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

Ewa Witek
Uniwersytet Ekonomiczny w Katowicach

WYKORZYSTANIE MIESZANEK ROZKŁADÓW POISSONA DO OCENY LICZBY PRZYZNANYCH PATENTÓW W KRAJACH UE

Streszczenie: W artykule przedstawiono zastosowanie mieszanek warunkowych rozkładów Poissona w regresji. Mieszanki tych rozkładów stosowane są wówczas, gdy zbiór obserwacji charakteryzuje się nadmiernym rozproszeniem, będącym wynikiem na przykład pominięcia jednej z ważnych zmiennych objaśniających. Celem referatu jest zbadanie wpływu wydatków na badania i rozwój na liczbę przydzielonych patentów w krajach Unii Europejskiej.

Słowa kluczowe: model mieszanek, niejednorodność, funkcja wiążąca.

1. Wstęp

Mieszanki rozkładów warunkowych, w których wyróżnia się wpływ zmiennych objaśniających na zmienną objaśnianą i w których możliwe jest uwzględnienie zmiennych towarzyszących (wpływających na przynależność obiektów do klas), zwane są mieszankami modeli GLM (*generalized linear mixture models*) [Leisch 2004]. Gdy analizowany zbiór obserwacji jest zbiorem niejednorodnym, mieszanki rozkładów wykorzystywane są w analizie regresji w celu wyodrębnienie kilku podzbiorów oraz oszacowania parametrów modeli dla każdego z nich.

W artykule przedstawiono zastosowanie mieszanek rozkładów Poissona do oceny liczby przyznanych patentów w krajach UE. Na podstawie danych dla 27 krajów Europy zbudowano model mieszanek rozkładów Poissona, wykorzystując pakiet `flexmix` programu **R**. Na podstawie kryteriów informacyjnych (np. BIC) określono optymalną liczbę klas oraz zinterpretowano parametry oszacowanego modelu mieszanek.

2. Mieszanka rozkładów Poissona

Zakłada się, że każda składowa mieszanki jest charakteryzowana przez warunkowy rozkład prawdopodobieństwa, a związek pomiędzy zmienną zależną i zmiennymi objaśniającymi jest określony za pomocą złożonego modelu, czyli uogólnionego

modelu liniowego (GLM). Liczba podzbiorów, tj. rozkładów składowych mieszanki (*components of mixture model*), może być rozumiana jako liczba segmentów rynku, w których prawdopodobieństwa zakupu w zależności od ceny czy reklamy są wyraźnie różne dla klientów każdego z wyodrębnionych segmentów. Zależnie od skali pomiaru zmiennej objaśnianej wśród mieszanek wyróżnić można mieszanki modeli logitowych oraz mieszanki regresji Poissona. Ponieważ mieszanki rozkładów Poisssona zostaną wykorzystane w przykładzie empirycznym, poniżej przedstawiono ogólną postać mieszanek tych rozkładów:

$$f(y_i \left| \mathbf{x}_i^j, \mathbf{x}_i^{\varpi}, \boldsymbol{\Theta} \right.) = \sum_{s=1}^{u} \pi_s(\mathbf{x}_i^{\varpi}, \boldsymbol{\alpha}) Poisson(y_i \left| \lambda_{is} \right.), \tag{1}$$

$$Poisson(y_i \left| \lambda_{is} \right.) = \frac{\lambda_{is}^{y_i}}{y_i!} \exp(-\lambda_{is}), \tag{2}$$

gdzie: $Poisson(y_i \left| \lambda_{is} \right.)$ – funkcja *s*-tego rozkładu Poissona,

y_i – realizacja zmiennej zależnej Y dla *i*-tej obserwacji,

$\mathbf{x}_i^j$ – wektor realizacji zmiennych objaśniających, $\mathbf{x}_i^j = [x_{i1,\ldots,} x_{im_1}]$,

$\mathbf{x}_i^{\varpi}$ – wektor realizacji zmiennych towarzyszących[1], $\mathbf{x}_i^{\varpi} = [x_{i1,\ldots,} x_{im_2}]$,

