Analiza regresji liniowej ilości zdobytych punktów na IMO 2015

1 Wstęp

W swojej pracę badam zależności pomiędzy ilością całkowitej liczby punktów zdobytych przez kraje na Międzynarodowej Olimpiadzie Matematycznej (IMO) w 2015 roku, a szeregiem 10 zmiennych objaśniających. Analizuję różne wskaźniki ekonomiczne, społeczne i edukacyjne dla poszczególnych krajów w celu zidentyfikowania czynników, które mogą mieć wpływ na wyniki matematyczne na IMO.

2 Podstawowe charakterystyki danych

- *imo2015_tasks_done*: Zmienna objaśniana. Opisuje ilość zdobytych punktów przez kraje na IMO 2015, co jest miarą ich sukcesu w konkursie.
- *GDP*: PKB kraju. Ten wskaźnik ekonomiczny odzwierciedla łączną wartość dóbr i usług wytwarzanych w kraju w ciągu roku. Służy jako miara wielkości i rozwoju gospodarczego kraju
- *population*: Populacja. Ta zmienna reprezentuje liczbę mieszkańców danego kraju.
- *high_tech_exports*: Eksport wysokich technologii($). Ta zmienna odnosi się do wartości eksportu zaawansowanych technologii przez dany kraj. Jest to wskaźnik, który może świadczyć o stopniu innowacyjności i zaangażowania kraju w sektor wysokich technologii.
- *migration*: Ilość migrantów. Ta zmienna wskazuje liczbę osób migrujących do danego kraju.
- *gov_educ_expenditure*: Procent wydatków publicznych na rozwój szkół (% od PKB).
- *area*: Powierzchnia kraju. Ta zmienna odnosi się do obszaru zajmowanego przez dany kraj.
- *prob_of_death*: Wskaźnik prawdopodobieństwa zginięcia. Ten wskaźnik wskazuje na poziom bezpieczeństwa i stabilności społecznej w danym kraju.
- *internet_user_percentage*: Wskaźnik dostępu do internetu. Ten wskaźnik odzwierciedla odsetek populacji, który ma dostęp do Internetu.
- *gross_enrollment_ratio*: Wskaźnik skolaryzacji. Ten wskaźnik mierzy stopień uczestnictwa w systemie edukacyjnym, które uczą się w stosunku do ogólnej populacji.
- *unemployment_rate*: Stopa bezrobocia. Ten wskaźnik mierzy odsetek osób bez pracy w stosunku do aktywnej siły roboczej.

Dane dla zmiennych objaśniających, zostały pozyskane z World Bank Open Data na rok 2015. Natomiast dane dotyczące zmiennej objaśnianej, czyli zdobytych punktów przez kraje, zostały uzyskane ze strony IMO 2015
Przedstawione podsumowanie danych(summary) zawiera informacje na temat próbki danych składającej się z 59 obserwacji. Wartości minimalne, maksymalne, mediany, średnie oraz kwartyle wskazują, że dane mają zróżnicowany zakres i rozrzut. Na przykład, *GDP* waha się od 1.055×10^{10} do 4.445×10^{12}, a *population* od 3.308×10^5 do 1.323×10^9. W naszym zbiorze danych mamy zarówno zmienne wyrażone jako liczby, np. *GDP*, które reprezentują konkretne wartości ekonomiczne, jak i zmienne wyrażone jako proporcje, np. *prob_of_death*, które są skalowane w zakresie od 0 do 1. W analizowanych danych nie występują braki danych.

