Data Enrichment for Incremental Reach Estimation

Aiyou Chen[1], Jim R. Koehler[1], Art B. Owen[2], Nicolas Remy[1], and Minghui Shi[1]

[1]Google Inc.
[2]Stanford University

Abstract

There is increasing interest in measuring the overlap and/or incremental reach of cross-media campaigns. The direct method is to use a cross-media panel but these are expensive to scale across all media. Typically, the cross-media panel is too small to produce reliable estimates when the interest comes down to subsets of the population. An alternative is to combine information from a small cross-media panel with a larger, cheaper but potentially biased single media panel. In this article, we develop a data enrichment approach specifically for incremental reach estimation. The approach not only integrates information from both panels that takes into account potential panel bias, but borrows strength from modeling conditional dependence of cross-media reaches. We demonstrate the approach with data from six campaigns for estimating YouTube video ad incremental reach over TV. In a simulation directly modeled on the actual data, we find that data enrichment yields much greater accuracy than one would get by either ignoring the larger panel, or by using it in a data fusion.

1 Incremental online reach

People are spending an increasing amount of time on the internet which motivates marketers to increase their online ad presence (The Nielsen Company, 2011). The value of online advertising is reflected not only in the reach of those ads, but also in their incremental reach: the number of consumers reached online but not reached by other media such as television. Measurement of online incremental reach can help advertisers optimize their mixed media campaigns. See for example Collins and Doe (2009), Doe and Kudon (2010), and Jin et al. (2012).

The ideal instrument to measure online incremental reach is a single source panel (SSP) of consumers for whom online and televised viewing data are both available along with demographic information (age, gender, income group and so on). Such single source panels can be very expensive to recruit and may then be unavailable or available with a smaller than desired sample size.

A common alternative is to combine two single media panels. One may measure demographics and television reach, while the other measures demographics and online reach. This approach is variously called statistical matching or data fusion. D'Orazio et al. (2006) provide a comprehensive overview. Data fusion can be approached as a missing data problem (Little and Rubin, 2009), for a merged data matrix in which data from a collection of panels are stacked one above the other (Rässler, 2004). One can, for example fit a model for TV viewership in one panel and use it to impute missing TV viewership information in the other. The output of data fusion is a synthetic table merging demographics, with both online and television exposures.

Strictly speaking, data fusion is impossible. We cannot deduce the joint distribution of three quantities from the joint distributions of two, or even three, pairs of those quantities. Data fusion is usually made under a strong assumption: the chance that a given person is exposed to an online ad may depend on their demographics, but conditionally on their demographics, it may not also depend on their television exposure. That is, the two exposure variables are conditionally independent given the subject's demographics. This is called the *conditional independence assumption*, or CIA.

An alternative to data fusion is to combine a small SSP with a larger panel measuring just one of the reach variables, such as online reach, but not television reach. This second panel may be recruited online and can be much larger than the SSP. We call it the broad reach panel, or BRP for short. In this setting the SSP gives some information on the conditional dependence of the reach variables given demographics, which one can use to leverage the information from the BRP. For example, Gilula et al. (2006) make use of such a strategy, parameterizing the conditional dependence by an odds ratio that they estimate from the SSP. This approach requires an assumption that the conditional dependence between reach variables that we estimate from the SSP be the same as the conditional dependence in the BRP. Given demographics, the two reach variables should have an identical distribution in SSP and BRP. We call this the *identical distribution assumption*, or IDA.

In the Google context, we have an SSP and a larger BRP. We have many campaigns to consider and both the CIA and IDA will fail for some of them, to varying and unknown degrees. To handle this we develop four estimators for the incremental lift in our population: an unbiased one based on the SSP alone, one that would be unbiased if the CIA held, one that would be unbiased under the IDA, and one that would be unbiased under both assumptions. For these alternative estimators to be accurate their bias does not have to be zero, it just has to be small. The difficulty is in knowing how small is small enough, when all we have to judge by are the data at hand. We make our predictions by using a data dependent shrinkage that pulls the potentially biased estimators towards the unbiased one.

When we deploy our algorithm, we also split our population into subpopulations and apply shrinkage within each resulting subgroup. The process of forming these subgroups from data induces some bias, even for the SSP only method. This is an algorithmic bias, not a sampling bias, and it is incurred as part of a bias-variance tradeoff.

We take advantage of recent theoretical results in Chen et al. (2013) showing that even biased data sets can always be used to improve regression models as long as there are at least 5 regression coefficients to estimate and at least 10 degrees of freedom for error. This is similar to the well known result of Stein (1956) in which the sample mean is inadmissible in 3 or more dimensions as an estimator of the center of a Gaussian distribution.

An outline of this paper is as follows. We describe the data sources, the notation and our problem formulation in Section 2. Section 3 shows the potential for variance reduction available from using the CIA, IDA or both. The potential variance reductions carry the risk of increased bias. Section 5 adapts the data enrichment procedure to the incremental reach setting in order to make a data driven bias-variance tradeoff. Section 6 shows in a realistic simulation that data enrichment improves the accuracy of incremental reach estimates over the main alternatives of ignoring the larger data set and using it for data fusion. Some technical arguments are placed in the Appendix.

2 Notation and example data sets

We begin with our notation. For person i in the target population, the vector X_i contains their demographic variables. These typically include gender, age, education, and so on. Age and education are usually coded as ordered categorical variables.

The variable Y_i equals 1 if person i was reached by a television campaign, and 0 otherwise. Similarly, the variable Z_i is 1 if person i was reached by the corresponding online campaign, and 0 otherwise. The derived variable $V_i = Z_i(1 - Y_i)$ takes the value 1 if and only if person i is an incremental reach.

Given values (X, Y, Z) for everybody in the population, we would know the online and televised reach as well as the incremental reach for every demographic grouping. In particular, a person contributes to the incremental reach if and only if $V = 1$, that is, the person saw the ad online but not on TV.

We use census data to get the proportions of people at each level of X in the target population, such as a whole country, region or a demographic subset, such as adults in New York state. We write $p(x)$ for the fraction of people in the target population with $X = x$. The possible values for x are given by the set $\mathcal{X}$. The census data has no information on Y and Z. If we did not know $p(x)$ then there would be some bias from the X distribution within the SSP and BRP being different from the population. Our analyses use p-weighted combinations of the data to remove this source of bias allowing us to focus on biases from CIA or IDA not holding.

Our goal is to estimate incremental reach for arbitrary subsets of the population, e.g. each demographic sector, that is $\mathsf{IR}(x) = \Pr(V = 1|X = x)$ for some or all $x \in \mathcal{X}$. We may factor $\mathsf{IR}(x)$ as

$$\mathsf{IR}(x) = \Pr(Z = 1 \mid X = x) \Pr(Y = 0 \mid X = x, Z = 1). \tag{1}$$

Both samples bear on the first factor in (1), while the second factor can only be estimated from the SSP.

The SSP has a sample size of n and the BRP has a sample size of N. It is useful to keep track of the fraction of our data values coming from each of these two samples. These are $f = n/(n+N)$ and $F = N/(n+N) = 1-f$ respectively.

