modelowanie częstości wypadków drogowych,
ujemny dwumianowy model Poissona, wariancja

KAMIŃSKA Joanna[1]

UJEMNY DWUMIANOWY ROZKŁAD POISSONA W MODELOWANIU WYPADKÓW DROGOWYCH

Problem modelowania (przewidywania) liczby wypadków drogowych na poszczególnych odcinkach dróg (sekcjach) pojawił się w rozważaniach naukowców około 30 lat temu. Przez ponad ćwierć wieku do przewidywania liczby wypadków drogowych w zależności od czynników ruchu drogowego stosowano bardzo różne modele oraz sposoby estymacji. W pracy opisano procedurę modelowania wypadków drogowych z zastosowaniem rozkładu ujemnego dwumianowego wraz z analizą wariancji. Analiza ta wykazała skuteczność modelu w wyjaśnianiu (opisywaniu) zmienności systematycznej (ponad 60%).

NEGATIVE BINOMIAL DISTIBUTION IN ACCIDENT PREDICTION MODELLING

The relationship between traffic accidents and traffic conditions has been the subject of research for about 30 years, mainly in the last 20 years. Researchers have developed many different models, types, conditions and functional forms depending on available data, local conditions and purpose. In paper the procedure of accident prediction modeling using negative binomial model was described. The analysis shows that model explain mode than 60% of systematic variance.

1. WSTĘP

Pierwotnie zależność między czynnikami zewnętrznymi, a liczbą wypadków opisywano deterministycznym równaniem regresji liniowej. Jednak, jak wykazali Jovanis i Chang (1986), zmienność natężenia ruchu w czasie oraz częste występowanie zerowej liczebności wypadków w sektorze sprawiają, że standardowa regresja liniowa może prowadzić do błędnych estymatorów parametrów, a nawet przewidywanych ujemnych wartości. W odpowiedzi na problem zastosowania regresji liniowej, rozpoczęto badania nad użyciem regresji Poissona (Joshua, Garber 1990), której zastosowanie jest możliwe dzięki temu, że dane dotyczące wypadków drogowych stanowią nieujemne całkowite wartości. W kolejnych latach pojawiały się rozszerzenia regresji Poissona, nowe modele oraz kolejne metody estymacji. Wśród najczęściej stosowanych modeli ekonometrycznych w przewidywaniu częstości wypadków drogowych w zależności od geometrii drogi oraz

[1]Uniwersytet Przyrodniczy we Wrocławiu, Wydział Inżynierii Kształtowania Środowiska i Geodezji, Katedra Matematyki, 50-357 Wrocław, ul. Grunwaldzka 53. Tel: +48 71 3205615, E-mail:Joanna.kaminska@up.wroc.pl

natężenia ruchu na określonych odcinkach, można wymienić (Lord, Mannering 2010): klasyczny model regresji Poissona, model gamma-Poissona (NB – negative binomial), jednowymiarowy model lognormalny Poissona (PLN), wielowymiarowy model lognormalny Poissona (MVPLN), model Conway-Maxwell-Poisson oraz wiele innych.

2. METODYKA
2.1 Model

Model regresji Poissona zakłada, że dla niezależnych par obserwacji (n_i, x_i), ($i=1,...,n$), prawdopodobieństwo wystąpienia n_i wypadków w segmencie i wynosi $P(n_i)$, gdzie

$$P(n_i) = \frac{e^{-\lambda_i}\lambda_i}{n_i!}, \tag{1}$$

λ_i jest parametrem Poissona równym wartości oczekiwanej liczby wypadków $E(n_i)$ w segmencie i. Regresja Poissona zakłada, że logarytm z parametru λ_i jest funkcją liniową zmiennych objaśniających $\lambda_i = \exp\left(\sum_j \beta_j X_{ji}\right)$. Dla tak zdefiniowanego rozkładu zachodzi $\mathrm{var}(Y_i) = \lambda_i = E(Y_i)$, co oznacza, że wariancja rozkładu powinna być równa jego wartości oczekiwanej. Dla danych empirycznych ten warunek zazwyczaj nie jest spełniony. Aby uniknąć tego ograniczenia wprowadzono rozszerzenie do funkcji definiującej parametr Poissona, w postaci dodania składnika losowego ε_i o zadanym rozkładzie. Zatem $\lambda_i = \exp\left(\sum_j \beta_j X_{ji} + \varepsilon_i\right)$.

