Kraków 2023



Wydział Matematyki i Informatyki

„Analiza czynników wpływających na wyniki reprezentantów krajów europejskich w Międzynarodowej Olimpiadzie Matematycznej z 2021 roku przy użyciu modelu regresji w języku R."

Inga Dyląg

**Cel analizy**

Celem analizy jest badanie czynników wpływających na sukcesy krajów europejskich 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 wzrostu gospodarczego
* landArea – powierzchnia w
* hdi – wskaźnik rozwoju społecznego, mierzący poziom edukacji, oczekiwaną długość życia i poziom dochodów
* education – średnia liczba lat edukacji
* gni – dochód narodowy brutto na mieszkańca
* 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 przestępczości, mierzący poziom przestępczości
* iq – średni poziom IQ mieszkańców

oraz zmiennej jakościowej:

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

**Model liniowy**

Po przeprowadzeniu wstępnej analizy modelu zostało odrzuconych 48 obserwacji ze względu na braki w nich zawarte.  
Po dokonaniu czyszczenia danych liczba obserwacji wyniosła 54, co przy 11 zmiennych daje 43 stopnie swobody.

**Analiza współczynników**

Współczynniki w tym modelu opisują wpływ każdej zmiennej niezależnej na zmienną zależną „ranking2021”. Szacunki współczynników, które są podane w podsumowaniu summary, pokazują oczekiwaną zmianę wartości zmiennej zależnej dla jednostkowego wzrostu zmiennej niezależnej, przy założeniu, że wartości pozostałych zmiennych niezależnych pozostają stałe.

Warto zwrócić uwagę na wartość t-statystyki dla każdego współczynnika i odpowiadające jej p-wartości, które wskazują, czy dany współczynnik jest istotny statystycznie, przy założeniu poziomu istotności równego . Wartości w kolumnie Pr(>|t|) wskazują na brak istotności statystycznej dobranych zmiennych.

Współczynnik wyrazu wolnego równy 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 wzrostem  
wartości tych zmiennych. Jednakże, wartości t-statystyki i odpowiadające im p-wartości sugerują, że wzrost ich wartości nie ma istotnego wpływu na „ranking2021”. Wszystkie wartości współczynników są względnie niskie i bliskie zera.

**Analiza dopasowania modelu**

Wartości reszt wahają się od do , co sugeruje, że model nie jest w pełni skuteczny w przewidywaniu rankingu. Wartość rozstępu ćwiartkowego (IQR) wynosi , co oznacza, że 50% wartości reszt mieści się w przedziale między a . 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 , zatem średni błąd predykcji modelu wynosi około . Niższe wartości błędu  
standardowego reszt wskazują na lepsze dopasowanie modelu, ponieważ oznacza to mniejszy błąd predykcji.  
  
Multiple R-squared wynosi , co oznacza, że zmienności zmiennej zależnej „ranking2021” może być wyjaśnione przez zmienne niezależne zawarte w modelu.

Adjusted R-squared wynosi, co oznacza, że około zmienności zmiennej zależnej może być wyjaśnione przez wszystkie zmienne niezależne, z wyłączeniem efektu zmiennej kategorycznej „Europe”.

F-statistic wynosi , a jej p-value wynosi , co jest mniejsze niż ustalony poziom istotności więc możemy  
odrzucić hipotezę zerową o braku zależności między zmiennymi niezależnymi a zmienną zależną i stwierdzić, że model jest  
statystycznie istotny.

**Analiza liniowej zależności i liniowej struktury**

**Obraz zawierający wykres

Opis wygenerowany automatycznie**Macierz korelacji pokazuje nam, że niektóre zmienne są ze sobą powiązane. Wyznacznik macierzy korelacji jest równy , jest to wartość bliska zeru, co oznacza, że zmienne są silnie skorelowane i mogą być liniowo zależne.

Obraz zawierający wykres

Opis wygenerowany automatycznieMacierz korelacji wskazuje na silną dodatnią korelację między zmiennymi "gni" oraz "hdi" o wartości równej . Sugeruje to odrzucenie jednej ze zmiennych “gni” lub “hdi” w celu uniknięcia efektu kolinearności, który może wpłynąć na stabilność  
modelu i trudności interpretacyjne. Korelacje tą widać dobrze na wykresie obok.

Wartość współczynnika Kappa powyżej 30 sugerują występowanie pewnej współliniowości w danych. W tym przypadku  
wartość Kappa wynosi , 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).

Wszystkie zmienne w modelu mają wartości VIF poniżej 6, co oznacza, że nie ma znaczących problemów z wieloliniowością. Warto jednak zauważyć, że wartość VIF dla zmiennej "hdi" wynosi blisko 10, co może sugerować pewną korelację z innymi zmiennymi w modelu. W takim przypadku warto dokładniej przyjrzeć się tym zmiennym i ich wpływowi na model.

Przeprowadzamy komendę raintest Rainbow test z hipotezą zerową o liniowości reszt i otrzymujemy wartość   
p-value .

W przypadku testu RESET wartość p-value sugeruje odrzucenie hipotezy zerowej, że model jest poprawnie  
zdefiniowany.

**Analiza normalności residuów**

Test Shapiro-Wilka wykazał wartość p-value równą , zatem nie ma podstaw do odrzucenia hipotezy zerowej o normal- ności rozkładu danych. Natomiast test Kolmogorova-Smirnova wskazuje, że p-value jest bardzo niskie (), co  
oznacza, że istnieje istotna różnica między rozkładem reszt a rozkładem normalnym.

Wynik testu Andersona-Darlinga wykazał p-value równe . Wysokie p-value (bliskie 1) sugeruje, że reszty pochodzą  
z populacji o rozkładzie normalnym.

Aby lepiej zbadać ten problem zobrazowałam go z pomocą komendy qqPlot.

Obraz zawierający wykres

Opis wygenerowany automatycznieObraz zawierający wykres

Opis wygenerowany automatycznieJak widać punkty skupiają się wokół linii prostej, tylko dwa punkty wychodzą poza obszar ufności, dlatego możemy stwierdzić, że rozkład reszt jest normalny.

Obraz zawierający wykres

Opis wygenerowany automatycznieTen fakt bardzo dobrze obrazuje również powyższy wykres (2) gęstości residuów.

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

**Stałość wariancji**

Przeprowadzamy komendę crPlots. Na podstawie wykresów możemy stwierdzić znaczne problemy ze stałością wariancji dla zmiennych „pop2023”, „growthRate”, „qol”. Przyjrzyjmy się bliżej zmiennej „pop2023” i porównajmy wariancje między dwoma podgrupami: kraje o populacji mniejszej niż i większej niż . Test jednorodności wariancji var.test nie znalazł istotnych różnic między tymi dwoma grupami w odniesieniu do wariancji, ponieważ p-value .

Przeprowadzamy komendę bptest test heteroskedastyczności z hipotezą zerową mówiącą o homoskedastyczności  
i otrzymujemy p-value .

Obraz zawierający wykres

Opis wygenerowany automatycznie

Jak widzimy na wykresie, krzywa dopasowania nie jest zbliżona do linii prostej, co może wskazywać na nieliniowe zależności między zmiennymi oraz nietypowy rozkład błędów.

**Autokorelacja residuów**

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.

Obraz zawierający wykres

Opis wygenerowany automatyczniePoniższy wykres zdecydowanie potwierdza naszą teorię.