Co wpływa na długość naszego życia?

Szymon Malec, Michał Wiktorowski

1. Wstęp

Celem niniejszej pracy jest przeanalizowanie, jak poszczególne czynniki zewnętrzne wpływają na długość naszego życia. Do analizy posłużymy się zbiorem danych charakteryzującym wiele państw świata pod kilkoma aspektami, takimi jak średni czas edukacji, spożycie alkoholu, czy powszechność szczepień na różne choroby. Dane dostępne są tutaj. Postaramy się odpowiedzieć na pytanie, co najlepiej wpływa na długość życia, a co wręcz przeciwnie. Wnioski wyciągnięte z analizy wyskorzystamy, by znaleźć dla Polski możliwe drogi do zwiększenia średniego czasu życia jej obywateli.

2. Opis danych

Poniżej zostały opisane wszystkie kolumny znajdujące się w badanych danych.

- **Country** - kolumna ta zawiera nazwy państw, które zostały uwzględnione w zbiorze danych. W sumie są to 193 różne państwa.
- **Year** - kolumna z latami. Zebrane dane pochodzą z lat 2000-2015.
- **Status** - odnosi się do stopnia rozwoju państw. Jest podzielona na dwie zasadnicze kategorie: developing, czyli państwa wciąż rozwijające się (stanowią 83% wszystkich wartości), oraz developed, czyli te już rozwinięte (pozostałe 17%).
- **Life expectancy** - oczekiwana długość życia. Podana jest w latach.
- **Adult mortality** - mówi nam o śmiertelności wśród osób dorosłych. Podane wartości oznaczają liczbę zgonów osób między 15, a 60 rokiem życia przypadającą na 1000 osób.
- **Infant deaths** - oznacza śmiertelność wśród niemowląt. Podana w liczbie zgonów na 1000 osób.
- **Alcohol** - przeciętna konsumpcja alkoholu u osób powyżej 15 roku życia w ciągu roku. Podana jako spożycie czystego alkoholu w litrach.
- **Percentage expenditure** - wydatki na zdrowie jako procent PKB na osobę.

- **Hepatitis B** - zasięg szczepień przeciw zapaleniu wątroby typu B wśród 1-latków podany w procentach.
- **Measles** - liczba zanotowanych przypadków odry na milion osób.
- **BMI** - przeciętny indeks BMI całej populacji danego kraju.
- **Under-five deaths** - liczba zgonów dzieci w wieku poniżej 5 lat na tysiąc osób.
- **Polio** - zasięg szczepień przeciw polio wśród 1-latków podany w procentach.
- **Total expenditure** - wydatki rządu na zdrowie jako procent wszystkich wydatków.
- **Diphtheria** - zasięg szczepień przeciw błonicy i krztuścowi wśród 1-latków podany w procentach.
- **HIV/AIDS** - ilość zgonów na HIV/AIDS wśród niemowlaków (0-4 lata) na 1000 osób.
- **GDP** - produkt krajowy brutto (PKB) podany w dolarach.
- **Population** - populacja danego kraju.
- **Thinness 1-19 years** - proporcja osób z niedowagą wśród osób w wieku 10-19 lat, wyrażona w procentach.
- **Thinness 5-9 years** - proporcja osób z niedowagą wśród osób w wieku 5-9 lat, wyrażona w procentach.
- **HDI** - wskaźnik rozwoju społecznego, podany jako liczba od 0 do 1.
- **Schooling** - średni czas trwania edukacji podany w latach.

3. Brakujące dane

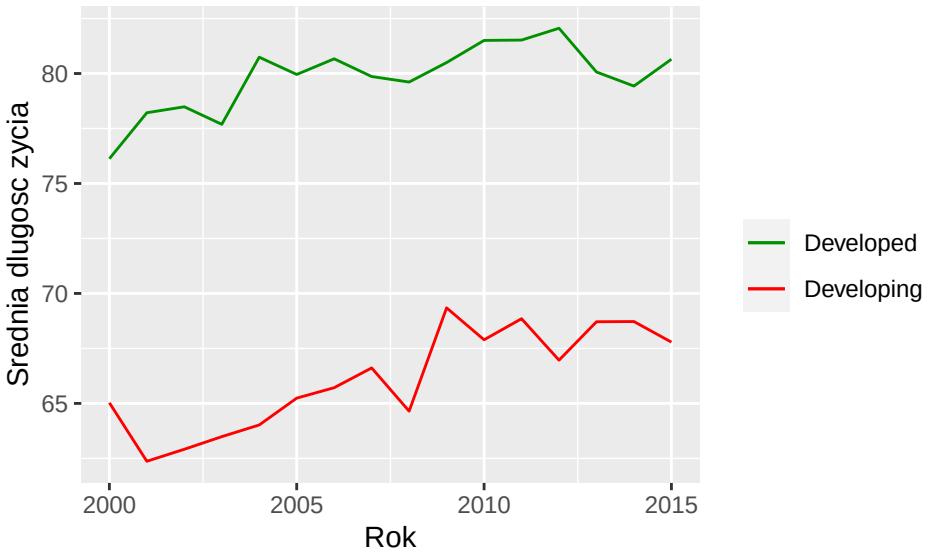
Zanim przejdziemy do analizy naszych danych, sprawdźmy, czy zawierają one jakieś braki. W tym celu zliczamy wszystkie brakujące wartości dla każdej zmiennej liczbowej w każdym roku.

