MARZENA NOWAKOWSKA

Politechnika Świętokrzyska
spimn@tu.kielce.pl

Modele regresji Bayesa w analizach bezpieczeństwa ruchu drogowego

Wnioskowanie statystyczne jest narzędziem stosowanym w analizach bezpieczeństwa ruchu drogowego (brd) od początku ich zaistnienia. Wnioskowanie to opiera się o dwie szkoły: częstościową (zwaną również klasyczną) oraz bayesowską (nieklasyczną). Spośród nich pierwsza zdobyła sobie znaczącą przewagą w pracach naukowych poświęconych badaniom zagrożeń na drodze. Jednakże od pewnego czasu, początkowo sporadycznie a następnie coraz częściej i chętniej, są stosowane metody bayesowskie. W szczególności, ostatnich kilka lat zaowocowało stosunkowo dużą liczbą zagranicznych (głównie amerykańskich, kanadyjskich i chińskich) publikacji z zakresu brd, w których dyskutowano modele regresyjne budowane w oparciu o teorię Bayesa. W modelach takich przyjmuje się, że parametry strukturalne nie są stałymi, tylko zmiennymi losowymi o pewnych rozkładach aposteriorycznych będących wynikiem wcześniejszej (apriorycznej) wiedzy na ich temat oraz uaktualnienia tej wiedzy poprzez informacje z danych empirycznych. Rozwój modelowania bayesowskiego stał się możliwy dzięki technikom numerycznym: metodom próbkowania i generowania łańcuchów Markowa Monte Carlo.

Większość wspomnianych prac została dedykowana budowaniu i dyskusji bayesowskich modeli GLM do prognozowania zdarzeń (wypadków) drogowych, w znacznej części w analizach typu „przed i po" (np. [1, 5, 7, 10, 11, 13]). Rzadziej, i w stosunkowo ograniczonym zakresie, pojawiały się publikacje poświęcone modelom regresyjnym do klasyfikowania cech jakościowych, takich jak zachowanie sprawcy, rodzaj czy status (ciężkość) zdarzenia drogowego (np. [4, 6]).

Modele regresji Bayesa są trudne koncepcyjnie i wymagające obliczeniowo. Niemniej jednak stwarzają nową jakość w zakresie rozwoju metod badań naukowych i umożliwiają elastyczne, choć niestandardowe, podejście w interpretacji wyników. Koncepcja takiego modelowania została omówiona w prezentowanym artykule. Wykorzystując tę nieklasyczną metodę zbudowano modele regresji logistycznej, w których jakościowa zmienna celu określała status (ciężkość) zdarzenia drogowego a zmiennymi objaśniającymi były wybrane cechy charakteryzujące kierującego sprawcę wypadku drogowego. Przedmiotem dyskusji było więc zagadnienie klasyfikacyjne w aspekcie modelowania bayesowskiego zastosowane do analiz brd.

Budowa modelu regresyjnego – podejście klasyczne i bayesowskie

Klasa uogólnionych modeli liniowych GLM jest rozszerzeniem teorii i metod modeli liniowych w odniesieniu do danych z odpowiedzią niekoniecznie podlegającą rozkładowi normalnemu [2, 12]. W modelu GLM wykorzystuje się transformację g znaną funkcją łączącą (ang. *link function*) stosowaną do składnika deterministycznego mającego postać kombinacji liniowej:

$$y_i = \mathrm{g}^{-1}(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}) + \varepsilon_i, \quad i = 1, ..., n \qquad (1)$$

w której: n jest liczbą obserwacji, k – liczbą parametrów modelu.

Uogólniony model liniowy można następująco scharakteryzować (pogrubioną czcionką zaznaczono wektory/macierze):

1. Składnik systematyczny Z modelu ma postać liniowej kombinacji zmiennych objaśniających (wejść): $Z = \boldsymbol{B} \times [1, \boldsymbol{X}] = B_0 + B_1 X_1 + \ldots + B_k X_k$, w której $\boldsymbol{B}$ są policzonymi z próby estymatorami prawdziwych wartości parametrów $\boldsymbol{\beta}$.
2. Funkcja łącząca g(·) odnosi się do modelu linowego poprzez zależność: $\mathrm{g}(\mu) = Z$, $\mu = \mathrm{E}(Y)$. Funkcja łącząca jest monotoniczna i odwracalna, dzięki czemu odpowiedź μ może być wyrażona poprzez odwrotność tej funkcji: $\mu = g^{-1}(Z)$.
3. Składnik losowy ε modelu jest opisany przez funkcję rozkładu należącą do grupy rozkładów wykładniczych.

W klasycznej analizie regresji zakłada się, że parametry strukturalne $\boldsymbol{\beta}$ modelu są wartościami stałymi. Szacuje się je zazwyczaj za pomocą metody największej wiarygodności MLE (ang. *Maximum Likelihood Estimation*). Niepewność dotycząca oszacowanych wartości parametrów wyraża się poprzez ich błędy standardowe (błędy standardowe współczynników regresji).

W podejściu bayesowskiem przyjmuje się, że parametry strukturalne są zmiennymi losowymi, co definiuje bayesowską analizę regresji. W takim przypadku estymacja opiera się na formule Bayesa i polega na oszacowaniu rozkładu parametrów strukturalnych, mając wiedzę wstępną na temat tych rozkładów i uaktualniając tę wiedzę w oparciu o dane [12]:

$$\begin{aligned} \mathrm{P}(\boldsymbol{B} \mid Y, \boldsymbol{X}) &= \mathrm{P}(B_0, \ldots, B_k \mid Y, \boldsymbol{X}) = \\ &= \mathrm{P}(B_0, \ldots, B_k) \cdot \mathrm{P}(Y, \boldsymbol{X} \mid B_0, \ldots, B_k) / \mathrm{P}(Y, \boldsymbol{X}) \propto \\ \mathrm{P}(B_0, \ldots, B_k) &\cdot \mathrm{L}(Y, \boldsymbol{X} \mid B_0, \ldots, B_k) = \mathrm{P}(\boldsymbol{B}) \cdot \mathrm{L}(Y, \boldsymbol{X} \mid \boldsymbol{B}) \end{aligned} \qquad (2)$$