Jeśli e^{ε_i} ma rozkład gamma otrzymany model nazywany jest gamma Poissona – „negative binomial" - NB (Maycock, Hall 1984; Hauer i in. 1988), jeśli natomiast e^{ε_i} ma rozkład log-normalny ($\varepsilon_i \sim N(0, \sigma_u^2)$), model nazywany jest lognormalnym-Poissona - PLN (Miaou i in. 2003, Miaou i in 2005).
Funkcja gęstości prawdopodobieństwa w modelu NB jest postaci:

$$P(Y_i = n_i \mid \lambda_i, \alpha) = \frac{\Gamma\left(n_i + \frac{1}{\alpha}\right)}{n_i!\,\Gamma\left(\frac{1}{\alpha}\right)} \cdot \left(\frac{\frac{1}{\alpha}}{\frac{1}{\alpha} + \lambda_i}\right)^{\frac{1}{\alpha}} \cdot \left(\frac{\lambda_i}{\frac{1}{\alpha} + \lambda_i}\right)^{\lambda_i} \tag{2}$$

Parametr α nazywany jest parametrem dyspersji (rozproszenia). Dla tego rozkładu wartość oczekiwana nadal równa jest parametrowi Poissona $\lambda_i = E(n_i)$, przy czym wariancja może być od niej różna (większa) i przedstawia ją zależność

$$\mathrm{var}(n_i) = \lambda_i + \alpha \cdot \lambda_i^{\,2}. \tag{3}$$

Dla parametru rozproszenia $\alpha = 0$ rozkład NB sprowadza się do klasycznego rozkładu Poissona.

Parametr dyspersji α na podstawie wzoru (3) można zapisać jako:

$$\alpha = \frac{\dfrac{\mathrm{var}(n_i)}{E(n_i)} - 1}{E(n_i)}. \tag{4}$$

Na podstawie wyestymowanej wartości parametru dyspersji oraz wartości oczekiwanej wyznaczono wartość wariancji z modelu i na jej podstawie określono podział zmienności na systematyczną oraz losową oraz określono „skuteczność" modelowania poprzez wyznaczenie udziału wariancji wyjaśnionej przez model.

2.2 Estymacja parametrów

Do wyznaczenia wartości parametrów modelu zastosowano najpowszechniej znaną oraz stosowana ze względu na łatwość oraz szybkość obliczeń metodę maksimum prawdopodobieństwa (maksimum likelihood method) z dyskretnym algorytmem Newtona (Winkelmann, 2008). W niniejszej pracy estymację parametrów modelu wykonano z zastosowaniem pakietu ekonometrycznego LIMDEP.

2.3 Ocena dopasowania

Aby ocenić dopasowanie modelu do danych empirycznych stosuje się wiele różnych metod. Najczęściej do porównania dopasowania różnych modeli stosowane są miary dopasowania w postaci wskaźników. Do oceny stopnia dopasowania modelu nie wystarczy jeden wskaźnik (Lord and Park, 2008). W pracy zastosowano 5 miar dopasowania, które opisano poniżej:

- Akaike's Information Criterion (AIC)

$$AIC = \frac{-2\ln L(M_j) - 2k}{N} \tag{5}$$

gdzie $\ln L(M_j)$ jest wartością logarytmu funkcji prawdopodobieństwa (log-likelihood value) modelu *j*, *k* jest liczbą parametrów oraz N – liczbą obserwacji (tutaj $N = 1671$).

- Bayesian Information Criteria (BIC)

$$BIC = D(M_j) + k \cdot \ln N = -2 \ln L(M_j) + k \cdot \ln N \tag{6}$$

gdzie $D(M_j)$ jest dewiacją modelu M_j

- Mean absolute deviation (MAD)

$$MAD = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i| \tag{7}$$

gdzie $\hat{y}_i$ jest przewidywaną liczbą wypadków w segmencie *i*, y_i jest obserwowaną liczbą wypadków w segmencie *i*.

- Mean squared prediction error (MSPE)

$$MSPE = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \tag{8}$$

- R^2

$$R^2 = \frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{9}$$

gdzie $\bar{y}$ jest średnią, obserwowaną liczbą wypadków.

Model jest lepiej dopasowany kiedy wartości wskaźników są mniejsze.

3. DANE EMPIRYCZNE

Dane wykorzystane w niniejszej pracy, udostępnione autorowi dzięki uprzejmości Rune Elvik'a z The Institute of Transport Economics w Oslo, zawierają informacje o 1671 sekcjach obejmujących odcinki dróg krajowych na terenie jednostki administracyjnej Hordaland w Norwegii. Dane dotyczące liczby wypadków drogowych w sektorach pochodzą z okresu 1993-2000. Analiza została wykonana dla całkowitej liczby odnotowanych wypadków. Dane obejmują 8 zmiennych opisujących charakterystyki geometryczne drogi, natężenie ruchu i dodatkowe informacje. W rozważanych segmentach odnotowano w badanym okresie 3175 wypadków, 113 (3,56%) z nich było śmiertelnych, w 790 (47%) segmentach nie odnotowano ani jednego wypadku w okresie 8 lat. Tabela 1 przedstawia podstawowe charakterystyki statystyczne rozważanych zmiennych.

