# EKONOMETRIA

26

## Zastosowanie matematyki w ekonomii

Redaktor naukowy **Janusz Łyko**

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2009

## Spis treści

## Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 76

Ekonometria 26 2009

**Kamil Fijorek**
Uniwersytet Ekonomiczny w Krakowie

# MODEL REGRESJI DLA CECHY PRZYJMUJĄCEJ WARTOŚCI Z PRZEDZIAŁU (0, 1) – UJĘCIE BAYESOWSKIE

**Streszczenie:** W artykule przedstawiono model regresji dla cechy, która przyjmuje wartości z obustronnie otwartego przedziału (0,1). Krótko omówiono wady powszechnie stosowanych metod modelowania tego typu danych. W tym kontekście zaprezentowano zreparametryzowany rozkład beta, a następnie na jego podstawie skonstruowano model regresji. W ramach ujęcia bayesowskiego przedstawiono estymację parametrów modelu, metody określania dobroci dopasowania oraz interpretacji parametrów modelu. W dalszej części dokonano bayesowskiego porównania modeli, zakładając, że rozkład zmiennej zależnej jest rozkładem beta, simplex lub normalnym. Opisaną metodologię zilustrowano przykładem.

**Słowa kluczowe:** beta regresja, ograniczona zmienna losowa, wnioskowanie bayerowskie.

## 1. Wstęp

Ogólnym celem przeprowadzania analizy regresji jest próba ilościowego ujęcia związku pomiędzy (najczęściej jedną) zmienną zależną (oznaczaną dalej symbolem $y$) a zmiennymi niezależnymi. W praktyce powszechnie stosowane są modele regresji dla ciągłej (nieograniczonej), licznikowej lub binarnej zmiennej zależnej. Jednakże modele regresji dla zmiennej, która przyjmuje wartości z przedziału (0,1), nie są powszechnie znane, co oznacza, że nie są powszechnie stosowane. Arbitralne założenie mówiące o tym, że zmienna zależna $y \in (0,1)$, nie jest szczególnie ograniczające, gdyż dla $y \in (a, b)$ (końce przedziału są znanymi stałymi) możliwe jest przekształcenie $(y-a)/(b-a) \in (0,1)$.

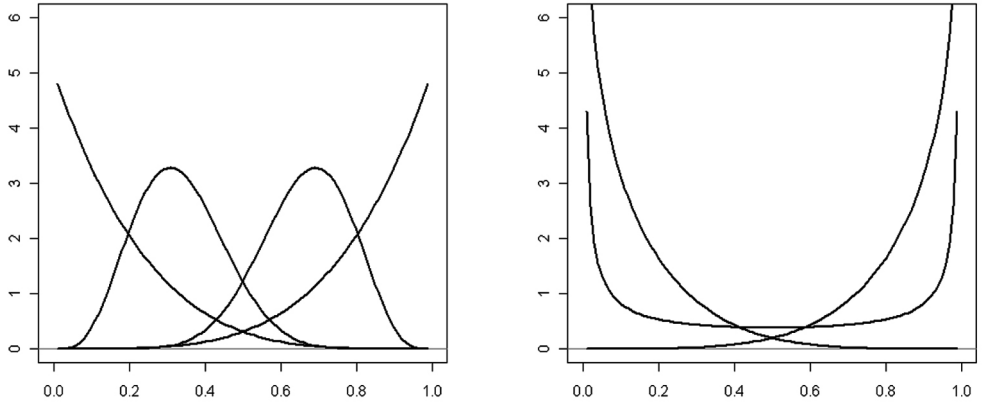
Kieschnick [2003] przeprowadził przegląd literatury, aby określić najpopularniejsze metody analizy rozważanego w artykule typu danych. Na pierwszym miejscu znalazła się (co nie jest szczególnym zaskoczeniem) klasyczna normalna regresja liniowa. Jednakże, ze względu na fakt, że zmienna zależna przyjmuje wartości z przedziału (0,1), założenie o normalności rozkładu nie może być spełnione. Ponadto wariancja ograniczonej zmiennej losowej jest funkcją wartości oczekiwanej, powodując, że założenie o stałej wariancji składnika losowego nie jest spełnione. Co więcej, zastosowanie tego podejścia może powodować generowanie przez mo-

