VOL. 55 STUDIA POLITOLOGICZNE
DOI: 10.33896/SPolit.2020.55.19

*Mateusz Wajzer**
*Monika Cukier-Syguła***

Analiza regresji z programem R – przykład użycia w badaniach politologicznych

Regression Analysis in Political Science Research Using the R Program

Słowa kluczowe: *analiza regresji, modele liniowe, modele nieliniowe, program R, politologia*

Keywords: *regression analysis, linear models, non-linear models, R program, political science*

Abstrakt: *Celem artykułu jest prezentacja podstawowych możliwości programu R użytego do budowy regresyjnych modeli opisujących zjawiska polityczne. W analizach wykorzystano zbiór danych ukazujących proces kształtowania się poziomu frekwencji wyborczej w wyborach do Kongresu Stanów Zjednoczonych w 2014 roku w zależności od wieku wyborców. Zastosowane procedury statystyczne (zbudowano dwa modele: liniowy oraz wielomianowy stopnia drugiego) szczegółowo omówiono, podając ścieżki odpowiadających im poleceń. Adresatami artykułu są głównie studenci i doktoranci politologii oraz dyscyplin pokrewnych, a także badacze, którzy nie pracowali wcześniej z programem R.*

Abstract: *The aim of the article is to present the basic functionalities of the R program for the creation of regression models describing political phenomena. A database of voter turnout during the 2014 U.S. Congress elections categorised according to voters' age was used for the analyses. The statistical procedures (linear and second-degree polynomial models) applied*

* ORCID ID: https://orcid.org/0000-0002-3108-883X, doktor, Instytut Nauk Politycznych, Uniwersytet Śląski w Katowicach. Zainteresowania badawcze: teoria gier, metodologia badań politologicznych, ewolucja społeczna. Email: mateusz.wajzer@us.edu.pl.

** ORCID ID: https://orcid.org/0000-0001-6211-3500, doktorantka, Wydział Nauk Społecznych, Uniwersytet Śląski w Katowicach. Zainteresowania naukowe: teorie demokracji, teorie elit, prawo samorządu terytorialnego. Email: monika.cukiersygula@onet.pl.

were discussed in detail, with paths to their respective commands being provided. The article is addressed primarily to postgraduate students in political science and related disciplines, as well as to researchers who have never used the R program before.

Wstęp

Termin „regresja" został wprowadzony do statystyki przez Francisa Galtona. W artykule „Regression towards mediocrity in hereditary stature" (1886) zaprezentował on wyniki badań nad dziedziczeniem wzrostu. Doszedł do wniosku, że dzieci bardzo wysokich rodziców są średnio od rodziców niższe, i na odwrót – dzieci bardzo niskich rodziców są średnio od rodziców wyższe. Zidentyfikowaną tendencję do dążenia do wartości średniej Galton nazwał regresją. Terminu tego nie używa się już w znaczeniu nadanym mu przez brytyjskiego polihistora. Obecnie analiza regresji stanowi jedną z najczęściej stosowanych metod w budowie modeli statystycznych. Umożliwia ona badanie zależności między zmienną objaśnianą (y) a zmiennymi objaśniającymi (x), predykcję wartości zmiennej objaśnianej przy zadanych wartościach zmiennych objaśniających oraz wskazanie wartości zmiennych istotnie wpływających na zmienną objaśnianą. Przyjmując jako kryterium podziału postać funkcji $f(x, \beta)$, możemy wyróżnić regresję liniową oraz regresję nieliniową. Do najczęściej stosowanych modeli regresyjnych zalicza się te opisujące zależności liniowe. Co jednak, gdy związki między cechami interesującymi badacza nie mają charakteru liniowego? W takich sytuacjach do danych empirycznych dopasowywana jest krzywa najlepiej opisująca obserwowany rozrzut wartości. Modele nieliniowe budowane są zatem na potrzeby konkretnych przypadków z zastosowaniem odpowiednio dobranych funkcji matematycznych[1].

