Praca domowa nr 2 – Ekonometria Bayesowska
Kacper Kalinowski 76975

Przedstawienie zagadnienia

Przedmiotem analizy były wyniki testów pewnych studentów. Dane wykorzystane do analizy bayesowskiej pochodzą ze strony kaggle.com. Analiza została przeprowadzona przy pomocy modelu N-G ze skrajnie nie informacyjnym rozkładem a priori i losowaniu z funkcji ważności.

Wykorzystane dane należą do pakietu SPSS v23 oraz zawierają informację na temat wyników pewnych studentów oraz ich charakterystyk takich jak:

- lokalizacja szkoły do której uczęszcza uczeń, trzy możliwe wyniki – szkoła znajduje się w mieście (nazwa zmiennej Urban), na przedmieściach (nazwa zmiennej Suburban) lub na wsi (zastosowano kodowanie referencyjne, kategorią bazową była szkoła na wsi),
- typ szkoły do której uczęszcza uczeń, dwa możliwe wyniki – szkoła publiczna (nazwa zmiennej Public) lub prywatna (zastosowano kodowanie referencyjne, kategorią bazową była szkoła prywatna),
- liczba studentów w klasie danego ucznia (nazwa zmiennej n_student),
- płeć ucznia (nazwa zmiennej gender).

Liczba obserwacji w zbiorze wynosi 2133. W celach badawczych wylosowano tylko 100 obserwacji ze zbioru.

Istnieją badania, które odnoszą się do różnic w osiąganych wynikach nauki przez uczniów uczęszczających do szkół publicznych oraz prywatnych (Sapelli, Vial, 2002), czy uczniów uczęszczających do szkół na wsi oraz w mieście (Palka, 2009, s. 237). Badacze koncentrują się również na analizie wpływu liczby uczniów w klasie a wynikami testów (Krueger, Whitmore, 2001) oraz wpływie płci (Caparo, Margaret, 2000).

Elicytacja parametrów a priori

Zastosowano całkowicie nie informacyjny rozkład Normalny-Gamma z nałożonymi dodatkowymi restrykcjami w postaci znaku parametrów beta w modelu. W związku z tym nastąpiło zawężenie dziedziny o zbiory, którym w modelu N-G przypisywano niezerową gęstość a priori.

Zważywszy na wybór wyżej wspomnianej specyfikacji, niezbędne było wybranie zmiennych objaśniających, a także znaku parametru przy zmiennych niezależnych. Proces ten bazował na informacjach z analiz pochodzących z literatury zagadnienia wyników testów studentów.

W przypadku zmiennej odnoszącej się do lokalizacji szkoły na podstawie literatury stwierdzono, że uczniowie szkół miejskich i podmiejskich osiągają lepsze rezultaty od uczniów ze szkół wiejskich (Palka, 2009, s. 237). Tym samym znak parametrów przy zmiennych Urban oraz Suburban będzie miał wartość dodatnią. Być może wiąże się to z faktem lepiej wykwalifikowanej i bardziej licznej kadry mieszkającej na zurbanizowanych terenach.

Na podstawie literatury zdiagnozowano, że uczniowie szkół publicznych osiągają gorsze wyniki od uczniów ze szkół prywatnych (Sapelli, Vial, 2002). Tym samym znak parametru przy zmiennej Public będzie miał wartość ujemną. Prawdopodobnie wiąże się to z faktem, że szkoły prywatne częściej oferują lepsze wynagrodzenie nauczycielom, co przyciąga lepiej wykwalifikowaną kadrę.

Liczba studentów w klasie również ma istotny wpływ na osiągane przez uczniów wyniki nauki. Studenci uczący się w mniejszych grupach osiągali lepsze rezultaty testów (Krueger, Whitmore, 2001, s. 26). Oznacza to że znak parametru przy zmiennej n_student będzie miał wartość ujemną.

Praca domowa nr 2 – Ekonometria Bayesowska
Kacper Kalinowski 76975

Nauczyciele w mniej licznych klasach poświęcają średnio więcej czasu uczniom, niż nauczyciele w bardziej licznych grupach.

Badacze podkreślają również, że płeć nie ma istotnego wpływu na osiągane wyniki testów (Caparo, Margaret, 2000, s. 3). Z tego względu zmienna gender nie wzięła udziału w procesie modelowania zjawiska.