Zgodnie z regułą Bayesa, rozkład aposterioryczny $\mathrm{P}(\boldsymbol{B} \mid Y, \boldsymbol{X})$ jest proporcjonalny do rozkładu apriorycznego $\mathrm{P}(\boldsymbol{B})$ i funkcji wiarygodności $\mathrm{L}(Y, \boldsymbol{X} \mid \boldsymbol{B})$. W zależności (2) $\mathrm{P}(Y, \boldsymbol{X})$ jest rozkładem brzegowym wektora obserwacji. Funkcja wiarygodności $\mathrm{L}(Y, \boldsymbol{X} \mid \boldsymbol{B})$ opisuje łączny rozkład prawdopodobieństwa n-elementowej próby losowej przy założeniu znanych wartości (znanego rozkładu) wektora parametrów $\boldsymbol{B}$:

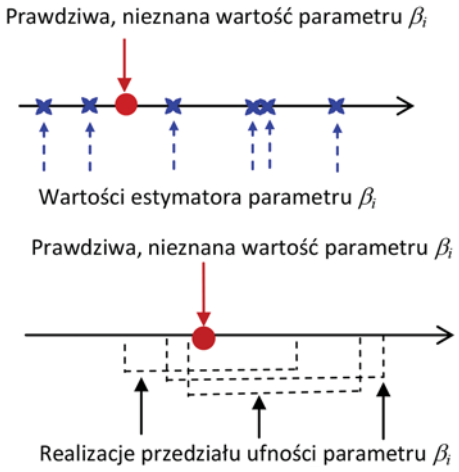
$$\begin{aligned} \mathrm{L}(Y, \boldsymbol{X} \mid \boldsymbol{B}) &= \mathrm{P}(Y = y_1 \mid \boldsymbol{B}^T \cdot \boldsymbol{x}_1) \cdot \\ \mathrm{P}(Y = y_2 \mid \boldsymbol{B}^T \cdot \boldsymbol{x}_2) &\cdot \ldots \cdot \mathrm{P}(Y = y_n \mid \boldsymbol{B}^T \cdot \boldsymbol{x}_n) \end{aligned} \qquad (3)$$

w której: $y_1, y_2, \ldots, y_n$ są wartościami zmiennej losowej Y w kolejnych obserwacjach, natomiast $\boldsymbol{x}_1, \boldsymbol{x}_2 \ldots, \boldsymbol{x}_n$ reprezentują wartości wektora zmiennych objaśniających $\boldsymbol{X}$ w kolejnych obserwacjach.

Dobrym kompendium wiedzy na temat bayesowskich modeli regresji jest opracowanie przygotowane przez SAS Institute [12]. Firma ta oferuje profesjonalne, komercyjne oprogramowanie umożliwiające wykonanie przedmiotowych analiz. Oprócz systemów komercyjnych, istnieje również oprogramowanie dostępne na licencji *freeware*, które użytkownik może pobrać z witryn internetowych, np.: http://www.mrc-bsu.cam.ac.uk/ software/ bugs/the-bugs-project-winbugs/.

Klasyczne i bayesowskie parametry modelu regresji

W klasycznej analizie regresji prawdziwa wartość parametru strukturalnego nie jest znana. Jest szacowana na podstawie estymatora wyznaczanego z próby losowej. Dla danego parametru β_i można wyznaczyć wiele różnych wartości estymatora, ponieważ do tego celu można wykorzystywać wiele różnych prób pobranych z populacji generalnej, co zilustrowano na rys. 1. Ponieważ zazwyczaj badacz dysponuje jedną próbą z populacji generalnej, więc może uzyskać jednopunktowe oszacowanie nieznanego parametru.



Rys. 1. Graficzna interpretacja estymacji parametru klasycznego modelu regresji

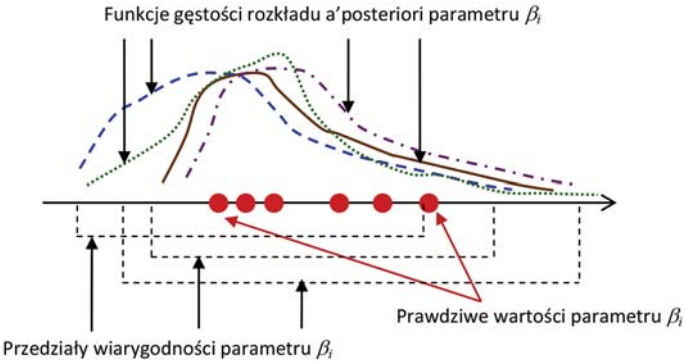
W regresji Bayesa parametr strukturalny jest zmienną losową. Może przyjmować wiele wartości, zgodnie z rozkładem, któremu podlega. Rozkład ten, zwany rozkładem a'posteriori, zawiera informację z dwóch źródeł: (1) danych oraz (2) wcześniejszą wiedzę o parametrze. Jest więc wyznaczany na podstawie próby losowej (dane), ale też dodatkowo na podstawie apriorycznego rozkładu parametru. Różne założenia dotyczące wcześniejszej wiedzy decydują o postaci rozkładu a'priori i, w konsekwencji, mają wpływ na rozkład a'posteriori (rys. 2). Każdorazowe uaktualnienie któregokolwiek z dwóch ww. źródeł może zmienić rozkład a'posteriori.

W ujęciu klasycznym, przedział ufności dla parametru β_i, to przedział losowy $[U1, U2]$, dla którego $\mathrm{P}(U1 \leq \beta_i \leq U2) = (1 - \alpha)$, gdzie $\alpha \in (0, 1)$. Otrzymaną z próby realizację $[u1, u2]$ przedziału ufności można określić jako jeden z tych przedziałów, które z prawdopodobieństwem $(1 - \alpha)$ zawierają β_i. Należy podkreślić, że zmieniać mogą się granice realizacji przedziału a nie wartość parametru. Nie można więc stwierdzać, że β_i zmienia się od $u1$ do $u2$.

