ALBANIAN JOURNAL OF MATHEMATICS
Volume 11, Number 1, Pages 35–71
ISSN: 1930-1235; (**2017**)

EVALUATION OF PROMOTIONAL CAMPAIGN EFFECTS WITH SELF-SELECTION OF PARTICIPATION - PROPENSITY SCORE APPLICATION

JIMMY CELA

Aspirent Consulting, LLC
400 Embassy Row NE 260, Atlanta, GA 30328

Abstract. Propensity score based methods applied to mitigate the bias in treatment effect estimation incurred by self-selection on observables, usually follow non-parametric matching approaches. Parametric estimation, performed by regressing solely on propensity scores, is suggested in theory, but is not generally applied. However, when appropriate, a parametric approach is preferable to a non-parametric or semi-parametric one as it provides more information, insight and inference on the same data set. We test parametric regression method through simulations, creating different scenarios of system-determined treatment assignment. It results that regressing only on propensity score, is not sufficient to properly mitigate the treatment estimation bias. We consider the propensity score as an omitted variable, which when added into the model, makes covariates and the binary treatment of interest conditionally independent. Propensity score enters the model as a generated regressor, because it is created in a separate modeling stage, and provides for unbiased and consistent estimation of treatment effects. This estimation is superior to the semi-parametric ones in our tests. Two real data with potential self-selection bias problems are analyzed to illustrate some application issues and to point out in particular the need for specific propensity scores application at any given situation.

Keywords: self-selection bias, propensity score, promotional campaign effect estimation, conditional independence, ignorability of treatment.

1. Introduction

A main problem in evaluating promotional campaign effect on individual responses is the violation of the sampling randomness principle. The campaign participation is not assigned at random to customers. It is rather a decision taken by them. Consequently, the estimate of campaign effect can, under given circumstances, be biased. Literature examples of applications correcting for self-selection bias are numerous and helpful (e.g., [1, 2] provide a good view on the issue), but they only work well within their context and, as pointed out by Heckman, "there is no context-free universal cure for the selection problem. There are as many cures as there are contexts" [3]. For the data analyst this usually creates both confusion

and the known dilemma about the usefulness of analysis after the results are delivered [4]. In fact, this confusion is a reflection of the discrepancies and disputes among the leading researchers regarding this topic (the book "Drawing inferences from self-selected samples" (2000) [3] is a clear illustration; see also Conniffe, Gash, and O'Connell for a view of disputes [5], or Smith and Todd [6], for a particular example). Meanwhile, promotional campaigns are in the order of the day in many business administration groups, which are in need for an established guideline of promotional campaign effects estimation.

The aim of this paper is to show our experiences in promotional campaign effect estimation, with some references to the hospitality industry. Many campaigns with different incentives are launched repeatedly over the business year. Their audience, including a loyalty program membership, is broad and campaign participation is not sporadic. Both participant and non-participant group sizes are considerable, in contrast with some clinical studies in which the self-selected treatment group is minuscule compared to the control. The nature of incentive directs the campaign appeal to certain groups of individuals, who for one reason or another, are more interested in what is offered. When exposed to the promotional offer, it is the individual that makes the decision to participate or not. We assume that both participants and non-participants have the same odds to be exposed to the offer. While conscious that this cannot hold all the time and for everybody true, daily advances of communication technology, through which these offers are conveyed, make this assumption not particularly strong. One basic campaign incentive is to offer some reward in exchange of some purchase. What the business administrators primarily want to know is the real effect of the campaign, which is the difference in sales between the actual figures and what those would have been in the absence of the promotion.

This conditional statement (both literally and statistically) translates to an estimation procedure adjustment, which is presented at some detail in the next section. In this paper we consider the indispensable assumption of treatment conditional independence and the related propensity score based methods, which usually estimate the average treatment effect with self-selection on observables, non-and-semi-parametrically [7]. Our aim is rather focused on the parametric method, i.e., regression analysis on propensity score (probability of participation given covariates), which shifts the idea of "conditioning on" into "regressing on" this variable.

We maintain a generalized linear model framework on a four-component system: response, covariates, treatment (which represents a promotional campaign in our context), and the generated propensity score. The participation decision model that generates the latter is outlined in its most common application, logistic regression, whose performance affects the overall analysis performance in an interesting way: logistic fit must be "good but not excellent".

The necessity of propensity score presence in the model is justified by its role in making the treatment of interest and covariates conditionally independent, as their co-dependence is a source of estimation bias. Meanwhile, the model based solely on generated propensity scores without covariates is, in fact, a regression on the latter already organized in a given functional form by the generating model, which does not allow for the parametric flexibility of a multivariate model.

To test the validity of our assumptions we simulate situations in which the dependence of the binary treatment variable on systematic components, i.e., the extent of self-selection, varies over a wide range, from almost completely deterministic to virtually stochastic, on a pretty wide interval of participation rates. In all these scenarios the known effect of treatment is maintained constant and independent, so that its estimation is not affected in any way by factors other than self-selection extent and participation rates of each scenario (controlled circumstances, like extensive simulations with ideal conditions and deviations from these remain the only verifying tool in estimation of treatment effects, in absence of social experiments; see, e.g., [8, 9]).

Simulation results show that adding the generated propensity score to a generalized linear model substantially increases the model ability to mitigate or eliminate bias in treatment effect estimation. In addition, propensity score adds a very important dimension to the interpretative power of the model, as it directly links the intention to participate in the promotional campaign with the gains from it, or in general terms, marketing with profit. This way of model conception should become a routine in estimation of campaign effects in observational data, not only for the possible bias in the effect estimate, but also as a way to get the proper insight and inferences in the two components of campaign success: participation rate and its gains at individual customer level. "Gains" here refer to the immediate return on campaign investment and not to the other forms of gain that are of long-term effect (e.g., building brand name or value). Having a "score" in the mean structure enables the customer segmentation directly, adding the categorical variable into the model in form of the separate intecepts for each segment, or in significant interaction with other covariates. This increases the model flexibility to explain the response variability by allowing the parameters to vary across segments.

We analyze two real data sets with promotional campaign to illustrate the benefits mentioned above, as well as some applicative issues.

2. Estimation of promotional campaign effect with non-random participation

Let y_{ik} denote the amount of sales (or "the performance") during the promotional period for customer "i", who participated in campaign or not as indicated by the participation indicator d: $d = k$, where $k = 1$ for participants and $k = 0$ otherwise. Population I of customers $\{i : i \in I\}$ is characterized by two conceptually distinct random variables, y_{i0} and y_{i1}. The conditional variable $y_{ik}|d$ implies four distinct subpopulations: $y_{i0}|d = 0$, $y_{i1}|d = 0$, $y_{i1}|d = 1$ and $y_{i0}|d = 1$, of which $(y_{i1}-y_{i0})|d = 1$ and $(y_{i1}-y_{i0})|d = 0$ are unobservable. The observed outcomes $y_{i(obs)} = (y_{i0}|d = 0, y_{i1}|d = 1)$ are expressed through the indicator d as:

$$y_{obs} = (1 - d) \cdot y_0 + d \cdot y_1 = y_0 + d \cdot (y_1 - y_0) \tag{1}$$

(henceforth the subscript i will be dropped, unless explicitly written). The random decision variable d is binomial; each individual decision on participation is a Bernoulli trial. (The effects of treatment with more than two distinct levels, be it categorical or even continuous, have been also studied [10, 11]). It is believed that the relation between d and y_{obs} is causal. The effect of d on y_{obs} is the average promotion effect (APE). The average promotion effect on participants APE1, is the effect of business

interest, and the expected value of the following conditional difference:

$$APE1 = E[y_1 - y_0|d = 1] \tag{2}$$

Analogously, $APE0 = E[y_1 - y_0|d = 0]$ estimates the change in response due to promotion for non-participants had they participated. The unconditional expectation $APE = E[y_1 - y_0]$ estimates the promotion effect on a randomly selected individual in I, participant or non-participant. Both APE0 and APE are not business relevant. Scenarios like "what would have happened if non-participants participated" ingrained in both APE0 and APE, calculate the possible gains at a better participation level, but in this case the promotion itself would not be the same. Under the given competition, it would either be with a better incentive, or better managed and conveyed to customers. APE is a "d probability weighted average" of APE1 and APE0:

$$APE = P[d = 1] \cdot APE1 + P[d = 0] \cdot APE0.$$

Clearly,

$$APE = APE0 + P[d = 1] \cdot (APE1 - APE0),$$

which shows that:

$$APE1 = APE0 \Leftrightarrow APE1 = APE0 = APE.$$

Another estimator, which relates the d effect with y level is "the quantile treatment effect" [12], which will not be treated here.

The events $d = 1$ and $d = 0$ are mutually exclusive at individual level. $y_1|d = 1$ and $y_0|d = 0$ are observable, or facts. Their counterparts, $y_0|d = 1$ and $y_1|d = 0$ cannot be observed; they are "counterfacts", providing a particular setting for causal inference [13, 14, 15]. The problems in estimation for such setting come from the counterfactual $y_0|d = 1$, which appears in the very basic relation (2):

$$APE1 = E[y_1 - y_0|d = 1] = E[y_1|d = 1] - E[y_0|d = 1]$$

Every cross sectional data, experimental or observational, faces this situation: the same experimental or observational unit cannot be observed simultaneously in both control (untreated) and treated state. In experiments, the treatment assignment is applied as a rule completely at random across units and the response variables (y_1, y_0) or covariates X (here defined as variables temporally prior to promotion [14]), do not affect treatment assignment. Therefore, we have y and d unconditionally independent: $P[d|y] = P[d]$.

Also, X and d are (by design) independent: $P[d|X] = P[d]$, or $y_k \perp d$. $y_k|d = k$ is a random sample from y_k. $E[y_k|d] = E[y_k]$ and most importantly, $(y_1 - y_0) \perp d$, so that:

$$APE1 = E[y_1 - y_0|d = 1] = E[y_1 - y_0] = APE \tag{3}$$

$APE1$ equals APE and $APE0$ if d is randomly assigned to individuals; $APE1$ in (3) in that case could be estimated without any correction for selection bias and non-treated individuals are readily controls for the treated. However, in observational data with self-selection, d is not unconditionally independent of y_k, i.e., $y_k|d = k$ is not a random sample from y_k, $E[y_k|d = k] \neq E[y_k]$ and most importantly, $E[y_1 - y_0|d = k] \neq E[y_1 - y_0]$. Symmetrically, $P[d|y] \neq P[d]$ and, when y is correlated with X, $P[d|X] \neq P[d]$.

Despite the crucial difference in variable independence, observational data follow the same analysis approach as experimental data. The untreated units are also used

in place of the counterfactual controls, but a straightforward use of them as control group, will not produce necessarily an unbiased estimator of treatment partial effect. The inequality $E[y_0|d=1] \neq E[y_0|d=0]$ incurs bias, which equals the difference between the two expectations. Identifying the counterfactual $E[y_0|d=1]$ by available social experimental data [16], serves as a reliable criterion of the truth for the otherwise unobservable counterfactual. This is a luxury the analysts normally do not have. The correlation between response and treatment is the first hurdle to overcome for unbiased and consistent estimation in observational data. This is realized by the conditional (on X) independence assumption or "ignorability of treatment" (henceforth IT) assumption as termed by Rosenbaum and Rubin (1983) in their seminal paper "The central role of the propensity score in observational studies for casual effects" [17]. This assumption states that conditional on X, d and y are independent. That is, citing Wooldridge (2002) [26], "even if (y_0, y_1) and d might be correlated, they are uncorrelated once we partial out X". This assumption is strong and requires d to be a deterministic function of (observable) covariates, which brings us to the other assumption term "selection on observables". IT, as a conditional independence assumption, requires:

$$P[d=1|y_0, y_1, X] = P[d=1|X] \tag{4}$$

The last term, framing a joint distribution of d and X, is called the propensity score:

$$P[X] \equiv P[d=1|X] = E[d|X] \tag{5}$$

A symmetrical expression of (4) is $P[y_0, y_1|d=1, X] = P[y_0, y_1|X]$. Loosely, the last expression is perceived as stratifying the observable $(q \times h)$ matrix $X = (x_1, x_2, \ldots, x_h)$ of a sample of size N, in q unique submatrices X_j, where $j \in [2, N]$ and $max(q) = N$, such that there is no difference in X within a given X_j. Let X_{jk} be a subset of X_j collecting observations with $d = k$. $q \times k$ groups of size n_{jk} are formed, such that $i \in group_{jk} \Leftrightarrow X_{ijk} = X_{jk}$. The respective responses y_{ijk} are then random draws from y_{jk}. Variables in $\{y_j, X_j, d_j\}$ are unconditionally independent. Even if y_{obs} is a deterministic function of X: $E[y_{obs}] = f(X)$ for some function f, within any given X_j, X cannot be the source of the differences in y_{ijk}; the only observable difference source is d. So, $E[y_1|X_j, d=1] - E[y_0|d=0] = E[y_{j1}] - E[y_{j0}] = \alpha_j$ estimates the effect of d given X_j. The practical problem with this estimation rests in getting reasonable subsets X_j. If at least one of the elements $x_g, g = 1, \ldots, h$, is continuous, then $max(q) = N$, and $q = N \Leftrightarrow max(n_{jk}) = 1$. Also, it is well possible that $n_{jk} = 0$ for some j. The latter makes $E[y_{jk}]$ inestimable. If x_g in X are all discrete variables, then $q = \prod_{g=1}^{h} l_g$, where l is the number of levels of g^{th} variable. The exact matching technique is based on the above logic. More generally, it is sufficient to group observations so that distribution of X is not statistically different within each group, which is a much more relaxed condition than exact matching. In order to have reliable estimates of $E[y_{jk}]$, a certain sample size n_{jk} is necessary, which translates into q getting smaller. Both components $E[y_{jk}]$ and the respective q, should be tested, as $H_0 : E[y_{jk}] = \mu_{yjk} + |c|$, for some c, where μ_{yjk} is estimated by $E[y_{jk}]$, i.e., the mean of $(y|X_j, d=k)$. The relation between sample size and detection power of the test represented by $|c|$, for a non-ratio variable is $n_{jk} = c^{-2} \cdot \left(z_{\frac{\alpha}{2}} + z_\beta\right)^2 \cdot \sigma_{jk}^2$, where $z_{\frac{\alpha}{2}}$ and z_β are standard normal variables with cumulative density function (cdf) equal to $(1-\frac{\alpha}{2})$ and $(1-\beta)$,

respectively, α is the significance of the test (probability of rejecting H_0 when H_0 is true is not larger than α), β is the power of the test (probability of accepting H_0 when H_0 is false is not larger than β), and σ^2_{jk} is the true variance of y_{jk}. Reasonably small $|c|$ are desirable, but this may require n_{jk} that cannot maintain equal distributions of X_{j1} and X_{j0}. Any increase in h decreases n_{jk} able to realize equal distributions of X_{q1} and X_{q0}. Small n_{jk} produce non-reliable $E[y_{jk}]$ (i.e., unacceptably large $|c|$) and at $n_{jk} = 0$, $E[y_{jk}]$ is inestimable. This is the "curse of dimensionality". Propensity score came as a solution to this problem: it reduces the dimension of matching from h to 1, and y stratification is based on $P[X]$ in place of X. Rubin and Rosenbaum [17] showed that $(y_1, y_0) \perp d|X \Rightarrow (y_1, y_0) \perp d|P[X]$. The proof can be derived by iterated expectations [25]. We prefer to give the following proof, which is probability-based. If ignorability of treatment holds, then:

$$y \perp d|X \Leftrightarrow P[y|d, X] = P[y|X],$$

or based on Dawid's symmetry rule of conditional independence [24], $y \perp d|X \Leftrightarrow P[d|y, X[= P[d|X]$.

