

Zastosowanie regresji logistycznej w studiach nad Unią Europejską

Adam Kirpsza

Streszczenie: Celem artykułu jest opisanie głównych założeń i sposobu praktycznego zastosowania regresji logistycznej w badaniach nad Unią Europejską. Jest to jedna z metod statystycznych, za pomocą której można w łatwy sposób weryfikować teorie rozumiane jako zbiór hipotez wraz z określeniem ich mocy predykcyjnej. W części pierwszej zdefiniowano procedurę statystycznego testowania hipotez, która stanowi fundament powyższej metody. W części drugiej scharakteryzowano w sposób matematyczny podstawowe założenia regresji logistycznej. W części trzeciej przedstawiono przebieg właściwej weryfikacji hipotez w oparciu o tę metodę. W ostatniej części zaprezentowano egzemplifikację zastosowania regresji logistycznej w badaniu wpływu określonych czynników na zaangażowanie ministrów w procesie podejmowania decyzji w Radzie.

„Holiści-pluraliści-behawiorysty-fizykaliści orzekli, iż jak wiadomo z fizyki, prawidłowość w naturze jest tylko statystyczna. Podobnie jak nie można z zupełną dokładnością przewidzieć drogi pojedynczego elektronu, tak też nie wiadomo na pewno, jak się będzie zachowywał pojedynczy kartofel. Dotychczasowe obserwacje pouczają, że miliony razy człowiek kopał kartofle, ale nie jest wykluczone, że jeden raz na miliard stanie się na odwrót, że kartofel będzie kopał człowieka”¹.

Stosowanie narzędzi statystycznych w analizach politologicznych w Polsce jest rzadkością. Na pierwszy rzut oka wydaje się, że zjawisko to wynika przede wszystkim z braku kompetencji matematycznych czy awersji badaczy nauk o polityce do modelowania liczbowego. Ale niechęć do statystyki ma również podłoże metodologiczne. Wielkie wrażenie, w szczególności na polskiej politologii bazującej w dużej mierze na filozofii i historii, wywarły pisma Karla Poppera i Imre Lakatosa, które poważnie zredefiniowały pojęcie i uprawianie nauki². Jednym z osiągnięć tych dwóch uczonych było wykazanie słabości probabilizmu, to jest programu badawczego, który zakładał, że choć teorie

¹ S. Lem, *Dzienniki gwiazdowe*, Warszawa 2008, s. 261.

² Vide K. Popper, *Logika odkrycia naukowego*, Warszawa 1977; idem, *Wiedza obiektywna. Ewolucyjna teoria epistemologiczna*, Warszawa 2002; I. Lakatos, *Pisma z filozofii nauk empirycznych*, Warszawa 1995.

naukowe są nieudowodnialne, to można im przypisać stopnie prawdopodobieństwa ze względu na dostępny materiał empiryczny³. Dowiedli oni bowiem, że każda teoria ma zerowe prawdopodobieństwo⁴. Ponieważ prawdopodobieństwo jest elementarną kategorią wnioskowania statystycznego, to jej zakwestionowanie spowodowało, że wielu politologów nabrało wątpliwości w stosunku do narzędzi statystycznych. W rezultacie, sposób uprawiania nauk o polityce pozostał teoretyczny, opisowy i postfaktyczny, pozbawiony mocy predykcyjnych, a matematycznie ograniczający się do obliczania średnich lub korelacji.

Jednakże ostatnie trendy w politologii, zwłaszcza w dziedzinie stosunków międzynarodowych, pokazują, że awersja do statystyki może w przyszłości ulec zmianie. Pojawiła się bowiem idea, która dezaktualizuje wnioski Poppera i Lakatosa, sprowadzająca się do postulatu „ukwantowania” nauk o polityce. Niektórzy badacze proponują, aby, tak jak w naukach ścisłych, odejść od analizy rzeczywistości w kategoriach newtonowskiej fizyki klasycznej i zaakceptować jej nowe, kwantowe rozumienie⁵. Cóż by to oznaczało? Otóż, jeśli przyjąć założenia mechaniki kwantowej, należałoby w naukach o polityce analizować wszelkie zjawiska w oparciu o metody statystyczne i koncepcję prawdopodobieństwa.

Wychodząc z powyższych obserwacji, autor niniejszego artykułu postawił sobie za cel przedstawienie jednej z metod statystycznych – regresji logistycznej. W części pierwszej zdefiniowano pojęcie weryfikacji hipotez statystycznych stanowiącej dział wnioskowania statystycznego. W części drugiej przedstawiono założenia regresji logistycznej, aby w części trzeciej opisać procedurę testowania przypuszczeń w oparciu o tę metodę. Natomiast w części czwartej zaprezentowano zastosowanie regresji logistycznej na przykładzie zaangażowania ministrów rządów państw członkowskich w proces decyzyjny w Radzie Unii Europejskiej.

Weryfikacja hipotez statystycznych

Weryfikacja hipotez statystycznych jest jednym z działów wnioskowania statystycznego, czyli uogólniania otrzymanych wyników badania próby

³ Vide R. Carnap, *Logical Foundations of Probability*, London 1971; H. Reichenbach, *The Theory of Probability: An Inquiry into the Logical and Mathematical Foundations of the Calculus of Probability*, Berkeley 1971.

⁴ Szerzej w: A. Grobler, *Metodologia nauk*, Kraków 2006, s. 37–60.

⁵ Vide A. Wendt, 'Social Theory' as Cartesian Science: An Auto-Critique from a Quantum Perspective, [w:] S. Guzzini, A. Leander (red.), *Constructivism and International Relations: Alexander Wendt and Its Critics*, New York 2006; D. Akrivoulis, *The 'Quantum Politics' Metaphor in International Relations: Towards a Hermeneutics of Political Metaphoricity*, paper presented at the 57th Political Studies Association Annual Conference, [University of Bath, UK, 11–13 April 2007; T. Becker, *Quantum Politics: Applying Quantum Theory to Political Phenomena*, New York 1991; C. Zorn, C. Smith jr, *Some Quantum-Like Features of Mass Politics in Two-Party Systems*, Cornell University, 4 April 2011, <http://arxiv.org/pdf/1107.0964v1.pdf>.

losowej na całą populację wraz z oszacowaniem błędów wynikających z tego uogólnienia⁶. Polega ona na postawieniu przypuszczeń dotyczących rozkładu zmiennych, a następnie sprawdzeniu ich prawidłowości poprzez dowodzenie nie wprost (*reductio ad absurdum*). Oznacza to, że prawdziwość przyjętej hipotezy jest wykazywana poprzez odrzucenie założenia o prawdziwości jej negacji. Weryfikacja przypuszczeń statystycznych *sensu stricto* sprowadza się zatem do trzech działań: po pierwsze, sformułowania hipotezy badawczej H_1 , którą chce się udowodnić, po drugie, postawienia negacji H_1 w postaci hipotezy zerowej H_0 i przyjęcia, że jest ona prawdziwa, po trzecie, weryfikacji H_0 , to jest sprawdzenia, do jakich konsekwencji prowadzi przyjęcie założenia o jej prawdziwości. Jeżeli rzeczony konsekwencje polegają na otrzymaniu wyniku mało prawdopodobnego, to można odrzucić założenie o prawdziwości H_0 . W rezultacie, stosując prawo wyłączonego środka, które stanowi, że z pary: zdanie i jego negacja tylko jedno jest prawdziwe, udowadnia się prawidłowość H_1 .⁷

Przenosząc powyższe założenia na szerszy wymiar, testowanie hipotez statystycznych *sensu largo* można sprowadzić do trzech etapów. Pierwszym jest konstrukcja odpowiedniego modelu, czyli testowalnej teorii. Na początku należy postawić pytanie badawcze formułujące zagadnienie, które ma być wyjaśnione. Następnie powyższe H_1, \dots, H_k , pytanie przekształca się w grupę hipotez badawczych (alternatywnych) gdzie k – numer hipotezy, prognozujących wpływ określonych czynników, zwanych zmiennymi niezależnymi, na badany problem, definiowany jako zmienna zależna. W kolejnym kroku należy sformułować negacje hipotez badawczych (H_0), które zakładają brak istotnego wpływu zmiennych niezależnych na zmienną zależną.

Drugim etapem jest testowanie hipotez zerowych poprzez dokonanie obliczeń prawdopodobieństw. Najpierw należy przyjąć metodę, za pomocą której zostanie przeprowadzona weryfikacja, czyli, w niniejszym artykule, regresję logistyczną. Następnie, wybiera się statystykę testową, przy użyciu której będzie badana istotność całego modelu i prognozowanych związków między zmiennymi. W regresji logistycznej jest to, co do zasady, statystyka Walda o rozkładzie chi-kwadrat (χ^2), o której będzie jeszcze mowa. Po określeniu próby losowej z populacji (np. 500 aktów prawnych przyjętych w latach 1999–2004) oblicza się wartość statystyki testowej, po czym należy wyznaczyć tzw. prawdopodobieństwo testowe (p-wartość, graniczny poziom istotności), to jest prawdopodobieństwo otrzymania wartości statystyki testowej większej lub równej niż otrzymanej przy założeniu prawdziwości H_0 , mówiąc prościej, prawdopodobieństwo, z jakim prawdziwa jest H_0 . Im p-wartość jest większa, tym bardziej H_0 jest prawdziwa i *vice versa* – mała p-wartość świadczy o fałszywości H_0 . Na tym etapie można jednak popełnić dwa błędy, to jest błąd

⁶ R. Szwed, *Metody statystyczne w naukach społecznych. Elementy teorii i zadania*, Lublin 2009, s. 24, 144.