3 Model liniowy w oparciu na dane

Residual Standart Error(RSE) wynosi 30.32, co oznacza, że błąd standardowy residuów wynosi około 30.32 jednostek. W przypadku naszej zmiennej objaśnianej, która przyjmuje wartości od 0 do 151, ta wartość RSE jest relatywnie wysoka, co wskazuje na pewien stopień niedopasowania modelu do danych.
Multiple R-squared wynosi 0.5513, co oznacza, że 55.13% zmienności Y jest wyjaśniane przez zmienne niezależne w tym modelu. Pozostałe 44.87% zmienności Y nie jest uwzględnione w tym modelu i może być spowodowane przez inne czynniki, błędy losowe lub nieuwzględnione zmienne.
Adjusted R-squared wynosi 0.4579, co oznacza, że 45.79% zmienności Y jest wyjaśniane przez zmienne niezależne w tym modelu, uwzględniając liczbę zmiennych niezależnych i stopnie swobody.
F-statistic wynosi 5.899, a p-value wynosi 9.57e-06. Na podstawie tych wyników możemy odrzucić hipotezę zerową na poziomie istotności 0.05. Hipoteza zerowa zakłada, że wszystkie współczynniki regresji są równe zero, co sugeruje brak wpływu zmiennych niezależnych na zmienną objaśnianą. Jednak ze względu na niską wartość p-value, mamy silne przekonanie, że przynajmniej jedna z niezależnych zmiennych ma istotny wpływ na zmienną objaśnianą. Oznacza to, że ten model regresji jest istotny i przynajmniej jedna z niezależnych zmiennych ma wpływ na zmienną objaśnianą

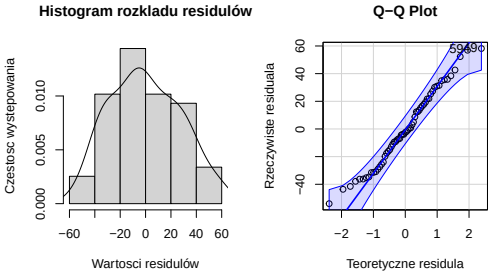
4 Analiza współczynników modelu

$$imo2015\_tasks\_done = (\mathbf{119.6}) + (\mathbf{-7.099 \cdot 10^{-12}}) \cdot GDP + (\mathbf{3.050 \cdot 10^{-8}}) \cdot population + (\mathbf{7.167 \cdot 10^{-10}}) \cdot high\_tech\_exports$$
$$+(\mathbf{1.301 \cdot 10^{-5}}) \cdot migration + (\mathbf{-490.9}) \cdot gov\_educ\_expenditure + (\mathbf{7.179 \cdot 10^{-6}}) \cdot area + (\mathbf{-796.1}) \cdot prob\_of\_death +$$
$$(\mathbf{-44.02}) \cdot internet\_user\_percentage + (\mathbf{32.02}) \cdot gross\_enrollment\_ratio + (\mathbf{-21.39}) \cdot unemployment\_rate$$

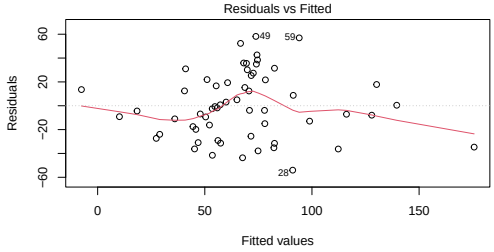
- *Intercept*: Wartość zmiennej objaśnianej, gdy wszystkie zmienne niezależne są równe zeru. W tym przypadku wynosi 119.6. Jest statystycznie istotnym współczynnikiem(p-value<0.05), co oznacza, że różni się znacząco od zera.
- *GDP*: Wpływ PKB na osiągnięcia podczas IMO 2015. Nie jest statystycznie istotną zmienną, ponieważ wartość p-value wynosi więcej niż 0.05.
- *population*: Wpływ populacji na osiągnięcia. Nie jest statystycznie istotną zmienną, ponieważ wartość p-value wynosi więcej niż 0.05.
- *high_tech_exports*: Wpływ eksportu wysokich technologii na osiągnięcia. Jest statystycznie istotną zmienną, ponieważ wartość p-value wynosi mniej niż 0.05. W tym przypadku wartość wynosi 7.167×10^{-10}. Interpretacja niematematyczna: jeśli zwiększymy eksport wysokich technologii o jednostkę w kraju, spodziewamy się, że ilość zdobytych punktów wzrośnie o 7.167×10^{-10}.
- *migration*: Wpływ migracji na osiągnięcia. Nie jest statystycznie istotną zmienną, ponieważ wartość p-value wynosi więcej niż 0.05, co sugeruje, że nie ma statystycznie istotnego wpływu na zmienną objaśnianą.
- *gov_educ_expenditure*: Wpływ wydatków publicznych(% od PKB) na rozwój szkół na osiągnięcia. Nie jest statystycznie istotną zmienną, ponieważ wartość p-value wynosi więcej niż 0.05
- *area*: Wpływ powierzchni kraju na osiągnięcia. Jest statystycznie istotną zmienną, ponieważ wartość p-value wynosi mniej niż 0.05. W tym przypadku wartość wynosi 7.179×10^{-6}. Interpretacja niematematyczna: jeśli zwiększymy powierzchnie kraju o $1km^2$ spodziewamy się, że ilość punktów wzrośnie o 7.179×10^{-6}
- *prob_of_death*: Wpływ prawdopobobieństwa śmieri w kraju na osiągnięcia. Jest statystycznie istotną zmienną, ponieważ wartość p-value wynosi mniej niż 0.05. W tym przypadku wartość wynosi -796.1. Interpretacja niematematyczna: jeśli zwiększymy prawdopodobieństwo zginięcia w kraju o 1% spodziewamy się, że ilość punktów spadznie o 7.961
- *internet_user_percentage*: Wpływ dostępności internetu na osiągnięcia. Nie jest statystycznie istotną zmienną, ponieważ wartość p-value wynosi więcej niż 0.05
- *gross_enrollment_ratio*: Wpływ wskaźnika skolaryzacji na osiągnięcia. Nie jest statystycznie istotną zmienną, ponieważ wartość p-value wynosi więcej niż 0.05
- *unemployment_rate*: Wpływ wskaźnika bezrobocia na osiągnięcia. Nie jest statystycznie istotną zmienną, ponieważ wartość p-value wynosi więcej niż 0.05