For our example, we study the incremental reach of online advertising over television advertising for six household products. The SSP has 6322 panelists recruited based on a probability sample from the target population. In addition to Y and Z we measure the following demographic variables: age, gender, education and occupation. The age variable is ordinal with six levels: under 20, 21–30, 31–40, 41–50, 51–60, and over 60. We label these 1 through 6. The education variable is also ordinal with 4 levels: below middle school, middle school, high school, and beyond high school. Gender was coded as 1 for female and 2 for male. Occupation was coded as 1 for employed and 0 for not employed.

In addition to the SSP, there is a larger panel of 12,821 panelists for whom we have the same demographic variables X, the online viewership Z, but not the television viewership Y. In many Google contexts, the second data set comes from logs or a panel recruited online. In this instance our BRP was from a second probability sample. As such it might have somewhat smaller bias relative to the BRP than what we would see in those other contexts. Our setting is similar to that of Singh et al. (1993), except that in their case the (X, Y, Z) sample is of lower fidelity to the population than the (X, Z) survey. Our data structure is different from Gilula et al. (2006) in that we do not have a TV-only panel which contains (X, Y).

We track six advertising campaigns for these two panels. Three of the campaigns are for laundry detergents. The others are a beer, a kind of salt and Google's own Chrome browser.

Table 2.1 summarizes overall web reach, TV reach and incremental reach statistics for each campaign based on both SSP and BRP separately. Web reach is much lower than TV reach, but the proportion of incremental reach among web reach is quite high from 20% to 50%.

3 Potential gains from assumptions

In this section, we analyze the potential variance reductions available using either the CIA, the IDA, or both of them at once. Making these assumptions brings more data to bear on the problem and reduces variance. To the extent that the assumptions are violated they introduce a bias. We will trade off bias versus variance using cross-validation.

Our goal throughout is to estimate $\theta = \Pr(V = 1)$ for an aggregate such as the entire target population, or a subset thereof. We do not attempt to predict which specific individuals have $V_i = 1$ when that is unknown. In our experience

Campaign	WEB_S	TV_S	IR_S	WEB_B	$(\text{IR}/\text{WEB})_\text{S}$
Soap 1	0.63	67.67	0.30	0.67	0.47
Soap 2	0.65	70.69	0.27	0.55	0.42
Soap 3	1.95	64.41	0.78	1.83	0.40
Beer	0.86	56.87	0.37	0.85	0.43
Salt	2.12	77.49	0.49	1.91	0.23
Chrome	11.87	67.87	3.49	12.32	0.29

Tab. 2.1: Summary statistics for 6 campaigns: web reach in SSP, TV reach in SSP, incremental reach in SSP, and web reach in BRP, as percentages. The final column shows the fraction of web reaches which are incremental.

it is very hard to make reliable predictions for individuals, and moreover, the marketing questions of interest have to do with aggregates.

Using the SSP alone we can estimate θ by $\hat{\theta}_\text{S} = \overline{V}_\text{S}$, the average of V_i for $i \in \text{S}$. This is the baseline method against which improvements will be measured. The variance using the SSP alone is $\theta(1-\theta)/n$. Here, and below, we sometimes abbreviate SSP to S and BRP to B which also have helpful mnemonics, S for small data set and B for big data set.

Our methods for improvement are based on the decomposition

$$\theta = \Pr(Z = 1)\Pr(Y = 0 \mid Z = 1).$$

Under the IDA, we can use the BRP data to get a better estimate of the first factor, $\Pr(Z = 1)$. Under the CIA, we can get a better estimate of the second factor $\Pr(Y = 0 \mid Z = 1)$.

3.1 Gain from the IDA alone

If the IDA holds, then we can estimate $\Pr(Z = 1)$ by the pooled average $(n\bar{Z}_\text{S} + N\bar{Z}_\text{B})/(N + n) = f\bar{Z}_\text{S} + F\bar{Z}_\text{B}$. Here $\bar{Z}_\text{S}$ and $\bar{Z}_\text{B}$ are averages of Z_i over $i \in \text{S}$ and $i \in \text{B}$ respectively. The IDA does not help us with $\Pr(Y = 0 \mid Z = 1)$ because there are no Y values in the BRP. We estimate that factor by $\bar{V}_\text{S}/\bar{Z}_\text{S}$, the fraction of online reaches in the SSP that are incremental reaches, arriving at the estimate

$$\hat{\theta}_\text{I} = (f\bar{Z}_\text{S} + F\bar{Z}_\text{B})\frac{\bar{V}_\text{S}}{\bar{Z}_\text{S}}.$$

We adopt the convention that $\bar{V}_\text{S}/\bar{Z}_\text{S} = 0$ in those cases where $\bar{Z}_\text{S} = 0$. Such cases automatically have $\bar{V}_\text{S} = 0$ too. This event has negligible probability when n is large enough and θ is not too small. When the data are partitioned into small subsamples (e.g., demographic groups) we may get some such 0s. In that case our convention biases what is usually a small number down to zero. This small bias is conservative.

We use the delta method (Taylor expansion) to derive an approximation $\widetilde{\mathrm{var}}(\hat\theta_{\mathrm{I}})$ to $\mathrm{var}(\hat\theta_{\mathrm{I}})$, in Section A.1 of the Appendix. The result is that

$$\frac{\widetilde{\mathrm{var}}(\hat\theta_{\mathrm{I}})}{\mathrm{var}(\hat\theta_{\mathrm{S}})} = 1 - F\frac{1-p_z}{p_z}\frac{\theta}{1-\theta},$$

where $p_z = \Pr_{\mathrm{S}}(Z=1) = \Pr_{\mathrm{B}}(Z=1)$. The IDA then affords us a big gain to the extent that the BRP is large (so F is large), incremental reach is high (so $\theta/(1-\theta)$ is large) and online reach is small (so $p_z/(1-p_z)$ is large).

Incremental reaches must also be online reaches, and so $\theta \leqslant p_z$. Therefore

$$\frac{\widetilde{\mathrm{var}}(\hat\theta_{\mathrm{I}})}{\mathrm{var}(\hat\theta_{\mathrm{S}})} \geqslant 1 - F\frac{\theta}{p_z} \geqslant 1 - F.$$

The first inequality will be close to an equality when p_z and hence θ is small. For our applications $1 - F\theta/p_z$ is a reasonable approximation to the variance ratio. The second inequality reflects the fact that pooling the data cannot possibly be better than what we would get with an SSP of size $n + N$.

From $\widetilde{\mathrm{var}}(\hat\theta_{\mathrm{I}})/\mathrm{var}(\hat\theta_{\mathrm{S}}) \approx 1 - F\theta/p_z$ we see that using the BRP is effectively like multiplying the SSP sample size n by $1/(1-F\theta/p_z)$. Our greatest precision gains come when a high fraction of online reaches are incremental, that is, when θ/p_z is largest. In our application this proportion ranges from 20% to 50% when aggregated to the campaign level. See Table 2.1 in Section 2.

3.2 Gain from the CIA alone

Here we evaluate the variance reduction that would follow from the CIA. In that case, we could take advantage of the Z–Y independence, and estimate θ by

$$\hat\theta_{\mathrm{C}} = \bar Z_S(1 - \bar Y_S).$$

It is shown in the Appendix that the delta method variance of $\hat\theta_{\mathrm{C}}$ satisfies

$$\frac{\widetilde{\mathrm{var}}(\hat\theta_{\mathrm{C}})}{\mathrm{var}(\hat\theta_{\mathrm{S}})} = 1 - \frac{p_y(1-p_z)}{1-\theta} \geqslant 1 - p_y, \tag{2}$$

when the CIA holds. This can represent a dramatic improvement, when the online reach p_z and incremental reach θ are both small while the TV reach p_y is large. If the CIA holds, our application data suggest the variance reduction can be from 50% to 80%. The reverse setting with tiny TV reach and large online reach would not be favorable to $\hat\theta_{\mathrm{C}}$, but our data are not of that type.