Podsumowanie nałożonych restrykcji znajduje się w poniższej tabeli:

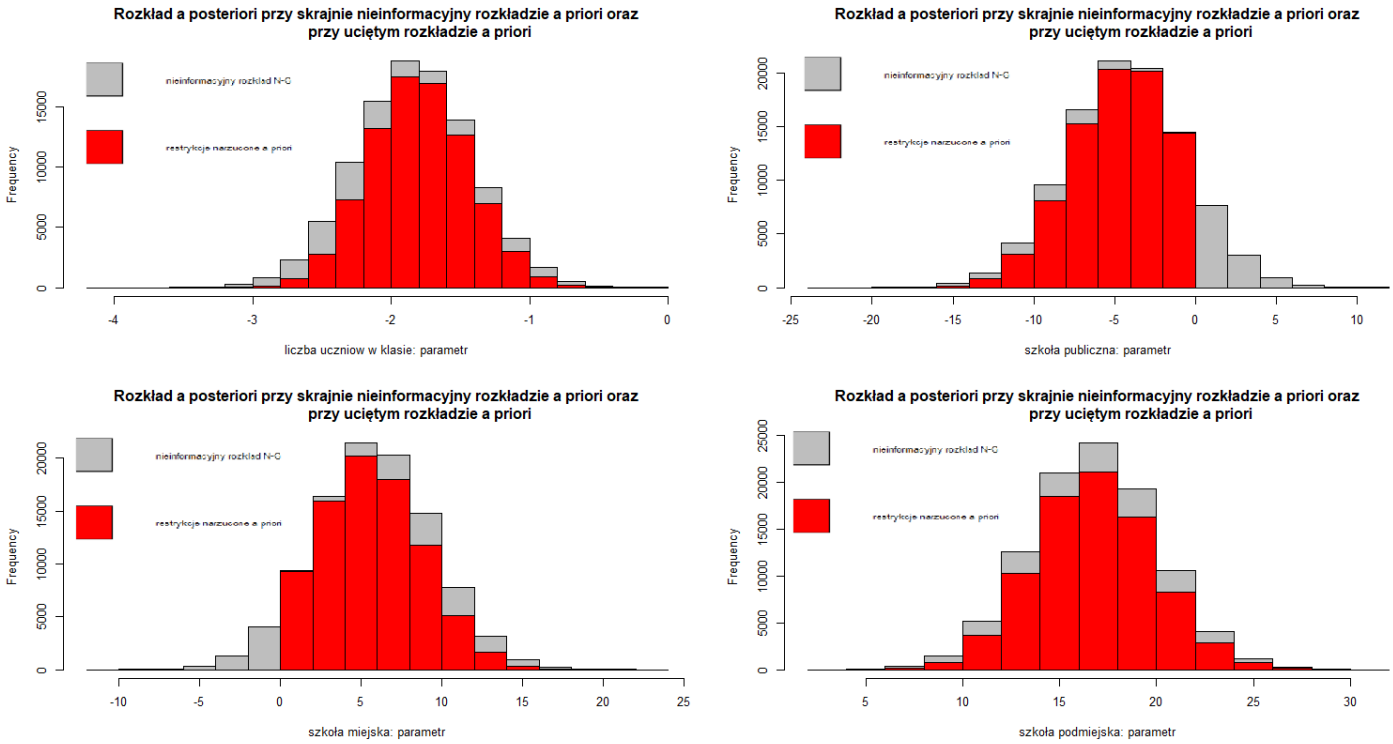
Zmienna	**Parametr**	**Wiedza a priori**	**Gęstość a priori**
Urban	β_1	>	$\mathbb{I}_{R+}$
Suburban	β_2	>	$\mathbb{I}_{R+}$
Public	β_3	<	$\mathbb{I}_{R-}$
N_student	β_4	<	$\mathbb{I}_{R-}$
Stała	β_0	-	$\mathbb{I}_{R}$
Precyzja	h	>	gamma

Wartość oczekiwana parametrów a posteriori

Zmienna	**Wartość oczekiwana parametru $\overline{\boldsymbol{\beta}}$**
N_student	-1.799
Public	-4.889
Urban	5.744
Suburban	16.79
Stała	103.54

Praca domowa nr 2 – Ekonometria Bayesowska
Kacper Kalinowski 76975

Rysunek 1. Rozkłady brzegowe parametrów a posteriori



W przypadku zmiennej odnoszącej się do liczby uczniów w klasie rozkład a posteriori nie różni się istotnie od rozkładu nie informacyjnego a priori. Spowodowane jest to faktem, że oszacowanie parametru w modelu OLS wyszło istotne przy bardzo niskim p-value w teście sprawdzającym istotność oszacowania. Z tego też względu wartość oczekiwana parametru a posteriori przy tej zmiennej nie różni się istotnie od oszacowania parametru bez ograniczeń.

