PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 176
Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania 2011

Robert Kapłon
Politechnika Wrocławska

O DWÓCH SPOSOBACH UWZGLĘDNIENIA NIEJEDNORODNOŚCI OBSERWACJI W MODELU CZĘSTOŚCI ZAKUPÓW

Streszczenie: Dane dotyczące częstości zakupów są obecnie łatwo dostępne za sprawą systemów transakcyjnych, które je zbierają i przechowują. Dlatego potrzebne są odpowiednie narzędzia pozwalające na ich analizę. Często wykorzystywany model Poissona, ze względu na niejednorodność danych, jest nieodpowiedni. W tej sytuacji należy poszukiwać bardziej złożonych i zarazem bardziej wiarygodnych modeli. W pracy zaprezentowano dwa konkurencyjne modele pozwalające na uchwycenie niejednorodności: mieszanki rozkładów Poissona oraz mieszane modele Poissona. Wychodząc natomiast od przesłanek teoretycznych i empirycznych, wskazano na podobieństwa i różnice między nimi.

Słowa kluczowe: mieszanki rozkładów, mieszane modele, model Poissona.

1. Zarys problemu

Częstość zakupów należy utożsamiać z liczbą dokonanych zakupów, którą najczęściej rozpatruje się w pewnym przedziale czasowym. Przykładem takich danych może być liczba: użyć karty kredytowej, tankowań na stacji benzynowej, transakcji dokonanych w sklepie internetowym, zakupionych produktów określonego typu itd. Rejestracja takich zachowań konsumentów odbywa się najczęściej z wykorzystaniem tzw. systemów transakcyjnych. Dzięki nim zgromadzone dane o dokonanych zakupach są wiarygodne, a dostępność do nich jest niemal natychmiastowa. To sprawia, że są one cennym i zarazem łatwo dostępnym źródłem informacji, pod warunkiem wykorzystania odpowiednich technik analitycznych.

Punktem wyjścia w analizie takich danych dyskretnych jest najczęściej rozkład Poissona:

$$\Pr[X = x \mid \lambda] = \frac{\exp(-\lambda)}{x!}\lambda^x, \qquad x = 0, 1, 2, ...$$

w którym $E(X) = \lambda, \quad Var(X) = \lambda.$

Ze względu na szczególną własność, równość wariancji i wartości oczekiwanej, cechuje go ograniczona możliwość aplikacyjna. Jeśli analizowana populacja wykazuje duże zróżnicowanie, co jest zjawiskiem dość częstym (por. [Agresti 2002, s. 7]), wtedy szczególnego znaczenia nabierają koncepcje i podejścia pozwalające taką niejednorodność uwzględnić.

Dlatego celem artykułu jest przedstawienie różnych podejść do analizy niejednorodnych obserwacji związanych z częstością zakupów. Zawiera on, oprócz rozważań teoretycznych, materiał statystyczny pozwalający na wskazanie podobieństw i różnic między nimi.

2. Skończone mieszanki Poissona

Podstawowe założenie, jakie w tym podejściu się przyjmuje, opiera się na spostrzeżeniu, że badaną populację G można podzielić na rozłączne podpopulacje:

$$G = G_1 \cup G_2 \cup \ldots \cup G_S \wedge \underset{s \neq c}{\forall} G_s \cap G_c = \varnothing, \quad s, c = 1, \ldots, S. \tag{1}$$

