

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	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
1.1.3. Transformacja zmiennych	9
1.2. Modele mieszane	9
1.2.1. Metody estymacji	10
1.2.2. Badanie istotności parametrów i wybór najlepszego modelu	13
1.2.3. Interpretacja parametrów modelu mieszanego	15
1.2.4. Predykcja z modelu mieszanego	15
Rozdział 2. Badania własne	17
2.1. Opis zbioru badawczego	17
2.2. Dyskusja wyników	18
2.2.1. Model 1: zależność między liczbą zachorowań a czasem	19
2.2.2. Model 2: zależność między liczbą zachorowań a liczbą wykonywanych testów na COVID-19	25
2.2.3. Model 3: zależność między liczbą zachorowań a oczekiwaną długością życia	31
2.2.4. Model 4: zależność między liczbą zachorowań a gęstością zaludnienia	34
2.2.5. Model 5: zależność między liczbą zachorowań a siłą obostrzeń	35
2.2.6. Model 6: zależność między liczbą zachorowań a wskaźnikiem rozwoju społecznego	40
2.2.7. Model 7: zależność między liczbą zachorowań a odsetkiem osób umierających na choroby serca	43
2.2.8. Model 8: zależność między liczbą zachorowań a powszechnością cukrzycy	46
2.2.9. Model 9: zależność między liczbą zachorowań a odsetkiem osób żyjących w skrajnej biedzie	49

2.2.10. Model 10: zależność między liczbą zachorowań a wysokością PKB na osobę	52
Podsumowanie i wnioski	55
Bibliografia	57
Spis rysunków	59
Spis tabel	61
Załączniki	63
Streszczenie (Summary)	65

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.

Przereagować wstęp jak już cała praca będzie napisana, dopisać o metodach analizy.
--

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:

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

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1.$$

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:

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k + \varepsilon,$$

lub w zapisie macierzowym

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

Dopisać o założeniach, czyli jakie mają być epsilony.

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 wiarygodności, ML (ang. *Maximum Likelihood*) polega na maksymalizacji wartości funkcji prawdopodobieństwa ze względu na β (w praktyce maksymalizujemy zwykle logarytm z tej funkcji)

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

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

Estymatory uzyskane tą metodą są asymptotycznie nieobciążone.

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

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

Estymatory uzyskane tą metodą są nieobciążone [1].

Opisać dokładniej REML, zastanowić się czy opisywać tu czy przy mieszanych

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. 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, nie licząc wyrazu wolnego).

Badanie efektów brzegowych poszczególnych zmiennych należy poprzedzić testem F (testem globalnym), który weryfikuje hipotezę

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

przeciwko hipotezie alternatywnej

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

1.1.3. Transformacja zmiennych

Opisać przekształcenie Boxa-Coxa

1.2. Modele mieszane

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

Modeli mieszanych używamy w przypadku powtarzanych pomiarów bądź hierarchicznej, czyli zagnieżdżonej struktury. Takie dane charakteryzują się korelacją między obserwacjami z tej samej grupy, co nie pozwala na użycie modelu liniowego z efektami stałymi, ponieważ założenie o braku seryjnej korelacji błędu modelu nie jest spełnione. 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].

komentarz do powyższego: nie do końca na tym polega różnica... Chodzi o to, że efekty stałe, to takie w których interesuje nas wielkość efektu (średnia), natomiast w efektach losowych bierzemy jedynie pod uwagę fakt, iż wprowadzona zmienna wnosi do modelu pewną zmienność (dokładniej pozwala od całkowitej zmienności, tą zmienność odjąć) i szacuje się tylko wariancję (lub odchylenie std.) a nie jakieś parametry rozkładu. Ponadto efektów losowych można się spodziewać wtedy gdy nie kontrolujemy wszystkich poziomów zmiennej niezależnej.

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 efekt konkretnego pacjenta, ale raczej wpływ leku na przeciętną osobę. Dodatkowo, traktujemy pacjentów jako losowo wybranych. Podejście modelu mieszanego będzie polegało na potraktowaniu wpływu pacjenta jako czynnik zakłócający.

Poprawić styl w ostatnim zdaniu.

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 wiarygodności.

1.2.1. Metody estymacji

Do oceny wartości parametrów modelu mieszanego można stosować metody ML (Największej Wiarygodności) oraz REML (Resztowej Największej Wiarygodności), wspomniane w tej pracy przy okazji modeli liniowych. W przypadku modeli mieszanych obydwojema 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 traktowane jako czynniki zakłócające.

Estymacja parametrów modelu mieszanego jest trudnym zagadnieniem. Można wyróżnić dwa podejścia, jedno z nich wykorzystuje własności macierzy ZDZ^T , a drugie polega na rozwijaniu metod numerycznych używanych do znalezienia maksimum funkcji wiarygodności. Przykładową metodą jest np. użycie algorytmu Newtona-Raphsona - iteracyjnej metody optymalizacji. To podejście jest dobre dla zbioru danych o praktycznie dowolnym rozmiarze, ale ze względu na złożoność pamięciową rzędu $O((p+q)^2)$, źle sprawdza się dla modelu z dużą liczbą parametrów do oszacowania. Pod tym względem bardziej efektywne jest rozwiązanie przy wykorzystaniu własności macierzy rzadkich. To podejście zostanie przedstawione poniżej.

Macierz rzadka jest to macierz, w której większość elementów ma wartość zero. Algorytmy dla macierzy rzadkich są zwykle szybsze niż analogiczne algorytmy dla macierzy gęstych. Zamiast przechowywać wszystkie wartości takiej macierzy, wystarczy zapisać w pamięci wartości i indeksy elementów, które są różne od zera. Macierze rzadkie w praktyce mają często tak wielki rozmiar, że niemożliwe by było opracowanie na nich zwykłymi algorytmami.

W modelu mieszanym macierz Z jest macierzą rzadką. Często również macierz X jest taką macierzą.

Z definicji modelu mieszanego $u \sim \mathcal{N}(0, \sigma^2 D)$, gdzie σ^2 to wariancja wektora ε . Niech

$$D = \Lambda \Lambda^T,$$

gdzie Λ to macierz trójkątna dolna (ponieważ D jest macierzą kowariancji, to zawsze można znaleźć taki rozkład). Macierz D (a przez to także macierz Λ) jest parametry-

zowana wektorem θ . Do tego obierzmy wektor w taki, że

$$u = \Lambda w.$$

Taki wektor w ma rozkład $\mathcal{N}(0, I_{q \times q})$.

Rozkład warunkowy $y|u$ jest rozkładem $\mathcal{N}(X\beta + Zu, \sigma^2 I)$. Ale ponieważ nie obserwujemy u , a jedynie y , to aby wnioskować o u , będziemy rozważać gęstość $u|y$. Z reguły Bayesa mamy

$$f_{u|y} = \frac{f_{y|u} f_u}{f_y}.$$

Zacniemy od wyznaczenia gęstości łącznej $f_{y,u} = f_{y|u} f_u$.

$$\begin{aligned} f_{y,u}(\beta, \theta, \sigma^2) &= f_{y|u}(\beta, \theta, \sigma^2) f_u(\beta, \theta, \sigma^2) = \\ &= \frac{\exp(-(y - X\beta - Z\Lambda u)^T (y - X\beta - Z\Lambda u) / (2\sigma^2))}{(2\pi\sigma^2)^{n/2}} \cdot \frac{\exp(-u^T u / (2\sigma^2))}{(2\sigma^2)^{q/2}} = \\ &= \frac{\exp(-(\|y - X\beta - Z\Lambda u\|^2 + \|u\|^2) / (2\sigma^2))}{(2\pi\sigma^2)^{(n+q)/2}}, \end{aligned}$$

gdzie $\|a\|^2$ to suma kwadratów współrzędnych wektora a .

Minimalizacja tej gęstości po u lub β jest równoważna minimalizacji sumy kwadratów reszt z karą za współczynniki u :

$$r^2(\theta, \beta, w) = \|y - X\beta - Z\Lambda w\|^2 + \|w\|^2.$$

Funkcję $r^2(\theta, \beta, w)$ nazywamy sumą kwadratów reszt z karą oznaczamy *PRSS* (*Penalized Residual Sum of Squares*). Przez r_θ określimy $\min_{w, \beta} r^2(\theta, \beta, w)$. Wartości minimalizujące *PRSS* po w i β można znaleźć, rozwiązując układ równań liniowych.

$$\begin{bmatrix} X^T X & X^T Z \Lambda \\ \Lambda^T Z^T X & \Lambda^T Z^T Z \Lambda + I \end{bmatrix} \begin{bmatrix} \beta \\ w \end{bmatrix} = \begin{bmatrix} X^T y \\ \Lambda^T Z^T y \end{bmatrix} \quad (1.1)$$

To zadanie da się rozwiązać efektywnie nawet dla bardzo dużych q , używając rzadkiej dekompozycji Choleskiego. Po lewej stronie równania przynajmniej macierz $\Lambda^T Z^T Z \Lambda + I$ jest macierzą rzadką. Można ją przedstawić jako

$$\begin{bmatrix} A & 0 \\ B & L \end{bmatrix} \begin{bmatrix} A^T & B^T \\ 0 & L \end{bmatrix}$$

gdzie A i B to nieduże macierze (o ile p jest nieduże), a L to rzadki pierwiastek Choleskiego macierzy $\Lambda^T Z^T Z \Lambda + I_{q \times q}$, czyli jest to rzadka macierz trójkątna dolna spełniająca warunek

$$LL^T = \Lambda^T Z^T Z \Lambda + I_{q \times q}.$$

Użycie dekompozycji Choleskiego macierzy rzadkich, która sama jest rzadka, jest kluczowym momentem pozwalającym na operowanie na dużych macierzach. W tym celu kolumny macierzy Z odpowiednio się permutuje. Dzięki temu można operować na macierzach, które w postaci pełnej nie mieściłyby się w pamięci.

Po wyznaczeniu dekompozycji Choleskiego łatwo rozwiązać układ równań 1.1. Co więcej

$$-2 \log l(\theta, \beta, \sigma|y) = \log(2\pi\sigma^2) + \log(|L|^2) + \frac{r_\theta^2}{\sigma^2}, \quad (1.2)$$

gdzie $|L|$ to wyznacznik macierzy L (która jest trójkątna, więc łatwo go policzyć). Minimalizując powyższe wyrażenie po σ^2 , otrzymujemy warunkowy estymator wariancji (dla zadanego θ):

$$\hat{\sigma}_\theta^2 = \frac{r_\theta^2}{n}$$

Podstawiając tą ocenę wariancji do równości 1.2, otrzymujemy funkcję wariancji sprofilowaną do parametru θ

$$-2 \log l(\theta|y) = \log(|L|^2) + n + n \log \left(\frac{2\pi r_\theta^2}{n} \right)$$

Wartość tej funkcji możemy wyznaczyć efektywnie, nawet dla dużych $p+q$, a sama funkcja wiarygodności zależy jedynie od parametru θ , którego wymiar g jest zazwyczaj nieduży (jest to liczba komponentów wariancyjnych). Maksymalizację tak opisaną funkcji wiarygodności po przestrzeni parametrów o niewielkim wymiarze wykonuje się standardowymi algorytmami numerycznymi.

Wyznaczywszy metodą ML (lub REML) ocenę parametru $\hat{\theta}$, możemy obliczyć oceny pozostałych parametrów modelu.

W powyższej metodzie macierz $V = \sigma^2(I + ZDZ^T)$ mogła mieć prawie dowolną postać. W wielu praktycznych sytuacjach macierz D , a tym samym macierz V , ma bardzo prostą strukturę.

Rozważmy model niezależnych g komponentów losowych, każdy komponent złożony z niezależnych q_i efektów takich, że $\sum_i q_i = q$. Dodatkowo oznaczmy wariancję kolejnych q_i przez $\sigma_i^2 = \sigma^2 \theta_i$. W takim modelu macierz D jest macierzą diagonalną

$$D = \begin{bmatrix} \theta_1 I_{q_1} & 0 & \cdots & 0 \\ 0 & \theta_2 I_{q_2} & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & \theta_g I_{q_g} \end{bmatrix}$$

Dla takiej macierzy D macierz wariancji y można wyrazić jako

$$Var(y) = V = \sigma^2(I + ZDZ^T) = I\sigma^2 + \sum_{i=1}^g \sigma_i^2 Z_i Z_i^T$$

gdzie Z_i to macierz złożona z kolumn macierzy Z odpowiadających tym efektom losowym, które mają wariancję σ_i^2 . Każda Z_i to jeden komponent wariancyjny.

