Kraków 2023

UNIWERSYTET JAGIELLOŃSKI
W KRAKOWIE

Wydział Matematyki i Informatyki

‘Analiza czynników wpływających na wynik rankingu w Międzynarodowej Olimpiadzie Matematycznej w 2021 roku przy użyciu modelu regresji w języku R.’

Inga Dyląg

Cel analizy

Celem analizy jest badanie czynników wpływających na sukcesy poszczególnych państw w dziedzinie matematyki. Zastanowimy się, które zmienne mają największy wpływ na umiejscowienie w rankingu krajów w międzynarodowej olimpiadzie matematycznej w 2021, a które ze zmiennych mają mniejsze lub żadne znaczenie w prognozowaniu wyników.

Opis danych

Dane wykorzystane w analizie pochodzą z baz danych: International Mathematical Olimpiad (imo-official.org), World Population Review (worldpopulationreview.com) oraz Human Development Reports (hdr.undp.org) i zostały przeze mnie połączone w jeden zbiór danych za pomocą narzędzia PowerQuery.

Zbiór danych składa się ze zmiennych ilościowych opisujących cechy danego kraju takie jak:

- ranking2021 – miejsce w rankingu w Międzynarodowej Olimpiadzie Matematycznej w 2021 roku
- pop2023 – liczba ludności na rok 2023
- growthRate – wskaźnik przyrostu naturalnego (%)
- landArea – powierzchnia kraju w km^2
- hdi – wskaźnik rozwoju społecznego, mierzący poziom edukacji, oczekiwaną długość życia i poziom dochodów na rok 2021
- education – średnia liczba lat edukacji w roku 2021
- gni – dochód narodowy brutto na mieszkańca w roku 2021 (USD)
- qol – wskaźnik jakości życia, mierzący poziom zadowolenia mieszkańców z różnych aspektów życia
- happiness – wskaźnik szczęścia, mierzący poziom subiektywnego szczęścia mieszkańców w roku 2021
- crime – wskaźnik mierzący poziom przestępczości (liczba przestępstw na 1000 mieszkańców)
- iq – średni poziom IQ mieszkańców na rok 2021

oraz zmiennej jakościowej:

- Europe – zmienna binarna, która przyjmuje wartość 1 dla krajów europejskich i 0 dla pozostałych krajów

Model liniowy

$$ranking2021 = 129.3 - 1.789 * 10^{-8} * \text{pop2023} + 72.89 * \text{growthRate} - 2.267 * 10^{-6} * landArea - 81.28 * hdi - 4.62 * education + 1.943 * 10^{-4} * gni + 0.1779 * qol + 13.81 * happiness + 0.01422 * crime - 0.8678 * iq + 4.136 * Europe$$

Po przeprowadzeniu wstępnej analizy modelu zostało odrzuconych 48 obserwacji ze względu na braki danych w nich zawarte. Po dokonaniu czyszczenia danych liczba obserwacji wyniosła 55, co przy 11 zmiennych daje 43 stopnie swobody.

Analiza współczynników

Współczynnik wyrazu wolnego równy 129.3 reprezentuje wartość oczekiwaną zmiennej objaśnianej, gdy wszystkie zmienne objaśniające są równe zero. Współczynniki dla zmiennych niezależnych ‘growthRate’, ‘gni’, ‘qol’, ‘happiness’, ‘crime’ oraz ‘Europe’ są dodatnie, co oznacza, że oczekuje się wzrostu ‘ranking2021’ wraz z wzrostem wartości tych zmiennych. Współczynniki dla ‘pop2023’, ‘landArea’, ‘hdi’, ‘education’ oraz ‘iq’ są ujemne, co oznacza, że oczekuje się spadku ‘ranking2021’ wraz ze wzrostemwartości tych zmiennych. Jednakże odpowiadające im p-value oraz wartości t-statystyki sugerują, że wzrost ich wartości nie ma istotnego wpływu na ‘ranking2021’.