3.3 Gain from the CIA and IDA

Finally, suppose that both the CIA and IDA hold. If we apply both assumptions, we can get the estimator $\hat\theta_{\mathrm{I,C}} = (f\bar Z_{\mathrm{S}} + F\bar Z_{\mathrm{B}})(1 - \bar Y_{\mathrm{S}})$. We already gain a lot from the CIA, so it is interesting to see how much more the IDA adds when the CIA holds. We show in the Appendix that under both assumptions,

$$\frac{\widetilde{\mathrm{var}}(\hat{\theta}_{\mathrm{I,C}})}{\widetilde{\mathrm{var}}(\hat{\theta}_{\mathrm{C}})} = \frac{f(1-p_y)(1-p_z)+p_yp_z}{(1-p_y)(1-p_z)+p_yp_z}.$$

If both reaches are high then we gain little, but if both reaches are small then we reduce the variance by almost a factor of f, when adding the IDA to the CIA. In our case we expect that the television reach is large but the online reach is small, fitting neither of these extremes. Consider a campaign with $f = 1/3$, $p_y = 2/3$ and $p_z = 99/100$, similar to the soap campaigns. For such a campaign,

$$\frac{\widetilde{\mathrm{var}}(\hat{\theta}_{\mathrm{I,C}})}{\widetilde{\mathrm{var}}(\hat{\theta}_{\mathrm{C}})} = \frac{(1/9)\times .99+(2/3)\times .01}{(1/3)\times .99+(2/3)\times .01} \doteq .34,$$

so the combined assumptions then allow a nearly three-fold variance reduction compared to CIA alone.

4 Example campaigns

Our data enrichment scheme is described in Section 5. Here we illustrate the results from that scheme on six marketing campaigns and discuss the differences among different algorithms.

In addition to data enrichment, we also show results from tree structured models. Those split the data into groups and recursively split the groups. More about tree fitting is in Section 5. One model fits a tree to the SSP data alone and another one works with the pooled SSP and BRP data.

For all three of those methods we have aggregated the predictions over the age variable, which takes six levels. In addition, we show the empirical results for age, which amount to recording the percentage of incremental reaches, that is, data with $Z(1-Y) = 1$, at each unique level of age in the SSP. There is no corresponding empirical prediction fully disaggregated by age, gender, income and education, because of the great many empty cells that would cause.

We found the age related patterns of incremental reach particularly interesting. Figure 4.1 shows estimated incremental reach for all three models and the empirical counts, on all six campaigns, averaged over age groups. The beer campaign is particularly telling. The empirical data show a decreasing trend of incremental reach with age. The tree fit to SSP-only data yields a fit that is constant in age. The tree model had to explore splitting the data on all four variables without a prior focus on age. There were only 23 incremental reach events for beer in the SSP data set. With such a small number of events and four predictors, there is considerable possibility of overfitting. Cross-validation lead to a model that grouped the entire SSP into one set, that is, the tree had no splits. Both pooling and data enrichment were able to borrow strength from the BRP as well as take advantage of approximate independence of television and web exposure. They then recover the trend with age.

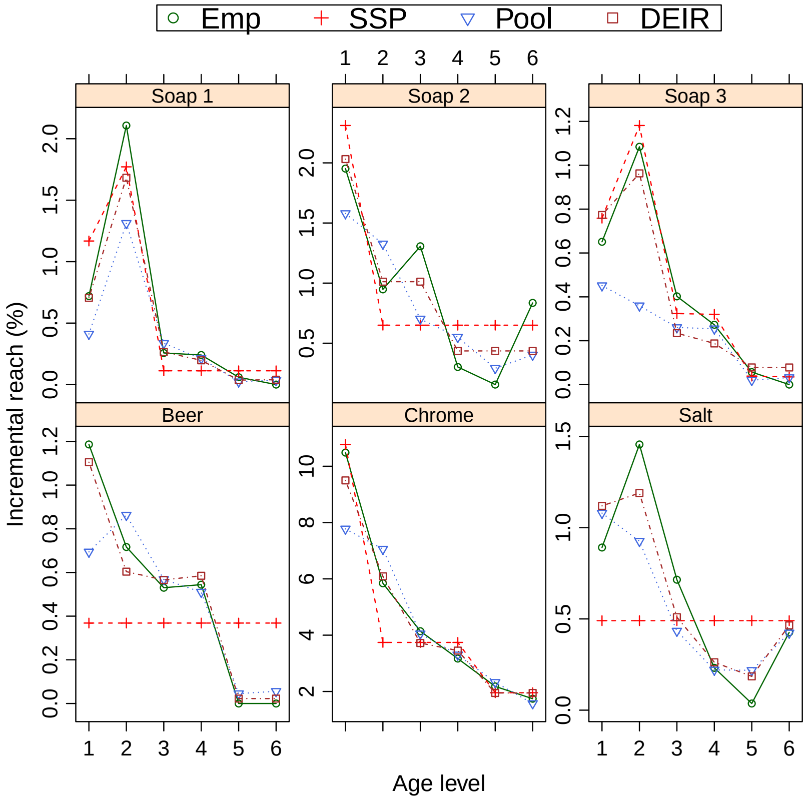

Fig. 4.1: Estimated incremental reach by age, for six campaigns and three models: SSP, Pooling and DEIR as described in the text. Empirical counts are marked by Emp.

The Salt campaign had a similarly small number of incremental reaches and once again the SSP only tree was constant. Fitting a tree to the SSP data always gave a flatter fit versus age than did DEIR which in turn was flatter than what we would get simply pooling the data. Section 6 gives simulations in which DEIR has greater accuracy than using pooling or SSP only.

5 Data enrichment for incremental reach

For a given sample we would like to combine incremental reach estimates $\hat{\theta}_S$, $\hat{\theta}_I$, $\hat{\theta}_C$ and $\hat{\theta}_{I,C}$ whose assumptions are: none, IDA, CIA and IDA+CIA, respectively. The latter three add some value if their corresponding assumptions are nearly true, but our information about how well those assumptions hold comes from the same data we are using to form the estimates.

The circumstances are similar to those in data enriched linear regression (Chen et al., 2013). In that problem there is a regression model $Y_i = X_i^\mathsf{T}\beta + \varepsilon_i$ which holds in the SSP and a biased regression model $Y_i = X_i^\mathsf{T}(\beta+\gamma) + \varepsilon_i$ holds in the BRP. The estimates are found by minimizing

$$S(\lambda) = \sum_{i\in \mathrm{S}}(Y_i - X_i^\mathsf{T}\beta)^2 + \sum_{i\in \mathrm{B}}(Y_i - X_i^\mathsf{T}(\beta+\gamma))^2 + \lambda\sum_{i\in \mathrm{S}}(X_i^\mathsf{T}\gamma)^2, \qquad (3)$$

over β and γ for a nonnegative penalty factor λ. The ε_i are independent with mean 0 and variance σ^2_S in the SSP and σ^2_B in the BRP.

