

Autorzy: Jacek Osiewalski¹, Jerzy Marzec²

Dwuwymiarowe zmienne licznikowe - bayesowskie modelowanie selekcji próby³

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, proponujemy 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.

¹ Katedra Ekonometrii i Badań Operacyjnych Uniwersytetu Ekonomicznego w Krakowie, ul. Rakowicka 27, 31-510 Kraków; e-mail: eeosiewa@cyf-kr.edu.pl.

² Katedra Ekonometrii i Badań Operacyjnych Uniwersytetu Ekonomicznego w Krakowie, ul. Rakowicka 27, 31-510 Kraków; e-mail: marzecj@uek.krakow.pl.

³ Praca sfinansowana ze środków przyznanych Wydziałowi Zarządzania Uniwersytetu Ekonomicznego w Krakowie, w ramach dotacji na utrzymanie potencjału badawczego.

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 mieszaniki rozkładów. Niestandardowym modelem jest warunkowy model Poissona, który zaproponowali Berkhout i Plug [2004]. Omówienie tych zagadnień, wraz z odwołaniami do literatury, można znaleźć m.in. w artykule Marzec [2012]. Warto podkreślić, że kontekście modeli dwuwymiarowych nie pojawia się kwestia selekcji próby.

Jako główną część składową proponowanego modelu statystycznego wykorzystujemy specyfikację ZIP-CP (ang. *zero inflated Poisson - conditional Poisson*), którą zaproponowali Marzec i Osiewalski [2012]. Jest to dwuwymiarowa regresja typu Poissona, ogólniejsza niż model P-CP (*Poisson – conditional Poisson*), który wprowadzili Berkhout i 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” (ang. *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 Osiewalski [2012]. Druga część naszego 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 (drugi) 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 omawiamy nasz model statystyczny, postać funkcji wiarygodności i analizę bayesowską tego modelu, zwracając uwagę na jego dwa przypadki szczególne. W czwartym, empirycznym punkcie pracy przedstawiamy 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 nie posiadających kart). W części piątej zawarto uwagi końcowe.

Nasz przykład empiryczny wpisuje się badania rozwoju obrotu bezgotówkowego w Polsce, które są prowadzone od kilku lat; zob. np. [Polasik 2009, 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 ($\Pr\{Y_1=0\}=1$) zawsze gdy $Y_3=0$, 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:

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

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

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

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

$$E(Y_2 \mid Y_3 = 1) = \lambda_2 \exp[\lambda_1 (e^\alpha - 1)], \quad (3)$$

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

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

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 częściej 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 wprowadzenie rozkładu typu ZIP na miejsce brzegowego rozkładu Poissona. Modele regresji dla skokowej zmiennej objaśnianej z nadmierną liczbą zer spopularyzował głównie Lambert [1992], a Cameron i Trivedi [1998, 2005] oraz 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 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 :

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

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

$$\Pr^*\{Y_1 = i | Y_3 = 1\} = g^*(i) = \begin{cases} \gamma & \text{dla } i = 0, \\ \frac{1-\gamma}{1-g(0)}g(i) & \text{dla } i \in N, \end{cases} \quad (7)$$

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 $\Pr^*\{Y_1 = i | Y_3 = 1\} = g^*(i) = g(i) = \Pr\{Y_1 = i | Y_3 = 1\}$ 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

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

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

$$E^*(Y_1 | Y_3 = 1) = (1 - g(0))^{-1}(1 - \gamma)\lambda_1, \quad (9)$$

$$E^*(Y_2 | Y_3 = 1) = (1 - g(0))^{-1}[(1 - \gamma)E(Y_2 | Y_3 = 1) + (\gamma - g(0))\lambda_2], \quad (10)$$

$$\text{Var}^*(Y_1 | Y_3 = 1) = \frac{1-\gamma}{1-g(0)}\lambda_1(1 + \frac{\gamma-g(0)}{1-g(0)}\lambda_1), \quad (11)$$

$$\text{Var}^*(Y_2 | Y_3 = 1) = \frac{1-\gamma}{1-g(0)}\{\text{Var}(Y_2 | Y_3 = 1) + \frac{\gamma-g(0)}{1-g(0)}[E(Y_2 | Y_3 = 1) - \lambda_2]^2 + \frac{\gamma-g(0)}{1-\gamma}\lambda_2\}, \quad (12)$$

