Jerzy P. Rydlewski[1], Daniel Kosiorowski[2]

Badanie możliwości zastosowania beta-regresji do modelowania związków pomiędzy stopą bezrobocia w województwach a innymi wskaźnikami makroekonomicznymi w latach 2004–2018[3]

1. Wstęp

W modelu beta-regresji zakładamy, że zmienna objaśniana ma rozkład beta. Model jest użyteczny w sytuacjach, gdy zmienna objaśniana przyjmuje dowolne wartości z pewnego ustalonego przedziału, np. z przedziału [0%, 100%]. W najprostszej sytuacji rozkład beta jest dwuparametrowy, jednakże można rozważać jego bardziej skomplikowaną postać, która pozwala na uwzględnienie zmienności w czasie niektórych parametrów rozkładu beta. Używając metody największej wiarygodności, można estymować parametry modelu.

Zastosowaliśmy metodę beta-regresji do modelowania związków pomiędzy stopą bezrobocia w polskich województwach a wybranymi wskaźnikami makroekonomicznymi w latach 2004–2018. Uwzględnienie modelu beta-regresji, w którym modelowana jest wartość oczekiwana zmiennej losowej o rozkładzie beta, pozwala na lepszy wgląd w naturę obserwowanego zjawiska.

[1] Akademia Górniczo-Hutnicza w Krakowie, Wydział Matematyki Stosowanej.

[2] Uniwersytet Ekonomiczny w Krakowie, Wydział Zarządzania.

[3] Jerzy P. Rydlewski uprzejmie dziękuje za wsparcie finansowe ze strony AGH w Krakowie, dotacja statutowa dla WMS grant numer 16.16.420.054. Daniel Kosiorowski uprzejmie dziękuje za wsparcie finansowe ze strony MNiSW w ramach „Regional Initiative of Excellence" Programme for 2019–2022. Project no.: 021/RID/2018/19. Total financing: 11 897 131,40 PLN. Daniel Kosiorowski dziękuje również UEK w Krakowie za wsparcie w postaci środków na utrzymanie potencjału badawczego przyznanych Wydziałowi Zarządzania na 2019 r.

2. Motywacje i cele

Modeli regresji używa się, by analizować zmienne powiązane ze sobą. Jak pokazuje praktyka, model regresji liniowej jest tym, który jest szczególnie popularny w różnego rodzaju zastosowaniach. Nie jest on jednakże właściwy, gdy z góry wiadomo, że zmienna zależna należy do pewnego przedziału, np. do przedziału [0,1], bądź opisuje frakcję, o której z góry wiadomo, że musi należeć do przedziału [0%,100%]. W przypadku zastosowania metody regresji liniowej do prognozowania takich danych może się wręcz zdarzyć, że uzyskana prognoza jest mniejsza od 0% albo większa od 100%, co czyni taką prognozę, rzecz jasna, bezwartościową.

Jednym z rozwiązań jest przekształcenie zmiennej zależnej tak, by przyjmowała dowolne wartości rzeczywiste, a następnie modelowanie wartości oczekiwanej tak przekształconej średniej jako liniowego predyktora zmiennych objaśniających. Takie podejście ma jednak pewne wady. Pierwsza z nich wynika z tego, że parametry modelu nie mają naturalnej interpretacji. Kolejna wada wiąże się z faktem, że miary proporcji zwykle wykazują asymetrię, tymczasem wnioskowanie „standardowe" bazuje na założeniu asymptotycznej normalności. W konsekwencji w przypadku modelowania danych wykazujących asymetrię oparte na założeniu asymptotycznej normalności „przedziały ufności" nie mogą być zdefiniowanymi poprawnie statystycznymi przedziałami ufności. Może to prowadzić do błędnej analizy statystycznej[4].

Jeszcze innym rozwiązaniem jest założenie, że analizowane dane pochodzą z rozkładu normalnego, a następnie można je przycinać. Niestety asymptotyczny rozkład estymatora największej wiarygodności takiego modelu jest nieznany, możliwe jest także, że estymator największej wiarygodności w tak sformułowanym modelu nie istnieje[5]. Wreszcie zauważmy, że istnieją modele, których nie da się odpowiednio zlinearyzować[6].

