POLITECHNIKA LUBELSKA
WYDZIAŁ PODSTAW TECHNIKI
Kierunek: MATEMATYKA

Praca inżynierska

Zastosowanie modeli mieszanych w analizie rozwoju pandemii choroby COVID-19 na świecie

The use of mixed-effects models in the analysis of the COVID-19 pandemic in the world

Praca wykonana pod kierunkiem:
dra Dariusza Majerka

Autor:
Alicja Hołowiecka
nr albumu: 89892

Lublin 2020

Spis treści

Wstęp

Pandemia choroby COVID-19 jest wydarzeniem, które wstrząsnęło całym światem w roku 2019. Właściwie nikt chyba nie może powiedzieć, że nie poczuł się dotknięty przez sytuację związaną z rozprzestrzenianiem się wirusa. Pierwsze przypadki pojawiły się pod koniec 2019 roku we wschodnich Chinach, w mieście Wuhan. Na początku 2020 roku chorowali już obywatele większości państw na świecie. Na moment pisania tej pracy, sytuacja nadal nie jest opanowana i nie wiadomo, jak się rozwinie.

Biorąc to pod uwagę, tym ważniejszy wydaje się temat poruszany w tej pracy. Wiele jednostek naukowych podejmuje próby znalezienia odpowiedniego modelu, aby przewidzieć rozwój pandemii. Przedstawione w tej pracy modele mieszane co prawda nie pozwalają na dokładną predykcję, ale są dobrym narzędziem, aby odkryć, które czynniki mają wpływ na rozwój pandemii w przeciętnym kraju.

Rozdział 1

Teoretyczne podstawy badań własnych

W tej części pracy przedstawimy metody matematyczne, które zostaną użyte w części praktycznej tej pracy. Zgodnie z tematem, będą to głównie modele mieszane.

1.1. Modele liniowe

Na początek przypomnimy podstawowe wiadomości o modelach liniowych.

Model regresji prostej ma postać

$$y = x\beta_1 + \beta_0 + \varepsilon$$

gdzie oszacowania parametrów β_1, β_0 obliczamy następująco:

$$\hat{\beta}_1 = \frac{Cov(x, y)}{Var(x)},$$
$$\hat{\beta}_0 = \overline{y} - \overline{x}\hat{\beta}_1.$$

Model interpretujemy w ten sposób, że jeżeli zmienna x wzrośnie o 1, to zmienna y zmieni się o β_1.

Zmienną y nazywamy zmienną zależną, a x - niezależną.

Jeżeli w modelu występuje więcej niż jedna zmienna niezależna, to mówimy o regresji wielorakiej (lub wielokrotnej). Wówczas model ma postać:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon,$$

lub w zapisie macierzowym

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

1.1.1. Metody estymacji parametrów modelu liniowego

1. Metoda najmniejszych kwadratów, OLS (ang. *Ordinary Least Squares*) - w metodzie tej minimalizujemy błąd kwadratowy, czyli sumę kwadratów reszt, którą oznaczamy RSS (ang. *Residual Sum of Squares*).

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Twierdzenie Gaussa-Markowa: taki estymator jest BLUE (Best Linear Unbiased Estimator), przy odpowiednich założeniach.

2. Metoda największej wiarogodności, ML (ang.`Maximum Likelihood`) polega na maksymalizacji wartości funkcji prawdopodobieństwa ze względu na β (w praktyce maksymalizujemy zwykle logarytm z tej funkcji)

$$\hat{\sigma}^2_{ML} = RSS/n$$

Estymując σ^2, maksymalizujemy funkcję wiarogodności zarówno ze względu na β, jak i σ^2.

Estymatory uzyskane tą metodą są asymptotycznie nieobciążone.

3. Resztowa metoda największej wiarogodności, REML (ang. `Residual/Restricted Maximum Likelihood Method`) - z estymacji parametru σ^2 usuwamy wpływ parametrów zakłócających β.

$$\hat{\sigma}^2_{REML} = RSS/(n-p)$$

Estymatory uzyskane tą metodą są nieobciążone [1].

1.1.2. Badanie istotności parametrów

Aby zbadać istotność współczynników modelu liniowego, weryfikujemy hipotezę postaci

$$H_0 : \beta_i = 0$$

przeciw hipotezie alternatywnej

$$H_1 : \beta_i \neq 0.$$

Do zweryfikowania tej hipotezy wykorzystujemy test Walda. Statystyka testowa ma postać

$$T = \hat{\beta}_i / se(\hat{\beta}_i)$$

i jest nazywana statystyką t (ang. `t-value`).

Przy założeniu prawdziwości H_0 statystyka ta ma rozkład t-Studenta o $n-k-1$ stopniach swobody (n - liczba obserwacji, k - liczba parametrów w modelu).

Można także użyć testu F (testu globalnego), aby zweryfikować hipotezę

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

przeciwko hipotezie alternatywnej

$$H_1 : \exists j \beta_j \neq 0$$

1.2. Modele mieszane

W powyżej opisanych modelach liniowych z efektami stałymi zakładamy niezależność kolejnych pomiarów, dlatego nie są to odpowiednie modele, kiedy mamy np. kilka pomiarów dla pojedynczego elementu. W takim przypadku możemy użyć modeli liniowych z efektami mieszanymi (stałymi i losowymi), które krótko nazywamy modelami mieszanymi.

Modeli mieszanych używamy w przypadu powtarzanych pomiarów bądź w przypadku hierarchicznej lub zagnieżdżonej struktury. Takie dane charakteryzują się korelacją między obserwacjami z tej samej grupy, co nie pozwala na użycie modelu liniowego z efektami stałymi. Dlatego do modelu wprowadza się czynnik losowy.

Czynnik stały jest pewnym parametrem, którego wartość estymujemy na podstawie próbki, natomiast czynnik losowy jest zmienną losową, dla której próbujemy oszacować parametry jej rozkładu [2].

