Data Mining and Knowledge Discovery (2018) 32:1275–1305
https://doi.org/10.1007/s10618-018-0576-8

Linear regression for uplift modeling

Krzysztof Rudaś[1,2] · **Szymon Jaroszewicz**[2]

Received: 10 December 2017 / Accepted: 20 June 2018 / Published online: 28 June 2018

Abstract
The purpose of statistical modeling is to select targets for some action, such as a medical treatment or a marketing campaign. Unfortunately, classical machine learning algorithms are not well suited to this task since they predict the results *after* the action, and not its causal impact. The answer to this problem is uplift modeling, which, in addition to the usual training set containing objects on which the action was taken, uses an additional control group of objects not subjected to it. The predicted true effect of the action on a given individual is modeled as the difference between responses in both groups. This paper analyzes two uplift modeling approaches to linear regression, one based on the use of two separate models and the other based on target variable transformation. Adapting the second estimator to the problem of regression is one of the contributions of the paper. We identify the situations when each model performs best and, contrary to several claims in the literature, show that the double model approach has favorable theoretical properties and often performs well in practice. Finally, based on our analysis we propose a third model which combines the benefits of both approaches and seems to be the model of choice for uplift linear regression. Experimental analysis confirms our theoretical results on both simulated and real data, clearly demonstrating good performance of the double model and the advantages of the proposed approach.

Keywords Uplift modeling · Linear regression · Causal discovery

Responsible editors: Jesse Davis, Elisa Fromont, Derek Greene, Björn Bringmann.

✉ Krzysztof Rudaś
k.rudas@mini.pw.edu.pl

Szymon Jaroszewicz
s.jaroszewicz@ipipan.waw.pl

1 Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland

2 Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

1 Introduction

Machine learning models are frequently used to select targets for an action such as a medical treatment or a marketing campaign. Typically, a sample is drawn from the population, individuals in the sample are subjected to the action and a model is trained to predict the outcomes. The model is then used to select cases from the general population for which the action is most profitable, e.g. the estimated recovery probability is above some threshold or predicted monetary gain from purchases made after being targeted by the campaign is high.

Unfortunately this approach is usually incorrect since it only takes into account what happens *after* the action, ignoring the outcome we would observe had a given person not been targeted (Radcliffe and Surry 2011; Rzepakowski and Jaroszewicz 2012; Lai 2006). To better understand the difference, consider sending e-mails with discount codes to selected customers. Suppose an average sale is 50 Euros and the model predicts that, after receiving a 25% discount code, customer C_1 will spend 150 Euros, so targeting her seems a good idea. This may however not be the case. If she would have bought the product even without the discount, at the full price of 200 Euros, the campaign would have actually resulted in a loss of 50 Euros. It is, however, profitable to target a customer C_2 who spent only 10 Euros after receiving the discount, but without the discount would not have bought anything at all.

Classical machine learning is unable to distinguish those two cases. An answer to this problem is *uplift modeling* which uses two training sets: a treatment group with objects subjected to an action and a control group with objects left untreated. The goal is to model the *difference* between responses in both groups conditional on the predictors, such that the benefit of taking the action is assessed against the background of not taking it (Radcliffe and Surry 2011; Lai 2006; Holland 1986).

In this paper we address the problem of uplift modeling with a numerical outcome variable, i.e. *uplift regression*. We are going to focus on linear models since they are very useful in practice and allow us to provide detailed analysis comparing the current approaches and new one proposed in this paper.

More formally, let x be a feature vector describing a customer, y^T the numerical outcome (e.g. purchase value) we would observe after targeting the customer, and y^C the numerical outcome we would observe had the customer not been targeted. Our purpose is to build a linear model of the form $x'\beta^U$ which predicts the quantity $y^T - y^C$, called the *uplift*. The goal of this paper is to find as good an estimator of β^U as possible.

What makes uplift modeling challenging is that, for each case, only one of the outcomes y^T or y^C is known, never both: once a person received an offer they cannot be made to forget it. This is known as the Fundamental Problem of Causal Inference (Holland 1986).

The contributions of the paper are as follows. First, we formally analyze two approaches to uplift regression: the double model approach and the outcome variable transformation approach which we here adapt to the problem of regression. To the best of our knowledge this is the first such detailed comparison of uplift models and one of very few theoretical results on uplift modeling. Our analysis contradicts the common belief (Radcliffe and Surry 2011; Kuusisto et al. 2014; Guelman et al.

2012) that the double model approach is usually inferior to dedicated uplift models. Based on the results of our analysis we propose a modified estimator which combines the benefits of both approaches and which we believe to be the model of choice for uplift linear regression.

1.1 Literature overview

We will first position uplift modeling within the broader field of causal discovery and later provide and overview of literature specific to this discipline.

Causal discovery aims not at predicting future outcomes, but, instead, to model the effects of interventions which directly change the values of some variables (Pearl 2009). Causal discovery can roughly be divided into two subdomains. One deals with the case in which the action was applied to some individuals (interventional data is available), the other performs causal discovery when only observational data is available (Pearl 2009).

Causal discovery from purely observational data is possible under additional assumptions. For example, it is possible to uncover parts of the causal structure provided that reliable conditional independence tests are available (Spirtes et al. 2000), or to determine the direction of causal influence by assuming a specific form of causal relationship (see e.g. additive noise models Hoyer et al. 2009). A good general overview of such methods can be found in Pearl (2009), Spirtes and Zhang (2016). Those methods are not related to uplift modeling which aims to characterize the influence of a specific action, not to discover unknown causal structure.

A class of techniques has been developed in economics, social sciences and medical statistics for causal discovery from interventional data, where a specific action has been performed on some of the objects. The aims of those methods are similar to uplift modeling, but the research problems addressed are different. The main focus is on making causal inferences when treatment assignment was not random. For example a treatment was applied based on doctor's decision. Such treatment assignment poses challenges for causal discovery; for example, the doctor might have applied the treatment only to healthier patients which have greater survival probability in the first place. Ignoring this fact will make the treatment look more beneficial than it is in reality (Robins and Hernán 2018). The most popular approach is based on so called propensity scores, that is estimates of the probability that a given individual receives the treatment. Treatment and control cases with similar scores are then matched together or assigned weights based on inverse treatment probabilities. A good overview of those methods can be found in Imbens and Rubin (2015) and Robins and Hernán (2018), some newer results from the machine learning community can be found in Johansson et al. (2016) and Shalit et al. (2017).

It should be noted that matching and weighting approaches are based on strong assumptions (such as all confounding variables being observed) which can not be tested directly from data (Robins and Hernán 2018). Such assumptions are not necessary when treatment assignment is random; randomized controlled trials are thus a gold standard in causal discovery (Imbens and Rubin 2015).

Uplift modeling is different from those approaches since it strives to get the best possible estimator from properly randomized data. Propensity score based methods use almost exclusively the double model approach, which we will analyze in this paper.

Another relevant method is g-estimation (Robins 1994; Robins and Hernán 2018) which also directly estimates uplift coefficients on full data. However the method itself is not statistically efficient (Robins 1994) and requires a correction, which amounts to subtracting predicted control outcomes from true treatment outcomes, making it in fact a double model estimator (Robins 1994). The true benefit of g-estimation is that it can be applied to complex multistage treatments which are beyond the scope of the current paper.

Uplift modeling literature is concentrated mainly on the problem of classification. Several approaches are based on decision trees (Rzepakowski and Jaroszewicz 2010, 2012; Radcliffe and Surry 2011) trained using modified splitting criteria which aim to maximize differences in responses between groups, or some related information theoretical measure. Several works investigate combining such trees into ensembles (Guelman et al. 2012; Sołtys et al. 2014). Work on linear uplift models includes approaches based on class variable transformation (Lai 2006; Jaśkowski and Jaroszewicz 2012; Kane et al. 2014; Pechyony et al. 2013) used with logistic regression and approaches based on Support Vector Machines (Kuusisto et al. 2014; Zaniewicz and Jaroszewicz 2013, Oct 2017). Those works only address the problem of classification and do not provide theoretical analyses which would clearly demonstrate the merits of each approach. See Radcliffe and Surry (2011) for a general overview of uplift modeling and its applications.

1.2 Notation and assumptions

We assume following notation: vectors are denoted with lowercase letters y, β, etc., and matrices with uppercase letters X, Σ, etc. Subscripts denote vector components, e.g. y_i denotes the ith component of y with the exception of x_i which will denote ith row of the data matrix X. An identity matrix is denoted with I and 0 will denote both zero vectors and scalars; transpose is denoted with $'$. $\mathcal{N}(\mu, \Sigma)$ denotes the multivariate normal distribution with mean vector μ and covariance matrix Σ.

Quantities related to the treatment group will be denoted with superscript T and those related to the control group with superscript C. For example n^T is the number of cases in the treatment group.

Vectors and matrices may be random variables. Expectations are denoted in the usual fashion, for example the expectation of y, the expectation of y with respect to a random variable X, and the expectation of y conditional on g will be denoted, respectively, as $\mathrm{E}\, y$, $\mathrm{E}_X\, y$ and $\mathrm{E}\, y|g$. Similar notation will be used for variances, where, by variance of a vector we understand its covariance matrix. Whenever we use limits involving random variables, convergence in probability will be assumed. Convergence in distribution will be denoted $\xrightarrow{d}$.

Performance of linear models is typically measured by how well their estimate $\hat{\beta^U}$ approximates the true parameter vector β^U in terms of mean squared error $\mathrm{E}\,\|\hat{\beta^U} - \beta^U\|^2$ (Heumann et al. 2013). For unbiased estimators this error is determined solely

by the estimator's covariance matrix (Heumann et al. 2013), which should be as small as possible. All models we will encounter will be unbiased, so we will only compare their variances. A covariance matrix Σ_1 is considered greater than a covariance matrix Σ_2 if $\Sigma_1 - \Sigma_2$ is positive definite.

Throughout, we will assume that there are p predictor variables and n data records arranged in a matrix $X \in \mathbb{R}^{n \times p}$. Let x_i denote the ith row of X. We assume that rows of X are generated independently from each other and follow the same distribution. Let g be a random vector of length n with $g_i \in \{T, C\}$. If $g_i = T$ ($g_i = C$) then the ith observation has been assigned to the treatment (control) group, respectively.

We also define the matrices $X^T \in \mathbb{R}^{n^T \times p}$ and $X^C \in \mathbb{R}^{n^C \times p}$ whose rows are rows from X assigned to treatment and control groups; n^T and n^C denote respectively the number of cases in the treatment and control training sets, $n = n^T + n^C$. Denote by $q^T = \frac{n^T}{n}$ and $q^C = \frac{n^C}{n}$ the proportions of cases in both groups.

Likewise, we define treatment and control response vectors $y^T \in \mathbb{R}^{n^T}$ and $y^C \in \mathbb{R}^{n^C}$ and denote by y the combined vector of all responses. We will assume that the expected outcome in the control group and the expected strength of the influence of the action (i.e. *uplift*) are linear functions of the predictors. As a result the response in the target group is also linear and the assumed relationships in the data can be written as (Heumann et al. 2013)

$$y^C = X^C \beta^C + \varepsilon^C, \tag{1}$$

$$y^T = X^T \beta^C + X^T \beta^U + \varepsilon^T = X^T \beta^T + \varepsilon^T, \tag{2}$$

where β^C are the true response coefficients in the untreated population, β^T is the respective treatment coefficient vector, and β^U is the vector of coefficients defining the strength and direction of the effect of the action on a given individual. Note that $\beta^T = \beta^C + \beta^U$ so the response in the treatment group equals the baseline plus the change in response caused by the action.

