prof. dr hab. inż. Małgorzata Bogdan
Instytut Matematyki
Uniwersytet Wrocławski

Wrocław 20.02.2022

RECENZJA ROZPRAWY DOKTORSKIEJ

Tytuł rozprawy: Linear Regression for Uplift Modeling

Autor rozprawy: mgr Krzysztof Rudaś

Promotor rozpawy: prof. dr hab. inż. Szymon Jaroszewicz

Cel i charakter rozprawy

Rozprawa doktorska mgr. Krzysztofa Rudasia dotyczy ważnego problemu estymacji współczynników w modelu różnicowym, który opisuje wpływ interwencji na zadaną zmienną ilościową. Celem jest przewidywanie reakcji danego obiektu na interwencję w zależności od wartości pewnych zmiennych objaśniających. Trudność polega na tym, że obiekty w zbiorze treningowym są obserwowane tylko w jednym z dwóch stanów: bez lub po interwencji. Autor zasadniczo koncentruje się na klasycznym modelu liniowym i analizuje dwa proponowane w literaturze rozwiązania: estymator podwójny, gdzie osobno analizuje się grupę kontrolną i zabiegową, oraz estymator różnicowy, gdzie rozwiązuje się jedno równanie regresji liniowej na odpowiednio przekształconym wektorze odpowiedzi. W wyniku matematycznej analizy tych rozwiązań autor proponuje własną metodę, która poprawia oba powyższe estymatory. Autor podaje asymptotyczne rozkłady tych trzech estymatorów dla zupełnej i prostej randomizacji, a także gdy zmienna odpowiedzi jest nieliniową funkcją zmiennych niezależnych. W tym przypadku konieczne jest jednak założenie, że funkcja odpowiedzi na interwencję jest funkcją liniową. Poza tym autor analizuje estymatory regularyzowane za pomocą regresji grzbietowej, a także za pomocą technik analogicznych do estymatora ściągającego Jamesa-Steina. W obu przypadkach autor proponuje własne rozwiązania o dobrych

własnościach teoretycznych i praktycznych. Częściowe wyniki rozprawy zostały opublikowane w czasopiśmie "Data Mining and Knowledge Discovery" (140 pt MNiSW).

Struktura rozprawy

Rozprawa doktorska napisana jest w języku angielskim. Składa się z 7 rozdziałów.

W pierwszym rozdziale wprowadzono do zagadnienia estymacji parametru różnicowego: zaproponowano odpowiedni model liniowy i wyjaśniono typowe techniki randomizacji.

W rozdziale 2 wprowadzono dwa podstawowe estymatory: podwójny i różnicowy. Udowodniono, że dla ustalonej macierzy eksperymentu X, estymator podwójny jest estymatorem nieobciążonym o minimalnej wariancji. Natomiast w przypadku gdy macierz X jest losowa podano przypadek w którym estymator różnicowy ma mniejszą wariancję. Ten szczególny przypadek wymaga aby współczynniki regresji w grupie kontrolnej i zabiegowej różniły się tylko znakiem, a zasadniczym powodem przewagi estymatora różnicowego jest bardziej efektywne wykorzystanie próby i większa liczba stopni swobody w rozkładzie Wisharta opisującym rozkład macierzy $X'X$. Porównanie rozkładów asymptotycznych obu estymatorów wskazuje, że wspomniany "symetryczny" przypadek jest odosobniony i w innych sytuacjach estymator różnicowy ma większą asymptotyczną wariancję niż estymator podwójny. Ta obserwacja jest punktem wyjścia do konstrukcji autorskiego estymatora, który polega na wstępnej estymacji ważonej średniej współczynników regresji w obu grupach a następnie na redukcji problemu do opisanej wyżej sytuacji symetrycznej poprzez odpowiednie przekształcenie zmiennej odpowiedzi. Twierdzenie 5 pokazuje, że nowo uzyskany estymator ma tą samą asymptotyczną wariancję co estymator podwójny ale spodziewamy się poprawy własności dla małych rozmiarów próby (ponownie ze względu na większą liczbę stopni swobody dla rozkładu macierzy $X'X$).

