

Streszczenie

W niniejszej rozprawie doktorskiej została przedstawiona problematyka estymacji współczynników modeli liniowych w modelowaniu różnicowym. Modelowanie różnicowe zajmuje się przewidywaniem przyczynowych skutków danego działania takiego jak kampania marketingowa czy terapia medyczna na poziomie pojedynczych obserwacji. Aby możliwe było odkrywanie zależności przyczynowych, modele budowane są na danych uzyskanych w badaniach randomizowanych, w których występują dwa zbiory uczące: eksperymentalny i kontrolny. Celem pracy jest uzyskanie możliwie najlepszych estymatorów współczynników liniowego modelu wpływu danego działania na konkretne jednostki.


Pracę można podzielić na trzy części. W pierwszej zawarty jest ogólny opis problemu, założenia modelowe, omówiona jest także problematyka losowego podziału obserwacji na grupę eksperymentalną i kontrolną (randomizacja).

W drugiej, najważniejszej części pracy przedstawione zostały nowe estymatory współczynników regresji w modelowaniu różnicowym. Część ta rozpoczyna się od statystycznej analizy dwóch podstawowych estymatorów stosowanych w regresji różnicowej: estymatora podwójnego i estymatora różnicowego przy założeniu liniowych odpowiedzi i randomizacji blokowej. W tym rozdziale został omówiony także nowy estymator, nazwany estymatorem korygowanym, łączący zalety dwóch wcześniejszych podejść. Następnie przeanalizowane zostało zachowanie estymatorów w przypadku nieliniowych odpowiedzi w grupie kontrolnej i eksperymentalnej. Okazało się, że, o ile wpływ działania pozostaje funkcją liniową, estymatory zachowują wiele pożądaných własności takich jak zgodność. Następnie przedstawione zostały własności opisywanych estymatorów w przypadku stosowania randomizacji prostej.

Kolejny rozdział dotyczy metod regularyzacji modeli różnicowych. Omówiono szereg różnych podejść, różniących się regularyzowanymi wektorami współczynników. W szczególności przedstawiono i przeanalizowano nową metodę regularyzacji pozwalającą na interpolację między estymatorem podwójnym i różnicowym. Dalej przedstawione zostały różnicowe estymatory ściągające. Zaproponowano cztery podejścia powstałe w wyniku zaadaptowania do estymatora podwójnego metod ściągania zwykłego estymatora liniowego.

Trzecia część pracy zawiera opisy i analizę eksperymentów numerycznych, w których potwierdzono teoretyczne wnioski z poprzednich rozdziałów. Ponadto pokazane zostało zachowanie wszystkich metod estymacji na dwóch rzeczywistych zbiorach danych.

Słowa kluczowe: modelowanie różnicowe, randomizacja, model liniowy, zła specyfikacja, regularyzacja, estymatory ściągające.



Abstract

This doctoral dissertation, addresses the issues related to estimating the coefficients of linear models in uplift modeling. Uplift modeling deals with the prediction of causal effects of a given action, such as a marketing campaign or a medical therapy at the level of individual observations. To make discovering causal relationships possible, models are built on data obtained from randomized trials involving two training sets: experimental and control. The purpose of this work is to obtain the best possible estimators of the coefficients of a linear model of the impact of the action on specific individuals.

The work can be divided into three parts. The first one contains a general description of the problem, model assumptions, and a discussion of randomization methods.

In the second, most important part of the work, new estimators of regression coefficients for uplift modeling are presented. This part begins with a statistical analysis of two well known estimators used in uplift regression: the double estimator and the uplift estimator under the assumptions of linear responses and complete (block) randomization. This chapter also introduces a new estimator called the corrected estimator, which combines the advantages of the two previous approaches. Next, the behavior of the estimators in the case of non-linear responses in the control and experimental groups is analyzed. It turned out that, as long as the effect of the action remains a linear function, the estimators retain many desirable properties such as consistency. Then, the properties of the estimators under simple randomization were analyzed.

The next chapter deals with regularization methods for uplift modeling. A number of different approaches were discussed which regularize different vectors of coefficients. In particular, we present and analyze a new method of regularization allowing for smooth interpolation between the double and the uplift estimator. Next, shrinkage uplift estimators are presented. Four such estimators have been obtained by applying classic shrinkage methods to the double regression estimator.

The third part of the work contains descriptions and analysis of the experimental evaluation, which confirmed the theoretical conclusions of the previous chapters. In addition, the behavior of all estimators was tested on two real life datasets.

Keywords: uplift modeling, randomization, linear model, misspecification, regularization, shrinkage estimators.