Analiza dopasowania modelu

Wartości reszt wahają się od -43.501 do 44.746. Wartość rozstępu ćwiartkowego (IQR) wynosi 28.29292, co oznacza, że 50% wartości reszt mieści się w przedziale między -14.98 a 13.313. Można to interpretować tak, że większość reszt mieści się w stosunkowo wąskim przedziale, co sugeruje, że model dobrze dopasowuje się do danych. Jednakże istnieją także wartości odstające, które mogą wpłynąć na jakość dopasowania modelu.

Residual standard error wynosi 23.79, zatem średni błąd predykcji modelu wynosi około 23.79.

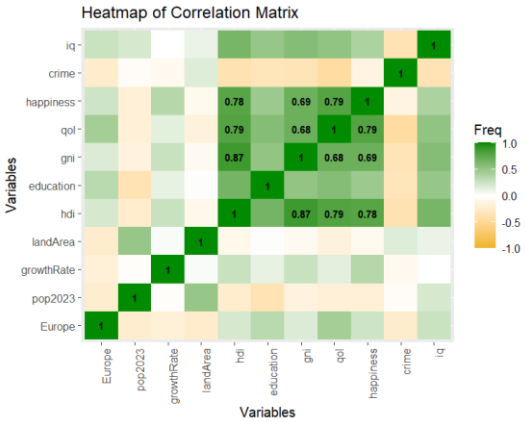
Multiple R-squared wynosi 0.4071, co oznacza, że 40.71% zmienności zmiennej zależnej ‘ranking2021’ może być wyjaśnione przez zmienne niezależne zawarte w modelu.

Adjusted R-squared wynosi 0.2555, co oznacza, że około 25,55% zmienności zmiennej zależnej może być wyjaśnione przez wszystkie zmienne niezależne, z wyłączeniem efektu zmiennej kategorycznej ‘Europe’.

Analiza liniowej zależności i liniowej struktury

W macierzy korelacji, zauważamy współliniowość pomiędzy zmiennymi ‘happiness’, ‘qol’ i ‘hdi’. Silna dodatnia korelacja występuje również między zmiennymi ‘gni’ oraz ‘hdi’ o wartości równej 0.8719, co jest spodziewane, ponieważ GNI jest jednym z kluczowych czynników uwzględnianych w obliczeniu HDI. Sugeruje to odrzucenie jednej z tych zmiennych w celu uniknięcia efektu kolinearności. W pozostałych przypadkach korelacja jest niewielka.

Wartość Kappa wynosi 7.936582, co oznacza, że nie ma silnych dowodów na występowanie współliniowości między zmiennymi w analizowanym zbiorze danych. Jednakże, wynik ten należy interpretować ostrożnie, w celu dokładniejszego zbadania związku między zmiennymi obliczmy czynnik inflacji wariancji (VIF). Wartość VIF dla zmiennej ‘hdi’ wynosi blisko 10, co może sugerować pewną korelację z innymi zmiennymi w modelu. Dla reszty zmiennych wartości VIF są poniżej 6.



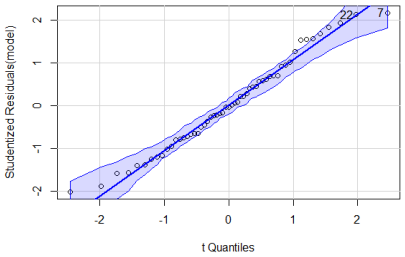
Wykres 1.1

Przeprowadzamy komendę `raintest` Rainbow test z hipotezą zerową o liniowości modelu i otrzymujemy wartość p-value = 0.9198. Natomiast w przypadku testu RESET wartość p-value = 0.01755 sugeruje odrzucenie hipotezy zerowej, że model jest poprawnie zdefiniowany. W celu wyciągnięcia odpowiednich wniosków na temat liniowej struktury modelu przyjrzyjmy się wykresowi Residuals vs Fitted (1.5). Widzimy, że rozproszenie punktów nie jest do końca losowe, a linia trendu jest daleka od linii prostej. Wskazuje to na nieliniowość i skłania do stwierdzenia, że model regresji liniowej nie jest odpowiedni.

