F O L I A O E C O N O M I C A C R A C O V I E N S I A
Vol. LIII (2012) PL ISSN 0071-674X

DWUWYMIAROWY MODEL TYPU ZIP-CP W ŁĄCZNEJ ANALIZIE ZMIENNYCH LICZNIKOWYCH

JERZY MARZEC
Katedra Ekonometrii i Badań Operacyjnych Uniwersytetu Ekonomicznego w Krakowie
e-mail: marzecj@uek.krakow.pl

JACEK OSIEWALSKI
Katedra Ekonometrii i Badań Operacyjnych Uniwersytetu Ekonomicznego w Krakowie
e-mail: eeosiewa@cyf-kr.edu.pl

ABSTRACT

J. Marzec, J. Osiewalski. *Bivariate ZIP-CP type model in the joint analysis of count variables*. Folia Oeconomica Cracoviensia 2012, 53: 5–20.

In the paper a generalization of the Berkhout and Plug (2004) bivariate Poisson regression model is proposed; in the Berkhout and Plug model one of the variables has the marginal Poisson distribution, while the other follows the conditional Poisson distribution. In the new model the marginal distribution is of the ZIP type and has the same parameterization as the hurdle model. Bayesian estimation of the model and the formal Bayesian comparison of its two alternative specifications are presented. The empirical example concerns joint modelling of the number of cash payments and bank card payments in Poland as well as inference on their correlation.

STRESZCZENIE

W pracy zaproponowano uogólnienie modelu dwuwymiarowej regresji poissonowskiej, który wprowadzili Berkhout i Plug (2004), przyjmujący brzegowy rozkład Poissona dla jednej zmiennej i warunkowy rozkład Poissona dla drugiej. W nowym modelu rozkład brzegowy jest typu ZIP i ma taką parametryzację jak w modelu płotkowym. Przedstawiono bayesowską estymację tego modelu i formalne bayesowskie porównanie dwóch alternatywnych jego specyfikacji. Przykład empiryczny dotyczy łącznego modelowania i wnioskowania o korelacji między liczbą płatności kartą i gotówką w Polsce.

KEY WORDS — SŁOWA KLUCZOWE

bivariate Poisson regression model, zero inflated Poisson model, bank card and cash payments

dwuwymiarowe modele regresji Poissona, model Poissona z nadwyżką zer, płatności kartą płatniczą i gotówką.

1. WPROWADZENIE

Regresja poissonowska jest podstawowym modelem analizy zmiennych licznikowych (tj. o wartościach całkowitych nieujemnych). Istnieją jej dwuwymiarowe uogólnienia; niektóre nakładają ograniczenia na korelację między zmiennymi, inne prowadzą do komplikacji natury statystyczno-numerycznej; zob. np. Kocherlakota i Kocherlakota (1992), Winkelman (2008). Na tym tle obiecujący jest model, który zaproponowali Berkhout i Plug (2004) przyjmując brzegowy rozkład Poissona dla jednej zmiennej oraz warunkowy rozkład Poissona dla drugiej (przy ustalonej pierwszej). Model P-CP (*Poisson — conditional Poisson*) jest łatwy w estymacji i dopuszcza korelację różnego znaku (dodatnią albo ujemną), ale znak ten zależy od znaku jednego parametru, a nie od zmiennych objaśniających; zob. także Marzec (2012). Model P-CP został użyty w Polsce do badania zależności między liczbą transakcji dokonywanych gotówką i liczbą transakcji dokonywanych kartą bankową (zob. Polasik, Marzec, Fiszeder, Górka (2012)); wbrew intuicji korelacja między nimi okazała się dodatnia, co skłania do ponowienia badań, ale po rozszerzeniu ograniczonej specyfikacji Berkhouta i Pluga.

Modele regresji dla skokowej zmiennej objaśnianej z nadmierną liczbą zer zostały spopularyzowane przede wszystkim przez artykuł Lamberta (1992). Cameron i Trivedi (1998, 2005) oraz Winkelman (2008) przedstawiają ekonometryczne modele danych licznikowych z przykładami ich zastosowań w ekonomii. W pracy rozważamy uogólnienie modelu P-CP, polegające na zastąpieniu brzegowego rozkładu Poissona pierwszej z dwóch zmiennych rozkładem typu ZIP (*zero inflated Poisson*), przy pozostawieniu warunkowego rozkładu Poissona drugiej zmiennej. Rozkład typu ZIP może mieć uzasadnienie w wielu sytuacjach praktycznych, gdyż bywa tak, że zerowa wartość zmiennej obserwowanej jest jakościowo odmienna od innych wartości. Proponowana w tej pracy parametryzacja odpowiada tzw. modelowi płotkowemu (ang. *hurdle model*). Osiewalski (2012) wprowadził skokowy rozkład dwuwymiarowy, nazwany ZIP-CP (*ZIP — conditional Poisson*) i podał jego momenty. Model dwuwymiarowej regresji poissonowskiej, oparty na rozkładzie ZIP-CP, prowadzi do znaku kowariancji zależnego od wartości zmiennych objaśniających. Głównym celem pracy jest omówienie estymacji i empirycznego wykorzystania tego modelu statystycznego, nazwanego też ZIP-CP (tak, jak leżący u jego podstaw typ rozkładu).

W następnej części pracy przedstawiamy zwięźle rozkłady P-CP i ZIP-CP, skupiając uwagę na momentach rzędu 1 i 2 oraz współczynniku korelacji. W trzeciej omawiamy model statystyczny typu ZIP-CP i jego ujęcie bayesowskie, zaś w czwartej prezentujemy przykład empiryczny.

