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 2021

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:

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

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

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

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

gdzie ε to niezależne (tzn. $Cov(\varepsilon_i, \varepsilon_j) = 0$ dla $i \neq j$) zakłócenie losowe o rozkładzie normalnym ze średnią 0 i wariancją σ^2.

1.1.1. Metody estymacji parametrów modelu liniowego

Aby oszacować wartości parametrów modelu liniowego, wykorzystujemy poniższe metody estymacji:

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 mówi, że taki estymator jest najlepszym (w sensie najmniejszej wariancji) liniowym nieobciążonym estymatorem (ang. *BLUE, Best Linear Unbiased Estimator*) przy założeniach, że $E(\varepsilon_i) = 0$ i $Var(\varepsilon_i) = \sigma^2$ dla każdego i oraz $Cov(\varepsilon_i, \varepsilon_j) = 0$ dla $i \neq j$.
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

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

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

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

przeciwko hipotezie alternatywnej

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

1.1.3. Interpretacja parametrów modelu

W modelu postaci $y = \beta_0 + \beta_1 x$ dodatnia wartość β_1 oznacza, że wzrostowi x towarzyszy wzrost y, a ujemna wartość β_1, że wraz ze wzrostem x, maleje y [22]. Jeżeli model jest dobrze dopasowany do danych, to możemy go interpretować w ten sposób, że wzrost zmiennej x o 1, powoduje zmianę zmiennej y o β_1. Podobnie w przypadku modelu regresji wielorakiej

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

wzrost zmiennej x_i o jedną jednostkę, przy niezmienionych poziomach pozostałych zmiennych, powoduje zmianę wartości y o β_i.

Miarami jakości dopasowania modelu do danych są m. in. współczynnik determinacji R^2 oraz skorygowany współczynnik determinacji. Współczynnik determinacji informuje o tym, jaka część zmienności (wariancji) zmiennej zależnej w próbie jest wyjaśniona zmiennością modelu. Przyjmuje wartości z przedziału $[0, 1]$. Jeżeli w modelu występuje wyraz wolny, a do estymacji wykorzystano metodę najmniejszych kwadratów, to współczynnik determinacji można interpretować jako procent wariancji zmiennej zależnej, która jest wyjaśniana przez model (więc dopasowanie jest tym lepsze, im wartość R^2 jest bliższa jedności [24]). Współczynnik determinacji jest wyrażony wzorem:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2},$$

gdzie y_i - i-ta obserwacja zmiennej zależnej, $\hat{y}_i$ - oszacowanie i-tej wartości zmiennej zależnej na podstawie modelu, $\bar{y}$ - średnia arytmetyczna zaobserwowanych empirycznie wartości zmiennej zależnej.

W przypadku modeli zawierających więcej niż jedną zmienną, zdarza się, że dodanie do modelu nowej zmiennej podniesie współczynnik R^2, mimo że faktycznie nie będzie poprawiała dopasowania modelu. Dlatego można także korzystać z miary nazywanej skorygowanym współczynnikiem determinacji, który określony jest wzorem:

$$\tilde{R}^2 = 1 - \frac{n-1}{n-k-1}(1 - R^2),$$

gdzie R^2 - współczynnik determinacji, n - liczba obserwacji, k - liczba zmiennych w modelu (nie licząc wyrazu wolnego) [25]. Interpretacja $\tilde{R}^2$ jest taka sama, jak R^2, ale jeśli te dwie wartości znacznie się od siebie różnią, to warto interpretować raczej skorygowany współczynnik determinacji niż zwykłe R^2.

1.1.4. Transformacja zmiennych

Jeżeli obserwacje charakteryzują się wariancją, która rośnie lub maleje wraz ze wzrostem zmiennej niezależnej, to przydatna może być transformacja zmiennych. Często używana jest na przykład transformacja logarytmiczna, która jest łatwa w interpretacji (zmiany wartości zlogarytmowanej odpowiadają zmianom procentowym w oryginalnej skali). Przekształcenie takie warto stosować, kiedy zaobserwowane wartości zmiennej charakteryzują się silną asymetrią prawostronną. Jednakże, nie można go stosować, jeśli pojawiają się wartości niedodatnie. Jeżeli oryginalne obserwacje oznaczymy jako $x_1, x_2, \ldots, x_n$, to przekształcone obserwacje $w_1, w_2, \ldots, w_n$ będą takie, że $w_i = \log(x_i)$ dla $i \in \{1, 2, \ldots, n\}$ [6].

Innym rodzajem transformacji są transformacje potęgowe, np. pierwiastki kwadratowe lub sześcienne. Te przekształcenia nie zawsze są tak proste w interpretacji jak logarytmiczne. Zapisujemy je jako $w_i = x_i^p$, gdzie $i \in \{1, 2, \ldots, n\}$, a p to potęga, jaką przekształcamy obserwacje [6].

Transformacja Boxa-Coxa jest rodziną transformacji zawierającą zarówno przekształcenia logarytmiczne, jak i potęgowe. Przekształcenie Boxa-Coxa ma postać:

$$g^{(\lambda)}(X) = \begin{cases} \log(X), & \text{gdy } \lambda = 0 \\ \frac{X^\lambda - 1}{\lambda}, & \text{gdy } \lambda \neq 0 \end{cases} .[23]$$

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]. W przypadku efektu stałego interesuje nas jego wielkość (średnia), natomiast przy efektach losowych bierzemy pod uwagę jedynie fakt, że wprowadzona zmienna wnosi do modelu pewną zmienność (a dokładniej, pozwala odjąć tą zmienność od całkowitej

zmienności) i szacujemy wariancję lub odchylenie standardowe, a nie 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. W modelu mieszanym, wpływ konkretnego pacjenta będzie traktowany jako czynnik zakłócający.

1.2.1. Metody estymacji

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. metody największej wiarogodności.

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 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 wiarogodności. Przykładową metodą jest użycie algorytmu Newtona-Rapshona - 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. Ta właśnie metoda 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 maja często tak wielki rozmiar, że niemożliwe by było opracowanie na nich zwykłymi algorytmami [27]. 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 parametryzowana 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 twierdzenia Bayesa [28] mamy

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

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

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

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 i oznaczamy $PRSS$ (*Penalized Residual Sum of Squares*). Przez r_θ określimy $\min\limits_{w,\beta} r^2(\theta,\beta,w)$. Wartości minimalizujące $PRSS$ po w i β można znaleźć, rozwiązując układ równań liniowych.

$$\begin{bmatrix} X^TX & X^TZ\Lambda \\ \Lambda^TZ^TX & \Lambda^TZ^TZ\Lambda + I \end{bmatrix} \begin{bmatrix} \beta \\ w \end{bmatrix} = \begin{bmatrix} X^Ty \\ \Lambda^TZ^Ty \end{bmatrix} \tag{1.1}$$

To zadanie da się rozwiązać efektywnie nawet dla bardzo dużych q, używając rzadkiej dekompozycji Choleskiego [8]. Po lewej stronie równania przynajmniej macierz $\Lambda^TZ^TZ\Lambda + I$ 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 $\Lambda^TZ^TZ\Lambda + I_{q\times q}$, czyli jest to rzadka macierz trójkątna dolna spełniająca warunek

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

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

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

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 wiarogodności zależy jedynie od parametru θ, którego wymiar g jest zazwyczaj nieduży (jest to liczba komponentów wariancyjnych). Maksymalizację tak opisanej funkcji wiarogodności po przestrzeni parametrów o niewielkim wymiarze wykonuje się stamdardowymi 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 protą 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 wiarogodności 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 wiarogodności 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].

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 opiszemy, które z nich są najskuteczniejsze [2].

1. Iloraz wiarygodności(ang. *likelihood ratio*) - tworzymy dwa zagnieżdżone modele: model 0 - niezawierający 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 (ang. *Log Likelihood*). Tego testu nie można używać do modeli wyznaczonych metodą REML [2].
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 [2].
3. Test permutacyjny -można go stosować do dokładniejszego wyznaczenia wartości p dla efektu stałego. Funkcja wiarogodności może być użyta jako statystyka testowa. Rozkład statystyki testowej otrzymujemy wykonując permutacje na tej kolumnie macierzy X, która odpowiada interesującemu nas efektowi [1]. Dla każdej permutacji wyliczamy logarytm funkcji wiarogodności i sprawdzamy, ile z nich przekroczyło logarytm funkcji wiarogodności dla modelu z niepermutowanymi kolumnami. Testy permutacyjne mają wiele zalet, między innymi, nie muszą być spełnione założenia dotyczące rozkładu normalnego danych w próbce. Przy dostatecznie dużej liczbie permutacji, zwykle dają dokładne wartości p, niezależnie od wielkości próby [5].
4. Kryteria informacyjne - służą do wyboru najlepszego spośród modeli. Najpopularniejszym jest Kryterium Informacyjne Akaikego (ang. *Akaike Information Criterion, AIC*). Jest ono zdefiniowane następującym wzorem:

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

   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 [2]. 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 dla efektów stałych test Walda, mimo jego gorszych właściwości statystycznych [1].

1.2.3. 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 [3].

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

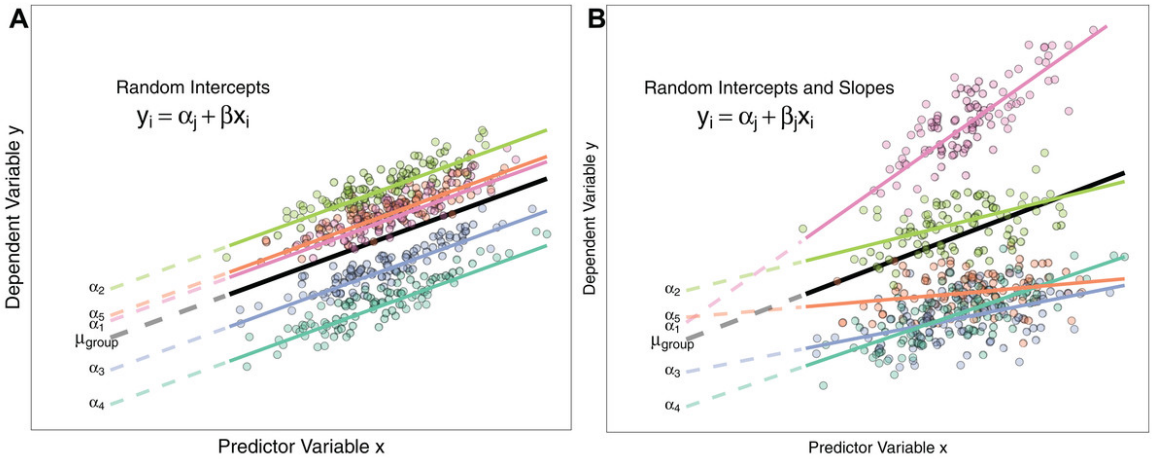
$$\widetilde{u} = DZ^TV^{-1}(y - X\beta),$$

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

1.2.4. 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 [12].

Można wyróżnić dwa główne rodzaje modeli mieszanych. Pierwszym typem jest model z losowym wyrazem wolnym (ang. *Random Intercept Model*). W takim modelu jedynie wyraz wolny różni się pomiędzy grupami obserwacji[9]. Przykład takiego modelu jest przedstawiony na rysunku 1.1 na wykresie *A*. Widać, że prosta regresji dla poszczególnych grup może być przesunięta w górę lub w dół w stosunku do średniej globalnej (μ_{group}). Innym rodzajem modelu mieszanego jest model, w którym także współczynniki przy niektórych zmiennych także różnią się pomiędzy grupami [9]. Przykładem takiego modelu jest model z losowym wyrazem wolnym i współczynnikiem nachylenia (ang. *Random Intercept and Slope*) pokazany na rysunku 1.1 na wykresie *B*. Oprócz zmian w wyrazie wolnym między grupami, widać, że prosta regresji może być nachylona do osi OX pod mniejszym lub większym kątem niż prosta regresji dla całej populacji (oznaczona kolorem czarnym.)



