Erasmus University Rotterdam

Master Thesis

Causal effects of binary, continuous and multivariate treatments using propensity score methods in a marketing context

Author:
Aniek Markus

Student number:
383550

Supervisor:
dr. Mikhail Zhelonkin

Second assessor:
dr. Andrea Naghi

Company supervisor:
dr. Bert de Bruijn

Business Analytics and Quantitative Marketing

Erasmus School of Economics

August 14, 2019

Abstract

In this thesis we aim to estimate the causal effect of multiple advertisements on the tune-in of a TV program. We contribute to the current literature by applying propensity score methods in a generalized problem setting and relatively new domain. Furthermore, we introduce a combination of average and treated dose-response functions, investigate estimating treatment effects for a low dimensional multivariate (instead of bivariate) treatment variable and introduce a smooth coefficient model for multivariate treatments. In particular, we study the effect of multiple TV advertisements on the tune-in of the season premiere of America's Got Talent in 2016. Starting from the most commonly studied case in the literature, binary treatments, we extend the analysis to continuous and multivariate treatments. We explore the use of many different methods and prefer *CBPS* as treatment assignment model for binary treatments and *Poisson regression* for continuous/multivariate treatments in our case. We find small treatment effects, depending on the treatment variable(s) used. Additional exposures to advertising have a positive impact on the probability of tune-in. Furthermore, the results suggest more recent advertisements have a higher impact and advertisements on the same channel (NBC) are most effective.

Keywords: propensity score methods, continuous treatments, multivariate treatments, promotion response, dose-response function, CBPS, smooth coefficient model.

Contents

1 Introduction

Companies typically spend a substantial part of their total revenue on marketing. Measuring and evaluating the effectiveness of marketing campaigns is an important part of the marketing process, that is expected to gain even more attention coming years (Nielsen, 2018). Marketing agencies want to get insight in the return on money spend and aim to find the optimal marketing strategy tailored to their campaign objective. It is thus of high importance to make good quality estimates of the effect of different media channels. Specifically, we want to isolate the effect that is due to the used marketing resources. In other words, the difference with what happened had the campaign not taken place. Hence, we are interested in the causal effect of advertising.

Causality is the relationship between two variables, connecting a cause and effect. It is generally understood that correlation does not imply causation, but what does is less clear. Whereas correlation is easy to find, causation is much harder to establish. The fundamental problem of causal inference is that we can observe at most one outcome for each person (Holland, 1986). Hence, causal inference studies the prediction of counterfactuals (i.e. the outcome(s) for the treatment(s) each person did not receive). The 'gold standard' to estimate causal effects is the randomized controlled trial (RCT), as the randomization involved minimizes the selection bias allowing for a direct comparison between the treated and non-treated group. However, RCTs are often costly or might not be allowed due to ethical concerns. In the marketing domain, we are also often limited to observational data, as will be the case in this thesis.

The research in this thesis is conducted on behalf of Pointlogic, a Nielsen Division. It is interested to measure the effect of media campaigns for customers. A typical problem set up is that multiple marketing channels are used simultaneously during a campaign period and individuals have different levels of exposure for each of the marketing channels because of differences in media consumption. In this thesis we focus on studying the effect of TV advertisements on the tune-in of a TV program. Here, the different marketing channels are the type of TV channels the show can be promoted on and the moment in time the advertisements take place. We define treatment as exposure to advertising.

Typically, researchers employ a simple, often inappropriate, regression-based approach to evaluate the effect size of marketing interventions (Rubin & Waterman, 2006). In this thesis, we examine the use of propensity score methods (PSM), introduced by Rosenbaum and Rubin (1983). PSMs are a widely applied set of techniques to study the effect of treatments based on the Neyman-Rubin causal model[1]. The Neyman-Rubin causal model defines causal effects as comparisons of potential outcomes. The potential outcomes are the possible values the outcome variable Y can take, the causal effect is defined as the difference between the potential outcomes. For example, let Y_{1i} be the outcome for individual i in case of treatment ($T_i = 1$) and Y_{0i} the outcome in the absence of treatment ($T_i = 0$, control group). Ideally, one would compare the difference $E[Y_{1i}|T_i = 1] - E[Y_{0i}|T_i = 0]$ for each individual i. However, we can never observe both for any individual i. Furthermore, a direct comparison across individuals between the treatment and control group is likely to be biased in a non-experimental setting. PSMs aim to correct this bias by comparing outcomes between individuals who are as similar as possible. For binary treatments, the propensity score is defined as the probability of receiving treatment given observed covari-

[1] Also known as the potential outcome framework.

ates (Rosenbaum & Rubin, 1983). Individuals with the same propensity score will have a similar distribution of observed covariates (Austin, 2011). The key idea of PSMs is to use the propensity score as a balancing score to reduce the bias due to lack of randomization.

Similar to conventional regression-based approaches, PSMs only reduce the non-random differences in treatment related to measured confounders[2]. However, PSMs have several advantages compared to the former (Austin, 2011). As we focus on modelling the treatment instead of the outcome variable, (1) it is simpler to do model checks and (2) we are not influenced by the impact of model changes on the effect we try to estimate. Using propensity scores, we can carefully check if the covariate distributions between different treatment levels are similar without using the outcome variable. This makes PSMs (3) much easier to analyse and interpret. Nevertheless, PSMs are not well studied to estimate the effect of marketing interventions (Rubin & Waterman, 2006). Moreover, the number of PSM studies analysing generalized treatments is still limited. That is, studies investigating the causal effect of categorical, ordinal, continuous or multivariate treatment variables.

Therefore, this thesis investigates the following main question:

> How can reliable estimates of the causal effect of multiple TV advertisements on the tune-in of a popular TV program be obtained using PSMs?

In particular, we look at the following subquestions:

- How can the treatment effect in the case of binary and continuous univariate treatment variables be identified and estimated using PSMs? How can this be extended to multivariate treatment variables?
- What is the bias introduced in the estimated treatment effect by not taking the non-random treatment assignment into account correctly (as often in practice)?
- What is our recommendation for causal inference under these circumstances?

Most causal inference literature studies the effect of treatment versus non-treatment, which is an example of the binary treatment case. Therefore, we start this thesis by exploring binary treatments, indicating whether an individual was exposed to (a certain level of) advertising or not. Ultimately, we are not interested in the effect of binary treatments. Instead, we want to estimate the effect of multiple treatments that can each differ in dose. Therefore, we extend the analysis to allow for different doses or intensities of advertising, which we call continuous treatments. Finally, we model the case where individuals can be exposed to multiple treatments simultaneously (multivariate treatments). This means they can be exposed to different types of advertising, each with a certain intensity. We contribute to the current literature by applying PSMs methods in a generalized problem setting and relatively new domain. This is relevant for future applications.

The remainder of this thesis is structured as follows. Section 2 provides some background on causal inference for the interested reader. In Section 3, we discuss applications of PSMs in the literature related to this research. Section 4 contains a description of the data set. Next, Section 5 describes the used methodology for binary, continuous and multivariate treatments respectively. The results are summarized in Section 6. Finally, Section 7 discusses the findings, limitations and angles for future research.

[2]Variables influencing both the outcome and treatment variable are called confounders.

2 Background

This section gives a short introduction to the domain of causal inference. First, Section 2.1 distinguishes causal inference from standard statistical analysis. Next, the conventional regression-based approach is put in perspective in Section 2.2. Finally, Section 2.3 discusses three common approaches to causal inference.

2.1 What is causal inference?

Causal inference aims to make cause-effect statements and studies the effect of arbitrary, hypothetical interventions. Standard statistical analysis (i.a. regression analysis) is concerned with estimating the parameters of a distribution from a sample of the distribution. This can be used to make claims about how certain events are regularly associated among each other under similar conditions (Pearl et al., 2009). The aim to predict the belief of events under changing instead of static conditions distinguishes causal inference from standard statistical analysis. To achieve this it is necessary to make causal assumptions and state these explicitly so that it is clear under which assumptions a certain results holds. Causal assumptions are basic beliefs on the relationship of variables, based on the knowledge of the researcher, that remain valid when external conditions change. These assumptions can in principle not be tested in observational data. For a more elaborate discussion on the concept of causal inference in relation to standard statistical analysis we refer to Holland (1986) and Pearl et al. (2009).

2.2 Conventional regression analysis

In regression analysis, the estimated effect of a variable represents the change in response associated with that variable, holding a given set of control variables constant. We can simply assume these estimates are causal. In fact, if we observe all relevant variables that could affect an outcome and compare units with the same value for all variables other than the treatment variable, we know the effect is indeed causal if the outcome changes. However, this is rarely the case and regression analysis would be redundant under these circumstances as we could compare individuals directly. In practice, we are almost always missing variables. As long as we observe enough variables, we can still assume treatment is as good as randomly assigned given these observed variables. The problem, however, is that we might not be able to hold all other variables constant while changing the variable of interest.

Standard covariate adjustment is often insufficient when the number of covariates is large and the covariate distribution varies substantially with treatment (Imbens, 2000). Furthermore, it is unknown which set of variables should be controlled for. In practice, researchers often try to correct for as many variables as possible (Pearl & Mackenzie, 2018). However, because not all variables are observed, this is not necessarily the best strategy as illustrated by Pearl's M-bias example in Figure 2.

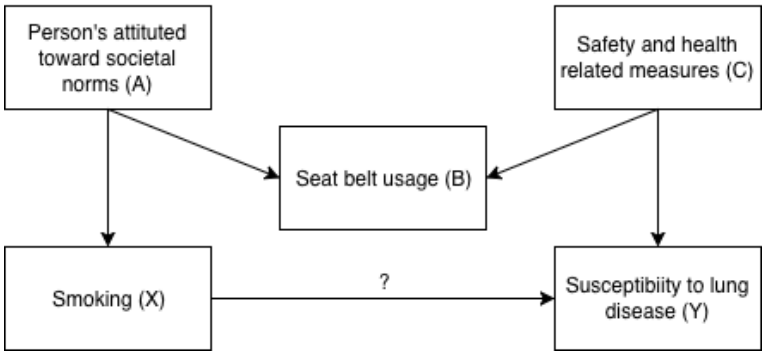


Figure 2: Pearl's M-bias example

We are interested in the effect of X on Y. In the above diagram, one can see X and Y are unconfounded (i.e. there is no common factor influencing both). However, if we control for B this becomes a confounder of X and Y. This is exactly what happened in a study as part of a tobacco litigation, because they found that B was correlated with both X and Y. If one would also control for A or C this is not a problem, however this is not possible as these variables are not observed. This is a real life example that shows controlling for all variables can result in undesired outcomes.

As the propensity score is a balancing score, we can use diagnostics for PSMs to check whether covariate balance is achieved. In conventional regression analysis, it is more difficult to determine whether the model is correctly specified (Austin, 2011). Moreover, it can be that there is a clear distinction in covariates between different treatment levels. In the regression based-approach this might not be noticed as results are extrapolated over groups with different treatment levels. PSMs allow the researcher to explicitly examine the underlying distribution of covariates between groups (Austin, 2011). Finally, we do not need to make linearity assumptions between the treatment and response variable in PSMs as we construct a separate model for the treatment variable.

Studies comparing treatment effects estimated by PSMs and conventional regression analysis find small differences, with PSMs resulting in slightly weaker associations (Shah, Laupacis, Hux, & Austin, 2005; Stürmer et al., 2006). However, using simulation Martens, Pestman, de Boer, Belitser, and Klungel (2008) show differences can be substantial and the size depends on several factors such as the number of influential variables, the magnitude of the treatment effect and the incidence proportion. Furthermore, they find PSMs tend to be closer to the true marginal effect.

2.3 Different causal frameworks

The fundamental question of causality arises in many fields. From these disciplines, different approaches have developed over time. Three common frameworks to formulate causal models are: the Neyman-Rubin causal model, structural equation models and the structural causal model. These frameworks enable the researcher to describe the cause-effect relationship using formal notation and make underlying assumptions explicit. The crucial distinction between the above frameworks is that structural equation models and the structural causal model do not require the assumption that selection takes place on observable variables only. However, this comes at the cost of other assumptions (e.g. distribution of the error terms).

The first framework, and the approach used in this thesis, is the Neyman-Rubin causal model (Rosenbaum & Rubin, 1983). The Neyman-Rubin causal model is the approach to causal analysis based on the potential outcome framework first introduced by Neyman (1923). Developed by statisticians, it defines causal effects as comparisons of potential outcomes (see Section 1). The problem is that we can never observe more than one outcome for each individual. The Neyman-Rubin causal model views causal inference essentially as a missing data problem (Rubin, 2005).

Next, structural equation models are the traditional econometric approach to causal analysis is. Contrary to the Neyman-Rubin causal model, structural equation models explicitly model the relationships between variables. Two main components are distinguished: the structural model and the measurement model. The first models the dependencies between the (possibly endogenous) explanatory variables and the outcome variable, the second the underlying latent structure of the endogenous variable(s). Together, they can model the selection bias resulting from a non-random observation of the dependent variable. An early application of this approach was discussed by Roy (1951), who studied workers choosing a hunting or fishing job based on their productivity. More formally, Heckman (1976) discussed the problem of selection bias and suggested a two-step method to solve the above problem. Models of self-selection fall into a broader class of switching-regression models (Maddala, 1983). The difference with endogenous switching regression models is that in these models both regimes are observed partially, whereas in Heckman's selection model only one observed regime is observed and of interest. The typical approach when modelling selection bias is based on assumptions about the error distribution. In the most standard case one assumes a normal model, but some more general distribution assumptions have been investigated. Nevertheless, this approach is criticised for its reliance on distributional assumptions and lack of robustness to departures from normality (Heckman, Tobias, & Vytlacil, 2000). An outline of this approach is presented in Appendix 9.4 as an angle for future research, but we focus on PSMs instead.

Finally, the structural causal model combines the potential outcome framework, graphical models and structural equation models. This approach originates from computer science. Although the representation differs, structural equation models and the structural causal models are basically two sides of the same coin. Instead of the equations, the structural causal model uses directed acyclic graphs (DAG). Causal graphs originate from path analysis introduced by Wright (1920) and express the conditional dependence structure by using a set of nodes and arcs. Nodes represent random variables and the (absence of) arcs conditional independence assumptions. Once the graphical model is developed, do-calculus can be used to express interventions (Pearl, 1995). In a graph, do(.) removes in-going edges into the target of intervention, while keeping out-going edges. This notation allows researchers to make a difference between observing ($P(Y|X, Z)$) and doing ($P(Y|\text{do}(X), Z)$), where the difference between the two is the result of hidden common causes (i.e. unmeasured confounding). For more details on this approach, we refer the interested reader to Pearl et al. (2009) as this thesis will not further investigate the use of this approach.

3 Literature review

This section provides an overview of the relevant applications of PSMs in the literature. First, we discuss PSMs applications related to marketing in Section 3.1. Next, Section 3.2 gives an overview of applications that focus on generalized treatments.

3.1 Propensity score methods in marketing

Since the introduction of PSMs by Rosenbaum and Rubin (1983), the number of new publications has increased substantially each year (Stürmer et al., 2006; Thoemmes & Kim, 2011). Although PSMs have been applied in many areas, the majority of research applications studies medical or epidemiological interventions (Rubin & Waterman, 2006). The popularity of PSMs in this field can be explained due to the fact that RCTs raise ethical questions in these situations and might take too long (e.g. for chronic diseases). Within social sciences, most PSMs studies were published in the field of education or public health (Thoemmes & Kim, 2011).

There are hardly any articles available studying promotion response as done in this thesis. Closely related, however, is the field of program evaluation. Although this focuses on questions about the response to policies and projects, one can argue a campaign or promotion is a specific type of program aimed to change the behaviour of individuals by sending a message. In the literature on program evaluation, there are various studies evaluating labour market and health promotion programs, of which some employ PSMs. For example, Lechner (2002) evaluates different active labour market policies in Zurich. They use propensity score matching to adjust for individual heterogeneity and use a multiple treatment approach to model a range of heterogeneous sub-programs (e.g. training, public employment programs or job counselling). Kluve, Schneider, Uhlendorff, and Zhao (2012) use the generalized propensity score to assess differences in treatment effects due to the duration of training in labour market programmes. Furthermore, Mills, Kessler, Cooper, and Sullivan (2007) look at the impact of health promotion programs in the workplace on health risks and work productivity. Nyman, Abraham, Jeffery, and Barleen (2012) evaluate the impact of another health promotion program, to analyse if the objective to lower the health care expenditures and reduce absenteeism is achieved. These studies evaluate the program over a longer time and use PSMs to ensure comparability between treatment and control groups.

Furthermore, some (marketing) campaigns in the public domain have been evaluated using PSMs. For example, the effects of the National Youth Anti-Drug Media Campaign in the U.S. are well-studied. Lu, Zanutto, Hornik, and Rosenbaum (2001) use it to illustrate multivariate matching with doses of treatment. They model ordinal treatments using McCullagh's ordinal logit model. Zanutto, Lu, and Hornik (2005) use a similar methodology but propose subclassification instead of matching. Yanovitzky, Zanutto, and Hornik (2005) investigate the case of both binary and ordinal treatments. The three studies all found little or no indication of a treatment effect. Moreover, Fong, Hazlett, Imai, et al. (2018) study the effect of political advertisements on campaign contributions using a continuous treatment. Remarkably, despite the parallels between the discussed problem settings and promotion response in marketing, there are no studies we know of that use the PSMs methodology for the latter purpose.

3.2 Applications of generalized treatments

Moreover, the number of studies applying generalized propensity scores is still limited. We focus on the applications related to marketing here. First, we consider studies of the discrete case with a limited number of treatment options. Lechner (2002) models categorical treatments representing labour market policies as a series of pairwise binary treatments. McCaffrey et al. (2013) investigate the effect of different outpatient treatment approaches, which is also categorical. Instead of pairwise comparisons, they use a multinomial logistic regression model with weighting estimators. The studies investigating the effect of the National Youth Anti-Drug Campaign all consider ordinal treatments (Lu et al., 2001; Zanutto et al., 2005; Yanovitzky et al., 2005). Categorical and ordinal treatments are sometimes referred to as multiple treatments in the literature. This is different from multivariate treatments, studied in this thesis. For a complete review of PSMs for categorical and ordinal treatments we refer to Lopez, Gutman, et al. (2017).

Now, the continuous case allows to express the intensity of treatment more precisely. It has a large number of applications in the medical domain (Hirano & Imbens, 2001). Kluve et al. (2012) use this to model the duration of training in labour market programs. Fong et al. (2018) consider political advertisements as continuous treatment variable. Imai and Van Dyk (2004) investigate the effect of smoking using a bivariate treatment variable, where both duration and smoking are continuous. We are not aware of any other bivariate or multivariate treatment applications. For a discussion of generalized treatment regimes, we refer to Imai and Van Dyk (2004).

4 Data

We investigate the tune-in on the season premiere of America's Got Talent (AGT) in 2016. It was aired on NBC on Tuesday May 31st at 08:00 PM EST. In particular, we are interested in the influence of the exposure to advertising for this show on the probability an individual watches this show. The data comes from the Nielsen Company TV panel and contains detailed information on who is watching what and when. In particular we use data from the Nielsen People Meter (NPM), which measures TV viewing on a respondent level. The campaign period investigated is May 11th - 31st. During this period, AGT has been marketed across different types of TV channels: on channel, off channel and a few smaller channels (cross channel, Dish, DirecTV, Local Cable). Here, on channel is defined as NBC, cross channel are different channels owned by NBC (USA Network, E!) and off channel are all other channels (i.a. TLC, Lifetime Television).

For an overview of the spread of the Gross Rating Points (GRP)[3] over time and across marketing channels, see Figure 3. In total the number of GRPs was 77.38, which is quite quite low and means that the reach of this campaign was not very high. Furthermore, the big majority (96%) of the GRPs were delivered on NBC (on channel) and off channel.

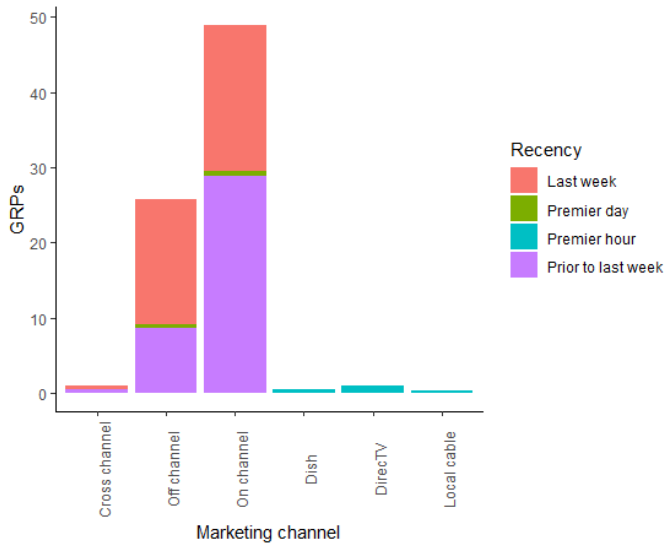


Figure 3: Overview GRPs over time and across different marketing channels.

The analysis in this thesis is focused on the audience 'Adults aged between 18-49', as this was the target audience of the marketing campaign. The raw data contains information on individual characteristics, general TV viewing behaviour, exposure to TV advertisements and tune-in of AGT. We remove 283 individuals from the dataset for who we do not have the tune-in of AGT (response variable). The resulting dataset contains information on $N = 25339$ individuals. Because the panel is not fully representative of the total U.S. population, we use sample weights r_i to correct for this. These weights are determined by Nielsen and indicate how many people in the U.S. a certain individual i represents, such that the resulting sample reflects the total population well in terms of sociodemographics such as age and gender. In the literature of PSMs very little is written on how

[3] A common measure to calculate marketing impact defined as the percentage of the target audience reached multiplied by the average exposure frequency.

to incorporate sample weights. Some researchers argue in favour of a certain approach, but recommendations are conflicting and lack experimental evidence. We follow the robust approach proposed by Ridgeway, Kovalchik, Griffin, and Kabeto (2015), who test various alternatives using simulation and real life data. They propose using sampling weights in all stages of the PSM: as weights in the model to estimate the propensity score and in the final output by multiplying the obtained weight with the sample weight r_i.

The response variable Y is a binary vector of length N indicating whether an individual i has seen at least 1 minute of the season premiere. The decision to model tune-in as binary response variable is motivated by the belief that advertisements influence the decision of an individual to start watching a program, but not the duration of watching. The number of minutes an individual is watching a program is influenced by other factors.

Next, we include a set of variables describing the sociodemographics and general viewing behaviour of individual i. We denote this set of variables by the N x C matrix X. Sociodemographics include age, gender, region, race and income. We measure the viewing behaviour in the two weeks prior to the season premiere (May 17th - 30th). This includes for example the total time watching TV, the time division across channels and the viewing behaviour during different times of the day and week. To capture non-linearities in the relation between explanatory variables and the response variable, we transform continuous variables into categorical variables[4]. We do this by making a separate group for zero values and split the remaining in three groups based on the quantiles. Categorical variables enter the model as dummy variables, where the number of included variables is equal to the number of categories minus one.

Finally, we discuss the definition of the treatment variables T. For the binary treatment case we consider two different definitions of the treatment variable: any exposure to advertising and a total number of exposures larger than a certain threshold. This second definition can be interesting to consider if one believes advertising only has influence after a few exposures. Based on plots displaying the number of exposures versus the level of tune-in, we decide to investigate a threshold of at least three exposures. Although the case of binary treatments is merely a starting point in this thesis for further generalizations, one can also see this as an alternative way of modelling ordinal/continuous treatments (Lopez, Gutman, et al., 2017). For the continuous treatment case we define the treatment variable as the number of exposures.

For the definition of multivariate treatment variables, it is important to note that there are only two major marketing channels (on and off channel). This makes it impossible to analyse the effect of all individual marketing channels separately. But, the effect of recency of advertisements is also interesting to analyse. Hence, we can alternatively create treatment variables based on the time until the season premiere (i.e. prior to last week, last week and premiere day[5]). To sum up, we consider two sets of multivariate treatment variables: one based on the time of advertisements and one based on the channel the advertisement was broadcast on. For the definition of the variable(s) of interest for the case of binary, continuous and multivariate treatments we refer to Table 1.

[4]By including different levels of a continuous explanatory variable separately, we make a simple approximation for different types of non-linear effects.

[5]We join the category premier hour with premier day, because the separate groups are too small.

Table 1: Definition treatment parameters for binary, continuous and multivariate treatments.

	Binary treatments	Continuous treatments	Multivariate treatments
Dimensions	Vector of length N, $T_i \in \{0, 1\}$.	Vector of length N, $T_i \in [0, \mathrm{T}_{i,max}]$.	Matrix of size N x M, $T_{mi} \in [0, \mathrm{T}_{mi,max}]$.
Definition	(a) Indicator if there has been any exposure to the media campaign (number of exposures > 0). (b) Indicator if the total number of exposures to the media campaign exceeds a certain threshold (number of exposures is > 2).	Count variable equal to the total number of exposures to the media campaign.	Count variables equal to the number of exposures to the media campaign. (a) Based on time: prior to last week, last week and on premier day ($M = 3$). (b) Based on channel: on channel, other channels ($M = 2$).

A table with a complete overview and description of the variables can be found in Appendix 9.1. In that table, it is also indicated which variables influence the outcome and treatment variables respectively. The choice of variables, as well as the type of variables included, is based on domain knowledge. As variable selection is not the focus of this thesis, the set of variables presented there will be the final set of variables used for the response model and the initial set of variables used for the treatment assignment model. However, we do check for multicollinearity using the variance inflation factor (VIF) and exclude four dummy variables that have a too high correlation (> 0.7) with other variables in the model.

In the obtained dataset, 6.5% of all individuals watched (part of) the show. However, of all individuals in the dataset there are many individuals who never watched TV (8.8%) or NBC (54.1%) during the two week period investigated. Among frequent NBC viewers AGT is more popular, with 27.2% watching (part of) the show. We define frequent NBC viewers here as the top two terciles constructed using the viewing minutes of NBC primetime from May 2nd - 16th[6]. To get insight in the relation between the various treatment variables and the level of tune-in, we made some exploratory data analysis plots (see Appendix 9.2). We indeed find increasing levels of tune-in for higher numbers of exposures. Further analysis using PSMs will verify the strength of the causal relation between these two.

[6]The terciles are constructed as follows: 1) multiply the viewing minutes and sample weights r_i, 2) sort the individuals i based on their viewing minutes, 3) split the dataset in three equal groups based on the cumulative sum of step 1.

5 Methodology

PSMs, proposed by Rosenbaum and Rubin (1983), aim to reduce bias introduced by confounders due to lack of randomization in observational data. An overview of the steps taken when doing causal inference with PSMs is shown in Figure 4[7]. These steps are the same for binary, continuous and multivariate treatments although the details of each step may differ. We first discuss the steps for binary treatments in Section 5.1, then continuous treatments in Section 5.2 and finally multivariate treatments in Section 5.3.