Taking $\lambda = 0$ amounts to fitting regressions separately in the two samples yielding an estimate $\hat\beta$ that does not use the BRP at all. The limit $\lambda \to \infty$ corresponds to pooling the two data sets, which would be optimal if there were no bias, i.e., if $\gamma = 0$. The specific penalty in (3) discourages the estimated γ from making large changes to the SSP; it is one of several penalties considered in that paper.

Varying λ from 0 to ∞ gives a family of estimators that weight the SSP to varying degrees. The optimal λ is unknown. An oracle that knew γ and the error variance in the two data sets would be able to compute the optimal λ under a mean squared error loss. Chen et al. (2013) get a formula for the oracle's λ and then plug estimates of γ and the variances into that formula. They show, under conditions, that the resulting plugin estimate gives better estimates of β than using the SSP only would. The conditions are that the Y values are normally distributed, and that the model have at least 5 regression parameters and 10 error degrees of freedom. The normality assumption allows a technical lemma due to Stein (1981) to be used and we believe that gains from using the BRP do not require normality.

In principle we might multiply the sum of squared errors in the BRP by $\tau = \sigma^2_\mathrm{S}/\sigma^2_\mathrm{B}$ if that ratio is known. If $\sigma^2_\mathrm{BRP} > \sigma^2_\mathrm{SSP}$ then we should put less weight on the BRP sample relative to the SSP sample. However the same effect is gained by increasing λ. Since the algorithm searches for optimal λ over a wide range it is less important to precisely specify τ. Chen et al. (2013) took $\tau = 1$, simply summing all squared errors, and we will generalize that approach.

For the present setting we must modify the method. First our responses are binary, not Gaussian. Second we have four estimators to combine, not two. Third, those estimators are dependent, being fit to overlapping data sets.

5.1 Modification for binary response

To address the binary response there are two reasonable choices. One is to employ logistic regression. The other is to use tree-structured regression and then pool the estimators at the leaves of the tree. Regarding prediction accuracy, there is no unique best algorithm. There will be data sets for which simple logistic regression outperforms tree based classifiers and vice versa.

For this paper we have adopted trees. Tree structured models have two practical advantages. First, the resulting cells that they select correspond to empirically determined market segments, which are then interpretable. Sec-

Data set	Source	Imputed V	Assumptions
D_0	SSP	$Z_S(1 - Y_S)$	none
D_1	BRP	$Z_B(1 - \widehat{Y}_{\text{SSP}}(X_B, Z_B))$	IDA
D_2	SSP	$\widehat{Z}_{\text{SSP}}(X_S)(1 - \widehat{Y}_{\text{SSP}}(X_S))$	CIA
D_3	SSP	$\widehat{Z}_{\text{SSP+BRP}}(X_S)(1 - \widehat{Y}_{\text{SSP}}(X_S))$	CIA & IDA

Tab. 5.1: Four incremental reach data sets and their imputed incremental reaches. The hats denote model-imputed values. For example $\widehat{Y}_{\text{SSP}}(X_B, Z_B)$ is a predictive model for Y based on values X and Z fit using data from SSP and evaluated at $X = X_B$ and $Z = X_B$ (from BRP).

ond, within any of those cells, the model is intercept-only. Then both logistic regression and least squares reduce to a simple average.

Each leaf of the regression tree defines a subset of the data that we call a cell. There are cells $1, \ldots, C$. The SSP has n_c observations in cell c and the BRP has N_c observations there.

For each cell and each set of assumptions we use a linear regression model relating an incremental reach quantity like $\widetilde{V}_i$ to an intercept. When there are no assumptions then $\widetilde{V}_i$ is the observed incremental reach for $i \in S$. Otherwise we may take advantage of the assumptions to impute values $\widetilde{V}_i$ using more of the data. The incremental reach values for each set of assumptions are given in Table 5.1. The predictive models shown there are all fit using `rpart`.

For $k = 0, 1, 2, 3$ let $\widetilde{V}_k$ be vector of imputed responses under any of the assumptions from Table 5.1 and $\widetilde{X}_k$ their corresponding predictors. The regression framework minimizes

$$\|\widetilde{V}_0 - \widetilde{X}_0\beta\|^2 + \sum_{k=1}^{3} \|\widetilde{V}_k - \widetilde{X}_k(\beta + \gamma_k)\|^2 + \sum_{k=1}^{3} \lambda_k \|\widetilde{X}_0\gamma_k\|^2. \tag{4}$$

over β and γ_k for penalties λ_k. In our setting each $\widetilde{X}_k$ is a column vector of ones of length m_k. For cell c, $m_{1k} = N_c$ and $m_{0k} = m_{2k} = m_{3k} = n_c$.

5.2 Search for λ_k

It is very convenient to search for suitable weights in the simplex

$$\Delta^{(K)} = \{(\omega_0, \omega_1, \ldots, \omega_K) \mid \omega_k \geqslant 0, \sum_{k=0}^{K} \omega_k = 1\}$$

because it is a bounded set, unlike the set $[0, \infty]^K$ of usable vectors $\lambda = (\lambda_1, \ldots, \lambda_K)$. Chen et al. (2013) remark that it is more reasonable to use a common set of λ_k over all cells, stemming from unequal sample sizes. The search we use combines the advantages of both approaches.

Our search strategy for the simplex is to choose a grid of weight vectors

$$\boldsymbol{\omega}_g = (\omega_{g0}, \omega_{g1}, \dots, \omega_{gK}) \in \Delta^{(K)}, \quad g = 1, \dots, G.$$

For each vector $\boldsymbol{\omega}_g$ we find a vector $\boldsymbol{\lambda}_g = (\lambda_1, \dots, \lambda_K)$ such that

$$\sum_{c=1}^{C} p_c \omega_{k,c} = \boldsymbol{\omega}_{gk}, \quad k = 0, 1, \dots, K,$$

where p_c is the proportion of our target population in cell c. That is, the population average weight of $\omega_{k,c}$ matches ω_{gk}. These weights give us the vector $\boldsymbol{\lambda}_g = (\lambda_{g1}, \dots, \lambda_{gK})$. Using $\boldsymbol{\lambda}_g$ in the penalty criterion (4) specifies the weights we use within each cell.

Our algorithm chooses the tree and the vector $\boldsymbol{\omega}$ jointly using cross-validation. It is computationally expensive to make high dimensional searches. With K factors there is a $K-1$ dimensional space of weights to search. Adding in the tree size gives a K'th dimension. As a result, combining all of our estimators requires us to search a 4 dimensional grid of values.

We have chosen to set one of the ω_k to 0 to reduce the search space from 4 dimensions to 3. We always retained the unbiased estimate $\hat{\theta}_S$ along with two others. In some computations reported in section A.4 of the Appendix we find only small differences among setting $\omega_1 = 0$, or $\omega_2 = 0$ or $\omega_3 = 0$. The best outcome was setting $\omega_1 = 0$. That has the effect of removing the estimate based on IDA only. As we saw in section 3, the IDA-only model had the least potential to improve our estimate. As a bonus, all three of the retained submodels have the same sample sizes and then common λ over cells coincides with common ω over cells.