Rozkład a priori zmiennej Public został ograniczony za pomocą restrykcji, ze względu na to, że spora część masy prawdopodobieństwa znajdowała się przy wartościach powyżej 0. Z tego też względu wartość oczekiwana parametru a posteriori różni się istotnie od oszacowanej wartości parametru z modelu OLS.

Rozkład a priori zmiennej Urban został ograniczony za pomocą restrykcji, ze względu na to, że część obserwacji wylosowanych z rozkładu nie informacyjnego Normalnego-Gamma ma wartości powyżej 0. Z tego też względu wartość oczekiwana parametru a posteriori różni się istotnie od oszacowanej wartości parametru z modelu OLS.

Rozkład zmiennej a posteriori Suburban nie różni się istotnie od rozkładu nie informacyjnego a priori. Spowodowane jest to faktem, że oszacowanie parametru przy tej zmiennej w modelu OLS wyszło istotne przy bardzo niskim p-value w teście sprawdzającym istotność oszacowania. Z tego też względu wartość oczekiwana parametru a posteriori przy tej zmiennej nie różni się istotnie od oszacowania parametru bez ograniczeń.

Praca domowa nr 2 – Ekonometria Bayesowska
Kacper Kalinowski 76975

Ocena zmiennych w modelu

Tabela 1. Oszacowania parametrów z modelu OLS oraz przedziały HPD

	Dolna granica HPD (95%)	Oszacowanie bez ograniczeń (model OLS)	Górna granica HPD (95%)
Stała	90.458	103.979	117.500
n_student	-2.661	-1.837	-1.012
Public	-11.529	-4.303	2.923
Urban	-1.409	5.757	12.924
Suburban	10.365	16.767	23.168

Tabela 2. Wartość oczekiwana parametrów a posteriori oraz przedziały HPD

	Dolna granica HPD (95%)	Oszacowanie z ograniczeniami	Górna granica HPD (95%)
Stała	90.965	103.542	116.596
n_student	-2.502	-1.799	-1.101
Public	-10.006	-4.889	-0.001
Urban	0.001	5.744	10.954
Suburban	10.923	16.791	22.769

Zmienne Public oraz Urban w modelu OLS są nieistotne statystycznie, ponieważ w 95% przedziałach HPD dla tych zmiennych zawiera się 0. Dzięki nałożeniu restrykcji zawężone zostały rozkłady parametrów – mniejsza masa prawdopodobieństwa znajduje się blisko 0. Dzięki temu zmienne są ważne w modelu.

W przypadku rozkładów zmiennych n_student oraz Urban wartość 0 występuje z prawdopodobieństwem niemal równym 0 (co widać na wykresach rozkładów a posteriori). Oznacza to, że obie zmienne mają duże znaczenie w procesie modelowania zmiennej objaśnianej.

Ocena błędu modeli

Zdolności predykcyjne obu modeli zostały sprawdzone na niezależnym zbiorze danych, który został wydzielony z oryginalnego zbioru. Wylosowane obserwacje nie brały udziału w procesie modelowania. Statystyką na podstawie której wnioskowano było RMSE (root mean squared error). Poniższa tabela przedstawia podsumowanie wyników.

RMSE OLS	**RMSE Bayesian model**
15.32	10.57

W przypadku modelu klasycznego dokonując predykcji na wylosowanym zbiorze mylimy się średnio o około 15 punktów testowych, natomiast w przypadku modelu z restrykcjami mylimy się średnio o około 10,5 punktów testowych (średni wynik testu w zbiorze wynosi 67 pkt).

Podsumowując, zastosowanie bayesowskiego podejścia do analizy regresji z restrykcjami jakościowymi skutecznie zwiększa dokładność prognoz wyników testu.

Literatura

Praca domowa nr 2 – Ekonometria Bayesowska
Kacper Kalinowski 76975

Sapelli C., Vial B., (2002), The performance of private and public schools in the Chilean voucher system.

Palka K. (2009), Wyniki testu gimnazjalnego uczniów ze środowiska wiejskiego i miejskiego a poziom ich aspiracji edukacyjnych i zawodowych.

Krueger A. B., Whitmore D. M., (2001) The effect of attending a small class in the early grades on collage-test taking and middle chool test results: evidence from project star, The Economic Journal.

Caparro, Margaret M., Robert M., Wiggins, Barret B., (2000), An Investigation of the Effects of Gender, Socioeconomic Status, Race and Grades on Standardized Test Scores.