Celem naszej pracy jest zaproponowanie użyteczności aplikacyjnej mniej popularnej metody analitycznej, jaką jest metoda beta-regresji, służąca do modelowania zmiennej zależnej przyjmującej wartości z pewnego określonego przedziału. Nasza hipoteza badawcza brzmi: proponowana metoda jest lepsza

---

4 B.H. Baltagi, *Econometric Analysis of Panel Data*, Wiley, Chichester 2005.

5 C. Kleiber, S. Kotz, *Statistical Size Distributions in Economics and Actuarial Sciences*, Wiley, New Jersey 2003.

6 J.P. Rydlewski, *Estymatory największej wiarygodności w uogólnionych modelach regresji nieliniowej*, rozprawa doktorska, Uniwersytet Jagielloński, 2010.

od metody regresji liniowej dla rozpatrywanego typu danych. Ponadto chcemy pokazać, jak modelować zmienną zależną zmienną w czasie, przyjmującą jednak wartości z pewnego określonego przedziału. Wykorzystując zaproponowaną metodę, chcemy modelować związki przyczynowe pomiędzy stopą bezrobocia a innymi wskaźnikami makroekonomicznymi, np. tymi, które odzwierciedlają stopień rozwoju cyfrowego Polski w ujęciu województw. Zauważmy, że im lepszy jest model opisujący powyższe relacje, tym bardziej efektywną politykę społeczno-ekonomiczną może prowadzić rząd. Zastosowanie takiej metody umożliwia ulepszenie polityki samorządowej bądź stworzenie adaptacyjnej polityki samorządowej wiążącej przewidywania stopy bezrobocia ze stopniem pozyskiwania środków unijnych, wielkością realizowanych inwestycji w regionach albo z miarami rozwoju cyfrowego regionów.

3. Model beta-regresji

Podstawowy model beta-regresji zakłada, że zmienna zależna ma rozkład beta. Gęstość takiej zmiennej dana jest wzorem

$$f(x,\varphi,r)=\frac{1}{B(p,q)}x^{p-1}(1-x)^{q-1},\ x\in[0,1], \tag{1}$$

gdzie parametry rozkładu $p>0$ i $q>0$ oraz B to funkcja beta. W innym zapisie ta funkcja gęstości wyraża się wzorem

$$f(x,\varphi,r)=\frac{1}{B(r\varphi,(1-\varphi)r)}x^{r\varphi-1}(1-x)^{(1-\varphi)r-1},\ x\in[0,1], \tag{2}$$

gdzie B to funkcja beta, natomiast parametry funkcji gęstości spełniają nierówności: $0<\varphi<1,\ r>0.$ W jeszcze innym, równoważnym zapisie gęstość ta wyraża się wzorem

$$f(x,\varphi,r)=\frac{\Gamma(r)}{\Gamma(r\varphi)\Gamma((1-\varphi)r)}x^{r\varphi-1}(1-x)^{(1-\varphi)r-1},\ x\in[0,1], \tag{3}$$

gdzie Γ to funkcja gamma. Jeżeli zmienna losowa ma rozkład beta, to jej wartość oczekiwana oraz wariancja wyrażają się wzorami

$$E(X)=\varphi \tag{4}$$

oraz

$$V(X) = \frac{\varphi(1-\varphi)}{1+r}, \tag{5}$$

przy czym r to nieznany parametr precyzji w tym sensie, że – jak pokazuje ostatni wzór – dla ustalonej wartości oczekiwanej im większy jest parametr r, tym mniejsza jest wariancja zmiennej losowej o rozkładzie beta.

Uogólniony model liniowy zaproponowany w pracach J. Neldera i R.W.M. Wedderburna[7] oraz P. McCullagha i J.A. Neldera[8] zakłada, że model ma postać

$$E(y \mid x) = \varphi_x \tag{6}$$

oraz

$$g(\varphi_x) = a + b^T x, \tag{7}$$

gdzie rozkład zmiennej losowej y należy do rodziny wykładniczej rozkładów. Funkcję g nazywamy funkcją łączącą (ang. *link function*). Linearyzacja modelu polega na przekształceniu nieliniowej zależności pomiędzy y i x tak, by uzyskać zależność liniową albo otrzymać rozkład błędów bliski rozkładowi normalnemu. Następnie można stosować metody uogólnionego modelu liniowego. Konkretna funkcja g umożliwia zwykle osiągnięcie tylko jednego z powyższych celów.

