Analiza ekonometryczna liczby zgonów na Covid-19 dla państw o wysokim HDI, które uruchomiły programy szczepień.

Przemysław Pietrzak 109064

Jakub Kołpa 109434

Spis treści

Wprowadzenie przedmiotu analizy.

Wciąż trwająca, choć w wielu państwach medialnie wygłuszona, pandemia Covid-19 przyniosła poważne konsekwencje dla światowego porządku gospodarczego. Jej konsekwencje rozszerzają się z płaszczyzny społecznej czy czysto biologicznej także na ekonomiczną poprzez liczne nieprzewidziane efekty i bezprecedensowe rozwiązania przyjęte w walce z rozprzestrzenianiem wirusa, a mające daleko idące konsekwencje ekonomiczne.

Powszechnie stosowane lockdowny przynosiły negatywne krótkookresowe skutki ekonomiczne[1]. Dały się zaobserwować silne fluktuacje na rynkach pracy, czemu obraz na przykładzie Stanów Zjednoczonych daje internetowa aplikacja[2] przygotowana przez National Bureau of Economic Research w Cambridge Massachusetts. Odwołując się do wspomnianych danych można zaobserwować spadki współczynników zatrudnienia do kwietnia 2020 między 10%-30% w zależności od grupy dochodowej w stosunku do poziomów ze stycznia.

Te niebezpieczne i wymagające natychmiastowych reakcji efekty idą w parze ze znaczącym uszczerbkiem na kondycji produkcji światowej[3], której wolumen spadł w 2020 roku o 3,3%.

Ten wybiórczy zarys konsekwencji pandemii Covid-19 daje podstawy, by uznać podjęty temat za istotny z perspektywy ekonomicznej.

Jego powiązanie z wymiarem ekonomicznym jest konsekwencją przyjmowanych środków zapobiegawczych, które stanowiły reakcje na zmieniające się statystyki dotyczące zagrożenia wywołanego pandemią. W szczególności jako miarę intensywności zjawiska przyjmuje się relatywną liczbę zgonów wywołanych dotychczas pandemią Covid-19.

Praca koncentruje się zatem wokół modelowania zmiennej określającej liczbę zgonów na milion mieszkańców.

---

[1] *World Economic Outlook*. Washington, D.C: International Monetary Fund, 2020, s.67

[2] Chetty, Raj. *How Did COVID-19 and Stabilization Policies Affect Spending and Employment?: A New Real-time Economic Tracker Based On Private Sector Data*. Cambridge, MA : National Bureau of Economic Research, 2020.

[3] World Bank. *World Bank national accounts data, and OECD National Accounts data files.*, The World Bank Group, 2022, https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG

Opis zbioru danych oraz zmiennych wykorzystanych przy konstrukcji modeli ekonometrycznych.

W pracy zdecydowano się użyć danych zgromadzonych w elektronicznej publikacji „Our World in Data" kompletującej dane o problemach o charakterze globalnym takich jak bieda, choroby, głód, zmiany klimatyczne, wojny oraz nierówności społeczne[4]. Jest to projekt autorstwa „Global Change Data Lab" – fundacji zarejestrowanie w Anglii i Walii założonej przez historyka i ekonomistę Maxa Rosera. Zespół badawczy ma swoją siedzibę na Uniwersytecie Oksfordzkim.

Zdecydowano, że modelowane dane będą miały charakter przekrojowy. W tabeli poniżej umieszczono potencjalnie wartościowe w procesie modelowania zmienne, których istotność w kontekście problemu została rozważona w oparciu o wstępne rozpatrzenie ich charakterystyk w kontekście rozważanego problemu.

Tabela 1. Opis najważniejszych zmiennych używanych w pracy

Nazwa zmiennej	Opis zmiennej
total_cases	Łączna liczba potwierdzonych przypadków COVID-19 podzielona przez milion mieszkańców w celu zapewnienia porównywalności danych. Liczba przypadków może obejmować przypadki prawdopodobne, jeśli zostały zgłoszone.
total_deaths	Całkowita liczba zgonów przypisanych do COVID-19 podzielona przez milion mieszkańców w celu zapewnienia porównywalności danych. Liczby mogą obejmować zgony prawdopodobne, jeśli zostały zgłoszone.

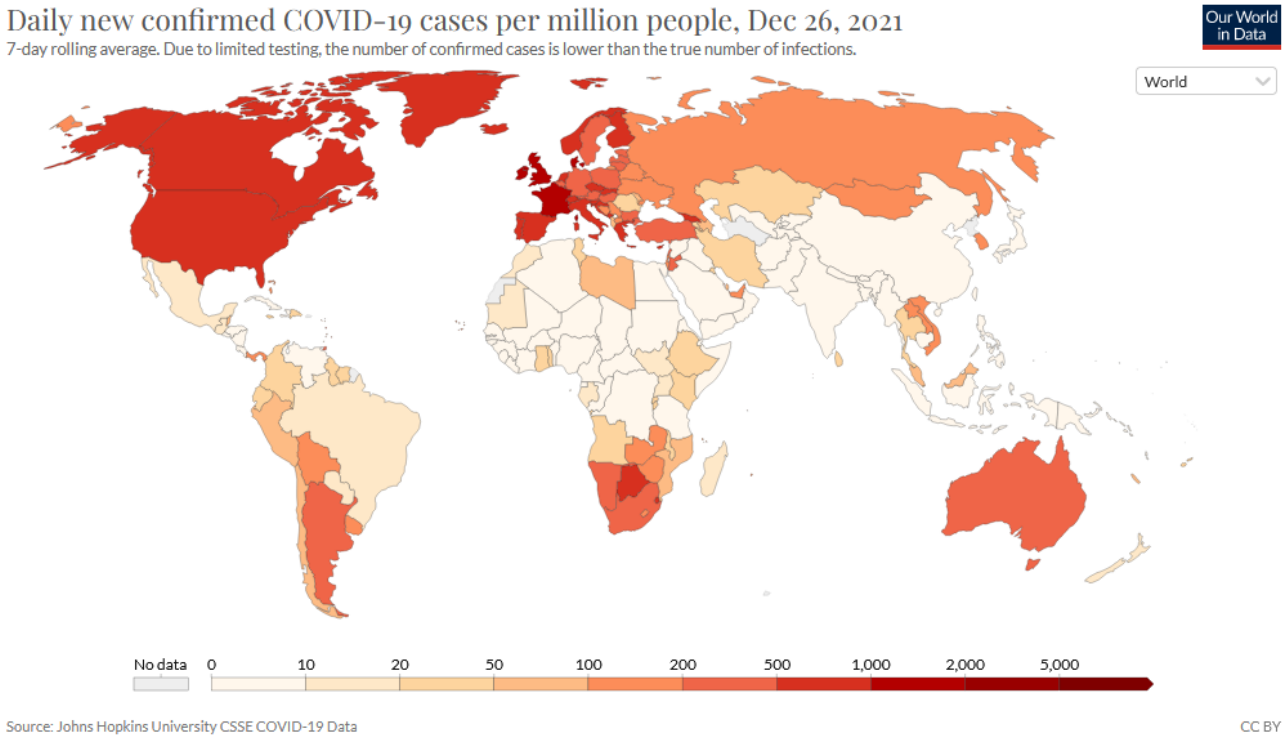
[4] Hannah Ritchie, Edouard Mathieu, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Esteban Ortiz-Ospina, Joe Hasell, Bobbie Macdonald, Diana Beltekian and Max Roser (2020) - "Coronavirus Pandemic (COVID-19)". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/coronavirus', pobrano: 23.04.2022.

total_tests	Liczba testów dla COVID-19 podzielona przez milion mieszkańców w celu zapewnienia porównywalności danych.
total_vaccinations	Całkowita liczba podanych dawek szczepionki COVID-19 podzielona przez milion mieszkańców w celu zapewnienia porównywalności danych.
population_density	Liczba osób podzielona przez powierzchnię kraju, mierzoną w kilometrach kwadratowych, ostatni dostępny rok.
aged_65_older	Odsetek ludności w wieku 65 lat i więcej, ostatni dostępny rok.
aged_70_older	Odsetek ludności w wieku 70 lat i więcej, ostatni dostępny rok.
gdp_per_capita	Produkt krajowy brutto według parytetu siły nabywczej (w cenach stałych z 2011 roku w dolarach międzynarodowych), ostatni dostępny rok
extreme_poverty	Odsetek ludności żyjącej w skrajnym ubóstwie, ostatni dostępny rok od 2010 r.
life expectancy	Oczekiwana długość życia w chwili urodzenia w 2019 r.
positive_rate	Udział pozytywnych testów COVID-19 wśród wszystkich zrobionych testów, podany jako średnia ruchoma z 7 dni.
stringency_index	Wskaźnik surowości reakcji rządu: miara złożona oparta na 9 wskaźnikach reakcji, w tym zamknięcia szkół, miejsc pracy i zakazów podróżowania, skalowana do wartości od 0 do 100 (100 = najostrzejsza reakcja).
population	Liczba ludności, ostatnie dostępne dane.
median_age	Mediana wieku ludności, prognoza ONZ na 2020 r.
female_smokers	Odsetek palących kobiet, ostatni dostępny rok.
male_smokers	Odsetek palących mężczyzn, ostatni dostępny rok.

hospital_beds_per_thousand	Łóżka szpitalne na 1 000 osób, ostatni dostępny rok od 2010 r.
human_development_index	Złożony wskaźnik mierzący średnie osiągnięcia w trzech podstawowych wymiarach rozwoju społecznego - długim i zdrowym życiu, wiedzy i godnym standardzie życia. Wartości na rok 2019, zaczerpnięte z http://hdr.undp.org/en/indicators/137506.

Część zmiennych przekształcono na wielkości względne w celu zagwarantowania porównywalności danych oraz z uwagi na taki charakter części zmiennych występujących w zestawie danych. Zmienne poddane przekształceniu polegającemu na podzieleniu przez liczbę mieszkańców to: total_deaths, total_cases, total_tests, total_vaccinations. W takim razie ich sens poddany został zmianie i zamiast określać całkowitą liczbę zgonów, przeprowadzonych testów, szczepień w danym społeczeństwie określają ich liczbę na milion mieszkańców.

