Wpływ danych makroekonomicznych na wyniki IMO 2015

Łukasz Świętochowski

12.06.2023

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

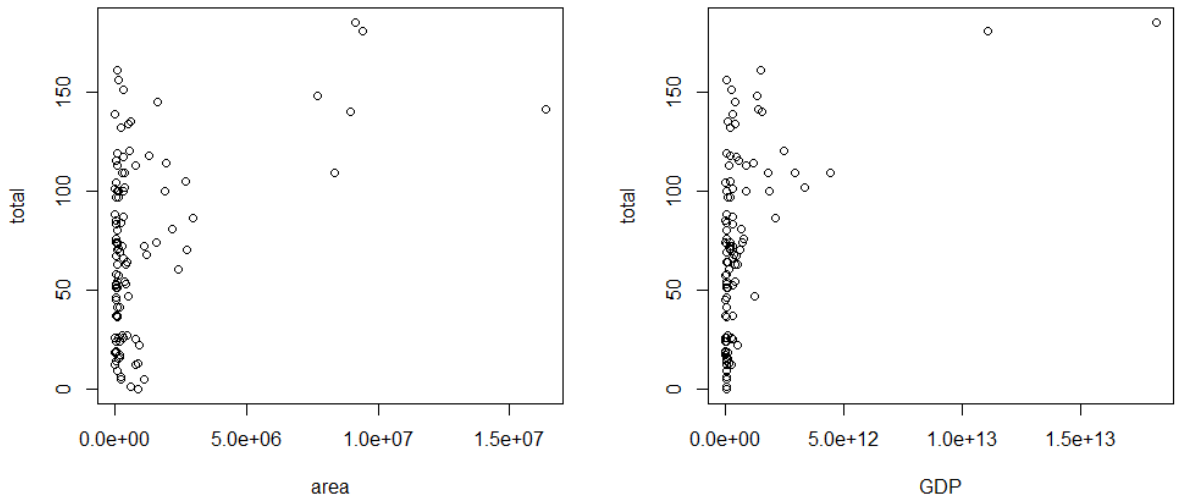
1. **total** - suma punktów danego kraju z wszystkich 6 zadań

Natomiast moimi zmiennymi objaśniającymi były

1. **population** - populacja państwa [liczba ludności w jednostkach]
2. **GDP** - GDP państwa liczone w USD [USD]
3. **GDP.per.cap** - GDP per capita liczone w USD [USD]
4. **mortality** - śmiertelność poniżej 5 roku życia na 1000 urodzeń $[\frac{\text{liczba}}{\text{tysiąc urodzeń}}]$
5. **internet** - % populacji posiadającej dostęp do internetu [% populacji]
6. **area** - powierzchnia państwa [km^2]
7. **life.expt** - oczekiwana długość życia przy narodzinach [lata]
8. **secondary** - wiek pójścia do szkoły drugiego stopnia [lata]
9. **primary** - długość szkoły podstawowej [lata]
10. **sex** - stosunek narodzin mężczyzn do kobiet [stosunek]

Pierwszą rzeczą jaką zrobiłem było użycie komendy *summary*, aby sprawdzić poprawność danych. Zauważyłem, że zmienna **internet** dla państwa Kosowo jest równa 0 oraz brak jednej obserwacji (**mortality** dla państwa Lichtenstein), niestety nie znalazłem brakujących informacji, więc zmienną **internet** dla Kosowo zmieniłem na brak obserwacji. Następnie wyliczyłem macierz korelacji (*cor*). Największą korelację z zmienną objaśnianą **total** wykazały **area** (0.4364) oraz **GDP** (0.454), więc spodziewam się iż będą kluczowe w prognozie, najmniejszą zaś **sex** (-0.0037) oraz **GDP.per.cap** (0.021) więc zakładam iż będą mało istotne dla modelu. Pozostałe korelacje natomiast oscylują między 0.1 a 0.3.