year	**00**	**01**	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**	**13**	**14**	**15**
life exp.	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0
adult mortality	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0
infant deaths	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
alcohol	1	1	1	1	1	2	1	1	1	1	1	1	1	2	1	177
perc. expend.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hepatitis B	98	88	70	52	45	36	32	24	20	17	15	13	13	11	10	9
measles	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BMI	2	2	2	2	2	2	2	2	2	2	2	2	2	4	2	2
under 5 deaths	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
polio	3	3	2	2	2	2	1	1	1	1	1	0	0	0	0	0
total expend.	4	4	4	3	3	3	3	3	3	3	3	3	2	2	2	181
diphtheria	3	3	2	2	2	2	1	1	1	1	1	0	0	0	0	0
HIV AIDS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GDP	29	28	28	28	27	27	27	27	27	27	27	27	29	33	28	29
population	40	40	40	40	40	40	40	40	40	40	40	40	41	49	41	41
thinness 1-19 y	2	2	2	2	2	2	2	2	2	2	2	2	2	4	2	2
thinness 5-9 y	2	2	2	2	2	2	2	2	2	2	2	2	2	4	2	2
HDI	10	10	10	10	10	10	10	10	10	10	10	10	10	17	10	10
schooling	10	10	10	10	10	10	10	10	10	10	10	10	10	13	10	10

Tablica 1: Liczba brakujących wartości dla każdej kolumny i dla każdego roku.

4. Porównanie krajów rozwiniętych i rozwijających się

Spodziewamy się, że w krajach lepiej rozwiniętych długość życia jest większa niż w krajach uboższych. Żeby przekonać się, czy w rzeczywistości tak jest, porównamy średnie dla obu typów krajów dla każdego roku. Ponieważ państwa zawarte w danych znacznie różnią się populacją, nie możemy traktować ich na równi. Z tego powodu skorzystamy z średniej ważonej, gdzie wagami będą populacje.



Wykres 1: Porównanie średniej długości życia dla krajów wysoko i słabo rozwiniętych