Z tym podziałem związana jest niepewność, gdyż *a priori* nie jest znana liczba podpopulacji (klas). Nie można również wskazać klasy, z której obserwacja pochodzi. Dlatego wprowadza się zmienną ukrytą Ψ, której rola sprowadza się do podziału badanej populacji na wzajemnie rozłączne, bezpośrednio nieobserwowalne podpopulacje zgodnie z koncepcją ujętą we wzorze (1). Zmienna ta przyjmuje wartości $\psi_1, \psi_2, \ldots, \psi_S$ odpowiednio z prawdopodobieństwami

$$\pi(\psi_1), \pi(\psi_2), \ldots, \pi(\psi_S).$$

Przy tych założeniach prawdopodobieństwo zaobserwowania wartości x pod warunkiem, że pochodzi ona z klasy G_s wynosi:

$$\Pr(X = x \mid \Psi = \psi_s) = \frac{\exp(-\lambda_s)}{x!} \lambda_s^x, \quad x = 0, 1, 2... \tag{2}$$

natomiast prawdopodobieństwo bezwarunkowe ma postać:

$$\begin{aligned} \Pr(X = x) &= \sum_{s=1}^{S} \pi(\psi_s) \Pr(X = x \mid \Psi = \psi_s) \\ &= \sum_{s=1}^{S} \pi(\psi_s) \frac{\exp(-\lambda_s)}{x!} \lambda_s^x. \end{aligned} \tag{3}$$

W tym miejscu należy podkreślić, że przedstawiony model bazuje na koncepcji klas ukrytych, dość często wykorzystywanych na gruncie nauk społecznych, a w szczególności w psychologii i socjologii (zob. [Clogg 1995; Kapłon 2002; McCutcheon 1987]). Natomiast w literaturze traktującej o skończonych mieszankach

rozkładów bardziej rozpowszechniona jest koncepcja, w której *explicite* nie odwołuje się do zmiennej ukrytej (zob. [McLachlan, Peel 2000; Titterington i in. 1985]). Nie brakuje również opinii wyrażających pogląd, że skończone mieszkanki rozkładów są synonimem analizy klas ukrytych [Vermunt 2004]. Choć w pewnych sytuacjach można mieć wątpliwości, o czym pisze Kapłon [2002], to jednak w wypadku prezentowanego modelu należy się z tym poglądem zgodzić. Z tego też względu zasadne wydaje się mówienie o skończonych mieszankach rozkładów Poissona, pomimo wprowadzenia zmiennej ukrytej.

Obecność zmiennej bezpośrednio nieobserwowalnej pozwala w łatwy sposób obliczyć wartość oczekiwaną i wariancję zmiennej X. Wykorzystując wzory (2) i (3), otrzymujemy:

$$\mu = E[X] = \sum_{s=1}^{S} \pi(\psi_s) E[X \mid \Psi = \psi_s] = \sum_{s=1}^{S} \pi(\psi_s)\lambda_s,$$

$$\sigma^2 = Var(X) = E[Var(X \mid \Psi)] + Var(E[X \mid \Psi])$$

$$= \mu + \sum_{s=1}^{S} (\lambda_s - \mu)^2 \pi(\psi_s).$$

Obliczone momenty pokazują, że w odróżnieniu od rozkładu Poissona, wartość oczekiwana jest mniejsza od wariancji, przy założeniu, że w populacji można wyróżnić przynajmniej dwie klasy. Stopień niejednorodności, mierzony wartościami estymowanych parametrów λ ma decydujący wpływ na wariancję − im większe ich zróżnicowanie, tym większa wariancja.

Choć model jest dość elastyczny i pozwala na uchwycenie niejednorodnych populacji, to jednak należy pamiętać, że jego cechą szczególną jest multimodalność, będąca konsekwencją podziału populacji na podpopulacje.

3. Mieszane modele Poissona

W odróżnieniu od poprzedniego modelu, w mieszanym modelu Poissona parametr λ ma charakter losowy. Przy tym założeniu rozkład częstości zakupów można wyrazić w ogólnej postaci (por. [Johnson i in. 2005, s. 345]):

$$\Pr(X = x) = E_{\lambda}[\Pr(X = x \mid \lambda)]. \tag{4}$$

W zależności od przyjętego rozkładu parametru λ − dyskretnego lub ciągłego − prawdopodobieństwo (4) przyjmuje odpowiednio postać:

$$\Pr(X = x) = \sum_{i \geq 0} \Pr(X = x \mid \lambda_i) p(\lambda_i)$$

oraz

$$\Pr(X = x) = \int_{\lambda} \Pr(X = x \mid \lambda) f(\lambda) d\lambda.$$

Niezależnie od przyjętego rozkładu, wartość oczekiwaną i wariancję można wyrazić ogólnie:

$$E[X] = E[E[X \mid \lambda]] = E[\lambda],$$

$$Var(X) = E[Var[X \mid \lambda]] + Var[E[X \mid \lambda]] = E[\lambda] + Var(\lambda).$$

