

dr hab. inż. Małgorzata Bogdan
Instytut Matematyki
Uniwersytet Wrocławski

Wrocław 26.08.2019

RECENZJA ROZPRAWY DOKTORSKIEJ

Tytuł rozprawy: Misspecification of Binary Regression Model: Properties and Inferential Procedures

Autor rozprawy: mgr Mariusz Kubkowski

Promotor rozprawy: prof. dr hab. Jan Mielniczuk

Cel i charakter rozprawy

Rozprawa doktorska mgr. Mariusza Kubkowskiego dotyczy bardzo ważnego i istotnego praktycznie zagadnienia identyfikacji zmiennych wpływających na cechę binarną w sytuacji gdy zastosowany model statystyczny odbiega od rzeczywistego rozkładu generującego dane. W ogólnym modelu binarnym zakłada się, że rozkład warunkowy binarnej zmiennej odpowiedzi $Y \in \{0, 1\}$ przy ustalonych wartościach wektora zmiennych niezależnych $x = (1, x_1, \dots, x_p)$ zadany jest wzorem

$$P(Y = 1|X = x) = q(x) \quad , \quad (1)$$

gdzie q jest pewną nieznaną funkcją. W rozprawie duży nacisk położono na analizę modelu semi-parametrycznego, w którym funkcja q ma postać

$$q(x) = q(x'\beta) \quad , \quad (2)$$

gdzie $\beta \in R^{p+1}$ jest wektorem parametrów opisujących wpływ poszczególnych zmiennych na zmienną odpowiedzi. Rozważono także uogólniony model semiparametryczny, w którym

$$q(x) = q(x'\beta_1, \dots, x'\beta_k) \quad . \quad (3)$$

Ponieważ funkcja q jest nieznana, istotność poszczególnych predyktorów szacuje się minimalizując funkcję ryzyka

$$R(b) = El(b, X, Y) , \quad (4)$$

gdzie $l(b, X, Y)$ jest pewną funkcją straty a wartość oczekiwana liczona jest względem rozkładu wektora losowego (Y, X) . Istotność poszczególnych zmiennych można mierzyć za pomocą wektora "rzutu", minimalizującego $R(b)$;

$$\beta^* = \operatorname{argmin}_b R(b) . \quad (5)$$

Model uznaje się za prawidłowo wyspecyfikowany, jeżeli funkcja straty jest zdefiniowana poprzez funkcję wiarygodności w modelu parametrycznym z którego rzeczywiście pochodzą dane. W przeciwnym przypadku model uznajemy za źle wyspecyfikowany.

W praktyce statystycznej należy założyć, że większość modeli jest źle wyspecyfikowanych. Mimo to szereg standardowo stosowanych modeli parametrycznych umożliwia poprawne ustalenie istotnych predyktorów i dość precyzyjną predykcję zmiennej odpowiedzi. W omawianej pracy doktorskiej zjawisko to wyjaśniono matematycznie za pomocą szeregu autorskich lub po autorsku udowodnionych twierdzeń.

W pierwszej części pracy autor koncentruje się na matematycznym opisie własności rzutu β^* i opisuje szereg sytuacji w których jego nośnik (zbiór niezerowych elementów) pokrywa się z nośnikiem rzeczywistego wektora parametrów β . W drugiej części pracy autor analizuje własności metod statystycznych służących do estymacji β^* . W szczególności podaje warunki przy których metoda *SSnet* zaproponowana przez Pokarowskiego i Mielniczuka (2015) jest zgodna z punktu widzenia estymacji nośnika β^* . W ostatniej części pracy autor demonstruje przykłady numerycznego wyznaczania wektora β^* dla logistycznej funkcji straty oraz symulacje ilustrujące działanie *SSnet* w źle wyspecyfikowanym modelu.

Struktura rozprawy

Rozprawa doktorska napisana jest w języku angielskim. Składa się z 6 rozdziałów i obszernego dodatku.

W pierwszym rozdziale wprowadzono podstawowe definicje, zaprezentowano podstawowe funkcje straty stosowane w analizie danych binarnych oraz przedstawiono przykłady modeli źle wyspecyfikowanych.

