Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie
Wydział Zarządzania
Informatyka i Ekonometria

Zależność oczekiwanej długości życia w krajach świata od wybranych czynników społeczno-gospodarczych

Mateusz Grzelik

18 czerwca 2024

1 Cel projektu i hipotezy badawcze

Celem projektu jest ustalenie, jakie czynniki oraz w jaki sposób wpływają na oczekiwaną długość życia. Wśród potencjalnych regresorów znajdują się wskaźniki demograficzne (np. gęstość zaludnienia), ekonomiczne (PKB per capita), środkowiskowe (Emisja CO2 per capita) oraz geograficzne (położenie w Europie lub Afryce). Główna hipoteza ma postać:

Oczekiwana długość życia jest zależna od pewnych czynników społecznych, gospodarczych, środowiskowych lub geograficznych.

W celu zweryfikowania hipotezy głównej postawimy następujące hipotezy szczegółowe:

1. **Populacja danego kraju wpływa na oczekiwaną długość życia.**
2. **Wartość PKB per capita wpływa pozytywnie na oczekiwaną długość życia.**
3. **Gęstość zaludnienia wpływa na oczekiwaną długość życia.**
4. **Procent populacji zurbanizowanej wpływa pozytywnie na oczekiwaną długość życia.**
5. **Emisja CO2 na mieszkańca wpływa na oczekiwaną długość życia.**
6. **Procent populacji z dostępem do podstawowych usług sanitarnych wpływa pozytywnie na oczekiwaną długość życia.**
7. **Położenie państwa w Europie wpływa pozytywnie na oczekiwaną długość życia.**
8. **Położenie państwa w Afryce wpływa negatywnie na oczekiwaną długość życia.**
9. **Stosunek liczby mężczyzn do kobiet wpływa na oczekiwaną długość życia.**

Jeżeli conajmniej jedna z hipotez okaże się prawdziwa, potwierdzi to hipotezę główną.

Ze względu na ustalony cel projektu (wskazanie zależności między zmiennymi), większą uwagę będziemy przykładać do tego jak dobrze model jest dopasowany do danych, niż do efektywności pod względem prognozowania.

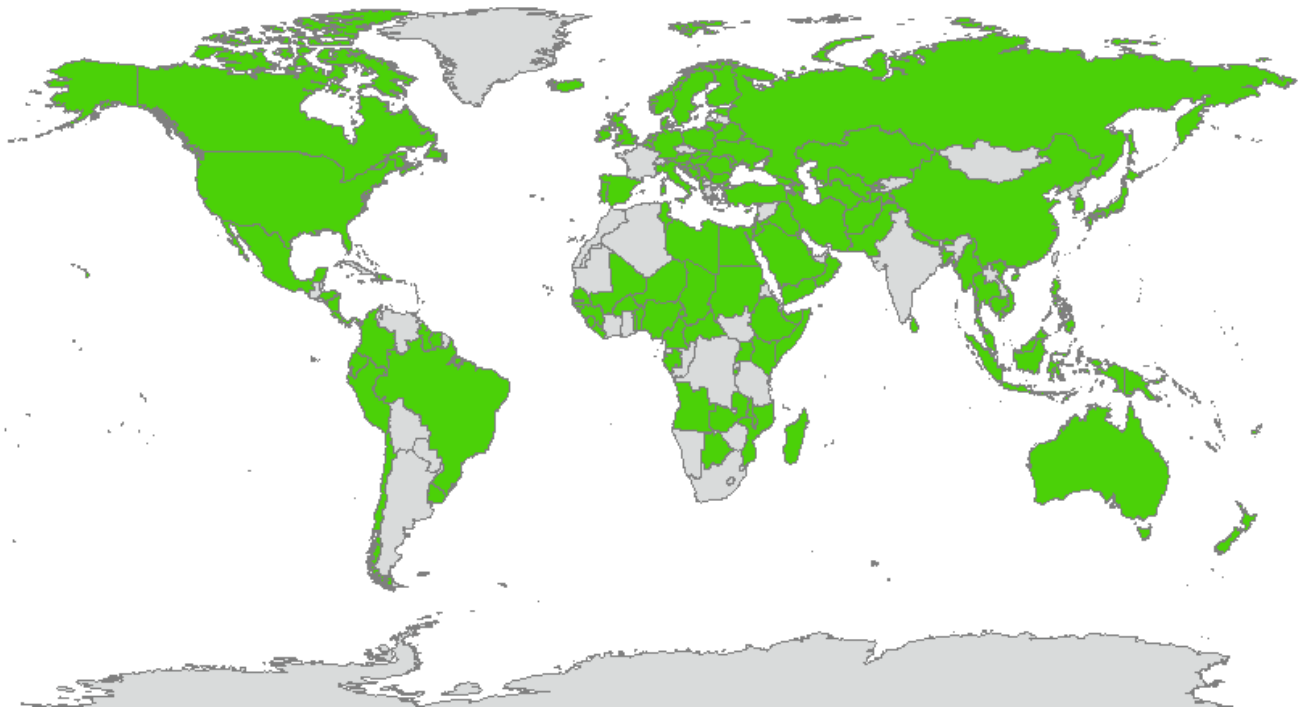
2 Źródła oraz opis danych

Dane pochodzą z następujących witryn: GapMinder, World Bank oraz Wikpedia. Poniżej znajdują się bezpośrednie linki do źródeł poszczególnych zmiennych. Zbiór opisuje 156 państw ze wszystkich kontynentów.

Zbiór danych zawiera 11 zmiennych:

- **Oczekiwana długość życia** - zmienna objaśniana. Wyrażona w latach z dokładnością do części dziesiątych. Dane pochodzą z roku 2019, z bazy danych GapMinder.
- **Populacja** - stan na rok 2019, dane pochodzą z bazy danych GapMinder.
- **PKB per capita w aktualnych dolarach** - stan na rok 2019, dane pochodzą z bazy danych World Bank.
- **Gęstość zaludnienia** - stan na rok 2019, ilość ludzi przypadająca na kilometr kwadratowy. Dane pochodzą z bazy danych World Bank.
- **Zurbanizowany procent populacji** - stan na rok 2019. Procent populacji zamieszkały w miastach. Dane pochodzą z bazy danych World Bank.
- **Emisja CO2 per capita** - stan na rok 2019, wyrażona w tonach na osobę. Dane pochodzą z bazy danych World Bank.
- **Procent populacji mający dostęp do co najmniej podstawowych usług sanitarnych** - stan na rok 2019.. Dane pochodzą z bazy danych World Bank.
- **Czy europejskie** - zmienna binarna, przyjmuje wartość 1 jeżeli państwo znajduje się w Europie, 0 w pozostałych przypadkach. W sumie 40 państw europejskich.
- **Czy afrykańskie** - zmienna binarna, przyjmuje wartość 1 jeżeli państwo znajduje się w Afryce, 0 w pozostałych przypadkach. W sumie 39 państw afrykańskich.
- **Stosunek mężczyzn do kobiet** - ilość mężczyzn przypadająca na jedną kobietę. Dane pochodzą z artykułu na Wikipedii. Ze względu na brak dostępności danych z roku 2019, wykorzystano dane z roku 2020.

We wszystkich przypadkach data dostępu to 06.06.2024.



Rysunek 1: Państwa obecne w zbiorze danych

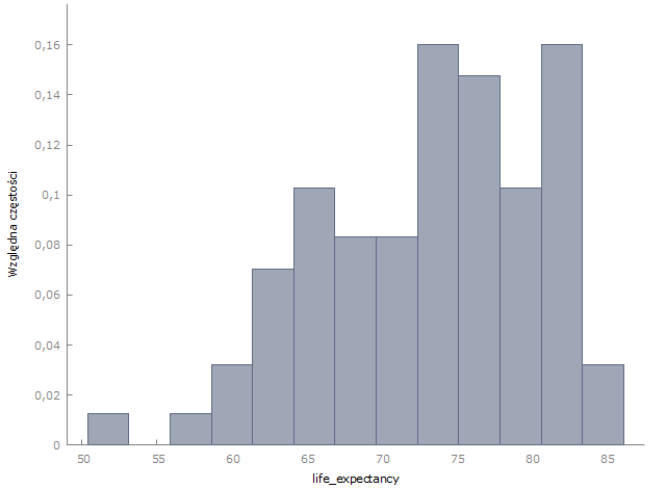
3 Statystyki opisowe oraz wykresy zależności

3.1 Oczekiwana długość życia

Średnia	73,230
Mediana	74,175
Minimalna	51,770
Maksymalna	84,850
Odchylenie standardowe	7,2549
Wsp. zmienności	0,099071
Skośność	-0,48638
Kurtoza	-0,41839
Percentyl 5%	60,920
Percentyl 95%	83,014
Zakres Q3-Q1	11,105
Brakujące obs.	0

Rysunek 2: Statystyki opisowe dla oczekiwanej długości życia

Ujemna różnica między wartością średniej oraz mediany jak i współczynnik skośności na poziomie około -0.5 sugerują, że rozkład jest lekko lewostronnie asymetryczny. Zilustrujmy dane na histogramie:



Rysunek 3: Histogram dla zmiennej objaśnianej

3.2 Populacja

Średnia	3,6807e+007
Mediana	8,7742e+006
Minimalna	10764,
Maksymalna	1,4338e+009
Odchylenie standardowe	1,2355e+008
Wsp. zmienności	3,3567
Skośność	9,5495
Kurtoza	103,52
Percentyl 5%	57897,
Percentyl 95%	1,4845e+008
Zakres Q3-Q1	2,7971e+007
Brakujące obs.	0

Rysunek 4: Statystyki opisowe dla populacji

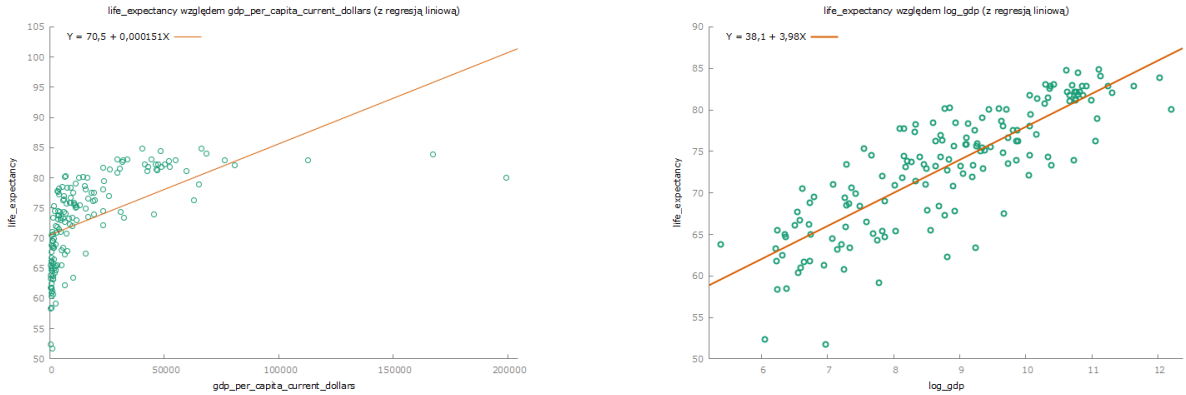
Rozkład tej zmiennej jest mocno asymetryczny prawostronnie - jest znacznie więcej państw o niewielkiej populacji. Poza tym w zbiorze danych znajduje się znacząca wartość odstająca - są to Chiny z populacją około 1,43 miliarda.

