Maszynopis artykułu: Marzec J., Wybrane dwuwymiarowe modele dla zmiennych licznikowych w ekonomii, [w:] Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie – Metody Analizy Danych" nr 884, Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków.

Jerzy Marzec
Adres e–mail: marzecj@uek.krakow.pl
Uniwersytet Ekonomiczny w Krakowie
Katedra: Katedra Ekonometrii i Badań Operacyjnych

Wybrane dwuwymiarowe modele dla zmiennych licznikowych w ekonomii[1]

1. Wstęp

Rozkład Poissona, ujemny rozkład dwumianowy i rozkład logarytmiczny są powszechnie używane do opisu zjawisk, gdy zmienna objaśniana jest mierzona na skali ilorazowej i jednocześnie przyjmuje wyłącznie wartości nieujemne. Pierwsze literaturowo udokumentowane zastosowanie rozkładu Poissona pochodzi z końca XIX wieku, a dotyczy badań przeprowadzonych przez rosyjskiego uczonego polskiego pochodzenia Władysława Bortkiewicza. W 1898 r. opublikował on wyniki analiz dotyczących śmiertelności żołnierzy, służących w dziesięciu korpusach kawalerii armii pruskiej w latach 1875-1894, spowodowanej kopnięciami przez konie. Początkowo rozkład Poissona był proponowany do probabilistycznego opisu rozkładu zdarzeń rzadkich, które pojawiają się z małym prawdopodobieństwem w ciągu nieskończonej liczby niezależnych powtórzeń tego samego doświadczenia o dwóch możliwych wynikach. Obecnie w literaturze naukowej istnieją liczne aplikacje z wykorzystaniem tego rozkładu w analizie danych przekrojowych, szeregów czasowych i danych longitudinalnych. W ekonomii modele Poissona znajdują zastosowanie także wówczas, gdy przedmiotem zainteresowania są wyniki zachowań jednostek podejmujących decyzje. Obok modeli dyskretnego wyboru są one podstawowymi narzędziami opisu zjawisk rozważanych na gruncie mikroekonometrii. Lista zastosowań regresji Poissona jest długa. W ekonomice zdrowia stosuje się je m.in. w analizie intensywności korzystania z różnych form usług opieki zdrowotnej, liczby wypadków w pracy lub chorób zawodowych. W zakresie ekonomiki pracy modele te służą badaniu absencji

[1] Artykuł powstał w ramach badań statutowych finansowanych przez Uniwersytet Ekonomiczny w Krakowie. Autor pragnie podziękować za merytoryczną dyskusję i cenne uwagi, które otrzymał od uczestników IV Konferencji Naukowej „Metody Ilościowe w Ekonomii" zorganizowanej przez WSB we Wrocławiu oraz podczas otwartych zebrań Katedry Ekonometrii i Badań Operacyjnych UEK w Krakowie.

w miejscu pracy i mobilności zawodowej ludności w wieku produkcyjnym. W ubezpieczeniach znajdują zastosowanie w analizie szkodowości w portfelu ubezpieczeń komunikacyjnych lub majątkowych. W transporcie modele zmiennych licznikowych umożliwiają ocenę intensywności wypadków komunikacyjnych. W bankowości zaś są jednym z narzędzi analizy spłacalności rat kapitalowo-odsetkowych przez kredytobiorców i pomiaru ryzyka kredytowego. W zakresie nauk o przedsiębiorstwach modele regresji Poissona mogą być dogodnym sposobem opisu zależności zmian koniunkturalnych (faz wzrostu bądź spadku gospodarczego) na skalę bankructwa przedsiębiorstw. W marketingu przedmiotem badania mogą być decyzje pojedynczych konsumentów dotyczące jednoczesnych zakupów określonej ilości różnych produktów lub usług. W demografii są podstawowym narzędziem opisu zjawisk dotyczących np. umieralności czy prokreacji. C. Cameron i P. Trivedi [1998, 2005] przedstawiają wyniki przykładowych badań empirycznych pochodzących z różnych dziedzin ekonomii, które zostały uzyskane za pomocą szczegółowych modeli danych licznikowych.

Celem niniejszego artykułu jest przedstawienie wybranej klasy modeli statystycznych opisujących zależność między dwiema zmiennymi licznikowymi. W literaturze są one określane terminem dwuwymiarowych modeli Poissona (ang. *bivariate Poisson models*). Przedmiotem zainteresowania w szczególności są te konstrukcje, w ramach których możliwa jest analiza zjawisk charakteryzujących się zarówno dodatnią jak i ujemną korelacją między zmiennymi endogenicznymi. Opracowanie ma zatem charakter przeglądowy.

W artykule rozważamy dwie propozycje. Pierwsza z nich to mieszanka rozkładu Poissona z rozkładem log-normalnym, która została zaproponowana przez J. Aitchisona i C. Ho w 1989 r. Druga specyfikacja nazwana przez P. Berkhouta i E. Pluga [2004] warunkowym modelem Poissona, jest prostszą konstrukcją. Posiada pewne restrykcje dotyczące własności, ale ma także zalety, m.in. łatwość estymacji w przeciwieństwie do modelu pierwszego.