Tab.1. Charakterystyki statystyczne zmiennych.

Nazwa zmiennej	Średnia	Odchylenie stand.	Min.	Max.
Zmienna zależna				
Liczba wypadków	1,90	4,44	0	55
Zmienne niezależne				
Średnie dzienne natężenie ruchu /1000 (AADT/1000)	2,96	4,80	0,04	53,15
Logarytm naturalny z minimalnej liczby pasów ruchu	1,01	0,06	0,69	1,61
Droga krajowa (1 jest TAK, 0 jeśli NIE)	0,26	0,44	0	1
Ograniczenie prędkości 50 km/h (1=tak, 0=p.p.)	0,15	0,35	0	1
Ograniczenie prędkości 60 km/h (1=tak, 0=p.p.)	0,15	0,36	0	1
Ograniczenie prędkości 70 km/h (1=tak, 0=p.p.)	0,02	0,13	0	1
Długość segmentu w iloczynie z czasem pomiaru w latach (km·rok)	7,51	1,08	2,7	8
Logarytm naturalny z liczby skrzyżowań + 1	0,27	0,50	0	2,71
Liczba obserwacji		1671		

4. WYNIKI

W celu dokonania wyboru modelu przeprowadzono testy chi kwadrat (χ^2) dla: klasycznego modelu Poissona oraz modelu ujemnego dwumianowego - negative binomial model (NB). Wyniki przedstawiono w tab.2.

Tab.2. Rozkład liczby sektorów dla liczby odnotowanych wypadków w trzech wariantach modeli

Liczba wypadków	Liczba sektorów		
	data	Poisson	NB
0	790	250	835
1	373	475	272
2	175	451	157
3	114	286	104
4	57	136	73
5	36	52	53
6	27	16	40
7	13	4	30
8	10	1	23
9	9	0	18
10	4	0	14
11	9	0	11
12	6	0	8
>12	48	0	32
χ^2		8764,3	233,7

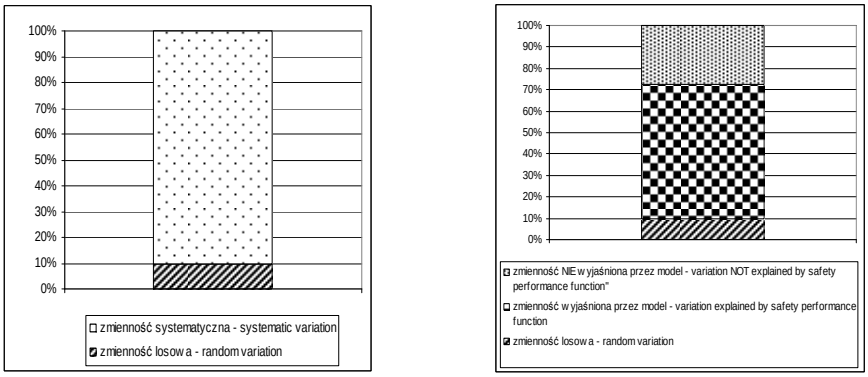
Na podstawie dokonanej analizy jako lepiej oddający rozkład liczby wypadków wybrano model NB. Wyniki estymacji parametrów modelu przedstawiono w tabeli 3.

Tab.3. Wartości oszacowanych parametrów wraz z błędem oszacowania oraz miary dopasowania dla liczby wypadków

Nazwa zmiennej	współczynnik	Błąd stand.
Stała	0,118	
AADT/1000	0,152	
Logarytm z minimalnej liczby pasów ruchu	-1,787	
Droga krajowa (1 jest TAK, 0 jeśli NIE)	0,171	
Ograniczenie prędkości 50 km/h (1=tak, 0=p.p.)	0,609	
Ograniczenie prędkości 60 km/h (1=tak, 0=p.p.)	0,430	
Ograniczenie prędkości 70 km/h (1=tak, 0=p.p.)	0,552	
Długość segmentu w iloczynie z czasem pomiaru	0,155	
Logarytm naturalny z liczby skrzyżowań + 1	0,430	
Alfa	0,662	
Miary dopasowania		
AIC	3,0	
BIC	5001	
MAD	2,2	
MSPE	14,3	
R^2	0,73	