del predykcji spoza przedziału określoności zmiennej zależnej. Drugim często spotykanym postępowaniem jest transformacja logitowa zmiennej zależnej (surowych danych). Następnie dla tak przekształconych danych wykonywana jest klasyczna regresja. Paolino [2001] w swoich badaniach symulacyjnych wykazał, że transformacja logitowa nie zawsze jest lepszym wyborem w porównaniu z klasyczną regresją liniową, gdyż m.in. niedoszacowuje błędów średnich szacunku. Problemem również jest to, że transformacja logitowa nie stabilizuje wariancji zmiennej zależnej. Inną metodą, już nie tak często stosowaną jak dwie poprzednie, jest wykorzystanie modelu tobitowego. To podejście również cierpi z powodu pewnych nieścisłości, gdyż przyczyną braku danych spoza przedziału (0,1) nie jest cenzorowanie (lub ucięcie), ale fakt, że takie wartości nie mogą wystąpić.

Naturalnym rozwiązaniem wspomnianych powyżej problemów związanych z modelowaniem wartości z przedziału (0,1) wydaje się bezpośrednie przyjęcie rozkładu prawdopodobieństwa, który będzie respektował ograniczenie zmiennej zależnej.

## 2. Rozkłady prawdopodobieństwa dla cechy o wartościach z przedziału (0, 1)

W niniejszym artykule założono, że zmienna zależna przyjmuje wartości z obustronnie otwartego przedziału (0,1). W przypadku, gdy przedział ten jest obustronnie (lub jednostronnie) domknięty, opisane metody nie znajdują bezpośredniego zastosowania. Pewne podstawy teoretyczne w celu uogólnienia metod na dyskretno-ciągły rozkład zmiennej zależnej poczynili autorzy prac [Lesaffre, Rizopoulos, Tsonaka 2004; Ospina, Ferrari 2008].




**Rys. 1.** Funkcja gęstości rozkładu beta w zależności od wartości parametrów kształtu

Źródło: opracowanie własne.

Najbardziej znanym rozkładem prawdopodobieństwa zdefiniowanym na przedziale (0,1) jest dwuparametrowy rozkład beta. Rozkład beta jest bardzo elastyczny. W zależności od wartości parametrów funkcja gęstości może być symetryczna, asymetryczna, *J*-kształtna, *L*-kształtna lub *U*-kształtna. Na rysunku 1 przedstawiono kilka przykładów funkcji gęstości rozkładu beta.

Innym proponowanym w literaturze rozkładem prawdopodobieństwa zdefiniowanym na przedziale (0,1) jest dwuparametrowy rozkład simplex [Barndorff-Nielsen 1991; Kieschnick 2003; Qiu, Song, Tan 2008]. Pomimo rozbudowanej bazy teoretycznej istniejącej dla tego rozkładu, jak wynika z badań symulacyjnych przeprowadzonych przez autora niniejszego opracowania, rozkład simplex jest mało elastyczny, tzn. funkcja gęstości może zmieniać kształt w ograniczonym zakresie. Z tego powodu w dalszej części pracy uwaga zostanie skupiona na modelu regresji, w którym warunkowy rozkład zmiennej zależnej to rozkład beta.

Funkcja gęstości rozkładu beta w standardowej parametryzacji ma postać:

$$f(y|p,q)=\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}y^{p-1}(1-y)^{q-1};\ 0<y<1;\ p>0,\ q>0, \tag{1}$$

gdzie $\Gamma(\bullet)$ oznacza funkcję gamma, natomiast $p$ oraz $q$ są parametrami kształtu. Wartość oczekiwana wynosi $E(y)=\frac{p}{p+q}$, natomiast wariancja $Var(y)=\frac{pq}{(p+q)^2(p+q+1)}$. W przypadku, gdy oba parametry kształtu są większe od jedności, rozkład beta ma wartość modalną. W przypadku, gdy oba parametry są równe 1, rozkład beta redukuje się do rozkładu jednostajnego.