W podejściu bayesowskim szacowany parametr β_i jest zmienną losową, więc podlega określonemu rozkładowi a'posteriori [12] :

$$\mathrm{P}(\beta_i \in C \mid Y, \boldsymbol{X}) = \int_C \mathrm{P}(\beta_i \mid Y, \boldsymbol{X})\, \mathrm{d}\beta_i \qquad (4)$$

Przyjmując, że prawdopodobieństwo jest równe $(1 - \alpha)$, gdzie $\alpha \in (0, 1)$, obszar C nosi nazwę bayesowskiego przedziału wiarygodności (ang. *Bayesian credible interwal*). Takich przedziałów może być wiele. Przedział wiarygodność reprezentuje informację, że β_i należy do C z określonym prawdopodobieństwem $(1 - \alpha)$. Jeżeli ze zbioru wszystkich przedziałów wiarygodności C wyznaczony zostanie taki C^*, że: (1) $\mathrm{P}(\beta_i \mid Y, X; \beta_i \in C^*) \geq \mathrm{P}(\beta_i \mid Y, X; \beta_i \in C)$ oraz (2) C^* jest najkrótszy, to nosi on nazwę przedziału największej gęstości prawdopodobieństwa a'posteriori *HPD* (ang. *highest posterior density interval*). Gdy funkcja gęstości a'posteriori nie jest funkcją jednostajną, to dla ustalonej wartości $(1 - \alpha)$ przedział ten jest unikalny. Przedział *HPD* jest do pewnego stopnia odpowiednikiem przedziału ufności. Może więc być wyznaczony dla zadanego α i wykorzystany do określenia istotności parametru strukturalnego modelu Bayesa.



Rys. 2. Graficzna interpretacja estymacji parametru bayesowskiego modelu regresji.

Wyznaczenie rozkładów a'posteriori parametrów modelu regresji

Rozkład a'priori może być sklasyfikowany do jednej z dwóch kategorii [12]:

- rozkład a'priori nieinformatywny (ang. *noninformative*) – rozkład płaski, w którym wystąpienie każdej wartości jest jednakowo prawdopodobne; jego wpływ na rozkład aposterioryczny jest minimalny,
- rozkład a'priori informatywny (ang. *informative*) – gdy jest dostępna pełna lub prawie pełna informacja o tym rozkładzie (postać i parametry); rozkład ten dominuje nad funkcją wiarygodności w zależności Bayesa a jego wpływ na rozkład aposterioryczny może być duży lub bardzo duży.

Jednym ze sposobów wyznaczenia rozkładów a'posteriori $\mathrm{P}(\boldsymbol{B} \mid Y, \boldsymbol{X})$ są metody próbkujące z wykorzystaniem łańcuchów Markowa Monte Carlo (MCMC) (ang. *Marcov Chains Monte Carlo*) [3, 12]. Metoda MCMC polega na generowaniu

dla każdego parametru ciągu liczb spełniających kryteria łańcucha Markowa, które mają rozkład zgodny z rozkładem a'posteriori. Najpopularniejszym algorytmem prowadzącym do uzyskania tego rozkładu jest algorytm Metropolisa oraz jego uogólnienie – algorytm Metropolisa-Hastingsa. Często stosowany jest również próbnik Gibbsa (ang. *Gibbs sampler*).

Wynik metody MCMC jest determinowany przez [12]: liczbę iteracji w łańcuchu, liczbę odrzuconych wartości początkowych, zwanych wartościami spalonymi (ang. *burn-in*), oraz wskaźnik przerzedzenia (ang. *thinning*) informujący o tym, co który element łańcucha jest pobierany do próby aposteriorycznej. Istotnym elementem procedury jest uzyskanie zbieżności łańcuchów Markowa, dzięki czemu próby wynikowe poszczególnych parametrów modelu pochodzą ze stacjonarnych rozkładów a'posteriori. Do oceny jakości łańcuchów Markowa wykorzystuje się testy diagnostyczne (np. Gelmana-Rubina, Geweke'a, Heidelbergera-Welcha) oraz wykresy diagnostyczne śladu i autokorelacji.

Eksperymenty badawcze

Przedmiotem analiz był klasyfikator statystyczny – model regresji logistycznej, którego zadaniem było klasyfikowanie statusu zdarzenia drogowego $StsZd$ do jednej z dwóch kategorii: wypadek lekki (WL) albo wypadek ciężki lub śmiertelny (WCS). Zmiennymi objaśniającymi $\boldsymbol{X}$ w modelu były cechy charakteryzujące kierującego sprawcę tego zdarzenia: rodzaj nieprawidłowego zachowania będącego przyczyną zdarzenia, wiek, płeć, obecność alkoholu lub innych środków odurzających we krwi kierującego. Nie brano pod uwagę innych zmiennych (jak np. cechy drogi), które mogą być dodatkowymi czynnikami wpływającymi na ciężkość wypadku. Głównym celem była bowiem dyskusja bayesowskiego podejścia do regresyjnych analiz w zakresie bezpieczeństwa ruchu drogowego, przy czym, w dyskusji tej, przykładowy zbiór wejść był definiowany tylko przez czynniki ludzkie.

W modelu logistycznym funkcją łączącą jest logit. W przeprowadzanym eksperymencie badawczym argumentem tej funkcji było prawdopodobieństwo warunkowe $\mathrm{P}(StsZd = WCS \mid X_1, \ldots, X_k)$ wypadku o statusie WCS [8, 9]:

$$\begin{aligned}&\operatorname{logit}\left(\mathrm{P}(StsZd = WCS \mid X_1, \ldots, X_k)\right) \\ &= \ln\left(\frac{\mathrm{P}(StsZd = WCS \mid X_1, \ldots, X_k)}{1 - \mathrm{P}(StsZd = WCS \mid X_1, \ldots, X_k)}\right) = \\ &= B_0 + B_1 X_1 + \cdots + B_k X_k\end{aligned} \tag{5}$$

W procesie estymacji wykorzystano rzeczywiste dane opisujące zdarzenia drogowe i ich uczestników, zarejestrowane na odcinkach drogi krajowej nr 7 przebiegającej przez województwo świętokrzyskie. Dane pobrano z policyjnego Systemu Ewidencji Wypadków i Kolizji (SEWIK) i przygotowano je do analiz następująco:

- uwzględniono tylko wypadki (zdarzenia drogowe z ofiarami wśród ludzi),
- wypadki miały miejsce na zamiejskich jednojezdniowych dwukierunkowych odcinkach drogi krajowej nr 7 (w tym z uwzględnieniem skrzyżowań),
- jeden pełnoletni kierujący był sprawcą wypadku – odrzucono przypadki, gdy kierujący był nieletni, gdy wypadek nie był spowodowany przez kierującego lub gdy było co najmniej dwóch kierujących sprawców,
- kierujący uczestnicy wypadków prowadzili pojazdy silnikowe,
- w wypadkach nie uczestniczyli piesi,
- odrzucono rekordy z niejednoznacznie zdefiniowanym zachowaniem sprawcy („inne zachowania" wg policyjnej karty zdarzenia drogowego) oraz zachowania rzadkie (nie przekraczające 3% ogólnej liczby obserwacji),
- odrzucono rekordy z brakami w danych.

