Predykcja wyników IMO za pomocą danych makroekonomicznych.

Łukasz Świętochowski

30.04.2023

1 Wstęp

Tematem przewodnim projektu jest pokazanie zależności między wynikami w IMO danego kraju a jego wskaźnikami makroekonomicznymi.
Wszystkie dane dotyczące IMO pochodzą z oficjalnej strony międzynarodowej olimpiady matematycznej, natomiast zródłem pozostałych danych jest DataBank od organizacji WorldBank. Wszelkie luki w danych uzupełniłem za pomocą wikipedii, stron rządowych krajów oraz artykółów naukowych. Dane pochodzą z 2015 roku.

2 Dane do modelu

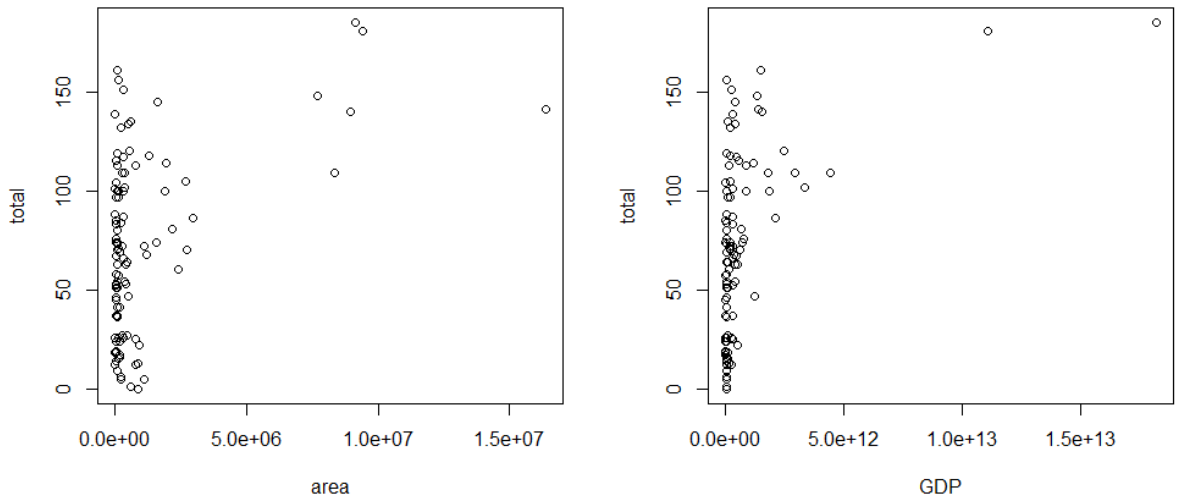
Zmienną objaśnianą będzie

1. **total** - suma punktów danego kraju z wszystkich 6 zadań

Natomiast moimi zmiennymi objaśniającymi były

1. **population** - populacja państwa
2. **GDP** - GDP państwa liczone w USD
3. **GDP.per.cap** - GDP per capita liczone w USD
4. **mortality** - śmiertelność poniżej 5 roku życia na 1000 urodzeń
5. **internet** - % populacji posiadającej dostęp do internetu
6. **area** - powierzchnia państwa
7. **life.expt** - oczekiwana długość życia przy narodzinach
8. **secondary** - wiek pójścia do szkoły drugiego stopnia
9. **primary** - długość szkoły podstawowej
10. **sex** - stosunek narodzin mężczyzn do kobiet

Pierwszą rzeczą jaką zrobiłem było użycie komendy summary, aby sprawdzić poprawność danych. Nie zauważyłem żadnych błędów w danych. Następnie wyliczyłem macierz korelacji. Największą korelację wykazały **area** oraz **GDP**, więc spodziewam się iż będą kluczowe w prognozie, najmniejszą zaś **sex**(-0.0037) oraz **GDP.per.cap** (0.0201) więc zakładam iż będą mało istotne dla modelu. Pozostałe korelacje natomiast oscylują między 0.1 a 0.3.
Zobaczę teraz wykres łącznego wyniku od powierzchni oraz od GDP



Na załączonych wykresach można zauważyć delikatny zarys trendu, niestety w obu przypadkach duży wpływ mają dane odstające.

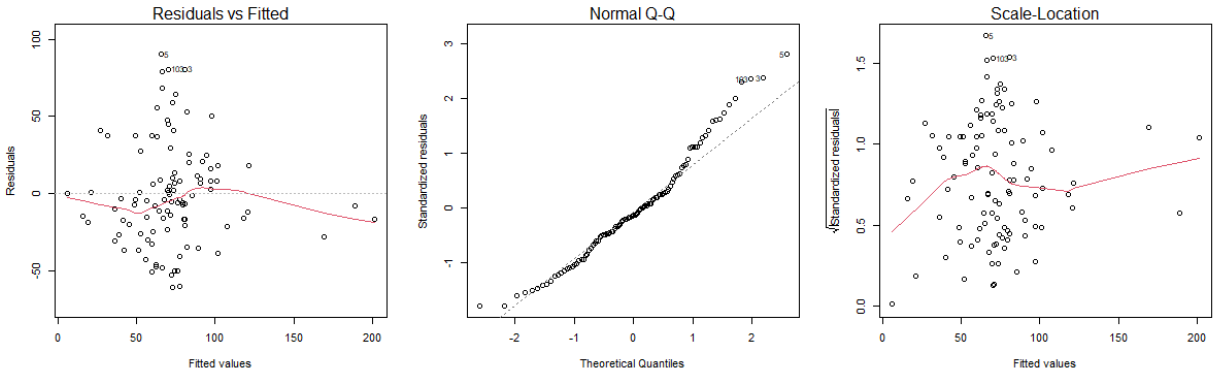
3 Model liniowy i jego analiza

Postanowiłem stworzyć model liniowy.Otrzymałem model liniowy o następujących współczynnikach:

$$-12.61 + 2.37 \cdot 10^{-8} X_{population} + 4.76 \cdot 10^{-12} X_{GDP} - 1.854 \cdot 10^{-4} X_{GDP.per.cap} - 0.201 X_{mortality} + 0.048 X_{internet} + 4.71 \cdot 10^{-6} X_{area} + 1.84 X_{life.expt} + 2.90 X_{secondary} - 17.51 X_{primary} + 2.808 \cdot 10^{2} X_{sex}$$

$$-12.61 + 2.37 \cdot 10^{-8} \textbf{population} + 4.76 \cdot 10^{-12} \textbf{GDP} - 1.854 \cdot 10^{-4} \textbf{GDP.per.cap} - 0.201 \textbf{mortality} + 0.048 \textbf{internet} + 4.71 \cdot 10^{-6} \textbf{area} + 1.84 \textbf{life.expt} + 2.90 \textbf{secondary} - 17.51 \textbf{primary} + 2.808 \cdot 10^{2} \textbf{sex}$$

Po stworzeniu modelu sprawdziłem jego charakterystyki komendą summary, co pozwoliło mi stwierdzić iż jedynymi istotnymi zmiennymi są **GDP**, **area**, **life.expt**, **primary** ze względu na małe p.value przy hipotezie, że są równe 0 (odp: 0.02805, 0.00751, 0.0527, 0.00917).
Wartość R^2 modelu wyniosła **0.4466599**, więc model nie jest dobrym objaśnieniem **total**. Następnym krokiem jest sprawdzenie wykresów modelu i weryfikacja założeń. Za pomocą funkcji plot uzyskałem następujące wykresy.



Na wykresie **Residuals vs Fitted** można zauważyć delikatny trend reszt modelu, co może sugerować nieodpowiedni dobór modelu.

Zweryfikuje to przeprowadzając test Harveya-Colliera komendą *harvtest*, otrzymując p.value = 0.02999, co jest wartością graniczą więc postanowiłem skorzystać z testu Ramseya RESET *resettest*, otrzymując p.value 0.01209 oraz z Rainbowtestu z p.value = 0.05479,stwierdzając w ten sposób **brak liniowości modelu**.
Wykres **Normal Q-Q** odchyla się od oczekiwanej prostej w prawym górnym rogu, co może oznaczać brak normalności reszt.
Ponieważ p.value w teście Shapiro-Wilka *shapiro.test* jest równe 0.03556, a w teście Jarque-Bera *jarque.bera.test* wynosi 0.1161, to biorąc po uwagę te wartości jak i zakrzywienie na wykresie **odrzucam założenie o normalności reszt**.
Wykres **Scale-Location** tutaj również można zauważyć delikatny tren co sugeruje odstępstwo od założenia o jednorodnej wariancji. Aby sprawdzić to założenie wykonam test Harrisona-McCabe komendą *hmctest* co dało mi p.value 0.47 oraz test Goldfelda-Quandta komendą *gqtest* otrzymując p.value równe 0.5224 więc **nie ma podstaw do odrzucenia założenia**.
Założenie o braku współliniowości sprawdzę komendą VIF. Zadna z wartości nie przekroczyła 10, więc nie ma podstaw do sądzenia, że zmienne są współliniowe.
Na koniec sprawdzę założenie o niezależności reszt. Użyję do tego testu Durbina-Watsona *dwtest*. Po przeprowadzeniu go otrzymałem p.value równe 0.08092, co jest wartością graniczną, dlatego skorzystam dodatkowo z testu Breuscha-Godfreya *bgtest*, tym razem z p.value 0.1712, więc za tym idzie **nie ma podstaw do odrzucenia hipotezy o niezależności reszt**.

4 Analiza niematematyczna

Model mówi nam, iż wraz z wzrostem wskaźników takich jak populacja, GDP, % populacji z dostępem do internetu, powierzchnia, oczekiwana długość życia przy narodzinach, wiek rozpoczęcia szkoły drugiego stopnia oraz stosunek narodzin mężczyzn do kobiet wzrasta łączny wynik na IMO. Z kolei wraz z wzrostem wskaźników typu GDP per capita, śmiertelność poniżej 5 roku życia oraz długość szkoły podstawowej zmniejsza się łączny wynik na IMO. Jak wspomniałem wcześniej jedynymi istotnymi statystycznie zmiennymi są GDP, powierzchnia, oczekiwana długość życia przy narodzinach oraz długość szkoły podstawowej, pozostałe nie mają znaczącego wpływu na model.

5 Podsumowanie

Stworzony przez nas model ma jednorodną wariancję, jest współliniowy oraz jego reszty są niezależne, natomiast nie opisuje w satysfakcjonującym stopniu łącznego wyniku kraju na IMO, nie jest liniowy, jego reszty nie mają rozkładu normalnego, Biorąc to wszystko pod uwagę mogę stwierdzić, iż model **nie jest dobrze dopasowany i wymaga pewnego rodzaju poprawek**.