Uwaga 1.1. Może się zdarzyć, że w wyniku estymacji otrzymamy ujemne oceny pewnych parametrów σ_i^2 . Oczywiście nie można takich wartości interpretować jako oceny wariancji. Problem z ujemnymi wartościami $\hat{\sigma}_i^2$ można rozwiązać na kilka sposobów, np.

- wartość ujemną zastąpić przez 0. Jest to metoda najprostsza, ale generuje obciążenie;
- zastosować optymalizację z ograniczeniami. Na przestrzeni parametrów zadajemy liniowe ograniczenia i szukamy maksimum funkcji wiarygodności na zbiorze ograniczonym do nieujemnych parametrów;
- zamiast σ_i^2 można używać innych parametrów, które da się przekształcić w nieujemne oceny σ_i^2 . Przykładowo dla nowej parametryzacji $\gamma_i = \log(\sigma_i^2)$ możemy optymalizować funkcję wiarygodności ze względu na parametry γ_i po całej prostej, a następnie otrzymane oceny $\hat{\gamma}_i$ możemy przekształcić na dodatnie oceny $\hat{\sigma}_i^2 = \exp(\hat{\gamma}_i)$. Wadą tego podejścia jest niemożliwość uzyskania oceny z brzegu przedziału, tzn. nie otrzymamy nigdy oceny $\hat{\sigma}_i^2 = 0$.

1.2.2. Badanie istotności parametrów i wybór najlepszego modelu

W modelach mieszanych konieczne jest zbadanie istotności dla efektów stałych oraz 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 opisujemy, które z nich są najskuteczniejsze [2].

1. Iloraz wiarygodności(ang. *likelihood ratio*) - tworzymy dwa zagnieżdżone 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 wówczas 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.

komentarz do tego podpunktu: Napisz dokładniej o co chodzi, bo to ważne. Dokładniej to robimy to tak... permutujemy wielokrotnie kolumnę/y zawierające interesujące nas efekty, po czym wyliczamy dla nich log funkcji wiarygodności a następnie sprawdzamy ile z nich przekroczyło log f. wiarygodności dla modelu z niepermutowanymi kolumnami.

4. Kryteria informacyjne - służą do wyboru najlepszego spośród modeli. Najpopularniejszym jest Kryterium Informacyjne Akaikego (ang. *Akaike Information Criterion, AIC*). Jest ono zdefiniowane następującym wzorem:

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

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 rozważanych 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].

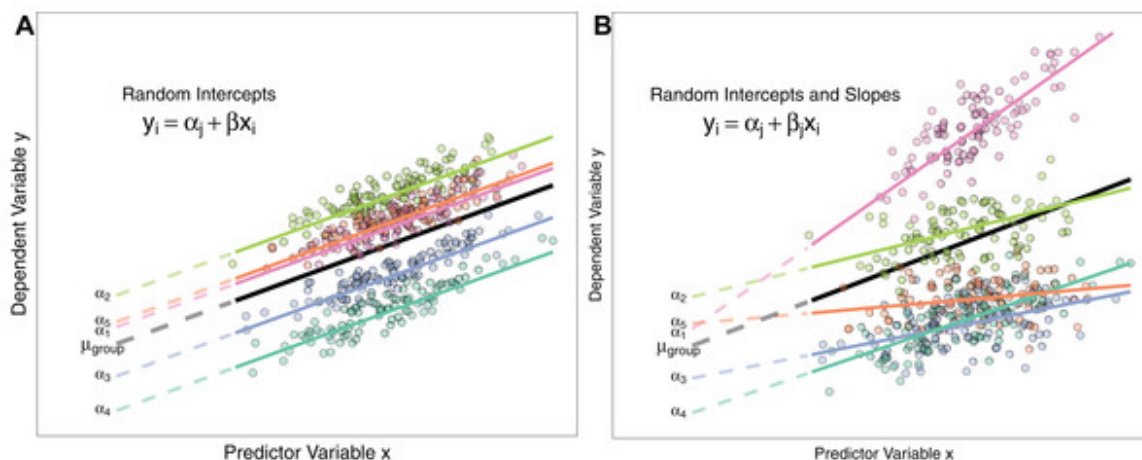
komentarz do ostatniego zdania: jeszcze taka uwaga że nie wszystkie metody nadają się do każdego efektu (nie ma testu Walda dla efektu losowego).

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 globalnej, a w drugim przypadku może dodatkowo być nachylona do osi OX pod mniejszym lub większym kątem.



Rysunek 1.1: Rodzaje modeli mieszanых

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 mleczności 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

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

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

Jeżeli chcielibyśmy dokonać predykcji dla nieznaney 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 \tilde{u} można wyznaczyć ze wzoru

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

gdzie V to macierz $\sigma^2(I + ZDZ^T)$ [1]. Wówczas predykcja będzie sumą $\hat{\mu}$ oraz oceny efektu losowego dla odpowiedniego osobnika.

komentarz: brakuje mi na zakończenie jakiegoś porównania efektów stałych i losowych.

Rozdział 2

Badania własne

2.1. Opis zbioru badawczego

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, 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 (**total cases per million**) - jest to liczba potwierdzonych przypadków koronawirusa w danym kraju od momentu rozpoczęcia epidemii. Zamiast wartości liczby zachorowań, będziemy używać liczby zachorowań na milion mieszkańców ,
- czas (**time**) - numer dnia od początku pandemii w danym kraju
- liczba wykonanych testów (**total tests per thousand**) - będziemy używać liczby wykonanych testów w przeliczeniu na tysiąc mieszkańców danego kraju,
- wskaźnik siły obostrzeń (**stringency index**) - wskaźnik tego, jak silne obostrzenia wprowadził rząd danego kraju. Jest to kombinacja dziewięciu zmiennych, m.in. zamykanie szkół, polityka wykonywania testów, ograniczenie kontaktów między-

- ludzkich itp. Może przyjmować wartości od 0 do 100, im większa wartość, tym silniejsze obostrzenia w danym kraju [4],
- gęstość zaludnienia (**population density**),
 - PKB 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 względu na tę 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.

Dane były zbierane do dnia 1 grudnia 2020 r.

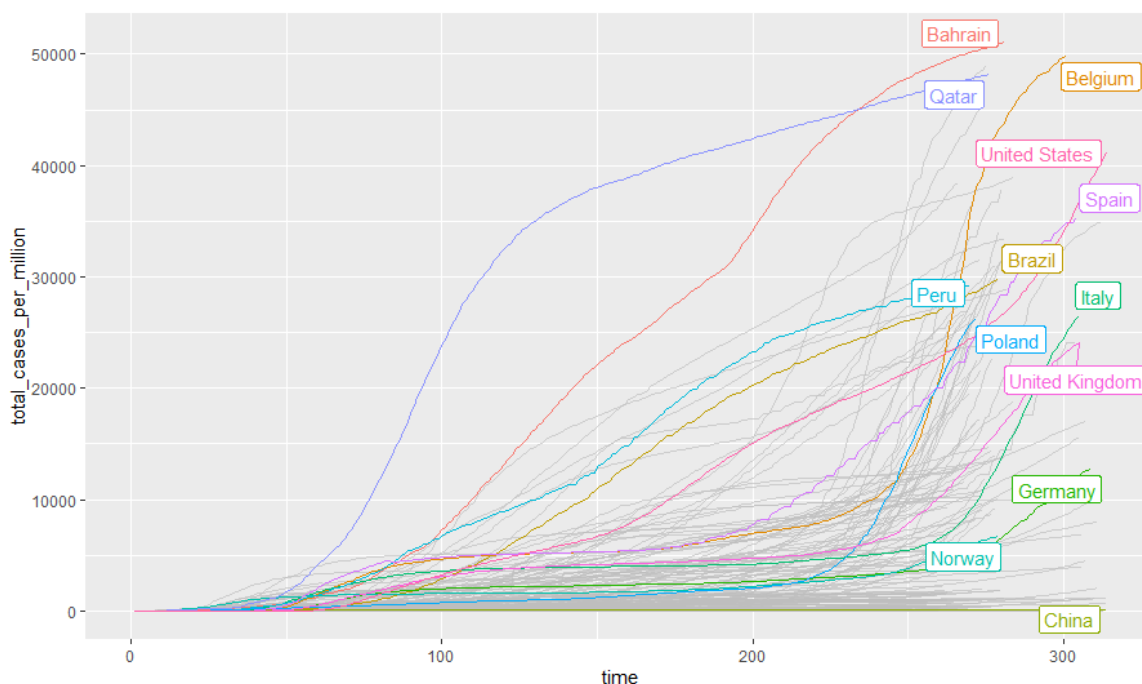
2.2. Dyskusja wyników

We wszystkich modelach mieszanych kraj jest czynnikiem losowym. Modele mieszane są budowane na podstawie całego zbioru danych. Modele liniowe są budowane na podstawie zbioru danych, gdzie znajdują się po maksymalnie cztery obserwacje dla każdego kraju: po 3, po 6, po 9 i po 12 miesiącach trwania epidemii.

Modele są dopasowywane przy użyciu środowiska R. Do dopasowania modeli mieszanych zostały wykorzystane pakiety **lme4** oraz **lmerTest**. Funkcja **lmer** z pakietu **lme4** wykorzystuje opisaną w części teoretycznej metodę oszacowania parametrów modelu mieszanego - algorytm używający macierzy rzadkich oraz dekompozycji Choleskiego [9]. Pakiet **lmerTest** umożliwia obliczenie wartości p dla parametrów modelu mieszanego [10]. W przypadku modeli liniowych użyto funkcji **lm** z pakietu bazowego R. Funkcja ta wykorzystuje metodę najmniejszych kwadratów estymacji parametrów modelu liniowego [11]. W modelach, gdzie dokonano transformacji zmiennych, została użyta funkcja **powerTransform** z pakietu **car**. Funkcja ta wykorzystuje transformację Boxa-Coxa [12].

2.2.1. Model 1: zależność między liczbą zachorowań a czasem

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.

ujednolicić nazwę zmiennej total cases/total cases per million

	Model 1
(Intercept)	−1941.35*** (337.72)
time	37.31*** (0.26)
AIC	806648.05
BIC	806682.57
Log Likelihood	−403320.03
Num. obs.	41287
Num. groups: location	151
Var: location (Intercept)	16967251.36
Var: Residual	17530563.89

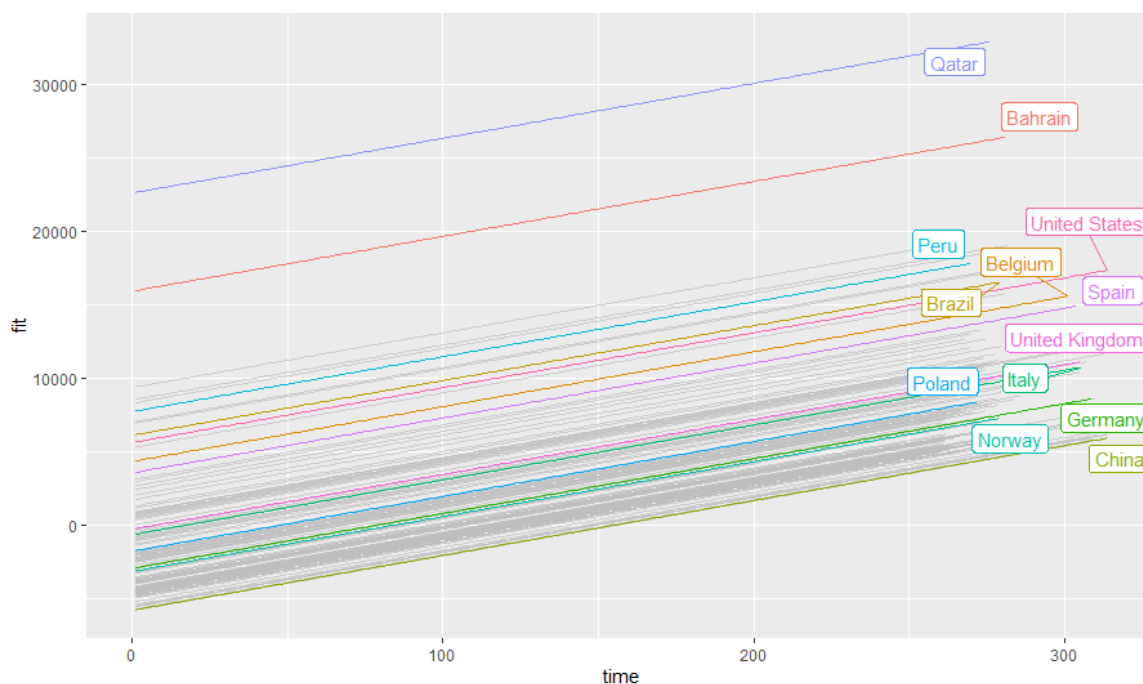
*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