⁷ G. Wieczorkowska, J. Wierziński, *Statystyka: od teorii do praktyki*, Warszawa 2011, s. 189–191.

I rodzaju polegający na odrzuceniu H_0 , która w rzeczywistości jest prawdziwa, i błąd II rodzaju, gdy przyjęta zostanie fałszywa H_0 . W celu ich minimalizacji należy określić tzw. poziom istotności α , uznaniowo wynoszący 0,05 (5 przypadków na 100), który wyznacza maksymalnie dopuszczalne prawdopodobieństwo ich popełnienia, inaczej mówiąc, maksymalne prawdopodobieństwo, dla którego badacz jest w stanie zaakceptować własną hipotezę jako prawdziwą. Ogółem przyjmuje się, że jeśli p – wartość $\leq \alpha$, to można H_0 odrzucić, a gdy p – wartość $> \alpha$, to nie ma podstaw do odrzucenia H_0 .

W rzeczywistości jednak obliczanie p-wartości jest bardzo skomplikowane, dlatego korzysta się z tablic rozkładu danej statystyki. W przypadku rozkładu chi-kwadrat, na którym oparta jest statystyka Walda, tablica zawiera w wierszach stopnie swobody ($df = k - 1$, gdzie k – liczba wartości w zmiennej zależnej, czyli w regresji logistycznej 2, bo 0 i 1), w kolumnach poziomy istotności, a jej wnętrzu stanowią wartości krytyczne, czyli liczby odcinające na wykresie rozkładu skrajne pole (zwane obszarem krytycznym), w którym znajdują się mało prawdopodobne wyniki statystyki dla danego poziomu istotności skutkujące odrzuceniem H_0 . Postępowanie wygląda następująco: najpierw należy odczytać wartość krytyczną chi-kwadrat dla przyjętego błędu I rodzaju (poziomu istotności) $\alpha=0,05$ ($\chi^2_{krytyczne}$), a następnie porównać ją z wcześniej otrzymanym wynikiem statystyki ($\chi^2_{otrzymane}$).

Trzecim etapem weryfikacji jest podjęcie decyzji w sprawie hipotez zerowych. Jeżeli:

- 1) $\chi^2_{otrzymane} \geq \chi^2_{krytyczne}$ – odrzuca się H_0 , wykazując prawdziwość H_k ;
- 2) $\chi^2_{otrzymane} < \chi^2_{krytyczne}$ – nie ma podstaw od odrzucenia H_0 , H_k jest fałszywa.

Takie rozumowanie wynika z założenia, że jeżeli wartość statystyki testowej obliczona na podstawie danych z próby znajduje się w obszarze krytycznym wyznaczonym przez α , to znaczy, że efekt zawarty w hipotezie ma szansę pojawić się w mniej niż 5 przypadkach na 100, a więc w przedziale przyjętym przez badacza. Inaczej mówiąc, jest istotny statystycznie.

Założenia regresji logistycznej

Przed omówieniem regresji logistycznej, kluczowym zadaniem jest zdefiniowanie pojęcia regresji. Otóż, jest to metoda statystyczna, której celem jest badanie zależności pomiędzy zmiennymi niezależnymi (objaśniającymi, wyjaśniającymi), zwanymi także kowariantami (*covariates*) lub predyktorami (*predictors*), a zmienną zależną (objaśnianą, wyjaśnianą). Formalnie rzecz ujmując, regresja pozwala estymować warunkową wartość oczekiwaną ($[E(Y|X)]$ – co do zasady średnią arytmetyczną) zmiennej zależnej dla określonych wartości lub wektora zmiennych niezależnych. Inaczej mówiąc, umożliwia przewidywanie wartości zmiennej zależnej na podstawie znanych wartości predyktorów. W praktyce rzeczona metoda polega na konstrukcji funkcji opisującej,

jak zależy wartość oczekiwana zmiennej zależnej od zmiennych niezależnych. Dlatego ogólną postać regresji można przedstawić następująco:

$$E(X|Y) = f(X, \beta) \quad (1.1)$$

$$Y = f(X, \beta) + \varepsilon \quad (1.2)$$

gdzie:

Y – zmienna zależna,

X – zmienne niezależne,

$E(X|Y)$ – warunkowa wartość oczekiwana Y przy danej wartości X ,

β – wektor współczynników regresji,

$f(X, \beta)$ – funkcja regresji,

ε – błąd losowy.

Najbardziej popularną regresją jest regresja liniowa, która dla wielu zmiennych niezależnych (regresja liniowa wielokrotna) przyjmuje postać:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon \quad (1.3)$$

gdzie:

β_0 – wyraz wolny (intercept, stała), punkt przecięcia z osią Y , gdy funkcja nie zawiera zmiennych ($X=0$),

β_k – współczynniki kierunkowe nachylenia linii regresji,

β_0 i β_k – współczynniki regresji,

k – liczba kolejnych zmiennych niezależnych.

Celem regresji liniowej jest konstrukcja na wykresie rozrzutu zmiennych X i Y linii prostej, która będzie najlepiej opisywać związek między nimi. Najpierw uzyskane z próby dane transformuje się w postaci zmiennych i wprowadza się na wykres. Następnie konstruuje się na ich podstawie prostą za pomocą metody najmniejszych kwadratów polegającej na znalezieniu jak najmniejszej sumy kwadratów odchyleń (czyli odległości) punktów obserwacji zmiennych od linii regresji. Uzyskane współczynniki prostej informują, jak wzrost lub spadek zmiennej X wpływa na Y . Na przykład, prosta $Y = 2000 + 397X$, gdzie X jest zmienną „wykształcenie” mierzoną w latach edukacji, a Y „dochodem” (w zł), oznaczałaby, że wzrost wykształcenia o jeden rok odpowiada wzrostowi zarobków o 397 zł. Ponadto, jeśli wstawi się w miejsce X liczbę lat edukacji danej osoby, np. 10, można wyliczyć, że jej średni dochód będzie wynosił 5970 zł. Regresja liniowa pozwala zatem nie tylko przewidywać wpływ predyktorów na zmienną zależną, ale także prognozować jej średnią wartość..

Jednakże powyższa procedura napotyka na trzy problemy, gdy jej zmienna zależna jest dychotomiczna (binarna, 1 lub 0). Po pierwsze, przy Y znajdującym się w przedziale $[0,1]$ regresja nie jest liniowa, lecz przybiera kształt litery S, ponieważ zmiana wielkości $E(X|Y)$ przy spadku lub wzroście X o jedną jednostkę jest progresywnie mniejsza, gdy warunkowa wartość oczekiwana zbliża się do 0 lub 1. Ponadto, nie może ona przyjąć wartości od $-\infty$ do $+\infty$ ⁸.

⁸ D. Hosmer, S. Lemeshow, *Applied Logistic Regression*, New York 2000, s. 5–6.

Po drugie, prosta wytyczona na bazie binarnej zmiennej zależnej łamie założenie o homoskedastyczności, czyli zasadę, że wariancja błędu losowego jest stała lub taka sama dla wszystkich wartości zmiennych niezależnych⁹. Po trzecie, przy skali dychotomicznej naruszona jest reguła o rozkładzie normalnym błędu losowego¹⁰. Dlatego badanie za pomocą regresji liniowej wpływu predyktorów na binarną zmienną objaśnianą jest niemożliwe.

Remedium na powyższe problemy stanowi właśnie regresja logistyczna. Można ją opisać za pomocą następujących równań:

$$P = \frac{e^{\text{logit}(p)}}{1 + e^{\text{logit}(p)}} = \frac{1}{1 + e^{-\text{logit}(p)}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)}} \quad (1.4)$$

gdzie:

e – stała Eulera = 2,71828..., baza logarytmu naturalnego,

$$\text{logit}(p) = \ln \frac{p}{1-p} = \ln(p) - \ln(1-p) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k \quad (1.5)$$

$$O = \frac{p}{q} = \frac{p}{1-p} = e^{(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)} \quad (1.6)$$

Powyższe reguły matematyczne generują cztery zasadnicze założenia regresji logistycznej. Po pierwsze, jak już wspomniano, zmienna zależna w tej procedurze statystycznej musi mieć rozkład dychotomiczny (binarny)¹¹, co oznacza, że może ona przyjąć tylko dwie wartości: albo 1, gdy dane zdarzenie ma miejsce, albo 0, gdy dane zdarzenie nie ma miejsca, np. zdrowy/chory, wygrana/przegrana. Rzeczona zmienna nie ma zatem, tak jak w przypadku regresji liniowej, rozkładu normalnego i nie jest ilościowa, lecz nominalna.