2. ROZKŁADY P-CP I ZIP-CP ORAZ ZWIĄZEK MIĘDZY ICH MOMENTAMI

Rozważamy łączny rozkład prawdopodobieństwa dwóch zmiennych losowych (Y_1, Y_2) — przyjmujących wartości całkowite nieujemne — i przedstawiamy go następująco:

$$\Pr\{Y_1 = i, Y_2 = j\} = \Pr\{Y_1 = i\} \Pr\{Y_2 = j \mid Y_1 = i\} = g(i)\, h(j,i), \quad (i, j \in N \cup \{0\}). \tag{1}$$

Jeśli rozkład brzegowy zmiennej Y_1 jest rozkładem Poissona o wartości oczekiwanej i wariancji λ_1, a rozkład warunkowy Y_2 przy ustalonej wartości zmiennej Y_1 jest rozkładem Poissona o wartości oczekiwanej i wariancji $\lambda_2 \exp(\alpha Y_1)$, czyli

$$g(i) = \exp(-\lambda_1)(\lambda_1)^i / i!, \quad h(j,i) = \exp[-\lambda_2 \exp(\alpha i)](\lambda_2)^j \exp(\alpha i j) / j!, \tag{2}$$

to mamy rozkład dwuwymiarowy P-CP (*Poisson — conditional Poisson*), który zaproponowali Berkhout i Plug (2004) i uzyskali dla niego wyniki m.in. w postaci następujących momentów:

$$E(Y_2) = \lambda_2 \exp[\lambda_1 (e^{\alpha} - 1)], \tag{3}$$

$$E[(Y_2)^2] = E(Y_2) + [E(Y_2)]^2 \exp[\lambda_1 (e^{\alpha} - 1)^2], \tag{4}$$

$$Var(Y_2) = E(Y_2) + [E(Y_2)]^2 \{\exp[\lambda_1 (e^{\alpha} - 1)^2] - 1\}, \tag{5}$$

$$E(Y_1 Y_2) = \lambda_1 e^{\alpha} E(Y_2). \tag{6}$$

Jeśli $\alpha \neq 0$, to bezwarunkowa wariancja (5) zmiennej Y_2 jest większa od wartości oczekiwanej (3). Zależność między obu zmiennymi sprawia, że brzegowy rozkład zmiennej Y_2 odpowiada empirycznie częstej sytuacji zwiększonej wariancji danych licznikowych. Brzegowy rozkład zmiennej Y_1, czyli rozkład Poissona, nie ma tej właściwości. Jest to pierwszy powód uogólnienia dwuwymiarowego rozkładu P-CP przez wprowadzenie rozkładu ZIP na miejsce brzegowego rozkładu Poissona. Należy też zauważyć, że znak kowariancji między Y_1 i Y_2, czyli znak wyrażenia

$$Cov(Y_1, Y_2) = E(Y_1 Y_2) - E(Y_1)E(Y_2) = \lambda_1 (e^{\alpha} - 1) E(Y_2), \tag{7}$$

zależy jedynie od znaku stałej rzeczywistej α, a nie od wielkości λ_1, λ_2, parametryzowanych głębiej (uzależnianych od zmiennych objaśniających) w statystycznych zastosowaniach tego modelu probabilistycznego. W tej części krótko przedstawimy uogólnienie, które wprowadził Osiewalski (2012), dopuszczające

związek znaku kowariancji i wielkości λ_1, co w analizach statystycznych stwarza możliwość uzmiennienia tego znaku, w zależności od wartości zmiennych objaśniających poziom λ_1.

Obecnie rozważamy inny, ogólniejszy niż (1) przypadek, w którym łączny rozkład prawdopodobieństwa $\Pr^*\{Y_1 = i, Y_2 = j\}$ zmiennych (Y_1, Y_2) o wartościach nieujemnych $(i, j \in N \cup \{0\})$ jest określony przez ten sam warunkowy rozkład Y_2 przy ustalonym Y_1:

$$\Pr^*\{Y_2 = j \mid Y_1 = i\} = h(j,i) = \Pr\{Y_2 = j \mid Y_1 = i\} \tag{8}$$

oraz rozkład brzegowy zmiennej Y_1, który odmiennie niż w (1) traktuje wartość 0:

$$\Pr^*\{Y_1 = i\} = g^*(i) = \begin{cases} \gamma & dla\ i = 0, \\ \dfrac{1-\gamma}{1-g(0)}\, g(i) & dla\ i \in N, \end{cases} \tag{9}$$

gdzie γ jest ustaloną liczbą z przedziału (0, 1), zaś funkcje g i h są takie same jak w (1). Jeśli $\gamma = g(0)$, to $\Pr^*\{Y_1 = i\} = g^*(i) = g(i) = \Pr\{Y_1 = i\}$ i oba rozkłady łączne są identyczne. Jeśli $\gamma \neq g(0)$ a funkcje g i h zadane są wzorami (2), czyli są poissonowskie, to brzegowy rozkład zmiennej Y_1 jest typu ZIP (*Zero Inflated Poisson*), zaś warunkowy dla Y_2 pozostaje rozkładem Poissona. Rozkład taki oznaczamy ZIP-CP, a jego momenty mają ogólną postać

$$E^*(Y_1^m Y_2^n) = (1-g(0))^{-1}[(1-\gamma)E(Y_1^m Y_2^n) + (\gamma - g(0))0^m E(Y_2^n \mid Y_1 = 0)], \tag{10}$$

gdzie wykorzystuje się znaną postać momentów rozkładu P-CP W szczególności:

$$E^*(Y_1) = (1-g(0))^{-1}(1-\gamma)E(Y_1) = (1-g(0))^{-1}(1-\gamma)\lambda_1, \tag{11}$$

$$E^*(Y_1^2) = (1-g(0))^{-1}(1-\gamma)E(Y_1^2) = (1-g(0))^{-1}(1-\gamma)\lambda_1(1+\lambda_1), \tag{12}$$

$$E^*(Y_2) = (1-g(0))^{-1}[(1-\gamma)E(Y_2) + (\gamma - g(0))\lambda_2], \tag{13}$$

$$E^*(Y_2^2) = (1-g(0))^{-1}[(1-\gamma)E(Y_2^2) + (\gamma - g(0))\lambda_2(1+\lambda_2)], \tag{14}$$

$$E^*(Y_1 Y_2) = (1-g(0))^{-1}(1-\gamma)E(Y_1 Y_2) = (1-g(0))^{-1}(1-\gamma)\lambda_1 \exp(\alpha)E(Y_2), \tag{15}$$

$$Var^*(Y_1) = \frac{1-\gamma}{1-g(0)}\lambda_1(1 + \frac{\gamma - g(0)}{1-g(0)}\lambda_1), \tag{16}$$