W Rozdziale 2 przedstawiono szereg interesujących wyników dotyczących własności rzutu β^* w modelu semiparametrycznym (2) i przy założeniu, że funkcja straty jest postaci

$$l(b, x, y) = \rho(b'x, y) . \quad (6)$$

Większość wyników udowodniono przy założeniu, że rozkład wektora zmiennych niezależnych spełnia warunek liniowych regresji $LCR(b)$ (strona 9 w doktoracie) w punkcie β . Ponadto zwykle zakłada się liniową niedegenerowalność X (warunek LND na stronie 10 doktoratu). Wydaje się, że podstawowym wynikiem jest tu Twierdzenie 2.6 w części 2.1, które podaje warunki przy których zachodzi relacja

$$\tilde{\beta}^* = \eta \tilde{\beta} ,$$

gdzie $\tilde{\beta} = (\beta_1, \dots, \beta_p)$ uzyskuje się poprzez usunięcie wyrazu wolnego β_0 z wektora β a η jest stałą zadaną równaniem (2.7) w doktoracie. Po tym Twierdzeniu następuje szereg wniosków opisujących jego zastosowanie w przypadku logistycznej i kwadratowej funkcji straty. Na tym etapie pewien niedosyt budzi fakt, że nie przedyskutowano warunków przy których $\eta = 0$, tzn. sytuacji gdy nośnik $\tilde{\beta}^*$ niekoniecznie jest równy nośnikowi $\tilde{\beta}$. Mankament ten zostaje częściowo usunięty w częściach 2.2 i 2.3, gdzie odpowiednie wersje Twierdzenia 2.6 zostały udowodnione specyficznie dla logistycznej i kwadratowej funkcji straty i w których można znaleźć dokładne wyrażenia na stałe proporcjonalności. Te wyniki wydają się naturalnie implikować interesujący i ważny wynik z części 2.4 pokazujący współkierunkowość rzutów względem logistycznej i kwadratowej funkcji straty. Interesujące dodatkowe przykłady z oszacowaniami na stałą η można znaleźć w części 2.6. Można tam też znaleźć ciekawy przykład ilustrujący istotność założenia $LCR(\beta)$.

W Rozdziale 3 zaprezentowano rozszerzenie Twierdzenia 2.6 i jego "logistycznej" wersji na przypadek uogólnionego modelu semiparametrycznego (3). W części 3.3 udowodniono interesujący wynik pokazujący współkierunkowość wektora $\tilde{\beta}^*$ uzyskanego przy pomocy logistycznej funkcji straty z pierwszym wektorem kanonicznym stosowanym w Liniowej Analizie Dyskryminacyjnej (LDA). Część 3.4. zawiera pewne wyniki techniczne dotyczące addytywnego modelu binarnego. W tej części brakuje mi trochę dyskusji przypadków gdy niektóre elementy wektora η mogą być równe zero.

Rozdziały 4 i 5 zawierają wyniki dotyczące statystycznych metod estymacji rzutu β^* i jego nośnika. W tych rozdziałach zakładamy, że dysponujemy

próbą składającą się z n niezależnych realizacji $(X_1, Y_1), \dots, (X_n, Y_n)$ wektora losowego (X, Y) z ogólnego modelu binarnego

$$P(Y = 1|X = x) = q(x) ,$$

gdzie elementy wektora losowego X mają rozkład sub-gausowski. Dodatkowo założono, że funkcja straty jest postaci (6) a estymacja parametrów odbywa się w oparciu o empiryczne ryzyko

$$R_n(b) = \frac{1}{n} \sum_{i=1}^n \rho(b' X_i, Y_i) .$$

W przypadku gdy liczba zmiennych p jest większa od liczby obserwacji n minimum funkcji $R_n(b)$ jest zwykle osiągane w nieskończenie wielu punktach. Aby zapobiec temu efektowi często stosuje się metody regularyzacyjne, w których do funkcji $R_n(b)$ dodaje się pewną funkcję regularyzującą. W rozdziale 5 przeanalizowano własności popularnej metody regularyzacyjnej LASSO;