Po drugie, lewą stronę równania (1.4) nie stanowi, tak jak w regresji liniowej, oczekiwana wartość zmiennej zależnej, lecz prawdopodobieństwo P . Oznacza to, że regresja logistyczna polega nie na objaśnianiu samej zmiennej zależnej, lecz na określeniu prawdopodobieństwa przyjęcia przez nią danej wartości, co do zasady 1, a więc prawdopodobieństwa zajścia zdarzenia opisywanego w tejże zmiennej. Należy jednak podkreślić, że regresja logistyczna zakłada specyficzną definicję tej kategorii. Klasyczne rozumienie prawdopodobieństwa zaproponowane przez Pierre'a de Laplace'a sprowadza się do stosunku liczby zdarzeń korzystnych (sukcesów) do liczby wszystkich prób, zakładając że wszystkie przypadki wzajemnie się wykluczają i są jednakowo możliwe. Natomiast regresja logistyczna bazuje na tzw. prawdopodobień-

⁹ S. Menard, *Applied Logistic Regression*, Thousand Oaks 2002, s. 5, 7.

¹⁰ D. Hosmer, S. Lemeshow, *Applied Logistic...*, s. 6–7.

¹¹ Warto jednak zaznaczyć, że w regresji logistycznej zmienna zależna może przybierać także więcej niż dwie wartości (tzw. wielomianowa regresja logistyczna – *multinomial logistic regression*). Vide D. Hosmer, S. Lemeshow, *Applied Logistic...*, s. 260–287.

stwie szansy (*odds* – O), czyli stosunku prawdopodobieństwa, że dane zdarzenie wystąpi (p) do prawdopodobieństwa, że dane zdarzenie nie wystąpi ($q = 1 - p$), co obrazuje (1.6).

Po trzecie, z (1.5) wynika, że regresja logistyczna ma postać funkcji logistycznej (logistycznej), nie zaś liniowej. Wynika to z faktu, że choć szansa posiada tę zaletę w stosunku do prawdopodobieństwa, że przyjmuje wartości w przedziale $(0, +\infty)$, podczas gdy $p \in (0,1)$, to jej poważną wadą jest dolna granica wynosząca 0 oraz brak normalnego i linearnego rozkładu w stosunku do zmiennych niezależnych. W rezultacie, niemożliwa jest interpretacja współczynników regresji (β). Aby rozwiązać ten problem, stosuje się logarytm naturalny szansy, który ma tę właściwość, że jest liniowo zależny od zmiennej objaśniającej X oraz przyjmuje wartości $(-\infty, +\infty)$. Dzięki temu, otrzymuje się postać znaną z regresji liniowej, co umożliwia łatwe obliczanie i interpretację współczynników regresji β . Po przekształceniu odwrotnym (1.5) pojawia się pełna postać modelu logistycznego (1.6).

Po czwarte, o ile współczynniki regresji liniowej były szacowane za pomocą metody najmniejszych kwadratów, o tyle w regresji logistycznej są one obliczane według metody największej wiarygodności. Polega ona na konstrukcji funkcji wiarygodności, jako funkcji nieznanymi parametrów β , która określa prawdopodobieństwo uzyskania zaobserwowanych danych X , a następnie obliczeniu jej maksimum, które będą poszukiwanymi współczynnikami regresji. Funkcję wiarygodności można przedstawić wzorem:

$$L(\beta) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i}$$

ale z uwagi na lepszą matematycznie postać wyraża się ją w formie jej logarytmu naturalnego:

$$\ln[L(\beta)] = \sum_{i=1}^n [y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)]$$

W celu znalezienia estymatorów największej wiarygodności β , należy podzielić powyższe równanie na β_0 i β_k i zrównać je do zera lub zastosować metody iteracyjne najmniejszych kwadratów¹².

Choć powyższe formuły matematyczne mogą okazać się skomplikowane, to z pomocą przychodzą programy statystyczne (np. SPSS, STATA, STATISTICA, SAS czy R), które automatycznie je wyliczają.

Testowanie modelu regresji logistycznej

Po zapoznaniu się z założeniami regresji logistycznej, można przejść do procedury weryfikacji hipotez za pomocą tej metody. Odbywa się ona na trzech

¹² D. Larose, *Metody i modele eksploracji danych*, Warszawa 2008, s. 166.

obszarach. Pierwszym jest testowanie słuszności całej teorii, a więc wszystkich przypuszczeń razem wziętych. Sprowadza się ono do zbadania jakości dopasowania modelu, czyli odpowiedzi na pytanie, w jaki stopniu dobrze wyjaśnia on zmienną zależną. Drugim obszarem jest testowanie poszczególnych hipotez polegające na identyfikacji istotności statystycznej przewidzianego wpływu predyktorów na zmienną zależną. Trzecim wymiarem weryfikacji jest analiza mocy predykcyjnej teorii, czyli poznania, jak dobrze model klasyfikuje obserwacje.

Testowanie jakości dopasowania całego modelu

Weryfikacja jakości dopasowania modelu (teorii) polega na przeprowadzeniu czterech testów: współliniowości, ilorazu wiarygodności, Hosmera-Lemeshowa i zbieżności.

Test współliniowości (*collinearity*) polega, z jednej strony, na sprawdzeniu, czy między zmiennymi niezależnymi nie ma silnej zależności liniowej, z drugiej zaś, czy jedna ze zmiennych nie jest liniową kombinacją kilku innych, co określa się zjawiskiem wielowspółliniowości (*multicollinearity*). Współliniowość generuje poważne problemy analityczne, ponieważ w sytuacji silnego skorelowania parametrów, uzyskują one wysokie wartości błędów standardowych (S.E.)¹³, przez co mogą być błędnie określone jako nieistotne statystycznie. Co więcej, silna zależność może spowodować zawyżone oszacowanie wielkości oraz zmianę kierunku współczynnika regresji, tym samym niepoprawnego określenia efektu predyktora. W rezultacie, następuje błędna weryfikacja hipotez.

W celu diagnozy powyższych zjawisk należy przeprowadzić dwie operacje. Pierwszą jest wyliczenie korelacji liniowej między wszystkimi zmiennymi niezależnymi, biorąc również pod uwagę wyraz wolny. Współczynnik korelacji jest określany mianem *r* Pearsona i definiowany wzorem:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

$$r \in \langle -1, 1 \rangle$$

gdzie:

x, *y* – zmienne korelowane,

i – próby losowe,

n – liczba obserwacji,

\bar{x} , \bar{y} – średnie wartości z prób *x* i *y*.

¹³ Błąd standardowy jest miarą błędu oszacowania, czyli różnicy między przewidywaną wartością zmiennej zależnej a jej rzeczywistą wartością. Sposoby jego obliczania zawiera: ibidem, s. 47–48, 174

W przypadku, gdy $r > 0,7$, można mówić o bardzo silnej zależności, co powinno prowadzić do usunięcia z analizy jednej z pary skorelowanych zmiennych lub jej transformacji¹⁴. Natomiast jeśli współliniowość dotyczy wyrazu wolnego, najlepszym rozwiązaniem jest jego pozostawienie, gdyż odgrywa on rolę swoistego „kosza na śmieci”, zawierając wszystkie niewyjaśnione wariacje zmiennych zawartych w modelu¹⁵.

Drugą procedurą diagnozy współliniowości jest obliczenie współczynnika *VIF* (*variance inflation factor*). Pokazuje on, o ile wariacje parametrów są zawyżone pod wpływem zależności liniowych w modelu, to jest, o ile większy jest błąd standardowy w porównaniu do jego wartość w sytuacji, gdy dana zmienna nie byłaby skorelowana z inną. Dla danego predyktora *VIF* można zdefiniować w następujący sposób:

$$VIF_j = \frac{1}{(1 - R_j^2)}$$

gdzie:

R_j^2 – współczynnik determinacji dla j -tej zmiennej.

W literaturze można spotkać opracowania, które w różny sposób określają poziom *VIF* świadczący o obecności współliniowości¹⁶, np. $VIF > 10$ lub $VIF \geq 5$. W niniejszej pracy przyjęto tę drugą wartość, gdyż oznacza ona przynajmniej dwukrotne powiększenie przedziału ufności z powodu korelacji zmiennych. W przypadku pojawienia się tej wartości, postępowanie jest podobne jak w procedurze pierwszej – daną zmienną należy wyeliminować lub przetransformować.

Test ilorazu wiarygodności¹⁷ ma za zadanie sprawdzenie, czy model ze zmiennymi jest lepszy od modelu bez żadnej zmiennej (tylko z wyrazem wolnym), inaczej mówiąc, czy dostarcza istotnie więcej informacji o zmiennej zależnej. W pierwszym kroku należy pomnożyć logarytm funkcji wiarygodności $\log(L)$ przez -2 , otrzymując równanie zwane dewiancją (D), która pozwala stwierdzić, czy model generuje wiarygodne wyniki, to jest, w jakim zakresie dobrze prognozuje wartości zmiennej zależnej. Im mniejsza dewiancja, tym większa wiarygodność modelu:

$$D = - 2 \log L$$

Następnie dokonuje się analizy dewiancji modelu bez zmiennych niezależnych (X_k) oraz modelu z predyktorami za pomocą statystyki G o rozkładzie

¹⁴ R. Szwed, op.cit., s. 313.

