

Uniwersytet Warszawski
Wydział Matematyki, Informatyki i Mechaniki

Anna Lis

Nr albumu: 234118

**Uogólnione modele addytywne
z parametrem położenia, skali
i kształtu**

**Praca magisterska
na kierunku MATEMATYKA
w zakresie MATEMATYKA STOSOWANA**

Praca wykonana pod kierunkiem
dra inż. Przemysława Biecka
Zakład Statystyki Matematycznej
MIM UW

Listopad 2011

Oświadczenie kierującego pracą

Potwierdzam, że niniejsza praca została przygotowana pod moim kierunkiem i kwalifikuje się do przedstawienia jej w postępowaniu o nadanie tytułu zawodowego.

Data

Podpis kierującego pracą

Oświadczenie autora (autorów) pracy

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Data

Podpis autora (autorów) pracy

Streszczenie

W pracy przedstawiona została teoria funkcji sklejanych, w tym m.in. wielomianowe funkcje sklejane, B-splajny i kubiczne wygładzone funkcje sklejane, które mają szerokie zastosowanie w regresji lokalnie wygładzanej. Przedstawione są również modele liniowe, uogólnione modele liniowe (GLM), uogólnione modele addytywne (GAM) oraz ich szersza klasa, czyli uogólnione modele addytywne z parametrem położenia, skali i kształtu (GAMLSS). Opis matematyczny modeli GAM i GAMLSS, oraz funkcji sklejanych został uzupełniony o charakterystykę odpowiadających im funkcji środowiska R, a także ich zastosowanie do zbioru danych pediatrycznych. Celem wykonywanych analiz było opracowanie polskich norm ciśnienia tętniczego dla dzieci i młodzieży osobno dla obu płci w zależności od wieku i wysokości ciała. Dane wykorzystane w przykładach i analizach pochodzą ze współpracy z Instytutem „Pomnik – Centrum Zdrowia Dziecka”.

Słowa kluczowe

regresja funkcji sklejanych, kubiczna funkcja sklejana, B-splajn, skośność, kurtoza, model liniowy, GLM, GAM, GAMLSS, kryterium AIC, kryterium BIC, worm plot, statystyka Q , centyl, rozkład BCCG

Dziedzina pracy (kody wg programu Socrates-Erasmus)

11.2 Statystyka

Klasyfikacja tematyczna

62H12 Estimation
G2J05 Linear regression
G2J99
G2J20 Diagnostics
G2J12 Generalized linear models
G2G99
62P10 Applications to biology and medical sciences

Tytuł pracy w języku angielskim

Generalized additive models for location, scale and shape (GAMLSS)

Spis treści

1. Funkcje sklejane	9
1.1. Funkcje kawałkami wielomianowe i funkcje sklejane	10
1.2. Naturalne funkcje sklejane	12
1.3. Wygładzone funkcje sklejane	13
1.4. Stopnie swobody i macierze wygładzeń	14
1.5. Wybór parametru wygładzenia	16
1.6. Jak zapisać funkcje sklejane za pomocą B-splajnów?	16
1.6.1. B-splajny	17
1.6.2. Jak zapisać wygładzone funkcje sklejane za pomocą B-splajnów?	18
2. Funkcje sklejane w pakiecie R i ich zastosowanie	21
2.1. Wielomianowe funkcje sklejane – <code>bs(splines)</code> i <code>ns(splines)</code>	21
2.1.1. Przykłady zastosowania funkcji <code>bs</code> i <code>ns</code>	22
2.2. Wygładzone funkcje sklejane - <code>smooth.spline(stats)</code>	27
2.2.1. Przykład zastosowania funkcji <code>smooth.spline</code>	28
3. Regresja liniowa, modele GLM i GAM	31
3.1. Wprowadzenie	31
3.2. Modele Liniowe	31
3.3. Uogólnione modele liniowe (GLM)	33
3.4. Uogólnione modele addytywne (GAM)	33
3.5. Wyznaczanie modelu addytywnego (GAM)	35
3.6. Uogólnione modele addytywne w R	37
3.6.1. Funkcje <code>gam</code>	37
3.6.2. Przykład wykorzystania funkcji <code>gam</code>	38
4. GAMLSS	41
4.1. Co to jest GAMLSS?	41
4.2. Postać GAMLSS	42
4.3. Estymacja modelu	44
4.4. Liniowy predyktor w GAMLSS	45
4.5. Pakiet <code>gamlss</code>	45
4.5.1. Różne funkcje w pakiecie <code>gamlss</code>	45
4.5.2. Funkcja <code>gamlss()</code>	46
4.5.3. Dostępne rozkłady	46
4.5.4. Addytywne składniki	49
4.5.5. Kilka słów o budowie modelu	51
4.6. Podsumowanie GAMLSS	51

5. Analiza danych medycznych z wykorzystaniem gamlss	53
5.1. Krzywe centylowe wzrostu dla chłopców	53
5.2. Model ciśnienia skurczowego dla chłopców bez nadwagi	63
5.3. Podsumowanie	73
A. Opis danych danemed	77
B. Wybrane definicje	79
B.1. Kryteria informacyjne – AIC i BIC	79
B.2. Skośność	80
B.3. Kurtoza	80
B.4. Wybrane rozkłady	81
B.4.1. Rozkład normalny (NO)	81
B.4.2. Rozkład log-normalny (LOGNO, LNO)	82
B.4.3. Rozkład Box'a-Cox'a-Cole'a-Green'a (BCCG)	82
Bibliografia	83

Wprowadzenie

W niemal wszystkich dziedzinach badań empirycznych mamy do czynienia ze złożonością zjawisk i procesów. W związku z tym z wykorzystaniem metod analizy danych są wykonywane ilościowe oceny relacji występujących pomiędzy różnymi aspektami badanych zjawisk i procesów. Bardzo popularną i chętnie stosowaną metodą statystyczną jest analiza regresji, której ogólną postać można zapisać jako:

$$Y|X \sim \mathcal{F}(\theta) \quad (1)$$

$$\mathbb{E}(Y|X) = f(X, \beta).$$

W modelu tym Y oznacza zmienną zależną zwaną także zmienną objaśnianą lub zmienną odpowiedzi, natomiast X to wektor zmiennych niezależnych zwanych zmiennymi objaśniającymi lub predyktorami. W analizie regresji istotnym zagadnieniem jest opisanie oczekiwanej wartości zmiennej Y za pomocą zmiennych objaśniających X , a więc wyznaczenie wektora parametrów modelu β , przy założonej postaci modelu opisanej funkcją $f(\cdot)$. Obserwacji nie podlegają wartości oczekiwane, ale wartości zmiennej losowej o rozkładzie z rodziną \mathcal{F} indeksowanej parametrem θ . W zależności od analizowanego zagadnienia np. rodzaju i liczby zmiennej objaśnianej, oraz zmiennych objaśniających stosuje się różne metody analizy regresji.

Niewątpliwie najprostszą i najbardziej popularną metodą analizy regresji jest regresja liniowa, której ogólny model przyjmuje postać:

$$Y = X\beta + \varepsilon, \quad (2)$$

gdzie ε to zakłócenie losowe o rozkładzie $\mathcal{N}(0, \sigma^2)$. W regresji liniowej pomiędzy zmiennymi objaśniającymi, a zmienną objaśnianą istnieje mniej lub bardziej wyrazista zależność liniowa, tzn. funkcja $f(\cdot)$ ze wzoru (1) jest funkcją liniową.

Jeżeli w analizowanych danych spodziewamy się wystąpienia nieliniowych zależności między zmiennymi niezależnymi i zmienną zależną, można postąpić na kilka sposobów: można mianowicie próbować dokonać transformacji zmiennych, tak aby w pewnym stopniu „uliniowić” model, lub też rozważyć zastosowanie regresji nieliniowej. Cechą wspólną tego rodzaju regresji jest to, że należy znać a priori (lub założyć na wstępnie analizy) jakąś matematyczną zależność wiążącą zmienne objaśniające ze zmienną objaśnianą (rozkład zmiennej odpowiedzi Y i funkcję $f(\cdot)$). W związku z tym te metody znane są pod nazwą *regresji parametrycznej*. W wielu przypadkach jednak określenie konkretnej postaci funkcji $f(\cdot)$ nie jest łatwe, a czasami wręcz niewykonalne. Zdarza się to na przykład, gdy analizowane dane wiążą bardzo złożoną zależność. W takich sytuacjach rozważa się często użycie metod wygładzania funkcjami sklejonymi, których główną ideą jest określenie klasy funkcji $f(\cdot)$ (np. założenie, że $f \in \mathbb{C}_1$), oraz wyznaczenie takiej liniowej kombinacji różnych funkcji danej klasy, która opisze relację pomiędzy X i Y . Omówienie idei funkcji sklejanych i przedstawienie ich zastosowania będzie pierwszym punktem niniejszej pracy.

Rozważając użyteczność standardowej metody regresji liniowej (2) można zauważać, że klasyczne założenia o normalności błędu, czy liniowej relacji między zmienną objaśnianą i zmiennymi objaśniającymi często okazują się niedostateczne. Wówczas alternatywą mogą być uogólnione modele liniowe (GLM) lub uogólnione modele addytywne (GAM), które zostały zaprojektowane w celu pokonania niektórych problemów, jakie napotyka się w prostej regresji liniowej. W pewnych przypadkach, szczególnie dla większych zbiorów danych, okazują się one także być niewystarczające [17].

Głównym celem tej pracy jest przedstawienie metody uogólnionych modeli addytywnych z parametrem położenia, skali i kształtu, nazywanej w skrócie GAMLSS. Metoda GAMLSS radzi sobie z większością ograniczeń modeli GLM i GAM, jednocześnie łączy różne typy modeli regresyjnych, oraz dodatkowo umożliwia modelowanie wszystkich parametrów rozkładu tj. średniej (położenia), wariancji (rozproszenia), a także parametrów kształtu – skośności i kurtozy.

Kolejnym punktem niniejszej pracy jest zaprezentowanie zastosowania funkcji `gamlss()` dostępnej w pakiecie `gamlss` środowiska R, która odpowiada za wyznaczenie modelu GAMLSS. Pewne możliwości tej metody zostały wykorzystane do opracowania norm ciśnienia tętniczego dla dzieci i młodzieży w Polsce, które będą mogły służyć rozpoznawaniu podwyższzonego ciśnienia tętniczego.

Opracowane w tej pracy zagadnienie norm ciśnienia dla dzieci i młodzieży polskiej posiada interesujące podłożę medyczne. Problem podwyższzonego ciśnienia tętniczego w wieku dziecięcym i młodzieńczym jest związany z większym ryzykiem rozwoju nadciśnienia tętniczego i choroby sercowo-naczyniowej w wieku dorosłym. Nadciśnienie tętnicze występujące w wieku rozwojowym prowadzi do uszkodzeń narządowych stwierdzanych już w momencie rozpoznania choroby. Ze względu na te uwarunkowania, pomiar ciśnienia tętniczego i porównanie wyniku z biologicznym układem odniesienia (zakresy referencyjne, normy) są niezmiennie istotne. W Polsce nie jest dostępny zakres referencyjny ciśnienia tętniczego wieku rozwojowego, który byłby opracowany na podstawie danych pochodzących z próby reprezentatywnej dla całej krajowej populacji. Wyniki pomiarów ciśnienia polskich dzieci są interpretowane na podstawie zakresów referencyjnych ciśnienia tętniczego dzieci i młodzieży ze Stanów Zjednoczonych, które zostały opracowane dla populacji istotnie różniącej się od populacji krajowej, zwłaszcza ze względu na rozpowszechnienie nadwagi i otyłości, oraz różnice rasowe. Ostatni amerykański raport został opublikowany w 2004 roku i podaje wartości ciśnienia tętniczego w zależności od płci, wieku i wzrostu. Raport ten budzi jednak istotne zastrzeżenia m.in. ze względu na fakt, że średnie ciśnienie tętnicze dzieci Afroamerykanów jest wyższe niż dzieci rasy białej. Częstość uzyskiwania wysokiego wyniku pomiaru ciśnienia tętniczego różni się między rasami, jakkolwiek w dużym stopniu jest to zależne od różnic etnicznych w występowaniu nadwagi i otyłości. Dlatego interpretując wynik pomiaru ciśnienia tętniczego dziecka pochodzącego z innej populacji, należy uwzględnić odrębności wynikające ze zróżnicowania etnicznego, [12]. Powyższe rozważania stanowią motywację do opracowania aktualnych polskich norm ciśnienia tętniczego dla dzieci i młodzieży.

Na potrzebę projektu OLAF¹ z 2010 roku została zebrana reprezentacyjna próba z udziałem ponad 17500 dzieci i młodzieży. We współpracy z doktorem Zbigniewem Kułagą – Kierownikiem Zakładu Zdrowia Publicznego IPCZD i Koordynatorem projektu OLAF – korzystając z danych pochodzących z projektu oraz wykorzystując nową metodę GAMLSS, dokonałam przeliczenia bezwzględnych wartości ciśnienia tętniczego na centylową reprezentację. Analizy zostały wykonane w oparciu o dane dla dzieci z prawidłową masą ciała, czyli z wyklucze-

¹Opracowanie norm ciśnienia tętniczego dzieci i młodzieży w Polsce OLAF(PL0080)

niem osób z nadwagą i otyłością². Wyniki z przeprowadzonych analiz w całości znajdują się w publikacji autorstwa m.in. doktora Zbigniewa Kułagi, *Oscillometric blood pressure percentiles for Polish normal-weight school-aged children and adolescents* [13], której jestem także współautorem. W niniejszej pracy została przedstawiona niewielka część wykonanych analiz, która obejmuje wyznaczenie norm ciśnienia skurczowego dla dzieci i młodzieży płci męskiej w wieku 7-18 z uwzględnieniem wieku i wzrostu.

Praca została podzielona na pięć rozdziałów, jednak koncepcyjnie składa się z dwóch części: pierwszej – zawierającej teorię i zastosowanie poparte przykładami funkcji sklejanych, oraz drugiej – opisu modelu GAMLSS, pakietu `gamlss` z funkcją o tej samej nazwie, która wyznacza model GAMLSS w środowisku R, jak również zastosowanie metody GAMLSS do wyznaczenia m.in. norm ciśnienia skurczowego.

W rozdziale pierwszym przedstawiony został jeden z głównych nurtów *nieparametrycznej*³ estymacji funkcji *regresji* – metody przybliżania funkcjami sklejonymi. Omówione zostały m.in. wielomianowe funkcje sklejane, naturalne funkcje sklejane, wygładzone funkcje sklejane oraz bazowe funkcje B-sklejane, tzw. B-splajny.

Drugi rozdział zawiera opis i przykłady wykorzystania funkcji `bs`, `ns` oraz `smoothsplines`, które w środowisku R wspomagają lub odpowiadają za generowanie modeli zawierających funkcje sklejane i wygładzone funkcje sklejane.

W trzecim rozdziale przedstawione są następujące klasy modeli statystycznych: modele liniowe, uogólnione modele liniowe i uogólnione modele addytywne, które stanowią niejako wprowadzenie do modelu GAMLSS. Dodatkowo rozdział ten zawiera charakterystykę funkcji programu R, które służą do budowy modelu GAM, oraz które stanowiły wzór podczas konstrukcji funkcji `gamlss()`. Opis ten został uzupełniony o przykład zastosowania funkcji `gam()`.

Rozdział czwarty zawiera szczegółowy opis postaci modelu GAMLSS. Oprócz teorii dotyczącej modelu, w tym rozdziale scharakteryzowany został także pakiet `gamlss`, który jest dostępny w środowisku R. Charakterystyka pakietu uwzględnia m.in. opis wybranych funkcji pakietu (w tym głównej funkcji `gamlss()`) i pewnych składników wykorzystywanych podczas estymacji, oraz przedstawienie listy rozkładów udostępnionych w pakiecie `gamlss`.

W piątym rozdziale zaprezentowane zostały przykłady zastosowania funkcji pakietu `gamlss` do wyznaczenia centyl wzrostu oraz polskich norm ciśnienia skurczowego dla dzieci i młodzieży płci męskiej. Wszystkie przykłady zawarte w pracy oparte zostały o opisaną w dodatku A bazę danych pediatrycznych `danemed`, która pochodzi ze współpracy z Instytutem „Pomnik – Centrum Zdrowia Dziecka”.

²Motywacja oraz sposób wykluczania osób z nadwagą zostały szczegółowo opisane w [13].

³Metoda *nieparametryczna* regresji stanowi alternatywną koncepcję dla *regresji parametrycznej*. Metody regresji nieparametrycznej nie zakładają, że estymowana funkcja $f(\cdot)$ jest znana z dokładnością do skończenie wielu estymowanych parametrów. Ogólnie, nieparametryczność polega na tym, że mechanizm badanego zjawiska traktuje się jako nieznany i w związku z tym nie zakłada się często żadnej postaci modelu $f(\cdot)$. Analiza oparta jest wyłącznie na danych i szukane są jedynie związki pomiędzy wielkościami wejściowymi, a wyjściowymi. (Harańczyk G., *Zastosowanie technik data mining w badaniach naukowych*, 2010.)

Rozdział 1

Funkcje sklejane

Rozdział ten w znacznej części opiera się na teorii zawartej w piątym rozdziale książki [7]. Zostaną w nim omówione popularne metody, które w modelowaniu statystycznym pozwalają wychodzić poza modele liniowe. W wielu zagadnieniach statystycznych modele liniowe nie wystarczają by dobrze opisać zależności występujące między zmiennymi objaśniającymi, a zmienną objaśnianą. W większości przypadków dla odpowiednich danych wejściowych zmienne wymagają wykonania pewnych transformacji, a następnie użycia modeli liniowych w nowej przestrzeni zmiennych.