Przykładową sytuacją, gdzie możemy użyć modelu mieszanego, jest badanie działania leku na grupie pacjentów, gdzie dokonujemy kilku pomiarów na danym pacjencie. W tym przypadku nie interesuje nas konkretny pacjent, ale raczej wpływ leku na przeciętnego pacjenta. Dodatkowo, traktujemy pacjentów jako losowo wybranych. Podejście modelu mieszanego będzie polegało na potraktowaniu wpływu pacjenta jako czynnik zakłócający.

Rozważamy model postaci

$$y = X\beta + Zu + \varepsilon$$

gdzie X - macierz zmiennych będących efektami stałymi, Z - macierz zmiennych będących efektami losowymi, β to wektor nieznanych efektów stałych, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_{n\times n})$ to zakłócenie losowe, a $u \sim \mathcal{N}(0, \sigma^2 D)$ to wektor zmiennych losowych odpowiadających efektom losowym [1].

Znając D, możemy estymować parametry β uogólnioną metodą najmniejszych kwadratóW. Do estymowania nieznanego D możemy użyć np. metodą największej wiarogodności.

1.2.1. Metody estymacji

Do oceny wartości parametrów modelu mieszanego można stosować metody ML (Największej Wiarogodności) oraz REML (Resztowej Największej Wiarogodności), wspomniane w tej pracy przy okazji modeli liniowych. W przypadku modeli mieszanych obydwoma metodami możemy uzyskać estymatory obciążone, ale to obciążenie jest zazwyczaj mniejsze w przypadku estymatorów uzyskanych metodą REML.

Różnica między metodą REML i ML polega na tym, że w metodzie REML najpierw usuwamy wpływ efektów stałych, które w modelach mieszanych są czynnikami zakłócającymi.

1.2.2. Badanie istotności parametrów i wybór najlepszego modelu

W modelach mieszanych konieczne jest zbadanie istotności dla efektów stałych oraz dla efektów losowych.

Dla efektów stałych testujemy hipotezę

$$H_0 : \beta_i = 0$$

przeciwko hipotezie alternatywnej

$$H_1 : \beta_i \neq 0,$$

a dla komponentów wariancyjnych weryfikujemy hipotezę

$$H_0 : \sigma_j^2 = 0$$

przy jednostronnej hipotezie alternatywnej

$$H_1 : \sigma_j^2 > 0.$$

Metody, które mają zastosowanie dla modeli liniowych z efektami stałymi, nie zawsze dają się zastosować w przypadku modeli mieszanych. Wymienimy teraz kilka metod doboru najlepszego modelu i opiszemy, które z nich są najskuteczniejsze [2].

1. Iloraz wiarygodności(ang. `likelihood ratio`) - tworzymy dwa modele, model 0, który nie zawiera elementów, których istotność chcemy zbadać, i model 1, który zawiera te elementy. Pozostałe zmienne muszą być takie same w obu modelach. Statystyka testowa wygląda następująco:

   $$2(l(\hat{\beta}_1, \hat{\sigma_1}, \hat{D}_1|y) - l(\hat{\beta}_0, \hat{\sigma_0}, \hat{D}_0|y)),$$

   gdzie l - logarytm z funkcji prawdopodobieństwa.

   Tego testu nie można używać do modeli wyznaczonych metodą `REML`.

2. Test F dla efektów stałych - metoda taka sama jak ta używana przy modelach z efektami stałymi. W przypadku modeli mieszanych może sprawiać problemy, ponieważ statystyka testowa niekoniecznie musi mieć rozkład F. Należy także wprowadzać poprawkę na liczbę stopni swobody.
   Na ogół ta metoda daje dobre rezultaty dla mniej skomplikowanych modeli, gdy układ jest zbalansowany (wszystkie grupy są równoliczne). Dla modeli bardziej skomplikowanych, lub kiedy brak równoliczności, wartości p oraz statystyki t mogą być błędne.
3. Test permutacyjny - można użyć metod *bootstrapowych*, aby znaleźć dokładniejsze wartości p-value. Należy wygenerować dane z modelu 0 (na podstawie oszacowanych parametrów) i obliczyć statystykę *likelihood ratio*. Tą procedurę powtarzamy wielokrotnie i oceniamy istotność.
4. Kryteria informacyjne - służą do wyboru najlepszego spośród modeli. Najpopularniejszym jest Kryterium Informacyjne Akaikego (ang. `Akaike Information Criterion, AIC`). Jest ono zdefiniowane następującym wzorem:

$$-2(\texttt{max log likelihood}) + 2p,$$

   gdzie p to liczba parametrów modelu.
   Można stosować to kryterium do modeli, które różnią się jedynie efektami stałymi, gdzie liczba efektów losowych jest identyczna dla wszystkich modeli, które porównujemy. Gdyby modele różniły się liczbą efektów losowych, należałoby rozważyć, w jaki sposób zliczyć liczbę parametrów p.
   Kryterium Akaikego jest miarą utraconej informacji, więc po obliczeniu go dla roważanych modeli, należy wybrać ten, gdzie otrzymana wartość jest najmniejsza.

Przy obliczeniach dotyczących stosunkowo małych zbiorów danych, można użyć każdej z tych metod, ale w przypadku dużej liczby obserwacji, niektóre obliczenia mogą zająć zbyt wiele czasu. Najmniej skomplikowany obliczeniowo jest test Walda, gdzie dokonujemy tylko jednej estymacji współczynników. Przy użyciu testu ilorazu wiarygodności, należy dokonać dwóch estymacji - dla modelu z i bez testowanego efektu. Stosując testy permutacyjne, musimy dokonać obliczeń setki lub tysiące razy. Dlatego w przypadku najbardziej skomplikowanych problemów zwykle stosuje się test Walda, mimo jego gorszych właściwości statystycznych [1].