5 Dalsza analiza modelu

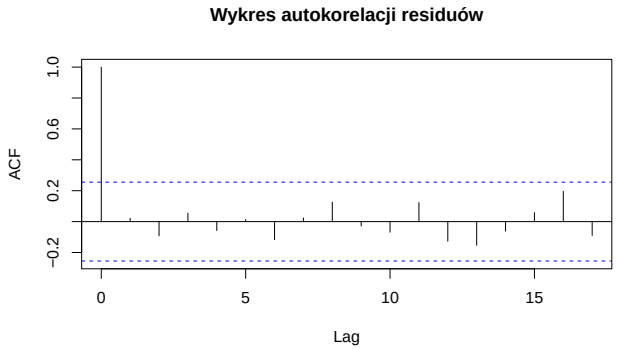
- Współliniowość: wartości VIF wszystkich zmiennych są poniżej 5, co wskazuje na brak lub niewielką obecność współliniowości między zmiennymi niezależnymi. Jednak wartość kappa, przekraczająca 1000, sugeruje potencjalną współliniowosc. Dodatkowo, macierz korelacji danych ujawnia wysoką korelację (powyżej 0.7 w wartościach bezwzględnych) między parami zmiennych: "*GDP*" - "*high_tech_exports*", "*internet_user_percentage*" - "*gross_enrollment_ratio*", oraz "*internet_user_percentage*" - "*prob_of_death*". W celu uniknięcia potencjalnej współliniowości, można rozważyć usunięcie zmiennych "*GDP*", "*internet_user_percentage*" oraz "*gross_enrollment_ratio*" (uwzględniając również korelację między "*gross_enrollment_ratio*" a "*prob_of_death*", która wynosi -0.59) z modelu.



- Normalny rozkład błędów: Po sprawdzeniu histogramu oraz wykresu QQ-Plot można zauważyć, że rozkład błędów w modelu jest bardzo zbliżony do rozkładu normalnego. Dodatkowo, przeprowadzone testy normalności (ad.test, shapiro.test, lillie.test) wykazują, że pvalue są większe od 0.05. Na podstawie tych wyników można wnioskować, że założenie o normalności rozkładu błędów jest spełnione.