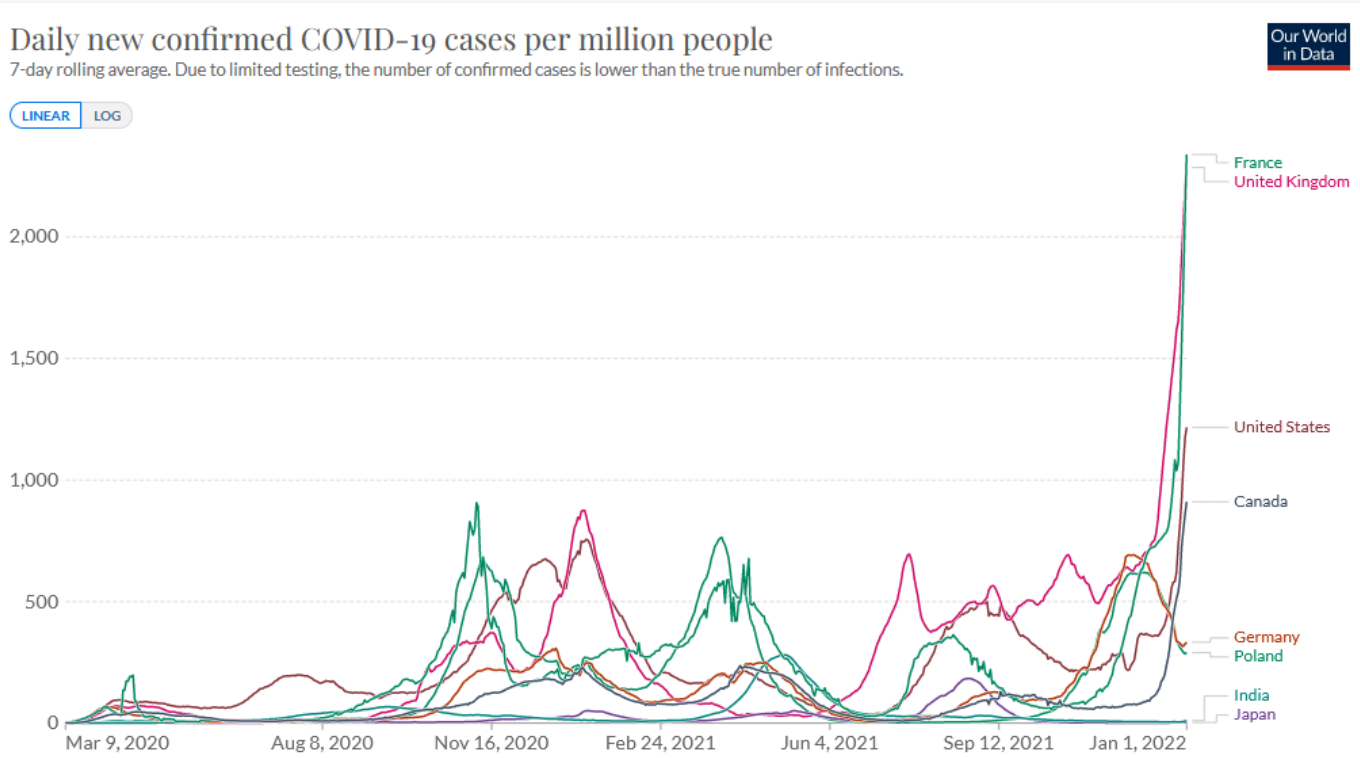
Dla następujących danych policzono średnią z okresu 90 dniowego począwszy od 01.10.2021. Jest to to przedział obejmujący okres wzmożonej liczby przypadków Covid-19, w obrębie licznych wysokorozwiniętych państw.



Rysunek 1. Nowe przypadki zachorowań na Covid-19 na milion mieszkańców 26.12.2021.

Źródło: Johns Hopkins University CSSE Covid-19 Data

Zmienne poddano uśrednieniu dla ujęcia dynamiki pandemii oraz reakcji rządów (zmienna *stringency_index*), która różniła się w czasie względem regionów (występowanie fal z opóźnieniem lub w innych okresach).



Rysunek 2. Nowe przypadki zachorowań na Covid-19 na milion mieszkańców dla 8 krajów od początku pandemii.

Źródło: Our World in Data

Eksploracja danych

Histogramy oraz wykresy typu scatterplot posłużyły do wstępnej oceny wpływu zmiennych objaśniających na zmienną objaśnianą i selekcji odpowiedniej, obiecującej grupy zarówno samych zmiennych jak i uwzględnionych w analizie państw o podobnych charakterystykach, aby znaczące różnice dotyczące poziomu rozwoju państw oraz świadczące o nim braki w danych nie przyczyniały się do zakłamania wyników.

W szczególności usunięto obserwacje odstające dla państw o wysokiej gęstości zaludniania, krajów które nie raportowały śmierci wywołanych przez Covid-19 oraz co najważniejsze krajów, których Human Development Index nie przekracza 0.7, a także takie, które nie wprowadziły programu szczepień. Ostatni krok, wydzielający państwa rozwinięte, wydaje się najistotniejszy w kontekście wyników analizy, bo z samym HDI wiążą się takie własności społeczeństw jak: *life_expectancy* - poziom usług służby ochrony zdrowia, struktura

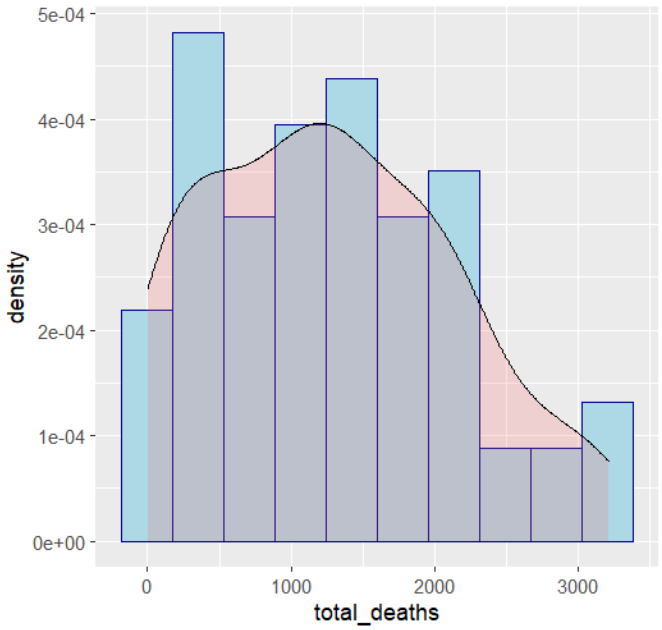
demograficzna społeczeństwa, zamożność – GNI per capita oraz edukacja. Ta operacja służy ujednoliceniu próby charakteryzującej się prawidłowościami, które zamierza się ująć w model ekonometryczny.

Poniżej umieszczone zostały histogramy obrazujące rozkłady zmiennych przed dokonaniem selekcji. Łatwo nich dostrzec silnie skośne rozkłady zawierające odstające obserwacje. Ta obserwacja wraz z analizą kierunków wpływu zmiennych pozwala poddać selekcji dane z uwagi na niezgodne z teorią obserwacje (kierunki wpływu części zmiennych).

Wykres 1. Histogramy rozkładów zmiennych przed dokonaniem selekcji.

Źródło: Opracowanie własne

Zauważa się zróżnicowanie danych w podgrupach, które uzasadnia się zróżnicowanym poziomem zaawansowania społeczeństw, które poza charakterystykami tych społeczeństw znajduje odzwierciedlenie w jakości dostarczanych danych. Dlatego też zdecydowano o ich wykluczeniu ze zbioru danych.



Wykres 2. Histogram zmiennej total_deaths.

Źródło: Opracowanie własne

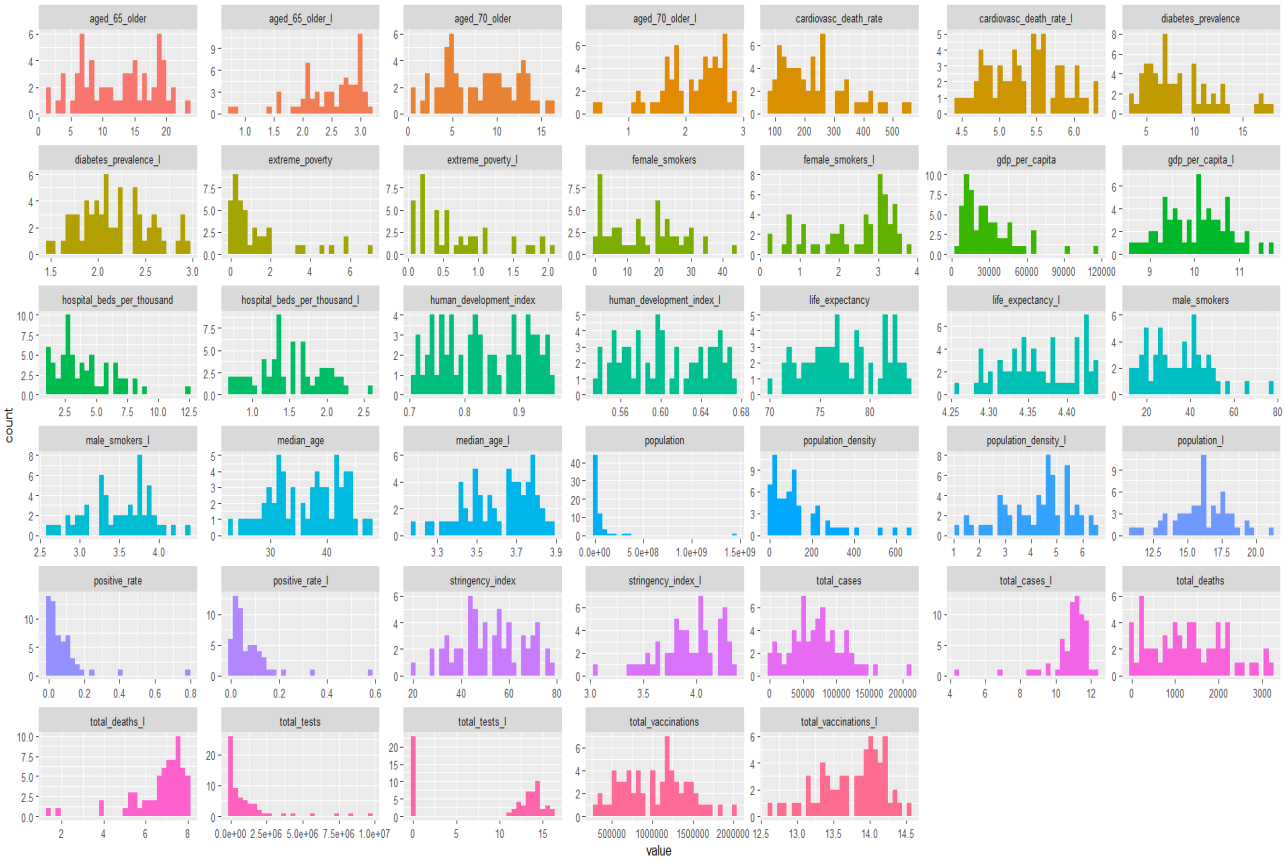
```
        Jarque Bera Test

        data:  df$total_deaths

        X-squared = 2.657, df = 2, p-value = 0.2649
```

Test Jarque Bera nie daje podstaw do odrzucenia hipotezy zerowej o normalności rozkładu zmiennej *total_deaths* przy poziomie istotności p=0.05. Co ułatwia skorzystanie z klasycznej postaci liniowego modelu ekonometrycznego. Dodatkowo przekształcenie logarytmiczne pozbawia tej własności rozkład zmiennej zależnej.