$$Var^*(Y_2) = \frac{1-\gamma}{1-g(0)}\{Var(Y_2) + \frac{\gamma - g(0)}{1-g(0)}[E(Y_2) - \lambda_2]^2 + \frac{\gamma - g(0)}{1-\gamma}\lambda_2\}, \quad (17)$$

$$Cov^*(Y_1, Y_2) = \frac{1-\gamma}{1-g(0)}\{Cov(Y_1, Y_2) + \frac{\gamma - g(0)}{1-g(0)}\lambda_1[E(Y_2) - \lambda_2]\}, \quad (18)$$

co prowadzi do współczynnika korelacji postaci

$$Corr^*(Y_1, Y_2) = \frac{Cov(Y_1, Y_2) + \frac{\gamma - g(0)}{1-g(0)}\lambda_1[E(Y_2) - \lambda_2]}{\sqrt{\lambda_1(1 + \frac{\gamma - g(0)}{1-g(0)}\lambda_1)\{Var(Y_2) + \frac{\gamma - g(0)}{1-g(0)}[E(Y_2) - \lambda_2]^2 + \frac{\gamma - g(0)}{1-\gamma}\lambda_2\}}}, \quad (19)$$

gdzie $E(Y_2)$, $Var(Y_2)$ i $Cov(Y_1, Y_2)$ są momentami rozkładu P-CP danymi w (3), (5) i (7). Równoważny zapis kowariancji w rozkładzie ZIP-CP to (po prostych przekształceniach)

$$\begin{aligned} Cov^*(Y_1, Y_2) &= (1-g(0))^{-2}(1-\gamma)\lambda_1[(1-g(0))\exp(\alpha)E(Y_2) - (1-\gamma)E(Y_2) - (\gamma - g(0))\lambda_2] \\ &= (1-\exp(-\lambda_1))^{-2}(1-\gamma)\lambda_1\lambda_2\{[(1-\exp(-\lambda_1))e^{\alpha} - (1-\gamma)]\exp(\lambda_1(e^{\alpha}-1)) - \gamma + \exp(-\lambda_1)\}. \end{aligned} \quad (20)$$

Widzimy, że zmienne losowe (Y_1, Y_2) o łącznym rozkładzie prawdopodobieństwa ZIP-CP

1) są skorelowane ujemnie, jeśli
$[(1-\exp(-\lambda_1))e^{\alpha} - (1-\gamma)]\exp(\lambda_1(e^{\alpha}-1)) < \gamma - \exp(-\lambda_1)$,
2) są skorelowane dodatnio, jeśli
$[(1-\exp(-\lambda_1))e^{\alpha} - (1-\gamma)]\exp(\lambda_1(e^{\alpha}-1)) > \gamma - \exp(-\lambda_1)$,
3) są nieskorelowane, jeśli
$[(1-\exp(-\lambda_1))e^{\alpha} - (1-\gamma)]\exp(\lambda_1(e^{\alpha}-1)) = \gamma - \exp(-\lambda_1)$.

W przypadku $\gamma = g(0) = \exp(-\lambda_1)$, czyli brzegowego rozkładu Poissona dla Y_1, który przyjęli Berkhout i Plug (2004), skomplikowana formuła kowariancji (18) i (20) sprowadza się do znacznie prostszej postaci (7), gdzie znak kowariancji zależy jedynie od znaku stałej α. W pozostałych przypadkach, tj. gdy brzegowy rozkład dla Y_1 jest typu ZIP, znak kowariancji (20) zależy od wartości przyjmowanych przez λ_1 i α (a nie tylko od znaku tej drugiej stałej). Oczywiście, konkretna wartość kowariancji w rozkładzie ZIP-CP (a nie sam jej znak) oraz wartość współczynnika korelacji (19) zależą od wszystkich stałych występujących w funkcji prawdopodobieństwa tego rozkładu, tj. od γ, λ_1, λ_2 i α.

Zauważmy też, że zwiększenie prawdopodobieństwa zerowej wartości Y_1 (w stosunku do rozkładu Poissona o wartości oczekiwanej i wariancji λ_1), czyli

przyjęcie rozkładu ZIP z $\gamma > g(0)$, prowadzi do wariancji (16) większej niż wartość oczekiwana (11). Zatem rozkład ZIP-CP umożliwia modelowanie zwiększonej wariancji obu obserwowanych zmiennych licznikowych, chociaż nie są one traktowane symetrycznie.

3. MODEL STATYSTYCZNY TYPU ZIP-CP

Rozważamy T stochastycznie niezależnych dwuwymiarowych zmiennych losowych (Y_{1t}, Y_{2t}; $t = 1,2,...,T$) o różnych rozkładach typu ZIP-CP postaci

$$\Pr^*\{Y_{1t} = i, Y_{2t} = j\} = g_t^*(i)h_t(j,i) \quad (i, j \in N \cup \{0\}), \tag{21}$$

$$\Pr^*\{Y_{1t} = i\} = g_t^*(i) = \begin{cases} \gamma_t & dla\ i = 0, \\ \dfrac{1-\gamma_t}{1-g_t(0)}\, g_t(i) \end{cases}$$

$$dla\ i \in N; \quad g_t(i) = \exp(-\lambda_{1t})(\lambda_{1t})^i / i!, \tag{22}$$

$$\Pr^*\{Y_{2t} = j \mid Y_{1t} = i\} = h_t(j,i) = \exp[-\lambda_{2t}\exp(\alpha i)](\lambda_{2t})^j \exp(\alpha ij)/j!, \tag{23}$$

$$\lambda_{1t} = \exp(x_t\beta_1), \quad \lambda_{2t} = \exp(w_t\beta_2),$$

$$\gamma_t = \exp(-e^{\delta}\lambda_{1t}) = \exp(-\exp(\delta + x_t\beta_1), \tag{24}$$

gdzie x_t i w_t są wierszami wartości zmiennych objaśniających, które mogą się pokrywać (w części lub w całości). Zmienne objaśniające określają prawdopodobieństwa pojawienia się poszczególnych wartości zmiennych Y_{1t} i Y_{2t}, a wpływ zmiennych objaśniających na te prawdopodobieństwa jest determinowany wielkością poszczególnych składowych kolumn β_1 i β_2 oraz wielkością parametru δ, przy czym parametr δ decyduje o wielkości odchylenia prawdopodobieństwa, że $Y_{1t}=0$, od wartości wynikającej z rozkładu Poissona. W tak określonym parametrycznym modelu statystycznym wektor parametrów θ jest kolumną grupującą δ, α, β_1 i β_2. Zauważmy, że momenty rozkładu łącznego pary (Y_{1t}, Y_{2t}), podane w poprzedniej części pracy, zależą teraz od zmiennych objaśniających.