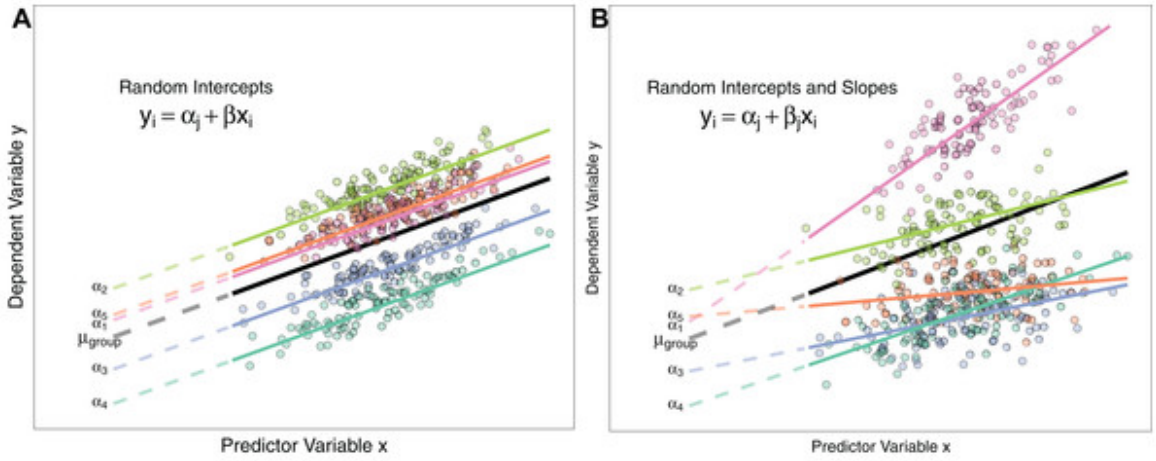
1.2.3. Interpretacja parametrów modelu mieszanego

W modelu mieszanym efekty stałe należy interpretować tak jak w przypadku regresji, analizy wariancji lub analizy kowariancji, w zależności od rodzaju zmiennej

niezależnej. Trzeba jednak pamiętać, że oszacowane wartości współczynników reprezentują wartość średnią dla całej populacji, a dla poszczególnych obiektów badania będą się różnić o wartość oceny efektu osobniczego.

Dla efektu losowego możemy oszacować jego wariancję. Informuje nas ona o tym, jak bardzo mogą się różnić współczynniki efektów stałych dla poszczególnych obiektów badania [7].

Można wyróżnić dwa rodzaje efektów losowych: losowy wyraz wolny (rysunek 1.1, wykres A) oraz losowy wyraz wolny i współczynnik nachylenia (rysunek 1.1, wykres B). W pierwszym przypadku, prosta regresji dla poszczególnych obiektów badań może być przesunięta w górę lub w dół w stosunku do średniej, a w drugim przypadku może także być nachylona do osi OX pod mniejszym lub większym kątem.



Rysunek 1.1: Rodzaje modeli mieszanych
Źródło: [8]

1.2.4. Predykcja z modelu mieszanego

Proces predykcji jest trudniejszy w przypadku modelu mieszanego niż dla zwykłego modelu liniowego. Musimy zdecydować, czy uwzględnić, czy wykluczyć efekt losowy z predykcji. Efekty losowe mogą mieć różny wkład w predykcję. Mogą być całkowicie pominięte, mogą być uśrednione lub mogą być na pewnym ustalonym poziomie. Uśrednienie efektów losowych powoduje predykcję zależną od wartości efektów losowych, które zostały zaobserwowane do tej pory. Pominięcie efektów losowych powoduje predykcję na poziomie średniej populacyjnej [6].

Aby lepiej przybliżyć zagadnienie predykcji z modelu mieszanego, posłużymy się przykładem badania mleczości krów. Dla 10 krów (oznaczonych literami od A do J) zmierzono ilość mleka wyprodukowaną przez każdą z nich w ciągu dnia. Pomiary powtórzono pięciokrotnie [1].

Model ma postać

$$y_{milk.amount} = \mu + Z_{cow}u_{cow} + \varepsilon,$$

$$u_{cow} \sim \mathcal{N}(0, \sigma^2_{cow}).$$

Jeżeli chcielibyśmy dokonać predykcji dla nieznanej lub do tej pory niezbadanej przez nas krowy, to wynikiem byłaby ocena średniej dla całej populacji, czyli $\hat{\mu}$.

Aby dokonać predykcji dla konkretnej krowy spośród tych przebadanych, potrzebne nam są oceny efektów osobniczych krów.

Znając macierz D i parametry β, predykcje efektów losowych $\widetilde{u}$ można wyznaczyć ze wzoru

$$\widetilde{u} = DZ^TV^{-1}(y - X\beta),$$

gdzie V to macierz $\sigma^2(I + ZDZ^T)$ [1].

Wówczas predykcja będzie sumą $\hat{\mu}$ oraz oceny efektu losowego dla odpowiedniego osobnika.

Rozdział 2

Badania własne

2.1. Zbiór danych i jego wstępne przygotowanie

Zbiór danych pochodzi z witryny internetowej Our World In Data [3], gdzie dane zostały zebrane z różnych źródeł, m. in. ze Światowej Organizacji Zdrowia (WHO) oraz Europejskiego Centrum ds. Zapobiegania i Kontroli Chorób (ECDC). W zbiorze znajduje się 210 krajów, dane dotyczące terytoriów międzynarodowych oraz łącznie dla całego świata. Mamy ponad 40 kolumn z różnymi parametrami - w dalszej części pracy opiszemy, które zmienne będą przez nas użyte.

W zbiorze znajdowało się wiele braków danych. Dla każdego kraju zostały usunięte dane sprzed rozpoczęcia się epidemii na jego terytorium (`total cases=0`), dni są numerowane kolejnymi liczbami całkowitymi.

Ze zbioru danych zostały usunięte wszystkie kraje o populacji poniżej miliona mieszkańców, ponieważ w większości były to nieduże wysepki, dla których dane były wybrakowane. Oprócz tego, kilka innych krajów zostało usuniętych, ponieważ mimo większej populacji, dane były niepełne.