Ze względu na dość złożoną postać funkcji prawdopodobieństwa (1) wystąpienia zadanej liczby wypadków oraz iloczynowo-wykładniczą postać parametru Poissona nie jest możliwa interpretacja ilościowa wartości współczynników otrzymywanych w procesie modelowania. Możliwa jest jedynie ocena jakościowa wpływu poszczególnych czynników na liczbę wypadków. Na podstawie tabeli 3 wyraźnie widoczny jest wpływ ograniczenia prędkości na liczbę wypadków. W każdym przypadku dodatnie wartości współczynnika oznaczają, że są to obszary zwiększonego ryzyka wystąpienia wypadku. Dodatnie są również współczynniki związane z natężeniem ruchu (AADT). Zwiększenie natężenia ruchu powoduje zwiększenie liczby wypadków. Podobnie liczba skrzyżowań jest czynnikiem, którego wzrost powoduje wzrost liczby wypadków. Długość odcinka oraz czas przez jaki zliczane są wypadki w sektorze w sposób oczywisty jest dodatnio skorelowana z ich liczbą. Redukujący wpływ na liczbę wypadków ma, wśród analizowanych zmiennych objaśniających, jedynie minimalna liczba pasów ruchu. Oznacza to, że zwiększenie liczby pasów jezdni znacznie zwiększa bezpieczeństwo.

Na podstawie wzoru (4) wyznaczono udział poszczególnych rodzajów zmienności (rys.1.)



a) b)

Rys.1. Rozdział wariancji a) w danych empirycznych, b) po modelowaniu

Wariancja systematyczna stanowi 90,4% całkowitej zmienności liczby wypadków drogowych. Zastosowany model pozwolił opisać - wyjaśnić 69,4% zmienności systematycznej, co oznacza, że zastosowanie modelu do szacowania liczby wypadków drogowych według przedstawionej metodyki obarczone jest jedynie 27,6% nieopisaną przez model zmiennością losową (residua).

5. PODSUMOWANIE I WNIOSKI

Modelowanie liczby wypadków drogowych jest bardzo ważną częścią badań nad bezpieczeństwem ogólnie pojętego transportu. W pracy, na podstawie danych pochodzących z Norwegii, wykazano, że do modelowania częstości wypadków drogowych model ujemny dwumianowy Poissona dobrze opisuje liczbę wypadków drogowych w sekcjach. Na podstawie przeprowadzonego modelowania można stwierdzić, że zwiększenie liczby pasów jezdni znacznie zmniejsza liczbę wypadków drogowych. Pozostałe rozpatrywane zmienne: natężenie ruchu, ograniczenie prędkości, liczba skrzyżowań w segmencie, rodzaj drogi oraz długość segmentu są pozytywnie skorelowane z liczbą wypadków drogowych, czyli powodują zwiększenie ich liczby. Analiza zmienności występującej w danych empirycznych oraz wartościach otrzymanych z modelu pozwala stwierdzić, że model wyjaśnia niemal 70% wariancji systematycznej, co stanowi 60,8% całkowitej zmienności liczby wypadków drogowych w poszczególnych sekcjach.

6. BIBLIOGRAFIA

[1] Hauer, E., Bamfo, J., 1997: *Two tools for finding what function links the dependent variable to the explanatory variables.* In: Proceedings of the ICTCT 1997 Conference, Lund, Sweden.

[2] Hauer, E., Ng, J.C.N., Lovell, J., 1988: *Estimation of safety at signalized intersections.* Transportation Research Record 1185, 1-10.

[3] Joshua, S.C., Gerber, N.J., 1990: *Estimating truck accident rate and envolvements using linear and Poisson regression models.* Transportation Planning and Technology 15(1), 41-58.

[4] Jovanis, P.P., Chang, H.L., 1996: *Modeling the relationship of accidents to Miles traveled.* Transportation Research Record 1068, 42-51.
[5] Lord, D., Mannering F., 2010: *The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives*. Transportation Research Par A 44, 291-305.
[6] Lord, D., Park, P.Y-J., 2008: *Investigating the effects of the fixed and varying dispersion parameters of Poisson-gamma models on empirical Bayes estimates*. Accident Analysis and Prevention 40, 1441-1457.
[7] Maycock, G., Hall, R.D., 1984: *Accident at 4-Arm Roundabouts. TRRL Laboratory Report 1120*, Transportation and Road Research Laboratory, Crowthorne, GB.
[8] Miaou S.-P., Bligh, R.P., Lord,D., 2005. *Developing median barrier installation guidelines: a benefit/cost analysis using Texas data*. Transportation Research Record 1984, 3-19.
[9] Miaou, S.-P., Song,J.J., Mallick, B.K., 2003.: *Roadway traffic crash mapping: a space-time modeling approach*. Journal of Transportation and Statistics 6(1), 33-57.
[10] Winkelmann, R., 2008.: *Econometric Analysis of Count Data. Fifth edition*, Springer, ISBN 978-3-540-77648-2.