2. Prosty model regresji Poissona

Rozważamy model statystyczny dla licznikowej zmiennej losowej Y, która przyjmuje wartości ze zbioru liczb całkowitych nieujemnych. Niech Y ma rozkład Poissona z parametrem λ, Y~Poisson(λ), więc funkcja prawdopodobieństwa ma postać

$$\Pr(Y = y) = p_Y(y|\lambda) = \frac{1}{y!}\exp(-\lambda)\cdot\lambda^y, \quad \lambda > 0, \tag{1}$$

gdzie *y* oznacza pojedynczą realizację tej zmiennej, która przyjmuje wartości 0,1,2,3,…. Parametr λ jest jednocześnie jej wartością oczekiwaną ($E(Y)$) i wariancją Y ($Var(Y)$). Zauważmy, że powyższy rozkład ma dużo ograniczeń, gdyż jeden parametr definiuje wszystkie momenty tej zmiennej. W zastosowaniach ekonometrycznych, gdy dostępne są dodatkowe informacje o badanych jednostkach w postaci *k*-wymiarowego wektora zmiennych objaśniających x_t, można osłabić założenie o homoscedastyczności. W przypadku próby prostej y_t (dla t=1,…,T) przyjmuje się, że

$$\lambda_t = \exp(x_t \cdot \beta), \tag{2}$$

gdzie β jest k-elementowym wektorem nieznanych parametrów informującym o kierunku i sile oddziaływania zmiennych objaśniających na charakterystyki rozkładu obserwowanej zmiennej. Szerzej o konstrukcji i własnościach tego modelu oraz podstawowej metodzie estymacji jego parametrów – metodzie największej wiarygodności (MNW) – piszą m.in. C. Cameron i P. Trivedi [1998, 2005]. Kolejnym krokiem w kierunku uogólnienia powyższej konstrukcji – osłabienia założenia o równości wartości oczekiwanej i wariancji – jest zastosowanie mieszanki rozkładów Poissona-gamma, która jest równoważna przyjęciu dla zmiennej *Y* dwuparametrycznego ujemnego rozkładu dwumianowego; zob. np. C. Cameron i P. Trivedi [1998, 2005].

3. Dwuwymiarowy model Poissona z korelacją dodatnią

W tej części przedstawimy podstawową klasę rozkładów dwuwymiarowych zmiennych licznikowych. Rozważamy dwuwymiarową zmienną losową $Y = [Y_1\ Y_2]$, czyli parę zależnych zmiennych. W monografii S. Kocherlakota i K. Kocherlakota [1992] znajdziemy różne propozycje rozkładów tej zmiennej, od najprostszych do bardzo złożonych. Jednakże w ramach standardowych koncepcji można modelować wyłącznie te zjawiska, które charakteryzują się nieujemną korelacją. Najczęściej rozważa się zmienną dwuwymiarową, której rozkład jest określony przez łączną funkcję prawdopodobieństwa postaci

$$\Pr(Y_1 = y_1, Y_2 = y_2) = \exp(-\lambda_1 - \lambda_2 - \lambda_3) \sum_{r=0}^{\min(y_1, y_2)} \frac{\lambda_1^{y_1 - r}}{(y_1 - r)!} \cdot \frac{\lambda_2^{y_2 - r}}{(y_2 - r)!} \cdot \frac{\lambda_3^{r}}{r!}, \tag{3}$$

gdzie parametry λ_1, λ_3 i λ_3 są dodatnie.

Historia tego rozkładu sięga lat 30 ubiegłego stulecia. Powyższy model jest konstrukcją czysto teoretyczną, nie mającą bezpośredniej interpretacji w zjawiskach empirycznych. Jednakże rozkład ten jest granicznym przypadkiem dwuwymiarowego rozkładu

dwumianowego. Ponadto funkcja prawdopodobieństwa określona formułą (3) reprezentuje łączny rozkład zmiennych będących sumą dwóch spośród trzech jednowymiarowych zmiennych o rozkładach Poissona. Jeżeli V_1, V_2 i V_3 są niezależnymi jednowymiarowymi zmiennymi o rozkładach opisanych parametrami λ_1, λ_2 i λ_3, to rozkład pary zmiennych $Y_1 = V_1+V_3$ i $Y_2 = V_2+V_3$ jest zdefiniowany wzorem (3). Zatem łatwo zauważyć, że wektor wartości oczekiwanych zmiennej Y składa się z elementów odpowiednio $E(Y_1) = \lambda_1 + \lambda_3$ i $E(Y_2) = \lambda_2 + \lambda_3$. Brzegowe wariancje tej zmiennej dwuwymiarowej są równe wartościom oczekiwanym. Ponadto, kowariancja zmiennych Y_1 i Y_2 jest równa $\text{cov}(Y_1, Y_2) = Var(V_3) = \lambda_3$, więc współczynnik korelacji dany jest wzorem

$$corr(Y_1, Y_2) = \frac{\lambda_3}{\sqrt{(\lambda_1 + \lambda_3)(\lambda_2 + \lambda_3)}}. \qquad (4)$$