Przez X zostanie oznaczona macierz eksperymentu rozmiaru $N \times P$, która zawiera P -zmiennych objaśniających oraz N obserwacji (w przypadku jednej zmiennej objaśniającej jest to wektor długości N). Niech $h_m(X) : \mathbb{R}^P \rightarrow \mathbb{R}$ będzie m -tą transformacją X , $m = 1, \dots, M$. Wówczas można zapisać:

$$f(X) = \sum_{m=1}^M \beta_m h_m(X) \quad (1.1)$$

w postaci rozszerzonego modelu liniowego transformowanych X . Otrzymana w wyniku M -przekształceń macierz $\mathbf{h}(X)$ nazywana będzie macierzą regresji. Główną ideą tego podejścia jest ustalenie a priori funkcji h_m oraz zbudowanie nowego modelu, który będzie liniowy w nowych zmiennych.

Przykłady funkcji h_m :

- $h_m(X) = X_m$, $m = 1, \dots, P$ – opisują oryginalny model liniowy i stanowią aproksymację $f(X)$ rozwinięciem Taylora pierwszego rzędu.
- $h_m(X) = X_j^2$ lub $h_m(X) = X_j X_k$ zwiększą w modelu wkład poszczególnych składników wielomianowych, co prowadzi do zwiększenia rzędu rozwinięcia Taylora. Liczba zmiennych rośnie wykładniczo w stosunku do stopnia wielomianu. Pełny model kwadratowy z P -zmiennymi objaśniającymi wymaga $O(P^2)$ kwadratowych i mieszanych składników, lub bardziej ogólnie $O(P^d)$ składników dla wielomianu stopnia d .
- $h_m(X) = \log(X_j)$, $\sqrt{X_j}$ oraz inne nieliniowe transformacje pojedynczych zmiennych.
- $h_m(X) = I(L_m \leq X_k < U_m)$ – indykatorka¹ dla pewnego zakresu zmiennej X_k . Podzielenie zakresu zmiennej X_k na M_k rozłącznych przedziałów umożliwia modelowanie

¹Indykatorem $x \in X_I$ nazywa się funkcję rzeczywistą $f : X_I \mapsto \mathbb{R}$ określoną następującym wzorem:

$$f(x) := \begin{cases} 1 & \text{gdy } x \in X_I \\ 0 & \text{gdy } x \notin X_I \end{cases}$$

rozkładu zmiennej X_k za pomocą funkcji kawałkami stałymi.

Czasami używane są popularne funkcje h_m , takie jak logarytmy lub funkcje potęgowe. Częściej jednak w celu osiągnięcia bardziej elastycznej reprezentacji $f(X)$ są wykorzystywane funkcje wielomianowe.

W pierwszych trzech podrozdziałach zostały przedstawione rodziny funkcji kawałkami wielomianowych i funkcji sklejanych (tj. wielomianowe funkcje sklejane, naturalne funkcje sklejane oraz wygładzone kubiczne funkcje sklejane). Następne podrozdziały zawierają opis matematycznych struktur przydatnych do wyznaczania estymatora utworzonego za pomocą kubicznych wygładzonych funkcji sklejanych.

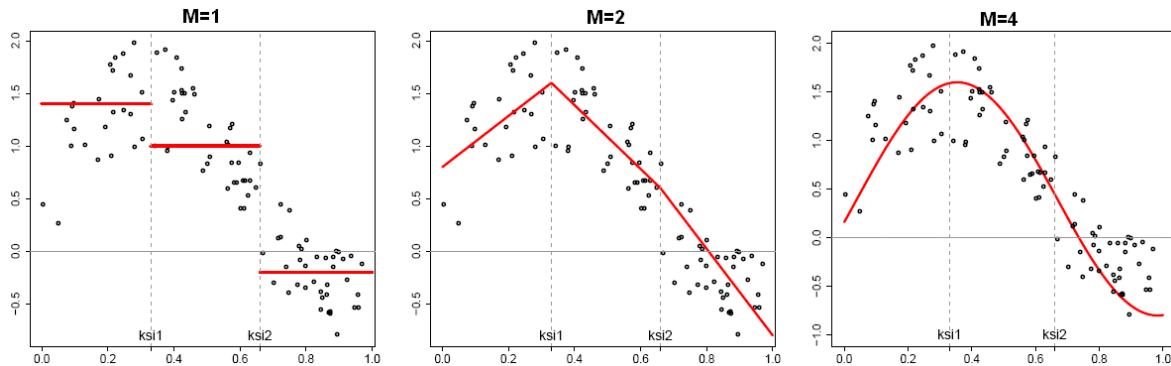
1.1. Funkcje kawałkami wielomianowe i funkcje sklejane

Aby w zrozumiały sposób omówić metody konstruowania kawałkami wielomianowych funkcji sklejanych na wstępnie przyjmuję założenie, że X to wektor długości N objaśniającej zmiennej ilościowej, która przyjmuje dowolne wartości rzeczywiste, np. wartości z określonego przedziału. Dodatkowo wartości w wektorze X są posortowane niemalejąco. Funkcję kawałkami wielomianową $f(X)$ uzyskuje się przez podzielenie dziedziny X na rozłączne przedziały, wyznaczone przez ciąg węzłów, czyli punktów wektora $\xi = (\xi_1, \dots, \xi_K)$, $K \leq N$. W każdym wyznaczonym przez kolejne węzły przedziale funkcja f będzie reprezentowana przez różne funkcje wielomianowe. Celem powyższych zabiegów jest opisanie zmiennej objaśnianej Y przy pomocy funkcji f kawałkami wielomianowymi.

Najprostszym przykładem funkcji wielomianowych są funkcje kawałkami stałe, które zostały przedstawione na lewym rysunku 1.1. Ustalam dwa węzły ξ_1, ξ_2 , które nie są węzłami brzegowymi tzn. nie są najmniejszą bądź największą wartością zawartą w X . Wówczas trzy bazowe funkcje, z których zostanie utworzona funkcja kawałkami sklejana, to:

$$h_1(X) = I(X < \xi_1), \quad h_2(X) = I(\xi_1 \leq X < \xi_2), \quad h_3(X) = I(\xi_2 \leq X), \quad (1.2)$$

gdzie $I(\cdot)$ jest indykatorem odpowiedniego przedziału. Po weryfikacji warunków h_i ($i = 1, 2, 3$) dla każdego z punktów X , przy pomocy metody najmniejszych kwadratów zostaje wyznaczony estymator $\hat{f}(X) = \sum_{m=1}^3 \hat{\beta}_m h_m(X)$, gdzie $\hat{\beta}_m = \bar{Y}_m$ jest średnią Y w m -tym przedziale.



Rysunek 1.1: Kolejne wykresy od lewej przedstawiają następujące dopasowania funkcjami: kawałkami stałymi, ciągłymi funkcjami liniowymi oraz kubicznymi funkcjami 3-ego stopnia.
Źródło: opracowanie własne.

Aby zapewnić ciągłość estymowanej funkcji, estymacja współczynników wielomianu w danym przedziale nie może odbywać się niezależnie od estymacji współczynników wielomianów

w przedziałach sąsiednich. Środkowy wykres 1.1 przedstawia dopasowanie wielomianami liniowymi, które tworzą ciągłą funkcję przedziałami liniową. Do funkcji wyrażonych wzorami (1.2) zostały dodane trzy dodatkowe funkcje bazowe: $h_{m+3} = h_m(X)X$, $m = 1, \dots, 3$. Ciągłość w węzłach ξ_1 i ξ_2 wymaga spełnienia dwóch warunków: $f(\xi_1^-) = f(\xi_1^+)$ i $f(\xi_2^-) = f(\xi_2^+)$, które implikują odpowiednio $\beta_1 + \xi_1\beta_4 = \beta_2 + \xi_1\beta_5$ oraz $\beta_2 + \xi_2\beta_5 = \beta_3 + \xi_2\beta_6$. Mając te dwa warunki można oczekiwac odzyskania dwóch parametrów, cztery parametry pozostaną wolne.

W celu uzyskania ciągłej liniowej funkcji sklejanej można także skorzystać z następującej reprezentacji funkcji bazowych:

$$h_1(X) = 1, \quad h_2(X) = X, \quad h_3(X) = (X - \xi_1)_+, \quad h_4(X) = (X - \xi_2)_+, \quad (1.3)$$

gdzie:

$$(X - \xi_j)_+ = \begin{cases} X - \xi_j & \text{jeżeli } X > \xi_j \\ 0 & \text{w p.p.} \end{cases}$$

W sytuacji gdy zależy nam na bardziej wygładzonym estymatorze, wówczas jego wyznaczenie wymaga zwiększenia stopnia lokalnych wielomianów. Prawy wykres rys. 1.1 przedstawia kawałkami wielomianową, ciągłą funkcję, która w węzłach ma ciągłą pierwszą i drugą pochodną. Taka funkcja znana jest pod nazwą *kubicznej funkcji sklejanej*.

Kubiczne funkcje sklejane z węzłami w punktach ξ_1 i ξ_2 mogą być reprezentowane przez następującą bazę:

$$\begin{aligned} h_1(X) &= 1, & h_2(X) &= X, & h_3(X) &= X^2, \\ h_4(X) &= X^3, & h_5(X) &= (X - \xi_1)_+^3, & h_6(X) &= (X - \xi_2)_+^3. \end{aligned} \quad (1.4)$$

Jest to sześć funkcji bazowych odpowiadających 6-wymiarowej przestrzeni funkcji liniowych. W łatwy sposób można obliczyć liczbę stopni swobody: $(3 \text{ przedziały}) \times (4 \text{ parametry dla danego przedziału}) - (2 \text{ węzły}) \times (3 \text{ warunki dla jednego węzła}) = 6$.

Zgodnie z oznaczeniami z książki [7] *funkcją sklejaną stopnia M*, inaczej *splajnem stopnia M* z węzłami ξ_j , $j = 1, \dots, K$ jest kawałkami wielomianowa funkcja stopnia M , która ma ciągłą pochodną do $(M - 2)$ -stopnia. Dla kubicznych funkcji sklejanych $M = 4$. W rzeczywistości kawałkami stałe funkcje z rysunku 1.1 są funkcjami sklejonymi stopnia $M = 1$, liniowe funkcje kawałkami ciągłe stanowią funkcje sklejane stopnia $M = 2$. Ogólnie, zbiór funkcji bazowych ustalonego rzędu można przedstawić przy pomocy składników:

$$h_j(X) = X^{j-1}, \quad j = 1, \dots, M \quad (1.5)$$

$$h_{M+l}(X) = (X - \xi_l)_+^{M-1}, \quad l = 1, \dots, K.$$

Postać funkcji sklejanej S stopnia M , czyli splajn stopnia M można zapisać w postaci szeregu potęgowego:

$$S(X) = \sum_{j=1}^M \beta_j X^{j-1} + \sum_{j=1}^K \gamma_j (X - \xi_j)_+^{M-1}.$$

Istnieje opinia [7], że kubiczne funkcje sklejane są funkcjami sklejonymi najniższego rzędu dla których nieciągłość w węzłach nie jest widoczna dla ludzkiego oka. Rzadko też w analizach statystycznych stosuje się funkcje sklejane wyższego stopnia, chyba, że z jakiś przyczyn autorowi zależy na gładkich pochodnych wyższego rzędu. W praktyce najczęściej wykorzystywany stopniami funkcji sklejanych są: $M = 1, 2$ oraz 4 .

Przytoczona metoda funkcji sklejanych z ustalonimi węzłami znana jest pod nazwą *regresji funkcji sklejanych* (ang. *regression splines*). Polega ona na ustaleniu stopnia funkcji sklejanych, liczby węzłów i ich położenia. Dodatkowo wymagane jest sparametryzowanie rodzin

funkcji sklejanych przez liczbę funkcji bazowych lub liczbę stopni swobody. W środowisku R do wyznaczania bazy funkcji sklejanych służy funkcja `bs`. Przykładowo, dla wektora \mathbf{x} zawierającego N obserwacji wyrażenie `bs(x, df = 7)` generuje tzw. *macierz regresji* w bazie kubicznych funkcji sklejanych wyznaczoną dla 4^2 wewnętrznych węzłów w odpowiednich percentylach \mathbf{x} , tutaj $\mathbf{x}(20, 40, 60 \text{ i } 80)$. Można również na wstępnie określić położenie węzłów oraz stopień funkcji sklejanych, np. `bs(x, degree=1, knots = c(0.2, 0.4, 0.6))` generuje macierz regresji rozmiaru $N \times 4$ w bazie liniowych funkcji sklejanych z trzema wewnętrznymi węzłami. Więcej informacji na temat funkcji `bs()` można znaleźć w rozdziale 2.1.

Dla przestrzeni funkcji sklejanych ustalonego rzędu z ustalonym ciągiem węzłów, podobnie jak dla zwykłych wielomianów, istnieje wiele równoważnych reprezentacji baz. Koncepcja wielomianów ustalonego rzędu jest bardzo prosta, jednak jej wadą jest numeryczna nieoptymalność. Rozwiązaniem dla tego problemu jest *baza B-splajnów* opisana w rozdziale 1.6, która pozwala na wykonywanie efektywnych obliczeń nawet gdy liczba węzłów K jest duża.

1.2. Naturalne funkcje sklejane

Zachowanie wielomianów dopasowanych do danych często bywa problematyczne poza ich brzegowymi węzłami. Wówczas jakakolwiek ekstrapolacja może okazać się niebezpieczna, gdyż wielomiany wyższych stopni mogą tam znacznie odbiegać od trendu przybliżanej funkcji. Te własności w znacznym stopniu obciążają estymator pogarszając w ten sposób dopasowanie funkcjami sklejonymi.

Pewnym rozwiązaniem dla funkcji sklejanych stopnia trzeciego są *naturalne funkcje sklejane*, które mają nałożone dodatkowe ograniczenia na funkcję poza brzegowymi węzłami. Zakładają bowiem, że funkcja sklejana dla X mniejszych niż ξ_1 i większych niż ξ_K jest funkcją liniową. Takie założenia wynikają z warunków $S''(\xi_1^+) = 0$ i $S''(\xi_K^-) = 0$, które zwalniają dodatkowe cztery stopnie swobody (po dwa warunki w dwóch węzłach). Bardziej opłacalne od założenia wyższego stopnia funkcji wielomianowej poza brzegowymi węzłami, gdzie zazwyczaj znajduje się najmniej informacji, może okazać się wybranie większej liczby węzłów wewnętrz zbioru danych. Zatem w tym wypadku założenie o liniowości funkcji poza brzegowymi węzłami często jest uzasadnione.

Naturalne kubiczne funkcje sklejane z K węzłami są reprezentowane przez K funkcji bazowych. Wyznaczając bazę dla tych funkcji można wyjść od bazy kubicznych funkcji sklejanych i zredukować ją przez nałożenie odpowiednich warunków brzegowych. Naturalną funkcję sklejaną trzeciego stopnia, która jest liniowa dla X mniejszych niż ξ_1 i większych niż ξ_K , można zapisać w postaci szeregu potęgowego:

$$S(X) = \beta_0 + \beta_1 X + \sum_{j=1}^K \gamma_j (X - \xi_j)_+^3. \quad (1.6)$$

Następujące ograniczenia:

$$\sum_{j=1}^K \gamma_j = 0 \quad \sum_{j=1}^K \gamma_j \xi_j = 0, \quad (1.7)$$

dla powyższej funkcji S wynikają z warunku zerowania się drugiej pochodnej, a ich rozwiązanie pozostawia K wolnych parametrów [16].

²Liczba węzłów wynika z prostego rachunku $7 - 3 = 4$, gdzie z `df=7`, natomiast 3 to domyślna wartość parametru `degree` w funkcji `bs`.

1.3. Wygładzone funkcje sklejane

W tym rozdziale została omówiona popularna metoda funkcji sklejanych, która rozwiązuje problem wyboru węzłów występujący w regresji funkcji sklejanych poprzez użycie maksymalnego zbioru węzłów. Stopień dopasowania natomiast jest kontrolowany przez pewien *współczynnik regulujący*. Rozważam następujące zagadnienie: spośród wszystkich funkcji f dwukrotnie różniczkowalnych, chcę znaleźć taką \hat{f} , która minimalizuje poniższą sumę:

$$RSS(f, \lambda) = \sum_{i=1}^N \{y_i - f(x_i)\}^2 + \lambda \int_a^b f''(t)^2 dt, \quad (1.8)$$

gdzie λ to tzw. *współczynnik wygładzający* lub *współczynnik kary* (ang. *smoothing parameter*), natomiast $[a, b]$ jest przedziałem na którym są określone węzły x_i ($i = 0, 1, \dots, N$) tzn. $a \leq x_0 \leq x_1 \leq \dots \leq x_N \leq b$. Pierwszy składnik prawej strony mierzy dopasowanie do danych, zaś drugi uwzględnia *krzywiznę*³ funkcji f . λ natomiast stanowi niejako kompromis między tymi dwoma. Drugi składnik kryterium często jest nazywany karą za niegładkość estymatora, przy czym brak gładkości nie oznacza w tym kontekście braku różniczkowalności estymatora, lecz gwałtowne oscylacyjne zmiany jego wartości. Im większa jest wartość współczynnika λ , tym kara za niegładkość jest większa i wynikowy estymator charakteryzuje się łagodniejszą zmiennością. Wprowadzenie kary opisanego typu nazywa się czasami regularyzacją estymatora funkcji f , [10]. Tak pomyślana regularyzacja pociąga za sobą wygładzenie estymatora, natomiast otrzymany estymator jest nazywany *wygładzoną funkcją sklejaną*.

