Zeszyty Naukowe Uniwersytet Ekonomiczny w Krakowie

5(965)

ISSN 1898-6447
Zesz. Nauk. UEK, 2017; 5(965): 31–49
DOI: 10.15678/ZNUEK.2017.0965.0503

Jacek Osiewalski
Jerzy Marzec

Dwuwymiarowe zmienne licznikowe – bayesowskie modelowanie selekcji próby*

Streszczenie

W artykule przedstawiono propozycję łącznego modelu statystycznego dwóch zmiennych licznikowych, z których jedna może być zdegenerowana w zerze. Rozważane jest modelowanie oparte na przełączaniu między dwu- i jednowymiarowym modelem regresji poissonowskiej, przy czym przełączanie zależy od zaobserwowanej wartości trzeciej, dychotomicznej zmiennej. Zalecana jest analiza bayesowska; w dwóch szczególnych przypadkach proponowanego modelu bayesowskiego sformułowano konsekwencje ważne dla wnioskowania. W części empirycznej rozważane jest łączne modelowanie liczby płatności gotówką i kartą w Polsce, z wykorzystaniem danych zarówno dla posiadaczy kart, jak i osób ich nieposiadających.

Słowa kluczowe: dwuwymiarowe modele regresji Poissona, przełączanie między rozkładem niezdegenerowanym i zdegenerowanym, faktoryzacja funkcji wiarygodności, płatności kartą płatniczą i gotówką.
Klasyfikacja JEL: C25, C24, C51.

Jacek Osiewalski, Uniwersytet Ekonomiczny w Krakowie, Wydział Zarządzania, Katedra Ekonometrii i Badań Operacyjnych, ul. Rakowicka 27, 31-510 Kraków, e-mail: eeosiewa@cyf-kr.edu.pl

Jerzy Marzec, Uniwersytet Ekonomiczny w Krakowie, Wydział Zarządzania, Katedra Ekonometrii i Badań Operacyjnych, ul. Rakowicka 27, 31-510 Kraków, e-mail: marzecj@uek.krakow.pl

* Artykuł stanowi wynik realizacji projektu sfinansowanego ze środków przyznanych Wydziałowi Zarządzania Uniwersytetu Ekonomicznego w Krakowie w ramach dotacji na utrzymanie potencjału badawczego.

1. Wprowadzenie

Przy łącznym modelowaniu zmiennych licznikowych można spotkać się z sytuacją, gdy jedna z nich jest z konieczności zerem dla wielu obserwowanych obiektów. Na przykład jeśli badamy determinanty i współzależność liczby przejazdów mieszkańców miasta transportem publicznym i własnymi samochodami, to dla osób bez samochodu liczba przejazdów tym środkiem jest stale równa zero. Powstaje pytanie, jakie są możliwości i konsekwencje wnioskowania o determinantach liczby przejazdów transportem publicznym oraz o zależności między oboma liczbami przejazdów na podstawie danych dotyczących wszystkich badanych mieszkańców miasta – wobec badania tych determinant i tej samej zależności na podstawie danych dotyczących tylko mieszkańców miasta posiadających samochód. Wykorzystanie tych ostatnich danych oznacza wstępną selekcję obserwacji i uniemożliwia przenoszenie wyników analizy na wszystkich mieszkańców. Aby wykorzystać cały zbiór obserwacji i umożliwić wyciąganie ogólniejszych wniosków, autorzy zaproponowali model statystyczny uwzględniający przełączanie między dwoma modelami zmiennych licznikowych: modelem dwuwymiarowym i jednowymiarowym; za przełączanie odpowiada dychotomiczny model stosownej zmiennej zero-jedynkowej (reprezentującej w przytaczanym przykładzie posiadanie samochodu). Takie podejście pozwala ująć różne sytuacje jako przypadki szczególne i sformułować kluczową testowalną hipotezę identyczności mechanizmu określającego generowanie (w dwóch grupach obiektów) wartości tej zmiennej licznikowej, która nigdy nie jest zdegenerowana.

Zasadniczą częścią składową omawianego w tej pracy modelu przełącznikowego jest dwuwymiarowy model zmiennych licznikowych, opisujący przypadek, w którym żadna ze zmiennych nie jest skoncentrowana w zerze. Regresja poissonowska jest znanym modelem analizy zmiennych licznikowych. Istnieją jej dwuwymiarowe uogólnienia, lecz większość z nich charakteryzuje się ograniczeniami dotyczącymi znaku współczynnika korelacji między zmiennymi, inne zaś prowadzą do komplikacji natury statystyczno-numerycznej (zob. m.in. [Kocherlakota i Kocherlakota 1992, Winkelman 2008]). Modele, które dopuszczają zarówno korelację dodatnią, jak i ujemną, można uzyskać wykorzystując np. kopule lub mieszanki rozkładów. Innym podejściem jest warunkowy model Poissona, który zaproponowali P. Berkhout i E. Plug [2004]. Omówienie tych zagadnień, wraz z odwołaniami do literatury, można znaleźć m.in. w artykule [Marzec 2012]. Warto podkreślić, że w kontekście modeli dwuwymiarowych nie pojawia się w literaturze kwestia selekcji próby.

Jako główną część składową proponowanego modelu statystycznego wykorzystano specyfikację ZIP-CP (*zero inflated Poisson – conditional Poisson*), którą zaproponowano w pracy [Marzec i Osiewalski 2012]. Jest to dwuwymia-

rowa regresja typu Poissona, ogólniejsza niż model P-CP (*Poisson – conditional Poisson*), który wprowadzili P. Berkhout i E. Plug [2004]. W modelu P-CP przyjmuje się brzegowy rozkład Poissona dla jednej zmiennej i warunkowy rozkład Poissona dla drugiej (przy ustalonej pierwszej); model ten 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. W modelu ZIP-CP dwuwymiarowej regresji typu Poissona zamiast brzegowego rozkładu Poissona pierwszej z dwóch zmiennych wprowadza się rozkład typu ZIP, w wersji „płotkowej" (*hurdle model*), co prowadzi do znaku kowariancji (między oboma zmiennymi licznikowymi) zależnego od wartości zmiennych objaśniających. Charakterystyki modelu ZIP-CP wynikają z własności dwuwymiarowego skokowego rozkładu ZIP-CP, który wprowadził i zbadał J. Osiewalski [2012]. Druga część proponowanego modelu przełącznikowego to jednowymiarowa regresja Poissona dla drugiej zmiennej – w przypadku gdy pierwsza jest zdegenerowana (skoncentrowana w zerze). Jak już wspomniano, trzecią częścią jest specyfikacja dychotomiczna, opisująca przełączanie między przypadkiem dwuwymiarowym (niezdegenerowanym) i jednowymiarowym (zdegenerowanym).

Następny punkt pracy poświęcony jest prezentacji probabilistycznych podstaw modelu, tj. rozkładów skokowych wykorzystywanych w budowie trzech części składowych tego modelu – w szczególności rozkładu ZIP-CP. W trzecim punkcie omówiono proponowany model statystyczny i postać funkcji wiarygodności oraz przedstawiono analizę bayesowską tego modelu, zwracając uwagę na jego dwa przypadki szczególne. W czwartym, empirycznym punkcie pracy zaprezentowano nowe wyniki, uzyskiwane na podstawie pełnego zbioru danych, w łącznym badaniu liczb transakcji dokonywanych kartą bankową i gotówką (zob. [Polasik, Marzec, Fiszeder i Górka 2012] oraz [Marzec i Osiewalski 2012]). Przykład ten ilustruje problemy modelowania i wnioskowania w sytuacji zmiennych licznikowych, z których jedna (liczba płatności kartą) jest zdegenerowana dla wielu badanych jednostek (osób nieposiadających kart). W piątym punkcie zawarto uwagi końcowe.