Conditioning on $P[X]$: $P[d|y, p[X]] = P[d|y, P[d|X]] = P[d|y, P[d|y, X]] = P[d|y] \Leftrightarrow d \perp y|X \Leftrightarrow y \perp d|X$. The use of propensity score is not a "stand alone" solution. It does not work well if IT assumption is violated. The violation extent, quantified by the differences $\Delta_d = P[y|d, X] - P[y|X]$ or $\Delta_d = P[d|y, X] - P[d|y]$, depends on X quality of information. Practically, it is more reasonable to expect mitigation than elimination of bias. The trade-off between the observable multidimensional X and the one-dimensional $P[X]$ is that we do not observe the latter. What we do not observe, we estimate. A consistent estimator of $P[X_j]$ is $\widehat{P}[X_j] = \dfrac{n_{j1}}{n_{j1} + n_{j0}}$. The size of n_{jk} poses again the problem of reliability of $\widehat{P}[X_j]$. Moreover, the IT assumption itself does not leave much of a choice; it is indeed indispensable for unbiased estimation. Principally, the unconditional independence among variables, which does not hold on the whole data set, is assumed to hold in data subsets. Rosenbaum and Robin defined propensity score as "the coarsest balancing score", and a balancing score as "a function $b(X)$ of the observable covariates X such that the conditional distribution of X given $b(X)$ is the same for the treated and control units": $P(X|p(X), d = 1) = P(X|p(X), d = 0)$, or $d \perp X|p(X)$ [17]. Intuitively, by grouping on $P[X]$, we try to filter out the role of X in y variability, keeping only that of d. Conditioning on propensity scores realizes the independence of treatment on (i) response y, and (ii) information X. $\widehat{P}[X]$ as function of X can take on as many distinct values as $v : v \in [max(l_{xk}), N]$, where l_{xk} is the number of distinct levels for covariate x_j in $X, j = 1, \cdots, k$; a continuous x_k can take on up to N different values. Therefore the exact matching on propensity scores might become impossible.

The estimation of $P[X]$ groups can be implemented in analogy to group matching. Let g denote the number of $\widehat{P}[X]$ strata, and stratum q an interval of $\widehat{P}[X]$ values $(p_{q1}, p_{q2}), q = 1, \cdots g$. $\Delta = p_{q2} - p_{q1}$ is the caliper breadth, and $i \in q$ if $\widehat{P}[X] \in [p_{q1}, p_{q2}]$. While all individuals sharing the same X_i are in the same group q, the reverse is not necessarily true: not all $\widehat{P}[X] \in [p_{q1}, p_{q2}]$ stem necessarily from the same X. It is even possible that exactly the same value of $\widehat{P}[X]$ is generated by more than one subset X_k of X: $\widehat{P}[X_i] = \widehat{P}[X_j] = \cdots = \widehat{P}[X_p]$, where $i \neq j \neq k \neq \cdots \neq p$. In a given stratum q might well reside differently distributed

subsets of X. However, the distributions of $X_{q1} = X_q|d = 1$ and $X_{q0} = X_q|d = 0$ are expected not to differ statistically. The following is a simple illustrating example (more details on a reproducible SAS code is given in the Appendix). Three discrete independent variables with each three levels "A", "B" and "C" determine, through a latent variable, a binary outcome. The propensity scores are calculated and presented in Table 1.

TABLE 1. Simulated propensity scores on 3 discrete variables (1, 2, 3) with 3 levels each (A, B, C).

Group "q" (No.)	Estimated P[Xq]	Variable Classes	Frequency	Total New Group Frequncy	New Group Participants (d = 1)
1	0.053	C,B,C	95	95	5
2	0.123	A,A,A	349	349	43
3	0.13	C,C,C	576	756	99
4	0.133	C,B,B	180		
5	0.15	B,A,A	220	220	33
6	0.238	B,B,B	450	450	107
7	0.315	B,A,B	130	130	41

The possible number of distinct $\widehat{P}[X_p]$ values is $3^3 = 27$, but only seven are observed. In each of the seven groups we hope $y_k|d = k$ to be a random draw from y_k. The independence of X and d given $\widehat{P}[X_p]$ is guaranteed: $P[d = 1|Group = q, \widehat{P}[X_q]] = P[d = 1|\widehat{P}[X_q]]$. We assume IT: $P[y_{qk}|d = k, \widehat{P}[X_q]] = P[y_{qk}|\widehat{P}[X_q]]$. The size of "treated subgroup" of group 1, i.e., "promotion participants", is expected to be $95 \cdot 0.053 \approx 5$. With such a small sample size of participants, the inference on this group would not be reliable, even if IT assumption holds. To reduce the number of propensity score strata, we could join together groups 3 and 4. $\widehat{P}[X_q] \approx 0.13$ acts as balancing score, because in the new group with X made of "C,C,C" and "C,B,B", the distribution of X will be statistically the same, disregarding d. In concrete figures: $P[X = "C, C, C"|NewGroup, d = 1] = \dfrac{576 \cdot 0.130}{99} \approx 0.75$, which is approximately the same as $P[X = "C, C, C"|NewGroup, d = 0] = \dfrac{576 \cdot (1 - 0.130)}{756 - 99} \approx 0.76$. We would add groups 2 and 5 as well, if the used $\widehat{P}[X_q]$ group interval were, say, $[0.10, 0.20)$. The same probabilities calculated above would be 42.8% and 43.6%, respectively, which are still pretty close. If we add group 6, these probabilities become 23.8% and 34.3%, respectively, which is not as close any more. The balancing property of propensity score is easily testable by conducting ANOVA analysis, where x_g is the dependent variable and the propensity score group is the independent categorical variable (with as many classes as groups). IT assumption, though, cannot be tested straightforwardly. To reach balanced X, the

strata might need to get narrower. Also, fine-tuning matching within a caliper with the help of additional adjustments (e.g., Mahalanobis distance between X elements [18]) is used to account for unbalanced X). In the oversimplified example above there was no propensity score, whose corresponding observations were not observed at both treated and untreated states. Practically, (in particular when at least one element of X is continuous) we encounter unmatched observations of given propensity score values falling everywhere in the ranked vector of propensity scores. Heckman et al. [16] define three components of bias based on X for participants and non-participants: (i) difference in support, (ii) difference in distribution and (iii) selection bias at common values of X. Matching on balanced X or trimming out parts of X that are not in common supports eliminates the first two components, while the third one remains. Non-exact matching and the fact that propensity score methods deal only with selection on observables, while selection on unobservables might be quite an important source of bias, are main reasons why propensity score method has been often (sharply) criticized [3, 5, 19, 20]. However, there is interesting evidence that the propensity scores can correct for selection bias on unobservables as well [21]. The additional assumption introduced by Rosenbaum and Rubin [17],

$$0 < P[X] < 1, \forall X \tag{6}$$

that turns the "ignorability of treatment" assumption into the "strong ignorability of treatment", states that for any given X there exists at least one individual, whose participation decision differs from that of the other individuals sharing the same X. However, in order to allow for a certain number of participants ($P[X]$ near 0) or non-participants ($p[X]$ near 1), a more reasonable formulation would be: $0 < m < P[X] < 1-m$, for some m. Implementation of conditioning on propensity score to estimate casual treatment effect leads to a relatively simple semi-parametric two-stage procedure: (i) Compute $\widehat{P}[X]$; (ii) stratify on intervals of ranked $\widehat{P}[X]$, change these intervals until no statistical difference between $\widehat{P}[X]$ middle scores and X of participants and non-participants is reached within each interval (caliper), and get $\widehat{APE}$ as the average difference between y means of treated and untreated observations. To get $\widehat{APE1}$, apply weighted average of differences, where weights equal the number of participants in each caliper [22, 19, 23, 25]. $\widehat{P}[X]$ is also used as a tool (weight) in other semi-parametric APE estimation methods. When conditioning on d and X, the semi-parametric formulas will contain the conditional density of d, which makes $\widehat{P}[X]$ a common term therein. Four semi-parametric formulas for APE1 estimation follow, as proposed by:

A. Wooldridge [26]:

$$\widehat{APE1} = \frac{\sum_{i=1}^{N} \frac{y_i \cdot (d_i - \widehat{P}[X]}{1-\widehat{P}[X]}}{\sum_{i=1}^{N} d_i}$$

B. Hirano, Imbens and Ridder [27]

$$\widehat{APE1} = \frac{\sum_i \widehat{P}[X_i] \cdot \left[\frac{y_i \cdot d_i}{\widehat{P}[X_i]} - \frac{y_i \cdot (1-d_i)}{1-\widehat{P}[X_i]}\right]}{\sum_i \widehat{P}[X_i]}$$

C. Ridgeway, McCaffrey, Morral and Lim [28]:

$$\widehat{APE1} = \frac{\sum y_1}{N_{y1}} - \frac{\sum y_0 \cdot w_0}{\sum w_0}$$

where

$$w_0 = (1-d) \cdot \frac{\widehat{P}[X]}{1-\widehat{P}[X]}$$

D. Hirano and Imbens [30]

$$\widehat{APE1} = \frac{\sum_{i=1}^{N} \frac{d_i \cdot y_i}{\widehat{P}[X_i]}}{\sum_{i=1}^{N} \frac{d_i}{\widehat{P}[X_i]}} - \frac{\sum_{i=1}^{N} \frac{(1-d_i) \cdot y_i}{1-\widehat{P}[X_i]}}{\sum_{i=1}^{N} \frac{1-d_i}{1-\widehat{P}[X_i]}}$$

The performance of these estimators is given in Table 8 (see Appendix). Parametric estimation of $\widehat{P}[X]$ and of $\widehat{APE1}(X)$ or $\widehat{APE}(X)$, involves related functional forms of X.

$$APE(X) = \int (E[y_1|X, d=1] - E[y_1|X, d=0]) dF_X(x)$$

and

$$[APE1(X) = \int (E[y_1|X, d=1] - E[y_0|X, d=0]) dF_{X|d=1}(x) \tag{7}$$

where F_X is the cdf of X [29]. Assuming IT and with $\widehat{P}[X] = E[d|X] = g(X)$, it results:

$$\begin{aligned} APE1(X) &= E_{X|d=1}\Big[E[y_1 - y_0|X, d=1\Big] = \\ &= E_X\Big[E[y_1 - y_0|X\Big] = \int E[y_1 - y_0|X] dF_{X|d=1})x) \\ &= \int f'(X) dF_{X|d=1}(x) = f(X) \end{aligned}$$

and

$$\begin{aligned} APE1(\widehat{P}[X]) &= E_{\widehat{P}[X]|d=1}\Big[E[y_1 - y_0|\widehat{P}[X], d=1\Big] \\ &= E_{\widehat{P}[X]}\Big[E[y_1 - y_0|\widehat{P}[X]\Big] = \int E[y_1 - y_0|\widehat{P}[X]] dF_{\widehat{P}[X]|d=1})(\widehat{P}[X]) \\ &= \int h'(\widehat{P}[X] dF_{\widehat{P}[X]|d=1}(\widehat{P}[X]) \\ &= \int h''(X) dF_{X|d=1}(g(x)) = h(X) \end{aligned}$$

for some functions g, f', f'', h', h'' and h. $\widehat{APE}$ and $\widehat{APE1}$ as functions of X or $\widehat{P}[X]$ vary across individual groups sharing different X or $\widehat{P}[X]$. Not only (y_1, y_0), but their difference $(y_1 - y_0)$ as well, is a function of X. Under the assumption of constant APE over the whole population, $E[y_1|X] = E[y_0|X] + APE$ and $APE = APE1$. The last assumption is very relaxed, but also very convenient for parametric modeling, leading to switching regression:

$$g(E[y_k|X, d=k]) = h(X) + \widehat{\alpha} \cdot d \tag{8}$$

for some functions h and g, where $\widehat{\alpha}$ is the parameter of interest. A widely used application of (8) is the generalized linear model:

$$g(E[y]) = l(X) + \widehat{\alpha} \cdot d \tag{9}$$

where $l(X) = \sum_{i=0}^{k} b_i \cdot f_i(X)$ is a linear combination of X, $f_0(X) = 1$ to guarantee an intercept, $f_i(X)$ is any function of X, g is some function, like identity, log etc., and $\widehat{APE1} = g^{-1}\Big(l(X)+\widehat{\alpha}\Big) - g^{-1}\Big(l(X)\Big)$. Note that $l(X)$ is linear in the coefficients of $f_i(X)$, whereas $f_i(X)$ can take any form, linear or non-linear in X. Model in (8) is referred to as the "kitchen sink regression" [26]. Its functional form h(X) is very flexible; this gives "kitchen sink regression" virtually the maximum prediction power for the given X. At the same time it is not exactly a convenience experimenting endlessly many forms of $h(X)$ until the "desired" result is obtained, which can add to the model a heavy dose of subjectivity. Based on the IT assumption, in order to to make y and d independent, conditioning on $P[X]$ is as good as conditioning on X. In analogy to matching based on $\widehat{P}[X]$, which substitutes for that on X, regressing on $\widehat{P}[X]$ is suggested as alternative to "kitchen sink regression" in two forms: regressing y_i on

$$1, d_i, \widehat{P}[X] \tag{10}$$

or

$$1, d_i, \widehat{P}[X], d_i \cdot (\widehat{P}[X] - \widehat{\mu_p}) \tag{11}$$

where $\widehat{\mu_p}$ is the mean of $\widehat{P}[X]$ [26], making two very strong assumptions: that $APE(X) = APE1(X)$, and that they are constant across X. Note that in the semi-parametric estimation of APE or APE1, it is not assumed that these are equal and constant across strata. The parameter estimate of d_i is expected to be a consistent $\widehat{APE}$ and $\widehat{APE1}$ [26].

