PROBABILISTYCZNE STRUKTURY REGRESJI

Agnieszka Szumera

STRESZCZENIE

Klasycznie pojmowane zagadnienie regresji polega na wyznaczaniu wielomianów W stopnia nieprzekraczającego zadanej liczby $n \in \mathbb{N}$, które są optymalnie dopasowane do zadanych skończonych ciągów liczbowych x i y, reprezentujących dane empiryczne. Jako kryterium optymalizacji przyjmuje się tutaj odchylenie kwadratowe. Zatem wielomiany W są wyznaczane metodą najmniejszych kwadratów, na podstawie danych empirycznych. Wielomiany regresyjne odgrywają istotną rolę w analizie empirycznych danych liczbowych. Najprostsze do wyznaczenia są wielomiany regresyjne stopnia nie przekraczającego 1, czyli tzw. regresje liniowe. Zależność liniowa jest również bardzo intuicyjna. Z tego powodu regresje liniowe są najczęściej stosowane w praktyce. Z drugiej strony takie podejście niesie za sobą nadmierne uproszczenie szukanej zależności funkcyjnej pomiędzy danymi. Trudno jest w ten sposób wychwycić bardziej złożone zależności.

Przedmiotem badań w rozprawie jest opracowanie matematycznie spójnej i możliwie ogólnej teorii regresji, stosowanej na przykład w rachunku prawdopodobieństwa, statystyce i ekonometrii, gdzie wyznaczanie najlepiej dopasowanych przebiegów teoretycznych do danych empirycznych jest oparte na koncepcji metody najmniejszych kwadratów. W pracy wprowadzono nowe pojęcie *probabilistycznej struktury regresji* $\mathfrak{P} := (A, B, \delta; x, y)$ nad przestrzenią probabilistyczną $\mathcal{P} := (\Omega, \mathcal{A}, P)$, gdzie

- A jest niepustym zbiorem, zaś $B = \mathbb{R}$ lub $B = \mathbb{C}$;
- funkcje $x\colon \Omega \to A$ i $y\colon \Omega \to B$ są określone na pewnym niepustym zbiorze zdarzeń elementarnych Ω, (są to tzw. funkcje próbkujące, czyli funkcje generowane poprzez dane eksperymentalne modelu regresji);
- funkcja $\delta\colon (\Omega \to B) \times (\Omega \to B) \to \overline{\mathbb{R}}$ interpretowana jest jako kryterium odchylenia funkcji teoretycznej od danych empirycznych. Jest ona definiowana poprzez przestrzeń probabilistyczną $\mathcal{P}$ zgodnie z klasyczną koncepcją metody najmniejszych kwadratów.

Dla zadanej probabilistycznej struktury regresji $\mathfrak{P}$ rozważana jest rodzina funkcji $\mathcal{F}$, nazwana *teoretycznym modelem funkcyjnym* probabilistycznej struktury regresji $\mathfrak{P}$, będąca podzbiorem rodziny wszystkich funkcji odwzorowujących zbiór A w zbiór B. Zakłada się, że klasa $\mathcal{F}$ tworzy rozmaitość liniową w przestrzeni funkcji $(A \to B)$ ze standardową strukturą dodawania funkcji i mnożenia funkcji przez stałą.

Naturalnym zagadnieniem dla zadanej probabilistycznej struktury regresji $\mathfrak{P}$ jest badanie i wyznaczanie optymalnych funkcji modelu funkcyjnego $\mathcal{F}$, które są w sensie kryterium δ, najlepiej dopasowane do danych empirycznych reprezentowanych przez funkcje x i y. Ujmując to w języku matematycznym rozważany jest problem ekstremalny polegający na wyznaczeniu i badaniu klasy $\mathrm{Reg}(\mathcal{F}, \mathfrak{P})$ wszystkich funkcji $f_0 \in \mathcal{F}$ minimalizujących funkcjonał $\mathcal{F} \ni f \to \delta(t \circ x, y) \in \overline{\mathbb{R}}$. Funkcje takie zostały nazwane *funkcjami regresji* w klasie $\mathcal{F}$ względem probabilistycznej struktury regresji

$\mathfrak{P}$. Tak postawiony problem ekstremalny został nazwany *problemem regresji* w klasie $\mathcal{F}$ dla struktury $\mathfrak{P}$.