3.3 PKB per capita w aktualnych dolarach

Średnia	17885,
Mediana	6896,8
Minimalna	216,97
Maksymalna	1,9938e+005
Odchylenie standardowe	27606,
Wsp. zmienności	1,5435
Skośność	3,4629
Kurtoza	16,377
Percentyl 5%	571,97
Percentyl 95%	65263,
Zakres Q3-Q1	21465,
Brakujące obs.	0

Rysunek 5: Statystyki opisowe dla PKB per capita

Mediana jest w tym przypadku znacznie mniejsza od średniej, zatem rozkład jest silnie asymetryczny prawostronnie, co potwierdza duży współczynnik skośności. Zmienną o takim rozkładzie warto spróbować zlogarytmować. Przedstawmy na wykresie związek ze zmienną objaśnianą dla PKB per capita przed i po zlogarytmowaniu.



(a) Bez logarytmu (b) Z logarytmem

Rysunek 6: Porównanie zależności dla PKB per capita przed i po zlogarytmowaniu

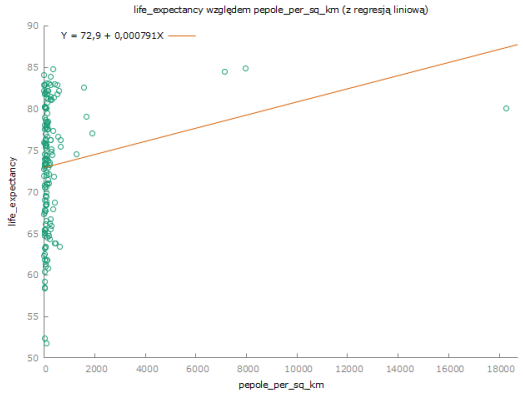
Współczynnik korelacji w obu przypadkach jest statystycznie istotny, jednak dla zmiennej zlogarytmowanej jest znacznie wyższy, wyższa jest również statstyka testowa (Bez logarytmu: $r = 0.57$, $t = 8.71$, z logarytmem: $r = 0.83$, $t = 18.76$). Powyższe wyniki sugerują użycie w modelu zmiennej zlogarytmowanej, o czym więcej w sekcji 4.

3.4 Gęstość zaludnienia

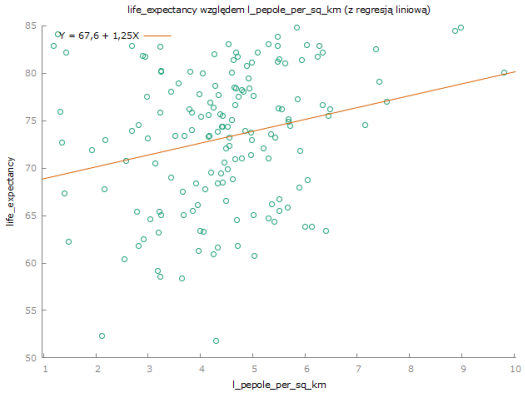
Średnia	383,14
Mediana	89,265
Minimalna	3,2937
Maksymalna	18270,
Odchylenie standardowe	1688,3
Wsp. zmienności	4,4065
Skośność	8,7065
Kurtoza	83,144
Percentyl 5%	6,4906
Percentyl 95%	744,58
Zakres Q3-Q1	175,08
Brakujące obs.	0

Rysunek 7: Statystyki opisowe dla gęstości zaludnienia

Ponownie, duża skośność sugeruje, że mamy do czynienia ze zmienną o prawostronnie asymetrycznym rozkładzie. Przedstawmy na wykresie związki ze zmienną objaśnianą przed i po zlogarytmowaniu:



(a) Bez logarytmu



(b) Z logarytmem

Rysunek 8: Porównanie zależności dla gęstości zaludnienia przed i po zlogarytmowaniu

Pomimo, że efekt nie jest tak uderzający jak w przypadku PKB per capita, logarytm poprawia wartość współczynnika korelacji oraz jego istotność (bez logarytmu: $r = 0.18, t = 2.32$, z logarytmem: $r = 0.25, t = 3.2$, w obu przypadkach $p < 0.05$), zatem jest to transformacja warta rozważenia przy wyborze ostatecznej postaci modelu.

3.5 Procent populacji zurbanizowanej

Średnia	60,070
Mediana	59,538
Minimalna	13,250
Maksymalna	100,00
Odchylenie standardowe	23,753
Wsp. zmienności	0,39542
Skośność	-0,15060
Kurtoza	-1,0322
Percentyl 5%	19,943
Percentyl 95%	97,469
Zakres Q3-Q1	38,890
Brakujące obs.	0

Rysunek 9: Statystyki opisowe dla procenta populacji zurbanizowanej

W tym przypadku rozkład okazuje się być z dobrym przybliżeniem symetryczny. Na uwagę zasługuje ujemna wartość kurtozy - rozkład jest platokurtyczny, czyli procent populacji zurbanizowanej jest szeroko rozłożony wokół średniej w krajach świata.

3.6 Emisja CO2 na mieszkańca

Średnia	4,1528
Mediana	3,3751
Minimalna	0,033715
Maksymalna	22,063
Odchylenie standardowe	4,1092
Wsp. zmienności	0,98950
Skośność	1,6720
Kurtoza	3,6073
Percentyl 5%	0,14062
Percentyl 95%	12,515
Zakres Q3-Q1	5,4549
Brakujące obs.	0

Rysunek 10: Statystyki opisowe dla emisji CO2 na mieszkańca

Rozkład ponownie okazuje się prawostronnie asymetryczny, jednak logarytmiczna transformacja czyni rozkład lewostronnie asymetrycznym. Ponadto współczynnik korelacji ze zmienną objaśnianą zarówno bez jak i z transformacją okazuje się nieistotny.

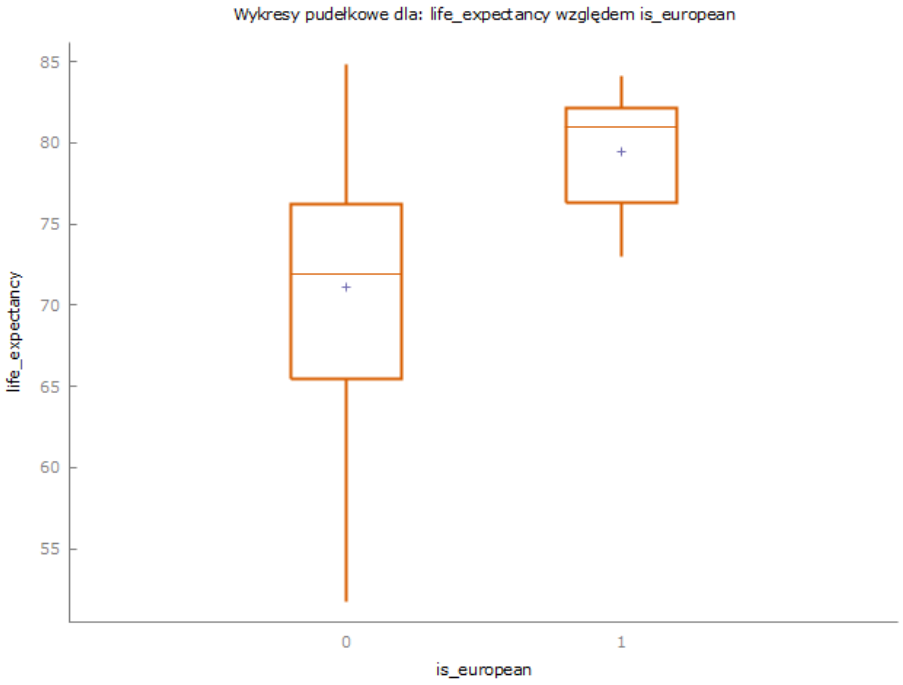
3.7 Procent populacji z dostępem do usług sanitarnych

Średnia	76,525
Mediana	90,339
Minimalna	8,6520
Maksymalna	100,00
Odchylenie standardowe	28,107
Wsp. zmienności	0,36729
Skośność	-1,0311
Kurtoza	-0,39642
Percentyl 5%	18,060
Percentyl 95%	100,00
Zakres Q3-Q1	44,338
Brakujące obs.	0

Rysunek 11: Statystyki opisowe dla procenta populacji z dostępem do usług sanitarnych

Rozkład w tym przypadku jest asymetryczny lewostronnie - w ponad połowie państw (dokładnie 80) w zbiorze danych procent populacji z dostępem do usług sanitarnych jest wyższy od 90%. W związku z powyższym występuje też niewielki współczynnik zmienności, który może mieć negatywny wpływ na interpretowalność parametrów modelu.