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

gdzie $E(Y_2 | Y_3 = 1)$ i $\text{Var}(Y_2 | Y_3 = 1)$ 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 $[(1 - e^{-\lambda_1})e^\alpha - (1 - \gamma)]\exp(\lambda_1(e^\alpha - 1)) < \gamma - e^{-\lambda_1}$,

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

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

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 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 (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, by można 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

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

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

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

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

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

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

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

$$E(Y_1^m Y_2^n) = p E^*(Y_1^m Y_2^n \mid Y_3 = 1) + (1-p) 0^m E(Y_2^n \mid Y_3 = 0, Y_1 = 0), \quad (17)$$

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

3. Model statystyczny

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

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

gdzie

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

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

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

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 poprzedniej części pracy, zależą teraz od zmiennych objaśniających.

W literaturze specyfikacja oparta na wzorze (19) jest nazywana modelem płotkowym – ang. *hurdle model*, zob. [Cameron i Trivedi 2005, s. 680]. Porównanie tej specyfikacji z oryginalnym modelem ZIP podaje Winkelmann [2008]. Głównymi zaletami naszej 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):

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

Jeśli $\beta_2 = \beta_{2,0}$, to $\Pr^*\{Y_{2t} = j | Y_{3t} = 1, Y_{1t} = 0\} = \Pr\{Y_{2t} = j | Y_{3t} = 0, Y_{1t} = 0\}$, 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

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

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 Osiewalski i 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 δ , α , β_1 , β_2 , β_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 $Y_{3t} = y_{3t}$ ($t = 1, 2, \dots, T$), to odpowiadająca tym wartościom funkcja wiarygodności ma postać

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

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 δ , α , β_1 , β_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 (ang. *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. $\text{Corr}(Y_{1t}, Y_{2t} | \theta)$, 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 $\text{Corr}(Y_{1t}, Y_{2t} | \theta)$ i warunkowego, określanego tylko na podstawie modelu ZIP-CP, tj. $\text{Corr}(Y_{1t}, Y_{2t} | Y_{3t}=1, \theta) = \text{Corr}^*(Y_{1t}, Y_{2t} | Y_{3t}=1, \delta, \alpha, \beta_1, \beta_2)$.

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 nie narzucającym takiej restrykcyj.

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

W celu ilustracji przydatności empirycznej 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]⁴. 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 roku 2010 albo w styczniu roku 2011. 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. Przy braku karty płatniczej, średnia liczba transakcji gotówką wynosi 22,5 ($\pm 19,8$) 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 $y_{3t}=1$.

⁴ Badanie to, obejmujące m.in. zebranie materiału statystycznego przez TNS Pentor, zostało sfinansowane przez Narodowy Bank Polski.

Wyniki zmodyfikowanego testu W^2 Andersona i Darlinga [1954]⁵ wskazują na silne niepodobieństwo tych rozkładów.

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

Liczba płatności	częstość ($y_{3t}=1$)	Struktura	częstość ($y_{3t}=0$)	Struktura
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.