Rysunek 1.1: Rodzaje modeli mieszanych
Źródło: [16]

Warto dodać, że model mieszany dla danych dotyczących zmian w czasie (ang. *longitudinal data*) lub dla danych grupowanych (ang. *clustered data*) może być uzyskany z odpowiedniego modelu dla danych przekrojowych (ang. *cross-sectional data*) poprzez wprowadzenie do niego efektów losowych. Oznacza to, że oprócz omawianych

w tej pracy modeli liniowych z efektami stałymi i losowymi (ang. *Linear Mixed Effects Models, LME models*) wyróżniamy także nieliniowe modele mieszane (ang. *Nonlinear Mixed Effects Models, NLME models*) oraz uogólnione liniowe modele mieszane (ang. *Generalized Linear Mixed Model, GLMM*), które uzyskujemy odpowiednio z modeli nieliniowej regresji oraz uogólnionych modeli liniowych (GLM). Z kolei w analizie przeżycia pojawiają się mieszane modele przeżycia nazywane *frailty models* [9]. W części praktycznej tej pracy będziemy mieli do czynienia jedynie z modelami liniowymi z efektami stałymi i losowymi.

Rozdział 2

Modele rozwoju pandemii

2.1. Opis zbioru badawczego

Zbiór danych pochodzi z witryny internetowej Our World In Data [13], 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: [29]:

— liczba zachorowań (*total_cases_per_million*) - jest to liczba potwierdzonych przypadków koronawirusa w danym kraju od momentu rozpoczęcia epidemii. Zamiast wartośći 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 międzyludzkich itp. Może przyjmować wartości od 0 do 100, im większa wartość, tym silniejsze obostrzenia w danym kraju [14],

— gęstość zaludnienia (*population_density*),
— PKB danego kraju na osobę (*GDP_per_capita*) - Produkt Krajowy Brutto, przeliczony na hipotetyczną walutę dolara międzynarodowego [15],
— część społeczeństwa żyjąca w skrajnym ubóstwie (*extreme_poverty*) - stan na możliwie aktualny rok po 2010
— śmiertelność z powodu chorób układu sercowo-naczyniowego (*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
— wskaźnik rozwoju społecznego, tzw. HDI (*human_development_index*) - miara opisująca stopień rozwoju społeczno-ekonomicznego poszczególnych krajów, do którego pomiaru służą m. in. oczekiwana długość życia, średnia liczba lat edukacji otrzymanej przez mieszkańców w wieku co najmniej 25 lat, oczekiwana liczba lat edukacji, PKB na osobę [30]
— oczekiwana długość życia (*life_expectancy*) - stan na 2019 r.

Dane były zbierane do dnia 1 grudnia 2020 r.

2.2. Wyniki

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, 6, 9 i 12 miesiącach trwania epidemii. W przypadku modeli liniowych zastosowano konieczne przekształcenia zmiennych za pomocą transformacji Boxa-Coxa.

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 [17]. Pakiet `lmerTest` umożliwia obliczenie p-value dla parametrów modelu mieszanego [18]. W przypadku modeli liniowych użyto funkcji `lm` z pakietu bazowego R. Funkcja ta wykorzystuje metodę najmniejszych kwadratów estymacji parametrów modelu liniowego [19]. W modelach, gdzie dokonano transformacji zmiennych, została użyta funkcja `powerTransform` z pakietu `car`. Funkcja ta wykorzystuje transformację Boxa-Coxa [20].

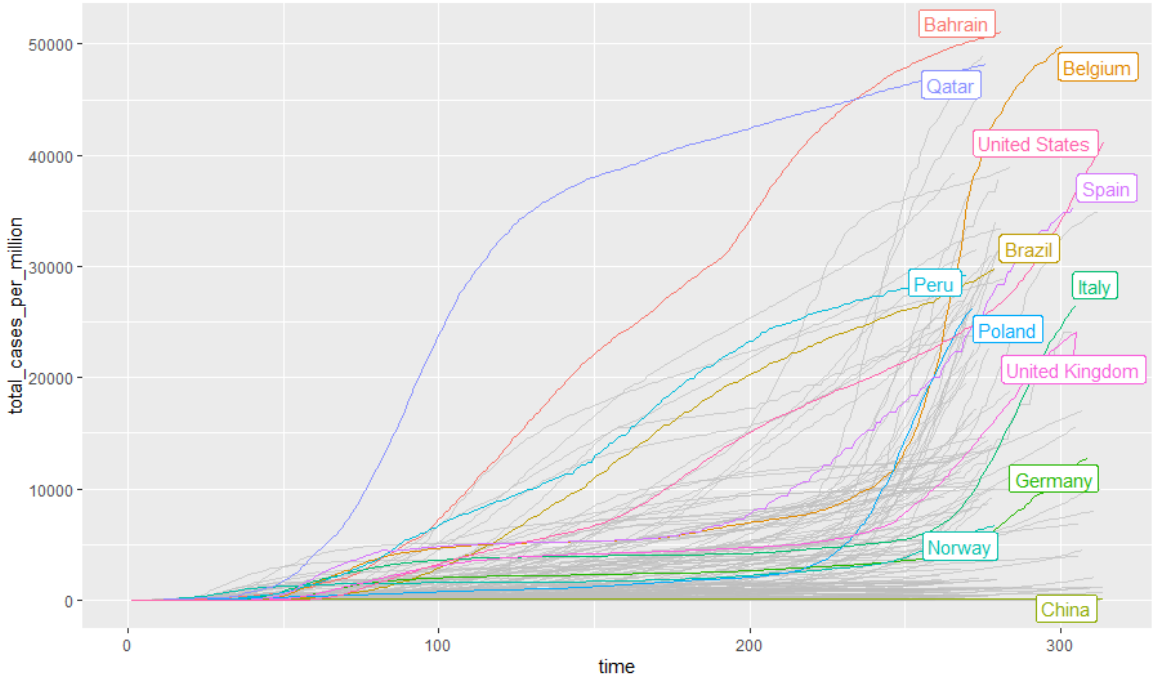
Dla uproszczenia zapisu, we wszystkich poniższych modelach nazwa *total_cases*

oznacza zmienną *total_cases_per_million*, a *total_tests* oznacza *total_tests_per_thousand*.

2.2.1. Model 1: zależność między liczbą zachorowań a czasem

Hipoteza 1: Czas ma istotny wpływ na liczbę zachorowań.

Na rysunku 2.1 jest pokazana zależność pomiędzy liczbą zachorowań a czasem (gdzie czas rozumiemy jako kolejne dni trwania epidemii).



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

Pierwszy model mieszany 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. Podsumowanie tego modelu jest przedstawione w tabeli 2.1.

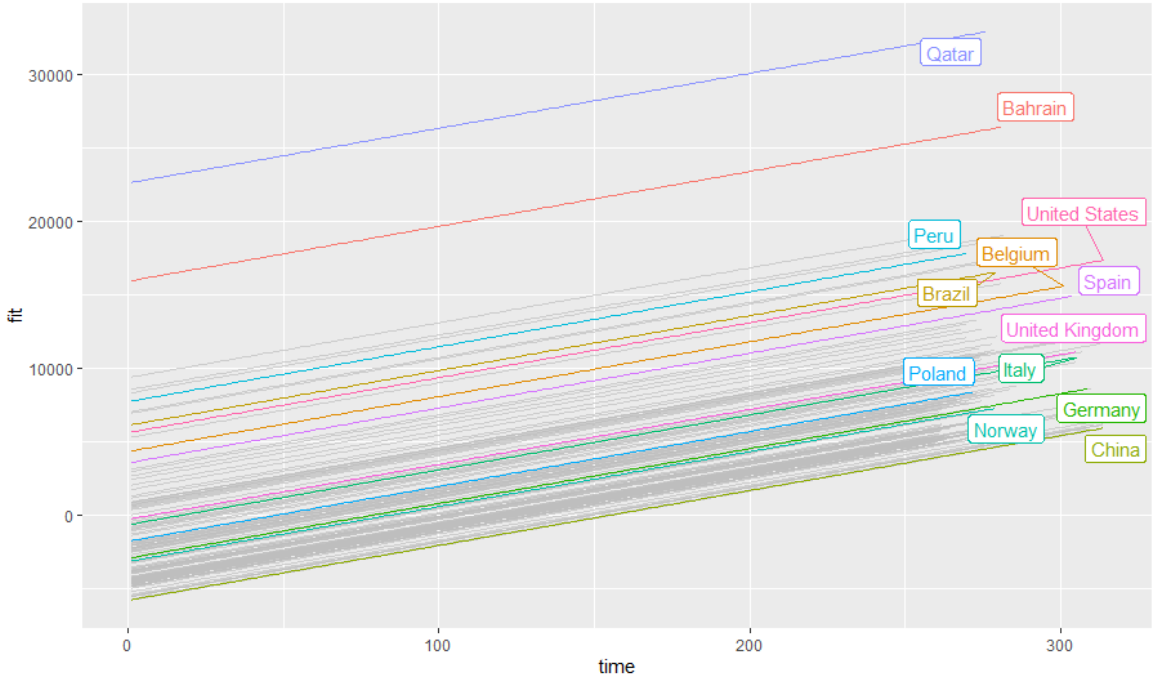
Widać, że efekt losowy jest odpowiedzialny za około połowę wariancji resztowej. Oznacza to, że zmienność liczby zachorowań dla danego kraju jest około dwukrotnie mniejsza niż zmienność liczby zachorowań dla różnych krajów.

Zarówno wyraz wolny, jak i współczynnik przy zmiennej `time`, są istotne statystycznie (p-value poniżej 0.001). Dodatkowo, β_{time} wynosi 37.31, jest dodatni, więc upływający czas sprawia, że liczba zachorowań rośnie.

	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



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:

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

$$+ Z_{time,location}u_{time,location} + \varepsilon$$

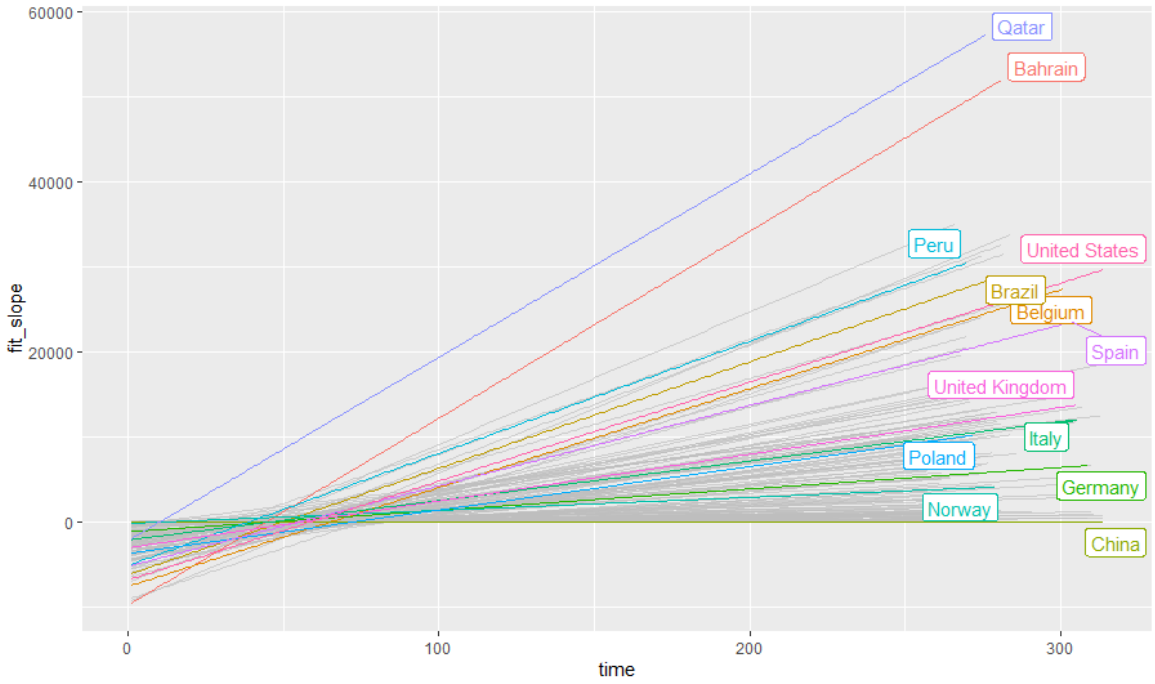
dla którego podsumowanie jest przedstawione w tabeli 2.2.