Zobaczę teraz wykres **total** od **area** 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

3.1 Model liniowy i jego charakterystyki

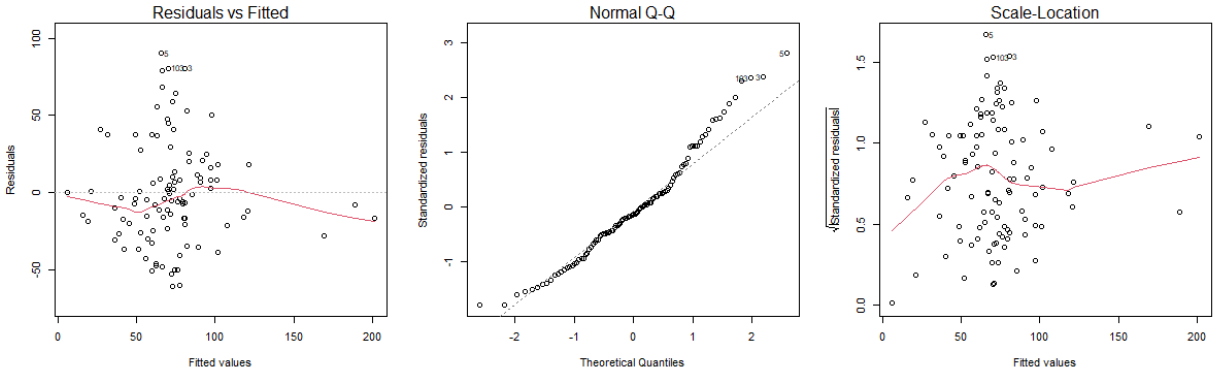
Postanowiłem stworzyć model liniowy za pomocą wszystkich zmiennych, otrzymując w ten sposób model o następujących współczynnikach:

$$-337.4+2.064\cdot 10^{-8}\textbf{population}+5.435\cdot 10^{-12}\textbf{GDP}-1.854\cdot 10^{-4}\textbf{GDP.per.cap}-0.2327\textbf{mortality}+0.07613\textbf{internet}+4.975\cdot 10^{-6}\textbf{area}+1.929\textbf{life.expt}+3.542\textbf{secondary}-14.79\textbf{primary}+2.808\cdot 10^{2}\textbf{sex}$$

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: 0.01460, 0.00527, 0.05552, 0.03288).
Wartość R^2 modelu wyniosła **0.3974196**, więc model nie jest dobrym objaśnieniem **total**.

3.2 Założenia modelu

Następnym krokiem będzie 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* w R), otrzymując p.value 0.01209 oraz z Rainbowtestu (*raintest*) 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**.

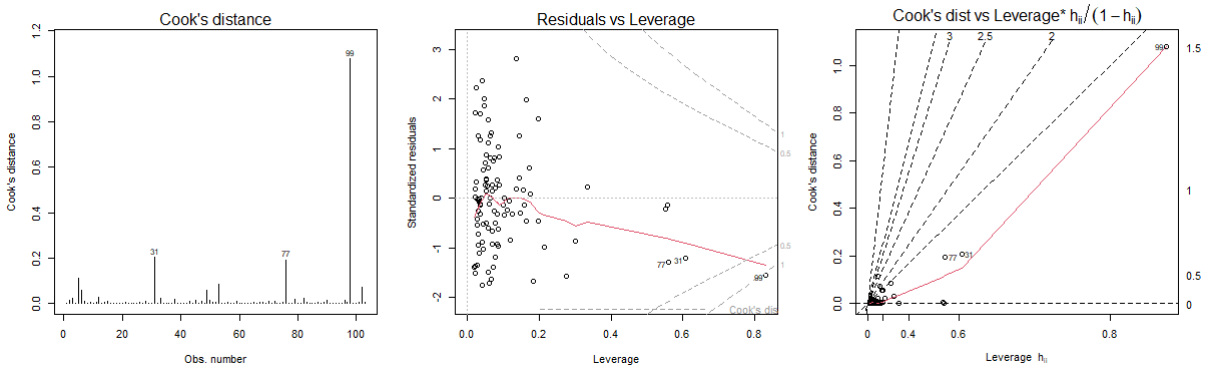
Wykres **Scale-Location** tutaj również można zauważyć delikatny tren co sugeruje odstępstwo od założenia o stałej wariancji. Aby sprawdzić to założenie wykonam test Harrisona-McCabe komendą *hmctest* co dało mi p.value 0.482 oraz test Goldfelda-Quandta komendą *gqtest* otrzymując p.value równe 0.5224 więc **nie ma podstaw do odrzucenia założenia**.