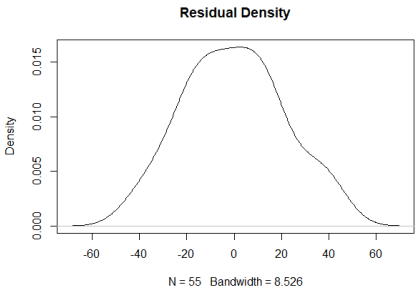
Analiza normalności residuów

Test Shapiro-Wilka wykazał wartość p-value równą 0.7923, zatem nie ma podstaw do odrzucenia hipotezy zerowej o normalności rozkładu danych. Natomiast test Kolmogorova-Smirnova wskazuje, że p-value jest bardzo niskie ($8.454 * 10^{-12}$), co oznacza, że istnieje istotna różnica między rozkładem reszt a rozkładem normalnym. Wynik

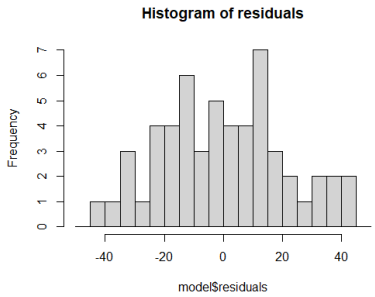
testu Andersona-Darlinga wykazał p-value równe 0.9245. Wysokie p-value (bliskie 1) sugeruje, że reszty pochodzą z populacji o rozkładzie normal-nym. Aby lepiej zbadać ten problem zobrazowałam go z pomocą komendy `qqPlot` (Wykres 1.2). Jak widać punkty skupiają się wokół linii prostej, tylko dwa punkty wychodzą poza obszar ufności, dlatego możemy stwierdzić, że rozkład reszt jest normalny.



Wykres 1.2



Wykres 1.3



Wykres 1.4

Ten fakt bardzo dobrze obrazuje również powyższy wykres (1.3) gęstości residuów.

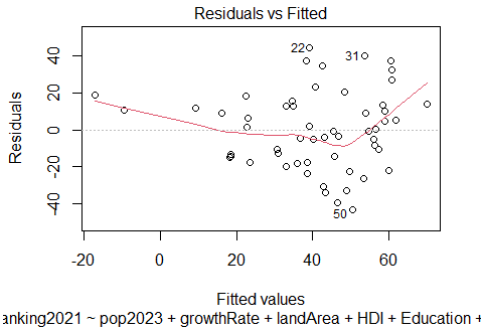
Histogram (1.4) natomiast, wykazuje pewne problemy z normalnością, ale mimo to przyjmujemy normalność.

Stałość wariancji

Przeprowadzamy komendę `crPlots`. Na podstawie wykresów możemy w większości przypadków stwierdzić, że punkty wydają się być rozłożone równomiernie względem przerywanej linii, a co za tym idzie, nie ma problemów ze stałością wariancji. Przyjrzyjmy się bliżej zmiennej 'pop2023', gdyż zagęszczenie punktów na wykresie wzbudz a pewne wątpliwości. Obkładamy zmienną logarytmem i porównuje my wariancje między dwoma podgrupami: kraje o populacji mniejszej niż $10^{7.2}$ i większej niż $10^{7.2}$. Test jednorodności wariancji `var.test` nie znalazł istotnich różnic między tymi dwoma grupami, ponieważ p-value = 0.8384.

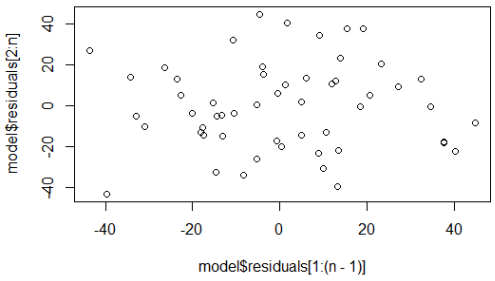
Przeprowadzamy komendę `bptest` test homoskedastyczności White'a i otrzymujemy p-value = 0.5903. Test Goldfeld-Quandt `gqtest` (p-value= 0.6172) oraz test Harrison-McCabe `hmctest` (p-value = 0.41) również wskazują na stałość wariancji.