Proponowany w tej pracy przykład empiryczny wpisuje się w badania rozwoju obrotu bezgotówkowego w Polsce, które są prowadzone od kilku lat (zob. np. [Polasik i Maciejewski 2009, Fiszeder i Polasik 2009, Polasik 2015, Polasik, Wisniewski i Lightfoot 2012, Górka 2013, Goczek i Witkowski 2015, 2016]). Z punktu widzenia banku centralnego interesującą kwestią jest określenie czynników motywujących do korzystania z kart płatniczych i identyfikacja tych barier utrudniających działalność przedsiębiorstw handlowych, które są związane z dodatkowymi opłatami *interchange* za transakcje dokonane przy użyciu kart. Dla gospodarki i finansów państwa wymierne korzyści rodzi ograniczenie transakcji gotówkowych między klientem detalicznym a sprzedawcą na rzecz

transakcji dokonywanych kartą, co częściowo przyczyniłoby się do zmniejszenia szarej strefy. Ważnym elementem badań wzbogacających obecny stan wiedzy na temat płatności kartą i gotówką są propozycje nowych modeli ekonometrycznych, opisujących złożone decyzje podejmowane przez konsumentów.

2. Probabilistyczne podstawy nowego modelu statystycznego

Rozważamy łączny rozkład prawdopodobieństwa trzech zmiennych losowych (Y_1, Y_2, Y_3), z których trzecia ma rozkład dwupunktowy (jest zmienną zero-jedynkową), druga może przyjąć dowolną wartość całkowitą nieujemną, a pierwsza ma rozkład jednopunktowy, gdy $Y_3 = 0 \left(\Pr\{ Y_1 = 0 \mid Y_3 = 0 \} = 1 \right)$, może zaś przyjąć dowolną wartość całkowitą nieujemną, gdy $Y_3 = 1$. Zatem przy $Y_3 = 0$ rozkład (warunkowy) pary (Y_1, Y_2) jest tożsamy z rozkładem pary $(0, Y_2)$, czyli odpowiada rozkładowi pojedynczej zmiennej Y_2. Jedynie przy $Y_3 = 1$ rozkład pary (Y_1, Y_2) jest dwuwymiarowym rozkładem na zbiorze wszystkich par liczb całkowitych nieujemnych. Temu ostatniemu poświęcamy specjalną uwagę, rozważając przypadek prostszy: P-CP (zob. [Berkhout i Plug 2004]) i ogólniejszy: ZIP-CP (zob. [Osiewalski 2012]).

Przy $Y_3 = 1$ rozkład prawdopodobieństwa pary (Y_1, Y_2) jest następujący:

$$\Pr\left\{ Y_1 = i, Y_2 = j \mid Y_3 = 1 \right\} = \Pr\left\{ Y_1 = i \mid Y_3 = 1 \right\} \Pr\left\{ Y_2 = j \mid Y_3 = 1, Y_1 = i \right\} = g(i)\, h(j,i), \qquad (1)$$

przy czym $i, j \in N \cup \{0\}$. Jeśli rozkład 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

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

to mamy rozkład dwuwymiarowy P-CP o momentach postaci [Berkhout i Plug 2004]:

$$E\left(Y_2 \mid Y_3 = 1 \right) = \lambda_2 \exp[\lambda_1 (e^{\alpha} - 1)], \qquad (3)$$

$$Var\left(Y_2 \mid Y_3 = 1 \right) = E\left(Y_2 \mid Y_3 = 1 \right) + \left[E\left(Y_2 \mid Y_3 = 1 \right) \right]^2 \{ \exp[\lambda_1 (e^{\alpha} - 1)^2] - 1 \}, \qquad (4)$$

$$Cov\left(Y_1, Y_2 \mid Y_3 = 1 \right) = \lambda_1 (e^{\alpha} - 1) E\left(Y_2 \mid Y_3 = 1 \right). \qquad (5)$$

Jeśli $\alpha \neq 0$, to wariancja (4) zmiennej Y_2 jest większa od wartości oczekiwanej (3). Zależność między obu zmiennymi sprawia, że rozkład zmiennej Y_2 odpowiada empirycznie częstej sytuacji zwiększonej wariancji danych licznikowych. Rozkład zmiennej Y_1, czyli rozkład Poissona, nie ma tej właściwości. Jest to pierwszy powód uogólnienia dwuwymiarowego rozkładu P-CP przez wprowa-

dzenie rozkładu typu ZIP na miejsce brzegowego rozkładu Poissona. Modele regresji dla skokowej zmiennej objaśnianej z nadmierną liczbą zer spopularyzował głównie D. Lambert [1992], a A.C. Cameron i P.K. Trivedi [1998, 2005] oraz R. Winkelman [2008] przedstawiają stosowne modele danych licznikowych z przykładami ich zastosowań w ekonomii.

Należy zauważyć, że znak kowariancji między Y_1 i Y_2, czyli znak wyrażenia (5), zależy jedynie od znaku stałej α, a nie od wielkości λ_1, λ_2, parametryzowanych głębiej (uzależnianych od zmiennych objaśniających) w statystycznych zastosowaniach tego modelu probabilistycznego. Uogólnienie, które zaproponował J. Osiewalski [2012], dopuszcza związek znaku kowariancji i wielkości λ_1. Ta ogólniejsza klasa rozkładów (oznaczana gwiazdką) jest określona przez ten sam warunkowy rozkład Y_2 przy ustalonym Y_1:

$$\Pr^*\left\{Y_2 = j \mid Y_3 = 1, Y_1 = i\right\} = h(j,i) = \Pr\left\{Y_2 = j \mid Y_3 = 1, Y_1 = i\right\} \tag{6}$$

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

$$\Pr^*\left\{Y_1 = i \mid Y_3 = 1\right\} = g^*(i) = \begin{cases} \gamma & \text{dla } i = 0, \\ \dfrac{1-\gamma}{1-g(0)} g(i) & \text{dla } i \in N, \end{cases} \tag{7}$$

gdzie γ jest ustaloną liczbą z przedziału (0, 1), funkcje g i h są zaś takie same jak w (1). Jeśli $\gamma = g(0)$, to $\Pr^*\left\{Y_1 = i \mid Y_3 = 1\right\} = g^*(i) = g(i) = \Pr\left\{Y_1 = i \mid Y_3 = 1\right\}$ i mamy przypadek (1). Jeśli $\gamma \neq g(0)$, a funkcje g i h zadane są nadal wzorami (2), to rozkład zmiennej Y_1 jest typu ZIP, zaś warunkowy dla Y_2 przy ustalonym Y_1 pozostaje rozkładem Poissona. Rozkład łączny to ZIP-CP, a jego momenty mają ogólną postać:

$$E^*\left(Y_1^m Y_2^n \mid Y_3 = 1\right) = \frac{(1-\gamma)E\left(Y_1^m Y_2^n \mid Y_3 = 1\right) + (\gamma - g(0))0^m E\left(Y_2^n \mid Y_3 = 1, Y_1 = 0\right)}{1-g(0)}, \tag{8}$$

gdzie wykorzystuje się znaną postać momentów rozkładu P-CP (dla $m = 0$ przyjmując $0^m = 1$). W szczególności otrzymujemy:

$$E^*\left(Y_1 \mid Y_3 = 1\right) = (1-g(0))^{-1}(1-\gamma)\lambda_1, \tag{9}$$

$$E^*\left(Y_2 \mid Y_3 = 1\right) = (1-g(0))^{-1}\left[(1-\gamma)E\left(Y_2 \mid Y_3 = 1\right) + (\gamma - g(0))\lambda_2\right], \tag{10}$$