Procedures (10) and (11) suggest a two-stage model, where the first stage estimates $p(X)$. Modeling $P[X]$ as the probability of participation conditioned on X can be realized by different models and assumptions. For example, Ridgeway suggests boosting algorithms (see his dissertation thesis [31] and other publication titles by this author), whereas Minkin [32] suggests semi-parametric methods using support vector machines. The usual parametric way models d through an underlying latent variable ν:

$$\nu_i = X\Gamma + \upsilon_i, d_i = I(\nu > 0) \tag{12}$$

where Γ is a vector of coefficients $\gamma_1, \gamma_2, \cdots, \gamma_k$, υ is a random error term and $I(\Delta)$ is the indicator function showing that d = 1 if $\nu > 0$, and $d = 0$ otherwise. The random component ν_i makes the decision process stochastic. It allows for introducing the individual specific unobserved characteristics that affect decision. The condition $\nu > 0$ is equivalent to $X\Gamma > -\upsilon$. The last inequality shows the weight of observables and unobservables, respectively, in explaining selection individually. The assumption that υ follows a standard logistic distribution is the most popular one [26]:

$$P[d = 1|X) = \widehat{P[X]} = P[\upsilon > 0] = P[\nu > -X\Gamma) = 1 - G(-X\Gamma) = G[X\Gamma) \tag{13}$$

where G is cdf of ν.

The "competition" between $X\Gamma$ and ν in explaining d reflects the extent at which the data can predict participation. A "good" data set in the sense of participation prediction does not let much to be explained by the unobservable ν, making d a deterministic function of X. IT assumption is satisfied. Otherwise, the participation is determined by the unobservable ν. The models in (10) and (11) include $\widehat{P[X]}$ as an independent variable and become very attractive from the business managerial perspective. While X often bears information with only descriptive value, the propensity of participation $P[X]$ acquires an economically well-defined meaning: it scores the activity rate of customers and put the latter in relation with the customer performance. The model in (11) can be written as:

$$f(E[y]) = \widehat{\lambda_0} + \widehat{\lambda_1} \cdot \widehat{P[X]} + \hat{\lambda_2} \cdot \left(\widehat{P[X]} - \mu_p\right) + \widehat{\lambda_3} \cdot d \tag{14}$$

$\widehat{\lambda_1}$ estimates the partial effect of $\hat{P[X]}$ on $f(E[y])$, that is the relationship between customer performance and their activation behavior. $\widehat{\lambda_2}$ depends on the change of customer performance across the propensity scores range. It allows for shape flexibility in $f(E[y])$ curve. The variable of interest $\widehat{\lambda_3}$, is the difference in $f(E[y])$ for two customers with the same $\widehat{P[X]}$ but different participation decision d. This condition is reached when logistic regression "does not work very well", as it erroneously predicts equally two different outcomes. A hypothetical perfectly working logistic regression, which predicts right every single observation, would mean failure to propensity scores method. A perfect participation prediction does not allow any estimation of counterfactual $E[y_0|d = 1]$ based on real data. As Heckman points out [33], missing data (unobserved counterfactuals) give rise to the problem of casual inference, but missing data (unobservables υ) are also required to solve the problem of casual inference.

3. Parametric estimation of APE1

Estimating $\widehat{P[X]}$ with a logistic regression, as given in (13), leads to the functional form: $\widehat{P[X]} = \dfrac{exp(\widehat{X\Gamma})}{1 + exp(\widehat{X\Gamma})}$, where X is the matrix of covariates and $\widehat{\Gamma}$ is the respective vector of parameter estimates. As such, the models in (10) and (12) are versions of "kitchen sink regression". For example the model in (10) turns out nonlinear in X coefficients:

$$\begin{aligned} E[y|X] &= \widehat{b_0} + \widehat{b_1} \cdot \widehat{P[X]} + \widehat{b_2} \cdot d \\ &= \widehat{b_0} + \widehat{b_1} \cdot \frac{exp(\widehat{X\Gamma})}{1 + exp(\widehat{X\Gamma})} + \widehat{b_2} \cdot d \\ &= \frac{\widehat{b_0} + (\widehat{b_0} + \widehat{b_1}) \cdot exp(\widehat{X\Gamma})}{1 + exp(\widehat{X\Gamma})} + \widehat{b_2} \cdot d \\ &= \frac{\widehat{b_0} + \widehat{c} \cdot exp(\widehat{X\Gamma})}{1 + exp(\widehat{X\Gamma})} + \widehat{b_2} \cdot d \end{aligned}$$

where $\widehat{c} = (\widehat{b_0} + \widehat{b_1})$.

The expression above is not necessarily the best functional form of X as a control function. $\widehat{\Gamma}$ vector is produced by an equation not related to y; we could get the

same $\widehat{\Gamma}$ for quite different y. Therefore regression on $\widehat{P[X]}$ lacks in flexibility. This is the price paid to dimension reduction.

Intuitively, condition on both X and $\widehat{P[X]}$ makes the model more flexible. Because the generalized linear models are by and large the most frequently used models, we will focus on the model (9): $g(E[y]) = l(X) + \widehat{\alpha} \cdot d$. Assuming IT, there are no simultaneity issues in the model. However, X and d are co-dependent: $E[d|X] \neq E[d]$. This might be a bias source in APE estimation. In analogy with the conditional independence between y and d given X, we seek a third variable, in presence of which X and d are conditionally independent. The best candidate, as mentioned above, is $\widehat{P[X]}$: $P(d|X, \widehat{P[X]}) = P(d|\widehat{P[X]})$, i.e., $d \perp X|P(d|X)$; $P(d|X, \widehat{P[X]}) = P(d|X)$, i.e., $d \perp \widehat{P[X]}|X$; $P[\widehat{P[X]}|X, d] = P[\widehat{P[X]}|X)$, i.e., $X \perp \widehat{P[X]}$. In the multivariate set $\{X, d, \widehat{P[X]}\}$, all components are conditionally independent, and each plays its role in the new model:

$$g(E[y] = l(X) + \gamma \cdot \widehat{P[X]} + \widehat{\alpha} \cdot d \tag{15}$$

X is the finest balancing score in Rosenbaum and Rubin definition [17]; d is the variable of interest; $\widehat{P[X]}$ makes X and d conditionally independent; IT assumption makes y and d independent through the presence of X. The equation in (15) becomes non-linear in X coefficients:

$$\begin{aligned} g(E[y]) &= l(X) + \widehat{\gamma} \cdot \widehat{P[X]} + \widehat{\alpha} \cdot d \\ &= \sum_{i=1}^{k} x_i \cdot \widehat{b_i} + \widehat{\gamma} \cdot \frac{exp\Big(\sum_{i=1}^{k} x_i \cdot \widehat{\delta_i}\Big)}{1 + exp\Big(\sum_{i=1}^{k} x_i \cdot \widehat{\delta_i}\Big)} + \widehat{\alpha} \cdot d \\ &= \frac{\sum_{i=1}^{k} x_i \cdot \widehat{b_i} + exp\Big((\sum_{i=1}^{k} x_i \cdot \widehat{\delta_i}) + ln(\sum_{i=1}^{k} x_i \cdot \widehat{b_i})\Big) + \gamma \cdot exp\Big(\sum_{i=1}^{k} x_i \cdot \widehat{\delta_i}\Big)}{1 + exp\Big(\sum_{i=1}^{k} x_i \cdot \widehat{\delta_i}\Big)} \end{aligned}$$

This nonlinear model is much more involved than the generalized linear model (15), which produces the same $\widehat{\alpha}$. Generalized linear model in (9) is a model with an omitted variable [34], whereas the proposed model (15) is a generated regressor model [35]. What is missing in (9) is a variable responsible for the relationship between the inclination to participate and the response, when this exists. The estimation bias is incurred if the absent variable is co-dependent with both response and participation status. Inclination to participate enters the equation as a scaled variable. The propensity score scales it from 0 to 1 at individual level. It characterizes the X groups and not the individual, bearing the attributes of a categorical variable, although it is a continuous one. Its partial effect in the model, after controlling for all covariates and participation, shows how the individual performance is related with the group inclination to participate, or which strata of customers are more interested in the campaign - a main concern for business managers for several reasons, one of which is giving incentives to the groups of customers that would have "performed" anyway. The $\widehat{P[X]}$ partial effect is expected to differ with the levels of $\widehat{P[X]}$ itself. This expectation is not reflected in the proposed model (15). Also, it is reasonable to think that the promotion effect also varies across $\widehat{P[X]}$

strata. Therefore, the model (15) can be adjusted to:

$$g(E[y]) = l(X) + \widehat{\gamma_C} \cdot I_C \cdot \widehat{P[X]} + \widehat{\alpha_C} \cdot I_C \cdot d \tag{16}$$

where I_C is an indicator variable of stratum C, which groups individuals based on their $\widehat{P[X]}$ (calipers). Then,

$$\hat{APE1}_C = g^{-1}\Big(l(X) + \gamma_C \cdot I_C \cdot \widehat{P[X]} + \widehat{\alpha_C} \cdot I_C \cdot d\Big) - g^{-1}\Big(l(X) + \gamma_C \cdot I_C \cdot \widehat{P[X]}\Big)$$

and

$$\widehat{APE1} = \sum_{C=1}^{p} \frac{APE1_C \cdot N_{C|d=1}}{N_1}$$

where $N_{C|d=1}$ is the number of participants of group C and N_1 is the total number of participants. $\widehat{P[X]}$ is also as a measurement of the randomness extent in participation decision process. In a completely random participation, d is X independent:

$$P(d = 1|X) = P[d] = \widehat{P[X]} = E[d|X] = \widehat{\mu_p}.$$

The distance $|\widehat{P[X_i]} - \widehat{\mu_p}|$ is proportional to this randomness. Note the use of this difference in (11). Plugging it in (15) renders:

$$g(E[y]) = l(X) + \widehat{\gamma} \cdot \Big(\widehat{P[X]} - \widehat{\mu_p}\Big) + \widehat{\alpha} \cdot d = l'(X) + \widehat{\gamma} \cdot \widehat{P[X]} + \widehat{\alpha}\cdot$$

where the constant $\widehat{\gamma} \cdot \widehat{\mu_p}$ is absorbed in the intercept term of $l'(X)$. Therefore we use $\widehat{P[X]}$ instead of $|\widehat{P[X_i]} - \widehat{\mu_p}|$ in (15); parameter estimates remain the same.

4. Logistic regression performance and propensity scores method

Propensity score methods require the fit of participation decision model not to be excellent. Some error in prediction is needed so that both participants and non-participants share the same $\widehat{P[X]}$. The main information derived by fit statistics in the decision model, is about the extent the campaign participation is a function of observables. A random decision for a participation model means the inability of available X to explain d. Decision itself might well be non-random. If factors that determine it are (partially) unobservable, then decision to participate will manifest its share of randomness. As such, the way we perceive the participation decision is quite a bit data dependent. The less information at individual level (e.g., lack of continuous variables), the less deterministic the outcome of the model. Here follows a brief review on logistic regression performance, which is currently the most common method in estimating $\widehat{P[X]}$. A performance index list can be found in [36]. Here follows a part of them with the respective authors:

A. ϕ_1 = 1- log(Lu)/ log(Lc) (McFadden,1974)
B. ϕ_2 = 1- (Lc/Lu)2/N (Cragg and Uhler, 1970)
C. ϕ_3 = 1 - (log(Lu)/ log(Lc))-(2/N)logLc (Estrella, 1998)
D. Maximum Rescaled R^2: $\phi_4 = \dfrac{\phi_2}{1 - Lc^{\frac{2}{N}}}$ (Nagelkerke , 1991) [37], where Lu and Lc are the unconstrained and constrained likelihoods, respectively, and N is the sample size. Correlation coefficient between binary response r and respective probability prediction $\widehat{r}$, has also been used:

E. $\phi_5 = \rho^2(r, \widehat{r})$ (Morrison, 1972; Goldberger 1973). The "c" statistic is another widely used index, which deserves special attention in applying propensity scores method:

F. $c = \dfrac{n_c + 0.5 \cdot (t - n_c - n_d)}{t}$, where t is the number of pairs of observations with different responses, n_c of which are concordant and the rest n_d discordant. This index equals the area under the curve of "receiver operating plot" (ROC), the plot of "Sensitivity" against "1 - Specificity". Sensitivity(z) is the ratio between the correctly predicted event responses and the actual number of events at a predicted probability cut point z, such that observation "i" is considered an event if $\widehat{P[X]}_i > z$, and a nonevent otherwise. 1 - Specificity(z) is the ratio between falsely predicted event responses at cut point z and the actual number of non-events. Equivalently, Sensitivity(z) is the percentage of registrants correctly predicted at z, whereas 1 - Specificity(z) is the percentage of non-registrants wrongly predicted as registrants at z. Both indexes involve a counting process:

$$\text{Sensitivity} = \frac{\sum_{i \in \text{Registrants}} I(\widehat{P[X]_i}) \geq z)}{n_{\text{Registrants}}}$$

and

$$1 - \text{Specificity} = \frac{\sum_{i \in \text{non-Registrants}} I(\widehat{P[X]_i}) \geq z)}{n_{\text{non-Registrants}}}$$

where $I(\cdot)$ is an indicator function and $n(\cdot)$ is the sample size for $(\cdot)$.