Na wykresie (1.5) zauważamy pewną zmienność rozrzutu reszt. Jednak, ponieważ wszystkie testy wykazały homoskedastyczność, stwierdzamy, że rozłożenie punktów wynika z nieliniowości modelu.



Wykres 1.5

Autokorelacja residuów



Wykres 1.6

Wykonujemy kolejno testy na autokorelację reszt: Durbin-Watson, Box-Pierce oraz Box-Ljunga i otrzymujemy wartości p-value 0.3912, 0.8976 oraz 0.8948. Wyniki sugerują brak istotnej autokorelacji reszt.

Wykres (1.6) zdecydowanie potwierdza naszą teorię.

Wyniki diagnostyki pierwszego modelu

Model posiada normalne residua, brak problemów z autokorelacją reszt jak i homoskedastycznością, natomiast model nie przeszedł testów na liniowość. Mamy też współliniowość zmiennych niezależnych. Co więcej w modelu 1 nie ma zmiennych istotnych statystycznie.

Wartości odstające

Badania dźwigni oraz odległości Cooka dla dopasowania metodą najmniejszych kwadratów pokazują, że w zbiorze danych jest w sumie pięć obserwacji, których dźwignia przekracza wyliczony próg 0.436.
Najbardziej wpływowa okazała się obserwacja Chiny. Wartość dźwigni dla tej obserwacji wynosi 0.9542, a odległości Cooka 7.5136. Po przeanalizowaniu prawdopodobnej przyczyny tych wartości, stwierdzamy, że kraj ten został zidentyfikowany jako obserwacja odstająca ze względu na dużo większą populację oraz mniejszą średnią liczbę lat edukacji. Jednak, z uwagi na poprawność danych, uznajemy te ekstremalne wartości za niosące istotne informacje dla naszej analizy. Wartość Cook.distance dla reszty obserwacji nie przekracza 0.6.