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 poszczególnych 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:

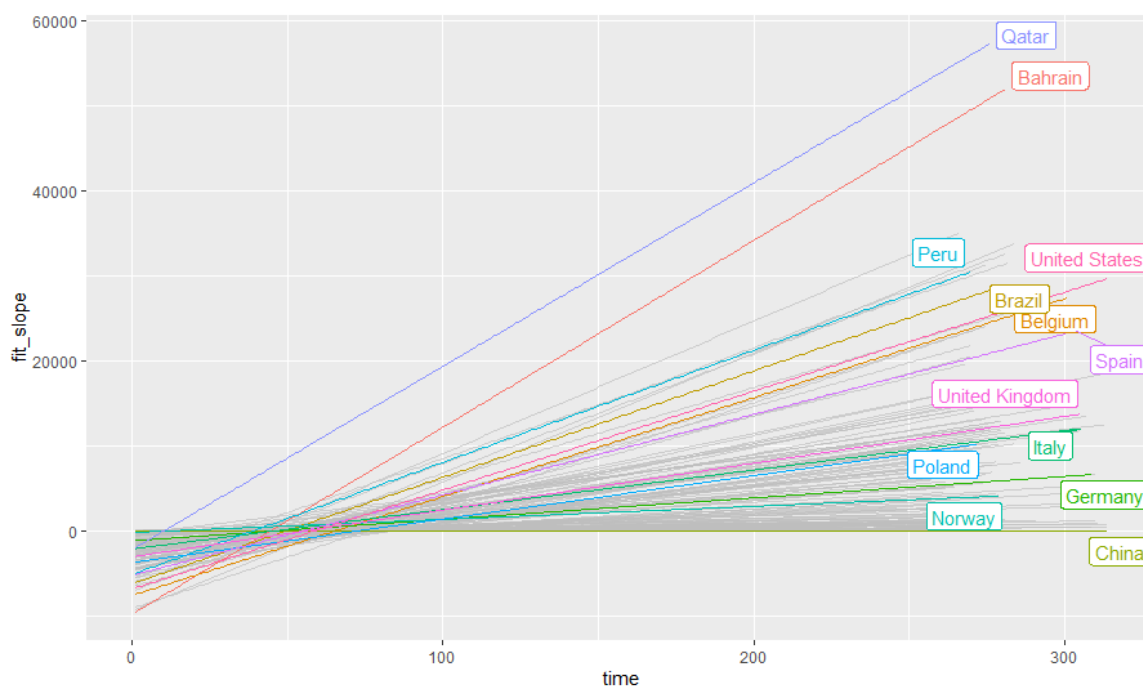
$$y_{total_cases} = \beta_0 + X_{time}\beta_{time} + Z_{0,location}u_{0,location} + \\ + Z_{time,location}u_{time,location} + \varepsilon$$

	Model 1
(Intercept)	−1784.78*** (145.44)
time	35.49*** (2.93)
AIC	759067.36
BIC	759119.13
Log Likelihood	−379527.68
Num. obs.	41287
Num. groups: location	151
Var: location (Intercept)	3113329.04
Var: location time	1296.95
Cov: location (Intercept) time	−53442.24
Var: Residual	5450792.56

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Tabela 2.2: Wyniki dla modelu 1

Źródło: Opracowanie własne



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.

W analogiczny sposób można zbudować model, gdzie zależność od czasu będzie funkcją kwadratową, to jest

$$y_{total_cases} = \beta_0 + X_{time}\beta_{time} + X_{time^2}\beta_{time^2} + \\ + Z_{0,location}u_{0,location} + Z_{time,location}u_{time,location} + \\ Z_{time^2,location}u_{time^2,location} + \varepsilon$$

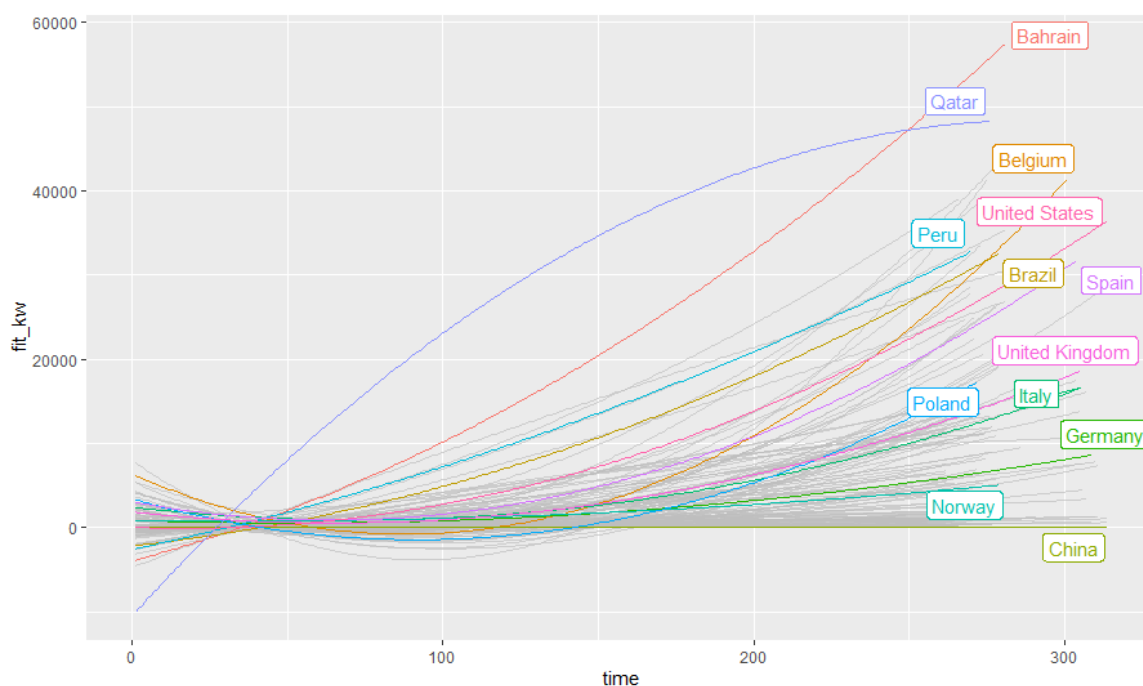
	Model 1
(Intercept)	383.14*** (107.34)
time	−10.66 (161.66)
time ²	0.16 (0.21)
AIC	721041.58
BIC	721127.86
Log Likelihood	−360510.79
Num. obs.	41287
Num. groups: location	151
Var: location (Intercept)	1671050.31
Var: location time	3946212.95
Var: location I(time ²)	6.52
Cov: location (Intercept) time	24812.00
Cov: location (Intercept) I(time ²)	279.78
Cov: location time I(time ²)	5030.38
Var: Residual	2045067.48

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Tabela 2.3: Wyniki dla modelu 1 z czynnikiem kwadratowym

Źródło: Opracowanie własne

Wizualnie model z drugą potęgą zmiennej *time* wydaje się być lepiej dopasowany do danych (rys. 2.4), jednakże w tym modelu ani pierwsza, ani druga potęga zmiennej niezależnej nie są istotne statystycznie.

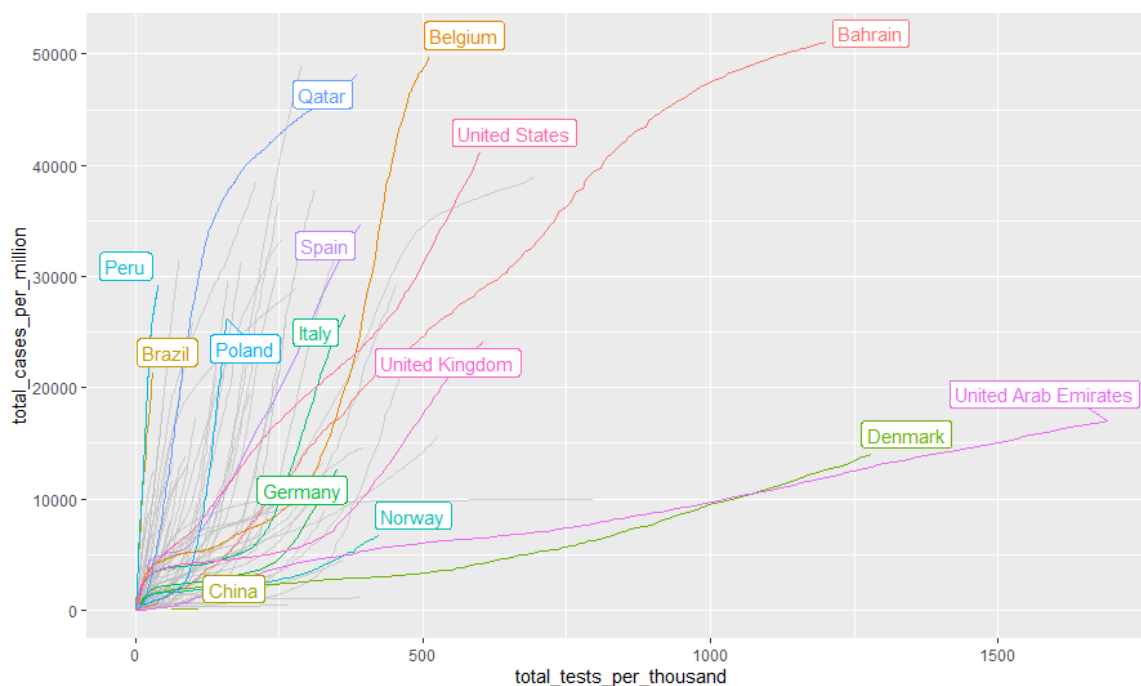


Rysunek 2.4: Wykres przedstawiający zależność między liczbą zachorowań a czasem oszacowaną za pomocą modelu, gdzie zarówno wyraz wolny, jak i współczynniki przy $time$ i $time^2$ zależą od kraju

Źródło: Opracowanie własne

2.2.2. Model 2: zależność między liczbą zachorowań a liczbą wykonywanych testów na COVID-19

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



Rysunek 2.5: Wykres przedstawiający zależność między liczbą zachorowań a liczbą wykonywanych testów w poszczególnych krajach

Źródło: Opracowanie własne

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:

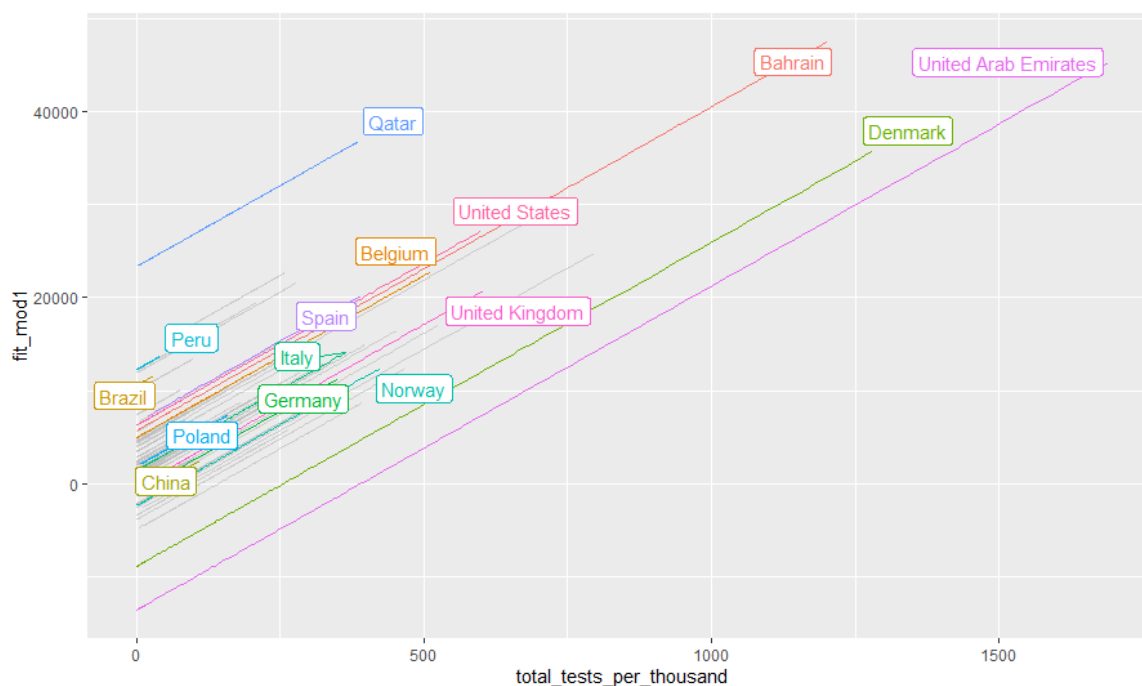
	Model 1
(Intercept)	1741.18*** (464.57)
total_tests_per_thousand	34.85*** (0.28)
AIC	430649.03
BIC	430681.01
Log Likelihood	−215320.52
Num. obs.	21907
Num. groups: location	97
Var: location (Intercept)	20699413.62
Var: Residual	19714018.35

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Tabela 2.4: 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.