Założenie modelu o braku współliniowości sprawdzę komendą VIF. Żadna 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 0.07722, co jest wartością graniczną, dlatego skorzystam dodatkowo z testu Breuscha-Godfreya *bgtest*, tym razem z p.value 0.1549, więc za tym idzie **nie ma podstaw do odrzucenia hipotezy o niezależności reszt**.

3.3 Obserwacje odstające

W celu analizy obserwacji odstających moim pierwszym krokiem było zastosowanie funkcji *plot* w celu przeanalizowania wykresów.



Jak widać na powyższych rysunkach obserwacje 99 (Stany Zjednoczone), 77 (Federacja Rosyjska) oraz 31 (Indie) odbiegają od reszty.

Najbardziej wpływową obserwacją są Stany Zjednoczone, przy której miara Cooka jest w okolicy 1. Po przeanalizowaniu danych dotyczących tego kraju nie znalazłem błędów. Duży wpływ może być spowodowany dużym łącznym wynikiem przy jednoczesnym dużym GDP oraz powierzchni.Może być również spowodowany brakiem liniowości modelu, i zredukuje się wraz z transformacją modelu.

Kolejnym odstającym krajem jest Federacja Rosyjska z miarą Cooka w okolicy 0.4. Jako kraj o największej powierzchni i łącznym wyniku równym 141 napewno posiada duży wpływ na zmienne. Po przeanalizowaniu czynników makroekonomicznych stwierdziłem brak błędów w obserwacjach.

Ostatnim krajem do sprawdzenia są Indie. W Indiach IMO jest mało znaczącym osiągnięciem, które nie gwarantuje nawet przyjęcia na uczelnie. Dużo ważniejszych egzaminem jest JEE (Joint Entrance Examination), który jest czymś w rodzaju matury dotyczącej inżynierskich kierunków. Ponieważ charakterystyka tego egzaminu jest zupełnie inna od IMO, dobre wyniki na jednym nie przenoszą się na wyniki w drugim. Z tego też powodu oraz faktu, iż współczynniki makroekonomiczne odstają od reszty krajów postanowiłem odrzucić tą obserwację.

4 Transformacja modelu

Ponieważ model nie spełnia wszystkich założeń oraz nie jest dobrze dopasowany postanowiłem wykonać transformacje modelu, w tym celu stworzyłem kilka pośrednich modeli, które po krótce omówię oraz model końcowy. W celu ułatwienie wykonywania transformacji zmieniłem wynik Tanzanii z 0 na 0.0001. Zmiana ta nie wywołała żadnej istotnej zmiany w modelu.

4.1 Transformacja logarytmiczna

W pierwszym modelu pośrednim postanowiłem wykonać transformacje logarytmiczną zmiennej total $\tilde{y} = log(y+\alpha)$ dla $\alpha = 5.9999$. Otrzymany model miał podobne R^2, lecz nie był liniowy, nie był również normalny oraz wykres diagnostyczny sugerował brak założenia o jednorodnej wariancji. Biorąc to pod uwagę wraz z dużą miarą Cooka Chin (około 8) postanowiłem odrzucić tą transformację.

4.2 Transformacja Boxa-Coxa

Następną transformacją będzie transformacja Boxa-Coxa zmiennej total $\tilde{y} = \frac{y^{\lambda}-1}{\lambda}$, dla $\lambda \neq 0$ oraz $\tilde{y} = log(y)$, dla $\lambda = 0$. Transformacje wykonałem dla $\lambda = 0.5454545$. Model miał delikatnie wyższe R^2, oraz spełniał założenie o normalności reszt, lecz wciąż można było zauważyć problemy z liniowością modelu, która jest kluczowym założeniem oraz bardzo dużą miarę Cooka Chin (około 7). Postanowiłem zostawić tą transformację i budować następne pośrednie modele bazując na tym.