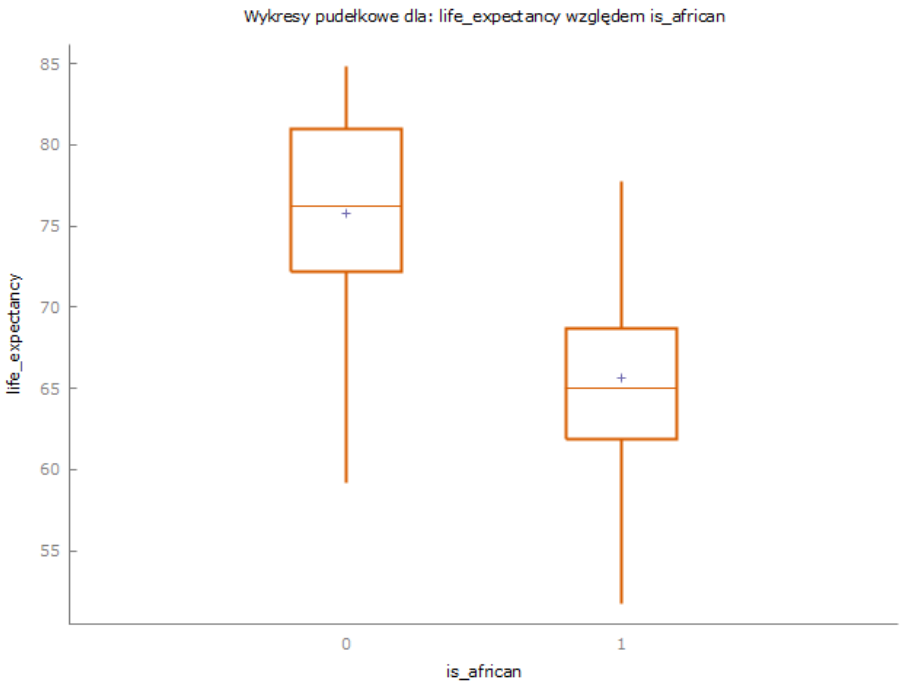
3.8 Czy państwo znajduje się w Europie



Rysunek 12: Wykres pudełkowy - oczekiwana długość życia w państwach Europejskich a reszcie świata

Powyższy wykres sugeruje, że państwa europejskie mają na ogół wyższą oczekiwaną długość życia.

3.9 Czy państwo znajduje się w Afryce



Rysunek 13: Wykres pudełkowy - oczekiwana długość życia w państwach Afrykańskich a reszcie świata

Powyższy wykres sugeruje, że państwa afrykańskie mają niższą oczekiwaną długość życia.

Adnotacja do wykresów pudełkowych dla zmiennych binarnych

Nie należy wyciągać pochopnych wniosków z wykresów - może się bowiem okazać, że parametr przy zmiennej binarnej w modelu nie będzie istotnie różny od zera. Oznaczałoby to, że widoczna na wykresie różnica jest w rzeczywistości wypadkową pozostałych uwzględnionych w modelu czynników. Nawet jeżeli parametry okażą się istotnie różne od zera, nie można przyjąć, że różnice są spowodowane jedynie przez lokalizację danego państwa *ceteris paribus*. Możliwe bowiem, że istnieją inne, pominięte w tym projekcie czynniki.

3.10 Stosunek mężczyzn do kobiet

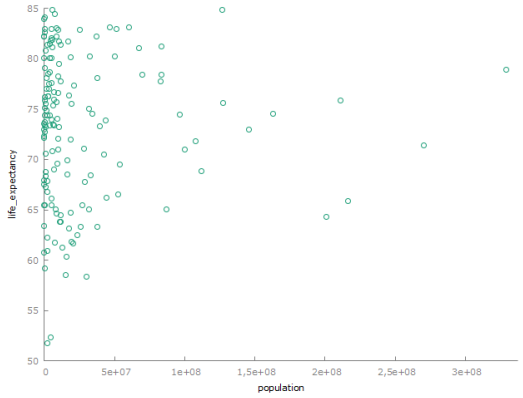
Średnia	1,0165
Mediana	0,98500
Minimalna	0,86000
Maksymalna	3,3900
Odchylenie standardowe	0,24118
Wsp. zmienności	0,23725
Skośność	7,9438
Kurtoza	68,988
Percentyl 5%	0,89700
Percentyl 95%	1,1130
Zakres Q3-Q1	0,057500
Brakujące obs.	0

Rysunek 14: Statystyki opisowe dla stosunku mężczyzn do kobiet

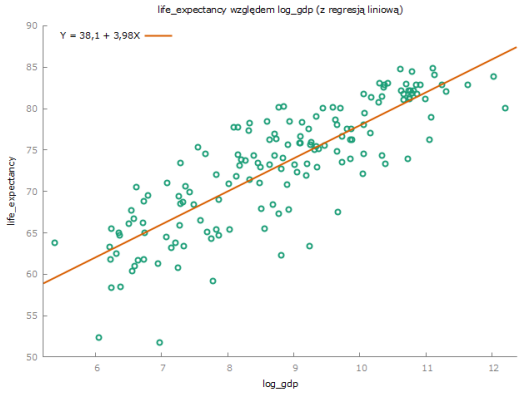
Niski współczynnik zmienności oznacza, że większość wartości zmiennej oscyluje wokół jedynki - różnice w liczbie kobiet i mężczyzn w większości państw są nieznaczne. Rozkład zaburzają obserwacje odstające, czyli Katar, gdzie na jedną kobietę przypada około 3.4 mężczyzny oraz Zjednoczone Emiraty Arabskie, gdzie wskaźnik ten wynosi około 2.56.

3.11 Wykresy zależności

Poniżej znajdują się wykresy zależności każdej ze zmiennych objaśniających ze zmienną objaśnianą. PKB per capita oraz gęstość zaludnienia zlogarytmowane. Ze zbioru danych zostały usunięte Katar oraz Zjednoczone Emiraty Arabskie (ze względu na odstający stosunek liczby mężczyzn do kobiet) oraz Chiny (ze względu na odstającą populację).