Mając na uwadze powyższą kategoryzację, omówiono dwa modele: liniowy oraz nieliniowy – wielomianowy stopnia drugiego (kwadratowy). Analizom poddano przykładowy zbiór danych dotyczących frekwencji wyborczej w wyborach do Kongresu Stanów Zjednoczonych z 2014 roku. W analizach użyto programu R[2].

[1] A. Sen, M. Srivastava, *Regression Analysis: Theory, Methods, and Applications*, New York 1990, s. 1–19.

[2] W języku polskim wybrane funkcje statystyczne R zostały omówione w następujących pracach: Ł. Komsta, *Wprowadzenie do środowiska R*, dokument w formacie pdf., 2004; T. Górecki, *Podstawy statystyki z przykładami w R*, Legionowo 2011; M. Walesiak, E. Gatnar (red.), *Statystyczna analiza danych z wykorzystaniem programu R*, Warszawa 2012;

Podstawowe informacje o programie R

R jest programem stosowanym głównie do obliczeń statystycznych i wizualizacji danych. Jego twórcami są Robert Gentleman i Ross Ihaka, statystycy z Uniwersytetu w Auckland. Od 1997 roku w rozwój projektu R jest zaangażowany międzynarodowy zespół programistów (R Core Team) wspierany finansowo przez Fundację R (ang. *The R Foundation for Statistical Computing*). Wkład w ciągłe usprawnianie działania programu oraz poszerzanie jego możliwości wnoszą także sami użytkownicy, tworząc biblioteki (pakiety) używane w wielu obszarach[3]. We wrześniu 2018 roku w repozytorium CRAN (ang. *The Comprehensive R Archive Network*) było dostępnych ponad 13 000 bibliotek. Pierwsza oficjalna wersja R pojawiła się w roku 2000 i od tego czasu program jest stale aktualizowany.

Do kluczowych cech programu R zaliczyć należy:

- niekomercyjność – R jest dostępny w ramach Powszechnej Licencji Publicznej GNU (ang. *GNU General Public License*, GNU GPL). W praktyce oznacza to, że użytkownicy mają swobodę w używaniu oprogramowania, w analizowaniu jego działania, mogą rozpowszechniać jego kopie (nieodpłatnie) oraz dokonywać dowolnych modyfikacji;
- funkcjonalność – R umożliwia tworzenie wielu nowych pakietów o zastosowaniach wykraczających poza standardowe obliczenia statystyczne. Przy jego użyciu możemy między innymi budować modele klasycznej (games, hop, GameTheory) i ewolucyjnej teorii gier (EvolutionaryGames), prowadzić analizę treści tweetów (twitteR, ROAuth, tm, topicmodels, sentimentr, ggplot2, igraph, wordcloud) czy generować raporty (knitr);
- zgodność – użytkownicy uzyskują dostęp do obiektów R z kodów takich języków, jak C lub C++.

P. Biecek, *Analiza danych z programem R. Modele liniowe z efektami stałymi, losowymi i mieszanymi*, Warszawa 2013; P. Biecek, *Przewodnik po pakiecie R*, Wrocław 2014; J. Brzezińska, *Analiza logarytmiczno-liniowa. Teoria i zastosowania z wykorzystaniem programu R*, Warszawa 2015. Marek Gągolewski napisał natomiast książkę poświęconą programowaniu w języku R. Zob. M. Gągolewski, *Programowanie w języku R. Analiza danych, obliczenia, symulacje*, Warszawa 2016.

[3] Użytkownicy R tworzą dobrze zintegrowane środowisko, czego najlepszymi dowodami są periodycznie organizowane konferencje (https://www.r-project.org/conferences.html) oraz aktywność w blogosferze (https://www.r-bloggers.com/).

Pliki instalacyjne R dla systemów Windows, Linux, Mac OS X można pobrać z serwerów, których adresy odnajdziemy na stronie: https://cran.r-project.org/mirrors.html. Pracę z R znacznie usprawnia bezpłatny edytor RStudio również dostępny dla systemów Windows, Linux i Mac OS X. RStudio pobieramy ze strony: https://www.rstudio.com/. Do analiz zaprezentowanych poniżej użyto programu R w wersji 3.4.2 wspomaganego edytorem RStudio w wersji 1.0.143.

