Badanie czynników wpływających na rezultat poszczególnych państw na Międzynarodowej Olimpiadzie Matematycznej

Jakub Sola

30 czerwca 2023

1 Opis projektu

Każdego roku najlepsi reprezentanci swoich państw biorą udział w Międzynarodowej Olimpiadzie Matematycznej. Projekt ma na celu zbadanie zależności pomiędzy ich wynikami, a wskaźnikami rozwoju tych państw. Dokonana analiza pozwoli stwierdzić, które czynniki i w jakim stopniu wpływają pośrednio na końcowy rezultat drużyn na Olimpiadzie. Dzięki temu możliwe będzie dokonanie predykcji i oszacowanie wyników przyszłych konkursów, a także wskazanie kierunku rozwoju państw mogącego zwiększyć ich szansę na odniesienie sukcesu.

2 Dane

Projekt zakłada badanie zależności między zmienną objaśnianą `total` - sumaryczą liczbą punktów zdobytą przez zawodników z danego państwa w 2021 roku na Międzynarodowej Olimpiadzie Matematycznej, a 10 zmiennymi objaśniającymi:

- `GDP` - PKB per capita [US$] (2021),
- `Pop` - populacja [mln] (2021),
- `PCR` - współczynnik ukończenia szkoły podstawowej [% odpowiedniej grupy wiekowej] (2020/2021),
- `Urban` - ludność zamieszkująca obszary miejskie [mln] (2021),
- `WiP` - procentowy udział kobiet w parlamencie (2021),
- `Area` - powierzchnia [1000 km^2] (2021),
- `Urb_area` - powierzchnia obszarów miejskich [1000 km^2] (2021),
- `Internet` - ludność z dostępem do internetu [mln] (2021),
- `Roads` - całkowita długość sieci drogowej [1000 km] (2018-2021),
- `Unemp` - bezrobocie [%] (2021).

Badanie przeprowadzono na 106 krajach rywalizujących w 62 Międzynarodowej Olimpiadzie Matematycznej w 2021 roku. Podstawowe statystyki wszystkich zmiennych odczytujemy po wprowadzeniu komendy `summary`:

```
> summary(DATA)
   Country             Code               Total             GDP                Pop                PCR              Urban       
 Length:106         Length:106         Min.   :  0.00   Min.   :   533.4   Min.   :   0.3725   Min.   : 52.67   Min.   :  0.350  
 Class :character   Class :character   1st Qu.: 27.00   1st Qu.:  4387.3   1st Qu.:   5.2175   1st Qu.: 94.35   1st Qu.:  3.232  
 Mode  :character   Mode  :character   Median : 61.00   Median : 10505.0   Median :  11.8362   Median : 98.89   Median :  8.515  
                                       Mean   : 67.46   Mean   : 22111.5   Mean   :  64.4505   Mean   : 96.50   Mean   : 37.909  
                                       3rd Qu.:105.00   3rd Qu.: 32368.5   3rd Qu.:  45.8425   3rd Qu.:101.87   3rd Qu.: 32.413  
                                       Max.   :208.00   Max.   :133590.1   Max.   :1412.3600   Max.   :121.92   Max.   :882.894  
      WiP              Area            Urb_area          Internet           Roads              Unemp       
 Min.   : 2.326   Min.   :    0.033   Min.   :  0.0178   Min.   :   0.3292   Min.   :   0.087   Min.   : 0.992  
 1st Qu.:19.467   1st Qu.:   65.370   1st Qu.:  1.5196   1st Qu.:   2.9269   1st Qu.:  27.216   1st Qu.: 4.513  
 Median :26.885   Median :  238.465   Median :  3.5188   Median :   7.8436   Median :  84.140   Median : 6.180  
 Mean   :27.846   Mean   : 1028.343   Mean   : 17.2230   Mean   :  45.1764   Mean   : 372.404   Mean   : 7.527  
 3rd Qu.:38.329   3rd Qu.:  619.262   3rd Qu.: 13.6057   3rd Qu.:  32.6395   3rd Qu.: 222.957   3rd Qu.: 9.463  
 Max.   :61.250   Max.   :17098.250   Max.   :522.3452   Max.   :1051.1400   Max.   :6803.479   Max.   :28.770  
```

Przed stworzeniem modelu regresji liniowej przyjrzyjmy się fragmentowi macierzy korelacji w celu zbadania zależności między zmienną objaśnianą a pozostałymi danymi.