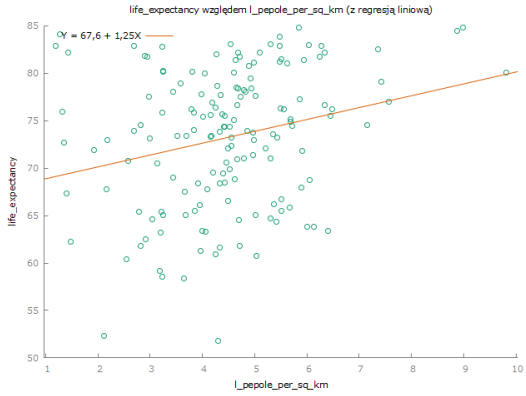


(a) Populacja

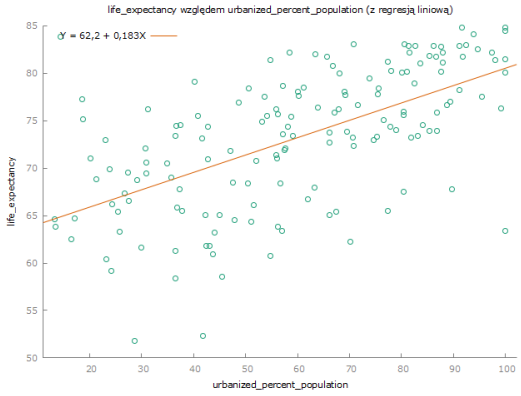


(b) PKB per capita

Rysunek 15: Wykresy zależności cz. I

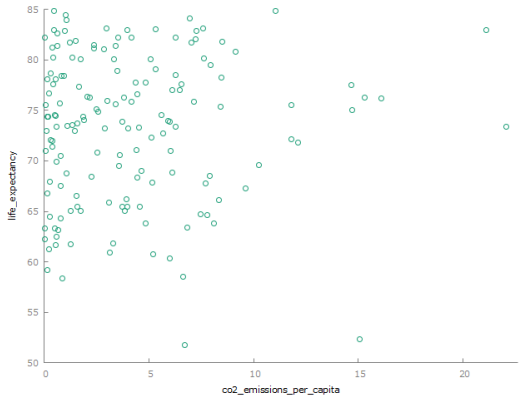


(a) Gęstość zaludnienia

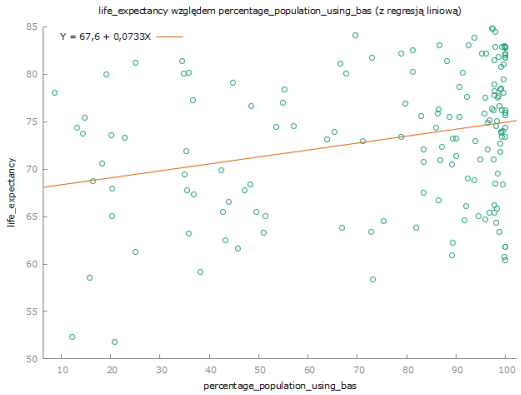


(b) Procent populacji zurbanizowanej

Rysunek 16: Wykresy zależności cz. II

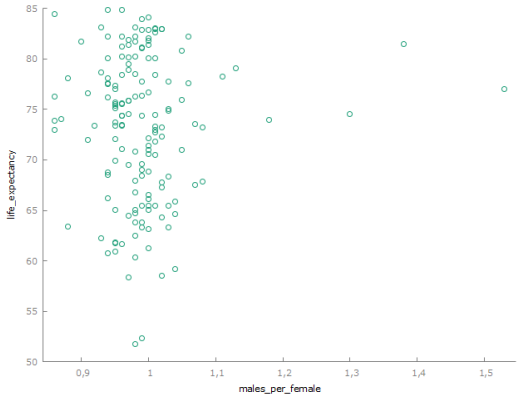


(a) Emisja CO2 na mieszkańca



(b) Procent populacji z dostępem do usług sanitarnych

Rysunek 17: Wykresy zależności cz. III

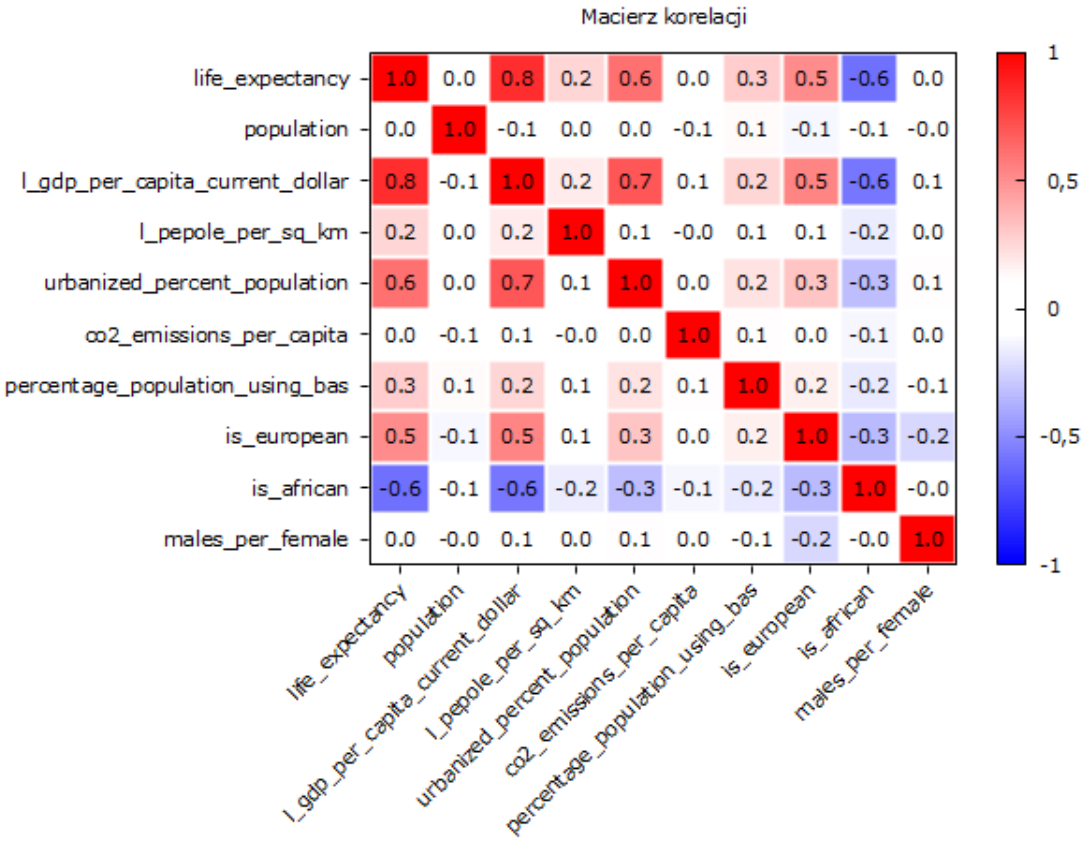


(a) Stosunek mężczyzn do kobiet

Rysunek 18: Wykresy zależności cz. IV

Wykresy dla logarytmu z PKB per capita oraz procenta populacji zurbanizowanej najbardziej przypominają zależność liniową, zatem spodziewamy się ich wysokiej istotności w modelu. Słabsza zależność jest widoczna w przypadku logarytmu z gęstości zaludnienia oraz procenta populacji z dostępem do usług sanitarnych - tutaj istnieje szansa na przynajmniej częściową istotność. Natomiast dla populacji, emisji CO2 oraz stosunku mężczyzn do kobiet zależność jest prawie niewidoczna - będą to zmienne które prawdopodobnie zostaną wykluczone z ostatecznej postaci modelu.

3.12 Macierz korelacji



Rysunek 19: Macierz korelacji dla zmiennych modelu

Macierz korelacji potwierdza wnioski płynące z wykresów - kandydatami na istotne regresory są PKB per capita ($r = 0.8$), procent populacji zurbanizowanej ($r = 0.6$), a nieco gorszymi gęstość zaludnienia ($r = 0.2$) oraz procent populacji z dostępem do usług sanitarnych ($r = 0.3$). Macierz korelacji ujawnia także dość silną zależność w przypadku zmiennych binarnych określających czy dane państwo leży w Europie lub w Afryce.

Uwagę zwracają wysokie korelacje w środku macierzy, pomiędzy zmiennymi objaśniającymi (np. dla PKB per capita i procent populacji zurbanizowanej $r = 0.7$). Możliwe zatem, że w modelu wystąpi współliniowość lub silny efekt katalizy. Aby zapobiec tym zjawiskom w kolejnej sekcji zredukujemy zbiór zmiennych objaśniających.

4 Wstępna postać modelu i redukcja zbioru zmiennych objaśniających

4.1 Wstępna postać modelu

Wyestymujemy model ściśle liniowy, z uwzględnieniem wszystkich zmiennych w zbiorze danych.

Model 7: Estymacja KMNK, wykorzystane obserwacje 1-153
Zmienna zależna (Y): life_expectancy

	współczynnik	błąd standardowy	t-Studenta	wartość p	
const	64,1177	5,22447	12,27	4,23e-024	***
population	1,93543e-09	7,19898e-09	0,2688	0,7884	
gdp_per_capita_c~	8,19779e-05	1,94922e-05	4,206	4,57e-05	***
pepole_per_sq_km	-0,000449549	0,000266532	-1,687	0,0938	*
urbanized_percen~	0,0981911	0,0175751	5,587	1,13e-07	***
co2_emissions_pe~	-0,0512877	0,0881978	-0,5815	0,5618	
percentage_popul~	0,0209598	0,0133424	1,571	0,1184	
is_european	2,35307	1,06224	2,215	0,0283	**
is_african	-5,88599	0,926250	-6,355	2,62e-09	***
males_per_female	1,40706	5,03912	0,2792	0,7805	

Średn.aryt.zm.zależnej	73,17693	Odch.stand.zm.zależnej	7,313039
Suma kwadratów reszt	2777,780	Błąd standardowy reszt	4,407384
Wsp. determ. R-kwadrat	0,658289	Skorygowany R-kwadrat	0,636783
F(9, 143)	30,60920	Wartość p dla testu F	3,19e-29
Logarytm wiarygodności	-438,8688	Kryt. inform. Akaike'a	897,7375
Kryt. bayes. Schwarza	928,0419	Kryt. Hannana-Quinna	910,0477

Rysunek 20: Wyniki estymacji modelu ze wszystkimi zmiennymi, bez logarytmów

Część zmiennych okazuje się nieistotna, wobec tego zredukujemy liczbę zmiennych objaśniających z użyciem metody Hellwiga oraz krokowej wstecznej. Rozważymy takze postać modelu z logarytmami PKB per capita i gęstości zaludnienia.

```
Model 8: Estymacja KMNK, wykorzystane obserwacje 1-153
Zmienna zależna (Y): life_expectancy

                    współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------------
  const                42,8070         4,86628          8,797     4,12e-015 ***
  population            5,66011e-09    6,19543e-09      0,9136    0,3625
  l_gdp_per_capita~     3,00415        0,368741         8,147     1,69e-013 ***
  l_pepole_per_sq_~     0,455762       0,215686         2,113     0,0363    **
  urbanized_percen~     0,0224861      0,0184457        1,219     0,2248
  co2_emissions_pe~    -0,0352095      0,0758915       -0,4639    0,6434
  percentage_popul~     0,0150975      0,0115044        1,312     0,1915
  is_european           1,42884        0,888721         1,608     0,1101
  is_african           -2,69581        0,895116        -3,012     0,0031    ***
  males_per_female     -0,231203       4,29994         -0,05377   0,9572

Średn.aryt.zm.zależnej  73,17693   Odch.stand.zm.zależnej  7,313039
Suma kwadratów reszt    2059,016   Błąd standardowy reszt  3,794563
Wsp. determ. R-kwadrat  0,746709   Skorygowany R-kwadrat   0,730767
F(9, 143)               46,84082   Wartość p dla testu F   2,47e-38
Logarytm wiarygodności -415,9628   Kryt. inform. Akaike'a  851,9256
Kryt. bayes. Schwarza   882,2300   Kryt. Hannana-Quinna    864,2358

Wyłączając stałą, największa wartość p jest dla zmiennej 11 (males_per_female)
```

Rysunek 21: Wyniki estymacji modelu ze wszystkimi zmiennymi, z logarytmami

4.2 Metoda Hellwiga

4.2.1 Dla wersji bez logarytmów

Dla modelu bez logarytmów podzbiór zmiennych objaśniających o najwyższej integralnej pojemności informacyjnej ($H = 0.632$) to:

- PKB per capita
- Procent zurbanizowanej populacji
- Czy państwo europejskie
- Czy państwo afrykańskie

4.2.2 Dla wersji z logarytmami

Dla modelu z logarytmami podzbiór zmiennych objaśniających o najwyższej integralnej pojemności informacyjnej ($H = 0.706$) zawiera jedynie zmienną PKB per capita (z logarytmem).

4.3 Metoda Krokowa Wsteczna

4.3.1 Dla wersji bez logarytmów

Dla modelu bez logarytmów podzbiór wybrany metodą krokową wsteczną to:

- PKB per capita
- Procent zurbanizowanej populacji
- Czy państwo europejskie
- Czy państwo afrykańskie

Zatem jest identyczny jak w sekcji 4.2.1.

4.3.2 Dla wersji z logarytmami

Dla modelu z logarytmami podzbiór wybrany metodą krokową wsteczną to:

- PKB per capita (logarytm)
- Gęstość zaludnienia (logarytm)
- Czy państwo afrykańskie

4.4 Porównanie modeli

Rozważymy trzy zdefiniowane powyżej zbiory zmiennych objaśniających i porównamy wyestymowane na ich podstawie modele.

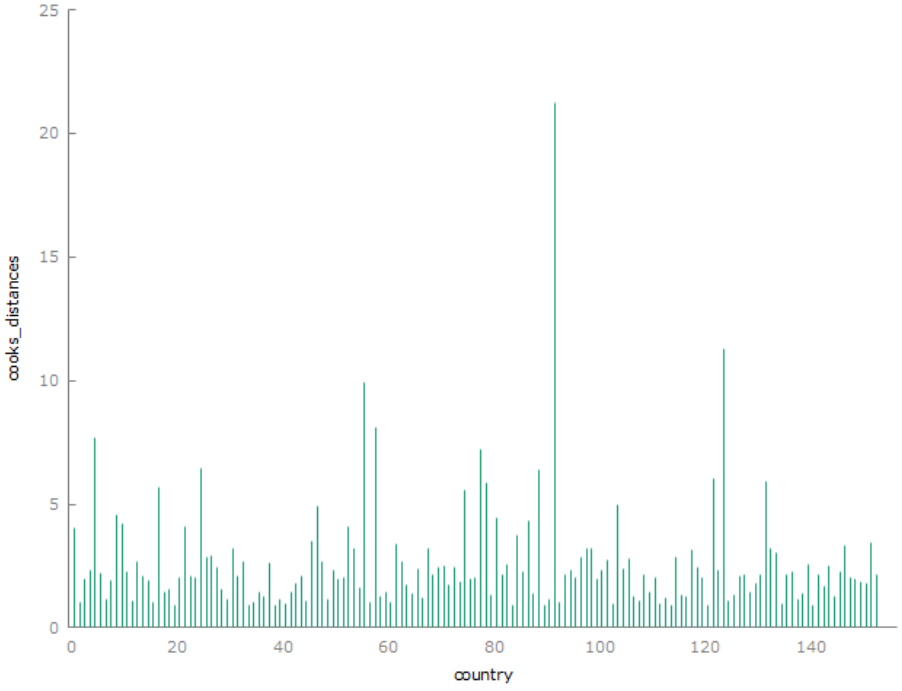
	Model 1	**Model 2**	**Model 3**
Metoda doboru	Hellwiga/Krokowa wsteczna	Hellwiga	Krokowa wsteczna
Współczynnik determinacji R^2	0.643	0.706	0.734
Liczba zmiennych	4	1	3
Skorygowany R^2	0.633	0.704	0.728
Kryterium informacyjne Akaike'a	894.42	858.71	847.65
P-value w teście Breuscha-Pagana	0.03	0.03	0.56

Tabela 1: Zestawienie wyestymowanych modeli

Model 3 (gdzie zmienne objaśniające to PKB per capita, gęstość zaludnienia oraz czy państwo jest afrykańskie) charakteryzuje się najwyższą wartością skorygowanego współczynnika determinacji oraz najniższą wartością kryterium informacyjnego Akaike'a. Ponadto, jako jedyny jest homoskedastyczny według testu Breuscha-Pagana.