Rysunek 2.6: Wykres przedstawiający dopasowanie modelu random intercept

Źródło: Opracowanie własne

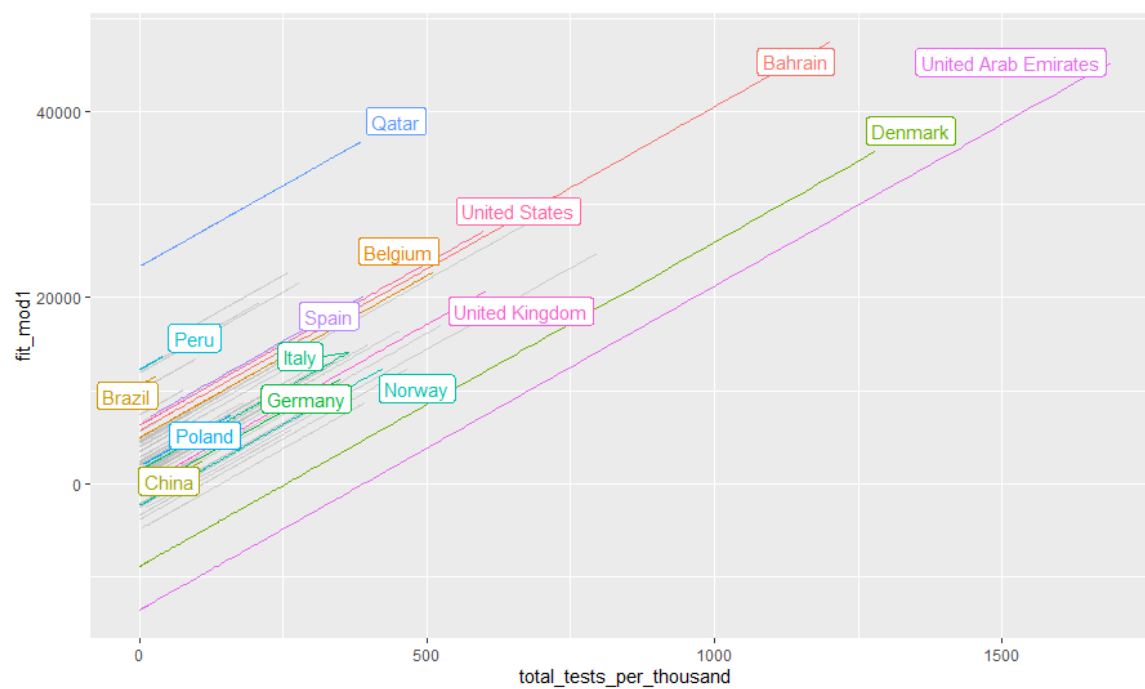
Można także dopasować model, w którym współczynnik nachylenia prostej zależy od kraju:

$$y_{total_cases} = \beta_0 + X_{total_tests} \beta_{total_tests} + \\ + Z_{0,location} u_{0,location} + Z_{total_tests,location} u_{total_tests,location} + \varepsilon$$

	Model 1
(Intercept)	−324.65 (175.92)
total_tests_per_thousand	109.17*** (13.83)
AIC	391507.90
BIC	391555.87
Log Likelihood	−195747.95
Num. obs.	21907
Num. groups: location	97
Var: location (Intercept)	2910697.88
Var: location total_tests_per_thousand	18059.27
Cov: location (Intercept) total_tests_per_thousand	−6245.44
Var: Residual	3206479.62

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Tabela 2.5: Wyniki dla modelu 2 z losowym współczynnikiem nachylenia



Rysunek 2.7: Wykres przedstawiający dopasowanie modelu random intercept and slope

Źródło: Opracowanie własne

2.2.3. Model 3: zależność między liczbą zachorowań a oczekiwaną długością życia

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

Trzeci model jest modelem liniowym:

$$y_{total_cases} = \beta_0 + X_{life_expectancy}\beta_{life_expectancy} + \varepsilon$$

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

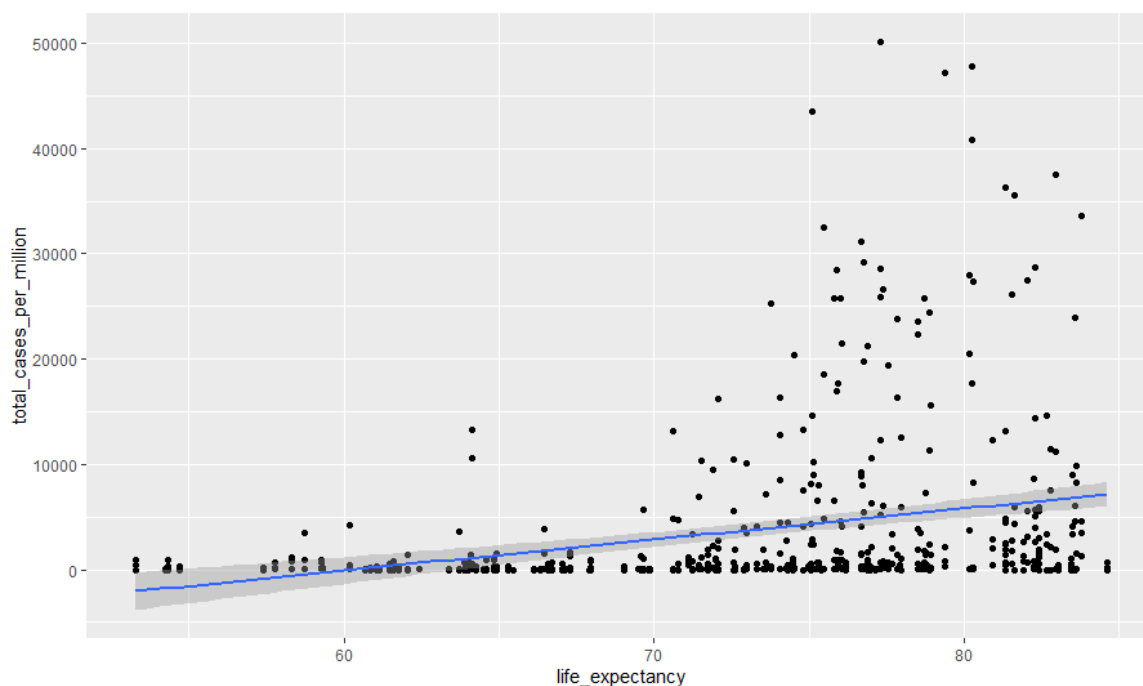
Dla modelu trzeciego otrzymujemy następujące wyniki:

	Model 1
(Intercept)	−14780.85*** (294.55)
life_expectancy	248.30*** (4.03)
R ²	0.08
Adj. R ²	0.08
Num. obs.	41287

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Tabela 2.6: Wyniki dla modelu 3

Źródło: Opracowanie własne



Rysunek 2.8: Wykres przedstawiający dopasowanie modelu liniowego bez przekształcenia zmiennych

Źródło: Opracowanie własne

Z analizy modelu liniowego wynika, że oczekiwana długość życia ma istotny wpływ na liczbę zachorowań. Im większa jest oczekiwana długość życia, tym więcej będzie potwierdzonych przypadków koronawirusa. Z wykresu oraz na podstawie niskiej wartości R^2 (około 8%) wynika, że model jest słabo dopasowany do danych. Można zastosować transformację zmiennych, aby poprawić dopasowanie.

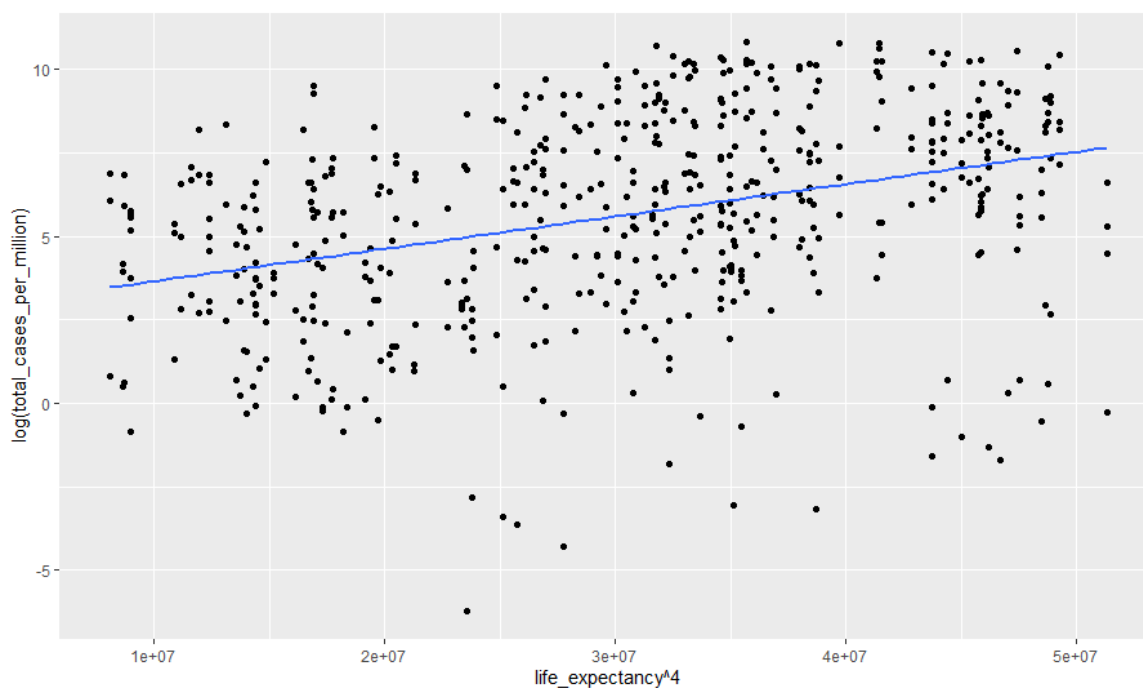
Z transformacji Boxa-Coxa otrzymujemy następujące potęgi dla zmiennych: 3.72 dla `life_expectancy` oraz 0.09 dla `total_cases_per_million`. W przybliżeniu przyjmujemy czwartą potęgę dla pierwszej cechy, a dla drugiej 0 czyli logarytm.

$$\log(y_{total_cases}) = \beta_0 + \beta_{life_expectancy} X_{life_expectancy}^4 + \varepsilon$$

	Model 1
(Intercept)	3.22***
	(0.04)
life_expectancy ⁴	0.00***
	(0.00)
R ²	0.10
Adj. R ²	0.10
Num. obs.	41287

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Tabela 2.7: Wyniki dla modelu 3 po przekształceniu zmiennych



Rysunek 2.9: Wykres przedstawiający dopasowanie modelu liniowego z przekształceniem zmiennych

Źródło: Opracowanie własne

Po przekształceniu, R^2 wzrosło nieznacznie, do 10%. Wniosek pozostaje ten sam - im wyższa oczekiwana długość życia, tym więcej osób chorych na COVID-19.