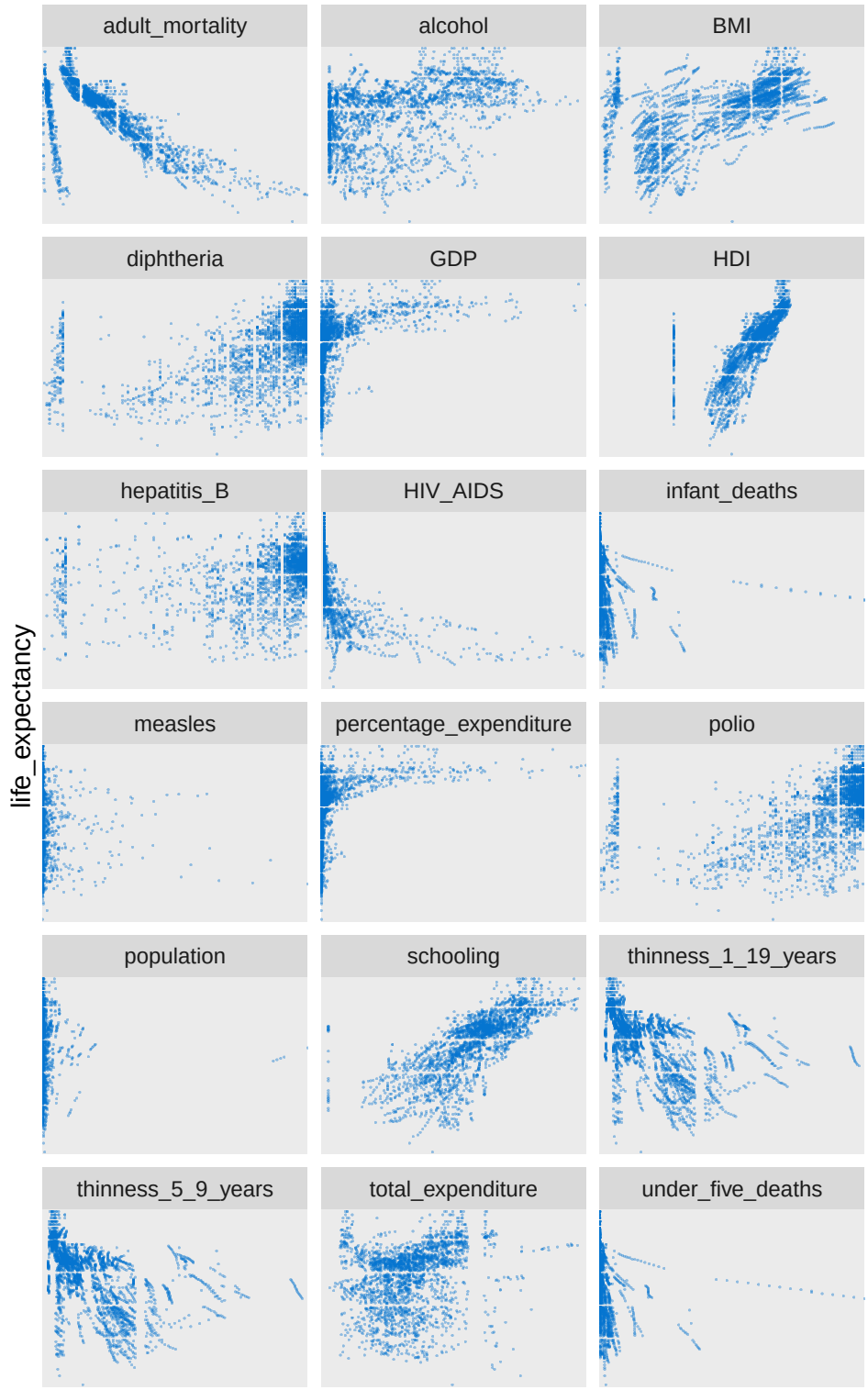
Jak możemy zauważyć na powyższym wykresie, krzywa dla państw rozwiniętych znajduje się znacznie wyżej niż dla państw rozwijających się. W obu przypadkach jednak dostrzec można w miarę podobny trend wzrostowy. Ponieważ w dalszej części pracy analizowane będą kolumny z danymi dotyczącymi nauczania oraz PKB, spójrzmy już teraz na wykresy punktowe tych wartości z zaznaczeniem, które wartości dotyczą krajów rozwiniętych, a które rozwijających się.

Wykres 2: Wykresy punktowe długości życia, po lewej od czasu nauczania, po prawej od PKB.

Na obu wykresach zauważyć można, że państwa lepiej rozwinięte wyróżniają się długim czasem nauczania, wysokim PKB oraz długim czasem życia. W dalszej części analizy państwa te traktowane będą jako wzór do osiągnięcia długowieczności.

5. Analiza zależności

Jednym z głównych celów raportu jest sprawdzenie, które czynniki mają największy wpływ na długość życia. Dobrym początkiem będzie narysowanie wykresów punktowych długości życia od poszczególnych zmiennych. Wykresy takie dają wgląd na to, jak zachowują się dane oraz pozwalają ocenić, czy występuje między nimi jakaś zależność. Dodatkowo skorzystamy ze współczynnika korelacji Spearmana, ponieważ jest to miara monotonicznej zależności, a właśnie takiej zależności szukamy. Wspomnianą korelację obliczymy między długością życia, a pozostałymi zmiennymi liczbowymi dla każdego roku z osobna.



Wykres 3: Wykresy punktowe oczekiwanej długości życia od poszczególnych zmiennych.

year	2000	2001	2002	2003	2004	2005	2006	2007
adult mortality	-0.56	-0.57	-0.58	-0.54	-0.74	-0.54	-0.59	-0.61
infant deaths	-0.58	-0.61	-0.62	-0.61	-0.62	-0.61	-0.62	-0.61
alcohol	0.43	0.41	0.44	0.43	0.44	0.44	0.46	0.45
percentage expenditure	0.52	0.49	0.53	0.54	0.51	0.54	0.53	0.49
hepatitis B	0.14	0.18	0.3	0.35	0.32	0.34	0.33	0.36
measles	-0.45	-0.46	-0.45	-0.39	-0.26	-0.2	-0.27	-0.22
BMI	0.6	0.64	0.65	0.6	0.56	0.48	0.61	0.62
under five deaths	-0.61	-0.63	-0.63	-0.63	-0.63	-0.62	-0.63	-0.63
polio	0.58	0.57	0.57	0.59	0.49	0.49	0.51	0.5
total expenditure	0.3	0.28	0.26	0.25	0.25	0.26	0.31	0.3
diphtheria	0.52	0.6	0.58	0.59	0.56	0.52	0.52	0.54
HIV AIDS	-0.72	-0.73	-0.71	-0.74	-0.76	-0.76	-0.76	-0.77
GDP	0.65	0.66	0.7	0.68	0.65	0.68	0.63	0.59
population	-0.14	-0.32	-0.13	-0.17	-0.07	-0.02	-0.17	-0.14
thinness 1-19 years	-0.54	-0.5	-0.58	-0.56	-0.59	-0.6	-0.65	-0.67
thinness 5-9 years	-0.54	-0.5	-0.55	-0.57	-0.6	-0.67	-0.7	-0.67
HDI	0.72	0.78	0.77	0.79	0.8	0.82	0.89	0.89
schooling	0.77	0.76	0.76	0.77	0.78	0.78	0.82	0.81