Szczególne przypadki:

- dla $\lambda \rightarrow 0$ kara nie jest nałożona i wynikowy estymator jest bardzo dopasowany do danych

³Krzywizna krzywej jest to miara stopnia odchylenia danej funkcji od wzorca, [8]. Dla danej krzywej o równaniu $y = g(x)$, $g \in \mathbb{C}^2[a, b]$, krzywizna w punkcie x jest równa:

$$\kappa(x) = \frac{g''(x)}{\sqrt{1 + g'(x)^2}}. \quad (1.9)$$

Funkcja $g \in \mathbb{C}^2[a, b]$ interpolująca punkty (x_i, y_i) ($i = 0, 1, \dots, n$) jest „najgładszą” w tym sensie, że spośród wszystkich funkcji z klasy $\mathbb{C}^2[a, b]$ interpolujących te punkty osiąga najmniejszą wartość tzw. krzywizny całkowej:

$$\int_a^b \kappa(x)^2 dx. \quad (1.10)$$

W praktyce trudno jest szukać rozwiązania zagadnienia interpolacji w postaci funkcji „najgładszej” z powodu skomplikowanego wyrażenia opisującego krzywiznę. Nietrudno zauważać, że dla małych wartości $|g'(x)|$ krzywizna w punkcie x jest w przybliżeniu równa $g''(x)$. Uproszczenie to dostarcza interesującego rozwiązania w postaci funkcji sklejanych trzeciego stopnia, które wystarcza w wielu zastosowaniach, nawet gdy wartości $|g'(x)|$ nie są małe. Kryterium minimalnej krzywizny sprowadza się wówczas do minimalizacji całki:

$$\int_a^b g''(x)^2 dx. \quad (1.11)$$

Będę opierać się na powyższych wnioskach i następującym twierdzeniu, którego dowód można znaleźć w [9]: Jeżeli $g \in \mathbb{C}^2[a, b]$, i s jest funkcją sklejaną trzeciego stopnia interpolującą g w węzłach x_i ($i = 0, 1, \dots, n$), to:

$$\int_a^b [s''(x)]^2 dx \leq \int_a^b [g''(x)]^2 dx. \quad (1.12)$$

Z powyższego można wywnioskować, że w mówiąc o wyborze funkcji „najgładszej” można mieć na myśli minimalizację prawej strony powyższej nierówności.

- dla $\lambda \rightarrow \infty$ kara dominuje i wynikowy estymator charakteryzuje się łagodniejszą zmiennością.

Te dwa przypadki znacznie się różnią, począwszy od bardzo dopasowanego do danych rozwiązania, a skończywszy na bardzo gładkim rozwiążaniu. Rozwiążując problem (1.8) chcę wyznaczyć funkcję f indeksowaną parametrem λ , która stanowi ciekawą klasę funkcji z $\lambda \in (0, \infty)$.

Kryterium (1.8) jest zdefiniowane na przestrzeni funkcji Sobolewa dla których drugi składnik jest określony. Zostało udowodnione [7], że (1.8) posiada wyraźne, skończenie wymiarowe minimum, które jest naturalną funkcją sklejaną trzeciego stopnia o jednoznacznych węzłach w punktach x_i $i = 1, 2, \dots, N$. Wydawałoby się, że ta rodzina funkcji jest nadal zbytnio sparametryzowana, ponieważ posiada N węzłów, które implikują N stopni swobody. Jednakże, składnik kary przekłada się na karę współczynników funkcji sklejanych, które zostaną odpowiednio zmniejszone na drodze liniowego dopasowania.

Rozwiążanie w postaci naturalnych funkcji sklejanych można zapisać następująco:

$$f(x) = \sum_{j=1}^N N_j(x) \theta_j, \quad (1.13)$$

gdzie $N_j(x)$ jest N -wymiarowym zbiorem bazowych funkcji reprezentujących rodzinę naturalnych funkcji sklejanych. Kryterium (1.8) można sprowadzić do postaci:

$$RSS(\theta, \lambda) = (\mathbf{y} - \mathbf{N}\theta)^T (\mathbf{y} - \mathbf{N}\theta) + \lambda \theta^T \boldsymbol{\Omega}_N \theta, \quad (1.14)$$

gdzie $\{\mathbf{N}\}_{ij} = N_j(x_i)$, $\{\boldsymbol{\Omega}_N\}_{ij} = \int N_j''(t) N_k''(t) dt$. Rozwiązaniem zagadnienia (1.14) jest:

$$\hat{\theta} = (\mathbf{N}^T \mathbf{N} + \lambda \boldsymbol{\Omega}_N)^{-1} \mathbf{N}^T \mathbf{y}. \quad (1.15)$$

Można postrzegać je jako rozwiązanie uogólnionej regresji grzbietowej. Dopasowaną w ten sposób wygładzoną funkcją sklejaną będzie wówczas funkcja postaci:

$$\hat{f}(x) = \sum_{j=1}^N N_j(x) \hat{\theta}_j. \quad (1.16)$$

Efektywne techniki obliczania estymatora wygładzonych funkcji sklejanych zostały omówione w podrozdziale 1.6.2.

1.4. Stopnie swobody i macierze wygładzeń

Jak do tej pory nie zostało jeszcze wskazane, w jaki sposób dla wygładzonych funkcji sklejanych jest wyznaczany parametr λ . W tym rozdziale przedyskutuję intuicyjne metody wyboru wielkości wygładzenia.

Wygładzone funkcje sklejane z ustalonym współczynnikiem kary λ są przykładem *linear smoother*, czyli w wolnym tłumaczeniu „liniowego wygładzacz”. Jest tak, ponieważ estymowane parametry w (1.15) są liniową kombinacją y_i . Niech $\hat{\mathbf{f}}$ będzie wektorem długości N , który dla x_i zawiera dopasowane wartości $\hat{f}(x_i)$. Wówczas:

$$\begin{aligned} \hat{\mathbf{f}} &= \mathbf{N}(\mathbf{N}^T \mathbf{N} + \lambda \boldsymbol{\Omega}_N)^{-1} \mathbf{N}^T \mathbf{y} \\ &= \mathbf{S}_\lambda \mathbf{y}. \end{aligned} \quad (1.17)$$

Dopasowanie jest liniowe ze względu na \mathbf{y} , a ograniczony liniowy operator \mathbf{S}_λ znany jest jako *smoother matrix*, czyli tzw. *macierz wygładzenia*. Jednym z wniosków wynikających z tej

liniowości jest sposób wyznaczenia wektora dopasowanych wartości $\hat{\mathbf{f}}$, który nie zależy od \mathbf{y} . \mathbf{S}_λ zależy tylko od x_i i λ .

Operatory liniowe znane są z bardziej tradycyjnego dopasowania metodą najmniejszych kwadratów. Założę, że \mathbf{B}_ξ jest $N \times M$ macierzą M bazowych kubicznych funkcji sklejanych określonych na N punktach x_i , ze zbiorem węzłów ξ , oraz $M \ll N$. Wówczas wektor dopasowanych wartości funkcji sklejanej dany jest przez:

$$\hat{\mathbf{f}} = \mathbf{B}_\xi (\mathbf{B}_\xi^T \mathbf{B}_\xi)^{-1} \mathbf{B}_\xi^T \mathbf{y} = \mathbf{H}_\xi \mathbf{y}. \quad (1.18)$$

Tutaj liniowy operator \mathbf{H}_ξ jest operatorem rzutu i znany jest w statystyce jako „macierz daszkowa” (ang. *hat matrix*). Występują pewne istotne podobieństwa i różnice między \mathbf{H}_ξ i \mathbf{S}_λ :

- obie są symetrycznymi, *dodatnio semi-określonymi*⁴ macierzami,
- $\mathbf{H}_\xi \mathbf{H}_\xi = \mathbf{H}_\xi$ jest *idempotentna*⁵, podczas gdy $\mathbf{S}_\lambda \mathbf{S}_\lambda \preceq \mathbf{S}_\lambda$ znaczy tyle, że prawa strona przewyższa lewą stronę dodatnią semi-określonością macierzy. To jest konsekwencją ściągającej natury \mathbf{S}_λ , która została omówiona poniżej,
- rząd macierzy \mathbf{H}_ξ jest równy M , zaś dla \mathbf{S}_λ wynosi N .

Wyrażenie $M = \text{tr}(\mathbf{H}_\xi)$ wyznacza wymiar przestrzeni rzutowej, który również jest liczbą bazowych funkcji, czyli także liczbą parametrów biorących udział w dopasowaniu. Poprzez analogię definiuje się *efektywne stopnie swobody* (ang. *effective degrees of freedom*) dla wygładzonych funkcji sklejanych:

$$df_\lambda = \text{tr}(\mathbf{S}_\lambda), \quad (1.19)$$

które są sumą diagonalnych elementów \mathbf{S}_λ . Ta bardzo pozytywna definicja pozwala w bardziej intuicyjny sposób parametryzować wygładzone funkcje sklejane. Przykładowo, jeśli zależy nam na uzyskaniu 12 stopni swobody dla dopasowanej krzywej, by wyznaczyć parametr λ należy rozwiązać równanie $\text{tr}(\mathbf{S}_\lambda) = 12$.

Aby potwierdzić jednoznaczność rozwiązania równania (1.19) przytoczę wcześniej kilka istotnych faktów. Jeśli macierz \mathbf{S}_λ jest symetryczna (oraz dodatnio semi-określona), to posiada rzeczywisty rozkład na wartości własne. Zanim dokonam kolejnych obliczeń, zapiszę \mathbf{S}_λ w wygodnej formie jaką zaproponował *Reinsch*, [7]:

$$\mathbf{S}_\lambda = (I + \lambda \mathbf{K})^{-1}, \quad (1.20)$$

gdzie \mathbf{K} nie zależy od λ . Warto podkreślić, że wiele implementacji wygładzonych funkcji sklejanych do zaoszczędzenia na przechowywaniu ogromnych macierzy w pamięci komputera, czy obliczaniu macierzy odwrotnych korzysta z ograniczonego charakteru tych macierzy. Ma to sens jeśli liczba węzłów n jest niewielka lub umiarkowana, [7].

Jeśli $\hat{\mathbf{f}} = \mathbf{S}_\lambda \mathbf{y}$ jest rozwiązaniem

$$\min_{\mathbf{f}} (\mathbf{y} - \mathbf{f})^T (\mathbf{y} - \mathbf{f}) + \lambda \mathbf{f}^T \mathbf{K} \mathbf{f}, \quad (1.21)$$

\mathbf{K} jest znana jako *macierz kary* (ang. *penalty matrix*), i forma kwadratowa \mathbf{K} ma reprezentację w terminach ważonej sumy kwadratów drugiej różniczki⁶. Wówczas rozkład na wartości

⁴Macierz A jest *dodatnio semi-określona* jeżeli $\forall x \in \mathbf{R}^n / \{0\} : f_A(x) \geq 0$ oraz $\exists x \in \mathbf{R}^n / \{0\} : f_A(x) = 0$.

⁵Macierz *idempotentna* to macierz kwadratowa A, która spełnia warunek $A^2 = A$.

⁶Reinsch opierając się na zagadnieniu (1.8) zapisał składnik „kary” jako formę kwadratową:

$$\int (f''(x))^2 dx = \mu^T K \mu,$$

własne macierzy \mathbf{S}_λ jest następujący:

$$\mathbf{S}_\lambda = \sum_{k=1}^N \rho_k(\lambda) \mathbf{u}_k \mathbf{u}_k^T, \quad \text{gdzie } \rho_k(\lambda) = \frac{1}{1 + \lambda d_k}, \quad (1.22)$$

natomiast d_k jest odpowiednią wartością własną \mathbf{K} . Jeśli relacja (1.19) jest monotoniczna ze względu na λ , można ją odwrócić, zatem dla ustalonego df istnieje jednoznaczne rozwiązanie dla λ .

1.5. Wybór parametru wygładzenia

W regresji funkcji sklejanych parametrami były: stopnień splajnów, liczba i rozmieszczenie węzłów. W przypadku wygładzonych funkcji sklejanych występuje tylko jeden parametr, którym jest współczynnik kary λ , ponieważ węzłami są wszystkie unikalne wartości X , a stopień funkcji jest prawie zawsze taki sam jak dla kubicznych funkcji sklejanych.

Ustalenie liczby stopni swobody w przypadku wygładzonych funkcji sklejanych zwykle odbywa się z wykorzystaniem relacji $df_\lambda = \text{tr}(S_\lambda)$. Jeśli jest ona monotoniczna ze względu na λ , można odwrócić relację i określić λ przez ustalenie df . W praktyce może to być osiągnięte z użyciem prostych metod numerycznych. Przykładowo, aby określić wielkość wygładzenia w środowisku R, można użyć funkcji `smooth.spline(x, y, df = 6)`. To pozwala w bardziej tradycyjny sposób wybierać model, a także umożliwia użycie kilku różnych wartości `df`, i wyboru jednej opartej np. na F-teście lub wykresach reszt. Wyznaczanie `df` w ten sposób zapewnia jednakowe podejście do porównania wielu różnych metod wygładzania. Jest to szczególnie przydatne w uogólnionych modelach addytywnych (rozdział 3.4) oraz uogólnionych modelach addytywnych z parametrem położenia, skali i kształtu (rozdział 4), gdzie w jednym modelu może zostać użyte jednocześnie kilka metod wygładzających.

1.6. Jak zapisać funkcje sklejane za pomocą B-splajnów?

W tym podrozdziale opiszę bazę *B-splajnów* tzn. *układ B-splajnów* reprezentujący wielomianowe funkcje sklejane, a następnie omówię jej użycie w wyznaczaniu wygładzonych funkcji sklejanych.

gdzie $\mu = f(x_i)$ to dopasowanie, $K = \Delta' W^{-1} \Delta$, Δ jest macierzą $(N - 2) \times N$ drugiej różniczki z elementami:

$$\Delta_{ii} = \frac{1}{h_i}, \quad \Delta_{i,i+1} = -\frac{1}{h_i} - \frac{1}{h_{i+1}}, \quad \Delta_{i,i+2} = \frac{1}{h_{i+1}},$$

W jest symetryczną, trójdziagonalną macierzą rzędu $N - 2$ z elementami

$$W_{i-1,i} = W_{i,i-1} = -\frac{h_i}{6}, \quad w_{i,i} = \frac{h_i + h_{i+1}}{3}$$

i $h_i = x_{i+1} - x_i$ to odległości pomiędzy kolejnymi wartościami x .

Stwierdzenie: W powyższych oznaczeniach wygładzoną funkcję sklejaną można zapisać:

$$\mathbf{f} = (I + \lambda K)^{-1} \mathbf{y} \quad **$$

Dowód: RSS można zapisać jako: $RSS = (y - \mu)^T (y - \mu) + \lambda \mu^T K \mu$. Różniczkując ze względu na μ dostajemy: $\frac{dRSS}{d\mu} = -2(y - \mu) + 2\lambda K \mu$. Przyrównując prawą stronę do zera otrzymujemy: $y = \hat{\mu} + \lambda K \hat{\mu} = (I + \lambda K) \hat{\mu}$. Po obustronnym przemnożeniu przez $(I + \lambda K)^{-1}$ otrzymujemy **, [16].

1.6.1. B-splajny

Zanim zdefiniuję postać funkcji B-sklejanych, wprowadzę kilka niezbędnych oznaczeń. Niech $\xi_0 < \xi_1 < \dots < \xi_K < \xi_{K+1}$ będą dwoma *brzegowymi węzłami* (ang. *boundary knots*) definiującymi przedział na którym konstruowane będą funkcje sklejane. Przez $B_{i,m}(x)$ ⁷ oznaczę i-tą bazową funkcję *B-sklejaną* (inaczej *B-splajn*) stopnia m dla ciągu węzłów ξ , $m \leq M$. B-splajny definiuje się następująco wykorzystując rekurencję:

$$B_{i,1}(x) = \begin{cases} 1 & \text{jeżeli } \xi_i \leq x \leq \xi_{i+1} \\ 0 & \text{w.p.p.} \end{cases} \quad (1.23)$$

dla $i = 1, \dots, K + 2M - 1$,

$$B_{i,m}(x) = \frac{x - \xi_i}{\xi_{i+m-1} - \xi_i} B_{i,m-1}(x) + \frac{\xi_{i+m} - x}{\xi_{i+m} - \xi_{i+1}} B_{i+1,m-1}(x) \quad (1.24)$$

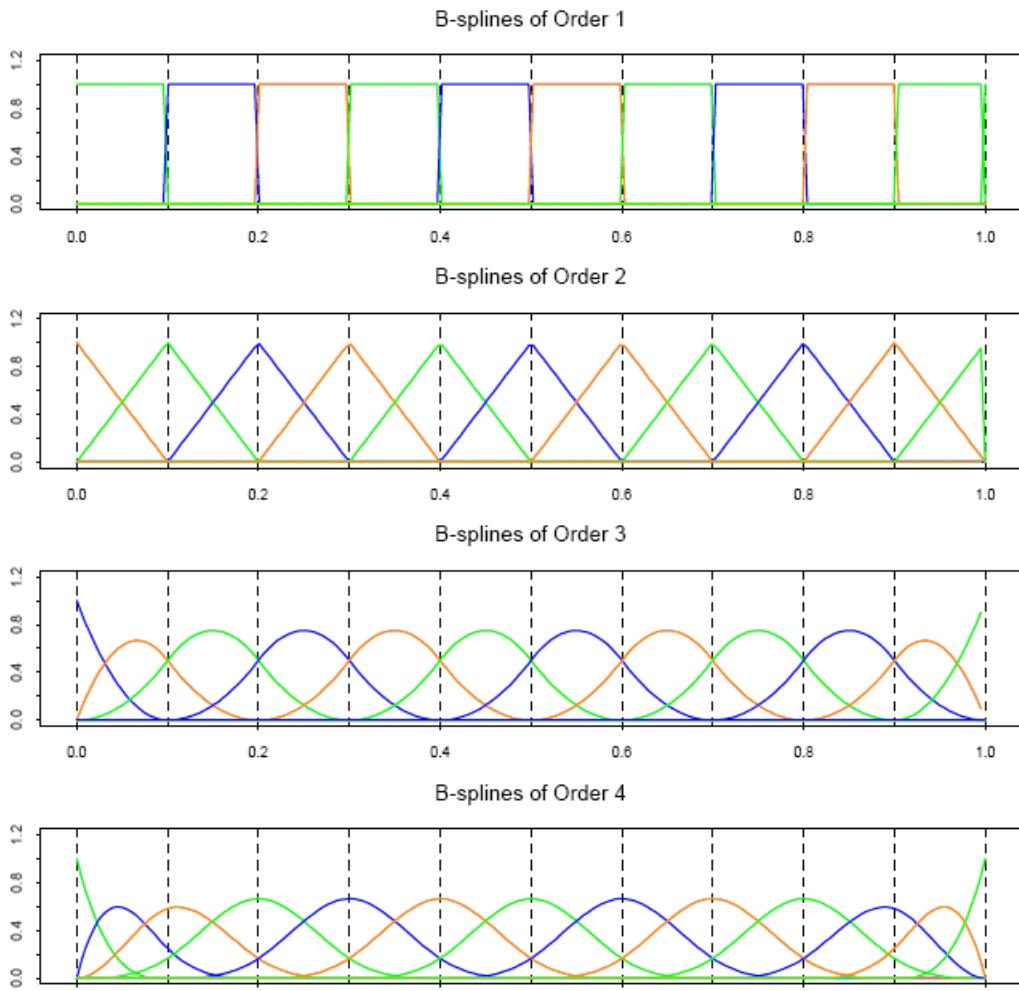
dla $i = 1, \dots, K + 2M - m$. Dla ciągu węzłów ξ oraz $M = 4$ w powyższych definicjach $B_{i,4}$, $i = 1, \dots, K + 4$ tworzą $(K + 4)$ -bazowe kubiczne funkcje B-splajnów. Kontynuując rekurencję można wygenerować bazę B-splajnów dla funkcji sklejanych dowolnego stopnia. Wykresy z rysunku 1.2 przedstawiają ciąg B-splajnów do czwartego stopnia z węzłami w punktach $0, 0.1, \dots, 0.9, 1$. Jeśli występują zduplikowane punkty, trzeba zachować ostrożność by uniknąć dzielenia przez zero w (1.24). Przyjmując, że $B_{i,1} = 0$ jeśli $\xi_i = \xi_{i+1}$, przez indukcję można pokazać, że $B_{i,m} = 0$ jeśli $\xi_i = \xi_{i+1} = \dots = \xi_{i+m}$. Warto zauważyć, że do konstrukcji bazy B-splajnów stopnia $m < M$ z węzłami ξ , wymagane są funkcje $B_{i,m}$, dla $i = M-m+1, \dots, M+K$.