4.5 Obserwacje odstające

W celu zidentyfikowania obserwacji odstających wyznaczymy odległości Cook'a dla każdej z obserwacji (więcej na temat metody w sekcji 6.6).



Rysunek 22: Odległości Cook'a - przed usunięciem outlierów

Nie ma konsensusu co do granicznej wartości odległości Cook'a. W niektórych opracowaniach [1] obserwacje o odległościach większych od trzykrotności średniej odległości są uważane za możliwe obserwacje odstające. W

naszym zbiorze danych średnia odległość Cook'a wynosi 2.62, jej trzykrotność - 7.86. Na wykresie widoczne są trzy obserwacje przekraczające ten próg:

- obserwacja 56 - Hong Kong ($D_i = 9.94, y = 84.45$)
- obserwacja 92 - Monako ($D_i = 21.24, y = 80.06$)
- obserwacja 124 - Singapur ($D_i = 11.28, y = 84.85$)

Obserwacje te to wysoko rozwinięte państwa-miasta (Hong Kong administracyjne jest częścią Chin, w praktyce jednak gospodarka jest w dużej części odrębna od gospodarki Chińskiej), o ponadprzeciętnych wartościach zmiennej objaśnianej jak i zmiennych objaśniających. W dalszej analizie pozbędziemy się ich ze zbioru danych.

5 Ostateczna postać modelu

Ostateczna postać modelu to:

$$y = 39.42 + 3.62 \cdot ln(x_1) + 0.6 \cdot ln(x_2) - 2.51 \cdot x_3$$

gdzie x_1 - PKB per capita, x_2 - gęstość zaludnienia i x_3 - zmienna binarna, określa czy państwo leży w Afryce.

Model 2: Estymacja KMNK, wykorzystane obserwacje 1-150
Zmienna zależna (Y): life_expectancy

	współczynnik	błąd standardowy	t-Studenta	wartość p	
const	39,4233	2,66127	14,81	6,91e-031	***
l_gdp_per_capita~	3,61875	0,258360	14,01	8,60e-029	***
l_pepole_per_sq_~	0,600878	0,239437	2,510	0,0132	**
is_african	-2,51483	0,877326	-2,866	0,0048	***

Średn.aryt.zm.zależnej	72,97807	Odch.stand.zm.zależnej	7,241005
Suma kwadratów reszt	2086,388	Błąd standardowy reszt	3,780255
Wsp. determ. R-kwadrat	0,732939	Skorygowany R-kwadrat	0,727451
F(3, 146)	133,5636	Wartość p dla testu F	1,15e-41
Logarytm wiarygodności	-410,2824	Kryt. inform. Akaike'a	828,5647
Kryt. bayes. Schwarza	840,6073	Kryt. Hannana-Quinna	833,4572

Rysunek 23: Estymacja parametrów ostatecznej postaci modelu

6 Analiza własności modelu

6.1 Współczynnik determinacji

Współczynnik determinacji w modelu wynosi $R^2 = 0.73$. Przeprowadzimy test istotności współczynnika determinacji (inaczej test istotności wszystkich zmiennych objaśniających) o następujących hipotezach:

$$H_0 : a_1 = a_2 = a_3 = 0$$

$$H_1 : a_1 \neq 0 \vee a_2 \neq 0 \vee a_3 \neq 0$$

Statystyka testowa postaci:

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - (k + 1)}{k}$$

ma rozkład F o k stopniach swobody licznika i $n - (k + 1)$ stopniach swobody mianownika.
Obliczona wartość statystyki:

$$F = 133.5636$$

Wartość krytyczna:

$$F_{0.05;3;146} \approx 2.66 \ (p = 1.15 \cdot 10^{-41})$$

Zatem współczynnik determinacji (zbiór zmiennych objaśniających) można uznać za istotny.

6.2 Efekt katalizy

W celu sprawdzenia, czy współczynnik determinacji nie jest fałszywie zawyżony, zbadamy natężenie efektu katalizy oraz sprawdzimy, czy wśród zmiennych nie ma katalizatorów.

Wektor R_0 oraz macierz R mają postać:

$$R_0 = \begin{bmatrix} 0.84 \\ 0.18 \\ -0.6 \end{bmatrix} \qquad R = \begin{bmatrix} 1 & 0.07 & -0.59 \\ 0.07 & 1 & -0.15 \\ -0.59 & -0.15 & 1 \end{bmatrix}$$

Po przekształceniu w regularną parę korelacyjną:

$$R_0 = \begin{bmatrix} 0.18 \\ 0.6 \\ 0.84 \end{bmatrix} \qquad R = \begin{bmatrix} 1 & 0.15 & 0.07 \\ 0.15 & 1 & 0.59 \\ 0.07 & 0.59 & 1 \end{bmatrix}$$

Zmienna X_i z pary $(X_i, X_j), i < j$ jest katalizatorem, jeżeli $r_{ij} < 0$ lub $r_{ij} > \frac{r_i}{r_j}$.

(i, j)	r_{ij}	$r_{ij} - \frac{r_i}{r_j}$
(1, 2)	0.15	-0.15
(1, 3)	0.07	-0.034
(2, 3)	0.59	-0.12

Tabela 2: Wartość wyrażeń determinujących zmienne-katalizatory

Zgodnie z powyższą tabelą, w modelu nie ma katalizatorów. Dodatkowo wyznaczymy względne natężenie efektu katalizy:

$$H \approx 0.66$$

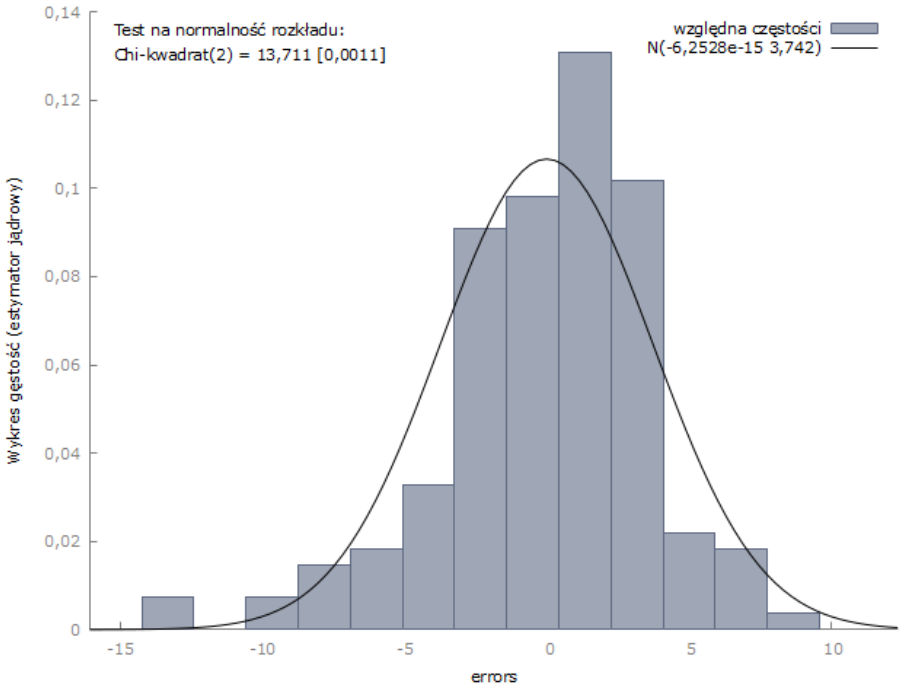
$$\eta = R^2 - H \approx 0.0716$$

$$W_\eta = \frac{\eta}{R^2} \cdot 100\% \approx 9.77\%$$

Wobec powyższych wyników, efekt katalizy nie zawyża znacząco współczynnika determinacji.

6.3 Normalność rozkładu składnika losowego

Przedstawmy empiryczny rozkład reszt modelu wraz z dopasowaną krzywą Gaussa:



Rysunek 26: Rozkład częstości reszt modelu

```
Test na normalność rozkładu errors:

 Test Doornika-Hansena = 13,7112, z wartością p 0,00105356

 Test Shapiro-Wilka = 0,96177, z wartością p 0,000357666

 Test Lillieforsa = 0,0744926, z wartością p ~= 0,04

 Test Jarque'a-Bera = 27,5235, z wartością p 1,05522e-006
```

Rysunek 27: Testy normalności rozkłady składnika losowego

Wszystkie cztery wykonane testy wskazały, że składnik losowy nie ma rozkładu normalnego. Pomimo tego, ze względu na dużą liczbę stopni swobody w modelu, estymatory parametrów mają asymptotyczne rozkłady normalne (na mocy CLT), zatem stosowanie testów t oraz F jest zasadne [2]. Problematyczne natomiast będzie stworzenie przedziału ufności dla prognozy, czym zajmiemy się w sekcji 6.14.

6.4 Istotność zmiennych

6.4.1 Logarytm z PKB per capita

Wykonamy test t-Studenta o następujących hipotezach:

$$H_0 : a_1 = 0$$

$$H_1 : a_1 \neq 0$$

Statystyka testowa:

$$t = \frac{a_1}{S(a_1)} \approx 14.01$$

Wartość krytyczna:

$$t_{0.05;146} \approx 1.98$$

$$|t_{obl}| > t_\alpha$$

Czyli parametr przy zmiennej jest istotnie różny od zera, zatem zmienna jest istotna.