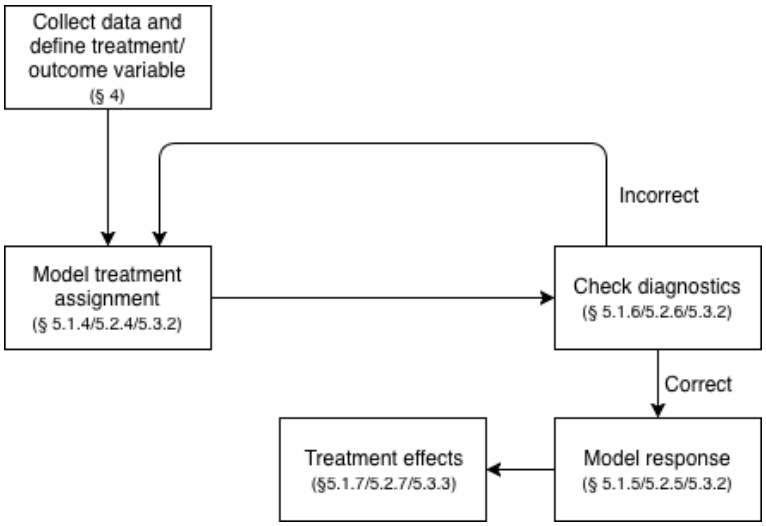


Figure 4: Flow of PSMs.

5.1 Binary treatments

5.1.1 Notation

First assume we have a binary treatment variable $T_i \in \{0, 1\}$, where $T_i = 1$ if an individual i has been exposed to any form of advertising or the number of exposures exceed a certain threshold. Each individual i has two potential outcomes:

$$Y_{0i}^* = X_i\beta_0 + \epsilon_{0i}, \qquad \text{for } T_i = 0, \tag{1}$$

$$Y_{1i}^* = X_i\beta_1 + \epsilon_{1i}, \qquad \text{for } T_i = 1, \tag{2}$$

where X_i is a 1 x C vector containing personal characteristics and TV viewing behaviour, β_t is a C x 1 vector of unknown parameters and ϵ_{ti} is the unexplained difference for individual i in $t \in \{0, 1\}$. These equations describe the relationship between the variables of interest in the possible states treatment and non-treatment separately.

Furthermore, we define the treatment variable T_i as follows:

$$T_i = \begin{cases} 1, & \text{if } T_i^* > 0, \\ 0, & \text{if } T_i^* \leq 0. \end{cases} \tag{3}$$

where T_i^* is the selection equation that describes the assignment of treatment. This is also dependent on personal characteristics and TV viewing behaviour Z_i, but this set is

[7]The relevant subsections for each step are indicated in parentheses.

possibly different from X_i:

$$T_i^* = Z_i\theta + \epsilon_{Ti}.$$

The observed outcome Y_i is typically defined as follows:

$$Y_i = \begin{cases} Y_{0i}^*, & \text{if } T_i = 0, \\ Y_{1i}^*, & \text{if } T_i = 1. \end{cases} \tag{4}$$

This model can be thought of as Rubin's potential outcome model or the switching regression model. Note, Y_0 and Y_1 are both partially observed, T^* is latent and T is known. However, in a marketing setting we often encounter limited-dependent variables. In this thesis we have a binary outcome variable Y, hence the observed outcome defined in Equation (4) is replaced by[8]:

$$Y_i = \begin{cases} Y_{0i} = 1, & \text{if } Y_{0i}^* > 0, \\ Y_{1i} = 1, & \text{if } Y_{1i}^* > 0, \\ Y_{0i} = 0, & \text{if } Y_{0i}^* \leq 0, \\ Y_{1i} = 0, & \text{if } Y_{1i}^* \leq 0. \end{cases} \tag{5}$$

We assume the observed data (Y_i, X_i, Z_i T_i) for $i = 1, ..., N$ is an i.i.d. sample from the target population of interest. We assume there is a random selection of individuals, that is representative given the sample weights r_i. However, we do not expect a random allocation of treatments to individuals.

5.1.2 Causal quantities of interest

Intuitively, the treatment effect is the difference for each individual i between the outcome under treatment ($T = 1$) and non-treatment ($T = 0$). Hence, the individual gain from treatment is $Y_{1i} - Y_{0i}$. However, we cannot observe both Y_{0i} and Y_{1i}. Therefore, the individual treatment effect is not identified. As an alternative, we look at different treatment parameters. Two of the most common treatment parameters are the average treatment effect (ATE):

$$\tau_{ATE} = E[Y_1 - Y_0] = E[Y_1] - E[Y_0],$$

and the average treatment effect on the treated (ATT):

$$\tau_{ATT} = E[Y_1 - Y_0|T = 1] = E[Y_1|T = 1] - E[Y_0|T = 1].$$

It is interesting to look at both, because a simple comparison of individuals with and without treatment (as measured in ATE) might give a misleading estimate of the treatment effect. The ATT computes the effect on individuals who are exposed to treatment, whereas the ATE estimates the effect of exposure on a randomly selected individual. In practice, we can only observe the observed average treatment effect (OTE):

$$\tau_{OTE} = E[Y|T = 1] - E[Y|T = 0] = E[Y_1|T = 1] - E[Y_0|T = 0].$$

The sample analogues of the ATE and ATT respectively are:

$$\tau_{ATE} = \frac{1}{N}\left[\sum_{i=1}^{N_1}(Y_{1i} - Y_{0i}|T_i = 1) + \sum_{i=1}^{N_0}(Y_{1i} - Y_{0i}|T_i = 0)\right],$$

$$\tau_{ATT} = \frac{1}{N_1}\sum_{i=1}^{N_1}(Y_{1i} - Y_{0i}|T_i = 1),$$

[8]Note we either observe Y_{0i}^* (for $T_i = 0$) or Y_{0i}^* (for $T_i = 1$), but never both.

where N_1 and N_0 are the number of individuals with $T = 1$ and $T = 0$ respectively ($N = N_1 + N_0$). As the counterfactuals $Y_0|T = 1$ and $Y_1|T = 0$ cannot be observed, the treatment parameters are not identified without further assumptions. We know that:

- $\tau_{ATE} = \tau_{OTE}$ if $E[Y_1|T = 1] = E[Y_1|T = 0]$ and $E[Y_0|T = 0] = E[Y_0|T = 1]$,
- $\tau_{ATT} = \tau_{OTE}$ if $E[Y_0|T = 0] = E[Y_0|T = 1]$.

Hence, for the ATT to be identified we only need to assume there is no difference in the average outcome between treatment and non-treatment group if they are not treated. For the ATE we need the extra assumption that there is also no difference in the average outcome between treatment and non-treatment group if they are treated. However, these assumptions are unlikely to hold for observational studies due to a lack of randomization that is usually present in experimental settings. Hence, we need to estimate the counterfactuals given certain identifying assumptions we are willing to make. We know the ATT is equal to the ATE if there is no selection bias.

5.1.3 Identification

PSMs estimate the counterfactuals and correct the bias in treatment effects by comparing outcomes between individuals who are as similar as possible. To establish this we use the propensity score π_i, defined as the probability of receiving treatment given observed covariates:

$$\pi_i = Pr(T_i = 1|Z_i) = E[T_i|Z_i]. \tag{6}$$

Hereby, we make the following assumptions:

1. Conditional independence (unconfoundedness): $((Y_{0i}, Y_{1i})) \perp T_i|Z_i$. In words, the potential outcomes are independent of treatment assignment conditional on the observed covariates. This states there is no hidden bias due to unmeasured confounders. If this assumption holds, then also $((Y_{0i}, Y_{1i})) \perp T_i|\pi_i$. Furthermore, exposure to treatment is random for a given propensity score $T_i \perp Z_i|\pi_i$.

2. Common support (overlap): $0 < Pr(T_i = 1|Z_i) < 1$. In words, this assumption states there should be a positive probability of treatment and non-treatment given the observed covariates for each individual.

3. Stable unit treatment value assumption (SUTVA): we assume an individual is not affected by the treatment other individuals receive and that treatments do not differ between individuals.

Rosenbaum and Rubin (1983) show that if these assumptions hold (i.e. treatment assignment is ignorable), one can obtain unbiased estimates of the treatment parameters. PSMs can only reduce selection bias due to observed confounding variables. It corrects unbalance in covariates and bias due to lack of overlap between treatment and non-treatment group (Heckman, Ichimura, Smith, & Todd, 1998). In practice, propensity scores are often unknown and must be estimated. However, it turns out using estimates of the propensity score can reduce the introduced efficiency loss (Rosenbaum, 1987; Rubin & Thomas, 1996; Hirano, Imbens, & Ridder, 2003). An intuitive explanation for this is that the true propensity score only adjusts for the systematic differences, whereas the estimated propensity score can correct for random sample differences too.

Given unconfoundedness, ATE and ATT are identified as this implies $E[Y_t|Z] = E[Y_t|T = t, Z] = E[Y|T = t, Z]$ for $t \in \{0, 1\}$. Although the assumption of unconfoundedness is strong, this assumption also underlies conventional regression analysis (i.e. exogeneity). Using PSMs, however, has the advantage that treatment effects can be estimated with smaller models (less parameters) and linearity assumptions do not need to be made.

5.1.4 Treatment assignment model

To estimate the propensity score π_i, we model the observed treatment T_i given the covariates Z_i. When doing this, we have to make two important choices: (1) which variables to include in the model and (2) which model is used for estimation.

Regarding variable selection there are three things to keep in mind (Caliendo & Kopeinig, 2008). First, the selected variables should make it credible that the assumption of unconfoundedness holds. Second, only variables influencing both the treatment assignment and the outcome should be included. Third, variables should be measured before treatment assignment. In practice, there are many possible sets of variables that can be included in the propensity score model. One can use economic theory, previous empirical findings or formal (statistical) tests to make the decision. We base the initial choice of variables on domain knowledge. As initial set of variables we use the outcome regressors that were measured before treatment assignment. However, in practice searching for an appropriate treatment assignment model specification is a repeated process of checking covariate balance (for definition, see Section 5.1.6) and adjusting the model. To improve the model, we consider adding/removing variables and including higher order covariates or interactions for covariates that are out of balance (Zanutto et al., 2005).

There is also a lot of choice in models to use for estimation. We investigate the following models to estimate the binary propensity score and compare their results[9]:

1. *Logistic regression*
   The most common way to estimate propensity scores in the binary setting is using a logistic regression. Hence, this is the first estimation method examined in this thesis:

   $$\text{logit}(\pi_i) = \ln(\frac{\pi_i}{1-\pi_i}) = \ln(\frac{Pr(T_i = 1|Z_i)}{1 - Pr(T_i = 1|Z_i)}) = Z_i\theta.$$

   Logistic regression is part of a larger family of generalized linear models. We use the `svyglm package` in R, which is especially designed for sampling weights and obtain $\hat{\theta}$ by iteratively re-weighted least squares (Lumley, 2019). We then use this result to compute:

   $$\hat{\pi}_i = \Pr(T_i = 1|Z_i) = \frac{\exp(Z_i\hat{\theta})}{1 + \exp(Z_i\hat{\theta})} = \frac{1}{1 + \exp(Z_i\hat{\theta})^{-1}}.$$

2. *Boosting*
   In practice, any standard probability model can be used to estimate propensity scores. Other popular alternatives in the literature are non-parametric methods such as generalized boosted models (GBM) (McCaffrey, Ridgeway, & Morral, 2004). A big advantage of these methods is that there is no need to specify possible interactions or polynomial terms among covariates as for parametric methods. This is often

[9] We include the sample weights r_i as weights in each of the treatment assignment models.

problematic as no standard procedures exist to do this. Instead, nonlinearities are automatically captured by GBM as they rely on many simple regression trees. Moreover, this method can work well with a large number of predictors. The performance of machine learning techniques to estimate propensity scores has been investigated by B. K. Lee, Lessler, and Stuart (2010). They find that boosted CART is most promising in comparison to other methods.

We implement this using the `gbm package` in R (Greenwell, Boehmke, & Cunningham, 2019). More specifically, we follow the recommendations of McCaffrey et al. (2004) for the tuning parameters. They suggest a maximum tree complexity of 4 (interaction.depth), a small shrinkage parameter or learning rate (0.0005) to ensure a smooth fit and subsampling 50% of the data in each iteration (bag.fraction). We use a binomial distribution. Furthermore, we choose a sufficient initial number of trees (n.trees = 3000) and optimize over the resulting GBM to find the number of trees for which the total covariate imbalance is lowest. We define total covariate imbalance as the average absolute standardized mean difference of all covariates (see Section 5.1.6).

3. *CBPS*
Finally, several improved methods have been suggested to automate the repeated process of trying different models and checking for covariate balance (Imai & Ratkovic, 2014; Hainmueller, 2012; Graham, de Xavier Pinto, & Egel, 2012; Tan, 2010). We investigate the covariate balancing propensity score (CBPS) of Imai and Ratkovic (2014) as it also extended to continuous treatments by Fong et al. (2018). This method directly optimizes the resulting balance in covariates instead of focusing on the accuracy of predicting treatment assignment. Thereby, it increases the robustness to misspecification of the propensity score model (Fong et al., 2018). It does so using generalized method of moments (GMM) estimation with moment conditions based on the covariate balance from maximum likelihood.

We follow the over-identified case presented by Imai and Ratkovic (2014), combining the covariate balancing conditions with score conditions. The covariate balancing conditions are operationalized using inverse propensity score weighting. These moment conditions, for ATE and ATT respectively, are:

$$E\left[\frac{T_i\tilde{Z}_i}{\pi_i} - \frac{(1-T_i)\tilde{Z}_i}{1-\pi_i}\right] = 0,$$
$$E\left[T_i\tilde{Z}_i - \frac{\pi_i(1-T_i)\tilde{Z}_i}{1-\pi_i}\right] = 0,$$

where $\tilde{Z}_i = f(Z_i)$ is a function specified by the researcher. Furthermore, the score condition (obtained from maximum likelihood of the logistic regression model[10]) is:

$$E\left[\frac{T_i\pi_i'}{\pi_i} - \frac{(1-T_i)\pi_i'}{1-\pi_i}\right] = 0.$$

So we use $f(Z_i) = \pi_i' = \frac{\partial \pi_i}{\partial \theta}$, thereby placing greater emphasis on observations with more predictive power (Imai & Ratkovic, 2014). Possible alternative specifications

[10] $\hat{\theta} = \arg\max_\theta \sum_{i=1}^{N} T_i \log(\pi_i) + (1-T_i)\log(1-\pi_i)$, where $\pi_i = \frac{\exp(Z_i\theta)}{1+\exp(Z_i\theta)}$.

of $f(Z_i)$ are proposed by Fan, Imai, Liu, Ning, and Yang (2016), but will not be investigated in this thesis.

The sample mean of the moment conditions is thus:

$$g(T,Z) = \frac{1}{N}\sum_{i=1}^{N} g(T_i, Z_i), \text{where } g(T_i, Z_i) = \begin{pmatrix} s(T_i, Z_i) \\ w(T_i, Z_i) \end{pmatrix},$$
$$s(T_i, Z_i) = \frac{T_i \pi_i'}{\pi_i} - \frac{(1-T_i)\pi_i'}{1-\pi_i},$$
$$w(T_i, Z_i) = \frac{T_i \tilde{Z}_i}{\pi_i} - \frac{(1-T_i)\tilde{Z}_i}{1-\pi_i} \text{ for ATE and}$$
$$w(T_i, Z_i) = \frac{N}{N_1}\left(T_i \tilde{Z}_i - \frac{\pi_i(1-T_i)\tilde{Z}_i}{1-\pi_i}\right) \text{ for ATT.}$$

Hence, we can estimate $\hat{\theta}$ as follows:

$$\hat{\theta} = \min\, g(T,Z)^T \Sigma(T,Z)^{-1} g(T,Z),$$

where $g(T,Z)$ is as defined above and $\Sigma(T,Z)^{-1}$ is computed using two step feasible GMM. We implement this using the `CBPS package` in R and compute $\hat{\pi}$ as (Fong, Ratkovic, Imai, & Hazlett, 2019):

$$\hat{\pi}_i = \frac{1}{1+\exp(Z_i\hat{\theta})^{-1}}.$$

5.1.5 Response model

There are also many methods available to model the response given the treatment. The most common techniques are: matching (Rosenbaum & Rubin, 1985), subclassification or stratification (Rosenbaum & Rubin, 1984), weighting (Rosenbaum, 1987; Hirano et al., 2003), regression or combinations of these. We focus on weighting methods for binary treatments. Firstly, because weighting includes all individuals in the analysis, which is not the case for matching. Secondly, it has the advantage that it can be combined with any statistical technique that accepts weights. The idea is to give weights to individuals in the observed sample in such a way that the groups become similar. We investigate inverse probability of treatment weighting methods, originally proposed by Horvitz and Thompson (1952). We now discuss (1) how to define the weights, (2) which outcome model is used for estimation and (3) which variables to include in the model.

The inverse probability of treatment weight is defined as follows for ATE and ATT:

$$w_{ATE,i} = \frac{T_i}{\hat{\pi}_i} + \frac{1-T_i}{1-\hat{\pi}_i} = \frac{T_i - \hat{\pi}_i}{\hat{\pi}_i(1-\hat{\pi}_i)}, \tag{7}$$
$$w_{ATT,i} = T_i + (1-T_i)\frac{\hat{\pi}_i}{1-\hat{\pi}_i} = \frac{T_i - \hat{\pi}_i}{1-\hat{\pi}_i}. \tag{8}$$

In words, we assign the weight $\frac{1}{\hat{\pi}_i}$ for those who receive treatment and $\frac{1}{1-\hat{\pi}_i}$ for those who do not receive treatment for ATE. Similarly, we assign the weights 1 and $\frac{\hat{\pi}_i}{1-\hat{\pi}_i}$ for ATT. Hence, individuals who are in the treatment group and are very likely to receive treatment get less weight than individuals who have a lower probability of receiving treatment. This

way the propensity score corrects for over- and under-sampled individuals in the treatment and control group.

Next, we multiply the obtained weight with the sample weight r_i to ensure representativeness of the sample (Ridgeway et al., 2015). Finally, we normalize $w_{ATE,i}$ and $w_{ATT,i}$ so that the average weight in the treatment and control group equals one (Hirano & Imbens, 2001). This is comparable to the situation without using balancing weights and results in the following weights:

$$w_{TE,i} = T_i \frac{r_i \cdot w_{TE,i} \cdot N_1}{\sum_{i,T_i=1}^{N} r_i \cdot w_{TE,i}} + (1 - T_i) \frac{r_i \cdot w_{TE,i} \cdot N_0}{\sum_{i,T_i=0}^{N} r_i \cdot w_{TE,i}} \text{ for } TE \in \{ATE, ATT\}, \quad (9)$$

where N_1 and N_0 are the number of individuals with $T = 1$ and $T = 0$ respectively ($N = N_1 + N_0$).

An appropriate outcome model should be selected based on the nature of the dependent variable (Austin, 2018). As we have a binary dependent variable Y, we use a logit model to model the response in this thesis:

$$Y_i = \Lambda(X_i\beta + T_i\gamma), \quad (10)$$

where $\Lambda(.)$ is the logistic distribution function, Y_i is the observed outcome, X_i is a 1 x C vector containing personal characteristics and TV viewing behaviour, β is a C x 1 vector of unknown parameters, γ is the effect of interest, ϵ_i is the unexplained difference for individual i. However, we also explore two estimation methods with a more general model specification (see IPW and AIPW in Section 5.1.7). Furthermore, we base the decision of which variables to include in the response model on domain knowledge (see Section 4). We include all variables in a linear form.

5.1.6 Diagnostics

Now we describe various model diagnostics to assess the performance of the treatment assignment model. This can be used to determine the quality of the final treatment assignment model and to determine if the model needs to be improved.

Propensity scores
First of all, we check the size of the propensity scores. Very small or large values of the propensity score lead to very large or small weights. Larger weights can lead to a larger standard error as one individual in the sample can affect the parameter estimates a lot. Propensity score values close to zero or one can cause problems for weighting methods. These values are often trimmed or truncated based on (low) quantiles of the distribution of the propensity scores (Austin & Stuart, 2015). We investigate the histogram of propensity scores and truncate these extreme values to reduce the variance of the estimates. As a threshold we choose to use the 2.5% and 97.5% percentiles of the propensity score.

Covariate balance
Next, we check if the treatment assignment model is adequately specified. There are two common ways to assess if the propensity score model performs well: evaluating the prediction of treatment assignment (e.g hit or miss, leave-one-out cross-validation) or assessing the covariate balance between the control and treatment group. As the ultimate goal of the treatment assignment model is to balance the covariates, we will focus on the

latter.

Comparing the similarity of the control and treatment group starts with a comparison of the means. The standardized mean difference (SMD) is a good measure to assess the balance in the marginal distributions of covariates (Caliendo & Kopeinig, 2008). For continuous variables this is defined as (Rosenbaum & Rubin, 1985):

$$SMD_j = \frac{\bar{X}_{1j} - \bar{X}_{0j}}{\sqrt{\frac{s_{1j}^2 + s_{0j}^2}{2}}},$$

where $\bar{X}_{1j}$ and $\bar{X}_{0j}$ are the weighted sample mean and s_{1j} and s_{0j} the weighted sample variance of continuous covariate j in the treatment ($T = 1$) and control ($T = 0$) group respectively.

Similarly, for binary variables (Austin, 2011):

$$SMD_j = \frac{\bar{X}_{1j} - \bar{X}_{0j}}{\sqrt{\frac{\bar{X}_{1j}(1-\bar{X}_{1j}) + \bar{X}_{0j}(1-\bar{X}_{0j})}{2}}},$$

where $\bar{X}_{1j}$ and $\bar{X}_{0j}$ are the weighted sample mean (or prevalance) of covariate j in the treatment ($T = 1$) and control ($T = 0$) group respectively. Categorical variables can be transformed to a set of binary variables to calculate the SMD.

We weigh the observations using the weights as defined in Equation (9) (Austin & Stuart, 2015). Hence, we present separate balance diagnostics for ATE and ATT. We compare this with the initial imbalance, which we obtain by using only the sample weight r_i. As the SMD is standardized, we can compare variables of different scales. Differences of less than 0.1 are considered negligible (Austin, 2011), hence covariate j is balanced if $|SMD_j| < 0.1$. Furthermore, we define total covariate imbalance as $\frac{1}{C}\sum_{j=1}^{C} |SMD_j|$.

Common support
A final step before estimating the treatment parameters is to check the region of common support. We do so by investigating the density distribution of the control and treatment group and by comparing the range (i.e. minima, maxima) of propensity scores for both groups (Caliendo & Kopeinig, 2008). If there is no overlap in distributions, there is no common support. If there are relatively few individuals in a certain part of the interval, this is called thin common support. Propensity score values close to zero or one are near violations of the common support assumption. There are several ways to deal with common support problems. The simplest would be to exclude observations that fall outside the common interval, however this can lead to a large data reduction. Furthermore, the estimated effect is not valid anymore for the entire population, but only for the remaining sub-sample. For other possible solutions we refer to the overview given by Lechner and Strittmatter (2014). In this thesis, we investigate the degree to which common support problems are present and take this into account when interpreting the results.

5.1.7 Treatment effects

As for the treatment assignment model, there are also many different methods to estimate the treatment effects. To estimate the effects of interest, ATE and ATT, we choose to

examine inverse propensity weighting (IPW), weighted least squares (WLS) and the augmented IPW (AIPW). Each of these methods will now be discussed in more detail.

Inverse propensity weighting
First, consider a well-known estimator that simply reweights the observations to make them representative of the population of interest. This gives an unbiased estimate of the treatment effect as:

$$\begin{aligned} E\left[\frac{TY}{\hat{\pi}}\right] &= E\left[E\left[\frac{TY}{\hat{\pi}}\middle|Z\right]\right] = E\left[E\left[\frac{TY_1}{\hat{\pi}}\middle|Z\right]\right] \text{ (using } Y = TY_1 + (1-T)Y_0) \\ &= E\left[\frac{E[T|Z]E[Y_1|Z]}{\hat{\pi}}\right] \text{ (by unconfoundedness)} \\ &= E[E[Y_1|Z]] = E[Y_1]. \end{aligned}$$

Similarly, $E[\frac{(1-T)Y}{1-\hat{\pi}}] = E[Y_0]$.

Hirano and Imbens (2001) propose to normalize the inverse probability weights of Horvitz and Thompson (1952) and use this to estimate the treatment effect directly. Incorporating the sample weights r_i, we obtain the following expressions to estimate ATE and ATT:

$$\hat{\tau}_{ATE,IPW} = \sum_{i=1}^{N} \frac{r_i T_i Y_i}{\hat{\pi}_i}\left(\sum_{i=1}^{N} \frac{r_i T_i}{\hat{\pi}_i}\right)^{-1} - \sum_{i=1}^{N} \frac{r_i (1-T_i) Y_i}{1-\hat{\pi}_i}\left(\sum_{i=1}^{N} \frac{r_i (1-T_i)}{1-\hat{\pi}_i}\right)^{-1},$$

$$\hat{\tau}_{ATT,IPW} = \sum_{i=1}^{N} r_i T_i Y_i\left(\sum_{i=1}^{N} r_i T_i\right)^{-1} - \sum_{i=1}^{N} r_i (1-T_i) Y_i \frac{\hat{\pi}_i}{1-\hat{\pi}_i}\left(\sum_{i=1}^{N} \frac{r_i (1-T_i)\hat{\pi}_i}{1-\hat{\pi}_i}\right)^{-1}.$$

Using the weights $w_{ATE,i}$ and $w_{ATT,i}$ derived in Section 5.1.5, we can alternatively write these estimators as:

$$\hat{\tau}_{ATE,IPW} = \frac{1}{N_1}\sum_{i=1}^{N} w_{ATE,i} T_i Y_i - \frac{1}{N_0}\sum_{i=1}^{N} w_{ATE,i}(1-T_i) Y_i,$$

$$\hat{\tau}_{ATT,IPW} = \frac{1}{N_1}\sum_{i=1}^{N} w_{ATT,i} T_i Y_i - \frac{1}{N_0}\sum_{i=1}^{N} w_{ATT,i}(1-T_i) Y_i.$$

Weighted least squares
Here, we combine weighting and regression adjustment by weighted coefficients. The idea is to use weights to put more importance on certain observations based on the propensity scores. If we only include T as covariate in a (linear) weighted regression (regressing Y on T), this is the same as the IPW estimator. However, WLS gives the possibility to include additional covariates in the model to adjust for.

We thus specify a model for Y. As a result, this approach gives consistent estimates if either the treatment assignment model or the response model is correctly specified (Kang, Schafer, et al., 2007). In the literature, this is commonly referred to as double robustness. We add the propensity score and outcome variables that are out-of-balance after applying the treatment assignment model as covariates to the response model. We estimate a logit model using weights $\hat{w}_{TE}$ for $\hat{\tau}_{TE,WLS}$ to find parameter estimates $\hat{\beta}_{TE}, \hat{\gamma}_{TE}$ for $TE \in \{ATE, ATT\}$. We then estimate the marginal effects as follows:

$$\hat{\tau}_{TE,WLS} = \frac{\partial Y}{\partial T} = \hat{\gamma}_{TE} \cdot \lambda(X\hat{\beta}_{TE} + T\hat{\gamma}_{TE}) \approx \hat{\gamma}_{TE} \cdot \frac{1}{N}\sum_{i=1}^{N} \lambda(X_i\hat{\beta}_{TE} + T_i\hat{\gamma}_{TE}),$$

where $\lambda(.)$ is the probability density function for the logistic distribution.