In the special case with $\omega_1 = 0$ we find after some calculus that the minimizer of (4) has

$$\hat{\beta}_c = \frac{\bar{V}_{0c} + \sum_{k\in\{2,3\}} \frac{\lambda_k}{1+\lambda_k} \bar{V}_{kc}}{1 + \sum_{k\in\{2,3\}} \frac{\lambda_k}{1+\lambda_k}} \equiv \sum_{k\in\{0,2,3\}} \omega_{kc}(\boldsymbol{\lambda}) V_{kc} \tag{5}$$

where $\bar{V}_{kc}$ is the simple average of $\widetilde{V}_k$ over $i \in S$ for cell c.

Our default grid takes all values of $\boldsymbol{\omega}$ whose coefficients are integer multiples of 10%. Populations D_0, D_2 and D_3 all have the sample size n and of these only D_0 is surely unbiased. An observation in D_0 is worth at least as much as an observation in D_2 or D_3 and so we require $\omega_0 \geqslant \max\{\omega_2, \omega_3\}$. Figure 5.1 shows this region and the set of 24 weight combinations that we use.

5.3 Search for tree size

Here we give a brief review of regression trees in order to define our algorithm. For a full description see the monograph by Breiman et al. (1985). The version we use is the function `rpart` (Therneau and Atkinson, 1997) in the R programming language (R Core Team, 2012).

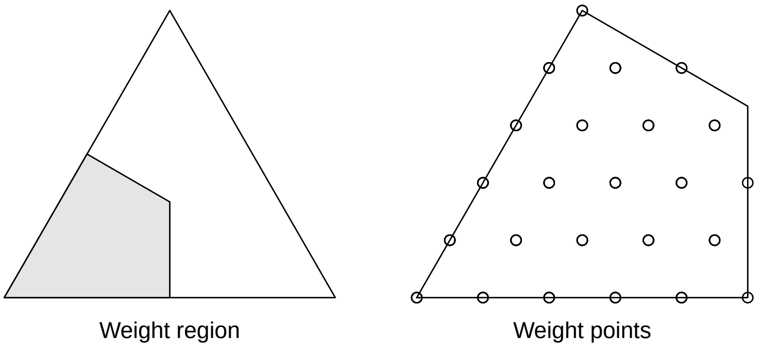

Fig. 5.1: The left panel shows the simplex of weights applied to data sets D_0, D_2 and D_3 with the unbiased data set D_0 in the lower left. The shaded region has the valid weights. The right panel shows that region with points for the 24 weights we use in our algorithm.

Regression trees are built from splits of the set of subjects. A split uses one of the features in X and creates two subsets based on the values of that feature. For example it might split males from females or it might split those with the two smallest education levels from the others. Such a split defines two subpopulations of our target population and it equally defines two subsamples of our sample.

A regression tree is a recursively defined set of splits. After the subjects are split into two groups based on one variable, each of those two groups may then be split again, using the same or different variables. Recursive splitting of splits yields a tree structure with subsets of subjects in the leaf nodes. Given a tree, we predict for subjects by a rule based on the leaf to which they belong. That rule uses the average within the subject's leaf node.

The tree is found by a greedy search that minimizes a measure of prediction error. In our case, the measure $R(T)$, is the sum of squared prediction errors. By construction any tree with more splits than T has lower error and this brings a risk of overfitting. To counter overfitting, `rpart` adds a penalty proportional to the number $|T|$ of leaves in tree T. The penalized criterion is $R(T) + \alpha|T|$ where the parameter $\alpha > 0$ is chosen by M-fold cross-validation. This reduces the potentially complicated problem of choosing a tree to the simpler problem of selecting a scalar penalty parameter α.

The `rpart` function has one option that we have changed from the default. That parameter is c_p, the complexity parameter. The default is 10^{-2}. The c_p parameter stops tree growing early if a proposed split improves $R(T)$ by less than a factor of c_p. We set $c_p = 10^{-4}$. Our choice creates somewhat larger trees to get more choices to use in cross-validation.

5.4 The algorithm

Here is a summary of the entire algorithm. First we make the following preprocessing steps.

1) Fit a large tree $\mathcal{T}$ by `rpart` relating observed incremental reaches V_i to predictor variables X_i in the SSP data. This tree returns a nested sequence of subtrees $T_0 \subset T_1 \subset \cdots \subset T_L \subset \mathcal{T}$. Each T_ℓ corresponds to a critical value α_ℓ of the penalty. Choosing α_ℓ from this list selects the tree T_ℓ. The value L is data-dependent, and chosen by `rpart`.

2) Specify a grid of values $\boldsymbol{\omega}_g$ for $g = 1, \ldots, G$. Here $\boldsymbol{\omega}_g = (\omega_{g0}, \omega_{g1}, \ldots, \omega_{gK})$ with $\omega_{gk} \geqslant 0$ and $\sum_{k=0}^{G} \omega_{gk} = 1$.

3) Randomly partition SSP data (X_i, Y_i, Z_i) into M folds S_m for $m = 1, \ldots, M$ each of roughly equal size n/M. For fold m the SSP will contain $\cup S_{m'}$ for all $m' \neq m$. We call this S_{-m}. The BRP for fold m is the entire BRP. We also considered using a bootstrap sample for the fold m BRP, but that was more expensive and less accurate in our numerical investigation as described in section A.4 of the Appendix.

After this precomputation, our algorithm proceeds to the cross-validation shown in Figure 5.2 to make a joint selection of the tree penalty parameter α_ℓ and the simplex grid point $\boldsymbol{\omega}_g$. Let the chosen values be α_* and $\boldsymbol{\omega}_*$. We select the tree T_* from step 1 above, corresponding to penalty parameter α_*. We treat each leaf node of T_* as a cell c. We translate $\boldsymbol{\omega}_*$ into the corresponding $\boldsymbol{\lambda}_c$ in every cell c of tree T_*. Then we minimize (4) using this $\boldsymbol{\lambda}_c$ and the resulting $\hat{\beta}_c$ is our estimate $\widehat{V}_c$ of incremental reach in cell c.

After choosing the tuning parameters $\boldsymbol{\omega}_g$ and α_ℓ by cross-validation, we use these parameters on the whole data set to make our final prediction.

6 Numerical investigation

In order to measure the effect of data enriched estimates on incremental reach, we conducted a simulation where we knew the ground truth. Our goal is to predict for ensembles, not for individuals, so we constructed two large populations in which ground truth was known to us, simulated our process of subsampling them, and scored predictions against the ground truth incremental reach probabilities.