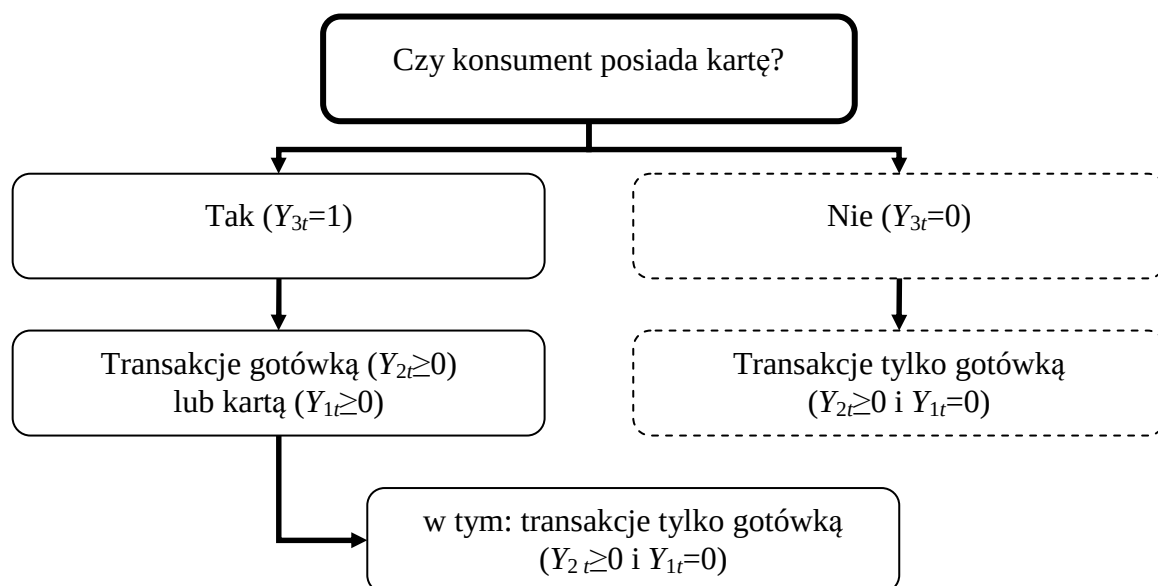
Wyniki uzyskane w modelu P-CP na podstawie danych obejmujących 1190 posiadaczy kart wskazywały na niewielką dodatnią korelację między liczbą płatności gotówką i kartą. Marzec i Osiewalski [2012] potwierdzili to stosując model ZIP-CP, ale jednocześnie pokazali, że jego redukcja do P-CP nie jest zasadna (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 identyfikacji właściwej kolejności zmiennych wynika z niesymetrycznej struktury modelu dwuwymiarowego.

⁵ Test w wersji dla zmiennych skokowych, dany formułą $W^2 = N \sum_{i=1}^N \frac{(F^{em}(a_{1,i}) - F^{em}(a_{0,i}))^2}{F^{em}(a_{0,i})(1 - F^{em}(a_{0,i}))} p^{em}(a_{0,i})$,

zastosowano dla dwóch szeregów składających się z $N=363$ obserwacji, gdzie F^{em} to dystrybuenta empiryczna, p^{em} to częstość, $a_{0,i}$ i $a_{1,i}$ są realizacjami zmiennych $Y_2|Y_3=0$ i $Y_2|Y_3=1$. 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).

Obecnie przedstawimy wyniki uzyskane na podstawie pełnego zbioru danych, uwzględniającego osoby bez kart płatniczych. Podobnie jak Marzec i Osiewalski [2012] wykorzystaliśmy dane surowe, bez wag określających stopień reprezentatywności poszczególnych obserwacji⁶. Na rysunku 1 pokazano schemat zależności między zmiennymi w modelu statystycznym. Rozważamy 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 posiadane charakterystyki ankietowanych konsumentów. Przyjęto, że w każdym z trzech modeli składowych występuje ten sam zestaw (potencjalnych) zmiennych objaśniających.

Wykres 1. Schemat zależności między zmiennymi Y_{1t} , Y_{2t} i Y_{3t} w pełnym modelu statystycznym.



Źródło: opracowanie własne.

W tabeli 2 prezentujemy 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.

⁶ Polasik Marzec, Fiszedler i Górka [2012] oraz Marzec, Polasik i Fiszedler [2013] użyli danych ważonych.

Więcej wniosków uzyskamy analizując wyniki zaprezentowane 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.

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

	$T=2518$	$T_1=1190$	$T_2=1328$
Zmienna objaśniająca	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)	3,5	3,5	1,75
Wykształcenie (lata nauki)	12,5	12,5	11
Posiadanie Internetu (1-tak, 0-nie)	1	1	0

Źródło: opracowanie własne.

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

	$T=2518$	$T_1=1190$	$T_2=1328$
Zmienna objaśniająca	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
Posiadanie Internetu (1-tak, 0-nie)	61	76	49

Źródło: opracowanie własne.