$\boldsymbol{\Theta}_s$ – wektor parametrów rozkładu składowego P_s,

$\boldsymbol{\Theta}$ – wektor parametrów mieszanki rozkładów, $\boldsymbol{\Theta} = (\pi_s, \boldsymbol{\alpha}_s, \boldsymbol{\Theta}_s)$,

π_s – prawdopodobieństwo *a priori* – wartość prawdopodobieństwa, że dana obserwacja należy do podpopulacji

$$P_s \ \ (\pi_s(\mathbf{x}_i^{\varpi}, \boldsymbol{\alpha}) \geq 0 \wedge \sum_{s=1}^{u} \pi_s(\mathbf{x}_i^{\varpi}, \boldsymbol{\alpha}) = 1), \boldsymbol{\Theta}_{\mathbf{s}} \neq \boldsymbol{\Theta}_l \forall s \neq l.$$

W praktyce zastosowań zakłada się, że wartość przeciętna jest równa wariancji, a zdarzenia są wzajemnie niezależne. Założenie równości wartości przeciętnej i wariancji są często naruszone, np. zbiór cechuje się niejednorodnością, nadmiernym zróżnicowaniem, rozproszeniem (*overdispersion*).

W przypadku, gdy zbiór danych jest zbiorem niejednorodnym, jednym z rozwiązań jest wykorzystanie mieszanek modeli Poissona, w których zakłada się, że:

a) zbiór wszystkich obserwacji złożony jest z *s* klas (podzbiorów),

[1] Zmienne towarzyszące wraz ze zmiennymi $X_1, ..., X_m$ biorą udział w szacowaniu parametrów modelu mieszanek, na podstawie którego można będzie dokonać klasyfikacji nowych obiektów bez udziału zmiennych objaśniających. Zmienne towarzyszące wykorzystywane są często w badaniach marketingowych, ekonomicznych, psychologicznych, w których pozyskanie zmiennych objaśniających jest bardzo kosztowne.

b) każda obserwacja z prawdopodobieństwem π_{is} przypisana jest do nieznanej klasy, podzbioru P_s. W literaturze można również spotkać się z określeniem, że obserwacja znajduje się w stanie s $(s = 1,\ldots,u)$ [Wang i in. 1998],

c) $(\mathbf{x}_i, y_i)$ to zbiór obserwacji $(i = 1,\ldots,n)$, y_i jest realizacją zmiennej losowej Y, pochodzącą z rozkładu Poissona, zaobserwowaną w danym okresie czasu, a wektor $\mathbf{x}_i = (\mathbf{x}_i^j, \mathbf{x}_i^{\varpi})$ tworzą odpowiednio m_1-wymiarowe, m_2-wymiarowe realizacje zmiennej objaśniającej i towarzyszącej,

d) dla zmiennej objaśnianej Y, zależnej od wektora zmiennych objaśniających, istnieje nieznana zmienna losowa Π, określająca klasę dla obserwacji $(\mathbf{x}_i, y_i)$, (Y, Π) są parami niezależne,

e) Π ma rozkład dyskretny, a prawdopodobieństwo $P(\Pi = s) = \pi_{is}$, gdzie $\sum_{s=1}^{u} \pi_{is} = 1$ dla *i*-tej obserwacji określone jest wzorem:

$$\pi_{is} \equiv \pi_s(\mathbf{x}_i^{\varpi}, \boldsymbol{\alpha}) = \frac{\exp(\boldsymbol{\alpha}_s' \mathbf{x}_i^{\varpi})}{1 + \sum_{l=1}^{u-1} \exp(\boldsymbol{\alpha}_s' \mathbf{x}_i^{\varpi})}, \; s = 1,\ldots,u-1, \tag{3}$$

$$\pi_{iu} \equiv \pi_u(\mathbf{x}_i^{\varpi}, \boldsymbol{\alpha}) = 1 - \sum_{s=1}^{u-1} \pi_{is}, \tag{4}$$

gdzie $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_{u-1})'$ oraz $\boldsymbol{\alpha}_s = (\alpha_{s1},\ldots,\alpha_{sl_1})'$ dla $1 \le s \le u-1$ są nieznanymi parametrami mieszanki,

f) jeżeli $\Pi = s$, to zmienna losowa Y ma rozkład Poissona, który można zapisać jako:

$$f_s(y_i \mid \mathbf{x}_i^j, \boldsymbol{\beta}_s) = Poisson(y_i \mid \lambda_{is}) = \frac{\lambda_{is}^{y_i}}{y_i!} \exp^{-\lambda_{is}}, \tag{5}$$

a wartość przeciętna rozkładu λ_{is} jest zależna od zmiennych objaśniających $\mathbf{X}^j$, zależność tę wyraża funkcja wiążąca (*link function*) dana równaniem:

$$\lambda_{is} \equiv \lambda_s(\mathbf{x}_i^j, \boldsymbol{\beta}_s) = \exp(\boldsymbol{\beta}'_s \, \mathbf{x}_i^j), \tag{6}$$

gdzie $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_u)'$ oraz $\boldsymbol{\beta}_s = (\beta_{s1},\ldots,\beta_{sl_2})'$ dla $1 \le s \le u$ są nieznanymi parametrami mieszanki.

Parametry modeli mieszanek szacuje się najczęściej za pomocą algorytmu EM [Dempster i in. 1977], a wyboru modelu optymalnego dokonuje się na podstawie kryteriów informacyjnych BIC, AIC i ICL [Frühwirth-Schnatter 2006].

3. Analiza empiryczna

Nad problemem zależności pomiędzy wydatkami na badania i rozwój (R & D) a liczbą przyznanych patentów zastanawiano się już wielokrotnie (m.in. [Cohen, Levin 1989; Mairesse, Sassenou 1991]). Wang, Cockuburn i Puterman [1996] zaproponowali, by do oceny parametrów modeli wyrażających zależność pomiędzy liczbą patentów a wydatkami na R & D, wykorzystać mieszanki rozkładów.

Celem przykładu jest pokazanie, jakie są relacje między wydatkami na badania i rozwój a liczbą udzielonych patentów w krajach Unii Europejskiej.

Zależność pomiędzy liczbą patentów a wydatkami na R&D można zapisać za pomocą „funkcji produkcji patentów" o postaci: $E(Y) = \exp(\boldsymbol{\beta}'\mathbf{X})$, gdzie Y oznacza liczbę przyznanych patentów, a $\mathbf{X}$ jest wektorem zmiennych objaśniających.

Parametry funkcji produkcji patentów mają ciekawą interpretację ekonomiczną, informują bowiem o efektach skali w zależności od nakładów na badania i rozwój. Trudność badania takiej zależności polega na tym, że dane bardzo często są nadmiernie rozproszone. Z problemem tym próbowano sobie poradzić, szacując parametry modelu Poissona i jego różnych modyfikacji [Hall i in. 1986; Gourieroux i in. 1984]. Za główną przyczynę tego problemu uznano brak „stabilności" (*instability problem*) między wydatkami na badania i rozwój a liczbą przyznanych patentów w pewnym przedziale czasowym.

W podanym dalej przykładzie założono, że zarówno współczynniki przy zmiennych objaśniających, jak i wyraz wolny mogą przyjmować wartości różne dla każdego z badanych krajów UE. Nakłada się tu jednak pewne ograniczenia, tj. każdy z krajów może się znaleźć w s ($s = 1, \ldots, u$) różnych klasach lub „stanach", odpowiadających różnym stopniom produktywności, np. „duży", „średni", „niski". Szacowane więc są „funkcje produkcji patentów" nie dla każdego krajów z osobna, lecz dla klas, do których należą. Przyjęcie tego rodzaju założeń można uzasadnić tym, że wszystkie kraje mają dostęp do takich samych technologii, ale różny, nieobserwowalny, potencjał innowacyjny (np. związany ze strukturą organizacji). Alternatywnie można by założyć, że każdy kraj ma taki sam potencjał innowacyjny, możliwości zaś technologiczne tych krajów są różne (niektóre z nich pracują w obszarach wysokiej technologii, a inne nie).

W badaniu analizowano najbardziej aktualne dane (z roku 2003) dotyczące zależności pomiędzy wydatkami na badanie i rozwój a liczbą patentów przyznanych 27 krajom Unii Europejskiej. Dane pochodzą z bazy Eurostatu [http://epp.eurostat.ec.europa.eu/].