6.4.2 Logarytm z gęstości zaludnienia

Wykonamy test t-Studenta o następujących hipotezach:

$$H_0 : a_2 = 0$$

$$H_1 : a_2 \neq 0$$

Statystyka testowa:

$$t = \frac{a_2}{S(a_2)} \approx 2.51$$

Wartość krytyczna:

$$t_{0.05;146} \approx 1.98$$

$$|t_{obl}| > t_\alpha$$

Czyli parametr przy zmiennej jest istotnie różny od zera, zatem zmienna jest istotna. W tym przypadku jeżeli poziom istotności wynosiłby 0.01, zmienna nie byłaby istotna.

6.4.3 Czy państwo afrykańskie

Wykonamy test t-Studenta o następujących hipotezach:

$$H_0 : a_3 = 0$$

$$H_1 : a_3 \neq 0$$

Statystyka testowa:

$$t = \frac{a_3}{S(a_3)} \approx -2.87$$

Wartość krytyczna:

$$t_{0.05;146} \approx 1.98$$

$$|t_{obl}| > t_\alpha$$

Czyli parametr przy zmiennej jest istotnie różny od zera, zatem zmienna jest istotna.

6.5 Testy dodanych i pominiętych zmiennych

6.5.1 Pominięcie gęstości zaludnienia

Wykonamy test restrykcji o następujacych hipotezach:

$$H_0 : \alpha_3 = 0$$

$$H_1 : \alpha_3 \neq 0$$

gdzie α_3 oznacza parametr przy logarytmie z gęstości zaludnienia. Statystyka testowa ma postać:

$$F = \frac{RRSS - URSS}{URSS} \cdot \frac{150 - 4}{1}$$

gdzie RRSS - suma kwadratów błędów w modelu bez gęstości zaludnienia, URSS - suma kwadratów błędów w modelu z gęstością zaludnienia. Statystyka ma rozkład F(1, 146).

```
Test porównawczy z Modelem 3

  Hipoteza zerowa: parametr regresji jest równy zero dla l_pepole_per_sq_km
  Statystyka testu: F(1, 146) = 6,29781, wartość p 0,0131802
  Pominięcie zmiennych poprawiło 0 z 3 kryteriów informacyjnych (AIC, BIC, HQC).

Model 4: Estymacja KMNK, wykorzystane obserwacje 1-150
Zmienna zależna (Y): life_expectancy

                     współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------------
  const                42,2333         2,45735         17,19     5,44e-037 ***
  l_gdp_per_capita~     3,60607        0,262924        13,72     4,34e-028 ***
  is_african           -2,80563        0,885171        -3,170    0,0019    ***

Średn.aryt.zm.zależnej  72,97807   Odch.stand.zm.zależnej  7,241005
Suma kwadratów reszt    2176,386   Błąd standardowy reszt  3,847772
Wsp. determ. R-kwadrat  0,721419   Skorygowany R-kwadrat   0,717629
F(2, 147)               190,3368   Wartość p dla testu F   1,60e-41
Logarytm wiarygodności -413,4497   Kryt. inform. Akaike'a  832,8994
Kryt. bayes. Schwarza   841,9313   Kryt. Hannana-Quinna    836,5688
```

Rysunek 28: Wyniki testu restrykcji dla gęstości zaludnienia

Przyjmując poziom istotności 0.05 odrzucimy hipotezę zerową, czyli model nie jest istotnie lepszy bez omawianej zmiennej. Należy zaznaczyć, że przy bardziej kategorycznym poziomie istotności konkluzja mogłaby być inna. Podobne wnioski wynikają z testu t-Studenta w sekcji 6.4.3.

6.5.2 Dodanie procenta populacji zurbanizowanej

Rozszerzymy nasz model o procent populacji zurbanizowanej i parametr przy tej zmiennej oznaczymy jako α_4. Wówczas hipotezy w teście restrykcji mają postać:

$$H_0 : \alpha_4 = 0$$

$$H_1 : \alpha_4 \neq 0$$

Statystyka testowa:

$$F = \frac{RRSS - URSS}{URSS} \cdot \frac{150 - 5}{1}$$

gdzie RRSS - suma kwadratów błędów w modelu bez procenta populacji zurbanizowanej, URSS - suma kwadratów błędów w modelu z procentem populacji zurbanizowanej. Statystyka ma rozkład F(1, 145).

```
  Hipoteza zerowa: parametr regresji jest równy zero dla urbanized_percent_population
  Statystyka testu: F(1, 145) = 1,5294, wartość p 0,218202
  Dodanie zmiennych poprawiło 0 z 3 kryteriów informacyjnych (AIC, BIC, HQC).

Model 6: Estymacja KMNK, wykorzystane obserwacje 1-150
Zmienna zależna (Y): life_expectancy

                     współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------------
  const                40,2903         2,74743         14,66     1,96e-030 ***
  l_gdp_per_capita~     3,35474        0,334788        10,02     2,77e-018 ***
  l_pepole_per_sq_~     0,631616       0,240293         2,629    0,0095    ***
  is_african           -2,63811        0,881395        -2,993    0,0032    ***
  urbanized_percen~     0,0226936      0,0183503        1,237    0,2182

Średn.aryt.zm.zależnej  72,97807   Odch.stand.zm.zależnej  7,241005
Suma kwadratów reszt    2064,611   Błąd standardowy reszt  3,773420
Wsp. determ. R-kwadrat  0,735726   Skorygowany R-kwadrat   0,728436
F(4, 145)               100,9183   Wartość p dla testu F   6,82e-41
Logarytm wiarygodności -409,4954   Kryt. inform. Akaike'a  828,9909
Kryt. bayes. Schwarza   844,0440   Kryt. Hannana-Quinna    835,1065

Wyłączając stałą, największa wartość p jest dla zmiennej 6 (urbanized_percent_population)
```

Rysunek 29: Wyniki testu restrykcji dla procenta populacji zurbanizowanej

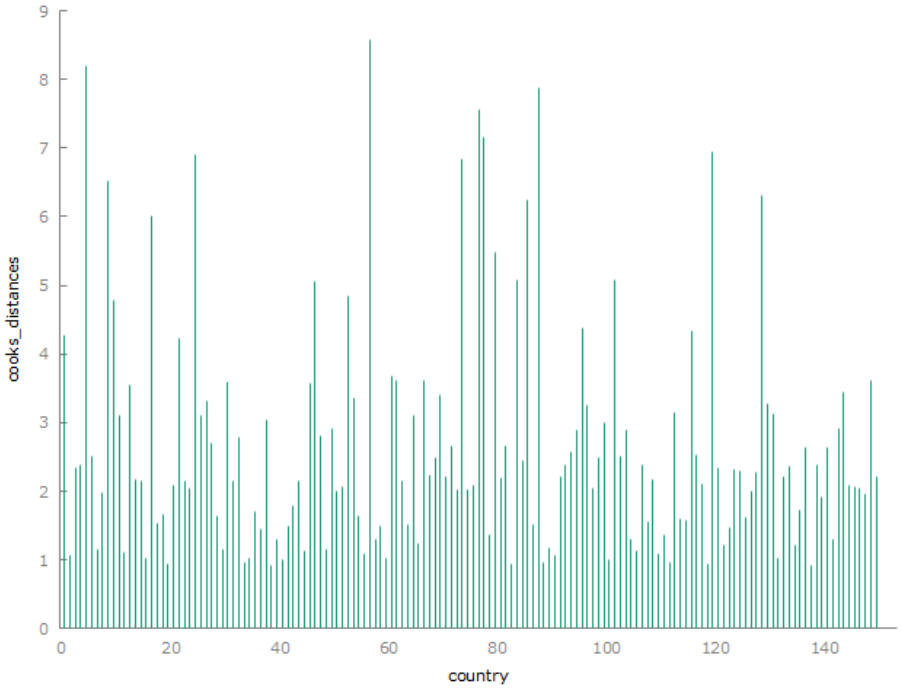
Nie ma podstaw do odrzucenia hipotezy zerowej, zatem nie można twierdzić, że dodanie omawianej zmiennej polepsza własności modelu.

6.6 Obserwacje odstające

Obserwacje odstające zindentyfikujemy na podstawie wyznaczenia odległości Cook'a dla każdej z obserwacji. Odległość Cook'a i-tej obserwacji dana jest wzorem [3]:

$$D_i = \frac{\sum_{j=1}^{n} \hat{y}_j - \hat{y}_{j(i)}}{(k+1)S^2}$$

gdzie $\hat{y}_j$ - przewidywanie wyjściowego modelu dla j-tej obserwacji, $\hat{y}_{j(i)}$ - przewidywanie modelu dla j-tej obserwacji estymowanego z pominięciem i-tej obserwacji, S^2 - wariancja reszt modelu wyjściowego.



Rysunek 30: Odległości Cook'a po usunięciu outlierów

Średnia odległość Cook'a wynosi w tym przypadku 2.65, jej trzykrotność - 7.95. W zbiorze danych nie ma obserwacji znacząco przekraczających ten pułap.

6.7 Test liczby serii

6.7.1 Dla logarytmu z PKB per capita

W celu sprawdzenia założenia o liniowej zależności między zmienną objaśnianą a objaśniającą wykonamy test serii Walda-Wolfowitza o następujących hipotezach:

$$H_0 : \textit{Postać modelu prawidłowa}$$

$$H_1 : \textit{Postać modelu nieprawidłowa}$$

Uporządkujemy próbę według rosnących wartości logarytmu z PKB per capita. Wówczas liczba serii wśród składnika losowego wynosi 83, liczba elementów dodatnich wynosi 79, a ujemnych 71. Oznaczmy liczbę serii jako L. Wtedy:

$$L \sim N\left(2 \cdot \frac{n_1 n_2}{n} + 1, \sqrt{\frac{2n_1 n_2(2n_1 n_2 - n)}{n^2(n-1)}}\right)$$

gdzie n_1 - liczba elementów dodatnich, n_2 - liczba elementów ujemnych, n - liczba obserwacji. W naszym przypadku:

$$L \sim N(75.79, 6.09)$$

Wyznaczmy 95% przedział ufności:

$$P(-1.96 < Z < 1.96) = 0.95$$

$$P(-1.96 < \frac{L - 76.94}{6.12} < 1.96) = 0.95$$

$$P(63.85 < L < 87.73) = 0.95$$

Ponieważ wyznaczona liczba serii wpada do tego przedziału, nie ma podstaw do odrzucenia hipotezy zerowej. Postać modelu można uznać za prawidłową.