Przyjmuje on wartości wyłącznie dodatnie i jest ograniczony od góry przez wartość $\lambda_3/(\lambda_3 + \min(\lambda_1, \lambda_2))$. Własności te są mocno restrykcyjne, co za tym idzie zastosowanie tego rozkładu w badaniach empirycznych jest ograniczone. Dodatkowe informacje o tej klasie modeli, a dotyczące m.in. formuł momentów wyższych rzędów rozkładu łącznego i warunkowego można znaleźć u Kocherlakota i Kocherlakota [1992]. Analogicznie w oparciu o powyższą koncepcję definiuje się rozkłady wielowymiarowych wektorów zmiennych losowych. Niestety wspomniane wady pozostają.

Powyższy model znalazł zastosowanie w marketingu. T. Brijs i in. [2004] zaprezentowali badania dotyczące zależności między zakupami wybranych produktów wykorzystywanych w kuchni. W badanym koszyku znalazły się sól zmiękczająca i detergent oraz dwa podstawowe składniki do wypieków domowych. Inna aplikacja pochodzi ze statystyki sportu. D. Karlis i I. Ntzoufras [2003] analizowali zależność między liczbą zdobytych i straconych goli w meczach piłki nożnej i wodnej. Natomiast K. Kockelman i J. Ma [2006] zastosowali powyższy model w przypadku badań dotyczących liczby osób, które uczestniczyły w wypadkach drogowych i mogły ponieść uszczerbek na zdrowiu. Rozważali trzy sytuacje: brak jakichkolwiek obrażeń, doznanie ciężkich obrażeń ciała i obrażenia śmiertelne.

4. Dwuwymiarowe modele z korelacją dodatnią lub ujemną

Jak wcześniej wspomniano powyższy model statystyczny zakłada wyłącznie dodatnią korelację. Propozycji rozkładów, które dopuszczają zarówno korelację dodatnią jak i ujemną

jest niewiele. Można je uzyskać wykorzystując funkcję kopula (zob. np. Ophem [1999]) lub mieszanki rozkładów. D. Karlis i E. Xekalaki [2005] zaprezentowali przegląd jedno i dwuwymiarowych mieszanek rozkładów Poissona. Ponadto zwracają uwagę, że tylko w przypadku pewnych mieszanek (tzw. drugiego rodzaju) można rozważać korelację ujemną. Przykładem takiego modelu o strukturze hierarchicznej jest mieszanka z wielowymiarowym rozkładem log-normalnym dla skorelowanych efektów losowych w równaniach definiujących parametry lambda [Aitchison i Ho 1989, Chib i Winkelmann 2001]. Niestandardowym modelem z dodatnią lub ujemną korelacją jest tzw. warunkowy model Poissona zaproponowany przez Berkhouta i Pluga [2004]. Jest to prostsza specyfikacja, ale dogodniejsza w estymacji. W niniejszym artykule omówimy obie propozycje.

Model Poissona log-normalny

Propozycja Aitchison i Ho [1989] polega na zbudowaniu rozkładu łącznego dla wektora losowego poprzez zastosowanie nieskończonej mieszanki w formie wielowymiarowego rozkładu log-normalnego ze złożoną strukturą korelacyjną. Rozkład mieszający jest określony dla efektów losowych występujących w równaniach definiujących wartości oczekiwane zmiennych obserwowanych o niezależnych rozkładach Poissona. W przypadku dwuwymiarowej zmiennej losowej łączna funkcja prawdopodobieństwa dla wektora $Y = [Y_1\ Y_2]$ ma postać[2]

$$\Pr(Y_1 = y_1, Y_2 = y_2) = \int_{R^+}\int_{R^+} p_{Y_1}(y_1|\lambda_1)\cdot p_{Y_2}(y_2|\lambda_2)\cdot p_{MLN}(\lambda|\mu_\lambda, \Sigma)d\lambda_1\, d\lambda_2, \qquad (5)$$

gdzie $p_{LnN}(\lambda|\,\mu_\lambda, \Sigma)$ jest funkcją gęstości zmiennej losowej $\lambda=[\lambda_1\ \lambda_2]$ o dwuwymiarowym rozkładzie log-normalnym z parametrem położenia $\mu_\lambda = [\mu_\lambda 1\ \mu_{\lambda 2}]$ i rozproszenia $\Sigma = [\sigma_{ij}]$ dla $i,j = 1, 2$. Wybór rozkładu log-normalnego jako rozkładu mieszającego był uzasadniony z dwóch powodów: a) charakteryzuje się silną prawostronną asymetrią, więc oddaje charakter rozkładu dla zmiennej obserwowanej, b) jego własności w przypadku wielowymiarowym i związki z innymi rozkładami są dobrze określone.