2.2.4. Model 4: zależność między liczbą zachorowań a gęstością zaludnienia

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

W czwartym modelu badamy zależność liczby zachorowań od gęstości zaludnienia:

$$y_{total_cases} = \beta_0 + X_{population_density}\beta_{population_density} + \varepsilon$$

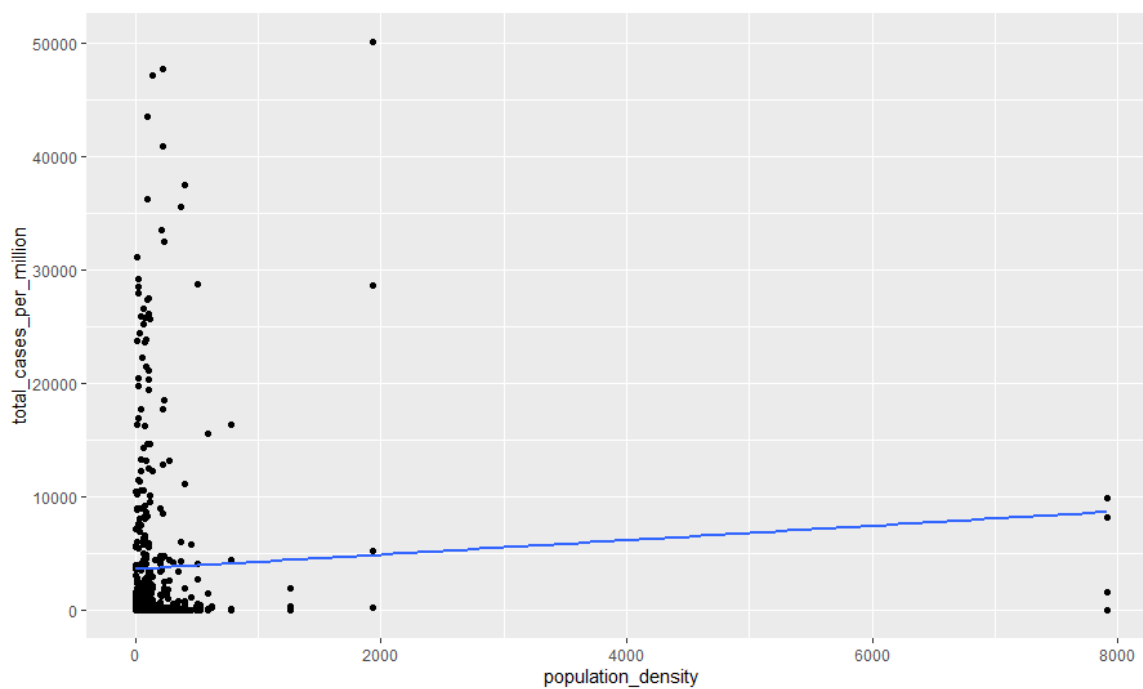
Wyniki są następujące:

	Model 1
(Intercept)	3606.05*** (356.36)
population_density	0.64 (0.49)
R ²	0.00
Adj. R ²	0.00
Num. obs.	537

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Tabela 2.8: Wyniki dla modelu 4

Gęstość zaludnienia nie jest czynnikiem istotnym statystycznie, a R^2 wynosi 0. Zatem liczba zachorowań na COVID-19 nie zależy od gęstości zaludnienia.

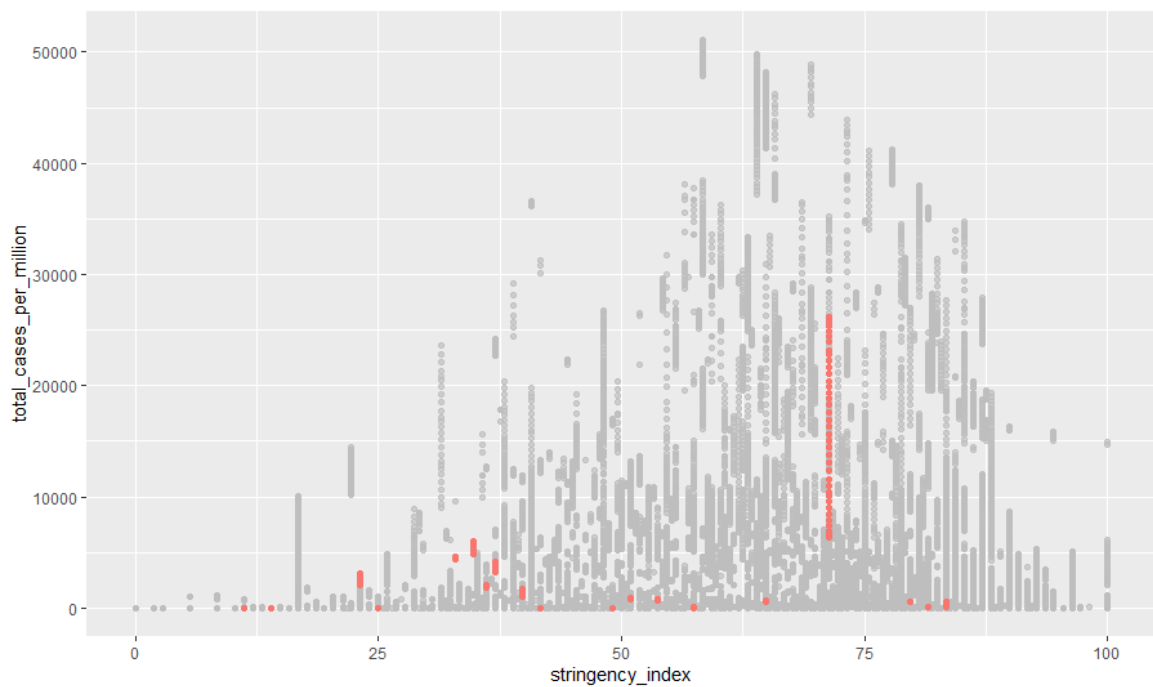


Rysunek 2.10: Wykres przedstawiający dopasowanie modelu liniowego

Źródło: Opracowanie własne

2.2.5. Model 5: zależność między liczbą zachorowań a siłą obostrzeń

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



Rysunek 2.11: Wykres przedstawiający zależność liczby zachorowań od współczynnika siły obostrzeń, kolorem zaznaczone dane dla Polski

Źródło: Opracowanie własne

Piąty 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:

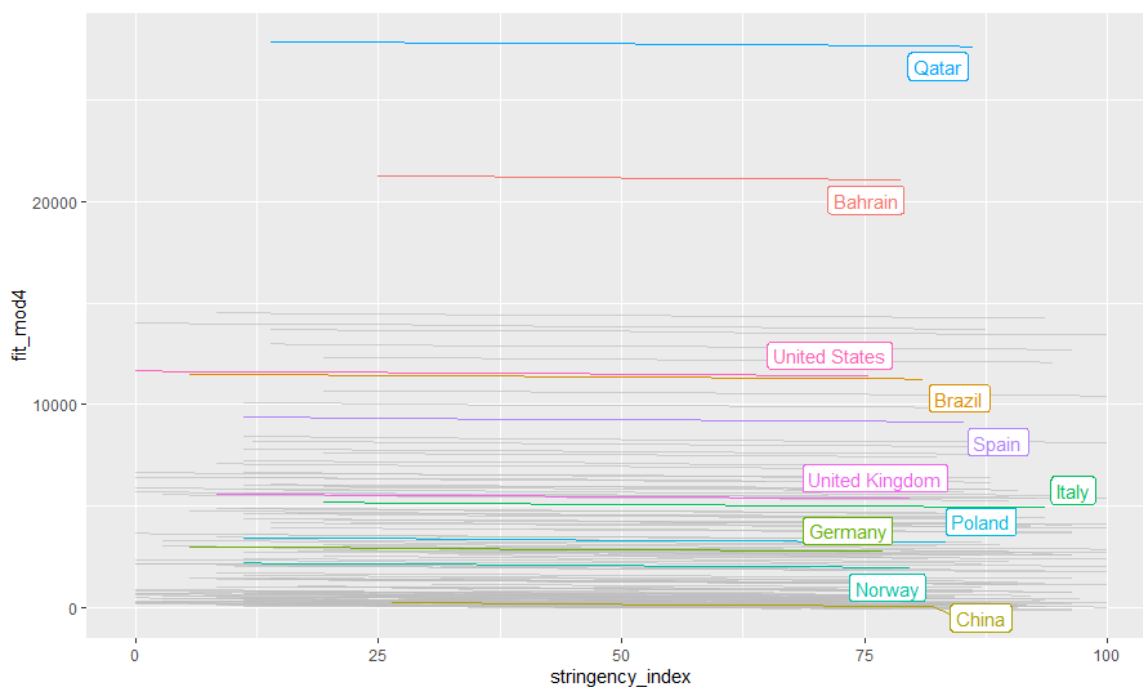
	Model 1
(Intercept)	3319.47*** (356.25)
stringency_index	−3.27* (1.43)
AIC	801139.90
BIC	801174.31
Log Likelihood	−400565.95
Num. obs.	40219
Num. groups: location	147
Var: location (Intercept)	17448313.08
Var: Residual	25718203.07

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Tabela 2.9: Wyniki dla modelu 5

Źródło: Opracowanie własne

Wskaźnik siły obostrzeń jest istotny statystycznie i ma wpływ ograniczający, co oznacza, że im silniejsze obostrzenia, tym mniejsza liczba zachorowań w danym kraju.



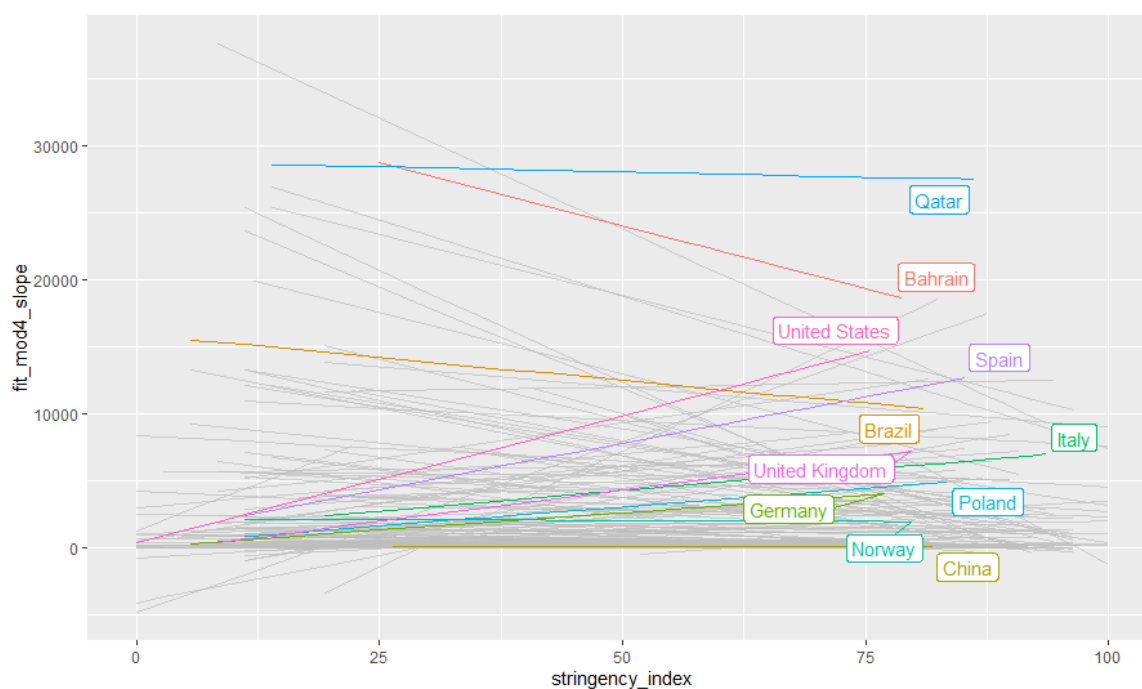
Rysunek 2.12: Wykres przedstawiający dopasowanie modelu random intercept

Źródło: Opracowanie własne

	Model 1
(Intercept)	4607.01*** (711.39)
stringency_index	-15.60 (8.46)
AIC	797733.02
BIC	797784.63
Log Likelihood	-398860.51
Num. obs.	40219
Num. groups: location	147
Var: location (Intercept)	71934195.57
Var: location stringency_index	9970.86
Cov: location (Intercept) stringency_index	-699926.27
Var: Residual	23308070.03

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Tabela 2.10: Wyniki dla modelu 5 z losowym współczynnikiem nachylenia



Rysunek 2.13: Wykres przedstawiający dopasowanie modelu random intercept and slope

Źródło: Opracowanie własne

Z wykresu 2.13 widać, że w niektórych krajach przy wzroście siły obostrzeń, liczba zachorowań maleje, a w innych rośnie. Dla przeciętnego kraju wpływ jest ograniczający (co wynika z tabeli 2.10).