	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 z uwzględnieniem wpływu kraju na wyraz wolny i przesunięcie linii regresji
Źródło: Opracowanie własne

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



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

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

$$y_{total\_cases} = \beta_0 + X_{time}\beta_{time} + X_{time^2}\beta_{time^2} +$$

$$+ Z_{0,location}u_{0,location} + Z_{time,location}u_{time,location} +$$

$$Z_{time^2,location}u_{time^2,location} + \varepsilon$$

Wyniki dla tego modelu znajdują się w tabeli 2.3.

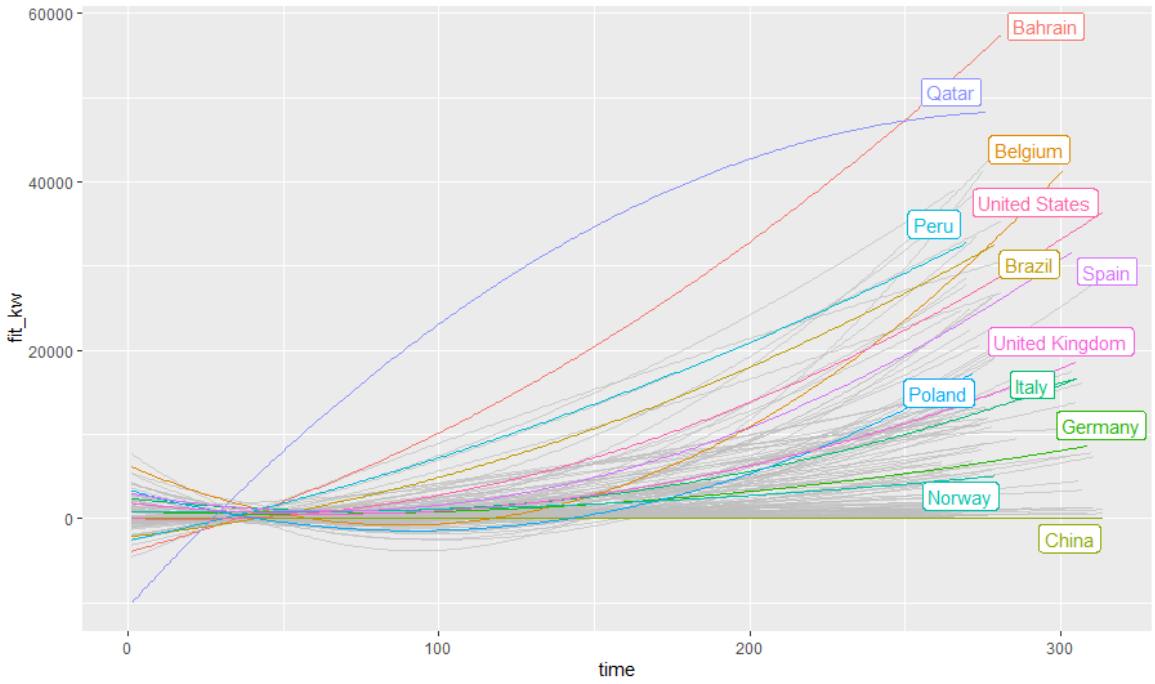
	Model 1
(Intercept)	383.14***
	(107.34)
time	−10.66
	(161.66)
time^2	0.16
	(0.21)
AIC	721041.58
BIC	721127.86
Log Likelihood	−360510.79
Num. obs.	41287
Num. groups: location	151
Var: location (Intercept)	1671050.31
Var: location time	3946212.95
Var: location I(time^2)	6.52
Cov: location (Intercept) time	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 wielomianowego mieszanego uwzględniającego wpływ kraju

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

Wracając do rysunku 2.1, można podejrzewać, że do danych będzie dobrze pasował model mieszany, w którym efekt stały czasu jest opisany wielomianem trzeciego stopnia, czyli

$$y_{total\_cases} = \beta_0 + X_{time}\beta_{time} + X_{time^2}\beta_{time^2} + X_{time^3}\beta_{time^3} +$$

$$+ Z_{0,location}u_{0,location} + Z_{time,location}u_{time,location} +$$

$$+ Z_{time^2,location}u_{time^2,location} + Z_{time^3,location}u_{time^3,location} + \varepsilon$$

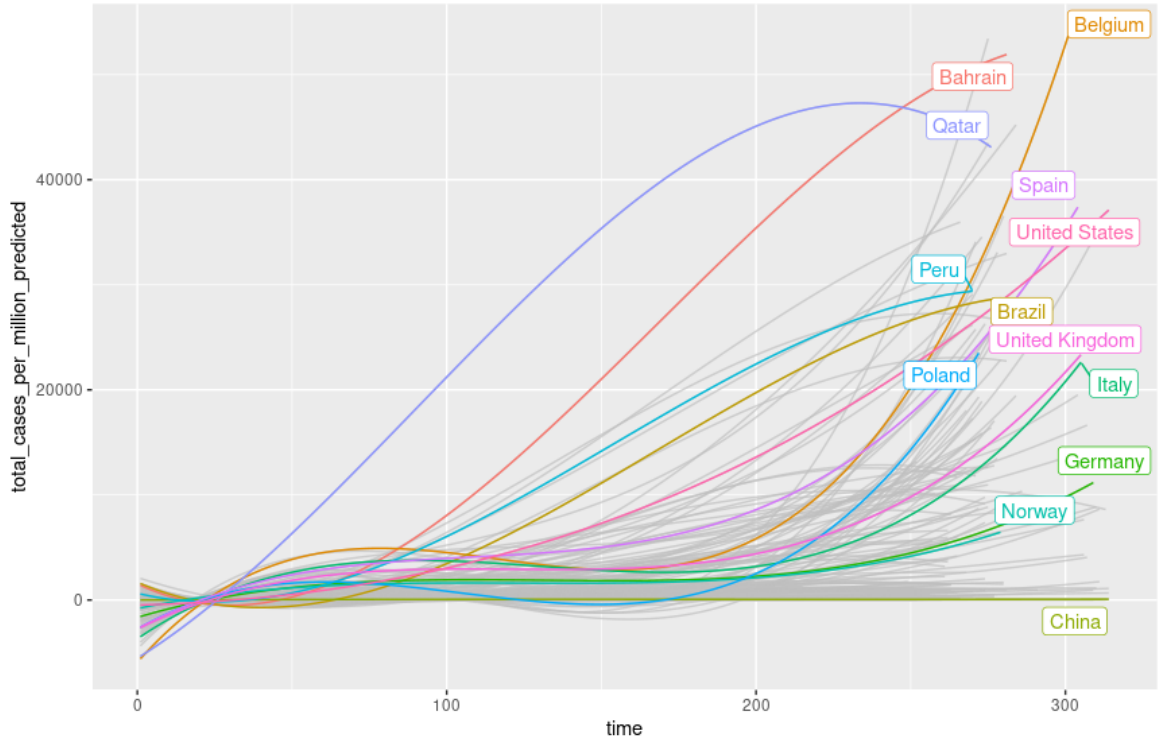
Podsumowanie tego modelu można odczytać z tabeli 2.4.

	Model 1
(Intercept)	3101.45***
	(320.37)
poly(time, 3)1	577597.25***
	(54939.11)
poly(time, 3)2	202728.96***
	(28196.37)
poly(time, 3)3	71609.14***
	(20647.78)
AIC	675017.16
BIC	675146.58
Log Likelihood	−337493.58
Num. obs.	41287
Num. groups: location	151
Var: location (Intercept)	15495330.05
Var: location poly(time, 3)1	455584118475.42
Var: location poly(time, 3)2	119854000709.20
Var: location poly(time, 3)3	64197657040.51
Cov: location (Intercept) poly(time, 3)1	2548609171.19
Cov: location (Intercept) poly(time, 3)2	329435355.67
Cov: location (Intercept) poly(time, 3)3	−105891364.41
Cov: location poly(time, 3)1 poly(time, 3)2	113956385885.80
Cov: location poly(time, 3)1 poly(time, 3)3	10930922943.81
Cov: location poly(time, 3)2 poly(time, 3)3	68483035447.11
Var: Residual	681134.37

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

Tabela 2.4: Wyniki dla modelu mieszanego wielomianowego stopnia trzeciego

Wszystkie trzy potęgi zmiennej *time* są istotne statystycznie (p-value poniżej 0.001), w dodatku przy każdej z nich współczynnik jest dodatni. Wykresy dopasowane na podstawie tego modelu są przedstawione na rysunku 2.5.

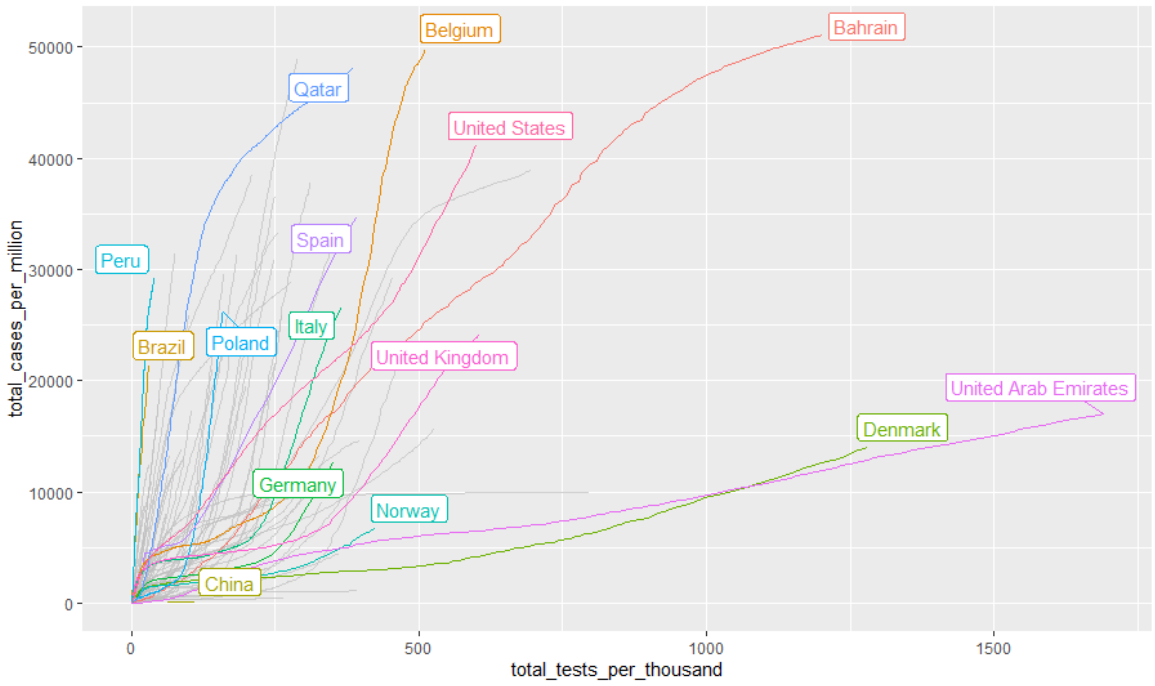


Rysunek 2.5: Model mieszany, gdzie zależność liczby zachorowań od czasu jest opisana wielomianem trzeciego stopnia

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.

Na rysunku 2.6 przedstawiona jest zależność liczby zachorowań od liczby wykonywanych testów.