Budowanie nowego modelu

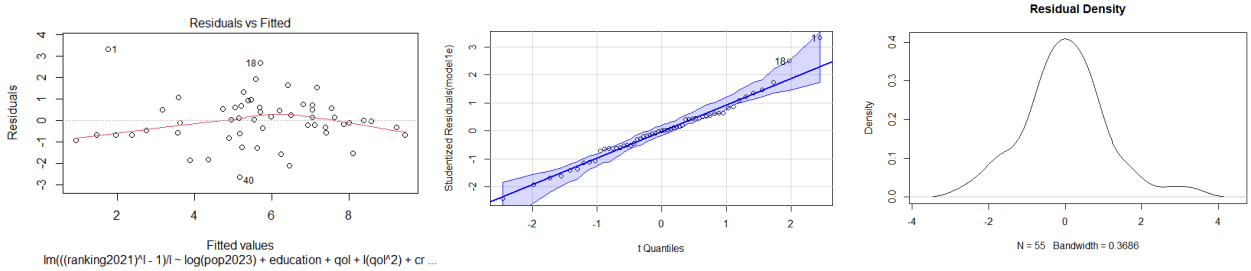
W modelu 1 wartość parametru VIF dla zmiennej ‘hdi’ wynosiła blisko 10. Sugerując się tym faktem, przystępujemy do konstrukcji modelu 1a z wyłączeniem zmiennej ‘hdi’. Wykorzystujemy kryterium informacyjne Akaike do porównania modeli 1 i 1a. Wartość AIC jest korzystniejsza dla modelu 1a. Wartość Kappa spadła do 5.8498, $Adjusted\ R^2$ wzrosło, a wartości VIF dla każdej zmiennej są teraz poniżej 6. Pozbyliśmy się problemu współliniowości, który wprowadzała zmienna ‘hdi’. Wciąż pozostaje silna korelacja między zmiennymi ‘happiness’ oraz ‘qol’ na poziomie 0.79. Konstruujemy dwa modele. W pierwszym z nich dokonujemy ekskluzji zmiennej 'happiness', natomiast w drugim wyłączamy zmienną 'qol'. Tym razem kryterium informacyjne Akaike sugeruje pozostawienie obu zmiennych w modelu. Wciąż występuje problem nieliniowości. Konstruujemy wykresy rozrzutu między zmienną zależną, a zmiennymi niezależnymi. Punkty na wykresach ‘pop2023’, ‘landArea’, ‘growthRate’, ‘qol’ wykazują nieliniowe wzorce. Wykładniczy kształt linii trendu przy zmiennej ‘pop2023’ sugeruje jej przekształcenie logarytmem. Konstruujemy w ten sposób model 1b. Wywołujemy komendę `crPlots` i stwierdzamy, że problem został naprawiony. Co więcej zwiększyła się wartość R^2 oraz $Adjusted\ R^2$ jak również poprawiła się istotność całego modelu jak i zmiennej ‘pop2023’. W tym modelu Intercept oraz zmienna ‘education’ również są istotne statystycznie. Wartość AIC spadła. Przekształcenie zmiennych ‘landArea’ oraz ‘growthRate’ nie dało pożądanych rezultatów, więc rezygnujemy z tej drogi, aby nie komplikować modelu.
Nadal występuje problem nieliniowości. Konstruujemy model 1c stosując wyższą potęgę zmiennej ‘qol’. Model 1c w ogólnej diagnostyce wypada lepiej niż 1b, więc decydujemy się na to przekształcenie.
Na tym etapie nasz model prezentuje się następująco:

```
model1c <- lm(ranking2021 ~ log(pop2023) + growthRate + landArea + iq +
                education + gni + qol + I(qol^2) + happiness + crime +
                Europe, data = data2)
```

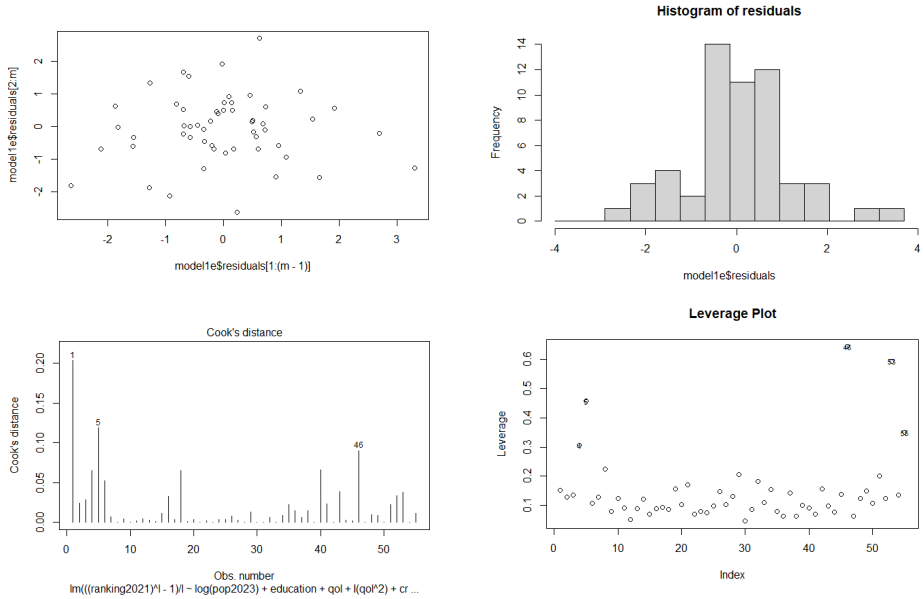
Wciąż występuje problem nieliniowości. Poprawiamy model stosując transformatę Box’a-Cox’a i otrzymujemy model 1d. Nowy model przechodzi testy na liniowość. Dokonujemy redukcji zmiennych za pomocą procedury AIC. Otrzymujemy model 1e, który pozytywnie przechodzi diagnostykę, zatem uznajemy go za model końcowy.