	Total	GDP	Pop	PCR	Urban	WiP	Area	Urb_area	Internet	Roads	Unemp
Total	1.00000000	0.12440838	0.32457287	0.28980227	0.40564925	-0.02997987	0.46277477	0.42174851	0.38257650	0.42199116	-0.27782920

Zmienne `Area`, `Urb_area`, `Roads` i `Urban` są najsilniej skorelowane ze zmienną objaśnianą `total` - ich korelacja jest wyższa niż 0.4. Najniższą korelację ze zmienną `total` - wynoszącą -0.03 - ma zmienna `WiP`. Przewidujemy, że zmienne z najwyższą korelacją odegrają kluczową rolę w naszym modelu.

3 Model liniowy w oparciu o dane

Budujemy model regresji liniowej w oparciu o zmienną objaśnianą `total` i współczynniki objaśniające. Przeprowadzamy podstawową analizę modelu, ponownie korzystając z funkcji `summary`.

```
> model = lm(Total~GDP+Pop+PCR+Urban+WiP+Area+Urb_area+Internet+Roads+Unemp)
> summary(model)
```

3.1 Residua

Zakres residuów wynosi od -63 do 86 z medianą -4.7. Może to sugerować brak normalności rozkładu reszt. Residua mają szeroki przedział i nie są scentralizowane wokół zera. Podejrzewamy, że istnieją pewne wartości odstające lub wpływowe punkty, które oddziałowywują na prognozę modelu. Ujemna wartość mediany wskazuje, że średnio model ma tendencję do przeszacowywania wartości zmiennych.

3.2 Współczynniki

Największy co do wartości bezwzględnej jes wyraz wolny oraz współczynniki przy zmiennych `Internet`, `Unemp`, `Pop` i `Urb_area`. Wyraz wolny ma zdecydowanie największą niepewność statystyczną - jego błąd standardowy wynosi 38. Pozostałe błędy standardowe są w większości przypadków niewielkie i nie przekraczają 0.4. Na poziomie istotności $\alpha = 0.05$ odrzucamy hipotezę zerową o zerowości współczynników dla `Internet`, `Unemp`, `Pop` i `Urb_area`. Te współczynniki mają największe `t value`, możemy więc stwierdzic, że są najistotniejsze statystycznie w naszym modelu. Zmienna `Roads`, z współczynnikiem i błędem standardowym bliskim 0 oraz wysokim `p value`, wydaje się być nieistotna w modelu i możliwa do pominięcia w przyszłych etapach projektu.