Histogramy niektórych zmiennych sugerują zastosowanie przekształcenia logarytmicznego dla uzyskania postaci bliższej rozkładowi normalnemu. Wpływ tego przekształcenia można zaobserwować na poniższych histogramach oraz wykresach pudełkowych. Jak widać transformacje nie pozwalają na uzyskanie widocznie „lepszych" postaci rozkładów dla większości zmiennych.



Wykres 3. Histogramy zmiennych i ich logarytmów.

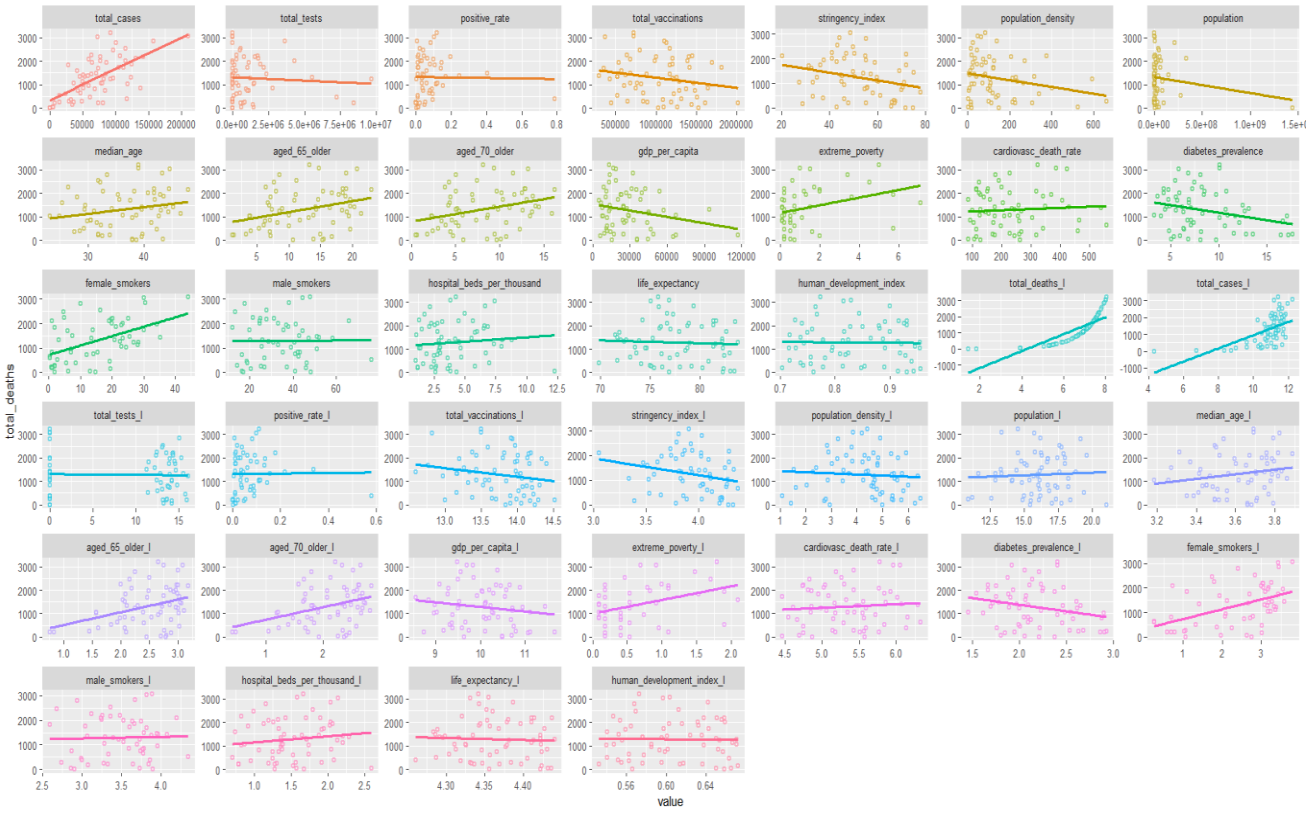
Źródło: Opracowanie własne

Co się tyczy wykresów pudełkowych to uwidaczniają one poprawę własności takich zmiennych jak *population_density* czy *gdp_per_capita*. Znaczenie tych zmian zostanie jednak ostatecznie zweryfikowanie podczas estymacji modeli. Inne zmienne na czele ze zmienną zależną *total_deaths* wykazuje się mniej praktyczną postacią rozkładu.

Wykres 4. Wykresy pudełkowe zmiennych i ich logarytmów.

Źródło: Opracowanie własne

Regresje pomocnicze umieszczone na wykresach typu scatterplot pozwalają usazadnić przewidywania co do kierunku wpływu takich kluczowych zmiennych jak: *total_vaccinations*, *stringency_index*, *gdp_per_capita*, *human_development_index* czy zmiennych odnoszących się do struktury demograficznej - *median_age*, *aged_65_older*, *aged_70_older* (co było brane pod uwagę podczas doboru próby), śmiertelnością na choroby układu krążenia oraz powszechność cukrzycy (częściowo charakter choroby cywilizacyjnej – stąd dodatnia korelacja). Natomiast inne jak: *population_density*, *total_beds_per_thousand* czy *positive_rate* dały mniej oczekiwane wyniki i prawdopodobnie wymagają interpretacji w szerszym kontekście.

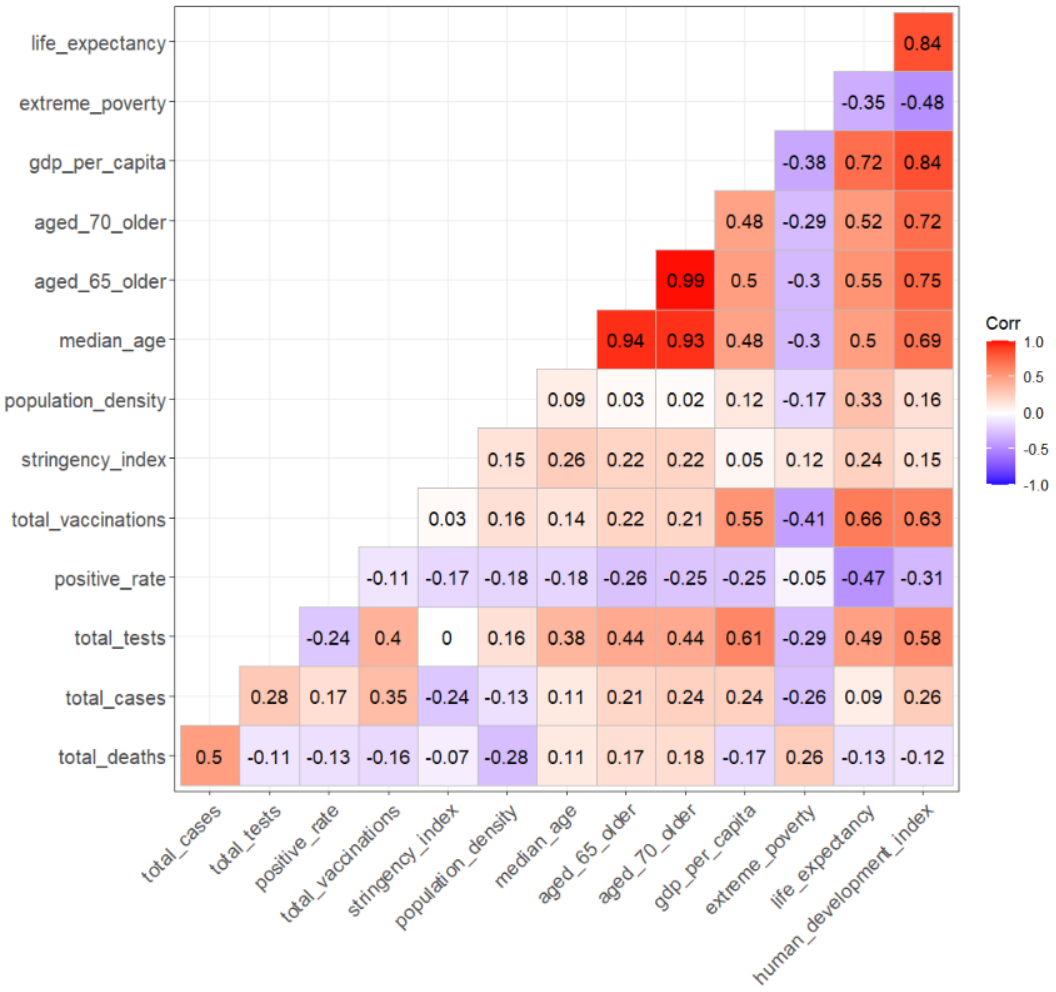


Wykres 5. Regresje pomocnicze zmiennych i ich logarytmów.

Źródło: opracowanie własne

Podczas rutynowego przeglądu danych warto również przyjrzeć się macierzy korelacji zmiennych, aby dokonać wstępnego przeglądu zależności między potencjalnymi zmiennymi objaśniającymi a zmienną objaśnianą. Dodatkowo warto zwrócić uwagę na siłę korelacji między pozostałymi zmiennymi. Jej wysoka wartość może sugerować wystąpienie problemu współliniowości stochastycznej rzutującej na efektywność estymatora (zwiększająca wariancję). Występowanie tego problemu zostanie jednak rozpatrzone po formalnym zdefiniowaniu modelu.

W szczególności warto zwrócić uwagę na skorelowanie zmiennych *total_vaccinations*, *gdp_per_capita* oraz *human_development_index* jako że wszystkie te zmienne zdają się posiadać potencjał przy wyjaśnianiu zmienności zmiennej *total_deaths*.



Wykres 6. Korelogram zmiennych.