2.2.6. Model 6: zależność między liczbą zachorowań a wskaźnikiem rozwoju społecznego

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

Szósty model przedstawia zależność liczby zachorowań od wskaźnika rozwoju społecznego:

$$y_{total_cases} = \beta_0 + X_{HDI}\beta_{HDI} + \varepsilon$$

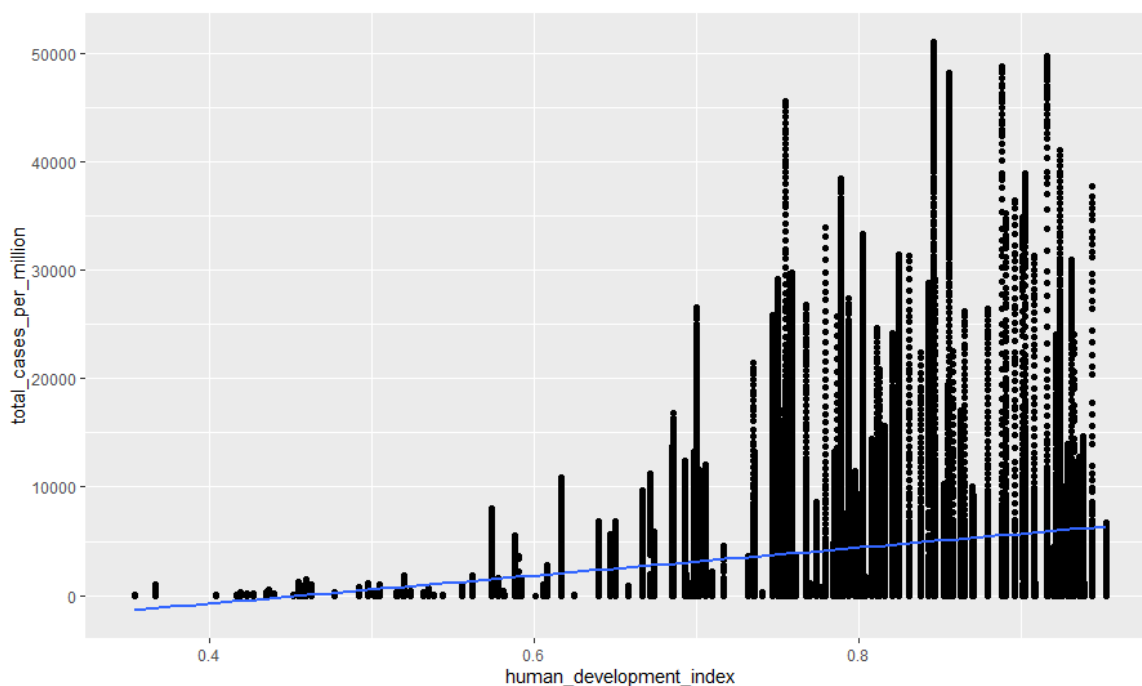
Z tego modelu mamy następujący wynik:

	Model 1
(Intercept)	−5916.47*** (145.15)
human_development_index	12894.77*** (198.99)
R ²	0.09
Adj. R ²	0.09
Num. obs.	41027

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Tabela 2.11: Wyniki dla modelu 6

HDI jest istotny i ma wpływ stymulujący. W krajach z wyższym wskaźnikiem rozwoju społecznego, zachorowań jest znacząco więcej.



Rysunek 2.14: Wykres przedstawiający dopasowanie modelu liniowego

Źródło: Opracowanie własne

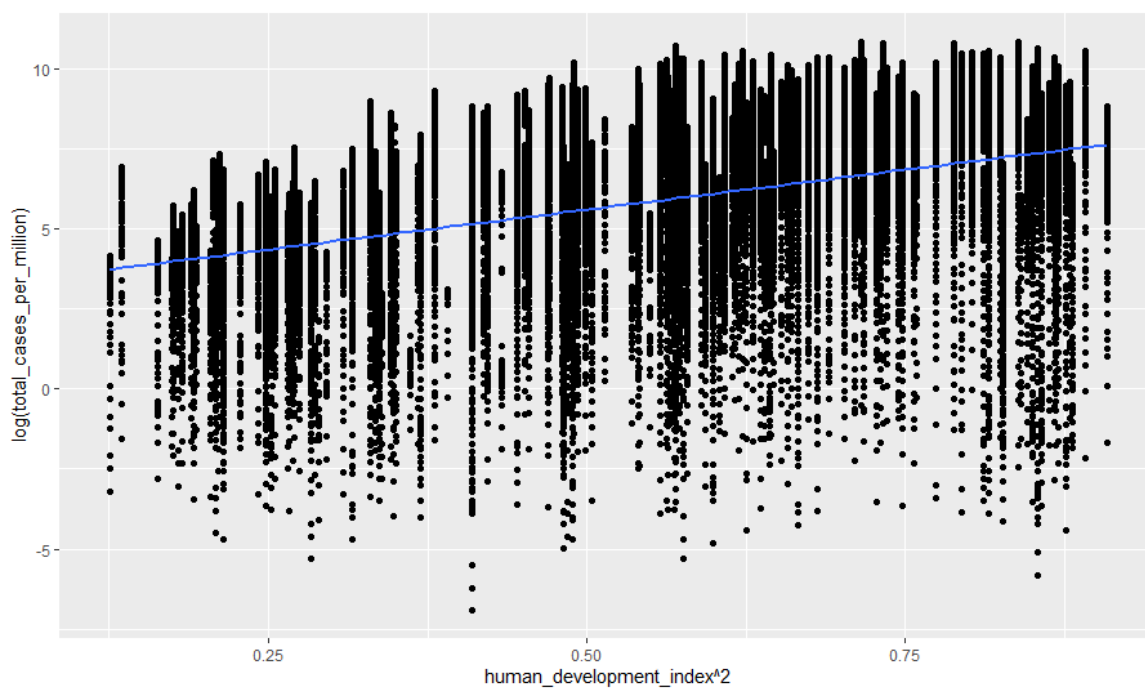
Z przekształcenia Boxa-Coxa otrzymujemy potęgę 2 dla zmiennej HDI oraz logarytm dla $total_cases_per_million$.

$$\log(y_{total_cases}) = \beta_0 + \beta_{HDI} X_{HDI}^2 + \varepsilon$$

	Model 1
(Intercept)	3.11***
	(0.04)
human_development_index ²	4.98***
	(0.07)
R ²	0.12
Adj. R ²	0.12
Num. obs.	41027

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Tabela 2.12: Wyniki dla modelu 6 po przekształceniu zmiennych



Rysunek 2.15: Wykres przedstawiający dopasowanie modelu liniowego z przekształceniem zmiennych

Źródło: Opracowanie własne

2.2.7. Model 7: zależność między liczbą zachorowań a odsetkiem osób umierających na choroby serca

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

Model siódmy wygląda następująco:

$$y_{total_cases} = \beta_0 + X_{cardiovasc_death_rate}\beta_{cardiovasc_death_rate} + \varepsilon$$

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

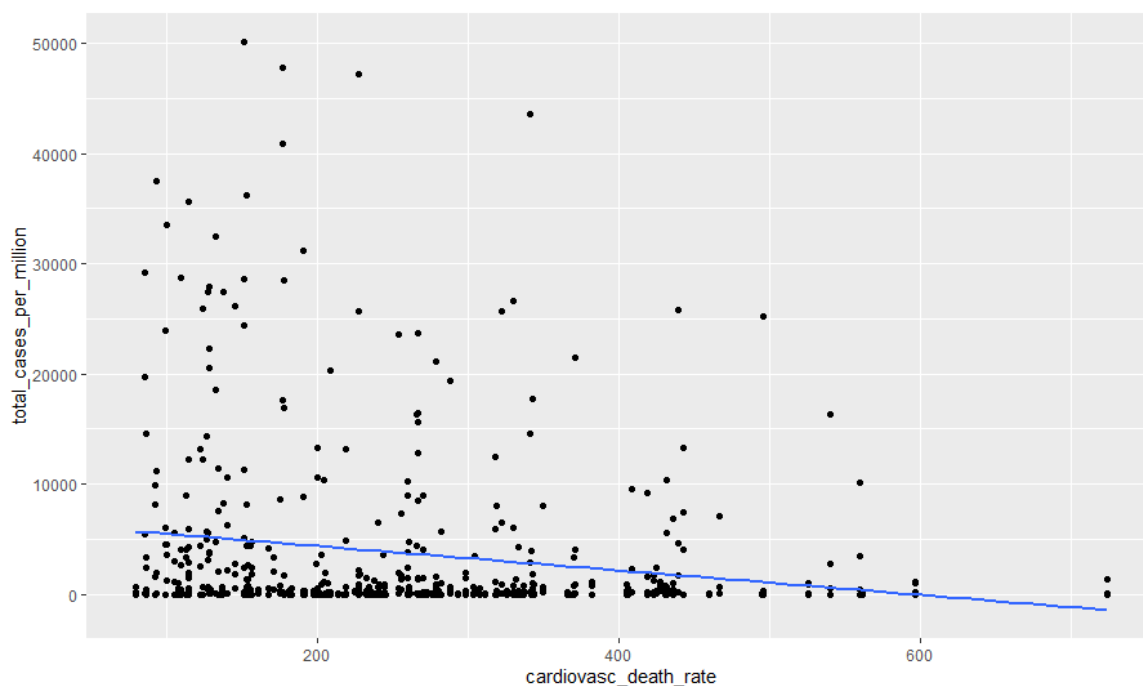
Otrzymujemy następujące podsumowanie:

	Model 1
(Intercept)	6602.10*** (792.12)
cardiovasc_death_rate	-11.06*** (2.76)
R ²	0.03
Adj. R ²	0.03
Num. obs.	537

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Tabela 2.13: Wyniki dla modelu 7

Źródło: Opracowanie własne



Rysunek 2.16: Wykres przedstawiający dopasowanie modelu liniowego

Źródło: Opracowanie własne

Efekt śmiertelności na choroby serca 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.

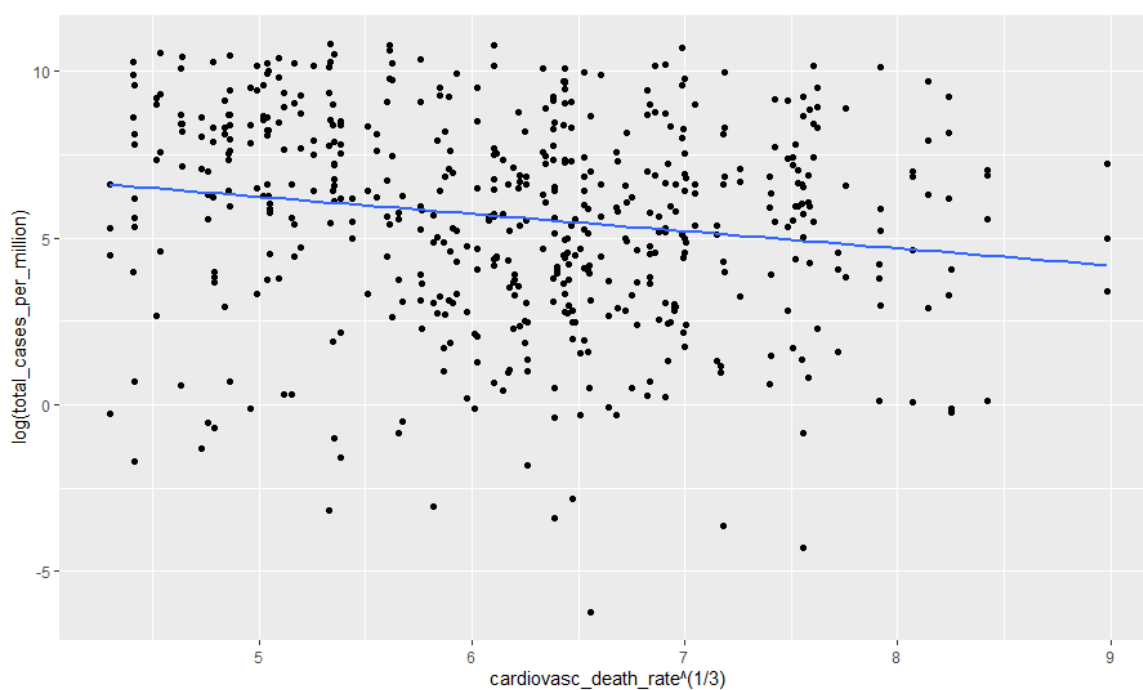
Na podstawie transformacji Boxa-Coxa otrzymujemy pierwiastek trzeciego stopnia ze zmiennej *cardiovasc_death_rate* oraz logarytm z *total_cases_per_million*.