4.3 Transformacje zmiennej population

Brak założenia o liniowości może również wynikać z braku dopasowania którejśc z zmiennych objaśniających. Postanowiłem stworzyć model w którym zamiast zmiennej mówiącej o populacji użyłem logarytmu tej zmiennej. Model ten posiadał znacząco wyższy współczynnik R^2 równy 0.5297, wykresy diagnostyczne nie sugerowały braku założeń oraz wszystkie testy potwierdzały brak problemów z założeniami.Nie było również problemu z wpływowymi obserwacjami. Otrzymaliśmy jednak wielę nieistotnych statystycznie zmiennych. Również postanowiłem zostwić tą transformację i budować na jej podstawie końcowy model.

4.4 Redukcja zmiennych

Ponieważ w poprzednim modelu pośrednim otrzymaliśmy nieistotne statystycznie zmienne, postanowiłem zredukować ich ilość stosując kryterium **BIC**. W tej pracy preferowałem użycie kryterium BIC zamiast AIC, gdyż nie chcę wykonać predykcji wyników, lecz zbadać wpływ różnych czynników na owe wyniki. Redukcji dokonałem używając funkcji *step*, otrzymując w ten sposób model z pięcioma statystycznie istotnymi zmiennymi oraz o wyższym współczynniku R^2 nie tracąc założeń z poprzedniego modelu przejściowego. Biorąc to wszystko pod uwagę uznaję ten model za model końcowy.

5 Model końcowy

Po przeprowadzeniu wszystkich zmian otrzymałem model o następujących współczynnikach

$$\tilde{\textbf{total}} = 112.1 + 2.067\tilde{\textbf{population}} + 5.418 \cdot 10^{-7}\textbf{area} + 0.5597\textbf{life.expt} - -2.065\textbf{primary} + 68.62\textbf{sex}$$

gdzie $\tilde{\textbf{total}} = \frac{\textbf{total}^{\lambda}-1}{\lambda}$ oraz $\tilde{\textbf{population}} = ln(\textbf{population})$

6 Porównanie modeli

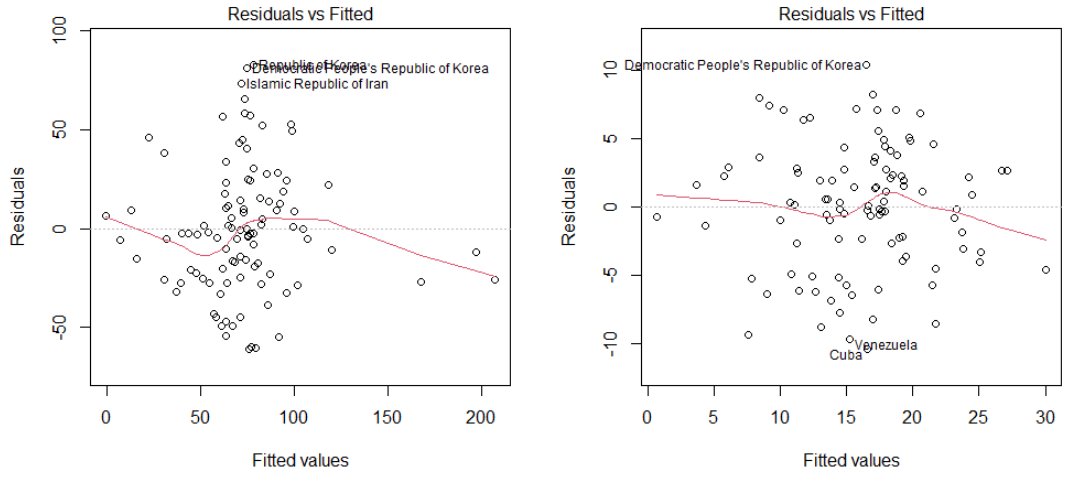
6.1 Istotność statystyczna zmiennych

W początkowym modelu istotnymi statystycznie zmiennymi były **GDP, area, life.expt, primary**, natomiast w modelu końcowym były to **population, area, life.expt, primary, sex**. Brak istotności zmiennej **GDP** w końcowym modelu mógł być spowodowany brakiem liniowości oraz wpływowymi obserwacjami początkowego modelu. Pojawienie się w końcowym modelu zmiennych **population** i **area** może być wynikiem transformacji Boxa-Coxa zmiennej objaśnianej.