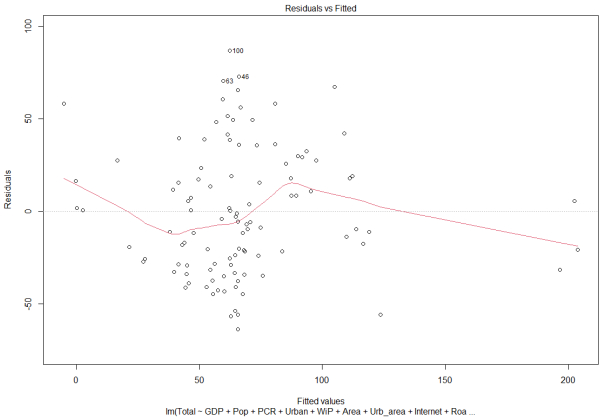
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.8374837 37.8951338   0.708  0.48055
GDP         -0.0001434  0.0001626  -0.881  0.38028
Pop         -1.5659898  0.3320764  -4.716 8.25e-06 ***
PCR          0.4423776  0.3820617   1.158  0.24982
Urban       -0.6914737  0.3826613  -1.807  0.07393 .
WiP          0.3248998  0.3243890   1.002  0.31909
Area         0.0037626  0.0020258   1.857  0.06636 .
Urb_area    -1.0286390  0.3904709  -2.634  0.00984 **
Internet     3.2986764  0.7427950   4.441 2.42e-05 ***
Roads       -0.0016819  0.0077480  -0.217  0.82862
Unemp       -2.0391746  0.8681815  -2.349  0.02091 *
```

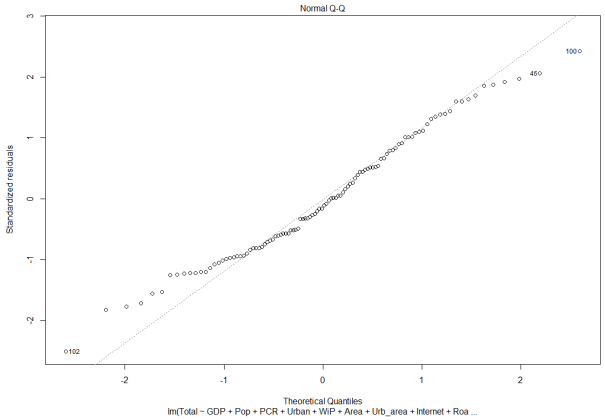
3.3 Błąd standardowy residuów

3.4 R^2

Współczynnik determinacji R^2 na poziomie 0.49 informuje nas, że model jest średnio dopasowany - połowa zmienności zmiennej objaśnianej nie została wyjaśniona przez regresję.

3.5 Wykresy





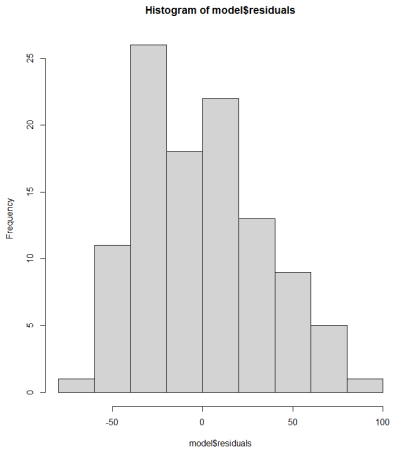
Pierwszy wykres `Residuals vs Fitted values` wskazuje na pewne naruszenie liniowości naszego modelu. Na wykresie `Normal Q-Q` punkty układają się w linię prostą, poza skrajnymi wartościami, więc nie możemy z całą pewnością stwierdzić normalności rozkładu reszt.

3.6 Testy

3.6.1 Liniowość modelu

Badamy liniowość modelu za pomocą trzech testów: `Harvey-Collier test, Rainbow Test oraz RESET Test`. Otrzymane `p values` to kolejno: 0.78, 0.42 i 0.04. Na podstawie `RESET testu` oraz wykresu `Residuals vs Fitted values` **nie można stwierdzić liniowości modelu**. Słusznym podejściem może okazać się przeprowadzenie transformacji zmiennych w modelu w celu ustalenia ich nieliniowych zależności.

3.6.2 Normalność reszt



Z histogramu residuów jasno wynika, że rozkład reszt jest prawostronnie skośny. Wykonamy dodatkowo cztery testy w celu zbadania normalności rozkładu residuów: `Test Shapiro-Wilka, Test Kołmogorowa-Smirnowa, Test Craméra-von Misesa` oraz `Anderson-Darling Test`. W każdym teście otrzymane `p value` nie przekroczyło 0.05, zatem definitywnie **odrzucamy hipotezę zerową o normalności rozkładu reszt**.

3.6.3 Współliniowość zmiennych

Przy pomocy komendy `vif` stwierdzamy, że **zmienne w modelu są współliniowe** - wartości dla `Internet, Pop, Urban` oraz `Urb_area` znacznie przekraczają dopuszczalne 10 (wynoszą kolejno 802, 342, 125 i 38). Korelację między tymi danymi obserwujemy również analizując macierz korelacji. Potencjalnym rozwiązaniem może być usunięcie zależnych zmiennych z modelu.

3.6.4 Homoskedastyczność

Aby przetestować stałość wariancji (homoskedastyczność) w modelu skorzystamy z testów `Breuscha-Pagana`, `Goldfelda-Quandta` i Harrisona-McCabea. We wszystkich testach otrzymaliśmy `p value > 0.45`, więc **zakładamy homoskedastyczność modelu**.

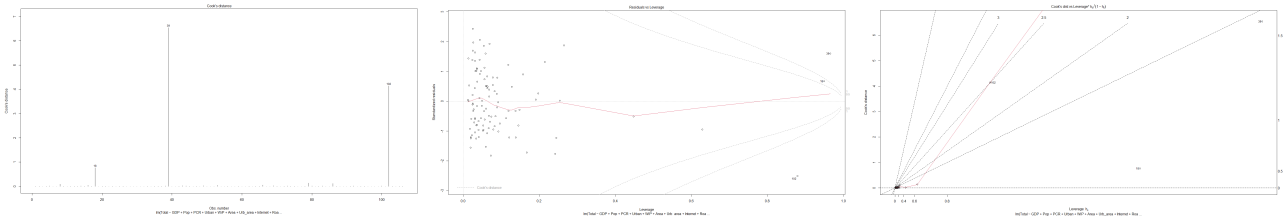
3.6.5 Niezależność reszt