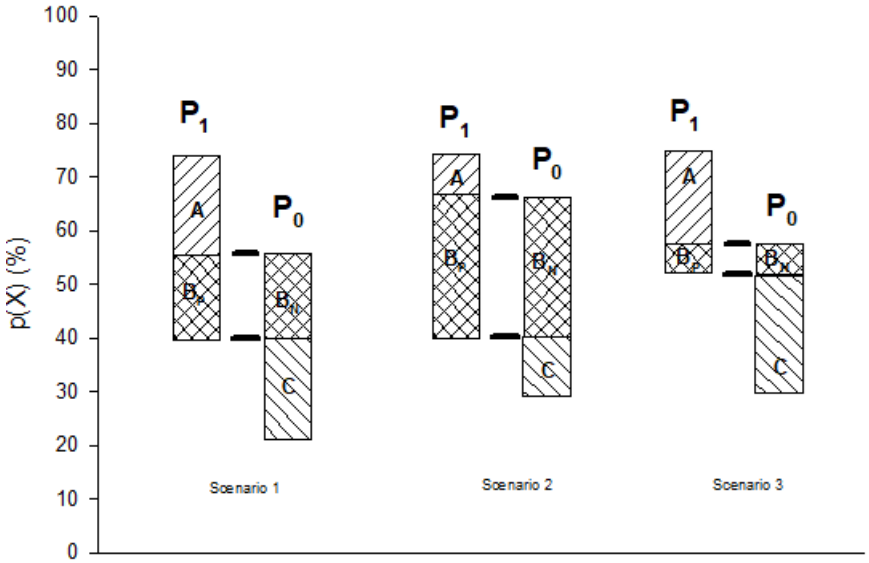
At any given z, Sensitivity and 1 - Specificity take on certain values (between 0 and 1). These values obtained over the whole range from 0 to 1 of the thresholds of predicted probability z, are coordinates for the points building ROC. For a small z (say 0.02), the likelihood to predict wrongly a registrant is minute, so sensitivity value would be 1 or very near 1. Also, the likelihood of assigning a non-registrant as registrant will be very high, so 1 - Specificity would be 1 or almost 1. With increasing z, both Sensitivity and 1 - Specificity will tend to decrease. The better a model predicts the slower is the decreasing rate upon increasing z. An ideal model, which predicts every single participant with $\widehat{P[X]} = 1$ and non-participant with $\widehat{P[X]} = 0$, would have a perfect square as ROC curve, whose area, and therefore c, equals 1. C is an index of predictive power of the model. Important in c is the way it is calculated:

$$c = \frac{n_c + 0.5 \cdot (t - n_c - n_d)}{t} = \frac{n_c + 0.5 \cdot n_{tied}}{t} = \frac{n_c}{t} + 0.5 \cdot \frac{n_{tied}}{t},$$

where n_{tied} is the number of tied pairs with different outcomes. The difference $1 - c$ equals the percentage of discordant pairs plus half of percentage of tied pairs. This detail on c composition is important: in (16) the significance of $\widehat{\gamma}_C$ depends on the concordant pairs percentage and that of $\widehat{\alpha}_C$ on the tied and discordant pairs percentages. Concordant pairs show that decision to participate systematically depends on X whereas the rest of pairs, discordant or tied, are needed to estimate the counterfactuals. If these pairs, summed up as $n_d + n_{tied}$ are absent or very few in number, then $\widehat{\alpha}_C$ will not be reliable. But also, if $n_d + n_{tied}$ grows much (of course, in expense of n_c), then selection is not being determined by observables.

A comparison of descending ranked of participants "P1" with non-participants "P0", as in Fig. 1, shows two regions A and C of sure concordance and a region B of mixed concordant and discordant observation pairs (denoted B_P for participants and B_N for non-participants). Region A is a subset A, $\{A : min(P_A) > max(P')\}$, where P_A denotes $\widehat{P[X]}$ in A and P' denotes $min(max(P_1, P_0))$. Similarly, subset C is defined as $\{C : max(PC) > min(P'')\}$, where $P'' = max(min(P1, P0))$. Depending on how well the logistic stage is predicting, the common support set $B = B_P \cup B_N$, becomes smaller or larger. Regions A and C often are trimmed out of data set and the analysis is continued on B. In parametric models, A and C are "counterfactual extrapolation regions" of $E[y_0|\widehat{P[X]}, d = 1]$ and $E[y_1|\widehat{P[X]}, d = 0]$, respectively. The number of observations per unit $\widehat{P[X]}$ differs on B and also between B_P and B_N for a given $\widehat{P[X]}$. The abundance of only one group associated with scarcity of the other, puts into question the reliability of $\widehat{\gamma}$ and $\widehat{\alpha}$ in (15).

FIGURE 1. $P[X]$ ranked grouped by participation status



There are three checking points in logistic regression stage, before continuing with stage two.

(i) Logistic overall fit through indexes ϕ_1 to ϕ_5 and c to detect either excellent fits or non-systematic decision mechanism. The first affects the reliability of estimates and the second makes bias correction redundant.

(ii) Relative size of common support B. If B is small compared to $A \cup C$ then estimation of $APE1$ is based on extrapolation rather than real data.

(iii) Sample sizes in $B|\widehat{P[X]}$.

Very small sample size of participants or of non-participants at a given $\widehat{P[X]}$ within the common support B affect reliability of $\widehat{APE1}|\widehat{P[X]}$. There is a relation

between (i), (ii) and (iii): an excellent prediction in (i) aggravates the problems in both (ii) and (iii), while "worsening" fit indices in (i) alleviates them.

5. Simulation results

Because of the counterfactual character of observational data, simulation becomes indispensable in evaluating the goodness of modeling approach to estimate APE1. The response variable of interest in our data, units sold, is practically an unbounded from above count random variable, assumed to follow the Poisson distribution and to be modeled as such [38]. Therefore we simulated a Poisson response variable with mean systematically determined by four variables: x_1, a continuous variable uniformly distributed with mean 2.5 and variance $\frac{25}{12}$; x_2, categorical variable randomly indexing half of observations; x_3, a categorical variable randomly indexing three equal parts of observations; and the participation binary variable of interest d. The mean structure for the response variable is:

$$\mu_{y0} = 0.2 \cdot x_1 + 0.05 \cdot I_1(x_2) + 0.3 \cdot I_2(x_3), \quad \text{if} \quad d = 0,$$

and

$$\mu_{y1} = APE1 \cdot \mu_{y0}, \quad \text{if } d = 1,$$

where $I_j(\cdot)$ are indicator functions of level j of categorical variables x_2 and x_3, which have two and three levels, respectively. APE1 equals 1.2 for promotion effect of 20% increase in expected participants response. APE1 equals 1 for null effect. Note that APE1 is multiplicative and not additive; as such, while it is assigned to have a constant effect of 20% increase in y_1, the increment in y_1 taken as the difference $(y_{i1}|d = 1) - (y_{i1}|d = 0)$ depends on the y_1 value. The multiplicative form of APE assigning is convenient in the generalized linear Poisson model used for its estimation. The Poisson regression is applied in all models used. The bias equals the difference between the estimated and simulated APE1. Different scenarios with respect to participation mechanism are simulated. This mechanism presumes a latent variable ν as given in (12) and (13). The systematic part of ν is $X\Gamma = 0.4 \cdot x_1 + 0.1 \cdot I_1(x_2) + (0.05 \cdot I_1(x_3) + 0.7 \cdot I_2(x_3))$. The random component υ in ν is a normally distributed variable with mean zero and variance σ^2 ranging from 0.02 to 2. b_0 characterizes the campaign incentive, which is independent of both X and ν. The more appealing a campaign, the larger the participation in it. Correspondingly, a larger b_0 in "$d = 1$ if $\upsilon > -b_0$" means higher participation. The participation mechanism has three components: (i) the observable individual features incorporated in X elements, (ii) the unobservable individual features represented by υ and (iii) campaign attractiveness, measured by b_0. While X and υ determine the randomness of decision process, b_0 determines the participation rate, given X and υ. Loosely, the same individual, who does not participate in a campaign, might participate if the offer were more attractive. Different participation decision scenarios are reflected in the logistic regression fit statistics, as given in Table 9 (see Appendix). The participation decision goes from very deterministic to very stochastic with the increase of σ^2 in υ from 0.02 to 2. Also, varying b_0 creates a wide range of participation rate. The total number of observations simulated is 50,000. Three models applied to estimate the participation effect were:

(i) a generalized linear Poisson model without correcting for selection bias in which the mean of y, μ_y, is linked to the linear predictor as $log(\mu_y) = x\widehat{B} + \widehat{\Theta} \cdot d$, where $\widehat{B}$ contains the intercept $\widehat{b_0}$;

(ii) a propensity scores model, in which a logistic regression estimates the propensity scores $\widehat{P[X]} = P[d = 1|X]$ and then in a second stage, μ_y is modeled as

$$log(\mu_y) = \widehat{b_0} + \widehat{b_1} \cdot \widehat{P[X]} + \widehat{b_2} \cdot \left(\widehat{P[X]} - \mu_{\widehat{P[X]}}\right) + \widehat{\alpha} \cdot d.$$

(iii) the model that combines (i) and (ii) by modeling μ_y as

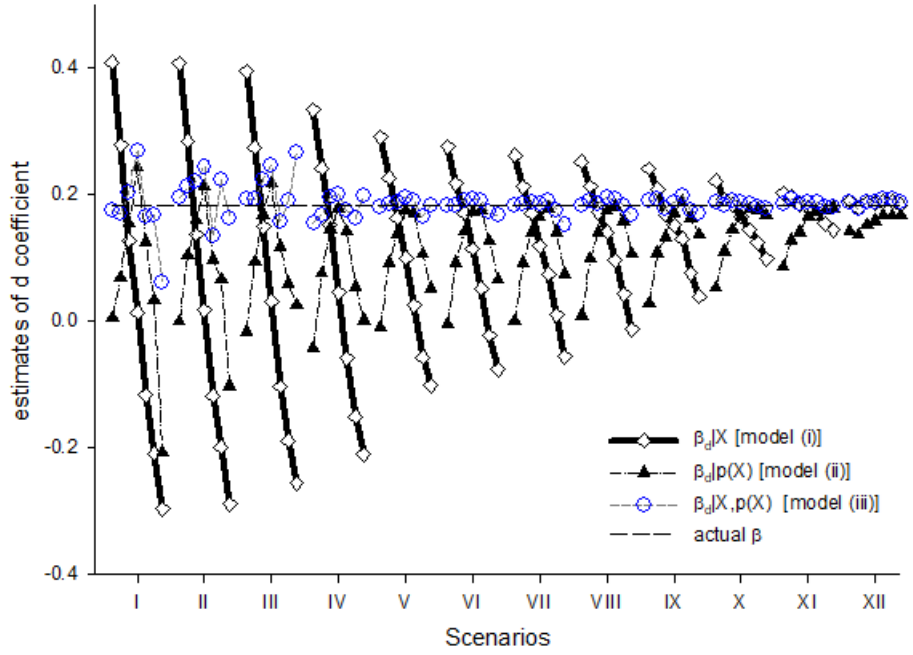
$$log(\mu_y) = X\widehat{B} + \widehat{\gamma_1} \cdot \widehat{P[X]} + \widehat{\alpha} \cdot d.$$

Henceforth these models will be named (i), (ii) and (iii) as ordered above. We believe that (i) and (ii) are the most likely choices the analysts would made if they were to analyze the data straightforwardly (without addressing the possible self-selection problem) or following the suggested model (11), respectively. The model outcome for the estimates of the coefficients of d is presented in Fig.2, where 12 different scenarios with respect to participation mechanism randomness are compared. The participation outcome gets more and more random (controlled by unobserved factors) from the left to the right of the graph. This is realized by increasing the variance of the random component υ in the latent participation variable ν. Note that υ is unconditionally independent of X or y, and its expectation remains 0. With increasing variance we realize a corresponding increased weight of unobservables in ν, i.e., participation. Within a given scenario different participation thresholds are tried, by applying different b_0, thus increasing participation rates within a given scenarios from the left to the right. In all, on graph 2 we move from a very systematic to a very stochastic participation mechanism, and from a low to a high participation rate, i.e., from an unattractive to a very attractive campaign incentive range of scenarios. Table 9 (see Appendix), summarizes the variance of υ and b_0 used in each scenario.

Model (iii) produces estimates of APE1 with much smaller bias than that of the other two. Evidently, model (ii) regressing on propensity scores mitigates bias compared to model (i) that does not correct for self-selection, but $\widehat{APE1}$ in (ii) is not robust to the ratio participants to non-participants that changes across a given scenario. With the participation mechanism getting more and more random, the bias in APE1 estimates becomes in general smaller and smaller, up to the rightmost scenario XII, in which apparently there is no need for any bias correction. The pattern of bias as a function of participation mechanism becomes clear by comparison between Fig. 2 and 3, the latter presenting the opposite selection mechanism: the condition for participation in Fig. 2 is the condition for non-participation in Fig. 3 and *vice versa*.

Model (iii) treatment effect estimate is more accurate (unbiased) and more robust to the extent of systemic weight in participation decision. The $\widehat{APE1}$ bias has different impacts in promotional evaluation practice. Let us assume that campaign effect is non-negative. A positive bias in an actual positive effect of campaign is a mild form of its impact. A negative bias, on the contrary, might reveal a good campaign as not worthy. If the actual effect of campaign is null, then a positive bias is more harmful than a negative one, because $\widehat{APE1} \leq 0$ is not only a suspicious result, but also a reason to stop the campaign or to consider it a failure, whereas

FIGURE 2. Estimated campaign participation effect by the three models. The participation mechanism is the opposite of the one shown in the previous figure 2.

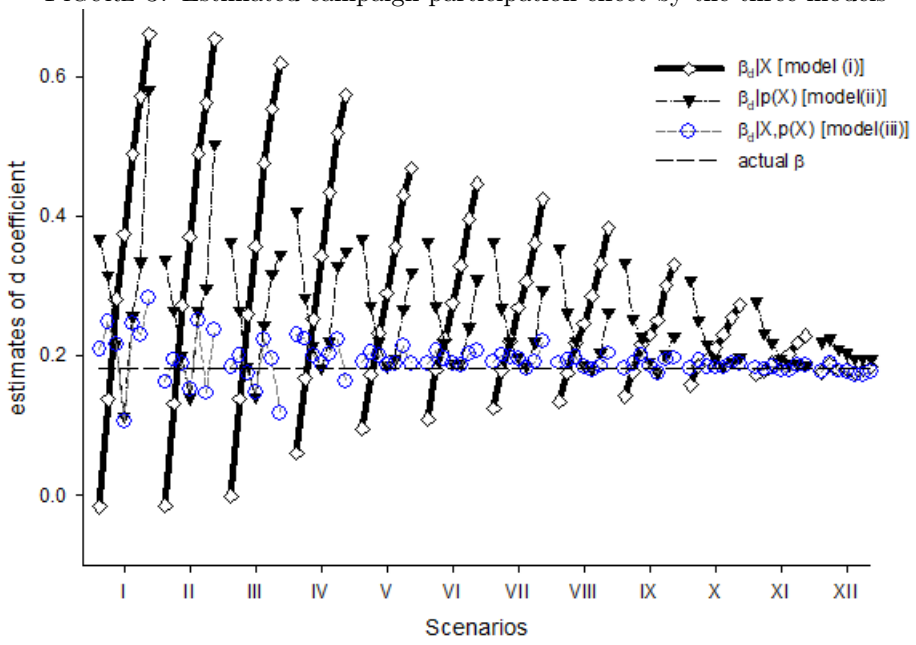


$\widehat{APE1} \geq 0$ when APE1 is in reality not larger than 0 leads to continuing a fruitless campaign or drawing wrong conclusions on profitability of a non-profitable action. To examine the performance of the three models for ineffective campaigns, we simulated the same participation scenarios as above with $APE = 0$. The estimates of APE1 obtained by the three models, given in the Table 10 (see Appendix), show that only model (iii) does not assign any significant effect to the campaign with real zero effect across all studied scenarios, as the p-values of the hypothesis test $H_0 : \widehat{APE1} = 0$ reveal. Including estimated propensity score in the generalized linear model estimating APE substantially improved the ability of model for a much more accurate $\widehat{APE1}$ compared to the same model without propensity score, or the same model without covariates.