Mimo niejawnej postaci funkcji (5) w przypadku niediagonalnej macierzy Σ Aitchison i Ho [1989] analitycznie wyznaczyli podstawowe charakterystyki zmiennej Y. Dysponujemy próbą $\{y_{t1}, y_{t2}\}$ dla $t=1,\ldots,T$. Kluczowy parametr – współczynnik korelacji, w tym przypadku wynosi

[2] Dla prostoty zapisu pominięto indeks t.

$$corr(Y_{t1}, Y_{t2}) = \frac{\exp(\sigma_{12}) - 1}{\sqrt{\left(\exp(\sigma_{11}) - 1 + \mu_{t1}^{-1}\right) \cdot \left(\exp(\sigma_{22}) - 1 + \mu_{t2}^{-1}\right)}}, \qquad (6)$$

gdzie $E(Y_{ti}) = \mu_{ti} = \exp(\mu_{\lambda i} + \sigma_{ii}/2)$ dla i=1, 2 oraz t=1,…,T. Ponadto wariancja jest równa $V(Y_{ti}) = \mu_{ti} + (\mu_{ti})^2 (\exp(\sigma_{ii}) - 1)$, a więc występuje tzw. zwiększenie rozproszenia (ang. *overdispersion*), co jest naturalne w przypadku zjawisk opisywanych przez model zmiennej skokowej, która przyjmuje wartości ze zbioru liczb naturalnych. Przypadek $\sigma_{12} = 0$ odpowiada nieskorelowanym zmiennym losowym, a $\sigma_{12} < 0$ ($\sigma_{12} > 0$) korelacji ujemnej (dodatniej). Oczywiście macierz Σ jest symetryczna i dodatnio określona. Liczba nieznanych parametrów wynosi pięć. Ponadto ze wzoru (6) i faktu, że współczynnik korelacji zmiennych o dwuwymiarowym rozkładzie log-normalnym jest równy $corr(\lambda_{t1}, \lambda_{t2}) = \left(e^{\sigma_{12}} - 1\right) \Big/ \sqrt{\left(e^{\sigma_{11}} - 1\right)\left(e^{\sigma_{22}} - 1\right)}$ wynika, iż $|corr(Y_{t1}, Y_{t2})| < |corr(\lambda_{t1}, \lambda_{t2})|$. Wniosek ten jest zgodny z intuicją, gdyż zmienne Y_{t1} Y_{t2} są pośrednio ze sobą zależne poprzez skorelowanie λ_{t1} i λ_{t2}. Zaletą powyższego modelu jest fakt, że a) można go zdefiniować w przypadku dowolnej liczby zmiennych zależnych, b) można mu nadać interpretację w nawiązaniu do rzeczywistych zjawisk przyrodniczych czy społecznych.

Powyższy model ma strukturę hierarchiczną. Zmienne Y_{t1} i Y_{t2} posiadają niezależne rozkłady Poissona z parametrami λ_{t1} i λ_{t2} warunkowe względem zmiennej $u_t = [u_{t1}\ u_{t2}]$. Na drugim szczeblu zdefiniowany jest układ równań regresji dla warunkowych wartości oczekiwanych λ_{t1} i λ_{t2} z multiplikatywnym składnikiem u_t o dwuwymiarowym rozkładzie normalnym z zerową wartością oczekiwaną i macierzą kowariancji Σ_u. W efekcie równoważna postać modelu jest następująca

$$\begin{cases} Y_{ti} | u_t \sim Poisson(\lambda_{ti}) & \text{dla} \quad i = 1, 2 \\ \lambda_{ti} = \exp(x_t \cdot \beta_i + u_{ti}) \\ u_t \sim N^{(2)}(0, \Sigma_u). \end{cases} \qquad (7)$$

Z zależności między wielowymiarowym rozkładem log-normalnym a wielowymiarowym rozkładem normalnym wynikają relacje pomiędzy elementami μ_λ i Σ a macierzą kowariancji Σ_u. Dodatkowo, po wprowadzeniu wektora zmiennych objaśniających x_t, otrzymuje się $\mu_{ti} = E(Y_{ti}) = \exp(x_t \cdot \beta_i + \sigma_{ii}/2)$. Parametrami tego modelu są β_1, β_2 i trzy swobodne elementy dodatnio określonej macierzy kowariancji Σ_u. Obie specyfikacje (5) i (7) tego modelu posiadają równoważną parametryzację.