W literaturze specyfikacja oparta na wzorze (22) jest nazywana modelem płotkowym — ang. *hurdle model*; zob. Cameron i Trivedi (2005), s. 680. Porównanie tej specyfikacji z oryginalnym modelem ZIP podaje Winkelman (2008). Głównymi zaletami naszej propozycji są prostota parametryzacji i stąd względna łatwość estymacji, a zwłaszcza prostota testowania zasadności redukcji nowego

modelu do standardowego modelu Poissona. Porównywanie oryginalnego modelu ZIP ze standardowym modelem Poissona nastręcza problemy związane ze specyfikacjami (hipotezami) niezagnieżdżonymi; zob. Winkelman (2008), str. 188.

Jeśli zaobserwowano $Y_{1t}=y_{1t}$ i $Y_{2t}=y_{2t}$ ($t=1,2,...,T$), to odpowiadająca tym wartościom funkcja wiarygodności ma postać

$$L^*(\theta;y)=\left[\prod_{t:y_{1t}=0}\gamma_t\,h_t(y_{2t},0)\right]\left[\prod_{t:y_{1t}>0}\frac{1-\gamma_t}{1-g_t(0)}\,g_t(y_{1t})h_t(y_{2t},y_{1t})\right], \qquad (25)$$

gdzie y oznacza macierz ($2xT$) zawierającą zaobserwowane wartości zmiennych Y_{1t} i Y_{2t}.

W empirycznych zastosowaniach tego modelu ważne jest nie tyle wnioskowanie o $\theta=(\delta,\ \alpha,\ \beta_1'\ \beta_2')'$, ile raczej o wielu nieliniowych funkcjach parametru θ — takich, jak prawdopodobieństwa łączne, brzegowe i warunkowe różnych wartości pary (Y_{1t}, Y_{2t}) oraz momenty i inne charakterystyki jej rozkładu. Małopróbkowe wnioskowanie zarówno o θ, jak i nieliniowych funkcjach θ, możliwe jest na gruncie statystyki bayesowskiej, której podstawy i przykłady zastosowań w empirycznych badaniach ekonomicznych prezentują np. Osiewalski (2001), Osiewalski i Pajor (2010).

Jak wiadomo, podejście bayesowskie sprowadza się do określenia na przestrzeni parametrów miary probabilistycznej (lub przynajmniej σ–skończonej) zwanej rozkładem *a priori*, a następnie wykorzystania funkcji wiarygodności do uzyskania rozkładu *a posteriori* parametrów (warunkowego względem danych i reprezentującego końcową wiedzę o θ). W szczególności ważnym zadaniem jest określenie kierunku i siły korelacji między Y_{1t} i Y_{2t}, czyli podanie (dla danego t) prawdopodobieństwa *a posteriori* ujemnej korelacji, tj. warunku $[(1-\exp(-\lambda_{1t}))e^{\alpha}-(1-\gamma_t)]\exp(\lambda_{1t}(e^{\alpha}-1))<\gamma_t-\exp(-\lambda_{1t})$., oraz prezentacja pełnego rozkładu *a posteriori* współczynnika korelacji o ogólnej postaci (19). Dodatkową możliwością jest formalne bayesowskie porównanie empirycznej adekwatności dwóch niezagnieżdżonych modeli ZIP-CP, odpowiadających zamianie kolejności zmiennych objaśnianych (czyli ich numeracji). Stwarza to nowe pole badań statystycznych. Badanie adekwatności prostszego modelu P-CP, który zaproponowali Berkhout i Plug (2004), sprowadza się w ramach specyfikacji (21)-(24) do testowania prostej hipotezy $\delta=0$; można to przeprowadzić formalnie — porównując czynniki Bayesa dwóch niezagnieżdżonych modeli z $\delta=0$ i $\delta\neq0$ — lub użyć nieformalnego, ale prostszego, testu typu Lindleya w ogólniejszym modelu, dopuszczającym dowolną rzeczywistą wartość δ. Dodajmy, że $\delta>0$ ($\delta<0$) oznacza prawdopodobieństwo zerowej wartości zmiennej Y_{1t} mniejsze (większe) niż w modelu Poissona. Zatem ważną kwestią jest obliczenie prawdopodobieństwa *a posteriori* takiej sytuacji.

Aby określić bayesowski model typu ZIP-CP, należy przyjąć rozkład *a priori* wektora θ. W pierwszej pracy dotyczącej takiego modelu proponujemy założyć

niezależność *a priori* parametrów i dla każdego indywidualnie przyjąć standardowy rozkład normalny $N(0, 1)$. Zerowe wartości oczekiwane *a priori* oznaczają, że największą szansę dajemy wstępnie najprostszemu modelowi, w którym $\{Y_{1t}\}$ i $\{Y_{2t}\}$ są niezależnymi od siebie próbami losowymi prostymi z dwóch rozkładów Poissona. Jednostkowe odchylenia standardowe *a priori* dają gwarancję, że specyfikacje odległe od tej najprostszej mają bardzo istotne wstępne szanse. Wydaje się, że taki prosty łączny rozkład a priori niesie słabą tylko wiedzę wstępną (nie jest bardzo informacyjny) i gwarantuje łatwość symulacji Monte Carlo z rozkładu *a posteriori*, ale jego konkretna rola informacyjna (w stosunku do funkcji wiarygodności) oraz wrażliwość rozkładu *a posteriori* są kwestiami empirycznymi, które należy badać odrębnie dla każdego analizowanego zestawu dwuwymiarowych danych licznikowych.