Wykorzystano następujące zmienne: *Patenty* – liczba przyznanych patentów w danym kraju; log(*R & D*) – logarytm wydatków na badania i rozwój wyrażony w milionach euro.

Szacowano parametry modelu mieszanek warunkowych rozkładów Poissona. Dla każdej składowej takiego modelu mieszanek szacowano więc parametry funkcji regresji Poissona, które są zarazem funkcjami elastyczności zdolności innowacyjnej krajów, w zależności od wydatków na badania i rozwój. Jeżeli współczynniki elastyczności przyjmują wartości większe niż jeden, wtedy przyrost wydatków na badania i rozwój powoduje bardziej niż proporcjonalny wzrost „produkcji" patentów (rosnące efekty skali).

Jednorodność zbioru obserwacji sprawdzono za pomocą statystyk Deana [1992, s. 451-457]. Uzyskane wyniki potwierdziły, że hipotezę zerową o jednorodności zbioru obserwacji należy odrzucić na każdym poziomie ufności[2].

W analizie uwzględniono także zmienną towarzyszącą (*PZ*) (*concomitant variable*), tj. procent naukowców oraz personelu działów R&D w stosunku do siły roboczej i zatrudnienia ogółem w danym kraju.

W badaniu przyjęto następujące założenia: liczba patentów uzyskanych przez *i*-ty kraj jest funkcją zależną od wektora:

$$\mathbf{x}_i = (\mathbf{x}_i^j, \mathbf{x}_i^{\varpi}),\ \mathbf{x}_i^j = (1, \log(R \,\&\, D)),\ \mathbf{x}_i^{\varpi} = (1, PZ),$$

– liczba uzyskanych patentów *i*-tego kraju jest niezależna od liczby uzyskanych patentów krajów pozostałych,
– zmienna zależna (liczba uzyskanych patentów *i*-tego kraju $(1 \le i \le 27)$) ma rozkład Poissona o wartości przeciętnej i prawdopodobieństwie *a priori*, określonych za pomocą wzorów (3)-(6).

Optymalną liczbę rozkładów składowych dla zbioru „patenty" wybrano za pomocą kryteriów informacyjnych BIC, AIC oraz ICL. Dla każdego z kryterium minimalną wartość uzyskano dla liczby klas równej trzy. Parametry oszacowanego modelu mieszanek funkcji regresji Poissona przedstawiono w tab. 1.

Tabela 1. Wyniki podziału dla trzech klas

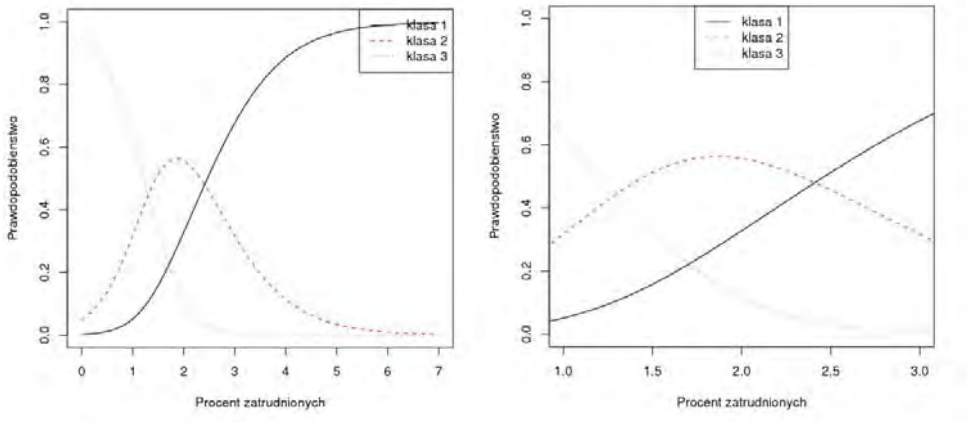
Klasa	π_s	Liczebność	λ_s
I	0,156	4	$\lambda_1 = \exp(-3,21 + 1,12\log(R \,\&\, D))$
II	0,376	10	$\lambda_2 = \exp(-5,12 + 1,25\log(R \,\&\, D))$
III	0,468	13	$\lambda_3 = \exp(-3,07 + 1,07\log(R \,\&\, D))$

Źródło: obliczenia własne.