Rysunek 2.6: Wykres przedstawiający zależność między liczbą zachorowań a liczbą wykonywanych testów w poszczególnych krajach
Źródło: Opracowanie własne

Model mieszany ma postać:

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

Badamy tutaj, czy liczba wykonywanych testów (w przeliczeniu na 1000 mieszkańców) ma wpływ na liczbę zachorowań. Dla tego modelu otrzymujemy wyniki przedstawione w tabeli 2.5:

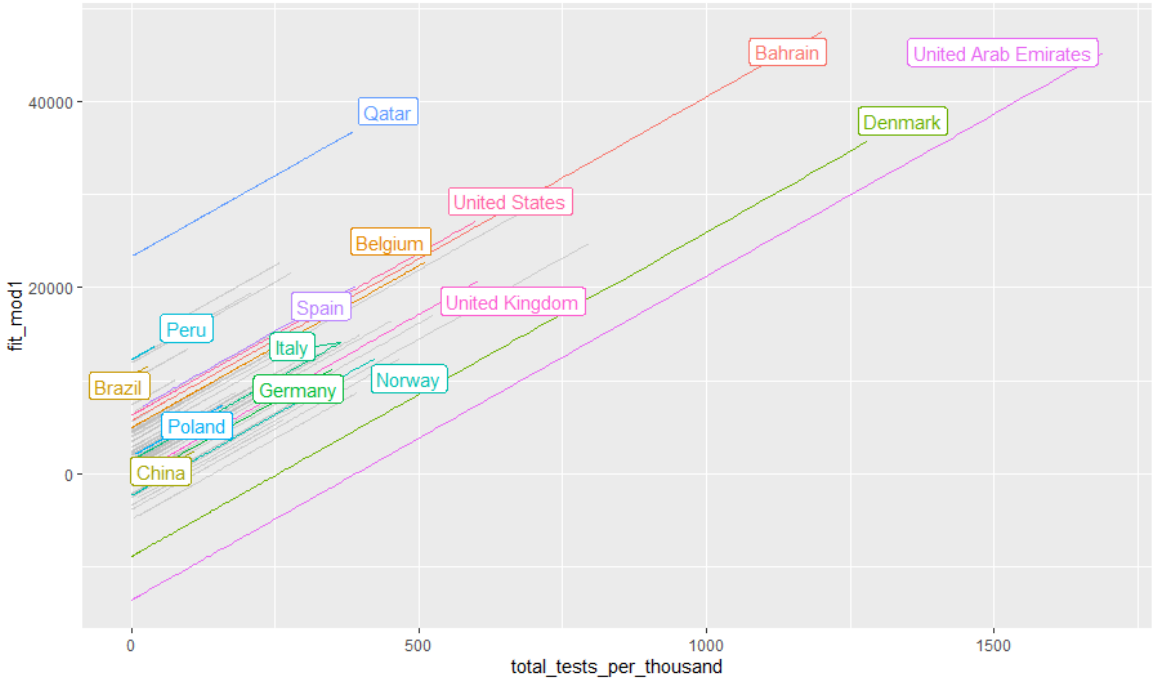
Widać po pierwsze, że efekt losowy jest odpowiedzialny za ponad połowę zmienności resztowej modelu. Po drugie, widać, że efekt stały liczby wykonywanych testów jest istotny statystycznie (p-value poniżej 0.001), i ma wpływ stymulujący na liczbę zachorowań (współczynnik β_{total_tests} wynosi 34.85). Dopasowanie tego modelu jest przedstawione na rysunku 2.7.

	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 2.5: Wyniki dla modelu 2

Źródło: Opracowanie własne



Rysunek 2.7: Wykres przedstawiający dopasowanie modelu mieszanego do zależności pomiędzy liczbą zachorowań a liczbą wykonywanych testów

Źródło: Opracowanie własne

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

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

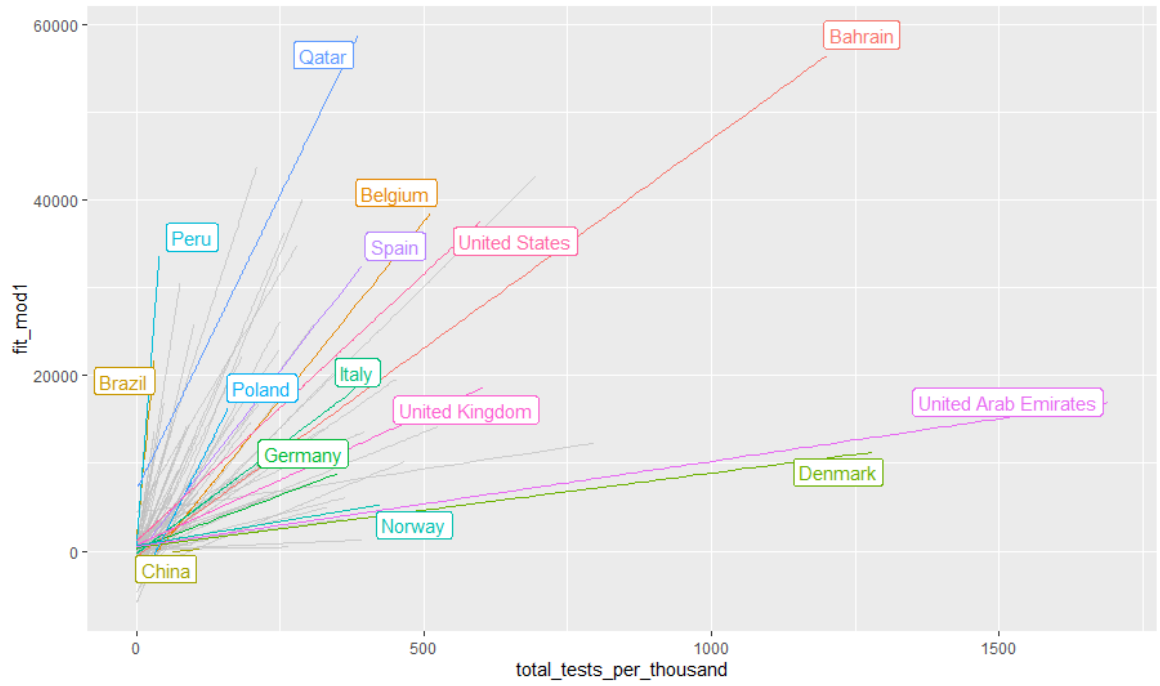
Dla takiego modelu otrzymujemy wyniki przedstawione w tabeli 2.6.

	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 2.6: Wyniki dla modelu 2 z losowym współczynnikiem nachylenia

Proste regresji dopasowane z tego modelu są widoczne na wykresie 2.8.



Rysunek 2.8: Wykres przedstawiający dopasowanie modelu typu *Random Intercept and Slope* do zależności pomiędzy liczbą zachorowań a liczbą wykonywanych testów
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, który prezentuje zależność liczby zachorowań od oczekiwanej długości życia w danym kraju.

Z transformacji Boxa-Coxa otrzymujemy następujące potęgi dla zmiennych: 3.72 dla $life\_expectancy$ oraz 0.09 dla $total\_cases\_per\_million$. W przybliżeniu przyjmiemy czwartą potęgę dla pierwszej zmiennej, a dla drugiej logarytm.

$$log(y_{total\_cases}) = \beta_0 + \beta_{life\_expectancy} X^4_{life\_expectancy} + \varepsilon$$

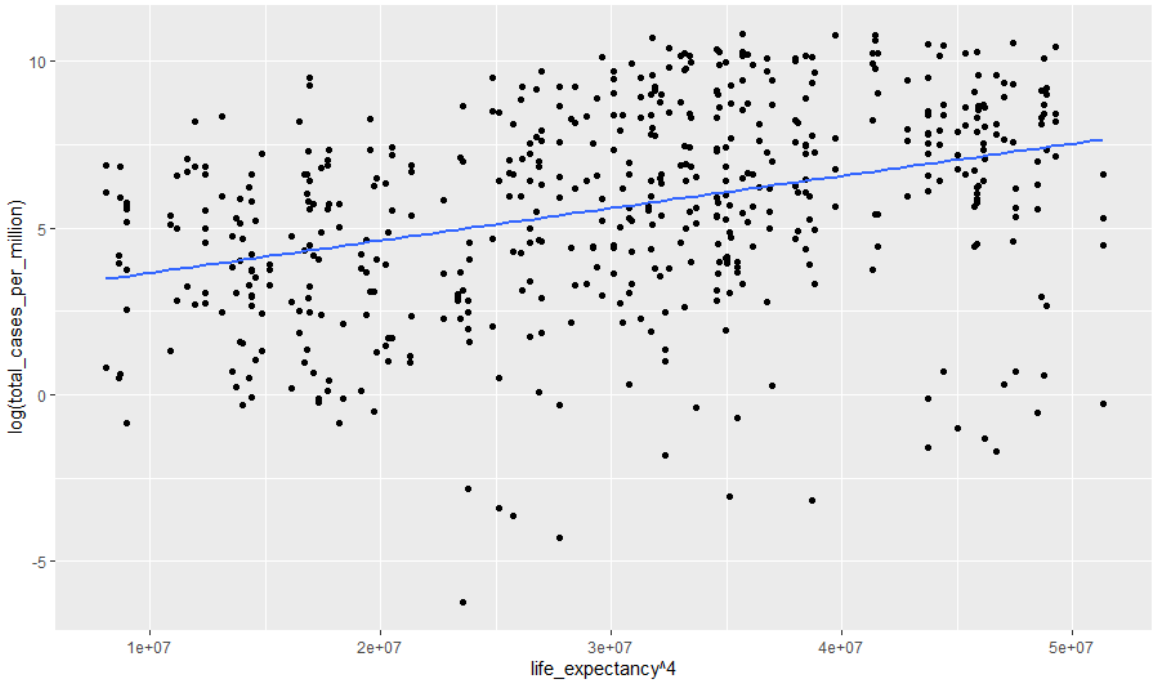
Podsumowanie tego modelu jest przedstawione w tabeli 2.7.

Efekt $life\_expectancy$ jest istotny statystycznie (p-value poniżej 0.001). Współczynnik $\beta_{life\_expectancy}$ wynosi około $8.54 \cdot 10^{-8}$. Jest on dodatni, co oznacza, że ze wzrostem oczekiwanej długości życia rośnie liczba zachorowań na COVID-19. Niska wartość tego współczynnika może wynikać z tego, że zmienna $life\_expectancy$ jest podniesiona do potęgi 4, więc obserwacje mają duże wartości. Współczynnik R^2 wynosi około 0.1, więc jest niski. Dopasowanie modelu liniowego do danych po przekształceniu jest przedstawione na rysunku 2.9.