Źródło: Opracowanie własne

Wybór zmiennych i postaci modelu

Przeprowadzono możliwie liczne i wyczerpujące kombinacje parametrów estymacji modeli liniowych przy pomocy napisanych skryptów, aby uprościć proces doboru zmiennych. Kombinacje zmiennych były rozpatrywane pod względem istotności, wpływu na reszty, testu RESET oraz współliniowości. Te podstawowe charakterystyki modeli oraz preferencja prostoty (jeżeli przekształcenie logarytmiczne nie gwarantuje znacznie lepszych rezultatów to nie stosuje się go) i interpretowalność wpływu (unika się kwadratów oraz zmiennych interakcyjnych, choć i takie zostały przetestowane).

Każdy dobór oscylował jednak wokół zasadniczych osi jakie wyznaczały zmienne związane z liczbą przypadków COVID-19 (z uwagi na niewątpliwe znaczenie tej zmiennej pozbawienie jej modelu mogłoby skutkować błędem pominiętej zmiennej), poziomem rozwoju społeczeństwa, ich strukturą demograficzną, czy charakterystykami zdrowotnymi. Zmienna, na którą szczególnie zwrócono uwagę są szczepienia wykonane na milion mieszkańców[5].

Na uzasadnienie doboru zmiennych składa się kilka następujących punktów:

(w nawiasach dla jasności zaznaczono spodziewany kierunek wpływu na zmienną *total_deaths*, bez szczegółowych przewidywań co do siły)

- liczba przypadków jest naturalnym i koniecznym dla poprawnej specyfikacji wyborem (*liczba_przypadków* +)
- charakterystyka choroby COVID-19, której przejście zazwyczaj wiąże się z poważniejszymi konsekwencjami dla osób starszych uzasadnia rozpatrywanie zmiennych związanych ze struktura demograficzną (*median_age, aged_65_older, aged_70_older, life_expectancy* +)
- charakterystyki dotyczące stanu zdrowia społeczeństwo odzwierciadlają stan służby zdrowia jak i poziom rozwoju społeczeństwa, który nie jest w pełni ujęty np. przez zmienną *gdp_per_capita* (*cardio_vasc_death_rate, diabeties prelevance*+)
- zmienne powiązane z zawansowaniem rozwoju społeczeństwa jak *human_development_index* czy *gdp_per_capita* pozwalają ująć poziom przygotowania sanitarnego do zwalczania pandemii
- szczepienia ochronne zapewniają średnio kilkumiesięczną redukcję prawdopodobieństwa ciężkiego przejścia COVID-19 (skuteczność zależy od dawki, typu oraz tego jak dawno podano dawkę)

Dalej postępując zgodnie z zarysowaną metodyką dokonano wyboru poniższej postaci modelu liniowej regresji. Tabela z wyestymowanymi modelami znajduje się poniżej:

[5] Z uwagi na postać danych (uśrednienie wartości z 90 dni), w których *total_deaths* odnoszą się do zgonów od początku pandemii, szczepienia być może nie znajdą w wynikach estymacji oszacowań parametrów właściwego odzwierciedlenia z uwagi na ich stosunkowo późne wprowadzenie.

	Dependent variable:											
	total_deaths											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
total_cases	0.017***	0.014***	0.014***	0.015***	0.015***	0.015***	0.015***	0.013***	0.013***	0.013***	0.013***	0.013***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
total_vaccinations	-0.001***	-0.0004*	-0.0004	-0.0004		-0.0005*			-0.0005**			-0.0005**
	(0.0002)	(0.0002)	(0.0002)	(0.001)		(0.0002)			(0.0002)			(0.0002)
positive_rate	-1,655.103**											
	(688.950)											
population_density			-0.632									
			(0.772)									
extreme_poverty		180.307***	175.271***									
		(56.549)	(57.116)									
I(total_vaccinations * total_vaccinations)				-0.000	-0.000***		-0.000*	-0.000**		-0.000**	-0.000**	
				(0.000)	(0.000)		(0.000)	(0.000)		(0.000)	(0.000)	
gdp_per_capita						-0.008*	-0.008*	-0.010**	-0.010**	-0.010**	-0.010**	-0.010**
						(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
aged_65_older								29.268**	30.793**			
								(13.081)	(13.023)			
aged_70_older										38.456**	38.456**	40.521**
										(18.513)	(18.513)	(18.438)
Constant	1,335.348***	567.543	636.969*	842.144*	644.379***	931.254***	723.428***	545.533**	758.989***	595.073***	595.073***	809.219***
	(254.208)	(340.116)	(351.894)	(472.551)	(174.576)	(228.324)	(175.936)	(206.896)	(247.362)	(199.737)	(199.737)	(242.406)
Observations	56	44	44	64	64	64	64	62	62	62	62	62
R^2	0.554	0.575	0.582	0.524	0.522	0.549	0.549	0.597	0.599	0.592	0.592	0.594
Adjusted R^2	0.528	0.543	0.539	0.500	0.506	0.526	0.527	0.568	0.571	0.563	0.563	0.566
Residual Std. Error	579.877 (df = 52)	592.642 (df = 40)	595.099 (df = 39)	606.617 (df = 60)	602.642 (df = 61)	590.494 (df = 60)	590.131 (df = 60)	559.359 (df = 57)	557.601 (df = 57)	562.503 (df = 57)	562.503 (df = 57)	561.022 (df = 57)
F Statistic	21.507*** (df = 3; 52)	18.044*** (df = 3; 40)	13.589*** (df = 4; 39)	21.981*** (df = 3; 60)	33.305*** (df = 2; 61)	24.305*** (df = 3; 60)	24.360*** (df = 3; 60)	21.066*** (df = 4; 57)	21.289*** (df = 4; 57)	20.673*** (df = 4; 57)	20.673*** (df = 4; 57)	20.857*** (df = 4; 57)
Note:												*p<0.1; **p<0.05; ***p<0.01

Wybrany model: $total\_deaths_i = \beta_0 + \beta_1 total\_cases_i + \beta_2 total\_vaccinations_i + \beta_3 gdp\_per\_capita_i + \beta_4 aged\_65\_older_i + \varepsilon_i$

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        758.989027 247.361585   3.068  0.00329 **
total_cases          0.013378   0.001803   7.419 6.37e-10 ***
total_vaccinations  -0.000492   0.000221  -2.226  0.02998 *
gdp_per_capita      -0.009603   0.004162  -2.307  0.02470 *
aged_65_older       30.792522  13.022925   2.364  0.02148 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 557.6 on 57 degrees of freedom
  (2 obserwacje zostały skasowane z uwagi na braki w nich zawarte)
Multiple R-squared:  0.599,      Adjusted R-squared:  0.5709
F-statistic: 21.29 on 4 and 57 DF,  p-value: 8.822e-11
```

Statystyka F uogólnionego testu Walda służącego do weryfikacji hipotezy o łącznej istotności parametrów wynosi 21.29, co stanowi podstawę do odrzucenia hipotezy zerowej o łącznej nieistotności oszacowań parametrów modelu przy poziomie istotności α=0.05.

W wybranym modelu wszystkie oszacowania parametrów można uznać za istotne statystycznie przy poziomie istotności α=0.05. Przed interpretacją wyników warto zaznaczyć, że liczba zgonów na milion określa liczbę zgonów licząc od początku pandemii, a nie wyłącznie w ostatnim okresie, z którego pobrano uśrednione wartości. Dlatego siła wpływu szczepień niekoniecznie znajdzie adekwatne odzwierciedlenie w wynikach. Ponadto stopień zaszczepienia może zawierać informacje dotyczące innych zmiennych w tym nieobserwowalnych, które zaburzają oszacowania (na przykład sugeruje się, że większy odsetek zaszczepiania może być powiązany z tym jak pandemia doświadczyła społeczeństwo przed wprowadzeniem programów szczepień).

Parametry otrzymane w wyniku estymacji możemy interpretować następująco:

- **stała $\boldsymbol{\beta_0}$** można stwierdzić średnio 758.989027 zgonów spowodowanych chorobą COVID-19 na milion mieszkańców od początku pandemii do okresu, którego dotyczą dane przy zerowym poziomie pozostałych zmiennych.
- **parametr $\boldsymbol{\beta_1}$** przy wzroście łącznej liczby potwierdzonych przypadków COVID-19 na milion mieszkańców o tysiąc jednostek, całkowita liczba zgonów przypisanych do COVID-19 na milion mieszkańców wzrasta średnio o 13.378 ceteris paribus.
- **parametr $\boldsymbol{\beta_2}$** przy wzroście łącznej liczby podanych dawek szczepionki COVID-19 na milion mieszkańców o tysiąc jednostek, całkowita liczba zgonów przypisanych do COVID-19 na milion mieszkańców spada średnio o 0.492 ceteris paribus.
- **parametr $\boldsymbol{\beta_3}$** przy wzroście produktu krajowego brutto per capita o tysiąc dolarów, całkowita liczba zgonów przypisanych do COVID-19 na milion mieszkańców spada średnio o 9.06 ceteris paribus.
- **parametr $\boldsymbol{\beta_4}$** przy wzroście odsetka ludności w wieku 65 lat i więcej o punkt procentowy, całkowita liczba zgonów przypisanych do COVID-19 na milion mieszkańców wzrasta średnio o 0.30792522 ceteris paribus.

Wyniki oszacowań są niemal zgodne z oczekiwaniami. Oczekiwano większego wpływu zmiennej *total_vaccinations,* jednak odwołanie się do wcześniej zaznaczonego charakteru zmiennych (statystyki zbierane od początku pandemii, a szczepienia w późniejszych fazach) pozwala łatwo uzasadnić jej mniejsze znaczenie. Sygnalizuje się istotną kwestię do rozwagi przy przyszłych próbach modelowania zjawiska jaką jest odpowiednie – uzależnione od przedmiotu analizy przetworzenie i przygotowanie danych.

Wartość miary dopasowania modelu R^2 pozwala stwierdzić, że zmienność zmiennej objaśnianej jest wyjaśniana przez zmienność zmiennych objaśniających w modelu w 59.9%.

Normalność rozkładu reszt. Test Jarque-Bara.

Normalność rozkładu reszt jest pomijalnym w przypadku stosunkowo dużej próby założeniem. Jednak warto przeprowadzić test, aby wiedzieć, czy korzystanie z testów jest uprawnione w mniejszych czy należy liczyć na asymptotyczne własności testów w dużych próbach.