Model końcowy

$$\frac{ranking2021^{\lambda}-\mathbf{1}}{\lambda} = 39.012 - 1.037 * \log(\text{pop2023}) - 0.529 * education - 0.119 * qol + 0.0005 * I(qol^2) + 0.038 * crime + 23.564 * \text{growthRate} - 0.069 * \text{iq}$$



Jak widzimy na wykresach (również tych na następnej stronie) model końcowy spisuje się bardzo dobrze. Spełnia założenia liniowości, normalności i autokorelacji residuów oraz homoskedastyczności. Wszystkie zmienne w modelu są istotne statystycznie. Model wyjaśnia 75,33% zmienności zmiennej zależnej. Jest istotny statystycznie, co potwierdza bardzo niskie p-value $= 2.562 \cdot 10^{-12}$.



W końcowym modelu mamy pięć obserwacji o dźwigni większej niż próg wynoszący 0.2909. Po identyfikacji tych obserwacji i przeanalizowaniu przyczyny dużego wpływu tych obserwacji na model, stwierdzamy, że pierwszą obserwacją jest Ukraina, dla której wskaźnik dźwigni wynosi 0.645, odstaje ze względu na zmienną 'growthRate', drugą jest Iran, dla którego wartość dźwigni wynosi 0.595, odstaje ze względu na wartość 'qol'. Następną z kolei obserwacją są Chiny, o których wspominałam już wcześniej, jednak w ostatecznym modelu dźwignia dla tego kraju spadła do 0.458. Następnie mamy Arabię Saudyjską i Hong Kong, dla których nie udało się zidentyfikować przyczyny potencjalnego odstawania. Żadna z tych obserwacji nie zawiera błędów w danych, w przypadku pierwszych trzech, wartości są rzeczywiście ekstremalne, ale niosą za sobą cenne informacje dla naszej analizy. Decydujemy się na zostawienie tych państw, ponieważ brak nam metod aby zredukować ich wpływ na model.

Porównanie diagnostyki pierwszego i końcowego modelu

Model początkowy, oparty na 55 obserwacjach i 11 zmiennych objaśniających, dysponował 43 stopniami swobody. Z kolei model końcowy, bazujący na 55 obserwacjach i 7 zmiennych objaśniających, charakteryzuje się 47 stopniami swobody. W modelu początkowym żadna ze zmiennych niezależnych nie była statystycznie istotna, natomiast w modelu końcowym wszystkie zmienne są istotne, zatem mają znaczący wpływ na zmienną zależną. Najistotniejsza okazała się być zmienna „pop2023". P-value dla tej zmiennej wyniosło 9.75^{-12}. W modelu końcowym błąd standardowy reszt jest znacznie mniejszy niż w pierwszym modelu, co wskazuje na mniejsze odchylenia między danymi rzeczywistymi a prognozowanymi. W modelu końcowym nie występuje problem współliniowości zmiennych objaśniających, który był obecny w modelu początkowym. Współczynnik Kappa zmalał do wartości 2.964. Średni błąd predykcji modelu początkowego wynosił 23.79, obserwujemy więc znaczną poprawę, ponieważ wartość RSE spadła do 1.193. Model końcowy jest bardziej skuteczny w przewidywaniu zmiennej zależnej.