	Model 1
(Intercept)	3.22^{***}
	(0.04)
life_expectancy^4	$8.54 \cdot 10^{-8}$ ***
	(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 do zależności pomiędzy liczbą zachorowań a oczekiwaną długością życia
Źródło: Opracowanie własne

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. Stosujemy przekształcenie logarytmiczne obu zmiennych.

$$\log(y_{total\_cases}) = \beta_0 + log(X_{population\_density})\beta_{population\_density} + \varepsilon$$

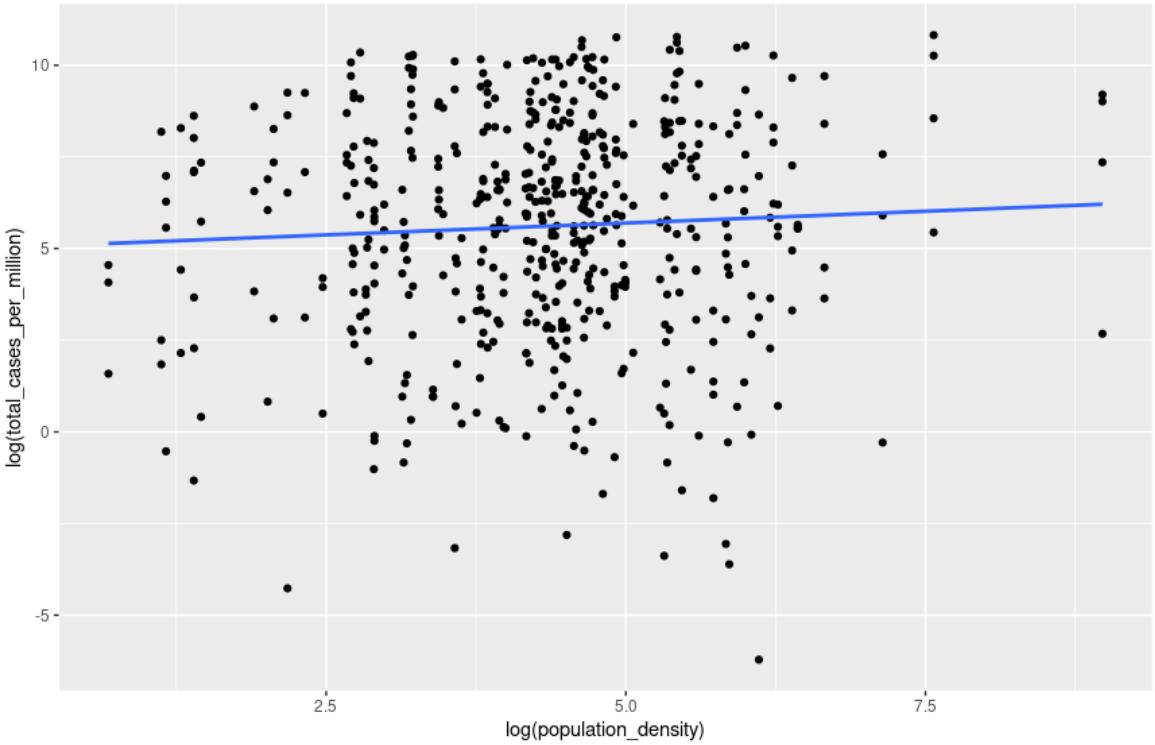
Wykres rozrzutu obu zmiennych po przekształceniu znajduje się na rysunku 2.10. Wyniki są przedstawione w tabeli 2.8.

	Model 1
(Intercept)	5.05***
	(0.46)
log(population_density)	0.13
	(0.10)
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 (p-value powyżej 0.05), a R^2 jest bliskie 0. Nie można mówić o istotnym związku pomiędzy gęstością zaludnienia a liczbą zachorowań.

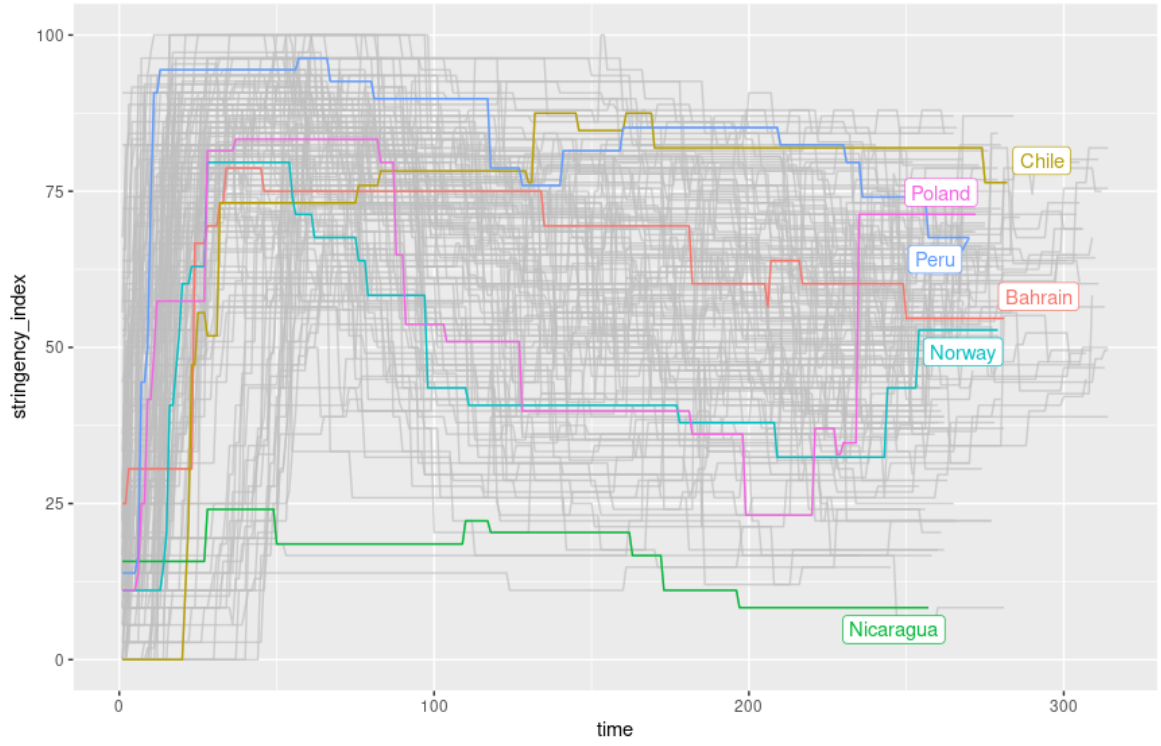


Rysunek 2.10: Wykres przedstawiający dopasowanie modelu liniowego do zależności pomiędzy liczbą zachorowań a gęstością zaludnienia po przekształceniu logarytmicznym
Źródło: Opracowanie własne

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

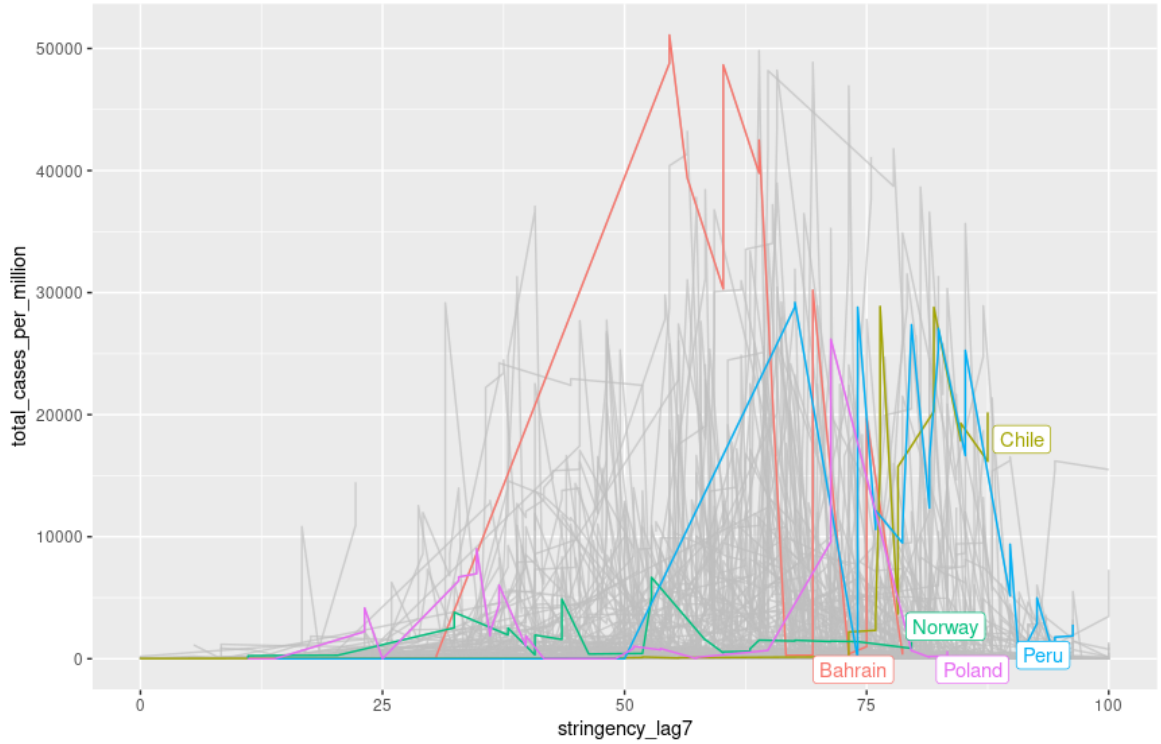
Hipoteza 5: Kraje różniące się siłą obostrzeń mają istotne różnice w liczbie zachorowań.

Siła obostrzeń jest mierzona za pomocą zmiennej *stringency_index*. Zarówno wartość średnia, jak i zmienność *stringency_index* bardzo się różnią pomiędzy krajami. Na rysunku 2.11 przedstawiono przebieg zmienności *stringency_index* w czasie. Jak widać, są kraje jak np. Nikaragua, gdzie siła obostrzeń jest na niskim poziomie i ma małą zmienność. Mamy też kraje takie jak Chile lub Peru, w których obostrzenia wzrastały silnie na początku pandemii, a potem zmieniały się już mniej znacznie. W Norwegii obostrzenia początkowo były wysokie, a potem znacząco spadły. W Polsce początkowo poziom siły obostrzeń zmieniał się podobnie jak w Norwegii, ale pod koniec badanego przedziału czasowego znacząco wzrósł.



Rysunek 2.11: Zmiany siły obostrzeń w poszczególnych krajach
Źródło: Opracowanie własne

Przed budową modelu, zostały przeanalizowane korelacje pomiędzy liczbą zachorowań, a zmienną *stringency_index* bez opóźnienia oraz z opóźnieniem kolejno 7, 14, 21, 28, 35 i 42 dni. Najwyższa korelacja wystąpiła przy opóźnieniu o 7 dni, więc tak opóźnionej zmiennej użyjemy do budowy modelu mieszanego. Rysunek 2.12 jest wykresem zależności liczby zachorowań od siły obostrzeń opóźnionej o 7 dni.



Rysunek 2.12: Zależność liczby zachorowań od siły obostrzeń opóźnionej o 7 dni
Źródło: Opracowanie własne

Na początku zbudujemy model postaci

$$y_{total\_cases} = \beta_0 + \beta_{stringency\_lag7} X_{stringency\_lag7} + Z_{location} u_{location} + \varepsilon$$

Wyniki dla tego modelu są przedstawione w tabeli 2.9. Zmienna *stringency_lag7* jest istotna statystycznie, choć jest na granicy istotności, gdyż p-value wynosi około 0.0464. Współczynnik $\beta_{stringency\_lag7}$ ma wartość -2.89, jest ujemny, co oznacza, że im wyższą wartość przyjmuje *stringency_lag7*, tym mniejsza liczba zachorowań.