W celu zweryfikowania pozytywnych opinii o zastosowaniu modelowania bayesowskiego dla małych prób (np. [5, 6, 7, 13]) dane do badań podzielono na dwa zbiory: z okresu 1999–2014, tworząc długi zbiór treningowy o liczbie obser-

Tabela 1. **Opis danych o zdarzeniach drogowych do eksperymentów badawczych**

Opis pozycji	Symbol	Okres 1999–2014	Okres 2008–2014
Liczebność zbioru		500	203
Status zdarzenia drogowego – zmienna objaśniana	StsZd		
Wypadek lekki	WL	49%	45%
Wypadek ciężki lub śmiertelny	WCS	36%+15%	42%+13%
Zachowanie kierującego – zmienna objaśniająca	ZchKr		
Niezachowanie bezpiecznej odległości między pojazdami	NzOd	9%	11%
Jazda po niewłaściwej stronie drogi	JzdNwSDr	4%	4%
Niedostosowanie prędkości do warunków ruchu	NdsPr	42%	33%
Nieudzielenie pierwszeństwa przejazdu	NdzPrwPrz	22%	27%
Nieprawidłowe skręcanie lub zawracanie	NprSkrZwr	6%	9%
Nieprawidłowe wymijanie lub wyprzedzanie	NprOmjWprz	11%	10%
Ograniczenie sprawności psychofizycznej*	OgSpPsFz	6%	6%
Grupa wiekowa kierującego – zmienna objaśniająca	GrWkK		
[18, 25)	01	22%	21%
[25, 35)	02	30%	28%
[35, 50)	03	28%	26%
[50, 65)	04	16%	19%
>= 65*	05	4%	6%
Płeć kierującego – zmienna objaśniająca	Plc		
Kobieta	K	11%	11%
Mężczyzna*	M	89%	89%
Obecność alkoholu lub innego środka odurzającego we krwi kierującego – zmienna objaśniająca	Alh		
Nie stwierdzono	N	92%	94%
Stwierdzono*	T	8%	6%

* Kategorie odniesienia w modelach logistycznych

wacji n = 500, oraz z okresu o połowę krótszego 2008–2014, tworząc krótszy zbiór treningowy o liczbie obserwacji n = 203. W porównaniu ze zbiorem 500-elementowym, zbiór o liczności 203 obserwacji można uznać za mało liczny, unikając jednocześnie w tym zbiorze zjawiska całkowitej lub pozornie całkowitej separacji danych w tablicy kontyngencji generowanej w procesie estymacji modelu [9]. Zestawienie informacji o tych danych zawarto w tabeli 1.

Modele Bayesa wyznaczono dla rozkładów a'priori parametrów:

- nieinformatywnych: rozkłady normalne o średniej 0 i odchyleniu sztandarowym 1E+6,
- informatywnych: rozkłady normalne o parametrach pobranych z modeli klasycznych otrzymanych za pomocą metody największej wiarygodności MLE [13]; dla każdej zmiennej objaśniającej budowano osobny klasyfikator, po czym estymator parametru strukturalnego i jego odchylenie standardowe wprowadzano do modelu Bayesa jako parametry rozkładów apriorycznych, w obu przypadkach (wstępne estymacje MLE oraz docelowa estymacja bayesowska) wykorzystując ten sam zbiór danych do obliczeń.

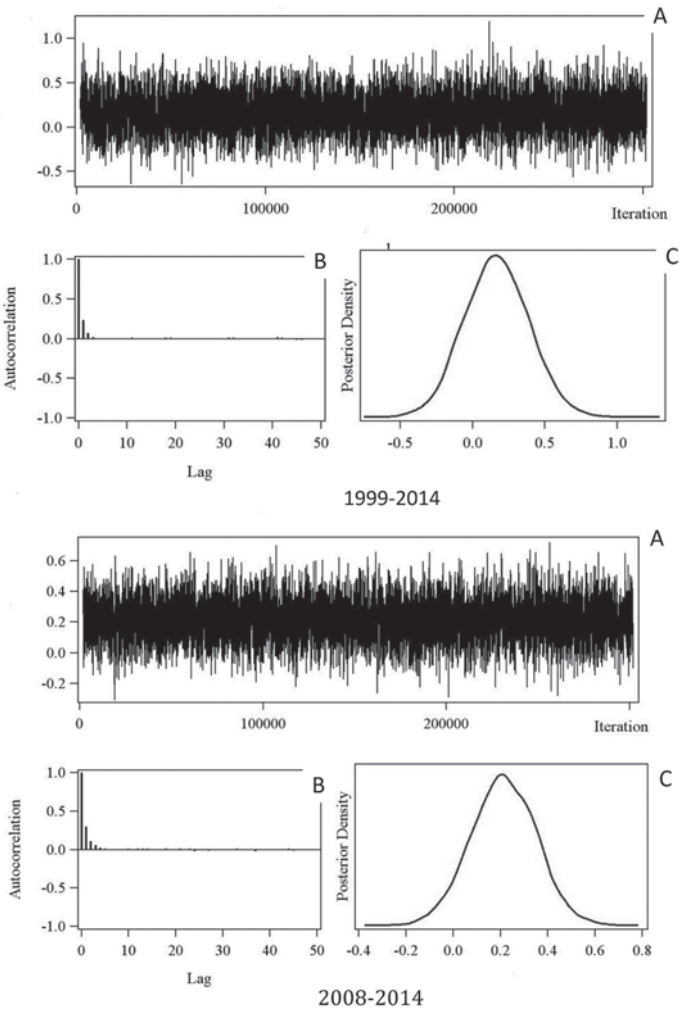
Wynikowe rozkłady a'posteriori powstały na podstawie 10000-elementowych łańcuchów Markowa wygenerowanych za pomocą algorytmu Metropolisa wg schematu: odrzucona liczba iteracji początkowych – 10000, liczba iteracji docelowych – 300000 ze wskaźnikiem przerzedzenia równym 30.