Augmented inverse propensity weighting
Another robust estimator is AIPW (Robins, Rotnitzky, & Zhao, 1994). Here, the original IPW estimator is augmented using two regression estimators. The AIPW estimator is also double robust (Scharfstein, Rotnitzky, & Robins, 1999). However, contrary to WLS, the outcome regression model is only used for prediction. This allows for more flexible modelling. The original AIPW estimator is defined as:

$$\hat{\tau}_{ATE,AIPW} = \frac{1}{N}\sum_{i=1}^{N}\frac{T_i Y_i}{\hat{\pi}_i} - \frac{T_i - \hat{\pi}_i}{\hat{\pi}_i}\hat{\mu}(1, X_i) - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-T_i)Y_i}{1-\hat{\pi}_i} + \frac{T_i - \hat{\pi}_i}{1-\hat{\pi}_i}\hat{\mu}(0, X_i) =$$
$$\frac{1}{N}\sum_{i=1}^{N}\hat{\mu}(1, X_i) + \frac{T_i(Y_i - \hat{\mu}(1, X_i))}{\hat{\pi}_i} - \frac{1}{N}\sum_{i=1}^{N}\hat{\mu}(0, X_i) + \frac{(1-T_i)(Y_i - \hat{\mu}(0, X_i))}{1-\hat{\pi}_i}.$$

We adjust this to incorporate the sample weights r_i and estimate:

$$\hat{\tau}_{TE,AIPW} = \frac{1}{N}\sum_{i=1}^{N}\hat{\mu}(1, X_i) + T_i w_{TE,i}(Y_i - \hat{\mu}(1, X_i)) -$$
$$\frac{1}{N}\sum_{i=1}^{N}\hat{\mu}(0, X_i) + (1-T_i) w_{TE,i}(Y_i - \hat{\mu}(0, X_i)),$$

where $\hat{\mu}(t, X_i) = E[Y_i|T_i = t, X_i] = X_i\hat{\beta}$ is the regression of the outcome on the covariates in group $t \in \{0, 1\}$ for $TE \in \{ATE, ATT\}$. The two response models are again estimated using a logit model. We do not explore more flexible modelling in this thesis, but refer to Glynn and Quinn (2010) for an application using generalized additive models (GAM).

5.1.8 Standard errors

As we need to incorporate the uncertainty resulting from the estimated propensity scores, naive standard errors are not valid. A common and general way to deal with this problem is to use bootstrapping to estimate the distribution of the treatment parameter (Caliendo & Kopeinig, 2008; Austin, 2016). However, for some cases we can also derive robust approximate sampling variances. As these analytical expressions do not include sample weights, we use a bootstrap variance estimator here. We verify the accuracy of the bootstrap variance estimator for the case without sample weights using the empirical sandwich estimator.

Bootstrap variance estimator
To incorporate uncertainty over the choice of weights, one can estimate the standard errors using bootstrapping where the results are re-estimated B times. The B obtained treatment effect parameters approximate the sampling distribution and can be used to approximate the standard error. In particular, we repeat the following procedure B times:

1. Resample the original data with replacement to obtain a bootstrap sample b of size N.

2. For each bootstrap sample b estimate the treatment effects $\hat{\tau}_{ATE}$ and $\hat{\tau}_{ATT}$ as outlined in Section 5.1.4-5.1.7.

Next, compute the standard errors as follows:

$$SE(\hat{\tau}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\tau}_b - \bar{\tau})^2},$$

where $\bar{\tau} = \frac{1}{B}\sum_{b=1}^{B}\hat{\tau}_b$. We use $B = 25$ iterations for computational feasibility, but advice to use more iterations for better accuracy.

Empirical sandwich estimator
Assuming π_i is correctly specified, the approximate sampling variances for the above treatment estimators can also be approximated via the empirical sandwich method. We compute the robust standard errors for the ATE using IPW and AIPW with logistic regression as treatment assignment model. For this, we look at the weighted estimators as a set of estimating equations and apply the theory of M-estimation (Lunceford & Davidian, 2004). Lunceford and Davidian (2004) present the following approximate sampling variances for propensity score models of the form $\pi_i = \Pr(T_i = 1|Z_i) = (1 + \exp(Z_i\theta)^{-1})^{-1}$:

$$V(\hat{\tau}_{ATE,IPW}) = \frac{1}{N^2}\sum_{i=1}^{N}\hat{I}^2_{IPW,i},$$

$$\hat{I}_{IPW,i} = \frac{T_i(Y_i - \hat{\mu}_{1,IPW})}{\hat{\pi}_i} - \frac{(1-T_i)(Y_i - \hat{\mu}_{0,IPW})}{1-\hat{\pi}_i} - (T_i - \hat{\pi}_i)\hat{H}_\theta^T \hat{E}_{\theta,\theta}^{-1} Z_i^T,$$

$$\hat{H}_\theta = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{T_i(Y_i - \hat{\mu}_{1,IPW})(1-\hat{\pi}_i)}{\hat{\pi}_i} - \frac{(1-T_i)(Y_i - \hat{\mu}_{0,IPW})\hat{\pi}_i}{1-\hat{\pi}_i}\right)Z_i^T,$$

$$\hat{E}_{\theta,\theta} = \frac{1}{N}\sum_{i=1}^{N}\hat{\pi}_i(1-\hat{\pi}_i)Z_i^T Z_i,$$

$$\hat{\mu}_{1,IPW} = \sum_{i=1}^{N}\frac{T_i Y_i}{\hat{\pi}_i}\left(\sum_{i=1}^{N}\frac{T_i}{\hat{\pi}_i}\right)^{-1} \text{ and } \hat{\mu}_{0,IPW} = \sum_{i=1}^{N}\frac{(1-T_i)Y_i}{1-\hat{\pi}_i}\left(\sum_{i=1}^{N}\frac{1-T_i}{1-\hat{\pi}_i}\right)^{-1}.$$

$$V(\hat{\tau}_{ATE,AIPW}) = \frac{1}{N^2}\sum_{i=1}^{N}\hat{I}^2_{AIPW,i},$$

$$\hat{I}_{AIPW,i} = \frac{T_i Y_i}{\hat{\pi}_i} - \frac{T_i - \hat{\pi}_i}{\hat{\pi}_i}\hat{\mu}(1, X_i) - \frac{(1-T_i)Y_i}{1-\hat{\pi}_i} + \frac{T_i - \hat{\pi}_i}{1-\hat{\pi}_i}\hat{\mu}(0, X_i) - \hat{\tau}_{ATE,AIPW}.$$

Here, $\hat{\mu}(t, X_i) = E[Y_i|T_i = t, X_i] = X_i\hat{\beta}$ is the regression of the outcome on the covariates in group $t \in \{0, 1\}$.

5.1.9 Sensitivity analysis

Finally, we want to assess the assumption of unconfoundedness (no unmeasured confounders). As hidden bias cannot be estimated, Rosenbaum (2002) recommends testing this by measuring the sensitivity of the results to potential unmeasured confounders. Hidden bias is present if individuals with the same baseline covariates have different odds of treatment. Although Rosenbaum's approach is developed for matching, it can be used to give an indication of the presence of hidden bias in our analysis.

In short, the idea is to assume there is an unobserved covariate u_i and analyse if we obtain the same results when the odds increase. We assume a logit model $\pi_i = \Lambda(Z_i\theta + u_i\eta)$ (where $\Lambda(.)$ is the logistic distribution function) and evaluate the results for different values of η. Rosenbaum (2002) proved this gives the following bounds on the odds ratio:

$$\frac{1}{\Gamma} \leq \frac{\pi_j/(1-\pi_j)}{\pi_k/(1-\pi_k)} \leq \Gamma,$$

where $\Gamma = \exp(\eta)$, π is the probability of treatment and $\pi/(1-\pi)$ the odds of receiving treatment for individuals j and k ($j \neq k$).

Now, suppose we know the results obtained remain unchanged for Γ^*. Then, we can interpret Γ^* as the factor by which an unobserved confounder should change the odds before it impacts the estimated results. Thus the higher Γ^*, the more resistant is the treatment effect to hidden bias. In social sciences, Γ^* is usually around 1.2. We examine Γ with an upper bound of 6 and increments of 0.1. We construct a simple matching and use the `rbounds package` in R for the sensitivity analysis (Keele, 2015).

5.2 Continuous treatments

5.2.1 Notation

Now we assume we have a continuous treatment variable $T_i \in \zeta$, where $\zeta = [0, T_{i,max}]$ as treatment exposures are non-negative. Strictly speaking T_i is the number of exposures to any form of advertising, but we interpret it as the intensity of exposure (or dose) and assume the variable can take on the full continuum of values. We revise the notation and definitions introduced for binary treatments in Section 5.1.1. Instead of Equations (1)-(2) each individual i now has infinitely many potential outcomes:

$$Y_{di}^* = X_i\beta_d + \epsilon_{di}, \qquad \text{for } d \in \zeta. \tag{11}$$

Furthermore, we replace Equation (3) and define the treatment variable T_i as:

$$T_i = Z_i\theta + \epsilon_{Ti}. \tag{12}$$

The observed outcome Y_i becomes (instead of Equation (5)):

$$Y_i = \begin{cases} Y_{di} = 1, & \text{if } Y_{di}^* > 0, \\ Y_{di} = 0, & \text{if } Y_{di}^* \leq 0. \end{cases} \tag{13}$$

As for binary treatments, we assume the observed data (Y_i, X_i, Z_i, T_i) for $i = 1, ..., N$ is an i.i.d. sample from the target population of interest. However, we do not expect that the treatment intensity (or dose) is randomly assigned.

5.2.2 Causal quantities of interest

Ideally, one would like to know the response of an individual i to each specific treatment dose d, resulting in the function $Y_i(d)$. However, we only observe the outcome Y_{di} for one randomly chosen treatment dose $T_i = d$. Hence, an individual dose-response function is not identified. As an alternative, we estimate the ATE τ_{ATE} and look at the average dose-response function (ADRF):

$$\mu(d) = E[Y_i(d)] \qquad \text{for } d \in \zeta.$$

This function gives the mean response for the population for a certain treatment dose d. This is not the same as $\mu(d) = E[Y_i(d)|T_i = d]$, which is done by conventional regression analysis. Once we have the ADRF $\mu(d)$, we can measure the total response in the population for a certain treatment dose. Moreover, one can examine the effect of a change in treatment dose, for example the difference between $\mu(d_1)$ for $T_i = d_1$ to $\mu(d_2)$ for $T_i = d_2$.

In addition to the ADRF, we investigate its counterpart, the treated dose-response function (TDRF). This function, indicated by $\tau(d)$, gives the mean response for individuals that received a certain treatment dose d. This is comparable to ATE versus ATT in the binary case. We define both dose-response functions as the conditional expectation of treatment dose d minus the base level of receiving no treatment. This is done to enable an easy interpretation.

5.2.3 Identification

In the literature for multi-valued treatments, the traditional definition of the propensity score no longer holds. Different generalizations of the binary propensity score are possible. In the early literature on generalized treatments, Imbens (2000) suggests computing a propensity score for each level of a categorical treatment variable, which results in different propensity scores for each dose level. Joffe and Rosenbaum (1999) consider propensity scores with doses of treatment and compute a single scalar propensity score instead. This subtle difference leads to two methods that generalize propensity scores:

- First, the propensity function (PF) is defined as the conditional density[11] of treatment given covariates (Imai & Van Dyk, 2004). We assume there exists a unique finite dimensional parameter $\pi_i(Z_i)$ such that $e(T_i|Z_i)$ only depends on Z_i through $\pi_i(Z_i)$, so $e(T_i|Z_i)$ equals $e(T_i|\pi_i)$.
- Second, the generalized propensity score (GPS) is equal to the treatment assignment model density function evaluated at the observed treatment variable and covariate for a particular individual (Hirano & Imbens, 2004). Hence, the GPS is defined as the probability to receive a certain dose of treatment d given observed covariates:

$$\pi_i(d, Z_i) = r(d, Z_i) = Pr(T_i = d|Z_i) = f_{T|Z}(d|Z_i).$$

Both methods are analogues to the propensity score in the binary case. A downside of the GPS is that propensity scores for different doses are different functions of covariates. Therefore, interpretation when comparing individuals with similar propensity scores but different doses is lacking. This makes the choice of response models more restricted for this method (e.g. subclassification is not possible). However, an advantage is that GPS is transformed to a probability scale (bounded between 0 and 1). Therefore, interpretation is more similar to the binary case and it can be used in combination with weighting methods.

We now revise two assumptions from Section 5.1.3:

1. Conditional independence (unconfoundedness): $Y_i \perp T_i|Z_i$ given $T_i = d$ for all $d \in \zeta$. In words, the potential outcomes are independent of treatment assignment conditional on the observed covariates given treatment dose d. This states there is no hidden bias due to unmeasured confounders. If this assumption holds, then also $Y_i \perp T_i|\pi_i$. Furthermore, exposure to treatment is random for a given propensity score $T_i \perp Z_i|\pi_i$.

[11] This is not a real probability, as the propensity function is not necessarily bounded between 0 and 1.

2. Positivity (common support): $Pr(T_i = d|Z_i) > 0$ for all $d \in \zeta$. In words, this assumption states that each individual i has a non-zero probability of receiving each treatment dose d given the observed covariates.

Assumption [3.] Stable unit treatment value assumption remains the same.

5.2.4 Treatment assignment model

As in the case of binary treatments, we begin by modelling the observed treatment T_i given the covariates Z_i. Variable selection is similar as the binary case, but we use different models for estimation. We investigate the following models to estimate the continuous propensity score and compare their results[12]:

1. *Linear regression*
   For continuous treatments, a common way to estimate the propensity score is by modelling the distribution of the treatment given covariates as normal density function (Imai & Van Dyk, 2004). We can then simply estimate the parameters $\hat{\theta}$ of a linear regression propensity score model using OLS:

   $$T_i = Z_i\theta + \epsilon_{Ti}.$$

   Although this might not be the most suitable model for our data as the treatment variable T_i is far from normal, we include it as a comparison and investigate its performance as propensity score model.

2. *Poisson regression*
   However, as treatment is defined as the number of exposures here, we strictly speaking have count data. Count data are typically modelled using a Poisson linear regression. Therefore, we suggest using a Poisson regression as treatment assignment model:

   $$log(T_i) = Z_i\theta + \epsilon_{Ti}.$$

   We implement this using the `svyglm package` in R (Lumley, 2019).

3. *Boosting*
   Alternatively, we can also use a non-parametric method such as gradient boosting in the continuous case (Zhu, Coffman, & Ghosh, 2015). Similarly to the binary case, we implement gradient boosting using the `gbm package` in R and optimize over the resulting GBM to find the number or trees for which the total covariate imbalance is lowest (Greenwell et al., 2019). We define total covariate imbalance as the average absolute standardized mean difference of all covariates over the different strata (see Section 5.2.6). We use a Poisson distribution with the same tuning parameters as before.

4. *CBGPS*
   Similarly, we use the CBGPS as in binary treatment case. This method has been generalized to continuous treatments by Fong et al. (2018) (CBGPS). CBGPS minimizes the correlation between the treatment variable T and covariates Z. It does so using generalized method of moments (GMM) estimation with moment conditions. Fong et al. (2018) developed a parametric and non-parametric method, where the latter does not require a correct model specification. Although advantageous, we use

[12] We include the sample weights r_i as weights in each of the treatment assignment models.

the parametric method due to the large computational cost of the non-parametric method. However, this method might suffer from numerical instability. We implement this using the `CBPS package` in R (Fong et al., 2019).

The fitted treatment model gives us the uniquely defined PF $\hat{\pi}_i(Z_i)$ for each of the above methods. To compute the GPS $\hat{\pi}_i(T_i, Z_i)$ we need the probability density function of the treatment variable. For the methods assuming the treatment variable T_i is normally distributed, we can use the normal density function to do this:

$$\hat{\pi}_i(T_i, Z_i) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(- \frac{\hat{\epsilon}_i}{2\hat{\sigma}^2} \right),$$

where $\hat{\epsilon}_i = T_i - Z_i\hat{\theta}$ and $\hat{\sigma}^2$ are estimated by a propensity score model. This is used to compute the GPS for the *Linear regression* and *CBGPS* method. Assuming the treatment variable T_i is Poisson distributed, as for the *Poisson regression* and *Boosting* method, we get:

$$\hat{\pi}_i(T_i, Z_i) = \frac{e^{-\hat{\psi}}\hat{\psi}^{T_i}}{T_i!},$$

where $\hat{\psi} = e^{Z_i\hat{\theta}}$.

5.2.5 Response model

Details for the response model are similar to the binary case in terms of the variable selection and the outcome model used (see Section 5.1.5). However, the weights are computed slightly different for ATE. We use the approach of Robins, Hernan, and Brumback (2000) who propose to use stabilized weights:

$$w_{ATE,i} = \frac{Pr(T_i = d)}{\pi_i(d, Z_i)} = \frac{f(T_i)}{f(T_i|Z_i)}.$$

Here, the numerator $Pr(T_i = d)$, the marginal density function of T_i, is included as stabilizing factor. This is necessary when T_i is continuous to avoid that some individuals get extremely large weights. A reasonable choice is often (Austin, 2018):

$$f(T_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{1}{2\sigma^2}(T_i - \mu) \right),$$

where μ and σ are the mean and variance of the treatment variable in the overall sample. However, if the normal assumption does not hold we should use another (possibly non-parametric) method to estimate $f(T_i)$ (Zhu et al., 2015). Hence, we use Kernel density estimation to estimate $f(T_i)$. We choose the bandwidth using Silverman's rule of thumb, Gaussian smoothing kernels and restrict the estimation such that there is no density below zero (treatment exposures are strictly positive).

We again multiply the obtained weight with the sample weight r_i to ensure representativeness of the sample (Ridgeway et al., 2015). We then normalize the obtained weights as in the binary case, but now make the average weight one over the full sample:

$$w_{ATE,i} = \frac{r_i \cdot w_{ATE,i} \cdot N}{\sum_{i=1}^{N} r_i \cdot w_{ATE,i}}.$$

5.2.6 Diagnostics

We also use similar model diagnostics to assess the performance of the treatment assignment model as in the binary treatment case, but adapt these slightly to the continuous setting.

Propensity scores
As before, create histograms of the propensity scores to check the size of the values and truncate extreme values to reduce the variance of the estimates. As a threshold we again use the 2.5% and 97.5% percentiles of the propensity score. We do this for both the PF and GPS.

Covariate balance
For binary treatment variables one can compare the balance of covariates in the treatment and control group before and after weighting. For continuous treatment variables this is more difficult. We assess the balance following the approach of Austin (2018). In short, the idea is to make groups and check the balance by comparing the SMD in a group with the other groups. To assess the covariate balance for the GPS, we take the following steps:

1. Split the treatment variable T in K strata S_1, ... S_K and compute the median of the treatment variable in each stratum t_1, ... t_K. We use the median as the treatment variable is a count variable.

Now, for each stratum $k \in 1, ..., K$:

2. Evaluate the GPS at the median t_k for the entire sample and use this to construct five quantiles.

3. Now define an indicator quantifying if the treatment variable T_i of individual i is in stratum k.

4. Compute the SMD as defined in Section 5.1.6 between individuals i in stratum k versus not in stratum k using the indicator for each of the five quantiles. This results in SMD_1^k, ..., SMD_5^k.

5. Take the average of the absolute values of SMD_1^k, ...,SMD_5^k to get SMD^k.

Compare this with the initial imbalance, which we obtain by comparing the SMD between individuals i in stratum k versus not in stratum k without incorporating GPS. We do not have to perform balance checks for PF and GPS separately as they both use the same propensity score models (Zhao, van Dyk, & Imai, 2018). Furthermore, we define total covariate imbalance as $\frac{1}{C}\frac{1}{K}\sum_{j=1}^{C}\sum_{k=1}^{K}|SMD_j^k|$.

Common support
Kluve et al. (2012) propose to use a similar tactic to assess the common support assumption as the covariate balance. Hence, we follow steps [1.]-[3.] as before and use the split between individuals i in stratum k versus not in stratum k to compare the overlap in distributions in a similar way as for binary treatments.

5.2.7 Treatment effects

Again there are many different methods to estimate the treatment effects. To estimate the effects of interest, ATE, ADRF and TDRF, we choose to use weighted least squares (WLS), subclassification (SC) and the smooth coefficient model (SCM). We use a weighting method for consistency with the binary treatment case. However, we also investigate

subclassification as it can be used to model multivariate treatments. The smooth coefficient model is an extension of subclassification. We compare the double robust version of these methods, including the outcome variables, to the case excluding additional covariates. Each of these methods will now be discussed in more detail.

Weighted least squares
Once we obtained the stabilized weights $\hat{w}_{ATE}$, we can estimate the causal effect using weighted least squares (Robins et al., 2000; Austin, 2018). As in the binary case, WLS gives the possibility to include additional covariates in the model, thereby specifying a model for Y. We add the GPS $\hat{\pi}_i(d, Z_i)$ and outcome variables that are out-of-balance after applying the treatment assignment model as covariates to the response model (the latter only in double robust case). We estimate the marginal effects as follows:

$$\hat{\tau}_{ATE,WLS} = \frac{\partial Y}{\partial T} = \hat{\gamma} \cdot \lambda(X\hat{\beta} + T\hat{\gamma}) \approx \hat{\gamma} \cdot \frac{1}{N} \sum_{i=1}^{N} \lambda(X_i\hat{\beta} + T_i\hat{\gamma}),$$

where $\lambda(.)$ is the probability density function for the logistic distribution and $\hat{\beta}$, $\hat{\gamma}$ are estimated using $\hat{w}_{ATE}$ in the logit model. We estimate the ADRF $\hat{\mu}_{WLS}(d)$ by averaging the response over all individuals conditional on the (recalculated) GPS for each treatment level d. For the TDRF $\hat{\tau}_{WLS}(d)$ we average the response over all individuals that received treatment level d.

Subclassification
We can approximate the causal effect by making subclasses based on the propensity score $\hat{\pi}_i(Z_i)$ (Imai & Van Dyk, 2004). The goal of making forming classes is that the baseline covariates in the treatment and control group are similar within each class, hence we use the PF instead of the GPS (not uniquely parameterized). Within each subclass, where individuals have similar propensity scores, we can then adequately estimate the effect of T on Y using the desired outcome model.

We subsequently compute the $\hat{\tau}_{ATE,SC}$ by a weighted sum of marginal effects with the relative proportion of observations that fall within each subclass. Additionally, we can include available covariates in the within-subclass model. We estimate a logit model and include the sample weights r_i as weights in the model (see Equation (10)). We use five equally sized subclasses, as this is suggested as a good strategy removing most of the initial imbalance by Zanutto et al. (2005). To compute the ADRF $\hat{\mu}_{SC}(d)$ we average the response of all individuals over all subclasses. For the TDRF $\hat{\tau}_{SC}(d)$ we do the same, but restrict to individuals that received treatment level d.

Smooth coeffficient model
Although estimating the effect of T on Y separately in each subclass is robust, it can be sensitive to the method of subclassification. Instead of making subclasses, we can allow this effect to vary as a function of $\hat{\pi}_i(Z_i)$ resulting in a SCM (Imai & Van Dyk, 2004). We fit this model using splines, flexible piece-wise functions constructed from polynomials. More specifically:

$$E[Y_i|T_i, \hat{\pi}_i(Z_i)] = f(\hat{\pi}_i(Z_i)) + g(\hat{\pi}_i(Z_i))T_i,$$

where we use penalized cubic regression splines with dimension five for f(.) and g(.) like Zhao et al. (2018). We obtain the ATE by averaging over the obtained individual treatment

effects, $\hat{\tau}_{ATE,SCM} = \sum_{i=1}^{N} \hat{g}(\hat{\pi}_i(Z_i))$.

For the ADRF $\hat{\mu}_{SCM}(d)$, we use the robust estimation method proposed by Zhao et al. (2018):

$$E[Y(d)] = E[E[Y(d)|\hat{\pi}_i(Z_i)]] = E[E[Y(T)|\hat{\pi}_i(Z_i), T = d],$$

where we estimate:

$$E[Y(T)|\hat{\pi}_i(Z_i), T = d] = h(\hat{\pi}_i(Z_i), T),$$

where h(.) is a smooth function of $\hat{\pi}_i(Z_i)$ and T. We construct h(.) using a tensor product and again use penalized cubic regression splines with dimension five. Then, we compute:

$$\hat{\mu}_{SCM}(d) = \frac{1}{N} \sum_{i=1}^{N} \hat{h}(\hat{\pi}_i(Z_i), d) - \hat{h}(\hat{\pi}_i(Z_i), 0).$$

For the TDRF $\hat{\tau}_{SC}(d)$ we do the same, but restrict to individuals that received treatment level d. We model both using `bam` of the `mgcv package` in R (Wood, 2019). This function can fit GAMs to large datasets. We use the alternative fitting approach, because of the increased computation speed due to parallelization (`discrete = TRUE`).

5.2.8 Standard errors

As for binary treatments, we need to incorporate the uncertainty resulting from estimating the propensity scores in the first stage. Therefore, the standard errors arising from the second stage estimation are not valid. We again use a bootstrap variance estimator as explained in Section 5.1.8 to compute the standard errors for the ATE, ADRF and TDRF.

5.3 Multivariate treatments

5.3.1 Notation

Now we assume we have multiple treatments $T_i = \{T_{1i}, T_{2i}, ..., T_{Mi}\}$. A case often considered in the literature is the case of choosing among a set of different treatments. This can be considered the categorical treatment case, where one receives one of the respective treatments. In this thesis, we allow individual i to receive multiple treatments, each with a certain intensity of exposure (or dose). This is called multivariate treatments. All treatment exposures are non-negative, $\zeta_m = [0, T_{mi,max}]$. We extend the notation and definitions introduced for continuous treatments in Section 5.2.1. Instead of Equation (11) each individual i now has infinitely many potential outcomes for each treatment m:

$$Y^*_{d_1,...,d_M i} = X_i \beta_{d_1,...,d_M} + \epsilon_{d_1,...,d_M i}, \qquad \text{for } d_1 \in \zeta_1, ..., d_M \in \zeta_M. \tag{14}$$

Furthermore, we replace Equation (12) and define the treatment variable T_{mi} for $m = 1, ..., M$ as:

$$T_{mi} = Z_i \theta_m + \epsilon_{Tmi}. \tag{15}$$

The observed outcome Y_i becomes (instead of Equation (13)):

$$Y_i = \begin{cases} Y_{d_1,...,d_M i} = 1, & \text{if } Y^*_{d_1,...,d_M i} > 0, \\ Y_{d_1,...,d_M i} = 0, & \text{if } Y^*_{d_1,...,d_M i} \leq 0. \end{cases} \tag{16}$$

As for binary and continuous treatments, we assume the observed data (Y_i, X_i, Z_i, T_i) for $i = 1, ..., N$ is an i.i.d. sample from the target population of interest. However, we do not expect that the treatment intensity (or dose) across the treatments is randomly assigned.

5.3.2 Univariate to multivariate

Now we extend the methodology from the case of univariate treatments to multivariate treatments. We do not discuss all separate subsections presented for binary and continuous treatments, but only the differences with continuous treatments (see Section 5.2.2-5.2.6).

Although PSMs are generalized to incorporate categorical, ordinal and continuous treatments, the analysis of multivariate treatments is uncommon in the literature. There is one exception, Imai and Van Dyk (2004) present a method to estimate the ATE for bivariate treatments based on subclassification. We are interested to estimate the causal effects for a low dimensional multivariate treatment variable, possibly larger than $M = 2$. We follow a similar approach, but extend it to compute the ADRF and TDRF. Furthermore, we investigate the use of a smooth coefficient model instead of only subclassification. We investigate additive treatment effects here. A combination of different treatments can potentially lead to smaller or larger treatment effects than the summed stand-alone effects, but we do not consider (possible) interaction effects in this thesis.