- Homoskedastycnośc: Po przeprowadzeniu testu Breuscha-Pagana, testu przyczynowości Grangera, testu Harrison-McCabe oraz analizie wykresu residułów vs wartości przewidywane, można stwierdzić, że istnieją dowody na heteroskedastyczność błędów w modelu. Wartości p-value testów Grangera oraz Harrison-McCabe są większe od 0.05, natomiast p-value testu Breuscha-Pagana jest równy 0.05 więc odrzucamy hipotezę zerową o homoskedastyczności błędów w modelu. Dodatkowo, wykres residułów nie pokazuje równomiernego rozproszenia wokół linii y=0, co sugeruje istnienie heteroskedastyczności



- Autokorelacja: Po przeprowadzeniu testów Durbina-Watsona, Ljung–Boxa oraz analizie wykresu funkcji autokorelacji (ACF) reszt modelu, można stwierdzić, że nie ma istotnych dowodów na występowanie autokorelacji. Wartości p-value większe niż 0.05 sugerują brak statystycznie istotnej autokorelacji. Ponadto, wykres ACF przedstawia wartości błędów modelu znajdujące się głównie w niebieskim przedziale, co też wskazuje na brak znaczącej autokorelacji.

- Liniowa struktura: Po przeprowadzeniu testu Rainbow oraz Harvey-Colliera otrzymujemy p-value większe od 0.05, co sugeruje brak istotnych dowodów przeciwko założeniu liniowości w modelu. Ale Ramsey's RESET zwraca pvalue równe 0.024, co oznacza to, że istnieją statystycznie istotne dowody przeciwko liniowości w modelu. Na podstawie tych wyników można stwierdzić, że istnieją pewne dowody na nieliniowość w modelu.

  W analizie modelu można stwierdzić, że istnieją pewne dowody na nieliniowość w relacjach między zmiennymi objaśniającymi a zmienną objaśnianą. Nie ma istotnych dowodów na występowanie autokorelacji w błędach modelu. Wykazano obecność heteroskedastyczności błędów. Rozkład błędów jest zbliżony do rozkładu normalnego. Wartości VIF są mniejsze od 5, co wskazuje na brak istotnej współliniowości. Analiza ta sugeruje potrzebę dalszej modyfikacji lub ulepszenia modelu.

6 Identyfikacja nietypowych obserwacji

W trakcie identyfikacji nietypowych obserwacji w modelu, obliczyłem dźwignie oraz statystyki Cook'a dla wszystkich obserwacji. Wyniki wskazują, że obserwacje 27, 32, 46 i 50 mają wysoką dźwignię(przekraczając średnią wartość dźwigni plus dwukrotność odchylenia standardowego), natomiast obserwacje 27 i 46 mają wysoką statystyka Cook'a. Jednakże, po dokładniejszej analizie, nie ma wyraźnych dowodów na to, że te obserwacje są całkowicie nierealne. W związku z tym, nie widzę konieczności usuwania tych obserwacji.

7 Redukowanie nieistotnych zmiennych

Przeprowadzając eliminację wsteczną, otrzymujemy model z zmiennymi: **high_tech_exports, area, prob_of_death i internet_user_percentage**.
W przypadku eliminacji na podstawie kryterium AIC oraz BIC, otrzymujemy taką samą listę zmiennych istotnych jak w przypadku eliminacji wstecz. Oznacza to, że ich usunięcie powoduje wzrost AIC oraz BIC oraz to, że te zmienne wnoszą istotny wkład w model.
Interpretując te wyniki, można stwierdzić, że zmienne *high_tech_exports, area, prob_of_death* i *internet_user_percentage* są istotne w wyjaśnianiu zmienności zmiennej objaśnianej (*imo2015_tasks_done*). Pozostałe zmienne (*gov_educ_expenditure, gross_enrollment_ratio, migration, population, unemployment_rate, GDP*) nie wnoszą istotnego wkładu w model i można je pominąć.

8 Przejście do drugiego modelu

Po odrzuceniu zmiennych nieistotnych, przechodzimy do drugiego modelu:

$$imo2015\_tasks\_done = (\mathbf{116.7}) + (\mathbf{5.798 \cdot 10^{-10}}) \cdot high\_tech\_exports + (\mathbf{7.880 \cdot 10^{-6}}) \cdot area +$$
$$+(\mathbf{-908.9}) \cdot prob\_of\_death + (\mathbf{-45.37}) \cdot internet\_user\_percentage$$