$$\log(y_{total_cases}) = \beta_0 + \beta_{cardiovasc_death_rate} \sqrt[3]{X_{cardiovasc_death_rate}} + \varepsilon$$

	Model 1
(Intercept)	8.81***
	(0.82)
cardiovasc_death_rate [^] (1/3)	-0.52***
	(0.13)
R ²	0.03
Adj. R ²	0.03
Num. obs.	537

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Tabela 2.14: Wyniki dla modelu 7 po przekształceniu zmiennych



Rysunek 2.17: Wykres przedstawiający dopasowanie modelu liniowego po przekształceniu zmiennych

Źródło: Opracowanie własne

2.2.8. Model 8: zależność między liczbą zachorowań a powszechnością cukrzycy

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

$$y_{total_cases} = \beta_0 + X_{diabetes_prevalence}\beta_{diabetes_prevalence} + \varepsilon$$

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

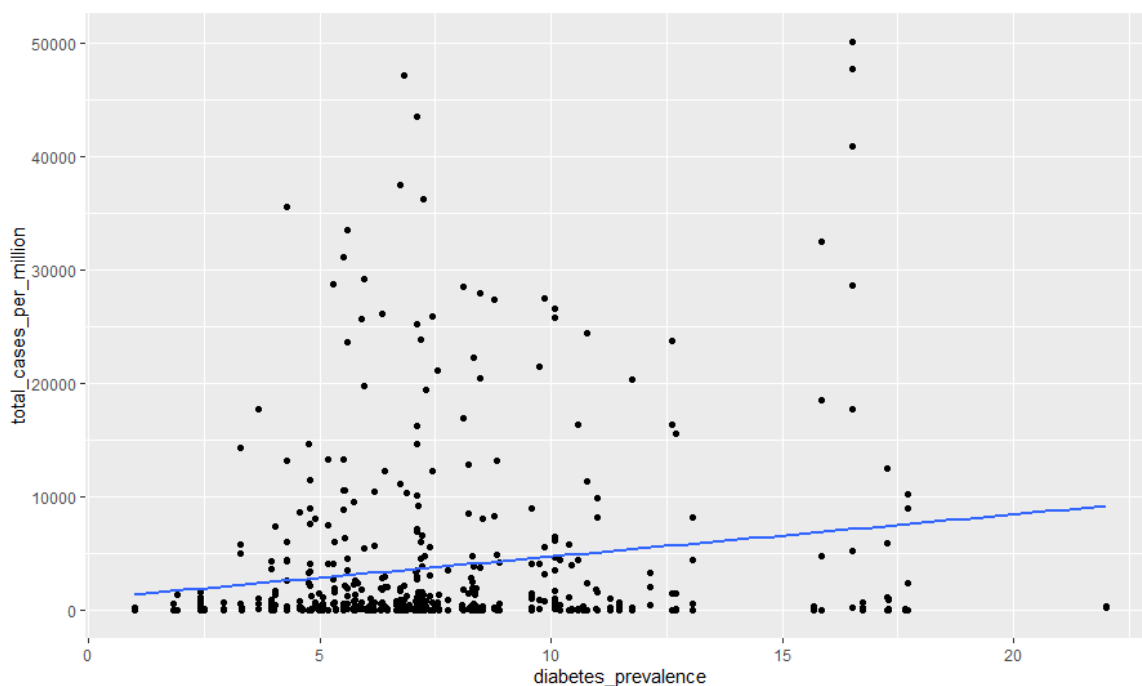
Otrzymujemy następujący wynik:

	Model 1
(Intercept)	991.31 (754.70)
diabetes_prevalence	372.06*** (91.52)
R ²	0.03
Adj. R ²	0.03
Num. obs.	537

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Tabela 2.15: Wyniki dla modelu 8

Źródło: Opracowanie własne



Rysunek 2.18: Wykres przedstawiający dopasowanie modelu liniowego

Źródło: Opracowanie własne

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.

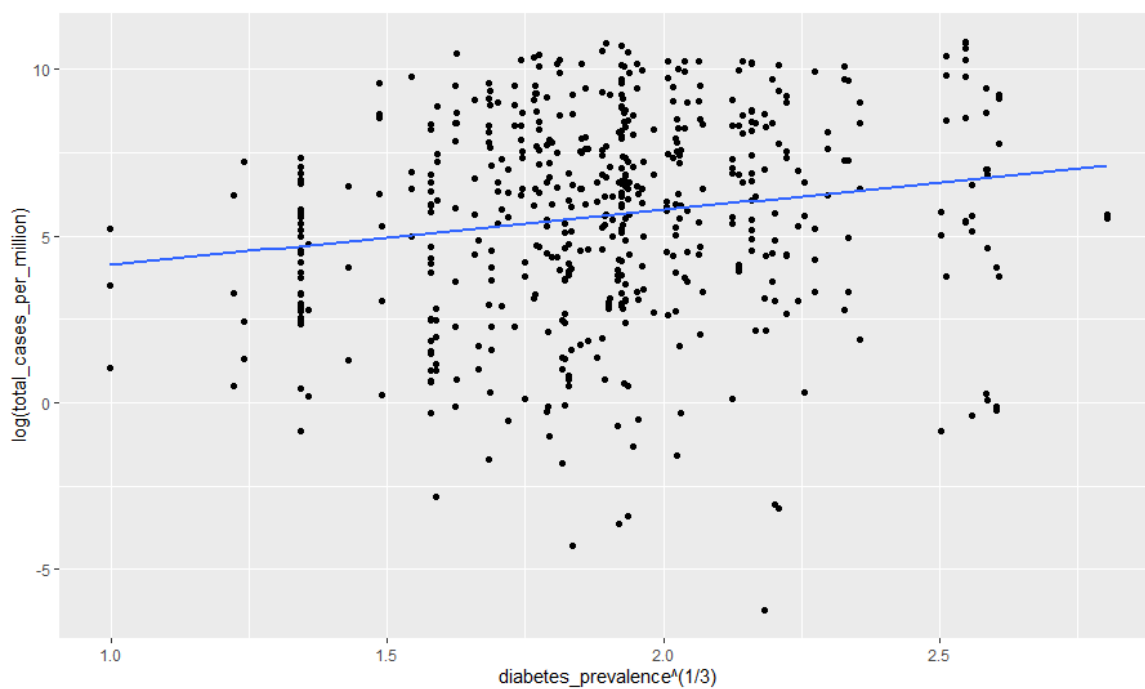
Na podstawie transformacji Boxa-Coxa otrzymujemy pierwiastek trzeciego stopnia ze zmiennej *diabetes_prevalence* oraz logarytm z *total_cases_per_million*.

$$\log(y_{total_cases}) = \beta_0 + \beta_{diabetes_prevalence} \sqrt[3]{X_{diabetes_prevalence}} + \varepsilon$$

	Model 1
(Intercept)	2.49** (0.79)
diabetes_prevalence [^] (1/3)	1.65*** (0.41)
R ²	0.03
Adj. R ²	0.03
Num. obs.	537

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Tabela 2.16: Wyniki dla modelu 8 po przekształceniu zmiennych



Rysunek 2.19: Wykres przedstawiający dopasowanie modelu liniowego z przekształceniem zmiennych

Źródło: Opracowanie własne

2.2.9. Model 9: zależność między liczbą zachorowań a odsetkiem osób żyjących w skrajnej biedzie

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:

$$y_{total_cases} = \beta_0 + X_{extreme_poverty}\beta_{extreme_poverty} + \varepsilon$$

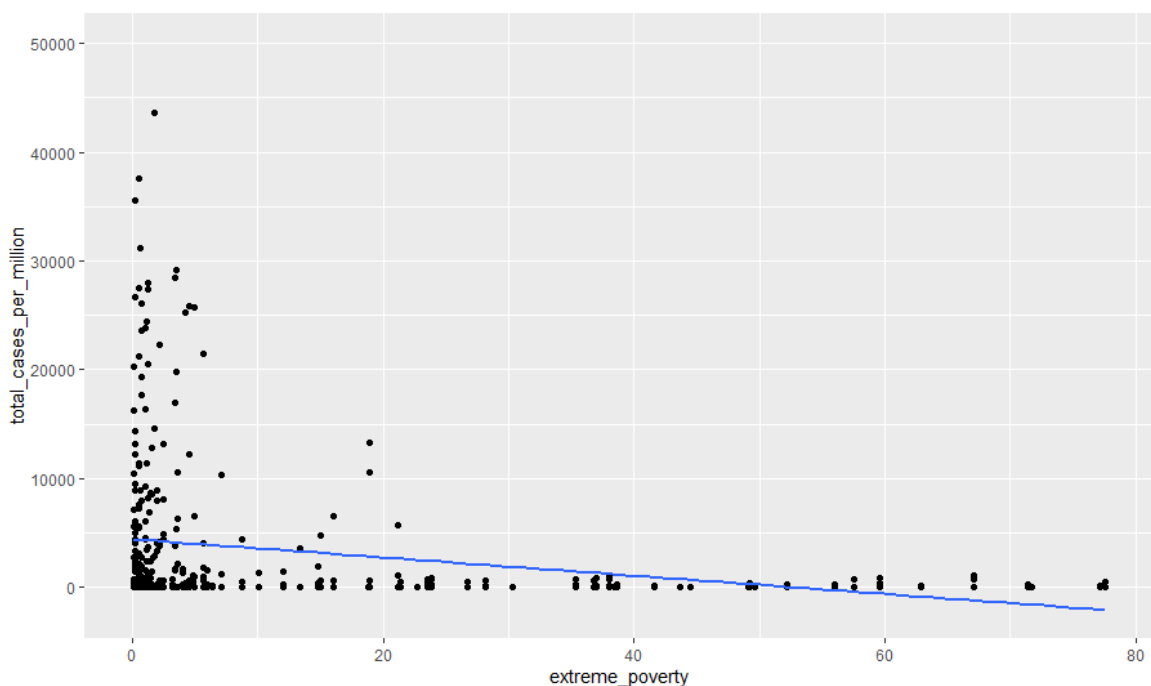
Model ten ma następujące podsumowanie:

	Model 1
(Intercept)	4433.28*** (406.29)
extreme_poverty	-83.75*** (17.02)
R ²	0.06
Adj. R ²	0.06
Num. obs.	389

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Tabela 2.17: Wyniki dla modelu 9

Źródło: Opracowanie własne



Rysunek 2.20: Wykres przedstawiający dopasowanie modelu liniowego

Źródło: Opracowanie własne

Czynnik *extreme_poverty* 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.

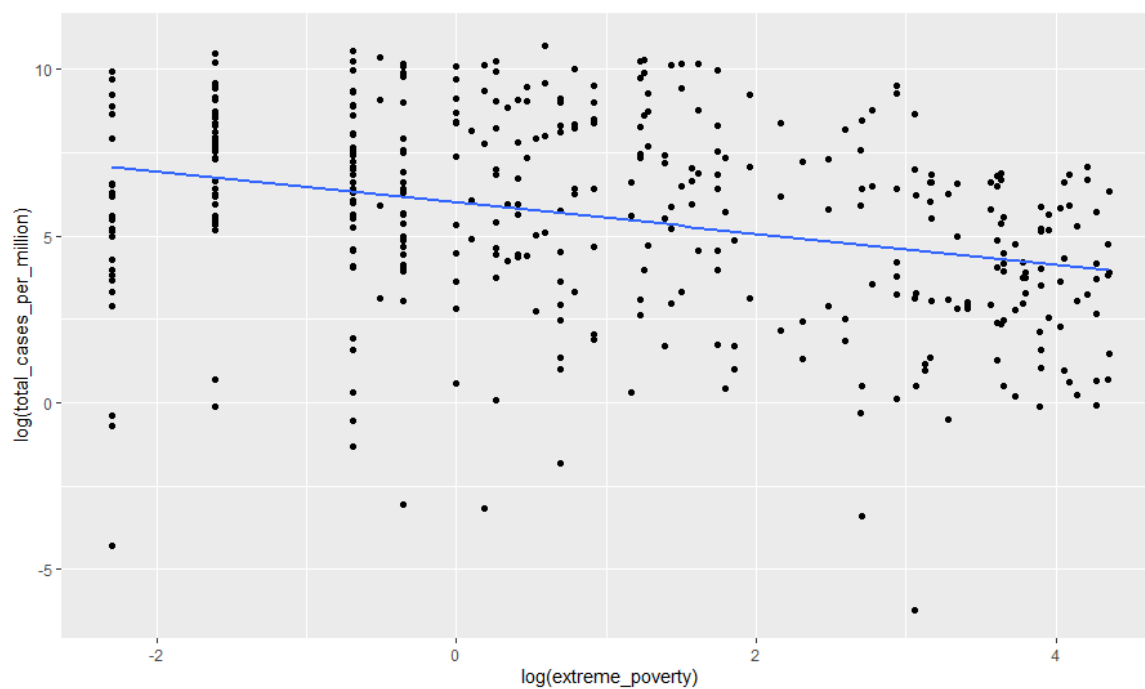
Z transformacji Boxa-Coxa otrzymujemy przekształcenie logarytmiczne dla obu zmiennych w modelu.

$$\log(y_{total_cases}) = \beta_0 + \beta_{extreme_poverty} \log(X_{extreme_poverty}) + \varepsilon$$

	Model 1
(Intercept)	6.00*** (0.16)
log(extreme_poverty)	−0.46*** (0.07)
R ²	0.09
Adj. R ²	0.09
Num. obs.	389

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Tabela 2.18: Wyniki dla modelu 9 z przekształceniem logarytmicznym obu zmiennych



Rysunek 2.21: Wykres przedstawiający dopasowanie modelu liniowego z przekształceniem zmiennych

Źródło: Opracowanie własne

2.2.10. Model 10: zależność między liczbą zachorowań a wysokością PKB na osobę

Hipoteza 10: Kraje o różnej wysokości PKB różnią się liczbą zachorowań.