```
        Jarque Bera Test

        data:  mod$residuals

        X-squared = 0.56295, df = 2, p-value = 0.7547
```

Statystyka testowa ma rozkład χ^2 z dwoma stopniami swobody, nie daje podstaw do odrzucenia hipotezy zerowej o normalności rozkładu składnika losowego modelu przy poziomie istotności $\alpha = 0.05$. Umożliwia to korzystanie z testów statystycznych weryfikujących pozostałe własności składnika losowego bez konieczności korzystania z asymptotycznych własności testów, czyli bez względu na rozmiar próby.

Poprawność postaci funkcyjnej modelu. Test RESET Ramseya.

Test RESET polega na dodaniu do równania kwadratów i iloczynów wszystkich par zmiennych oprócz zmiennych zero-jedynkowych. Zdarza się, że dołącza się również sześciany regresorów. Przeprowadzenie testu RESET Ramseya zasadza się na oszacowaniu regresji pomocniczej z uwzględnieniem kwadratów i interakcji zmiennych. Hipoteza zerowa testu odnosi się do nieistotności statystycznej oszacowań parametrów przy dodatkowych zmiennych. W takim razie przeprowadza się test Walda o statystyce z rozkładu F-Snedecora. Jej wartość nie daje podstaw do odrzucenia hipotezy zerowej o poprawnej postaci funkcyjnej modelu przy poziomie istotności $\alpha=0.05$. Pozwala na uznanie poprawności specyfikacji modelu. Sugeruje to również, że

nie występuje problem pominiętej zmiennej, nie ma także potrzeby dodawania kwadratów zmiennych ani zmiennych interakcyjnych.

```
RESET test

data:  mod

RESET = 0.57671, df1 = 2, df2 = 55, p-value = 0.5651
```

Współliniowość

Współliniowość występuje, gdy niespełnione jest jedno z założeń Klasycznej Metody Najmniejszych Kwadratów, a konkretnie:

$$rank(X) = (K + 1) < N$$

Jeśli jedna z kolumn X jest liniową kombinacją innych kolumn występuje dokładna współliniowość. W praktyce jednak problem ten jest rzadko spotykany i wynika zazwyczaj z niepoprawnego przygotowania danych. Zdecydowanie częściej bada się przybliżoną współliniowość, czyli sytuację, w której zmienne objaśniające związane są silną zależnością (są ze sobą silnie skorelowane). W takiej sytuacji występują następujące problemy:

- niewielkie zmiany w danych powodują duże zmiany w estymatorach otrzymanych w modelu (pogorszenie efektywności estymatora, niestabilność oszacowań)
- oszacowania parametrów mają duże błędy standardowe, a także mogą być statystycznie nieistotne nawet jeżeli są łącznie istotne, a współczynnik R^2 jest wysoki
- w konsekwencji testy istotności zmiennych nie są wiarygodne

W celu wykrycia przybliżonej współliniowości oblicza się czynnik inflacji wariancji VIF dla każdej zmiennej objaśniającej x_k:

$$\mathrm{VIF}_k = \frac{1}{1 - R_k^2}$$

Jeżeli obliczone wartości są większe od 10 to uznajemy, że występuje poważny problem przybliżonej współliniowości zmiennych.

total_cases	total_vaccinations	gdp_per_capita	aged_65_older
1.106164	1.548380	1.553555	1.107760

Jak wskazuje powyższa tabela analiza VIF nie daje podstaw do podejrzeń o przybliżoną współliniowość zmiennych objaśniających z uwagi na wartości znacznie mniejsze od umownej wartości granicznej 10.

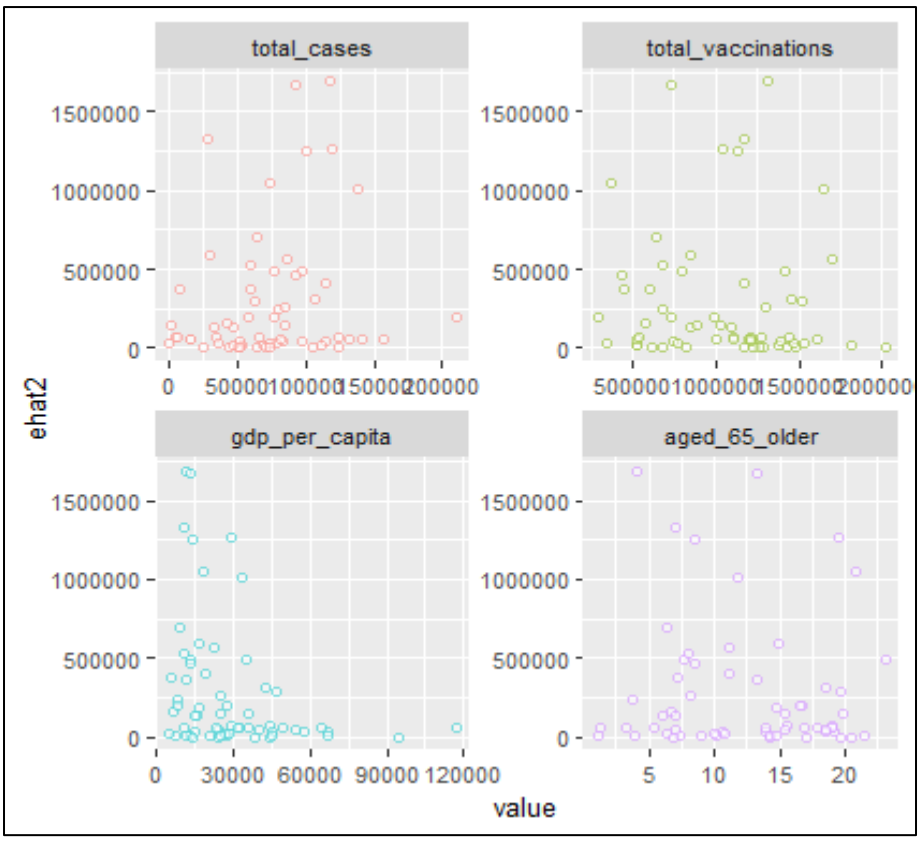
W przypadku wykrycia współliniowości można usunąć zmienne o największej wartości statystyki VIF, które są przyczyną problemu, może to jednak doprowadzić do problemu pominiętych zmiennych, w przypadku gdy usuwana zmienna jest istotna.

Inne rozwiązanie stanowi transformacja zmiennych lub zmiana metody estymacji. Przykładem jest użycie regresji grzbietowej, której estymator jest obciążony, lecz bardziej efektywny.

Weryfikacja założenia o sferyczności macierzy wariancji-kowariancji składnika losowego.

Założenie o sferyczności macierzy wariancji-kowariancji składnika losowego jest szczególnie istotne w kontekście wyprowadzenia estymatora macierzy wariancji-kowariancji estymatora parametrów klasycznej metody najmniejszych kwadratów.

W przypadku modeli opartych o dane przekrojowe standardową przyczyną niesferyczności jest heteroskedastyczność składnika losowego, czyli zróżnicowanie jego zmienności na przestrzeni próby. Do wstępnego zlokalizowania występowania zjawiska heteroskedastyczności reszt, które odzwierciedla się w postaci macierzy wariancji kowariancji poprzez odmienne wartości na diagonali, wykorzystano wykresy kwadratów reszt modelu od zmiennych objaśniających.

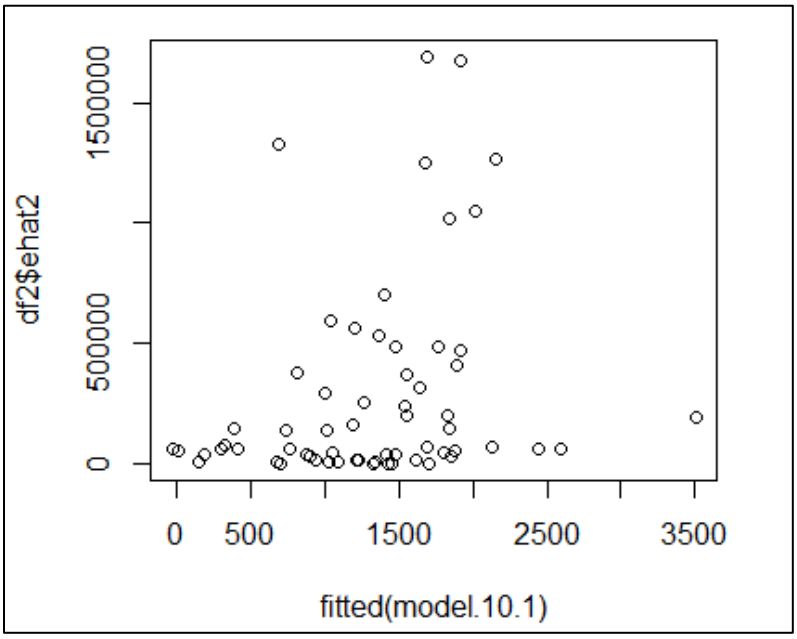


Wykres 7. Wykresy kwadratów reszt modelu od zmiennych objaśniających.

Źródło: Opracowanie własne

- Wykres kwadratów reszt z modelu względem zmiennej objaśniającej *total_cases* sugeruje występowanie heteroskedastyczności.
- Wykresy kwadratów reszt względem zmiennych objaśniających *total_vaccinations* oraz *aged_65_older* nie dają jednoznacznego sygnału dotyczącego występowania heteroskedastyczności.

- Wykres kwadratów reszt z modelu względem zmiennej objaśniającej *gdp_per_capita* sugeruje, że występuje w nim heteroskedastyczność.



Wykres 8. Wykres kwadratów reszt z modelu względem jego dopasowanych wartości.

Źródło: Opracowanie własne

Wykres kwadratów reszt z modelu względem dopasowanych wartości z modelu sugeruje występowanie heteroskedastyczności składnika losowego. Daje się spostrzec zróżnicowanie zmienności na przestrzeni wartości oszacowanych przez model.

Dalszą analizę zjawiska oraz rozwiązania przedstawia się w kolejnej części.

Test Goldfelda-Quandta względem zmiennej objaśniającej total_cases

Przeprowadzając test dzieli się próbę na dwie części według wartości zmiennej, którą podejrzewa się o związek z występowaniem heteroskedastyczności. Test pozwala sprawdzić, czy wariancja w obu grupach jest równa czy istotnie statystycznie różna. Opiera się on o iloraz kwadratów reszt z regresji oszacowanych dla obu grup z uwzględnieniem rozmiaru próby, gdzie licznik stanowi SEE/(N_1-K) z N_1 – liczebność grupy o większej wariancji, a K - liczba parametrów.

Przeprowadzony test Goldfelda-Quandta wskazuje, że hipoteza zerowa o homoskedastyczności składnika losowego powinna zostać odrzucona na rzecz hipotezy alternatywnej o heteroskedastyczności składnika losowego przy poziomie istotności $\alpha = 0.05$. Można swobodnie korzystać z tego testu, ponieważ składnik losowy ma rozkład normalny (czyli $SSE \sim \chi^2$). Dzięki temu wiadomo, że test Goldfelda-Quandta ma rozkład F i z dużym prawdopodobieństwem nie daje mylnych wyników dla małych prób.

Wybraną do testu zmienną jest *total_cases*, ponieważ osadzony na niej wykres kwadratów reszt sugeruje jej wpływ na zmienność reszt.