Eksperymenty badawcze wykonano w środowisku systemu SAS® wykorzystując jego procedury statystyczne oraz własne programy napisane we wbudowanym języku 4GL.

Wyniki eksperymentów badawczych

Wszystkie łańcuchy Markowa dla parametrów modeli Bayesa osiągnęły stacjonarność, co zostało zweryfikowane za pomocą wykresów śladu i autokorelacji oraz za pomocą testów Heidelbergera-Welcha. Uzyskane rozkłady aposterioryczne były jednomodalne. Na rys. 3 przedstawiono przykładowe wykresy śladu (A) i autokorelacji (B) łańcuchów Markowa oraz rozkłady aposteriori (C) dla parametru strukturalnego stojącego przy zmiennej *ZchKr_NdzPrwPrz* wyznaczone dla danych z okresu 1999–2014 (parametr nieistotny statystycznie) oraz dla danych z okresu 2008–2014 (parametr istotny statystycznie).

W tabeli 2 zestawiono wyniki estymacji modeli regresji logistycznej: estymatory parametrów strukturalnych dla regresji klasycznej (E) oraz wartości średnie rozkładów aposterioryczych (SR) reprezentujące parametry w modelu regresji Bayesa. Niepewność dotyczącą wartości parametru wyraża się w modelu klasycznym poprzez błąd standardowy BS, który jest oszacowaniem średniej rozbieżności między parametrem a jego możliwymi ocenami. W modelu Bayesa odpowiadającą mu miarą jest aposterioryczne odchylenie standardowe OS. Ponieważ trudno jest ocenić wartości BS i OS, w nawiasach umieszczono odpowiadające im wskaźniki względne – iloraz wspomnianej miary dyspersji i oszacowanego parametru. Wartość bezwzględna tego wskaźnika precyzji bliska 100% (i więcej) wskazuje na niezadowalającą dokładność. Wartość ponad 50% powinna zwrócić uwagę na inne miary oceny modelu. Wykresy miar BS i OS dla wyznaczonych modeli umieszczono na dole tabeli. Dodatkowo podano końce 95% przedziałów ufności PU dla modeli klasycznych i 95% przedziałów największej wiarygodności HPD dla modeli bayesowskich (obecność zera w HPD wskazuje na brak istotności odpowiadającego mu parametru).

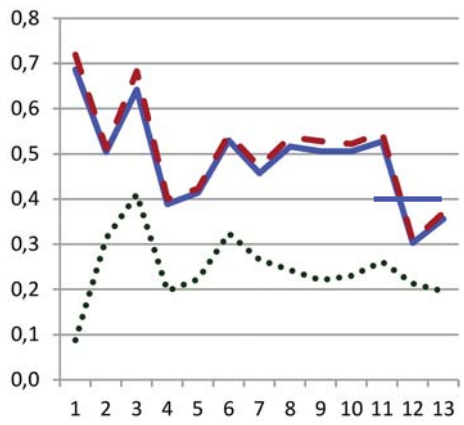
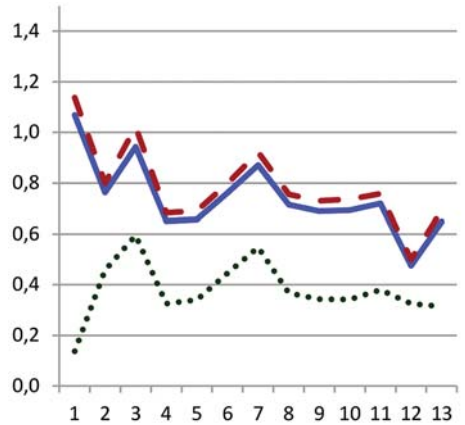


Rys. 3. Wykresy diagnostyczne dla parametru strukturalnego zmiennej ZchKr_NdzPrwPrz

Kolorowym tłem zaznaczono parametry istotne statystycznie (α = 0,05); ich wskaźniki precyzji nie przekraczały (co do wartości bezwzględnej) 50%. Pozostałe (niezaznaczone) parametry okazały się nieistotne statystycznie: dla modeli klasycznych ich 95% przedziały ufności zawierały zera, a dla modeli Bayesa zera wystąpiły w 95% przedziałach największej gęstości prawdopodobieństwa.

- W przypadku modeli grupy 1999–2014, różnice ($|E\text{-}SR|$) między parametrami modelu klasycznego i nieinformatywnego modelu Bayesa były stosunkowo małe – z jednym wyjątkiem (*ZchKr_JzdNwSDr*) nie przekroczyły 0,05, a ich średnia wynosiła 0,03. Większe różnice wystąpiły między parametrami modelu klasycznego i informatywnego modelu Bayesa – z jednym wyjątkiem (*Plc_K*) przekroczyły 0,11, a średnia różnic była równa 0,23. Współczynnik zmienności różnic był w przypadku modelu nieinformatywnego nieco mniejszy (67%) niż w przypadku modelu informatywnego (78%).