$$Var^*\left(Y_1 \mid Y_3 = 1\right) = \frac{1-\gamma}{1-g(0)}\lambda_1\left(1 + \frac{\gamma - g(0)}{1-g(0)}\lambda_1\right), \tag{11}$$

$$Var^*\left(Y_2 \mid Y_3 = 1\right) =$$

$$= \frac{1-\gamma}{1-g(0)}\left\{Var\left(Y_2 \mid Y_3 = 1\right) + \frac{\gamma - g(0)}{1-g(0)}\left[E\left(Y_2 \mid Y_3 = 1\right) - \lambda_2\right]^2 + \frac{\gamma - g(0)}{1-\gamma}\lambda_2\right\}, \tag{12}$$

$$Cov^*\left(Y_1, Y_2 \mid Y_3 = 1\right) = \\ = \frac{\lambda_1\lambda_2(1-\gamma)}{(1-\exp(-\lambda_1))^2}\left\{\left[(1-e^{-\lambda_1})e^{\alpha} - (1-\gamma)\right]\exp(\lambda_1(e^{\alpha}-1)) - \gamma + e^{-\lambda_1}\right\}, \tag{13}$$

gdzie $E\left(Y_2 \mid Y_3 = 1\right)$ i $Var\left(Y_2 \mid Y_3 = 1\right)$ są momentami rozkładu P-CP danymi w (3) i (4). Widzimy, że zmienne tworzące parę (Y_1, Y_2) o rozkładzie prawdopodobieństwa ZIP-CP:

1) są skorelowane ujemnie, jeśli $\left[(1-e^{-\lambda_1})e^{\alpha} - (1-\gamma)\right]\exp(\lambda_1(e^{\alpha}-1)) < \gamma - e^{-\lambda_1}$,

2) są skorelowane dodatnio, jeśli $\left[(1-e^{-\lambda_1})e^{\alpha} - (1-\gamma)\right]\exp(\lambda_1(e^{\alpha}-1)) > \gamma - e^{-\lambda_1}$,

3) są nieskorelowane, jeśli $\left[(1-e^{-\lambda_1})e^{\alpha} - (1-\gamma)\right]\exp(\lambda_1(e^{\alpha}-1)) = \gamma - e^{-\lambda_1}$.

W przypadku $\gamma = g(0) = e^{-\lambda_1}$, tj. rozkładu Poissona dla Y_1 (przy $Y_3 = 1$), złożona formuła kowariancji (13) sprowadza się do znacznie prostszej postaci (5), gdzie znak kowariancji zależy jedynie od znaku stałej α. W pozostałych przypadkach, tj. gdy rozkład Y_1 jest typu ZIP, znak kowariancji (13) 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 – w konsekwencji – wartość współczynnika korelacji zależą od wszystkich stałych występujących w funkcji prawdopodobieństwa tego rozkładu, tj. od $\gamma, \lambda_1, \lambda_2$ 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 (11) większej niż wartość oczekiwana (9). Rozkład ZIP-CP umożliwia modelowanie zwiększonej wariancji obu obserwowanych zmiennych licznikowych, chociaż nie są one traktowane symetrycznie.

Powyższe rozważania dotyczyły jedynie rozkładu warunkowego pary (Y_1, Y_2) przy $Y_3 = 1$, czyli bardziej złożonej części specyfikacji trójwymiarowej. Rozkład zmiennej Y_2 przy $Y_3 = 0$ – i tym samym przy jedynej wartości Y_1 (równej 0) – przyjmujemy tak, aby można było badać identyczność rozkładu warunkowego zmiennej Y_2 przy $Y_1 = 0$ w obu sytuacjach: $Y_3 = 0$ i $Y_3 = 1$. Zakładamy zatem, że jest to rozkład Poissona o funkcji prawdopodobieństwa:

$$\Pr\left\{Y_2 = j \mid Y_3 = 0, Y_1 = 0\right\} = h_0(j) = \exp(-\lambda_{2,0})(\lambda_{2,0})^j / j!, \tag{14}$$

z parametrem $\lambda_{2,0}$ niekoniecznie równym λ_2.

Podsumowując dotychczas przyjęte założenia, wprowadzamy następujący łączny rozkład trzech zmiennych skokowych:

$$\Pr\{Y_1 = i, Y_2 = j, Y_3 = l\} = \begin{cases} pg^*(i)h(j,i), & i, j \in N \cup \{0\}, l = 1, \\ (1-p)h_0(j), & i = 0, j \in N \cup \{0\}, l = 0, \\ 0, & i \in N, j \in N \cup \{0\}, l = 0, \end{cases} \tag{15}$$

gdzie $p = \Pr\{Y_3 = 1\}$. Brzegowy rozkład pary (Y_1, Y_2) jest swoistą mieszanką dwuwymiarowego rozkładu ZIP-CP i jednowymiarowego rozkładu Poissona:

$$\Pr\{Y_1 = i, Y_2 = j\} = pg^*(i)h(j,i) + (1-p)I_{(0)}(i)h_0(j), \quad i, j \in N \cup \{0\}, \tag{16}$$

gdzie $I_A(.)$ oznacza funkcję charakterystyczną zbioru A; jego momenty można zapisać jako:

$$E(Y_1^m Y_2^n) = pE^*\left(Y_1^m Y_2^n \mid Y_3 = 1\right) + (1-p)0^m E\left(Y_2^n \mid Y_3 = 0, Y_1 = 0\right), \tag{17}$$

przy czym $E^*\left(Y_1^m Y_2^n \mid Y_3 = 1\right)$ to moment zwykły lub mieszany rzędu (m, n) w rozkładzie ZIP-CP, dany ogólnym wzorem (8), zaś $E\left(Y_2^n \mid Y_3 = 0, Y_1 = 0\right)$ to moment zwykły rzędu n w rozkładzie Poissona z parametrem $\lambda_{2,0}$.

3. Model statystyczny

Rozważamy T trójwymiarowych zmiennych losowych $(Y_{1t}, Y_{2t}, Y_{3t}; t = 1, 2, \ldots, T)$, gdzie Y_{3t} są zmiennymi zero-jedynkowymi. Przy $Y_{3t} = 1$, pary (Y_{1t}, Y_{2t}) mają różne rozkłady typu ZIP-CP:

$$\Pr^*\{Y_{1t} = i, Y_{2t} = j \mid Y_{3t} = 1\} = g_t^*(i)h_t(j,i), \quad i, j \in N \cup \{0\}, \tag{18}$$

gdzie

$$\Pr^*\{Y_{1t} = i \mid Y_{3t} = 1\} = g_t^*(i) = \begin{cases} \gamma_t & \text{dla } i = 0, \\ \dfrac{1-\gamma_t}{1-g_t(0)} g_t(i) & \text{dla } i \in N;\ g_t(i) = e^{-\lambda_{1t}}(\lambda_{1t})^i / i!, \end{cases} \tag{19}$$

$$\Pr^*\{Y_{2t} = j \mid Y_{3t} = 1, Y_{1t} = i\} = h_t(j,i) = \exp[-\lambda_{2t}\exp(\alpha i)](\lambda_{2t})^j \exp(\alpha ij)/j!, \tag{20}$$

$$\lambda_{1t} = \exp(x_t\beta_1), \quad \lambda_{2t} = \exp(w_t\beta_2), \quad \gamma_t = \exp(-e^{\delta}\lambda_{1t}) = \exp(-\exp(\delta + x_t\beta_1)); \tag{21}$$

x_t i w_t są wierszami wartości zmiennych objaśniających, które mogą się pokrywać (w części lub w całości). Zmienne te określają prawdopodobieństwa pojawienia się poszczególnych par wartości Y_{1t} i Y_{2t}; wpływ x_t i w_t na te prawdopodobieństwa jest determinowany wielkością poszczególnych składowych kolumn β_1 i β_2, wielkością parametru zależności α oraz wielkością parametru δ, który decyduje o odchyleniu prawdopodobieństwa, że $Y_{1t} = 0$, od wartości wynikającej z rozkładu Poissona. Zauważmy, że momenty rozkładu pary (Y_{1t}, Y_{2t}), podane w poprzednim punkcie pracy, zależą teraz od zmiennych objaśniających.