4. PRZYKŁAD EMPIRYCZNY

W celu ilustracji empirycznej przydatności zaproponowanego modelu statystycznego typu ZIP-CP oraz możliwości, jakie daje jego analiza bayesowska, wykorzystamy dane, które Polasik, Marzec, Fiszeder i Górka (2012) badali stosując model prostszy (P-CP), szacowany metodą największej wiarygodności. Dane przedstawiają liczbę płatności gotówką i kartą płatniczą dokonanych (w miesiącu) przez $T = 1190$ osób, które w październiku i listopadzie roku 2010 oraz w styczniu roku 2011 ankietował *Pentor*. Wymienieni autorzy uzyskali i analizowali te dane w ramach projektu badawczego finansowanego przez Narodowy Bank Polski w roku 2010. Wyniki te wskazywały na dodatnią korelację między liczbą płatności gotówką i kartą płatniczą. Obecnie sprawdzimy, czy zastąpienie brzegowego rozkładu Poissona jednej zmiennej rozkładem typu ZIP jest empirycznie zasadne, a uzmiennienie w ten sposób możliwego znaku korelacji między zmiennymi wskaże na ujemną korelację między liczbą płatności gotówką i kartą (dla przynajmniej części respondentów). W niniejszych badaniach, o charakterze przede wszystkim metodycznym, wykorzystujemy dane surowe, tzn. bez indywidualnych wag uwzględniających reprezentatywność poszczególnych obserwacji (respondentów) wchodzących w skład próby; Polasik, Marzec, Fiszeder i Górka (2012) użyli danych ważonych.

W Tabeli 1 podajemy zmienne objaśniające i ich typowe wartości, tj. średnie w przypadku zmiennych ciągłych i najczęstsze dla zmiennych dychotomicznych.

W Tabeli 2 przedstawiamy dwuwymiarowy rozkład empiryczny liczby płatności gotówką i kartą oraz jego rozkłady brzegowe. Średnia liczba płatności gotówką wynosi 20,5 (wariancja jest równa 299), średnia liczba płatności kartą wynosi 5 (przy wariancji 45), korelację empiryczną zaś charakteryzuje współczynnik równy 0,008, wskazujący na brak liniowej zależności między liczbą płat-

Tabela 1

Informacje sumaryczne o zmiennych objaśniających

Zmienna objaśniająca	Średnia/ modalna
Płeć (1-mężczyzna, 0-kobieta)	0
Wiek (w latach)	40
Stan cywilny (1-żonaty lub zamężna, 0-nie)	1
Miejsce zamieszkania (1-miasto, 0 — wieś)	1
Miesięczny dochód w rodzinie (w tys. zł)	3,5
Wykształcenie (lata nauki)	12,5
Czy posiada internet (1-tak, 0-nie)	1

Źródło: opracowanie własne.

ności kartą (Y_1) i gotówką (Y_2). Dla obu zmiennych obserwujemy empiryczną wariancję zwiększoną w stosunku do średniej. Ponadto można zauważyć różnice między rozkładami brzegowymi, tj. empiryczny rozkład Y_2 jest przesunięty na prawo (na osi nośnika rozkładu) względem $p^{emp}(y_1)$, tzn. wartość modalna i mediana dla liczby transakcji gotówką są większe niż dla płatności kartą. Dla tej ostatniej formy płatności obserwuje się dużą frakcję zer (34%), która kontrastuje z niskim (około 0,007) prawdopodobieństwem zera, obliczonym z rozkładu Poissona o wartości oczekiwanej równej średniej z próby (czyli 5). Dla płatności gotówką frakcja zer wynosi około 2%, co przewyższa prawdopodobieństwo z rozkładu Poissona równe zaledwie 10^{-9}. W obu przypadkach wskazuje to na potrzebę zastosowania rozkładów z nadwyżką zer.

Uwzględniamy te same zmienne objaśniające dla obu zmiennych licznikowych, a zatem (biorąc pod uwagę wyrazy wolne w regresjach poissonowskich) β_1 i β_2 są kolumnami 8-wymiarowymi, natomiast wektor wszystkich parametrów θ jest kolumną 18-wymiarową. Przypomnijmy, że θ ma łączny normalny rozkład *a priori* o wartościach oczekiwanych 0 i jednostkowej macierzy kowariancji. Próbę zależną z 18-wymiarowego rozkładu *a posteriori* symulujemy za pomocą sekwencyjnego łańcucha Metropolisa i Hastingsa (M-H), tj. metody z grupy MCMC (*Markov Chain Monte Carlo*). W przypadku obu modeli (M_1: Y_{1t} oznacza liczbę płatności kartą a Y_{2t} — liczbę płatności gotówką, M_2: na odwrót) przeprowadzono 500 tysięcy losowań traktowanych jako próba z rozkładu *a posteriori*. Wcześniej wykonano kilka milionów losowań wstępnych (spalonych), badając m.in. wrażliwość algorytmu M-H na jego punkty startowe w przestrzeni parametrów.

Na Wykresie 1 przedstawiono zbieżność przyjętego łańcucha M-H do rozkładu *a posteriori* w modelu M_1, która jest zadawalająca z uwagi na szybko stabilizujący się dla wszystkich parametrów przebieg tzw. standaryzowanych statystyk sum skumulowanych (CuSum), tzn. średnich arytmetycznych (z poszczególnych

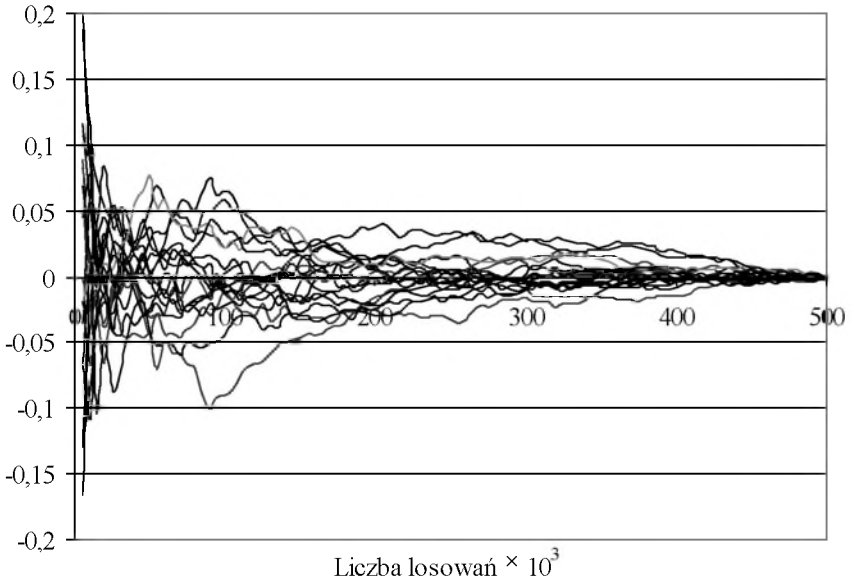
Tabela 2

Empiryczny łączny rozkład liczby płatności gotówką i kartą oraz jego rozkłady brzegowe