Tabela 2. **Wyniki estymacji modeli logistycznych klasyfikujących status zdarzenia drogowego**

Opis pozycji	Model klasyczny MLE		Model Bayesa nieinformatywny		Model Bayesa informatywny	
	E (BS/E)	95% PU	SR (OS/SR)	95% HPD	SR (OS/SR)	95% HPD
Modele grupy 1999-2014; N = 500						
Wyraz wolny	0,86 (80%)	–0,49; 2,21	0,90 (80%)	–0,48; 2,30	0,09 (96%)	–0,08; 0,27
ZchKr_NzOd	–1,22 (–41%)	–2,21; –0,23	–1,27 (–40%)	–2,24; –0,23	–1,10 (–28%)	–1,71; –0,49
ZchKr_JzdNwSDr	0,96 (67%)	–0,30; 2,23	1,04 (65%)	–0,23; 2,43	1,11 (37%)	0,33; 1,93
ZchKr_NdsPr	0,02 (2057%)	–0,75; 0,78	0,01 (2798%)	–0,75; 0,81	0,20 (101%)	–0,21; 0,57
ZchKr_NdzPrwPrz	–0,01 (–5765%)	–0,82; 0,81	–0,01 (–6538%)	–0,81; 0,85	0,17 (134%)	–0,25; 0,62
ZchKr_NprSkrZwr	0,03 (1976%)	–1,01; 1,07	0,03 (1723%)	–1,05; 1,08	0,22 (147%)	–0,39; 0,86
ZchKr_NprOmjWp	0,39 (117%)	–0,51; 1,29	0,40 (118%)	–0,50; 1,33	0,53 (50%)	0,01; 1,05
GrWkK_02	–0,28 (–187%)	–1,29; 0,74	–0,30 (–177%)	–1,30; 0,80	–0,04 (–673%)	–0,54; 0,41
GrWkK_03	–0,65 (–77%)	–1,65; 0,34	–0,68 (–77%)	–1,74; 0,32	–0,38 (–58%)	–0,82; 0,05
GrWkK_04	–0,50 (–100%)	–1,50; 0,49	–0,53 (–98%)	–1,56; 0,47	–0,26 (–89%)	–0,73; 0,17
GrWkK_05	–0,10 (–540%)	–1,13; 0,94	–0,11 (–484%)	–1,15; 1,01	0,12 (215%)	–0,39; 0,64
Plc_K	–0,70 (–43%)	–1,30; –0,11	–0,73 (–42%)	–1,35; –0,15	–0,73 (–29%)	–1,14; –0,31
Alh_N	–0,33 (–108%)	–1,03; 0,37	–0,34 (–110%)	–1,08; 0,37	0,02 (862%)	–0,37; 0,39
Wskaźniki oceny	–2LogL–R = 33,08 –2LogL(MP)/n = 1,32		DIC = 212,69		DIC = 204,34 (spadek: 8,35)	
Modele grupy 2008-2014; N = 203						
Wyraz wolny	0,73 (146%)	–1,38; 2,84	0,86 (132%)	–1,39; 3,09	0,21 (65%)	–0,07; 0,47
ZchKr_NzOd	–0,70 (–109%)	–2,21; 0,81	–0,75 (–106%)	–2,24; 0,85	–0,67 (–68%)	–1,53; 0,26
ZchKr_JzdNwSDr	0,97 (97%)	–0,90; 2,84	1,13 (90%)	–0,76; 3,23	1,13 (53%)	–0,10; 2,25
ZchKr_NdsPr	0,64 (102%)	–0,65; 1,93	0,69 (99%)	–0,63; 2,04	0,70 (46%)	0,07; 1,34
ZchKr_NdzPrwPrz	0,70 (94%)	–0,60; 2,00	0,76 (91%)	–0,64; 2,02	0,79 (43%)	0,15; 1,48
ZchKr_NprSkrZwr	0,14 (543%)	–1,36; 1,65	0,17 (459%)	–1,36; 1,75	0,23 (191%)	–0,60; 1,14
ZchKr_NprOmjWp	2,08 (42%)	0,36; 3,80	2,30 (40%)	0,50; 4,10	2,17 (25%)	1,16; 3,29
GrWkK_02	–0,64 (–112%)	–2,05; 0,78	–0,73 (–104%)	–2,17; 0,78	–0,33 (–110%)	–1,08; 0,38
GrWkK_03	–0,58 (–119%)	–1,94; 0,78	–0,67 (–109%)	–2,09; 0,78	–0,27 (–126%)	–0,93; 0,40
GrWkK_04	–0,80 (–86%)	–2,18; 0,57	–0,91 (–81%)	–2,29; 0,60	–0,49 (–70%)	–1,16; 0,18
GrWkK_05	–0,15 (–466%)	–1,58; 1,27	–0,22 (–346%)	–1,69; 1,28	0,15 (250%)	–0,60; 0,88
Plc_K	–0,71 (–67%)	–1,65; 0,23	–0,77 (–65%)	–1,78; 0,19	–0,72 (–45%)	–1,40; –0,11
Alh_N	–0,50 (–129%)	–1,79; 0,78	–0,59 (–118%)	–1,97; 0,73	–0,36 (–88%)	–0,97; 0,25
Wskaźniki oceny	–2LogL–R = 23,45 –2LogL(MP)/n = 1,26		DIC = 139,79		DIC = 129,21 (spadek: 10,58)	
Błędy oszacowania parametrów	Grupa 1999–2014			Grupa 2008–2014		
BS - MLE OS - Bayes nieinformatywny OS - Bayes informatywny						