	Model 1
(Intercept)	3377.31***
	(364.95)
stringency_lag7	−2.89*
	(1.45)
AIC	780586.43
BIC	780620.73
Log Likelihood	−390289.21
Num. obs.	39190
Num. groups: location	147
Var: location (Intercept)	18340804.07
Var: Residual	25666097.17

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

Tabela 2.9: Wyniki dla modelu 5 z losowym wyrazem wolnym

Spróbujemy także zbudować model, gdzie nachylenie prostej regresji także będzie zależało od kraju

$$y_{total\_cases} = \beta_0 + \beta_{stringency\_lag7} X_{stringency\_lag7} + Z_{location} u_{location} +$$

$$+ Z_{stringency\_lag7,location} u_{stringency\_lag7,location} + \varepsilon$$

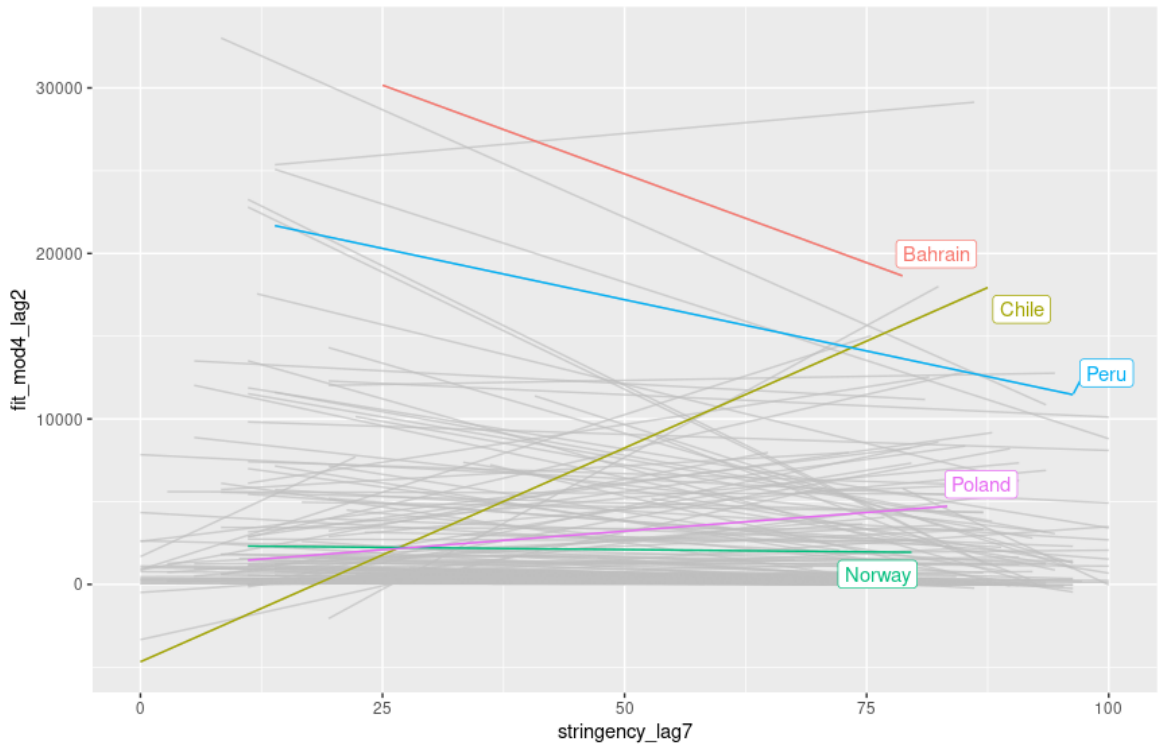
Podsumowanie tego modelu można odczytać z tabeli 2.10.

	Model 1
(Intercept)	4570.86***
	(403.65)
stringency_lag7	−14.86*
	(6.13)
AIC	777617.12
BIC	777668.58
Log Likelihood	−388802.56
Num. obs.	39190
Num. groups: location	147
Var: location (Intercept)	21876819.86
Var: location stringency_lag7	5020.82
Cov: location (Intercept) stringency_lag7	−200474.29
Var: Residual	23567356.48

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

Tabela 2.10: Wyniki dla modelu 5, w którym wyraz wolny oraz $\beta_{stringency\_lag7}$ zależą od kraju

Współczynnik przy *stringency_lag*7 jest istotny statystycznie (p-value około 0.0162) i wynosi −14.86, co oznacza, że w przeciętnym kraju siła obostrzeń wpływa na zmniejszenie liczby zachorowań po 7 dniach. Proste regresji dopasowane za pomocą tego modelu widać na rysunku 2.13. Z tego wykresu widać, że w niektórych krajach (np. Bahrain i Peru) zależność faktycznie jest malejąca. Z kolei np. w Norwegii zależność jest prawie nieistniejąca. Mamy też kraje takie jak Polska lub Chile, gdzie silniejszym obostrzeniom towarzyszy wyższa liczba zachorowań.



Rysunek 2.13: Proste regresji dopasowane do zależności między opóźnioną o 7 dni siłą obostrzeń a liczbą zachorowań
Źródło: Opracowanie własne

Ponieważ z wykresu 2.12 widać, że zależność między opóźnioną siłą obostrzeń a liczbą zachorowań raczej nie jest liniowa, to dopasujemy model mieszany z drugą potęgą zmiennej *stringency_lag7*

$$y_{total\_cases} = \beta_0 + \beta_{stringency\_lag7} X_{stringency\_lag7} + \beta_{stringency\_lag7^2} X_{stringency\_lag7^2} +$$

$$+ Z_{location} u_{location} + Z_{stringency\_lag7,time} u_{stringency\_lag7,time} +$$

$$+ Z_{stringency\_lag7^2,location} u + _{stringency\_lag7^2,location} + \varepsilon$$

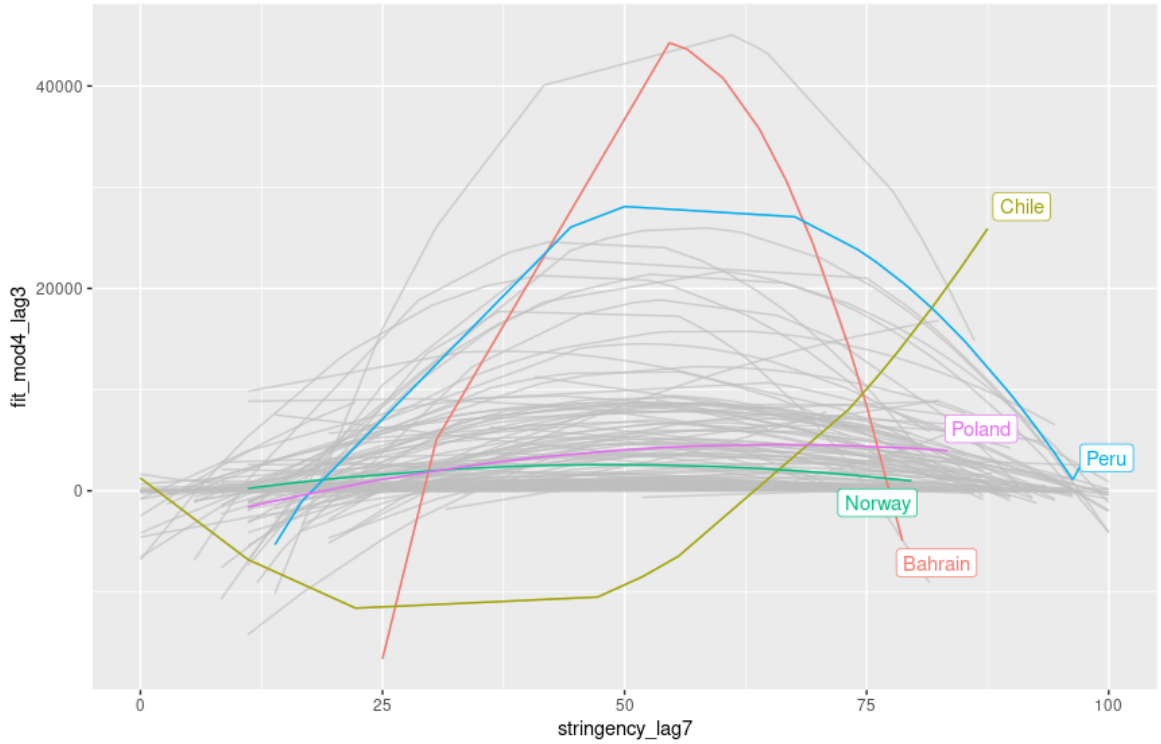
Podsumowanie tego modelu widać w tabeli 2.11. Współczynnik przy drugiej potędze *stringency_lag7* jest istotny statystycznie (p-value poniżej 0.001) i wynosi -402699.65. Efekt pierwszej potęgi *stringency_lag7* nie jest istotny, ale nie możemy go usunąć z modelu, ponieważ jest istotny efekt wyższego rzędu.

	Model 1
(Intercept)	3163.02***
	(378.38)
poly(stringency_lag7, 2)1	−21460.51
	(37503.93)
poly(stringency_lag7, 2)2	−402699.65***
	(77604.12)
AIC	763894.16
BIC	763979.92
Log Likelihood	−381937.08
Num. obs.	39190
Num. groups: location	147
Var: location (Intercept)	20524092.27
Var: location poly(stringency_lag7, 2)1	188332308392.00
Var: location poly(stringency_lag7, 2)2	869155957279.70
Cov: location (Intercept) poly(stringency_lag7, 2)1	168631325.28
Cov: location (Intercept) poly(stringency_lag7, 2)2	−1952065450.06
Cov: location poly(stringency_lag7, 2)1 poly(stringency_lag7, 2)2	125443897377.66
Var: Residual	16253620.55

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

Tabela 2.11: Wyniki dla modelu 5 z drugą potęgą zmiennej *stringensy_lag*7

Wykres zależności liczby zachorowań od *stringency_lag*7 i drugiej potęgi tej zmiennej widać na rysunku 2.14. W większości krajów współczynnik przy drugiej potędze *stringency_lag*7 rzeczywiście jest ujemny, co oznacza, że po przekroczeniu pewnego punktu (a dokładniej wierzchołka paraboli) liczba zachorowań zaczyna spadać. Wyjątkiem jest tu np. Chile, gdzie zależność jest odwrotna.



Rysunek 2.14: Proste regresji dopasowane do zależności między pierwszą i drugą potęgą opóźnionej o 7 dni siły obostrzeń a liczbą zachorowań
Źródło: Opracowanie własne

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 bada zależność liczby zachorowań od wskaźnika rozwoju społecznego. Z przekształcenia Boxa-Coxa otrzymujemy potęgę 2 dla zmiennej HDI oraz logarytm dla $total\_cases\_per\_million$.

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

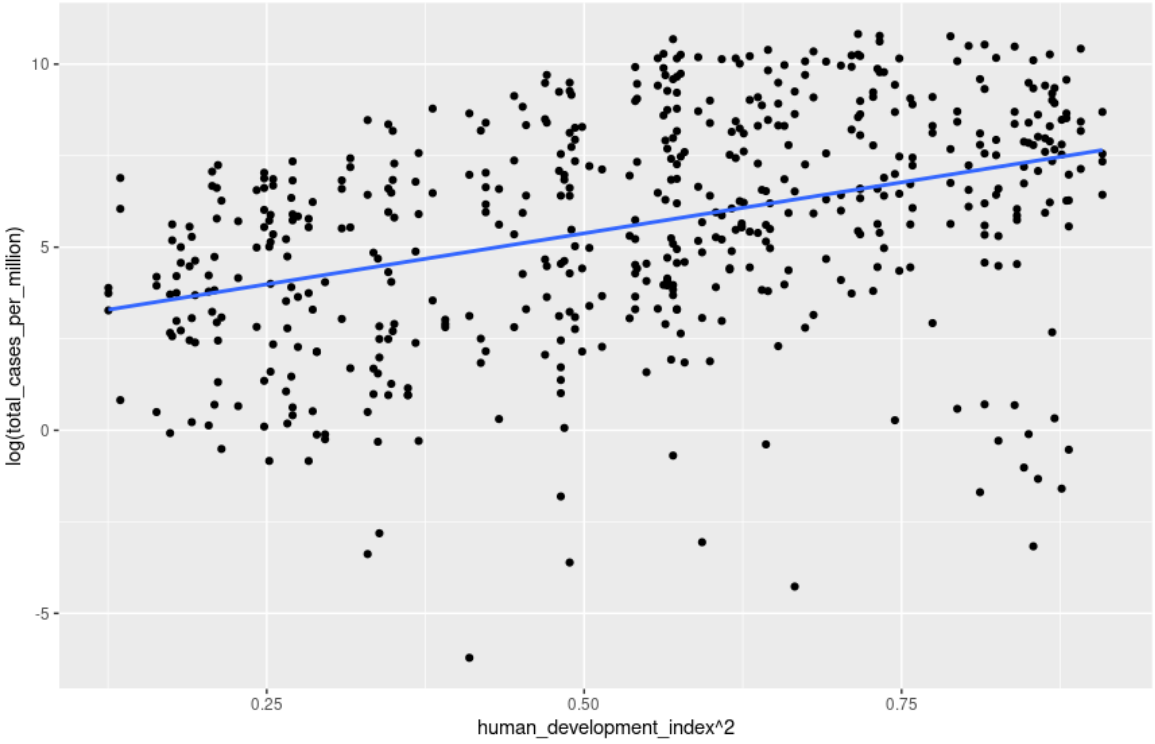
Podsumowanie modelu wygląda jak w tabeli 2.12.