Random vectors ε^T and ε^C denote random components of the responses in the treatment and control groups. It is assumed that X, ε^T and ε^C are independent of each other. Moreover elements of ε^T are assumed to be independent and identically distributed, with $\mathrm{E}\,\varepsilon_i^T = 0$ and $\mathrm{Var}\,\varepsilon_i^T = (\sigma^T)^2$ for some $\sigma^T > 0$. Analogous assumptions are made for ε^C with $\mathrm{Var}\,\varepsilon_i^C = (\sigma^C)^2$. Note that the elements of both vectors may follow different distributions. Let ε denote the combined vector of random components in both groups.

For simplicity, in all theorems and proofs we will assume that the matrices X, X^T and X^C are of full rank, and thus the nonsingularity of $X'X$, $X^{T'}X^T$ and $X^{C'}X^C$ (Heumann et al. 2013). Since singular matrices form a subset of measure zero, we may expect those facts to hold with probability one if $n^T, n^C \geq p$ and the distribution of x_i's is continuous. For binary matrices this is not the case and the estimators may fail to exist with nonzero probability. This case will only be analyzed in an example.

Another simplification which we are going to make is assuming that the predictors have zero mean $\mathrm{E}\, x_i = 0$. This assumption can easily be dropped from all asymptotic results but we retain it for notational simplicity.

1.2.1 Randomization

For the uplift model to have a causal interpretation it is necessary for the distributions of the predictors in both groups to be identical. In practice this means that the assignment of cases to the treatment and control groups should be random and independent of the predictors (Robins and Hernán 2018).[1] There are, however, several ways to achieve such randomization. In this work we will assume *complete randomization* (Imbens and Rubin 2015). This means that the proportions q^C, q^T of cases in both groups are fixed in advance but the actual assignment is random. The actual randomization may be performed by sorting the cases in random order and then using the first nq^C cases as controls and assigning the remaining ones to the treatment group. As a result, pairs (x_i, y_i) are:

1. identically distributed within both treatment and control groups,
2. no longer independent, but independent conditional on g.

To see this note that g_i's become dependent after fixing n^T so pairs (x_i, y_i) become dependent through g_i's and n^T. Conditioning on g isolates n^T from (x_i, y_i)'s which again become independent.

2 Theoretical analysis of uplift regression models

In this section we provide theoretical analysis of two frequently used types of uplift models: the double model and the model based on target variable transformation.

2.1 The double model estimator

The easiest and most intuitive way of estimating the vector β^U is to estimate β^T and β^C using two independent linear models and subtract their predictions. We will refer to this approach as the *double model* approach.

Definition 1 A vector $\hat{\beta_d^U}$ given by the formula

$$\hat{\beta_d^U} = (X^{T\prime}X^T)^{-1}X^{T\prime}y^T - (X^{C\prime}X^C)^{-1}X^{C\prime}y^C \tag{3}$$

is called the *double model estimator* of the parameter vector β^U.

Notice, that the estimator can be written as $\hat{\beta_d^U} = \hat{\beta^T} - \hat{\beta^C}$ where $\hat{\beta^T} = (X^{T\prime}X^T)^{-1}X^{T\prime}y^T$ and $\hat{\beta^C} = (X^{C\prime}X^C)^{-1}X^{C\prime}y^C$ are classical Ordinary Least Squares (OLS) estimators (Heumann et al. 2013) built independently in both groups.

[1] Methods for handling dependent assignment such as propensity score matching are beyond the scope of this work, see Sect. 1.1

Now we will look at the properties of the estimator in case when the matrix X is fixed (constant) which is typically considered in statistical textbooks (Heumann et al. 2013; Greene 2003). The case is of some practical importance in designed experiments, where the matrix X is under full control of the researcher, however for us it will serve mainly to illustrate some of the points we are going to make, since in most data mining applications the matrix X is random.

Theorem 1 *Assume that the predictor matrix X is fixed. Then the double model estimator $\hat{\beta}_d^U$ is BLUE (Best Linear Unbiased Estimator).*

The proof of this and the remaining theorems can be found in the Appendix.

From the theorem we can conclude that for fixed predictors the double model approach is in a certain sense optimal. Therefore, in order to find better estimators for uplift regression, the case of random predictors must be considered. Let us first give some results on the behavior of the double model estimator in this case. The following theorem shows that $\hat{\beta}_d^U$ is unbiased and asymptotically normal. The theorem thus extends typical statistical analysis of linear models (see e.g. Heumann et al. 2013; Greene 2003) to the case of uplift regression.

Theorem 2 *Assume that the predictor matrix X is random, $\mathrm{E}\, x_i = 0$, and $\mathrm{Var}\, x_i = \Sigma$. Assume further, that complete randomization was used. Then*

1. *$\hat{\beta}_d^U$ is unbiased, i.e. $\mathrm{E}\, \hat{\beta}_d^U = \beta^U$,*
2. *if, in addition, each row x_i of matrix the matrix X follows the normal distribution $\mathcal{N}_p(0, \Sigma)$ then $\mathrm{Var}\, \hat{\beta}_d^U = \left(\frac{(\sigma^T)^2}{n^T - p - 1} + \frac{(\sigma^C)^2}{n^C - p - 1} \right) \Sigma^{-1}$,*
3. *if $n \to \infty$ with the proportions q^T, q^C fixed, then $\sqrt{n}\left(\hat{\beta}_d^U - \beta^U\right) \xrightarrow{d} \mathcal{N}\left(0, \left(\frac{(\sigma^T)^2}{q^T} + \frac{(\sigma^C)^2}{q^C} \right) \Sigma^{-1}\right)$.*

The third part of the theorem states that for large n the estimator is approximately normally distributed and its covariance matrix is approximately equal to

$$\mathrm{Var}\, \hat{\beta}_d^U \approx \frac{1}{n} \left(\frac{(\sigma^T)^2}{q^T} + \frac{(\sigma^C)^2}{q^C} \right) \Sigma^{-1}.$$

Asymptotic variance is a convenient and frequently used tool to compare statistical models (Greene 2003).

For finite samples the covariance matrix can easily be computed only for normally distributed x_i's as shown in part 2. When the distribution of the covariates is not normal the situation may be very different as shown in the following example.

Example 1 Consider two binary predictor variables, independent of each other, each taking the value of 1 with probability $\frac{1}{2}$. Construct a data matrix $X \in \mathbb{R}^{n \times 2}$ by taking an n element sample from the joint distribution of the two variables. It is easy to see that the columns of X are equal with probability $\frac{1}{2^n}$. As a result, with probability $\frac{1}{2^n}$ the matrix $X'X$ is singular and an OLS estimator using X as predictors is undefined. Of course, even if the singularity is avoided the columns may become almost identical leading to an ill-conditioned $X'X$ matrix.

The double model estimator uses two separate models, each trained on a subset of the full dataset. Splitting the training data makes problems described in the above example worse (Muirhead 2005). The question therefore arises whether one can build a better model by estimating β^U directly on the whole dataset. Developing such models is currently the main research area in uplift modeling. In the next section we present and analyze one such estimator based on an outcome variable transformation.

2.2 The uplift regression estimator

The idea of using outcome variable transformations to build uplift models has first been proposed for classification. The idea is quite simple: the class in the control training set is reversed, both training sets are concatenated (possibly with some weighting of cases) and a single classifier is built on such a combined dataset. The approach was first mentioned by Lai (2006), rediscovered and formally justified in Jaśkowski and Jaroszewicz (2012), some additional analyses were provided in Kane et al. (2014).

As it turns out, the method can be adapted to the regression setting. Consider the following transformation $\tilde{y}$ of the target variable y:

$$\tilde{y_i} = \begin{cases} \frac{1}{q^T} y_i & \text{if } g_i = T, \\ -\frac{1}{q^C} y_i & \text{if } g_i = C \end{cases} \tag{4}$$

and define

Definition 2 A vector $\hat{\beta}_z^U$ given by the formula

$$\hat{\beta}_z^U = \left(X'X\right)^{-1} X'\tilde{y} \tag{5}$$

is called the *uplift regression estimator* of the parameter vector β^U.

The outcome in the control group is reversed similarly to the case of classification, but special weighting factors are required to obtain correct estimates. For example, when $q^T = q^C = \frac{1}{2}$ then the target needs to be multiplied by 2. The reason for this scaling becomes clear in the proof of Theorem 3 below, but we will first provide an intuitive explanation.

Take random variables $y_i \in \mathbb{R}, x_i \in \mathbb{R}^p$ and $g_i \in \{T, C\}$ understood as the response, predictors, and group assignment of case i. Define $\tilde{y}_i$ as in (4). We have

$$\begin{aligned} \mathrm{E}\,\tilde{y}_i|x_i &= \mathrm{E}_{g_i}\,\mathrm{E}(\tilde{y}_i|g_i, x_i) \\ &= P(g_i = T)\,\mathrm{E}\,(\tilde{y}_i|g_i = T, x_i) + P(g_i = C)\,\mathrm{E}\,(\tilde{y}_i|g_i = C, x_i) \\ &= q^T\,\mathrm{E}\left(\frac{1}{q^T} y_i|g_i = T, x_i\right) + q^C\,\mathrm{E}\left(-\frac{1}{q^C} y_i|g_i = C, x_i\right) \\ &= \mathrm{E}\,(y_i|g_i = T, x_i) - \mathrm{E}\,(y_i|g_i = C, x_i)\,, \end{aligned}$$

where the first equation follows from the law of total expectation (Billingsley 1995, Chapter 34) and the second from the definition of $\tilde{y}_i$. It can be seen that the expectation

of the transformed target variable equals the desired expected difference between responses in the treatment and control groups for an individual described by a feature vector x_i.

Another way to look at this estimator is to rewrite it as

$$\hat{\beta}_z^U = \frac{1}{q^T}\left(X'X\right)^{-1} X^{T\prime} y^T - \frac{1}{q^C}\left(X'X\right)^{-1} X^{C\prime} y^C,$$

which reveals that it is the double model estimator where the matrices $(X^{T\prime}X^T)^{-1}$ and $(X^{C\prime}X^C)^{-1}$ are replaced with an estimate $(X'X)^{-1}$ made on the full dataset (Heumann et al. 2013). Recall that due to randomization the true underlying covariance matrices are identical in the treatment and control groups (Robins and Hernán 2018), so one can expect that this estimate will be better than both $(X^{T\prime}X^T)^{-1}$ and $(X^{C\prime}X^C)^{-1}$. For example, looking at the binary data from Example 1 one concludes that the probability of observing a singular predictor matrix in the uplift regression estimator is $\frac{1}{2^n}$ compared to $\frac{1}{2^{\min\{n^T, n^C\}}}$ for the double model. This is a very significant difference in favor of $\hat{\beta}_z^U$. See Muirhead (2005) for quantitative results for normally distributed data.

The following theorem shows that, usually, those gains are, unfortunately, offset by other negative factors.

Theorem 3 *Assume that the predictor matrix X is random, $\mathrm{E}\, x_i = 0$, and $\mathrm{Var}\, x_i = \Sigma$. Assume further, that complete randomization was used. Then*

1. *$\hat{\beta}_z^U$ is unbiased,*
2. *if in addition $x_i \sim \mathcal{N}(0, \Sigma)$ and $n \to \infty$ with the proportions q^T, q^C fixed, then*

$$\sqrt{n}(\hat{\beta^U} - \beta^U) \xrightarrow{d} \mathcal{N}\left(0, \Sigma^{-1}\left(\frac{(\sigma^T)^2}{q^T} + \frac{(\sigma^C)^2}{q^C}\right) + bb' + \Sigma^{-1}\,\mathrm{Tr}(bb'\Sigma)\right),$$

where $b = \sqrt{\frac{q^C}{q^T}}\beta^T + \sqrt{\frac{q^T}{q^C}}\beta^C$.