Do formułowania hipotez i budowania modeli będziemy się posługiwać następującymi zmiennymi:

— liczba zachorowań - jest to liczba potwierdzonych przypadków koronawirusa w danym kraju od momentu rozpoczęcia epidemii. Zamiast wartośći liczby zachorowań, będziemy używać liczby zachorowań na milion mieszkańców (`total cases per million`),
— liczba wykonanych testów - będziemy używać liczby wykonanych testów w przeliczeniu na tysiąc mieszkańców danego kraju (`total tests per thousand`),
— wskaźnik siły obostrzeń (`sringency index`) - wskaźnik tego, jak silne obostrzenia wprowadził rząd danego kraju. Jest to kombinacja dziewięciu innych zmiennych, m.in. zamykanie szkół, polityka wykonywania testów, ograniczenie kontaktów międzyludzkich itp. Może przyjmować wartości od 0 do 100, im większa wartość, tym silniejsze obostrzenia w danym kraju [4],

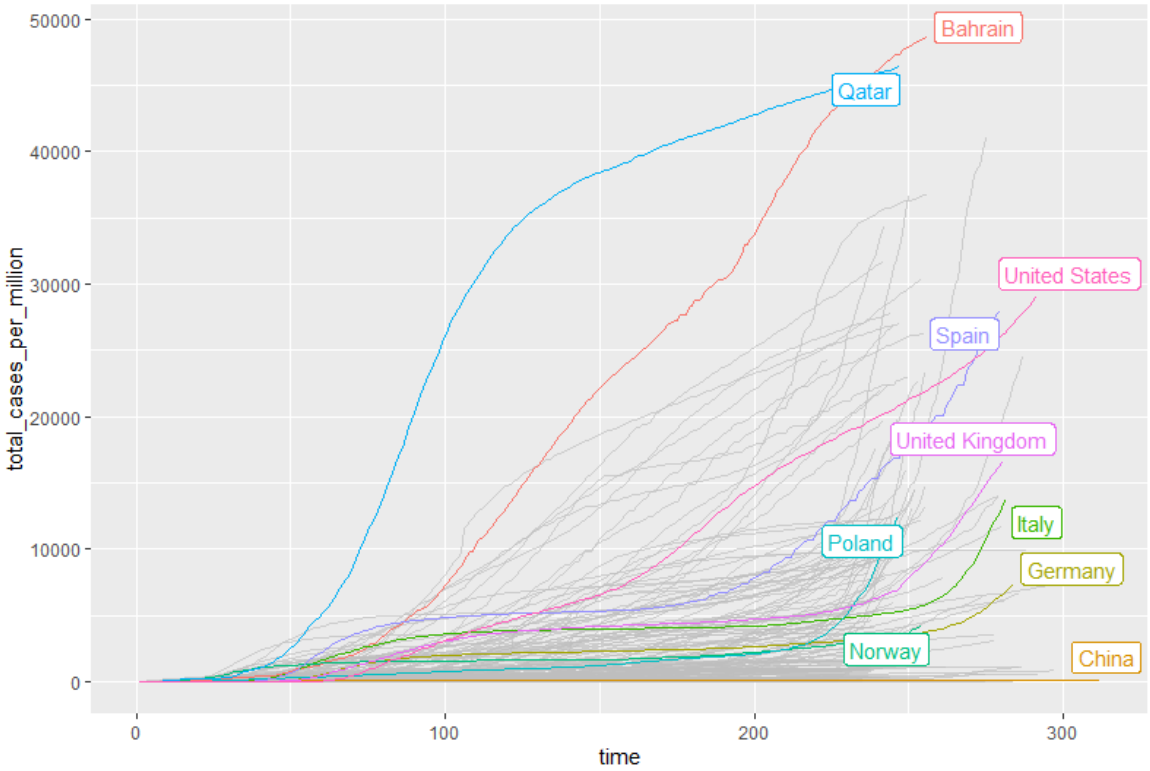
— gęstość zaludnienia (`population density`),
— PKP danego kraju na osobę (`GDP per capita`) - Produkt Krajowy Brutto, przeliczony na hipotetyczną walutę dolara międzynarodowego [5],
— część społeczeństwa żyjąca w skrajnym ubóstwie (`extreme poverty`)
— śmiertelność z powodu chorób sercowych (`cardiovasc death rate`) - stan na rok 2017
— powszechność występowania cukrzycy (`diabetes prevalence`) - odsetek populacji z cukrzycą, brane pod uwagę są osoby w wieku od 20 do 70 lat, stan na rok 2017
— oczekiwana długość życia (`life expectancy`) - kraje zostaną podzielone na kategorie ze wzgędu na tą zmienną, wyróżnimy kraje, w których oczekiwana długość życia jest poniżej 50 lat, między 50 a 54, między 55 a 59 i tak dalej aż do grupy krajów z oczekiwaną długością życia powyżej 80 lat.

2.2. Dyskusja wyników

We wszystkich opisanych poniżej modelach, kraj (`location`) jest efektem losowym.

2.2.1. Model 1

Hipoteza 1: Czas ma istotny wpływ na liczbę zachorowań.



Rysunek 2.1: Wykres przedstawiający zależność liczby zachorowań na milion mieszkańców w zależności od czasu z podziałem na kraje
Źródło: Opracowanie własne

Pierwszy model ma postać

$$y_{total\_cases} = \beta_0 + X_{time}\beta_{time} + Z_{location}u_{location} + \varepsilon$$

a więc przedstawia zależność liczby zachorowań od czasu, a kraj jest efektem losowym.