To model the observed treatment T_{mi} for $m = 1, ..., M$ given the covariates Z_i, we use the treatment assignment models for the continuous propensity score presented in Section 5.2.4. Hence, for each treatment m we obtain a propensity score model $\hat{\pi}_{mi}$, which results in M independent models. Although different estimation methods can be used for each of the M treatment assignment models, we restrict to the case where we use the same model for all treatments here.

Following the flow of PSMs, we subsequently check diagnostics. As we obtained a treatment assignment model for each treatment m, we can directly use the diagnostics described in Section 5.2.6. Hence, we assess the size of propensity scores, covariate balance and common support for each treatment separately. Details for the response model are again similar as for the binary and continuous case (see Section 5.1.5).

5.3.3 Treatment effects

To estimate the effects of interest, ATE, ADRF and TDRF, we again choose to use subclassification (SC) and the smooth coefficient model (SCM). We propose the SCM for multivariate treatments for similar reasons as in the continuous treatment case. Due to the increased computation time for double robust methods, we limit ourselves to the case excluding additional covariates for multivariate treatments. We now discuss both methods for multivariate treatments.

Subclassification
As for continuous treatments, we make subclasses based on the propensity score $\hat{\pi}_i(Z_i)$. However, instead of subclassifying on a one-dimensional variable, we use a low-dimensional variable. To avoid that subclasses become to small, we split the propensity score of each treatment in three equally sized subclasses. For $M = 2$ this results in 3-by-3 grid of treatment assignment models. In addition, we check if the size of each individual subclass is at least 100 invididuals. If this is not the case, we combine the subclass with the smallest subclass that is closest to the subclass (i.e. horizontal or vertical neighbouring subclass in grid). If there is still no variation in Y after these changes for a specific subclass, we assume the treatment effect for those individuals equals zero. This may happen for the lowest propensity score segments. We compute the ADRF $\hat{\mu}_{SC,m}(d)$ and TDRF $\hat{\tau}_{SC,m}(d)$

per treatment m as in the continuous case (see Section 5.2.7).

Smooth coefficient model
Again, we can allow this effect to vary as a function of $\hat{\pi}_{mi}(Z_i)$ resulting in a SCM (Imai & Van Dyk, 2004). Hence, we propose to fit the following model using splines, flexible piece-wise functions constructed from polynomials. More specifically:

$$E[Y_i|T_i, \hat{\pi}_i(Z_i)] = f(\hat{\pi}_{1i}(Z_i), ..., \hat{\pi}_{Mi}(Z_i)) + \sum_{m=1}^{M} g_m(\hat{\pi}_{1i}(Z_i), ..., \hat{\pi}_{Mi}(Z_i))T_{mi},$$

where we choose to use a slightly less flexible function and opt for splines of dimension three for f(.) and $g_m(.)$. Then we define $\hat{\tau}_{ATE,SCM,m} = \sum_{i=1}^{N} \hat{g}_m(\hat{\pi}_{1i}(Z_i), ..., \hat{\pi}_{Mi}(Z_i))$.

For the ADRF $\hat{\mu}_{SCM,m}(d)$, we extend the robust estimation method proposed by Zhao et al. (2018) to multivariate treatments:

$$\begin{aligned} E[Y(d)] &= E[E[Y(d)|\hat{\pi}_{1i}(Z_i), ..., \hat{\pi}_{Mi}(Z_i)]] = \\ &\quad E[E[Y(T)|\hat{\pi}_{1i}(Z_i), ..., \hat{\pi}_{Mi}(Z_i), T_{1i} = d_1, ..., T_{Mi} = d_M], \end{aligned}$$

Hence, we estimate:

$$E[Y(T)|\hat{\pi}_i(Z_i), T = d] = \sum_{m=1}^{M} h_m(\hat{\pi}_{1i}(Z_i), ..., \hat{\pi}_{Mi}(Z_i), T_m),$$

where $h_m(.)$ is a smooth function of $\hat{\pi}_i(Z_i)$ for $m = 1, ..., M$ and T_m. Then, we compute:

$$\hat{\mu}_{SCM,m}(d) = \frac{1}{N} \sum_{i=1}^{N} \hat{h}_m(\hat{\pi}_{1i}(Z_i), ..., \hat{\pi}_{Mi}(Z_i), d) - \hat{h}_m(\hat{\pi}_{1i}(Z_i), ..., \hat{\pi}_{Mi}(Z_i), 0).$$

Again, we choose to use splines of dimension three. For the TDRF $\hat{\tau}_{SCM,m}(d)$ we do the same, but restrict to individuals that received treatment level d. We again model using the `mgcv package` in R (Wood, 2019).

5.3.4 Standard errors

We again use a bootstrap variance estimator as explained in Section 5.1.8 to compute the standard errors for the ATE, ADRF and TDRF.

6 Results

6.1 Binary treatments

6.1.1 Diagnostics

In this section we examine the performance of the different treatment assignment models for binary treatments (*Logistic regression*, *Boosting* and *CBPS*). In particular, we examine 1) the size of the propensity scores, 2) the covariate balance and 3) the common support assumption.

Propensity scores
First, we analyse the size of the propensity scores. We observe some very small and large values of the propensity score for all treatment assignment models, which is undesirable as weighting methods are sensitive to these extreme values. Therefore, we chose to truncate these values. The histograms of the propensity scores before and after truncation are included in Appendix 9.5.1. Here, we summarize the results in Table 2.

Table 2: Descriptive statistics propensity scores binary treatment variables.

	Any exposure			Exposures > 2		
	Min	Max	Mean	Min	Max	Mean
Logistic regression (GLM)	0.0178	0.9718	0.2774	0.0020	0.7559	0.0936
Boosting (GBM)	0.0138	0.9158	0.2765	0.0042	0.7045	0.0929
CBPS	0.0344/ 0.0297	0.9547/ 0.9360	0.2891/ 0.2763	0.0047/ 0.0025	0.7265/ 0.6997	0.1004/ 0.0919

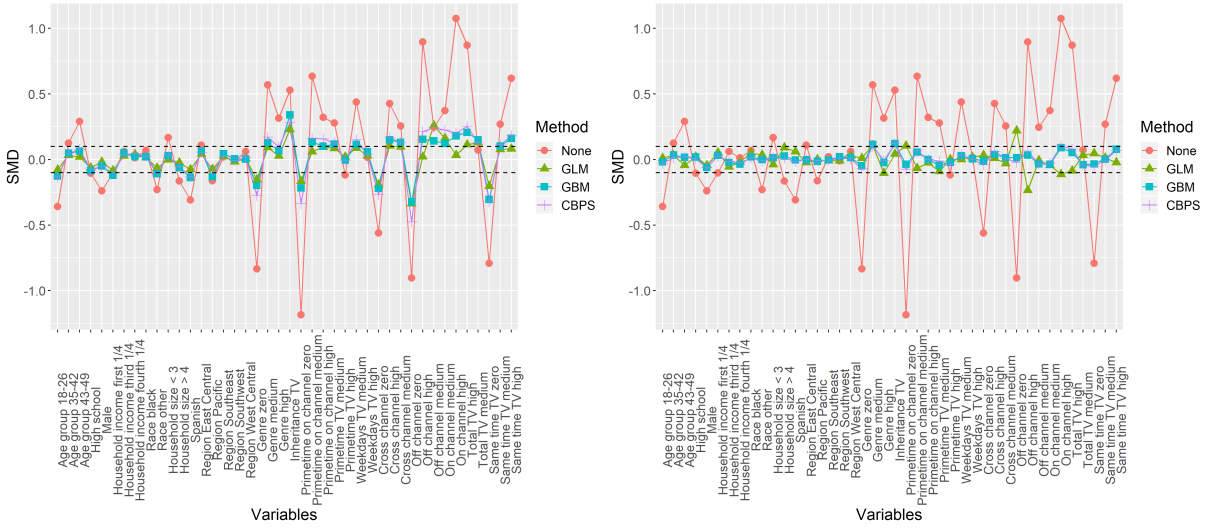
The minimum, maximum and mean of the propensity scores after truncation are reported. For CBPS the propensity scores for ATE/ATT are reported respectively.

We predict many values close to zero. This is clear from the histograms, but also because the minimum value of the propensity scores after truncation is still small (especially for *Exposures > 2*). *Boosting* has especially many values close to zero, as can be seen from the high peak in the histograms. This can be due to the relatively good prediction performance of this method. However, we do not want a treatment assignment model that perfectly predicts treatment as this results in an unequal spread of propensity scores. The resulting probabilities preferably cover the entire range from zero to one. If this is the case, the sample contains a diverse set of individuals. *Logistic regression* and *CBPS* result in probabilities with a more equal spread over the interval, especially in the lower part of the interval (0 - 0.2). This spread is better for *Any exposure* than *Exposures > 2*, as the proportion of individuals receiving treatment versus non-treatment is more balanced for this treatment variable.

Moreover, it is good to notice the relatively low maximum propensity score for *Exposures > 2*. This is the direct result of the choice of truncation level combined with the fact that there are relatively few individuals in the treatment group for this treatment variable ($N = 2379$). These individuals, who have higher probabilities (on average), are relatively few compared to the trimming level. This leads to a relatively large decrease.

Covariate balance

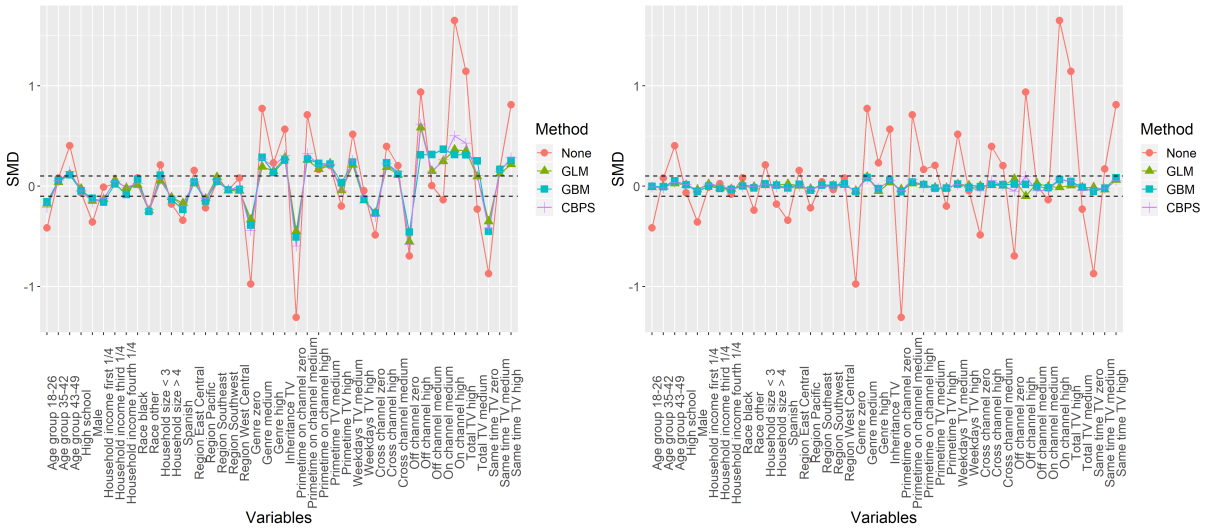
The goal of the treatment assignment model is to balance the covariates important for the response model. In Figures 5 and 6, we present the covariate balance for the two binary treatment variables *Any exposure* and *Exposures > 2*. The figures show the SMD (on the y-axis) for each of the covariates included in the outcome model (on the x-axis) for ATE and ATT respectively. The different lines belong to the different treatment assignment models. We first examine the initial imbalance, without applying a balancing propensity score. This is shown by the red line. We notice the SMD of general viewing behaviour variables is usually larger in absolute terms than that of sociodemographics, which means baseline viewing behaviour differs more between the treatment and non-treatment group than other variables. Furthermore, we know variables with a negative SMD are overrepresented in the non-treatment group. This is for example the case for the zero groups of *Genre*, *Primetime on channel*, *Cross/Off channel* and *Same time TV*, which is as expected a priori. Similarly, we see a large positive SMD for the high quantile group of these variables, implying we observe larger values for these variables in the treatment group.



Balance ATE. Balance ATT.

Figure 5: Covariate balance for binary treatment variable (a) *Any exposure* measured by SMD. The dotted line indicates |SMD| = 0.1.

Using the initially proposed variable selection for the treatment assignment model, we find that the total covariate balance improved for all methods. All variables that were initially out-of-balance are more in balance and many variables have negligible differences ($|\text{SMD}| < 0.1$) between the treatment and non-treatment group. However, the extend to which the balance improved differs for the two binary treatment variables (a, b) and causal effects of interest (ATE, ATT). To summarize the covariate balance, we present the total covariate imbalance metric in Table 3.



Balance ATE. Balance ATT.

Figure 6: Covariate balance for binary treatment variable (b) *Exposure* > 2 measured by SMD. The dotted line indicates $|\text{SMD}| = 0.1$.

Table 3: Total covariate imbalance binary treatment variables.

	Any exposure		Exposures > 2	
	ATE	ATT	ATE	ATT
Logistic regression (GLM)	0.0954	0.0855	0.2200	0.0294
Boosting (GBM)	0.1421	0.0516	0.2332	0.0360
CBPS	0.1508	0.0487	0.2425	0.0365

The initial imbalances are 0.3694 and 0.3995.

We see the balance is generally better for ATT than for ATE after weighting. This is the case for all examined treatment assignment models and both binary treatment variables. Furthermore, the balance of ATE is much better for *Any exposure* than for *Exposures* > 2. However, this is not true for ATT. A possible reason for this is that the proportion of treated versus non-treated individuals is more problematic for ATE and the fact that the methods perform worse if the balance of treated versus non-treated deteriorates. Although the methods still perform reasonably, the results for *Exposures* > 2 suggest balancing for a treatment variable that much less than 10% of the sample is exposed to is difficult.

Several attempts to further improve the balance of out-of-balance covariates did not lead to noticeable improvements. Given the selection of variables for the response model, we aim to find an appropriate treatment assignment model to balance these covariates. We first opted to include only the most influential outcome regressors that were measured before treatment assignment. We determined this by a simple OLS regression using significance level < 0.1. Next, we tried adding interaction terms (e.g. between *Total TV* and *On/Cross/Off channel*), but this mainly resulted in more multicollinearity problems. This approach is probably more lucrative if the model does not solely contain binary esplanatory variables. Furthermore, to avoid that the result would be influenced by variables that did not matter, we also tried to include only variables that where out-of-balance initially ($|\text{SMD}| > 0.25$). As these changes in the treatment assignment model did not significantly

improve the balance, we decided to stick to the (simple) initial variable selection.

Finally, we compare the achieved covariate balance of the different treatment assignment models. The differences for this set of treatment and outcome variables are not very large and all methods perform quite well. However, from experimenting with other sets of variables, we have seen the performance is sensitive to the variables included in the model. We find *Logistic regression* performs significantly worse when adding/removing variables, while the performance of *Boosting* and *CBPS* remains unchanged. Also, if there are continuous explanatory variables included in the model, *Boosting* or *CBPS* usually performs best in terms of covariate balance. Because *CBPS* has a better spread of values than *Boosting* (as earlier concluded), we prefer to rely on *CBPS* as treatment assignment model for binary treatments here.

Common support
Finally, to estimate a population treatment effect it is important that there is sufficient overlap (common support) between the treatment and non-treatment group. Histograms of the propensity scores of individuals in the treatment and non-treatment group are included in Appendix 9.5.1. We find that the minima and maxima of the propensity scores are similar for both groups. However, the propensity scores are on average (much) lower for the non-treatment group than for the treatment group. This is especially the case for the non-treatment group of *Exposures* $>$ *2*. Moreover, the estimated treatment effect for *Exposures* $>$ *2* is limited to the region 0-0.7. Based on the graphs, we conclude the common support assumption is sufficiently satisfied as similar individuals are observed in the treatment and non-treatment group. However, we do note the support is relatively thin.

6.1.2 Treatment effects

In this section we present and discuss the estimated treatment effects using inverse propensity weighting (IPW), weighted least squares (WLS) and the augmented IPW (AIPW). We first discuss the results for *Any exposure* and then *Exposures* $>$ *2*. After, we discuss the performance of the bootstrap variance estimator and investigate the robustness of the results to the assumption of unconfoundedness.

In Table 4 we present the estimated causal effects of *Any exposure*. The range of estimates for ATE and ATT are 0.005 - 0.038 and 0.006 - 0.024 respectively. This is quite a wide range of estimates, but all methods estimate a positive treatment effect. The estimated effects are, however, not significant for the majority of methods. IPW gives higher treatment effect estimates than WLS and AIPW. This suggest the treatment assignment model might not be correctly specified and correcting for additional covariates in the response model is necessary (as done by WLS or AIPW). However, we also know AIPW is very sensitive to weights, in particular it can underestimate the results if the support is thin (Glynn & Quinn, 2010). In the previous section we already concluded this is the case and we indeed find small treatment effects in comparison to the other methods. Therefore, we consider the results obtained using WLS most trustworthy here. Taking into account we considered *CBPS* the best treatment assignment model, we thus find an estimate of 0.7% and 0.5% for ATE and ATT respectively. In general, we find that the estimates for ATE are larger than (or equal) to ATT, but the difference is not significant. The estimated standard errors are similar in size for the methods, but somewhat larger for ATT than for ATE.

Table 4: Causal effects of binary treatment variable (a) *Any exposure* on tune-in AGT.

	IPW		WLS		AIPW	
	ATE	ATT	ATE	ATT	ATE	ATT
Logistic regression (GLM)	0.023*	0.012	0.009*	0.009*	0.008	0.012
	(0.008)	(0.016)	(0.003)	(0.004)	(0.006)	(0.009)
Boosting (GBM)	0.030*	0.023*	0.003	0.003	0.005	0.006
	(0.006)	(0.011)	(0.004)	(0.004)	(0.006)	(0.007)
CBPS	0.038*	0.024*	0.007	0.005	0.008	0.009
	(0.007)	(0.011)	(0.004)	(0.004)	(0.006)	(0.007)

Standard errors are given in parentheses, computed using bootstrap variance estimator. Effects significant at 5% significance level are marked with (*).

In Table 5 we present the estimated causal effects of *Exposures > 2*. The range of estimates for ATE and ATT are 0.004 - 0.072 and 0.001 - 0.059 respectively. Hence, we have slightly higher and wider estimates than for *Any exposure*, which is in line with our expectations. Furthermore, the majority of methods now find significant positive coefficients (both for ATE and ATT). Using the same reasoning as before, we choose to rely on the estimates with *CBPS* as treatment assignment model and WLS as response model here. Hence, we find a positive significant estimate of 1.0% and 1.2% for ATE and ATT respectively.

Table 5: Causal effects of binary treatment variable (b) *Exposures > 2* on tune-in AGT.

	IPW		WLS		AIPW	
	ATE	ATT	ATE	ATT	ATE	ATT
Logistic regression (GLM)	0.056*	0.056*	0.010*	0.014*	0.005	0.002
	(0.007)	(0.015)	(0.003)	(0.003)	(0.004)	(0.007)
Boosting (GBM)	0.050*	0.057*	0.005	0.009*	0.004	0.001
	(0.009)	(0.014)	(0.003)	(0.004)	(0.004)	(0.007)
CBPS	0.072*	0.059	0.010*	0.012*	0.005	0.001
	(0.007)	(0.014)	0.003)	(0.003)	(0.004)	(0.006)

Standard errors are given in parentheses, computed using bootstrap variance estimator. Effects significant at 5% significance level are marked with (*).

To conclude, we find a significant positive ATE and ATT when an individual is exposed to advertisements at least three times (*Exposures > 2*). However, we do not find an effect different from zero when we define treatment as being exposed at least once (*Any exposure*). The methods give a wide range of estimates, but consistently positive. The size of the effect can be a topic of debate, but we conclude that the advertising campaign as a whole has had a small positive impact on the probability of tune-in of AGT.

Now, we compare the standard errors obtained using bootstrapping with the empirical sandwich estimator. The results are shown in Table 6. The relative size of the standard errors is the same for both estimation approaches. Furthermore, despite the relatively small number of bootstrap replications ($B = 25$), the errors obtained using the bootstrap sandwich error are very similar to the empirical sandwich estimator. We expect this to converge even further for a higher number of iterations B. This provides confidence in the accurateness of the bootstrap errors presented in this thesis and we argue bootstrap errors are a good alternative to the empirical sandwich estimator.

Table 6: Comparison standard errors using bootstrap and empirical sandwich estimator.

		Bootstrap standard error	Sandwich standard error
Any exposure	IPW	6.659×10^{-3}	6.155×10^{-3}
	AIPW	4.807×10^{-3}	5.841×10^{-3}
Exposures > 2	IPW	1.174×10^{-2}	6.771×10^{-3}
	AIPW	5.693×10^{-3}	6.073×10^{-3}

Bootstrap standard errors are computed without sample weights r_i for comparability to sandwich standard errors.

Finally, we investigate how sensitive the results are to unmeasured confounders. Although the described sensitivity analysis is developed for matching and cannot be used to prove unmeasured confounders are not a problem in weighting methods, we argue we can use it as an indicator of how large the problem of unobserved confounding is, as it is not a property of the chosen methods but rather of the problem at hand.

In Table 7, we see the minimum value obtained is 1.4. This means the unobserved confounders need to increase the odds at least 1.4 times for the results to change. All other Γ^* values are higher, meaning unobserved confounders need to bring across even more change to influence the results. This suggests the assumption of unobserved confounding is reasonable to make. Furthermore, we observe *Any exposure* is more robust to the assumption of unmeasured confounders than *Exposures > 2*. We recommend further research to investigate sensitivity analysis for weighting methods to give stronger evidence for the correctness of this assumption.

Table 7: Robustness to unconfoundedness assumption.

	Any exposure	Exposures > 2
Logistic regression (GLM)	2.0	1.4
Boosting (GBM)	2.0	1.5
CBPS	2.0	1.4

This table contains the Γ^* values for which the treatment effect becomes insignificant (significance level 0.05).

6.2 Continuous treatments

6.2.1 Diagnostics

Next, we examine the performance of the different treatment assignment models for continuous treatments (*Linear regression*, *Poisson regression*, *Boosting* and *CBGPS*). In particular, we examine 1) the size of the propensity scores, 2) the covariate balance and 3) the common support assumption.

Propensity scores
We again investigate the propensity scores before and after truncation. The histograms are included in Appendix 9.5.2 and the summarized results are presented in Table 8. We note there is much more difference between treatment assignment models for continuous treatments. The range of the propensity scores, as well as the mean, differ substantially between methods. The range of estimated GPS is really small for the two methods based

on the normal distribution (*Linear regression* and *CBGPS*). This makes sense, as our treatment variable is far from normally distributed. The *Poisson regression* and *Boosting* are thus more appropriate to model the treatment variable. Hence, despite the popularity of the *Linear regression* we see it is important to look at the distribution of the treatment variable at hand. Finally, we see the distribution of the *Poisson regression* is much less skewed than the distribution obtained by *Boosting*. The propensity scores obtained by the *Poisson regression* span the entire range from zero to one.

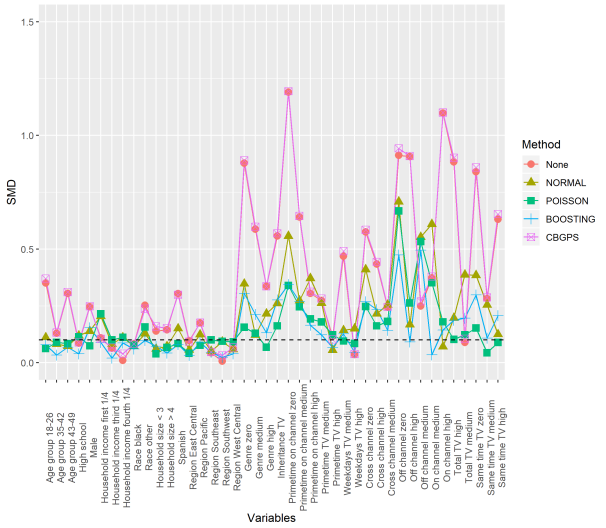
Table 8: Descriptive statistics GPS continuous treatment variable.

	Number of exposures		
	Min	Max	Mean
Linear regression (Normal)	0.0162	0.2599	0.2234
Poisson regression (Poisson)	0.0084	0.9534	0.6211
Boosting	0.0039	0.7792	0.5744
CBGPS	0.1211	0.1631	0.1588

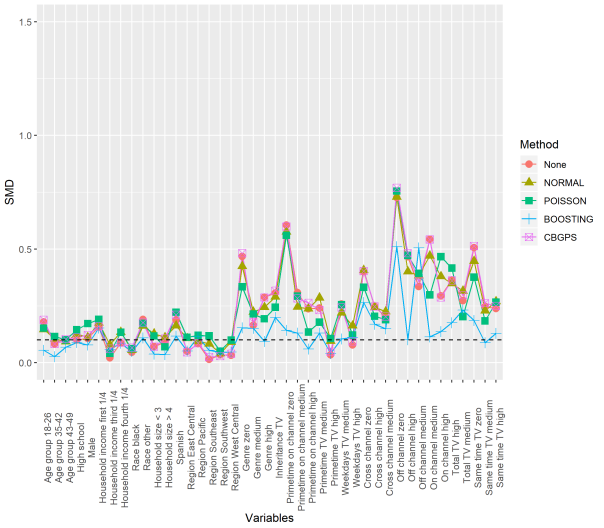
The minimum, maximum and mean of the GPS after truncation are reported.

Covariate balance
Now, we examine the covariate balance for continuous treatments. The lowest stratum (k = 1) contains individuals with zero exposures, the middle stratum (k = 2) contains individuals with one exposure and the high stratum (k = 3) contains individuals with more than one exposure. The comparison group are all remaining individuals not in stratum k. We show the covariate balance for the continuous treatment variable in each of the strata in Figure 7. First, we look at the initial imbalance and note it is lowest in the middle stratum. We observe similar patterns in the initially imbalanced variables compared to the binary case. In particular we find initial imbalance in the zero groups of *Genre*, *Primetime on channel*, *Off channel* and *Same time TV*, but also in the high group of On channel and *Total TV*.

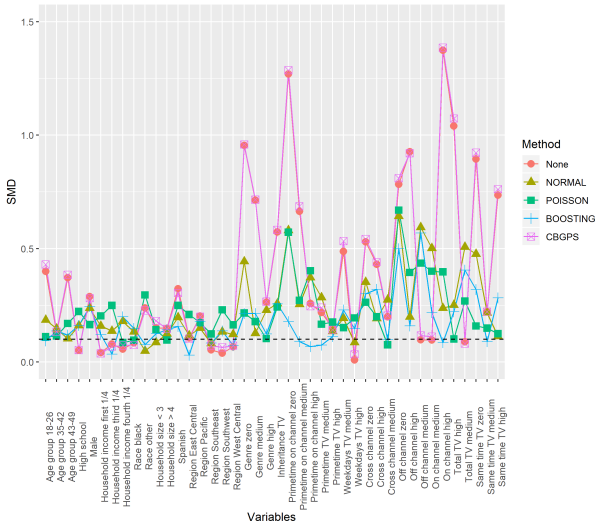
The total covariate balance improved for all methods in the low/high stratum, except for *CBGPS*. We find that *CBGPS* generally performs poorly and that it sometimes suffers from numerical instability. This is related to, but not due to, the chosen variable selection as this is also the case for other variable selections. *CBGPS* is thus very sensitive to model specification. For this reason, we do not recommend using this as treatment assignment model here despite its promising appearance. The non-parametric version of *CBGPS* might be worth investigating to overcome this problem. However, we do not investigate this further as the method is too computationally expensive. Furthermore, we observe *Boosting* is the only method that can improve the relatively good initial balance in the middle stratum. To further compare the methods, we look at the total covariate imbalance metric per stratum in Table 9.