Okazuje się, podobnie jak w poprzednim modelu, że wariancja jest większa od wartość oczekiwanej. Co więcej, jej wartość jest zależna od przyjętego rozkładu dla parametru λ. Tak więc od jego elastyczności zależy stopień uchwycenia niejednorodności.

Wśród wielu propozycji dotyczących rozkładu parametru λ, najbardziej rozpowszechniony jest rozkład gamma, gdyż jego przyjęcie prowadzi do ujemnego rozkładu dwumianowego (*NBD*). Znany jest on jako model Ehrenberga i dość często wykorzystuje się go w analizie danych dotyczących częstości zakupów [Ehrenberg 2004]. Z kolei złożenie, że parametr w *NBD* ma rozkład Poissona, prowadzi do powstania rozkładu Poissona–Pascala, co można zapisać w następującej sekwencji [Johnson i in. 2005]:

$$\underbrace{Poisson(\lambda) \underset{\lambda}{\wedge} Gamma(Y, P)}_{Negative.Bionomial} \underset{Y/k}{\wedge} Poisson(\theta)$$

Nie zawsze przyjęcie rozkładu λ prowadzi do otrzymania znanego rozkładu. Co więcej, powstały rozkład może nie mieć jawnej postaci. Przykładem jest rozkład log-normalny, który zdaniem Winkelmanna [2008] jest bardziej elastyczny niż *NBD*, jednak mało popularny w zastosowaniach praktycznych.

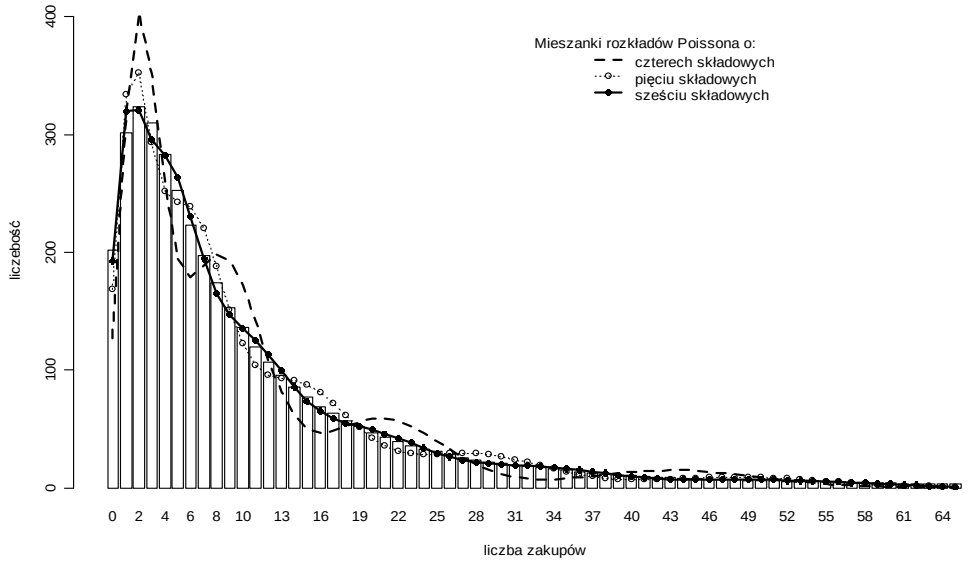
Z powyższych ustaleń wyłania się następujący wniosek. Otóż, w większości wypadków przyjęcie rozkładu dla parametru prowadzi do innego znanego rozkładu. Niejednorodność obserwacji można więc uchwycić poprzez zwiększoną elastyczność powstałego rozkładu (zazwyczaj pojawia się dodatkowy parametr). Ta elastyczność pozwala też uwzględnić tzw. długie ogony. Niestety, jeśli rozkład obserwacji wskazuje na multimodalność, wtedy żaden z bardzo wielu rozkładów mieszających (kilkanaście takich rozkładów prezentują Johnson, Kemp i Kotz [2005]) nie sprawi, że powstały ze złożenia rozkład nie będzie jednomodalny. Prawdziwy jest również wniosek ogólny: jeśli rozkład mieszający jest jednomodalny, to powstały rozkład również taki będzie. Dowiódł tego Holgate [1970, za: Gupta, Ong 2005].