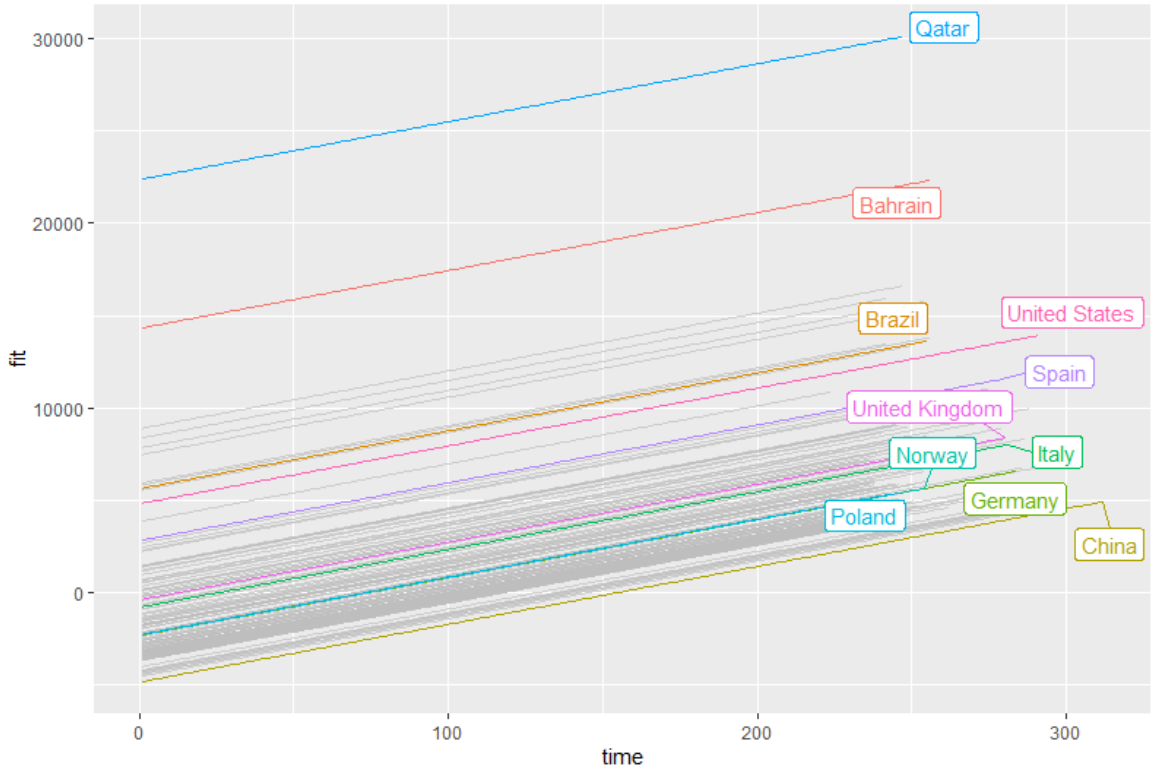
	Model 1
(Intercept)	-1321.86^{***}
	(294.61)
time	31.27^{***}
	(0.24)
AIC	715285.90
BIC	715320.03
Log Likelihood	−357638.95
Num. obs.	37529
Num. groups: location	152
Var: location (Intercept)	13018644.39
Var: Residual	10840327.89

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Tabela 2.1: Wyniki dla modelu 1
Źródło: Opracowanie własne

Widać, że efekt losowy jest odpowiedzialny za ponad połowę wariancji resztowej. Oznacza to, że zmienność liczby zachorowań dla danego kraju jest ponad dwukrotnie mniejsza niż zmienność liczby zachorowań dla różnych krajów.

Zarówno wyraz wolny, jak i współczynnik przy zmiennej `time`, są istotne statystycznie. Dodatkowo, korelacja pomiędzy liczbą zachorowań a czasem jest dodatnia, więc wraz z upływem czasu liczba zachorowań rośnie.



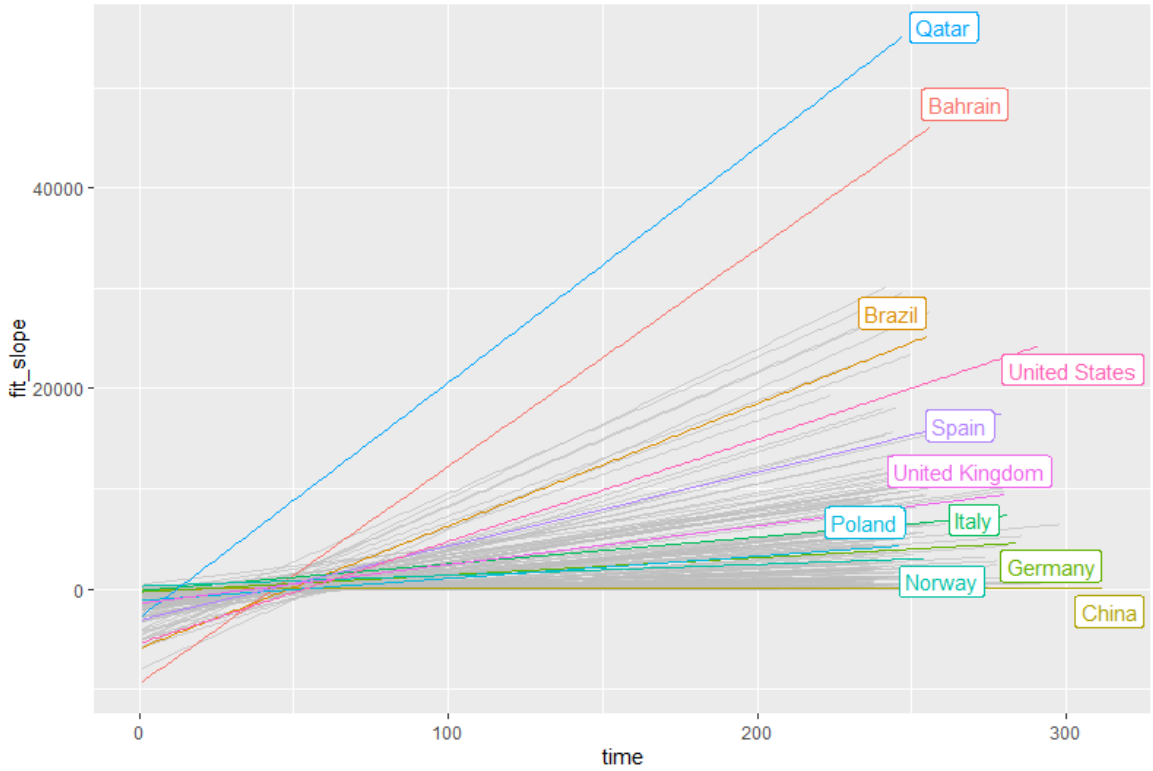
Rysunek 2.2: Wykres przedstawiający zależność między liczbą zachorowań a czasem oszacowaną za pomocą Modelu 1