Aitchison i Ho [1989] przedstawili ciekawą interpretację mieszanki danej wzorem (5). Rozważają wektor *d*-wymiarowy zmiennych skokowych, które reprezentują liczebności różnych gatunków motyli, odżywiających się nektarem kwiatowym z *d*-gatunków roślin rosnących na łące. Każdy gatunek motyli żeruje na swoim ulubionym gatunku roślin kwiatowych, więc nie konkurują między sobą o pożywienie. Czy można zatem oczekiwać, że zmienne losowe reprezentujące liczebności poszczególnych gatunków motyli są niezależne? Tak, ale gdy przyjmiemy, że rozważamy rozkład warunkowy względem ustalonej (danej) obfitości (ilość) roślin kwiatowych. Liczebność tych ostatnich może być dodatnio skorelowana, gdy rośliny konkurują ze sobą o dostęp do słońca, wody i minerałów zawartych w glebie. W innym przypadku, na wskutek działania czynników atmosferycznych (pogodowych) liczebności kwiatów i owoców różnych gatunków roślin są skorelowane dodatnio. Zatem zmienna u_t reprezentuje m. in. warunki pogodowe i konkurencyjność między poszczególnymi gatunkami roślin. Te czynniki środowiskowe są związane wyłącznie z miejscem życia motyli i łącznie oddziałują na przeciętne liczebności wszystkich badanych gatunków. Taka hierarchiczna zależność może więc być opisana modelem przywołanym powyżej. Na gruncie ekonomii istnieją analogiczne przykłady, np. dotyczące liczby przyjętych studentów na I rok studiów różnych uczelni. W wybranym mieście, np. Krakowie, znajdują się szkoły wyższe o odmiennym profilu, tj. technicznym, ekonomicznym, rolniczym, humanistycznym, teologicznym i artystycznym. Załóżmy, że studiowanie na każdej z tych uczelni jest tak samo trudne. Uczelnie te nie konkurują ze sobą, gdyż kształcą i przygotowują studentów do pracy w różnych zawodach. Jednakże sumaryczna liczba kandydatów na studia wyższe silnie zależy od liczby maturzystów w danym roku szkolnym. Niż albo wyż demograficzny dotyka wszystkie szkoły i uczelnie bez względu na profil. Ponadto, istnieją czynniki społeczne, które wpływają na zainteresowania młodych ludzi określonymi kierunkami studiów. Często są to moda, wpływ rodziców lub starszych kolegów, perspektywa dobrze płatnej i interesującej pracy. W ostatnim dziesięcioleciu brak matematyki na maturze mógł być barierą wejścia na uczelnie techniczne. Te ukryte czynniki reprezentowane przez skorelowane składniki u_t mają wpływ na liczbę nowoprzyjętych studentów na poszczególne uczelnie o rożnym profilu. Zatem badane zjawisko należy rozważać łącznie, a nie osobno dla każdej ze szkół akademickich.

Zastosowanie powyższego modelu w ekonomice zdrowia zaprezentowali R. Riphahn i in. [2003]. Analizowali oni liczbę wizyt u lekarza i okres pobytu w szpitalu (przynajmniej jedną dobę) w przypadku niemieckich gospodarstw domowych w latach 1984-1995. Z kolei

S. Chib i in. [1998] wykorzystali tą konstrukcję do budowy modelu dla danych panelowych ze skorelowanymi efektami losowymi. Przedmiotem badań była liczba patentów zgłoszonych przez firmy amerykańskie w latach 1975-1979. Chib i Winkelmann [2001] analizowali natomiast popyt na usługi medyczne w przypadku osób starszych. Badali rozkład liczby wizyt pacjentów m.in. u lekarza w przychodni, szpitalu i w izbie pogotowia ratunkowego. Rozważali model dla sześciu zmiennych licznikowych. Kolejna aplikacja powyższego modelu dotyczyła bezpieczeństwa komunikacji. J. Ma i in. [2008] zastosowali go do analizy liczby pięciu rodzajów wypadków drogowych na różnych odcinkach dróg z dwoma pasami ruchu poza terenem zabudowanym w stanie Waszyngton w USA. Przykładem badania z zakresu marketingu w turystyce jest artykuł J. Hellströma [2006]. Na podstawie ankiet przeprowadzonych w szwedzkich gospodarstwach domowych badał on zależność między liczbą wycieczek a liczbą wykupionych noclegów podczas tych podróży.

Niewątpliwą wadą log-normalnego modelu Poissona jest niejawna postać rozkładu próbkowego dla wektora obserwacji *Y*. Obecność całki wielokrotnej we wzorze (5) powoduje trudności estymacyjne. R. Riphahn i in. [2003] zastosowali w tym celu klasyczne metody całkowania numerycznego (kwadratury Gaussa–Hermite'a i Gaussa–Legendre'a), aby następnie procedurą quasi–Newtona poszukać maksimów funkcji wiarygodności. J. Hellström [2006] wykorzystał symulacyjną metodę największej wiarygodności (ang. *simulated maximum likelihood*).