6.2 R^2

Początkowy model posiadał współczynnik R^2 w wysokości **0.3974196**, natomiast w końcowym modelu udało się zwiększyć to do wartości **0.5437656** co jest znaczącą zmianą która informuje nas, że nowy model dużo lepiej opisuje zjawisko od poprzedniego.

6.3 Liniowość modelu



Lewy wykres diagnostyczny przedstawia stary model, prawy wykres zaś nowy. Jak możemy zauważyć na prawym rysunku linia dużo bardziej trzyma się poziomu zero, i nie widać żadnego trendu.

Test/p-value	Pierwszy model	Model ostateczny
Harvey-Collier test	0.02999	0.2749
RESET test	0.01209	0.3015
Rainbow test	0.05479	0.4774

Jak możemy zauważyć, nowy model zdał wszystkie testy diagnostyczne, więc nie ma problemu z założeniem o liniowości.

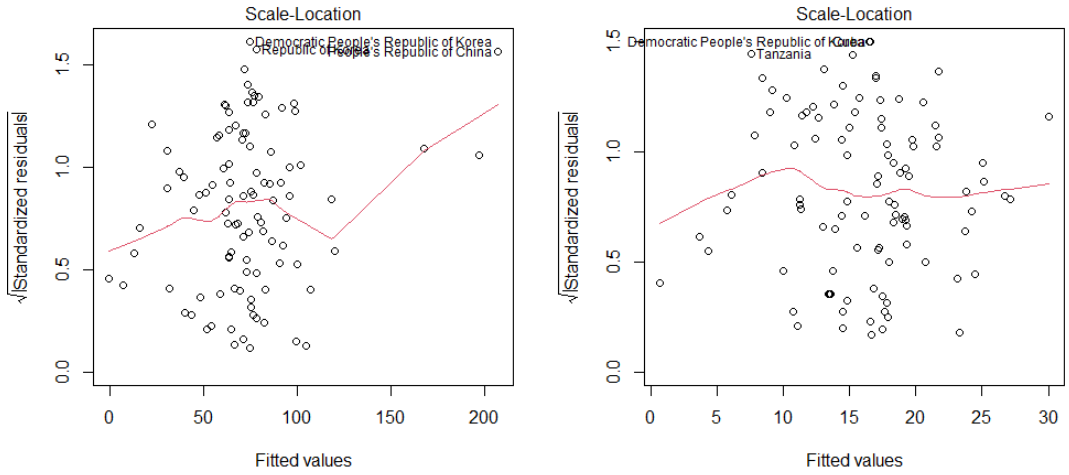
6.4 Normalność błędów

W tej diagnostyce z powodu braku zauważalnych różnic w wykresach diagnostycznych postanowiłem nie odwoływać się do nich, zamiast tego skupiłem się na testach statystycznych.

Test	Pierwszy model	Model ostateczny
Shapiro-Wilk test	0.03556	0.3041
Jarque Bera test	0.1161	0.4358

W pierwszym modelu test Shapiro Wilka odrzucił założenie o normalności reszt a test Jarque Bera nie odrzucił założenia. Powyższe testy w końcowym modelu nie odrzuciły założenia o braku normalności reszt w modelu. W związku z tym nie ma podstaw odrzucać owe założenie w końcowym modelu.

6.5 Jednorodna wariancja



Na lewej grafice dotyczącej pierwotnego modelu możemy zauważyć wachania wariancji w zależności od wartości, które natomiast na rysunku z prawej strony dotyczącym nowego modelu są znikome co sugeruje jednorodną wariancję.

Test	Pierwszy model	Model ostateczny
Goldfeld-Quandt test	0.5224	0.1807
Harrison-McCabe test	0.482	0.168

Po przeprowadzeniu testów statystycznych dla obu modeli mogżemy stwierdzić, że nie ma podstaw do odrzucenia założenia o jednorodnej wariancji.