Wyniki testów statystycznych:

	MODEL POCZĄTKOWY	MODEL KOŃCOWY
P-VALUE DLA TESTU:		
RAINBOW	0.9198	0.4938
RESET	0.01755	0.5828
SHAPIRO-WILKA	0.7923	0.4079
KOLMOGOROVA-SMIRNOVA	$8.454 \cdot 10^{-12}$	0.9507
ANDERSONA-DARLINGA	0.9245	0.3013
BREUSCHA-PAGANA	0.5903	0.7908
GOLDFELDA-QUANDTA	0.6172	0.7095
HARISSONA-MC-CABE'A	0.41	0.808
DURBINA-WATSONA	0.3912	0.1819
BOX-PIERCE'A	0.8976	0.878

BOX-LJUNGA	0.8948	0.8746

Model końcowy przechodzi większość testów statystycznych z lepszymi wynikami, co wskazuje na poprawę pod względem normalności reszt, stabilności wariancji i braku autokorelacji. Model początkowy miał problem nieliniowości, który udało się rozwiązać.

Przy pomocy walidacji krzyżowej obliczyliśmy pierwiastek błędu średniokwadratowego, który dla modelu początkowego wyniósł 41.93, dla modelu końcowego wynosi 16.99. Biorąc pod uwagę dwukrotny spadek, stwierdzamy że model końcowy daje przewidywania bliższe rzeczywistości niż model początkowy. Ponadto wartość parametru R^2 wzrosła z wartości 0.4071 do 0.7533, a $Adjusted\ R^2$, które wynosiło 0.2555, wynosi teraz 0.7165. Obserwujemy więc znaczną poprawę jakości modelu.

Podsumowując, model końcowy wydaje się znacznie lepszy od modelu początkowego, biorąc pod uwagę wyniki wszystkich kluczowych wskaźników i testów statystycznych.

PODSUMOWANIE

Celem naszej pracy była analiza czynników wpływających na plasowanie się pewnych krajów wyżej w rankingu Międzynarodowej Olimpiady Matematycznej w 2021 roku. Opracowany model przewiduje ranking na podstawie czynników takich jak: populacja kraju, przyrost naturalny, średnia liczba lat edukacji w danym kraju, jakość życia i wskaźnik przestępstw. Zidentyfikowaliśmy *populację kraju* jako najważniejszy czynnik wpływający na pozycję w rankingu, sugerując, że większa populacja może zwiększać prawdopodobieństwo wybitnych talentów matematycznych wśród obywateli. Następnie zauważyliśmy, że *liczba lat edukacji* ma duże znaczenie. Dłuższy okres szkolnictwa, wpływa pozytywnie na pozycję w rankingu. To pokazuje, jak ważna jest edukacja.
Wskaźnik jakości życia ma nieco skomplikowany wpływ na ranking kraju. Na początku, nawet mały wzrost tego wskaźnika skutkuje niższą pozycją w rankingu, ale, kiedy jakość życia osiąga pewien poziom, trend się odwraca i dalszy wzrost jakości życia sprawia, że kraje zaczynają radzić sobie coraz lepiej. Ten efekt kwadratowy sugeruje, że istnieje optymalny poziom jakości życia, po przekroczeniu którego, lepsza jakość życia może przyciągać do kraju talenty matematyczne. *Wskaźnik przestępczości* ma negatywny wpływ na ranking kraju. Więcej przestępstw oznacza gorszą pozycję kraju w rankingu, co sugeruje, że kraje o wyższym wskaźniku przestępczości mogą być mniej atrakcyjne dla utalentowanych matematyków. *Przyrost naturalny* ma również negatywny wpływ na pozycję w rankingu. Może to sugerować, że kraje o szybkim wzroście populacji mogą mieć trudności w utrzymaniu dobrego poziomu edukacji, co wpływa na ich ranking w olimpiadzie. Na koniec, wyższy *średni poziom IQ* w kraju wiąże się z jego lepszym wynikiem. To pokazuje, że kraje, w których ludzie są ogólnie mądrzejsi, mogą być bardziej konkurencyjne w takim konkursie jak olimpiada matematyczna.

Rezultaty badania okazały się być koherentne i zgodne z przewidywaniami, co sugeruje, że nasz model skutecznie rozpoznaje kluczowe czynniki wpływające na pozycję danego kraju w rankingu Międzynarodowej Olimpiady Matematycznej. Ta spójność między wynikami a naszymi oczekiwaniami potwierdza wiarygodność przeprowadzonej analizy.