Rozkład beta w standardowej parametryzacji nie jest dogodny do skonstruowania na jego podstawie modelu regresji. W tym kontekście Ferrari i Cribari-Neto [2004] zaproponowali zreparametryzowany rozkład beta. Wyszli oni z założenia, że typowe dla analizy regresji jest modelowanie parametru rozkładu prawdopodobieństwa odpowiedzialnego za wartość oczekiwaną. Przyjmując następującą parametryzację $\mu=\frac{p}{p+q};\ \phi=p+q;\ p=\mu\phi;\ q=(1-\mu)\phi;\ 0<\mu<1;\ \phi>0,$ uzyskano zmodyfikowaną wersję rozkładu beta, której funkcja gęstości ma następującą postać:

$$f(y|\mu,\phi)=\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}. \tag{2}$$

W tym przypadku wartość oczekiwana ma postać $E(y)=\mu,$ natomiast wariancja $Var(y)=\frac{V(\mu)}{1+\phi},$ gdzie $V(\mu)=\mu(1-\mu).$ Parametr $\phi$ może być interpretowany

jako parametr precyzji, gdyż dla ustalonego $\mu$ zwiększenie wartości $\phi$ powoduje zmniejszenie wariancji $y$.

## 3. Model regresji dla cechy o wartościach z przedziału (0,1)

Niech będzie danych $n$ niezależnych obserwacji $(y_i), i = 1, ..., n$ takich, że rozkład $y_i$ jest postaci $y_i \,|\, \mu_i, \phi \sim Beta\big(\mu_i\phi, \phi(1-\mu_i)\big)$. Model regresji jest uzyskany przez założenie, że wartość oczekiwana $y_i$ może być zapisana jako pewna monotoniczna transformacja liniowej kombinacji $k$ zmiennych niezależnych $x_i = (x_{i1}, ..., x_{ik})$:

$$g(\mu_i) = \sum_{j=1}^{k} x_{ij}\beta_j = \eta_i; \;\; \beta = (\beta_1, ..., \beta_k), \; \beta \in \mathbb{R}^k. \tag{3}$$

Biorąc pod uwagę, że zmienna zależna przyjmuje wartość z przedziału (0,1), należy rozważyć tylko takie transformacje liniowej kombinacji zmiennych niezależnych $g(\bullet)$, które przyjmują wartości z przedziału (0,1). Najprostszym wyborem jest przekształcenie logitowe, tj. $g(\mu_i) = \ln\left(\dfrac{\mu_i}{1-\mu_i}\right)$. W niektórych przypadkach preferowane jest jednak przyjęcie innej transformacji. Na przykład gdy prawdopodobne jest wystąpienie obserwacji nietypowych jako funkcję transformującą można wykorzystać dystrybuantę rozkładu *t*-Studenta o małej liczbie stopni swobody. Istniejące badania symulacyjne wskazują, że w typowych sytuacjach nie ma dużej korzyści ze stosowania innej niż logitowa transformacji [Kieschnick 2003].

Nic nie stoi na przeszkodzie, aby oprócz modelowania wartości oczekiwanej zmiennej zależnej również modelować parametr precyzji $\phi$ jako funkcję zmiennych niezależnych. Jednakże w tej pracy $\phi$ jest traktowane jako parametr zakłócający, niebędący przedmiotem bezpośredniego zainteresowania.

Po uwzględnieniu wszystkich przyjętych założeń możliwe jest wyznaczenie funkcji wiarygodności, a konkretnie jej logarytmu: $\ell(\beta, \phi) = \sum_{i=1}^{n} \ell_i(\mu_i, \phi)$, gdzie $\ell_i(\mu_i, \phi) = \ln\Gamma(\phi) - \ln\Gamma(\mu_i\phi) - \ln\Gamma\big((1-\mu_i)\phi\big) + (\mu_i\phi - 1)\ln y_i + \big[(1-\mu_i)\phi - 1\big]$ $\ln(1-y_i)$ oraz $\mu_i = \left(1 + e^{-x_i'\beta}\right)^{-1}$. W badaniach symulacyjnych wykazano, że numeryczna maksymalizacja logarytmu funkcji wiarygodności nie nastręcza szczególnych trudności [Smithson, Verkuilen 2005].