Własności funkcji $B_{i,M}$:

1. Nośnik funkcji $B_{i,M}$, czyli zbiór tych wszystkich x , gdzie $B_{i,M}(x) > 0$ jest przedziałem (ξ_i, ξ_{i+M}) .
2. $B_{i,M}(x) \geq 0, \forall i$ oraz $\forall x$.
3. $\sum_{i=1}^{K+M} B_{i,M}(x) = 1 \quad \forall x \in [\xi_0, \xi_{K+1}]$.
4. $B_{i,M}$ jest funkcją kawałkami wielomianową stopnia $M - 1$ z węzłami w ξ_1, \dots, ξ_K .
5. Dla $k \geq 1$ funkcje $B_{i,M}$ należą do klasy $C^{m-2}(R)$.

Zastosowanie B-splajnów należy osobno rozważyć w przypadku występowania powielonych wewnętrznych węzłów. Jeśli jeden z wewnętrznych węzłów w konstrukcji ξ zostanie podwojony i następnie jak wcześniej wygenerowany zostanie ciąg B-splajnów, wówczas otrzymana baza funkcji będzie rozpinała przestrzeń kawałkami wielomianowych funkcji, które w zduplikowanym punkcie mają nieciągłą pochodną niższego rzędu o jeden w porównaniu do przypadku bez powielonego węzła. Ogólnie, jeśli r_j -krotnie ($1 \leq r_j \leq M$) zostanie dołączony wewnętrzny węzeł ξ_j , wtedy pochodną najniższego rzędu, nieciągłą w $x = \xi_j$ będzie pochodna rzędu $M - r_j$. Zgodnie z powyższym, kubiczne funkcje sklejane bez powielonych punktów $r_j = 1, j = 1, \dots, K$, w każdym wewnętrzny węzle mają nieciągłą trzecią pochodną (4-1). Powielenie j-tego węzła trzykrotnie prowadzi do nieciągłości pierwszej pochodnej, natomiast 4-krotne nadpisanie go prowadzi do nieciągłości zerowej pochodnej, i wówczas funkcja

⁷ $B_{i,1}(x)$ dla $i = 1, \dots, K + 2M - 1$ znane są jako funkcje bazy Haara.



Rysunek 1.2: Kolejne B-splajny do czwartego stopnia dla równo rozmieszczonych dziesięciu węzłów na odcinku $[0, 1]$. Widać, że lokalnie są niezerowe na przestrzeni rozpiętej przez $M + 1$ węzłów. Źródło: [4].

będzie nieciągła w $x = \xi_j$. Jest to dokładnie ten sam przypadek, który pojawia się w brzegowym węźle po M -krotnym nadpisaniu tego węzła. Funkcja sklejana staje się nieciągła w brzegowym węźle (tzn. nie jest zdefiniowana poza brzegiem).

Baza B-splajnów ma ważne obliczeniowo implikacje, szczególnie gdy liczba węzłów K jest duża. Metoda najmniejszych kwadratów z N obserwacjami i $K + M$ zmiennymi (funkcjami bazowymi) wykonuje $O(N(K + M)^2 + (K + M)^3)$ operacji. Jeśli K jest znacznie mniejsze niż N , to algorytm staje się nie do przyjęcia dla dużych $N - O(N^3)$. Jeśli będziemy opisywać za pomocą $K + M$ -bazowych funkcji B-splajnów N -posortowanych rosnąco unikalnych obserwacji (N -węzłów), to otrzymamy macierz regresji rozmiaru $N \times (K + M)$, która będzie mieć wiele zer, a to z kolei może przyczynić się do zredukowania złożoności obliczeniowej do $O(N)$. Ta własność została wykorzystana w następnym podrozdziale.

1.6.2. Jak zapisać wygładzone funkcje sklejane za pomocą B-splajnów?

Chociaż naturalne funkcje sklejane tworzą bazę dla wygładzonych funkcji sklejanych, obliczeniowo wygodniej jest operować na większej przestrzeni B-splajnów (rozdział 1.3). Wygła-

dzoną funkcję sklejaną można zapisać $f(x) = \sum_{j=1}^{N+4} \gamma_j B_j(x)$, gdzie γ_j są współczynnikami i B_j to baza kubicznych funkcji B-sklejanych. Rozwiążanie wygląda tak samo jak poprzednio:

$$\hat{\boldsymbol{\gamma}} = (\mathbf{B}^T \mathbf{B} + \lambda \boldsymbol{\Omega}_B)^{-1} \mathbf{B}^T \mathbf{y}. \quad (1.25)$$

Jednym wyjątkiem jest zastąpienie macierzy \mathbf{N} rozmiaru $N \times N$, macierzą \mathbf{B} o rozmiarze $N \times (N+4)$, i podobnie macierz kary $\boldsymbol{\Omega}_B$ rozmiaru $(N+4) \times (N+4)$ zastępuje się macierzą $\boldsymbol{\Omega}_N$ o rozmiarze $N \times N$. Tak jak wcześniej można również zdefiniować efektywne stopnie swobody przez $df_\lambda = \text{tr}(\mathbf{W})$, gdzie $\mathbf{W} = \mathbf{B}(\mathbf{B}^T \mathbf{B} + \lambda \boldsymbol{\Omega}_B)^{-1} \mathbf{B}^T$.

Jeśli kolumny macierzy \mathbf{B} są wyznaczone przez B-splajny, które w kolejności od lewej do prawej, zostały oszacowane na podstawie posortowanych wartości X , oraz kubiczne B-splajny mają lokalnie nośnik, wówczas macierz \mathbf{B} jest dolnie 4-trójkątna. W konsekwencji macierz $\mathbf{M} = (\mathbf{B}^T \mathbf{B} + \lambda \boldsymbol{\Omega})$ jest dolnie 4-trójkątną macierzą zatem łatwo można dokonać jej rozkładu Cholesky'ego $\mathbf{M} = \mathbf{L}\mathbf{L}^T$. Rozwiążując $\mathbf{L}\mathbf{L}^T \boldsymbol{\gamma} = \mathbf{B}^T \mathbf{y}$ przez podstawienie otrzymujemy $\boldsymbol{\gamma}$ i rozwiązanie $\hat{\mathbf{f}}$ w $O(N)$ operacjach.

W praktyce gdy N jest duże, nie jest konieczne używanie wszystkich N wewnętrznych węzłów lecz rozsądne „przerzadzenie” pozwoli zaoszczędzić na obliczeniowości, a także będzie mieć niewielki wpływ na dopasowanie. Dla przykładu funkcja `smooth.spline` środowiska R korzysta z następującej funkcji obliczającej liczbę węzłów:

```
> n.kn <- function(n) {
+   if (n < 50L)
+     n
+   else trunc(
+     a1 <- log(50, 2)
+     a2 <- log(100, 2)
+     a3 <- log(140, 2)
+     a4 <- log(200, 2)
+     if (n < 200L) 2^(a1 + (a2 - a1) * (n - 50)/150)
+     else if (n < 800L) 2^(a2 + (a3 - a2) * (n - 200)/600)
+     else if (n < 3200L) 2^(a3 + (a4 - a3) * (n - 800)/2400)
+     else 200 + (n - 3200)^0.2
+   )
+ }
```

Zgodnie z powyższym, dla danej liczby `n` unikalnych wartości wektora `x`, `n.kn(n)` stanowi liczbę węzłów używaną w estymowaniu wygładznej kubicznej funkcji sklejanej, np. `n.kn(49) = 49` lub `n.kn(5000) = 204`, [23].

Rozdział 2

Funkcje sklejane w pakiecie R i ich zastosowanie

W tym rozdziale zostały przedstawione wybrane funkcje środowiska R wykorzystywane do wyznaczania funkcji sklejanych i wygładzonych funkcji sklejanych. Została omówiona budowa funkcji **bs** i **ns** oraz **smooth.spline**, przyjmowane przez nie argumenty, zwracane wartości, a także zastosowanie tych funkcji do rzeczywistych danych medycznych.

2.1. Wielomianowe funkcje sklejane – **bs(splines)** i **ns(splines)**

Do wyznaczenia wielomianowych funkcji sklejanych w bazie B-splajnów można wykorzystać funkcję **bs()**, natomiast za wyznaczenie naturalnych kubicznych funkcji sklejanych w bazie B-splajnów odpowiada funkcja **ns()**. Obie te funkcje znajdują się w pakiecie **splines**.

Deklaracje funkcji **bs(splines)** i **ns(splines)** wyglądają następująco:

```
bs(x, df = NULL, knots = NULL, degree = 3, intercept = FALSE,  
Boundary.knots = range(x))
```

```
ns(x, df = NULL, knots = NULL, intercept = FALSE, Boundary.knots =  
range(x))
```

Argumenty dla obu funkcji opisane zostały w tabeli (2.1).

Obie powyższe funkcje zwracają w postaci listy następujące wartości: macierz regresji rozmiaru $\text{length}(x) \times df$ oraz **knots** i **Boundary.knots**, które wykorzystywane są przez funkcje **predict.bs** i **predict.ns**. Gdy argumenty **df** lub **knots** są podane, wówczas dla funkcji **bs()** obowiązuje zależność:

$$df = \text{length}(knots) + \text{degree} + \mathbb{1}_{\text{intercept}=TRUE},$$

natomiast dla **ns()** mam:

$$df = \text{length}(knots) + 1 + \mathbb{1}_{\text{intercept}=TRUE}.$$

Funkcja **bs** generuje macierz regresji w bazie B-splajnów reprezentującą rodzinę kawałkami wielomianowych funkcji sklejanych z wyszczególnionymi wewnętrznymi węzłami i stopniem wielomianu. Funkcja **ns** generuje macierz regresji w bazie B-splajnów reprezentującą rodzinę kawałkami kubicznych funkcji sklejanych, z wyszczególnionym zbioremewnętrznych węzłów, i naturalnymi warunkami brzegowymi. Dodatkowo funkcja **ns** wymusza ograniczenie, aby wynikowa funkcja była liniowa poza brzegowymi węzłami.

Zazwyczaj funkcje **bs()** i **ns()** wykorzystuje się tworząc formułę modelu liniowego przez bezpośrednie dodanie składnika **bs()** lub **ns()** do formuły danego modelu.

Tabela 2.1: Argumenty funkcji `bs()` i `ns()`, Źródło [23].

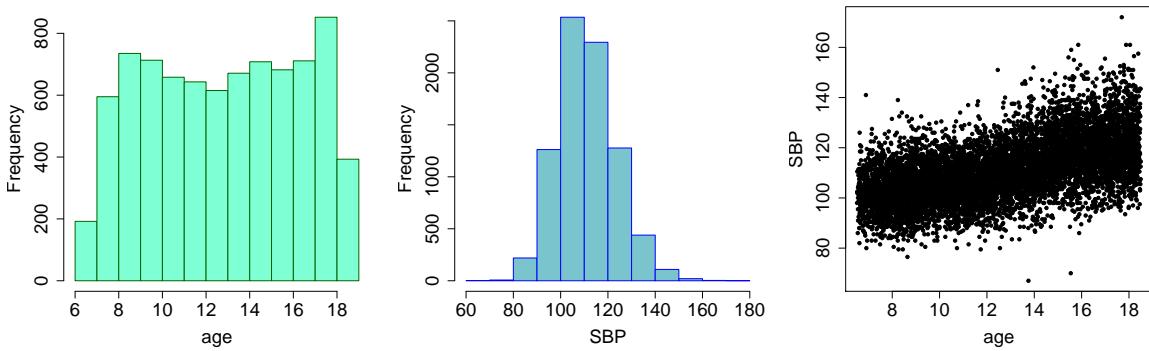
<code>x</code>	predyktor
<code>df</code>	odpowiada liczbie stopni swobody; w przypadku funkcji <code>bs</code> , gdy podana jest wielkość <code>df</code> zamiast <code>knots</code> , liczbę <code>knots</code> wyznacza zależność $\text{knots} = \text{df} - \text{degree} - 1$ dla <code>TRUE</code> ; dla <code>ns</code> , gdy podawana jest wielkość <code>df</code> , <code>ns</code> wyznacza liczbę <code>knots</code> = $\text{df} - 1 - 1$ dla <code>TRUE</code> ; dla obu funkcji węzły są rozmieszczone w odpowiednich kwantylach <code>x</code> .
<code>knots</code>	dla <code>bs</code> są to pewne wewnętrzne punkty definiujące splajn wybrane z uporządkowanego niemalejąco zbioru punktów zawartych w <code>x</code> . Domyslnie jest to <code>NULL</code> , co daje podstawę do zwykłej wielomianowej regresji. Dla <code>ns</code> są to wewnętrzne punkty uporządkowanego <code>x</code> ; domyslnie jest to <code>NULL</code> , i razem z naturalnymi warunkami brzegowymi daje podstawę do zwykłej liniowej regresji na <code>x</code> . Zazwyczaj są to wartości: średnia lub mediana dla jednego węzła, kwantyle dla większej liczby węzłów.
<code>degree</code>	odpowiada stopniowi funkcji kawałkami wielomianowej, domyslnie jak dla kubicznych funkcji sklejanych <code>degree=3</code> .
<code>intercept</code>	jeśli <code>TRUE</code> , stała jest dodana do bazy; domyslnie <code>FALSE</code> .
<code>Boundary.knots</code>	w przypadku <code>bs</code> punkty brzegowe, czyli punkty w zakresie których znajdują się wszystkie węzły dla bazy B-splajnów (domyslnie zakres danych). W przypadku <code>ns</code> są to brzegowe punkty z narzuconymi warunkami brzegowymi dla naturalnych splajnów i węzłami w bazie B-splajnów (domyslnie zakres danych). Dla obu funkcji jeśli oba parametry <code>knots</code> i <code>Boundary.knots</code> są podane, bazowe parametry nie zależą od <code>x</code> . Dane mogą być przedłużone na <code>Boundary.knots</code> .

2.1.1. Przykłady zastosowania funkcji `bs` i `ns`

Pokażę zastosowanie funkcji `bs` i `ns` do danych `datemed` opisanych na końcu pracy w dodatku A. Korzystając z danych dla chłopców (`sex=1`), zweryfikuję zależność ciśnienia skurczowego (SBP) od wieku (`age`). Obie zmienne `age` oraz SBP uwzględnione w analizie są zmiennymi ciągłymi. Opisany zbiór liczy 8168 przypadków, zmienna `age` w badanej grupie osób jest zawarta w przedziale od 6.5 – 18.5, natomiast ciśnienie skurczowe waha się od 67 do 172. Wykres 2.1 przedstawia kolejno histogramy zmiennej `age` i SBP oraz zależność SBP od `age`.

Dla zmiennej `age` wywołuję funkcję `bs` z parametrem `df=5`:

```
> bs(age, df = 5)
      1          2          3          4
5
[1,] 0.3163598 0.02331078 3.859730e-04 0.000e+00 0.000e+00
[2,] 0.2816437 0.01764678 2.498074e-04 0.000e+00 0.000e+00
...
[8167,] 0.000e+00 1.19831e-05 2.64252e-03 1.24006e-01 8.73338e-01
[8168,] 0.000e+00 4.21258e-04 2.66376e-02 3.47029e-01 6.25911e-01
attr(",degree")
[1] 3
attr(",knots")
33.33333% 66.66667%
10.74606 14.89938
```



Rysunek 2.1: a). Histogram zmiennej `age` b). Histogram zmiennej `SBP` c). `age` vs. `SBP`.

```
attr(,"Boundary.knots")
[1] 6.513347 18.496920
attr(,"intercept")
[1] FALSE
attr(,"class")
[1] "bs"      "basis"   "matrix"
```

Powyższe wywołanie odpowiada za utworzenie macierzy regresji rozmiaru 8168×5 . Zgodnie ze wzorem na liczbę wewnętrznych węzłów zostały ustalone 2 wewnętrzne węzły. Znajdują się one w kwantylach rzędu $1/3$ i $2/3$ oraz odpowiadają wartościom 10.74, 14.89. Brzegowe węzły znajdują się w 6.51 i 18.49.