4. Przykład

W kontekście powyższych rozważań wiadomo, że jeśli rozkład częstości zakupów jest przynajmniej dwumodalny, to mieszane modele Poissona nie są w stanie tego uchwycić. Z drugiej strony, jeśli niejednorodność ma odzwierciedlenie w wysokiej wariancji, bo występują tzw. długie ogony, wtedy przyjęcie odpowiedniego rozkładu mieszającego daje szansę na uwzględnienie tej niejednorodności w modelu. Rodzi się więc pytanie, jak w wypadku takiego typu danych zachowuje się model Poissona oparty na mieszankach rozkładów.

Aby na nie odpowiedzieć, wygenerowano dane z rozkładu Poissona przy założeniu, że parametr λ ma rozkład lognormalny. Wybrano ten rozkład ze względu na relatywnie długi ogon. Następnie oszacowano parametry mieszanego rozkładu Poissona, przyjmując od czterech do sześciu klas. Przyjęcie mniejszej liczby podpopulacji całkowicie dyskwalifikowało modele, biorąc pod uwagę na przykład statystykę chi-kwadrat. Obliczenia wykonano w środowisku **R**.

Analiza rys. 1 pokazuje, że model o 4 klasach niezbyt dokładnie odzwierciedla rozkład częstości. Widoczne są cztery wartości modalne, chociaż rozkład obserwacji wyraźnie wskazuje na jedną. Zwiększenie o dodatkową klasę poprawiło stopień dopasowania, jednak dalej widoczne są miejsca, w których występuje wyraźne niedoszacowanie lub przeszacowanie wartości. Dopiero model o 6 klasach/podpopulacjach dobrze opisuje rozkład częstości zakupów, co dodatkowo potwierdza wartość statystyki chi-kwadrat.



Rys. 1. Częstość zakupów podlegająca rozkładowi Poissona-log-normalnemu (słupki) oraz oszacowania na podstawie mieszanek rozkładów Poissona

Źródło: opracowanie własne.

Okazuje się więc, że model oparty na mieszankach rozkładów Poissona jest bardzo elastyczny i w zależności od liczby przyjętych klas można osiągnąć prawie dowolny stopień dopasowania do danych. Należy jednak pamiętać, że jego cechą szczególną jest multimodalność. Z tego względu, gdy występuje duży stopień niejednorodności danych, należy przyjąć relatywnie wiele klas, aby „pozbyć" się wartości modalnych. Wydaje się, że w takiej sytuacji bardziej naturalnym podejściem jest wykorzystanie mieszanych modeli Poissona.