Tablica 2: Tabela korelacji Spearmana między oczekiwaną długością życia, a poszczególnymi zmiennymi dla lat 2000-2007.

year	2008	2009	2010	2011	2012	2013	2014	2015
adult mortality	-0.78	-0.71	-0.71	-0.72	-0.67	-0.69	-0.73	-0.74
infant deaths	-0.61	-0.6	-0.59	-0.6	-0.59	-0.58	-0.58	-0.58
alcohol	0.45	0.43	0.44	0.44	0.57	0.56	0.59	0.2
percentage expenditure	0.46	0.51	0.49	0.46	0.51	0.45	0.42	0
hepatitis B	0.39	0.36	0.38	0.33	0.33	0.39	0.42	0.47
measles	-0.14	-0.21	-0.26	-0.15	-0.2	-0.2	-0.21	-0.23
BMI	0.6	0.61	0.55	0.55	0.55	0.53	0.51	0.55
under five deaths	-0.62	-0.63	-0.62	-0.63	-0.62	-0.6	-0.6	-0.6
polio	0.53	0.43	0.52	0.5	0.52	0.54	0.53	0.56
total expenditure	0.19	0.31	0.34	0.17	0.3	0.36	0.37	-1
diphtheria	0.52	0.47	0.52	0.49	0.51	0.56	0.52	0.55
HIV AIDS	-0.77	-0.77	-0.76	-0.78	-0.76	-0.77	-0.77	-0.78
GDP	0.59	0.62	0.59	0.6	0.66	0.61	0.58	0.57
population	-0.17	0.06	0.01	-0.03	0.01	-0.13	-0.14	-0.01
thinness 1-19 years	-0.68	-0.67	-0.62	-0.61	-0.58	-0.62	-0.62	-0.64
thinness 5-9 years	-0.67	-0.66	-0.62	-0.61	-0.61	-0.63	-0.66	-0.66
HDI	0.88	0.87	0.89	0.9	0.9	0.9	0.9	0.91
schooling	0.79	0.78	0.81	0.82	0.82	0.83	0.84	0.84

Tablica 3: Tabela korelacji Spearmana między oczekiwaną długością życia, a poszczególnymi zmiennymi dla lat 2008-2015.

Możemy zauważyć, że rozrzut punktów na poszczególnych wykresach jest bardzo zróżnicowany. Na jednych wydaje się on być całkowicie losowy, natomiast na innych widać pewną zależność. Z tabeli korelacji odczytać możemy, że najsilniejszą ujemną zależność z oczekiwanym czasem życia mają zmienne związane ze śmiertelnością oraz niedowagą. Natomiast pozytywną korelację przejawiają zmienne takie jak BMI, zasięg szczepień na błonicę oraz polio, PKB, a w szczególności średni czas nauczania i wskaźnik rozwoju społecznego. Wysoka korelacja tego ostatniego nie jest zaskoczeniem, ponieważ wskaźnik ten jest wyliczany m. in. z oczekiwanego czasu życia, a więc wartości te naturalnie są od siebie zależne. Interesująca może być za to zależność pomiędzy czasem życia, a nauczaniem. Na wykresie punktowym zauważyć można między nimi liniową korelację. Z kolei, jeśli spojrzymy na wykres dotyczący PKB, na myśl przychodzi nam zależność logarytmiczna. Aby dokładniej zbadać te zależności, poddamy wspomniane zmienne głębszej analizie.