W modelu dziesiątym pojawia się zależność liczby zachorowań od PKB na osobę:

$$y_{total_cases} = \beta_0 + X_{GDP}\beta_{GDP} + \varepsilon$$

Wyniki są następujące:

	Model 1
(Intercept)	1174.04*
	(462.01)
gdp_per_capita	0.13***
	(0.02)
R ²	0.11
Adj. R ²	0.11
Num. obs.	531

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

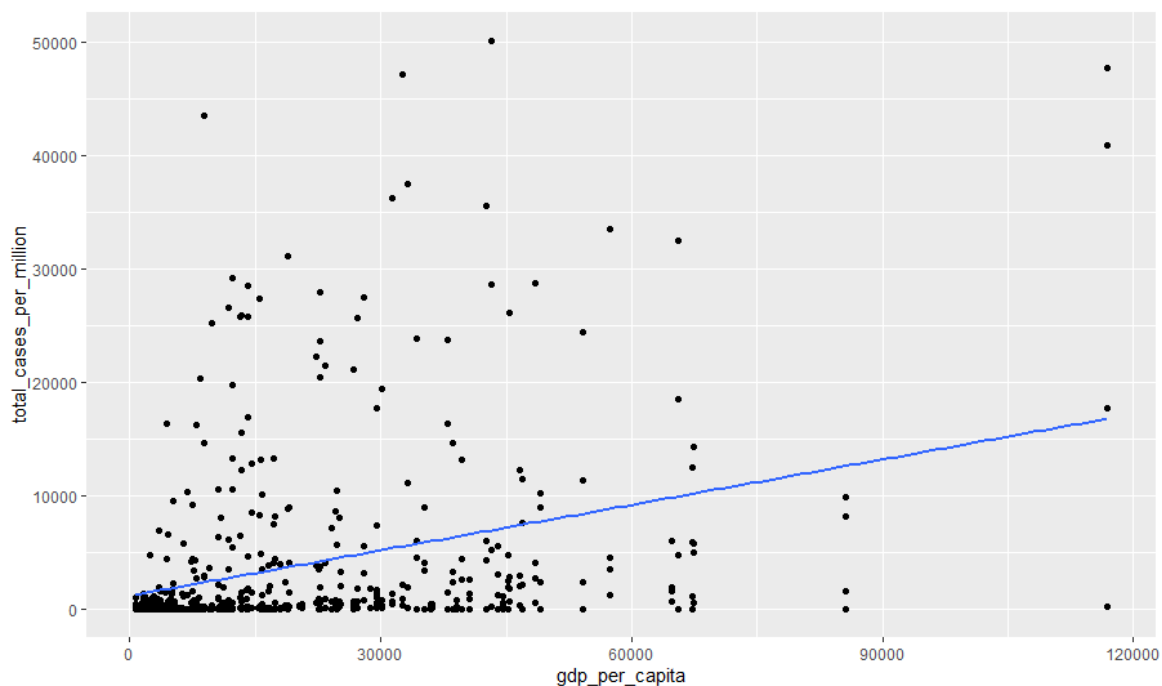
Tabela 2.19: Wyniki dla modelu 10

Źródło: Opracowanie własne

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

Na podstawie przekształcenia Boxa-Coxa otrzymujemy pierwiastek piątego stopnia z gdp_per_capita oraz logarytm z $total_cases_per_million$.

$$\log(y_{total_cases}) = \beta_0 + \beta_{GDP}\sqrt[5]{X_{GDP}} + \varepsilon$$



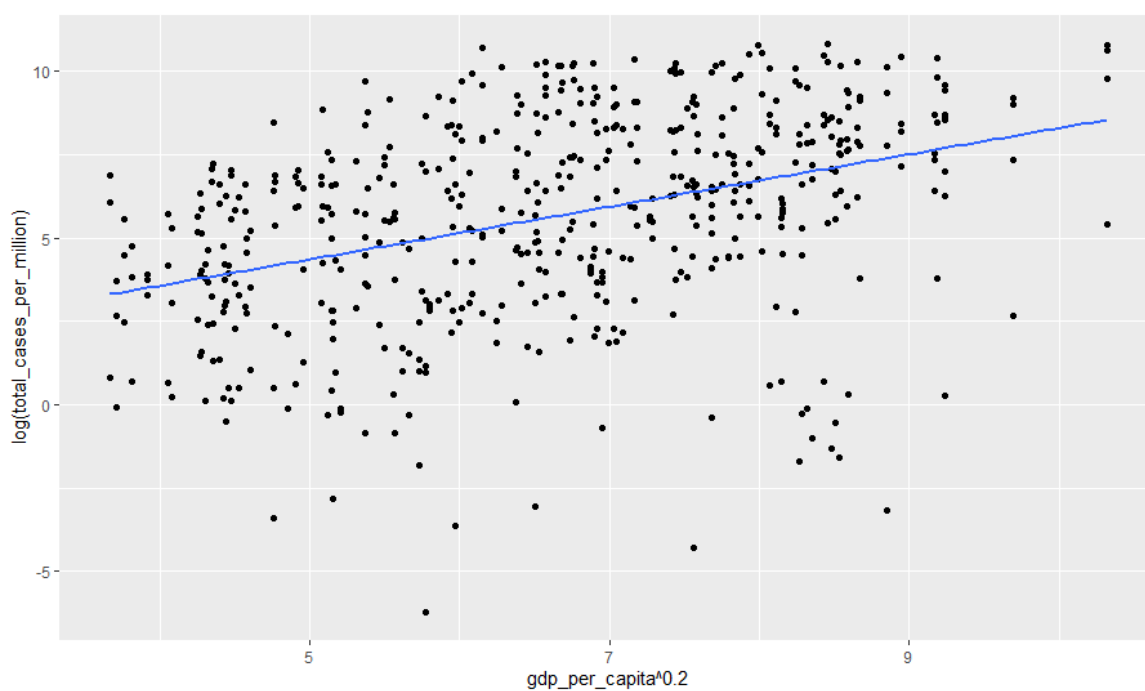
Rysunek 2.22: Wykres przedstawiający dopasowanie modelu liniowego

Źródło: Opracowanie własne

	Model 1
(Intercept)	0.42 (0.55)
$\text{gdp_per_capita}^{\wedge} 0.2$	0.79*** (0.08)
R^2	0.15
Adj. R^2	0.15
Num. obs.	531

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Tabela 2.20: Wyniki dla modelu 10 z przekształceniem zmiennych



Rysunek 2.23: Wykres przedstawiający dopasowanie modelu liniowego z przekształceniem zmiennych

Źródło: Opracowanie własne

Podsumowanie i wnioski

Podsumowanie wymaga solidnych poprawek, ale zostawić to na koniec.

W tabeli 2.21 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 2.21: 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 2.22: 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 2.22, 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.

Ułożyć alfabetycznie bibliografię, dowiedzieć się jak z linkami do stron

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)
- [9] <https://cran.r-project.org/web/packages/lme4/index.html> (dostęp: 03.01.2021)
- [10] <https://cran.r-project.org/web/packages/lmerTest/index.html> (dostęp: 03.01.2021)
- [11] <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/lm> (dostęp: 03.01.2021)
- [12] <https://www.rdocumentation.org/packages/car/versions/3.0-10/topics/powerTransform> (dostęp: 03.01.2021)

Spis rysunków

1.1	Rodzaje modeli mieszanych	15
2.1	Wykres przedstawiający zależność liczby zachorowań na milion mieszkańców w zależności od czasu z podziałem na kraje	19
2.2	Wykres przedstawiający zależność między liczbą zachorowań a czasem oszacowaną za pomocą Modelu 1	21
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	23
2.4	Wykres przedstawiający zależność między liczbą zachorowań a czasem oszacowaną za pomocą modelu, gdzie zarówno wyraz wolny, jak i współczynniki przy $time$ i $time^2$ zależą od kraju	25
2.5	Wykres przedstawiający zależność między liczbą zachorowań a liczbą wykonywanych testów w poszczególnych krajach	26
2.6	Wykres przedstawiający dopasowanie modelu random intercept	28
2.7	Wykres przedstawiający dopasowanie modelu random intercept and slope	30
2.8	Wykres przedstawiający dopasowanie modelu liniowego bez przekształcenia zmiennych	32
2.9	Wykres przedstawiający dopasowanie modelu liniowego z przekształceniem zmiennych	33
2.10	Wykres przedstawiający dopasowanie modelu liniowego	35
2.11	Wykres przedstawiający zależność liczby zachorowań od współczynnika siły obostrzeń, kolorem zaznaczone dane dla Polski	36
2.12	Wykres przedstawiający dopasowanie modelu random intercept	37
2.13	Wykres przedstawiający dopasowanie modelu random intercept and slope	38
2.14	Wykres przedstawiający dopasowanie modelu liniowego	41
2.15	Wykres przedstawiający dopasowanie modelu liniowego z przekształceniem zmiennych	42
2.16	Wykres przedstawiający dopasowanie modelu liniowego	44

2.17	Wykres przedstawiający dopasowanie modelu liniowego po przekształceniu zmiennych	45
2.18	Wykres przedstawiający dopasowanie modelu liniowego	47
2.19	Wykres przedstawiający dopasowanie modelu liniowego z przekształceniem zmiennych	48
2.20	Wykres przedstawiający dopasowanie modelu liniowego	50
2.21	Wykres przedstawiający dopasowanie modelu liniowego z przekształceniem zmiennych	51
2.22	Wykres przedstawiający dopasowanie modelu liniowego	53
2.23	Wykres przedstawiający dopasowanie modelu liniowego z przekształceniem zmiennych	54

Spis tabel

2.1	Wyniki dla modelu 1	20
2.2	Wyniki dla modelu 1	22
2.3	Wyniki dla modelu 1 z czynnikiem kwadratowym	24
2.4	Wyniki dla modelu 2	27
2.5	Wyniki dla modelu 2 z losowym współczynnikiem nachylenia	29
2.6	Wyniki dla modelu 3	31
2.7	Wyniki dla modelu 3 po przekształceniu zmiennych	33
2.8	Wyniki dla modelu 4	34
2.9	Wyniki dla modelu 5	37
2.10	Wyniki dla modelu 5 z losowym współczynnikiem nachylenia	38
2.11	Wyniki dla modelu 6	40
2.12	Wyniki dla modelu 6 po przekształceniu zmiennych	41
2.13	Wyniki dla modelu 7	43
2.14	Wyniki dla modelu 7 po przekształceniu zmiennych	45
2.15	Wyniki dla modelu 8	46
2.16	Wyniki dla modelu 8 po przekształceniu zmiennych	47
2.17	Wyniki dla modelu 9	49
2.18	Wyniki dla modelu 9 z przekształceniem logarytmicznym obu zmiennych	50
2.19	Wyniki dla modelu 10	52
2.20	Wyniki dla modelu 10 z przekształceniem zmiennych	53
2.21	Porównanie istotności i wpływu różnych czynników na liczbę zachorowań w przeciętnym kraju	55
2.22	Porównanie indeksów Akaike dla poszczególnych modeli	56

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.