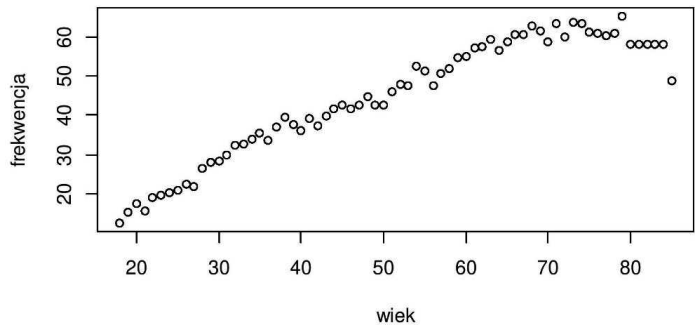
Przykład: zależność między wiekiem a poziomem frekwencji wyborczej

Odwołajmy się do danych dotyczących frekwencji wyborczej w wyborach do Kongresu Stanów Zjednoczonych z 2014 roku. Zostały one zgromadzone oraz udostępnione przez United States Census Bureau (USCB) z siedzibą w Suitland, Maryland (http://census.gov/data/tables/time-series/demo/voting-and-registration/p20-577.html). Dysponujemy zbiorem 86 obserwacji dla dwóch zmiennych ilościowych: wieku wyborców oraz poziomu frekwencji wyborczej. Sprawdźmy, czy między zmiennymi występuje zależność, tzn. czy wraz z wiekiem zarysowuje się jakakolwiek tendencja uczestnictwa wyborczego. Do graficznej prezentacji danych użyjemy polecenia

```
> plot(wiek, frekwencja)
```

wpisanego bezpośrednio w linii komend konsoli R. Funkcja plot() jest funkcją graficzną pakietu graphics[4].

Rysunek 1. Ilustracja zbioru danych dotyczących wieku oraz poziomu frekwencji wyborczej w wyborach do Kongresu Stanów Zjednoczonych z 2014 roku



Na osi odciętych zaznaczono wiek wyborców, na osi rzędnych – poziom frekwencji wyborczej w procentach.

Źródło: opracowanie własne.

[4] Podstawowy pakiet graficzny ładowany po uruchomieniu programu.

Już na podstawie pobieżnej analizy powyższego rysunku możemy zaryzykować tezę, że zależność między rozpatrywanymi cechami lepiej niż liniowa funkcja trendu opisywać będzie kwadratowa funkcja trendu. Jednak w pierwszej kolejności dopasujemy do danych model regresji liniowej

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{1}$$

gdzie y jest wektorem zawierającym wartości frekwencji wyborczej, x to wektor wartości odpowiadających wiekowi wyborców, β_0 i β_1 to nieznane współczynniki, zaś ε stanowi wektor składników losowych.

Przy użyciu poniższych poleceń wyznaczymy oceny współczynników modelu liniowego.

```
> model1 <- lm(frekwencja ~ wiek, data = kongres)
> summary(model1)
```

```
Call:
lm(formula = frekwencja ~ wiek, data = kongres)

Residuals:
   Min      1Q  Median      3Q     Max
–19.995  –2.749   1.195   2.979   6.500

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.09816    1.65218   4.296  5.84e-05 ***
wiek         0.72702    0.02998  24.252   < 2e–16 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.852 on 66 degrees of freedom
Multiple R-squared: 0.8991,     Adjusted R-squared: 0.8976
F-statistic: 588.2 on 1 and 66 DF, p-value: < 2.2e–16
```

Otrzymano następujące oceny współczynników: $\hat{\beta}_0 = 7{,}09816$ i $\hat{\beta}_1 = 0{,}72702$. Po podstawieniu ich do modelu opisanego równaniem (1) uzyskujemy następującą formułę do liniowej predykcji poziomu frekwencji wyborczej

STUDIA I ANALIZY / SP Vol. 55 400

$$\hat{y} = 7,09816 + 0,72702x. \qquad (2)$$

W *R* wartości teoretyczne zmiennej objaśnianej można uzyskać dzięki funkcji predict().