6. Wpływ edukacji na długość życia

Jako pierwszy przeanalizujemy wpływ czasu edukacji na oczekiwaną długość życia. W tym przypadku rozważymy dane wyłącznie z 2015 roku, czyli te najbardziej aktualne. Na wykresie 4 można zauważyć liniową zależność danych, zatem do zbadania korelacji możemy użyć współczynnika Pearsona. Przyjmuje on wartość $R \approx 0{,}752$, więc jest to dość mocna

korelacja. Dopasujemy teraz do danych prostą regresji korzystając z metody najmniejszych kwadratów. Przyjmijmy model

$$Y_i = ax_i + b + \epsilon_i,$$

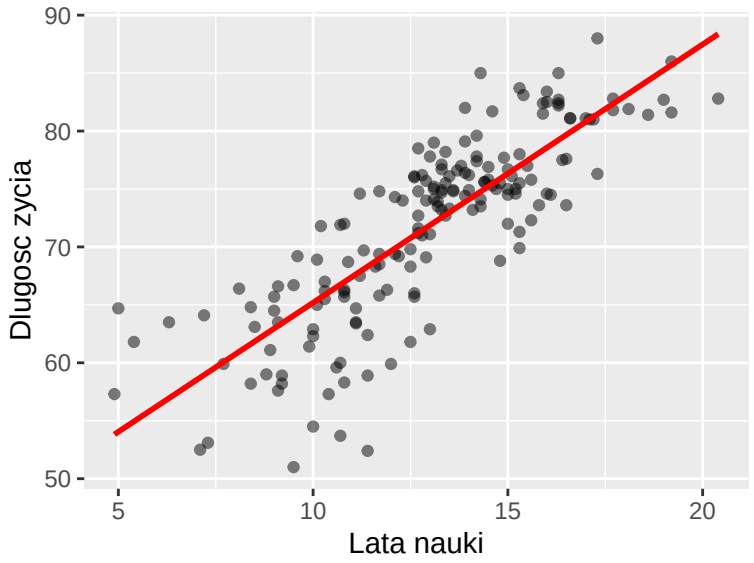
gdzie x_i to dane dotyczące czasu nauczania, a ϵ_i są i.i.d. ze średnią równą 0 i skończoną wariancją. Oznaczmy dane z czasem życia jako y_i. Wspomniana metoda polega na znalezieniu takich współczynników a, b dla których funkcja

$$S(a,b) = \sum_{i=1}^{n}(y_i - ax_i - b)^2$$

przyjmuje wartość najmniejszą. Rozwiązaniem jest para estymatorów

$$\begin{cases} \hat{a} = R\frac{S_y}{S_x} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\ \hat{b} = \bar{y} - a\bar{x} \end{cases}$$

gdzie R jest współczynnikiem korelacji Pearsona, a S_x, S_y są próbkowymi odchyleniami standardowymi.

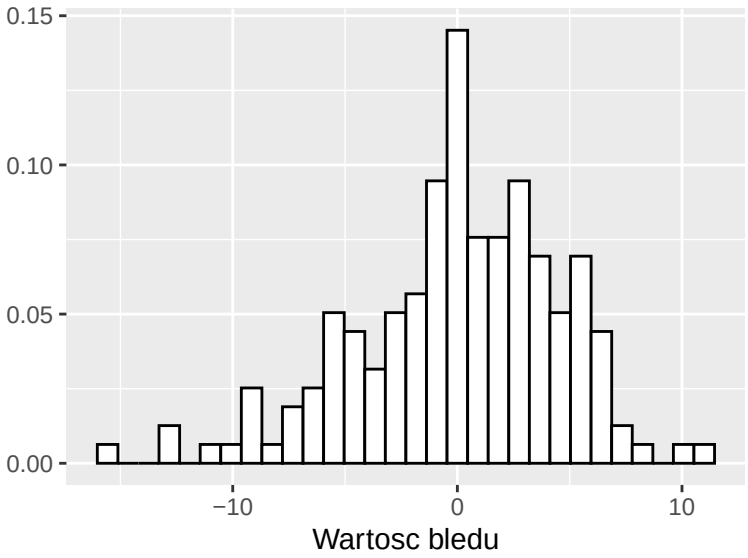


Wykres 4: Prosta regresji wyznaczona dla danych.

Kolejnym punktem będzie analiza residuów (błędów)

$$e_i = y_i - \hat{y}_i \ ,$$

gdzie $\hat{y}_i = \hat{a}x_i + \hat{b}$. W celu zbadania rozkładu residuów, spójrzmy na ich histogram



Wykres 5: Histogram residuów.