## 4. Bayesowska estymacja modelu regresji dla cechy o wartościach z przedziału (0, 1)

Buckley [2002] oraz Branscum, Johnson, Thurmond [2007] jako pierwsi podjęli się bayesowskiej estymacji modelu regresji dla cechy o wartościach z przedziału (0,1). Obaj autorzy założyli dla zmiennej zależnej rozkład beta oraz wykonali obliczenia w programie WinBUGS (*Bayesian Inference Using Gibbs Sampling*). Zastosowanie gotowego środowiska obliczeniowego, jakim jest WinBUGS, przyspiesza proces budowania modelu, aczkolwiek nie pozwala wyjść poza możliwości przewidziane przez autora oprogramowania. Oznacza to niezmiernie utrudnione wykorzystanie rozkładu zmiennej zależnej innego niż rozkład beta, a tym samym praktycznie wykluczona zostaje możliwość porównywania konkurencyjnych modeli. Ponadto, biorąc pod uwagę znacznie ograniczony zakres aspektów wnioskowania bayesowskiego poruszonych przez wspomnianych autorów, celowe wydają się dalsze badania.

Zastosowanie podejścia bayesowskiego w estymacji omawianego modelu regresji pozwala uwzględnić wstępną wiedzę badacza w postaci nałożonego na parametry modelu rozkładu *a priori* oraz umożliwia bardziej intuicyjną (w porównaniu z wnioskowaniem klasycznym) interpretację przedziałów ufności. Zdaniem autora są to ważniejsze (choć nie jedyne) zalety wnioskowania bayesowskiego. Znaczną wadą jest natomiast konieczność przeprowadzenia względnie skomplikowanych i często czasochłonnych obliczeń.

Wnioskowanie bayesowskie sprowadza się (w zasadzie) do wyznaczenia rozkładu warunkowego parametrów przy ustalonych obserwacjach, nazywanego rozkładem *a posteriori* [Osiewalski 2001, s. 16-17]. Funkcję gęstości rozkładu *a posteriori* parametrów uzyskuje się na podstawie wzoru Bayesa:

$$p(\beta,\phi|y)=\frac{L(\mu(\beta),\phi)\,p(\beta,\phi)}{\int L(\mu(\beta),\phi)\,p(\beta,\phi)\,d\beta d\phi}\propto\prod_{i=1}^{n}L_i(\mu_i(\beta),\phi)\;p(\beta,\phi), \qquad (4)$$

gdzie $f(y|\mu(\beta),\phi)=L(\mu(\beta),\phi|y)=\prod_{i=1}^{n}L_i(\mu_i(\beta),\phi|y)$ to funkcja wiarygodności dla $n$ niezależnych obserwacji, a $p(\beta,\phi)$ to rozkład *a priori* parametrów. W empirycznej części opracowania dla wszystkich parametrów przyjęto niewłaściwe rozkłady *a priori*.

Łączny rozkład parametrów (jak również ich rozkłady warunkowe) nie przyjmuje znanej postaci. Wyklucza to bezpośrednie metody symulacji z rozkładu *a posteriori* oraz próbkowanie Gibbsa. W tej sytuacji wykorzystano uniwersalny algorytm Metropolisa-Hastingsa z błądzeniem przypadkowym [Lynch 2007, s. 108-115] w celu wygenerowania próby z rozkładu *a posteriori* (wykonywano 100 000 losowań, pierwsze 10 000 uznawano za losowania spalone). Ponadto, w celu zbadania zbieżności do rozkładu *a posteriori*, algorytm Metropolisa-

-Hastingsa rozpoczynano z różnych punktów startowych oraz obserwowano, czy zbiega on do tego samego obszaru przestrzeni parametrów.