Balance stratum low (k = 1).



Balance stratum middle (k = 2).



Balance stratum high (k = 3).

Figure 7: Covariate balance for continuous treatment variable *Number of exposures* measured by SMD. The dotted line indicates SMD = 0.1.

Table 9: Total covariate imbalance continuous treatment variable.

	k = 1	k = 2	k = 3
Linear regression (Normal)	0.2117	0.2349	0.2387
Poisson regression (Poisson)	0.1561	0.2235	0.2243
Boosting	0.1453	0.1275	0.1796
CBGPS	0.3297	0.2507	0.3816

The initial imbalances are 0.3694, 0.2254 and 0.3767.

As in the binary treatment case, we tried to improve the balance of out-of-balance covariates by using different sets of regressors. Although this was not beneficial for most methods, the results presented for *CBGPS* in this section are obtained by adding only the most influential outcome regressors that were measured before treatment assignment. For the remaining methods we stick to the (simple) initial variable selection.

We conclude *Boosting* is the best method to achieve covariate balance for continuous treatments in our dataset. This method achieves the lowest covariate imbalance for all strata. It is followed in performance by the *Poisson regression*. This method mainly beats the other methods in the low stratum, but still performs relatively well.

Common support
Finally, we examine the common support assumption. Histograms of the propensity scores of individuals in a certain stratum versus not are included in Appendix 9.5.2. Note that the interpretation of the propensity scores is different here than for the binary treatment case. Loosely speaking, the GPS measures how likely an individual was to receive the observed treatment dose. The common support assumption is more problematic here than for binary treatments. It highly depends on the method, but also for the two best covariate balancing methods (*Poisson regression* and *Boosting*) there is a quite a difference in propensity scores in and out of stratum k. This should be taken into account when interpreting the estimated treatment effects as it can bias the results. Common support is even worse for *Boosting* than *Poisson regression*. Hence, we prefer to rely on *Poisson regression* as treatment assignment model for continuous treatments here.

6.2.2 Treatment effects

In this section we present and discuss the estimated treatment effects using weighted least squares (WLS), subclassification (SC) and the smooth coefficient model (SCM). We compare the double robust version of these methods (DR) to the case excluding additional covariates (not DR). We first discuss the results for ATE. After, we present the ADRF and TDRF using the *Poisson regression* as treatment assignment model.

In Table 10 we present the estimated causal effects of *Number of exposures*. The range of estimates for ATE is -0.003 - 0.009. These effect sizes are fairly small, but one should keep in mind this is the estimated change in probability per treatment unit and an individual can receive more than one unit. Hence, it cannot be compared to the estimates of the binary treatment case. The majority of the effects obtained using non-double robust estimation methods have significant positive coefficients. There is one exception, the *Poisson regression* gives a significant negative coefficient for WLS. However, we see that depending on the treatment assignment model, WLS results in both positive and negative

treatment effects for *Number of exposures*. These contradictory estimates might be the result of a too restrictive model that badly describes the underlying dynamics. In constrast to the other methods, WLS only estimates a single coefficient for the treatment variable. SC and the SCM are more flexible and allow this (and other coefficients) to be different for different levels of propensity scores. We observe the estimated effects for SC and SCM are smaller in general, but more conclusive (all positive).

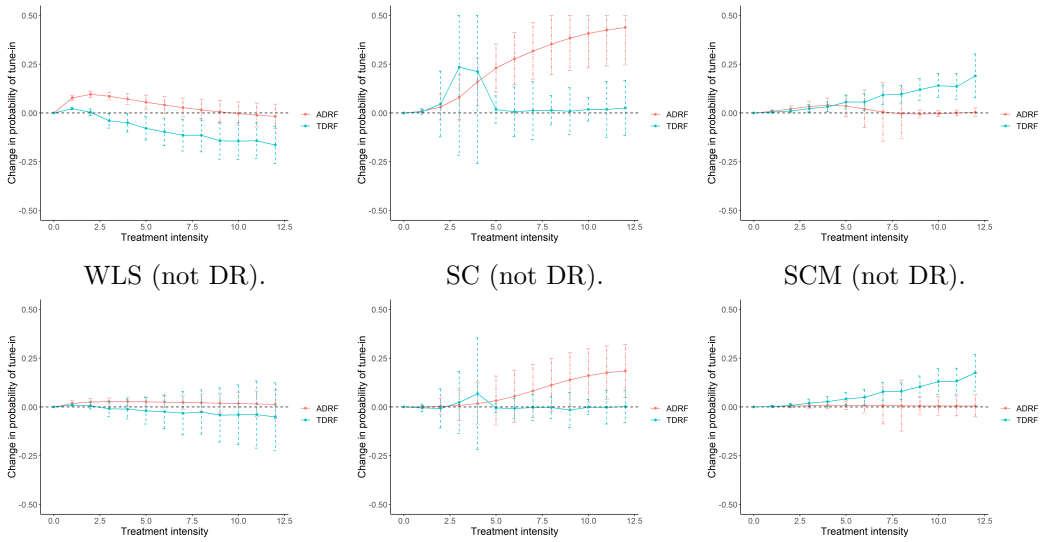
The effects obtained using double robust estimation are somewhat smaller. As these added covariates influence the estimated treatment effect and using our knowledge that the covariates are not perfectly balanced after the treatment assignment model, we believe the double robust results are more thrustworthy. Moreover, based on the diagnostics discussed in the previous section, we prefer the *Poisson regression* as treatment assignment model. Taking into account that the SCM is a generalization of SC and this method can be sensitive to the number of subclasses, we choose to rely on the results of the former. This suggests a positive significant ATE of 0.1% for *Number of exposures*.

Table 10: Causal effects of continuous treatment variable *Number of exposures* on tune-in AGT.

	Not double robust			Double robust		
	WLS	SC	SCM	WLS	SC	SCM
Linear regression (Normal)	0.002*	0.003*	0.002*	0.002*	0.001	0.001
	(0.001)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)
Poisson regression (Poisson)	-0.003*	0.003*	0.001*	-0.001	0.001	0.001*
	(0.001)	(0.000)	(0.000)	(0.002)	(0.001)	(0.000)
Boosting	-0.001	0.002	0.001*	-0.003	0.000	0.001
	(0.000)	(0.001)	(0.000)	(0.002)	(0.006)	(0.000)
CBGPS	0.009*	0.003*	0.003*	0.005*	0.001	0.001
	(0.002)	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)

The values represent the estimated ATE. Standard errors are given in parentheses, computed using bootstrap variance estimator. Effects significant at 5% significance level are marked with (*).

Next, we look at the dose-response functions depicted in Figure 8. The curves can directly be interpreted as the change in the probability of tune-in (on the y-axis) due to treatment dose d (on the x-axis) compared to receiving no treatment. The obtained dose-response functions differ quite a lot between the methods. The ADRF is higher than the TDRF for WLS and SC. This is also the case for the SCM for low treatment values (where we have most observations). This means the that additional exposures to advertising are less influential for individuals who are exposed to treatment than for a random selected individual in the population. Furthermore, we see a small increase in the ADRF followed by a stagnation or (weak) decrease. Hence, the effectiveness of additional treatment units decreases at some point (or even has negative impact). Both WLS and SCM show small significant increases in the ADRF for low treatment values. SC shows large significant increases for high treatment values. The results for the TDRF are inconclusive between the methods (both decreasing and increasing trends visible). This is also clear from the estimates for ATE. Like for the ATE, the effect sizes are smaller for the double robust curves. The curves for *Boosting* are presented in Appendix 9.5.2, but are very similar to the *Poisson regression*.



WLS (not DR). SC (not DR). SCM (not DR).

WLS (DR). SC (DR). SCM (DR).

Figure 8: Dose-response functions for continuous treatment variable *Number of exposures* on tune-in AGT using *Poisson regression*. The error bars represent the 95% confidence interval of the curves. The error bars are cut off if the endings fall outside the figure.

To conclude, we find a positive significant ATE when an individual is exposed to an additional treatment unit. The dose-response functions show additional exposures to advertising are most beneficial for randomly selected individuals in the population. This provides the insight that the effectiveness of the marketing campaign can be increased by targeting a slightly different group of people, as advertising is more effective for them. Furthermore, we can learn until when additional exposures are beneficial from the shape of the curves. Together with the current number of exposures for individuals, we then know if we should focus more on the reach (more individuals with some exposures) or the frequency (more exposures per individual). The ADRF curve shows the added value lies mainly in the first number of exposures. This provides the insight that we should focus on increasing the reach of the marketing campaign to increase the effectiveness of the marketing campaign.

6.3 Multivariate treatments

6.3.1 Diagnostics

Here, we examine the performance of the different treatment assignment models for multivariate treatments (*Linear regression*, *Poisson regression*, *Boosting* and *CBGPS*). For each of the separate treatment variables, we find that the findings for 1) the size of the propensity scores, 2) the covariate balance and 3) the common support assumption are similar to the continuous treatment case. Therefore, we present only the main findings for multivariate treatments here.

First, we note that the propensity scores of all methods for the separate treatment variables are quite unequally spread over the interval from zero to one. As in the continuous treatment case, we find the propensity scores obtained by *Poisson regression* have the most equal spread and least skewed distribution. Next, we summarize the total covariate imbalance metric for multivariate treatment variables (a, b) in Table 11 and 12. The covariate balance plots are included in Appendix 9.5.3. Based on the covariate balance,

Boosting is the preferred method for all treatment variables followed by *Poisson regression*. The difference in performance between the methods differs per treatment variable, but *Boosting* always has a (slightly) lower covariate imbalance. Finally, we again find that the common support assumption is problematic (little overlap). As before, it is slightly better for *Poisson regression* than for *Boosting*. Hence, as for continuous treatments, we choose to rely on *Poisson regression* as treatment assignment model for multivariate treatments here.

Table 11: Total covariate imbalance multivariate treatment variable (a) based on time.

	Prior to last week	Last week	Premier day
Linear regression (Normal)	0.2417	0.2403	0.2320
Poisson regression (Poisson)	0.1895	0.1979	0.2150
Boosting	0.1550	0.1741	0.2034
CBGPS	0.3069	0.2835	0.3402

The initial imbalances are 0.3332, 0.3212 and 0.3278. The values in the table are the average of the total covariate imbalance in the different strata (K = 3).

Table 12: Total covariate imbalance multivariate treatment variable (b) based on channel.

	On channel	Other channels
Linear regression (Normal)	0.2751	0.2925
Poisson regression (Poisson)	0.2317	0.2542
Boosting	0.1517	0.2180
CBGPS	0.3512	0.3262

The initial imbalances are 0.3475 and 0.3238. The values in the table are the average of the total covariate imbalance in the different strata (K = 2).

6.3.2 Treatment effects

In this section we present and discuss the estimated treatment effects using subclassification (SC) and the smooth coefficient model (SCM). We show the results of the non-double robust version of these methods here due to the increased computation time of the methods when adding additional covariates in a multivariate treatment setting. We first discuss the results for ATE. After, we present the ADRF and TDRF using the *Poisson regression* as treatment assignment model.

First, we investigate the influence of the recency of advertisements by considering advertising at different points in time as multiple treatments. In Table 13 we present the estimated causal effects of *Prior to last week*, *Last week* and *Premier day*. The ATE estimated using the SCM is (close to) zero for all treatment variables. Using SC we find that the ATE of *Premier day* is highest with 0.6%, followed by 0.2% for *Last week* and *Prior to last week*. However, none of these estimated effects is significantly different from zero.

Table 13: Causal effects of multivariate treatment variable (a) based on time on tune-in AGT.

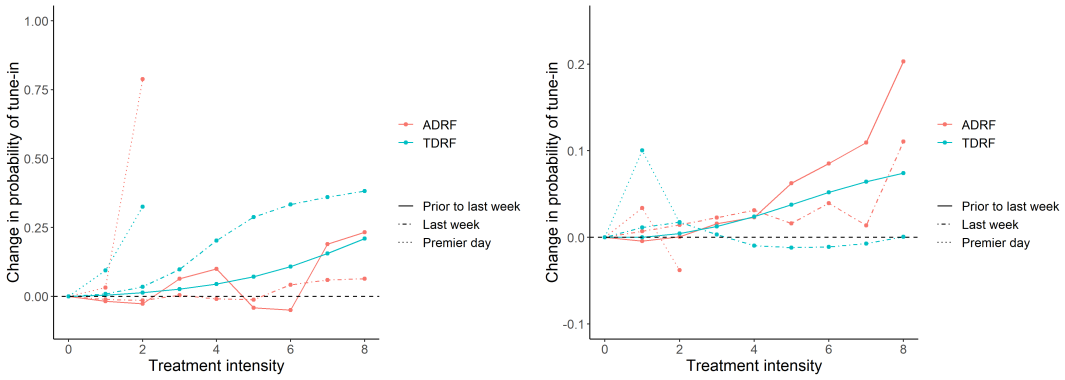
	Prior to last week		Last week		Premier day	
	SC	SCM	SC	SCM	SC	SCM
Poisson regression (Poisson)	0.002	0.000	0.002	0.000	0.006	0.000
	(0.003)	(0.000)	(0.001)	(0.000)	(0.011)	(0.000)
Boosting	0.004	0.001	0.001	0.001	0.000	0.000
	(0.002)	(0.000)	(0.001)	(0.000)	(0.008)	(0.001)

Standard errors are given in parentheses, computed using bootstrap variance estimator. Effects significant at 5% significance level are marked with (*).

The dose-response functions using SC and SCM for the different recencies are presented in Figure 9. The figure shows the ADRF and TDRF for *Prior to last week*, *Last week* and *Premier day* combined. We focus on this representation as it allows for a direct comparison between treatments and present the curves per method including error bars in Appendix 9.5.3 (same as for continuous treatments). The dose-response functions for SC and SCM differ just like the estimated ATE. The curves using the SC show an extremely large impact for *Premier day*. The SCM, on the other hand, shows a negative impact for two exposures of *Premier day*. Both findings are unrealistic and this is probably due to the fact that we have very few individuals with this number of exposures in our dataset. Furthermore, SC generally shows a higher potential (i.e. maximum change in probability of tune-in) than SCM. For the low range of treatment values, with many observations, the SC and SCM do give similar results.

If we use SC for multivariate treatments, we experience the problem of no variation in treatment within subclasses. Hence, these individuals cannot be taken into account constructing the curves. This is not a problem in SCM. Hence, we consider the curves obtained using SC less trustworthy than the curves obtained using SCM. We therefore choose to focus on the results of the latter.

For low treatment exposures, we see that exposures on *Premier day* are most impactful, followed by *Last week* and *Prior to last week*. For each of the treatment variables, the TDRF is larger than the ADRF in this range. This is also the case for the curves estimated using *Boosting* (included in Appendix 9.5.3). For high treatment exposures *Prior to last week* results in the biggest increase in the probability of tune-in. Furthermore, we then find the ADRF is larger than TDRF for *Prior to last week* and *Last week*.



SC (Poisson). SCM (Poisson).

Figure 9: Dose-response functions for multivariate treatment variable (a) *Prior to last week*, *Last week* and *Premier day* on tune-in AGT. Note these figures have a different y-scale.

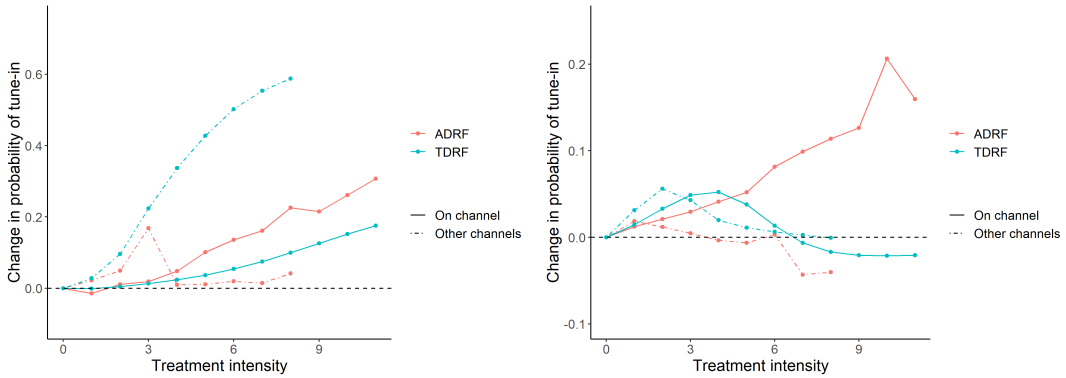
Now, we dive into the influence of the channel choice on the effectiveness of advertising. In Table 14 we present the estimated causal effects of *On channel* and *Other channels*. The estimated ATE are somewhat larger for *Other channels* than for *On channel*, but also more uncertain. We find a positive significant ATE of 0.4% for *On channel* using SC. The ATE of *Other channels* is 0.9% (not significant). The ATE estimated using the SCM are 0.1% and 0.0% for *On channel* and *Other channels*.

Table 14: Causal effects of multivariate treatment variable (b) based on channel on tune-in AGT.

	On channel		Other channel	
	SC	SCM	SC	SCM
Poisson regression (Poisson)	0.004*	0.001	0.009	0.000
	(0.002)	(0.000)	(0.024)	(0.000)
Boosting	0.004	0.002*	0.011	0.000
	(0.002)	(0.000)	(0.026)	(0.000)

Standard errors are given in parentheses, computed using bootstrap variance estimator. Effects significant at 5% significance level are marked with (*).

The dose-response functions using SC and SCM for the different channel choices are presented in Figure 10. We again focus on the curves obtained using SCM. We see *Other channels* has mainly potential for low treatment intensities. For higher treatment values, *On channel* has more impact on the probability of tune-in. Furthermore, the TDRF is higher than ADRF for low treatment exposures of *On channel*, but lower for exposures larger than four. The TDRF is always larger than ADRF for *Other channels*. The curves estimated using *Boosting* are again included in Appendix 9.5.3. The SCM curves using *Boosting* clearly show *On channel* has more effect than *Other channels*.



SC (Poisson). SCM (Poisson).

Figure 10: Dose-response functions for multivariate treatment variable (b) *On channel* and *Other channels* on tune-in AGT. Note these figures have a different y-scale.

To conclude, we find a positive significant ATE when an individual is exposed to an additional treatment unit *On channel*. The remaining ATE were not significantly different from zero. We find that the estimated curves differ, but predominantly in the range of treatment values with relatively few observations. Regarding the influence of the recency of advertisements, the estimates and curves suggest the exposures on *Premier day* are most impactful, followed by *Last week* and *Prior to last week* for low treatment exposures. Furthermore, the TDRF is higher than the ADRF for these treatment variables in this range, which means the advertisements are especially effective for the targeted audience. However, the order of ADRF and TDRF reverses for high treatment values. For the channel choice, we find some exposures of *Other channels* can be effective, but *On channel* has a much higher potential for higher number of exposures. The TDRF is again higher than the ADRF for low treatment values, but ADRF is higher than TDRF for high exposures of *On channel*.

7 Discussion

This thesis examines the use of propensity score methods (PSM) to estimate causal effects for generalized treatments in a marketing context. When studying other practical applications, we find researchers sometimes leave out crucial parts of information (e.g. how variables are chosen, the form of the variables included, if trimming/truncation is applied). We hope to contribute to future applications by giving a transparant application of all relevant steps. Furthermore, we explore the boundaries of PSMs for generalized treatments. In particular, we introduce a combination of dose-response functions (ADRF and TDRF), investigate estimating treatment effects for a low dimensional multivariate (instead of bivariate) treatment variable and extend the SCM to multivariate treatments.

In particular, we investigated the effect of different types of exposure to advertising on tune-in for the season premiere of AGT. We find small treatment effects, depending on the treatment variable(s) used. Additional exposures to advertising have a positive impact on the probability of tune-in. Furthermore, the results suggest more recent advertisements have a higher impact and advertisements on the same channel (NBC) are most effective. Moreover, compared to the conventional regression-based approach we find slightly smaller, but not substantially different estimated effects[13]. We argue, however, that the PSMs presented in this thesis allow for a more transparent and open discussion about the performance of the methods and consequently, the correctness of the causal claim.

We explore the use of many different methods and highlight their advantages and disadvantages. Taking into account the size of propensity scores, covariate balance and common support, we conclude *CBPS* is preferred as treatment assignment model for binary treatments and *Poisson regression* for continuous/multivariate treatments. Furthermore, we find the treatment assignment model removed some, but not all, initial imbalance in the sample. However, we emphasise we cannot draw any general conclusions based on the performance of the treatment assignment models here. This will be dependent on the particular problem instance. We subsequently use different response models to estimate the desired causal effects. Hereby, using double robust methods is advisable, but it is still no guarantee that the estimates will be correct. The investigated methods often result in different findings. Although the results are conflicting on some aspects, the main findings are similar and in line with expectations.

Furthermore, we find that some of the methods are very dependent on model specification. Both the variable selection and the type of variables included have a large influence on the estimation results in all steps of PSMs. Standard statistical variable selection algorithms for treatment assignment models fail, because the objective of these methods is not to optimize the covariate balance. To estimate causal effects, excellent domain knowledge is therefore required. We need to know which factors influence the treatment and outcome variables. This will allow the researcher to specify a better treatment assignment model, resulting in estimates closer to the true treatment effects.

There are two important data limitations in this study when using PSMs. First of all, we observe a low proportion of treated individuals in comparison to non-treated individuals. This is typical for this problem setting, where we measure the exposures of individuals for

[13]The estimated results using conventional regression analysis and a short comparison to the results obtained using PSMs can be found in Appendix 9.3.

different types of advertising. Hence, this is a general problem related to applying these methods to study promotion response. The large number of zeros makes treatment effect estimation more difficult and the presented dose-response curves less reliable on higher parts of the interval. For the methods to give reliable results, the proportion of treated versus non-treated should not be too unbalanced and we need enough variation in the non-zero exposures. Second, as mentioned earlier, variables included in the treatment assignment model should be measured before treatment assignment. However, in the current dataset, the variables measuring general viewing behaviour are measured simultaneously with treatment. Although these variables influence treatment assignment, it would be better to use information from the period before. Now people who seldom watch TV are never exposed to advertisements, whereas in fact one would like there to be a small chance. This is something that should be taken into account when collecting the data and could not be changed anymore for this study. However, this should be taken into account for future applications.

Furthermore, there are some restrictions to the methods employed in this thesis. First, PSMs only correct for selection bias based on observed confounders. Hence, it is important that we believe the assumption of unconfoundedness holds. We use a very commonly used method of sensitivity analysis to check this assumption, but this is not tailored to weighting methods. Furthermore, it can only be used in the binary treatment case. Extending this sensitivity analysis to make it more complete can add to the trustworthiness of the results. Second, we only focus on marginal covariate balance and not on the balance of higher order covariates. Ideally, one also considers joint distributions of variables as the entire covariate distribution should be the same across treatment and non-treatment groups. The distribution of continuous covariates can for example be checked using quantile-quantile plots or side-by-side empirical density plots. Third, we do not investigate the number of subclasses for SC or the choice of dimension for the SCM. We do find that SCM is sensitive to the choice of dimension, just like SC is sensitive to the number of subclasses. Examining the optimal number of subclasses and dimensions in this context is a topic for further investigation.

Finally, we have some recommendations for further research. First, we would like to encourage researchers to develop new methods and improve existing methods for generalized treatments. More promising methods that are available for binary treatments, such as entropy balancing of Hainmueller (2012), should be extended to continuous treatments. Moreover, methods available for continuous treatments should be adapted to non-normal distributions (e.g. CBGPS with Poisson distribution) and non-parametric methods should be improved. Second, it should be investigated how treatment effects for high dimensional multivariate treatment variables can be estimated using PSMs. The analysis presented in this thesis only works for low dimensional multivariate treatment variables as there is a limit to the number of subclasses (or dimensions) one can create. However, in practice there is a large number of marketing channels and such a high dimensional strategy is necessary. Third, as it is difficult to assess the validity of the causal claim, it is interesting to investigate how simulation studies can help practical applications. If one can closely resemble the real problem setting in a simulation (e.g. by generating the outcome variable using the real explanatory variables and an artificial data generating process), this can potentially be used as guidance to find out which estimates are closest to the true treatment effects.

To conclude, we encourage researchers in the marketing domain to use PSMs as a tool to evaluate the effectiveness of marketing campaigns. Although methods for generalized treatments still need to be developed further to be applicable to all problem settings, the big advantage of PSMs is that the model checks are independent of the outcome variable. Moreover, these checks are easy to interpret, which enables a more transparent and open discussion of the results. For trustworthy estimates, we emphasise problem understanding and the quality of the data is key.

8 References

Aakvik, A., Heckman, J. J., & Vytlacil, E. J. (2005). Estimating treatment effects for discrete outcomes when responses to treatment vary: An application to Norwegian vocational rehabilitation programs. *Journal of Econometrics*, *125*(1-2), 15–51.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, *46*(3), 399–424.

Austin, P. C. (2016). Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Statistics in medicine*, *35*(30), 5642–5655.

Austin, P. C. (2018). Assessing covariate balance when using the generalized propensity score with quantitative or continuous exposures. *Statistical methods in medical research*, 1974–1894.

Austin, P. C. & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, *34*(28), 3661–3679.

Caliendo, M. & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, *22*(1), 31–72.

Fan, J., Imai, K., Liu, H., Ning, Y., & Yang, X. (2016). *Improving covariate balancing propensity score: A doubly robust and efficient approach*. Technical report, Princeton University.

Fong, C., Hazlett, C., Imai, K., et al. (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, *12*(1), 156–177.

Fong, C., Ratkovic, M., Imai, K., & Hazlett, C. (2019). Package 'cbps'. *R package version*, 0.20.

Glynn, A. N. & Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political analysis*, *18*(1), 36–56.

Graham, B. S., de Xavier Pinto, C. C., & Egel, D. (2012). Inverse probability tilting for moment condition models with missing data. *The Review of Economic Studies*, *79*(3), 1053–1079.

Greenwell, B., Boehmke, B., & Cunningham, J. (2019). Package 'gbm'. *R package version*, 2.1.5.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, *20*(1), 25–46.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of economic and social measurement, volume 5, number 4* (pp. 475–492). NBER.

Heckman, J. J., Ichimura, H., Smith, J., & Todd, P. (1998). *Characterizing selection bias using experimental data*. National bureau of economic research.

Heckman, J. J., Tobias, J. L., & Vytlacil, E. (2000). *Simple estimators for treatment parameters in a latent variable framework with an application to estimating the returns to schooling*. National Bureau of Economic Research.

Hirano, K. & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology*, *2*(3-4), 259–278.

Hirano, K. & Imbens, G. W. (2004). The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, *226164*, 73–84.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161–1189.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*(396), 945–960.

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, *47*(260), 663–685.

Imai, K. & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(1), 243–263.

Imai, K. & Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, *99*(467), 854–866.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, *87*(3), 706–710.

Joffe, M. M. & Rosenbaum, P. R. (1999). Invited commentary: Propensity scores. *American journal of epidemiology*, *150*(4), 327–333.