Współczynnik β_{HDI} wynosi około 5.57 i jest istotny statystycznie (p-value poniżej 0.001). Wartość tego współczynnika jest dodatnia, co oznacza, że im wyższy wskaźnik rozwoju społecznego w danym kraju, tym więcej potwierdzonych przypadków COVID-19. Współczynnik determinacji wynosi około 15%, więc jest dość niski. Dopasowanie modelu do danych jest ukazane na rysunku 2.15.

	Model 1
(Intercept)	2.60***
	(0.33)
human_development_index^2	5.57***
	(0.57)
R^2	0.15
Adj. R^2	0.15
Num. obs.	534

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

Tabela 2.12: Wyniki dla modelu 6 po przekształceniu zmiennych



Rysunek 2.15: Wykres przedstawiający dopasowanie modelu liniowego do zależności pomiędzy liczbą zachorowań a wskaźnikiem rozwoju społecznego po przekształceniu zmiennych

Źródło: Opracowanie własne

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

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

Model ten jest analogiczny do poprzedniego, z tym że zamiast chorób sercowych

mamy tu odsetek chorych na cukrzycę. Na podstawie transformacji Boxa-Coxa otrzymujemy pierwiastek trzeciego stopnia ze zmiennej *diabetes_prevalence* oraz logarytm z *total_cases_per_million*.

$$\log(y_{total\_cases}) = \beta_0 + \beta_{diabetes\_prevalence} \sqrt[3]{X_{diabetes\_prevalence}} + \varepsilon$$

Wyniki znajdują się w tabeli 2.13.

	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 2.13: Wyniki dla modelu 8 po przekształceniu zmiennych

Wartość współczynnika $\beta_{diabetes\_prevalence}$ jest równa 1.65. Efekt ten jest istotny statystycznie (p-value poniżej 0.001). Ponieważ wartość współczynnika jest dodatnia, to wraz ze wzrostem odsetka osób chorujących na cukrzycę, rośnie liczba zachorowań. Dopasowanie modelu liniowego do danych po przekształceniu jest zaprezentowane na rysunku 2.16.

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

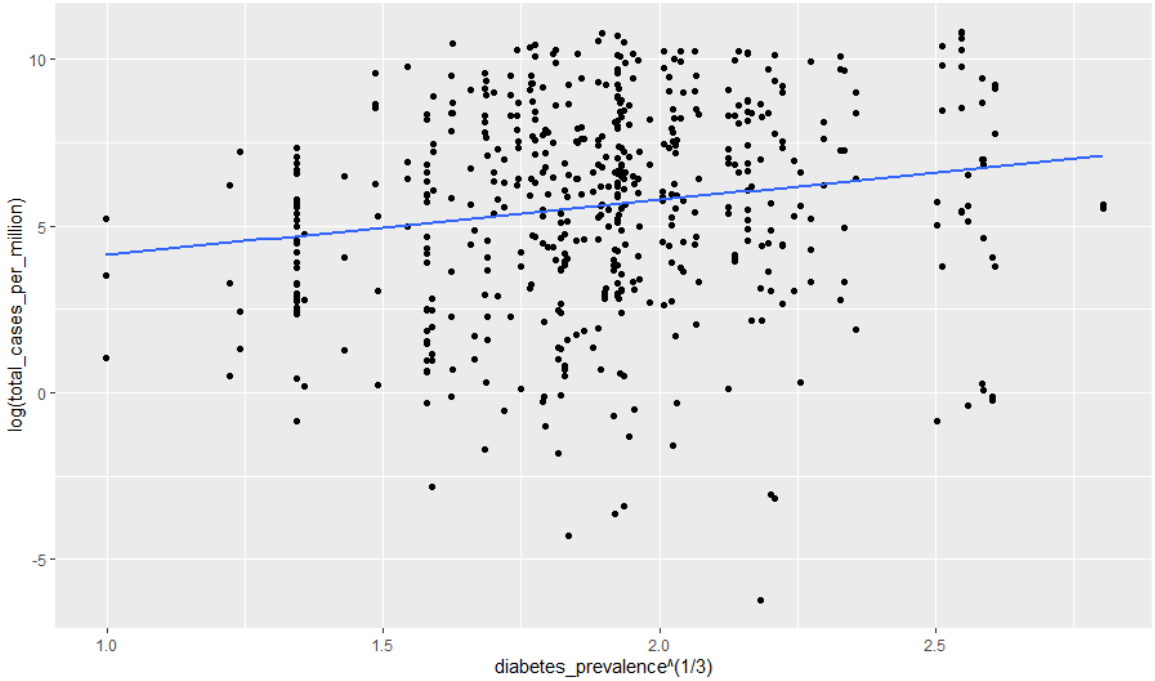
Hipoteza 9: Kraje o różnej wysokości odsetka osób żyjących w skrajnej biedzie różnią się liczbą zachorowań.

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

Stosując transformację Boxa-Coxa, otrzymujemy przekształcenie logarytmiczne dla obu zmiennych w modelu:

$$\log(y_{total\_cases}) = \beta_0 + \beta_{extreme\_poverty} \log(X_{extreme\_poverty}) + \varepsilon$$

Podsumowanie dla modelu 9 po przekształceniu zmiennych znajduje się w tabeli 2.14. Efekt $\log(extreme\_poverty)$ jest istotny statystycznie i ma wartość -0.46, co oznacza,



Rysunek 2.16: Wykres przedstawiający dopasowanie modelu liniowego z przekształceniem zmiennych do zależności pomiędzy odsetkiem osób chorych na cukrzycę a liczbą zachorowań na COVID-19
Źródło: Opracowanie własne

że wraz ze wzrostem zmiennej niezależnej maleje liczba zachorowań. R^2 wzrosło do 9%, a więc nieznacznie w stosunku do modelu przed przekształceniem zmiennych. Dopasowanie modelu liniowego do danych po przekształceniu widać na rysunku 2.17.

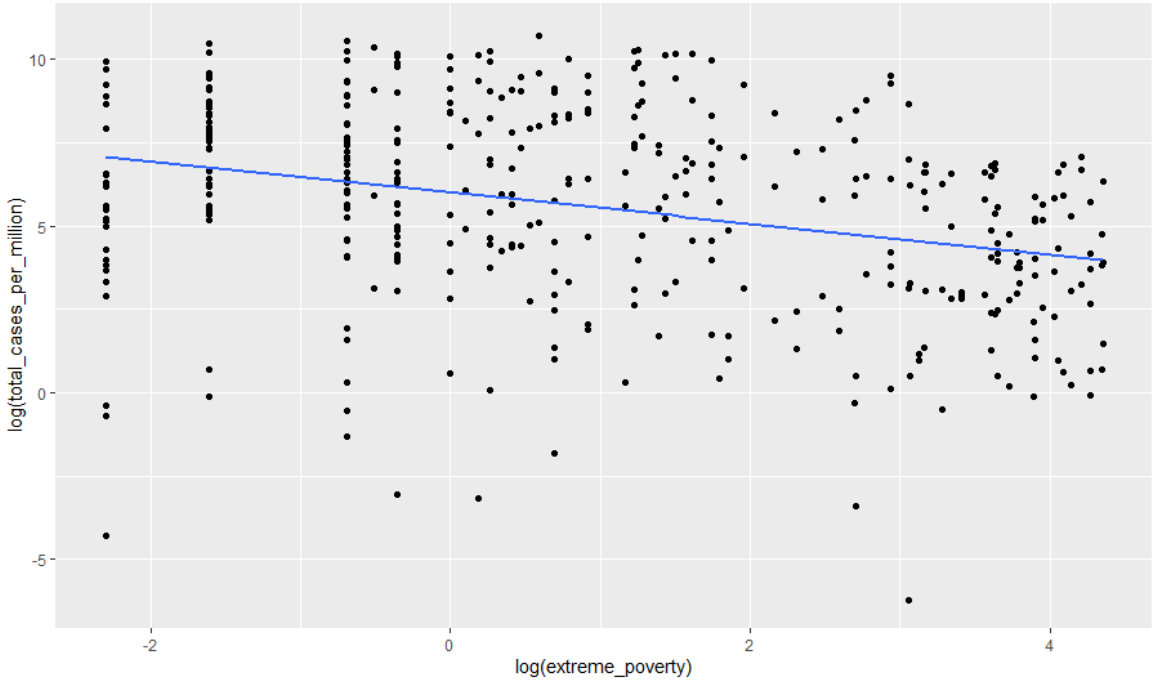
	Model 1
(Intercept)	6.00^{***}
	(0.16)
log(extreme_poverty)	-0.46^{***}
	(0.07)
R^2	0.09
Adj. R^2	0.09
Num. obs.	389

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

Tabela 2.14: Wyniki dla modelu 9 z przekształceniem logarytmicznym obu zmiennych

Czynnik *extreme_poverty* jest istotny statystycznie (p-value poniżej 0.001). Współczynnik $\beta_{extreme\_poverty}$ wynosi -0.46. Jest on ujemny, więc im większy jest w danym kraju odsetek osób żyjących w biedzie, tym niższa liczba zachorowań. Praw-

dopodobnie jest to spowodowane mniejszą dostępnością do służby zdrowia w biedniejszych krajach i mniejszą liczbą wykonywanych testów. R^2 wynosi około 9%, co jest niską wartością. Na rysunku 2.17 przedstawiona jest zależność między przekształconymi logarytmicznie liczbą zachorowań a odsetkiem osób żyjących w skrajnej biedzie, z dopasowanym modelem liniowym.



Rysunek 2.17: Wykres przedstawiający dopasowanie modelu liniowego z przekształceniem zmiennych dla zależności pomiędzy liczbą zachorowań a odsetkiem osób żyjącym w skrajnej biedzie

Źródło: Opracowanie własne

2.2.9. 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.

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

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

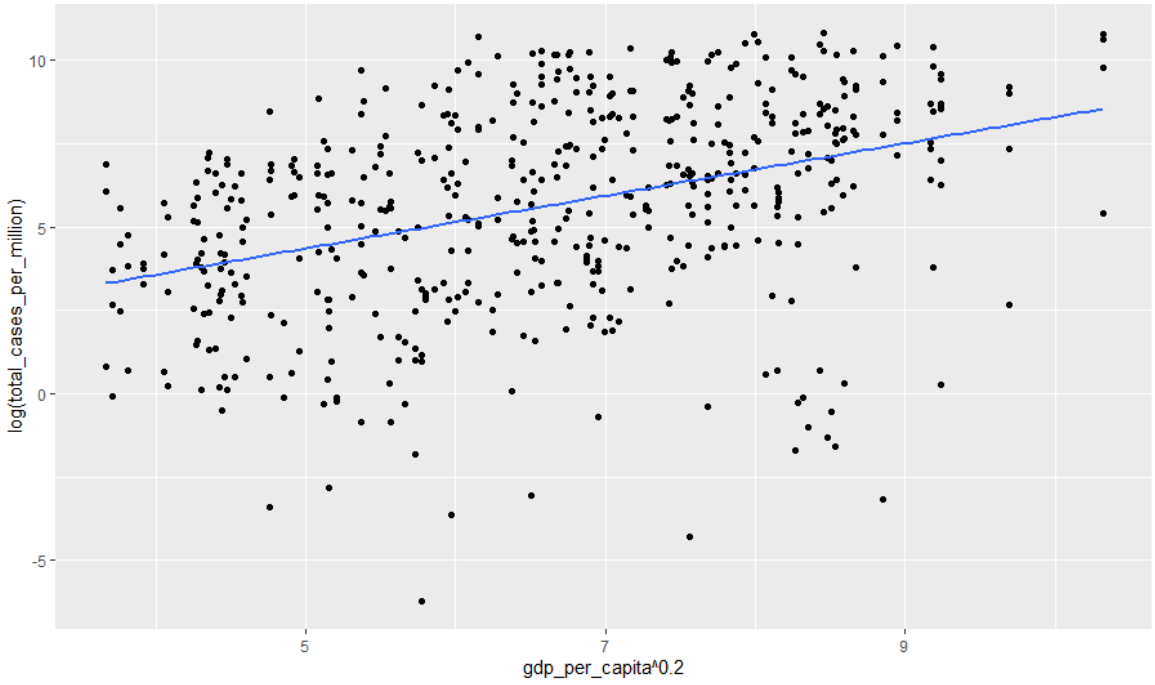
W tabeli 2.15 widać podsumowanie modelu 10 po przekształceniu zmiennych.