Standardową metodą analizy dopasowania modelu do danych jest wyznaczenie funkcji gęstości rozkładu predyktywnego (rozkładu przyszłych obserwacji) dla każdej z $n$ oryginalnych obserwacji. W przypadku dobrego dopasowania danych do modelu, tzn. gdy model adekwatnie opisuje proces generujący dane, przyszłe obserwacje powinny być podobne do rzeczywiście zaobserwowanych. Rozkład predyktywny uzyskuje się z następującego wyrażenia:

$$p\left(y^p \mid y\right) = \int p\left(y^p \mid \beta, \phi\right) L\left(\mu(\beta), \phi\right) p(\beta, \phi)\, d\beta d\phi. \tag{5}$$

Graficzna inspekcja dopasowania modelu do danych polega na naniesieniu na wykres funkcji gęstości rozkładu predyktywnego rzeczywistej realizacji zmiennej zależnej. Jeżeli obserwacja znajduje się w centrum rozkładu predyktywnego, można stwierdzić dobre dopasowanie, w przeciwnym razie, gdy obserwacja znajduje się w ogonach rozkładu, można mówić o złym dopasowaniu [Lynch 2007, s. 155--156]. Omówiona technika jest szczególnie przydatna, gdy liczba zmiennych niezależnych jest większa od 1.

Po określeniu dobroci dopasowania należy przejść do interpretacji kluczowych parametrów modelu ($\beta$). Ze względu na fakt, że wartość oczekiwana rozkładu zmiennej zależnej jest nieliniową funkcją zmiennych niezależnych, ich bezpośrednia interpretacja jest utrudniona. Aby ułatwić interpretację, wyznacza się efekty krańcowe dla poszczególnych zmiennych niezależnych, przyjmując, że pozostałe zmienne znajdują się na przeciętnym poziomie. Efekt krańcowy dla $j$-tej zmiennej zależnej (w przypadku transformacji logitowej) wyraża się następującym wzorem:

$$\frac{\partial g(x)}{\partial x_j} = \frac{\beta_i \exp\left(-x'\beta\right)}{\left[1 + \exp\left(-x'\beta\right)\right]^2}. \tag{6}$$

Na gruncie wnioskowania bayesowskiego możliwe jest bezpośrednie porównywanie konkurujących ze sobą modeli w celu określenia „najlepszego" modelu. Bayesowska idea porównywania modeli sprowadza się do wyznaczenia brzegowej gęstości wektora obserwacji przy założeniu danego modelu $p\left(y \middle| M_g\right) = \int L\left(\mu(\beta), \phi \middle| M_g\right) p\left(\beta, \phi \middle| M_g\right) d\beta d\phi$, gdzie $M_g$ oznacza $g$-ty model. Iloraz gęstości brzegowych dla dwóch konkurujących modeli nazywany jest czynnikiem Bayesa (BF – *Bayes Factor*). Wartość czynnika Bayesa większa od 1 przemawia na korzyść pierwszego modelu. W praktyce wartości większe od 3 uznaje się za znaczące. Na podstawie opisanej powyżej metodologii w dalszej części pracy zostaną porównane modele regresji zakładające, że rozkład zmiennej zależnej jest rozkładem beta, simplex lub normalnym.

Obliczenie gęstości brzegowej wektora obserwacji nie jest zadaniem prostym. W rozważanym w części empirycznej przypadku rozmiar przestrzeni parametrów nie jest duży, dlatego też możliwe było wyznaczenie prawdopodobieństw brzegowych za pomocą próbkowania z funkcją ważności – $q(\bullet)$. Zadanie to sprowadza się do zastosowania poniższych formuł:

$$p\left(y\middle|M_g\right)=\int\frac{L\left(\mu(\beta),\phi\right)p(\beta,\phi)}{q(\beta,\phi)}q(\beta,\phi)\,d\beta d\phi \tag{7}$$

$$BF=\frac{\sum_r w_r\left(M_1\right)}{\sum_r w_r\left(M_2\right)},\ \text{gdzie}\ w_r\left(M_g\right)=\frac{p_r\left(y\middle|\beta,\phi\right)p_r(\beta,\phi)}{q_r(\beta,\phi)}.$$

Podstawowe zalecenia odnośnie do konstruowania funkcji ważności $q(\bullet)$ wskazują na wykorzystanie wielowymiarowego rozkładu *t*-Studenta o niskiej liczbie stopni swobody, którego wektor wartości oczekiwanych oraz macierz kowariancji wyznacza się na podstawie wyników próbkowania z rozkładu *a posteriori* [Rossi, Allenby, McCulloch 2005, s. 162-166; Congdon 2006, s. 30-32].