Kształt histogramu jest zbliżony do krzywej gaussowskiej. Średnia wartość residuów wynosi $\mu_e = 0$, a ich wariancja $\sigma_e^2 = 20{,}81$. Posłużymy się testem Kołmogorowa-Smirnova w celu zbadania normalności rozkładu błędów. Przedstawmy hipotezy:

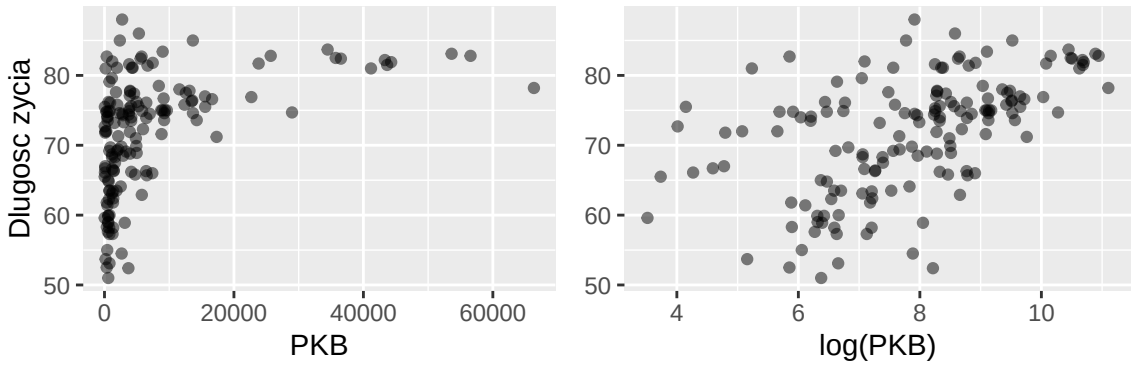
- $\mathcal{H}_0$: wartości residuów są z rozkładu normalnego $\mathcal{N}(0, 20.81)$
- $\mathcal{H}_1$: wartości residuów nie są z rozkładu normalnego $\mathcal{N}(0, 20.81)$

Wyznaczona p-wartość wynosi $p = 0{,}2532$. Ponieważ otrzymany wynik jest wystarczająco duży, to nie mamy podstaw do odrzucenia hipotezy zerowej i możemy przyjąć, że dane pochodzą z rozkładu normalnego $\mathcal{N}(0, 20.81)$.

Analiza residuów jest niezwykle istotna, kiedy decydujemy się robić predykcję danych. Znając rozkład błędów, możemy wyznaczyć przedziały ufności o danym poziomie istotności dla przewidywanych wyników. Innymi słowy możemy wyznaczyć prawdopodobieństwo z jakim predykowana wartość zmieści się w konkretnym przedziale. Wyniki, które otrzymaliśmy mogą być podstawą do wykonania takiej predykcji, jednak wcześniej należałoby jeszcze sprawdzić, czy residua są od siebie niezależne oraz czy ich wariancja jest stała.

7. PKB państwa, a długość życia

Podobnie jak w przypadku długości edukacji, będziemy analizować dane dotyczące PKB wyłącznie z 2015 roku. W przeciwieństwie do poprzedniego przypadku, dane te nie są zależne liniowo. Ich wykres punktowy kształtem przypomina bardziej zależność logarytmiczną. Sprawdźmy zatem, czy w rzeczywistości tak jest nakładając logarytm na wartości PKB.



Wykres 6: Wykresy punktowe z surowych danych po lewej i z przetransformowanych po prawej.

Możemy zauważyć, że dane po transformacji przypominają bardziej zależne liniowo, choć są dość mocno rozrzucone. Współczynnik korelacji Pearsona dla przetransformowanych danych wynosi 0.52. Przyjmnijmy model

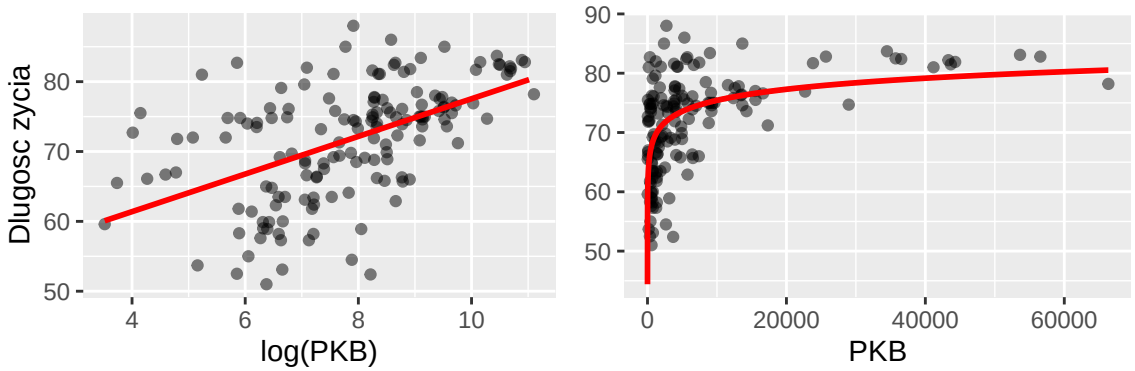
$$Y_i = a\log(x_i) + b + \epsilon_i,$$

gdzie x_i są danymi z PKB, a ϵ_i są i.i.d. o średniej równej 0 i skończonej wariancji. Aby jednak otrzymać model lioniowy jak w punkcie 6, podstawiamy $z_i = \log(x_i)$ i otrzymujemy