Naturalnym podejściem do estymacji modelu hierarchicznego jest wnioskowanie bayesowskie. Chib i in. [1998] oraz Chib i Winkelmann [2001] zaproponowali bayesowską estymację tej klasy modeli z wykorzystaniem narzędzi MCMC (ang. *Markov Chain Monte Carlo*). Pojawiły się już kolejne artykuły, w których autorzy sięgają po te narzędzia estymacji i wnioskowania statystycznego, zob. np. [Ma i in. 2008].

Warunkowy model Poissona

Inne, prostsze podejście do konstrukcji modelu zmiennych licznikowych proponują Berkhout i Plug [2004]. Rozważali warunkowy model Poissona dla dwóch skorelowanych zmiennych skokowych. Konstrukcja ta dopuszcza ujemną jak i i dodatnią korelację przy zachowaniu prostoty i elegancji. Niech Y_1 i Y_2 będą zależnymi zmiennymi skokowymi. Ich rozkład łączny można przedstawić jako iloczyn rozkładu warunkowego i brzegowego. W przypadku dwóch zmiennych rozważamy dwa modele statystyczne

$$\Pr(Y_1 = y_1, Y_2 = y_2) = g_{Y_1|Y_2}(y_1|y_2) \cdot g_{Y_2}(y_2), \tag{8}$$

$$\Pr(Y_1 = y_1, Y_2 = y_2) = g_{Y_2|Y_1}(y_2|y_1) \cdot g_{Y_1}(y_1). \tag{9}$$

Oba modele nie są równoważne, a zamiana numerów zmiennych nie prowadzi do otrzymania równoważnych konstrukcji statystycznych. Zauważmy, że wraz ze wzrostem liczby zmiennych skokowych rośnie liczba możliwych dekompozycji rozkładu łącznego, która z kolei jest równa liczbie permutacji zbioru złożonego ze numerów tych zmiennych. W przypadku wektora *m* zmiennych otrzymamy *m*! modeli statystycznych. Na gruncie klasycznym (niebayesowskim) może to rodzić problemy z wyborem modelu, który najlepiej opisuje badane zjawisko.

Przejdźmy do przedstawienia, za artykułem Berkhouta i Pluga [2004], założeń dotyczących specyfikacji rozkładów brzegowego i warunkowego oraz wynikających z tego charakterystyk opisujących badane zjawisko, tj. wartości oczekiwanych, wariancji i korelacji zmiennych w rozkładzie łącznym. Rozkład brzegowy dla jednej ze zmiennych np. obserwacji Y_{t1}, jest jednowymiarowym rozkładem Poissona z parametrem λ_{t1}

$$g(y_{t1}) = \frac{1}{y_{t1}!}\exp(-\lambda_{t1}) \cdot (\lambda_{t1})^{y_{t1}}, \text{ gdzie } \lambda_{t1} = \exp(x_t \cdot \beta_1). \tag{10}$$

Kluczową kwestią jest określenie rozkładu dla drugiej zmiennej Y_{t2} pod warunkiem zaobserwowania y_{t1} ($Y_{t1} = y_{t1}$), o którym przyjmuje się, że także jest rozkładem Poissona z parametrem λ_{t2}. Parametr ten jest ciągłą zmienną losową (jak w przypadku nieskończonych mieszanek), bo jest funkcją zmiennej Y_{t1}. Berkhout i Plug [2004] przyjęli, że wspomniany rozkład warunkowy dla Y_{t2} ma postać

$$g(y_{t2}|y_{t1}) = \frac{1}{y_{t2}!}\exp(-\lambda_{t2}) \cdot (\lambda_{t2})^{y_{t2}}, \text{ gdzie } \lambda_{t2} = \exp(x_t \cdot \beta_2 + \alpha \cdot y_{t1}). \tag{11}$$

Z powyższego równania wynika, że parametr α odgrywa kluczową rolę, gdyż jest odpowiedzialny za korelację. Znak tego parametru określa znak współczynnika korelacji między obiema zmiennymi. Zauważmy, że taka konstrukcja rozkładu łącznego zakłada, że rozkład brzegowy zmiennej Y_{t1} jest standardowym rozkładem Poissona, a więc wartość oczekiwana i wariancja są sobie równe. Powyższy model jest konstrukcją, która nie traktuje obu zmiennych symetrycznie. Może to być postrzegane jako wada.

W przypadku modelu statystycznego określonego przez funkcje prawdopodobieństwa (10) i (11) Berkhout i Plug [2004] podali formułę momentu silniowego dla łącznego rozkładu zmiennej dwuwymiarowej Y=[Y_1 Y_2]. Pozwoliło to na wyznaczenie z rozkładu łącznego charakterystyk brzegowego rozkładu zmiennej Y_2 i współczynnika korelacji między Y_1 i Y_2.