6. ANALYSIS OF TWO CAMPAIGNS

6.1. **Campaign 1.** A campaign launched some time ago offered an incentive to participants in exchange of some purchase. In order to participate in the campaign, the individuals should be member of a loyalty club and they should register. Registration was free of charge. It only showed a preliminary interest of participants. A registrant was not obliged to purchase. Customers that joined the club attracted by promotion increased the loyalty club membership size. There were 201,107 registrants out of a total of 1,164,742 club members who did purchase during the promotional period. 203,931 individuals enrolled and became club members

FIGURE 3. Estimated campaign participation effect by the three models



during the registration phase and 9,534 of them participated in the campaign. The other 960,811 individuals were already club members and 191,573 of them registrants. Observable explanatory variables are membership class in the beginning of promotional period (categorized as "Platinum", "Gold" and "Club" members), collector of loyalty benefits type (two categories, "Air Miles" or "Point" collectors), time since enrollment in club (years), time since the last purchase (months), enrollment tenure ("new" enrollees considered those who enrolled in campaign registration time, otherwise "old"), demographic data as average income, population age, percentage of females, businesses and population per square mile in the ZIP Code area of member home address, total purchases per individual during the year before promotion, point balance after redemptions, total points and total miles earned per individuals and the brand(s) of purchased article. Data contains mostly categorical and demographic variables − both of which do not provide information at individual level. The only variable with important individual information is the total purchases per individual during the year before promotion.

Participation is analyzed by logistic regression (see Table 11 in Appendix).

It is noteworthy the negative effect of time variables like "time since enrollment" and "time since the last purchase" on participation rate, which is backed up by the dramatic effect of enrollment tenure, altogether showing that the vital core of participation is the new membership. The unconstrained log-likelihood of the model is −479,849, whereas the constrained log-likelihood is −413,577. This small difference is to be expected, as covariates are mostly not individual specific. These likelihood values produce an R^2 as low as 0.129 and a maximum rescaled R^2 of 0.204. The c statistic is 0.758. Out of 147,215,549,944 pairs with different participation status, 75.3% were concordant, 23.8% discordant and 0.8% tied. Clearly, an excellent

fit of participation model is not a problem here and participation is perceived at a considerable extent as non-systemic. Parameter estimates of model (i) and (iii) are not much different (Table 12, see Appendix) and the main change by incorporating propensity scores in model (iii) consists in the different partial effect of the purchased brands. $\widehat{APE1}$ is virtually the same in both models. Application of model (ii) of regressing on propensity scores produces a stronger effect of participation, as shown in Table 2.

TABLE 2. Parameter estimates of model (ii), example 1.

Parameter	Estimate	StdErr	P-value
Intercept	0.4994	0.0011	<.0001
Propensity score	2.9335	0.0035	<.0001
Participation × (Propensity score - Mean)	-0.4027	0.0053	<.0001
Participation	0.28	0.0015	<.0001

The estimated overall campaign effect on participants, based on model (iii) is $100 \cdot [exp(0.2272) - 1] = 25.5\%$. However, it not realistic to expect the same APE1 in all participant groups. Interacting the categorical variable(s) of interest with the participation variable gives the APE1 estimates across the levels of these categorical variable(s). It is of a primary interest to estimate APE1 across the different strata of participation as this directly links the response to campaign with the returned value on its investment. The response to campaign (registration) does not imply a real participation, but an intention to do so. Propensity score is a scaled measure of this intention. Other overlapping in time campaigns launched by competitors give to customers, who can be registrants in several campaigns at the same time, the luxury of choosing the "right offer" for them.

The histogram of propensity score estimates, presented in Fig. 4, suggests roughly four groups with respect to the registration rates. Interacting a categorical variable indicating these four groups with the participation variable evaluates APE1 in each group. In doing so one must be aware of collinearity problems that can arise, because the new categorical variable can bear similar information with other variables already in the model, like say, membership class.

The results of interaction presented in Table 3, indicate that the campaign effect lift in the group "j" (CELj), calculated as $100\% \cdot [exp(PE_j) - 1]$, where PE_j is the parameter estimate for group j, differs across groups. Note that the number of expected purchases gained per customer (HNG) does not necessarily follow the pattern of estimated effect. It is calculated as $HNG_j = HN_j \cdot \left(\frac{1}{1+CEL_j}\right)$; it is depends on HNj, the observed mean individual response of participants in the group j.

6.2. **Campaign 2.** Another similar promotional campaign launched also in the past covered, in contrast with the example above, only one brand to be purchased. Out of 1,278,278 active loyalty club members during the promotional period, 240,858 participated in campaign. The parameter estimates of logistic stage are presented in the Table 4.

FIGURE 4. Propensity Score distribution in Example 1.

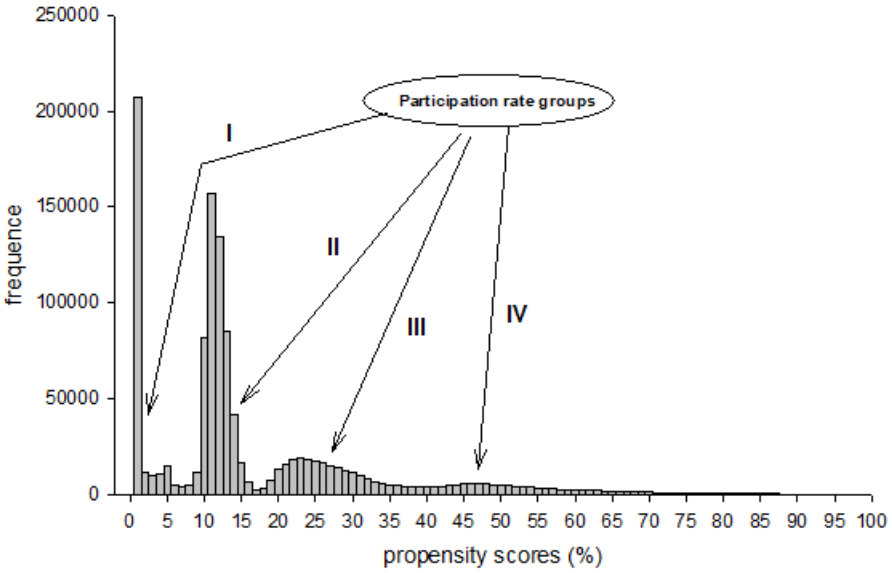


TABLE 3. APE1 estimates across individual response rate groups.

Group	Parameter estimate	StdErr	P-value	Estimated effect (%)
1	0.2339	0.0065	<.0001	26.40
2	0.2653	0.0026	<.0001	30.40
3	0.2968	0.002	<.0001	34.60
4	0.1259	0.0019	<.0001	13.40

The main difference in participation process in this example is that new enrollees have participated less than the older ones. There were 249,870,906,360 pairs with different outcomes, 73.4% of which concordant, 25.6% discordant and 1% tied. Statistic c is 0.74. The outcome for models (i) and (iii) in Table 5 shows a slight decrease of $\widehat{APE1}$ in (iii).

Including propensity scores in model (iii) is associated with two relevant changes in the partial effect of covariates: Total purchases before the promotional period changes from negative in model (i) into positive in model (iii) and the effect of new enrollees becomes much more important in model (iii). Both these changes are in concordance with previous experiences and also with the common sense. Model (ii), whose parameter estimates are presented in Table 7, did not produce reasonable estimates for the campaign effect. Its $\widehat{APE1}$ was negative. The estimated overall campaign effect on participants is $100 \cdot [exp(0.1184) - 1] = 12.6\%$.

TABLE 4. Parameter estimates of logistic regression for campaign participation, example 2. (The "baselines" of categorical variables, "CLUB" for membership class, "Old" for enrollment age and "Points" for collector type, whose the estimates are zero, are omitted)

Parameter	Class level	Estimate	StdErr	P-Value
Intercept		-1.0641	0.00793	<.0001
Membership	GOLD	0.3154	0.00544	<.0001
	OTHER	0.4402	0.0121	<.0001
	PLATINUM	0.8919	0.0101	<.0001
	SIGNATURE	-0.4337	0.00533	<.0001
Enrollment tenure	New	-0.2296	0.00312	<.0001
Total purchases pre-campaign		0.00875	0.00047	<.0001
Collector type	Miles	-0.00595	0.00325	0.0668
ZIP Income		0.0000013	0.00000011	<.0001
Population per square mile		-0.00000052	0.00000026	0.0429
Days sine last purchase		-0.00107	0.000019	<.0001
Day since enrollment		0.000039	0.00000121	<.0001

TABLE 5. Parameter estimates of models (i) and (iii), example 2.

Parameter	Class	Model i			Model ii		
		Estimate	StdErr	P-Value	Estimate	StdErr	P-Value
Intercept		0.5686	0.0035	<.0001	0.0923	0.0119	<.0001
Membership Class	GOLD	1.2619	0.003	<.0001	0.6634	0.0146	<.0001
	OTHER	1.1398	0.0069	<.0001	0.4555	0.0178	<.0001
	PLATINUM	2.0326	0.0039	<.0001	1.009	0.0247	<.0001
	SIGNATURE	0.5871	0.0025	<.0001	0.405	0.005	<.0001
Enrollee tenure	New	0.0984	0.0025	<.0001	0.3773	0.0071	<.0001
Total purchases prior to promotion		-0.039	0.0028	<.0001	0.0294	0.0032	<.0001
Total purchases prior to promotion by Membership Class	GOLD	0.0632	0.0028	<.0001	-0.0128	0.0033	<.0001
	OTHER	0.0805	0.0029	<.0001	0.0043	0.0033	0.198
	PLATINUM	0.0536	0.0028	<.0001	-0.0209	0.0033	<.0001
	SIGNATURE	0.0549	0.0028	<.0001	-0.012	0.0031	0.0001
Collector type	Miles	-0.0692	0.0022	<.0001	-0.0625	0.0022	<.0001
ZIP Income		0	0	<.0001	0	0	<.0001
Population per square mile		0	0	<.0001	0	0	<.0001
Days sine last purchase		0	0	<.0001	0.0003	0	<.0001
Day since enrollment		-0.0001	0	<.0001	-0.0001	0	<.0001
Propensity scores					2.9	0.0691	<.0001
Participation		0.1255	0.0018	<.0001	0.1184	0.0018	<.0001

Following the same procedure with propensity score histogram as in the first example, five groups of individuals can be visually detected, as presented in Fig. 5 (see Appendix).

The interaction of the five groups with the participation variable produces some counterintuitive negative estimates of participation in campaign, as given in Table 13 (see Appendix). It is very unlikely the campaign effect to be significantly negative.

The partial effect of propensity scores increases substantially towards lower group levels, with low participation rate. Something inherently different should characterize lower and higher groups of participation rates given in Fig. 5. Intuitively, by bringing together the results of logistic regression and second Poisson regression stage, one can notice that while the new enrollees do have lower registration rates, in the same time they have a larger response in terms of purchases. The enrollment tenure parameter estimate is -0.02296 in logistic and 0.3773 in the Poisson regression. Therefore it is worth examining the composition of propensity score histogram across enrollee loyalty program membership tenure, as shown in Fig. 6 (see Appendix).

Fig. 6 is the result of two logistic regressions on two separate data sets: one with the new enrollees and the other with the older ones. Evidently, the "newer" enrollees fill up the lower range of propensity cores (group 1 in Fig. 5 is comprised mostly of them). Additionally, the conspicuous difference in the group sizes ("newer" enrollees are a much bigger group) justifies creating of two data sets on enrollee tenure and of two independent models on them. "Newer" enrollees do not have heterogeneous groups on participation rates, whereas the more tenured enrollees are distributed in three groups, as shown in Fig. 6. Application of the two-stage model as applied in the examples above is straightforward. A logistic regression produces the propensity scores needed in stage two. In the newer enrollees model, the important variable of purchasess in the year prior to campaign is eliminated, as it is zero. This brings about the decrease in the logistic fitness indexes. For example the c statistic is 0.606 in the model with new enrollees and 0.674 in the model with more tenured enrollees. The results are summarized in Table 6.

New enrollees show a significantly higher sensitivity to the campaign appeal, and among the old enrollees, the group with low participation rate has the least gain from the campaign at individual level. In both examples shown above, propensity scores changed the partial effects of some of the other covariates. We are inclined to believe that that change is in the right direction. We do consider propensity score as an omitted variable in the model (i). Adding propensity score in the model extensively increased the insight to the very important relationship between participation rate and immediate gains from promotional campaign.

TABLE 6. APE1 estimates across different enrollment age groups.

Enrollment age	Participation rate group	Estimate	StdErr	P-Value	Estimated effect (%)
"Newer"		0.295	0.0047	<.0001	34.30
"Older"	1	0.0412	0.0048	<.0001	4.20
	2	0.2102	0.0033	<.0001	23.40
	3	0.2007	0.0049	<.0001	22.20

Table 7. Parameter estimates of model (ii), example 2.

Parameter	Estimate	StdErr	P-value
Intercept	0.3742	0.0019	<.0001
Propensity score	3.6946	0.0068	<.0001
Participation× (Propensity score - Mean)	0.8406	0.0123	<.0001
Participation	-0.0299	0.003	<.0001

7. Conclusions

The inclusion of estimated propensity score as an independent variable in models estimating promotional campaign effect helps in mitigating or eliminating estimation bias that stems from violation of randomness principle by self-selectivity in campaign participation. It results to an effective parametric method of estimation that outperforms the non-and-semi-parametric competitors. Also, the superiority to the parametric models regressing either on covariates or on propensity score alone is evident. The model gains considerably in interpretative power and related conclusions, as well. The reason for these improvement in model performance is that propensity score is generally an omitted variable in regression. While this is not a problem in experimental data, where propensity score is a constant by design and is absorbed in the intercept terms, in the observational data this does not hold true any more and, if not accounted for, brings about bias in parameter estimates.

APPENDIX - (SAS CODE, TABLES, FIGURES).