Źródło: Opracowanie własne

Na rysunku 2.2 widzimy linie dopasowane do liczby zachorowań w porzczególnych krajach. Każda prosta ma taki sam współczynnik kierunkowy, jedynie punkt przecięcia z osią OY (`Intercept`) różni się pomiędzy poszczególnymi krajami. Z tego wykresu możemy odczytać, jak różnią się średnie poziomy liczby zachorowań między krajami.

Oprócz powyższego modelu, w którym tylko wyraz wolny różni się pomiędzy krajami, można rozważyć także model, gdzie współczynnik nachylenia prostej także będzie zależał od efektu losowego, czyli model postaci:

$$y_{total\_cases} = \beta_0 + X_{time}\beta_{time} + Z_{0,location}u_{0,location} +$$
$$+ Z_{time,location}u_{time,location} + \varepsilon$$



Rysunek 2.3: Wykres przedstawiający zależność między liczbą zachorowań a czasem oszacowaną za pomocą modelu, gdzie zarówno wyraz wolny, jak i współczynnik nachylenia prostej różnią się pomiędzy krajami

Źródło: Opracowanie własne

Na rysunku 2.3 można zaobserwować, jak różnią się tendencje rozwojowe pandemii w poszczególnych krajach.

2.2.2. Model 2

Hipoteza 2: Liczba wykonywanych testów na COVID-19 ma związek z liczbą zachorowań.

Drugi model to:

$$y_{total\_cases} = \beta_0 + X_{total\_tests}\beta_{total\_tests} + Z_{location}u_{location} + \varepsilon$$

Badamy tutaj, czy liczba wykonywanych testów (w przeliczeniu na 1000 mieszkańców) ma wpływ na liczbę zachorowań.

Dla tego modelu otrzymujemy następujące wyniki:

	Model 1
(Intercept)	1319.61***
	(399.12)
total_tests_per_thousand	31.95***
	(0.28)
AIC	363631.09
BIC	363662.51
Log Likelihood	−181811.54
Num. obs.	19045
Num. groups: location	95
Var: location (Intercept)	14965179.24
Var: Residual	11168063.61

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Tabela 2.2: Wyniki dla modelu 2
Źródło: Opracowanie własne

Widać po pierwsze, że efekt losowy jest odpowiedzialny za ponad połowę zmienności resztowej modelu. Po drugie, widać, że efekt stały liczby wykonywanych testów jest istotny statystycznie, i ma wpływ stymulujący na liczbę zachorowań.

2.2.3. Model 3

Hipoteza 3: kraje o różnej oczekiwanej długości życia różnią się liczbą zachorowań.

Trzeci model wygląda następująco:

$$y_{total\_cases} = \beta_0 + X_{age}\beta_{age} + Z_{location}u_{location} + \varepsilon$$

Prezentuje on zależność liczby zachorowań od oczekiwanej długości życia w danym kraju.

Dla modelu trzeciego otrzymujemy następujące wyniki:

	Model 1
(Intercept)	640.36
	(1343.96)
age60-64	−41.47
	(1502.59)
age65-69	−248.44
	(1551.85)
age70-74	1509.97
	(1476.43)
age75-79	3319.36*
	(1436.70)
age80 and above	4104.39**
	(1480.87)
agebelow 55	−371.40
	(1993.64)
AIC	729555.23
BIC	729632.03
Log Likelihood	−364768.62
Num. obs.	37529
Num. groups: location	152
Var: location (Intercept)	10768986.56
Var: Residual	15929745.12

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Tabela 2.3: Wyniki dla modelu 3
Źródło: Opracowanie własne

Ponownie efekt losowy odpowiada za mniej niż połowę wariancji resztowej. Dla grupy wiekowej 75-79 różnica w średniej liczbie zachorowań jest na granicy istotności statystycznej. Dopiero dla krajów o oczekiwanej długości życia powyżej 80 lat pojawia się istotna różnica - liczba zachorowań w tych krajach jest największa. W pozostałych grupach wiekowych nie można mówić o istotnych różnicach.

2.2.4. Model 4

Hipoteza 4: Kraje o różnej gęstości zaludnienia różnią się liczbą zachorowań.

W czwartym modelu badamy zależnośc liczby zachorowań od gęstości zaludnienia:

$$y_{total\_cases} = \beta_0 + X_{population\_density}\beta_{population\_density} + Z_{location}u_{location} + \varepsilon$$

Wyniki są następujące:

	Model 1
(Intercept)	2402.87***
	(306.53)
population_density	0.78
	(0.44)
AIC	729672.88
BIC	729707.01
Log Likelihood	−364832.44
Num. obs.	37529
Num. groups: location	152
Var: location (Intercept)	13106475.48
Var: Residual	15929750.91

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Tabela 2.4: Wyniki dla modelu 4
Źródło: Opracowanie własne

Zatem gęstośc zaludnienia nie jest czynnikiem istotnie różnicującym liczbę zachorowań w krajach.

2.2.5. Model 5

Hipoteza 5: Kraje różniące się siłą obostrzeń mają istotne różnice w liczbie zachorowań.