$$\hat{\beta}_L = \arg \min_b R_n(b) + \lambda \|\tilde{b}\|_1 ,$$

gdzie $\|\tilde{b}\|_1 = \sum_{i=1}^p |\tilde{b}_i|$. Twierdzenia 4.9 i 4.15 podają oszacowania na odległość między $\hat{\beta}_L$ a β^* w przypadku logistycznej i ogólnej funkcji straty. Oba te twierdzenia umożliwiają określenie asymptotycznych warunków przy których LASSO umożliwia prawidłową separację między rzeczywistymi predyktorami a nieistotnymi zmiennymi niezależnymi. Własność ta mówi, że dla odpowiednio dużego n elementy wektora $\hat{\beta}$ odpowiadające istotnym predyktorom będą co do wartości bezwzględnej większe od elementów odpowiadających nieistotnym zmiennym niezależnym. W celu identyfikacji rzeczywistych predyktorów pozostaje teraz ustalić odpowiedni punkt odcięcia. W Rozdziale 5 ten problem został zaadresowany poprzez podanie warunków na zgodność Uogólnionego Kryterium Informacyjnego *GIC* (Generalized Information Criterion) oraz procedury *SSnet* łączącej LASSO i *GIC*.

W części rozprawy dotyczącej *GIC* i *SSnet* moje wątpliwości budzi założenie, że $k_n \log p_n = o(a_n)$, gdzie k_n oznacza maksymalny rozmiar przeszukiwanych modeli a a_n jest karą w kryterium *GIC*. Wydaje się, że w interesującej sytuacji gdy $p_n/n \rightarrow \infty$ i $k_n \rightarrow \infty$ eliminuje ono przypadki gdy $a_n = \log n$ (BIC) lub $a_n = \log n + 2 \log p_n$ (EBIC1), zaproponowane i przeanalizowane w części symulacyjnej. Będę wdzięczna za bliższe skomentowanie tego założenia i zastosowanych kryteriów na obronie.

Rozdział 6 zawiera obszerne wyniki symulacji komputerowych proponowanych rozwiązań statystycznych oraz propozycje numerycznego wyznaczania rzutu β^* dla dowolnej funkcji q .

Dodatek zawiera szereg wyników technicznych. W moim odczuciu część A.1 dotycząca istnienia i jednoznaczności wektora $\hat{\beta}$ jest bardzo ważna i powinna się znaleźć w głównej części rozprawy.

Ocena pracy doktorskiej

W moim odczuciu jest to niewątpliwie wyróżniająca rozprawa doktorska. Zaprezentowanymi wynikami matematycznymi i statystycznymi można by wypełnić co najmniej trzy standardowe rozprawy. Bardzo spodobał mi się Rozdział 2, który ilustruje odporność standardowych rozwiązań parametrycznych (logistyczna i kwadratowa funkcja straty) w modelu semiparametrycznym. Chciałabym też zwrócić uwagę na stosunkowo dużą liczbę konkretnych przykładów ilustrujących wprowadzane pojęcia statystyczne i matematyczne. Czytanie poszczególnych rozdziałów było ułatwione dzięki wstępom sygnalizującym najważniejsze twierdzenia z ich krótkim wyjaśnieniem i umiejscowieniem w stosunku do podobnych wyników w literaturze. Muszę jednak przyznać, że pomimo tych starań czytanie rozprawy było dla mnie dość trudne. W wyniku przyjętego "liniowego" sposobu prezentacji mniej istotne techniczne Lematy przeplatają się z głównymi Twierdzeniami. Poza tym dążenie do zachowania pełnej ogólności Twierdzeń prowadzi do dużej liczby technicznych założeń. Wydaje się, że w wielu miejscach można byłoby istotnie rozszerzyć dyskusję tych założeń. Te zastrzeżenia dotyczące prezentacji wyników oczywiście nie wpływają na moją bardzo wysoką ocenę tej rozprawy i z pełnym przekonaniem stwierdzam, że

rozprawa doktorska mgr Mariusza Kubkowskiego spełnia wymogi ustawy o stopniach i tytułach naukowych i wnioskuję o dopuszczenie jej do publicznej obrony. Wnioskuję również o jej wyróżnienie.

Z wyrazami szacunku,

Małgorzata Bogdan