Kang, J. D., Schafer, J. L. et al. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, *22*(4), 523–539.

Keele, L. J. (2015). Package 'rbounds'. *R package version*, 2.1.

Kluve, J., Schneider, H., Uhlendorff, A., & Zhao, Z. (2012). Evaluating continuous training programmes by using the generalized propensity score. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *175*(2), 587–617.

Lechner, M. (2002). Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. *Review of Economics and Statistics*, *84*(2), 205–220.

Lechner, M. & Strittmatter, A. (2014). Practical procedures to deal with common support problems in matching estimation. *Econometric Reviews*, *38*(2), 193–207.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in medicine*, *29*(3), 337–346.

Lee, L.-F. (1982). Some approaches to the correction of selectivity bias. *The Review of Economic Studies*, *49*(3), 355–372.

Lee, L.-F. (1983). Generalized econometric models with selectivity. *Econometrica: Journal of the Econometric Society*, 507–512.

Lopez, M. J., Gutman, R. et al. (2017). Estimation of causal effects with multiple treatments: A review and new ideas. *Statistical Science*, *32*(3), 432–454.

Lu, B., Zanutto, E., Hornik, R., & Rosenbaum, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association*, *96*(456), 1245–1253.

Lumley, T. (2019). Package 'survey'. *R package version*, 3.36.

Lunceford, J. K. & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in medicine*, *23*(19), 2937–2960.

Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*. Cambridge university press.

Martens, E. P., Pestman, W. R., de Boer, A., Belitser, S. V., & Klungel, O. H. (2008). Systematic differences in treatment effect estimates between propensity score methods and logistic regression. *International journal of epidemiology*, *37*(5), 1142–1147.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in medicine*, *32*(19), 3388–3414.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, *9*(4), 403.

Mills, P. R., Kessler, R. C., Cooper, J., & Sullivan, S. (2007). Impact of a health promotion program on employee health risks and work productivity. *American Journal of Health Promotion*, *22*(1), 45–53.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. [Translated in Statistical Science (1990)]. *Annals of Agricultural Sciences*, *10*, 1–51.

Nielsen. (2018). The Nielsen CMO report 2018.

Nyman, J. A., Abraham, J. M., Jeffery, M. M., & Barleen, N. A. (2012). The effectiveness of a health promotion program after 3 years: Evidence from the university of minnesota. *Medical Care*, 772–778.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688.

Pearl, J. et al. (2009). Causal inference in statistics: An overview. *Statistics surveys*, *3*, 96–146.

Pearl, J. & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Allen Lane.

Ridgeway, G., Kovalchik, S. A., Griffin, B. A., & Kabeto, M. U. (2015). Propensity score analysis with survey weighted data. *Journal of Causal Inference*, *3*(2), 237–249.

Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. LWW.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, *89*(427), 846–866.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, *82*(398), 387–394.

Rosenbaum, P. R. (2002). Overt bias in observational studies. In *Observational studies* (pp. 71–104). Springer.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rosenbaum, P. R. & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, *79*(387), 516–524.

Rosenbaum, P. R. & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33–38.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford economic papers*, *3*(2), 135–146.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*(469), 322–331.

Rubin, D. B. & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, 249–264.

Rubin, D. B. & Waterman, R. P. (2006). Estimating the causal effects of marketing interventions using propensity score methodology. *Statistical Science*, 206–222.

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, *94*(448), 1096–1120.

Shah, B. R., Laupacis, A., Hux, J. E., & Austin, P. C. (2005). Propensity score methods gave similar results to traditional regression modeling in observational studies: A systematic review. *Journal of clinical epidemiology*, *58*(6), 550–559.

Stürmer, T., Joshi, M., Glynn, R. J., Avorn, J., Rothman, K. J., & Schneeweiss, S. (2006). A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of clinical epidemiology*, *59*(5), 437–e1.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, *97*(3), 661–682.

Thoemmes, F. J. & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate behavioral research*, *46*(1), 90–118.

Wood, S. (2019). Package 'mgcv'. *R package version*, 1.8–28.

Wright, S. (1920). The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. *Proceedings of the National Academy of Sciences*, *6*(6), 320–332.

Yanovitzky, I., Zanutto, E., & Hornik, R. (2005). Estimating causal effects of public health education campaigns using propensity score methodology. *Evaluation and program planning*, *28*(2), 209–220.

Zanutto, E., Lu, B., & Hornik, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics*, *30*(1), 59–73.

Zhao, S., van Dyk, D. A., & Imai, K. (2018). *Propensity-score based methods for causal inference in observational studies with fixed non-binary treatments*.

Zhu, Y., Coffman, D. L., & Ghosh, D. (2015). A boosting algorithm for estimating generalized propensity scores with continuous treatments. *Journal of causal inference*, *3*(1), 25–40.

9 Appendix

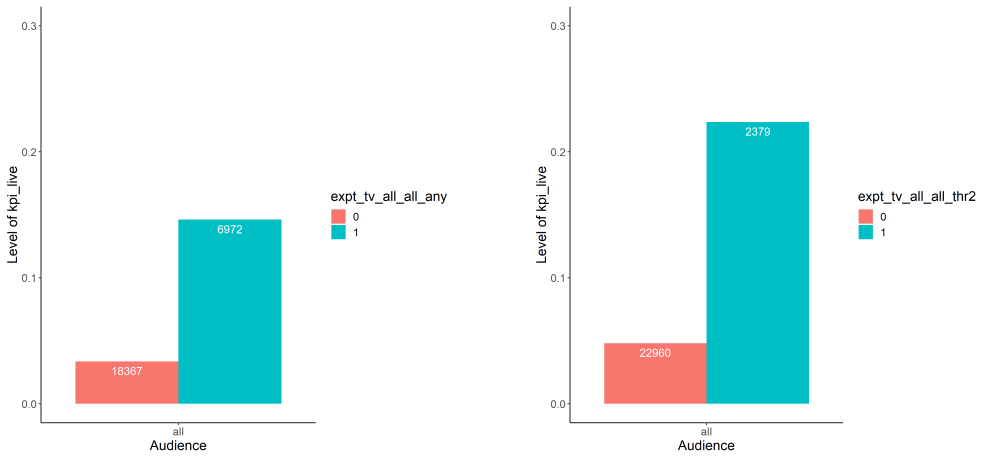
9.1 Variable overview

	Variable name	Definition	Type	Influencing outcome?	Influencing treatment?
Dependent variable	Tune-in	Indicator if the number of minutes watched is at least 1 minute.	Binary	-	-
Sociodemographics	Age groups	Age in years for respondent: [18-27), [27-35), [35-43), [43-49].	Categorical	Yes	Yes
	Male	Indicator if the respondent is male.	Binary	Yes	Yes
	Region	Different regions in the U.S.: East Central, Northeast, Pacific, Southeast, Southwest, West Central.	Categorical	Yes	Yes
	Household size	The size of the household of the respondent: 1-2 persons, 3-4 persons, 5 or more persons.	Categorical	Yes	Yes
	Household income	Income in dollars (x 1000).	Categorical	Yes	Yes
	High school	Indicator if the education level of the respondent is high school or less.	Binary	Yes	Yes
	Race	Ethnic group of respondent: black, white or other.	Categorical	Yes	Yes
	Spanish	Indicator if respondent speaks some or more Spanish.	Binary	Yes	Yes
General viewing behaviour	Primetime on channel	The proportion of time watching NBC during primetime.	Continuous	Yes	Yes
	Primetime TV*	The proportion of time watching TV during primetime.	Continuous	Yes	Yes
	Weekdays TV*	The proportion of time watching TV during weekdays.	Continuous	Yes	Yes
	Genre	The proportion of time watching the same genre as AGT (participation variety).	Continuous		
	Inheritance TV	Indicator if respondent watched TV during the 15 min before the show started.	Binary	Yes	No
	On channel*	The proportion of total time watching NBC.	Continuous	Yes	Yes
	Cross channel	The proportion of total time watching cross channel.	Continuous	Yes	Yes
	Off channel	The proportion of total time watching off channel.	Continuous	Yes	Yes
	Total TV*	The proportion of total time watching TV.	Continuous	Yes	Yes
	Same time TV	The proportion of total time watching TV at weekdays during primetime.	Continuous	Yes	Yes

Treatment variables	Any exposure	Indicator if there has been any exposure to the media campaign.	Binary	Yes	-
	Exposures > 2	Indicator if the total number of exposures to the media campaign exceeds a certain threshold.	Binary	Yes	-
	Total number of exposures	Count variable equal to the total number of exposures to the media campaign.	Continuous	Yes	-
	Number of exposures - prior to last week	Count variable equal to the number of exposures to the media campaign prior to last week.	Continuous	Yes	-
	Number of exposures - last week	Count variable equal to the number of exposures to the media campaign last week.	Continuous	Yes	-
	Number of exposures - premiere day	Count variable equal to the number of exposures to the media campaign on premier day.	Continuous	Yes	-
	Number of exposures - on channel	Count variable equal to the number of exposures to the media campaign on channel.	Continuous	Yes	-
	Number of exposures - other channels	Count variable equal to the number of exposures to the media campaign off channel, cross channel and other channels.	Continuous	Yes	-

Note: continuous variables are modelled as categorical variable constructed by quantiles (use three groups and create a separate group for zero). Furthermore, only one (set of) treatment variable(s) is included at the same time, see Section 4 for more details. We exclude the zero group of some variables to avoid multicollinearity issues (marked with *).

9.2 Exploratory data analysis



(a) Number of exposures > 0. (b) Number of exposures > 2.

Figure 11: Level of tune-in for binary treatment variables (a) and (b).

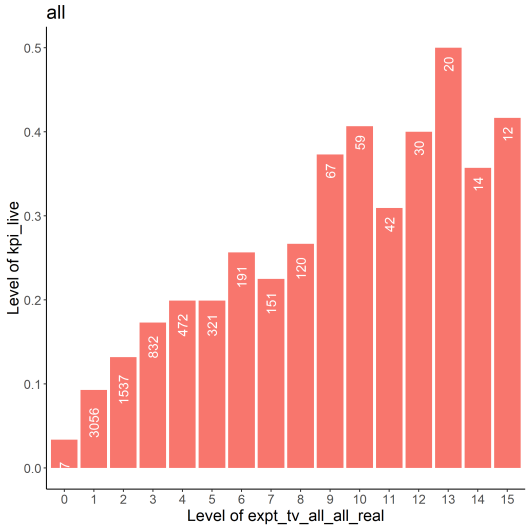


Figure 12: Level of tune-in for continuous treatment variable.

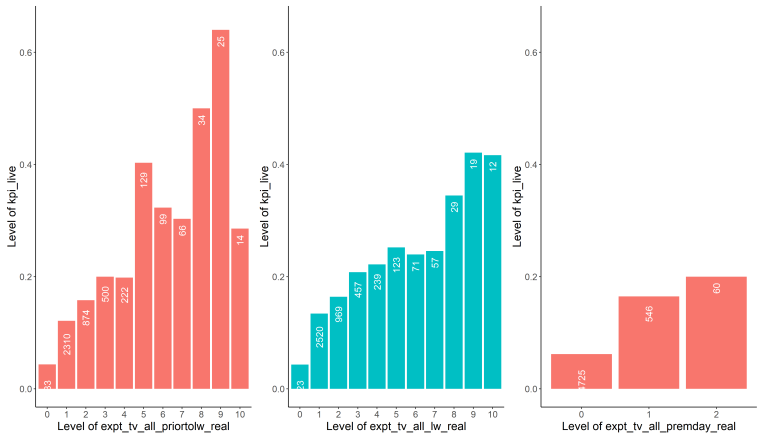


Figure 13: Level of tune-in for multiple treatment variables (a) based on time.

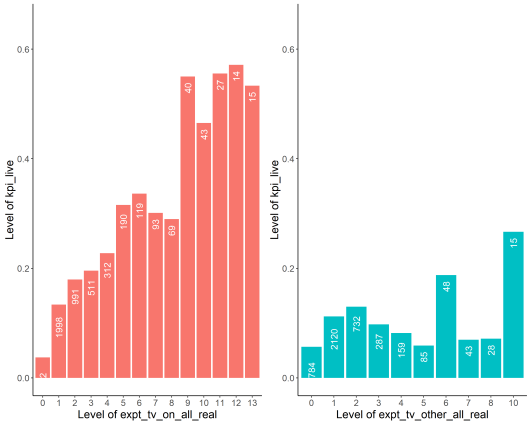


Figure 14: Level of tune-in for multiple treatment variables (b) based on channel.

9.3 Conventional regression analysis

To compare the results of causal inference with standard statistical analysis, we present the estimated effects by two commonly used binary choice models here. The outcome of these models is typically very similar, but not identical. Here, we introduce the notation necessary and discuss the results obtained using standard statistical analysis.

First, we define the logistic regression or logit model:

$$\begin{aligned} \text{logit}(Y_i) = \Lambda^{-1}(Y_i) &= X_i\beta + T_i\gamma, \\ Y_i = \Lambda(X_i\beta + T_i\gamma) &= \frac{\exp(X_i\beta + T_i\gamma)}{1 + \exp(X_i\beta + T_i\gamma)}, \end{aligned}$$

where $\Lambda(.)$ is the logistic distribution function. We approximate the marginal effect of X_j by the average of sample marginal effects:

$$\frac{\partial Y}{\partial T} = \hat{\gamma} \cdot \lambda(X\hat{\beta} + T\hat{\gamma}) \approx \hat{\gamma} \cdot \frac{1}{N} \sum_{i=1}^{N} \lambda(X_i\hat{\beta} + T_i\hat{\gamma}).$$

where $\lambda(.)$ is the probability density function for the logistic distribution.

Second, we define the probit model:

$$Y_i = \Phi(X_i\beta + T_i\gamma),$$

where $\Phi(.)$ is the normal distribution function. We approximate the marginal effect of X_j by the average of sample marginal effects:

$$\frac{\partial Y}{\partial T} = \hat{\gamma} \cdot \phi(X\hat{\beta} + T\hat{\gamma}) \approx \hat{\gamma} \cdot \frac{1}{N} \sum_{i=1}^{N} \phi(X_i\hat{\beta} + T_i\hat{\gamma}).$$

where $\phi(.)$ is the probability density function for the normal distribution.

We use the same selection of variables as for the response model of the PSMs. The results are shown below in Table 15. We see (almost) all estimated coefficients are positive using standard statistical analysis. Not all effects are significantly different from zero, but the estimates for *Exposures > 2*, *Number of exposures*, *Prior to last week* and *On channel* suggest a positive impact on the tune-in of AGT.

These findings are overall quite similar to the results obtained using the PSMs presented in this thesis. However, the estimated effects using the standard statistical methods are somewhat larger than for the PSMs. An exception to this are the estimated effects for the binary treatment variable *Any exposure*, where the reverse is true. The effects are also more often found to be significant. This is especially the case for the estimated effects of both multivariate treatment variables (a, b). The relative performance of the different treatments also slightly differs between the two approaches for the multivariate treatment case.

Table 15: Estimated effects of treatment variables on tune-in AGT.

	UM	LPM	Logit	Probit
Binary - any exposure	0.113*	0.001	0.001	0.000
	(0.004)	(0.007)	(0.005)	(0.005)
Binary - exposures > 2	0.175*	0.052*	0.016*	0.011*
	(0.007)	(0.011)	(0.005)	(0.003)
Continuous - number of exposures	-	0.009*	0.002*	0.002*
		(0.002)	(0.001)	(0.001)
Multiple - prior to last week	-	0.015*	0.004*	0.003*
		(0.003)	(0.001)	(0.001)
Multiple - last week	-	0.005	0.001	0.001
		(0.003)	(0.001)	(0.001)
Multiple - premier day	-	0.005	0.003	0.002
		(0.014)	(0.005)	(0.004)
Multiple - on channel	-	0.014*	0.003*	0.002*
		(0.003)	(0.001)	(0.001)
Multiple - other channels	-	-0.001	0.000	0.000
		(0.003)	(0.002)	(0.001)

Standard errors are given in parentheses, computed using bootstrap variance estimator. Effects significant at 5% significance level are marked with (*).

9.4 Endogenous switching regression

As an alternative to PSMs, one can use an endogenous switching regression model. This technique does not require the assumption that selection takes place on observable variables only. Hence, it also corrects for hidden bias due to unmeasured confounders. However, this comes at the cost of strong distributional assumptions of the error terms and parameter identification can be problematic (Heckman et al., 2000). As we focus on generalized treatments in PSMs, this is outside the scope of this thesis. Here, we give an outline of this method for binary treatments in a similar way as we did for PSMs, as a starting point for further research.

9.4.1 Identification

Heckman corrections correct selection bias by modelling the endogeneity. It models the dependencies in the error term and corrects for the resulting bias explicitly. We make the following assumptions[14]:

- $\epsilon_{ti} \sim N(0, \sigma_t^2)$ for $t \in \{0, 1\}$,
- $\epsilon_{Ti} \sim N(0, \sigma_T^2)$,
- $\sigma_T^2 = 1$ (we normalize the variance parameter to unity for identification).

This results in the following error distribution:

$$\begin{bmatrix} \epsilon_{0i} \\ \epsilon_{1i} \\ \epsilon_{Ti} \end{bmatrix} \sim N(0, \Sigma), \text{ where } \Sigma = \begin{bmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{0T} \\ \sigma_{01} & \sigma_1^2 & \sigma_{1T} \\ \sigma_{0T} & \sigma_{1T} & 1 \end{bmatrix}.$$

We call this an endogenous switching model as we know ϵ_T is possibly correlated with ϵ_{ti} for $t \in \{0, 1\}$. Hence, $\sigma_{0T} \neq 0$ and $\sigma_{1T} \neq 0$ (equal would be exogenous switching model). We now discuss two methods to estimate $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\theta}$, $\hat{\sigma}_0^2$, $\hat{\sigma}_1^2$, $\hat{\sigma}_{0T}$ and $\hat{\sigma}_{1T}$. Note: σ_{01} is not estimatable.

Formally, the model is identified by the above stated normality assumptions. However, identification can be weak if there are only a limited number of observations in the tails where we expect substantial non-linearity in the inverse Mills ratio. Hence, for better identification an exclusion restriction is often used. This means that at least one explanatory variable should be included in the selection equation with a non-zero coefficient that does not appear in the outcome equations. This variable is usually selected guided by economic theory.

9.4.2 Two-step method

Continuous outcome
Most easy and common way to estimate this framework is using the two-stage method initially proposed by Heckman (1976). However, this method has shortcomings for the case of binary observed outcome variables like we have (similar to using a linear probability model instead of a binary choice model). First, the estimated probabilities are

[14] Heckman et al. (2000) show the normal results can be extended to more general, strictly increasing, continuous distribution functions. L.-F. Lee (1982, 1983) allow the error terms to be jointly distributed according to the Student-t distribution with varying degrees of freedom v, which is especially attractive for fat-tailed data.

not explicitly bounded, which may give probabilities smaller than zero or larger than one. Second, the error terms are not normally distributed. Nevertheless, it could be used to compute final estimates or to produce initial values for an iterative maximum likelihood solution.

To derive the bias we need to obtain the expected values of the residuals in Equation (1) and (2). For $t \in \{0,1\}$, let $\eta_{ti} = \epsilon_{ti} - \sigma_{tT}\epsilon_{Ti}$. Rewriting gives $\epsilon_{ti} = \eta_{ti} + \sigma_{tT}\epsilon_{Ti}$. Then we have $E[\eta_{ti}\epsilon_{Ti}] = E[(\epsilon_{ti} - \sigma_{tT}\epsilon_{Ti})\epsilon_{Ti}] = E[\epsilon_{ti}\epsilon_{Ti} - \sigma_{tT}\epsilon_{Ti}^2] = \sigma_{tT} - \sigma_{tT} = 0$. Next,

$$\begin{aligned} E[X_i\beta_1 + \epsilon_{1i}|T_i = 1] &= X_i\beta_1 + E[\epsilon_{1i}|\epsilon_{Ti} > -Z_i\theta] \\ &= X_i\beta_1 + E[\eta_{1i} + \sigma_{1T}\epsilon_{Ti}|\epsilon_{Ti} > -Z_i\theta] \\ &= X_i\beta_1 + \sigma_{1T}E[\epsilon_{Ti}|\epsilon_{Ti} > -Z_i\theta] \\ &= X_i\beta_1 + \sigma_{1T}\frac{\phi(Z_i\theta)}{\Phi(Z_i\theta)}. \end{aligned}$$

Here, $\phi(.)$ and $\Phi(.)$ are the standard normal density and cumulative distribution function respectively as we assume the error terms are independent and normally distributed. Correlation $\epsilon_t, \epsilon_T = \rho_t = \frac{\sigma_{tT}}{\sigma_t\sigma_T}$. Similarly we can find[15]:

$$E[X_i\beta_0 + \epsilon_{0i}|T_i = 0] = X_i\beta_0 - \sigma_{0T}\frac{\phi(Z_i\theta)}{1 - \Phi(Z_i\theta)}.$$

Now estimation goes as follows:

1. In the first step we estimate the bias correction terms.
   Fit a probit model using all N observations to estimate $\hat{\theta}$:

   $$\begin{aligned} \Pr(T_i = 1|X_i) &= \Pr(T_i^* > 0|X_i) = Pr(Z_i\theta + \epsilon_{Ti} > 0|X_i) \\ &= Pr(\epsilon_{Ti} > -Z_i\theta|X_i) = 1 - F(-Z_i\theta) = F(Z_i\theta), \end{aligned}$$

   where F(.) is the normal cumulative distribution function $\Phi(.)$ if we assume the error terms are independent and normally distributed.

   Then we can compute the selection-correction terms evaluated at $\hat{\theta}$ for each individual i: $\frac{\phi(Z_i\hat{\theta})}{\Phi(Z_i\hat{\theta})}$ and $\frac{\phi(Z_i\hat{\theta})}{1-\Phi(Z_i\hat{\theta})}$.

2. Perform regression (OLS) for the subsamples created by treatments separately including the selection-correction terms as additional regressor to estimate $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}_{0T}$ and $\hat{\sigma}_{1T}$:

   $$\begin{aligned} Y_{0i} &= X_i\beta_0 - \sigma_{0T}\frac{\phi(Z_i\hat{\theta})}{1 - \Phi(Z_i\hat{\theta})} + \eta_{0i}, && \text{for } T_i = 0, \\ Y_{1i} &= X_i\beta_1 + \sigma_{1T}\frac{\phi(Z_i\hat{\theta})}{\Phi(Z_i\hat{\theta})} + \eta_{1i}, && \text{for } T_i = 1. \end{aligned}$$

9.4.3 Maximum likelihood estimation

Continuous outcome
The two-step method was initially useful because computers were not very powerfull. However, we know an iterative approach will always be more efficient if the error assumptions

[15] $E[X_i\beta_1 + \epsilon_{1i}|T_i = 0] = X_i\beta_1 - \sigma_{1T}\frac{\phi(Z_i\theta)}{1-\Phi(Z_i\theta)}$ and $E[X_i\beta_0 + \epsilon_{0i}|T_i = 1] = X_i\beta_0 + \sigma_{0T}\frac{\phi(Z_i\theta)}{\Phi(Z_i\theta)}$.

we made are met. Furthermore, we can adapt it to the case of binary observed outcome variables. Therefore, we propose to use this as final estimation method (with the outcome of the two-step method as initial values).

First for continuous observed outcome variables. We know Y_i is a mixed distribution of Y_{0i} and Y_{1i}, depending on the treatment parameter T_i. Hence, the complete data likelihood function is defined as:

$$\begin{aligned}\mathcal{L} &= \prod_{i=1}^{N} \{f(Y_{1i}|T_i=1)\ \text{Prob}(T_i=1)\}^{T_i}\ \{f(Y_{0i}|T_i=0)\ \text{Prob}(T_i=0)\}^{1-T_i},\\ &= \prod_{i=1}^{N} \{f(Y_{1i})\ \text{Prob}(T_i=1|Y_{1i})\}^{T_i}\ \{f(Y_{0i})\ \text{Prob}(T_i=0|Y_{0i})\}^{1-T_i}.\end{aligned}$$

From the above assumptions we know $f(Y_{0i}) = \frac{1}{\sigma_0}\phi(\frac{Y_{0i}-X_i\beta_0}{\sigma_0})$ and $f(Y_{1i}) = \frac{1}{\sigma_1}\phi(\frac{Y_{1i}-X_i\beta_1}{\sigma_1})$, where $\phi(.)$ is the standard normal density function. Furthermore, from the properties of the multivariate normal distribution we know:

$$\epsilon_{Ti}|Y_{1i} \sim N(\frac{\rho_1}{\sigma_1}(Y_{1i}-X_i\beta_1), 1-\rho_1^2),$$

where $\rho_1 = \frac{\sigma_{1T}}{\sigma_1}$. Hence, the following:

$$\begin{aligned}\Pr[T_i=1|Y_{1i}] &= \Pr[\epsilon_{Ti} > -Z_i\theta|Y_{1i}]\\ &= \Pr[\frac{\epsilon_{Ti}-\frac{\rho_1}{\sigma_1}(Y_{1i}-X_i\beta_1)}{\sqrt{1-\rho_1^2}} > \frac{-Z_i\theta-\frac{\rho_1}{\sigma_1}(Y_{1i}-X_i\beta_1)}{\sqrt{1-\rho_1^2}}]\\ &= 1-\Phi(\frac{-Z_i\theta-\frac{\rho_1}{\sigma_1}(Y_{1i}-X_i\beta_1)}{\sqrt{1-\rho_1^2}})\\ &= \Phi(\frac{Z_i\theta+\frac{\rho_1}{\sigma_1}(Y_{1i}-X_i\beta_1)}{\sqrt{1-\rho_1^2}}).\end{aligned}$$

Similarly, we know:

$$\epsilon_{Ti}|Y_{0i} \sim N(\frac{\rho_0}{\sigma_0}(Y_{0i}-X_i\beta_0), 1-\rho_0^2),$$

where $\rho_0 = \frac{\sigma_{0T}}{\sigma_0}$. And:

$$\begin{aligned}\Pr[T_i=0|Y_{0i}] &= \Pr[\epsilon_{Ti} \leq -Z_i\theta|Y_{0i}]\\ &= \Pr[\frac{\epsilon_{Ti}-\frac{\rho_0}{\sigma_0}(Y_{0i}-X_i\beta_0)}{\sqrt{1-\rho_0^2}} \leq \frac{-Z_i\theta-\frac{\rho_0}{\sigma_0}(Y_{0i}-X_i\beta_0)}{\sqrt{1-\rho_0^2}}]\\ &= \Phi(\frac{-Z_i\theta-\frac{\rho_0}{\sigma_0}(Y_{0i}-X_i\beta_0)}{\sqrt{1-\rho_0^2}})\\ &= 1-\Phi(\frac{Z_i\theta+\frac{\rho_0}{\sigma_0}(Y_{0i}-X_i\beta_0)}{\sqrt{1-\rho_0^2}}).\end{aligned}$$

This results in the following log likelihood function:

$$\begin{aligned}\log\mathcal{L} = -\frac{N}{2}\log 2\pi + \sum_{i=1}^{N} T_i\{-\frac{1}{2}(\frac{Y_{1i}-X_i\beta_1}{\sigma_1})^2 - \log(\sigma_1) + \log(\Phi(\frac{Z_i\theta+\frac{\rho_1}{\sigma_1}(Y_{1i}-X_i\beta_1)}{\sqrt{1-\rho_1^2}}))\}+\\ (1-T_i)\{-\frac{1}{2}(\frac{Y_{0i}-X_i\beta_0}{\sigma_0})^2 - \log(\sigma_0) + \log(1-\Phi(\frac{Z_i\theta+\frac{\rho_0}{\sigma_0}(Y_{0i}-X_i\beta_0)}{\sqrt{1-\rho_0^2}}))\}.\end{aligned}$$

To find the maximum likelihood estimates we optimize the log likelihood function. There are many optimization algorithms available, but not all can handle constraints. We suggest to use the limited memory Broyden–Fletcher–Goldfarb–Shanno algorithm for constrained optimization problems (L-BFGS-B).