S.L.P. Ferrari i F. Cribari-Neto[9] zaproponowali model beta-regresji, który choć jest podobny do skrótowo wyżej opisanego uogólnionego modelu liniowego, to jednak nie jest jego szczególnym przypadkiem. Założyli oni, że zmienna losowa y ma rozkład beta z parametrami φ i r. Struktura modelująca zależność zmiennej zależnej od zmiennych niezależnych opisywana jest wzorem

$$g(\varphi(t)) = \sum_{i=1}^{k} x_{ti} b_i, \tag{8}$$

gdzie $b = (b_1,\ldots,b_k)^T$ to wektor nieznanych parametrów regresji, $x_{t1},\ldots,x_{tk}$ to wektor k zmiennych zależnych oraz $k < n$, przy czym zwykle przyjmuje się $x_{i1} = 1$, aby model zawierał wyraz wolny. Zakłada się ponadto, że funkcja g jest ściśle monotoniczna, podwójnie różniczkowalna, odwzorowująca przedział (0,1) na zbiór liczb rzeczywistych. Zastosowanie idei funkcji łączącej pozwala na pewną

---

7 J. Nelder, R.W.M. Wedderburn, *Generalized Linear Models*, „Journal of the Royal Statistical Society" series A, 1972, vol. 135(3), s. 370–384.

8 P. McCullagh, J.A. Nelder, *Generalized Linear Models*, Chapman and Hall, New York 1983.

9 S.L.P. Ferrari, F. Cribari-Neto, *Beta regression for modelling rates and proportions*, „Journal of Applied Statistics" 2004, vol. 31(7), s. 799–815.

elastyczność modelu, ponieważ rozwiązując konkretne zagadnienie, można wybrać funkcję łączącą, która umożliwia najlepsze dopasowywanie do danych. Funkcjami łączącymi mogą być np.:

- funkcja logitowa: $g(\varphi) = \log\frac{\varphi}{1-\varphi}$;
- funkcja probitowa: $g(\varphi) = \Phi^{-1}(\varphi)$, gdzie $\Phi(\cdot)$ to dystrybuanta standardowego rozkładu Gaussa;
- funkcja log-log: $g(\varphi) = -\log(-\log\varphi)$.

W zaproponowanym modelu parametry regresji opisują średnią zmiennej zależnej, model jest heteroskedastyczny i ze względu na „plastyczność" rozkładu beta pozwala rozważać przypadki asymetrycznej zależności pomiędzy zmiennymi.

S.L.P. Ferrari i F. Cribari-Neto[10] konstruują następnie dla tak zdefiniowanego modelu funkcję wiarygodności i obliczają estymatory największej wiarygodności (ENW). Korzystają przy tym milcząco z założenia, że ENW parametrów modelu istnieją oraz są wyznaczone jednoznacznie. W swojej pracy udowadniają ponadto, że ENW w modelu beta regresji, o ile spełniają pewne naturalne założenia, są zgodne oraz asymptotycznie normalne.

Formalny dowód faktu, że założenia umożliwiające korzystanie z modelu beta-regresji są poprawne, to znaczy, że estymatory największej wiarygodności rzeczywiście istnieją i są jednoznaczne, znajduje się w pracach J.P. Rydlewskiego[11] oraz J.P. Rydlewskiego i D. Mielczarka[12].

Po estymowaniu parametrów modelu metodą największej wiarygodności ważne jest przeprowadzenie analiz diagnostycznych w celu sprawdzenia jakości dopasowanych estymatorów. Można rozważać globalną miarę wariancji wyjaśnianej oddziaływaniem zmiennych objaśniających występujących w modelu. W tym celu autorzy wprowadzają współczynnik pseudo R^2 równy kwadratowi współczynnika korelacji próbkowej pomiędzy $g(y)$ oraz $g(\hat{\varphi})$.