¹⁵ Y.H. Chan, *Biostatistics 202. Logistic regression analysis*, „Singapore Medicine Journal” 2004, nr 4, s. 152.

¹⁶ Np. Grażyna Wieczorkowska i Jerzy Wierzbński proponują $VIF \geq 10$, podczas gdy Daniel Larose skłania się raczej ku $VIF \geq 5$. *Vide* G. Wieczorkowska, J. Wierzbński, op.cit., s. 310; D. Larose, op.cit., s. 125.

¹⁷ D. Hosmer, S. Lemeshow, *Applied Logistic...*, s. 145–147.

chi-kwadrat z k stopniami swobody. W tym celu stawia się H_0 , że model ze zmiennymi niezależnymi jest nieistotny statystycznie oraz hipotezę alternatywną H_1 , że model z danymi zmiennymi jest istotny statystycznie. Można to wyrazić w sposób następujący:

$$G = D[\text{model bez } X_k] - D[\text{model z } X_k] = -2 \log L[\text{model bez } X_k] + 2 \log L[\text{model z } X_k] = -2 \log \left[\frac{L[\text{model bez } X_k]}{L[\text{model z } X_k]} \right]$$

Otrzymaną wartość statystyki G analizuje się za pomocą tablicy rozkładu chi-kwadrat i jeżeli jest ona większa lub równa $\chi^2_{krytyczne}$ dla $\alpha = 0,05$ (p-wartość $\leq \alpha$), należy odrzucić H_0 i uznać, że model ze zmiennymi jest lepszy od zerowego. W innym przypadku nie ma podstaw do odrzucenia H_0 .

Test Hosmera-Lemeshowa (test H-L, *goodness-of-fit test*)¹⁸ analizuje jakość dopasowania modelu i polega na wyznaczeniu dziesięciu podgrup danych modelu i porównaniu liczności zdarzeń zaobserwowanych z częstością zdarzeń oczekiwanych. Modele, dla których są one bliskie, uznaje się za dobrze dopasowane. Test H-L wyraża się wzorem:

$$HL = \sum_{g=1}^n \frac{(O_g - E_g)^2}{N_g \pi_g (1 - \pi_g)}$$

gdzie:

O_g – zdarzenia zaobserwowane,

E_g – zdarzenia prognozowane,

N_g – liczba obserwacji,

π_g – prognozowane ryzyko,

n – liczba podgrup.

Przed przystąpieniem do weryfikacji należy postawić H_0 : $O_g = E_g$ dla wszystkich podgrup oraz hipotezę alternatywną H_1 : $O_g \neq E_g$ dla przynajmniej jednej podgrupy. Test H-L bazuje na statystyce o rozkładzie chi-kwadrat z dwoma stopniami swobody. Uzyskaną wartość testu należy zatem zanalizować w oparciu o tablicę tego rozkładu i porównać z wartością krytyczną dla α . Jeśli:

1) $\chi^2_{otrzymane} \geq \chi^2_{krytyczne}$ ($p \leq \alpha$) – odrzuca się H_0 , co oznacza, że model jest źle dopasowany, gdyż obserwowane wartości zdarzeń nie są bliskie wartościom prognozowanym;

2) $\chi^2_{otrzymane} < \chi^2_{krytyczne}$ ($p > \alpha$) – nie ma podstaw od odrzucenia H_0 , model jest dobrze dopasowany.

Test zbieżności polega na obliczeniu współczynnika determinacji (zbieżności) R-kwadrat informującego o proporcji zmienności (wariancji) Y wyjaśnianej

¹⁸ D. Hosmer, S. Lemeshow, *Goodness-of-fit tests for the multiple logistic regression model*, „Communications in Statistics – Theory and Methods” 1980, nr 10, s. 1043–1069; S. Lemeshow, D. Hosmer, *A review of goodness-of-fit statistics for use in the development of logistic regression models*, „American Journal of Epidemiology” 1982, nr 1, s. 92–106.

przez zmienność (wariancję) X . Jest on powszechnie stosowany w regresji liniowej i przyjmuje wartość z przedziału $[0,1]$. Uważa się, że im wielkość R^2 jest bliższa 1, tym lepsze dopasowanie modelu. Współczynnik determinacji wyraża się wzorem¹⁹:

$$R^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{\sum_{i=1}^n (Y - \bar{Y})^2}$$

gdzie:

Y_i – przewidywana (teoretyczna) wartość zmiennej zależnej Y ,

Y – rzeczywista wartość zmiennej zależnej Y ,

\bar{Y} – średnia arytmetyczna rzeczywistych wartości zmiennej zależnej Y .

Jak już wspomniano, w regresji logistycznej zamiast metody najmniejszych kwadratów jest stosowana funkcja wiarygodności, przez co powyższa wersja R^2 nie może być zastosowana. Proponuje się w zamian inną postać współczynnika determinacji, określaną mianem pseudo-R-kwadrat²⁰:

$$R_L^2 = \frac{L_0 - L_p}{L_0} = 1 - \frac{L_p}{L_0},$$

gdzie:

L_0 – wartość funkcji wiarygodności dla modelu bez zmiennych niezależnych,
 L_p – wartość funkcji wiarygodności dla modelu zawierającego p zmiennych niezależnych.

Wykazano jednak, że w regresji logistycznej *pseudo* R^2 nigdy nie przyjmuje wartości 1 i jest wrażliwy na ilość zmiennych w modelu²¹. Dlatego zaproponowano jego poprawione wersje, takie jak R^2 Cox-Snella²² czy R^2 Nagelkerke'a²³. Ten ostatni zostanie zastosowany w niniejszej pracy. Można go zobrazować następującym równaniem:

$$R_{Nagelkerke}^2 = \frac{1 - \left(\frac{L_0}{L_p}\right)^{2/n}}{1 - (L_0)^{2/n}},$$

¹⁹ W literaturze proponowany jest również skorygowany R^2 , który uwzględnia liczbę obserwacji oraz relacje między zmiennymi niezależnymi. Vide G. Wieczorkowska, J. Wierbiński, op.cit., s. 305–306.

²⁰ Zwany również, od nazwiska jego twórcy, R^2 McFaddena. Vide D. McFadden, *Conditional logit analysis of qualitative choice behavior*, [w:] *Frontiers in Econometrics*, red. P. Zarembka, New York 1974, s. 105–142.

²¹ D. Hosmer, S. Lemeshow, *Applied Logistic...*, s. 167.

²² D. Cox, E. Snell, *The Analysis of Binary Data*, London 1989.

²³ N. Nagelkerke, *A Note on a General Definition of the Coefficient of Determination*, „*Biometrika*” 1991, nr 3, s. 691–692.

gdzie:

n – liczba obserwacji.

Trzeba jednak podkreślić, że stosowanie współczynnika determinacji przy badaniu modeli regresji logistycznej jest krytykowane, stąd jego wartość należy oceniać sceptycznie²⁴.

Badania istotności statystycznej predyktorów

Po analizie jakości dopasowania modelu można przejść do weryfikacji poszczególnych hipotez. W regresji logistycznej stosuje się w tym celu dwie metody: test Walda (Z_w) i iloraz szans (OR – *odds ratio*). Pierwsza analizuje istotność statystyczną wpływu predyktorów na zmienną zależną i polega najpierw na oszacowaniu współczynnika regresji β oraz wartości błędu standardowego ($S.E.$ – *standard error*) dla danej zmiennej niezależnej X , a następnie obliczeniu na ich podstawie statystyki Walda:

$$Z_w = \frac{\beta}{S.E.(\beta)}$$

Statystyka Walda ma rozkład chi-kwadrat, dlatego otrzymaną jej wartość dla X należy porównać z odczytaną z tablic rozkładu wartością krytyczną dla $\alpha = 0,05$. Jeżeli:

1) $\chi^2_{otrzymane} \geq \chi^2_{krytyczne}$ ($p \leq \alpha$) – odrzuca się H_0 , co oznacza, że wpływ predyktora na zmienną zależną jest istotny statystycznie, a postawiona hipoteza badawcza jest prawdziwa;

2) $\chi^2_{otrzymane} < \chi^2_{krytyczne}$ ($p > \alpha$) – nie ma podstaw od odrzucenia H_0 , wpływ predyktora na zmienną zależną jest nieistotny, hipoteza badawcza okazała się fałszywa.

Iloraz szans natomiast informuje, jak zmienia się szansa zajścia zdarzenia zawartego w zmiennej zależnej (wartość 1) w przypadku wzrostu danej zmiennej niezależnej o jedną jednostkę. Wykorzystując (1.6), oblicza się go w następujący sposób:

$$OR = \frac{O(A)}{O(B)} = \frac{P(A)}{1 - P(A)} \div \frac{P(B)}{1 - P(B)}$$

gdzie:

$P(A)$ – prawdopodobieństwo zajścia zdarzenia (wartość 1 w zmiennej zależnej) w klasie obserwacji A (wartość 1 zmiennej niezależnej);

$P(B)$ – prawdopodobieństwo zajścia zdarzenia (wartość 1 w zmiennej zależnej) w klasie obserwacji B (wartość 0 zmiennej niezależnej).