The participation decision process is thought of as the realization of a latent continuous variable, which "switches on" whenever it exceeds a given threshold. This is the basic idea in implementing generic code of a systemic participating decision, like the SAS code producing the propensity scores in Table 1, which follows:

```
data s;
do i=1 to 2000;
if i < 350 then a="A";
if i>=350 and i<1150 then a="B";
if i>=1150 then a="C";
if i<700 then b="A";
if i>=700 and i< 1425 then b="B";
if i>=1425 then b="C";
if i<570 then c = "A";
if i>=570 and i<1330 then c = "B";
if i>=1330 then c = "C";
if a="A" then IA_A=1;else IA_A=0;
if a="B" then IA_B=1; else IA_B=0;
if a="C" then IA_C=1; else IA_C=0;
if b="A" then IB_A=1; else IB_A=0;
if b="B" then IB_B=1; else IB_B=0;
if b="C" then IB_C=1; else IB_C=0;
if c="A" then IC_A=1; else IC_A=0;
if c="B" then IC_B=1; else IC_B=0;
if c="C" then IC_C=1; else IC_C=0;
V=0.2 - 0.4 · IA_A - 0.3 · IA_B - 0.7 · IA_C +
0.9 · IB_A + 0.1 · IB_B + 0.5 · IB_C -
1.1 · IC_A + 0.1 · IC_B - 0.5 · IC_C +
1.5 · rannor(-1);

if V < 1.3 then y = 0; else y = 1;
output; end;
run;

proc logistic data=s descending;
class a b c;
model y = a b c;
output out=out_data p=pred_prob;
run;

data out_data;
set out_data;
description =a||b||c;
run;

proc means data=out_data n;
var y;
```

```
class pred_prob description;
output out=final;
run;
```

The binary outcomes y are realizations of a latent variable V, determined by a threshold (in this case 1.3). Also note that for X made of only discrete covariates, the expected probabilities of $y = 1$ (propensity scores) equal simply the means of y across all classes formed by discrete covariates, which are consistent estimators for propensity scores. The SAS code (in place of proc logistic) is:

```
proc means data=out_data mean;
var y;
class description;
output out=final1;
run;
```

This is the non-parametric version of getting the propensity score. The parametric version though renders much more information on selection mechanism, and most importantly, it specifies the partial effect of each covariate.

Table 8: Semi-parametric estimates of APE1:
Wooldridge (2002), Hirano, Imbens and Ridder (2002) HIR, Ridgeway, McCaffrey, Morral and Lim (2002) RMML, Hirano and Imbens (2002) HI, Caliper matching − CM, Actual APE1 − APE1.

(Scenario)	b0	Wooldridge	HIR	RMML	HI	CM	APE1
0.02 (I)	1.8	0.80	0.79	0.52	0.79	0.66	0.25
	1.5	0.56	0.55	0.33	0.77	-0.11	0.23
	1.3	0.37	0.37	0.30	0.76	0.30	0.22
	1.1	0.31	0.31	0.58	0.75	0.63	0.20
	0.9	0.21	0.21	0.50	0.75	0.61	0.19
	0.7	0.12	0.12	0.45	0.73	0.80	0.18
	0.5	0.12	0.12	0.71	0.77	0.46	0.18
0.05 (II)	1.8	1.18	1.17	0.36	0.73	0.26	0.25
	1.5	1.01	1.01	0.27	0.70	0.17	0.23
	1.3	0.93	0.93	0.40	0.72	0.48	0.22
	1.1	0.87	0.86	0.48	0.70	0.30	0.20
	0.9	0.69	0.68	0.58	0.70	0.68	0.19
	0.7	0.51	0.51	0.57	0.70	0.43	0.18
	0.5	0.36	0.35	0.63	0.73	0.51	0.18
0.10 (III)	1.8	0.91	0.89	0.34	0.63	0.27	0.25
	1.5	1.05	1.03	0.46	0.62	0.56	0.23
	1.3	1.07	1.07	0.51	0.63	0.42	0.22
	1.1	1.07	1.07	0.52	0.61	0.50	0.20
	0.9	1.03	1.02	0.50	0.61	0.56	0.19
	0.7	0.99	0.99	0.56	0.63	0.51	0.18
	0.5	0.83	0.83	0.56	0.63	0.44	0.18
0.20 (IV)	1.8	0.68	0.63	0.40	0.54	0.39	0.24
	1.5	0.73	0.70	0.39	0.48	0.35	0.22
	1.3	0.81	0.80	0.39	0.49	0.42	0.21
	1.1	0.83	0.82	0.39	0.46	0.36	0.20
	0.9	0.90	0.89	0.50	0.51	0.48	0.19
	0.7	0.92	0.91	0.53	0.51	0.53	0.18
	0.5	0.92	0.92	0.55	0.54	0.49	0.18
0.35 (V)	1.8	0.23	0.16	0.23	0.35	0.23	0.24
	1.5	0.45	0.41	0.33	0.34	0.32	0.22
	1.3	0.51	0.48	0.33	0.32	0.26	0.21
	1.1	0.59	0.57	0.37	0.33	0.37	0.20
	0.9	0.64	0.63	0.36	0.32	0.28	0.19
	0.7	0.71	0.70	0.43	0.38	0.40	0.18
	0.5	0.77	0.77	0.49	0.44	0.38	0.18
0.40 (VI)	1.8	0.19	0.11	0.20	0.32	0.21	0.23
	1.5	0.39	0.34	0.31	0.30	0.29	0.22
	1.3	0.43	0.39	0.30	0.27	0.25	0.21
	1.1	0.51	0.49	0.32	0.27	0.27	0.20
	0.9	0.56	0.54	0.34	0.29	0.31	0.19
	0.7	0.62	0.61	0.37	0.33	0.33	0.18
	0.5	0.72	0.71	0.46	0.42	0.36	0.18

Continuation of Table 8

(Scenario)	b0	Wooldridge	HIR	RMML	HI	CM	APE1
0.45 (VII)	1.8	0.19	0.11	0.19	0.28	0.20	0.23
	1.5	0.25	0.19	0.24	0.25	0.24	0.22
	1.3	0.37	0.33	0.29	0.24	0.28	0.21
	1.1	0.46	0.43	0.33	0.27	0.26	0.20
	0.9	0.49	0.47	0.31	0.26	0.31	0.19
	0.7	0.56	0.55	0.34	0.30	0.29	0.18
	0.5	0.64	0.63	0.40	0.36	0.38	0.18
0.50 (VIII)	1.8	0.19	0.11	0.21	0.26	0.21	0.23
	1.5	0.19	0.13	0.20	0.21	0.22	0.21
	1.3	0.32	0.28	0.26	0.21	0.25	0.20
	1.1	0.39	0.36	0.29	0.23	0.25	0.20
	0.9	0.46	0.43	0.31	0.26	0.28	0.19
	0.7	0.50	0.48	0.31	0.27	0.27	0.18
	0.5	0.58	0.57	0.36	0.32	0.30	0.18
0.60 (IX)	1.8	0.16	0.08	0.19	0.22	0.20	0.22
	1.5	0.19	0.13	0.19	0.17	0.18	0.21
	1.3	0.23	0.18	0.22	0.19	0.23	0.20
	1.1	0.30	0.26	0.25	0.19	0.23	0.19
	0.9	0.33	0.31	0.25	0.21	0.21	0.19
	0.7	0.41	0.39	0.30	0.25	0.27	0.18
	0.5	0.45	0.44	0.31	0.27	0.27	0.18
0.75 (X)	1.8	0.17	0.08	0.19	0.18	0.20	0.21
	1.5	0.19	0.12	0.20	0.17	0.20	0.20
	1.3	0.19	0.14	0.18	0.16	0.18	0.20
	1.1	0.21	0.16	0.19	0.16	0.18	0.19
	0.9	0.25	0.21	0.21	0.18	0.21	0.19
	0.7	0.30	0.28	0.25	0.21	0.23	0.18
	0.5	0.30	0.28	0.23	0.21	0.22	0.18
1.00 (XI)	1.8	0.18	0.08	0.19	0.17	0.19	0.20
	1.5	0.17	0.10	0.18	0.15	0.19	0.19
	1.3	0.18	0.12	0.18	0.15	0.18	0.19
	1.1	0.19	0.14	0.19	0.16	0.19	0.19
	0.9	0.19	0.15	0.18	0.16	0.18	0.18
	0.7	0.21	0.17	0.19	0.17	0.18	0.18
	0.5	0.22	0.19	0.20	0.17	0.19	0.17
2.00 (XII)	1.8	0.17	0.08	0.17	0.15	0.18	0.18
	1.5	0.18	0.09	0.18	0.16	0.18	0.18
	1.3	0.17	0.09	0.17	0.15	0.17	0.18
	1.1	0.17	0.10	0.17	0.15	0.17	0.17
	0.9	0.16	0.10	0.16	0.15	0.17	0.17
	0.7	0.16	0.11	0.16	0.15	0.16	0.17
	0.5	0.16	0.12	0.16	0.15	0.16	0.17

End of Table8

Table 9: Different participation decision scenarios and their respective fit statistics.

σ^2 of υ Scenario	b0	% Participants	LogLc	LogLu	c	ϕ_2	ϕ_4	ϕ_5	ϕ_1	ϕ_2	ϕ_3
	1.8	75	-28101	-934	0.9994	0.6627	0.9817	0.9845	0.9668	0.6627	0.9782
	1.5	60	-33606	-924	0.9996	0.7295	0.9867	0.9879	0.9725	0.7295	0.992
	1.3	50	-34657	-896	0.9996	0.7409	0.9878	0.9889	0.9741	0.7409	0.9937
0.02 (I)	1.1	40	-33702	-952	0.9995	0.7302	0.9864	0.9876	0.9718	0.7302	0.9918
	0.9	30	-30558	-914	0.9995	0.6945	0.9845	0.9862	0.9701	0.6945	0.9863
	0.7	21	-25594	-626	0.9996	0.6317	0.9858	0.9882	0.9756	0.6317	0.9776
	0.5	14	-20449	-609	0.9996	0.5478	0.9805	0.9845	0.9702	0.5478	0.9436
	1.8	75	-28104	-2292	0.9966	0.6439	0.9538	0.9614	0.9184	0.6439	0.9403
	1.5	60	-33608	-2300	0.9973	0.7142	0.966	0.9695	0.9316	0.7142	0.9728
	1.3	50	-34657	-2253	0.9975	0.7264	0.9686	0.9713	0.935	0.7264	0.9774
0.05 (II)	1.1	40	-33697	-2403	0.9971	0.714	0.9646	0.9682	0.9287	0.714	0.9715
	0.9	30	-30568	-2298	0.9969	0.6772	0.9598	0.9651	0.9248	0.6772	0.9577
	0.7	21	-25662	-1646	0.998	0.6174	0.962	0.9689	0.9359	0.6174	0.9404
	0.5	14	-20481	-1467	0.9979	0.5326	0.9524	0.9619	0.9284	0.5326	0.8847
	1.8	75	-28116	-4488	0.9869	0.6114	0.9054	0.922	0.8404	0.6114	0.873
	1.5	60	-33634	-4558	0.9895	0.6875	0.9296	0.939	0.8645	0.6875	0.9321
	1.3	50	-34657	-4591	0.9897	0.6996	0.9328	0.9411	0.8675	0.6996	0.9393
0.1 (III)	1.1	40	-33676	-4583	0.9894	0.6877	0.9293	0.9384	0.8639	0.6877	0.9319
	0.9	30	-30586	-4487	0.9883	0.6479	0.9181	0.9304	0.8533	0.6479	0.9045
	0.7	21	-25805	-3458	0.9912	0.5909	0.9179	0.9342	0.866	0.5909	0.8744
	0.5	14	-20474	-2998	0.9911	0.5029	0.8995	0.9206	0.8536	0.5029	0.7927
	1.8	75	-28239	-8227	0.9554	0.5509	0.8139	0.851	0.7087	0.5509	0.7517
	1.5	60	-33642	-9105	0.9579	0.6252	0.8453	0.8739	0.7294	0.6252	0.8277
	1.3	50	-34657	-9079	0.9598	0.6405	0.854	0.8791	0.738	0.6405	0.8439
0.2 (IV)	1.1	40	-33692	-8983	0.9589	0.6278	0.8482	0.8751	0.7334	0.6278	0.8316
	0.9	31	-30769	-8448	0.9584	0.5905	0.8341	0.8646	0.7254	0.5905	0.7963
	0.7	22	-26257	-7265	0.9619	0.5322	0.8186	0.8581	0.7233	0.5322	0.7406
	0.5	15	-20686	-6134	0.9627	0.4413	0.784	0.832	0.7035	0.4413	0.6343
	1.8	74	-28909	-13444	0.8816	0.4613	0.6731	0.7478	0.5349	0.4613	0.5874
(V)	1.5	60	-33698	-14944	0.8849	0.5277	0.7129	0.7786	0.5565	0.5277	0.6658
	1.3	50	-34657	-15262	0.8849	0.5397	0.7196	0.7833	0.5596	0.5397	0.6792
0.35	1.1	41	-33786	-14970	0.8849	0.5289	0.7136	0.7791	0.5569	0.5289	0.6671
	0.9	31	-31096	-13942	0.8871	0.4965	0.6976	0.7675	0.5517	0.4965	0.6313
	0.7	23	-27051	-12411	0.8901	0.4432	0.6704	0.7487	0.5412	0.4432	0.5696
	0.5	16	-21824	-10495	0.8921	0.3644	0.6258	0.7088	0.5191	0.3644	0.4722
	1.8	73	-29183	-14943	0.8541	0.4343	0.6304	0.7176	0.488	0.4343	0.5422
	1.5	60	-33745	-16618	0.8568	0.496	0.6696	0.7484	0.5076	0.496	0.6156
	1.3	50	-34657	-17097	0.8546	0.5046	0.6728	0.7512	0.5067	0.5046	0.6245
0.4 (VI)	1.1	41	-33822	-16762	0.855	0.4946	0.667	0.7469	0.5044	0.4946	0.6131
	0.9	32	-31296	-15683	0.8571	0.4645	0.6505	0.7352	0.4989	0.4645	0.5789
	0.7	24	-27337	-13892	0.8627	0.416	0.6256	0.7156	0.4918	0.416	0.523
	0.5	17	-22413	-11850	0.8646	0.3446	0.5821	0.675	0.4713	0.3446	0.4352
	1.8	72	-29451	-16371	0.825	0.4074	0.5886	0.688	0.4441	0.4074	0.4993
	1.5	59	-33762	-18268	0.8258	0.4619	0.6235	0.7171	0.4589	0.4619	0.5637
	1.3	50	-34657	-18752	0.8239	0.4707	0.6276	0.7205	0.4589	0.4707	0.5732
0.45 (VII)	1.1	41	-33861	-18383	0.8245	0.4616	0.6222	0.7166	0.4571	0.4616	0.5628
	0.9	32	-31454	-17210	0.8275	0.4343	0.6068	0.705	0.4528	0.4343	0.5317
	0.7	24	-27638	-15380	0.832	0.3876	0.5794	0.682	0.4435	0.3876	0.4769
	0.5	17	-22947	-13165	0.835	0.3238	0.5391	0.6417	0.4263	0.3238	0.3995
	1.8	72	-29725	-17703	0.796	0.3818	0.5489	0.66	0.4044	0.3818	0.46
	1.5	59	-33809	-19776	0.7946	0.4295	0.5794	0.6872	0.415	0.4295	0.5158
	1.3	50	-34657	-20293	0.792	0.437	0.5827	0.6901	0.4145	0.437	0.5238
0.5 (VIII)	1.1	41	-33901	-19908	0.7928	0.4286	0.5774	0.6864	0.4128	0.4286	0.5141
	0.9	33	-31646	-18662	0.797	0.4051	0.5642	0.6754	0.4103	0.4051	0.4875
	0.7	25	-28013	-16760	0.8014	0.3625	0.5379	0.6517	0.4017	0.3625	0.4376
	0.5	18	-23498	-14364	0.8068	0.3061	0.5023	0.6132	0.3887	0.3061	0.3704
	1.8	71	-30251	-19963	0.7423	0.3374	0.4807	0.6119	0.3401	0.3374	0.3953
	1.5	59	-33928	-22328	0.7341	0.3712	0.4999	0.6322	0.3419	0.3712	0.4332
	1.3	50	-34657	-22850	0.7315	0.3764	0.5019	0.6352	0.3407	0.3764	0.4387
0.6 (IX)	1.1	42	-33975	-22428	0.7326	0.3699	0.4978	0.6312	0.3399	0.3699	0.4313
	0.9	34	-31968	-21283	0.7335	0.3478	0.482	0.6174	0.3342	0.3478	0.4056
	0.7	26	-28701	-19199	0.741	0.3162	0.4631	0.5978	0.3311	0.3162	0.3697
	0.5	19	-24620	-16672	0.7479	0.2723	0.4347	0.5644	0.3228	0.2723	0.3188
(X)	1.8	69	-31003	-23071	0.6566	0.2719	0.3826	0.5409	0.2558	0.2719	0.3068
	1.5	58	-34053	-25276	0.6502	0.2961	0.398	0.5588	0.2578	0.2961	0.3337
	1.3	50	-34657	-25738	0.648	0.3001	0.4001	0.5613	0.2573	0.3001	0.338
0.75											