Następnie wywołuję funkcję `ns()` dla zmiennej `age` z parametrem `df=5`:

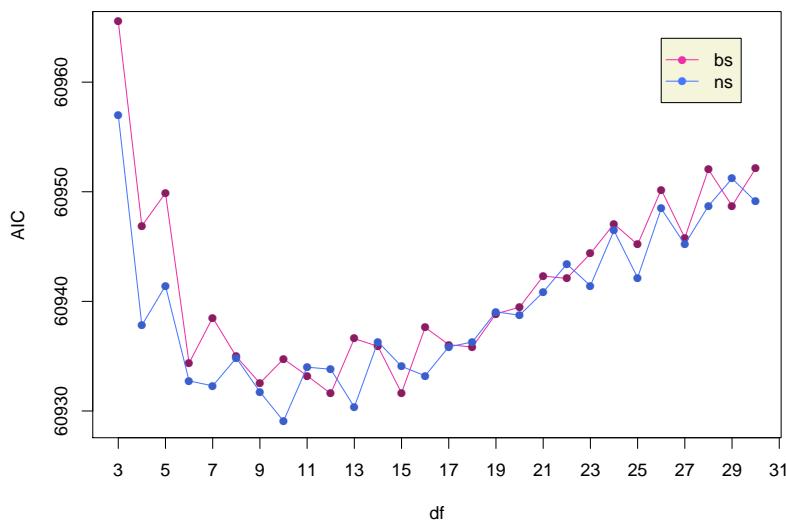
```
> ns(age, df = 5)
           1          2          3          4          5
[1,] 0.001610455 0.0000e+00 -0.04996063 0.15844657 -0.10848594
[2,] 0.001042310 0.0000e+00 -0.04338198 0.13758286 -0.09420087
...
[8167,] 0.0000e+00 6.599770e-05 -5.76155e-02 0.41902256  0.63852694
[8168,] 0.0000e+00 2.320094e-03  1.20156e-01 0.37912975  0.49839382
attr(,"degree")
[1] 3
attr(,"knots")
 20%      40%      60%      80%
9.14716 11.59699 14.12731 16.48460
attr(,"Boundary.knots")
[1] 6.513347 18.496920
attr(,"intercept")
[1] FALSE
attr(,"class")
[1] "ns"      "basis"   "matrix"
```

Powyższe wywołanie także odpowiada za utworzenie macierzy regresji rozmiaru 8168×5 . Porównując to wywołanie z wywołaniem funkcji `bs()`, teraz wzrosła liczba wewnętrznych węzłów (o dwa węzły). Zostały uzyskane 4 węzły wewnętrzne i znajdują się one w kwantylach rzędu $1/5$, $2/5$, $3/5$ i $4/5$, czyli odpowiednio w 9.14, 11.59, 14.12 oraz 16.48. Brzegowe węzły podobnie jak poprzednio to: 6.51 i 18.49. Można zauważyć, że uwzględnienie tej samej liczby stopni swobody (`df = 5`), dla `bs` i `ns`, sprawia, że w przypadku `ns` można uzyskać cztery

wewnętrzne węzły, natomiast dla `bs` – tylko dwa. Większa liczba węzłów to zwykle lepsze dopasowanie modelu.

Tworzę modele liniowe, w których zmienną odpowiedzi jest SBP, zaś zmienną objaśniającą będzie wynik wywołania funkcji `bs(age)` i `ns(age)` z różnymi wartościami parametru `df`. Parametry będą wybierane spośród kolejnych liczb całkowitych z przedziału od 3 do 30. Dokonam zatem przybliżenia zmiennej SBP z wykorzystaniem kubicznych funkcji sklejanych i naturalnych kubicznych funkcji sklejanych, które zależeć będą od zmiennej `age`. Porównam dopasowanie tych modeli przy pomocy kryterium AIC, (dodatek B.1).

```
> bm3 <- lm(SBP ~ bs(age, df=3), data = danemed)
> bm4 <- lm(SBP ~ bs(age, df=4), data = danemed)
...
> bm30 <- lm(SBP ~ bs(age, df=30), data = danemed)
> ns3 <- lm(SBP ~ ns(age, df=3), data = danemed)
> ns4 <- lm(SBP ~ ns(age, df=4), data = danemed)
...
> ns30 <- lm(SBP ~ ns(age, df=30), data = danemed)
```



Rysunek 2.2: Wartości kryterium AIC dla modeli utworzonych z wykorzystaniem funkcji `bs` i `ns` z uwzględnieniem różnych wartości dla parametrów `df`. Źródło: opracowanie własne.

Na podstawie rys. 2.2 można wybrać „najlepszy” model. Kierując się kryterium AIC, dla metody kubicznych funkcji sklejanych model z `df=12` wydaje się być najlepiej dopasowany.

```
> summary(lm(SBP ~ bs(age, df=12), data = danemed))

Call:
lm(formula = SBP ~ bs(age, df = 12), data = danemed)

Residuals:
    Min      1Q  Median      3Q     Max 
-49.0667 -6.9538 -0.5463  6.5036 50.7335 

Coefficients:
            Min      1Q  Median      3Q     Max 
-49.0667 -6.9538 -0.5463  6.5036 50.7335
```

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept) 100.8724   1.4227  70.904 < 2e-16 ***
bs(age, df = 12)1  2.4036   2.5512  0.942  0.34615
bs(age, df = 12)2 -0.3146   1.4539 -0.216  0.82867
bs(age, df = 12)3  4.3864   1.7841  2.459  0.01397 *
bs(age, df = 12)4  4.5698   1.5541  2.941  0.00329 **
bs(age, df = 12)5  5.5185   1.6833  3.278  0.00105 **
bs(age, df = 12)6  9.4868   1.6185  5.862  4.76e-09 ***
bs(age, df = 12)7 11.7423   1.6440  7.143  9.95e-13 ***
bs(age, df = 12)8 19.5873   1.6357 11.975 < 2e-16 ***
bs(age, df = 12)9 17.0905   1.6849 10.143 < 2e-16 ***
bs(age, df = 12)10 20.1888   1.7771 11.360 < 2e-16 ***
bs(age, df = 12)11 21.3043   1.8589 11.461 < 2e-16 ***
bs(age, df = 12)12 19.8831   1.9774 10.055 < 2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.07 on 8155 degrees of freedom
Multiple R-squared: 0.3259, Adjusted R-squared: 0.3249
F-statistic: 328.5 on 12 and 8155 DF, p-value: < 2.2e-16

```

W ramce współczynników dla modelu z $df=12$ zostały wypisane estymatory $\hat{\beta}$ dla 12 nowych zmiennych zapisanych przy pomocy bazy kubicznych B-splajnów razem z wynikiem testu istotności każdego z nich. Na analizowany obszar składało się 10 przedziałów, które wydzielone zostały przez 9 wewnętrznych węzłów. Na podstawie p-wartości można uznać istotność nowych zmiennych zapisanych w bazie kubicznych funkcji sklejanych.

Korzystając z metody naturalnych kubicznych funkcji sklejanych i kryterium AIC za najlepszy model uznaję model z $df=10$:

```

> summary(lm(SBP ~ ns(age, df=10), data = danemed))

Call:
lm(formula = SBP ~ ns(age, df = 10), data = danemed)

Residuals:
    Min      1Q      Median      3Q      Max 
-49.073 -6.927 -0.533  6.531 50.799 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 102.0758   0.8490 120.227 < 2e-16 ***
ns(age, df = 10)1  2.7756   0.9299  2.985 0.002844 ** 
ns(age, df = 10)2  3.5970   1.2425  2.895 0.003804 ** 
ns(age, df = 10)3  4.1724   1.1188  3.730 0.000193 *** 
ns(age, df = 10)4  8.3882   1.2054  6.959 3.69e-12 *** 
ns(age, df = 10)5 10.4361   1.1445  9.119 < 2e-16 *** 
ns(age, df = 10)6 18.5294   1.1722 15.807 < 2e-16 *** 
ns(age, df = 10)7 15.6320   1.1638 13.432 < 2e-16 *** 
ns(age, df = 10)8 19.6345   0.9184 21.379 < 2e-16 *** 
ns(age, df = 10)9 18.8722   2.0547  9.185 < 2e-16 *** 
ns(age, df = 10)10 19.7454   0.8483 23.275 < 2e-16 *** 
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.07 on 8157 degrees of freedom

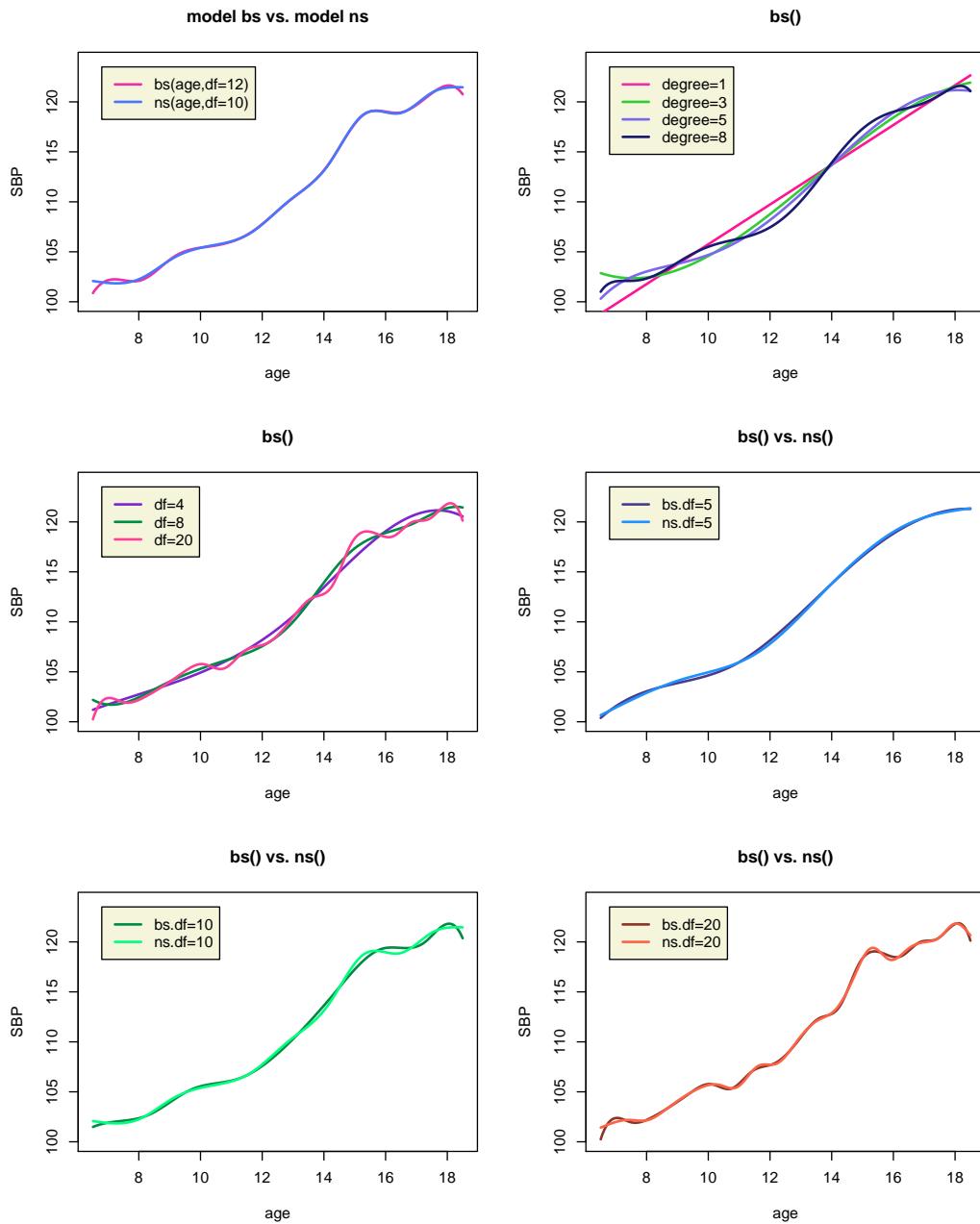
```

```

Multiple R-squared:  0.3258 ,   Adjusted R-squared:  0.3249
F-statistic: 394.1 on 10 and 8157 DF,  p-value: < 2.2e-16

```

Tutaj na analizowany obszar składa się 10 przedziałów wydzielonych przez 9 wewnętrznych węzłów. P-wartości potwierdzają istotność nowych zmiennych utworzonych przez transformację przy pomocy kubicznych funkcji B-sklejanych. Rysunek 2.3. (lewy górnny) przedstawia porównanie dopasowanych wartości dla powyższych modeli.



Rysunek 2.3: Porównanie dopasowanych wartości dla modeli ze składnikami `bs` i `ns` wywołanych z różnymi parametrami. Źródło: opracowanie własne.

Wykresy zawarte na rys. 2.3 zawierają wyniki dopasowania uzyskane z modeli, w których zostały użyte funkcje `bs` i `ns` uwzględniające różne wartości parametrów dla tych funkcji.

Na podstawie wykresów można stwierdzić, że w krańcowym zakresie danych naturalne kubiczne funkcje klejane zazwyczaj zachowują się bardziej łagodnie niż zwykłe kubiczne funkcje klejane. Zwiększając liczbę stopni swobody wzrasta zmienność dopasowanej krzywej. Dla danych dotyczących ciśnienia skurczowego chłopców nie widać jednak znaczących różnic pomiędzy krzywymi otrzymanymi dla różnych parametrów **df**.

2.2. Wygładzone funkcje klejane - **smooth.spline(stats)**

Smoothing spline, czyli wygładzanie splajnami jest to metoda polegająca na dopasowaniu wygładzonej krzywej do zbioru pewnych zakłóconych obserwacji za pomocą kubicznych funkcji B-sklejanych. Podejście matematyczne do wyznaczenia wygładzonego estymatora za pomocą funkcji klejanych zostało omówione w podrozdziałach 1.3-1.5. Popularną funkcją środowiska R wykorzystywaną w tej metodzie jest dostępna w bibliotece **stats** funkcja **smooth.spline**.

Deklaracja funkcji **smooth.spline** wygląda następująco:

```
smooth.spline(x, y = NULL, w = NULL, df, spar = NULL, cv = FALSE,
all.knots = FALSE, nknots = NULL, keep.data = TRUE, df.offset = 0,
penalty = 1, control.spar = list())
```

Wybrane parametry tej funkcji zostały opisane w tabeli 2.2.

Tabela 2.2: Argumenty funkcji **smooth.spline**, Źródło: [23]

x	predyktor, jeśli x jest dwukolumnową macierzą, to zawiera x i y
y	wektor zmiennej odpowiedzi
w	wektor wag, domyślnie w = 1
df	liczba stopni swobody, inaczej ślad <i>macierzy wygładzenia</i>
spar	parametr wygładzenia przyjmujący zazwyczaj wartości z przedziału $(0, 1]$; wyznaczenie współczynnika kary λ opiera się na obliczeniu całki kwadratu drugiej pochodnej, która jest monotoniczną funkcją argumentu spar , szczebeli poniżej
cv	gdy cv=TRUE zwykła walidacja krzyżowa leave-one-out, natomiast domyślnie cv=FALSE oznacza uogólnioną walidację krzyżową (GCV)
all.knots	jeśli przyjmuje wartość TRUE , każdy punkt x jest traktowany jako węzeł; domyślne all.knots=FALSE oznacza, że węzłami jest pewien podzbiór zbioru x
nknots	liczba węzłów, jeśli all.knots=F , domyślnie liczba mniejsza niż wymiar x
penalty	współczynnik kary dla stopni swobody w kryterium GCV (domyślnie 1)

Współczynnik kary λ występujący w kryterium (1.8) z rozdziału 1.3 wyrażony jest tutaj jako funkcja zmiennej **spar**:

$$\lambda = r256^{3spar-1}, \quad (2.1)$$

gdzie $r = \frac{tr(\mathbf{B}^T \mathbf{WB})}{tr(\Omega)}$, \mathbf{B} to macierz wyznaczona przez $\{\mathbf{B}\}_{ij} = B_j(x_i)$, w której $B_j(\cdot)$ oznacza j-ty B-splajn, \mathbf{W} jest macierzą diagonalną z wagami o śladzie równym liczbie obserwacji czyli n , Ω_B jest macierzą $\{\Omega_B\}_{ij} = \int B_i''(t) B_j''(t) dt$.

Kryterium (1.8) zostało tutaj uogólnione i rozszerzone do postaci ważonej sumy kwadratów reszt ze współczynnikiem kary:

$$WRSS(\theta, \lambda) = (y - \mathbf{B}\theta)^T \mathbf{W}(y - \mathbf{B}\theta) + \lambda \theta^T \Omega_B \theta, \quad (2.2)$$

gdzie $\mathbf{B}\theta$ stanowi reprezentację bazy B-splajnów, θ to rozwiązanie równania regresji grzbietowej:

$$(\mathbf{B}^T \mathbf{W} \mathbf{B} + \lambda \Omega_B) \theta = \mathbf{B}^T \mathbf{W} y. \quad (2.3)$$

Wektor θ zawiera współczynniki funkcji sklejanej.

Jeśli w wywołaniu funkcji `smooth.spline` nie został podany argument `spar` lub `spar=NULL`, wartość wygładzenia zostanie wyznaczona z wykorzystaniem zmiennej `df`. Jeśli żaden z tych argumentów nie został podany, do wyznaczenia λ wykorzystuje się metodę zwyczajnej walidacji krzyżowej *leave-one-out* (ang. *ordinary*) lub uogólnionej (ang. *generalized*). Rodzaj walidacji krzyżowej uwzględnia parametr `cv`.

Problematycznym zagadnieniem w funkcji `smooth.spline` jest występowanie zduplikowanych punktów w wektorze `x` połączone z wyznaczaniem parametru wygładzenia metodą walidacji krzyżowej *leave-one-out*. Podczas gdy metoda uogólnionej walidacji krzyżowej działa poprawnie, walidacja krzyżowa *leave-one-out* nie radzi sobie ze zduplikowanymi punktami – procedura pakietu R używa wówczas przybliżenia, które wiąże się z pominięciem zbioru powielonych punktów. W takich przypadkach należy unikać parametru `cv=TRUE`.

2.2.1. Przykład zastosowania funkcji `smooth.spline`

W tym przykładzie wykorzystam ten sam zbiór danych, który uwzględniony był w przykładzie z podrozdziału 2.1.1. Dla dzieci i młodzieży płci męskiej wyznaczę estymator ciśnienia skurczowego (SBP) z wykorzystaniem wygładzonych funkcji sklejanych, który będzie zależał od wieku (`age`). Lewy rysunek 2.4 przedstawia wykres zależności ciśnienia skurczowego od zmiennej (`age`), która zawiera wiek dziecka wyrażony zmienną ciągłą z dokładnością do części dziesiętnych. Wywołując procedurę `smooth.spline` z domyślnymi parametrami, `df=40` i `df=3` tworzę trzy modele wyznaczające wygładzoną funkcję sklejaną dla zmiennej niezależnej `age` i objaśnianej SBP.

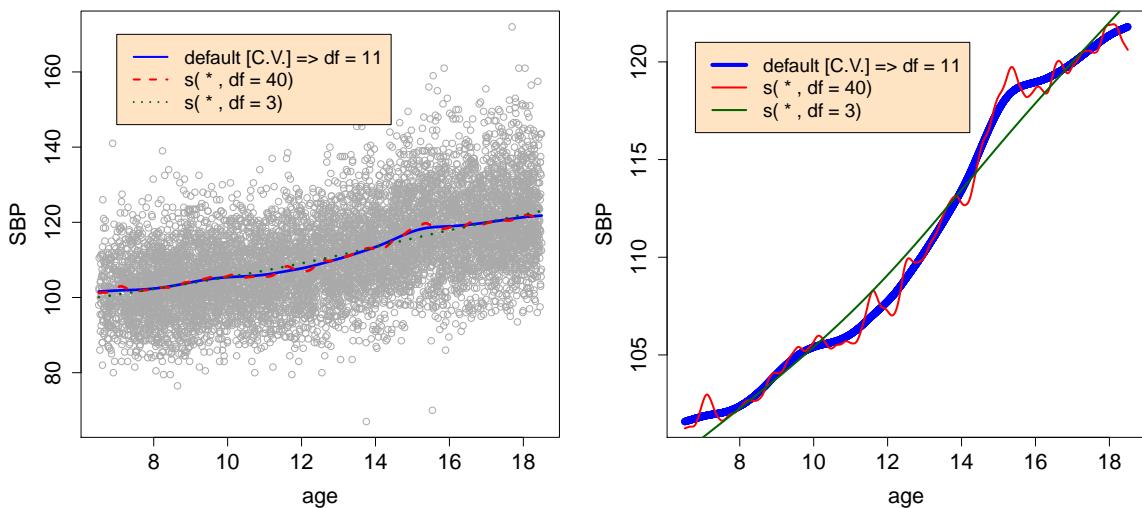
```
> (boys.spl <- smooth.spline(danemed$age, danemed$SBP))
Call:
smooth.spline(x = danemed$age, y = danemed$SBP)