W literaturze specyfikacja oparta na wzorze (19) jest nazywana modelem płotkowym – zob. [Cameron i Trivedi 2005, s. 680]. Porównanie tej specyfikacji z oryginalnym modelem ZIP przedstawia R. Winkelman [2008]. Głównymi zaletami przedstawionej w niniejszym artykule propozycji są prostota parametryzacji

i względna łatwość estymacji oraz prostota testowania zasadności redukcji specyfikacji (19) do standardowego modelu Poissona.

Przy $Y_{3t} = 0$, pary $(Y_{1t}, Y_{2t}) = (0, Y_{2t})$ mają rozkłady zdegenerowane (bo zmienne Y_{1t} mają rozkład jednopunktowy), zaś jako warunkowe rozkłady Y_{2t} przyjmujemy różne rozkłady Poissona – przez analogię do (20):

$$\Pr\left\{Y_{2t}=j \,\middle|\, Y_{3t}=0, Y_{1t}=0\right\} = h_{0,t}(j) = \exp[-\lambda_{2t,0}](\lambda_{2t,0})^j / j!, \quad \lambda_{2t,0} = \exp(w_t \beta_{2,0}). \tag{22}$$

Jeśli $\beta_2 = \beta_{2,0}$, to $\Pr^*\left\{Y_{2t}=j \,\middle|\, Y_{3t}=1, Y_{1t}=0\right\} = \Pr\left\{Y_{2t}=j \,\middle|\, Y_{3t}=0, Y_{1t}=0\right\}$, czyli sposób generowania wartości zmiennej Y_{2t} przy $Y_{1t} = 0$ jest identyczny bez względu na wartość zero-jedynkowej zmiennej Y_{3t}. Do weryfikacji hipotezy, że $\beta_2 = \beta_{2,0}$, potrzebny jest trójwymiarowy model statystyczny, tj. parametryczna klasa rozkładów postaci:

$$\Pr\{Y_{1t}=i, Y_{2t}=j, Y_{3t}=l; \theta\} = \begin{cases} p_t g_t^*(i) h_t(j,i), & i,j \in N \cup \{0\}, l=1, \\ (1-p_t) h_{0,t}(j), & i=0, j \in N \cup \{0\}, l=0, \\ 0, & i \in N, j \in N \cup \{0\}, l=0, \end{cases} \tag{23}$$

gdzie $p_t = \Pr\{Y_{3t}=1\} = 1 - F(-z_t \beta_3)$, z_t jest wektorem zmiennych objaśniających, zaś F jest dystrybuantą reprezentującą dychotomiczny model zmiennej Y_{3t}. W badaniu empirycznym przyjmujemy model logitowy, czyli zakładamy dystrybuantę rozkładu logistycznego. Warto byłoby w przyszłości rozważyć modele dychotomiczne oparte na dystrybuancie skośnego rozkładu t Studenta, które J. Osiewalski i J. Marzec [2004a, 2004b] wprowadzili jako alternatywę dla dwóch podstawowych specyfikacji: logitowej i probitowej. W modelu statystycznym trójki zmiennych (Y_{1t}, Y_{2t}, Y_{3t}) wektor parametrów θ jest kolumną grupującą $\delta, \alpha, \beta_1, \beta_2, \beta_3$ i $\beta_{2,0}$. Zakładamy, że przy dowolnie ustalonym θ trójwymiarowe obserwacje są niezależne.

Jeśli zaobserwowano $Y_{1t} = y_{1t}$, $Y_{2t} = y_{2t}$ i $Y_{3t} = y_{3t}$ $(t = 1, 2, \ldots, T)$, to odpowiadająca tym wartościom funkcja wiarygodności ma postać:

$$\begin{aligned} L(\theta; y) = &\left[\prod_{t:\, y_{3t}=1,\, y_{1t}=0} \gamma_t h_t(y_{2t}, 0)\right]\left[\prod_{t:\, y_{3t}=1,\, y_{1t}>0} \frac{1-\gamma_t}{1-g_t(0)} g_t(y_{1t}) h_t(y_{2t}, y_{1t})\right] \\ &\left[\prod_{t:\, y_{3t}=1,\, y_{1t}=0} h_{0,t}(y_{2t})\right]\left[\prod_{t:\, y_{3t}=1} p_t\right]\left[\prod_{t:\, y_{3t}=0} (1-p_t)\right] = L_1(\beta_1, \beta_2, \alpha, \delta) L_2(\beta_{2,0}) L_3(\beta_3), \end{aligned} \tag{24}$$

gdzie y oznacza macierz $(3 \times T)$ zawierającą zaobserwowane wartości zmiennych Y_{1t}, Y_{2t} i Y_{3t}. Dwa pierwsze czynniki funkcji wiarygodności L odpowiadają dwuwymiarowej składowej mieszanki i tworzą funkcję L_1 parametrów $\delta, \alpha, \beta_1, \beta_2$; kolejny czynnik odpowiada składowej jednowymiarowej i stanowi funkcję L_2 parametru $\beta_{2,0}$; ostatnie dwa czynniki odpowiadają dychotomicznej zmiennej przełącznikowej i tworzą funkcję L_3 parametru β_3. Jeśli brak jest związków między trzema wyróżnionymi grupami parametrów, czyli charakteryzują się one swobodą

zmienności (*variation freeness*), to wnioskowanie o każdej z nich prowadzi się odrębnie, na podstawie odpowiedniego czynnika L_r ($r = 1, 2, 3$), a nie na podstawie pełnej funkcji wiarygodności. Zauważmy, że faktoryzacja funkcji wiarygodności nie wynika z konkretnej postaci rozkładów, które przyjmujemy, lecz z samej struktury modelu statystycznego i ze swobody zmienności poszczególnych grup parametrów. Na potrzeby wnioskowania brak związków między parametrami (lub ich obecność) formalizuje się ściśle na gruncie bayesowskim, gdzie wprowadza się rozkład *a priori* (miarę probabilistyczną) na przestrzeni parametrów i można wtedy rozważać stochastyczną niezależność grup (wektorów) parametrów. W tej pracy skupiamy się na dwóch przypadkach: niezależności *a priori* trzech wyróżnionych grup parametrów i równości $\beta_2 = \beta_{2,0}$.