Rozważmy teraz model kwadratowy

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon, \qquad (3)$$

wyznaczając następnie oceny poszczególnych współczynników regresji. Do tego celu użyjemy komend

```
> model2 <- lm(frekwencja ~ wiek + I(wiek^2), data = kongres)
> summary(model2)
```

```
Call:
lm(formula = frekwencja ~ wiek + I(wiek^2), data = kongres)

Residuals:
    Min        1Q      Median     3Q       Max
–11.6156   –1.9355    0.3952   1.6889   4.9853

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
(Intercept)   –1.868e+01   2.432e+00   –7.678    1.09e–10 ***
wiek           1.898e+00   1.031e-01   18.407     < 2e–16 ***
I(wiek^2)     –1.137e-02   9.869e-04  –11.521     < 2e–16 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.803 on 65 degrees of freedom
Multiple R-squared: 0.9668,     Adjusted R-squared: 0.9658
F-statistic: 947.4 on 2 and 65 DF, p-value: < 2.2e–16
```

Poszukiwane parametry przyjęły wartości: $\hat{\beta}_0 = -18,676$, $\hat{\beta}_1 = 1,898$, $\hat{\beta}_2 = -0,01137$. Podstawiając je do równania (3), otrzymujemy model w postaci:

$$\hat{y} = -18,676 + 1,898x - 0,01137x^2. \qquad (4)$$

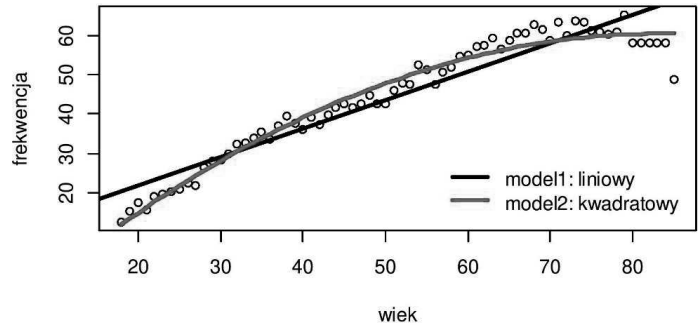
Rysunek 2 ilustruje dopasowanie modeli (2) i (4) do danych. Linie regresji (graficzne odwzorowanie modeli) zostały dodane przy użyciu następujących poleceń

```
> abline(model1, lwd = 3)
> lines(smooth.spline(wiek, predict(model2)), col = "red", lwd = 3)
```

natomiast legenda

```
> legend(56, 30, legend = c("model1: liniowy", "model2: kwadratowy"),
col = c("black", "red"), lty = c("solid", "solid"), box.lty = "blank", lwd = 3)
```

Rysunek 2. Modele opisujące zależność między wiekiem a poziomem frekwencji wyborczej na przykładzie wyborów do Kongresu Stanów Zjednoczonych z 2014 roku



Źródło: opracowanie własne.