Smoothing Parameter spar= 0.9108295 lambda= 0.01244968 (10 iter.)
Equivalent Degrees of Freedom (Df): 10.9875
Penalized Criterion: 362251.2
GCV: 101.6346
> (boys40.spl<-smooth.spline(danemed$age, danemed$SBP, df=40))
Call:
smooth.spline(x = danemed$age, y = danemed$SBP, df = 40)

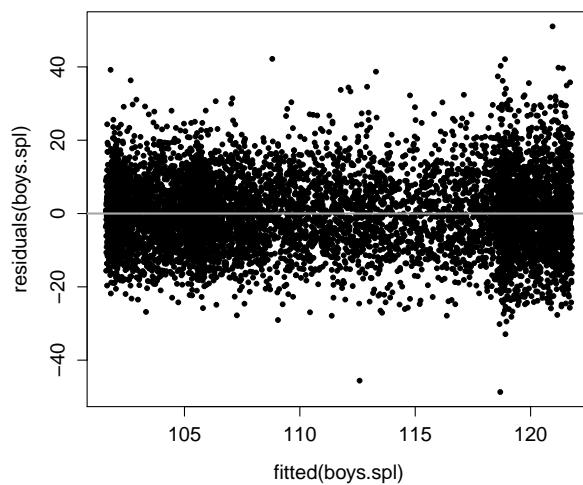
Smoothing Parameter spar= 0.5831793 lambda= 5.345368e-05 (14 iter.)
Equivalent Degrees of Freedom (Df): 40.0064
Penalized Criterion: 358541.9
GCV: 101.9030
> (boys3.spl<-smooth.spline(danemed$age, danemed$SBP, df=3))
Call:
smooth.spline(x = danemed$age, y = danemed$SBP, df = 3)