```
Goldfeld-Quandt's test
F = 2.349618, df1 = df2 = 25, p-value = 0.01858861
```

Przy poziomie istotności $\alpha = 0.05$ istnieją podstawy do odrzucenia hipotezy zerowej o homoskedastyczności składnika losowego co skutkuje naruszeniem założenia o sferyczności macierzy wariancji-kowariancji składnika losowego

Test Breuscha-Pagana

Test Goldfelda-Quandta pozwala wykryć zależność wariancji składnika losowego z jedną ze zmiennych, natomiast hetereskedastyczność może mieć bardziej złożoną naturę – może występować z uwagi na wiele zmiennych.

Test Breuscha-Pagana zakłada, że wariancja składnika resztowego jest liniową funkcją zmiennych objaśniających modelu. Statystyka testu Breuscha-Pagana oparta jest na rozkładzie χ^2, w wybranym modelu o czterech stopniach swobody. Wartość statystyki testowej wskazuje podstawy do odrzucenia hipotezy zerowej o homoskedastyczności składnika losowego na rzecz hipotezy alternatywnej o heteroskedastyczności składnika losowego przy poziomie istotności $\alpha=0.05$.

```
studentized Breusch-Pagan test
data:  mod
BP = 10.602, df = 4, p-value = 0.03142
```

Test White'a

Test White'a to najogólniejszy z wykorzystanych testów na heteroskedastyczność, bo zawiera najmniej założeń co do jej charakteru. Stanowi on szczególną postać testu Breuscha-Pagana. Odróżnia go obecność w regresji pomocniczej kwadratów oraz interakcji (iloczynów) wszystkich zamiennych objaśniających. Statystyka testu White'a ma rozkład χ^2. W modelu pomocniczym występuje czternaście zmiennych objaśniających – stąd czternaście stopni swobody. Wartość tej statystyki implikuje odrzucenie hipotezy zerowej o homoskedastyczności składnika losowego na rzecz hipotezy alternatywnej o heteroskedastyczności składnika losowego przy poziomie istotności α=0.05.

```
White's test

LM = 24.88569, df = 14, p-value = 0.03571276
```

Testy na heteroskedastyczność – podsumowanie

Powyższe testy wskazują na problem heteroskedastyczności. Nie jest spełnione założenie o stałej wariancji składnika losowego dla wszystkich obserwacji. Estymator OLS nadal jest nieobciążony, zgodny, ale przestaje być najefektywniejszy w klasie nieobciążonych estymatorów (szczególnie w małych próbach), co oznacza względnie wysokie wartości oszacowań błędów standardowych estymowanych parametrów. Estymator macierzy wariancji-kowariancji jest obciążony i niezgodny, więc poprawność wnioskowania statystycznego może być podważona.

Aby uwzględnić problem na poziomie estymacji błędów standardowych, bez ingerencji w postać modelu, można zastosować odporne estymatory macierzy wariancji-kowariancji oszacowanych parametrów.

Odporne estymatory macierzy wariancji-kowariancji

Zastosowanie odpornych błędów standardowych, do estymacji macierzy wariancji-kowariancji estymatora OLS, oparte jest na założeniu pewnej postaci macierzy wariancji-kowariancji składnika losowego $\hat{\Sigma}$.

$$Var[\widehat{\beta}] = (X^T X)^{-1} X^T \hat{\Sigma} X \left(X^T X\right)^{-1}$$

W pracy stosuje się odporne estymatory HC_0 (White'a) o następującej postaci: $Estymator\ HC_0 = \omega_i = \hat{e}_i^2$. Elementy ω_i stanowią diagonalne wartości macierzy:

$$\hat{\Sigma} = \begin{bmatrix} e_1^2 & 0 & \dots & 0 \\ 0 & e_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & e_n^2 \end{bmatrix}$$

```
t test of coefficients:

                      Estimate  Std. Error t value  Pr(>|t|)
(Intercept)         7.5899e+02  2.1010e+02  3.6125 0.0006419 ***
total_cases         1.3378e-02  1.6952e-03  7.8917  1.04e-10 ***
total_vaccinations -4.9196e-04  2.1873e-04 -2.2491 0.0283848 *
gdp_per_capita     -9.6028e-03  3.1877e-03 -3.0125 0.0038599 **
aged_65_older       3.0793e+01  1.2516e+01  2.4603 0.0169383 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Coefficients:

                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        758.989027 247.361585   3.068  0.00329 **
total_cases          0.013378   0.001803   7.419 6.37e-10 ***
total_vaccinations  -0.000492   0.000221  -2.226  0.02998 *
gdp_per_capita      -0.009603   0.004162  -2.307  0.02470 *
aged_65_older       30.792522  13.022925   2.364  0.02148 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Z uwagi na ingerencje w strukturę macierzy wariancji-kowariancji składnika losowego jedynie przy wyznaczaniu macierzy wariancji-kowariancji estymatora OLS brak różnicy w wynikach oszacowań parametrów modelu, w którym wykorzystano odporne błędy standardowe.

Zmianie uległy natomiast standardowe błędy szacunku parametrów na korzyść wyższej istotności oszacowań parametrów modelu z odpornymi błędami standardowymi.

Przy przyjęciu innego estymatora wariancji-kowariancji składnika losowego otrzymuje się następujące rezultaty. Są one analogiczne w stosunku do odpornych estymatorów White'a z dokładnością do klasyfikacji istotności zawartej w dolnej części ramki.

$$Estymator\ HC_1 = \omega_i = \frac{N}{N-k}\hat{e}_i^2$$

```
t test of coefficients:

                      Estimate  Std. Error t value  Pr(>|t|)
(Intercept)          7.5899e+02  2.1912e+02  3.4638  0.001018 **
total_cases          1.3378e-02  1.7680e-03  7.5668 3.613e-10 ***
total_vaccinations -4.9196e-04  2.2812e-04 -2.1565  0.035274 *
gdp_per_capita     -9.6028e-03  3.3245e-03 -2.8885  0.005464 **
aged_65_older       3.0793e+01  1.3053e+01  2.3590  0.021776 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ważona MNK

Kolejnym z zastosowanych rozwiązań problemu niesferyczności macierzy wariancji-kowariancji składnika losowego są modele oparte o Uogólnioną Metodę Najmniejszych Kwadratów. Uchyla ona założenie o sferyczności na rzecz ogólniejszej postaci macierzy wariancji-kowariancji składnika losowego, która jest dodatnio określona i symetryczna (te własności umożliwiają jej rozkład na pomocniczą macierz określającą przekształcenie jakie należy przyłożyć do danych w celu doprowadzenie do sferycznej postaci macierzy wariancji-kowariancji składnika losowego).

Ważona OLS jest szczególną odmianą GLS. W przypadku heteroskedastyczności opiera się (podobnie jak odporne estymatory) na przyjęciu pewnej postaci $E[\epsilon\epsilon^T]$ pozwalającej wyrugować występowanie zróżnicowania wariancji składnika losowego w podgrupach dla nowej,

skorygowanej o założoną relację składnika losowego ze zmiennymi objaśniającymi, postaci modelu.

Standardowo, w pierwszej kolejności, przyjmuje się liniową zależność wariancji składnika losowego względem jednej ze zmiennych objaśniających. Modele oparte o taką zależność obejmują numery 1-2 w poniższej tabeli (zmienne *total_cases* i *gdp_per_capita*). Z kolei w modelu 3 wagi stanowią eksponenty oszacowanych wartości kwadratów reszt z pomocniczego modelu objaśniającego logarytm naturalny kwadratów reszt pierwotnej regresji (z uwagi na niemożność wykorzystania oszacowań dla kwadratów reszt jako wag – wystąpienie ujemnych wartości). W modelu pomocniczym także wykorzystano zmienne *total_cases* i *gdp_per_capita* objaśniające, których potencjalne znaczenie w kontekście objaśniania zmienności reszt uznano na podstawie przedstawionych wyżej wykresów oraz przeprowadzonego testu.

	Dependent variable:		
	total_deaths		
	(1)	(2)	(3)
total_cases	0.018***	0.013***	0.015***
	(0.001)	(0.002)	(0.001)
total_vaccinations	-0.0003***	-0.0003	-0.0005***
	(0.0001)	(0.0003)	(0.0002)
gdp_per_capita	-0.010***	-0.013	-0.009***
	(0.003)	(0.008)	(0.002)
aged_65_older	15.257	36.136**	18.004**
	(10.907)	(17.499)	(8.152)
Constant	482.353**	637.191**	780.978***
	(186.940)	(249.151)	(223.111)
Observations	62	62	62
R^2	0.842	0.496	0.786
Adjusted R^2	0.831	0.460	0.771
Residual Std. Error (df = 57)	2.229	4.549	1.758
F Statistic (df = 4; 57)	75.743***	14.003***	52.226***
Note:			*p<0.1; **p<0.05; ***p<0.01

Tabela 1. Modele estymowane metodą Ważonej Metody Najmniejszych Kwadratów.

Źródło: Opracowanie własne przy użyciu pakietu 'stargazer'

Jak daje się spostrzec, na podstawie wyników testu Breuscha-Pagana, zastosowane metody pozwoliły wyeliminować problem heteroskedastyczności.