## 5. Przykład empiryczny

Przedstawiona metodologia zostanie zilustrowana na podstawie zbioru danych zawierającego informację o dochodzie całkowitym gospodarstwa domowego (zmienna niezależna) oraz o odsetku wydatków na żywność (zmienna zależna). Obserwacje pochodzą z losowej próby 38 gospodarstw domowych z dużego miasta w Stanach Zjednoczonych (zob.: [Griffiths, Hill, Judge 1993, tab. 15.4]). Wybór tego stosunkowo prostego zbioru danych jest podyktowany faktem, że kilka spośród dotychczas opublikowanych opracowań traktujących o analizie regresji zmiennej zależnej o wartościach z przedziału (0,1) wykorzystuje go w celach ilustracyjnych [Ferrari, Cribari-Neto 2004; Branscum, Johnson, Thurmond 2007].

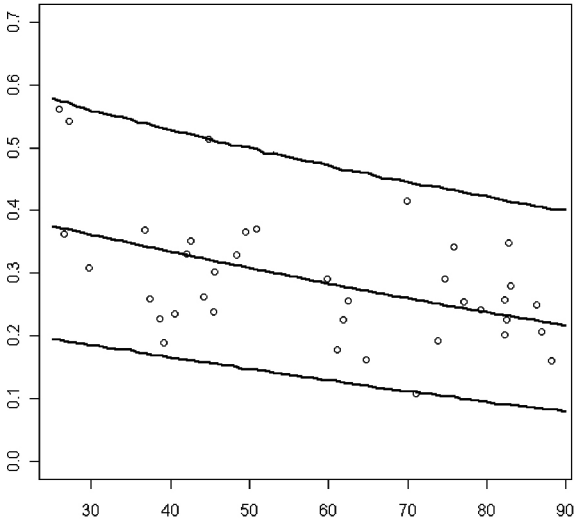
**Tabela 1.** Wyniki estymacji modelu regresji

| Parametr | Ocena punktowa | 95-procentowy przedział ufności | Efekty krańcowe | 95-procentowy przedział ufności |
|---|---|---|---|---|
| $\beta_0$ | –0,211 | (–0,626; 0,208) | – | – |
| $\beta_1$ | –0,119 | (–0,0189; –0,0049) | –0,00244 | (–0,00386; –0,00104) |
| $\phi$ | 27,51 | (16,63; 41,18) | – | – |

Źródło: opracowanie własne.

W tabeli 1 zaprezentowano podstawowe charakterystyki rozkładu *a posteriori* parametrów modelu, tj. wartości przeciętne, które uzupełniono o 95-procentowe przedziały ufności. Dodatkowo umieszczono tam punktową oraz przedziałową ocenę efektu krańcowego zmiany dochodu całkowitego gospodarstwa domowego (przy założeniu, że dochód znajduje się na przeciętnym dla próby poziomie).

Na rysunku 2 przedstawiono wykres rozrzutu danych wraz z naniesioną na niego funkcją regresji oraz dolną i górną granicą predykcji (95-procentowy przedział predykcji uzyskany na podstawie rozkładu predyktywnego). Na podkreślenie zasługuje obserwacja, że uzyskane przedziały predykcji ściśle odzwierciedlają naturę ograniczonej zmiennej zależnej, tzn. są one asymetryczne (uwzględnienie skośności rozkładu zmiennej zależnej) oraz ich długość zmniejsza się w miarę zbliżania się do krańców przedziału określoności zmiennej zależnej (uwzględnienie zależności wariancji zmiennej zależnej od jej wartości oczekiwanej).