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

W tabeli 4 zaprezentowano wartości oczekiwane i odchylenia standardowe *a posteriori* parametrów. Postulowany w modelu M_1 wpływ poszczególnych zmiennych objaśniających na liczbę płatności gotówką, gdy konsument korzysta równocześnie z karty, został potwierdzony przez dane. Natomiast 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		$E(\beta y)$	$D(\beta y)$		$E(\beta y)$	$D(\beta y)$
„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 . Dla czterech zmiennych objaśniających (na siedem możliwych), tj. 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 F. Niech $\kappa = \beta_2 - \beta_{2,0}$; opierając się na idei klasycznego testu F, dla zmiennej wielowymiarowej κ rozważa się formę kwadratową postaci (por. Marzec, Osiewalski, 2008):

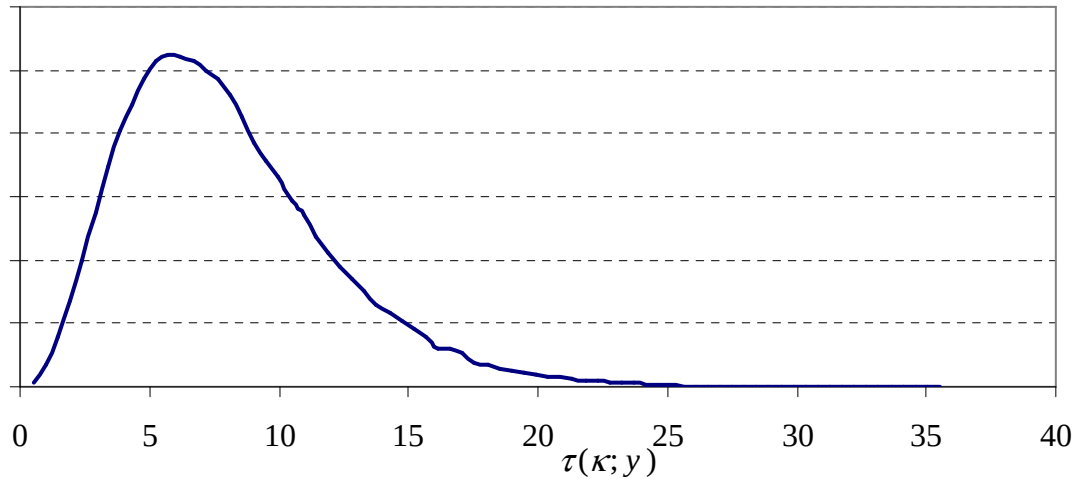
$$\tau = \tau(\kappa; y) = (\kappa - E(\kappa | y))' (V(\kappa | y))^{-1} (\kappa - E(\kappa | y)), \quad (25)$$

gdzie $E(\kappa | y) = E(\beta_2 | y) - E(\beta_{2,0} | y)$ i $V(\kappa | y) = V(\beta_2 | y) + V(\beta_{2,0} | y)$, 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 $p(\tau | y)$. Testowanie hipotezy $\kappa = 0$ sprowadza się do zbadania, czy wartość $\tau(0; y)$ leży w obszarze największej gęstości $p(\tau | y)$, 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 grupy.

Dalsze rezultaty badań otrzymane za pomocą tego testu zaprezentowano na wykresie 2, który przedstawia gęstość *a posteriori* zmiennej losowej $\tau(\kappa; y)$. 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); estymacja parametrów nie jest obciążona „błędem selekcji próby”.

Wykres 2. Rozkład a posteriori zmiennej $\tau(\kappa; y)$.



Źródło: opracowanie własne.

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 pod-modeli, 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 $\text{Corr}(Y_{1t}, Y_{2t} | \theta)$, 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. W modelu ZIP-CP – tylko dla posiadaczy karty płatniczej – średnia wartość oczekiwana *a posteriori* warunkowego współczynnika korelacji $\text{Corr}(Y_{1t}, Y_{2t} | Y_{3t}=1, \theta)$ 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.

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

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

Źródło: opracowanie własne.

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. Nasz model statystyczny polega na zastosowaniu zmiennej dychotomicznej (zero-jedynkowej) do przełączania między dwoma modelami zmiennych licznikowych: dwu- i jednowymiarowym, przy czym jednowymiarowy jest otrzymany z dwuwymiarowego przez odpowiednie warunkowanie. O ile przedstawiony schemat modelowania ma walor ogólności, to wybór konkretnych klas modeli składowych może podlegać zmianom. Wybierając specyfikację ZIP-CP dla dwuwymiarowej zmiennej licznikowej i logistyczną dla zmiennej dychotomicznej kierowaliśmy się prostotą obu, prowadzącą do prostego modelu trójwymiarowego, a także zbadanymi dobrymi własnościami modelu ZIP-CP. Zmiana specyfikacji logistycznej na inną, np. opartą na dystrybucie skośnego rozkładu Studenta i wykorzystującą interakcje między zmiennymi objaśniającymi (tzw. model II rzędu, zob. [Osiewalski i Marzec 2004b]), 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 naszego 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 naszym 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 Pajor [2016].