²⁴ D. Hosmer i S. Lemeshow krytykują ten wskaźnik, podczas gdy S. Menard go popiera. Por. D. Hosmer, S. Lemeshow, *Applied Logistic...*, s. 167; S. Menard, *Applied Logistic...*, s. 26.

Przyjmuje się, że jeżeli:

1) $OR > 1$, to czynnik zawarty w X ma stymulujący wpływ na wystąpienie danego zdarzenia. Wtedy OR informuje, o ile wzrasta szansa na wystąpienie wartości 1 w zmiennej zależnej, gdy predyktor wzrasta o jedną jednostkę;

2) $OR < 1$, to czynnik zawarty w X negatywnie wpływa na wystąpienie danego zdarzenia. Wtedy OR informuje, o ile spada szansa na wystąpienie wartości 1 w zmiennej zależnej, gdy predyktor wzrasta o jedną jednostkę;

3) $OR = 1$, to czynnik zawarty w X nie ma wpływu na wystąpienie danego zdarzenia.

Przy analizie ilorazu szans wyznacza się również dla każdego predyktora tzw. przedział ufności (CI – *Confidence Interval*) – pojęcie wprowadzone przez statystyka polskiego pochodzenia Jerzego Spławę-Neymana²⁵. Jest to przedział liczbowy, który z określonym z góry przez badacza prawdopodobieństwem $1-\alpha$, zwanym poziomem ufności, będzie zawierał nieznaną wartość szacowanego parametru populacji. Inaczej mówiąc, CI określa minimalny i maksymalny iloraz szans dla danej zmiennej niezależnej. W celu obliczenia krańcowych granic przedziału ufności dla OR stosuje się następujący wzór:

$$e^{[\beta - z_{1-\alpha/2} S.E.(\beta), \beta + z_{1-\alpha/2} S.E.(\beta)]}$$

gdzie:

β – współczynnik regresji,

$z_{1-\alpha/2}$ – kwantyl rozkładu normalnego $N(0,1)$ rzędu $1 - \alpha/2$. Wielkość tę można zastąpić wartością $t(\alpha, \infty)$ odczytana z tablic statystycznych rozkładu t -Studenta.

Biorąc pod uwagę fakt, że w niniejszej pracy przyjęto $\alpha = 0,05$, to poziom ufności będzie wynosił 0,95, a według standardowej formuły wyrażającej tę wielkość w formie procentowej – $[100(1 - \alpha)\%]=95\%$. Oznacza to, że po obliczeniu dolnej i górnej granicy przedziału ufności będzie można z prawdopodobieństwem równym 95% oczekiwać, że iloraz szans dla danego predyktora będzie nie mniejszy i nie większy niż wartości dolnej i górnej granicy CI .

Test mocy predykcyjnej modelu

Weryfikacja możliwości predykcyjnych modelu (teorii) sprowadza się do dokonania dwóch testów w oparciu o tabelę klasyfikacyjną i krzywą ROC.

W przypadku regresji logistycznej standardowa **tabela klasyfikacyjna** jest macierzą kwadratową o wymiarach 2×2 , w której kolumny zawierają klasy decyzji prognozowane przez model, a wiersze klasy decyzji zaobserwowane realnie (zob. Tabela nr 1). Istotnym elementem analizy tabeli jest procen-

²⁵ J. Neyman, *Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability*, „Philosophical Transactions of the Royal Society of London. Series A. Mathematical and Physical Sciences” 1937, nr 767, s. 333–380.

towe określenie trzech wielkości: trafności (ACC), czułości (C – *sensitivity*) i specyficzności (S – *specificity*). Pierwsza określa, jaki procent w całej liczbie obserwacji stanowią przypadki poprawnie pozytywnie i negatywnie sklasyfikowane przez model (np. ilu łącznie chorych i zdrowych zostało poprawnie wskazanych jako chorych i zdrowych) i wyraża się wzorem:

$$ACC = 100 \left[\frac{TP + TN}{TP + TN + FP + FN} \right] \%$$

Czułość informuje o procentowym udziale przypadków poprawnie i pozytywnie sklasyfikowanych przez model we wszystkich przypadkach pozytywnie zaobserwowanych (np. ilu chorych zostało poprawnie wskazanych jako chorych) i przybiera postać:

$$C = 100 \left[\frac{TP}{TP + FN} \right] \%$$

Natomiast specyficzność określa procentowy udział przypadków poprawnie, lecz negatywnie przewidzianych przez model w ogólnej liczbie przypadków negatywnie zaobserwowanych (np. ilu zdrowych zostało poprawnie wskazanych jako zdrowych):

$$S = 100 \left[\frac{TN}{FP + TN} \right] \%$$

Powyższe wskaźniki są wyliczane przy tzw. punkcie odcięcia prawdopodobieństwa przewidywanego (co do zasady 0,5) i informują, w jakim stopniu model prawidłowo przewiduje wartości zmiennej zależnej. Należy jednak podkreślić, że analiza tabel klasyfikacyjnych posiada poważne ograniczenia, przez co powinna być stosowana jako narzędzie komplementarne w stosunku do bardziej rygorystycznych metod²⁶.

Tabela nr 1. Przykład tabeli klasyfikacyjnej

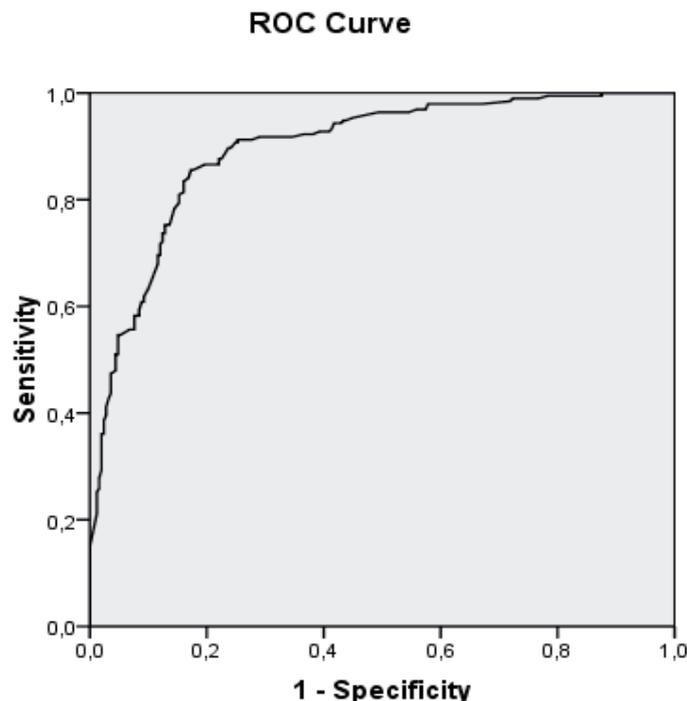
		Decyzje prognozowane	
		Pozytywna	Negatywna
Decyzje zaobserwowane	Pozytywna	Prawdziwie pozytywna (TP – <i>True Positive</i>)	Fałszywie Negatywna (FN – <i>False Negative</i>)
	Negatywna	Fałszywie pozytywna (FP – <i>False Positive</i>)	Prawdziwie negatywne (TN – <i>True Negative</i>)

²⁶ D. Hosmer, S. Lemeshow, *Applied Logistic...*, s. 160.

Analiza ROC (*Receiver Operating Characteristics*) polega na konstrukcji krzywej (zwanej ROC) ilustrującej związek między dwoma, opisanymi powyżej, współczynnikami klasyfikacyjnymi: czułością i specyficzną. Dla każdego z możliwych punktów odcięcia mieszczących się w przedziale $[0,1]$ należy obliczyć C i S , a następnie otrzymane rezultaty nanieść na układ współrzędnych, na którego osi odciętych (x) znajduje się specyficznosc (w postaci $1 - \text{Specyficznosc}$), zaś na osi rzędnych (y) – czułość. Uzyskane punkty łączy się w celu wykreślenia krzywej.

Wykres nr 1 obrazuje przykład krzywej ROC, w której zmienną zależną jest sukces postulatów Parlamentu (1-sukces; 0-porażka). Jeśli pokrywa się ona z przekątną ($y=x$), to oznacza, że model nic nie wnosi do analizy, gdyż dokonane przez niego klasyfikacje są tak samo dobre jak decyzje losowa. Im bardziej krzywa zbliża się do lewej górnej ćwiartki wykresu, tym lepsze są wskazania modelu. Wysoki poziom czułości (bliski 1) oznacza bowiem, że model prawidłowo rozpoznaje dane przypadki (chorych), podczas gdy niska wartość $1-S$ świadczy o tym, że niewiele przypadków negatywnych klasyfikuje jako pozytywne (zdrowych jako chorych). Pożądane jest zatem: wysoka wartość C i niska $1-S$.

