lift and total campaign lift, because control cookies may have seen campaign ads using their other cookies, users may answer surveys before seeing all the ads they will see, and measurement for each cookie only reflects the marginal contribution of the cookie and the effects of multiple ad exposures are typically sublinear.

6.2.3 Impressions after the Survey

Users may see additional campaign ads after the survey; the estimates we produce only reflect the effect of ads prior to the survey.

6.2.4 Filtering

We use several layers of filtering to improve data quality. For example, we filter out short term cookies, cookies who answered the same question more than once, and responses with very short or a very long response time (defined as the time between solicitation and responding to the survey).

6.2.5 Scope

We are able to collect a large number of responses, to reduce the random variation of the estimates.

6.3 Estimation Details

Signals available for everyone include (estimated) age and gender, device (desktop, mobile, tablet), number of ads seen on different platforms in the week before the first ad impression (this is useful as a proxy for level of internet activity), number of virtual impressions for campaign ads, and some campaign-specific variables, e.g. indicator variables when an advertiser uses multiple campaigns. Signals only available for solicited users include the solicitation gap (time between ad exposure and solicitation), and number of campaign impressions before solicitation. Signals only available for the treatment group include how long a user watches a campaign video.

We transform some variables, e.g. $\log(1 + x)$ for ad count variables. We have special handling for rare levels of categorical variables for privacy reasons, to avoid producing unreliable slice/dice estimates, and to avoid undefined estimates, e.g. extrapolating from respondents to a larger group when the larger group contains a level not present among respondents. Options include merging those levels into “Other”, or using regularization, combined with not reporting on rare slices.

For consistency with industry practice, we determine statistical significance of lift using one-sided tests with significance level 0.1, and produce 80% confidence intervals.

6.4 Example Study

Here is a case study based on a campaign by an anonymous advertiser. The question funnel stage is ad recall; the question text is: “Which of the following have you seen online video advertising for recently?” $y = 1$ if the user selects the campaign advertiser from a multiple-choice list. The table below shows additive lift measured in four ways:

- Raw: Treatment positive rate minus control positive rate for respondents
- Corrected Respondents: Corrected for differences in covariates, for respondents
- Extrapolate to Solicited: Corrected, for solicited users
- Extrapolate to All: Corrected, all campaign users

	Treatment	Baseline	Δ	SE	Conf. Interval
Raw(Treatment - Control)	0.169	0.136	0.033	0.011	(0.018, 0.048)
Corrected Respondents	0.169	0.139	0.030	0.012	(0.015, 0.045)
Extrapolate to Solicited	0.162	0.133	0.029	0.014	(0.011, 0.047)
Extrapolate to All	0.124	0.097	0.027	0.016	(0.007, 0.048)

All lifts in the table are significantly positive. Standard errors are larger when extrapolating to solicited or All, due to model uncertainty.

We also report lift and confidence intervals for slices (age, gender, impression frequency, solicitation gap etc.). From the frequency slice plot (Figure 2), we see that both impression frequency 1 and 2+ have significant lift. In this particular campaign, we do not find evidence that 2+ frequencies has higher lift than 1 frequency. The standard error for frequency 2+ is larger because there are fewer users with 2+ impressions.

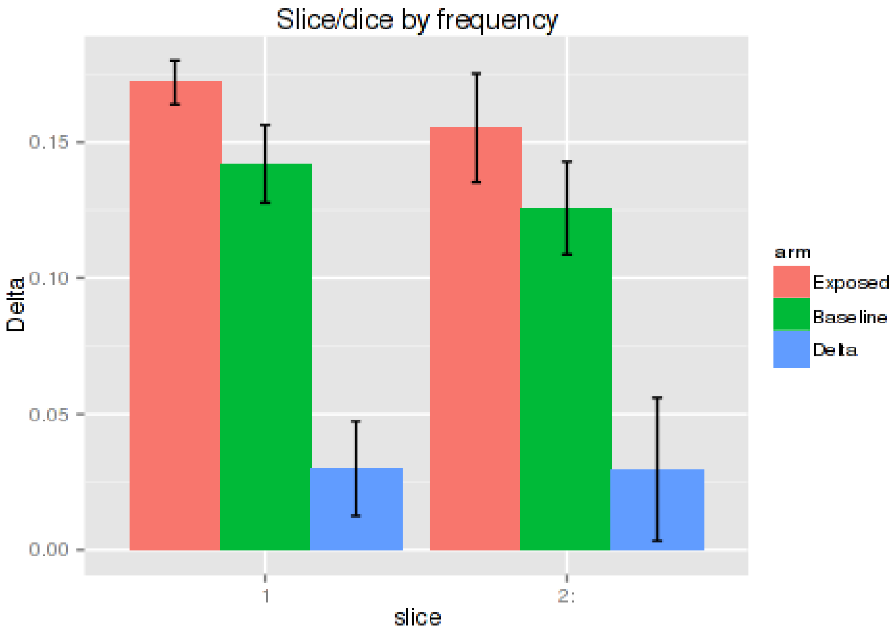

Figure 2: Slice/dice by frequency.

7. Conclusion

We present a method for measuring the brand lift through randomized experiments and online surveys. The experiment setup enables estimation of the causal effect of the ad on brand metrics. We describe methods to correct for imbalance in covariates, solicitation and response biases. We produce slice estimates — lift broken down by different dimensions such as device, demographics, and impression frequency. Confidence intervals use t intervals for additive lift and Fieller intervals for relative lift, using bootstrap standard errors and covariances.

Acknowledgments

We thank David Broockman, Jim Dravillas, Tony Fagan, Ayres Fan, Jessica Hwang, Jim Koehler, Randall Lewis, Daniel Popescu, Samir Pradhan, Mehdi Sharifzadeh, Mark Thomas, Mike Wurm, Yunwen Yang, and Sen Yuan for helpful comments, and many others for their contributions.

References

Bloniarz, A., H. Liu, C.-H. Zhang, J. S. Sekhon, and B. Yu (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences 113*(27), 7383–7390.

Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation review 8*(2), 225–246.

Callegaro, M., R. P. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, and P. J. Lavrakas (2014). *Online panel research: A data quality perspective*. John Wiley & Sons.

Chan, D., R. Ge, O. Gershony, T. Hesterberg, and D. Lambert (2010). Evaluating online ad campaigns in a pipeline: causal models at scale. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 7–16. ACM.

Efron, B. and R. J. Tibshirani (1993). An introduction to the bootstrap.

Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society. Series B (Methodological) 16*(2), 175–185.

Greenland, S. and M. A. Mansournia (2015). Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in medicine 34*(23), 3133–3143.

Hanna, R., A. Rohm, and V. L. Crittenden (2011). Were all connected: The power of the social media ecosystem. *Business horizons 54*(3), 265–273.

Hesterberg, T., N. H. Choi, L. Meier, C. Fraley, et al. (2008). Least angle and l1 penalized regression: A review. *Statistics Surveys 2*, 61–93.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman and Hall/CRC.

Sostek, K. and B. Slatkin (2017). How google surveys works. Technical report, White paper, Google Inc.