W rozdziale 3 przeanalizowano rozkłady trzech omówionych powyżej estymatorów w sytuacji gdy w modelu pojawia się człon nieliniowy. Model ten jednak w dalszym ciągu zakłada, że różnica funkcji odpowiedzi w grupie kontrolnej i zabiegowej jest liniową funkcją zmiennych niezależnych. Udowodniono asymptotyczny brak obciążenia i asymptotyczną normalność rozważanych estymatorów i wskazano, że ceną za błędną specyfikację modelu jest zwiększona asymptotyczna wariancja.

W rozdziale 4 przebadano powyższe estymatory w sytuacji prostej randomizacji. Udowodniono, że oba klasyczne estymatory mają zwykle większą wariancję w przypadku prostej randomizacji niż dla randomizacji zupełnej, a wariancja asymptotyczna autorskiego skorygowanego estymatora jest taka sama dla obu przypadków randomizacji.

W rozdziale 5 omówiono wykorzystanie regresji grzbietowej do estymacji parametru różnicowego. Zaproponowano naturalną regularyzację estymatora podwójnego i różnicowego a także autorskie podejście polegające na zastosowaniu osobnej penalizacji dla sumy i różnicy parametrów regresji w obu grupach. Udowodniono, że taki estymator ma mniejszą wariancję niż standardowy estymator podwójny i umożliwia na interpolację między regularyzowanymi wersjami obu standardowych estymatorów.

W rozdziale 6 omówiono zastosowanie regularyzacji typu Jamesa-Steina w kontekście estymacji parametru różnicowego. Zaproponowano dwie bezpośrednie modyfikacje estymatora podwójnego oraz własną konstrukcję, zmierzającą do minimalizacji błędu predykcji. Przy bardzo ograniczających warunkach (ortogonalne macierze planu) udowodniono, że błąd predykcji tego estymatora jest nie większy niż klasycznego estymatora podwójnego.

W rozdziale 7 zamieszczono wyniki symulacji ilustrujących wyniki teoretyczne i analizy danych rzeczywistych. Analizy te pokazują szczególne dobre własności estymatorów uzyskanych z zastosowaniem regularyzacji za pomocą regresji grzbietowej.

Ocena pracy doktorskiej

Rozprawa doktorska mgr. Rudasia dotyczy ważnego zagadnienia. Zawiera kilka interesujących własnych konstrukcji estymatorów parametru różnicowego i szereg twierdzeń i spójnych analiz teoretycznych ułatwiających zrozumienie własności analizowanych estymatorów. Jest napisana z dużą starannością i dbałością o czytelnika. Twierdzenia dotyczą zwykle klasycznej asymptotyki i wykorzystują standardowe techniki, ale ich przeprowadzenie wymagało precyzji i opanowania szerokiej wiedzy ze statystyki. Moje zasadnicza uwaga krytyczna dotyczy bardzo zdawkowego potraktowania podejścia opartego na modelu z interakcjami (wzory (1.3) i (1.4) w pracy doktorskiej). Model ten wykorzystano jedynie w kontekście podejścia regularyzacyjnego, gdzie poddano go krytyce ze względu na trudność wyboru między jego dwoma formami. Nie rozważono natomiast naturalnej symetrycznej formy tego modelu

$$y = X\gamma + TX\beta + \epsilon \ ,$$

gdzie kolumny macierzy TX są uzyskane poprzez przemnożenie kolumn macierzy X przez kolumnę zawierającą informację o grupie: $T_i = -1$ gdy dany obiekt jest w grupie kontrolnej i $T_i = 1$ gdy jest on w grupie zabiegowej. Jak łatwo sprawdzić, w tak utworzonym modelu $\gamma = \frac{\beta^C + \beta^T}{2}$ a $\beta = \frac{\beta^U}{2}$ i estymator parametru różnicowego dostaje się równocześnie z estymatorem parametru γ (bardzo zbliżonym w naturze do parametru $\beta^\star$ z autorskiego estymatora skorygowanego) stosując klasyczne metody wyznaczania parametrów w modelach liniowych. Mam nadzieję na obronie usłyszeć kilka słów dotyczących porównania tego podejścia do metody uzyskiwania skorygowanego estymatora z pracy doktorskiej.

Poza ta uwagą nie mam innych większych zastrzeżeń. i z pełnym przekonaniem stwierdzam, że **rozprawa doktorska mgr Krzysztofa Rudasia spełnia wymogi ustawy o stopniach i tytułach naukowych i wnioskuję o dopuszczenie jej do publicznej obrony.**

Z wyrazami szacunku,

Małgorzata Bogdan

Małgorzata Bogdan