Wykres nr 1. Krzywa ROC – zdjęcie krzywej otrzymane w programie SPSS



Diagonal segments are produced by ties.

Źródło: opracowanie własne na podstawie własnej bazy danych.

Najważniejszą jednak własnością interpretacyjną ROC jest pole powierzchni znajdującej się pod krzywą, oznaczane jako AUC (*Area Under Curve*). Przyjmuje

ono wartości w przedziale $[0,1]$ i świadczy o jakości klasyfikacyjnej modelu²⁷. Im wyższy jest wskaźnik AUC , tym lepszy model. Przyjmuje się, że jeśli²⁸:

- a) $AUC=0,5$ – brak klasyfikacji,
- b) $0,7 \leq AUC < 0,8$ – klasyfikacja jest akceptowalna,
- c) $0,8 \leq AUC < 0,9$ – klasyfikacja jest bardzo dobra,
- d) $AUC \geq 0,9$ – klasyfikacja jest najlepsza.

Przy analizie ROC stawia się hipotezę zerową H_0 : $AUC = 0,5$ oraz hipotezę alternatywną H_1 : $AUC \neq 0,5$. Następnie wyznacza się na podstawie statystyki testowej wartość p , którą porównujemy z poziomem istotności α . Jeśli $p \leq \alpha$, to odrzucamy H_0 i przyjmujemy H_1 , to znaczy model jest lepszy niż model losowy. Natomiast jeśli $p > \alpha$, to nie ma podstaw do odrzucenia H_0 , czyli zaproponowany model jest słaby pod względem prognostycznym.

Na analizowanym wykresie AUC wynosi 0,895, co oznacza, że w przypadku 89% par poprawek zaproponowanych przez Parlament, z których jedna została w pełni zaakceptowana (1), a druga odrzucona (0), model z danymi zmiennymi przypisuje większe prawdopodobieństwo poprawkom, które posłowie przeforsowali. Ponieważ $AUC > 0,5$ można stwierdzić, że model bardzo dobrze klasyfikuje przypadki, czyli ma zdecydowanie większą moc prognostyczną niż losowość.

Przykład zastosowania regresji logistycznej

Obszarem szczególnie eksplorowanym w literaturze poświęconej Unii Europejskiej jest proces podejmowania decyzji w Radzie. Liczne badania pokazują, że ok. 80%–90% aktów prawodawczych jest formalnie przyjmowanych na niższych szczeblach tej instytucji, to jest w grupach roboczych i w COREPER, co oznacza, iż tylko 10–20% projektów jest rozstrzyganych na spotkaniach ministerialnych (tzw. punkty B)²⁹. W rzeczywistości jednak ministrowie znacznie częściej angażują się w proces decyzyjny. Obserwacje funkcjonowania Rady pokazują, że jeszcze gdy projekt jest analizowany na niższych poziomach, organizowane są wielokrotnie specjalne spotkania ministerialne, na których debatuje się nad nim lub przyjmuje tzw. polityczne porozumienia, stanowiące potem wytyczne dla grup roboczych i COREPER w dalszych pracach nad

²⁷ J. Hanley, B. McNeil, *The Meaning and Use of the Area under a Receiving Operating Characteristics (ROC) Curve*, „Radiology” 1982, nr 1, s. 29–36; idem, *Receiver Operating Characteristics (ROC) Methodology: The State of the Art*, „Critical Reviews in Diagnostic Imaging” 1989, nr 3, s. 307–335.

²⁸ D. Hosmer, S. Lemeshow, *Applied Logistic...*, s. 162.

²⁹ F. Häge, *Who Decides in the Council of the European Union?*, „Journal Common Market Studies” 2008, nr 3, s. 546; M.P.M.C. van Schendelen, *„The Council Decides”: Does the Council Decide?*, „Journal of Common Market Studies” 1996, nr 4, s. 538; F. Hayes-Renshaw, H. Wallace, *The Council of Ministers*, Basingstoke 1997, s. 40, 78.

danym aktem prawnym³⁰. W tym kontekście pojawia się pytanie badawcze, jakie czynniki powodują powyższe zaangażowanie ministrów.

W celu rozstrzygnięcia tej kwestii zostanie przeprowadzona regresja logistyczna. Pierwszym etapem jest postawienie hipotez badawczych określających wpływ czynników (zmiennych niezależnych) na zaangażowanie ministrów (zmienna zależna). Powinny one logicznie wynikać z założeń określonych teorii lub programów badawczych, np. z racjonalizmu czy konstruktywizmu. Jednak z racji ograniczeń ilościowych artykułu, problem ten zostanie pominięty, toteż hipotezy będą miały charakter własny, intuicyjny, bez szerszego rodowodu teoretycznego. Nie wnikając zatem w uzasadnienie, postawiono następujące hipotezy:

H₁: Zaangażowanie ministrów w proces decyzyjny jest bardziej prawdopodobne, jeżeli głosowanie nad projektem legislacyjnym opiera się na zasadzie jednomyślności.

H₂: Zaangażowanie ministrów w proces decyzyjny jest bardziej prawdopodobne, jeżeli projekt jest przyjmowany w ramach współdecydowania/zwykłej procedury ustawodawczej.

H₃: Zaangażowanie ministrów w proces decyzyjny jest bardziej prawdopodobne, jeżeli negocjowany projekt legislacyjny zawiera klauzulę komitologiczną³¹.

H₄: Zaangażowanie ministrów w proces decyzyjny jest bardziej prawdopodobne, jeżeli projekt jest przyjmowany po rozszerzeniu UE z 1 maja 2004 r.

Powyższe hipotezy tworzą model regresji logistycznej, który, bazując na (1.4.), można wyrazić w sposób następujący:

$$P(Z) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \text{Sposób_głosowania} + \beta_2 \text{Procedura_legislacyjna} + \beta_3 \text{Komitologia} + \beta_4 \text{Rozszerzenie})}}$$

gdzie:

$P(Z)$ – prawdopodobieństwo zaangażowania ministrów (zmienna binarna).

Następnie należy postawić hipotezy zerowe, stanowiące negację powyższych przypuszczeń. Zakładają one kolejno brak istotnego wpływu jednomyślności, współdecydowania, komitologii i wschodniego rozszerzenia na zaangażowanie ministrów.

W drugim etapie ma miejsce testowanie hipotez zerowych. W tym celu należy zbudować i zakodować odpowiednią próbę losową z populacji (bazę danych), na której będą przeprowadzone testy statystyczne. W analizie zostanie wykorzystana ogólnodostępna baza zaprojektowana przez Pierpaolo Settembri'ego, która zawiera dane na temat wszystkich aktów prawnych

³⁰ A. Kirpsza, *Rada Ministrów bez ministrów? Wpływ struktury organizacyjnej Rady Unii Europejskiej na kształt procesu podejmowania decyzji: perspektywa konstruktywizmu społecznego*, „Studia Europejskie” 2011, nr 4, s. 14.

³¹ Jest to każdy przepis, który umożliwia przyjmowanie w ramach procedury komitologii aktów wykonawczych do danego aktu prawnego.

(rozporządzeń, decyzji, dyrektyw, ale także instrumentów dawnego II i III filara) uchwalonych przez Radę (samodzielnie lub wspólnie z Parlamentem Europejskim) w okresie od 1 stycznia do 31 grudnia 2003 r. oraz od 1 lipca 2005 r. do 30 czerwca 2006 r.³² Luka czasowa między 2003 a 2005 r. została wprowadzona, aby lepiej uchwycić efekt wschodniego rozszerzenia UE z 1 maja 2004 r. W celu eliminacji aktów nielegislacyjnych, to jest instrumentów II i III filara, o których informacje są niekompletne i które w większości przypadków były przyjmowane w procedurach innych niż prawodawcze, baza danych została zredukowana tylko do projektów zaproponowanych przez Komisję Europejską. W rezultacie, uzyskano 659 regulacji, z których zostało dalej wyłączonych 192 aktów z powodu braku danych o choćby jednej ze zmiennych zawartych w hipotezach.

W tak skonstruowanej próbie dokonano kodowania predyktorów. Zmienna zależna, „Zaangażowanie ministrów” jest dychotomiczna i wynosi 1, gdy przed przyjęciem danego aktu prawnego w Radzie ministrowie debatowali nad projektem lub przyjęli polityczne porozumienie (choćby jeden punkt B w toku całej procedury) albo 0, gdy dana regulacja nie była przed jej formalnym uchwaleniem przedmiotem debaty lub takiego porozumienia (brak punktów B). Zmienna niezależna o nazwie „Sposób głosowania” posiada wartość dychotomiczną i jest równa 1, gdy do przyjęcia danego aktu prawnego była wymagana jednomyślność albo 0, gdy został on uchwalony w Radzie kwalifikowaną większością głosów. Zmienna niezależna „Procedura legislacyjna” jest także binarna i wynosi 1, gdy dany akt prawny został uchwalony w ramach współdecydowania albo 0, jeżeli był przyjęty w ramach innej procedury (np. konsultacji, zgody, współpracy). Zmienna niezależna „Komitologia” posiada wartość dychotomiczną i jest równa 1, gdy w danym akcie prawnym była zawarta klauzula komitologiczna albo 0, gdy takiej klauzuli nie było. Zmienna niezależna „Rozszerzenie” jest także dychotomiczna i wynosi 1, jeżeli dany akt prawny został uchwalony po 1 maja 2004 r. albo 0, gdy uchwalono go przed tą datą. Skonstruowano tylko jeden model regresji logistycznej zawierający wszystkie powyższe zmienne.