Sprawdzimy autokorelację reszt w naszym modelu, używając w tym celu (Testu Durbina-Watsona oraz `Testu Breuscha-Godfreya`. W rezultacie otrzymaliśmy `p value` równe kolejno 0.63 i 0.60. Co za tym idzie, **nie ma żadnych podstaw do odrzucenia hipotezy o niezależności reszt**.

4 Poprawa modelu

4.1 Obserwacje odstające

Użyjemy 3 wykresów: `Cook's distance, Residuals vs Leverage i Cook's dist vs Leverage`, aby zlokalizować odstające wartości w naszym zbiorze danych.



Zauważamy, że obserwacje 18, 39 oraz 102 (odpowiadające kolejno Chinom, Indiom oraz Stanom Zjednoczonym) wyróżniają się na tle pozostałych.
Dystans Cooka dla Indii jest największy i bliski 7. Wynika to z faktu, że na tle pozostałych państw Indie osiągnęły przeciętny wynik na IMO, biorac pod uwagę olbrzymią populację, powierzchnię, ludność z dostępem do internetu oraz sieć dróg. Z artykułu dostępnego na stronie medium.com jasno wynika, że w Indiach znacznie więcej uwagi poświęca się na inne egzaminy i olimpiady, zatem ich wynik na Międzynarodowej Olimpiadzie Matematycznej jest znacznie zaniżony. W związku z powyższym postanowiłem usunąć tą obserwację.
Dla USA miara Cooka przekroczyła 4. Prawdopodobnie spowodowane jest to faktem, że tak samo jak w przypadku Indii Stany powinny osiągnąć lepszy rezultat na IMO zakładając ich olbrzymią powierzchnię, populację i znaczną przewagę nad innymi państwami pod względem pozostałych czynników. Ich rezultat punktowy wyniósł 165, dzięki czemu plasują się w czołówce stawki, jednak jest to niewiele w porównaniu z np. Chinami, które otrzymały 208 punktów. Na tej podstawie odrzucam również tą obserwację.
W przypadku Chin dystans Cooka również jest duży (w okolicach 1), jednak w tym przypadku najwyższy wynik na Olimpiadzie jest spowodowany również wysokimi wskaźnikami badanych czynników, przez co wyróżnia się spośród innych. Postanowiłem zachować tą obserwację w modelu.

4.2 Upraszczanie modelu

Na początku postaramy się poprawić nasz model usuwając zmienne mało istotne statystycznie, aby osiągnąć lepsze dopasowanie do danych i zmniejszyć złożoność modelu. W tym celu użyjemy funkcji `step()`, która przeprowadzi schodkową

metodę eliminacji zmiennych na podstawie kryterium informacyjnego Akaike (AIC). W wyniku działania funkcji wyeliminowane zostały kolejno zmienne `WiP`, `Area`, `PCR`, `Urb_area` oraz `Urban`. Kryterium AIC zmniejszyło się z 1048.08 do 1042.12, w porównaniu z pierwotnym modelem. Jednocześnie w nowym modelu odnotowaliśmy nieznaczne zmniejszenie się standardowego błędu residuów (o 0.24).

4.3 Usuwanie współliniowości

Wcześniejsze testy wykazały silną zależność między niektórymi zmiennymi objaśniającymi, co zmniejsza wiarygodność naszego modelu. Pierwszym nasuwającym się pomysłem jest pozbycie się najbardziej skorelowanych zmiennych. Po usunięciu zmiennej `Pop`, współczynniki obliczone za pomocą funkcji `vif` mieszczą się w akceptowalnym przedziale - są mniejsze niż 10.