$$Y_i = az_i + b + \epsilon_i.$$

Teraz już możemy skorzystać z metody najmniejszych kwadratów, aby dopasować prostą regresji. W ten sposób otrzymujemy estymatory postaci

$$\begin{cases} \hat{a} = \frac{\sum_{i=1}^{n}(z_i-\bar{z})(y_i-\bar{y})}{\sum_{i=1}^{n}(z_i-\bar{z})^2} \\ \hat{b} = \bar{y} - a\bar{z} \end{cases}$$

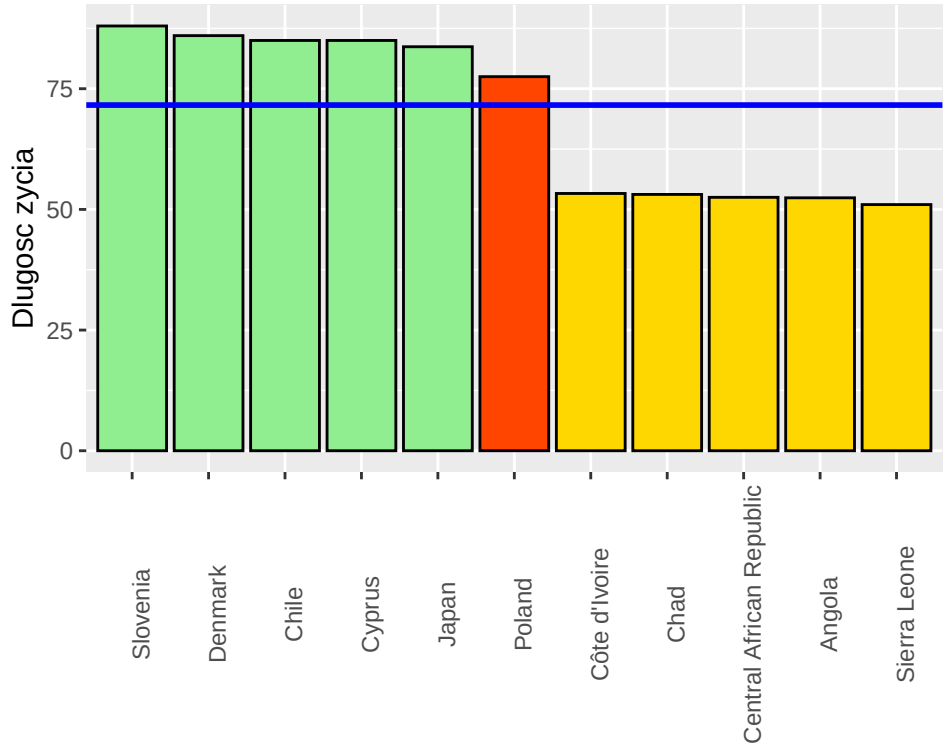


Wykres 7: Po lewej prosta regresji dla przetransformowanych danych. Po prawej krzywa regresji dopasowana do oryginalnych danych.

Jak możemy zobaczyć, krzywa logarytmiczna w miarę pokrywa się z danymi, więc można przyjać ją za przybliżenie zależności pomiędzy dwoma zmiennymi. Ważną obserwacją jest to, że największy wzrost mamy dla bardzo małych wartości PKB - im jest większe, tym wolniej rośnie długość życia. Oznacza to, że dla państw z małym PKB, nawet niewielki jego wzrost może przyczynić się do znacznego zwiększenia długości życia, zaś dla państw lepiej rozwinietych, wzrost PKB ma już na to nieznaczny wpływ.

8. Możliwości dla Polski

Wyniki przeprowadzonych analiz wykorzystamy, by określić co pozwoliłoby Polsce wydłużyć średni czas życia. Zacznijmy od porównania Polski z resztą świata. Rozważymy dane z 2015 roku - najnowsze, którymi dysponujemy. Średnia długość życia na globie wyniosła wtedy 71,62 lata. Mianem najdłużej żyjących ludzi mogli się wtedy poszczycić Słoweńcy z imponującą średnią życia aż 88 lat. Natomiast najkrócej żyjącymi ludźmi okazali się być obywatele Sierra Leony z wynikiem zaledwie 51 lat. Polacy osiągneli wynik 77.5 lat, co stawia nas na 42 miejscu w rankingu - wynik dobry, ale nie najlepszy.