Smoothing Parameter spar= 1.296075 lambda= 7.558774 (16 iter.)
Equivalent Degrees of Freedom (Df): 3.000394
Penalized Criterion: 369932.8
GCV: 102.3770
```

Na podstawie powyższych wyników można odczytać wysokości współczynników wygładzenia dla kolejnych modeli: `boys.spl` $df \approx 11$, $\lambda = 0.012$, dla `boys40.spl` $df=40$, $\lambda = 0.00005$, natomiast dla `boys3.spl` $df=3$, $\lambda \approx 7.5$. Do wykresu z analizowanym podzbiorem danych zostały dodane trzy krzywe wygładzone. Niebieska funkcja została wyznaczona z powyższego modelu dla domyślnych parametrów, czerwona dla modelu w którym $df=40$, natomiast zielona dla $df=3$. Dodatkowo na podstawie prawego przeskalowanego wykresu dopasowanych wartości z modeli `boys.spl`, `boys40.spl` i `boys3.spl`, można zauważyc, że zielona krzywa jest niemal liniowa, niebieska krzywa jest znacznie bardziej gładka niż czerwona, która charakteryzuje się dużą zmiennością. Dla modelu `boys.spl`, który w porównaniu z pozostałymi dwoma modelami wydaje się być najlepiej dopasowany, na rysunku 2.5 zostały przedstawione otrzymane reszty.



Rysunek 2.4: Krzywe wygładzone dopasowane do danych.



Rysunek 2.5: Uzyskane reszty dla modelu `boys.spl` z dopasowanymi parametrami.

Rozdział 3

Regresja liniowa, modele GLM i GAM

3.1. Wprowadzenie

W tym rozdziale rozpoczynam dyskusję na temat pewnej szczególnej metody uczenia pod nadzorem – metody uogólnionych modeli addytywnych (GAM), która stanowi jedną z najbardziej wszechstronnych procedur dla modeli regresji nieparametrycznej. Idea tej metody jest oparta o dużo większą elastyczność niż tradycyjne parametryczne metody modelowania takie jak modele liniowe, czy uogólnione modele liniowe, które zostały krótko przedstawione w pierwszych podrozdziałach tego rozdziału. Ostatni podrozdział zawiera natomiast charakterystykę funkcji środowiska R, które są odpowiedzialne za wyznaczanie modeli GAM oraz przedstawia zastosowanie funkcji `gam()` do danych pediatrycznych `danemed`.

Rozdział ten powstał w znacznej części na podstawie opisu teorii z dziewiątego rozdziału książki [7], piątego rozdziału [10] oraz artykułu [17]. Stanowi on w całości wprowadzenie do modeli GAMLSS opisanych w rozdziale 4.

3.2. Modele Liniowe

Wśród technik modelowania statystycznego jednym z najprostszych i najczęściej stosowanych narzędzi są modele regresji liniowej. Liniowa funkcja regresji ma postać:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_P x_{Pi} + \varepsilon_i, \quad (3.1)$$

gdzie Y_i dla $i = 1, \dots, N$ są zmiennymi losowymi objaśnianymi, (x_{1i}, \dots, x_{Pi}) dla $i = 1, \dots, N$ są zaobserwowanymi wartościami N -danych obserwacji dla P zmiennych objaśniających. ε_i to tzw. *błędы* lub inaczej *zakłócenia*, które dla $i = 1, \dots, N$ są z założenia niezależnymi zmiennymi losowymi o tym samym rozkładzie z zerową średnią i stałą wariancją.

Model (3.1) można wygodniej zapisać w formie macierzowej:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (3.2)$$

gdzie \mathbf{Y} i $\boldsymbol{\varepsilon}$ są wektorami $N \times 1$, natomiast \mathbf{X} jest znaną macierzą eksperymentu $N \times P$, a $\boldsymbol{\beta}$ to wektor $P \times 1$. Nieznane wielkości (3.2), czyli parametry $\boldsymbol{\beta}$ można estymować minima-lizując sumę kwadratów błędów:

$$\boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^N \varepsilon_i^2. \quad (3.3)$$

Wynikiem minimalizacji (3.3) ze względu na β jest estymator najmniejszych kwadratów dla β :

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}. \quad (3.4)$$

Rozwiążanie najmniejszych kwadratów (3.4) dostarcza estymator dla współczynników β , lecz nie umożliwia przeprowadzania testów istotności tych współczynników. Odbywa się to z dodatkowym założeniem, że $\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_N)$, gdzie \mathbf{I}_N jest $N \times N$ macierzą identyczności. Podsumowując, model regresji liniowej będzie dany przez:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \quad \text{gdzie } \varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_N). \quad (3.5)$$

Po uwzględnieniu faktu, że liniowa funkcja zmiennej losowej o rozkładzie normalnym jest zmienną losową o rozkładzie normalnym, można obliczyć oczekiwane wartości w równaniu (3.5) i zapisać model (3.5) jako:

$$\mathbf{y} \sim N(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_N), \quad \text{gdzie } \boldsymbol{\mu} = \mathbf{X}\beta. \quad (3.6)$$

Postać (3.6) jest lepsza niż (3.5), gdyż łatwiej można rozszerzyć model na rozkłady inne niż rozkład normalny. Warto zauważyć, że w obu powyższych sformułowaniach wartości oczekiwane są warunkowane wartościami obserwowanymi zmiennych objaśniających, to znaczy, że zmienną objaśnianą modeluje się danymi zawartymi w znanej macierzy \mathbf{X} . Z (3.6) można wykazać liniowy związek pomiędzy średnią Y , $\mathbb{E}(Y) = \mu$, a x 'ami.

Funkcja wiarygodności modelu jest prawdopodobieństwem zaobserwowania próby, więc w przypadku (3.6) jest postaci:

$$L(\beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta) \right\} \quad (3.7)$$

z logarytmem funkcji wiarygodności:

$$l(\beta, \sigma^2) = \frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta). \quad (3.8)$$

Warto zauważyć, że maksymalizacja log-wiarygodności (3.8) ze względu na β jest równoważona minimalizacji najmniejszych kwadratów wielkości $(\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta)$ w (3.8). Zatem w tym przypadku estymator największej wiarygodności (MLE) i estymator najmniejszych kwadratów dla β w (3.4) są identyczne. MLE dla σ^2 dany jest przez

$$\hat{\sigma}^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\beta})^T (\mathbf{Y} - \mathbf{X}\hat{\beta})}{N}. \quad (3.9)$$

MLE dla σ^2 , $\hat{\sigma}^2$ jest estymatorem obciążonym.

Estymatory β i σ^2 podaje się przez podstawienie obserwowanych wartości $\mathbf{y} = (y_1, y_2, \dots, y_N)^T$ dla zmiennych losowych $\mathbf{Y} = (Y_1, Y_2, \dots, Y_N)^T$ dając np. estymator $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ dla β .

Specyfikacja modelu

W praktyce podczas wyboru modelu liniowego można spotkać się z następującymi problemami:

- relacja pomiędzy zmienną objaśnianą i zmiennymi objaśniającymi nie ma liniowego charakteru,
- zmienne losowe ε_i i w konsekwencji zmienna objaśniana nie mają rozkładu normalnego,

- zmienne losowe ε_i nie są niezależne,
- wariancja ε_i , a więc także wariancja zmiennej objaśnianej nie jest stała dla wszystkich obserwacji.

Uogólnione modele liniowe (GLM) oraz uogólnione modele addytywne (GAM) opisane w następnych podrozdziałach częściowo stanowią rozwiązanie dla pierwszego i drugiego problemu. Większość powyższych problemów rozwiąże natomiast metoda GAMLSS omawiana w rozdziale 4.

3.3. Uogólnione modele liniowe (GLM)

Równanie (3.6) modelu liniowego pozwala na rozszerzenie go do uogólnionych modeli liniowych (GLM). W uogólnionym modelu liniowym we wzorze (3.6) rozkład normalny zmiennej Y_i jest zastąpiony przez wykładowiczą rodzinę rozkładów. Ponadto zostaje wprowadzona monotoniczna funkcja wiążąca $g(\cdot)$ (ang. *link function*), która opisuje związek wartości oczekiwanej zmiennej objaśnianej Y_i oznaczonej μ_i z liniowym predyktorem η_i będącym kombinacją liniową zmiennych objaśniających:

$$g(\mu_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}. \quad (3.10)$$

W formie wektorowej powyższy zapis przyjmuje postać:

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}^T \boldsymbol{\beta}. \quad (3.11)$$

Rozkład zmiennej losowej Y należy do rodziny rozkładów wykładowiczych i może być zdefiniowany przez funkcję gęstości prawdopodobieństwa $f_Y(y; \mu, \phi)$ postaci:

$$f_Y(y; \mu, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{\phi} + c(y, \phi) \right\}, \quad (3.12)$$

gdzie $E(Y) = \mu = b'(\theta)$ i $V(Y) = \phi V(\mu)$, z funkcją wariancji $V(\mu) = b''[\theta(\mu)]$. Postać (3.12) obejmuje wiele ważnych rozkładów, tj. rozkład normalny, Poissona, gamma, odwrotny gaussowski, rozkład Tweediego (Tweedie, 1984), dla których funkcja wariancji jest równa odpowiednio $V(\mu) = 1, \mu, \mu^2, \mu^3$ i μ^p dla $p < 0$ lub $p > 1$, a także dwumianowy oraz ujemny dwumianowy z funkcjami wariancji równymi odpowiednio $V(\mu) = \frac{\mu(1-\mu)}{N}$ i $V(\mu) = \mu + \frac{\mu}{\phi}$.

3.4. Uogólnione modele addytywne (GAM)

Tradycyjne modele liniowe oraz uogólnione modele liniowe w wielu sytuacjach okazują się narzędziem niedostatecznym, ponieważ w sytuacjach opisujących rzeczywistość wiele zjawisk ma bardziej złożony charakter. W poprzednich rozdziałach zostały przedstawione techniki stosowane do predefiniowania funkcji bazowych, dzięki którym można osiągać nieliniowe estymatory. W tym rozdziale zostanie wskazana alternatywa dla modeli liniowych i modeli GLM – *uogólnione modele addytywne*, w skrócie *GAM*, (ang. *Generalized Additive Models*). Modele GAM zostały opracowane w 1990 roku przez Trevor'a Hastie and Rob'a Tibshirani. Zaproponowali oni estymację dla wielowymiarowych zmiennych przy pomocy addytywnymi funkcjami „nieparametrycznymi”, które mogą być estymowane np. przez wygładzone kubiczne funkcje klejane.

W terminach regresji oraz oznaczeniach z poprzednich podrozdziałów GAM mają postać:

$$E(Y_i|x_{i1}, x_{i2}, \dots, x_{iP}) = \alpha + f_1(x_{i1}) + f_2(x_{i2}) + \dots + f_P(x_{iP}), \quad (3.13)$$

gdzie α jest stałą, f_j , $j = 1, 2, \dots, P$ to nieznane funkcje j -tej zmiennej wyjaśniającej estymowane m.in. przy pomocy regresji lokalnie wielomianowej¹ lub wygładzonych kubicznych funkcji sklejanych². Estymacja funkcji f_j , odbywa się wspólnie dla $j = 1, \dots, P$ przy pomocy pewnej iteracyjnej procedury – *backfitting algorithm* – zwanej w literaturze [10] *algorytmem wielokrotnego dopasowania*. Algorytm ten został opisany w rozdziale 3.5.

Przykładowo, w modelu regresji logistycznej średnia dla zmiennej binarnej Y , $\mu = \mathbb{P}(Y = 1|X)$ jest powiązana z predyktorami za pomocą modelu regresji liniowej i funkcji *logit*:

$$\log\left(\frac{\mu}{1-\mu}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_P x_P. \quad (3.14)$$

W modelu *addytynym* regresji logistycznej każdy liniowy składnik jest zastępowany ogólniejszą postacią funkcyjną:

$$\log\left(\frac{\mu}{1-\mu}\right) = \alpha + f_1(x_1) + f_2(x_2) + \dots + f_P(x_P), \quad (3.15)$$

gdzie f_j , dla $j = 1, 2, \dots, P$, zgodnie z wyżej przedstawioną notacją są nieznanymi funkcjami wyznaczanymi za pomocą algorytmu wielokrotnego dopasowania.

Ogólnie, w modelach GAM średnia zmiennej wyjaśnianej Y warunkowana zmiennymi wyjaśniającymi ozn. $\mu = \mathbb{E}(Y|X)$, jest modelowana przy pomocy addytywnych funkcji f_j , $j = 1, 2, \dots, P$, zmiennych wyjaśniających. Podobnie jak w GLM można określić funkcję g , która będzie wiązać μ z addytywnymi funkcjami zmiennych wyjaśniających:

$$g(\mu) = \alpha + f_1(x_1) + f_2(x_2) + \dots + f_P(x_P). \quad (3.16)$$

Popularne przykłady funkcji wiążących:

- $g(\mu) = \mu$ jest tożsamościową funkcją wiążącą wykorzystywaną w liniowych i addytywnych modelach dla zmiennej wynikowej o rozkładzie Gaussa.
- $g(\mu) = \text{logit}(\mu)$ dla modeli z binarną zmienną Y .
- $g(\mu) = \text{probit}(\mu)$ dla modeli o dwumianowym rozkładzie prawdopodobieństwa zmiennej Y . *Probit* jest to funkcja odwrotna do dystrybuantu Gaussa: $\text{probit}(\mu) = \phi^{-1}(\mu)$.
- $g(\mu) = \log(\mu)$ dla log-liniowych lub log-addytywnych modeli, gdzie zmienna Y ma rozkład Poissona.

Powysze trzy rozkłady należą do rodziny rozkładów wykładniczych. W modelach GAM podobnie jak dla GLM zmienna Y należy do rodziny rozkładów wykładniczych.

¹W środowisku R odpowiada za to np. funkcja `loess(stats)`, która dopasowuje lokalnie wygładzone wielomiany drugiego stopnia, [23].

²Za dopasowywanie wygładzonymi kubicznymi funkcjami sklejanymi w programie R odpowiada funkcja `smooth.splines(splines)`, rozdział 2.2.

3.5. Wyznaczanie modelu addytywnego (GAM)

W tym podrozdziale zostanie przedstawiony modułowy schemat algorytmu, który wyznacza model addytywny. Blokowa budowa tego algorytmu w sposób elastyczny umożliwia estymowanie funkcji $f_j(\cdot)$ występujących w poniższym wzorze (3.17) przez zastosowanie ustalonej metody estymacji nieparametrycznej np. estymatora opartego na funkcjach sklejanych, czy estymatora LOESS, [10]. W celu uproszczenia schematu algorytmu zakładam, że rozpatrywana zmienna odpowiedzi posiada rozkład gaussowski oraz funkcja $g(\cdot)$ jest identycznościową funkcją wiążącą. Dodatkowo jako metodę estymacji wybieram wygładzone kubiczne funkcje sklejane.

Uwzględniając powyższe założenia, addytywny model ma postać:

$$Y_i = \alpha + \sum_{j=1}^P f_j(x_{ij}) + \varepsilon_i, \quad (3.17)$$

gdzie ε_i jest błędem losowym o średniej równej 0. Dla danych obserwacji x_{ij} i y_i , $i = 1, 2, \dots, N$, $j = 1, 2, \dots, P$, kryterium sumy kwadratów z karą zawartą w składniku:

$$\lambda_j \int_{b^{(j)}}^{a^{(j)}} \hat{f}_j''(t_j)^2 dt_j, \quad (3.18)$$

można zapisać w następujący sposób:

$$PRSS(\hat{\alpha}, \hat{f}_1, \hat{f}_2, \dots, \hat{f}_P) = \sum_{i=1}^N \left\{ y_i - \hat{\alpha} - \sum_{j=1}^P \hat{f}_j(x_{ij}) \right\}^2 + \sum_{j=1}^P \lambda_j \int_{b^{(j)}}^{a^{(j)}} \hat{f}_j''(t_j)^2 dt_j. \quad (3.19)$$

W powyższych wzorach (3.18) i (3.19) $\lambda_j \geq 0$ jest parametrem wygładzającym, natomiast $[a^{(j)}, b^{(j)}]$ to przedział $a^{(j)} \leq x_1^{(j)} \leq x_2^{(j)} \leq \dots \leq x_N^{(j)} \leq b^{(j)}$, na którym jest określonych N węzłów j -tej zmiennej objaśniającej. Podane zagadnienie mówi, że minimalizacja sumy kwadratów błędów ze współczynnikiem kary oznacza znalezienie odpowiedniej wartości parametru $\hat{\alpha}$, oraz odpowiednich P -funkcji $\hat{f}_j(\cdot)$, dla których $\hat{f}_j(\cdot)$ jest funkcją j -tej zmiennej objaśniającej x_j . Zostało udowodnione [7], że minimum (3.19) istnieje i można zapisać je w następującej postaci:

$$\hat{\mathbf{f}}(\mathbf{x}) = \hat{\alpha} + \sum_{j=1}^P \hat{f}_j(x_j). \quad (3.20)$$

Każda z funkcji $\hat{f}_j(\cdot)$, która estymuje odpowiednią funkcję $f_j(\cdot)$, ma jednoznacznie wyznaczone węzły w wartościach ze zbioru x_{ij} , $i = 1, 2, \dots, N$. Poszukiwane funkcje nieparametryczne są określone na prostej, a nie na oryginalnej przestrzeni P -wymiarowej. Tym sposobem korzystając z modeli addytywnych unikamy konieczności rozwiązywania zadania estymacji nieparametrycznej w przestrzeni wielowymiarowej.

Znajdowanie $\hat{f}_j(\cdot)$ przebiega iteracyjnie z wykorzystaniem *algorytmu wielokrotnego dopasowania* (ang. *backfitting algorithm*). Przyjmuje się, że estymowana funkcja regresji jest postaci:

$$\mathbf{f}(\mathbf{x}) = \alpha + \sum_{j=1}^P f_j(x_j). \quad (3.21)$$

Aby uniknąć niejednoznaczności związanej z wyznaczeniem stałych w modelu (3.21) należy nałożyć dodatkowo na funkcję (3.20) i rozwiązanie (3.19) odpowiednie ograniczenia. Dla danych zmiennych \mathbf{X} i Y zostaje wprowadzony warunek $\mathbb{E}f_j(X_j) = 0$, który pociąga za sobą

$\mathbb{E}Y = \alpha$. Próbkowe odpowiedniki tych warunków odniesione do modelu (3.20) mają postać:

$$\frac{1}{N} \sum_{i=1}^N \hat{f}_j(x_{ij}) = 0 \quad (3.22)$$

oraz

$$\hat{\alpha} = \frac{1}{N} \sum_{i=1}^N y_i. \quad (3.23)$$

W procesie poszukiwania rozwiązania (3.20) warunki (3.22) i (3.23) są nakładane algorytmicznie. Jeżeli funkcja regresji jest postaci (3.21), to stosuje się iteracyjny sposób znajdowania rozwiązania (3.19) nazywany *wielokrotnym dopasowaniem*. Sposób ten polega na iteracyjnym dopasowaniu funkcji kolejnych zmiennych objaśniających. Nazwa – *wielokrotne dopasowanie* – pochodzi stąd, że każda funkcja $f_j(\cdot)$ jest estymowana wielokrotnie za każdym razem na podstawie innych aktualizowanych reszt, [10]. Kroki algorytmu zostały zapisane w postaci następującego schematu:

ALGORYTM WIELOKROTNEGO DOPASOWANIA DLA MODELU ADDYTYWNEGO:

1. INICJALIZACJA

Dla $i = 1, \dots, N$ ustalamy:

$$\hat{\alpha}^{(0)} = \frac{1}{N} \sum_{i=1}^N y_i$$

oraz dla $j = 1, \dots, P$ przyjmujemy początkowe oceny $\hat{f}_j^{(0)}(\cdot)$ funkcji $f_j(\cdot)$,
(np. $\hat{f}_j^{(0)}(\cdot) \equiv 0$ dla $j = 1, \dots, P$, [7]).

2. ITERACYJNIE

* PĘTŁA ZEWNĘTRZNA:

Dla $r = 0, 1, \dots$ do uzyskania zbieżności wykonujemy:

* PIERWSZA PĘTŁA WEWNĘTRZNA:

Dla $j = 1, 2, \dots, P$, oraz dla ustalonej metody estymacji nieparametrycznej S_j wykonujemy:

* DRUGA PĘTŁA WEWNĘTRZNA:

Dla $i = 1, \dots, N$ wyznaczamy j -tą ocenę $\hat{f}_j(\cdot)$

$$\hat{f}_j^{(r)} \leftarrow S_j \left[\left(y_i - \hat{\alpha}^{(r)} - \sum_{k \neq j} \hat{f}_k^{(r)}(x_{ik}) \right) \right]$$

* KONIEC DRUGIEJ PĘTLI WEWNĘTRZNEJ.

* KONIEC PIERWSZEJ PĘTLI WEWNĘTRZNEJ.

Dla wyznaczonej oceny w r -tym kroku $\hat{f}_j^{(r)}$, dokonujemy jej aktualizacji:

$$\hat{f}_j^{(r+1)} \leftarrow \hat{f}_j^{(r)} - \frac{1}{N} \sum_{i=1}^N \hat{f}_j^{(r)}(x_{ij}).$$

Funkcje \hat{f}_j wymagają modyfikacji dopóki zmieniają się o mniej niż założony próg, tzn. do czasu stwierdzenia zbieżności algorytmu (lub zatrzymania go).

* KONIEC ZEWNĘTRZNEJ PĘTLI

Warto dodać, że jeżeli rozważamy operację S_j na całym zbiorze danych, to reprezentację operatora \mathbf{S}_j będzie można zapisać w postaci macierzy o rozmiarze $N \times N$, (rozdział 1.4). Poprzez analogię ze stopniami swobody do omawianych operatorów wygładzających z rozdziału 1, stopnie swobody dla j-tego czynnika będą wyznaczane ze wzoru $df_j = \text{tr}(\mathbf{S}_j) - 1$, [7].

Powyższy algorytm dobrze spisuje się w praktyce i jest wykorzystywany w wielu modelach. W tym rozdziale został przedstawiony model, gdzie zmienna Y należała do gaussowskiej rodziny rozkładów i przyjmowała identycznościową funkcję wiążącą. Naturalnym rozszerzeniem tego modelu jest zastąpienie funkcji wiążącej inną funkcją, oraz gaussowskiej rodziny rozkładów przez inny rozkład. Wówczas estymacja uogólnionych modeli addytywnych będzie odbywać się za pomocą odpowiedniej modyfikacji przedstawionego w ramce algorytmu wielokrotnego dopasowania. Pewną modyfikację tego algorytmu wykorzystuje metoda GAMLSS, o której mowa w następnym rozdziale.

3.6. Uogólnione modele addytywne w R

W tym podrozdziale scharakteryzuje sposób konstruowania uogólnionych modeli addytywnych w środowisku R. Pokażę także przykład wykorzystania GAM do problemu opisanego we wstępie pracy. Korzystając z funkcji `gam(gam)` spróbuje utworzyć model, który pozwoliłby na estymowanie ciśnienia skurczowego za pomocą zmiennych wiek i wzrost.

3.6.1. Funkcje `gam`

Uogólnione modele addytywne w środowisku R mogą być wyznaczone za pomocą funkcji `gam` dostępnej w bibliotece z 1990 roku o nazwie `gam`, której autorami są Hastie i Tibshirani. Estymacja `gam(gam)` opiera się na wykorzystywaniu iteracyjnych metod z użyciem algorytmu wielokrotnego dopasowania oraz metod wygładzających („*linear smoother*”) opartych na regresji lokalnie wielomianowej, i wygładzonych funkcjach sklejanych. Nowszą wersją tej funkcji `gam` jest funkcja Wood'a z 2000/2001 roku pochodząca z biblioteki `mgcv`³, która wyznacza klasę modeli GAM korzystając z wygładzonych funkcji sklejanych z jednociennym automatycznym wyborem parametrów wygładzenia⁴. Uogólnione modele addytywne, jak wynika z nazwy, nie są ograniczone do modeli z błędem o rozkładzie normalnym i identycznościową funkcję wiążącą. W obu przypadkach `gam(gam)` i `gam(mgvc)` za odpowiednie określenie funkcji wiążących i rodziny rozkładu zmiennej objaśnianej, podobnie jak w `glm`, odpowiada argument `family`. W metodach wyznaczania modeli GAM za pomocą funkcji `gam(gam)` i `gam(mgvc)` występują pewne różnice. Korzystając z zamieszczonego w <http://www.mail-archive.com/r-help@stat.math.ethz.ch/msg31741.html> porównania pakietów `gam` i `mgcv`, poniżej przedstawiam ich krótką charakterystykę.

Funkcja z pakietu `gam` opiera się na podejściu GAM prezentowanym w podrozdziale 3.5. Estymacja odbywa z wykorzystaniem algorytmu wielokrotnego dopasowania. Ta wersja `gam`

³Nazwa biblioteki `mgcv` pochodzi od nazwy metody, jaką są wybierane parametry wygładzenia *multiple generalized cross-validation*.

⁴Parametry wygładzenia są estymowane razem z resztą modelu, z wykorzystaniem m.in. minimalizacji kryterium uogólnionej walidacji krzyżowej (ang. *generalized cross-validation*)

$$\frac{n\hat{\sigma}^2}{n - df_{mod}},$$

gdzie $\hat{\sigma}^2$ jest estymatorem wariancji błędu, df_{mod} odpowiada liczbie stopni swobody dla modelu, włączając parametryczne i wygładzone składniki modelu. W uogólnionych modelach addytywnych, (np. dla addytywnych modeli regresji logistycznej) estymator rozproszenia $\hat{\phi}$ zastępuje wariancję błędu $\hat{\sigma}$, [4].

udostępnia szeroką gamę wygładzeń – obecnie dostępna jest metoda wygładzania przy pomocy funkcji sklejanych oraz metoda regresji lokalnie wielomianowej. Za wygładzanie przy pomocy regresji lokalnie wielomianowej odpowiada składnik `lo()`, natomiast `s()` wspomaga `gam` w wygładzaniu funkcjami sklejonymi. Obie funkcje `lo()` i `s()` zostały wbudowane w funkcję `gam`. Parametr wygładzenia dla danego składnika (tj. `span` dla `lo()`, lub `df` dla `s()`) jest określony bezpośrednio, a nie jak w pakiecie `mgcv` metodą uogólnionej walidacji krzyżowej.

Funkcja `gam()` z pakietu `mgcv` opiera się na *penalized regression splines* czyli w wolnym tłumaczeniu *regresji funkcji sklejanych ze współczynnikiem kary*. Estymacja odbywa się bezpośrednio przez maksymalizację wiarygodności ze współczynnikiem kary, która zintegrowana jest z estymacją parametru wygładzenia za pomocą np. kryterium GCV. Bardzo ważnym składnikiem w tym podejściu jest składnik `s()`, który w przeciwieństwie do swojego odpowiednika zawartego w pakiecie `gam`, może być funkcją więcej niż jednej zmiennej (możnałączyć gładkie składniki zawierające interakcje dwóch lub więcej predyktorów). W formule modelu można dodawać liniowe składniki jak i nieparametryczne.Więcej na temat funkcji `gam(gam)` i `gam(mgvc)` można przeczytać w [4], [23].

Podsumowując, jeśli zależy nam np. na otrzymaniu parametrycznej reprezentacji modelu, automatycznym wyznaczeniu parametru wygładzenia i/lub zawarciu w modelu składników gładkich interakcji zmiennych wyjaśniających, warto skorzystać z funkcji `gam` dostępnej w pakiecie `mgcv`. Jeśli chcemy natomiast wykorzystać lokalnie wielomianową regresję i preferujemy raczej podejście iteracyjne, wówczas lepszy jest pakiet `gam`.

3.6.2. Przykład wykorzystania funkcji `gam`

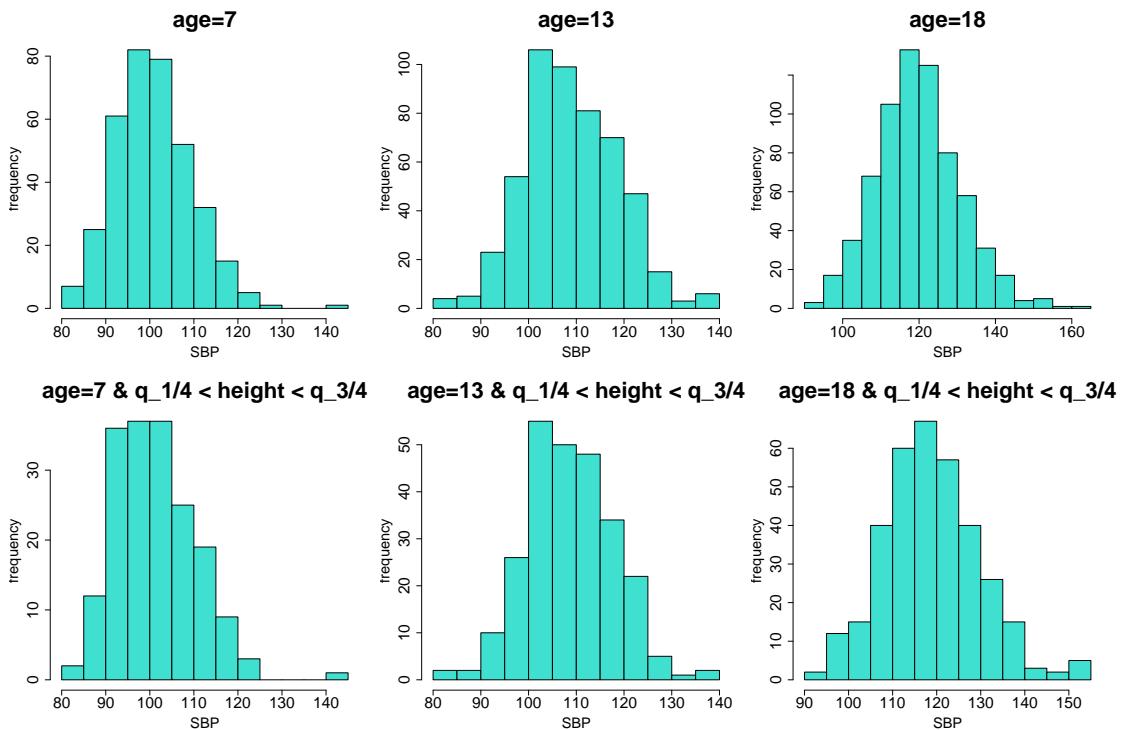
Metodę GAM z funkcją `gam(gam)` spróbuję zastosować do problemu przedstawionego we wstępie pracy, a mianowicie ciśnienie skurczowe (SBP) będę chciał estymować na podstawie wieku i wysokości, czyli zmiennych `age` i `height`. Do analiz wybieram 6627 obserwacji ze zbioru `danemed` dla osób płci męskiej, wykluczając osoby z nadwagą i otyłością.

Wykres 3.1 przedstawia histogramy zmiennej `SBP` dla wybranych trzech grup wiekowych: 7, 13 i 18 lat. Pierwsze trzy wykresy zostały utworzone dla wszystkich obserwacji z próby⁵. Kolejne trzy wykresy powstały dla tych samych grup wiekowych z uwzględnieniem obserwacji mieszczących się w 1/4 – 3/4 kwartylach zmiennej `height`. W każdym z tych przypadków widać, że dane cechują się prawostronną skośnością (dodatek B.2).

Jak już zostało wspomniane, w modelach GAM mamy do czynienia z modelowaniem średniej zmiennej odpowiedzi warunkowanej zmiennymi wyjaśniającymi. Pozostałe parametry rozkładu są traktowane jako stałe. Natomiast zmienna odpowiedzi pochodzi z ustalonego rozkładu należącego do rodziny rozkładów wykładniczych. Rozważając trafność wyboru modelu GAM do analizowanego problemu, powołam się na źródło *WHO Child Growth Standards* z 2006 roku oraz opinię autorów metody GAMLSS – Stasinopoulos'a i Rigby'ego, która została zawarta w pracy [19]. Oba te źródła wskazują GAMLSS jako odpowiednią metodę do wyznaczenia krzywych rozwoju dzieci i młodzieży. Można zatem przypuszczać, iż modele GAM mogą nie pasować do rozważanego zagadnienia.

Aby to zweryfikować konstruuję modele używając funkcji `gam` z domyślną wartością dla parametru `family=gaussian`, tzn. identycznościową funkcją wiążącą oraz zmienną wyjaśnianą pochodząjącą z rozkładu normalnego. W dwóch pierwszych modelach do opisania nieliniowości korzystam z regresji lokalnie wielomianowej zawartej w składniku `lo`. Parametry składnika `lo` równe `degree=1` oraz `degree=2` wskazują na stopień lokalnych wielomianów biorących

⁵Do wyznaczenia odpowiednich grup wiekowych została wykorzystana zmienna `wiek.kat`, która kategoryzuje ciągłą zmienną `age`.



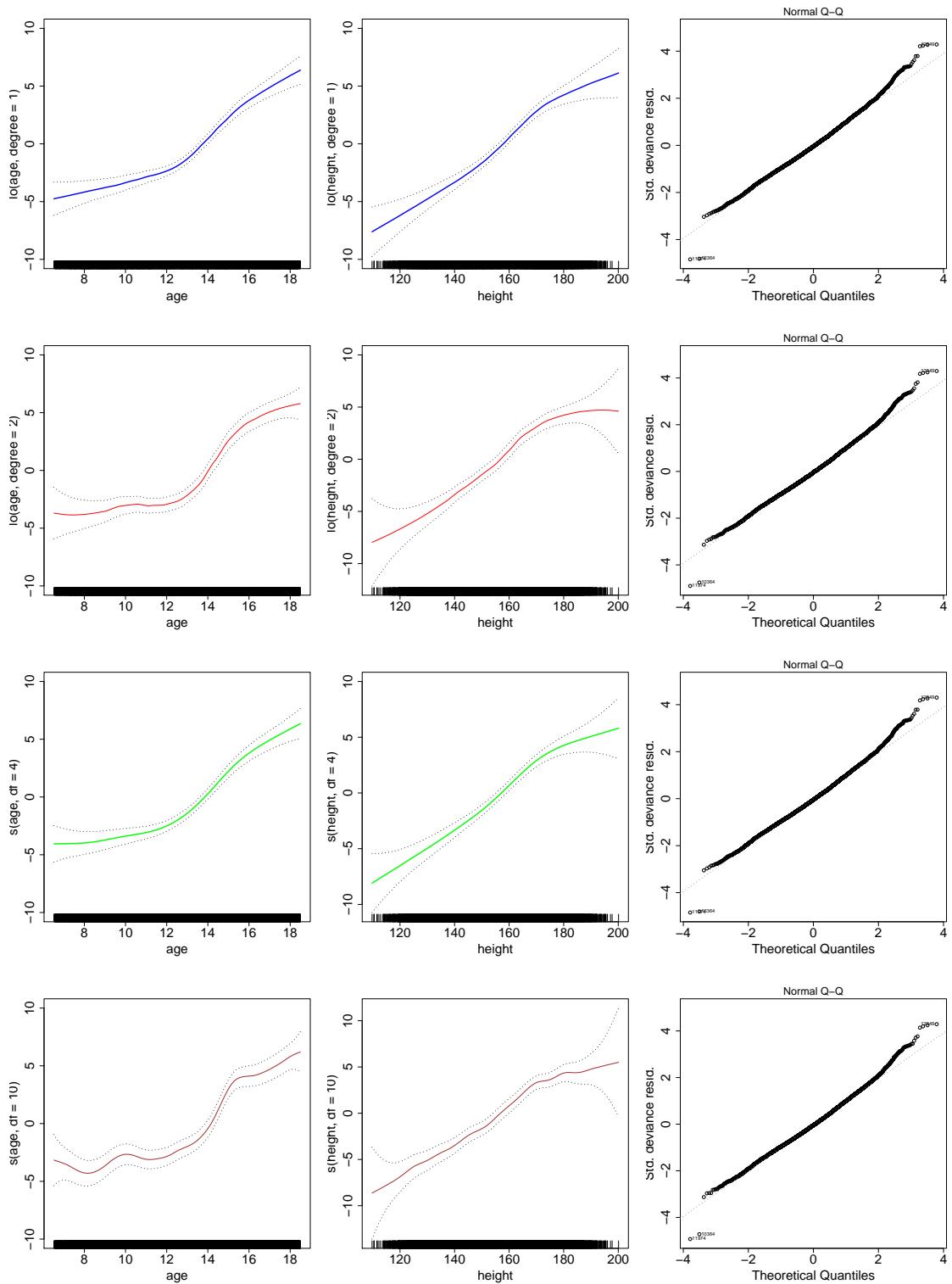
Rysunek 3.1: Histogramy zmiennej objaśnianej SBP dla wybranych trzech grup wiekowych.

udział w dopasowaniu. W kolejnych modelach zostało zastosowane wygładzanie przy pomocy kubicznych funkcji sklejanych ze stopniami swobody $df=4$ oraz $df=10$.