Można także rozważać różnego rodzaju reszty. Zwykłe residua postaci $r_t = y_t - \hat{\varphi}_t$ nie mają uzasadnienia ze względu na heteroskedastyczność, będącą nieodłączną częścią modelu. Dlatego też S.L.P. Ferrari i F. Cribari-Neto[13] rozważali standaryzowane residua (residua Pearsona)

---

[10] Ibidem.

[11] J.P. Rydlewski, *Beta-regression model for periodic data with a trend*, „Universitatis Iagellonicae Acta Mathematica" 2007, t. 45, s. 211–222; J.P. Rydlewski, *Estymatory największej*…, op.cit.

[12] J.P. Rydlewski, D. Mielczarek, *On the maximum likelihood estimator in the generalized beta regression model*, „Opuscula Mathematica" 2012, vol. 32, s. 761–774.

[13] S.L.P. Ferrari, F. Cribari-Neto, op.cit.

$$r_t = \frac{y_t - \hat{\varphi}_t}{\sqrt{\hat{V}(y_t)}}. \tag{9}$$

Ponieważ nie jest znany rozkład tych reszt, to autorzy zaproponowali metodę graficzną, która polega na analizie wykresu standaryzowanych reszt względem obserwacji. Więcej szczegółów znajduje się w artykule S.L.P. Ferrari i F. Cribari-Neto[14].

F. Cribari-Neto i A. Zeileis[15] zaprezentowali *betareg* – darmowy pakiet środowiska R, który umożliwia przeprowadzanie obliczeń obejmujących zastosowania beta-regresji. W tym pakiecie wykonano większość obliczeń zaprezentowanych w tej pracy.

W pracach J.P. Rydlewskiego[16] oraz J.P. Rydlewskiego i D. Mielczarka[17] znajduje się także opis i formalne uzasadnienie poprawności modelu beta-regresji, w którym parametr φ zmienia się w czasie. Ponieważ parametr φ wyraża wartość oczekiwaną, to zaproponowany model pozwala na modelowanie zmiennej losowej o rozkładzie beta, której wartość oczekiwana zmienia się w czasie. Tym samym powstaje możliwość rozważania zagadnienia beta-regresji, w którym błędy modelu są charakteryzowane przez rozkład beta, którego wartość oczekiwana zmienia się w czasie, i w dodatku jesteśmy w stanie modelować postać funkcyjną wartości oczekiwanej.

Przejdźmy teraz do opisu zmiennego w czasie modelu beta-regresji. Niech $x_1,\ldots, x_n$ będą niezależnymi zmiennymi losowymi o rozkładzie beta. Ze względu na możliwość przeskalowania danych, ograniczonych do z góry zadanego przedziału, wystarczy, że rozważamy dane z przedziału [0,1]. Zakładamy, że wartość oczekiwana zmiennej zależnej jest modelowana następująco:

$$E(x_j) = \varphi(t_j) = \sum_{k=1}^{m} a_k f_k(t_j), \tag{10}$$

to znaczy jest sumą pewnej ustalonej liczby ciągłych funkcji liniowo niezależnych; t_j można traktować jako moment dokonywania pomiaru.

Ponieważ parametry rozkładu beta możemy zapisać następująco: $p=r\varphi$ oraz $A = (A_1,\ldots,A_m)$, to prawdziwy jest wzór

---

[14] Ibidem.

[15] F. Cribari-Neto, A. Zeileis, *Beta regression in R*, „Journal of Statistical Software" 2010, vol. 34, s. 1–24.

[16] J.P. Rydlewski, *Beta-regression model…*, op.cit.; J.P. Rydlewski, *Estymatory największej…*, op.cit.

[17] J.P. Rydlewski, D. Mielczarek, op.cit.

$$p(A,t) = \sum_{k=1}^{m} A_k f_k(t). \tag{11}$$

Warto rozważyć model, w którym funkcja ze wzorów (7) oraz (8) jest funkcją okresową, a parametrami są jej współczynniki Fouriera[18]. Dodatkowo można i warto do takiego modelu dołączyć funkcję opisującą trend. Układem funkcji modelujących parametr $p(A,t)$ może być, przykładowo,

$$(\tau_1, \sin(\cdot), \sin 2(\cdot), ..., \sin m(\cdot), 1, \cos(\cdot), \cos 2(\cdot), ..., \cos m(\cdot)). \tag{12}$$