Wykres 8: Porównanie oczekiwanej długości życia w Polsce z pięcioma najwyższymi wartościami, pięcioma najniższymi wartościami oraz średnią z całego świata (oznaczoną niebieską linią).

Zdefiniujmy statystykę Z w następujący sposób

$$Z_X(x_0) = \frac{\#\{x \in X : x < x_0\}}{\#X}.$$

Mówi ona jaka część liczb ze zbioru X jest poniżej wartości x_0. Wykorzystamy ją, by policzyć jaka część państw jest poniżej wartości każdej zmiennej dla Polski. Statystyki te wyliczymy dla roku 2014, ponieważ w roku 2015 występuje sporo braków danych, co widoczne jest w tabeli 1.

Zmienna	**Wartość statystyki** Z
life expectancy	0.77
adult mortality	0.07
infant deaths	0.5
alcohol	0.94
percentage expenditure	0.6
hepatitis B	0.73
measles	0.37
BMI	0.77
under five deaths	0.47
polio	0.55
total expenditure	0.55
diphtheria	0.87
HIV AIDS	0.63
GDP	0.83
population	0.64
thinness 1-19 years	0.35
thinness 5-9 years	0.36
HDI	0.85
schooling	0.91

Tablica 4: Wartości statystyki Z dla każdej zmiennej w przypadku Polski.

Przypomnijmy, że zmienne wpływające pozytywnie na długość życia to BMI, zasięg szczepień na błonicę oraz polio, PKB, średni czas nauczania i wskaźnik rozwoju społecznego. Na powyższej tabeli zauważymy, że z wymienionych zmiennych najniższą wartość ma statystyka obliczona dla szczepień na polio i wynosi ona 0.55, co stawia nasz kraj lekko powyżej połowy. Stąd, aby zwiększyć długość życia Polaków, państwo mogłoby zwiększyć zasięg szczepień na wspomnianą chorobę.

Natomiast zmiennymi skorelowanymi ujemnie są zmienne związane ze śmiertelnością oraz niedowagą. W tabeli zobaczymy, że śmiertelność u dorosłych w Polsce jest bardzo niska, jednak śmiertelność w przypadku niemowląt i dzieci poniżej 5-tego roku życia stawia nasze państwo mniej więcej po środku w stosunku do reszty świata. Z kolei watości statystyki Z dla kolumn dotyczących niedowagi mieszczą się w okolicach 0.35, co także nie jest najlepszym wynikiem. Dobrą strategią byłoby zatem zapobieganie zgonów u małych dzieci w pierwszej kolejności, a następnie walka z niedowagą.

9. Podsumowanie

Zestawienie danych, które postanowiliśmy przeanalizować okazało się być bardzo różnorodne, a do rozpatrzenia było wiele czynników. Nie wszystkie jednak były aż tak istotne. Określiliśmy to na podstawie wykresów punktowych między oczekiwaną długością życia, a poszczególnymi zmiennymi (wykres 3). Na jednych wykresach była widoczna pewna zależność w danych, za to na innych dane były mocno rozproszone. Potwierdziliśmy nasze przypuszczenia dotyczące zależności wyliczając odpowiednie współczynniki korelacji. Jednym z czynników najbardziej skorelowanych z oczekiwaną długością życia okazał się być średni czas edukacji. Na podstawie wyników analizy, wraz ze wzrostem czasu poświęconego na naukę, wzrasta też oczekiwana długość życia.

Następnie porównaliśmy średnią długość życia Polaków z obywatelami pozostałych krajów. W 2015 roku osiągneliśmy wynik 77,5 lat - większy od średniej o niemal 6 lat i całościowo jest to 42 najwyższy wynik (wykres 8). Trochę nam jednak brakuje do liderów rankingu jakim są Słoweńcy z wynikiem 88 lat. Przeanalizowaliśmy jakie rezultaty w poszczególnych kategoriach osiąga Polska na tle reszty państw i na tej podstawie pokazaliśmy, co można byłoby zrobić, aby poprawić nasze wyniki.

Okazało się, że statystyki związane ze śmiertelnością u dzieci, oraz niedowagą są w Polsce bardzo niekorzystne. Rozwiązaniem, które mogłoby poprawić nasz wynik byłaby więc walka z niedowagą, oraz zredukowanie śmiertelności u dzieci. Są też takie charakterystyki, gdzie osiągamy naprawdę dobre rezultaty, jak chociażby długość edukacji. Nie jest to jednak powód do tego, aby rezygnować z inwestowania w oświatę, ponieważ została pokazana ścisła zależność między oczekiwaną długością życia, a szkolnictwem. Stąd inwestowanie w naukę jest również niezwykle istotne.