6.7.2 Dla logarytmu z gęstości zaludnienia

Przedział ufności pozostaje taki sam jak w poprzedniej sekcji. Po uporządkowaniu próby według gęstości zaludnienia rosnąco liczba serii wynosi 74. Wartość ta należy do przedziału ufności, zatem nie ma podstaw do odrzucenia hipotezy zerowej - postać modelu można uznać za prawidłową.

6.8 Test RESET

W celu ustalenia, czy dobrana postać analityczna modelu jest poprawna, wykonamy test RESET o następujących hipotezach:

$$H_0 : \textit{postać modelu prawidłowa}$$

$$H_1 : \textit{postać modelu nieprawidłowa}$$

Statystyka testowa postaci:

$$F = \frac{R^2_{II} - R^2_1}{1 - R^2_{II}} \cdot \frac{n - (k + 3)}{2}$$

ma rozkład F o 2 stopniach swobody licznika i $n - (k + 3) = 144$ stopniach swobody mianownika. R^2_{II} oznacza współczynnik determinacji w modelu do którego dodano obliczone na podstawie bazowego modelu wartości zmiennej objaśnianej w kwadracie oraz w sześcianie jako zmienne objaśniające. R^2_I to współczynnik determinacji bazowego modelu.

```
Pomocnicze równanie regresji dla testu specyfikacji RESET
Estymacja KMNK, wykorzystane obserwacje 1-150
Zmienna zależna (Y): life_expectancy

                    współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------------
  const              -7,64114        293,952          -0,02599    0,9793
  l_gdp_per_capita~  -5,36792         68,6720         -0,07817    0,9378
  l_pepole_per_sq_~  -0,876947        11,3865         -0,07702    0,9387
  is_african          3,98736         47,7293          0,08354    0,9335
  yhat^2              0,0391419        0,262758        0,1490     0,8818
  yhat^3             -0,000200751      0,00120794     -0,1662     0,8682

Statystyka testu: F = 0,148688,
z wartością p = P(F(2,144) > 0,148688) = 0,862
```

Rysunek 31: Wyniki testu RESET

Test nie daje podstaw do odrzucenia hipotezy zerowej, zatem nie ma potrzeby zmiany postaci analitycznej modelu.

6.9 Heteroskedastyczność

Wykonamy test Breuscha-Pagana na heteroskedastyczność. Test ten polega na zbudowaniu pomocniczego równania regresji, w którym za pomocą zmiennych objaśniających wyjaśniamy kwadraty błędów z naszego modelu. Niech α oznacza wektor parametrów pomocniczego równania regresji. Hipotezy testu mają postać:

$$H_0 : \alpha_2 = \alpha_3 = \alpha_4 = 0$$

$$H_1 : \neg H_0$$

Mnożnik Lagrange'a będący statystyką testową jest równy połowie wyjaśnionej sumy kwadratów i ma asymptotyczny rozkład χ^2, w tym przypadku o 3 stopniach swobody [4].

```
Test Breuscha-Pagana na heteroskedastyczność
Estymacja KMNK, wykorzystane obserwacje 1-150
Zmienna zależna (Y): standaryzowane uhat^2

                    współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------------
  const                2,94318         1,29900         2,266      0,0249   **
  l_gdp_per_capita~   -0,187720        0,126108       -1,489      0,1388
  l_pepole_per_sq_~   -0,0627729       0,116872       -0,5371     0,5920
  is_african          -0,117800        0,428232       -0,2751     0,7836

  Wyjaśniona suma kwadr. = 10,707

Statystyka testu: LM = 5,353514,
z wartością p = P(Chi-kwadrat(3) > 5,353514) = 0,147667
```

Rysunek 32: Wyniki testu Breuscha-Pagana

Test Breuscha-Pagana nie daje podstaw do odrzucenia hipotezy H_0 zatem model jest homoskedastyczny.

6.10 Test Chowa

W celu przeprowadzenia testu stabilności parametrów modelu podzielimy próbę na dwie podpróby - w pierwszej znajdą się państwa, w których procent populacji mający dostęp do podstawowych usług sanitarnych jest większy od 90%, a w drugiej pozostałe. Niech α oznacza wektor parametrów modelu dla pierwszej podpróby, a β dla drugiej. Zweryfikujemy hipotezy:

$$H_0 : \alpha = \beta$$

$$H_1 : \alpha \neq \beta$$

```
Pomocnicze równanie regresji dla testu Chowa
Estymacja KMNK, wykorzystane obserwacje 1-150
Zmienna zależna (Y): life_expectancy

                    współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------------
  const               41,1600          3,70836        11,10      5,34e-021 ***
  l_gdp_per_capita~    3,34998         0,376937        8,887     2,55e-015 ***
  l_pepole_per_sq_~    0,667336        0,315449        2,116     0,0361    **
  is_african          -3,49708         1,18032        -2,963     0,0036    ***
  dummy_sanitary      -1,83035         5,41174        -0,3382    0,7357
  du_l_gdp_per_cap~    0,387612        0,521442        0,7433    0,4585
  du_l_pepole_per_~   -0,213462        0,486685       -0,4386    0,6616
  du_is_african        2,08811         1,76680         1,182     0,2392

Średn.aryt.zm.zależnej  72,97807   Odch.stand.zm.zależnej  7,241005
Suma kwadratów reszt    2013,171   Błąd standardowy reszt  3,765270
Wsp. determ. R-kwadrat  0,742311   Skorygowany R-kwadrat   0,729608
F(7, 142)               58,43583   Wartość p dla testu F   1,01e-38
Logarytm wiarygodności -407,6031   Kryt. inform. Akaike'a  831,2062
Kryt. bayes. Schwarza   855,2913   Kryt. Hannana-Quinna    840,9912

Test Chowa na strukturalne różnice poziomów ze względu na zmienną: dummy_sanitary
  F(4, 142) = 1,2911 z wartością p 0,2764
```

Rysunek 33: Wyniki testu Chowa

Zgodnie z wydrukiem, nie ma podstaw do odrzucenia hipotezy zerowej, zatem w obu podpróbach zależności można modelować takim samym równaniem.

6.11 Współliniowość

W celu zbadania przybliżonej współliniowości obliczmy wskaźniki VIF dla każdej ze zmiennych.

$$VIF = \frac{1}{1 - R^2}$$

gdzie R^2 jest współczynnikiem determinacji w modelu, gdzie badana zmienna objaśniająca jest uzależniona od pozostałych.

```
Ocena współliniowości VIF(j) - czynnik rozdęcia wariancji
VIF (Variance Inflation Factors) - minimalna możliwa wartość = 1.0
Wartości > 10.0 mogą wskazywać na problem współliniowości - rozdęcia wariancji

l_gdp_per_capita_current_dollar    1,528
             l_pepole_per_sq_km    1,023
                     is_african    1,554

VIF(j) = 1/(1 - R(j)^2), gdzie R(j) jest współczynnikiem korelacji wielorakiej
pomiędzy zmienną 'j' a pozostałymi zmiennymi niezależnymi modelu.
```

Rysunek 34: Wskaźniki rozdęcia wariancji (VIF) dla zmiennych modelu

Najwyższy VIF występuje dla PKB per capita. Przekłada się on na współczynnik determinacji $R^2 \approx 0.35$. Wartość ta nie wskazuje na problem współliniowości.

6.12 Koincydencja

Model jest koincydentny, jeżeli dla każdego zmiennej objaśniającej znak współczynnika korelacji liniowej jest zgodny ze znakiem parametru przy tej zmiennej w modelu. Zestawimy w tabeli wartości funkcji signum dla korelacji oraz parametru przy każdej zmiennej:

	$\text{sgn}(r_i)$	$\text{sgn}(a_i)$
PKB per capita	+	+
Gęstość zaludnienia	+	+
Czy afrykańskie	-	-

Tabela 3: Wartości funkcji signum dla współczynników korelacji i parametrów

Zatem model jest koincydentny.

6.13 Interpretacja parametrów modelu

Przypomnijmy postać modelu:

$$y = 39.42 + 3.62ln(x_1) + 0.6ln(x_2) - 2.51x_3$$

6.13.1 PKB per capita

Wzrost PKB per capita o 1% powoduje wzrost oczekiwanej długości życia o około 0.036 roku *ceteris paribus*.

$$y_1 = 39.42 + 3.62ln(1.01 \cdot x_1) + 0.6ln(x_2) - 2.51x_3 = 39.42 + 3.62ln(1.01) + 3.62ln(x_1) + 0.6ln(x_2) - 2.51x_3$$

$$y_1 = y_0 + 3.62ln(1.01) \approx y_0 + 0.036$$

Analogicznie, wzrost PKB per capita o 10% spowoduje wydłużenie oczekiwanej długości życia o około 0.35 roku ($3.62 \cdot ln(1.1)$) [5].

Ponieważ zależność jest logarytmiczna, oczekiwana długość życia dla niskich PKB per capita rośnie szybko, a dla dużych wolniej. Zatem można wysunąć wniosek, że w państwach słabiej rozwiniętych oczekiwana długość życia rośnie szybciej niż w krajach o silnej, ustabilizowanej gospodarce - przy małym PKB per capita wzrost o 1% przekłada się na niższą nominalnie wartość niż przy większym.

6.13.2 Gęstość zaludnienia

Wzrost gęstości zaludnienia o 1% powoduje wydłużenie oczekiwanej długości życia o około 0.006 roku ($\approx$ 2.19 dnia) *ceteris paribus*. Wzrost gęstości zaludnienia o 10% powoduje wzrost oczekiwanej długości życia w przybliżeniu o 0.057 roku. Pozytywny wpływ gęstości zaludnienia jest prawdopodobnie związany z lepszą dostępnością opieki zdrowotnej, lepszą komunikacją, wymianą informacji itd.

6.13.3 Czy państwo afrykańskie

Położenie państwa w Afryce skraca oczekiwaną długość życia o około 2.5 roku względem państw spoza Afryki przy takim samym PKB per capita i tej samej gęstości zaludnienia. Najprawdopodobiej nie jest to jednak skutek jedynie geograficznego położenia, ale także nieuwzględnionych w tym projekcie czynników.