Po wyborze próby i jej zakodowaniu należy przejść do obliczeń mających na celu weryfikację modelu. Oznacza to przeprowadzenie testów w trzech obszarach: jakości dopasowania modelu, istotności poszczególnych hipotez i mocy predykcyjnej teorii. Rozpoczynając od pierwszego, należy najpierw dokonać testu współliniowości. Tabela nr 2 przedstawia krzyżowe korelacje między predyktorami, włączając w to wyraz wolny, oraz wskaźnik *VIF*. Można dojść do wniosku, że zależności są słabe, gdyż nie przekraczają granicy 0,7. Ponadto, *VIF* dla wszystkich zmiennych jest poniżej krytycznej wartości ($VIF < 3$). Można zatem wykluczyć niepożądane zjawisko współliniowości.

³² Vide P. Settembri, *The surgery succeeded. Has the patient died? The impact of enlargement on the European Union*, „Jean Monnet Working Paper” 2004, nr 7. Baza jest dostępna pod adresem: <http://www.councildata.cergu.gu.se/cdeu/index.php/research-data/settembri>.

Tabela nr 2. Korelacje między zmiennymi niezależnymi

	Stała	Sposób głosowania	Procedura legislacyjna	Komitologia	Rozszerzenie UE	VIF
Stała	1,000	-0,602	0,514	-0,497	-0,409	-
Sposób głosowania		1,000	0,433	0,237	0,022	1,146
Procedura legislacyjna			1,000	-0,097	-0,083	1,201
Komitologia				1,000	0,113	1,149
Rozszerzenie					1,000	1,027

Źródło: opracowanie własne na podstawie bazy danych P. Settembri'ego.

Tabela nr 3 obrazuje wyniki trzech kolejnych testów jakości dopasowania modelu: ilorazu wiarygodności, Hosmera-Lemeshowa oraz zbieżności. Wartość statystyki G dla dewiancji wynosi 81,536, podczas gdy $\chi^2_{krytyczne}$ dla $\alpha = 0,05$ i siedmiu stopni swobody jest równe 14,067, co oznacza, że wynik testu jest istotny statystycznie nawet na poziomie 0,01 ($\chi^2_{krytyczne}$ dla 0,01=18,4765). Można zatem odrzucić H_0 i przyjąć, że model jest zdecydowanie lepszy od losowego. Idąc dalej, wynik testu H-L nie jest istotny na poziomie 0,05, gdyż $\chi^2_{krytyczne}$ dla α i czterech stopni swobody wynosi 9,4877. Oznacza to, że nie ma podstaw do odrzucenia H_0 , czyli model jest dobrze dopasowany. Natomiast współczynnik determinacji pokazuje, że ok. 23% wariacji zmiennej zależnej jest wyjaśniana przez wariację zmiennych niezależnych. Jak pokazano powyżej, wartość R^2 w regresji logistycznej nie jest wysoka, należy jednak uznać powyższy wynik za niesatysfakcjonujący. Oznacza on bowiem, że istnieje jeszcze duży procent wariacji, który wyjaśniają inne predyktory nieujęte w modelu. Reasumując, teoria zdała test jakości dopasowania, gdyż jest lepsza od modelu losowego i podaje nowe wyjaśnienie zachowania zmiennej zależnej, choć tylko w małym zakresie rzeczywistości.

Tabela nr 3. Współczynniki jakości dopasowania modelu

	-2log(L)	Chi-kwadrat	Test Hosmera-Lemeshow'a	R^2 Nagelkerke
Wartość współczynnika	476,460	81,536***	4,660	0,230 (23%)
Poziom istotności – p	-	0,000	0,701	-
Stopnie swobody	-	7	4	-

Źródło: opracowanie własne na podstawie bazy danych P. Settembri'ego.

Po teście jakości dopasowania modelu można przejść do weryfikacji poszczególnych hipotez. Tabela nr 4 przedstawia współczynniki regresji, błędy standardowe, wyniki testów Walda, stopnie swobody, ilorazy szans i przedziały ufności. Zmienna „Sposób głosowania” jest istotna statystycznie na najwyższym poziomie ($p < 0,01$), co pozwala odrzucić H_0 o braku wpływu tego predyktora na zmienną zależną i przyjąć za prawdziwą hipotezę alternatywną ($\chi^2_{krytyczne}$ dla α przy jednym stopniu swobody jest równe 3,8415, a dla 0,01 – 6,6349). Iloraz szans wynosi 3,304, czyli jest większy od 1, świadcząc o stymulującym i dodatnim wpływie zasady głosowania na zaangażowanie ministrów. Wartość tę można zinterpretować następująco: szansa, że ministrowie będą debatować nad aktem legislacyjnym przed jego formalnym przyjęciem jest trzy razy większa (3 do 1) w przypadku projektów przyjmowanych w ramach jednomyślności niż uchwalanych w oparciu o regułę kwalifikowanej większości głosów. Przedziały ufności omawianej zmiennej mówią, że z prawdopodobieństwem 0,95 (95%) można oczekiwać, że iloraz szans zaangażowania ministrów w przypadku projektów przyjmowanych jednomyślnie nie będzie niższy niż 1,783 i wyższy niż 6,112. Podobna interpretacja dotyczy dwóch kolejnych predyktorów. Zmienne „Procedura legislacyjna” i „Komitologia” są istotne statystycznie i dodatnie, co oznacza odrzucenie hipotez zerowych, tym samym potwierdzenie hipotez alternatywnych nr 2 i 3. Gdy dany akt prawny jest uchwalany w ramach współdecydowania/ZPU, to szansa zaangażowania ministrów zwiększa się aż 4,7-krotnie w porównaniu do innych procedur legislacyjnych, przy czym maksymalnie może wzrosnąć nawet 7,8-krotnie (zob. przedział ufności). Natomiast jeśli negocjowany projekt zawiera delegację komitologiczną, to szansa udziału ministrów w procesie decyzyjnym jest ponad trzykrotnie, a w przypadkach maksymalnych nawet pięciokrotnie, większa niż w sytuacji braku takiej klauzuli. Odmienna interpretacja dotyczy jednak zmiennej „Rozszerzenie”. Jest ona istotna statystycznie tylko na poziomie $0,05 < p < 0,1$, tymczasem w niniejszej analizie jako α przyjęto 0,05, dlatego wynik ten nie jest zgodny z przewidywaniami modelu. Nie ma zatem podstaw do odrzucenia H_0 o braku wpływu rozszerzenia na zaangażowanie ministrów. Ponadto, współczynnik omawianej zmiennej jest ujemny, co oznacza, że udział ministrów jest bardziej prawdopodobny przed rozszerzeniem UE, co ponownie stoi w sprzeczności z hipotezą nr 4. Wreszcie, wartość ilorazu szans jest poniżej zera, co świadczy o destymulującym wpływie tej zmiennej, inaczej mówiąc, szansa na zaangażowanie ministrów w przypadku projektów uchwalonych po 1 maja 2004 r. jest o 0,646 (między 0,412 a 1,020) mniejsza niż przed tą datą. Reasumując, regresja logistyczna wykazała prawdziwość trzech z czterech hipotez, a badany model przyjmuje ostateczną postać:

$$P(\text{zaangażowanie ministrów}) = \frac{1}{1 + e^{-z}}$$

gdzie:

$$Z = - 2.008 + 1,195 \textit{ Spos\o{b_g\l{o}sowania} + 1,555 \textit{ Procedura_legislacyjna} + 1,116 \textit{ Komitologia} - 0,433 \textit{ Rozszerzenie}$$

Tabela nr 4. Wyniki regresji logistycznej

Zmienna zależna: zaangażowanie ministrów (1 – polityczne porozumienie, 0 – brak politycznego porozumienia)							
Zmienne niezależne	β	S.E.	Statystyka Walda	Stopnie swobody	Ilo-raszans	Przedział ufności (95% CI)	
						Dolna granica	Górna granica
Sposób głosowania	1,195***	0,315	14,419	1	3,304	1,783	6,112
Procedura legislacyjna	1,555***	0,260	35,731	1	4,734	2,843	7,882
Komitologia	1,116***	0,242	21,335	1	3,054	1,902	4,904
Rozszerzenie	-0,433*	0,231	3,517	1	0,648	0,412	1,020
Stała, wyraz wolny	-2,008***	0,244	67,790	1	0,134		

Poziomy istotności: *** – $p < 0,01$; ** – $p < 0,05$; * – $p < 0,1$. Współczynniki istotne statystycznie zostały pogrubione. Poziomy istotności zostały obliczone poprzez porównanie otrzymanych wartości statystyki Walda i wartości krytycznych rozkładu chi-kwadrat dla $p=0,1$, $0,5$ i $0,01$.