Rozdział pierwszy zawiera zarys historyczny teorii regresji oraz syntetyczny opis asynchronicznych (ARS) i synchronicznych (SRS) struktur regresji. Pojęcia te zostały zaczerpnięte z artykułu *Generalized approach to the problem of regression* (DOI 10.1007/s13324-014-0096-7) będącego próbą ujednolicenia teorii regresji. Ukazał się on w 2015 roku i stał się inspiracją do napisania niniejszej pracy doktorskiej. W rozdziale drugim przedstawiono pomocnicze fakty dotyczące całkowania funkcji zespolonych, wykorzystywane w dalszej części dysertacji. Udowodniono między innymi twierdzenia o całkowaniu przez podstawianie oraz o aproksymacji zespolonych funkcji całkowalnych przez funkcje proste.

Rozdział trzeci stanowi zasadniczą część pracy, w której zdefiniowano probabilistyczną strukturę regresji $\mathfrak{P} := (A, B, \delta; x, y)$ nad przestrzenią probabilistyczną $\mathcal{P} := (\Omega, \mathcal{A}, P)$. Głównym wynikiem jest tutaj twierdzenie 3.8 charakteryzujące klasę funkcji regresji. Wynika z niego szereg podstawowych własności funkcji regresji opisanych w dalszej części pracy. W szczególności podano warunki zapewniające jednoznaczność funkcji regresji. Opisano również regresję pierwszego rodzaju w terminach probabilistycznych struktur regresji. Poruszono także zagadnienie zależności pomiędzy rzeczywistymi i zespolonymi probabilistycznymi strukturami regresji. W szczególności zaprezentowany został przykład uzasadniający rozważanie zespolonych probabilistycznych struktur regresji. Rozdział czwarty przedstawia zagadnienie efektywnego wyznaczenia funkcji regresji w terminach zadanej skończonej bazy rozmaitości liniowej $\mathcal{F}$. Została tutaj omówiona struktura klasy regresji oraz procedura wyznaczania funkcji regresji. Ponadto rozwiązano problem regresji dla jednowymiarowych, dwuwymiarowych i trójwymiarowych modeli teoretycznych $\mathcal{F}$. Został poruszony także przypadek ogólny rozwiązania problemu regresji dla wielowymiarowych liniowych modeli teoretycznych. Powyższe rozważania zostały zilustrowane przykładami. Poruszono również zagadnienie transformacji funkcji regresji spowodowane zastąpieniem funkcji próbkującej y przez złożenie $g \circ y$, gdzie g jest wielomianem stopnia pierwszego.

Rozdział piąty przedstawia ilustrację praktyczną rozważań teoretycznych. Na podstawie twierdzenia 4.10 opracowano metodę numeryczną obliczania funkcji regresji dla dowolnie zadanego układu funkcji bazowych. Zaprezentowano również realne przykłady wyznaczania funkcji regresji oparte na implementacji tej metody w arkuszu kalkulacyjnym MS Excel przy użyciu języka programowania VBA. Kod programu został zamieszczony w załączniku. Ponadto znajdują się tam wynikowe zrzuty ekranu z wykonanych obliczeń numerycznych dla realnych przykładów zamieszczonych w rozdziale piątym.

Słowa kluczowe: Przestrzeń probabilistyczna, funkcja regresji, struktura regresji, regresja liniowa, regresja wielomianowa, optymalizacja wielomianowa.

Klasyfikacja Tematyczna Pracy(2020): 28A10, 46E99, 60A10, 60B99, 62J05, 62J20, 65K10, 90C23.