- W grupie 2008–2014 różnice między parametrami modelu klasycznego i modeli Bayesa były nieco inne niż w grupie 1999–2014: średnia wynosiła 0,09 w przypadku modelu nieinformatywnego i 0,19 w przypadku modelu informatywnego. Zmienność tych różnic w przypadku modelu nieinformatywnego wynosiła 56%, a w przypadku modelu informatywnego 84%.
- W grupie 1999–2014 istotny statystycznie we wszystkich modelach był parametr dla *ZchKr_NzOd*. W modelu informatywnym Bayesa istotne statystycznie okazały się dodatkowo parametry dla zmiennych: *ZchKr_JzdNwSDr*, *ZchKr_NprOmjWp* oraz *Plc_K*.
- W grupie 2008–2014 istotny statystycznie we wszystkich modelach był parametr dla *ZchKr_NprOmjWp*. W modelu informatywnym Bayesa jako istotne statystycznie zostały ponadto zidentyfikowane parametry dla: *ZchKr_NdsPr*, *ZchKr_NdzPrwPrz* oraz *Plc_K*.
- Wykresy wskaźników precyzji informują, że błędy oszacowania parametrów w przypadku modeli MLE i nieinformatywnych modeli Bayesa były podobne w obu grupach, chociaż nieco większe różnice można zauważyć w grupie 2008–2014. W przypadku informatywnych modeli Bayesa zaznaczyły się wyraźnie mniejsze wartości tych błędów dla wszystkich parametrów, zarówno w grupie 1999–2014, jak i w grupie 2008–2014.
- Wskaźniki precyzji oszacowania parametrów były generalnie mniejsze w modelach grupy 1999–2014 niż w odpowiadających im modelach grupy 2008–2014 – por. linie łamane o tych samych kolorach i stylach na obu wykresach w tabeli 2.
- Do określenia jakości dopasowania modeli klasycznych wykorzystano wskaźnik oceny wewnętrznej *-2LogL* [8]. Im większa różnica *-2LogL-R* między wartością tej miary dla modelu ze stałą *M0* (bez zmiennych objaśniających) i dla modelu nasyconego *MP* (ze stałą i zmiennymi objaśniającymi), tym lepsza jakość modelu nasyconego. Do porównania jakości modeli z obu okresów zastosowano iloraz wskaźnika *-2LogL* dla *MP* i liczby obserwacji *n* – im mniejszy ten iloraz, tym lepszy model. Oba nasycone modele klasyczne okazały się lepsze niż odpowiadające im modele ze stałą. Jakość klasyfikacji tych modeli była podobna: 1,32 i 1,26.
- W przypadku modeli regresji Bayesa miarą oceny wewnętrznej jest *DIC* [1, 6, 7]. Wskaźnik służy do porównywania modeli szacowanych na tym samym zbiorze treningowym – wartość *DIC* mniejsza o co najmniej 5 wskazuje na lepszy model. Klasyfikatory Bayesa wyznaczone przy założeniu apriorycznych rozkładów informatywnych były lepsze niż klasyfikatory z rozkładami nieinformatywnymi, zarówno dla dłuższego, jak i krótszego zbioru treningowego; miara *DIC* była mniejsza odpowiednio o ponad 8 i o ponad 10.

Modele Bayesa z informatywnymi rozkładami a'priori parametrów strukturalnych okazały się najlepszymi klasyfikatorami, zarówno w przypadku dłuższego, jak i krótszego zbioru danych. Liczba istotnych statystycznie zmiennych objaśniających w tych modelach była większa niż w odpowiadających im modelach klasycznych i modelach bayesowskich z nieifnormatywnymi rozkładami apriorycznymi. Interpretując otrzymane wyniki wykorzystuje się iloraz szans i określa, jak zmieniała się szansa, że wypadek drogowy był ciężki lub śmiertelny w skutkach wraz ze wskazaną zmianą wartości zmiennej objaśniającej, przy ustalonych pozostałych wejściach [8, 9]. Podejście bayesowskie uwzględnia niepewność we wnioskowaniu poprzez dopuszczalność zmian wartości parametrów strukturalnych.

Zgodnie z informatywnym modelem Bayesa wyznaczonym do okresu analizy 1999–2014 szansa, że wypadek drogowy był ciężki lub śmiertelny w skutkach:
- dla zmiennej opisującej zachowanie kierującego sprawcy wypadku była w porównaniu z ograniczeniem sprawności psychofizycznej:
  - mniejsza średnio o 67% w przypadku niezachowania bezpiecznej odległości między pojazdami (przedział zmiany wartości: od 39% do 82%),
  - średnio trzy razy większa w przypadku jazdy po niewłaściwej stronie drogi (przedział zmiany wartości: od 1,4 do 6,9 razy),
  - większa średnio o 70% w przypadku nieprawidłowego wyprzedzania lub omijania (przedział zmiany wartości: od 1% do 185%),
- dla zmiennej informującej o płci kierującego sprawcy wypadku była mniejsza dla kobiet niż dla mężczyzn średnio o 52% (przedział zmiany wartości: od 27% do 68%).

Zgodnie z informatywnym modelem Bayesa wyznaczonym dla okresu analizy 2008–2014 szansa, że wypadek drogowy był ciężki lub śmiertelny w skutkach:
- dla zmiennej opisującej zachowanie kierującego sprawcy wypadku była w porównaniu z ograniczeniem sprawności psychofizycznej:
  - średnio dwa razy większa w przypadku niedostosowania prędkości do warunków ruchu (przedział zmiany wartości: od nieco ponad 1 do 3,8 razy),
  - średnio ponad dwa razy większa w przypadku nieudzielenia pierwszeństwa przejazdu (przedział zmiany wartości: od 1,2 do 4,4 razy),
  - średnio ponad osiem i pól razy większa w przypadku nieprawidłowego wyprzedzania lub omijania (przedział zmiany wartości: od 3,2 do 26,8 razy),
- dla zmiennej informującej o płci kierującego sprawcy wypadku była mniejsza dla kobiet niż dla mężczyzn średnio o 52% (przedział zmiany wartości: od 11% do 75%).

Podsumowanie

Techniki bayesowskie umożliwiają łączenie systematycznej wiedzy dotyczącej szacowanych parametrów oraz informacji pochodzącej z danych i pozwalają na otrzymanie rozkładów tych parametrów, a nie tylko ich ocen punktowych lub przedziałowych. Coraz większe możliwości wykorzystania metod bayesowskich wynikają z ich uniwersalności oraz rozwoju technik obliczeniowych. Mimo złożoności, zastosowanie teorii Bayesa do modelowania staje się więc coraz popularniejsze, zwłaszcza w zagranicznych badaniach dotyczących analiz bezpieczeństwa ruchu drogowego. W artykule przedstawiono koncepcję bayesowskiego modelu regresji, a następnie zilustrowano omówioną metodę badawczą na przykładzie modelu logistycznego do klasyfikacji statusu wypadku drogowego. Wykonane eksperymenty badawcze

pozwoliły na potwierdzenie już istniejących opinii na temat metody oraz wyciągnięcie nowych wniosków:

- taką samą istotność parametrów i stosunkowo niewielkie różnice w ich wartościach uzyskano dla modeli: klasycznego i nieinformatywnego Bayesa, zarówno w przypadku dłuższego, jak i krótszego zbioru danych treningowych,
- modele Bayesa z informatywnymi rozkładami a'priori miały większą liczbę parametrów istotnych statystycznie niż odpowiadające im (szacowane na tym samym zbiorze danych) modele klasyczne oraz nieinformatywne Bayesa,
- modele Bayesa z informatywnymi rozkładami a'priori uzyskały najlepszą precyzję oszacowania parametrów, niezależnie od liczebności zbioru treningowego; modele te miały również lepszą miarę oceny wewnętrznej niż modele Bayesa z rozkładem nieinformatywnym,
- wprowadzenie w modelu Bayesa rozkładu informatywnego wpłynęło pozytywnie na jakość klasyfikatorów, również w przypadku krótszego zbioru danych uaktualniających informacje aprioryczne,
- zastosowanie metody Bayesa do wyznaczania modelu regresji może być uzasadnione, jeżeli dysponuje się nielicznym zbiorem danych do analiz; w tym przypadku istotne znaczenie ma dobór właściwego rozkładu a'priori (np. będącego wynikiem wcześniejszych analiz).