We see that an additional quantity now appears in the asymptotic variance. To gain some intuition on how large it can become, assume $\Sigma = I$ and look at the trace of the asymptotic covariance, i.e. the sum of variances of all the coefficients. We have (noting that $\mathrm{Tr}(I) = p$ and $\mathrm{Tr}(bb') = \|b\|^2$, see e.g. Heumann et al. 2013)

$$\begin{aligned}\mathrm{Tr}\,\mathrm{Var}\,\hat{\beta}_z^U &\approx \frac{1}{n}\,\mathrm{Tr}\left(I\left(\frac{(\sigma^T)^2}{q^T} + \frac{(\sigma^C)^2}{q^C}\right) + bb' + I\,\mathrm{Tr}(bb')\right)\\ &= \frac{p}{n}\left(\frac{(\sigma^T)^2}{q^T} + \frac{(\sigma^C)^2}{q^C}\right) + \frac{(p+1)\|b\|^2}{n}.\end{aligned}$$

Thus the trace of the variance exceeds that of the double model estimator by $(p+1)\|b\|^2/n$, where the vector b is given in Theorem 3. Since the vector b can be arbitrarily large, the variance of $\hat{\beta}_z^U$ can be arbitrarily worse than that of the double model approach.

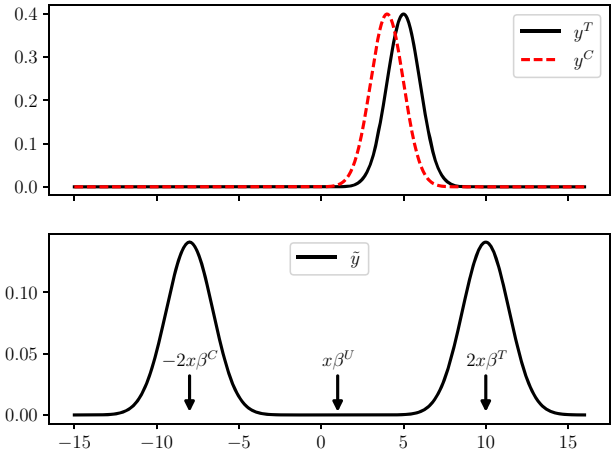

Fig. 1 Probability density functions of response variables in the double model (upper) and uplift regression (lower) conditional on a fixed feature vector x

Another issue to note is that, unlike other results in this paper and classical OLS analysis (Greene 2003), the second part of the theorem requires normally distributed predictors even in the asymptotic case. If the distribution is not normal the variance may be even larger depending on the fourth moments of x_i (for details see the proof, where a quartic form of x_i appears).

There is however a special case when uplift regression performs well:

Observation 1 *If in Theorem 3 the vector b is zero, the asymptotic variance of uplift regression is identical to that of the double model approach. It is the case, for example, if* $q^T = q^C = \frac{1}{2}$ *and* $\beta^T = -\beta^C$.

The experimental section will show that in this special (and rather unusual) case the method is actually superior to the other two methods analyzed in this paper. In the next section we provide an improved estimator, whose derivation is based on the above observation. But first let us give an intuition behind the increased variance of the uplift regression estimator and identify cases when it is most useful.

Figure 1 gives the intuition behind the increased variance of $\hat{\beta}_z^U$. The figure shows example probability density functions of response variables in the double model (upper) and uplift regression (lower) for a given fixed feature vector x. The plots can also be interpreted as densities of random error terms ε^T, ε^C, ε_z^U centered on the predicted expected values. The chart assumes $q^T = q^C = \frac{1}{2}$, $x\beta^T = 5$, $x\beta^C = 4$. In the double model approach both error terms have narrow densities with variances $(\sigma^T)^2$ and $(\sigma^C)^2$ respectively. On the other hand, the density of the error term in the uplift regression model is bimodal and much wider than both the treatment and control errors. This is a result of reversing the sign of $x\beta^C$ which brings it further apart from $x\beta^T$, as marked by the arrows in the picture. Notice however that the expected value coincides with $x\beta^U$ as desired.

In classical linear regression analysis the variance of the error term directly influences the covariance of estimated coefficients (Heumann et al. 2013; Greene 2003),

and the wider bimodal distribution clearly has higher variance. Indeed, using the law of total variance (Billingsley 1995, Chapter 34) we get

$$\begin{aligned}\operatorname{Var}\tilde{y}|x &= \operatorname{E}_g \operatorname{Var}\tilde{y}|x,g + \operatorname{Var}_g \operatorname{E}\tilde{y}|x,g \\ &= q^T \frac{1}{(q^T)^2}(\sigma^T)^2 + q^C \frac{1}{(q^C)^2}(\sigma^C)^2 + \operatorname{Var}_g \begin{cases} \frac{1}{q^T} x\beta^T & \text{if } g = T \\ -\frac{1}{q^C} x\beta^C & \text{if } g = C \end{cases} \\ &= \frac{(\sigma^T)^2}{q^T} + \frac{(\sigma^C)^2}{q^C} + xq^T q^C \left(\frac{\beta^T}{q^T} + \frac{\beta^C}{q^C}\right)^2,\end{aligned}$$

where g is the random variable representing group assignment, and the last quantity is the variance of a two-point distribution. Note that the last term is equal to xb where b is the vector defined in Theorem 3, so this simple argument gives an intuitive justification of its form.

To summarize this section, we list three cases when uplift regression estimator can be useful; otherwise the double model estimator should be used, or better, the corrected uplift regression estimator proposed in the next section.

1. n is small (e.g. close to p) or the matrix X is ill conditioned (see e.g. Example 1) so that the gains in better estimation of $(X'X)^{-1}$ outweigh the negative effects of increased error term variance.
2. Error variances $(\sigma^T)^2$ and $(\sigma^C)^2$ are huge compared to the actual responses such that the variance in Theorem 3 is dominated by the first term.
3. When $q^C\beta^T \approx -q^T\beta^C$. This however is rarely seen in practice where uplift is typically small and we have $\beta^T \approx \beta^C$ (Radcliffe and Surry 2011).

3 The corrected uplift regression estimator

In this section we introduce a new estimator which combines the benefits of the double regression model (low asymptotic variance) with the benefits of the uplift regression model (good estimation of the $(X'X)^{-1}$ matrix for small samples). Experiments in the next section suggest that this is the model of choice for linear uplift regression.

The main idea is to introduce corrections, such that the increase in variance described in Theorem 3 is diminished as much as possible. Looking at Eqs. 1 and 2 it is easy to see that subtracting some vector β^* from for both β^T and β^C does not affect β^U which is the quantity of interest.

Pick $\beta^* = q^C\beta^T + q^T\beta^C$. After subtracting β^* from β^T and β^C the vector b from Theorem 3 becomes

$$\begin{aligned}&\sqrt{\frac{q^C}{q^T}}(\beta^T - \beta^*) + \sqrt{\frac{q^T}{q^C}}(\beta^C - \beta^*) \\ &\quad = \sqrt{\frac{q^C}{q^T}}(\beta^T - q^C\beta^T - q^T\beta^C) + \sqrt{\frac{q^T}{q^C}}(\beta^C - q^C\beta^T - q^T\beta^C) = 0,\end{aligned}$$

so after making such a correction the increased variance of the uplift regression model would not be observed.

Unfortunately we cannot directly modify coefficient vectors β^T and β^C. We do not even know their exact values needed to compute β^*. To solve the second problem we will estimate β^* from data. Define:

$$y_i^* = \begin{cases} \frac{q^C}{q^T} y_i, & \text{if } g_i = T \\ \frac{q^T}{q^C} y_i, & \text{if } g_i = C, \end{cases}$$

and use the following estimator (i.e. the classical least squares estimator Heumann et al. 2013) applied to variable y^*

$$\hat{\beta}^* = (X'X)^{-1} X' y^*. \tag{6}$$

Since we cannot modify the true coefficient vectors β^T and β^C, we will instead modify (*correct*) the target vector y by subtracting $X\hat{\beta}^*$ from it. As a result we obtain the following two-stage estimator.

Definition 3 A vector $\hat{\beta}_c^U$ given by the formula

$$\hat{\beta}_c^U = (X'X)^{-1} X' \widetilde{y_c}$$

where

$$y_c = y - X\hat{\beta}^*$$

and

$$\hat{\beta}^* = (X'X)^{-1} X' y^*$$

is called the *corrected uplift regression estimator* of the parameter β^U.

The $\widetilde{y_c}$ is the result of applying Eq. 4 to y_c, and the estimator is in fact the uplift regression estimator from Definition 2 with the target vector y replaced by a *corrected* target vector y_c.

It is not obvious that such a two step regression procedure will give an improved estimator, since the correction may increase the variance of individual error terms (Heumann et al. 2013; Greene 2003). The following theorem shows that we do indeed obtain an improvement.

Theorem 4 *Assume that the predictor matrix X is random,* $\mathrm{E}\, x_i = 0$, *and* $\mathrm{Var}\, x_i = \Sigma$. *Assume further that complete randomization was used. Then*

1. *$\hat{\beta}^*$ is an unbiased estimator of β^*,*
2. *$\hat{\beta}_c^U$ is an unbiased estimator of β^U,*

3. *if* $n \to \infty$ *with the proportions* q^T, q^C *fixed, then* $\sqrt{n}(\hat{\beta^U} - \beta^U) \xrightarrow{d} N\left(0, \left(\frac{\sigma^{T^2}}{q^T} + \frac{\sigma^{C^2}}{q^C}\right)\Sigma^{-1}\right)$.

It can be seen that the corrected estimator has the same asymptotic distribution as the double regression estimator, thus recovering its low asymptotic variance. Furthermore, both estimators $\hat{\beta}^*$ and $\hat{\beta_c^U}$ are computed based on the full dataset using better estimates of $(X'X)^{-1}$, as does the uplift regression estimator. Notice also that the assumption of normality is no longer needed to compute asymptotic variance, in line with classical OLS (Greene 2003).

Finite sample variance of the corrected estimator proved difficult to compute but the experiments in the next section demonstrate that for small n it is superior to both double model and uplift regression estimators. Its performance becomes identical the double model approach as the sample size becomes large. Since the cost of computing the corrected estimator is not much larger than that of computing the other two estimators, we believe it is the estimator of choice for uplift linear regression.

4 Experimental evaluation

We will now evaluate the three proposed estimators on synthetic data and on two real life datasets. This way we are going to verify the theoretical results presented in the previous sections, and analyze the estimators' behavior on finite samples and real data which may not follow theoretical assumptions.

4.1 Synthetic data

We first evaluate the three estimators on synthetic data in order to illustrate their behavior in a controlled setting. We begin by describing our data generation procedure.

For each $p = 5, 20, 100$ we generated a random matrix X for growing values of n. Each row x_i of X is generated from the multivariate normal distribution $\mathcal{N}(0, I)$, with zero mean and unit covariance matrix. Each sample is assigned to treatment or control group using complete randomization with equal group proportions $q^T = q^C = \frac{1}{2}$. If n is odd, we arbitrarily choose $n^T = n^C + 1$. The outcome variables are then generated based on Eqs. 1 and 2 with $\varepsilon_i^T, \varepsilon_i^C \sim \mathcal{N}(0, 1)$.

The vectors β^T and β^C are generated randomly, in such a way that $\|\beta^T\| = \|\beta^C\| = 1$ and the angle between the two vectors has a specified value. This way we are able to perform tests for large and small uplift (compared to the responses) and simulate the special case from Observation 1. We tested with three different angles: $\frac{\pi}{10}$, corresponding to the case of small uplift frequently encountered in practice; π, corresponding to an ideal case for the uplift regression model, and an intermediate value of $\frac{\pi}{2}$.

To measure model performance we compare the estimated $\hat{\beta^U}$ with the true $\beta^U = \beta^T - \beta^C$ using the squared error

$$\|\hat{\beta^U} - \beta^U\|^2.$$

Since in this experiment we use synthetic data, the true β^U is known, and the error measure can be computed directly using coefficients estimated from the model. For each value of the angle, each estimator and each value of n the experiment has been repeated 100 times and the results averaged. We have only used $n \geq 2p$ to avoid having less records than variables in the double model. The results are shown in Table 1 and Figs. 2, 3 and 4.