To make our large samples realistic, we built them from our real data. We created S- and B-populations by replicating our SSP (respectively BRP) records 100 times each. Then in each simulation, we form an SSP by drawing 6000 observations at random from the S-population, and a BRP by drawing 13,000 observations at random from the B-population.

```
for $\ell = 1, \dots, L$ do   // initialize error sum of squares
    for $g = 1, \dots, G$ do
        $\mathrm{SSE}_{\ell,g} \leftarrow 0$
for $m = 1, \dots, M$ do   // folds
    construct Table 5.1 for fold $m$, using $S_{-m}$ and $B$
    fit tree $\mathcal{T}_m$ for fold $m$ by rpart
    prune tree $\mathcal{T}_m$ to $T_{1,m}, \dots, T_{L,m}$,  tree $T_{\ell,m}$ uses $\alpha_\ell$
    for $\ell = 1, \dots, L$ do   // tree sizes
        define cells $S_{-m,c}$ and $B_c$, $c = 1, \dots, C$ from leaves of $T_{\ell,m}$
        for $g = 1, \dots, G$ do   // simplex weights
            convert $\boldsymbol{\omega}_g$ into $\boldsymbol{\lambda}_g$
            for $c = 1, \dots, C$ do   // cells
                compute $\widetilde{V}_k$ for $k = 0, 2, 3$ in cell $c$
                get $\widehat{V}_c = \hat{\beta}_c$ from the weighted average (5)
                $V_c \leftarrow \frac{1}{|S_{m,c}|} \sum_{i \in S_{m,c}} V_i$   // held out incr. reach
                $p_c \leftarrow$ fraction of true S population in cell $c$
                $\mathrm{SSE}_{\ell,g} \leftarrow \mathrm{SSE}_{\ell,g} + p_c(\widehat{V}_c - V_c)^2$
```

Fig. 5.2: Data enrichment for incremental reach (deir) algorithm. After precomputation described on page 13 we run this cross-validation algorithm to choose the complexity parameter α_ℓ and the weights ω_g, as the joint minimizers ℓ^* and g^* of $\mathrm{SSE}_{\ell,g}$. The values p_c come from a census or from the SSP if the census does not have the variables we need. We use $M = 10$.

For each campaign, we apply deir with this sample data to estimate the incremental reach $\hat{V}(x)$. We used 10–fold cross-validation. The mean square estimation error (MSE) is $\sum_x p(x)(\hat{V}(x) - V(x))^2$. This sum is taken over all x values in the SSP.

The simulation above was repeated 1000 times. The root mean square error was divided by the true incremental reach to get a relative RMSE.

We consider two comparison methods. The first is to use the SSP only. That method computes $\hat{\theta}_S$ within the leaves of a tree. The tree is found by `rpart`. The second comparison is a tree fit by `rpart` to the pooled SSP and BRP data and using both CIA and IDA. We do not compare to the empirical fractions because many of them are from empty cells.

Figure 6.1 compares the relative errors in the SSP only method to data

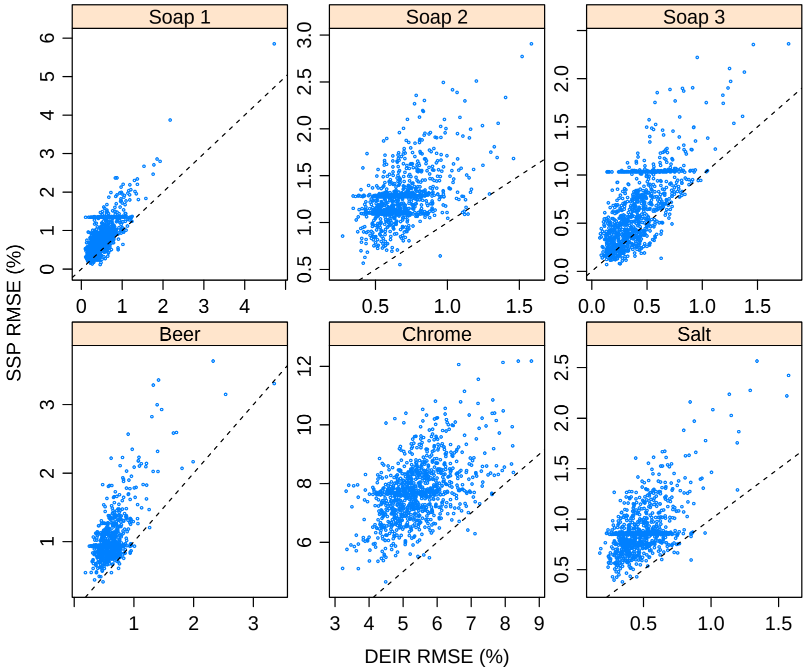

Fig. 6.1: Performance comparison, SSP only versus data enrichment, predictive relative mean square errors. There is one panel for each of 6 campaigns with one point for each of 1000 replicates. The reference line is the forty-five degree line.

enrichment. Data enrichment is consistently better over all 6 campaigns we simulated in the great majority of replications. It is clear that the populations are similar enough that using the larger data set improves estimation of incremental reach.

Under the IDA we can pool the SSP and BRP together using `rpart` on the combined data to estimate $\Pr(Z = 1 \mid X)$. Under the CIA we can multiply this estimate by $\Pr(Y = 0 \mid X)$ fit by `rpart` to the SSP, see Table 5.1 under the assumption CIA & IDA. This method, as an implementation of statistical matching, uses two separate applications of `rpart` each with their own built in cross-validation.

Figure 6.2 compares the relative errors of statistical matching to data enrichment. Data enrichment is consistently better over all 6 campaigns we simulated in the great majority of replications.

We also investigate for each estimator, how much of the predictive error is contributed by bias. It is well known that predictive mean square error can be decomposed as the sum of variance and squared bias. These quantities are typically unknown in practice, but can be evaluated in simulation studies.

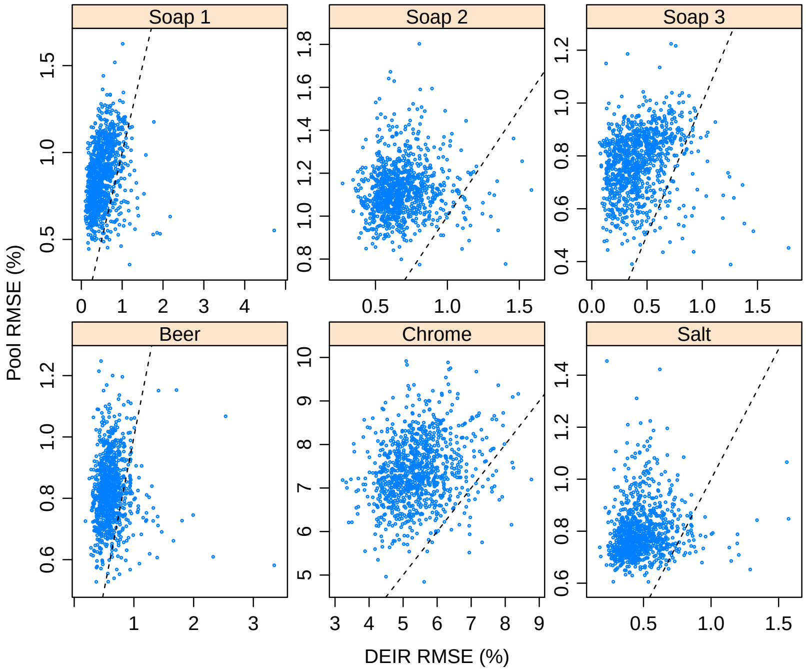

Fig. 6.2: Performance comparison, statistical matching (data pooling) versus data enrichment, predictive relative mean square errors. There is one panel for each of 6 campaigns with one point for each of 1000 replicates. The reference line is the forty-five degree line.