Efekt PKB na osobę jest istotny statystycznie (p-value poniżej 0.001). Współczynnik β_{GDP} wynosi 0.79. Jest on dodatni, więc im wyższe PKB danego kraju, tym wyższa liczba zachorowań. Współczynnik R^2 wynosi 0.15, co oznacza, że zmienna *gdp_per_capita* wyjaśnia około 15% zmienności modelu. Procent wyjaśnionej zmienności jest więc niski. Na rysunku 2.18 widać wykres rozrzutu zmiennej *total_cases_per_million* w zależności od *gdp_per_capita* po przekształceniu, z dopasowanym modelem liniowym.

	Model 1
(Intercept)	0.42
	(0.55)
gdp_per_capitaˆ 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 2.15: Wyniki dla modelu 10 z przekształceniem zmiennych



Rysunek 2.18: Wykres przedstawiający dopasowanie modelu liniowego z przekształceniem zmiennych do zależności pomiędzy liczbą zachorowań a PKB na osobę

Źródło: Opracowanie własne

Dyskusja wyników i wnioski

W tabeli 2.16 znajduje się podsumowanie istotności poszczególnych czynników, których wpływ na liczbę zachorowań był badany w tej pracy.

Cecha	Wpływ na liczbę zachorowań
Czas	istotny, można do tej zależności dopasować wielomian trzeciego stopnia
Liczba wykonywanych testów na COVID-19	istotny, wraz ze wzrostem liczby testów rośnie liczba zachorowań
Oczekiwana długość życia	istotny, ze wzrostem oczekiwanej długości życia rośnie liczba zachorowań
Gęstość zaludnienia	nieistotny
Wskaźnik siły obostrzeń	istotny jest wskaźnik opóźniony o 7 dni, można do tej zależności dopasować model kwadratowy, w większości krajów przy wysokiej sile obostrzeń jest mniej zachorowań
Wskaźnik rozwoju społecznego	istotny, w krajach o wysokim wskaźniku rozwoju jest więcej zachorowań
Śmiertelność z powodu chorób serca	istotny, im wyższy jest ten współczynnik, tym mniej zachorowań na COVID-19
Powszechność występowania cukrzycy	istotny, im wyższy jest ten współczynnik, tym więcej przypadków koronawirusa
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 2.16: Porównanie istotności i wpływu różnych czynników na liczbę zachorowań w przeciętnym kraju

Źródło: Opracowanie własne

W niniejszej pracy wpływ różnych czynników na liczbę zachorowań na COVID-19 w około 150 krajach został zbadany przy pomocy siedmiu modeli liniowych oraz trzech mieszanych. Z modeli liniowych otrzymaliśmy cztery czynniki, których wzrost powoduje wyższą liczbę zachorowań. Są to: oczekiwana długość życia, wskaźnik rozwoju społecznego, powszechność występowania cukrzycy oraz PKB na osobę. Takie wnioski wydają się dość oczywiste. Wiele źródeł mówi, że na COVID-19 są szczególnie narażone osoby starsze oraz osoby z obniżoną odpornością lub chorobami współistniejącymi [26]. Szczególnie wiek, nadwaga i cukrzyca są wymieniane jako czynniki zwiększające prawdopodobieństwo zakażenia i ciężkiego przebiegu choroby [7]. Z kolei wysokie PKB oraz wskaźnik rozwoju społecznego sugerują, że w danym kraju opieka medyczna jest dobrze rozwinięta, co skutkuje większą liczbą wykonywanych testów, pozwalając wykryć więcej przypadków wirusa. Dodatkowo, w krajach wysoko rozwiniętych większą popularnością cieszą się podróże międzynarodowe (zarówno prywatne, jak i służbowe), co sprzyja rozprzestrzenianiu się pandemii.

Ograniczająco na liczbę wykrytych przypadków koronawirusa wpływa odsetek populacji żyjący w skrajnym ubóstwie. Jednakże najprawdopodobniej ubóstwo nie wpływa bezpośrednio na zmniejszenie liczby chorych. Można przypuszczać, że mniejsza liczba zachorowań jest raczej spowodowana słabiej rozwiniętą opieką medyczną i niższą liczbą wykonywanych testów, więc znaczna część osób chorych nie jest diagnozowana. Drugim czynnikiem ograniczającym jest śmiertelność z powodu chorób serca, co jest zaskakującą zależnością. Można było przypuszczać, że wyższa śmiertelność z powodu chorób serca oznacza, że w danym kraju więcej osób doświadcza takich schorzeń, a osoby z chorobami układu sercowo-naczyniowego są bardziej narażone na COVID-19 [26] [7]. Możliwym wytłumaczeniem wniosku o ograniczającym wpływie śmiertelności osób chorych na serce na liczbę chorych na koronawirusa jest, że kraje z wysoką śmiertelnością na serce to kraje o słabo rozwiniętej opiece medycznej, i wtedy mamy do czynienia z takim samym przypadkiem, jak przy krajach o dużym ubóstwie.

Czynnikiem niemającym wpływu na liczbę przypadków koronawirusa okazała się gęstość zaludnienia. Jest to zaskakujący wniosek, ponieważ zdawałoby się, że w krajach o większej gęstości zaludnienia wirus może łatwiej się rozprzestrzeniać, więc zachorowań powinno być więcej. Brak zależności może być spowodowany tym, że gęstość zaludnienia jest podawana dla całego kraju, a przecież może bardzo się różnić wewnątrz terytorium danego państwa.

Za pomocą modeli mieszanych w tej pracy zbadano wpływ czasu, liczby wykonywanych testów oraz wskaźnika siły obostrzeń na liczbę zachorowań. Zależność liczby zachorowań od czasu została opisana wielomianem trzeciego stopnia, którego współczynniki różnią się pomiędzy krajami. Wpływ liczby testów na liczbę wykrytych

przypadków koronawirusa jest stymulujący, czego można było się spodziewać, ponieważ liczba wykrytych przypadków to nic innego jak liczba testów z pozytywnym wynikiem (w niektórych krajach można także potwierdzić przypadek COVID-19 na podstawie samych objawów). Model badający zależność liczby zachorowań od wskaźnika siły obostrzeń dawał niejednoznaczne wyniki, generalnie silniejsze obostrzenia sprawiały, że liczba zachorowań w danym kraju malała (po około tygodniu od wprowadzenia silniejszych obostrzeń), ale w niektórych państwach zależność ta nie zachodziła.

Wszystkie modele mieszane 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] 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.

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

[5] Tim Hesterberg, Shaun Monaghan, David S. Moore, Ashley Clipson, Rachel Epstein, *Bootstrap Methods and Permutation Tests. Companion Chapter 18 to the Practice of Business Statistics*, W. H. Freeman and Company, New York, 2003

[6] Hyndman, R.J., Athanasopoulos, G. *Forecasting: principles and practice, 3rd edition*, OTexts: Melbourne, Australia, 2019, `OTexts.com/fpp3` (dostęp: 07.01.2021)

[7] Jerzy Duszyński, Aneta Afelt, Anna Ochab-Marcinek, Radosław Owczuk, Krzysztof Pyrć, Magdalena Rosińska, Andrzej Rychard, Tomasz Smiatacz, *Zrozumieć COVID-19. Opracowanie zespołu ds. COVID-19 przy Prezesie Polskiej Akademii Nauk*, Polska Akademia Nauk, 14 września 2020 r.

[8] Taboga, Marco (2017). *Cholesky decomposition, Lectures on matrix algebra.* `https://www.statlect.com/matrix-algebra/Cholesky-decomposition` (dostęp: 09.01.2021)

[9] Piet de Jong, Gillian Z. Heller, *Generalized Linear Models for Insurance Data*, Cambridge University Press, New York, 2008

[10] Lang Wu, *Mixed Effects Models for Complex Data*, University of British Columbia, Vancouver, Canada, 2010

[11] Alain F. Zuur, Elena N. Ieno, Neil J. Walker, Anatoly A. Saveliev, Graham M. Smith, *Mixed Effects Models and Extensions in Ecology with R*, Springer, New York 2009

[12] `https://ourworldindata.org/coronavirus` (dsotęp: 30.11.2020)

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

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

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

[16] https://cran.r-project.org/web/packages/lme4/index.html (dostęp: 03.01.2021)

[17] https://cran.r-project.org/web/packages/lmerTest/index.html (dostęp: 03.01.2021)

[18] https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/lm (dostęp: 03.01.2021)

[19] https://www.rdocumentation.org/packages/car/versions/3.0-10/topics/powerTransform (dostęp: 03.01.2021)

[20] https://www.rdocumentation.org/packages/faraway/versions/1.0.7/topics/pulp (dostęp: 04.01.2021)

[21] https://www.naukowiec.org/wiedza/statystyka/interpretacja-wykresow-rozrzutu_769.html (dostęp: 07.01.2021)

[22] https://pl.wikipedia.org/wiki/Przekszta%C5%82cenie_Boxa-Coxa (dostęp: 07.01.2021)

[23] https://pl.wikipedia.org/wiki/Wsp%C3%B3%C5%82czynnik_determinacji (dostęp: 08.01.2021)

[24] https://pl.wikipedia.org/wiki/Model_statystyczny#Skorygowany_wsp%C3%B3%C5%82czynnik_determinacji (dostęp: 08.01.2021)

[25] https://www.medonet.pl/koronawirus-pytania-i-odpowiedzi/sars-cov-2,kto-jest-w-najwiekszej-grupie-ryzyka-koronawirusa-,artykul,48398767.html (dostęp: 09.01.2021)

[26] https://pl.wikipedia.org/wiki/Macierz_rzadka (dostęp: 08.01.2021)

[27] https://pl.wikipedia.org/wiki/Twierdzenie_Bayesa (dostęp: 08.01.2021)

[28] https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-codebook.csv (dostęp: 09.01.2021)

[29] https://pl.wikipedia.org/wiki/Wska%C5%BAnik_rozwoju_spo%C5%82ecznego (dostęp: 09.01.2021)

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

Ta praca przedstawia zastosowanie modeli liniowych oraz modeli mieszanych do analizy rozwoju pandemii COVID-19 na świecie. Pierwszy rozdział jest poświęcony teorii matematycznej. Na początku omówione zostają zagadnienia dotyczące modeli liniowych. Następnie te pojęcia są poszerzane o modele liniowe z efektami stałymi i losowymi. Przedstawione zostały problemy takie jak: metody estymacji oraz badanie istotności parametrów modelu, predykcja z modelu mieszanego, jak również interpretacja modelu. W rozdziale drugim przeprowadzono badania na zbiorze danych dotyczącym zachorowań na COVID-19. Zbudowano dziesięć modeli, które mają na celu zidentyfikowanie czynników istotnie wpływających na liczbę zachorowań na koronawirusa w różnych krajach. W ostatnim rozdziale zostały przedstawione wnioski wyciągnięte z badań.

The Use of Mixed-Effects Models in the Analysis of the COVID-19 Pandemic in the World

This paper presents the use of linear and mixed-effects models in analysis of the development of the COVID-19 pandemic worldwide. The first chapter is dedicated to mathematical theory. At the beginning, issues concerning linear models are discussed. Then these concepts are extended with linear models with fixed and random effects. Problems such as: estimation methods and testing the significance of model parameters, prediction from a mixed-effects model, as well as model interpretation are presented. In the second chapter, a study was conducted on the COVID-19 dataset. Ten models were built to identify the factors significantly affecting the number of coronavirus cases in different countries. The last chapter presents the conclusions drawn from the study.