Table 1 shows the actual values of squared errors for various estimators, angles between β^T and β^C, and sample sizes n. Let us first look at angles $\pi/10$ and $\pi/2$. It can be seen that for very small samples the double estimator has an extremely large error. This is the result of poor estimation of $(X^{T\prime}X^T)^{-1}$ and $(X^{C\prime}X^C)^{-1}$ since n^T and n^C are only slightly larger than p or even equal to it (see Muirhead 2005 for exact distributions of those matrices). The uplift regression estimator is better than the double estimator for relatively small n but quickly becomes worse. This shows that its applicability is in fact limited. On the other hand, the corrected uplift regression estimator is consistently better than both competitors even for thousands of training records and becomes comparable with the double estimator only for $n \approx 10^4$. Moreover, for $p = 100$ it maintains its lead for all n.

The last three columns of Table 1 show the case when $\beta^T = -\beta^C$ (angle equal to π) which should be optimal for the uplift regression model by Observation 1. We see that this is indeed the case: this model performs much better than the double estimator for a wide range of sample sizes. The corrected model is, however, not much worse and it actually performs better for very small n. Moreover the case of such a strong uplift rarely occurs in practice (Radcliffe and Surry 2011).

Overall the results fully confirm our theoretical analysis given in the previous two sections:

1. The uplift regression estimator outperforms the double estimator for small n due to better estimation of the matrix $(X'X)^{-1}$, for larger values of n it becomes significantly worse. In Table 1 its error is often more than twice as large.
2. The corrected uplift regression estimator combines the benefits of the two other estimators and outperforms them both for all n. One exception is the case $\beta^T = -\beta^C$, which is artificial and unlikely to occur in practice.
3. As n tends to infinity the errors of all models tend to zero. They all consistently estimate β^U.

One last issue requires a comment. Notice that, in the table, errors for the double estimator are exactly the same regardless of the angle between β^T and β^C. To see why this is the case, rewrite

$$\begin{aligned}\hat{\beta_d^U} - \beta^U &= (X^{T\prime}X^T)^{-1}X^{T\prime}y^T - (X^{C\prime}X^C)^{-1}X^{C\prime}y^C - \beta^T + \beta^C \\ &= (X^{T\prime}X^T)^{-1}X^{T\prime}(X^T\beta^T + \varepsilon^T) - (X^{C\prime}X^C)^{-1}X^{C\prime}(X^C\beta^C + \varepsilon^C) \\ &\quad - (X^{T\prime}X^T)^{-1}X^{T\prime}X^T\beta^T + (X^{C\prime}X^C)^{-1}X^{C\prime}X^C\beta^C \\ &= (X^{T\prime}X^T)^{-1}X^{T\prime}\varepsilon^T - (X^{C\prime}X^C)^{-1}X^{C\prime}\varepsilon^C\end{aligned}$$

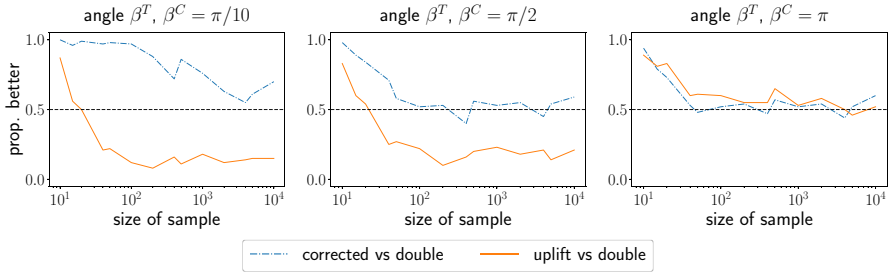

Fig. 2 Fraction of cases for which uplift and corrected uplift estimators are better than the double model estimator in terms of $\|\hat{\beta^U} - \beta^U\|^2$ for $p = 5$

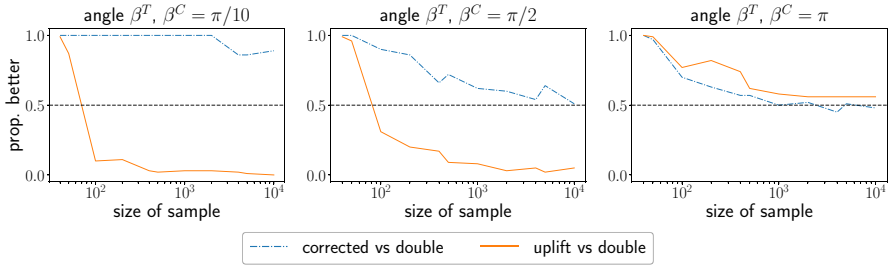

Fig. 3 Fraction of cases for which uplift and corrected uplift estimators are better than the double model estimator in terms of $\|\hat{\beta^U} - \beta^U\|^2$ for $p = 20$

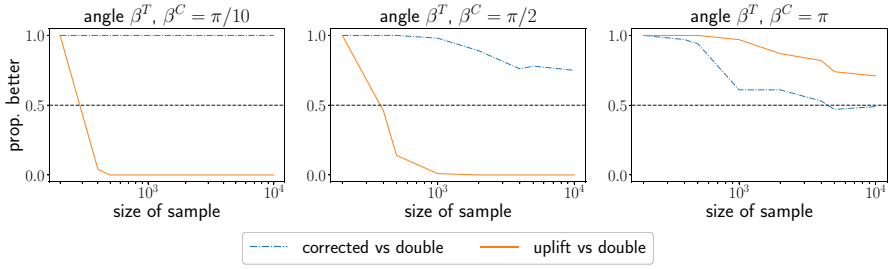

Fig. 4 Fraction of cases for which uplift and corrected uplift estimators are better than the double model estimator in terms of $\|\hat{\beta^U} - \beta^U\|^2$ for $p = 100$

and note that the resulting quantity does not depend on β^T, β^C. Since we use the same random seed for each angle, the double estimator was trained on the same X and ε in each case, yielding identical estimates.

Another kind of comparison is shown in Figures 2, 3 and 4, where we compare the fractions of random samples where uplift and corrected uplift estimators are better than the double model estimator in terms of the squared error $\|\hat{\beta^U} - \beta^U\|^2$. The fractions are shown as functions of n for various angles between β^T and β^C, for $p = 5, 20, 100$. When a plot hovers around 0.5 (horizontal line), the models are comparable. The figure confirms our earlier analysis: for small n both estimators outperform the double model. For larger n the uplift regression estimator performs poorly except for the special case

Table 1 Squared errors $\|\hat{\beta^U} - \beta^U\|^2$ averaged over 100 experiments

p	n	Angle between β^T and β^C								
		$\pi/10$			$\pi/2$			π		
		Double	Uplift	Corrected	Double	Uplift	Corrected	Double	Uplift	Corrected
5	10	8770.7	14.4907	**3.2398**	8770.7	13.8421	**4.5513**	8770.7	9.7570	**6.5290**
	15	6.7159	5.3407	**1.9347**	6.7159	4.5389	**2.3728**	6.7159	**2.9406**	3.1489
	20	2.9230	2.8976	**1.0411**	2.9230	2.4046	**1.4736**	2.9230	**1.4663**	1.7642
	40	0.8180	1.6082	**0.6113**	0.8180	1.2894	**0.6895**	0.8180	**0.7342**	0.7847
	50	0.5121	0.9844	**0.4091**	0.5121	0.7954	**0.4780**	0.5121	**0.4511**	0.5390
	100	0.1868	0.4296	**0.1673**	0.1868	0.3162	**0.1764**	0.1868	**0.1756**	0.1839
	200	0.0890	0.2037	**0.0841**	0.0890	0.1533	**0.0861**	0.0890	**0.0871**	0.0888
	400	0.0419	0.0980	**0.0408**	**0.0419**	0.0778	0.0426	0.0419	**0.0412**	0.0434
	500	0.0402	0.0868	**0.0393**	0.0402	0.0646	**0.0398**	0.0402	**0.0388**	0.0402
	1000	0.0207	0.0409	**0.0205**	0.0207	0.0318	**0.0205**	0.0207	0.0204	0.0204
	2000	0.0102	0.0222	**0.0101**	0.0102	0.0161	**0.0101**	0.0102	0.0101	0.0101
	4000	0.0047	0.0111	**0.0046**	0.0047	0.0078	**0.0046**	0.0047	0.0046	0.0046
	5000	0.0040	0.0084	0.0040	0.0040	0.0065	0.0040	0.0040	0.0040	0.0040
	10,000	0.0017	0.0040	0.0017	0.0017	0.0029	0.0017	0.0017	0.0017	0.0017

Table 1 continued

p	n	Angle between β^T and β^C								
		$\pi/10$			$\pi/2$			π		
		Double	Uplift	Corrected	Double	Uplift	Corrected	Double	Uplift	Corrected
20	40	16,787.2	6.8272	**1.9052**	16,787.2	4.7506	**2.8541**	16787.2	3.9995	**3.6350**
	50	8.6742	4.1472	**1.5253**	8.6742	3.1359	**2.2192**	8.6742	**2.4466**	2.8407
	100	1.3789	2.0805	**0.8543**	1.3789	1.6313	**1.0398**	1.3789	**1.0968**	1.2173
	200	0.5168	0.8324	**0.4066**	0.5168	0.6488	**0.4433**	0.5168	**0.4462**	0.4851
	400	0.2268	0.3797	**0.2033**	0.2268	0.2916	**0.2136**	0.2268	**0.2147**	0.2208
	500	0.1790	0.3152	**0.1646**	0.1790	0.2460	**0.1707**	0.1790	**0.1720**	0.1758
	1000	0.0887	0.1715	**0.0853**	0.0887	0.1262	**0.0871**	0.0887	**0.0870**	0.0891
	2000	0.0424	0.0812	**0.0415**	0.0424	0.0628	**0.0418**	0.0424	**0.0420**	0.0427
	4000	0.0202	0.0406	**0.0200**	0.0202	0.0308	**0.0202**	0.0202	**0.0201**	0.0205
	5000	0.0158	0.0319	**0.0156**	0.0158	0.0240	**0.0157**	0.0158	**0.0157**	0.0158
	10,000	0.0081	0.0157	0.0081	0.0081	0.0117	0.0081	0.0081	0.0081	0.0081
100	200	250,184.4	7.4751	**2.0098**	250,184.4	5.7524	**2.9794**	250184.4	3.9564	**3.8201**
	400	2.0006	2.6848	**1.0107**	2.0006	2.0632	**1.2525**	2.0006	**1.3549**	1.5111
	500	1.3370	1.9676	**0.7989**	1.3370	1.5406	**0.9566**	1.3370	**0.9854**	1.1140
	1000	0.4977	0.8861	**0.3988**	0.4977	0.6685	**0.4376**	0.4977	**0.4446**	0.4782
	2000	0.2252	0.4217	**0.2021**	0.2252	0.3151	**0.2111**	0.2252	**0.2129**	0.2225
	4000	0.1094	0.2123	**0.1040**	0.1094	0.1593	**0.1066**	0.1094	**0.1064**	0.1094
	5000	0.0813	0.1614	**0.0781**	0.0813	0.1210	**0.0800**	0.0813	**0.0795**	0.0820
	10,000	0.0410	0.0801	**0.0401**	0.0410	0.0609	**0.0406**	0.0410	**0.0406**	0.0411

Best models are highlighted in bold

The models used are: the double regression model (*double*), the uplift regression model based on target variable transformation (*uplift*) and the proposed corrected uplift regression model (*corrected*)

$\beta^T = -\beta^C$. The proposed, corrected uplift regression estimator performs well in all cases, especially in the practically important case of small uplift, when $\beta^T \approx \beta^C$. In this case the advantage does not vanish even for $n = 10{,}000$ and is especially pronounced for $p = 100$, where the corrected model is better in all cases.