5. Podsumowanie

Jeśli częstości zakupów wykazują się pewnym stopniem niejednorodności, wtedy można wykorzystać jedno z przedstawionych podejść. Należy jednak pamiętać, że mieszane modele Poissona mają zastosowanie tam, gdzie rozkład częstości jest jednomodalny. Wynika to z następującej własności: jeśli przyjęty rozkład parametru w modelu Poissona ma jedną modę, wtedy i rozkład złożony będzie miał jedną. Tę klasę modeli można szczególnie zarekomendować w wypadku rozkładów o długich ogonach. Jeśli oczekuje się, że istnieją przeciętne częstości zakupów, wokół których gromadzi się znaczna część obserwacji – co wskazywałoby na istnienie wielu wartości modalnych – wtedy mieszanki rozkładów Poissona są naturalnym wyborem. Jak pokazała analiza numeryczna, wspomniany model jest bardzo elastyczny w tym sensie, że przyjęcie dostatecznie dużej liczby podpopulacji pozwala na bardzo dobre dopasowanie do danych, których rozkład jest jednomodalny i cechuje się długim ogonem. Tym samym model ten radzi robie bardzo dobrze w sytuacjach, które wymagałyby zastosowania mieszanych modeli Poissona − odwrotne stwierdzenie, że te ostatnie mogą stać się substytutem dla mieszanek Poissona, nie jest prawdziwe.

W kontekście powyższych uwag pojawia się pytanie natury aplikacyjnej: czy zwiększanie liczby podpopulacji/klas ma sens praktyczny, czy tylko jest sztucznym zabiegiem, mającym na celu poprawę dopasowania modeli? Wydaje się, że dalsze badania, zmierzające do odpowiedzi na tak postawione pytanie, powinny być prowadzone.

Literatura

Agresti A., *Categorical Data Analysis*, Wiley-Interscience Publication, New Jersey 2002.

Clogg C.C., *Latent Class Models*, [w:] G. Arminger, C.C. Clogg, M.E. Sobel (red.), *Handbook of Statistical Modelling for Social and Behavioural Science*, Plenum, New York 1995, s. 311-359.

Ehrenberg A., *My Research in Marketing: How It Happened*, „Marketing Research" 2004, vol. 16, s. 36-41.

Gupta R.C., Ong S.H., *Analysis of long-tailed count data by Poisson mixtures*, „Communications in Statistics − Theory and Methods" 2005, no. 34, s. 557-573.

Holgate P., *The modality of some compound Poisson distributions*, „Biometrika" 1970, no. 57, s. 666-667.
Johnson N.L., Kemp A.W., Kotz S., *Univariate Discrete Distributions*, Hoboken, Wiley-Interscience, N.J. 2005.
Kapłon R., *Analiza danych dyskretnych za pomocą metody LCA*, [w:] K. Jajuga, M. Walesiak (red.), *Klasyfikacja i analiza danych – teoria i zastosowania*, Taksonomia nr 9, Prace Naukowe Akademii Ekonomicznej we Wrocławiu, Wrocław 2002.
McCutcheon A.L., *Latent Class Analysis*, Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-064, Thousand Oaks 1987.
McLachlan G.J., Peel. D., *Finite Mixture Models*, Wiley, New York 2000.
R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna 2009, URL: http://www.R-project.org.
Titterington D.M., Smith A.F.M., Markov U.E., *Statistical Analysis of Finite Mixture Distributions*, John Wiley & Son, New York 1985.
Vermunt J.K., *Mixture Model*, [w:] M. Lewis-Beck, A. Bryman, T.F. Liao (red.), *Encyclopedia of Research Methods for the Social Sciences*, Sage Publications, New Bury Park 2004, s. 653.
Winkelmann R., *Econometric Analysis of Count Data*, Springer, Berlin 2008.

TWO APPROACHES FOR MODELLING DATA HETEROGENEITY IN A PURCHASE FREQUENCY MODEL

Summary: Purchase frequency data are easily available with respect to the transactional systems that collect and store them. Thus, one needs suitable tools allowing for analyzing such data. Frequently used Poisson model on account of the data heterogeneity is unsuitable. This implies that the more complex and more reliable models are needed. In the paper we describe two competitive models allowing for capturing heterogeneity: finite mixture Poisson and mixed Poisson model. We also show similarities and differences based on theoretical as well as empirical background.