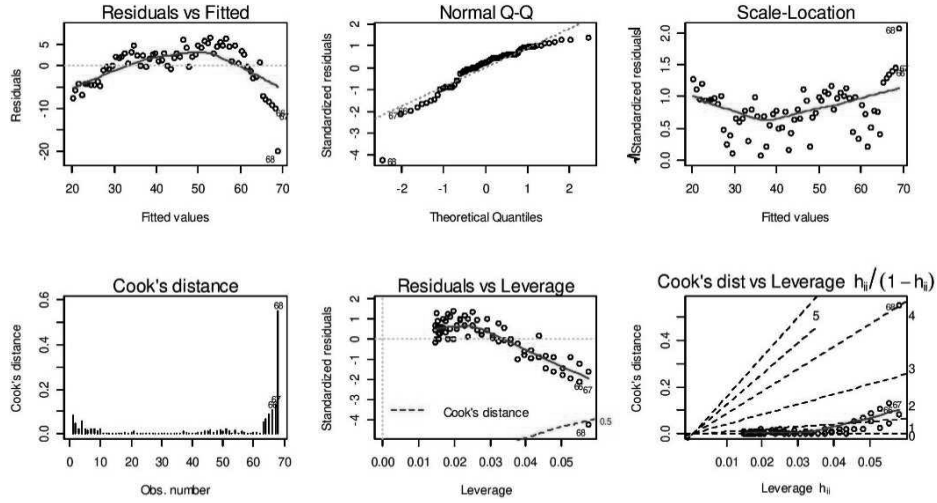
Oceniając jakość dopasowania modeli do danych można się kierować różnymi kryteriami. Jedno z nich stanowi porównanie wartości współczynnika R^2. W modelu liniowym wyniosła ona 0,8991, co oznacza, że ok. 90% wariancji zmiennej objaśnianej jest wyjaśnione przez model, zaś w modelu kwadratowym – 0,9668, co oznacza, że ok. 97% wariancji zmiennej objaśnianej wyjaśnia model. Biorąc pod uwagę, że modele nie różnią się liczbą predyktorów wyższa wartość współczynnika determinacji wskazuje na model lepiej dopasowany. Konstatację tę potwierdza interpretacja sześciu wykresów diagnostycznych wygenerowanych dla każdego modelu z zastosowaniem komend

```
> par(mfrow = c(2, 3), pty = "m")
> plot(model1, which = 1:6)
> par(mfrow = c(2, 3), pty = "m")
> plot(model2, which = 1:6)
```

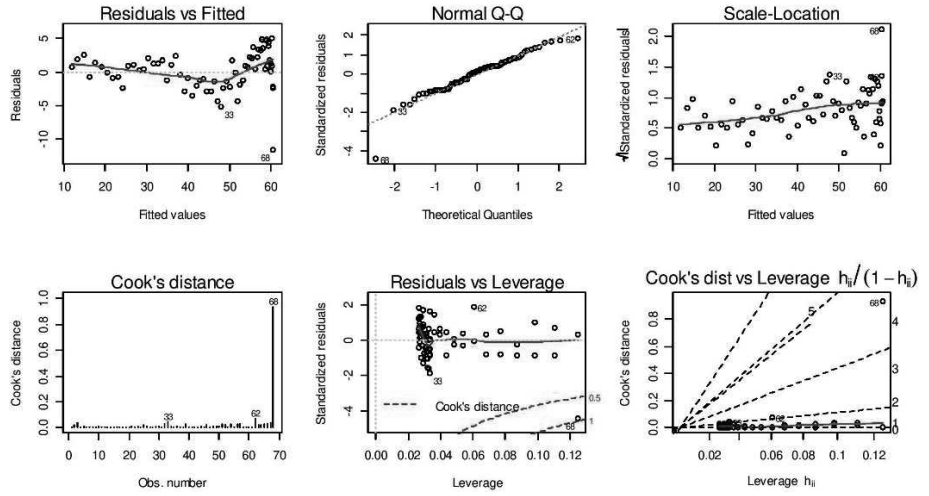

STUDIA I ANALIZY / SP Vol. 55 402

Rysunek 3. Wykresy diagnostyczne dla modelu liniowego (a) i dla modelu kwadratowego (b)

(a)



(b)



Źródło: opracowanie własne.

Analiza statystyk graficznych pokazuje, że w przypadku modelu liniowego występuje wyższa zależność funkcyjna reszt od zmiennej objaśnianej. Świadczy to o gorszym dopasowaniu modelu.