Overall, we believe that Figures 2, 3 and 4 demonstrate that it is beneficial to use the corrected model, even if the reduction in the error is small (Table 1, large values of n). Its computational complexity is not much larger than that of the double model and the results are consistently better, especially for larger values of p.

4.2 Real data: a social program

In this section we describe the experiments we performed on the first real dataset. We begin by describing the dataset and the experimental methodology, later we present actual results.

4.2.1 Description of the dataset

The dataset we use comes from a program (called OSNAP) to motivate children to higher physical activity (Giles et al. 2016). The data describes 401 children who have been randomly split into two groups: these who participated in the OSNAP program, and remaining ones who did not. Children's activity was measured using special equipment. We will focus on the outcome variable `tot_dur_mv1`, which describes daily minutes of overall moderate to vigorous physical activity. We chose this variable since this is the principal indicator in the study. Because each child was observed for several days we average the observations for each child.

We first removed all post-randomization variables: `valdays`, `mn_avgfree_outdoor`, `mn_avgfree_indoor`, `mn_avgstruc_outdoor`, `mn_avgstruc_indoor`, `allmean_min_wr`, `mn_avgfreepa`. Further, we removed the variables `mn_pctdavg_mean_temp26p` and `pctdavg_mean_temp26p` since they were identical to a third variable. This dataset is called OSNAP_1. Additionally, we created a set OSNAP_2 by removing the following variables which were highly correlated with each other: `mtype`, `mn_avgpa`, `mn_avginpa`, `mn_avgstrucpa`, `mn_avgfree_indoor`, `grade` (we did keep indicators for each specific grade), `gradeK_2`, `mn_avgoutpa`, and `pair`. There are 34 variables in the OSNAP_1 dataset and 24 variables in OSNAP_2. This way we obtained two training sets with differing levels of collinearity (Heumann et al. 2013).

4.2.2 Methodology

Since for real data we do not have access to true coefficient vectors, we had to use test sets. The treatment and control groups have both been randomly split into two parts, creating treatment and control train and test sets. The models were of course built on the training sets. The use of the test sets requires additional discussion.

Since we do not know the true model coefficients, we decided to compare model's predictive performance. This, however, is also difficult since, due to the Fundamental

Problem of Causal Inference (see Sect. 1 and Holland 1986), for each case we only know its outcome after the treatment or without it. As a result, the true gain with which models' prediction should be compared is never known at the level of individual data records (Radcliffe and Surry 2011; Rzepakowski and Jaroszewicz 2012). To solve this problem we will compare model predictions with actual outcomes on *subsets* of records. Similar ideas have been used to test other approaches to causal inference (Shalit et al. 2017) and to test classification uplift models (Radcliffe and Surry 2011; Rzepakowski and Jaroszewicz 2012).

Let $S \subseteq \{1, \ldots, n\}$ be indices of cases in such a subset. Of course S should only contain test cases. Define $S^T = \{i \in S : g_i = T\}$ and $S^C = \{i \in S : g_i = C\}$. We will use the following absolute error measure to assess model performance

$$\left| \frac{1}{|S^T|} \sum_{i \in S^T} x_i \hat{\beta}^U - \left(\frac{1}{|S^T|} \sum_{i \in S^T} y_i^T - \frac{1}{|S^C|} \sum_{i \in S^C} y_i^C \right) \right|, \tag{7}$$

that is we compare average uplift over all cases in S^T, as predicted by the model, with uplift over S^T computed based on actual outcomes in this subset. We pick subsets which are continuous slices of given width along a random projection. To select a subset we first generate a random vector and project the data onto this vector. Let $p \in [0, 1]$ be the proportion of cases which should be included in S. We pick for S indices of all cases whose projections fall between quantiles r and $r + q$, where r is generated uniformly at random from $[0, 1 - p]$. We chose three possible values for p: 10, 50 and 100%; the last value amounts to comparing true and predicted uplift on the full test set.

The whole process was repeated 1000 times and the results averaged to make the experiments repeatable.

4.2.3 Results

The results for OSNAP_1 are shown in Fig. 5 and for OSNAP_2 in Fig. 6. The training set contained 110 cases in both treatment and control groups. The test sets were of similar size.

It can be seen that the size of the test subset does not affect the results. It is evident that when highly correlated variables are present the double model fails completely. This is due to the fact that some of those variables are categorical, and, even though $X'X$ is nonsingular, either $X^{T'}X^T$ or $X^{C'}X^C$ may be; see Example 1. Both other estimators provide meaningful results with the proposed corrected estimator giving vastly superior estimates.

When highly correlated variables are removed (OSNAP_2 dataset, Fig. 6) the performance of the double model improves dramatically. It is now much better than the uplift estimator, however the proposed corrected estimator maintains its lead.

While testing on full data (rightmost figures results for other test set sizes are similar), the averaged absolute error (Eq. 7) of the double model was 2.96 and of the corrected model 2.42. This is a decrease of about 18%. Moreover, the corrected

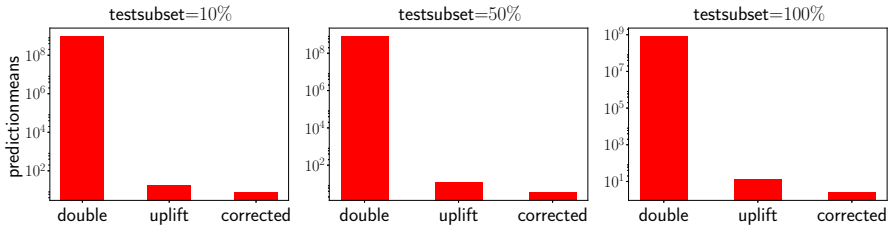

Fig. 5 Differences between predicted and true uplift for three estimators on the $OSNAP_1$ data computed on random subsets of size 10, 50, and 100%. $n^T = n^C = 110$. The models used are the double regression model (*double*), the uplift regression model based on target variable transformation (*uplift*) and the proposed corrected uplift regression model (*corrected*)

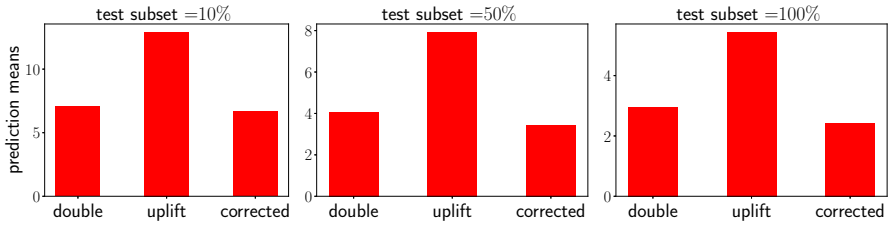

Fig. 6 Differences between predicted and true uplift for three estimators on the $OSNAP_2$ data computed on random subsets of size 10, 50, and 100%. $n^T = n^C = 110$. The models used are the double regression model (*double*), the uplift regression model based on target variable transformation (*uplift*) and the proposed corrected uplift regression model (*corrected*)

uplift regression model was more accurate than the double regression model in 73.4% of the subsamples. This shows that on the $OSNAP_2$ data, the proposed approach is clearly superior. The corrected model is not much more expensive to build, but offers systematically lower prediction errors.

4.2.4 Interpretation of the models

To further validate the presented approach, in this section we analyze coefficients of the three proposed models on the $OSNAP_2$ data.

The coefficients were obtained by building the models on the full $OSNAP_2$ dataset, but to evaluate the significance of the coefficients we used subsampling estimates (Politis et al. 1999) of standard deviations of the coefficients. To this end we sampled 66% of the data 1000 times, built the three models on each subsample and computed standard deviations of the coefficients over the subsamples. Table 2 lists coefficients which were at least one standard deviation away from zero for at least one model.

It can be seen that the double model and the corrected uplift regression produced very similar coefficients, with almost identical significance patterns. Uplift regression coefficients have much larger absolute values, and more of them are significant (but usually only at 1 standard deviation). Overall, however, the signs of coefficients were identical (a significant exception being `mn_mean_temp26p` discussed below) suggesting that all models discovered similar phenomena.

Table 2 Averaged coefficients for variables in the $OSNAP_2$ data

Variable	Double		Corrected		Uplift	
	Coeff.	Std.	Coeff.	Std.	Coeff.	Std.
Intercept	− 35.74*	18.15	− 36.32**	13.95	84.42**	41.09
gtmeter	− 17.38*	8.71	− 9.53*	7.9	− 0.91	11.84
mn_mean_temp26p	0.61**	0.28	0.56**	0.23	− 1.07*	0.64
ndays_precip	0.9	2.6	1.29	2.07	8.6*	5.77
sex=female	4.28*	4.2	4.04*	3.84	5.52	10.14
grade=2	− 5.75	10.66	− 4.97	7.34	− 35.15*	23.65
grade=3	− 7.32	10.11	− 5.23	7.29	− 39.85*	22.67
grade=4	− 7.65	10.62	− 5.58	7.45	− 30.68*	22.71
grade=5	− 4.84	10.44	− 2.65	7.8	− 34.9*	23.49
race=1	− 1.66	10.49	− 1.99	8.3	− 35.52*	21.05
race=2	− 7.56*	5.88	− 7.23*	5.08	− 31.45**	13.87
race=4	8.0	10.08	4.97	8.44	24.22*	21.36

Standard deviation estimates are based on resampling. Only variables significant in at least one model are shown. Stars show statistical significance of the variables, the number of stars is the number of standard deviations away from zero

We tried to verify the coefficients using external sources. The original OSNAP paper (Giles et al. 2016) did little stratified analysis, indicating only that sex did not significantly influence outcomes and that children in the first grade responded more positively. We found that sex has a mild influence on program success, but the results are not strongly significant; the second statement is reflected in Table 2, indicators for `grade` > 2 all have negative coefficients, but the results are only mildly significant.

We have also found two papers (Harrison et al. 2017; Edwards et al. 2015) discussing the effect of weather on children's physical activity. While they are purely observational studies, both indicate strong influence of outside temperature. Our models indicate that temperature is important also when one tries to actively encourage children to exercise (variable `mn_mean_temp26p`: mean temperature between 2 and 6 p.m.). Surprisingly the uplift regression model gave negative coefficient value in this case. We were unable to explain this phenomenon. The `ndays_precip` variable indicates the number of rainy days during the study period. In Harrison et al. (2017) and Edwards et al. (2015) it was shown that precipitation is negatively correlated with exercise, but our models suggest, that in those cases encouragement can actually be more effective.

The `gtmeter` variable which was found significant by the double model and the corrected uplift regression is related to the type of physical activity meter used. We find it plausible that the type of meter influences the measurements. This may indicate that using different types of meters in the study might have affected the final results.

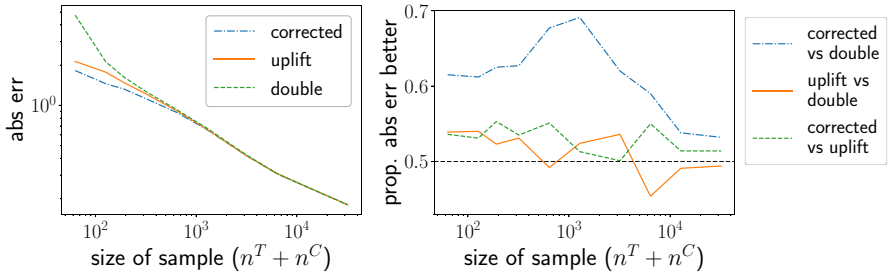

Fig. 7 Model performance on a direct marketing campaign dataset. Sample size is the sum of treatment and control group sizes, treatment group size being 2/3 of the total. The models used are the double regression model (*double*), the uplift regression model based on target variable transformation (*uplift*), and the proposed corrected uplift regression model (*corrected*)

4.3 Real data: a marketing campaign

In this section we evaluate model performance on a real life marketing dataset obtained from Kevin Hillstrom's MineThatData blog (Hillstrom 2008). It contains results of an e-mail campaign for an Internet based retailer. The dataset contains information about 64,000 customers and includes basic features such as the amount of money spent in the previous year or when the last purchase was made. A total of 10 variables is available. The customers have been randomly split into three groups: the first received an e-mail campaign advertising men's merchandise, the second, a campaign advertising women's merchandise, and the third was kept as control. Our target variable was the amount spent by each person during the two weeks following the campaign (Hillstrom 2008). We combined both e-mail groups into a single treatment group.

