POLITECHNIKA LUBELSKA
WYDZIAŁ PODSTAW TECHNIKI
Kierunek: MATEMATYKA

Praca inżynierska

Zastosowanie modeli mieszanych w analizie rozwoju pandemii wywołanej wirusem 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.

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

$$H_0 : \beta_i = 0$$

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.

1.2.2. Badanie istotności parametrów

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

Te same metody co dla efektów stałych

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.

OPISAĆ KTÓRE KRAJE I DLACZEGO ZOSTAŁY USUNIĘTE ZE ZBIORU, ZWRÓCIĆ UWAGĘ NA ZMIENNE KTÓRYCH BĘDZIEMY UŻYWAĆ W DALSZEJ CZĘŚCI PRACY

2.2. Probelmy szczegółowe i cele

2.2.1. Hipoteza 1

Wpływ kraju (efektu losowego) jest większy niż wpływ czasu (czynnika stałego) w modelu mieszanym.

2.2.2. Hipoteza 2

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

2.2.3. Hipoteza 3

Ze względu na zmienną `life_expectancy` dzielimy kraje na grupy co 5 lat: poniżej 50, 50-54, 55-59, ..., 80 i więcej.

Hipoteza: kraje w różnych grupach ze względu na oczekiwaną długość życia różnią się liczbą zachorowań

2.2.4. Hipoteza 4

Kraje o różnej gęstości zaludnienia różnią się liczbą zachorowań.

2.2.5. Hipoteza 5

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

2.2.6. Hipoteza 6

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

2.2.7. Hipoteza 7

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

2.2.8. Hipoteza 8

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

2.2.9. Hipoteza 9

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

2.2.10. Hipoteza 10

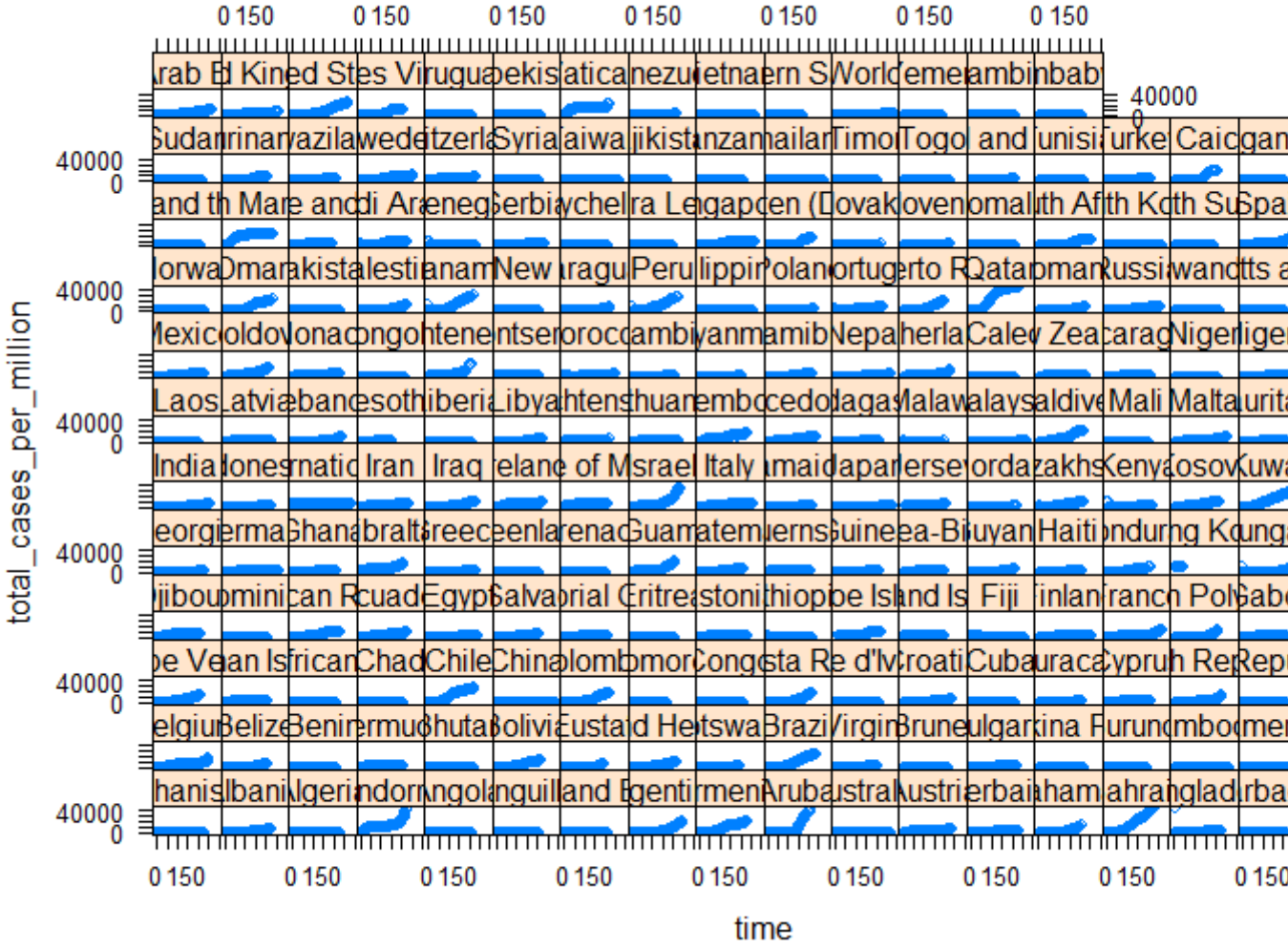
Kraje o różnej wysokości PKB różnią się liczbą zachorowań.