Po weryfikacji hipotez można przejść do analizy mocy predykcyjnej modelu. Tabela nr 5 obrazuje zdolności klasyfikacyjne teorii. Celność modelu (ACC) przy punkcie odcięcia 0,5 wynosi 75,6% [100(55+298/467)%] i w porównaniu do modelu tylko z wyrazem wolnym (71,5%) jest nieco wyższa, co oznacza, że posiada on wartość dodaną. Inne współczynniki, to jest czułość i specyficzność, wynoszą kolejno: 41,35% [100(55/133)]% i 89,2% [100(298/334)%]. Model sklasyfikował zatem poprawnie 76% klas decyzyjnych, z czego prawidłowo pogrupowanych przypadków pozytywnych jest 41,35%, a negatywnych – 89,2%. Widać zatem, że o ile teoria potrafi bardzo dobrze prognozować akty prawne, w odniesieniu do których zaangażowanie ministrów nie będzie miało miejsca, to ma problem z predykcją projektów stanowiących szczególne zainteresowanie ministrów. W tym kontekście, jego pozytywna wartość predykcyjna jest słaba.

Tabela nr 5. Tabela klasyfikacyjna

		Prognozowane zaangażowanie ministrów		
		Nie (brak udziału)	Tak (udział)	Suma
Zaobserwowane zaangażowanie ministrów	Nie (brak udziału)	298	36	334
	Tak (udział)	78	55	133
	Suma	376	91	467

Źródło: opracowanie własne na podstawie bazy danych P. Settembri'ego.

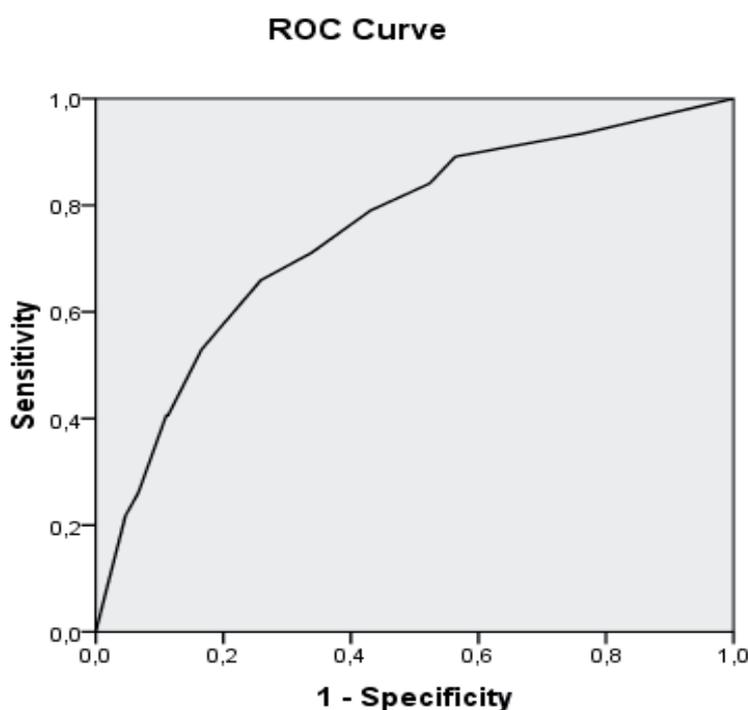
Na podstawie czułości i specyficzności skonstruowano krzywą ROC, co pokazuje Wykres nr 2. Pole powierzchni pod krzywą (*AUC*) wyniosło **0,751***** i jest to wynik istotnie większy od $\alpha = 0,05$. Oznacza to, że model klasyfikuje przypadki znacznie lepiej niż model oparty na losowym zgadywaniu. Ponadto, można stwierdzić, że w 75,1% wszystkich możliwych par aktów prawnych zaproponowanych przez Komisję, z których jeden cechował się zaangażowaniem ministrów (1), a drugi nie (0), model przypisze większe prawdopodobieństwo dla aktu prawnego przyjętego z udziałem ministrów. Tabela nr 6 pokazuje również wartości *C* i *1-S* dla określonych punktów odcięcia prawdopodobieństwa. Chcąc polepszyć słabą pozytywną moc predykcyjną modelu wynikającą z tabeli klasyfikacyjnej (punkt odcięcia – 0,5), można wybrać najbardziej odpowiedni punkt odcięcia na krzywej ROC, który stanowiłby optymalną podstawę dla prognoz modelu. Chodzi tu o taki punkt, który jest najbliższy miejscu (0,1) na wykresie, zawierający pożądaną niską wartość *1-S* i wysoką *C*. Tabela nr 6 jak i Wykres nr 2 pokazują, że tym punktem jest 0,299, dla którego czułość wynosiłaby już 0,659 (66%), a specyficzność 0,741 (1-0,259). Reasumując, choć wartość *AUC* nie jest bardzo wysoka, to należy uznać, że zdolności klasyfikacyjne modelu są akceptowalne, a przyjęcie punktu odcięcia prawdopodobieństwa 0,299 zamiast 0,5 pozwala zwiększyć możliwości predykcyjne modelu.

Tabela nr 6. Punkty odcięcia prawdopodobieństwa na krzywej ROC

Punkty odcięcia	0	0,099	0,164	0,216	0,257	0,291	0,299	0,347	0,428	0,512	0,566	0,617	1
<i>C</i>	1	0,935	0,891	0,841	0,790	0,710	0,659	0,529	0,406	0,406	0,261	0,217	0
<i>1-S</i>	1	0,765	0,564	0,523	0,430	0,337	0,259	0,166	0,113	0,110	0,067	0,047	0

Źródło: opracowanie własne na podstawie bazy danych P. Settembri'ego.

Wykres nr 2. Krzywa ROC obliczona w programie SPSS



Diagonal segments are produced by ties.

Źródło: opracowanie własne na podstawie bazy danych P. Settembri'ego.

Konkluzje

Metody statystyczne są ciekawym narzędziem analizy rzeczywistości. Choć, co do zasady, są wykorzystywane w badaniach medycznych, biostatystyce czy socjologii, to równie dobrze mogą być eksploatowane w naukach o polityce, w szczególności w studiach nad Unią Europejską. Dowodzi tego niniejszy artykuł, w którym pokazano, jak przy użyciu jednej z nich – regresji logistycznej – budować i weryfikować modele teoretyczne definiujące wpływ określonych zmiennych na proces podejmowania decyzji w UE. Oczywiście, nie należy zapominać, że przedmiot nauk politycznych jest tak skomplikowany, że jego wyjaśnianie za pomocą jednej z metod ilościowych jest niewystarczające, tym bardziej, że są one obciążone ryzykiem błędu. Rozwiązaniem powyższego problemu może być tylko przyjęcie procedury triangulacyjnej, czyli weryfikowanie hipotez za pomocą kilku metod ilościowych lub jakościowych jednocześnie. Nie zmienia to jednak faktu, że regresja logistyczna jako procedura statystyczna pozwala w dużym stopniu nie tylko wyjaśniać zjawiska polityczne, które już miały miejsce, ale także prognozować ich kształt.

Z tego też powodu częstsze stosowanie metod probabilistycznych przez polskich badaczy Unii Europejskiej i politologów stanowiłoby poważną wartość dodaną. Po pierwsze, otworzyłoby wciąż tkwiącym na poziomie opisu naukom

o polityce drzwi do wyższych poziomów nauki, to jest eksplanacji i predykcji. Obecnie, stosowanie tych kategorii odbywa się tylko na poziomie teoretycznym, w przyszłości mogłoby nastąpić także na gruncie empirycznym. Po drugie, porzucenie awersji do statystyki zmusiłoby badaczy polityki do przyjęcia metody hipotetycznej, to jest najpierw stawiania hipotez, potem ich empirycznego uprawdopodobnienia poprzez wnioskowanie nie wprost, a następnie przenoszenia wyników na populację. Biorąc pod uwagę fakt, że współczesna polska politologia ciągle opiera się na opisie i formułowaniu przypuszczeń nieweryfikowanych lub trudnych do sprawdzenia, byłby to znaczny postęp. Po trzecie wreszcie, spojrzenie łaskawszym okiem na statystykę pozwoliłoby zbudować grunt dla przyjęcia kwantowego paradygmatu nauk społecznych, bazującego właśnie na kategoriach probabilistycznych. Byłby to ważny krok prowadzący do wyciągnięcia nauk politycznych z kryzysu, w jakim tkwią obecnie. Biorąc pod uwagę fakt, że ogólny przedmiot tej dyscypliny – interakcje polityczne między ludźmi i ich efekty – jest chaotyczny, ciężko definiowalny i trudny do intersubiektywnego zmierzenia, a więc podobny do praw mechaniki kwantowej, to wprowadzenie do niej kategorii fizyki nieklasycznej skonstruowałoby silne podstawy eksplanacyjno-predykcyjne.