[2] Dla każdej z trzech statystyk Deana *p-value* < 2,2e-16.

Obserwacje zostały przypisane do klas o najwyższym prawdopodobieństwie *a posteriori*. Do klasy pierwszej przypisano: Niemcy, Luksemburg, Holandię i Finlandię. Do drugiej: Czechy, Danię, Estonię, Grecję, Hiszpanię, Francję, Polskę, Portugalię, Szwecję i Wielką Brytanię. Z kolei do klasy trzeciej pozostałe kraje UE, tj. Belgię, Bułgarię, Irlandię, Cypr, Litwę, Łotwę, Węgry, Maltę, Austrię, Rumunię, Słowację, Słowenię.

Wpływ zmiennej towarzyszącej na prawdopodobieństwo przynależności do klas zilustrowano na rys. 1.



a) dla pełnego zakresu PZ b) dla PZ z przedziału (1; 3)

Rys. 1. Prawdopodobieństwo przynależności firm krajów UE do trzech klas w zależności od zmiennej towarzyszącej

Źródło: obliczenia własne.

Na rysunkach 1a i b można zauważyć, że gdy procent osób pracujących nad postępem technicznym firm danego kraju jest niski, dominują firmy krajów klasy trzeciej, co oznacza, że gdy procent osób zaangażowanych w badania jest niski, prawdopodobieństwo przynależności do klasy trzeciej jest największe. Wraz ze wzrostem zatrudnienia prawdopodobieństwo przynależności firm do klasy trzeciej spada prawie do zera. Gdy procent osób zatrudnionych zawiera się w przedziale $1{,}4 \leq PZ \leq 2{,}4$, prawdopodobieństwo przynależności do klasy drugiej jest największe. Należy stwierdzić, że wraz ze wzrostem liczby osób pracujących nad postępem technicznym wzrasta prawdopodobieństwo przynależności firm krajów UE do klasy pierwszej, gdy $PZ > 2{,}4$, prawdopodobieństwo przynależności do tej klasy jest największe (kraje należące do tej klasy to: Niemcy, Luksemburg, Holandia, Finlandia).

W celu określenia jakości dopasowania model mieszanek porównany został z klasycznym modelem GLM (funkcją regresji Poissona). Model ten ma następującą postać:

$$\lambda = \exp(-4{,}76 + 1{,}24 \log(R \& D)),$$

a wartości kryteriów AIC, BIC oraz ICL są znacznie wyższe niż w przypadku modelu mieszanek. Obliczono także reszty Pearsona. Na podstawie rysunków reszt, których nie zamieszczono ze względu na ograniczenia objętościowe pracy, widać wyraźnie, że jakość dopasowania dla modelu mieszanek jest znacznie wyższa niż w przypadku pojedynczego modelu GLM.

Na podstawie oszacowanego modelu mieszanek wyznaczono również oczekiwaną liczbę patentów dla dwóch krajów, w których wydatki na badania i rozwój wynoszą odpowiednio: 292 (mln euro) i 1296 (mln euro). Jako hipotetyczne wartości wydatków przyjęto wartości wydatków na badania i rozwój dwóch krajów kandydujących: Chorwacji oraz Turcji. Z prawdopodobieństwem 0,6 Chorwacja przydzielona została do klasy 3, a oczekiwana wartość patentów, oszacowana na podstawie modelu mieszanek, wynosi 21. Turcja z kolei z prawdopodobieństwem równym 1 przydzielona została do klasy 3, a oczekiwaną wartość patentów oszacowano na poziomie 42.

Ponieważ rzeczywista liczba patentów w badanym roku jest znana (dostępna w bazie Eurostatu), policzono także błędy prognozy. W przypadku Chorwacji obserwuje się niedoszacowanie o 7, natomiast w przypadku Turcji − przeszacowanie o 21 patentów. Oczekiwana liczba patentów wyznaczona na podstawie funkcji regresji Poissona jest znacznie bardziej niedoszacowana niż dla modelu mieszanek, tj. oczekiwana liczba patentów wynosi: 2 dla Chorwacji oraz 4 dla Turcji (niedoszacowanie odpowiednio o 26 i 17)[3]. Należy jednak pamiętać, że kraje te nie należą jeszcze do Unii Europejskiej. Muszą spełnić szereg wymagań, by sprostać oczekiwaniom oraz kryteriom stawianym krajom wspólnoty. Dowodem tego jest również liczba patentów przyznawana w tych krajach − nadal niższa (przypadek Turcji) niż wartość oczekiwana, oszacowana na podstawie modelu mieszanek.