2.3. Dyskusja wyników

2.3.1. Model 1

Pierwszy model ma postać

```
mod <- lme(total_cases_per_million~time,
random = ~1|location,
data = covid)
```



Rysunek 2.1: Wykres przedstawiający rozwój pandemii we wszystkich krajach, tak, wiem, że nic na nim nie widać

Źródło: Opracowanie własne

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

Dla efektu losowego otrzymujemy następujący wynik:

```
Random effects:
Formula: ~1 | location
         (Intercept) Residual
StdDev:    3241.434 4262.032
```

Widać zatem, że efekt losowy jest odpowiedzialny za około 45% wariancji całego modelu.

Dla efektów stałych mamy następujący wynik:

```
Fixed effects: total_cases_per_million ~ time
            Value Std.Error    DF  t-value p-value
(Intercept) 763.1050 225.19114 51584  3.38870   7e-04
time         15.7823   0.25565 51584 61.73324   0e+00
```

A więc zarówno wyraz wolny, jak i współczynnik przy zmiennej Czas, są istotne statystycznie. Dodatkowo, korelacja pomiędzy liczbą zachorowań a czasem jest dodatnia, więc wraz z upływem czasu liczba zachorowań rośnie dla przeciętnego kraju.

2.3.2. Model 2

Drugi model to:

```
mod1 <- lme(total_cases_per_million~time+total_tests_per_thousand,
random = ~1|location,
data = covid_na)
```

Badamy tutaj, czy czas oraz liczba wykonywanych testów mają wpływ na liczbę zachorowań, jeżeli kraj traktujemy jako czynnik losowy.

Dla tego modelu otrzymujemy następujące wyniki:

```
Random effects:
Formula: ~1 | location
         (Intercept) Residual
StdDev:    3677.593 2901.592

Fixed effects: total_cases_per_million ~ time + total_tests_per_thousand
                             Value Std.Error    DF  t-value p-value
(Intercept)              -558.9119  384.0912 17724 -1.45515  0.1456
time                       17.5806    0.4263 17724 41.23543  0.0000
total_tests_per_thousand   22.5888    0.3550 17724 63.63620  0.0000
```

Widać po pierwsze, że efekt losowy jest odpowiedzialny za ponad połowę zmienności modelu. Po drugie, widać, że oba efekty stałe są istotne statystycznie, i oba mają wpływ stymulujący na liczbę zachorowań.