Since the data is fairly large, but contains few variables, we compare performance of the three analyzed models for increasing sample sizes. We used the same testing methodology as for the OSNAP data. The test sets were randomly selected 50% subsets of treatment and control groups. Training sets of various sizes were sampled from the remaining records. The train/test selection procedure was repeated 1000 times and the results averaged. The outcomes are shown in Fig. 7. The left chart shows the averaged absolute error (Eq. 7), with S^T and S^C being full treatment and control test sets. The right chart shows the proportion of subsamples on which one of the models was better than another for all three pairs of models.

The left chart shows that all models' errors become almost identical for $n \approx 1000$, but for lower n the differences are more pronounced. The right chart shows, that even for larger n, the differences remain detectable. Overall the advantage of corrected uplift regression model over the double model is clear and quite stable. Its advantage over uplift regression is smaller but still visible.

Good performance of the uplift regression model (which consistently outperforms the double model until n reaches about 1000) requires a comment since it seems to contradict our earlier results. We believe this to be a result of a specific target variable distribution: most customers made no purchase at all. Out of 64,000 customers, only 578 made a purchase and, as a result, the distribution of the target variable has a sharp

peak at 0. Most target values are thus unaffected by the outcome variable transformation given in Eq. 4 and the phenomenon shown in Fig. 1 does not occur.

5 Conclusions, limitations and future research

The paper presented a detailed analysis of current approaches to uplift regression: the double model approach and the approach based on outcome variable transformation. Contrary to popular belief (Radcliffe and Surry 2011; Kuusisto et al. 2014; Guelman et al. 2012), we found that the double model approach usually performs much better than the outcome transformation approach. Based on our analysis we also proposed a third estimator, which combines the benefits of both methods, and which we believe to be the estimator of choice for uplift regression.

A limitation of the proposed method is that the theoretical framework currently only applies to linear regression. We believe this to be an important first step, future work will involve extending the above analysis to other kinds of machine learning models. Another limitation is that, currently, the influence of regularization on model performance is not included in the analysis. Future research will address designing optimal regularization strategies for uplift regression.

Appendix: Proofs

Proof (of Theorem 1) From classical theory of ordinary least squares (Heumann et al. 2013) we know that $\mathrm{E}\,\hat{\beta}^T = \beta^T$ and $\mathrm{E}\,\hat{\beta}^C = \beta^C$ so $\mathrm{E}\,\hat{\beta}_d^U = \mathrm{E}\,\hat{\beta}^T - \mathrm{E}\,\hat{\beta}^C = \beta^T - \beta^C = \beta^U$ proving that $\hat{\beta}_d^U$ is indeed unbiased.

The rest of the proof roughly follows the proof of Gauss–Markov theorem (Greene 2003, Chapter 4) in both groups. Any linear estimator of β^U has the form $\hat{\beta}^U = Ay$ for some matrix A. W.l.o.g. it can be rewritten as

$$\hat{\beta}^U = \left[\begin{array}{c|c} (X^{T\prime}X^T)^{-1}X^T + D_1 & D_2 \\ \hline D_3 & -(X^{C\prime}X^C)^{-1}X^C + D_4 \end{array}\right] \left[\begin{array}{c} y^T \\ \hline y^C \end{array}\right]$$

Denoting $D^T = D_1 + D_3$, $D^C = -(D_2 + D_4)$ we get

$$\hat{\beta^U} = \left[(X^{T\prime}X^T)^{-1}X^{T\prime} + D^T\right] y^T - \left[(X^{C\prime}X^C)^{-1}X^{C\prime} + D^C\right] y^C.$$

Using Eqs. 1 and 2 and the fact that $\mathrm{E}\,\varepsilon^T = \mathrm{E}\,\varepsilon^C = 0$ we get

$$\mathrm{E}\,\hat{\beta^U} = [(X^{T\prime}X^T)^{-1}X^{T\prime} + D^T](X^T\beta^T + \mathrm{E}\,\varepsilon^T)$$

$$
\begin{aligned}
&- [(X^{C\prime}X^C)^{-1}X^{C\prime} + D^C](X^C\beta^C + \mathrm{E}\,\varepsilon^C)\\
&= \beta^T - \beta^C + D^T X^T \beta^T - D^C X^C \beta^C.
\end{aligned}
$$

For $\hat{\beta^U}$ to be unbiased this expectation must equal β^U for all possible values of β^T, β^C giving $D^T X^T = 0$ and $D^C X^C = 0$. $\hat{\beta^U}$ is the sum of two parts, one based on the control the other on treatment data. Since the parts are independent, the variance of $\hat{\beta^U}$ is their sum. We have

$$
\begin{aligned}
&\mathrm{Var}[(X^{C\prime}X^C)^{-1}X^{C\prime} + D^C]y^C = [(X^{C\prime}X^C)^{-1}X^{C\prime} + D^C](\mathrm{Var}\,y^C)[(X^{C\prime}X^C)^{-1}X^{C\prime} + D^C]'\\
&= (\sigma^C)^2(X^{C\prime}X^C)^{-1}X^{C\prime}X^C(X^{C\prime}X^C)^{-1} + (\sigma^C)^2(X^{C\prime}X^C)^{-1}(D^C X^C)'\\
&\quad + (\sigma^C)^2 D^C X^C (X^{C\prime}X^C)^{-1} + (\sigma^C)^2 D^C D^{C\prime}\\
&= (\sigma^C)^2((X^{C\prime}X^C)^{-1} + D^C D^{C\prime}).
\end{aligned}
$$

Since $D^C D^{C\prime}$ is nonnegative definite the above variance cannot be less than $(\sigma^C)^2(X^{C\prime}X^C)^{-1}$. Repeating the above steps for the second part we see that the variance of $\hat{\beta^U}$ cannot be lower than $(\sigma^C)^2(X^{C\prime}X^C)^{-1} + (\sigma^T)^2(X^{T\prime}X^T)^{-1}$, that is the variance achieved by the double model estimator. □

Proof (of Theorem 2) We have $\mathrm{E}\,\hat{\beta^T} = \mathrm{E}(X^{T\prime}X^T)^{-1}X^{T\prime}y^T = (X^{T\prime}X^T)^{-1}X^{T\prime}X^T\beta^T + (X^{T\prime}X^T)^{-1}X^{T\prime}\,\mathrm{E}\,\varepsilon^T$. Since X^T and ε^T are independent and $\mathrm{E}\,\varepsilon^T = 0$ the second term vanishes giving $\mathrm{E}\,\hat{\beta^T} = \beta^T$ also for random X^T. Analogous result for the control group allows us to reuse the reasoning from the proof of part 1 of Theorem 1.

Using the law of total variance

$$
\mathrm{Var}\,\hat{\beta^T} = \mathrm{Var}_{X^T}\,\mathrm{E}(\hat{\beta^T}|X^T) + \mathrm{E}_{X^T}\,\mathrm{Var}(\hat{\beta^T}|X^T) = \mathrm{Var}_{X^T}\,\beta^T + (\sigma^T)^2\,\mathrm{E}_{X^T}(X^{T\prime}X^T)^{-1}.
$$

The first term is equal to zero since β^T is a constant. Since x_i is assumed to follow multivariate normal distribution with zero mean $(X^{T\prime}X^T)^{-1}$ follows the inverse Wishart distribution whose variance is $\Sigma^{-1}/(n-p-1)$ (Muirhead 2005). Noting that $\hat{\beta^T}$ and $\hat{\beta^C}$ are independent we get

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta_d^U}) &= \mathrm{Var}(\hat{\beta^T}) + \mathrm{Var}(\hat{\beta^C})\\
&= (\sigma^T)^2\frac{\Sigma^{-1}}{n^T - p - 1} + (\sigma^C)^2\frac{\Sigma^{-1}}{n^C - p - 1}.
\end{aligned}
$$

For the proof of the last part we will use classical asymptotic results for ordinary least squares (Greene 2003, Chapter 4) which state that

$$
\sqrt{n^T}(\hat{\beta^T} - \beta^T) \xrightarrow{d} \mathcal{N}\left(0, (\sigma^T)^2\Sigma^{-1}\right) \quad \sqrt{n^C}(\hat{\beta^C} - \beta^C) \xrightarrow{d} \mathcal{N}\left(0, (\sigma^C)^2\Sigma^{-1}\right),
$$

implying

$$\begin{aligned}\sqrt{n}(\hat{\beta}_d^U - \beta^U) &= \frac{\sqrt{n}}{\sqrt{n^T}}\sqrt{n^T}(\hat{\beta^T} - \beta^T) + \frac{\sqrt{n}}{\sqrt{n^C}}\sqrt{n^C}(\hat{\beta^C} - \beta^C) \\ &\xrightarrow{d} \frac{1}{\sqrt{q^T}}\mathcal{N}\left(0, (\sigma^T)^2\Sigma^{-1}\right) + \frac{1}{\sqrt{q^C}}\mathcal{N}\left(0, (\sigma^C)^2\Sigma^{-1}\right) \\ &= \mathcal{N}\left(0, \left(\frac{(\sigma^T)^2}{q^T} + \frac{(\sigma^C)^2}{q^C}\right)\Sigma^{-1}\right).\end{aligned}$$

□

Before proving the next two theorems let us introduce a useful lemma.

Lemma 1 *Suppose the data was generated using a model given in Eqs. 1 and 2, and that complete randomization was used. The predictor matrix X is assumed to be random with $\mathrm{E}\, x_i = 0$, $\mathrm{Var}\, x_i = 0$. Let γ, δ be two constants and d be a random vector defined as*

$$d = (X'X)^{-1}[\gamma X^{T'} y^T + \delta X^{C'} y^C].$$

Then

1. $\mathrm{E}(X'X)^{-1} X^{T'} X^T = q^T I$ *and* $\mathrm{E}(X'X)^{-1} X^{C'} X^C = q^C I$,
2. $\mathrm{E}\, d = q^T \gamma \beta^T + q^C \delta \beta^C$.
3. *As $n \to \infty$ with q^T, q^C held fixed*

$$\sqrt{n}(d - \mathrm{E}\, d) \xrightarrow{d} \mathcal{N}\left(0, \Sigma^{-1}(q^T(\gamma\sigma^T)^2 + q^C(\delta\sigma^C)^2) + q^C q^T \Sigma^{-1} \Delta \Sigma^{-1}\right)$$

   where $\Delta = \mathrm{Var}(x_i' x_i \breve{b})$ is the covariance matrix of the random vector $x_i' x_i \breve{b}$ and $\breve{b} = \gamma\beta^T - \delta\beta^C$
4. *If in addition $x_i \sim \mathcal{N}(0, \Sigma)$, then*

$$\Delta = \Sigma \breve{b}\breve{b}'\Sigma + \Sigma\, \mathrm{Tr}(\breve{b}\breve{b}'\Sigma)$$

Proof To prove the first statement of the lemma write $(X'X)^{-1} X^{T'} X^T = \sum_{i:g_i=T} (X'X)^{-1} x_i x_i'$. Notice now that random matrices $(X'X)^{-1} x_i x_i'$ are identically distributed (but not independent) regardless of g_i due to symmetry and randomization. They therefore have identical expectations and

$$\begin{aligned}\mathrm{E}(X'X)^{-1} X^{T'} X^T &= \sum_{i:g_i=T} \mathrm{E}(X'X)^{-1} x_i x_i' = n^T\, \mathrm{E}(X'X)^{-1} x_1 x_1' = \frac{n^T}{n} n\, \mathrm{E}(X'X)^{-1} x_1 x_1' \\ &= q^T \sum_{i=1}^{n} \mathrm{E}(X'X)^{-1} x_i x_i' = q^T\, \mathrm{E}(X'X)^{-1} \sum_{i=1}^{n} x_i x_i' = q^T\, \mathrm{E}(X'X)^{-1} X'X = q^T I.\end{aligned}$$

The proof for the control group is analogous. Now, using Eqs. 1 and 2 and noting $\mathrm{E}\, \varepsilon^T = \mathrm{E}\, \varepsilon^C = 0$ we get

$$\mathrm{E}\, d = \mathrm{E}(X'X)^{-1}[\gamma X^{T'} y^T + \delta X^{C'} y^C]$$

$$= \mathrm{E}(X'X)^{-1}[\gamma X^{T\prime} X^T \beta^T + \gamma X^{T\prime} \varepsilon^T + \delta X^{C\prime} X^C \beta^C + \delta X^{C\prime} \varepsilon^C]$$
$$= \gamma\, \mathrm{E}(X'X)^{-1} X^{T\prime} X^T \beta^T + \delta\, \mathrm{E}(X'X)^{-1} X^{C\prime} X^C \beta^C = q^T \gamma \beta^T + q^C \delta \beta^C,$$

where the last equality follows from the first part.