Przy separowalności funkcji wiarygodności niezależność *a priori* $(\delta, \alpha, \beta_1, \beta_2)$, $\beta_{2,0}$ i β_3 prowadzi do niezależności *a posteriori* tych parametrów (czyli niezależności warunkowej przy znanych obserwacjach). Oznacza to całkowitą odrębność wnioskowania o każdej grupie parametrów oraz zasadność wykorzystania tylko danych z $y_{3t} = 1$ w estymacji $(\delta, \alpha, \beta_1, \beta_2)$ i tylko danych z $y_{3t} = 0$ w estymacji $\beta_{2,0}$. Oczywiście, wnioskowanie o funkcjach wektora θ wszystkich parametrów – takich jak np. $Corr\left(Y_{1t}, Y_{2t} \mid \theta \right)$, tj. bezwarunkowy współczynnik korelacji między pierwszymi dwoma elementami trójki (Y_{1t}, Y_{2t}, Y_{3t}) – musi być oparte na rozkładzie *a posteriori* całego wektora θ, wykorzystującym pełną funkcję wiarygodności i kompletne dane. Empirycznie interesujące może być porównanie dwóch współczynników korelacji – bezwarunkowego $Corr\left(Y_{1t}, Y_{2t} \mid \theta \right)$ i warunkowego, określanego tylko na podstawie modelu ZIP-CP, tj. $Corr\left(Y_{1t}, Y_{2t} \mid Y_{3t} = 1, \theta \right) = Corr^{*}\left(Y_{1t}, Y_{2t} \mid Y_{3t} = 1, \delta, \alpha, \beta_1, \beta_2 \right)$.

W przypadku gdy zakładamy $\beta_2 = \beta_{2,0}$, czyli niezależny od Y_{3t} mechanizm generowania Y_{2t}, czynników L_1 i L_2 funkcji wiarygodności nie da się rozważać odrębnie, gdyż oba zależą od tego samego wektora β_2. Wnioskowanie o samych parametrach (a nie tylko o takich ich funkcjach, jak bezwarunkowy współczynnik korelacji) musi wykorzystywać pełną funkcję wiarygodności, opartą na wszystkich obserwacjach. Ograniczenie się we wnioskowaniu jedynie do danych z $y_{3t} = 1$ oznaczałoby „błąd selekcji próby". Testowanie hipotezy $\beta_2 = \beta_{2,0}$ możliwe jest oczywiście tylko w przypadku nienarzucającym takiej restrykcji.

Pełna specyfikacja bayesowskiego modelu statystycznego z rozkładem próbkowym postaci (23), prowadzącym do funkcji wiarygodności (24), wymaga przyjęcia konkretnego rozkładu *a priori* wektora θ. 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 modelowi bez zmiennych objaśniających. 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 wiedzę wstępną, gwarantując zarazem 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 danych.

4. Łączne modelowanie liczby płatności kartą i gotówką

Aby zilustrować przydatność empiryczną zaproponowanego modelu statystycznego, w szczególności analizy konsekwencji selekcji próby, wykorzystamy dane, które zgromadzono w celu prowadzenia badania opisanego w pracach [Polasik, Marzec, Fiszeder i Górka 2012] oraz [Marzec, Polasik i Fiszeder 2013][1]. Dane te zawierają m.in. informacje o posiadaniu karty płatniczej (y_{3t}) oraz o liczbie płatności gotówką (y_{2t}) i kartą (y_{1t}) dokonanych (w miesiącu) przez $T = 2518$ osób, które były ankietowane w jednym spośród trzech miesięcy: w październiku, listopadzie 2010 r. albo w styczniu 2011 r. Osoby nieposiadające karty stanowiły 52,7% wszystkich badanych.

Z analizy dwuwymiarowego rozkładu empirycznego dla liczby płatności gotówką i kartą warunkowego względem $y_{3t} = 1$ (w tym jego rozkładów brzegowych) wynika, że w sytuacji posiadania karty płatniczej średnia liczba płatności gotówką wynosi 20,5 (odchylenie standardowe jest równe 17,3), a średnia liczba płatności kartą wynosi 5 (przy odchyleniu standardowym 6,7). Współczynnik warunkowej korelacji empirycznej między y_{1t} a y_{2t} (przy $y_{3t} = 1$) kształtuje się na poziomie 0,008, co można interpretować jako brak nawet przybliżonej zależności liniowej między liczbą płatności kartą i gotówką. Jednowymiarowe rozkłady empiryczne przy $y_{3t} = 1$ sugerują potrzebę zastosowania modelu, w którym obie zmienne skokowe charakteryzują się rozkładem z nadwyżką zer (zob. [Marzec i Osiewalski 2012]).

Szeregi rozdzielcze liczby płatności gotówką przy $y_{3t} = 1$ i $y_{3t} = 0$ podano w tabeli 1. W przypadku braku karty płatniczej średnia liczba transakcji gotówką wynosi 22,5 (±19,8) i jest wyższa niż w przypadku posiadania karty. Także mediana empirycznego rozkładu y_{2t} przy $y_{3t} = 0$ jest przesunięta na prawo w stosunku do mediany rozkładu y_{2t} przy $y_{3t} = 1$. Wyniki zmodyfikowanego

[1] Badanie to, obejmujące m.in. zebranie materiału statystycznego przez TNS Pentor, zostało sfinansowane przez Narodowy Bank Polski.

testu W^2 Andersona i Darlinga [1954][2] wskazują na silne niepodobieństwo tych rozkładów.

Tabela 1. Empiryczne rozkłady liczby płatności gotówką y_{2t}, warunkowe względem posiadania ($y_{3t} = 1$) lub nieposiadania ($y_{3t} = 0$) karty płatniczej

Liczba płatności	Częstość ($y_{3t} = 1$)	Struktura (w %)	Częstość ($y_{3t} = 0$)	Struktura (w %)
0	24	2	0	0
(0; 5]	126	11	60	5
(5; 10]	248	21	275	21
(10; 15]	196	16	224	17
(15; 20]	148	12	208	16
(20; 25]	108	9	151	11
(25; 30]	85	7	123	9
(30; 35]	66	6	73	5
(35; 40]	55	5	57	4
(40; 45]	32	3	55	4
(45; 50]	32	3	26	2
> 50	70	6	76	6
Łącznie	1190	100	1328	100
Średnia	20,5	–	22,5	–
Mediana	16	–	18	–

Źródło: opracowanie własne.

Wyniki uzyskane w modelu P-CP na podstawie danych obejmujących 1190 posiadaczy kart wskazywały na niewielką dodatnią korelację między liczbą płatności gotówką i kartą. J. Marzec i J. Osiewalski [2012] potwierdzili to, stosując model ZIP-CP, ale jednocześnie pokazali, że jego redukcja do P-CP nie jest zasadna (zob. też wyniki dla parametrów α i δ prezentowane w tabeli 4). Korzystając z formalnego bayesowskiego porównywania modeli (poprzez czynnik Bayesa) ustalono ponadto, że w modelu ZIP-CP zmienna Y_{1t} musi wyrażać liczbę transakcji kartą, a Y_{2t} liczbę transakcji gotówką (nie na odwrót). Konieczność

[2] Test w wersji dla zmiennych skokowych, dany formułą:

$$W^2 = N\sum_{i=1}^{N} \frac{\left(F^{em}(a_{1,i}) - F^{em}(a_{0,i})\right)^2}{F^{em}(a_{0,i})\left(1 - F^{em}(a_{0,i})\right)} p^{em}(a_{0,i}),$$

zastosowano dla dwóch szeregów składających się z $N = 363$ obserwacji, gdzie F^{em} to dystrybuanta empiryczna, p^{em} to częstość, $a_{0,i}$ i $a_{1,i}$ są realizacjami zmiennych $Y_2 \mid Y_3 = 0$ i $Y_2 \mid Y_3 = 1$. Wartość statystyki W^2 wyniosła 10,9, a kwantyl rzędu 0,9999 rozkładu statystyki przy H_0 wynosi nie więcej niż 6,17 (dla 100 obserwacji).

identyfikacji właściwej kolejności zmiennych wynika z niesymetrycznej struktury modelu dwuwymiarowego.