Table 6.1 reports the fractions of squared bias in predictive mean square errors for each method in all six studies. We see there that the error for statistical matching (data pooling) is dominated by bias while the error for SSP only is dominated by variance. These results are not surprising because the SSP only method has no sampling bias (only algorithmic bias) while the pooled data set has maximal sampling bias. The proportion of bias for DEIR is in between these extremes. Here we have less population bias than a typical data fusion situation because the TV and online-only panels were recruited in the same way. The bottom of Table 6.1 shows that DEIR is able to trade off bias and variance more effectively than SSP only or data pooling, because DEIR attains the smallest predictive mean squared error.

Conclusions

Predictions of incremental reach can be improved by making use of additional data. That improvement comes only if certain strong assumptions are true or at

bias^2/mse	Beer	Chrome	Salt	Soap 1	Soap 2	Soap 3
SSP	0.35	0.42	0.26	0.12	0.28	0.12
Pool	0.88	0.82	0.88	0.88	0.88	0.93
DEIR	0.49	0.59	0.47	0.33	0.47	0.39
mse	Beer	Chrome	Salt	Soap 1	Soap 2	Soap 3
SSP	1.02	7.76	0.89	0.84	1.26	0.66
Pool	0.82	7.39	0.80	0.86	1.12	0.78
DEIR	0.61	5.42	0.48	0.52	0.68	0.42

Tab. 6.1: The upper rows show the fraction bias^2/mse of the mean squared prediction error due to bias for 3 methods to estimate incremental reach in 6 campaigns. The lower rows show the total mse, that is $\text{bias}^2 + \text{var}$.

least approximately true. Our only guide to the accuracy of those assumptions may come from the data themselves. Our data enriched incremental reach estimate uses a shrinkage strategy to pool estimates using different assumptions. Cross-validating the level of pooling gave us an algorithm that worked better than either ignoring the additional data or treating it the same as the unbiased data.

Acknowledgment

This project was not part of Art Owen's Stanford responsibilities. His participation was done as a consultant at Google. The authors would like to thank Penny Chu, Tony Fagan, Yijia Feng, Jerome Friedman, Yuxue Jin, Daniel Meyer, Jeffrey Oldham and Hal Varian for support and constructive comments.

References

Breiman, L., Friedman, J. H. Olshen, R. A., and Stone, C. J. (1985). *Classification and Regression Trees.* Chapman & Hall/CRC, Baton Rouge, FL.

Chen, A., Owen, A. B., and Shi, M. (2013). Data enriched linear regression. Technical report, Google. http://arxiv.org/abs/1304.1837.

Collins, J. and Doe, P. (2009). Developing an integrated television, print and consumer behavior database from national media and purchasing currency data sources. In *Worldwide Readership Symposium, Valencia.*

Doe, P. and Kudon, D. (2010). *Data integration in practice: connecting currency and proprietary data to understand media use.* ARF Audience Measurement 5.0.

D'Orazio, M., Di Zio, M., and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. Wiley, Chichester, UK.

Gilula, Z., McCulloch, R. E., and Rossi, P. E. (2006). A direct approach to data fusion. *Journal of Marketing Research*, XLIII:73–83.

Jin, Y., Shobowale, S., Koehler, J., and Case, H. (2012). The incremental reach and cost efficiency of online video ads over TV ads. Technical report, Google.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*. Springer, New York, Third edition.

Little, R. J. A. and Rubin, D. B. (2009). *Statistical Analysis with Missing Data*. John Wiley & Sons Inc., Hoboken, NJ, 2nd edition.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rässler, S. (2004). Data fusion: identification problems, validity, and multiple imputation. *Austrian Journal of Statistics*, 33(1&2):153–171.

Singh, A. C., Mantel, H., Kinack, M., and Rowe, G. (1993). Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption. *Survey Methodology*, 19:59–79.

Stein, C. M. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, volume 1, pages 197–206.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151.

The Nielsen Company (2011). The cross-platform report. Quarter 2, U.S.

Therneau, T. M. and Atkinson, E. J. (1997). An introduction to recursive partitioning using the RPART routines. Technical Report 61, Mayo Clinic.

A Appendix

A.1 Variance reduction by IDA

Recall that $f = n/(n+N)$ and $F = N/(n+N)$ are sample size proportions of the two data sets. Under the IDA we may estimate incremental reach by

$$\hat{\theta}_{\mathrm{I}} = (f\bar{Z}_{\mathrm{S}} + F\bar{Z}_{\mathrm{B}})\frac{\bar{V}_{\mathrm{S}}}{\bar{Z}_{\mathrm{S}}} = \bar{V}_{\mathrm{I}}\Big(f + F\frac{\bar{Z}_{\mathrm{B}}}{\bar{Z}_{\mathrm{S}}}\Big).$$

By the delta method (Lehmann and Romano, 2005), $\mathrm{var}(\hat\theta_{\mathrm{I}})$ is approximately

$$\begin{aligned}\widetilde{\mathrm{var}}(\hat\theta_{\mathrm{I}}) &= \mathrm{var}(\bar V_{\mathrm{S}})\Bigl(\frac{\partial\hat\theta_{\mathrm{I}}}{\partial\bar V_{\mathrm{S}}}\Bigr)^2 + \mathrm{var}(\bar Z_{\mathrm{B}})\Bigl(\frac{\partial\hat\theta_{\mathrm{I}}}{\partial\bar Z_{\mathrm{B}}}\Bigr)^2 + \mathrm{var}(\bar Z_{\mathrm{S}})\Bigl(\frac{\partial\hat\theta_{\mathrm{I}}}{\partial\bar Z_{\mathrm{S}}}\Bigr)^2\\ &\quad + 2\mathrm{cov}(\bar V_{\mathrm{S}}, \bar Z_{\mathrm{S}})\frac{\partial\hat\theta_{\mathrm{I}}}{\partial\bar V_{\mathrm{S}}}\frac{\partial\hat\theta_{\mathrm{I}}}{\partial\bar Z_{\mathrm{S}}},\end{aligned}$$

with partial derivatives evaluated with expectations $\mathbb{E}(\bar V_{\mathrm{S}})$, $\mathbb{E}(\bar Z_{\mathrm{S}})$, and $\mathbb{E}(\bar Z_{\mathrm{B}})$ replacing the corresponding random quantities. The other two covariances are zero because the S and B samples are independent.

From the binomial distribution we have $\mathrm{var}(\bar V_{\mathrm{S}}) = \theta(1-\theta)/n$, $\mathrm{var}(\bar Z_{\mathrm{B}}) = p_z(1-p_z)/N$ and $\mathrm{var}(\bar Z_{\mathrm{S}}) = p_z(1-p_z)/n$. Also

$$\mathrm{cov}(\bar V_{\mathrm{S}}, \bar Z_{\mathrm{S}}) = \frac{1}{n}\bigl(\mathbb{E}(V_iZ_i) - \mathbb{E}(V_i)\mathbb{E}(Z_i)\bigr) = \theta(1-p_z)/n.$$

After some calculus,

$$\begin{aligned}\widetilde{\mathrm{var}}(\hat\theta_{\mathrm{I}}) &= \frac{\theta(1-\theta)}{n} + \frac{p_z(1-p_z)}{N}\frac{\theta^2F^2}{p_z^2} + \frac{p_z(1-p_z)}{n}\frac{\theta^2F^2}{p_z^2} - 2\frac{\theta(1-p_z)}{n}\frac{\theta F}{p_z}\\ &= \mathrm{var}(\hat\theta_{\mathrm{S}}) + \frac{\theta^2F(1-p_z)}{p_z}\Bigl(\frac{F}{N} + \frac{F}{n} - \frac{2}{n}\Bigr)\\ &= \mathrm{var}(\hat\theta_{\mathrm{S}}) - \frac{\theta^2F(1-p_z)}{p_z}\frac{1}{n}\\ &= \mathrm{var}(\hat\theta_{\mathrm{S}})\Bigl(1 - F\frac{1-p_z}{p_z}\frac{\theta}{1-\theta}\Bigr).\end{aligned}$$