```
> vif(model3)
     GDP Internet    Roads    Unemp
1.146474 6.386452 6.242720 1.132934
```

Po zmianie, uprościliśmy model do 4 zmiennych objasniających: `GDP, Internet, Roadsi Unemp`. W porównaniu do pierwotnego modelu, statystyka R^2 spadła z 0.49 do 0.31, a RSE wzrósł z 36.16 do 40.1 co sprawia, że mimo pozbycia się liniowości (a właściwie zredukowania jej do akceptowalnej), nasz nowy model jest zdecydowanie słabiej dopasowany i gorzej objaśnia zmienną `Total`.
Inną testowaną metodą było połączenie zmiennych `Pop` oraz `Internet` i utworzenie nowej zmiennej `IP = Internet / Pop`, która reprezentuje procent ludności danego kraju z dostępem do internetu. Okazuje się, że zastępując obie zmienne ich kombinacją pozbywamy się współliniowości, a nasz model jest lepiej dopasowany niż ten, który uzyskaliśmy poprzez usunięcie zmiennych - tym razem statystyka R^2 wyniosła 0.41, a standardowy błąd residuów 37.11. Zmienna IP jest istotna statystycznie, a jej współczynnik w modelu wynosci 90.

```
> vif(model4)
     GDP       IP    Roads    Unemp
1.492042 1.468118 1.035255 1.077357
```

Jak widać, wszystkie współczynniki wygenerowane funkcją `vif` oscylują w okolicach 1, co sprawia, że rozwiązaliśmy problem współliniowości zmiennych w naszym modelu.

4.4 Transformacja Boxa-Coxa

Na początku przeprowadziłem transformację Boxa-Coxa zmiennej objaśnianej `Total`, poprzedzając ją zmianą wyniku Botswany z 0 na 0.001, aby móc przeprowadzić przekształcenie, co nie ma widocznego wpływu na model regresji. Dla $\lambda = 0.5454$ i $\hat{Total} = \frac{Total^{\lambda}-1}{\lambda}$ otrzymujemy model ze statystyką R^2 wynoszącą 0.37 i standardowym błędem residuów 6.17. Dodatkowo, wszystkie testy na liniowość modelu nie wskazały podstaw do odrzucenia hipotezy zerowej o liniowości. Jako, że znacząco zmalał wskaźnik RSE oraz rozwiązaliśmy problem liniowości, postanowiłem zachować tą zmianę i dalsze przekształcenia stosować w odniesieniu do ptransformacji Boxa-Coxa.

4.5 Transformacja logarytmiczna

Sprawdziłem, czy model będzie lepiej dopasowany, jeśli zamiast zmiennych objaśniających w modelu użyję funkcji logarytmicznych tych zmiennych. Zastosowanie tej metody dla zmiennej `Unemp` przyniosło oczekiwaną zmianę. Nieznacznie wzrósł współczynnik R^2 i zmalał wskaźnik RSE.

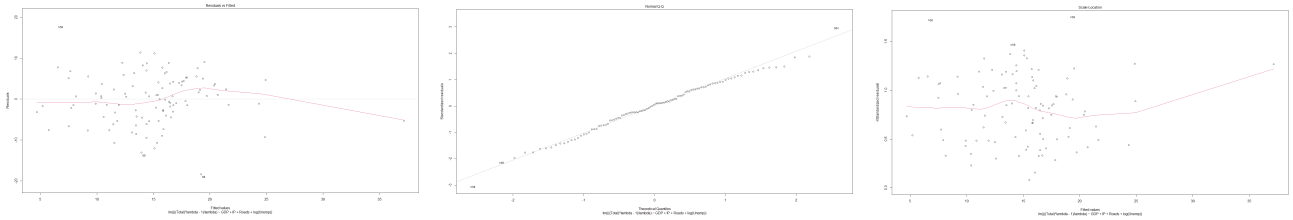
5 Ostateczny model

W wyniku transformacji i działań wykonanych na pierwotnym modelu, końcowym rezultatem naszych działań jest model

$$\hat{Total} = 9.613 - 5.869 \cdot 10^{-5} GDP + 0.1628 IP + 4.023 * 10^{-3} Roads - 3.132 log(Unemp),$$

gdzie $\hat{Total} = \frac{Total^{\lambda}-1}{\lambda}$, $\lambda = 0.5454$, $IP = Internet/Pop$.