Rzecz jasna, można dodać kolejną funkcję trendu, byleby tylko była ona ciągła i tworzyła układ funkcji liniowo niezależnych z pozostałym funkcjami. Otrzymujemy wtedy następujący układ funkcji

$$(\tau_1(\cdot), \tau_2(\cdot), \sin(\cdot), \sin 2(\cdot), ..., \sin m(\cdot), 1, \cos(\cdot), \cos 2(\cdot), ..., \cos m(\cdot)). \tag{13}$$

Dla tego ostatniego przypadku układu funkcji połóżmy $A = (A_{-1}, A_{-2}, A_1, ..., A_m, B_0, B_1, ..., B_m)$ i wtedy „parametr" p rozkładu beta przybiera postać następującej funkcji:

$$p(A,t) = A_{-1}\tau_1(t) + A_{-2}\tau_2(t) + B_0 + \sum_{k=1}^{m} (A_k \sin kt + B_k \cos kt). \tag{14}$$

Funkcja wiarygodności w takim modelu przyjmuje postać:

$$L(t_1, ...t_n, x_1, ..., x_n, A, r) = \prod_{j=1}^{n} \frac{1}{B(p(A,t_j), r - p(A,t_j))} x_j^{p(A,t_j)-1} (1-x_j)^{r-p(A,t_j)-1}, \tag{15}$$

natomiast logarytm naturalny funkcji wiarygodności w takim modelu ma postać

$$\ln L(t_1, ...t_n, x_1, ..., x_n, A, r) = \sum_{j=1}^{n} -\ln B(p(A,t_j), r - p(A,t_j)) +$$
$$+ (p(A,t_j) - 1)\ln x_j + (r - p(A,t_j) - 1)\ln(1 - x_j). \tag{16}$$

W pracach J.P. Rydlewskiego[19] oraz J.P. Rydlewskiego i D. Mielczarka[20] udowodniono, że w powyższym modelu beta-regresji istnieje dokładnie jeden

[18] J.P. Rydlewski, *Estymatory największej…*, op.cit.
[19] J.P. Rydlewski, *Beta-regression model…*, op.cit.; ibidem.
[20] J.P. Rydlewski, D. Mielczarek, op.cit.

estymator największej wiarygodności rozpatrywanych parametrów. Wykazano także, że estymatory największej wiarygodności, o ile spełniają pewne naturalne założenia, są zgodne oraz asymptotycznie normalne.

4. Modelowanie stopy bezrobocia

Zastosowaliśmy model beta-regresji, by analizować powiązanie stopy bezrobocia w województwach z pewnymi wybranymi wskaźnikami makroekonomicznymi w latach 2004–2018. Interesowały nas związki przyczynowe pomiędzy stopą bezrobocia a tymi wskaźnikami makroekonomicznymi, które odzwierciedlają stopień rozwoju cyfrowego Polski w ujęciu województw. Posłużyliśmy się danymi z Głównego Urzędu Statystycznego (GUS)[21]. Za zmienną objaśnianą posłużyły nam dane GUS opisujące roczne stopy bezrobocia rejestrowanego w województwach Polski od 2004 r. do 2018 r.

Stopę bezrobocia rejestrowanego GUS oblicza jako stosunek liczby bezrobotnych zarejestrowanych do liczby cywilnej ludności aktywnej zawodowo, przy czym bez osób odbywających czynną służbę wojskową oraz pracowników jednostek budżetowych prowadzących działalność w zakresie obrony narodowej i bezpieczeństwa publicznego. Stopę bezrobocia podaje się z uwzględnieniem pracujących w gospodarstwach indywidualnych w rolnictwie.

Sprawdzaliśmy, czy i jak stopa bezrobocia rejestrowanego jest powiązana z danymi opisującymi nakłady inwestycyjne na jednego mieszkańca, wykorzystaniem technologii informatycznych w firmach czy też nakładami na działalność badawczą i rozwojową. Analizowaliśmy różne zmienne, jednakże ze względu na braki w różnego rodzaju danych dostarczanych przez GUS oraz niewielką zmienność pośród niektórych innych zmiennych zdecydowaliśmy się na zastosowanie modelu beta-regresji, w którym zmiennymi objaśniającymi będą nakłady inwestycyjne na jednego mieszkańca oraz nakłady wewnętrzne na działalność badawczą i rozwojową na jednego mieszkańca każdego z województw.