6.14 Predykcja

6.14.1 Prognoza ex post

Wymienimy najistotniejsze błędy średnie prognoz ex post:

- **ME** = $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i}) = 7.4 \cdot 10^{-18} \approx 0$ - błąd bliski zeru świadczy o tym, że prognoza nie jest obciążona
- **MAE** = $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}| \approx 2.86$ - o tyle prognoza przeciętnie rózni się od wartości rzeczywistej
- **RMSE** = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2} \approx 3.73$ - o tyle przeciętnie różni się prognoza od wartości rzeczywistej z większą wagą na wartościach skrajnych
- **MAPE** = $\frac{1}{n}\sum_{i=1}^{n}|\frac{y_i - \hat{y_i}}{y_i}| \cdot 100\% \approx 4.08\%$ - o tyle procent przeciętnie prognoza różni się od wartości rzeczywistej

6.14.2 Prognoza ex ante

W naszym zbiorze danych dla Polski PKB per capita wynosi 15 700, gęstość zaludnienia 124, a oczekiwana długość życia 78.1. Wyznaczymy prognozowaną oczekiwaną długość życia, jeżeli PKB per capita wzrośnie do 25 000 USD, a gęstość zaludnienia do 150 osób na kilometr kwadratowy. Wówczas:

$$x_\tau = \begin{bmatrix} 1 & ln(25000) & ln(150) & 0 \end{bmatrix}$$

$$y_\tau^p = 79.08$$

6.15 Bootstrap - przedział ufności i błąd standardowy

Ze względu na brak normalności składnika losowego, do wyznaczenia 95% przedziału ufności posłużymy się metodą bootstrapową.

6.15.1 Teoretyczny wstęp

Prognozowaną wartość y przy danym x_0 można zapisać jako [6]:

$$y(x_0) = \phi(x_0) + \epsilon(x_0)$$

gdzie ϕ jest postacią modelu, a ϵ składnikiem losowym. Dodajmy do równania wartość teoretyczną $\hat{y}(x_0)$

$$y(x_0) = \hat{y}(x_0) + \phi(x_0) - \hat{y}(x_0) + \epsilon(x_0)$$

$$y(x_0) = \hat{y}(x_0) + \eta(x_0) + \epsilon(x_0)$$

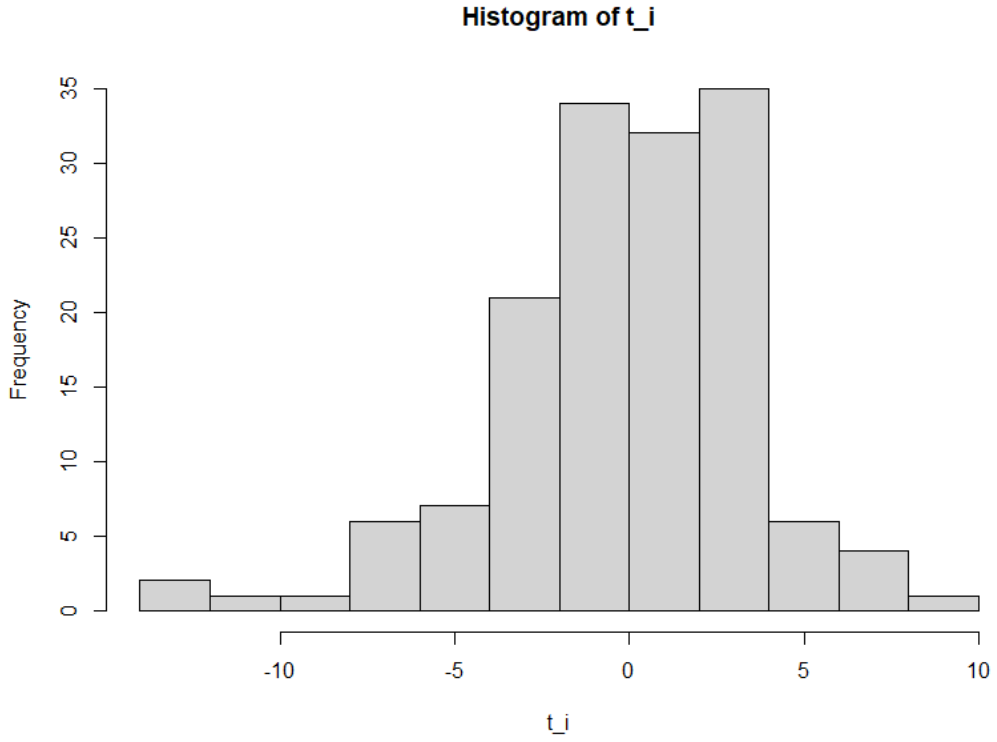
gdzie $\eta(x_0) = \phi(x_0) - \hat{y}(x_0)$ jest błędem wynikającym z postaci modelu. Ponieważ $\hat{y}(x_0)$ jest deterministyczne, w celu zbadania rozkładu $y(x_0)$ musimy zbadać rozkłady $\eta(x_0)$ - błędu wynikającego z postaci modelu oraz $\epsilon(x_0)$ - składnika losowego. Wyznaczymy empiryczne rozkłady tych zmiennych w użyciem metod bootstrapowych.

6.15.2 Symulacja

Przebieg symulacji:

1. Estymujemy model na podstawowej próbie, wyznaczamy reszty: $e_i = y_i - \hat{y}_i$ (czyli realizacje zmiennej losowej ϵ)
2. Tworzymy $n = 150$ prób złożonych ze 150 obserwacji, powstałych poprzez losowanie ze zwracaniem z ze zbioru danych (zatem mogą się pojawić duplikaty). Estymujemy model na podstawie każdej próby, oraz obliczamy prognozowaną wartość $y_{\hat{B}_i}$ na podstawie tego modelu.
3. Estymujemy oczekiwaną wartość bootstrapowej prognozy $\hat{\mu} = \frac{\sum_{i=1}^{n} y_{\hat{B}_i}}{n}$
4. Obliczamy $m_i = y_{\hat{B}_i} - \hat{\mu}$ (czyli realizacje zmiennej losowej η)
5. Obliczamy $t_i = m_i + e_i$, czyli realizacje sumy zmiennych losowych η i ϵ, całkowitego błędu prognozy
6. Wyznaczamy kwantyle rzędu 0.025 oraz 0.975 rozkładu zmiennej t_i, i dodajemy do nich teoretyczną wartość $\hat{y}(x_0)$ otrzymując 95% przedział ufności.

Wyznaczymy przedział ufności dla wektora $x_\tau = \begin{bmatrix} 1 & ln(25000) & ln(150) & 0 \end{bmatrix}$ z poprzedniej sekcji.



Rysunek 35: Histogram błędu prognozy

Wyznaczmy średni absolutny błąd tej prognozy:

$$S_\tau^p = \frac{\sum_{i=1}^{n} |t_i|}{n} \approx 2.84$$

$$\left| \frac{S_\tau^p}{y_\tau^p} \right| \approx 3.6\%$$

Oraz przedział ufności:

$$P(70.54 < y_\tau^p < 85.46) = 0.95$$

Czyli jeżeli PKB per capita wzrośnie do 25000, gęstość zaludnienia do 150 osób na kilometr kwadratowy, to oczekiwana długość życia w Polsce na 95% znajdzie się w przedziale od 70.54 do 85.46 lat. Jest to bardzo szeroki przedział ufności, więc wnioski jakie można z niego wyciągnąć nie są satysfakcjonujące.

7 Podsumowanie

7.1 Model od strony technicznej

Zbudowany model zawiera statystycznie istotne zmienne, jest homoskedastyczny, jego współczynnik determinacji jest wysoki i nie jest sztucznie zawyżony w wyniku efektu katalizy. Nie jest on jednak pozbawiony mankamentów. Należą do nich m.in. brak normalności składnika losowego oraz słabo interpretowalne przedziały ufności dla prognoz. Biorąc pod uwagę cel projektu, czyli ukazanie zależności między zmiennymi, a nie prognozowanie wartości, aspekty te schodzą na dalszy plan. Niemniej jednak wymienione problemy sugerują, że w dalszych badaniach warto rozszerzyć zbiór danych o pominięte w tym projekcie zmienne.

7.2 Hipotezy badawcze

W projekcie udało się potwierdzić następujące hipotezy badawcze:

- **Wartość PKB per capita wpływa pozytywnie na oczekiwaną długość życia.**
- **Gęstość zaludnienia wpływa na oczekiwaną długość życia.**
- **Położenie państwa w Afryce wpływa negatywnie na oczekiwaną długość życia.**

Potwierdza to naszą hipotezę główną - oczekiwana długość życia jest zależna od wybranych czynników ekonomicznych, społecznych oraz geograficznych. Model wyjaśnia ponad 70% zmienności oczekiwanej długości życia w krajach świata. Zdecydowanie największy wpływ na zmienną objaśnianą ma PKB per capita - parametr jest na moduł największy (tu warto przypomnieć, że model z samym PKB per capita charakteryzował się dość wysokim współczynnikiem determinacji).

7.3 Dalsze badania

Na rozwinięcie w dalszych badaniach zasługuje przede wszystkim zmienna binarna określająca, czy państwo leży w Afryce. Zmienna ta może zostać uznana za istotną nawet na bardzo restrykcyjnym poziomie istotności. Najprawdopodobniej w jej skład wchodzą pominięte w tym projekcie czynniki, których niskie/wysokie wartości są charakterystyczne dla państw afrykańskich. Podmiana tej zmiennej binarnej na inne, istotne (najlepiej ciągłe lub quasi-ciągłe) wskaźniki polepszyłaby znacznie użyteczność modelu. Można byłoby wówczas wskazać, na jakich czynnikach należy się skupić w celu wydłużenia oczekiwanej długości życia w państwach słabiej rozwiniętych.

Bibliografia

1. Boussiala, M. *Cook's Distance* 2020. https://www.researchgate.net/publication/344522968_Cook's_Distance.
2. Lumley, T., Diehr, P., Emerson, S. & Chen, L. The importance of the normality assumption in large public health data sets. *Annual Review of Public Health,* 153. https://courses.washington.edu/b511/handouts/Lumley%20Normality%20Assumption.pdf (2002).
3. Cook, R. D. Detection of Influential Observation in Linear Regression. *Technometrics* **19.** ISSN: 00401706. http://www.jstor.org/stable/1268249 (1977).
4. Breusch, T. S. & Pagan, A. R. A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica* **47,** 1287–1294. ISSN: 00129682, 14680262. http://www.jstor.org/stable/1911963 (1979).
5. Kenneth, B. Linear Regression Models with Logarithmic Transformations. https://kenbenoit.net/assets/courses/ME104/logmodels2.pdf (2011).
6. Kumar, S. & Srivastava, A. Bootstrap prediction intervals in non-parametric regression with applications to anomaly detection. https://ntrs.nasa.gov/api/citations/20130014367/downloads/20130014367.pdf (2012).

Spis rysunków

Spis tabel

Spis treści