Continuation of Table 9

σ^2 of υ Scenario	b0	% Participants	LogLc	LogLu	c	ϕ_2	ϕ_4	ϕ_5	ϕ_1	ϕ_2	ϕ_3
	1.1	43	-34098	-25437	0.6463	0.2928	0.3934	0.5558	0.254	0.2928	0.3295
	0.9	35	-32391	-24240	0.649	0.2782	0.3831	0.5447	0.2516	0.2782	0.3131
	0.7	28	-29734	-22378	0.6526	0.2549	0.3665	0.5249	0.2474	0.2549	0.2868
	0.5	22	-26281	-19813	0.6631	0.228	0.3504	0.5027	0.2461	0.228	0.2569
	1.8	66	-32046	-26699	0.5394	0.1926	0.2665	0.448	0.1669	0.1926	0.2086
	1.5	57	-34226	-28489	0.5346	0.2051	0.275	0.4602	0.1676	0.2051	0.2221
	1.3	50	-34657	-28881	0.5324	0.2063	0.2751	0.4613	0.1667	0.2063	0.2233
1 (XI)	1.1	44	-34229	-28548	0.5328	0.2033	0.2726	0.4583	0.166	0.2033	0.22
	0.9	37	-33007	-27566	0.5347	0.1956	0.2669	0.4504	0.1649	0.1956	0.2117
	0.7	31	-31066	-26027	0.5362	0.1825	0.2566	0.4363	0.1622	0.1825	0.1974
	0.5	26	-28524	-23878	0.5452	0.1696	0.2492	0.4229	0.1629	0.1696	0.1836
	1.8	59	-33797	-32151	0.2969	0.0637	0.086	0.2538	0.0487	0.0637	0.0653
	1.5	54	-34522	-32853	0.2943	0.0645	0.0862	0.2553	0.0483	0.0645	0.0661
	1.3	50	-34657	-32990	0.2938	0.0645	0.086	0.2552	0.0481	0.0645	0.0661
2 (XII)	1.1	46	-34512	-32838	0.2954	0.0647	0.0865	0.2557	0.0485	0.0647	0.0663
	0.9	43	-34088	-32428	0.2967	0.0642	0.0863	0.2547	0.0487	0.0642	0.0658
	0.7	39	-33407	-31787	0.2975	0.0627	0.0851	0.2517	0.0485	0.0627	0.0643
	0.5	35	-32459	-30825	0.3044	0.0633	0.087	0.2528	0.0503	0.0633	0.0649

End of Table9

Table 10: Estimates of APE1 and the p-values of the hypothesis test that APE1 effect is not significant in a simulation with the real APE1 value of 0.

Scenario	b0	Model i		Model ii		Model iii	
		$\hat{APE}1$	P-Value	$\hat{APE}1$	P-Value	$\hat{APE}1$	P-Value
	1.8	-0.2174	<.0001	0.1775	0.0072	0.018	0.7577
	1.5	-0.071	0.0001	0.0999	0.1083	0.0117	0.8563
	1.3	0.0801	<.0001	0.0449	0.5027	0.0246	0.7283
scenario I	1.1	0.1837	<.0001	-0.0562	0.4294	-0.0705	0.3321
	0.9	0.3082	<.0001	0.0839	0.2954	0.0611	0.4453
	0.7	0.3919	<.0001	0.1363	0.225	0.0098	0.9248
	0.5	0.4841	<.0001	0.4154	0.0039	0.1438	0.2238
	1.8	-0.2153	<.0001	0.1758	<.0001	-0.0164	0.6608
	1.5	-0.0778	<.0001	0.0713	0.0757	-0.0206	0.6158
	1.3	0.0702	0.0002	0.0262	0.5286	-0.0085	0.8472
scenario II	1.1	0.1794	<.0001	-0.0376	0.3823	-0.0468	0.3079
	0.9	0.3093	<.0001	0.0882	0.084	0.0645	0.2012
	0.7	0.3825	<.0001	0.1082	0.1253	-0.0485	0.4557
	0.5	0.4778	<.0001	0.3224	0.0003	0.0505	0.5067
	1.8	-0.1976	<.0001	0.1995	<.0001	0.0045	0.8642
	1.5	-0.0707	<.0001	0.0788	0.0064	-0.0046	0.8752
	1.3	0.06	0.0011	0.0164	0.5781	-0.0141	0.6514
scenario III	1.1	0.167	<.0001	-0.036	0.2524	-0.0463	0.1615
	0.9	0.297	<.0001	0.0712	0.0492	0.0428	0.2337
	0.7	0.3744	<.0001	0.1318	0.0061	0.0036	0.9366
	0.5	0.4412	<.0001	0.1629	0.0072	-0.0734	0.1664
	1.8	-0.1357	<.0001	0.2267	<.0001	0.0435	0.0255
	1.5	-0.0362	0.0313	0.1047	<.0001	0.0299	0.1484
	1.3	0.0512	0.0033	0.0341	0.1062	0.0021	0.9244
scenario IV	1.1	0.1536	<.0001	0.0055	0.8079	-0.0007	0.9773
	0.9	0.2532	<.0001	0.0452	0.0798	0.021	0.4153
	0.7	0.3391	<.0001	0.1434	<.0001	0.0329	0.2805
	0.5	0.3944	<.0001	0.1705	<.0001	-0.0189	0.6081
	1.8	-0.0976	<.0001	0.1904	<.0001	0.0067	0.6595
	1.5	-0.0272	0.069	0.0881	<.0001	0.0091	0.5701
	1.3	0.0369	0.0176	0.0426	0.0091	0.0082	0.6253
scenario V	1.1	0.0984	<.0001	0.007	0.685	-0.0021	0.9064

Continuation of Table 10

Scenario	b0	Model i		Model ii		Model iii	
		$\hat{APE}1$	P-Value	$\hat{APE}1$	P-Value	$\hat{APE}1$	P-Value
	0.9	0.1728	<.0001	0.0187	0.3381	0.0062	0.7539
	0.7	0.2505	<.0001	0.0872	0.0002	0.0309	0.1703
	0.5	0.289	<.0001	0.1384	<.0001	0.0056	0.8342
scenario VI	1.8	-0.0832	<.0001	0.1861	<.0001	0.0052	0.7212
	1.5	-0.018	0.214	0.0896	<.0001	0.0124	0.4121
	1.3	0.0296	0.048	0.0409	0.008	0.0053	0.7345
	1.1	0.0846	<.0001	0.0108	0.5066	0.0002	0.9885
	0.9	0.1466	<.0001	0.0132	0.4654	0.0049	0.7875
	0.7	0.2145	<.0001	0.0616	0.0046	0.0196	0.3513
	0.5	0.2662	<.0001	0.1278	<.0001	0.0238	0.3361
scenario VII	1.8	-0.0705	<.0001	0.1807	<.0001	0.0041	0.7685
	1.5	-0.0153	0.2707	0.089	<.0001	0.0101	0.4818
	1.3	0.0277	0.0553	0.0413	0.0052	0.0074	0.618
	1.1	0.0788	<.0001	0.0185	0.2319	0.0086	0.5885
	0.9	0.124	<.0001	0.009	0.5994	0.0027	0.877
	0.7	0.1796	<.0001	0.0413	0.0404	0.0102	0.6055
	0.5	0.2442	<.0001	0.1138	<.0001	0.0368	0.11
scenario VIII	1.8	-0.0594	<.0001	0.172	<.0001	0.0035	0.7913
	1.5	-0.0175	0.1947	0.0822	<.0001	0.0036	0.796
	1.3	0.0261	0.0618	0.0423	0.0029	0.0093	0.5166
	1.1	0.0565	0.0001	0.0078	0.5958	-0.0029	0.8476
	0.9	0.1038	<.0001	0.0064	0.6914	0.0011	0.9464
	0.7	0.1507	<.0001	0.0278	0.1398	0.0061	0.7422
	0.5	0.2021	<.0001	0.0817	0.0004	0.0207	0.3374
scenario IX	1.8	-0.0499	<.0001	0.1522	<.0001	-0.0038	0.7633
	1.5	-0.0143	0.2647	0.0744	<.0001	-0.0003	0.9825
	1.3	0.0263	0.0459	0.0485	0.0003	0.0141	0.289
	1.1	0.0427	0.002	0.0124	0.3689	0.0009	0.9464
	0.9	0.0673	<.0001	-0.0007	0.9627	-0.0049	0.7442
	0.7	0.1199	<.0001	0.0249	0.14	0.016	0.3409
	0.5	0.1481	<.0001	0.0476	0.0174	0.0141	0.4641
scenario X	1.8	-0.0308	0.0101	0.1274	<.0001	-0.002	0.8667
	1.5	-0.0039	0.7457	0.0701	<.0001	0.0046	0.703
	1.3	0.0063	0.6108	0.0379	0.0026	-0.0011	0.9284
	1.1	0.0265	0.0392	0.0176	0.1717	0.0013	0.9221
	0.9	0.0468	0.0006	0.0068	0.6181	0.0026	0.8479
	0.7	0.0735	<.0001	0.0137	0.3577	0.0112	0.4542
	0.5	0.0901	<.0001	0.0169	0.3159	0.0063	0.7066
scenario XI	1.8	-0.0141	0.2096	0.0954	<.0001	-0.0006	0.9595
	1.5	-0.0088	0.4338	0.0539	<.0001	-0.0047	0.6748
	1.3	0.0042	0.7112	0.041	0.0006	0.0003	0.9763
	1.1	0.01	0.3984	0.0188	0.1199	-0.0025	0.8318
	0.9	0.0186	0.1347	0.0095	0.447	-0.0024	0.8436
	0.7	0.0377	0.0043	0.0117	0.3733	0.0078	0.5521
	0.5	0.0471	0.001	0.0083	0.5577	0.0067	0.6378
scenario XII	1.8	-0.0058	0.5778	0.0399	0.0005	-0.0038	0.7166
	1.5	0.0077	0.4555	0.0457	<.0001	0.0085	0.4135
	1.3	-0.0019	0.8562	0.0307	0.0063	-0.0028	0.792
	1.1	-0.0008	0.9394	0.0247	0.0282	-0.0035	0.7368
	0.9	-0.0044	0.6848	0.016	0.1571	-0.0083	0.4363
	0.7	-0.0005	0.9624	0.0182	0.1135	-0.0059	0.5904
	0.5	0.0044	0.6952	0.0186	0.1124	-0.0021	0.8542

End of Table 10

Table 11: Parameter estimates of logistic regression for campaign participation, example 1.

Variable	Class	Estimate	StdErr	P-Value
Intercept		-0.6131	0.169	0.0003
Membership pre-promotion	Club	-0.7696	0.017	<.0001
	Gold	0.0197	0.014	0.1633
	Platinum	0	.	.
Collector	Miles	0.0609	0.009	<.0001
	Points	0	.	.
Enrollment tenure	New	0.8901	0.112	<.0001
	Old	0	.	.
Years since enrollment		-0.00049	0.001	0.3383
Months since last purchase		-0.0416	0.001	<.0001
ZIP income		0	0	<.0001
ZIP Median Age		-0.0045	0.001	<.0001
ZIP % female		0.3894	0.051	<.0001
Businesses per square mile		-0.00001	0	0.0013
Population per square mile		0	0	0.0775
Total purchases in the prioryear		0.0161	0	<.0001
Point balance (in 10,000)		0.0699	0.002	<.0001
Points Pre-promotion (in 10,000)		-0.0545	0.002	<.0001
Miles earned		-0.00145	0.004	0.7202
Brand purchased	1	-0.8468	0.169	<.0001
	2	-0.7752	0.168	<.0001
	3	-0.1233	0.171	0.4707
	4	0.5463	0.273	0.045
	5	0.5943	0.793	0.4538
	6	8.8199	76.021	0.9076
	7	0.0861	0.181	0.6332
	8	0.6679	0.293	0.0228
	9	-0.8245	0.168	<.0001
	10	-0.1334	0.169	0.4296
	11	8.9889	42.896	0.834
	12	-0.0542	0.168	0.7469
	13	0.3409	0.17	0.0443
	14	0.1469	0.53	0.7816
	15	-5.7235	48.133	0.9053
	16	0.5941	0.224	0.0081
	17	0.5258	0.279	0.0598
	18	-0.0193	1.067	0.9856
	19	0.5033	0.176	0.0043
	20	-0.1975	0.224	0.3781
	21	-0.4049	0.54	0.4538
	22	1.642	1.21	0.1749
	23	-0.0198	0.176	0.9102
	24	0.0563	0.221	0.7985
	25	-1.1322	0.186	<.0001
	26	-0.4927	0.343	0.1504
	27	-0.9363	0.835	0.2624
	28	-7.4144	29.066	0.7987
	29	-0.6495	0.172	0.0002
	30	0	.	.
End of Table11				

Table 12: Parameter estimates of models (i) and (iii), example 1.