4. Podsumowanie

Na podstawie oszacowanego modelu mieszanek zaobserwowano rosnące efekty skali (współczynnik elastyczności dla zmiennej $\log(R \& D)$ przyjmuje wartości większe od jeden) dla trzech oszacowanych „funkcji produkcji patentów". Najwyższe efekty skali zaobserwowano w przypadku krajów klasy drugiej (Czechy, Dania, Estonia, Grecja, Hiszpania, Francja, Polska, Portugalia, Szwecja, Wielka Brytania). Kraje klasy trzeciej (Bułgaria, Belgia, Irlandia, Włochy, Cypr, Litwa, Łotwa, Węgry, Malta, Austria, Rumunia, Słowacja, Słowenia) charakteryzują się najniższymi

[3] Rzeczywista liczba patentów w roku 2003 wynosi: 28 dla Chorwacji oraz 21 dla Turcji.

efektami skali. Można by nawet nazwać je krajami o stałych efektach skali (współczynnik dla zmiennej $\log(R\&D)$ wynosi 1,07).

Największy wpływ na produkcję patentów w krajach UE ma stopa zwrotu z wydatków na badania i rozwój w krajach klasy drugiej. Nie bez znaczenia pozostaje również liczba osób pracujących nad postępem technicznym. Zmienna ta (towarzysząca) ma zdecydowanie największy wpływ na przynależność krajów do klasy pierwszej (Niemcy, Luksemburg, Holandia, Finlandia).

Literatura

Dean D.B., *Testing for overdispersion in Poisson and binomial regression model*, „Journal of the American Statistical Association" 1992, 87, s. 451-457.

Dempster A.P., Laird N.M., Rubin D.B., *Maximum likelihood for incomplete data via the EM algorithm (with discussion)*, „Journal of the Royal Statistical Society" 1977, B 39, s. 1-38.

Cohen W., Levin R., *Empirical Studies of Innovation and Market Structure*, [w:] R. Schmalensee, R. Willig (red.), *Handbook of Industrial Organization*, North-Holland, Amsterdam 1989, s. 1059-1107.

Frühwirth-Schnatter S., *Finite Mixture and Markov Switching Models*, Springer Series in Statistics, Hardcover, Berlin 2006.

Gourieroux C., Monfort A., Trognon A., *Pseudo maximum likelihood methods: Applications to Poisson models*, „Econometrica" 1984, 52, s. 701-720.

Hall B.H., Griliches Z., Hausman J.A., *Patents and R&D: Is there a lag*, „International Economic Review" 1986, 27, s. 265-283.

Leisch R., *FlexMix*: *A general framework for finite mixture models and latent class regression in R*, „Journal of Statistical Software" 2004, 11(8), s. 1-18.

Mairesse J., Sassenou M., *R & D and productivity*: *A survey of econometric studies at the firm level*, „STI Review" 1991, 8, s. 9-43.

Wang P., Puterman M.L., Cockburn I., Le N., *Mixed Poisson regression models with covariate dependent rates*, „Biometrics" 1996, 52, s. 381-400.

Wang P., Cockburn I.M., Puterman M.L., *Analysis of patent data – a mixed-Poisson-regression-model approach*, „Journal of Business & Economic Statistics" 1998, 16 (1), s. 27-41.

Źródło internetowe

http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/.

THE USE OF POISSON MIXTURE MODELS IN THE EVALUATION OF THE NUMBER OF PATENTS GRANTED IN THE EUROPEAN UNION COUNTRIES

Summary: The paper focuses on Poisson mixture models and their application in the regression. These models are often used to capture overdispersion in the data which can occur for example if important covariates are omitted in the regression. It is then assumed that the influence of these covariates can be captured by allowing a random distribution for the intercept. The goal of this paper is to analyze the relationship among patents and research and development spending in the European Union countries.