```
studentized Breusch-Pagan test
data: w1m1
BP = 0.00024622, df = 4, p-value = 1


studentized Breusch-Pagan test
data: w3m3
BP = 0.0019492, df = 4, p-value = 1

studentized Breusch-Pagan test
data: w5m5
BP = 0.00017318, df = 4, p-value = 1
```

Transformacja zmiennych

Kontrola wpływu zmiennych objaśniających na zmienność reszt modelu może odbywać się również za pośrednictwem transformacji zmiennych.

Proponuje się zastosowanie przekształcenia logarytmicznego zmiennej *total_cases* będącej przyczyną heteroskedastyczności reszt (test Goldfelda-Quandta). Takie rozwiązanie pozwala skondensować wartości transformowanej zmiennej, która wpływała na reszty. Wtedy, jeśli przed transformacją reszty zależały od zmienności pierwotnej zmiennej to po transformacji ta zmienność

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -2.782e+03  7.285e+02  -3.819 0.000333 ***
log(total_cases)     3.784e+02  6.418e+01   5.896 2.13e-07 ***
total_vaccinations -2.040e-04  2.434e-04  -0.838 0.405437
gdp_per_capita     -1.203e-02  4.624e-03  -2.602 0.011791 *
aged_65_older       4.504e+01  1.402e+01   3.212 0.002166 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 616.2 on 57 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.5104,     Adjusted R-squared:  0.476
F-statistic: 14.86 on 4 and 57 DF,  p-value: 2.25e-08
```

przyjmuje inny, wyrażony przez zmianę logarytmu, w przybliżeniu procentowy charakter[6]. Dlatego zastosowanie transformacji logarytmicznej pozwala zmniejszyć wielkość wpływu zmiennej na reszty.

Zaproponowane przekształcenie pozwala wyeliminować heteroskedastyczność kosztem utracenia istotności jednej ze zmiennych (dla α=0.05).

```
        studentized Breusch-Pagan test

        data:  mod

        BP = 6.0699, df = 4, p-value = 0.194
```

Endogeniczność. Metoda Zmiennych Instrumentalnych.

Występowanie endogeniczności jest równoznaczne z naruszeniem jednego z założeń MNK – $E(\varepsilon|X) \neq 0$. Oznacza ono, że występuje zależność między zmienną objaśniającą a składnikiem losowym. Skutkuje ona obciążeniem (systematycznym przeszacowaniem lub niedoszacowaniem) i niezgodnością (braku zbieżności stochastycznej) estymatora MNK.

Rozpatrując problem endogeniczności można uzasadnić występowanie zjawiska współzależności między liczbą zgonów a poziomem zaszczepienia. Jako prawdopodobny można rozważyć scenariusz, w którym społeczeństwa szczególnie dotknięte pandemią (stosunkowo duża liczba zgonów) przyjmują stosunkowo większą liczbę dawek szczepionki jako rozsądna reakcja obronna na łatwo dostrzegalne zagrożenie. Nie jest to jednak podstawne przypuszczenie z uwagi na ujemną korelację tych zmiennych. Być może taka prawidłowość dałaby się lepiej uzasadnić przy rozważeniu postaci panelowej danych pozwalającej odnaleźć dynamiczną zależność zmiennych.

Inną przyczyną występowania endogeniczności może być błąd pomiaru w tym przypadku odnoszący się do zmiennej *total_cases*. Można przypuszczać, że nie w każdym państwie wykrywalność występowania wirusa była na porównywalnym poziomie z uwagi na odmienne procedury związane z testowaniem, raportowaniem czy przede wszystkim z racji znacznych różnic w zrelatywizowanej liczbie wykonywanych testów. Nie rozpoznaje się jednak wśród dostępnych

[6] $\log(x)+\Delta=\log(xe^{\Delta})$, e^{Δ} oznacza proporcjonalną zmianę x. Oczywiście zmiana procentowa wyraża się przez $(e^{\Delta}-1)*100$. Zatem zmiana x o $(e^{\Delta}-1)*100\%$ przekłada się na zmianę zmiennej zależnej o $\Delta*\beta$

zmiennych potencjalnej zmiennej instrumentalnej. Zatem słuszność podejrzenia nie zostanie poddana weryfikacji.

Z drugiej strony zdaje się, że występowanie innej przyczyny endogeniczności – błąd pominiętej zmiennej również daje się uzasadnić i tym razem zweryfikować. Jeśli przyjrzeć się kwestii odpowiedzialności i dyscypliny społecznej, poziomowi antynaukowości, któremu wyraz daje nasilenie ruchów antyszczepionkowych i tzw. koronasceptyzmowi rzutującemu na nieodpowiedzialną społecznie postawę wobec zagrożenia pandemicznego to w modelu nie znajdzie się zmienna dająca im odzwierciedlenie. W dosyć ogólnym sensie zmienna *gdp_per_capita* jest powiązana z wymienionymi potencjalnie znaczącymi zmiennymi. Mimo braku silnych podstaw do wnioskowania o endogeniczności (brak sygnału dotyczącego endogeniczności w postaci znacznego skorelowania zmiennych ze składnikiem losowym) i odwołując się do wyżej podanego uzasadnienia stosuje się procedury szacowania modelu metodą zmiennych instrumentalnych.

Na potrzebę MZI zastosujmy jako instrumenty *gdp_per_capita* zmienną *human_development_index* oraz *stringency_index* które mogą stanowić pewną niedoskonałą miarę zarówno zamożności społeczeństwa ujętej przez *gdp_per_capita* jak i pominiętej zmiennej odnoszącej się do wymienionych w poprzednim akapicie, a nie ujętych modelem, zjawisk (uśrednione *stringency_index* ma oddawać to w jakim stopniu społeczeństwo demokratyczne jest w stanie poświęcić wygodę dla zniwelowania zaraźliwości, z kolei *human_development_index* ma ujmować poziom rozwoju społeczeństwa; w tym jego wyedukowanie).

Dla formalności prezentuje się wyniki wykonanych obliczeń. Począwszy od korelacji „endogenicznej" zmiennej objaśniającej ze składnikiem losowym i instrumentami oraz ilorazu współczynników korelacji ukazujących moc i egzogeniczność instrumentu; przechodząc przez test mocy instrumentu oraz test Hausmana pozwalający wykryć endogeniczność, a skończywszy na teście nadmiarowych restrykcji.

Korelacje zmiennych i składnika losowego. Moc i egzogeniczność z perspektywy korelacji.

Wartości kolejnych mierników kształtują się następująco:

- współczynnik korelacji zmiennej *gpd_per_capita* podejrzanej o endogeniczność z resztami: ρ(*gpd_per_capita*, *e*) = 8.964316e-17
- współczynnik korelacji zmiennej instrumentalnej *stringency_index* z resztami: ρ(*stringency_index*, *e*) = -0.02282265
- współczynnik korelacji zmiennej instrumentalnej *human_development_index* z resztami: ρ(*human_development_index*, *e*) = -0.07304828
- współczynnik korelacji zmiennej instrumentalnej *human_development_index* ze zmienną *gdp_per_capita* ρ(*human_development_index*, *gpd_per_capita*) = 0.7136542
- współczynnik korelacji zmiennej instrumentalnej *stringency_index* ze zmienną *gdp_per_capita* ρ(*stringency_index*, *gpd_per_capita*) = 0.01730316
- $\rho(human\_development\_index, e)\ /\rho(gpd\_per\_capita,\ human\_development\_index) = -0.1023581$
- $\rho(stringency\_index, e)\ /\rho(gpd\_per\_capita,\ stringency\_index) = -1.318987$

Opierając się na powyższych miarach nie można uznać, że instrumenty wykazują w większości oczekiwane własności. Zauważalna jest korelacja instrumentów z resztami modelu. Zasadniczo, korelacja składnika losowego i zmiennej gpd_per_capita nie wskazuje na endogeniczność; co więcej instrumenty wykazują większe skorelowanie z resztami niż sama zmienna „endogeniczna".

Podniesione kwestie zostaną poddane dalszej analizie w kolejnej części.

Weryfikacja mocy instrumentu na podstawie pierwszego kroku 2SLS. Badanie występowania endogeniczności zmiennej z wykorzystaniem testu Hausmana dla OLS i 2SLS.

Przejdźmy mimo wszystko do kolejnego kroku zachowując obie zmienne instrumentalne dla potrzeb poznawczych. Poniżej przeprowadzono weryfikację mocy instrumentów z wykorzystaniem pierwszego kroku dwustopniowej metody najmniejszych kwadratów.

Metoda ta polega na (1) oszacowaniu regresji pomocniczych – form zredukowanych, w których zmienne instrumentalne wraz ze zmiennymi egzogenicznymi objaśniają zmienne

endogeniczne, (2) wykorzystaniu oszacowanych wartości teoretycznych z form zredukowanych w regresji o postaci strukturalnej (pierwotnej).

Oto wyniki przeprowadzenia dwustopniowej metody najmniejszych kwadratów dla wybranych instrumentów.

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         6.574e+02  2.521e+02   2.608   0.0118 *
total_cases         1.383e-02  2.001e-03   6.912 5.73e-09 ***
total_vaccinations -3.257e-04  2.514e-04  -1.295   0.2008
aged_65_older       3.185e+01  1.296e+01   2.458   0.0172 *
gdp_per_capita     -1.417e-02  5.692e-03  -2.490   0.0159 *

Diagnostic tests:
                 df1 df2 statistic  p-value
Weak instruments   2  53    28.659 3.66e-09 ***
Wu-Hausman         1  53     1.621    0.209
Sargan             1  NA     0.010    0.921
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 548.3 on 54 degrees of freedom
Multiple R-Squared: 0.5651,      Adjusted R-squared: 0.5329
Wald test: 18.19 on 4 and 54 DF,  p-value: 1.634e-09
```

Przed przystąpieniem do głębszej analizy wyników należy sprawdzić, czy nie występują standardowe problemy OLS.

Jeżeli chodzi o współliniowość to analiza czynników inflacji wariancji nie pozwala rozpoznać występowania tego problemu (wszystkie VIF < 10).

total_cases	total_vaccinations	aged_65_older	gdp_per_capita
1.137575	1.969567	1.116006	2.006374

Dochodzi natomiast (tak jak w oryginalnym modelu) do naruszenia założenia OLS o sferyczności macierzy wariancji-kowariancji składnika losowego. Przeprowadzony test Breuscha-Pagana daje podstawy do odrzucenia hipotezy zerowej o homoskedasyczności składnika resztowego przy poziomie istotności α=0.05.

```
        studentized Breusch-Pagan test

        BP = 11.823, df = 4, p-value = 0.01872
```

Z tego względu powiela się rozwiązanie wykorzystane przy analizie modelu uzyskanego klasyczną metodą najmniejszych kwadratów – stosuje się WOLS z założeniem o zależności wariancji reszt ze zmienną *total_cases* (wagi = *1/total_cases*).

