Predykcja wyników IMO danymi makroekonomicznymi

Łukasz Świętochowski

30.04.2023

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 roku 2015.
Zmienną objaśnianą będzie

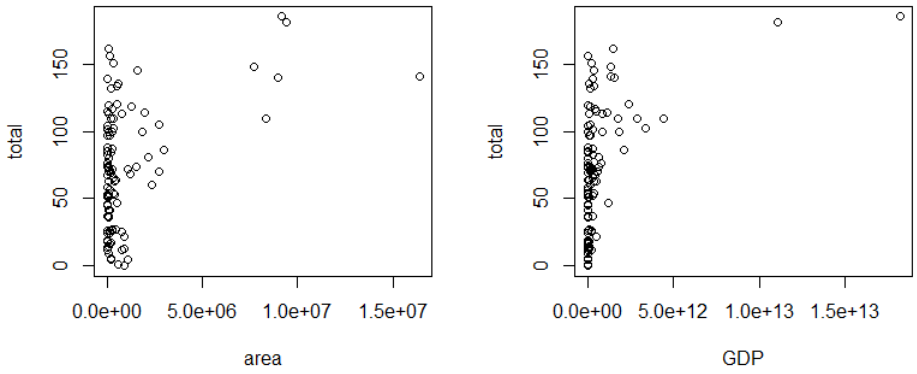
1. **total** - suma punktów danego kraju z wszystkich 6 zadań

Natomiast moimi zmiennymi objaśniającymi były

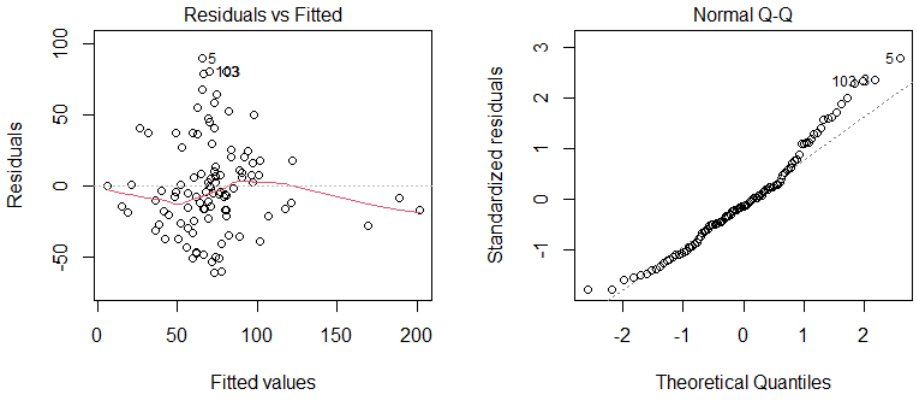
1. **population** - populacja państwa
2. **GDP** - GDP państwa liczone w USD
3. **GDP.per.cap** - GDP per capita liczone w USD
4. **mortality** - śmiertelność poniżej 5 roku życia na 1000 urodzeń
5. **internet** - % populacji posiadającej dostęp do internetu
6. **area** - powierzchnia państwa
7. **life.expt** - 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 przy narodzinach

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 **powierzchnia państwa** oraz **GDP**, więc

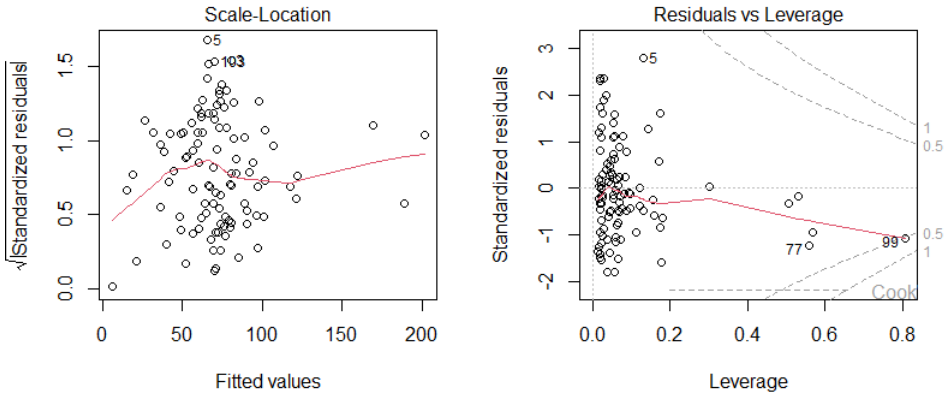
spodziewam się iż będą kluczowe w prognozie, najmniejszą zaś **stosunek narodzin mężczyzn do kobiet**(-0.0037) oraz **GDP per capita** (0.0201). Pozostałe korelacje natomiast oscylują między jedną dziesiątą a trzema dziesiętnymi.
Zobaczę teraz wykres łącznego wyniku od powierzchni oraz od GDP



Na załączonych wykresach można zauważyć delikatny zarys trendu, niestety w obu przypadkach duży wpływ mają dane odstające.
Postanowiłem stworzyć model liniowy bez użycia **GDP.per.cap** oraz **sex** ze względu na niską korelację.Otrzymałem model liniowy o następujących współczynnikach: $-12.61+2.37\cdot10^{-8}$**population** $+4.76\cdot10^{-12}$**GDP** -0.201**mortality** $+0.048$**internet** $+$ $4.71\cdot10^{-6}$**area** $+1.84$**life.expt** $+2.90$**secondary** -17.51**primary**
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 **0.4335074**, więc model nie jest dobrym objaśnieniem wyniku. 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 = 0.0815, co jest wartością graniczą więc postanowiłem skorzystać z testu Ramseya RESET *resettest*, otrzymując p.value 0.02129 stwierdzam brak liniowości modelu. Wnioskując **GDP** i **area** rozwiązaniem mogłaby być transformacja logarytmiczna zmiennej objaśnianej. 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 0.0094, więc odrzucam hipotezę 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 0.366 więc nie ma podstaw do odrzucenia założenia. Z wykresu **Residuals vs Leverage** można wyczytać, iż obserwacje numer 59 (Mongolia),77 (Rosja) oraz 5 (Korea Południowa) są odstające od reszty. Po sprawdzeniu poprawności danych mogę stwiedzić, iż ich odstawanie nie jest spowodowane błędem. 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.
Model mówi nam, iż wraz z wzrostem wskaźników takich jak populacja, GDP, % populacji z dostępem do internetu, powierzchnia, oczekiwana długość życia przy narodzinach oraz wiek rozpoczęcia szkoły drugiego stopnia wzrasta łączny wynik na IMO. Z kolei wraz z wzrostem wskaźników typu 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.
Biorąc to wszystko pod uwagę mogę stwierdzić, iż model **nie jest** dobrze dopasowany i wymaga pewnego rodzaju poprawek.