2.3.3. Model 3

Trzeci model wygląda następująco:

```
mod2 <- lme(total_cases_per_million~time+age,
random=~1|location,
```

```
data=covid)
```

Prezentuje on zależność liczby zachorowań od czasu i od oczekiwanej długości życia w danym kraju. Występuje także efekt losowy kraju.

Dla modelu trzeciego otrzymujemy następujące wyniki:

```
Random effects:
Formula: ~1 | location
        (Intercept) Residual
StdDev:     3128.64 2993.161

Fixed effects: total_cases_per_million ~ time + age
                    Value Std.Error    DF   t-value p-value
(Intercept)     -2608.462  1280.159 35550  -2.03761  0.0416
time               28.980     0.231 35550 125.46483  0.0000
age60-64          -69.561  1430.972   145  -0.04861  0.9613
age65-69         -324.578  1477.891   145  -0.21962  0.8265
age70-74         1172.910  1406.069   145   0.83418  0.4056
age75-79         2617.920  1368.246   145   1.91334  0.0577
age80 and above  3195.777  1410.326   145   2.26598  0.0249
agebelow 55      -184.197  1898.558   145  -0.09702  0.9228
```

Ponownie efekt losowy odpowiada za większą część wariancji. Czas ponownie jest istotny i ma wpływ stymulujący. 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.

2.3.4. Model 4

W czwartym modelu badamy zależnośc liczby zachorowań od czasu i gęstości zaludnienia, a kraj jest czynnikiem losowym.

```
mod3 <- lme(total_cases_per_million~time+population_density,
random=~1|location,
data = covid)
```

2.3.5. Model 5

Piaty model ma następującą postać:

```
covid_si <- drop_na(covid, stringency_index)
mod4 <- lme(total_cases_per_million~time+stringency_index,
random=~1|location,
data = covid_si)
```

W tym modelu sprawdzamy zależność liczby zachorowań od czasu i siły obostrzeń, kraj jest czynnikiem losowym.

Otrzymujemy następujące wyniki:

```
Random effects:
Formula: ~1 | location
         (Intercept) Residual
StdDev:    3172.804 2854.192

Fixed effects: total_cases_per_million ~ time + stringency_index
                       Value Std.Error    DF   t-value p-value
(Intercept)      -1710.7939 269.33057 33756  -6.35202       0
time                27.9374   0.23445 33756 119.16023       0
stringency_index    10.1151   0.89377 33756  11.31725       0
```

Tak jak w poprzednich modelach, czynnik losowy odpowiada za największą część zmienności. Czas jest istotny statystycznie. Wskaźnik siły obostrzeń także jest istotny i ma wpływ stymulujący, co oznaczałoby, że im silniejsze obostrzenia, tym więcej zachorowań. Ta interpretacja prawdopodobnie jest niepoprawna, można się domyślać, że raczej zachodzi odwrotna zależność - w krajach z największą liczbą zachorowań są wprowadzane najsurowsze obostrzenia.

2.3.6. Model 6

Szósty model przedstawia zależność liczby zachorowań od czasu i wskaźnika rozwoju społecznego, a kraj jest czynnikiem losowym.

```
covid_hdi <- drop_na(covid, human_development_index)
mod5 <- lme(total_cases_per_million~time+human_development_index,
random=~1|location,
covid_hdi)
```

Z tego modelu mamy następujący wynik:

```
Random effects:
Formula: ~1 | location
         (Intercept) Residual
StdDev:    3130.083 2984.162

Fixed effects: total_cases_per_million ~ time + human_development_index
                            Value Std.Error    DF   t-value p-value
(Intercept)             -6632.736 1173.8596 35117  -5.65037       0
time                       28.795    0.2315 35117 124.40143       0
human_development_index  7807.594 1623.1648   148   4.81011       0
```

Efekt losowy jest odpowiedzialny za nieznacznie większą część wariancji niż wszystkie pozostałe czynniki. Efekty stałe są istotne i oba mają wpływ stymulujacy. W krajach z wyższym wskaźnikiem rozwoju społecznego, zachorowań jest znacząco więcej.