Nakłady inwestycyjne na jednego mieszkańca zawierają informacje o nakładach inwestycyjnych, tj. „nakładach finansowych lub rzeczowych, których celem jest stworzenie nowych środków trwałych, rozbudowa lub modernizacja istniejących obiektów majątku trwałego”[22].

---

21 https://bdl.stat.gov.pl/BDL/dane/podgrup/tablica (odczyt: 05.04.2019).

22 Ibidem.

Drugą zmienną objaśniającą są nakłady wewnętrzne na działalność badawczą i rozwojową na jednego mieszkańca. Według opisu GUS: „Działalność badawczo-rozwojowa obejmuje pracę twórczą podejmowaną w sposób metodyczny w celu zwiększenia zasobów wiedzy – w tym wiedzy o rodzaju ludzkim, kulturze i społeczeństwie – oraz w celu tworzenia nowych zastosowań dla istniejącej wiedzy”[23].

Na początku rozważaliśmy model beta-regresji w postaci zaproponowanej przez S.L.P. Ferrari i F. Cribari-Neto[24], w którym zmienną zależną jest stopa bezrobocia rejestrowanego w latach 2004–2018, natomiast zmiennymi niezależnymi są nakłady inwestycyjne na jednego mieszkańca (b_2) oraz nakłady wewnętrzne na działalność badawczą i rozwojową na jednego mieszkańca (b_3). Model ten zawiera wyraz wolny (b_1). Zastosowaliśmy logitową funkcję łączącą. Z powyższego opisu wynika, że model beta-regresji ma następującą postać

$$g(\varphi(t)) = b_1 + b_2 x_{t2} + b_3 x_{t3}. \tag{17}$$

Oznaczmy ten model jako model nr 1. Korzystając z pakietu *betareg*, otrzymaliśmy estymację parametrów modelu dla całej Polski oraz dla każdego z województw. Ze względu na ograniczoną ilość miejsca zaprezentujemy wyniki dla Polski, rozpatrywanej jako całość (zob. tabela 1) oraz dla wybranych województw – dolnośląskiego oraz podkarpackiego (zob. tabele 2 i 3).

Tabela 1. Estymacja parametrów w modelu nr 1, gdy rozpatrujemy dane dla całej Polski

Parametr	Estymator	Błąd standardowy	Statystyka z	Wartość p
b_1	–1,254	1,610e-01	–7,788	6,78e-15
b_2	–7,951e-05	4,612e-05	–1,724	0,0847
b_3	–1,338e-03	5,601e-04	–2,389	0,0169
r	242,84	88,68	2,738	0,00618

Źródło: opracowanie własne.

Współczynnik pseudo R^2 wyniósł 0,6699.

[23] Ibidem.

[24] S.L.P. Ferrari, F. Cribari-Neto, op.cit.

Tabela 2. Estymacja parametrów w modelu nr 1, gdy rozpatrujemy dane dla województwa dolnośląskiego

Parametr	Estymator	Błąd standardowy	Statystyka z	Wartość p
b_1	–1,008	1,744e-01	–5,778	7,58e-09
b_2	–1,027e-04	4,313e-05	–2,382	0,01722
b_3	–1,751e-03	6,573e-04	–2,664	0,00771
r	190,69	69,65	2,738	0,00618

Źródło: opracowanie własne.

Współczynnik pseudo R^2 wyniósł 0,7626.

Tabela 3. Estymacja parametrów w modelu nr 1, gdy rozpatrujemy dane dla województwa podkarpackiego

Parametr	Estymator	Błąd standardowy	Statystyka z	Wartość p
b_1	–1,487	1,778e-01	–8,366	<2e-16
b_2	–2,799e-05	6,607e-05	–0,424	0,672
b_3	–8,381e-04	5,671e-04	–1,478	0,139
r	240,0	87,6	2,74	0,00614

Źródło: opracowanie własne.