	Transakcje kartą (Y_1)									
Transakcje gotówką (Y_2)	$p^{emp}(y_1,y_2)$	0	(0;5]	(5;10]	(10;15]	(15;20]	(25;30]	>30	$p^{emp}(y_2)$	struktura
	0	0	2	11	6	2	2	1	24	2%
	(0;5]	13	46	39	18	6	1	3	126	11%
	(5;10]	69	114	38	16	8	3	0	248	21%
	(10;15]	76	55	38	9	7	8	3	196	16%
	(15;20]	57	52	27	5	4	2	1	148	12%
	(20;25]	40	36	19	6	3	2	2	108	9%
	(25;30]	46	20	9	7	3	0	0	85	7%
	(30;35]	26	17	12	5	4	1	1	66	6%
	(35;40]	21	17	7	3	1	1	5	55	5%
	(40;45]	13	8	3	4	4	0	0	32	3%
	(45;50]	11	9	5	4	1	1	1	32	3%
	>50	37	15	3	4	4	2	5	70	6%
	$p^{emp}(y_1)$	409	391	211	87	47	23	22	1190	
	struktura	34%	33%	18%	7%	4%	2%	2%		

Źródło: opracowanie własne.

losowań) standaryzowanych końcowymi wartościami średnich i odchyleń standardowych. W przypadku drugiego modelu zastosowany algorytm także okazał się efektywnym narzędziem numerycznym.



Źródło: opracowanie własne.

Wykres 1. Zbieżność statystyk CuSum w modelu M_1

Pierwszą kwestią, którą należy poddać empirycznej weryfikacji, jest wybór jednej z dwóch alternatywnych specyfikacji (M_1, M_2) modelu statystycznego typu ZIP-CP. Warto przypomnieć, że prawdopodobieństwo *a posteriori* modelu M_i ($i = 1,2$) wyraża, zgodnie z wzorem Bayesa, formuła

$$p(M_i|\mathbf{y}) = \frac{p(\mathbf{y}|M_i)\cdot p(M_i)}{p(\mathbf{y}|M_1)\cdot p(M_1) + p(\mathbf{y}|M_2)\cdot p(M_2)}. \tag{26}$$

Można przyjąć równe szanse a priori, $p(M_i) = 0,5$, bo brak jest teoretycznych przesłanek do faworyzowania któregoś modelu. Do porównania wystarczy więc czynnik Bayesa, czyli iloraz brzegowych gęstości wektora obserwacji $BF = p(y|M_2)/\, p(y|M_1)$; zob. Osiewalski (2001), Wróblewska (2009). Wyniki prezentujemy w Tabeli 3.

Model M_1 jest kilkaset rzędów wielkości lepszy od M_2 i skupia prawie całą masę prawdopodobieństwa *a posteriori*; prawdopodobieństwo *a posteriori* modelu M_2

wynosi praktycznie zero. Przewaga modelu M_1 w opisie badanego zjawiska jest zdecydowana. W uzupełnieniu podajemy dla obu modeli wartości funkcji wiarygodności $L^*\left(\hat{\theta}_{NW}; y\right)$, zob. wzór (25), dla ocen największej wiarygodności, które zostały wyznaczone w ramach numerycznej realizacji algorytmu M-H. Dla modelu M_1 otrzymano $L^*\left(\hat{\theta}_{NW}; y\right) = 55\ 235$, a dla drugiej specyfikacji największa wartość funkcji wiarygodności była niższa, bowiem wyniosła 54 161. Z niebayesowskiego punktu widzenia wynik porównania modeli oparty na kryterium informacyjnym (którymkolwiek) także wskazuje na adekwatność modelu M_1 (w kontekście M_2). Warto wspomnieć, że z uwagi na niestandardową postać modelu (21)–(24) zastosowanie deterministycznych procedur optymalizacji funkcji wiarygodności spotkało się z ogromnymi problemami obliczeniowymi. Numeryczne narzędzia analizy bayesowskiej okazują się zatem przydatne także w estymacji metodą największej wiarygodności.

Tabela 3

Brzegowe gęstości wektora obserwacji i prawdopodobieństwa a posteriori obu modeli

Model	M_1: Y_{1t} liczba płatności kartą, Y_{2t} — gotówką	M_2: Y_{1t} liczba płatności gotówką, Y_{2t} — kartą
$\ln p\left(y\middle\vert M_i\right)$	55218,3	54142,8
Log_{10} BF	–	–467
Czynnik Bayesa (BF)	–	≈0
$p\left(M_i\right)$	0,5	0,5
$p\left(M_i\middle\vert y\right)$	≈1	≈0

Źródło: opracowanie własne.

Z uwagi na wyniki porównań modeli, dalsze rozważania natury interpretacyjnej będą opierać się na M_1, a wyniki dla drugiego modelu będą miały charakter uzupełniający. W Tabeli 4 podano wartości oczekiwane i odchylenia standardowe *a posteriori* parametrów naszej dwuwymiarowej regresji typu ZIP-CP W M_1 wszystkie zmienne objaśniające istotnie wpływają na liczbę płatności gotówką, natomiast tylko posiadanie internetu, wykształcenie i dochód powodują znaczące zróżnicowanie liczby płatności kartą. Oceny parametrów i błędy szacunku, które podają Polasik, Marzec, Fiszeder i Górka (2012), są bardzo zbliżone do bayesowskich wartości oczekiwanych i odchyleń standardowych *a posteriori* prezentowanych w tej pracy — mimo, że w naszych badaniach liczba zmiennych objaśniających jest ponad dwukrotnie mniejsza. Brzegowy rozkład *a posteriori* parametru δ (Wykres 2) pokazuje, że redukcja modelu ZIP-CP do P-CP jest bezza-

sadna, gdyż prawdopodobieństwo zerowej liczby płatności gotówką jest istotnie większe niż wynikałoby to z rozkładu Poissona. Wartość oczekiwana *a posteriori* dla δ wynosi -1,876 przy odchyleniu standardowym 0,041.