For the proof of the third part introduce the vector $\breve{\varepsilon}$, which plays the role of the random component if d is used as an estimator of regression coefficients. It is defined as

$$\breve{\varepsilon}_i = \begin{cases} \gamma y_i - x_i(q^T \gamma \beta^T + q^C \delta \beta^C) & \text{if } g_i = T \\ \delta y_i - x_i(q^T \gamma \beta^T + q^C \delta \beta^C) & \text{if } g_i = C. \end{cases} \tag{8}$$

To show asymptotic normality we need to count the expectation and variance of $x_i \breve{\varepsilon}_i$ conditional on g_i. First apply Eqs. 1 and 2 to y_i and recall that $\mathrm{E}\,\varepsilon_i^T = \mathrm{E}\,\varepsilon_i^C = 0$ to get

$$\mathrm{E}(\breve{\varepsilon}_i | x_i, g_i) = \begin{cases} \gamma x_i \beta^T - x_i(q^T \gamma \beta^T + q^C \delta \beta^C) = x_i(\gamma \beta^T (1 - q^T) - q^C \delta \beta^C) & \text{if } g_i = T \\ \delta x_i \beta^C - x_i(q^T \gamma \beta^T + q^C \delta \beta^C) = x_i(-q^T \gamma \beta^T + (1 - q^C) \delta \beta^C) & \text{if } g_i = C \end{cases}$$
$$= x_i(\gamma \beta^T - \delta \beta^C) \begin{cases} q^C & \text{if } g_i = T \\ -q^T & \text{if } g_i = C \end{cases} = x_i \breve{b} \begin{cases} q^C & \text{if } g_i = T \\ -q^T & \text{if } g_i = C. \end{cases} \tag{9}$$

Clearly $\mathrm{E}(x_i' \breve{\epsilon}_i | x_i, g_i) = x_i' \mathrm{E}(\breve{\epsilon}_i | x_i, g_i)$ and from the law of total expectation we now get

$$\mathrm{E}(x_i' \breve{\epsilon}_i | g_i) = \mathrm{E}_{x_i}(\mathrm{E}(x_i' \breve{\epsilon}_i | x_i, g_i) | g_i)$$
$$= \mathrm{E}_{x_i}(x_i' x_i | g_i) \breve{b} \begin{cases} q^C & \text{if } g_i = T, \\ -q^T & \text{if } g_i = C \end{cases} = \Sigma \breve{b} \begin{cases} q^C & \text{if } g_i = T, \\ -q^T & \text{if } g_i = C, \end{cases}$$

where the last equality follows from the fact that the distribution of x_i does not depend on g_i (randomization) and that the mean of x_i is zero so $\mathrm{E}\, x_i' x_i$ is simply the covariance matrix of x_i. From Eqs. 1 and 2 it follows that

$$\mathrm{Var}(\breve{\epsilon}_i | x_i, g_i) = \begin{cases} (\gamma \sigma^T)^2 & \text{if } g_i = T \\ (\delta \sigma^C)^2 & \text{if } g_i = C, \end{cases}$$

since after conditioning the only random components are ε^T and ε^C. This implies $\mathrm{Var}(x_i' \breve{\epsilon}_i | x_i, g_i) = x_i' x_i \,\mathrm{Var}(\breve{\epsilon}_i | x_i, g_i)$. From the law of total variance we get

$$\mathrm{Var}(x_i' \breve{\epsilon}_i | g_i = T) = \mathrm{E}_{x_i}(\mathrm{Var}(x_i' \breve{\epsilon}_i | x_i, g_i = T) | g_i = T) + \mathrm{Var}_{x_i}(\mathrm{E}(x_i' \breve{\epsilon}_i | x_i, g_i = T) | g_i = T)$$
$$= (\gamma \sigma^T)^2 \mathrm{E}_{x_i}(x_i' x_i | g_i = T) + \mathrm{Var}_{x_i}(x_i' x_i \breve{b} q^C | g_i = T)$$
$$= \Sigma (\gamma \sigma^T)^2 + (q^C)^2 \mathrm{Var}_{x_i}(x_i' x_i \breve{b} | g_i = T) = \Sigma (\gamma \sigma^T)^2 + (q^C)^2 \Delta,$$

where the last equality is a consequence of randomization. After obtaining an analogous result for $g_i = C$ we can write

$$\operatorname{Var}(x_i'\check{\epsilon}_i|g_i) = \begin{cases} \Sigma(\gamma\sigma^T)^2 + \Delta(q^C)^2 & \text{if } g_i = T, \\ \Sigma(\delta\sigma^C)^2 + \Delta(q^T)^2 & \text{if } g_i = C. \end{cases}$$

The argument below is very similar to classical derivation of asymptotic normality of OLS (Greene 2003). To obtain the limiting distribution of d write

$$\begin{aligned} \lim\sqrt{n}(d - \mathrm{E}\, d) &= \lim\sqrt{n}\left((X'X)^{-1}[X^{T'}\gamma y^T + X^{C'}\delta y^C] - (X'X)^{-1}[X^{T'}X^T + X^{C'}X^C]\,\mathrm{E}\, d\right) \\ &= \lim\sqrt{n}(X'X)^{-1}\left(X^{T'}(\gamma y^T - X^T\,\mathrm{E}\, d) + X^{C'}(\delta y^C - X^C\,\mathrm{E}\, d)\right) \\ &= \lim\sqrt{n}(X'X)^{-1}X'\check{\epsilon} = \lim\left(\frac{1}{n}X'X\right)^{-1}\frac{1}{\sqrt{n}}X'\check{\epsilon}. \end{aligned}$$

From the law of large numbers and $\mathrm{E}\, x_i = 0$ we have $\lim\frac{1}{n}X'X = \Sigma$ and by continuous mapping theorem (Greene 2003) $\lim(X'X/n)^{-1} = \Sigma^{-1}$. Since $(X'X/n)^{-1}$ converges to a constant, we can apply Slutsky's theorem (Greene 2003) to move it before the limit obtaining

$$\lim\sqrt{n}(d - \mathrm{E}\, d) = \Sigma^{-1}\lim\frac{1}{\sqrt{n}}X'\check{\epsilon} = \Sigma^{-1}\lim\frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_i\check{\epsilon}_i. \tag{10}$$

Computing the last limit is tricky since x_i's are not independent due to complete randomization (see Sect. 1.2) so the Central Limit Theorem cannot be directly applied. However, $x_i\check{\epsilon}_i$'s are independent conditional on g, and are identically distributed within each group. So we will compute the limit separately in the treatment and control group using only conditional variances and expectations such that the Central Limit Theorem can be applied. Notice first that

$$q^T\,\mathrm{E}(x_i'\check{\epsilon}_i|g_i = T) + q^C\,\mathrm{E}(x_i'\check{\epsilon}_i|g_i = C) = q^T\Sigma\check{b}q^C - q^C\Sigma\check{b}q^T = 0$$

so we may write

$$\begin{aligned} \frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_i'\check{\epsilon}_i &= \frac{1}{\sqrt{n}}\left(\sum_{g_i=T}x_i'\check{\epsilon}_i + \sum_{g_i=C}x_i'\check{\epsilon}_i\right) - \frac{n}{\sqrt{n}}\left(q^T\,\mathrm{E}(x_i'\check{\epsilon}_i|g_i = T) + q^C\,\mathrm{E}(x_i'\check{\epsilon}_i|g_i = C)\right) \\ &= \frac{1}{\sqrt{n}}\left(\sum_{g_i=T}x_i'\check{\epsilon}_i - n^T\,\mathrm{E}(x_i'\check{\epsilon}_i|g_i = T) + \sum_{g_i=C}x_i'\check{\epsilon}_i - n^C\,\mathrm{E}(X_i'\check{\epsilon}_i|g_i = C)\right) \\ &= \frac{\sqrt{n^T}}{\sqrt{n}}\frac{1}{\sqrt{n^T}}\sum_{g_i=T}(x_i'\check{\epsilon}_i - \mathrm{E}(x_i'\check{\epsilon}_i|g_i = T)) + \frac{\sqrt{n^C}}{\sqrt{n}}\frac{1}{\sqrt{n^C}}\sum_{g_i=C}(x_i'\check{\epsilon}_i - \mathrm{E}(x_i'\check{\epsilon}_i|g_i = C)) \\ &\xrightarrow{d}\sqrt{q^T}\mathcal{N}(0, \operatorname{Var}(x_i'\check{\epsilon}_i|g_i = T)) + \sqrt{q^C}\mathcal{N}(0, \operatorname{Var}(x_i'\check{\epsilon}_i|g_i = C)) \\ &= \mathcal{N}(0, q^T\operatorname{Var}(x_i'\check{\epsilon}_i|g_i = T) + q^C\operatorname{Var}(x_i'\check{\epsilon}_i|g_i = C)), \end{aligned}$$

where the limit in distribution follows from the Central Limit Theorem. Noting

$$q^T \operatorname{Var}(x_i' \breve{\epsilon}_i | g_i = T) + q^C \operatorname{Var}(x_i' \breve{\epsilon}_i | g_i = C) = \Sigma(q^T(\gamma\sigma^T)^2 + q^C(\delta\sigma^C)^2) + \Delta q^C q^T$$

and substituting into Eq. 10 we get

$$\begin{aligned}\lim \sqrt{n}(d - \mathrm{E}\, d) &\xrightarrow{d} \Sigma^{-1}\mathcal{N}(0, \Sigma(q^T(\gamma\sigma^T)^2 + q^C(\delta\sigma^C)^2) + \Delta q^C q^T) \\ &= \mathcal{N}(0, \Sigma^{-1}(q^T(\gamma\sigma^T)^2 + q^C(\delta\sigma^C)^2) + q^C q^T \Sigma^{-1}\Delta\Sigma^{-1}).\end{aligned}$$

To prove part 4 rewrite the expression for the variance as (using a vector analogue of the formula $\operatorname{Var} x = \mathrm{E}\, x^2 - (Ex)^2$)

$$\begin{aligned}\operatorname{Var}(x_i' x_i \breve{b}) &= \mathrm{E}_{x_i}(x_i' x_i \breve{b}\breve{b}' x_i' x_i | g_i) - \mathrm{E}_{x_i}(x_i' x_i \breve{b} | g_i)\, \mathrm{E}_{x_i}(x_i' x_i \breve{b} | g_i)' \\ &= \mathrm{E}_{x_i}(x_i' x_i \breve{b}\breve{b}' x_i' x_i | g_i) - \Sigma \breve{b}\breve{b}' \Sigma\end{aligned}$$