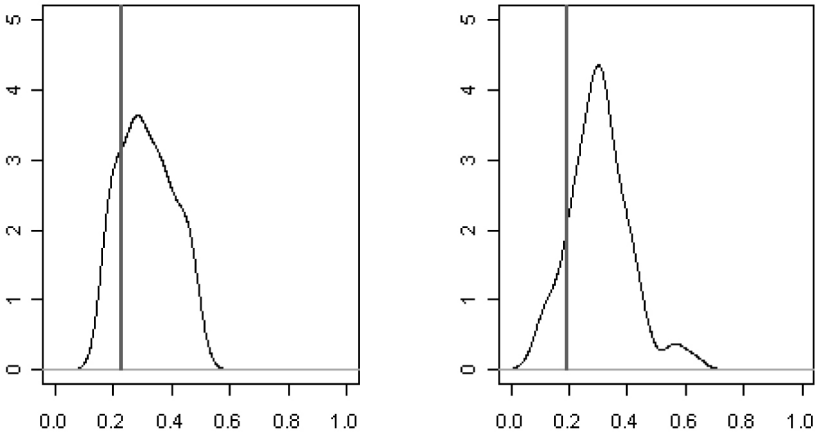

**Rys. 2.** Wykres rozrzutu danych wraz z dopasowaną funkcją regresji oraz 95-procentowymi przedziałami predykcji

Źródło: opracowanie własne.

Na rysunku 3 przedstawiono wykres funkcji gęstości predyktywnej dla dwóch przykładowych obserwacji. Wykres prezentuje rzeczywistą realizację zmiennej zależnej (pionowa kreska) oraz rozkład prawdopodobieństwa dla przyszłych realizacji wartości zmiennej zależnej. Lewa część wykresu obrazuje sytuację, w której przyszłe obserwacje generowane przez model znajdują się w zgodzie z zaobserwowaną wartością. Natomiast prawa część wykresu wskazuje sytuację, w której przyszłe obserwacje częściej będą większe niż zaobserwowana wartość.

W rozważanym przypadku (tylko 1 zmienna niezależna) informacja zawarta na rys. 3 znajduje się w bezpośredniej korespondencji z informacją przedstawioną na

rys. 2. Jednakże w sytuacji dużej liczby zmiennych niezależnych, gdy niemożliwe jest ich jednoczesne przedstawienie na wykresie rozrzutu, wykresy gęstości predyktywnej nadal dostarczają informacji o jakości dopasowania modelu do danych.




**Rys. 3.** Funkcja gęstości predyktywnej dla 2 przykładowych obserwacji

Źródło: opracowanie własne.

**Tabela 2.** Porównanie konkurencyjnych modeli regresji – czynniki Bayesa

| Rozkład | Beta | Simplex | Normalny |
|---|---|---|---|
| Beta | 1 | 0,82 | 210,6 |
| Simplex | 1,22 | 1 | 256,7 |
| Normalny | 0,0047 | 0,0039 | 1 |

Źródło: opracowanie własne.

W tabeli 2 zaprezentowano wyniki porównania konkurencyjnych specyfikacji modeli, w których kolejno założono, że rozkład zmiennej zależnej jest rozkładem beta, simplex lub normalnym. W wyniku stwierdzono, że dane przemawiają za rozkładem simplex, jednakże różnica pomiędzy nim a rozkładem beta jest zaniedbywalna. Istotna jest obserwacja, że dane bardzo silnie odrzucają model o warunkowym rozkładzie normalnym na korzyść dwóch pozostałych modeli.

## 6. Dyskusja

Interesującym, aczkolwiek mało znanym rozkładem prawdopodobieństwa zdefiniowanym na przedziale (0,1) jest dwuparametrowy rozkład Kumaraswamy. Jest

on równie elastyczny jak rozkład beta [Mitnik 2008]. Wadą tego rozkładu w porównaniu z rozkładem beta jest brak prostej formuły na wartość oczekiwaną oraz wariancję. Zaletą jest posiadanie dystrybuanty w postaci analitycznej. Fakt ten otwiera możliwość zbudowania modelu regresji na podstawie mediany.

