WARSAW UNIVERSITY OF TECHNOLOGY

Faculty of Mathematics and Information Science

# Ph.D. Thesis

Mariusz Kubkowski, M.Sc.

Misspecification of binary regression model: properties and inferential procedures

**Supervisor**

Prof. Jan Mielniczuk, Ph.D., D.Sc.

WARSAW, 2019

# Streszczenie

W poniższej rozprawie doktorskiej została przedstawiona problematyka złej specyfikacji modelu regresji binarnej. Pracę możemy podzielić zasadniczo na 4 części. W pierwszej części, którą stanowi Rozdział 1, został zawarty ogólny opis tego problemu oraz przykłady sytuacji, w których zła specyfikacja może wystąpić.

W drugiej części omówiono własności wektora współczynników teoretycznych $\boldsymbol{\beta}^*$ w dopasowanym modelu - wyniki zawarte w tej części stanowią uogólnienie wyników zawartych w pracach Kubkowski, Mielniczuk (2017) (Rozdział 2) oraz Kubkowski, Mielniczuk (2018) (Rozdział 3) do przypadku wypukłej funkcji straty. W Rozdziale 2 zbadano własności nośnika $s^*$ wektora współczynników teoretycznych w dopasowanym modelu w przypadku spełnienia warunku liniowych regresji i w przypadku niespełnienia tego warunku. W Rozdziale 3 jest rozważany ponadto addytywny model binarny.

Trzecia część, składająca się z Rozdziałów 4 i 5, skupia się na estymacji wektora $\boldsymbol{\beta}^*$ oraz zbioru $s^*$ dla losowych predyktorów subgaussowskich (także w przypadku, gdy liczba predyktorów jest większa od liczby obserwacji). W Rozdziale 4 pokazano wyniki dotyczące metody Lasso oparte o idee zawarte w pracach Fan i in. (2014a) oraz Bühlmann, van de Geer (2011). W Rozdziale 5 omówiono minimalizację Uogólnionego Kryterium Informacyjnego (GIC) w pewnej rodzinie $\mathcal{M}$, do której należy $s^*$. W Rozdziale 5 przedstawiono także procedurę dwustopniową SS (Screening - Selection) służącą do znajdowania estymatora $s^*$, która opiera się w swoim działaniu o metodę Lasso (pierwszy etap) i minimalizację GIC (drugi etap). W Rozdziale 5 zaprezentowano także rezultaty teoretyczne dotyczące jej działania.

Czwarta część (Rozdział 6) zawiera opisy i analizę eksperymentów numerycznych, w których zbadano procedury będące modyfikacjami procedury SS dla próby losowej oraz zaprezentowano procedurę numerycznego przybliżenia $\boldsymbol{\beta}^*$ i sprawdzono numerycznie jej działanie.

**Słowa kluczowe:** zła specyfikacja, binarny model regresyjny, regresja logistyczna, Lasso, Uogólnione Kryterium Informacyjne, zbiory aktywnych predyktorów, selekcja zmiennych, regresja wysoko-wymiarowa.

# Abstract

In this doctoral dissertation problem of misspecification of binary regression model is discussed. This dissertation consists of four parts. In the first part, consisting of Chapter 1, general description of this problem and examples of situations, where misspecification occurs, are given.

In the second part, we discuss properties of vector of theoretical coefficients $\boldsymbol{\beta}^*$ in fitted model. Results presented in this part generalize results contained in Kubkowski, Mielniczuk (2017) (Chapter 2) and Kubkowski, Mielniczuk (2018) (Chapter 3) to the case of convex loss function. In Chapter 2 we study properties of support $s^*$ of $\boldsymbol{\beta}^*$ in fitted model in the case when linear regressions condition is satisfied and in the case when this condition is not satisfied. We consider additionally additive binary model in Chapter 3.

In third part, consisting of Chapters 4 and 5, we focus on estimation of vector $\boldsymbol{\beta}^*$ and set $s^*$ for random subgaussian predictors (also in the case when number of predictors is greater than number of observations). In Chapter 4 several novel results concerning Lasso are shown. The results are based on ideas contained in papers of Fan et al (2014a) and Bühlmann, van de Geer (2011). In Chapter 5 minimization of Generalized Information Criterion over family $\mathcal{M}$ (to which $s^*$ belongs) is discussed. In Chapter 5 two-stage SS (Screening - Selection) procedure of finding estimator of $s^*$ is presented and its selection consistency is discussed. The procedure consists of screening based on Lasso in the first stage and GIC minimization in the second stage. In Chapter 5 theoretical results concerning SS procedure are presented.

Fourth part (Chapter 6) contains description and analysis of numerical experiments, in which we study properties of procedures which are modifications of SS procedure. We also present in this chapter procedure approximating $\boldsymbol{\beta}^*$ numerically and we check its performance.

**Key words:** misspecification, binary regression model, logistic regression, Lasso, Generalized Information Criterion, sets of active predictors, variable selection, high-dimensional regression.