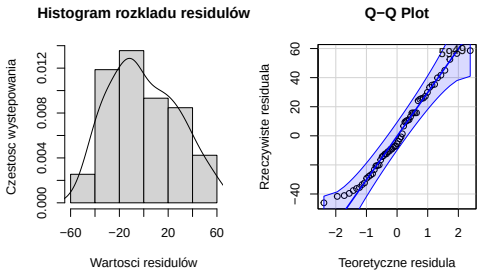
Wartości współczynników w drugim modelu nie uległy znaczącej zmianie w porównaniu do poprzedniego modelu. Głównym problemem drugiego modelu jest niewielkie naruszenie założenia normalności residuów (p-value testu Anderlinga-Darlinga oraz testu Shapiro-Wilka wynoszą odpowiednio 0.042 i 0.049). W celu poprawy tego aspektu(oraz potencjalnemu polepszeniu wartości AIC), przechodzimy do trzeciego modelu.

9 Przejście do trzeciego modelu

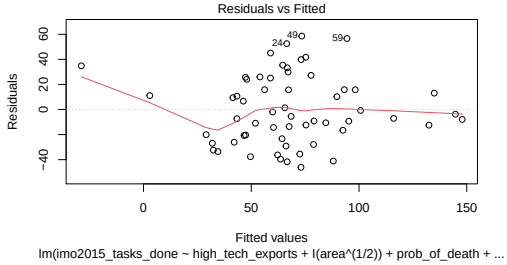
$$imo2015\_tasks\_done = (\mathbf{136.4}) + (\mathbf{5.207 \cdot 10^{-10}}) \cdot high\_tech\_exports + (\mathbf{0.02994}) \cdot \sqrt{area} +$$
$$+(\mathbf{-1024}) \cdot prob\_of\_death + (\mathbf{-28.74}) \cdot e^{internet\_user\_percentage}$$

Wszystkie zmienne są statystycznie istotne, przechodzimy więc do sprawdzania założeń modelu:

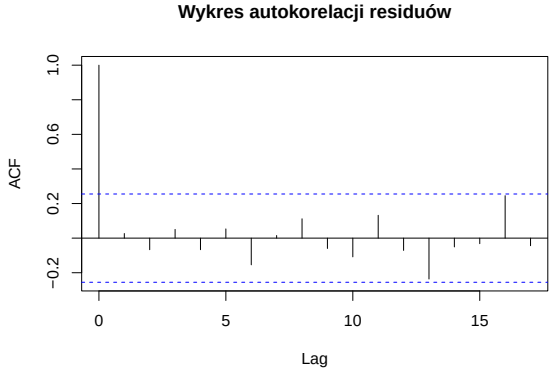
- Współliniowość: Wcześniej nie mieliśmy dużych problemów z występowaniem współliniowości, a w przypadku drugiego modelu sytuacja pozostaje niezmieniona. Wartości VIF dla zmiennych objaśniających w trzecim modelu też są mniejsze niż 5, co wskazuje na brak znaczącej współliniowości między zmiennymi.



- Normalny rozkład błędów: Wcześniej nie napotkaliśmy problemu z założeniem dotyczącym normalności rozkładu reszt, a teraz sytuacja również jest korzystna. Przeprowadzone testy na normalność zwracają wartości p-value znacznie większe od 0.05 lub bardzo bliskie tej wartości (np. test Anderlinga-Darlinga - 0.21, test Shapiroa-Wilka - 0.12, test Kolmogorova-Smirnova - 0.227). Ponadto, histogram residułów oraz QQ-plot przedstawiają rozkład reszt, który jest zgodny z rozkładem normalnym. Na podstawie tych wyników można wnioskować, że założenie dotyczące normalności rozkładu residułów jest spełnione.



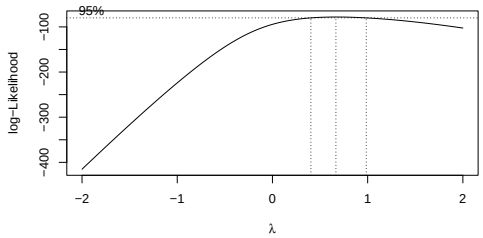
- Homoskedastycnośc: Wcześniej mieliśmy problem z homoskedastycznością, o czym świadczyło p-value testu Breuscha-Pagana wynoszące 0.03. Teraz jednak, p-value wszystkich przeprowadzonych testów (Breuscha-Pagana, testu przyczynowości Grangera oraz testu Harrison-McCabe) są większe od 0.05(p-value testu Breuscha-Pagana wynosi teraz 0.08, co jest niewielką ale pozytywną różnicą w porównaniu do wartości 0.05). Natomiast, warto zauważyć, że wykres residuów nie wykazuje równomiernego rozproszenia wokół linii y=0, co sugeruje obecność heteroskedastyczności. Niemniej jednak, ponieważ p-value wszystkich testów jest większe od 0.05, możemy przyjąć założenie o homoskedastyczności.