Współczynnik pseudo R^2 wyniósł 0,3971. Podobne wyniki jak dla województwa podkarpackiego, tzn. brak statystycznej istotności parametrów b_2 i b_3, uzyskaliśmy dla województw: kujawsko-pomorskiego, łódzkiego, lubelskiego, opolskiego, podlaskiego, śląskiego, świętokrzyskiego, warmińsko-mazurskiego oraz zachodniopomorskiego. W przypadku pozostałych województw wyniki są podobne do wyników województwa dolnośląskiego.

Zastosowaliśmy metodę regresji liniowej do danych rozpatrywanych w modelu nr 1. Zarówno dla całej Polski, jak i dla każdego z województw uzyskaliśmy brak statystycznej istotności parametrów. Innymi słowy, nie da się dokonać uzasadnionego w ramach metod statystyki dopasowania parametrów modelu liniowego dla rozpatrywanych danych.

Następnie badaliśmy model, w którym zmienną zależną jest stopa bezrobocia rejestrowanego w latach 2004–2018, natomiast zmienną niezależną są wyłącznie nakłady inwestycyjne na jednego mieszkańca (model nr 2), oraz kolejny model, w którym zmienną niezależną są nakłady wewnętrzne na działalność badawczą i rozwojową na jednego mieszkańca (model nr 3). Oba modele zawierają wyraz wolny. Zastosowaliśmy logitową funkcję łączącą. Model beta-regresji ma więc w każdym z tych dwóch przypadków postać

$$g(\varphi(t)) = b_1 + b_2 x_{t2}. \tag{18}$$

Tabela 4. Estymacja parametrów w modelu nr 2, gdy rozpatrujemy dane dla całej Polski

Parametr	Estymator	Błąd standardowy	Statystyka z	Wartość p
b_1	–1,259	1,874e-01	–6,722	1,79e-11
b_2	–1,561e-04	3,812e-05	–4,095	4,22e-05
r	171,32	62,57	2,738	0,00618

Źródło: opracowanie własne.

Współczynnik pseudo R^2 wyniósł 0,5257.

Tabela 5. Estymacja parametrów w modelu nr 2, gdy rozpatrujemy dane dla województwa dolnośląskiego

Parametr	Estymator	Błąd standardowy	Statystyka z	Wartość p
b_1	–0,991862	0,2079602	–4,769	1,85e-06
b_2	–0,0001788	0,0000372	–4,806	1,54e-06
r	124,95	45,64	2,738	0,00619

Źródło: opracowanie własne.

Współczynnik pseudo R^2 wyniósł 0,6429. Podobną statystyczną istotność uzyskaliśmy dla każdego z województw.

Tabela 6. Estymacja parametrów w modelu nr 3, gdy rozpatrujemy dane dla całej Polski

Parametr	Estymator	Błąd standardowy	Statystyka z	Wartość p
b_1	–1,4561099	–0,1284655	–11,34	< 2e-16
b_2	–0,0019775	0,0004346	–4,55	5,35e-06
r	202,73	74,03	2,738	0,00618

Źródło: opracowanie własne.

Współczynnik pseudo R^2 wyniósł 0,6263.

Tabela 7. Estymacja parametrów w modelu nr 3, gdy rozpatrujemy dane dla województwa dolnośląskiego

Parametr	Estymator	Błąd standardowy	Statystyka z	Wartość p
b_1	–1,3383247	0,1371426	–9,759	< 2e-16
b_2	–0,0027496	0,0005562	–4,944	7,67e-07
r	136,82	49,98	2,738	0,00619

Źródło: opracowanie własne.

Współczynnik pseudo R^2 wyniósł 0,6867. Podobną statystyczną istotność uzyskaliśmy dla każdego z województw.

Zastosowaliśmy metodę regresji liniowej do danych rozpatrywanych w modelu nr 2 oraz w modelu nr 3. Zarówno dla całej Polski, jak i dla każdego z województw uzyskaliśmy tym razem istotne statystycznie dopasowanie. Jednakże za każdym razem skorygowane R^2 było mniejsze niż dla modelu beta-regresji oraz wartość kryterium AIC uzyskana dla modelu beta-regresji była mniejsza niż wartość AIC uzyskana dla modelu regresji liniowej. Dla zobrazowania uzyskanych wyników poniżej zamieszczamy porównanie AIC dla modelu beta-regresji i dla modelu regresji liniowej zastosowanych wobec danych z modelu nr 3.