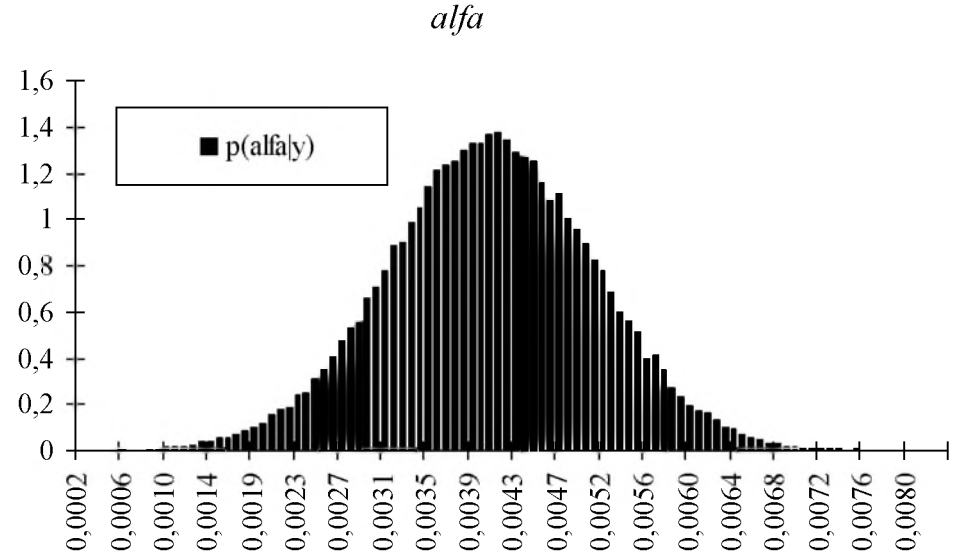
Tabela 4

Wartości oczekiwane i odchylenia standardowe a posteriori parametrów modeli

	Model	M_1		M_2	
	Zmienna/parametr	E(θ|y)	D(θ|y)	E(θ|y)	D(θ|y)
płatności kartą	„1"	0,909	0,098	-0,259	0,101
	Płeć	-0,045	0,025	0,006	0,026
	Wiek	-0,002	0,001	-0,007	0,001
	Stan cywilny	-0,047	0,029	0,056	0,031
	Miejsce zamieszkania	-0,007	0,028	0,077	0,030
	Dochód	0,051	0,010	0,094	0,011
	Wykształcenie	0,056	0,006	0,089	0,006
	Internet	0,360	0,039	0,558	0,042
płatności gotówką	„1"	2,826	0,049	2,803	0,048
	Płeć	-0,102	0,013	-0,093	0,013
	Wiek	0,008	0,001	0,008	0,001
	Stan cywilny	-0,158	0,015	-0,152	0,014
	Miejsce zamieszkania	0,145	0,015	0,133	0,015
	Dochód	0,016	0,006	0,019	0,005
	Wykształcenie	-0,008	0,003	-0,004*	0,003
	Internet	-0,085	0,016	-0,066	0,016
	α	0,004	0,001	0,0023	0,0007
	δ	-1,876	0,041	-1,638	0,053

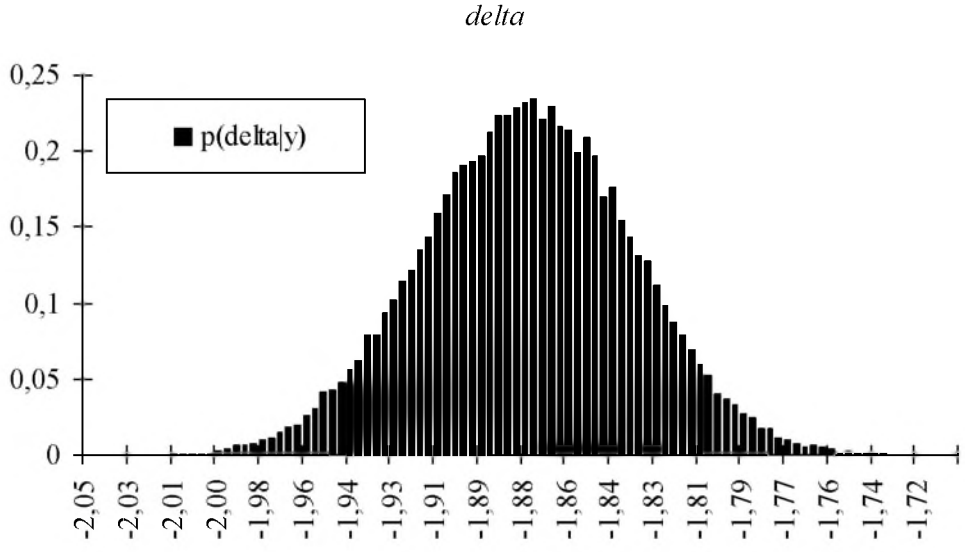
Źródło: opracowanie własne.

Zatem rozkład ten uległ znacznemu przesunięciu w stosunku do rozkładu *a priori* i jednocześnie zmniejszyło się jego rozproszenie. Dla α charakterystyki te wynoszą odpowiednio 0,004 i 0,001, wskazując na istotnie dodatnią zależność warunkowej średniej liczby płatności kartą od liczby płatności gotówką. Brzegowy rozkład *a posteriori* parametru α (Wykres 3) jest praktycznie ograniczony do przedziału (0,0006; 0,0076), czyli zawiera się w przedziale o wysokiej gęstości *a priori*, jednakże informacje z próby spowodowały uzyskanie rozkładu o znacząco mniejszym rozproszeniu.