Wszystkie zmienne w nowym modelu są istotne statystycznie. Zredukowaliśmy standardowy błąd residuów do 6.14 (w początkowym modelu wynosił on 35.18). Statystyka R^2 zmalała z 0.49 na 0.38, co spowodowane zostało usunięciem znaczącej liczby zmiennych objasniających w modelu. Możemy zatem stwierdzić, że końcowy model słabo objaśnia zmienną `Total`, jednak jest dobrze dopasowany.

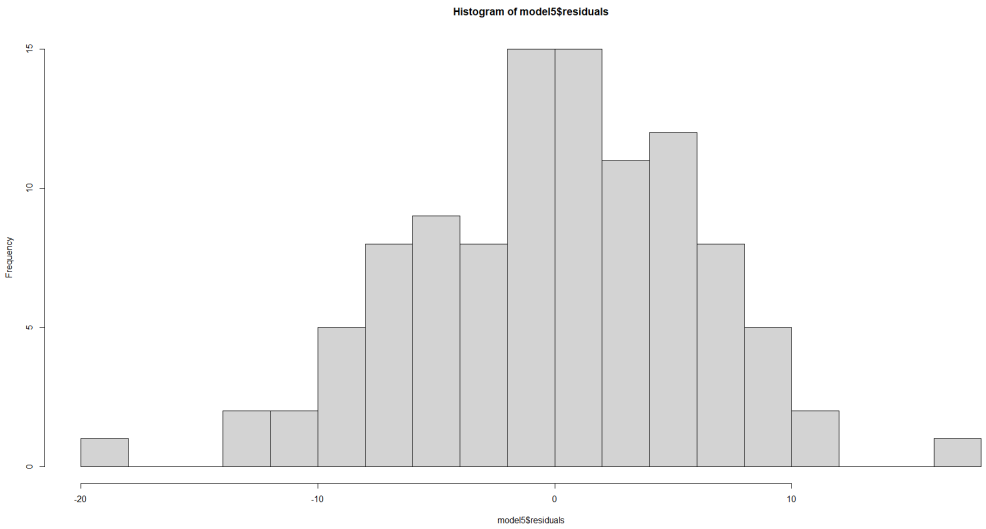
5.1 Wykresy



Jak widać, w porównaniu z pierwotnym modelem, wykres `Residuals vs Fitted values` przedstawia linię dużo bardziej zbliżoną do poziomej osi 0, co pozwala nam stwierdzić poprawę liniowości modelu. Zdecydowana wiełszość punktów na wykresie `Normal Q-Q` leży na linii skośnej, możemy odnotować znaczącą zmianę względem pierwszego modelu. Nie możemy odrzucić hipotezy o normalności rozkładu reszt. Ostatni wykres `Scale-Location` przedstwia linię, która poza skrajną wartością po prawo, przypomina prostą. Nie możemy na podstawie tego wykresu odrzucić hipotezy o homoskedastyczności reszt.

5.2 Testy

Użyjemy tych samych testów co dla pierwotnego modelu w celu zbadania charakterystyk nowego modelu.
Wszystkie testy `Harvey-Collier, Rainbow oraz RESET` wykazały brak podstaw do odrzucenia hipotezy zerowej o liniowości modelu (`p values` to kolejno: 0.75, 0.84 i 0.16), a zatem w połączeniu z analizą wykresu stwierdzamy, że model **spełnia założenie liniowości**.



`P value` dla testu `Shapiro-Wilka` wynosi 0.91, podczas gdy dla testów `Kołmogorowa-Smirnowa`, `Craméra-von Misesa` oraz `Anderson-Darling` są bliskie 0. Na podstawie wykresu `Normal Q-Q`, testu `Shapiro-Wilka` oraz histogramu **stwierdzam normalność reszt** w modelu.
Na podstawie `p values` z testów `Breuscha-Pagana`, `Goldfelda-Quandta` i Harrisona-McCabea, nie mieszczących się w standardowym przedziale ufności, a także analizując wykres `Scale-Location`, nie ma podstaw do odrzucenia hipotezy o **homoskedastyczność modelu**.

Wartości otrzymane po wywołaniu funkcji `vif` oscylują w okolicach 1, co za tym idzie zakładam **brak współliniowości zmiennych**.
Testy `Durbina-Watsona` oraz `Breuscha-Godfreya` nie wykazały podstaw do odrzucenia hipotezy o **braku zależności między resztami**.

6 Interpretacja

Liczba punktów uzyskanych na olimpiadzie przez poszczególne kraje jest wprost proporscjonalna do procentu ludności z dostępem do internetu i logarytmu z poziomu bezrobocia, natomiast maleje wraz ze wzrostem GDP i sieci drogowej.

7 Podsumowanie

Uzyskany model, mimo gorszego objaśniania zmiennej `Total`, jest lepiej dopasowany i spełnia założenia o liniowości, normalności reszt, homoskedastyczności, braku współliniowości zmiennych oraz niezależności residuów. To sprawia, że w mojej ocenie jest lepszy od pierwotnego modelu. Ponadto w nowym modelu nie występują obserwacje odstające. Oczywiście w przyszłości można poprawić model, dobierając inne zmienne objasniające i obierając inne transformacje.