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 5

Rozdział 1. Teoretyczne podstawy badań własnych 7

1.1. Modele liniowe 7

1.1.1. Metody estymacji parametrów modelu liniowego 7

1.1.2. Badanie istotności parametrów 8

1.2. Modele mieszane 8

1.2.1. Metody estymacji 9

1.2.2. Badanie istotności parametrów 9

Rozdział 2. Badania własne 11

2.1. Probelmy szczegółowe i cele 11

2.1.1. Hipoteza 1 11

2.1.2. Hipoteza 2 11

2.1.3. Hipoteza 3 11

2.1.4. Hipoteza 4 11

2.1.5. Hipoteza 5 11

2.1.6. Hipoteza 6 11

2.1.7. Hipoteza 7 12

2.1.8. Hipoteza 8 12

2.2. Zbiór danych i jego wstępne przygotowanie 12

2.3. Modele 12

2.4. Dyskusja wyników 13

Podsumowanie i wnioski 15

Bibliografia 17

Spis rysunków 19

Spis tabel 21

Załączniki 23

Streszczenie (Summary) 25

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. Probelmy szczegółowe i cele

2.1.1. Hipoteza 1

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

2.1.2. Hipoteza 2

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

2.1.3. Hipoteza 3

Kraje o różnej gęstości zaludnienia różnią się liczbą zachorowań.

2.1.4. Hipoteza 4

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

2.1.5. Hipoteza 5

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

2.1.6. Hipoteza 6

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

2.1.7. Hipoteza 7

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

2.1.8. Hipoteza 8

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

2.1.9. Hipoteza 9

Kraje o różnej wysokości PKB różnią się liczbą zachorowań.

2.2. 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. W przypadku zmiennych takich jak liczba zachorowań, zostały one wypełnione poprzez przepisanie danych z poprzedniego dnia. Dla każdego kraju zostały usunięte dane sprzed rozpoczęcia się epidemii na jego terytorium (`total cases=0`), dni są numerowane kolejnymi liczbami całkowitymi.

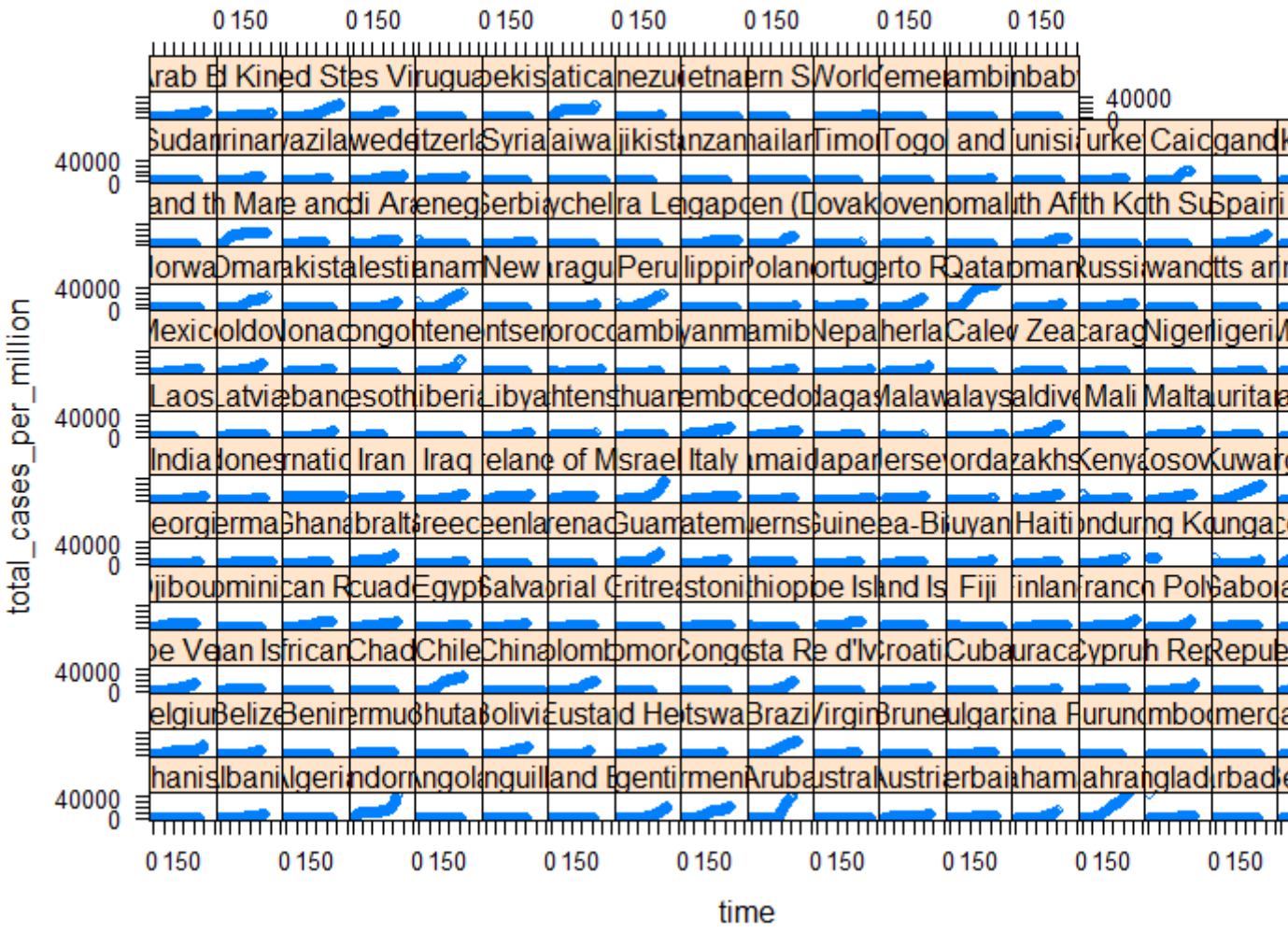
2.3. Modele

Pierwszy model, jaki przetestujemy, to zależność liczby zachorowań na milion mieszkańców w zależności od czasu, gdzie efektem losowym jest kraj.

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

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

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

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

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

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

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

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

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

2.4. Dyskusja wyników



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

Źródło: Opracowanie własne

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

2.1 Wykres przedstawiający rozwój pandemii we wszystkich krajach, tak, wiem, że nic na nim nie widać 14

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.