Wektor wartości oczekiwanych zmiennej dwuwymiarowej Y składa się z następujących elementów

$$\begin{aligned} E(Y_{t1}) &= \lambda_{t1} \\ E(Y_{t2}) &= \exp(\lambda_{t1} \cdot (\exp(\alpha)-1)) \cdot \exp(x_t \cdot \beta_2). \end{aligned} \tag{12}$$

Wariancje zmiennych Y_1 i Y_2 wynoszą odpowiednio

$$\begin{aligned} Var(Y_{t1}) &= \lambda_{t1} \\ Var(Y_{t2}) &= E(Y_{t2}) + E(Y_{t2})^2 \cdot \left(\exp\left(\lambda_{t1} \cdot (\exp(\alpha)-1)^2\right)-1\right). \end{aligned} \tag{13}$$

Zauważmy, że skoro $\lambda_{t1} > 0$, to wariancja zmiennej Y_2 jest zawsze większa od jej wartości oczekiwanej, $Var(Y_{t2}) > E(Y_{t2})$. Natomiast kluczową charakterystykę zależności między obiema zmiennymi – korelację – opisuje poniższa formuła

$$corr(Y_{t1}, Y_{t2}) = \frac{\lambda_{t1} \cdot E(Y_{t2}) \cdot (e^{\alpha} - 1)}{\sqrt{Var(Y_{t1}) \cdot Var(Y_{t2})}}. \tag{14}$$

Znak współczynnika korelacji zależy od parametru α. Gdy α jest dodatnie (ujemne), to korelacja jest także dodatnia (ujemna). Oczywiście przypadek α=0 oznacza, że kowariancja (licznik wzoru (14)) wynosi zero, więc obie zmienne losowe są nieskorelowane i niezależne.

W odróżnieniu do modelu mieszanki estymacja warunkowego modelu Poissona nie wymaga wyrafinowanych metod. Berkhout i Plug [2004] zaproponowali metodę największej wiarygodności. Podali analityczne formuły równań definiujące ten estymator. Jego własności asymptotyczne nie zostały jeszcze zbadane. Analityczna macierz drugich pochodnych ułatwi optymalizację numeryczną, ale wyznaczenie asymptotycznej macierzy kowariancji estymatora MNW jest utrudnione. Z uwagi na brak symetryczności w traktowaniu obu zmiennych Y_1 i Y_2 pojawia się problem wyboru modelu w kontekście danych. Koncepcja *maximum maximorum* jest nieformalnym rozwiązaniem problemu wyboru między $\Pr(Y_1 = y_1, Y_2 = y_2) = g_{Y_1|Y_2}(y_1|y_2) \cdot g_{Y_2}(y_2)$ a $\Pr(Y_1 = y_1, Y_2 = y_2) = g_{Y_2|Y_1}(y_2|y_1) \cdot g_{Y_1}(y_1)$.

Berkhout i Plug [2004] zastosowali powyższy model w celu zdiagnozowania sposobu spędzania wolnego czasu przez mieszkańców Holandii. Analizą objęto uczestnictwo w wydarzeniach kulturalnych (wizyty w teatrze lub kinie, na koncertach) oraz wizyty w miejscach atrakcyjnych turystycznie (np. w zoo, parku rozrywki, w ekspozycjach). Wspomniany model nie znalazł jeszcze szerszego zastosowania w rzeczywistych badaniach. Jedną z tych nielicznych aplikacji zaprezentowano w opracowaniu Polasik i in. [2011], które przedstawia badanie substytucji między liczbą transakcji gotówką i kartą bankową w

płatnościach detalicznych na podstawie danych z polskiego rynku. Dalsze uogólnienia metodologiczne są prezentowane w artykułach: J. Osiewalski (2012) oraz J. Marzec i J. Osiewalski (2012).

5. Podsumowanie

Oba modele, model Poissona log-normalny i warunkowy model Poissona pozwalają analizować zależności o dodatnim i ujemnym skorelowaniu. Pierwszemu można nadać interpretację w kategoriach obserwowanego zjawiska, drugi stanowi wyłącznie artefakt. Model mieszanki w naturalny sposób opisuje zależności między wieloma zmiennymi endogenicznymi, ale jest trudny w estymacji. Propozycja Berkhouta i Pluga jest z kolei prostsza, ale za cenę restrykcji dotyczących własności rozkładu jednej ze zmiennych. Jednakże uogólnienie modelu warunkowego, tj. rozszerzenie na przypadek wielu zmiennych, jest możliwe, ale rodzi komplikacje. Obie konstrukcje są modelami względem siebie niezagnieżdżonymi, co na gruncie niebayesowskim utrudnia ich testowanie. Estymacja parametrów i wzajemne testowanie obu modeli jest możliwe w ujęciu bayesowskim, co będzie przedmiotem dalszych, pogłębionych badań.

6. Literatura

Aitchison J., C. H. Ho [1989], *The multivariate Poisson-log normal distribution*, "Biometrika", vol. 76 (4), s 643-653.

Berkhout P., E. Plug [2004], *A bivariate Poisson count data model using conditional probabilities*, "Statistica Neerlandica", vol. 58, nr 3, s. 349-364.