- Autokorelacja: Pod względem autokorelacji, wcześniej nie napotkaliśmy na żadne problemy, a sytuacja pozostaje niezmieniona. Wartości p-value z testów Durbina-Watsona oraz Ljung–Boxa na autokorelację są większe od 0.05, co wskazuje na brak istotności autokorelacji. Analiza za pomocą wykresu ACF potwierdza te wyniki, gdzie wszystkie wartości autokorelacji zawierają się w przedziałach ufności. Możemy więc stwierdzić, że nie występuje problem autokorelacji w trzecim modelu.

- Liniowa struktura: Wcześniej zaobserwowano problem z liniową strukturą modelu na podstawie p-value testu Ramsey's RESET, które wynosiło 0.024, sugerując nieliniową relację. Jednak teraz p-value tego testu wynosi 0.16. Dodatkowo, wyniki pozostałych testów(Rainbow oraz Harvey-Colliera) mają p-value istotnie większe od 0.05. Można zatem wnioskować, że model spełnia założenie o liniowej strukturze
  Podsumowując: w niektórych przypadkach założenia uległy poprawie, na przykład p-value testu Breuscha-Pagana wzrosło z 0.03 do 0.08, co poprawia homoskedastyczność, a wartość testu Ramsey's RESET wzrosła z 0.024 do 0.16, co sugeruje większe dostosowanie do założenia o liniowej strukturze. W innych przypadkach założenia nie uległy istotnej zmianie, jak w przypadku autokorelacji, współliniowości oraz normalności rozkładu błędów.

10 Próba przejścia do czwartego modelu

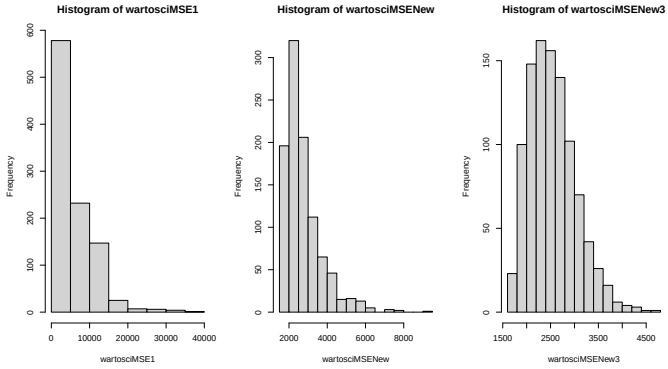


Próbując polepszyć homoskedastyczność trzeciego modelu, zastosowano transformację Boxa-Coxa. Jednak, na podstawie wykresu funkcji log-wiarygodności, można zauważyć, że 1 albo wpada w przedział ufności, albo jest bardzo bliska (nie jest to jednoznacznie widoczne). W związku z tym, nie widzę sensu stosować tą transformację.

11 Porównywanie modeli

W tym dziale przeprowadzimy porównanie w głownej mierze między trzecim a pierwszym modelem. Skoncentrujemy się na ocenie kilku ważnych miar jakości modelu, takich jak R-kwadrat, skorygowany R-kwadrat (adjusted R-kwadrat), kryterium informacyjne Akaike (AIC), kryterium informacyjne Bayesa (BIC) oraz przeprowadzimy walidację krzyżową.