Inną metodę oceny jakości dopasowania modeli oferuje częściowy test F implementowany w R

```
> anova(model1, model2)
```

Oto jego wyniki:

```
Analysis of Variance Table

Model 1: frekwencja ~  wiek
Model 2: frekwencja ~ wiek + I(wiek^2)
 Res.Df    RSS Df Sum  of Sq    F Pr(>F)
1    66 1553.8
2    65 510.8 1    1043 132.72 < 2.2e–16 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Wartość statystyki częściowego testu F wynosi 132,72, a odpowiadająca jej *p*-wartość – $2{,}2 \cdot 10^{-16}$. Na ich podstawie możemy odrzucić hipotezę zerową o braku istotnych statystycznie różnic między modelami i przyjąć, że model kwadratowy jest istotnie lepiej dopasowany do danych niż model liniowy. Procedurę tę możemy oczywiście powtórzyć, uwzględniając w modelu kwadratowym kolejny element, tj. zmienną objaśniającą podniesioną do potęgi trzeciej.

Zakończenie

Modele matematyczne znajdują współcześnie coraz szersze zastosowanie, także w tych dziedzinach nauki, w których powątpiewano w ich przydatność. Tendencji tej bez wątpienia sprzyja pojawianie się nowych narzędzi analitycznych i doskonalenie już istniejących. Jednym z nich jest program R, którego przykładowe zastosowanie w budowie regresyjnych modeli opisujących zjawiska polityczne zaprezentowano w niniejszym artykule. Analizom poddano zbiór danych ukazujących proces kształtowania się poziomu frekwencji wyborczej w zależności od wieku wyborców. Dane pochodziły z wyborów do Kongresu Stanów Zjednoczonych, które odbyły się w 2014 roku. W toku analiz zbudowano dwa modele: liniowy oraz wielomianowy stopnia drugiego. Każdy krok, począwszy od budowy modeli po ich diagnostykę, został szczegółowo opisany wraz z podaniem składni właściwych komend. Za użyciem R przemawia kilka istotnych

STUDIA I ANALIZY / SP Vol. 55 404

przesłanek. Po pierwsze, R jest programem bezpłatnym, co może mieć znaczenie dla osób nieuprawiających nauki zawodowo oraz dla doktorantów i studentów. Po drugie, R znajduje zastosowanie w wielu często skrajnie odmiennych dziedzinach nauki. Po trzecie, język R odznacza się elastyczną składnią, co umożliwia tworzenie własnych funkcji. Ponadto w R można korzystać z funkcji napisanych w innych językach. Po czwarte wreszcie, zarówno R, jak i jego poszczególne biblioteki mają dobrze opracowaną dokumentację, która znacznie usprawnia pracę. W tym kontekście niezwykle ważne jest również doświadczenie, którym dzielą się użytkownicy programu na forach internetowych oraz w blogach.

Bibliografia

P. Biecek, *Analiza danych z programem R. Modele liniowe z efektami stałymi, losowymi i mieszanymi*, Warszawa 2013.

P. Biecek, *Przewodnik po pakiecie R*, Wrocław 2014.

J. Brzezińska, *Analiza logarytmiczno-liniowa. Teoria i zastosowania z wykorzystaniem programu R*, Warszawa 2015.

F. Galton, *Regression towards mediocrity in hereditary stature*, «Journal of the Anthropological Institute of Great Britain and Ireland» 1886, nr 15, s. 246–263. doi: 10.2307/2841583.

M. Gągolewski, *Programowanie w języku R. Analiza danych, obliczenia, symulacje*, Warszawa 2016.

T. Górecki, *Podstawy statystyki z przykładami w R*, Legionowo 2011.

Ł. Komsta, *Wprowadzenie do środowiska R*, 2004. https://cran.r-project.org/doc/contrib/Komsta-Wprowadzenie.pdf (8.09.2018).

R: Conferences. https://www.r-project.org/conferences.html (15.09.2018).

A. Sen, M. Srivastava, *Regression Analysis: Theory, Methods, and Applications*, New York 1990. doi: 10.1007/978-1-4612-4470-7.

M. Walesiak, E. Gatnar (red.), *Statystyczna analiza danych z wykorzystaniem programu R*, Warszawa 2012.

STUDIA I ANALIZY / SP Vol. 55 405