Piaty model ma następującą postać:

$$y_{total\_cases} = \beta_0 + X_{stringency\_index}\beta_{stringency\_index} + Z_{location}u_{location} + \varepsilon$$

W tym modelu sprawdzamy zależność liczby zachorowań od siły obostrzeń. Otrzymujemy następujące wyniki:

	Model 1
(Intercept)	2932.54***
	(301.93)
stringency_index	−9.12***
	(1.16)
AIC	687061.29
BIC	687095.20
Log Likelihood	−343526.65
Num. obs.	35524
Num. groups: location	148
Var: location (Intercept)	12678743.39
Var: Residual	14372604.68

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Tabela 2.5: Wyniki dla modelu 5
Źródło: Opracowanie własne

Tak jak w poprzednim modelu, czynnik losowy odpowiada za niecałą część zmienności resztowej. Wskaźnik siły obostrzeń także jest istotny statystycznie i ma wpływ ograniczający, co oznacza, że im silniejsze obostrzenia, tym mniejsza liczba zachorowań w danym kraju.

2.2.6. Model 6

Hipoteza 6: Kraje o różnej wysokości wskaźnika rozwoju społecznego (HDI) różnią się liczbą zachorowań.

Szósty model przedstawia zależność liczby zachorowań od wskaźnika rozwoju społecznego:

$$y_{total\_cases} = \beta_0 + X_{HDI}\beta_{HDI} + Z_{location}u_{location} + \varepsilon$$

Z tego modelu mamy następujący wynik:

	Model 1
(Intercept)	-4489.15^{***}
	(1237.25)
human_development_index	9957.74^{***}
	(1711.04)
AIC	720398.99
BIC	720433.07
Log Likelihood	−360195.50
Num. obs.	37070
Num. groups: location	150
Var: location (Intercept)	10865235.65
Var: Residual	15798887.42

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Tabela 2.6: Wyniki dla modelu 6
Źródło: Opracowanie własne

Efekt stały jest istotny i ma wpływ stymulujacy. W krajach z wyższym wskaźnikiem rozwoju społecznego, zachorowań jest znacząco więcej.

2.2.7. Model 7

Hipoteza 7: Kraje o różnej wysokości odsetka śmierci z powodu chorób sercowych różnią się liczbą zachorowań.

Model siódmy wygląda następująco:

$$y_{total\_cases} = \beta_0 + X_{cardiovasc\_death\_rate}\beta_{cardiovasc\_death\_rate} + \\ + Z_{location}u_{location} + \varepsilon$$

i zawiera zależność od odsetka śmierci spowodowanych chorobami sercowymi w danym kraju.

Otrzymujemy następujące podsumowanie:

	Model 1
(Intercept)	4982.81***
	(672.06)
cardiovasc_death_rate	−9.25***
	(2.32)
AIC	729657.48
BIC	729691.61
Log Likelihood	−364824.74
Num. obs.	37529
Num. groups: location	152
Var: location (Intercept)	12094158.57
Var: Residual	15929749.88

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Tabela 2.7: Wyniki dla modelu 7
Źródło: Opracowanie własne

Czynnik losowy zachowuje się podobnie jak we wszystkich poprzednich modelach. Efekt stały jest istotny statystycznie. Co ciekawe, odsetek śmierci spowodowanych chorobami serca wpływa ograniczająco na liczbę zachorowań. Może to być związane z tym, że osoby chore na serce bardziej uważają, aby się nie zarazić, tym samym zmniejszają liczbę zachorowań w danym kraju.

2.2.8. Model 8

Hipoteza 8: Kraje o różnej wysokości odsetka osób chorych na cukrzycę różnią się liczbą zachorowań.

$$y_{total\_cases} = \beta_0 + X_{diabetes\_prevalence}\beta_{diabetes\_prevalence} + Z_{location}u_{location} + \varepsilon$$

Model ten jest analogiczny do poprzedniego, z tym że zamiast chorób sercowych mamy tu odsetek chorych na cukrzycę.

Otrzymujemy nastepujący wynik:

	Model 1
(Intercept)	93.18
	(612.83)
diabetes_prevalence	337.40***
	(74.81)
AIC	729646.52
BIC	729680.65
Log Likelihood	−364819.26
Num. obs.	37529
Num. groups: location	152
Var: location (Intercept)	11775109.20
Var: Residual	15929749.48

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Tabela 2.8: Wyniki dla modelu 8
Źródło: Opracowanie własne

Czynnik losowy ma taką samą istotność jak poprzednio. Rozpowszechnienie cukrzycy wpływa stymulująco na liczbę zachorowań. Może to być spowodowane tym, że osoby chore na cukrzycę mają słabszy organizm i są bardziej narażone na zakażenie.

2.2.9. Model 9

Hipoteza 9: Kraje o różnej wysokości odsetka osób żyjących w skrajnej biedzie różnią się liczbą zachorowań.

W tym modelu sprawdzamy zależność liczby zachorowań od odsetka osób żyjących w skrajnym ubóstwie:

$$y_{total\_cases} = \beta_0 + X_{extreme\_poverty}\beta_{extreme\_poverty} + Z_{location}u_{location} + \varepsilon$$

Model ten ma następujące podsumowanie:

	Model 1
(Intercept)	3074.37***
	(306.39)
extreme_poverty	−55.02***
	(11.98)
AIC	520055.39
BIC	520088.22
Log Likelihood	−260023.70
Num. obs.	27083
Num. groups: location	110
Var: location (Intercept)	6944468.09
Var: Residual	12546263.85

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Tabela 2.9: Wyniki dla modelu 9
Źródło: Opracowanie własne

Czynnik stały jest istotny statystycznie. Im większy jest w danym kraju odsetek osób żyjących w biedzie, tym niższa liczba zachorowań. Prawdopodobnie jest to spowodowane mniejszą dostępnością do służby zdrowia w biedniejszych krajach i mniejszą liczbą wykonywanych testów.