Brijs, T., D. Karlis, G. Swinnen, K. Vanhoof, G. Wets, P. Marchanda [2004], *A multivariate Poisson mixture model for marketing applications*, "Statistica Neerlandica", vol. 58, nr3, s. 322–348.

Cameron A.C., P.K. Trivedi [1998], *Regression analysis of count data*, Cambridge University Press, New York.

Cameron A.C., P.L. Trivedi [2005], *Microeconometrics*: *Methods and Application*, Cambridge University Press, New York.

Chib S., R. Winkelmann [2001], *Markov chain monte carlo analysis of correlated count data*, "Journal of Business and Economic Statistics", vol. 19 nr 4, s. 428-435.

Chib, S., E. Greenberg, R. Winkelmann [1998]. *Posterior simulation and Bayes factor in panel count data models*, "Journal of Econometrics", nr 86, s. 33-54.

Greene W. H. 2007, *Correlation in the bivariate Poisson regression model*, "Working Paper Series", Leonard N. Stern School of Business Paper No. ISSN 1547-3651, dostępny w Internecie: http://ssrn.com/abstract=990011, dostęp 30 października 2010 r.

Hellström J. [2006], *A bivariate count data model for household tourism demand*, "Journal of Applied Econometrics", vol. 21 s. 213–226.

Ma J., K. Kockelman, P. Damien [2008], *A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using bayesian methods*, "Accident Analysis and Prevention", nr 40, s. 964-975

Kockelman K., J. Ma 2006], *Bayesian multivariate Poisson regression for models of injury count, by severity*, "Transportation Research Record", nr 1950, s. 24-34.

Karlis D., E. Xekalaki [2005], *Mixed Poisson distributions*, "International Statistical Review", vol. 78, s. 35-58.

Karlis, D., I. Ntzoufras [2003], *Analysis of sports data using bivariate Poisson models*, "Journal of the Royal Statistical Society: Series D", vol. 52 (3), s. 381-393.

Kocherlakota S., K. Kocherlakota [1992], *Bivariate discrete distributions*, Marcel Dekker, New York.

Marzec J, Osiewalski J. (2012), *Dwuwymiarowy model typu ZIP-CP w łącznej analizie zmiennych licznikowych*, Folia Oeconomica Cracoviensia, vol. LIII, w druku.

Ophem van H. [1999], *A general method to estimate correlated discrete random variables*, "Econometric Theory", 15, s. 228-237.

Osiewalski J. (2012), *Dwuwymiarowy rozkład ZIP-CP i jego momenty w analizie zależności między zmiennymi licznikowymi*, [w:] „Spotkania z królową nauk (Księga jubileuszowa dedykowana Profesorowi Edwardowi Smadze)", red. A. Malawski i J. Tatar, Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków 2012, s.147-154.

Polasik M., J. Marzec, P. Fiszeder, J. Górka [2012], *Modelowanie wykorzystania metod płatności detalicznych na rynku polskim*, „Materiały i Studia NBP" nr 265, Warszawa.

Riphahn R.T., A. Wambach [2003], A. Million, *Incentive effects in the demand for health care: A bivariate panel count data estimation*, "Journal of Applied Econometrics", vol. 18, nr 4, s. 387-405.

Winkelmann R. [2008], *Econometric analysis of count data*, Springer-Verlag.

Maszynopis artykułu: Marzec J., Wybrane dwuwymiarowe modele dla zmiennych licznikowych w ekonomii, [w:] Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie – Metody Analizy Danych" nr 884, Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków.

Wybrane dwuwymiarowe modele

dla zmiennych licznikowych w ekonomii

Streszczenie

Głównym celem niniejszego artykułu jest prezentacja wybranych modeli dla dwuwymiarowych zmiennych licznikowych. Omawiane są podstawowe problemy związane z własnościami standardowego dwuwymiarowego rozkładu Poissona, który zakłada wyłącznie dodatnią korelację. Następnie przedstawia się konstrukcję, porównuje się zalety i wady dwóch innych modeli, które uwzględniają zarówno dodatnią jak i ujemną korelację, tj. modelu Poissona log-normalnego i warunkowego modelu Poissona. Prezentuje się także wybrane praktyczne zastosowania tychże modeli w ekonomii.

Słowa kluczowe: dwuwymiarowe modele regresji Poissona, skorelowane zmienne licznikowe, rozkład Poissona log-normalny, warunkowy model Poissona.

Title

A bivariate count data models in the economics

Summary

This paper presents an overview of the selected bivariate count data regressions. We describe the properties of the standard bivariate Poisson distribution and we draw attention that the main limitation of this model is that it assumes only positive correlation between two count variables. Next two alternative approach Poisson lognormal model and conditional Poisson model are presented. We discuss some of their merits and compare the properties of each of the two models which allow for a flexible correlation structure, i. e. both negative and positive value of the correlation coefficient. Many examples demonstrate that one can use these models in various economics areas.

Key words: bivariate Poisson regression models, correlated count variables, Poisson lognormal distribution, conditional Poisson model.