Initial values of the parameters are needed to start the algorithm. For this we can use the estimates obtained by a two-step estimation method. In the first step, we fit a probit model to the selection equation and in the second step we perform OLS regression for the two regime equations separately including an additional selection-correction term. For more details see Section 9.4.2. Then we maximize the function log $\mathcal{L}$ to get new parameter estimates $\hat{\theta}$, $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}_{0T}$, $\hat{\sigma}_{1T}$, $\hat{\sigma}_0^2$, $\hat{\sigma}_1^2$.

Binary outcome
In this case, two-step estimation is no longer suitable. Instead, we opt we estimate the parameters using Maximum Likelihood estimation. We know Y_i is a mixed distribution of Y_{0i} and Y_{1i}, depending on the treatment parameter T_i. Hence, the likelihood function is defined as:

$$\begin{aligned}\mathcal{L} = \prod_{i=1}^{N} &\{f(Y_{1i}=1, T_i=1)\}^{T_i Y_i} \ \{f(Y_{1i}=0, T_i=1)\}^{T_i(1-Y_i)} \\ &\{f(Y_{1i}=1, T_i=0)\}^{(1-T_i)Y_i} \ \{f(Y_{1i}=0, T_i=0)\}^{(1-T_i)(1-Y_i)}.\end{aligned}$$

where,

$$\begin{aligned}f(Y_{1i}=0, T_i=1) = \Pr(Y_{1i}^* < 0, T_i^* > 0) = \\ \Pr(\epsilon_{1i} < -X_i\beta_1, \epsilon_{Ti} > -Z_i\theta) = \\ \Pr(\epsilon_{1i} < -X_i\beta_1, \epsilon_{Ti} < Z_i\theta) = \\ \Phi_2(-X_i\beta_1, Z_i\theta, -\rho_1).\end{aligned}$$

here $\Phi_2(.)$ is the cumulative bivariate normal normal distribution.

This results in the following log likelihood function:

$$\begin{aligned}\log \mathcal{L} = \sum_{i=1}^{N} &T_i Y_i \log(\Phi_2(X_i\beta_1, Z_i\theta, \rho_1)) + T_i(1-Y_i)\log(\Phi_2(-X_i\beta_1, Z_i\theta, -\rho_1)) + \\ &(1-T_i)Y_i\log(\Phi_2(X_i\beta_0, -Z_i\theta, -\rho_0)) + (1-T_i)(1-Y_i)\log((\Phi_2(-X_i\beta_0, -Z_i\theta, \rho_0)).\end{aligned}$$

9.4.4 Treatment effects

Continuous outcome
For continuous outcome variables, we can estimate ATE and ATT as follows using the estimated parameters (Heckman et al., 2000):

$$\begin{aligned}\hat{\tau}_{ATE} &= \frac{1}{N}\sum_{i=1}^{N} X_i\hat{\beta}_1 - \frac{1}{N}\sum_{i=1}^{N} X_i\hat{\beta}_0, \\ \hat{\tau}_{ATT} &= \frac{1}{N}\sum_{i=1}^{N} X_i\hat{\beta}_1 - \frac{1}{N}\sum_{i=1}^{N} X_i\hat{\beta}_0 + \frac{1}{N}\sum_{i=1}^{N}(\hat{\sigma}_{1T} - \hat{\sigma}_{0T})\frac{\phi(Z_i\hat{\theta})}{\Phi(Z_i\hat{\theta})}.\end{aligned}$$

We know that ATT is bigger than ATE if $\text{Cov}(\epsilon_1 - \epsilon_0, \epsilon_T) = \text{Cov}(\epsilon_1, \epsilon_T) - \text{Cov}(\epsilon_0, \epsilon_T) = \hat{\sigma}_{1T} - \hat{\sigma}_{0T} > 0$. In this case treatment will produce greater benefit under self-selection than under random assignment.

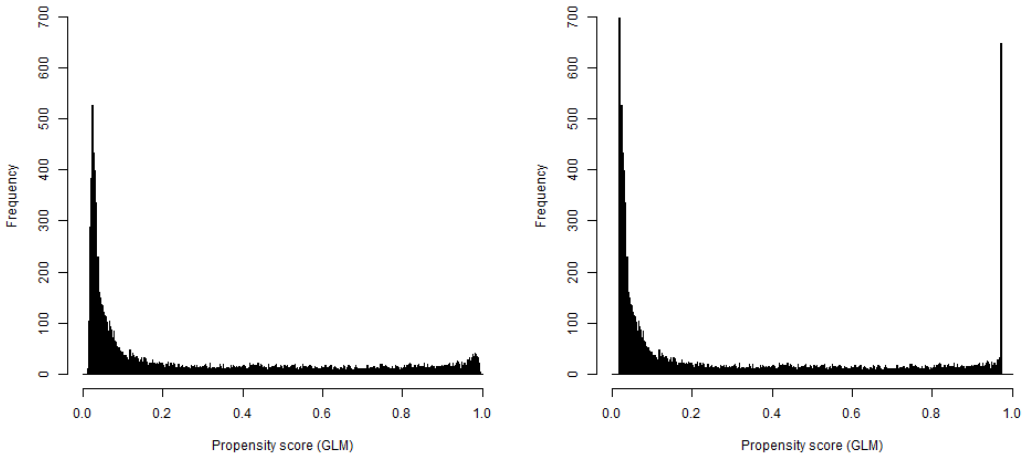
Binary outcome
For binary outcome variables, we can estimate ATE and ATT as follows (Aakvik, Heckman, & Vytlacil, 2005):

$$\begin{aligned}
\hat{\tau}_{ATE} &= \frac{1}{N}\sum_{i=1}^{N}\{\Pr(Y_{1i}=1|X_i) - \Pr(Y_{0i}=1|X_i)\} \\
&= \frac{1}{N}\sum_{i=1}^{N}\{F_{\epsilon_{1i}}(X_i\beta_1) - F_{\epsilon_{0i}}(X_i\beta_0)\}, \\
\hat{\tau}_{ATT} &= \frac{1}{N}\sum_{i=1}^{N}\{\frac{1}{F_{\epsilon_{Ti}(Z_i\theta)}}[F_{\epsilon Ti,\epsilon 1i}(Z_i\theta, X_i\beta_1) - F_{\epsilon Ti,\epsilon 0i}(Z_i\theta, X_i\beta_0)]\}.
\end{aligned}$$

9.5 Output diagnostics

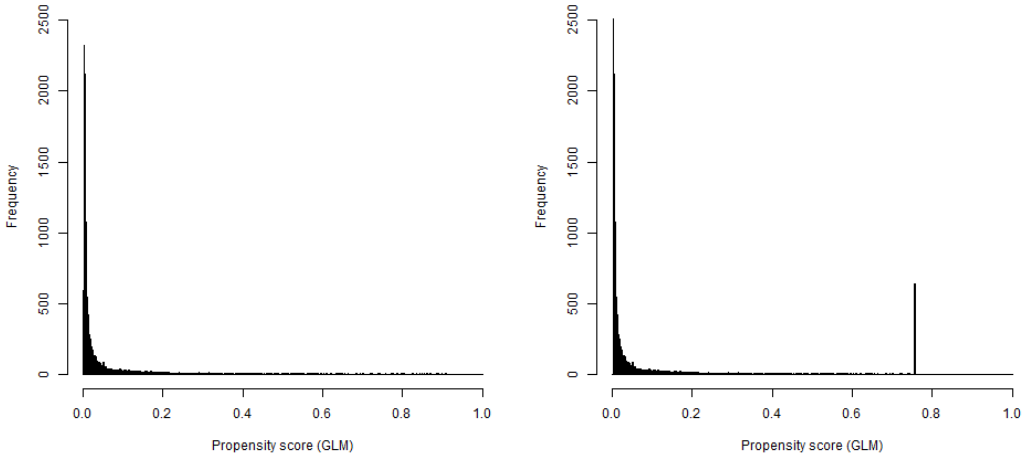
9.5.1 Binary treatments

Propensity scores



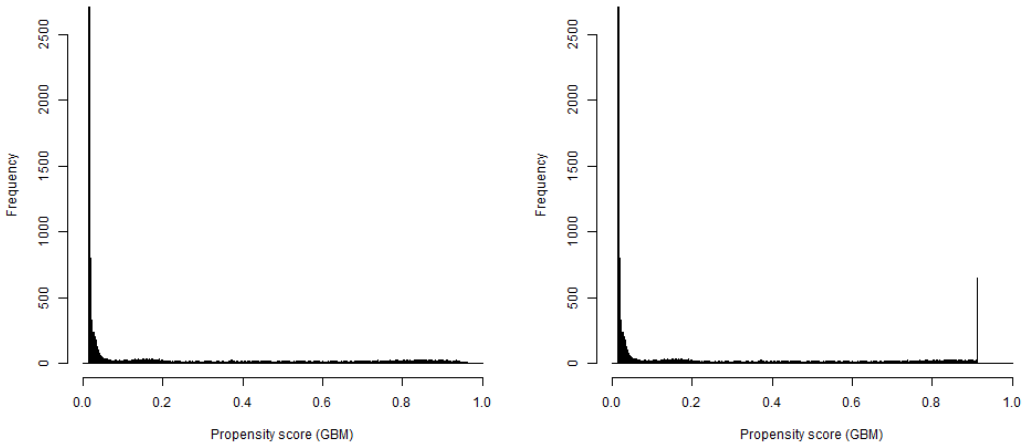
(a) Before truncation. (b) After truncation.

Figure 15: Propensity scores for binary treatment variable (a) *Any exposure* using *Logistic regression.*



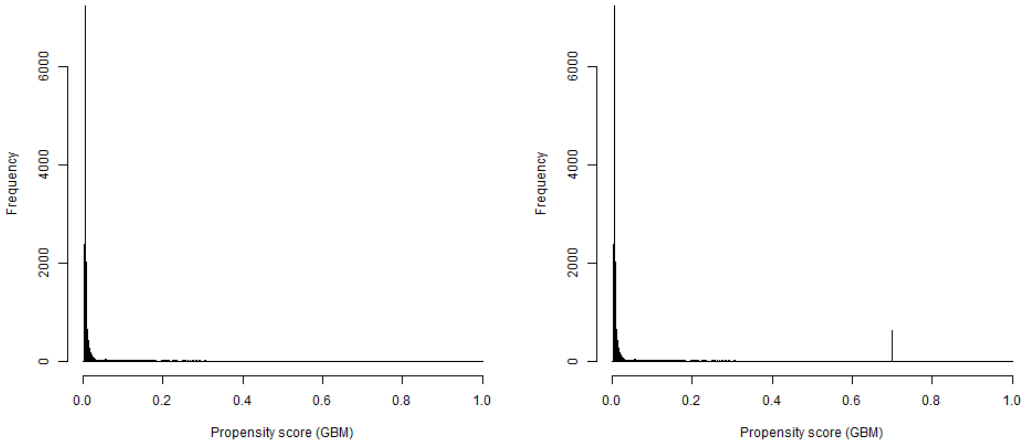
(a) Before truncation. (b) After truncation.

Figure 16: Propensity scores for binary treatment variable (b) *Exposures > 2* using *Logistic regression.*



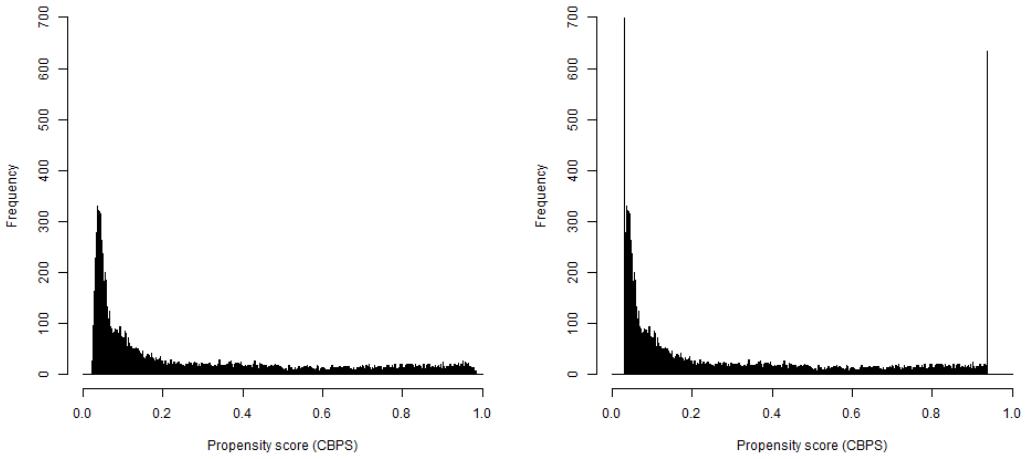
(a) Before truncation. (b) After truncation.

Figure 17: Propensity scores for binary treatment variable (a) *Any exposure* using *Boosting*.



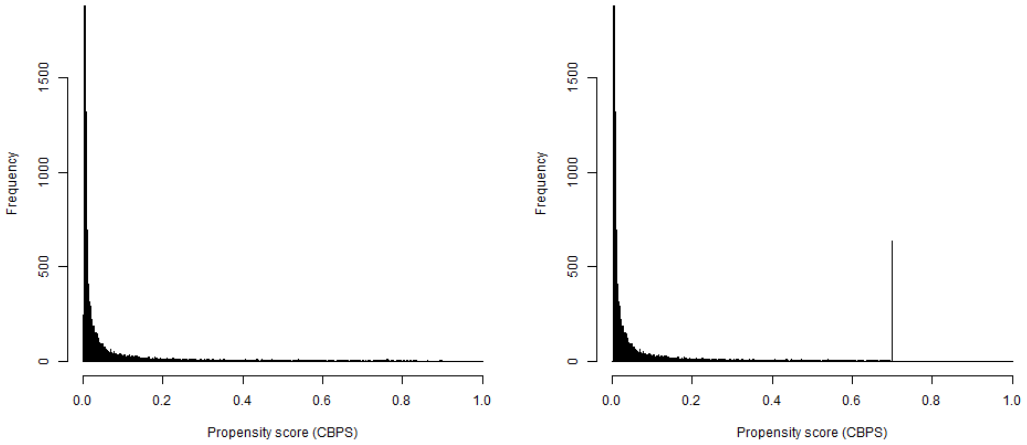
(a) Before truncation. (b) After truncation.

Figure 18: Propensity scores for binary treatment variable (b) *Exposures* $>$ *2* using *Boosting*.



(a) Before truncation. (b) After truncation.

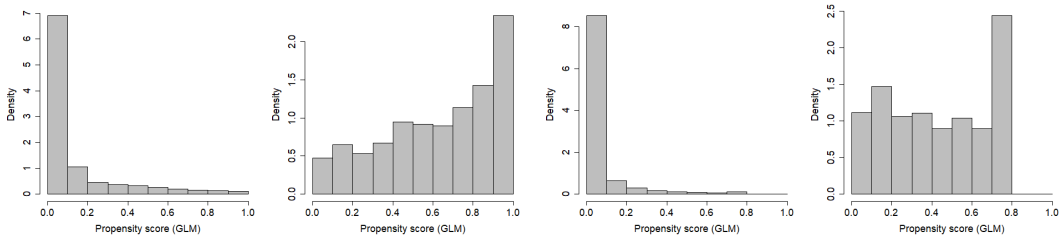
Figure 19: Propensity scores for binary treatment variable (a) *Any exposure* using *CBPS*.



(a) Before truncation. (b) After truncation.

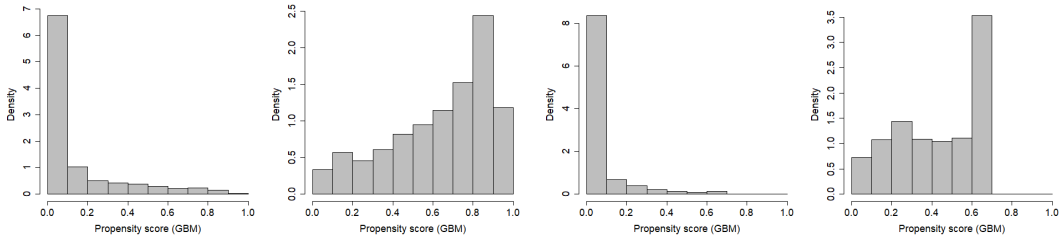
Figure 20: Propensity scores for binary treatment variable (b) *Exposures > 2* using *CBPS*.

Common support



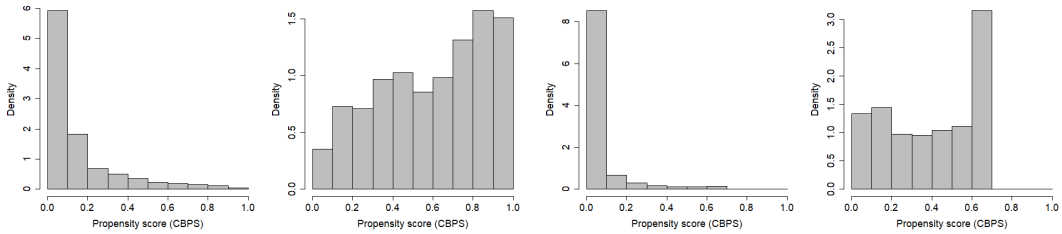
(a) $T_i = 0$. (a) $T_i = 1$. (b) $T_i = 0$. (b) $T_i = 0$.

Figure 21: Common support for binary treatment variable (a) *Any exposure* and b) *Exposure > 2* using *Logistic regression*.



(a) $T_i = 0$. (a) $T_i = 1$. (b) $T_i = 0$. (b) $T_i = 0$.

Figure 22: Common support for binary treatment variable (a) *Any exposure* and b) *Exposure > 2* using *Boosting*.

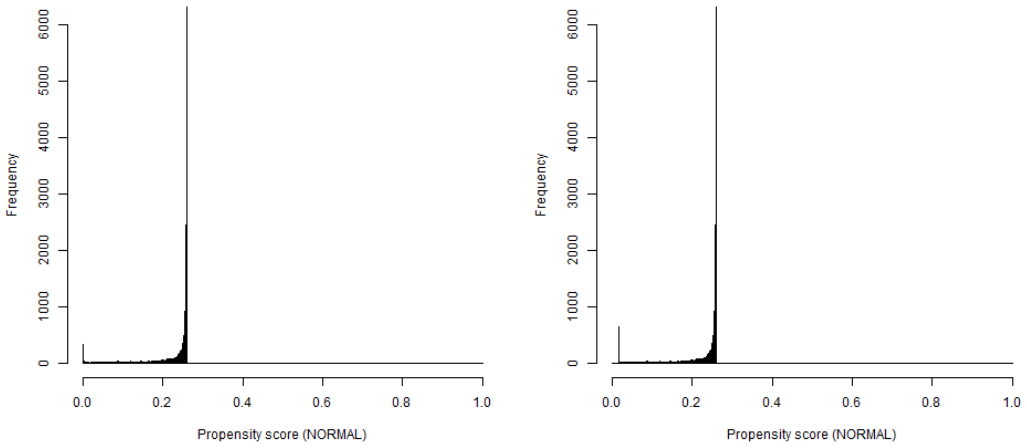


(a) $T_i = 0$. (a) $T_i = 1$. (b) $T_i = 0$. (b) $T_i = 0$.

Figure 23: Common support for binary treatment variable (a) *Any exposure* and b) *Exposure > 2* using *CBPS*.

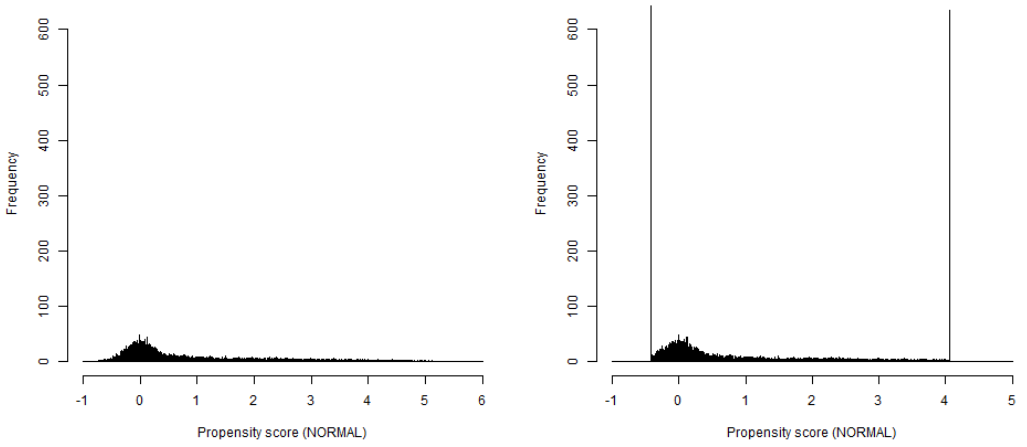
9.5.2 Continuous treatments

Propensity scores



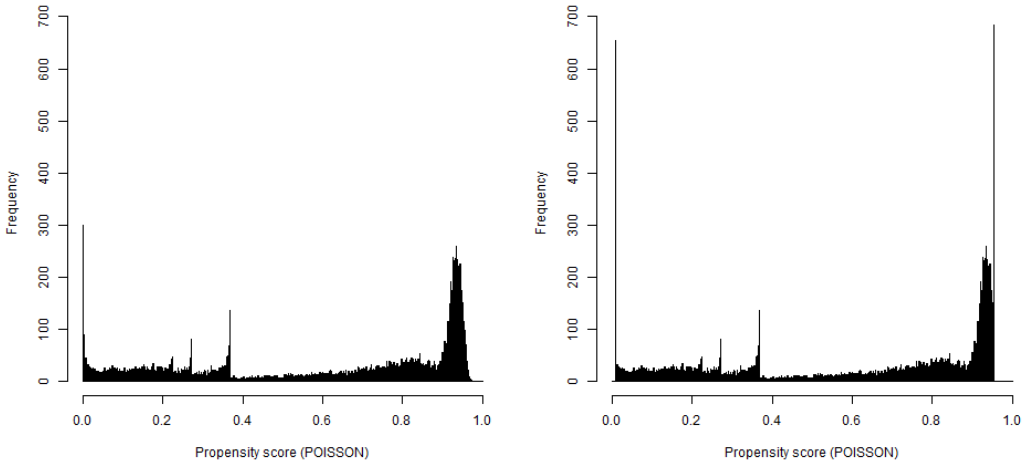
Before truncation. After truncation.

Figure 24: Propensity scores (GPS) for continuous treatment variable *Number of exposures* using *Linear regression*.



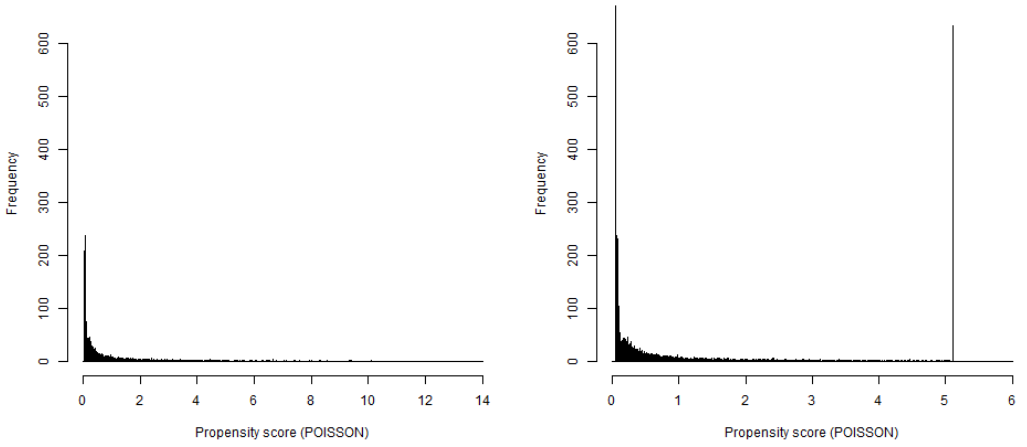
Before truncation. After truncation.

Figure 25: Propensity scores (PF) for continuous treatment variable *Number of exposures* using *Linear regression*.



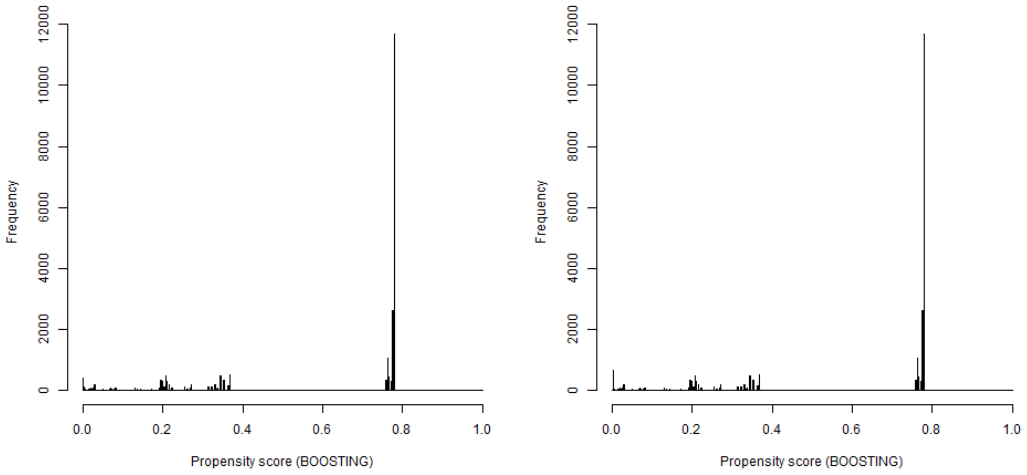
Before truncation. After truncation.

Figure 26: Propensity scores (GPS) for continuous treatment variable *Number of exposures* using *Poisson regression*.



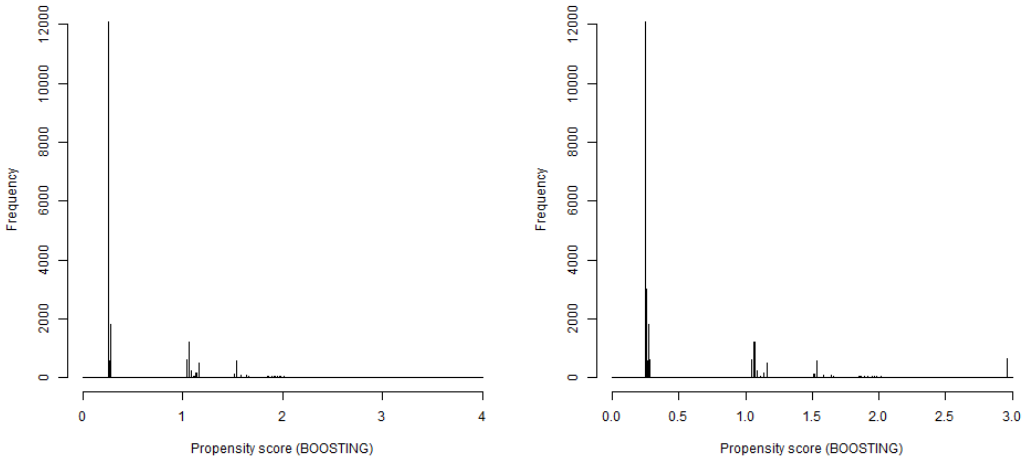
Before truncation. After truncation.