- Zauważamy, że wartości R^2, AIC i BIC praktycznie się nie zmieniły. Na przykład, AIC wynosiło 582 w pierwszym modelu, a teraz wynosi 569, podobnie BIC zmieniło się z 607 do 582. Jednak, co ciekawe, skorygowany R^2 w trzecim modelu (0.52) jest wyższy niż w pierwszym modelu (0.45). Oznacza to, że trzeci model lepiej uwzględnia liczbę zmiennych niezależnych i dostarcza bardziej skorygowanej miary dopasowania do danych.
- W celu pełniejszej analizy, przeprowadziłem również porównanie z drugim modelem, aby mieć pełniejszy obraz różnic między trzema modelami, ale po przejrzeniu się wartościam AIC, BIC oraz skorygowanego R^2, zauważamy, że te wyniki mieszczą się pomiędzy pierwszym a trzecim modelem.



- Po przeprowadzeniu walidacji krzyżowej, analizując histogramy wartości błędu średniokwadratowego (MSE), zauważamy, że histogram dla trzeciego modelu jest najbardziej przesunięty w lewo. Oznacza to, że ten model osiąga najmniejszą wartość błędu na zbiorze testowym w porównaniu z innymi modelami. Dodatkowo, obliczając średnią z wszystkich wartości MSE dla zbiorów testowych, obserwujemy, że trzeci model posiada najniższą średnią wartość błędu. Te wyniki wskazują na lepszą predykcyjną wydajność trzeciego modelu w porównaniu z innymi modelami poddawanymi walidacji krzyżowej.

12 Podsumowanie

Podsumowując ten projekt, przeprowadziłem analizę za pomocą modelu regresji liniowej, aby zrozumieć, jak różne czynniki opisujące państwa biorące udział w Międzynarodowej Olimpiadzie Matematycznej w 2015 roku wpływają na ilość rozwiązanych zadań. Badane zmienne obejmowały czynniki takie jak PKP, populacja, eksport wysokich technologii, ilość migrantów, procent wydatków publicznych na rozwój szkół, powierzchnia kraju, wskaźnik prawdopodobieństwa zginięcia, wskaźnik dostępu do internetu, wskaźnik skolaryzacji oraz stopa bezrobocia.

Większość zmiennych objaśniających nie miała istotnego wpływu na wynik w olimpiadzie. Przeprowadzając redukcję zmiennych nieistotnych, model został zoptymalizowany i zawierał zmienne takie jak eksport wysokich technologii, powierzchnię kraju, wskaźnik prawdopodobieństwa zginięcia oraz wskaźnik dostępu do internetu. Okazało się, że po zastosowaniu pierwiastka z powierzchni kraju oraz zamianie wskaźnika dostępu do internetu na jego eksponencjalną postać: exp(wskaźnik dostępu do internetu), model osiągnął lepsze wyniki w testach statystycznych oraz wykazywał poprawę w miarę różnych wskaźników oceny, takich jak AIC, BIC, skorygowany R-kwadrat oraz przy porównywaniu krzyżowym. To oznacza, że transformacje te przyczyniły się do lepszego dopasowania modelu do danych, a także poprawy jego jakości predykcyjnej.

Zwiększając eksport wysokich technologii oraz powierzchnię kraju, można się spodziewać wzrostu wyniku na olimpiadzie. Jest intuicyjnie, że większy eksport wysokich technologii może mieć korzystny wpływ na wynik, ponieważ świadczy o rozwiniętym sektorze naukowo-technologicznym, który może wpływać na lepsze kształcenie matematyczne. Z kolei większa powierzchnia kraju, jako wskaźnik rozwoju gospodarczego, może również pozytywnie wpływać na wynik, ze względu na większe zasoby i możliwości edukacyjne.

Natomiast przy zmniejszaniu wskaźnika prawdopodobieństwa zginięcia można oczekiwać zwiększenia ilości rozwiązanych zadań na olimpiadzie. Niższe ryzyko śmierci może przyczynić się do większej koncentracji i lepszych wyników uczestników. Warto zauważyć, że współczynnik przy zmiennej exp(internet_user_percentage) jest ujemny, co może wydawać się nieintuicyjne. Może to sugerować, że w przypadku tego konkretnego modelu większy odsetek użytkowników internetu nie jest bezpośrednio związany z wyższymi wynikami na olimpiadzie matematycznej. Jednak należy mieć świadomość, że opisane wyniki są jedynie częścią szerszego obrazu. Istnieje wiele innych czynników, które mogą odgrywać istotną rolę w osiągnięciach uczestników, jednak nie zostały uwzględnione w analizowanym modelu