Obecnie przedstawione zostaną wyniki uzyskane na podstawie pełnego zbioru danych, uwzględniającego osoby bez kart płatniczych. Podobnie jak w pracy [Marzec i Osiewalski 2012] wykorzystano dane surowe, bez wag określających stopień reprezentatywności poszczególnych obserwacji[3]. Rozważamy model statystyczny omówiony w poprzedniej części pracy, na który składają się odrębne modele zmiennych licznikowych dla $T_1 = 1190$ par (Y_{1t}, Y_{2t}) przy $Y_{3t} = 1$ i dla $T_2 = 1328$ zmiennych Y_{2t} przy $Y_{3t} = 0$ oraz łączący je model zmiennej dychotomicznej Y_{3t} (dla $T = T_1 + T_2 = 2518$). Jako zmienne objaśniające wykorzystano główne cechy ankietowanych konsumentów. Przyjęto, że w każdym z trzech modeli składowych występuje ten sam zestaw (potencjalnych) zmiennych objaśniających.

W tabeli 2 przedstawiono zmienne objaśniające i ich typowe wartości, tj. średnie w przypadku zmiennych ciągłych i najczęstsze dla zmiennych dychotomicznych. Warto zwrócić uwagę, że – na prezentowanym poziomie agregacji informacji z badania ankietowego – głównymi czynnikami określającymi posiadanie karty wydają się: deklarowany dochód w rodzinie, wykształcenie, stan cywilny i dostęp do Internetu. Więcej wniosków uzyskamy, analizując wyniki podane w tabeli 3. W całym zbiorze badanych kobiety stanowiły 56% ankietowanych, odsetek osób będących w formalnym związku wyniósł 56%, a 61% ankietowanych posiadało dostęp do Internetu. Czynnikiem wyjaśniającym zróżnicowanie liczby transakcji kartą bądź gotówką może być miejsce zamieszkania. W miastach mieszkało 63% wszystkich badanych. Wśród posiadaczy kart płatniczych 71% było mieszkańcami miast. Odsetek osób bez kart, a mieszkających w miastach, był niższy i wynosił 56%. Bezwarunkowa częstość posiadania karty była równa 47,3%, jednak częstość posiadania karty przez klienta pod warunkiem, że mieszka w mieście, wyniosła 53%. Natomiast udział posiadaczy karty wynosił: 49% wśród mężczyzn, 55% wśród zamężnych albo żonatych oraz 58% wśród osób posiadających dostęp do Internetu.

Uzyskany przy założeniu niezależności *a priori* rozkład *a posteriori* parametrów trójwymiarowego modelu statystycznego, danego wzorem (23), próbkowano stosując metody MCMC (Monte Carlo typu łańcuchów Markowa); zastosowano sekwencyjną wersję algorytmu Metropolisa i Hastingsa. Wykorzystując niezależność *a posteriori*, wynikającą z separowalności funkcji wiarygodności (24) i niezależności *a priori* poszczególnych wektorów parametrów, dokonano osobno estymacji parametrów każdego z trzech modeli składowych, tj. β_1, β_2, α i δ w modelu ZIP-CP (M_1), $\beta_{2,0}$ w modelu Poissona dla liczby transakcji gotówką

[3] W badaniach opisanych w pracach [Polasik, Marzec, Fiszeder i Górka 2012] oraz [Marzec, Polasik i Fiszeder 2013] użyto danych ważonych.

w przypadku braku karty (M_2), oraz β_3 w modelu logitowym posiadania karty (M_3). Łączna liczba parametrów wyniosła 34.

Tabela 2. Przeciętne (średnie lub najczęstsze) wartości zmiennych objaśniających

Zmienna objaśniająca	$T = 2518$	$T_1 = 1190$	$T_2 = 1328$
	łącznie	posiadający kartę	bez karty
Płeć (1 – mężczyzna, 0 – kobieta)	0	0	0
Wiek (w latach)	41,2	40,1	42,2
Stan cywilny (1 – żonaty lub zamężna, 0 – nie)	1	1	0
Miejsce zamieszkania (1 – miasto, 0 – wieś)	1	1	1
Miesięczny dochód w rodzinie (w tys. zł)	2,9	3,3	2,5
Wykształcenie (lata nauki)	12,3	13,2	11,5
Dostęp do Internetu (1 – tak, 0 – nie)	1	1	0

Źródło: opracowanie własne.

Tabela 3. Udział wartości 1 w przypadku dychotomicznych zmiennych objaśniających (w %)

Zmienna objaśniająca	$T = 2518$	$T_1 = 1190$	$T_2 = 1328$
	łącznie	posiadający kartę	bez karty
Płeć (1 – mężczyzna, 0 – kobieta)	44	45	42
Stan cywilny (1 – żonaty lub zamężna, 0 – nie)	56	65	48
Miejsce zamieszkania (1 – miasto, 0 – wieś)	63	71	56
Dostęp do Internetu (1 – tak, 0 – nie)	61	76	49

Źródło: opracowanie własne.

W tabeli 4 zaprezentowano wartości oczekiwane i odchylenia standardowe *a posteriori* parametrów. Postulowany w modelu M_1 wpływ wszystkich zmiennych objaśniających na liczbę płatności gotówką, gdy konsument korzysta równocześnie z karty, został potwierdzony przez dane. Natomiast tylko posiadanie przez konsumenta dostępu do Internetu, jego wykształcenie i dochód powodują znaczące zróżnicowanie liczby płatności kartą. W czystym modelu Poissona (M_2) płeć konsumenta i jego dochód wydają się nie mieć wpływu na zróżnicowanie liczby transakcji gotówką w sytuacji braku karty. W modelu logitowym (M_3) determinantami posiadania karty płatniczej okazują się wszystkie zmienne objaśniające z wyjątkiem wieku, który wyraźnie nie ma znaczenia.

Tabela 4. Wartości oczekiwane i odchylenia standardowe *a posteriori* parametrów (β) modeli

Zmienna/parametr	Model	$E(\beta \mid y)$	$D(\beta \mid y)$	Model	$E(\beta \mid y)$	$D(\beta \mid y)$
„1”	M_1: płatności kartą	0,911	0,098	M_3: model logitowy posiadania karty	–5,455	0,330
Płeć		–0,044	0,025		0,181	0,092
Wiek		–0,002	0,001		0,001	0,003
Stan cywilny		–0,048	0,029		0,596	0,102
Miejsce zamieszkania		–0,007	0,028		0,483	0,096
Dochód		0,051	0,010		0,185	0,039
Wykształcenie		0,056	0,006		0,297	0,024
Internet		0,361	0,039		0,622	0,106
„1”	M_1: płatności gotówką	2,825	0,050	M_2: model Poissona – transakcje gotówką (gdy brak karty)	2,133	0,051
Płeć		–0,101	0,013		–0,014	0,013
Wiek		0,008	0,001		0,002	0,001
Stan cywilny		–0,158	0,015		0,082	0,015
Miejsce zamieszkania		0,145	0,015		0,126	0,015
Dochód		0,016	0,006		–0,009	0,006
Wykształcenie		–0,008	0,003		0,062	0,003
Internet		–0,085	0,016		0,152	0,016
α	–	0,004	0,001	–	–	–
δ	–	–1,876	0,041	–	–	–

Źródło: opracowanie własne.