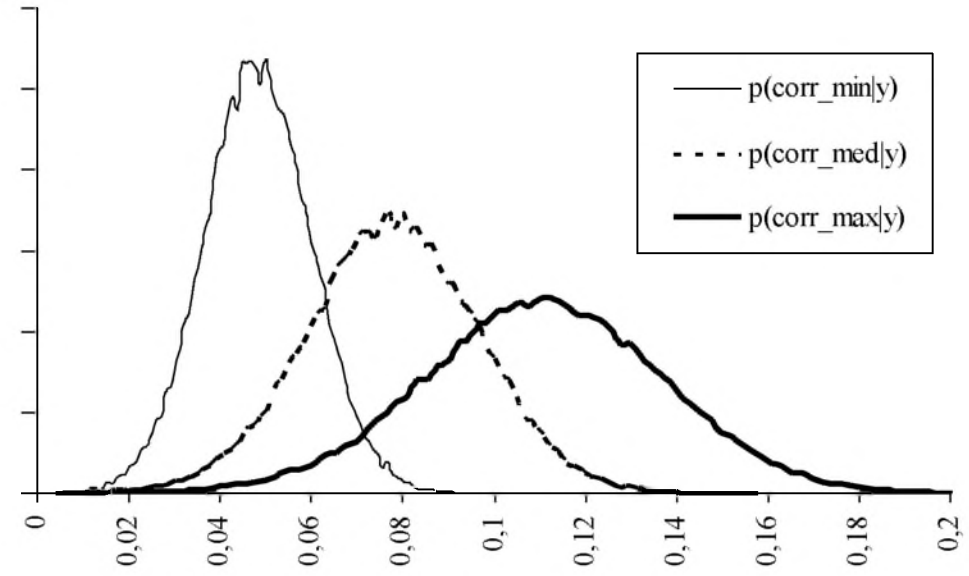
Źródło: opracowanie własne.

Wykres 2. Brzegowy rozkład a posteriori parametru α



Źródło: opracowanie własne.

Wykres 3. Brzegowy rozkład a posteriori parametru δ



Źródło: opracowanie własne.

Wykres 4. Rozkłady a posteriori próbkowych korelacji dla wybranych obserwacji

Osiewalski (2012) dowodzi, że w modelu typu ZIP-CP dodatniość parametru α nie musi oznaczać dodatniej korelacji próbkowej zmiennych objaśnianych (jak to jest w modelu P-CP). Rozkłady *a posteriori* próbkowych korelacji trzech par (Y_{1t}, Y_{2t}) — tych, dla których wartość oczekiwana *a posteriori* współczynnika korelacji jest najmniejsza, przeciętna w sensie mediany i największa — są pokazane na Wykresie 4. Dowodzą one słabej, ale jedynie dodatniej korelacji między liczbami płatności gotówką i kartą. Zastosowanie modelu bardziej adekwatnego, tj. typu ZIP-CP zamiast P-CP, nie zmienia (pod tym względem) wymowy wyników, które podali Polasik, Marzec, Fiszeder i Górka (2012).

5. PODSUMOWANIE

Zaproponowane uogólnienie modelu P-CP okazało się uzasadnione w przypadku wstępnych badań dotyczących preferencji polskich konsumentów w wyborze metod płatności. Wskazuje to na adekwatność modeli typu ZIP-CP w sytuacjach, gdy obserwuje się nadwyżkę (bądź deflację) obserwacji zerowych lub gdy dwie zmienne licznikowe, oddające rezultaty decyzji konsumentów, są ze sobą potencjalnie skorelowane (ujemnie albo dodatnio). Podejście bayesowskie pozwoliło na estymację parametrów rozważanych modeli bez odwoływania się do aproksymacji asymptotycznych. Bayesowskie porównywanie mocy wyja-

śniającej konkurencyjnych (niezagnieżdżonych) modeli formalnie potwierdziło wstępne wnioski uzyskane we wcześniejszych badaniach, a dotyczące wyboru jednej z dwóch alternatywnych specyfikacji statystycznych w kontekście zaobserwowanych danych.

Interesującym kierunkiem dalszych badań jest zastosowanie dwuparametrowej rodziny rozkładów Poissona (*generalized Poisson distribution*; zob. Consul i Jain (1973), Famoye i Singh (2006)) dla brzegowego rozkładu zmiennej Y_1 bądź także dla rozkładu warunkowego drugiej zmiennej.

BIBLIOGRAFIA

Berkhout P., Plug E. (2004), *A bivariate Poisson count data model using conditional probabilities*, "Statistica Neerlandica" vol. 58, 349–364.

Cameron A. C., Trivedi P. K. (1998), *Regression Analysis of Count data*, Cambridge University Press, New York.

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

Consul P. C., Jain G. C. (1973), *A Generalization of the Poisson Distribution*, "Technometrics" 15, s. 791–799.

Famoye F., Singh K. P. (2006), *Zero-Inflated Generalized Poisson Regression Model with an Application to Domestic Violence Data*, "Journal of Data Science", 4, s. 117–130.

Kocherlakota S., Kocherlakota K. (1992), *Bivariate Discrete Distributions*, Marcel Dekker, New York.

Lambert D. (1992), *Zero-inflated Poisson regression, with an application to defects in manufacturing*, "Technometrics" 34, s. 1–14.

Marzec J. (2012), *Wybrane dwuwymiarowe modele dla zmiennych licznikowych w ekonomii*, „Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie — Metody Analizy Danych" nr 884, s. 59–70.

Osiewalski J. (2001), *Ekonometria bayesowska w zastosowaniach*, Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków.

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.

Osiewalski J., Pajor A. (2010), *Bayesian Value-at-Risk for a Portfolio: Multi- and Univariate Approaches Using MSF-SBEKK Models*, "Central European Journal of Economic Modelling and Econometrics" 2, s. 253–277.

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

Winkelman R. (2008), *Econometric Analysis of Count Data*, Springer-Verlag, Berlin Heidelberg.

Wróblewska J. (2009), *Bayesian Model Selection in the Analysis of Cointegration*, "Central European Journal of Economic Modelling and Econometrics" 1, s. 57–69.