Przedmiotem dalszych prac będzie próba wykorzystania bayesowskiego uśredniania modeli w celu uwzględnienia niepewności o prawdziwej postaci rozkładu zmiennej zależnej, tzn. tego, czy jest to rozkład beta, simplex czy rozkład Kumaraswamy. W przypadku omawianej klasy modeli jest to obszar dotychczas niezbadany.

## Literatura


Barndorff-Nielsen O., *Some Parametric Models on the Simplex*, „Journal of Multivariate Analysis" 1991 vol. 39, s. 106-116.

Branscum A., Johnson W., Thurmond M., *Bayesian Beta Regression: Application to Household Expenditure Data and Genetic Distance between Foot-and-mouth Disease Viruses*, „Australian & New Zealand Journal of Statistics" 2007 vol. 49, no 3, s. 287-301.

Buckley J., *Estimation of Models with Beta-Distributed Dependent Variables: A Replication and Extension of Paolino (2001)*, „Political Analysis" 2002 vol. 11, s. 1-12.

Congdon P., *Bayesian Statistical Modelling*, Wiley, 2006.

Ferrari S., Cribari-Neto F., *Beta Regression for Modelling Rates and Proportions*, „Journal of Applied Statistics" 2004 vol. 31(7), s. 799-815.

Griffiths W., Hill R., Judge G., *Learning and Practicing Econometrics*, Wiley, 1993.

Kieschnick R., *Regression Analysis of Variates Observed on (0,1): Percentages, Proportions and Fractions*, „Statistical Modelling" 2003 vol. 3, no 3, s. 193-213.

Lesaffre E., Rizopoulos D., Tsonaka S., *The Logistic-transform for Bounded Outcome Scores*, Technical Report 0448, http://www.stat.ucl.ac.be/IAP, 2004.

Lynch S., *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*, Springer, 2007.

Mitnik P., *The Kumaraswamy Distribution: a Median Dispersion Reparametrization for Regression Modeling and Simulation-based Estimation*, Working Paper, http://ssrn.com/abstract=1231587, 2008.

Osiewalski J., *Ekonometria bayesowska w zastosowaniach*, AE, Kraków, 2001.

Ospina R., Ferrari S., *Inflated Beta Distributions*, Statistical Papers, Springer, 10.1007/s00362-008-0125-4, 2008.

Paolino P., *Maximum Likelihood Estimation of Models with Beta-Distributed Dependent Variables*, „Political Analysis" 2001 vol. 9, no 4, s. 325-346.

Qiu Z., Song P., Tan M., *Simplex Mixed-Effects Models for Longitudinal Proportional Data*, „Scandinavian Journal of Statistics" 2008 vol. 35, s. 577-596.

Rossi P., Allenby G., McCulloch R., *Bayesian Statistics and Marketing*, Wiley, 2005.

Smithson M., Verkuilen J., *A Better Lemon Squeezer? Maximum-Likelihood Regression With Beta-Distributed Dependent Variables*, „Psychological Methods" 2006 vol. 11, no 1, 54-71.

Smithson M., Verkuilen J., *Beta Regression: Practical Issues in Estimation*, http://psychology.anu.edu.au/people/smithson/details/betareg/Readme.pdf, 2005.

# REGRESSION MODEL FOR DATA RESTRICTED TO THE INTERVAL (0,1) – BAYESIAN APPROACH

**Summary:** This article presents a regression framework for a dependent variable which is restricted to the open interval (0,1). The main drawbacks of widely used methods of modelling this type of data (e.g. linear regression model) have been briefly discussed. In this context, the beta distributed dependent variable is presented on the basis of which a regression model is constructed. The estimation of the model parameters as well as graphical methods for assessing the goodness of fit and the interpretation of model parameters are shown within the Bayesian framework. Next the Bayesian comparison of three competing models assuming the beta, simplex or normal distribution of a dependent variable is conducted. The model comparison results are presented in terms of the Bayes Factors. Theoretical results are applied to a small dataset on food expenditure and income. Future research work will investigate, among others, the application of the Kumaraswamy distribution for a dependent variable and the application of the Bayesian model averaging.