2.3.7. Model 7

Model siódmy wygląda następująco:

```
mod6 <- lme(total_cases_per_million~time+cardiovasc_death_rate,
random=~1|location,
data= covid)
```

i oprócz zależności liczby zachorowań od czasu zawiera także zależność od odsetka śmierci spowodowanych chorobami sercowymi. Kraj jest traktowany jako efekt losowy.

Otrzymujemy następujące podsumowanie:

```
Random effects:
Formula: ~1 | location
       (Intercept) Residual
StdDev:    3241.687 2993.161

Fixed effects: total_cases_per_million ~ time + cardiovasc_death_rate
                          Value Std.Error    DF   t-value p-value
(Intercept)           830.9013  626.6099 35550   1.32603  0.1848
time                   28.9831    0.2310 35550 125.48108  0.0000
cardiovasc_death_rate  -7.4612    2.1624   150  -3.45041  0.0007
```

Czynnik losowy zachowuje się podobnie jak we wszystkich poprzednich modelach. Oba czynniki stałe są istotne. 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.

2.3.8. Model 8

```
mod7 <- lme(total_cases_per_million~time+diabetes_prevalence,
random=~1|location,
data= covid)
```

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

Otrzymujemy nastepujący wynik:

```
Random effects:
Formula: ~1 | location
        (Intercept) Residual
StdDev:    3173.523 2993.161

Fixed effects: total_cases_per_million ~ time + diabetes_prevalence
                         Value Std.Error    DF   t-value p-value
(Intercept)          -3320.595  566.8844 35550  -5.85762       0
time                    28.984    0.2310 35550 125.48575       0
diabetes_prevalence    300.581   69.1237   150   4.34845       0
```

Czynnik losowy ma taką samą istotność jak poprzednio, czas także. 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.

2.3.9. Model 9

W tym modelu sprawdzamy zależność liczby zachorowań od czasu i odsetka osób żyjących w skrajnym ubóstwie. Kraj jest czynnikiem losowym.

```
covid_ep <- drop_na(covid, extreme_poverty)
mod8 <- lme(total_cases_per_million~time+extreme_poverty,
random=~1|location,
data= covid_ep)
```

Model ten ma następujące podsumowanie:

```
Random effects:
Formula: ~1 | location
         (Intercept) Residual
StdDev:    2418.049 2562.526

Fixed effects: total_cases_per_million ~ time + extreme_poverty
                     Value Std.Error    DF   t-value p-value
(Intercept)     -479.1335 282.20423 25650  -1.69783  0.0896
time              26.5944   0.23335 25650 113.96617  0.0000
extreme_poverty  -42.5325  10.97942   108  -3.87384  0.0002
```

Czynnik losowy jest odpowiedzialny za niecałą połowę zmienności. Oba czynniki stałe są istotne. Im większy jest w danym kraju odsetek osób żyjących w biedzie, tym wyższa liczba zachorowań (miejmy nadzieję, że nasz rząd tego nie usłyszy).

2.3.10. Model 10

W modelu dziesiątym pojawia się zależność liczby zachorowań od czasu i PKB. Kraj jest efektem losowym.

```
covid_gdp <- drop_na(covid, gdp_per_capita)
mod9 <- lme(total_cases_per_million~time+gdp_per_capita,
random=~1|location,
data= covid_gdp)
```

Podsumowanie i wnioski

Badania jednoznacznie pokazują, że efekt kraju jako czynnika zakłócającego jest bardzo istotny, bardziej niż jakikolwiek inny czynnik stały (np. czas).

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.

W mojej pracy nie zajmowałam się badaniem, w jaki sposób te czynniki wpływają na wzrost lub spadek liczby zachorowań, chcę jedynie zasygnalizować, że mogą być istotne, skoro wykazany został wpływ efektu kraju na liczbę zachorowań.

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`

Spis rysunków

Spis tabel

Załączniki

1. Płyta CD z niniejszą pracą w wersji elektronicznej.

Streszczenie (Summary)

Zastosowanie modeli mieszanych w analizie rozwoju pandemii wywołanej wirusem 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.