Variable	Class	Model i			Model iii		
		Estimate	StdErr	P-value	Estimate	StdErr	P-value
Intercept		2.077	0.0268	<.0001	1.7216	0.0276	<.0001
Membership pre-promotion	Club	-0.59	0.0025	<.0001	-0.412	0.0041	<.0001
	Gold	-0.2114	0.0022	<.0001	-0.1634	0.0023	<.0001
	Platinum	0	0	.	0	0	.
Collector type	Miles	-0.0236	0.0017	<.0001	-0.0274	0.0017	<.0001
	Points	0	0	.	0	0	.
Enrollee tenure	New	0.755	0.0186	<.0001	0.6172	0.0188	<.0001
	Old	0	0	.	0	0	.
Time since enrollment (years)		-0.0145	0.0001	<.0001	-0.0147	0.0001	<.0001
Time since last purchase (months)		-0.0009	0.0001	<.0001	0.0009	0.0001	<.0001
ZIP income		0	0	<.0001	0	0	<.0001
ZIP Median Age		0.0009	0.0001	<.0001	0.0015	0.0001	<.0001
ZIP percent female		-0.0835	0.0098	<.0001	-0.1456	0.0098	<.0001
Businesses per square mile		0	0	0.0106	0	0	0.8973
ZIP Population per square mile		0	0	<.0001	0	0	<.0001
Total articles purchased in year before promotion		0.0075	0	<.0001	0.0053	0	<.0001
Membership point balance (in 10,000)		0.0565	0.0002	<.0001	0.0488	0.0002	<.0001
Membership points before promotion (in 10,000)		-0.0548	0.0002	<.0001	-0.0489	0.0002	<.0001
Miles earned		0.0174	0.0006	<.0001	0.0178	0.0006	<.0001
Brand purchased	1	-0.5715	0.0267	<.0001	-0.4508	0.0268	<.0001
	2	-0.6556	0.0266	<.0001	-0.543	0.0267	<.0001
	3	-0.1786	0.0273	<.0001	-0.1672	0.0273	<.0001
	4	-0.1941	0.051	0.0001	-0.3	0.051	<.0001
	5	0.0437	0.125	0.7266	-0.0759	0.125	0.5439
	6	-0.5258	0.448	0.2405	-1.0075	0.4481	0.0245
	7	0.0334	0.0288	0.2457	0.0052	0.0288	0.8574
	8	0.0743	0.0449	0.0983	-0.0777	0.045	0.0843
	9	-0.5574	0.0266	<.0001	-0.4389	0.0267	<.0001
	10	-0.1336	0.0268	<.0001	-0.1259	0.0268	<.0001
	11	0.0347	0.2198	0.8746	-0.3912	0.22	0.0754
	12	-0.0981	0.0266	0.0002	-0.1083	0.0266	<.0001
	13	0.0516	0.0268	0.0542	-0.0385	0.0268	0.1512
	14	0.0381	0.0767	0.6198	-0.0254	0.0767	0.7404
	15	1.1344	0.2686	<.0001	1.3102	0.2686	<.0001
	16	0.17	0.0315	<.0001	0.0459	0.0316	0.1468
	17	0.1152	0.0447	0.01	-0.0028	0.0448	0.9507
	18	0.0673	0.1111	0.5447	0.0616	0.1111	0.579
	19	0.0961	0.0274	0.0005	-0.008	0.0275	0.7697
	20	-0.206	0.0386	<.0001	-0.1821	0.0386	<.0001
	21	-0.1161	0.0854	0.1738	-0.0679	0.0854	0.4266
	22	0.3162	0.1584	0.0459	-0.0013	0.1585	0.9936
	23	0.0156	0.0278	0.5739	0.0043	0.0278	0.8772
	24	0.0618	0.034	0.0693	0.0207	0.034	0.5425
	25	-0.659	0.0305	<.0001	-0.4962	0.0306	<.0001

Continuation of Table 12							
Variable	Class	Model i			Model iii		
		Estimate	StdErr	P-value	Estimate	StdErr	P-value
	26	-0.007	0.0569	0.9023	0.0744	0.0569	0.1913
	27	0.502	0.099	<.0001	0.6378	0.099	<.0001
	28	-0.0955	0.2596	0.7128	0.3198	0.2597	0.2182
	29	-0.0586	0.0271	0.0309	0.042	0.0272	0.1225
	30	0	0	.	0	0	.
Propensity score					0.8561	0.0158	<.0001
Participation		0.2272	0.0013	<.0001	0.2206	0.0013	<.0001
End of Table 12							

TABLE 13. Parameter estimates of propensity scores and participation interacted with participation rate groups.

Variable	Group	Estimate	StdErr	P-Value
Propensity score·group	1	11.6011	0.2057	<.0001
	2	9.1195	0.1448	<.0001
	3	6.6581	0.1086	<.0001
	4	5.2507	0.0899	<.0001
	5	4.5788	0.0818	<.0001
participation·group	1	-0.1374	0.0073	<.0001
	2	-0.0433	0.0051	<.0001
	3	0.0802	0.0038	<.0001
	4	0.2008	0.0031	<.0001
	5	0.2204	0.0045	<.0001

FIGURE 5. Propensity Score distribution in Example 2.

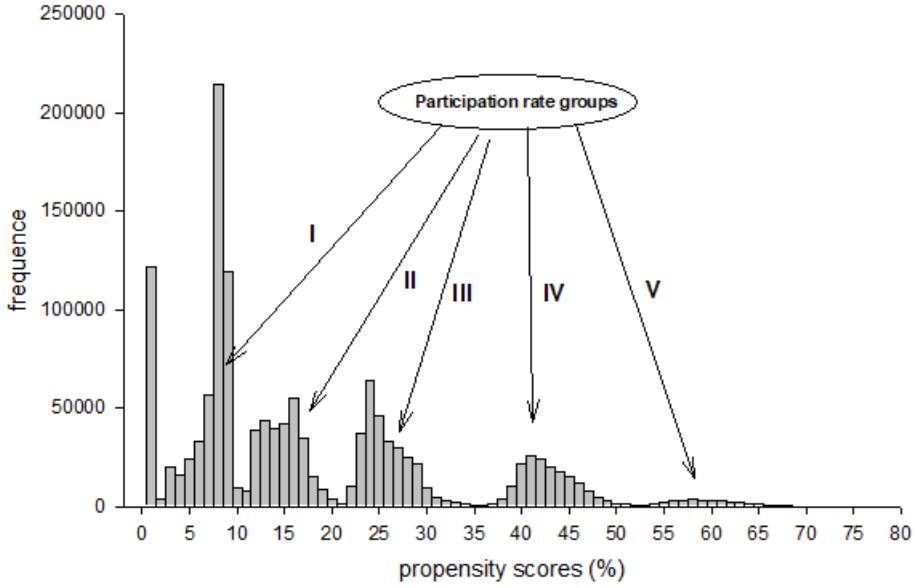
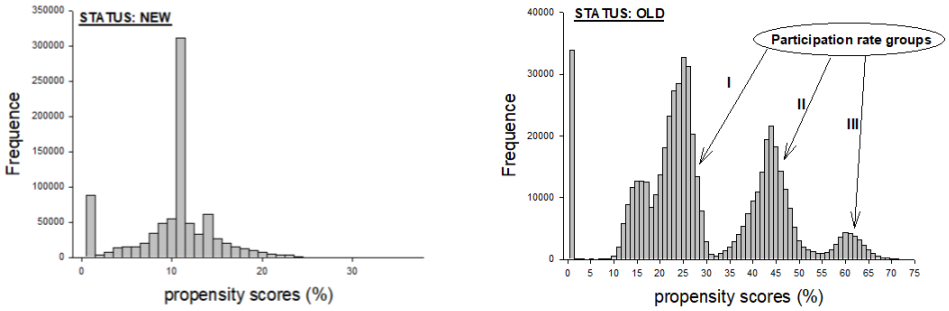


FIGURE 6. Propensity Score distribution in Example 2 across enrollment tenure.



References

[1] Bryson, A, Dorsett, R. and Purdon, S. (2002) The use of propensity score matching in the evaluation of active labour market policies, Working Paper No. 4, *Policy Studies Institute and National Centre for Social Research*, UK. URL: http://www.dwp.gov.uk/asd/asd5/WP4.pdf

[2] Keeble Claire, Law Graham Richard, Barber Stuart, Baxter Paul D., Participation Bias Assessment in Three High-Impact Journals, *SAGE Open* October-December, 2013: 1âĂŞ5.

[3] Heckman, J., Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes, *Drawing inferences from self-selected samples* 2000, (Ed. Howard Wainer), Lawrence Erlbaum Associates.

[4] Keeble Claire, Law Graham Richard, Barber Stuart, Baxter Paul D., Choosing a Method to Reduce Selection Bias: A Tool for Researchers, *Open Journal of Epidemiology*, 2015: 5, 155-162.

[5] Conniffe, D., Gash, V. and OConnell, P. Evaluating state programmes Natural experiments and propensity scores, 2000 *The Economic and social review*, 31, 283-308.

[6] Smith, J. and Todd, P. Reconciling conflicting evidence on the performance of propensity-score matching methods, *American Economic Review*, 2001, 91, 112-118.

[7] Heckman, J., Varieties of selection bias in Selectivity bias: new developments, *The American Economic Review*, 1990, 80, 313-318.

[8] Danielson, S. The propensity scores and estimation in non-random surveys - an overview. 2002. http://www.statistics.su.se/modernsurveys/publ/11.pdf

[9] DiPrete, T. and Engelhart, H. Estimating Causal Effects with Matching Methods in the Presence and Absence of Bias Cancellation, 2002. http://www.wjh.harvard.edu/~winship/cfa_papers/RESJune12.pdf

[10] Imai, K. and van Dyk, V. A. Causal Inference with general treatment regimes: Generalizing the propensity score, 2003. http://www.people.fas.harvard.edu/~kimai/files/pscore.pdf

[11] Plesca, M. and Smith, J. (2001) Evaluating multi-treatment programs: theory and evidence from the U.S. Job Training Partnership Act experiment *J. Empirical Economics (2007) 32: 491.* https://doi.org/10.1007/s00181-006-0095-0

[12] Firpo, S. Efficient semiparametric estimation of quantile treatment effects (draft), 2003 http://ist-socrates.berkeley.edu/~firpo/qte_firpo_jan_2003.pdf

[13] Winship, C. and Morgan, S. The estimation of causal effects from observational data, Annual Revue of Sociology, 1999, 25, 659-707.

[14] Sobel, M. Casual inference in the social and behavioral science, *Handbook of statistical modeling for the social and behavioral science*, (ed. Arminger, G., Clogg, C. and Sobel, M), Plnum Press, New York, 1995.

[15] Shonkoff, J. and Phillips, D. A. (editors), Defining and estimating casual effects, *From Neurons to Neighborhoods: The Science of Early Childhood Development*, 2000, 545-548, National Academy Press, Washington, D.C.

[16] Heckman, J., Ichimura, H., Smith, J. and Todd, P. Characterizing selection bias using experimental data, 1998, *Econometrica*, 66, 1017 - 1098

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

[18] Rubin, D. B. and Thomas, N., Combining propensity score matching with additional adjustments for prognostic covariates, 2000 *Journal of the American Statistical Association*, 95, 573-585.

[19] Riphahn, R. Matching methoden zur programmevaluation (lecture), 2002, http://www.unibas.ch/wwz/stat/lehre/riphahn/kap_3.pdf

[20] Rubin, D. Estimation from nonrandomized treatment comparisons using subclassification on propensity scores, *Proceedings of the International Conference on Nonrandomized Comparative Clinical Studies in Heidelberg*, April 10 -11, 1997.

[21] Froelich, M. A generalization of the balancing property of the propensity score, Discussion paper, 2002 http://www.fgn.unisg.ch/org/fgn/web.nsf/SysWebRessources/VWA_2002_08/$FILE/dp08froelich_ganz.pdf

[22] Zanutto, E. A comparison of propensity score and linear regression analyses of gender gaps in computer systems analysts careers, 2002, *Proceedings of the American Statistical Association, Social Statistics Section, Alexandria, VA: American Statistical Association.*

[23] Vuri, D. Propensity score estimates of the effect of fertility on marital dissolution, *The 2001 British Household Panel Survey Research Conference*, 5-7 July 2001, Colchester UK. www.iser.essex.ac.uk/activities/conferences/bhps2001/docs/pdf/papers/vuri.pdf
[24] Dawid, A. P. Conditional independence in statistical theory, *Journal of the Royal Statistical Society*, Series B 41, 1-31, 1979.
[25] Dehejia, R. and Wahba, S. Propensity score matching methods for non-experimental causal studies, *Review of Economics and Statistics*, 2002, 84, 151-161.
[26] Wooldridge, J. M., Econometric analysis of cross section and panel data, 2002, The MIT Press.
[27] Hirano, K., Imbens, G. and Ridder, G. Efficient estimation of average treatment effects using the estimated propensity score, Draft November 2002, http://web.mit.edu/afs/athena/org/e/econometrica/EconAcc/2094.pdf
[28] Ridgeway, G., McCaffrey, D., Morral, A. and Lim, N. An importance sampling framework for the analysis of observational data, 2002. http://www.i-pensieri.com/gregr/papers/propscorestalk.pdf
[29] Froelich, M. What is the value of knowing the propensity score for estimating average treatment effects? *Unpublished Work, UniversitÃđt St. Gallen*, 2002.
[30] Hirano, K. and Imbens, G., (2002) Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization, *Health Services and Outcomes Research Methodology 2:259âĂŞ278, 2001.* http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.452.5378&rep=rep1&type=pdf
[31] Ridgeway, G. 1999 Generalization of boosting algorithms and application of Bayesian inference for massive data sets, (dissertation of Ph.D. thesis), University of Washington. http://www.i-pensieri.com/gregr/papers/thesis.pdf
[32] Minkin, A. A non-parametric propensity score matching method for treatment effect estimation in observational studies, 2002. http://www.i-pensieri.com/gregr/papers/thesis.pdf
[33] Heckman, J., Ichimura, H. and Todd, P. Matching As An Econometric Evaluation Estimator, *Review of Economic Studies*, 65, 261 - 294, 1998.
[34] Jargowski, P. Omitted Variable Bias, *Kimberly Kempf-Leonard, ed., The Encyclopedia of Social Measurement, Vol. 2. San Diego, California: Academic Press.*, 919-924, 2005
[35] Pagan, A. Econometric Issues in the Analysis of Regressions with Generated Regressors *International Economic Review Vol. 25, No. 1*, 221-247, 1984
[36] Estrella, A. A new measure of fit for equations with dichotomous dependent variables, *Journal of Business and Economic Statistics, 16, 198-205*, 1998.
[37] Nagelkerke, N.J.D. A Note on a General Definition of the Coefficient of Determination, *Biometrika, 78, 691 âĂŞ692*, 1991.
[38] Cameron, A. C., and Trivedi, P. K., *Regression analysis of count data, Cambridge University Press*, 1998.