6.6 Współliniowość

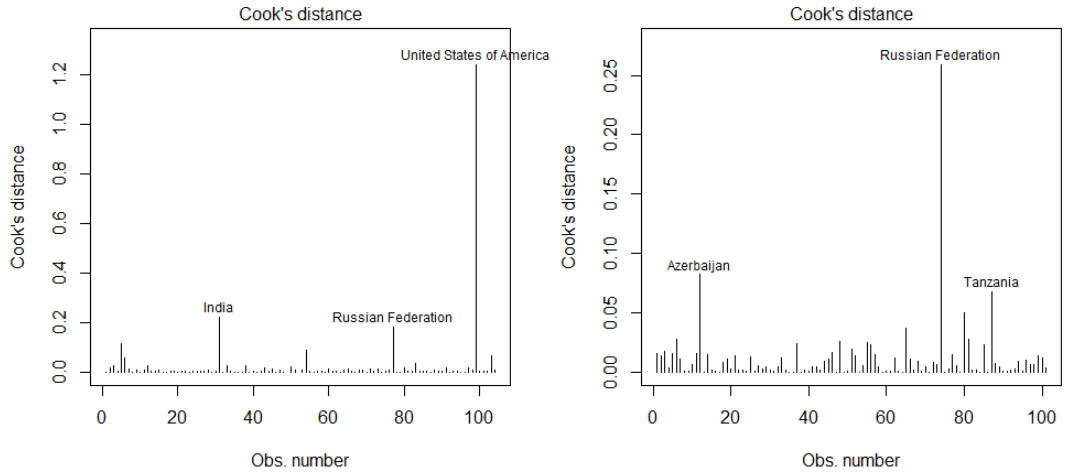
Początkowe dane nie były skorelowane. W obu modelach współczynnik **VIF** nie przekracza 10, więc oba modele nie są współliniowe.

6.7 Niezależność reszt

Test	Pierwszy model	Model ostateczny
Durbin-Watson test	0.07722	0.6092
Breusch-Godfrey test	0.1549	0.731

Jak możemy zauważyć w tabelce powyżej wyniki testów w obu modelach wskazują, iż oba modele posiadają niezależne reszty. Niemniej jednak w początkowym modelu p-value w teście Durbin Watson jest wartością graniczną, natomiast w końcowym modelu nie jest.

6.8 Odstające obserwacje



Jak możemy zauważyć na wykresie z lewej strony, pierwszy model miał jedną bardzo wpływową obserwację, dwie wpływowe w mniejszym stopniu oraz resztę o małym wpływie. W modelu końcowym natomiast możemy zauważyć, iż miara Cooka największej obserwacji jest w okolicy 0.25, reszta obserwacji natomiast nie jest bez znaczenia tylko ma wpływ na model.

6.9 Wnioski

Końcowy model lepiej opisuje zjawisko, wszystkie zmienne są istotne statystycznie, nie posiada zmiennych o zbyt dużym wpływie na model oraz spełnia wszystkie założenia modelu regresji liniowej.

7 Analiza niematematyczna

Model mówi, iż wraz z **wzrostem** populacji, powierzchni, oczekiwanej długości życia oraz stosuneku mężczyzn do kobiet **wzrasta** łączny wynik na IMO. Z kolei **wzrost** długości szkoły podstawowej **zmniejsza** łączny wynik na IMO. Populacja ma logarytmiczny wpływ na model, czyli dla dużych wartości nie ma zbyt dużego wpływu na wynik. Pozwala to na uwzględnienie wpływu populacji dla krajów o jej małej ilości przy jednoczesnym braku faworyzacji krajów o dużym zaludnieniu.

8 Podsumowanie

Stworzony przez nas model spełnia wszystkie założenia standardowego modelu regresji liniowej oraz opisuje w satysfakcjonującym stopniu wpływ zmiennych na łączny wynik kraju na IMO, lecz nie na tyle, aby można było wykonać predykcję wyników. Biorąc to wszystko pod uwagę mogę stwierdzić, iż model **jest dobrze dopasowany**.