Since $x_i \sim \mathcal{N}(0, \Sigma)$, from (Petersen and Pedersen 2012, Section 8.2.4) we have

$$\mathrm{E}_{x_i}(x_i' x_i \breve{b}\breve{b}' x_i' x_i | g_i) = 2\Sigma \breve{b}\breve{b}' \Sigma + \Sigma \operatorname{Tr}(\breve{b}\breve{b}' \Sigma)$$

which completes the proof. □

Proof (of Theorem 3) To prove part 1 take $\gamma = 1/q^T$ and $\delta = -1/q^C$ and apply part 2 of Lemma 1. Part 2 follows from Lemma 1 parts 3 and 4 after taking $\gamma = 1/q^T$ and $\delta = -1/q^C$ and noting $q^C q^T \Sigma^{-1}\Delta\Sigma^{-1} = bb' + \Sigma^{-1}\operatorname{Tr}(bb'\Sigma)$. □

Proof (of Theorem 4) To prove part 1 take $\gamma = q^C/q^T$ and $\delta = q^T/q^C$ and apply Lemma 1. For parts 2 and 3 let us rewrite the estimator as follows

$$\hat{\beta}_c^U = \frac{1}{q^T}(X'X)^{-1}(X^{T\prime}(y^T - X^T\hat{\beta}^*)) - \frac{1}{q^C}(X'X)^{-1}(X^{C\prime}(y^C - X^C\hat{\beta}^*)) \quad (11)$$

and expand the first term of the sum using Eq. 2

$$\begin{aligned}&\frac{1}{q^T}(X'X)^{-1}(X^{T\prime}(y^T - X^T\hat{\beta}^*)) \\ &\quad = \frac{1}{q^T}(X'X)^{-1}(X^{T\prime}X^T\beta^T + X^{T\prime}\varepsilon^T - X^{T\prime}X^T\beta^* - X^{T\prime}X^T(\hat{\beta}^* - \beta^*)) \\ &\quad = \frac{1}{q^T}(X'X)^{-1}X^{T\prime}(X^T(\beta^T - \beta^*) + \varepsilon^T) - \frac{1}{q^T}(X'X)^{-1}X^{T\prime}X^T(\hat{\beta}^* - \beta^*)).\end{aligned}$$

Using an analogous expansion for the second term in Eq. 11 and defining $\breve{y}^T = X^T(\beta^T - \beta^*) + \varepsilon^T$ and $\breve{y}^C = X^C(\beta^C - \beta^*) + \varepsilon^C$ we can write $\hat{\beta}_c^U = \beta_1 + \beta_2$ where

$$\beta_1 = \frac{1}{q^T}(X'X)^{-1}X^{T\prime}\breve{y}^T - \frac{1}{q^C}(X'X)^{-1}X^{C\prime}\breve{y}^C$$

$$\beta_2 = -\frac{1}{q^T}(X'X)^{-1}X^{T'}X^T(\hat{\beta}^* - \beta^*) + \frac{1}{q^C}(X'X)^{-1}X^{C'}X^C(\hat{\beta}^* - \beta^*).$$

Taking $\gamma = 1/q^T$, $\delta = -1/q^C$, by Lemma 1 part 2 we get $\mathrm{E}\,\beta_1 = \beta^T - \beta^C = \beta^U$ and by part 3, noting that $\breve{b} = 0$, and consequently $\Delta = \mathrm{Var}(x_i'x_i\breve{b}) = 0$ we have

$$\sqrt{n}(\beta_1 - \beta^U) \xrightarrow{d} \mathcal{N}\left(0, \Sigma^{-1}\left(\frac{(\sigma^T)^2}{q^T} + \frac{(\sigma^C)^2}{q^C}\right)\right) \tag{12}$$

so to complete the proof is suffices to show that β_2 has zero expectation and that its variance vanishes as $n \to \infty$.

Let us begin with demonstrating $\mathrm{E}\,\beta_2 = 0$. Let ε^* be defined as in Eq. 8 with $\gamma = q^C/q^T$ and $\delta = q^T/q^C$. Notice that by the law of total expectation and Eq. 9

$$\begin{aligned}\mathrm{E}\,\varepsilon_i^*|x_i &= \mathrm{E}_{g_i}\,\mathrm{E}(\varepsilon_i^*|x_i, g_i) = P(g_i = T)x_i\breve{b}q^C - P(g_i = C)x_i\breve{b}q^T \\ &= q^T x_i\breve{b}q^C - q^C x_i\breve{b}q^T = 0.\end{aligned}$$

Rewrite

$$\begin{aligned}(X'X)^{-1}X^{T'}X^T(\hat{\beta}^* - \beta^*) &= (X'X)^{-1}X^{T'}X^T\left[(X'X)^{-1}X'y^* - (X'X)^{-1}X'X\beta^*\right] \\ = (X'X)^{-1}X^{T'}X^T(X'X)^{-1}X'\left[X\beta^* + \varepsilon^* - X\beta^*\right] &= (X'X)^{-1}X^{T'}X^T(X'X)^{-1}X'\varepsilon^*.\end{aligned}$$

Using total expectation again

$$\mathrm{E}(X'X)^{-1}X^{T'}X^T(X'X)^{-1}X'\varepsilon^* = \mathrm{E}_X(X'X)^{-1}X^{T'}X^T(X'X)^{-1}X'\,\mathrm{E}(\varepsilon^*|X) = 0.$$

Repeating the above argument for $(X'X)^{-1}X^{C'}X^C(\hat{\beta}^* - \beta^*)/q^C$ completes the proof of this part. We will now prove the desired asymptotic behavior of β_2.

$$\begin{aligned}\sqrt{n}\beta_2 &= n(X'X)^{-1}\left(-\frac{1}{n^T}X^{T'}X^T + \frac{1}{n^C}X^{C'}X^C\right)\sqrt{n}(\hat{\beta}^* - \beta^*) \\ &= \left(\frac{1}{n}X'X\right)^{-1}\left(-\frac{1}{n^T}X^{T'}X^T + \frac{1}{n^C}X^{C'}X^C\right)\sqrt{n}(\hat{\beta}^* - \beta^*).\end{aligned}$$

We know (see proof of part 3 of Lemma 1) that $\left(\frac{1}{n}X'X\right)^{-1} \to \Sigma^{-1}$, $\frac{1}{n^T}X^{T'}X_T \to \Sigma$, and $\frac{1}{n^C}X^{C'}X_C \to \Sigma$. Therefore

$$\left(\frac{1}{n}X'X\right)^{-1}\left(-\frac{1}{n^T}X^{T'}X^T + \frac{1}{n^C}X^{C'}X^C\right) \to 0.$$

By Lemma 1 part 3 $\sqrt{n}(\hat{\beta}^* - \beta^*)$ converges in distribution to a multivariance normal. Applying Slutsky's theorem to the product of the two quantities we get $\beta_2 \to 0$ completing the proof of part 3. □

References

Billingsley P (1995) Probability and measure. Wiley, New York

Edwards NM, Myer GD, Kalkwarf HJ (2015) Outdoor temperature, precipitation, and wind speed affect physical activity levels in children: a longitudinal cohort study. J Phys Act Health 12(8):1074–1081

Giles C, Cradock A, Barrett J et al (2016) Promoting physical activity with the out of school nutrition and physical activity (OSNAP) initiative: a cluster-randomized controlled trial. JAMA Pediatr 170(2):155–162

Greene WH (2003) Econometric analysis. Pearson Education, New York

Guelman L, Guillén M, Pérez-Marín AM (2012) Random forests for uplift modeling: an insurance customer retention case. In: Modeling and simulation in engineering, economics and management, volume 115 of Lecture Notes in Business Information Processing (LNBIP), pp 123–133. Springer

Harrison F, Goodman A, van Sluijs EMF (2017) Weather and childrens physical activity; how and why do relationships vary between countries? Int J Behav Nutr Phys Act 14:74

Heumann C, Nittner T, Rao CR, Scheid S, Toutenburg H (2013) Linear models: least squares and alternatives. Springer, New York

Hillstrom K (2008) The MineThatData e-mail analytics and data mining challenge. MineThatData blog. http://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html. Retrieved on 19.04.2018

Holland PW (1986) Statistics and causal inference. J Am Stat Assoc 81(396):945–960

Hoyer PO, Janzing D, Mooij JM, Peters J, Schölkopf B (2009) Nonlinear causal discovery with additive noise models. In: Advances in neural information processing systems, pp 689–696

Imbens GW, Rubin DB (2015) Causal inference for statistics, social, and biomedical sciences: an introduction. Cambridge University Press, New York

Jaśkowski M, Jaroszewicz S (2012) Uplift modeling for clinical trial data. In: ICML 2012 workshop on machine learning for clinical data analysis, Edinburgh

Johansson FD, Shalit U, Sontag D (2016) Learning representations for counterfactual inference. In: Proceedings of the 33rd international conference on international conference on machine learning, ICML'16, pp 3020–3029

Kane K, Lo VSY, Zheng J (2014) Mining for the truly responsive customers and prospects using true-lift modeling: comparison of new and existing methods. J Market Anal 2(4):218–238

Kuusisto F, Costa VS, Nassif H, Burnside E, Page D, Shavlik J (2014) Support vector machines for differential prediction. In: ECML-PKDD

Lai LY-T (2006) Influential marketing: a new direct marketing strategy addressing the existence of voluntary buyers. Master's thesis, Simon Fraser University

Muirhead RJ (2005) Aspects of multivariate statistical theory. Wiley-Interscience, Hoboken

Pearl J (2009) Causality. Cambridge University Press, Cambridge

Pechyony D, Jones R, Li X (2013) A joint optimization of incrementality and revenue to satisfy both advertiser and publisher. In: WWW 2013 companion

Petersen KB, Pedersen MS (2012) The matrix cookbook. Version 20121115

Politis DN, Romano JP, Wolf M (1999) Subsampling. Springer, New York

Radcliffe NJ, Surry PD (2011) Real-world uplift modelling with significance-based uplift trees. Portrait Technical Report TR-2011-1, Stochastic Solutions

Robins J (1994) Correcting for non-compliance in randomized trials using structural nested mean models. Commun Stat Theory Methods 23(8):2379–2412

Robins JM, Hernán MA (2018) Causal inference. Chapman & Hall/CRC, Boca Raton (forthcoming)

Rzepakowski P, Jaroszewicz S (2012) Decision trees for uplift modeling with single and multiple treatments. Knowl Inf Syst 32:303–327

Rzepakowski P, Jaroszewicz S (2010) Decision trees for uplift modeling. In: Proc. of the 10th IEEE international conference on data mining (ICDM), pp 441–450, Sydney, Australia

Shalit U, Johansson FD, Sontag D (2017) Estimating individual treatment effect: generalization bounds and algorithms. In: Proceedings of the 34th international conference on machine learning, vol 70 of Proceedings of machine learning research, pp 3076–3085, Sydney, Australia, 06–11

Sołtys M, Jaroszewicz S, Rzepakowski P (2014) Ensemble methods for uplift modeling. Data Min Knowl Discov 1–29. online first

Spirtes P, Glymour CN, Scheines R (2000) Causation, prediction, and search. MIT press, Cambridge

Spirtes P, Zhang K (2016) Causal discovery and inference: concepts and recent methodological advances. In: Applied informatics, 3(1)
Zaniewicz Ł, Jaroszewicz S (2013) Support vector machines for uplift modeling. In: The first IEEE ICDM workshop on causal discovery (CD 2013), Dallas
Zaniewicz Ł, Jaroszewicz S (2017) l_p-support vector machines for uplift modeling. Knowl Inf Syst 53(1):269–296