Modelowanie związków między różnymi zmiennymi mogącymi odgrywać istotną rolę w opisywaniu zagrożeń bezpieczeństwa ruchu drogowego w praktyce umożliwia ilościową ocenę, czy i w jakim stopniu określona cecha (np. wybrana cecha kierującego sprawcy wypadku drogowego) wpływa na wartość zmiennej objaśnianej (ciężkość wypadku). Podejście bayesowskie pozwala na łączenie w modelu wcześniejszej albo uogólnionej wiedzy o zagrożeniach i wiedzy pochodzącej z danych. Ponadto, dopuszcza zmienność wpływu analizowanej cechy na ciężkość wypadku w zakresie wartości zidentyfikowanych za pośrednictwem modelu. Dzięki temu uzyskuje się szerokie pole do analiz zagrożeń brd (również porównawczych) w kontekście wybranych zmiennych niezależnych (np. wg cech sprawcy lub cechy drogi). To z kolei umożliwia doprecyzowanie ukierunkowania działań naprawczych lub prewencyjnych.

Bibliografia

[1] El-Basyouny K., Barua S., Islam M. T. Investigation of time and weather effects on crash types using full Bayesian multivariate Poisson lognormal models. Accident Analysis and Prevention 201; 73: 91-99.
[2] Hand D., Manilla H., Padhraic S. Eksploracja danych. Warszawa: WN-T, 2005.
[3] Häggström O. Finite Markov Chains and Algorithmic Applications. Cambridge University Press (Virtual Publishing) 2003.
[4] Helai H., Chor C.H., Haque M.M. Severity of driver in jury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. Accident Analysis and Prevention 2008; 40: 45-54.
[5] Heydari S., Miranda-Moreno L.F., Lord D., Fu L. Bayesian methodology to estimate and update safety performance functions under limited data conditions: A sensitivity analysis. Accident Analysis and Prevention 2014; 64: 41-51.
[6] Huang H., Abdel-Aty M. Multilevel data and Bayesian analysis in traffic safety. Accident Analysis and Prevention 2010; 42: 1556-1565.
[7] Mitra S., Washington S. On the nature of over-dispersion in motor vehicle crash prediction models. Accident Analysis and Prevention 2007; 39: 459-468.
[8] Nowakowska M. Logistic models in the crash severity classification on the basis of chosen road characteristics. Transportation Research Record, Journal of the Transportation Research Board, Highway Safety Data, Analysis, and Evaluation, Volume 2, Washington D.C. 2010; 2148: 16-26.
[9] Nowakowska M. Modelowanie związków między cechami drogi a zagrożeniami w ruchu na drogach zamiejskich. Warszawa: Oficyna Wydawnicza Politechniki Warszawskiej, 2013.
[10] Pei X. Wong S.C., Sze N.N. A joint probability approach to crash prediction models. Accident Analysis and Prevention 2011; 43: 1160-1166.
[11] Persaud B., Lan B., Lyon C., Bhim R. Comparison of empirical Bayes and full Bayes approach for before-after road safety evaluations. Accident Analysis and Prevention 2010; 42: 38-43.
[12] SAS/STAT® 9.2 User's Guide (Introduction to Bayesian Analysis Procedures). Second Edition, Cary, NC, USA: SAS Institute Inc., 2009.
[13] Yu R., Abdel-Aty M. Investigation different approaches to develop informative priors in hierarchical Bayesian safety performance functions. Accident Analysis and Prevention 2013; 56: 51-58.

Serwis GDDKiA • Aktualności

Nastąpił wybór wykonawcy S7 Lubień-Naprawa

Wybrano wykonawcę drogi ekspresowej S7 Kraków–Rabka Zdrój na odcinku Lubień–Naprawa. Jest nim polsko-ukraińskie konsorcjum przedsiębiorstw: IDS-BUD S.A z Warszawy i Korporacja ALTIS-HOLDING z Kijowa. Wykonawca ten zaoferował, że za 521 519 095,35 złotych, w ciągu 22 miesięcy od daty zawarcia umowy wybuduje ten odcinek drogi ekspresowej dając przy tym 10-letnią gwarancję jakości. W czwartek, 29 października 2015 r., w siedzibie GDDKiA Oddział Kraków nastąpiło otwarcie ofert wykonawców na budowę odcinka drogi ekspresowej S7 Lubień–Naprawa o długości 7 km 590 m. Oferty złożyło 12 firm i konsorcjów. Najniższa zaproponowana cena wynosiła ponad 521,5 mln zł, najwyższa ponad 828 mln zł. Większość firm (8) zaproponowało cenę pomiędzy ponad 521 mln a ponad 592 mln zł.

Zgodnie z przepisami w ciągu 10 dni mogą wpłynąć ewentualne odwołania od innych wykonawców. Zgodnie z Ustawą Prawo zamówień publicznych, zamówienie to podlega kontroli uprzedniej Prezesa Urzędu Zamówień Publicznych, w związku z tym umowa z wybranym konsorcjum może być zawarta po doręczeniu do GDDKiA Oddział w Krakowie informacji o wyniku tej kontroli. Na 7,6 km odcinku drogi ekspresowej S7 Lubień–Naprawa

oprócz dwujezdniowej drogi klasy S, wybudować trzeba m.in. dwa MOP-Y: Lubień i Krzeczów oraz 10 obiektów inżynierskich i 6 małych mostków na potoku Krzywańskim.

26-01-2016