Warto zauważyć, że występują duże różnice w wartościach oczekiwanych *a posteriori* parametrów opisujących liczbę transakcji gotówką w modelach M_1 i M_2. Aż dla czterech (z siedmiu) zmiennych objaśniających (stan cywilny, dochód, wykształcenie i Internet) znaki tych charakterystyk są przeciwne – co oznacza, że kierunek wpływu danej zmiennej na liczbę wykonanych transakcji gotówką jest inny w zależności od tego, czy konsument posiada dodatkowy instrument płatności w postaci karty. Odchylenia standardowe *a posteriori* większości parametrów są stosunkowo małe. Spostrzeżenia te sugerują, że równość $\beta_2 = \beta_{2,0}$ nie zachodzi. Oznaczałoby to, że we wnioskowaniu o samych parametrach selekcja próby nie powoduje negatywnych konsekwencji i można ograniczyć się do każdego modelu oddzielnie (dla odpowiednich podzbiorów danych).

W celu zweryfikowania hipotezy $\beta_2 = \beta_{2,0}$ zastosowano bayesowski odpowiednik testu chi-kwadrat. Niech $\kappa = \beta_2 - \beta_{2,0}$; opierając się na idei testu nie-

bayesowskiego, dla zmiennej wielowymiarowej κ rozważa się formę kwadratową postaci (por. [Marzec, Osiewalski 2008]):

$$\tau = \tau(\kappa; y) = (\kappa - E(\kappa \mid y))'(V(\kappa \mid y))^{-1}(\kappa - E(\kappa \mid y)), \tag{25}$$

gdzie $E(\kappa \mid y) = E\left(\beta_2 \mid y\right) - E\left(\beta_{2,0} \mid y\right)$ i $V(\kappa \mid y) = V\left(\beta_2 \mid y\right) + V\left(\beta_{2,0} \mid y\right)$, a sumowanie macierzy kowariancji *a posteriori* wynika z niezależności *a posteriori* obu porównywanych wektorów parametrów, zachodzącej w proponowanym modelu ogólnym (bez restrykcji). Jednowymiarowa zmienna τ jest losowa jako funkcja zarówno obserwacji, jak i parametrów modelu; we wnioskowaniu na podstawie danych interesuje nas jej rozkład *a posteriori*, czyli warunkowy względem danych, o gęstości $p(\tau \mid y)$. Testowanie hipotezy $\kappa = 0$ sprowadza się do zbadania, czy wartość $\tau(0; y)$ leży w obszarze największej gęstości $p(\tau \mid y)$, przy wysokim, ustalonym prawdopodobieństwie *a posteriori* $(1 - \alpha)$ tego obszaru. Jeśli tak, to nie odrzucamy hipotezy $\kappa = 0$ i przechodzimy do analizy modelu z tą restrykcją, uniemożliwiającą odrębne traktowanie dwóch podzbiorów obserwacji. Jeśli wartość $\tau(0; y)$ znajduje się poza obszarem wysokiej gęstości *a posteriori*, to równość $\kappa = 0$ jest nieuzasadniona w świetle dostępnych danych i ją odrzucamy, pozostając przy wnioskach z modelu ogólnego, umożliwiającego wstępny podział obserwacji na dwie grupy i osobne traktowanie każdej z nich.

Wyniki uzyskane za pomocą tego testu świadczą przeciwko hipotezie $\kappa = 0$. Rozkład *a posteriori* zmiennej losowej $\tau(\kappa; y)$ jest jednomodalny i prawostronnie asymetryczny, a jego modalna wynosi 5,7. Przedział (0, 20) zawiera $\tau(\kappa; y)$ z prawdopodobieństwem *a posteriori* równym 0,99, zaś wartość $\tau(0; y)$ wynosi 973,85. Zatem wartość ta znajduje się bardzo daleko w prawym ogonie rozkładu *a posteriori* dla τ, czyli założenie równości wektorów β_2 i $\beta_{2,0}$ nie znajduje uzasadnienia. Wystarczająca jest estymacja parametrów z wykorzystaniem separowalności funkcji wiarygodności danej wzorem (24); taka estymacja parametrów nie jest obarczona „błędem selekcji próby".

Na koniec prezentujemy wyniki dla współczynników korelacji między liczbą transakcji oboma instrumentami płatniczymi (Y_{1t}, Y_{2t}). Przypomnijmy, że współczynnik ten jest funkcją wszystkich parametrów trzech podmodeli, więc bez względu na wynik wcześniej prezentowanego testu wyznaczenie jego charakterystyk *a posteriori* jest możliwe tylko w modelu łącznym. Syntetyczne wyniki estymacji zostały pokazane w tabeli 5. Dla wszystkich obserwacji (T = 2518) otrzymano rozkłady *a posteriori* dla $Corr\left(Y_{1t}, Y_{2t} \mid \theta\right)$, czyli bezwarunkowego współczynnika korelacji, skupione blisko zera – ale wyłącznie po stronie wartości dodatnich; charakteryzowały się one małym odchyleniem standardowym. Średnia wartość oczekiwana *a posteriori* wyniosła 0,072, przy czym najmniejsza 0,031, a największa 0,16; korelacja jest więc bardzo słaba, ale dodatnia.

Tabela 5. Uśrednione (po obserwacjach) wartości oczekiwane *a posteriori* współczynników korelacji pary (Y_{1t}, Y_{2t})

Współczynnik korelacji	Średnia ocena	
	gdy konsument posiada kartę ($Y_{3t} = 1$)	gdy brak karty ($Y_{3t} = 0$)
$Corr\left(Y_{1t}, Y_{2t} \mid \theta\right)$	0,072 (dla $T = 2518$)	
$Corr\left(Y_{1t}, Y_{2t} \mid \theta\right)$	0,065 (dla $T_1 = 1190$)	0,079 (dla $T_2 = 1328$)
$Corr\left(Y_{1t}, Y_{2t} \mid Y_{3t} = 1, \theta\right)$	0,073 (dla $T_1 = 1190$)	–

Źródło: opracowanie własne.

W modelu ZIP-CP – tylko dla posiadaczy karty płatniczej – średnia wartość oczekiwana *a posteriori* warunkowego współczynnika korelacji $Corr\left(Y_{1t}, Y_{2t} \mid Y_{3t} = 1, \theta\right)$ wyniosła 0,073. Średnia ocena korelacji bezwarunkowej (między liczbą transakcji kartą i gotówką) jest praktycznie taka sama jak korelacji warunkowej przy $Y_{3t} = 1$, choć średnie cząstkowe korelacji bezwarunkowej (liczone dla posiadaczy karty i dla osób bez karty) są odmienne.

5. Podsumowanie

Omawiany trójwymiarowy rozkład skokowy i zbudowany na tej podstawie bayesowski model statystyczny zaproponowano w celu łącznego modelowania dwóch zmiennych licznikowych, z których pierwsza może być zdegenerowana w zerze. Zaproponowany przez autorów model statystyczny polega na zastosowaniu zmiennej dychotomicznej (zero-jedynkowej) do przełączania między dwoma modelami zmiennych licznikowych: dwu- i jednowymiarowym, przy czym model jednowymiarowy jest otrzymywany z dwuwymiarowego przez odpowiednie warunkowanie. O ile przedstawiony schemat modelowania ma walor ogólności, o tyle wybór konkretnych klas modeli składowych może podlegać zmianom. Wybierając specyfikację ZIP-CP dla dwuwymiarowej zmiennej licznikowej i logistyczną dla zmiennej dychotomicznej kierowano się prostotą obu, prowadzącą do prostego modelu trójwymiarowego, a także zbadanymi dobrymi własnościami modelu ZIP-CP. Zmiana specyfikacji logistycznej na inną, np. opartą na dystrybuancie skośnego rozkładu Studenta i wykorzystującą interakcje między zmiennymi objaśniającymi (tzw. model II rzędu, zob. [Osiewalski i Marzec 2004a]), nie jest trudna i może podnieść jakość modelu – choć nie musi (jest to kwestia empiryczna). Trudne będzie zastąpienie specyfikacji ZIP-CP, głównej części modelu przełącznikowego, innym modelem. Użycie w tym celu

alternatywnych specyfikacji dla dwóch powiązanych zmiennych licznikowych będzie przedmiotem dalszych badań autorów.