2.2.10. Model 10

Hipoteza 10: Kraje o różnej wysokości PKB różnią się liczbą zachorowań.
W modelu dziesiątym pojawia się zależność liczby zachorowań od PKB na osobę:

$$y_{total\_cases} = \beta_0 + X_{GDP}\beta_{GDP} + Z_{location}u_{location} + \varepsilon$$

Wyniki są następujące:

	Model 1
(Intercept)	546.41
	(338.86)
gdp_per_capita	0.11***
	(0.01)
AIC	720911.67
BIC	720945.75
Log Likelihood	−360451.83
Num. obs.	37057
Num. groups: location	150
Var: location (Intercept)	8906070.23
Var: Residual	16132368.91

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Tabela 2.10: Wyniki dla modelu 10
Źródło: Opracowanie własne

Efekt stały jest istotny statystycznie. Im wyższe PKB danego kraju, tym wyższa liczba zachorowań.

Podsumowanie i wnioski

W tabeli 2.11 znajduje się podsumowanie istotności poszczególnych czynników, których wpływ na liczbę zachorowań był badany w tej pracy.

Cecha	Wpływ na liczbę zachorowań
Czas	istotny, wraz z upływem czasu rośnie liczba zachorowań
Liczba wykonywanych testów na COVID-19	istotny, wraz ze wzrostem liczby testów rośnie liczba zachorowań
Oczekiwana długość życia	istotny, o ile ta wartość przekracza 80 lat, wówczas zachorowań jest więcej niż dla krajów o krótszej oczekiwanej długości życia
Gęstość zaludnienia	nieistotny
Wskaźnik siły obostrzeń	istotny, im wyższy, tym mniej zachorowań
Wskaźnik rozwoju społecznego	istotny, w krajach o wysokim wskaźniku rozwoju jest więcej zachorowań
Śmiertelność z powodu chorób sercowych	istotny, im wyższy jest ten współczynnik, tym mniej zachorowań na COVID-19
Powszechność występowania cukrzycy	istotny, im wyższy jest ten współczynnik, tym więcej zachorowań na COVID-19
Część populacji żyjąca w skrajnym ubóstwie	istotny, im większa jest część mieszkańców żyjąca w biedzie, tym mniej zachorowań
PKB na osobę	istotny, im wyższe PKB, tym więcej zachorowań

Tabela 2.11: Porównanie istotności i wpływu różnych czynników na liczbę zachorowań w przeciętnym kraju

Źródło: Opracowanie własne

Nr modelu	AIC
1.1	715285
1.2	655232
2	363631
3	729555
4	729672
5	687061
6	720399
7	729657
8	729646
9	520055
10	720911

Tabela 2.12: Porównanie indeksów Akaike dla poszczególnych modeli
Źródło: Opracowanie własne

Indeks Akaike jest miarą utraconej informacji. Im mniejszy jest indeks Akaike, tym lepiej model wyjaśnia badane zjawisko. Widzimy z tabeli 2.12, że najmniejsze AIC występuje dla modelu nr 2. Jest to model, gdzie występuje zależność między liczbą zachorowań a liczbą wykonywanych testów. Ta zależność jest bardzo oczywista, ponieważ liczbę zachorowań zlicza się na podstawie tego, ile testów dało wynik pozytywny. Modelem z drugim najniższym AIC jest model nr 9, gdzie badamy istotność wskaźnika części populacji żyjącej w skrajnym ubóstwie.

Wszystkie modele jednoznacznie pokazują, że efekt kraju jako czynnika zakłócającego jest bardzo istotny, w wielu przypadkach wyjaśnia ponad połowę wariancji resztowej modelu, więc zmienność liczby zachorowań pomiędzy różnymi krajami jest około dwukrotnie większa niż zmienność liczby zachorowań w pojedynczym kraju.

Na to, co jest nazywane w tej pracy „efektem kraju", składa się tak naprawdę wiele innych czynników, m. in. gęstość zaludnienia, sytuacja ekonomiczna danego kraju, odsetek osób z chorobami towarzyszącymi, rozkład wieku, jak również przyjęta strategia walki z koronawirusem, na którą z kolei składają się m. in. liczba wykonywanych testów, przepisy w sprawie zamykania szkół, miejsc publicznych, ograniczenie kontaktów międzyludzkich, i wiele innych.

Bibliografia

[1] Przemysław Biecek, *Analiza danych z programem R. Modele liniowe z efektami stałymi, losowymi i mieszanymi*, Wydawnictwo Naukowe PWN, Wydanie II, Warszawa 2013

[2] Julian J. Faraway, *Extending the Linear Model with R. Generalized Linear, Mixed Effects and Nonparametric Regression Models. Second Edition*, CRC Press Taylor & Francis Group, 2016

[3] `https://ourworldindata.org/coronavirus`

[4] `https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker` (dostęp 31.10.2020)

[5] `https://ourworldindata.org/what-are-ppps` (dostęp 31.10.2020)

[6] Welham, Sue & Cullis, Brian & Gogel, Beverley & Gilmour, A.R. & Thompson, Robin. *Prediction in linear mixed models.* Australian & New Zealand Journal of Statistics. vol. 46. (2004). p. 325 - 347. 10.1111/j.1467-842X.2004.00334.x.

[7] Howard J. Seltman, *Experimental Design and Analysis*, `http://www.stat.cmu.edu/~hseltman/309/Book/`

[8] `https://peerj.com/articles/4794/` (dostęp: 11.11.2020)

Spis rysunków

Spis tabel

Załączniki

1. Płyta CD z niniejszą pracą w wersji elektronicznej.

Streszczenie (Summary)

Zastosowanie modeli mieszanych w analizie rozwoju pandemii choroby COVID-19 na świecie

W tej pracy przedstawione są pojęcia związane z modelami liniowymi z efektami stałymi i losowymi. Następnie opisane są badania własne na zbiorze danych dotyczącym rozprzestrzeniania się choroby COVID-19 w różnych krajach na świecie.

The use of mixed-effects models in the analysis of the COVID-19 pandemic in the world

In this paper, concepts related to linear models with fixed and random effects are presented. Then, our own research is described on the dataset on the spread of COVID-19 in various countries around the world.