Figure 27: Propensity scores (PF) for continuous treatment variable *Number of exposures* using *Poisson regression*.



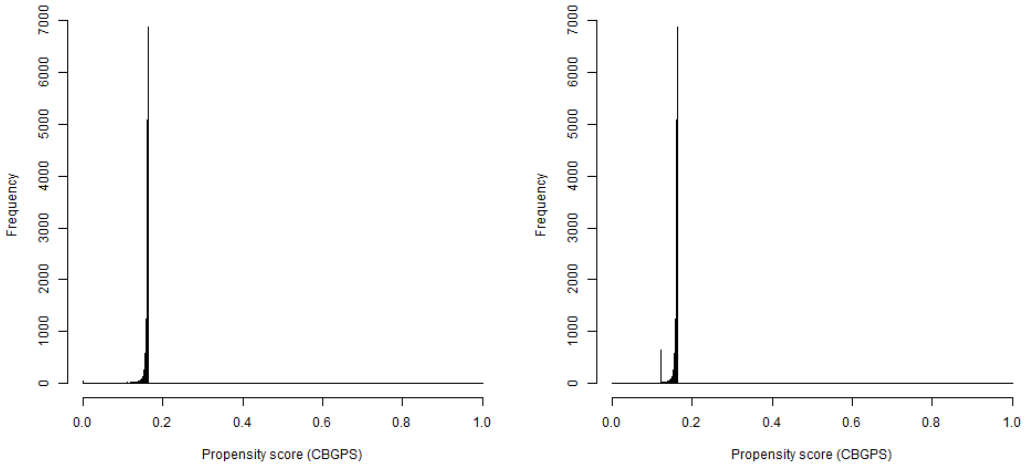
Before truncation. After truncation.

Figure 28: Propensity scores (GPS) for continuous treatment variable *Number of exposures* using *Boosting*.



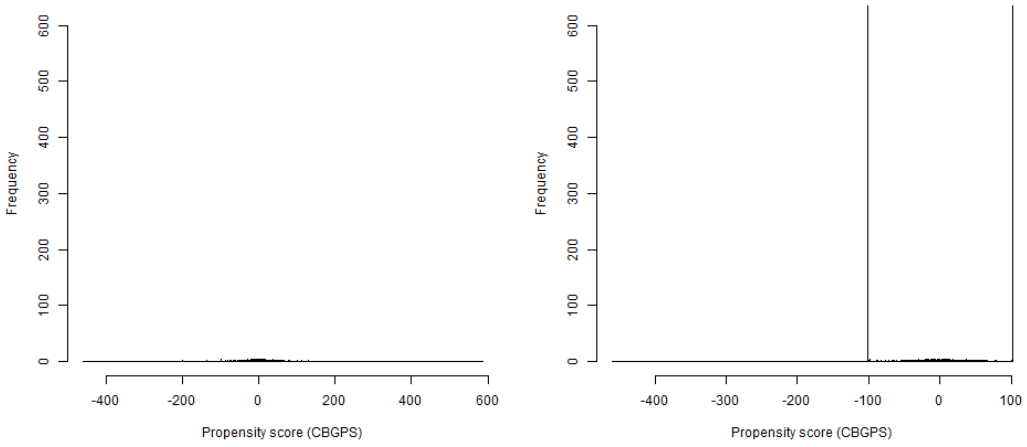
Before truncation. After truncation.

Figure 29: Propensity scores (PF) for continuous treatment variable *Number of exposures* using *Boosting*.



Before truncation. After truncation.

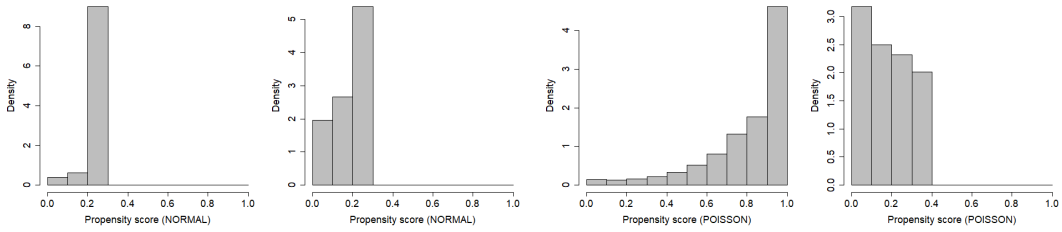
Figure 30: Propensity scores (GPS) for continuous treatment variable *Number of exposures* using *CBGPS*.



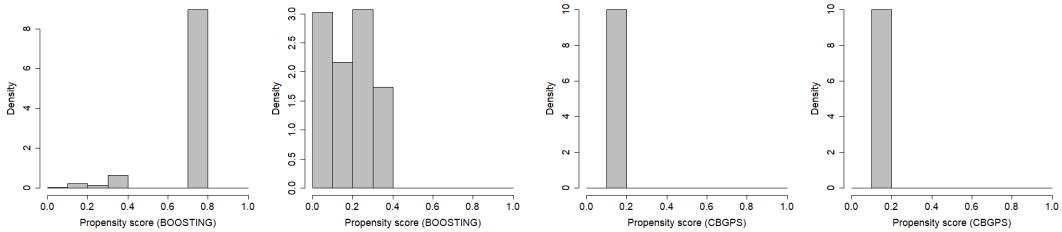
Before truncation. After truncation.

Figure 31: Propensity scores (PF) for continuous treatment variable *Number of exposures* using *CBGPS*.

Common support

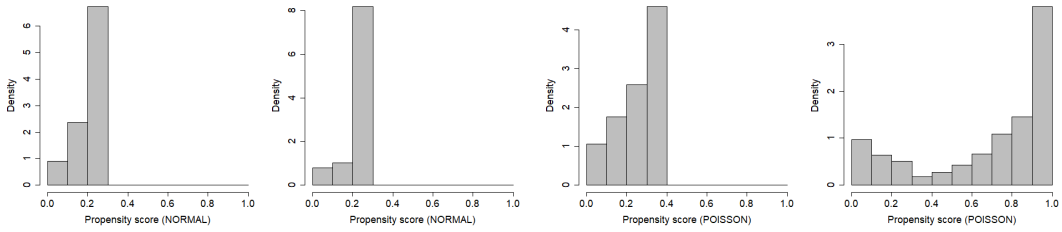


Normal - in k. Normal - not in k. Poisson - in k. Poisson - not in k.

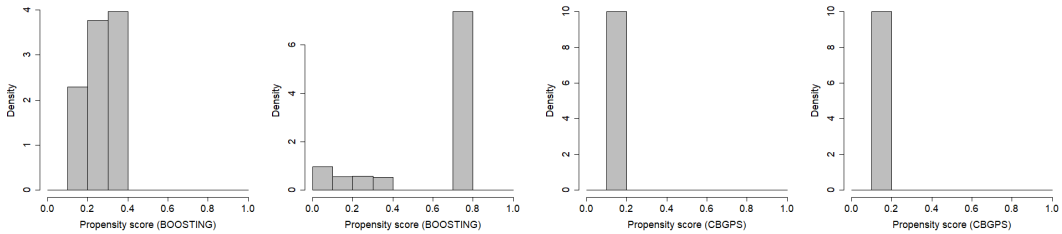


Boosting - in k. Boosting - not in k. CBGPS - in k. CBGPS - not in k.

Figure 32: Common support $k = 1$ for continuous treatment variable *Number of exposures*.

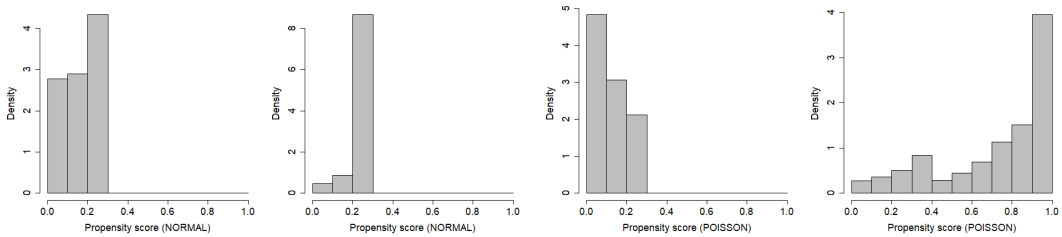


Normal - in k. Normal - not in k. Poisson - in k. Poisson - not in k.

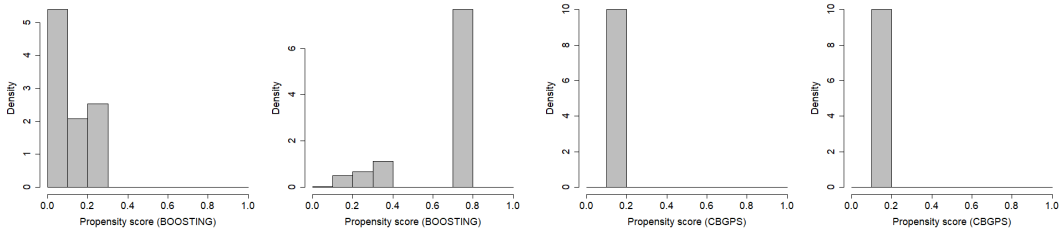


Boosting - in k. Boosting - not in k. CBGPS - in k. CBGPS - not in k.

Figure 33: Common support $k = 2$ for continuous treatment variable *Number of exposures*.



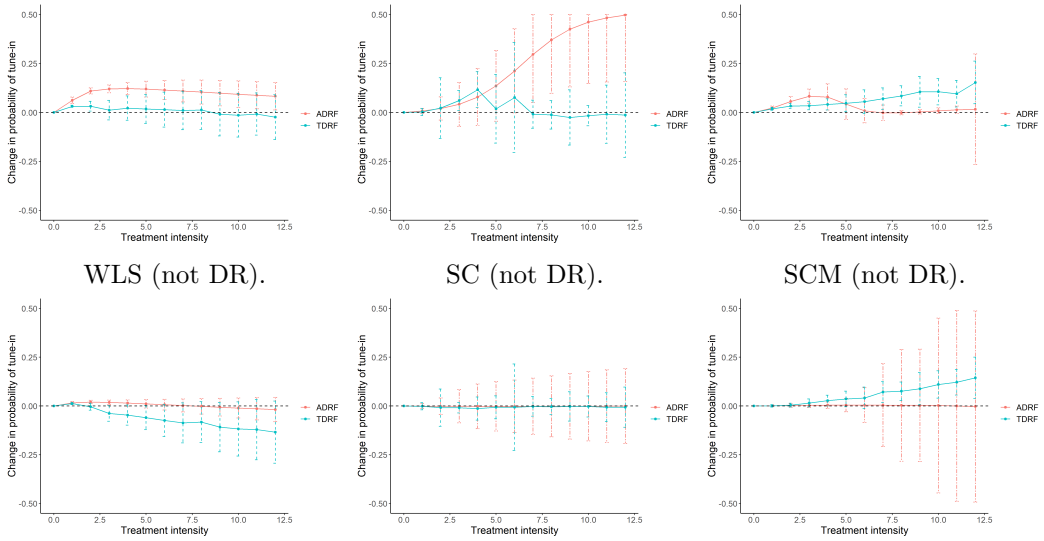
Normal - in k. Normal - not in k. Poisson - in k. Poisson - not in k.



Boosting - in k. Boosting - not in k. CBGPS - in k. CBGPS - not in k.

Figure 34: Common support $k = 3$ for continuous treatment variable *Number of exposures*.

Treatment effects



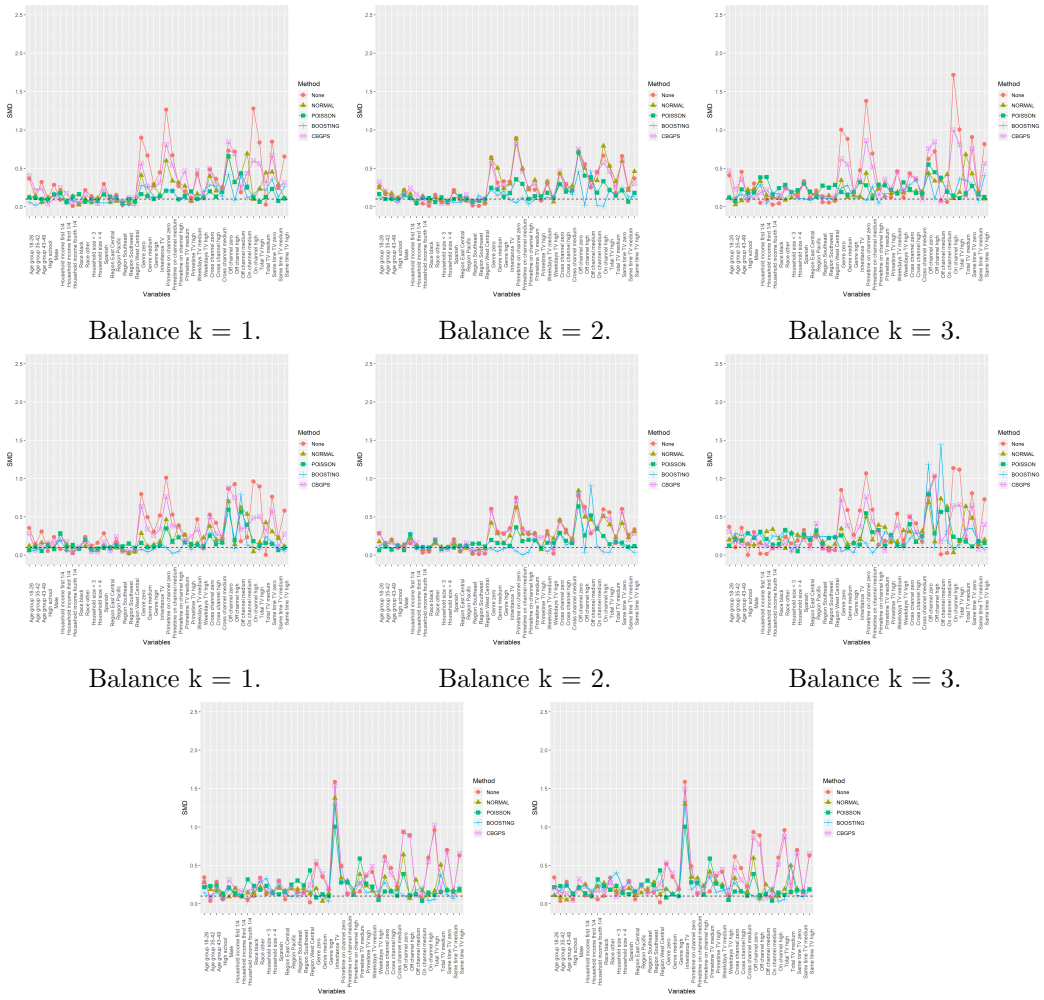
WLS (not DR). SC (not DR). SCM (not DR).

WLS (DR). SC (DR). SCM (DR)

Figure 35: Dose-response functions for continuous treatment variable *Number of exposures* on tune-in AGT using *Boosting*. The error bars represent the 95% confidence interval of the curves. The error bars are cut off if the endings fall outside the figure.

9.5.3 Multivariate treatments

Covariate balance

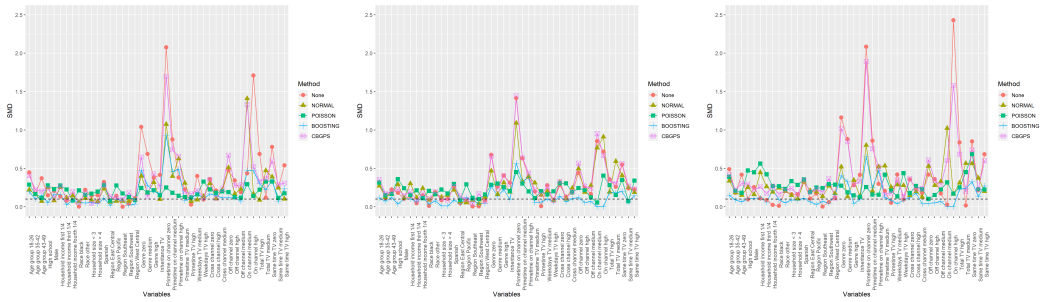


Balance k = 1. Balance k = 2. Balance k = 3.

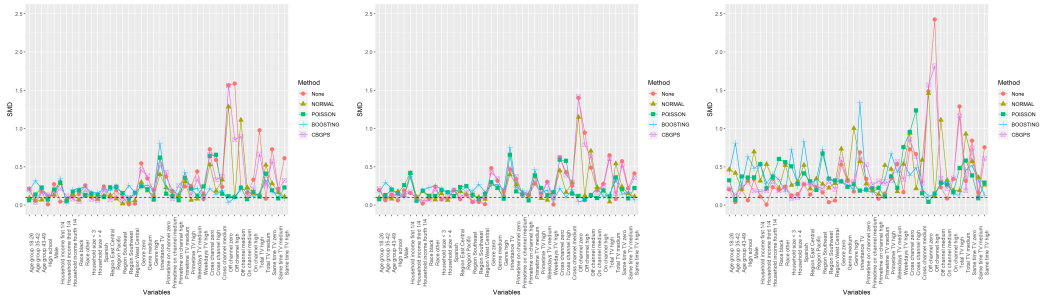
Balance k = 1. Balance k = 2. Balance k = 3.

Balance k = 1. Balance k = 2.

Figure 36: Covariate balance for multivariate treatment variable (a) *Prior to last week*, *Last week*, *Premier day* measured by SMD (from top to bottom). The dotted line indicates SMD = 0.1.



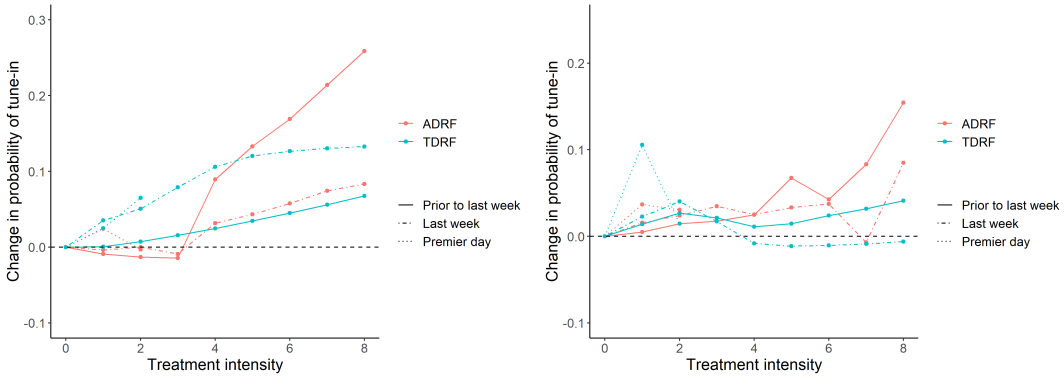
Balance k = 1. Balance k = 2. Balance k = 3.



Balance k = 1. Balance k = 2. Balance k = 3.

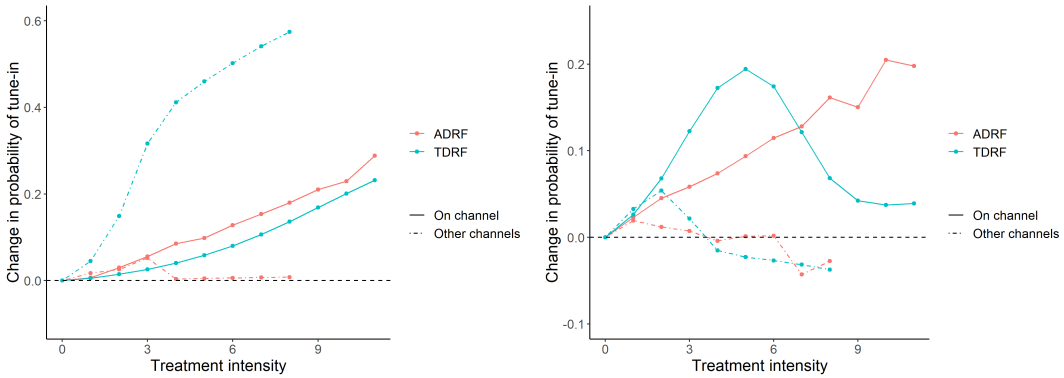
Figure 37: Covariate balance for multivariate treatment variable (b) *On channel* and *Other channel* measured by SMD (top versus bottom). The dotted line indicates SMD = 0.1.

Treatment effects



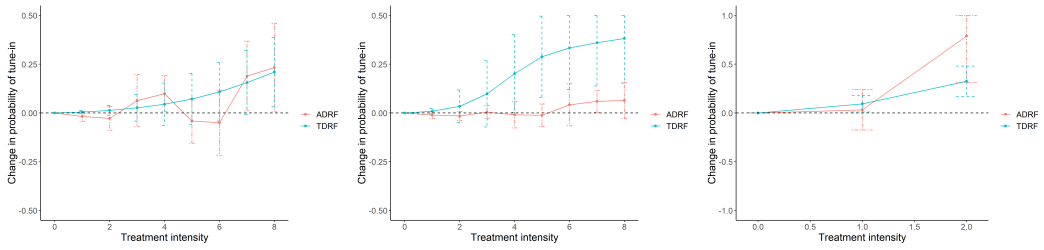
(a) SC (Boosting). (b) SCM (Boosting).

Figure 38: Dose-response functions for multivariate treatment variable (a) *Prior to last week*, *Last week* and *Premier day* on tune-in AGT.

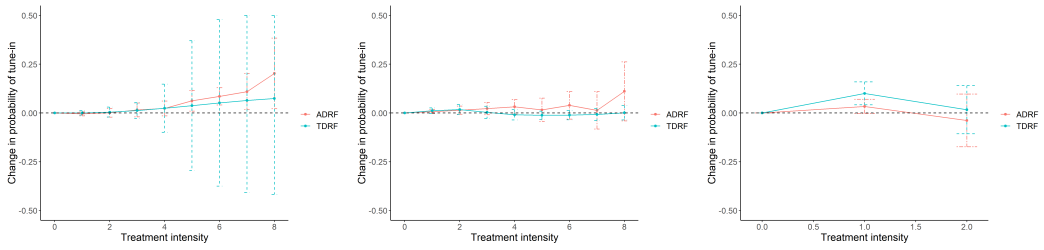


(a) SC (Boosting). (b) SCM (Boosting).

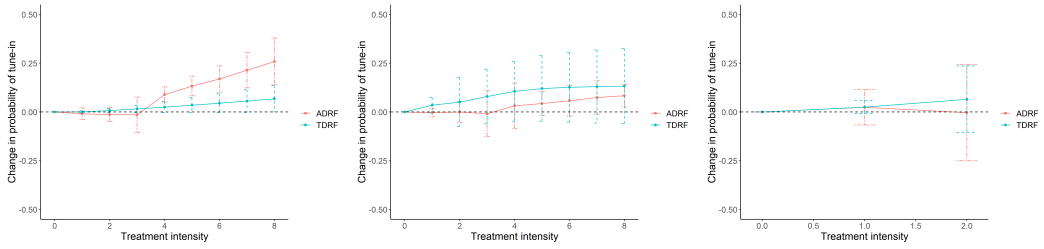
Figure 39: Dose-response functions for multivariate treatment variable (b) *On channel* and *Other channels* on tune-in AGT.



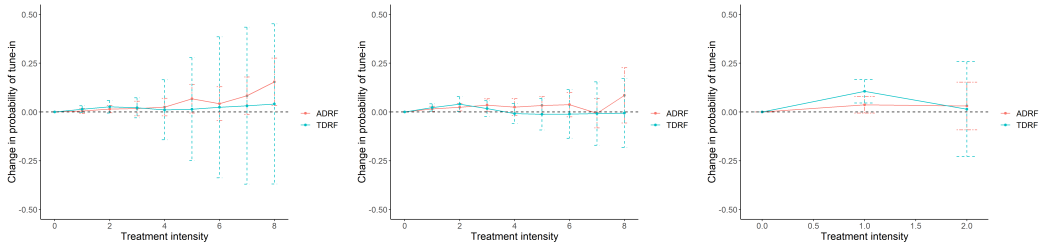
SC (Poisson).



SCM (Poisson).

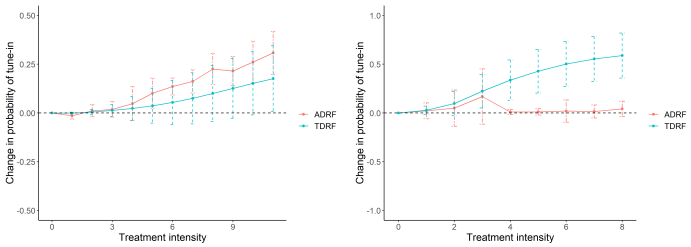


SC (Boosting).

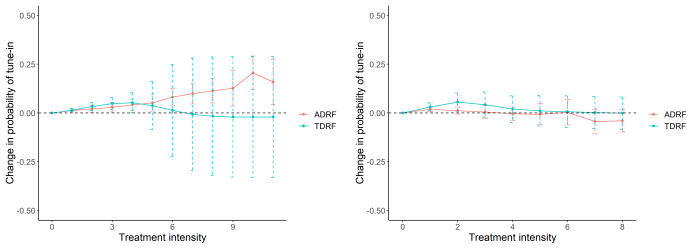


SCM (Boosting).

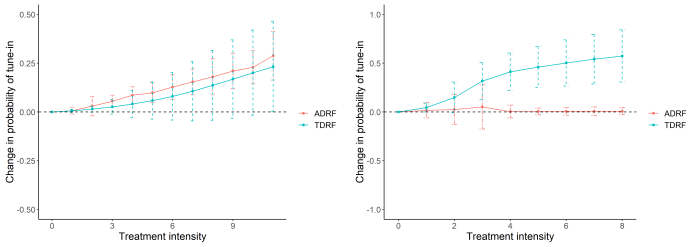
Figure 40: Dose-response functions for multivariate treatment variable (a) *Prior to last week*, *Last week* and *Premier day* on tune-in AGT (from left to right). The error bars represent the 95% confidence interval of the curves. The error bars are cut off if the endings fall outside the figure.



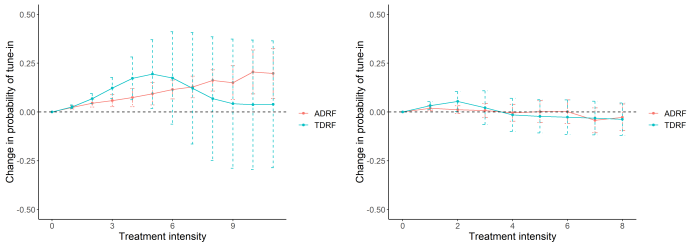
SC (Poisson).



SCM (Poisson).



SC (Boosting).



SCM (Boosting).

Figure 41: Dose-response functions for multivariate treatment variable (b) *On channel* and *Other channel* on tune-in AGT (left versus right). The error bars represent the 95% confidence interval of the curves. The error bars are cut off if the endings fall outside the figure.