A.2 Variance reduction by CIA

Applying the delta method to $\hat\theta_{\mathrm{C}} = \bar Z_S(1-\bar Y_S)$, we find that

$$\begin{aligned}\widetilde{\mathrm{var}}(\hat\theta_{\mathrm{C}}) &= \mathrm{var}(\bar Z_{\mathrm{S}})\Bigl(\frac{\partial\hat\theta_{\mathrm{C}}}{\partial\bar Z_{\mathrm{S}}}\Bigr)^2 + \mathrm{var}(\bar Y_{\mathrm{S}})\Bigl(\frac{\partial\hat\theta_{\mathrm{C}}}{\partial\bar Y_{\mathrm{S}}}\Bigr)^2 + 2\mathrm{cov}(\bar Y_{\mathrm{S}}, \bar Z_{\mathrm{S}})\frac{\partial\hat\theta_{\mathrm{C}}}{\partial\bar Y_{\mathrm{S}}}\frac{\partial\hat\theta_{\mathrm{C}}}{\partial\bar Z_{\mathrm{S}}}\\ &= \mathrm{var}(\bar Z_{\mathrm{S}})(1-p_y)^2 + \mathrm{var}(\bar Y_{\mathrm{S}})p_z^2 + 2\mathrm{cov}(\bar Y_{\mathrm{S}}, \bar Z_{\mathrm{S}})(1-p_y)p_z.\end{aligned}$$

Here $\mathrm{var}(\bar Z_{\mathrm{S}}) = p_z(1-p_z)/n$, $\mathrm{var}(\bar Z_{\mathrm{S}}) = p_z(1-p_z)/n$, and under conditional independence $\mathrm{cov}(\bar Y_{\mathrm{S}}, \bar Z_{\mathrm{S}}) = 0$. Thus

$$\begin{aligned}\widetilde{\mathrm{var}}(\hat\theta_{\mathrm{C}}) &= \frac{1}{n}\bigl(p_z(1-p_z)(1-p_y)^2 + p_y(1-p_y)p_z^2\bigr)\\ &= \frac{1}{n}\bigl(p_z(1-p_z)(1-p_y)^2 + p_y(1-p_y)p_z^2\bigr)\\ &= \frac{p_z(1-p_y)}{n}\bigl((1-p_z)(1-p_y) + p_yp_z\bigr).\end{aligned}$$

When the CIA holds, $\theta = p_z(1-p_y)$. Note that $\mathrm{var}(\hat\theta_{\mathrm{S}}) = \theta(1-\theta)/n$. After some algebraic simplification we find that

$$\frac{\widetilde{\mathrm{var}}(\hat\theta_{\mathrm{C}})}{\mathrm{var}(\hat\theta_{\mathrm{S}})} = 1 - \frac{p_y(1-p_z)}{1-\theta}.$$

A.3 Variance reduction by CIA and IDA

When both assumptions hold we can estimate θ by

$$\hat\theta_{\mathrm{I,C}} = (f\bar Z_{\mathrm{S}} + F\bar Z_{\mathrm{B}})(1-\bar Y_{\mathrm{S}}).$$

Under these assumptions, $\bar Z_{\mathrm{S}}$, $\bar Z_{\mathrm{B}}$ and $\bar Y_{\mathrm{S}}$ are all independent, and $\widetilde{\mathrm{var}}(\hat\theta_{\mathrm{I,C}})$ equals

$$\begin{aligned}
&\mathrm{var}(\bar Z_{\mathrm{S}})\Bigl(\frac{\partial\hat\theta_{\mathrm{I,C}}}{\partial\bar Z_{\mathrm{S}}}\Bigr)^2 + \mathrm{var}(\bar Z_{\mathrm{B}})\Bigl(\frac{\partial\hat\theta_{\mathrm{I,C}}}{\partial\bar Z_{\mathrm{B}}}\Bigr)^2 + \mathrm{var}(\bar Y_{\mathrm{S}})\Bigl(\frac{\partial\hat\theta_{\mathrm{I,C}}}{\partial\bar Y_{\mathrm{S}}}\Bigr)^2\\
&= \frac{p_z(1-p_z)}{n}f^2(1-p_y)^2 + \frac{p_z(1-p_z)}{N}F^2(1-p_y)^2 + \frac{p_y(1-p_y)}{n}p_z^2\\
&= \frac{p_z(1-p_y)}{n}\bigl(f(1-p_y)(1-p_z)+p_yp_z\bigr)
\end{aligned}$$

after some simplification. As a result

$$\frac{\widetilde{\mathrm{var}}(\hat\theta_{\mathrm{I,C}})}{\widetilde{\mathrm{var}}(\hat\theta_{\mathrm{C}})} = \frac{f(1-p_y)(1-p_z)+p_yp_z}{(1-p_y)(1-p_z)+p_yp_z}.$$

A.4 Alternative algorithms

We faced some design choices in our algorithm. First, we had to decide which estimators to include in our algorithm. We always include the unbiased choice $\hat\theta_S$ as well as two others. Second, we had to decide whether to use the entire BRP or to bootstrap sample it. We ran all six choices on simulations of all six data sets where we knew the correct answer. Table A.1 shows the mean squared errors for the six possible estimators on each of the six data sets. In every case we divided the mean squared error by that for the estimator combining $\hat\theta_S$, $\hat\theta_C$, and $\hat\theta_{I,C}$ without the bootstrap. We only see small differences, but the evidence favors choosing $\lambda_I = 0$ as well as not bootstrapping.

Our default method is consistently the best in this table, although only by a small amount. We saw that data enrichment is consistently better than either pooling the data or ignoring the large sample, and by much larger amounts than we see in Table A.1. As a result, any of the data enrichment methods in this table would make a big improvement over either pooling the samples or ignoring the BRP.

Estimators	$\hat{\theta}_S$, $\hat{\theta}_I$, $\hat{\theta}_C$		$\hat{\theta}_S$, $\hat{\theta}_I$, $\hat{\theta}_{I,C}$		$\hat{\theta}_S$, $\hat{\theta}_C$, $\hat{\theta}_{I,C}$	
BRP	All	Boot	All	Boot	All	Boot
Beer	1.02	1.02	1.00	1.01	1	1.01
Chrome	1.04	1.04	1.01	1.01	1	1.00
Salt	1.04	1.04	1.01	1.01	1	1.01
Soap 1	1.04	1.05	1.01	1.02	1	1.00
Soap 2	1.05	1.05	1.01	1.03	1	1.01
Soap 3	1.02	1.02	1.01	1.00	1	1.00

Tab. A.1: Relative performance of our estimators on six problems. The relative errors are mean squared prediction errors normalized to the case that uses $\hat{\theta}_S$, $\hat{\theta}_C$, $\hat{\theta}_{I,C}$ without bootstrapping. The relative error for that case is 1 by definition.