Poza testowaniem ważnej restrykcji, nasz 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), wykazaliśmy, 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. Pokazaliśmy 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, iss. 268, s. 765–769.
- Berkhout P., Plug E. [2004], *A bivariate Poisson count data model using conditional probabilities*, “Statistica Neerlandica” 58, iss. 3, s. 349–364, <http://dx.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.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, <http://dx.doi.org/10.1080/00401706.1973.10489112>.
- Famoye F., Singh K. P. [2006], *Zero-Inflated Generalized Poisson Regression Model with an Application to Domestic Violence Data*, "Journal of Data Science", vol. 4, no.1, s. 117-130.
- 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, s. 93–104, http://dx.doi.org/10.12775/AUNC_ECON.2009.029.
- Goczek Ł., Witkowski B. [2016], *Determinants of Card Payments*, "Applied Economics", vol. 48, s. 1530-1543, <http://dx.doi.org/10.1080/00036846.2015.1102846>.
- Goczek Ł., Witkowski B. [2015], *The determinants of cash-free transactions*, "The National Bank of Poland Working Paper Series", no. 146.
- Górka J. [2013], *Efektywność instrumentów płatniczych w Polsce*. Warszawa: Wydawnictwo Naukowe Wydziału Zarządzania Uniwersytetu Warszawskiego.
- 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, s. 1-14, <http://dx.doi.org/10.2307/1269547>.
- Marzec J. [2012], *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, s. 59-70.
- Marzec J., Osiewalski J. [2008], *Bayesian inference on technology and cost efficiency of bank branches*, "Bank i Kredyt", vol. 39, nr 9, s. 29-43.
- Marzec J., Osiewalski J. [2012], *Dwuwymiarowy model typu ZIP-CP w łącznej analizie zmiennych licznikowych*, „Folia Oeconomica Cracoviensia”, nr 53, s. 5-20.
- 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, s. 375-402.
- 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 2012, s.147-154.
- Osiewalski J., Marzec J. [2004a], *Uogólnienie dychotomicznego modelu probitowego z wykorzystaniem skośnego rozkładu Studenta*, „Przegląd Statystyczny”, tom 51, s. 13-24.
- Osiewalski J., Marzec J. [2004b], *Model dwumianowy II rzędu i skośny rozkład Studenta w analizie ryzyka kredytowego*, „Folia Oeconomica Cracoviensia”, nr 45, 63-83.
- Pajor A. [2016], *Estimating the marginal likelihood using the arithmetic mean identity*, "Bayesian Analysis" (forthcoming), <http://dx.doi.org/10.1214/16-BA1001>.
- 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. [2012a], *Modelowanie wykorzystania metod płatności detalicznych na rynku polskim*, „Materiały i Studia NBP”, nr 265.
- Polasik M., Wisniewski T.P., Lightfoot G. [2012b], *Modelling Customers' Intentions to Use Contactless Cards*, “International Journal of Banking, Accounting and Finance”, vol. 4, no. 3, s. 203-231, <http://dx.doi.org/10.1504/IJBAAF.2012.051590>.
- Polasik M. [2015], *Stan i potencjał rozwoju sieci akceptacji kart płatniczych w Polsce*, „Acta Universitatis Nicolai Copernici, Ekonomia”, vol. 46, s. 23–58, http://dx.doi.org/10.12775/AUNC_ECON.2015.002.
- Winkelmann R. [2008], *Econometric Analysis of Count Data*, Springer-Verlag, Berlin Heidelberg.

KEY WORDS - SŁOWA KLUCZOWE

bivariate Poisson regression models, switching between non-degenerate and degenerate distributions, likelihood factorization, bank card and cash payments
dwuwymiarowe modele regresji Poissona, przełączanie między rozkładem niezdegenerowanym i zdegenerowanym, faktoryzacja funkcji wiarygodności, płatności kartą płatniczą i gotówką