Tabela 8. Porównanie wartości kryterium AIC dla modelu beta-regresji oraz dla modelu regresji liniowej zastosowanych wobec danych z modelu nr 3 w przypadku, gdy rozpatrujemy dane dla całej Polski oraz dla województwa dolnośląskiego

	Model beta-regresji	Model regresji liniowej
Polska	AIC = –65,83	AIC = –65,26
Województwo dolnośląskie	AIC = –60,50	AIC = –57,16

Źródło: opracowanie własne.

Kryterium AIC, za każdym razem mniejsze dla modelu beta-regresji niż dla modelu regresji liniowej, pokazuje, że dla rozpatrywanych danych model beta-regresji dokonuje lepszego dopasowania niż model regresji liniowej.

5. Podsumowanie

Zaproponowana metoda oferuje – zgodnie z powszechnymi na gruncie ekonometrii miernikami jakości estymacji, takimi jak kryterium informacyjne Akaike (AIC) – lepsze dopasowanie od metody regresji liniowej. Ponadto jest niejako dopasowana do danych, wobec których zmienna zależna przybiera wartości z ustalonego przedziału domkniętego. Jej stosowanie do tego typu danych jest ponadto uzasadnione metodami statystyki matematycznej. W przyszłości planujemy dokonanie obliczeń z wykorzystaniem modelu beta-regresji uwzględniającego czas, gdzie możemy uwzględnić czynnik trendu oraz okresowość. Dla tego modelu nie istnieje na razie pakiet ułatwiający dokonanie obliczeń[25].

[25] Por. D. Kosiorowski, Z. Zawadzki, *DepthProc: An R Package for Robust Exploration of Multidimensional Economic Phenomena*, „Journal of Statistical Software" 2019 (w druku).

Bibliografia

Baltagi B.H., *Econometric Analysis of Panel Data*, Wiley, Chichester 2005.

Cribari-Neto F., Zeileis A., *Beta regression in R*, „Journal of Statistical Software" 2010, vol. 34, s. 1–24.

Ferrari S.L.P., Cribari-Neto F., *Beta regression for modelling rates and proportions*, „Journal of Applied Statistics" 2004, vol. 31(7), s. 799–815.

Kleiber C., Kotz S., *Statistical Size Distributions in Economics and Actuarial Sciences*, Wiley, New Jersey 2003.

Kosiorowski D., Zawadzki Z., *DepthProc: An R Package for Robust Exploration of Multidimensional Economic Phenomena*, „Journal of Statistical Software" 2019 (w druku).

McCullagh P., Nelder J.A., *Generalized Linear Models*, Chapman and Hall, New York 1983.

Nelder J., Wedderburn R.W.M., *Generalized Linear Models*, „Journal of the Royal Statistical Society" series A, 1972, vol. 135(3), s. 370–384.

Rydlewski J.P., *Beta-regression model for periodic data with a trend*, „Universitatis Iagellonicae Acta Mathematica" 2007, t. 45, s. 211–222.

Rydlewski J.P., *Estymatory największej wiarygodności w uogólnionych modelach regresji nieliniowej*, rozprawa doktorska, Uniwersytet Jagielloński, 2010.

Rydlewski J.P., Mielczarek D., *On the maximum likelihood estimator in the generalized beta regression model*, „Opuscula Mathematica" 2012, vol. 32, s. 761–774.

Źródła sieciowe

https://bdl.stat.gov.pl/BDL/dane/podgrup/tablica (odczyt: 05.04.2019).

* * *

Beta-regression application to model relationships between the unemployment rate in voivodships and other macroeconomic indicators for the years 2004–2018

Summary

Beta-regression method has been applied to model relationship between the unemployment rate in voivodships and other macroeconomic indicators of digital development, i.e., e.g. intramural expenditures on R&D for the years 2004–2018. Taking into account the beta regression model, in which the mean of beta distributed random variable is modelled, allows for a better insight into the nature of the observed phenomenon.

Keywords: Beta-regression, unemployment rate modelling, digital economy in Poland.