```
        Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
        (Intercept)          4.544e+02  1.872e+02   2.427 0.018590 *
        total_cases          1.750e-02  1.243e-03  14.083  < 2e-16 ***
        total_vaccinations -3.058e-04  9.216e-05  -3.318 0.001627 **
        aged_65_older        1.629e+01  1.118e+01   1.457 0.150906
        gdp_per_capita      -1.022e-02  2.933e-03  -3.485 0.000986 ***

        Diagnostic tests:
                        df1 df2 statistic p-value
        Weak instruments   2  53   111.021  <2e-16 ***
        Wu-Hausman         1  53     0.128   0.722
        Sargan             1  NA     0.796   0.372
        ---
        Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        Residual standard error: 2.196 on 54 degrees of freedom
        Multiple R-Squared: 0.8378,       Adjusted R-squared: 0.8258
        Wald test: 69.35 on 4 and 54 DF,  p-value: < 2.2e-16
```

Jak daje się zauważyć statystyka testu Walda pochodząca z rozkładu F przekracza umowną granicę 10 co daje podstawy do odrzucenia hipotezy zerowej o łącznej słabości instrumentów. W takim razie strata efektywności estymatora nie jest bardzo poważna.

Z kolei do weryfikacji hipotezy dotyczącej występowania w modelu zjawiska endogeniczności posłużono się testem Hausmana o statystyce z rozkładu χ^2. Test Hausmana pozwala weryfikować hipotezę zerową mówiącą o zgodności dwóch estymatorów o różnej efektywności, przy czym hipoteza alternatywna zakłada brak zgodności estymatora o wyższej efektywności. W tym przypadku estymatorem o większej wariancji jest estymator metody zmiennych instrumentalnych zatem, jako że endogeniczność powoduje niezgodność estymatora, test ten pozwala zweryfikować jej występowanie. Wyniki widoczne na wydruku z konsoli nie dają podstaw do odrzucenia hipotezy zerowej o egzogeniczności przy przyjętym poziomie istotności 0.05. Zatem wnioski pozostają analogiczne względem tych jakie można by wysnuć na podstawie wyników 2SLS bez uwzględniania zjawiska heteroskedastyczności składnika resztowego.

Test nadmiarowych restrykcji. Weryfikacja zasadności restrykcji.

Z uwagi na wykorzystanie dwóch instrumentów, co skutkuje nadmierną identyfikowalnością, przeprowadza się test na nadmiarowe restrykcje.

Test Sargana-Hansena o statystyce z rozkładu χ^2 wykorzystano do zweryfikowania hipotezy o słuszności nadmiarowych restrykcji. Tak jak test Hausman korzysta on z drugiego kroku podwójnej metody najmniejszych kwadratów. Następnie konstruuje się regresję objaśniającą reszty z oszacowanego modelu pierwszego kroku z wykorzystaniem zmiennych egzogenicznych.

Dokładniej rzecz ujmując hipoteza zerowa testu Sargana dotyczy łącznej egzogeniczności instrumentów. Dlatego rekomendowane jest przyjęcie wysokiego poziomu istotności np. $\alpha=0.25$ (w celu zminimalizowania prawdopodobieństwa przyjęcia modelu ze zmiennymi endogenicznymi). Dla takiego poziomu istotności można przyjąć słuszności hipotezy zerowej mówiącej o egzogeniczności nadmiarowych instrumentów.

Metoda zmiennych instrumentalnych podsumowanie.

Charakterystyka zmiennej podejrzanej o endogeniczność w kontekście modelu nie pozwala potwierdzić występowania zjawiska endogeniczności. Zarówno analiza korelacji jak i wyniki testu Hausmana odwodzą od stwierdzenia występowania endogeniczności w rozważanym w pracy modelu. Można zatem postulować zgodność estymatora OLS, jeśli uznać dotychczasowe testy za wystarczające. Nie rekomenduje się zatem zastosowania oszacowania metody zmiennych instrumentalnych, ponieważ jej estymator ma mniejszą efektywność, której skutkiem może być utrata istotności statystycznej zmiennych.

Podsumowanie

Zasadniczo udało się skonstruować prosty model ekonometryczny unikając i przeciwdziałając poważnym problemom uniemożliwiającym poprawne wnioskowanie statystyczne. Wszystkie brane pod uwagę zmienne są istotne, a kierunek ich wpływu daje się rozsądnie uzasadnić. Zastosowane techniki wystarczyły, aby wyeliminować problem niesferycznosci macierzy wariancji-kowariancji w postaci heteroskedastyczności. Ponadto istnieją podstawy do uznania, że nie występuje problem pominiętej zmiennej. Co więcej analiza wskazuje na brak endogeniczności zmiennych objaśniajacych. Kwestią otwartą pozostają zdolności prognostyczne pozwalające precyzyjniej określić jakość modelu. Niżej znajduje się skrótowy opis przebiegu analizy.

Przebieg analizy

Pracę na postacią modelu standardowo poprzedziła analiza charakterystyk zbioru danych. Wykorzystano podstawowe techniki wizualizacyjne w celu rozpoznania fundamentalnych własności rozkładów zmiennych. Proces ten pozwolił wykryć nieprawidłowości dotyczące jakości dostępnych dla pewnych państw danych oraz zróżnicowania ze względu na poziom rozwoju państw. Zdobyto również przydatne informacje dotyczące zależności zmiennych.

Kolejnym krokiem, którego nieodłącznym elementem była wstępna estymacja, był wybór postaci modelu. Wyniki estymacji oraz teoretyczne podwaliny pozwoliły wybrać prostą – wolną

od transformacji postać modelu. Wybrane zmienne były zarówno łącznie jak i indywidualnie istotne. Analiz czynników inflacji wariancji nie dała podstaw do podejrzewania współliniowości. Modelowi nie można również zarzucić niepoprawnej postaci funkcyjnej jak wskazały wyniki testu RESET Ramseya. Na dodatek model okazał się niemal w 60% objaśniać zmienność zmiennej zależnej.

Co się tyczy własności reszt modelu, to można uznać je za istotne na podstawie wyników testu Jarque Bera. Z kolei macierz wariancji-kowariancji reszt pierwotnie nie miała sferycznej postaci. Jednak zastosowane metody: dwa wybrane odporne estymatory błędów, WLS dla rożnych założeń o zależności zmienności reszt ze zmiennymi objaśniającymi oraz transformacja logarytmiczna zmiennej; pozwoliły skorygować model o wykryte występowanie heterogeniczności.

Analizę ekonometryczną zakończono na rozważeniu problemu endogeniczności w kontekście opisanego modelu. Rozpatrzono trzy możliwe przyczyny występowania endogeniczności, z których jedna pozwoliła na dalszą analizę. Wykorzystano metodę zmiennych instrumentalnych z nadmierną identyfikacją parametru zmiennej endogenicznej. Przeprowadzone testy doprowadziły do wniosków o statystycznie istotnej mocy instrumentów i niewystępowaniu endogeniczności. Następnie test Sargana-Hansena wskazał zasadność użycia dwóch instrumentów.

Wnioski i rekomendacje

Wyniki pracy obfitują w liczne uwagi i spostrzeżenia dotyczące poprawy i rozszerzenia analizy zjawiska. Dla zachowania spójności przedstawione rekomendacje będą dotyczyły przyjętego w pracy typu analizy. Proponowane rozwiązania potencjalnie poprawiające rzetelność analizy oprą się głównie na podejściu do przygotowania danych a mniej na procesie modelowania, który silnie zależy od tego pierwszego.

Jako pierwszą rekomendację związaną z selekcją i obróbką danych wyróżnia się uwzględnienie dynamicznego charakteru pandemii. Przedstawiony w pracy model pozwala objaśniać zmienność liczby zgonów od początku pandemii do wybranego końcowego okresu w oparciu o uśrednione charakterystyki z dziewięćdziesięciodniowego okna począwszy od

01.10.2021. Takie podejście pozwoliło ująć w pewien sposób dynamiczny charakter zjawiska[7]. Niemniej jest ono dalekie od doskonałego, dlatego proponuje się przeprowadzenie analizy w oparciu o dane dotyczące liczby śmierci w konkretnych okresach nie zaś o całkowitą liczbę zgonów. W szczególności, pozostawszy przy modelu przekrojowym, zaleca się uwzględnienie średniego czasu przebiegu choroby oraz opóźniania wpływy obostrzeń czy szczepień na liczbę zgonów. Takie podejście pozwoliłoby na zwiększenie liczebności próby poprzez kilkukrotne uwzględnienie obserwacji dla każdego państwa przykładowo dla wszystkich fal pandemii. To zwiększenie próby daje możliwość dalszej fragmentacji zestawu danych bez obaw na ograniczenie zbyt małej liczby obserwacji. Jednocześnie takie podejście wymagałoby znacznej uważności wykluczającej zanadto selektywny dobór próby.

Analiza przeprowadzona w pracy dotyczyła rozwiniętych państw, które wprowadziły programy szczepień. Sugeruje się przeprowadzenie analizy dla ogólniejszej grupy państw lub dla różnych grup z osobna np. ze względu na liczbę przypadków lub strukturę demograficzną. Takie rozwiązania pozwoliłyby na zestawienie wyników i wysnucie konkretniejszych wniosków na temat zależności między zmiennymi. Rozszerzenie analizy może w niektórych przepadkach okazać się jednak kłopotliwe z uwagi na słabą jakość danych raportowanych przez niektóre nierozwinięte lub niewielkie państwa.

Rozsądnym kierunkiem rozwoju analizy zdaje się również rozbudowa zbioru danych o dodatkowe zmienne oraz funkcjonalna ich transformacja. Można skorzystać z dodatkowych źródeł w celu uzupełninia modelu o zmienne odnoszące się do stany stan służby zdrowia w sposób bezpośredni lub ważnej kwestii mobilności społeczeństwa. Można także przyjrzeć się kwestii edukacji społeczeństw, związanej z reakcją na stosowane metody redukujące rozprzestrzenianie się wirusa. Co więcej można stworzyć wskaźniki lub zmienne binarne określające dynamikę reakcji rządów np. w oparciu o zmienną *stringency_index* (ujęcie konsekwencji stosowanych środków).

[7] Zmienność obostrzeń oraz liczby przypadków ujętych przez *stringency_index* i *total_cases*.

Bibliografia

- Chetty, Raj. *How Did COVID-19 and Stabilization Policies Affect Spending and Employment?*: A New Real-time Economic Tracker Based On Private Sector Data. Cambridge, MA : National Bureau of Economic Research, 2020.

- Ritchie H., Mathieu E., Rodés-Guirao L., Appel C., Giattino C., Ortiz-Ospina E., Hasell J., Macdonald B., Beltekian D., Roser M., *Coronavirus Pandemic (COVID-19)*, 2020, Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/coronavirus', pobrano: 23.04.2022.

- World Bank. World Bank national accounts data, and OECD National Accounts data files., The World Bank Group, 2022, https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG

- *World Economic Outlook.*, 2020, Washington, D.C: International Monetary Fund, s.67