Jeśli chodzi o specyficznie bayesowski element zaproponowanego modelu – rozkład *a priori*, to jego postać może oczywiście podlegać zmianom, ale należy zwrócić uwagę na dwa kluczowe elementy. Separowalność funkcji wiarygodności względem parametrów modeli składowych może być w pełni wykorzystana tylko przy niezależności *a priori* tych grup parametrów, więc nie należy z niezależności rezygnować. Z kolei szczególna postać rozkładu *a priori* (normalny o średniej 0 i wariancji 1), przyjęta przez nas dla każdego indywidualnego parametru, nie ma znaczenia, jeśli liczba obserwacji jest bardzo duża (jak w zaproponowanym przykładzie empirycznym). Oczywiście, przy małej liczbie obserwacji warto dokonać analizy wrażliwości wyników na rozkład *a priori* w ramach szerszej klasy (np. rozkładów Studenta).

W zaproponowanym modelu bayesowskim można łatwo zastosować test typu Lindleya, tj. bayesowski odpowiednik testu F bądź chi-kwadrat, by zbadać zasadniczą restrykcję identyczności parametrów opisujących w dwóch sytuacjach tę zmienną licznikową, która jest niezdegenerowana (jest nietrywialnie obserwowana) dla obu wartości zmiennej zero-jedynkowej. W dalszych badaniach warto zastosować bayesowskie porównywanie mocy wyjaśniającej konkurencyjnych modeli poprzez ich prawdopodobieństwa *a posteriori*, co wymaga odpowiedniej metody obliczenia brzegowej gęstości wektora obserwacji w każdym z modeli. W przypadku stosowania metod Monte Carlo łańcuchów Markowa (w celu próbkowania rozkładu *a posteriori* co najwyżej kilkudziesięciu parametrów modeli) właściwym narzędziem jest skorygowany estymator średniej arytmetycznej, który zaproponowała A. Pajor [2017].

Poza testowaniem ważnej restrykcji, proponowany model trójwymiarowy pozwala badać skutki wstępnej selekcji obserwacji, polegającej na usunięciu tych, dla których obserwowana jest tylko jedna zmienna licznikowa. W przykładzie empirycznym, dotyczącym modelowania liczby płatności dwoma instrumentami płatniczymi (kartą i gotówką), wykazano, że samo wnioskowanie o poszczególnych parametrach nie było zagrożone „błędem selekcji próby", gdyż odrzucono restrykcję wiążącą parametry opisujące liczbę płatności gotówkowych w sytuacjach posiadania karty płatniczej i jej braku. Pokazano też, że głębsze wnioskowanie o korelacji między liczbami płatności kartą i gotówką, odróżniające korelację warunkową (względem posiadania karty) od bezwarunkowej, możliwe jest dopiero na gruncie modelu pełnego (trójwymiarowego). Empirycznie, oba rodzaje korelacji okazały się podobne co do wartości.

Literatura

Anderson T.W., Darling D.A. [1954], *A Test of Goodness of Fit*, „Journal of the American Statistical Association", vol. 49, nr 268.

Berkhout P., Plug E. [2004], *A Bivariate Poisson Count Data Model Using Conditional Probabilities*, „Statistica Neerlandica", vol. 58, nr 3, https://doi.org/10.1111/j.1467-9574.2004.00126.x.

Cameron A.C., Trivedi P.K. [1998], *Regression Analysis of Count Data*, Cambridge University Press, New York.

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

Fiszeder P., Polasik M. [2009], *Modelowanie liczby transakcji dokonywanych przy użyciu gotówki i kart płatniczych na rynku polskim*, „Acta Universitatis Nicolai Copernici – Ekonomia", vol. 39, https://doi.org/10.12775/AUNC_ECON.2009.029.

Goczek Ł., Witkowski B. [2015], *The Determinants of Cash-free Transactions*, „The National Bank of Poland Working Paper Series", nr 146.

Goczek Ł. Witkowski B. [2016], *Determinants of Card Payments*, „Applied Economics", vol. 48, https://doi.org/10.1080/00036846.2015.1102846.

Górka J. [2013], *Efektywność instrumentów płatniczych w Polsce*, Wydawnictwo Naukowe Wydziału Zarządzania Uniwersytetu Warszawskiego, Warszawa.

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", vol. 34, https://doi.org/10.2307/1269547.

Marzec J. [2012], W*ybrane dwuwymiarowe modele dla zmiennych licznikowych w ekonomii* „Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie – Metody Analizy Danych", nr 884.

Marzec J., Osiewalski J. [2008], *Bayesian Inference on Technology and Cost Efficiency of Bank Branches*, „Bank i Kredyt", vol. 39, nr 9.

Marzec J., Osiewalski J. [2012], *Dwuwymiarowy model typu ZIP-CP w łącznej analizie zmiennych licznikowych*, „Folia Oeconomica Cracoviensia", nr 53.

Marzec J., Polasik M., Fiszeder P. [2013], *Wykorzystanie gotówki i karty płatniczej w punktach handlowo-usługowych w Polsce: zastosowanie dwuwymiarowego modelu Poissona*, „Bank i Kredyt" vol. 44, nr 4.

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*, Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków.

Osiewalski J., Marzec J. [2004a], *Model dwumianowy II rzędu i skośny rozkład Studenta w analizie ryzyka kredytowego*, „Folia Oeconomica Cracoviensia", nr 45.

Osiewalski J., Marzec J. [2004b], *Uogólnienie dychotomicznego modelu probitowego z wykorzystaniem skośnego rozkładu Studenta*, „Przegląd Statystyczny", t. 51.

Pajor A. [2017], *Estimating the Marginal Likelihood Using the Arithmetic Mean Identity*, „Bayesian Analysis", vol. 12, nr 1, https://doi.org/10.1214/16-BA1001.

Polasik M. [2015], *Stan i potencjał rozwoju sieci akceptacji kart płatniczych w Polsce*, „Acta Universitatis Nicolai Copernici, Ekonomia", vol. 46, https://doi.org/10.12775/AUNC_ECON.2015.002.

Polasik M., Maciejewski K. [2009], *Innowacyjne usługi płatnicze w Polsce i na świecie*, „Materiały i Studia NBP", nr 241.
Polasik M., Marzec J., Fiszeder P., Górka J. [2012], *Modelowanie wykorzystania metod płatności detalicznych na rynku polskim*, „Materiały i Studia NBP", nr 265.
Polasik M., Wisniewski T.P., Lightfoot G. [2012], *Modelling Customers' Intentions to Use Contactless Cards*, „International Journal of Banking, Accounting and Finance", vol. 4, nr 3, https://doi.org/10.1504/IJBAAF.2012.051590.
Winkelman R. [2008], *Econometric Analysis of Count Data*, Springer-Verlag, Berlin Heidelberg.

Bivariate Count Variables – Bayesian Modelling of Sample Selection

(Abstract)

The article presents a joint statistical model of two count variables, one of which can be degenerated at zero. We consider a modelling framework based on switching between a bivariate Poisson regression model and a univariate one, where the switching depends on the observed outcome of the third, dichotomous variable. Bayesian analysis is advocated; in two special cases of our Bayesian model, important consequences for inference are stated. In the empirical section we consider joint modelling of the number of cash and bank card transactions in Poland with the use of data for both cardholders and non-holders.

Keywords: bivariate Poisson regression models, switching between non-degenerate and degenerate distributions, likelihood factorisation, bank card and cash payments.