```
> mod1.gam = gam(SBP ~ lo(age, degree=1) + lo(height, degree=1))
> mod2.gam = gam(SBP ~ lo(age, degree=2) + lo(height, degree=2))
> mod3.gam = gam(SBP ~ s(age, df=4) + s(height, df=4))
> mod4.gam = gam(SBP ~ s(age, df=10) + s(height, df=10))
```

Dla każdego z modeli na rysunku 3.2 przedstawione zostały wykresy obrazujące addytywny wpływ zmiennych `age` i `height` na zmienną odpowiedzi `SBP`. Kolejne wiersze odpowiadają modelom: `mod1.gam`, `mod2.gam`, `mod3.gam` i `mod4.gam`. Na podstawie dwóch pierwszych modeli można zaobserwować wzrost stopnia dopasowania modelu do danych ze wzrostem stopnia lokalnych wielomianów. Natomiast, porównując `mod3.gam` i `mod4.gam` widać znaczny wzrost zmiенноśc dopasowanych krzywych w przypadku zwiększenia liczby stopni swobody.

W celu sprawdzenia adekwatności modeli rysuję wykresy kwantylowe, które sprawdzają normalność rozkładu reszt. Na osi poziomej przedstawione zostały kwantyle rozkładu normalnego odpowiadające resztom, zaś na osi pionowej kwantyle empiryczne dla standaryzowanych reszt. We wszystkich modelach pojawia się ten sam problem – można zauważać znaczne odstępstwa od normalności, co sugeruje nieadekwatność modeli, a także upoważnia do modyfikacji modelu poprzez np. transformację zmiennych. Lepszym sposobem na ominięcie problemów z dopasowaniem modelu GAM jest zastosowanie modelu GAMLSS.



Rysunek 3.2: Modele GAM przedstawiające addytywny wpływ odpowiednich zmiennych na wysokość ciśnienia SBP.

Rozdział 4

GAMLSS

Niniejszy rozdział w znacznej części opiera się na pracach [19], [21].

Pierwsze dwa podrozdziały stanowią wprowadzenie do modelu GAMLSS – zawierają charakterystykę, postać i różne typy modelu w zależności od wprowadzonych ograniczeń do ogólnej postaci GAMLSS. Podrozdział 4.3 zawiera opis metody estymacji, w podrozdziale 4.4 zostały wymienione różne metody za pomocą których można opisywać liniowy predyktor modelu. Następny podrozdział stanowi wprowadzenie do modeli GAMLSS w środowisku R. Kolejne części tego podrozdziału zawierają opis funkcji pakietu `gamlss`, listę dostępnych rozkładów, addytywnych składników funkcji `gamlss()`, oraz kilka słów o budowie modelu.

4.1. Co to jest GAMLSS?

Uogólnione modele addytywne z parametrem położenia, skali i kształtu, w skrócie GAMLSS (ang. *Generalized Additive Models for Location Scale and Shape*) zostały opracowane przez Rigby'ego i Stasinopoulos'a w 2005 roku jako elastyczna metoda mająca na celu pokonanie niektórych ograniczeń związanych z uogólnionymi modelami liniowymi oraz uogólnionymi modelami addytywnymi. W przeciwieństwie do popularnych metod GLM i GAM, w modelach GAMLSS nie jest wymagane, by objaśniana zmienna należała do rodziny rozkładów wykładniczych. Metoda GAMLSS jest szczególnie przydatna, gdy zmienna zależna jest *od-datnio* lub *ujemnie skośna, leptokurtyczna* lub *platykurtyczna*¹, czy też nadmiernie rozproszona oraz niejednorodna, tzn. gdy skala i kształt rozkładu zmiennej odpowiedzi zmienia się wraz ze zmiennymi objaśniającymi. Modele GAMLSS zostały rozszerzone o możliwość modelowania nie tylko średniej ale i innych parametrów rozkładu zmiennej y (tj. wariancji, skośności lub kurtozy) wykorzystując do tego liniowe funkcje parametryczne, addytywne nieparametryczne funkcje zmiennych objaśniających lub efekty losowe. Modele GAMLSS można zaliczyć do grupy modeli regresji typu „semi-parametrycznego”[11]. Są one parametryczne, ponieważ z założenia wymagają parametrycznego rozkładu zmiennej y , natomiast ich „semi-parametryczność” wynika z możliwości modelowania parametrów rozkładu z wykorzystaniem nieparametrycznych metod wygładzania funkcji.

Istnieje kilka pakietów w środowisku R, które swoją metodologią i sposobem implementacji są zbliżone do pakietu `gamlss`. Należą do nich m.in. scharakteryzowane w rozdziale 3 oryginalny pakiet `gam` oraz jego nowsza wersja `mgcv`. Omawiane tam metody dotyczą jednak modelowania wyłącznie średniej zmiennej objaśnianej, której rozkład należy do rodziny rozkładów wykładniczych.

¹ Wykresy funkcji gęstości prawdopodobieństwa dla przykładowych rozkładów przedstawiają rysunki B.1(a), B.1(b), B.1(c), B.1(d), zawarte na końcu pracy w dodatku B.2.

4.2. Postać GAMLSS

GAMLSS zakłada, że $\forall i = 1, \dots, n$ obserwacje $y_i|\theta^i$ są niezależne, oraz funkcja gęstości prawdopodobieństwa $f(y_i|\theta^i)$ jest warunkowana $\theta^i = (\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i}) = (\mu_i, \sigma_i, \nu_i, \xi_i)$ wektorem czterech parametrów rozkładu, gdzie każdy może być funkcją zmiennych objaśniających. Oznaczę $Y_i|\theta^i \sim \mathcal{D}(\theta^i)$, tzn. $Y_i|(\mu_i, \sigma_i, \nu_i, \xi_i) \sim \mathcal{D}(\mu_i, \sigma_i, \nu_i, \xi_i)$ niezależnie dla $i = 1, 2, \dots, n$, gdzie \mathcal{D} reprezentuje rozkład Y . Do $(\mu_i, \sigma_i, \nu_i, \xi_i)$ można odnieść się jak do wektora parametrów rozkładu. Pierwsze dwa parametry μ_i i σ_i często są charakteryzowane jako parametry położenia i skali, pozostałe mogą być uznane odpowiednio za parametry kształtu, np. skośność i kurtozę. Model także może być uogólniony na więcej niż 4 parametry.

Niech $\mathbf{y}^T = (y_1, y_2, \dots, y_n)$ będzie wektorem zmiennej odpowiedzi długości n . Dla $k = 1, 2, 3, 4$ niech $g_k(\cdot)$ będzie znaną monotoniczną funkcją wiążącą parametr θ_k ze zmiennymi objaśniającymi poprzez „semi-parametryczny” model addytywny w następujący sposób:

$$\begin{aligned} g_1(\boldsymbol{\mu}) &= \boldsymbol{\eta}_1 = \mathbf{X}_1 \boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} h_{j1}(\mathbf{x}_{j1}) \\ g_2(\boldsymbol{\sigma}) &= \boldsymbol{\eta}_2 = \mathbf{X}_2 \boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} h_{j2}(\mathbf{x}_{j2}) \\ g_3(\boldsymbol{\nu}) &= \boldsymbol{\eta}_3 = \mathbf{X}_3 \boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} h_{j3}(\mathbf{x}_{j3}) \\ g_4(\boldsymbol{\tau}) &= \boldsymbol{\eta}_4 = \mathbf{X}_4 \boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} h_{j4}(\mathbf{x}_{j4}) \end{aligned} \quad (4.1)$$

gdzie $\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}, \boldsymbol{\eta}_k$ i \mathbf{x}_{jk} dla $j = 1, 2, \dots, J_k$ i $k = 1, 2, 3, 4$ są wektorami długości n . Cztery kolejne zmienne oznaczają parametry rozkładu, natomiast $\boldsymbol{\eta}_k$ to predyktor k -tego parametru rozkładu. Dla $\boldsymbol{\eta}_k$ definiuję zbiór zmiennych objaśniających t_k , przy pomocy których będą modelowane dane parametry rozkładu zmiennej objaśnianej. Następnie \mathbf{x}_{jk} to z założenia dany wektor zmiennej objaśniającej. Funkcja h_{jk} jest nieparametryczną addytywną funkcją zmiennej objaśniającej X_{jk} oraz $\mathbf{h}_{jk} = h_{jk}(\mathbf{x}_{jk})$ jest wektorem, który stanowi wartość funkcji h_{jk} w \mathbf{x}_{jk} . Dodatkowo \mathbf{X}_k dla $k = 1, 2, 3, 4$ to tzw. macierze eksperymentu. Z powyższej postaci (4.1) modelu GAMLSS, nazywanej semi-parametryczną postacią modelu GAMLSS będę korzystać w rozdziale 5.

Ogólna postać modelu GAMLSS umożliwia występowanie składników efektów losowych² w modelach dla $\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}$. Teoria modeli z efektami losowymi wykracza znacznie poza zakres tej pracy. Szczegóły związane z ogólną postacią modelu GAMLSS można znaleźć w literaturze [17]. Zgodnie z artykułem [17] model GAMLSS w ogólnej postaci jest następujący:

$$\begin{aligned} g_k(\boldsymbol{\mu}) &= \boldsymbol{\eta}_1 = \mathbf{X}_1 \boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} \mathbf{Z}_{j1} \boldsymbol{\gamma}_{j1} \\ g_k(\boldsymbol{\sigma}) &= \boldsymbol{\eta}_2 = \mathbf{X}_2 \boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} \mathbf{Z}_{j2} \boldsymbol{\gamma}_{j2} \end{aligned} \quad (4.2)$$

²Szczegółowy opis modeli z efektami losowymi wraz z ciekawymi przykładami ich zastosowań można znaleźć w książce z 2011 roku pt. *Modele liniowe z efektami stałymi, losowymi i mieszanymi*, której autorem jest P. Biecek.

$$g_k(\boldsymbol{\nu}) = \boldsymbol{\eta}_3 = \mathbf{X}_3 \boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} \mathbf{Z}_{j3} \boldsymbol{\gamma}_{j3}$$

$$g_k(\boldsymbol{\tau}) = \boldsymbol{\eta}_4 = \mathbf{X}_4 \boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} \mathbf{Z}_{j4} \boldsymbol{\gamma}_{j4},$$

gdzie \mathbf{Z}_{jk} jest znaną macierzą rozmiaru $n \times q_{jk}$ zmiennych objaśniających (traktowanych jako losowe), dla $j = 1, 2, \dots, J_k$, $k = 1, 2, 3, 4$, $\boldsymbol{\gamma}_{jk}$ to wektory efektów losowych długości q_{jk} . Wektory te z założenia mają rozkład normalny $\boldsymbol{\gamma}_{jk} \sim N_{q_{jk}}(0, \mathbf{G}_{jk}^{-1})$, gdzie \mathbf{G}_{jk}^{-1} jest odwracalną, symetryczną macierzą $q_{jk} \times q_{jk}$, $\mathbf{G}_{jk} = \mathbf{G}_{jk}(\boldsymbol{\lambda}_{jk})$, która zależy od wektora hiperparametrów $\boldsymbol{\lambda}_{jk}$. Jeśli \mathbf{G}_{jk} jest osobliwa, wówczas przyjmuje się, że $\boldsymbol{\gamma}_{jk}$ posiada funkcję gęstości proporcjonalną do $\exp(-\frac{1}{2}\boldsymbol{\gamma}_{jk}^T \mathbf{G}_{jk} \boldsymbol{\gamma}_{jk})$.

Postać (4.2) modelu GAMLSS umożliwia modelowanie każdego parametru rozkładu jako liniowe funkcje zmiennych objaśniających i/lub jako liniowe funkcje składników losowych (efekty losowe). Rzadko jednak wykonuje się modelowanie wszystkich parametrów rozkładu przy pomocy zmiennych objaśniających, [17].

Z ogólnej postaci modelu GAMLSS (4.2) po nałożeniu pewnych ograniczeń można wyrowadzić kilka ważnych i znanych modeli. Poniżej zostały przedstawione przykłady takich modeli.

Niech $\mathbf{Z}_{jk} = \mathbf{I}_n$, gdzie \mathbf{I}_n jest macierzą identyczności $n \times n$ oraz $\boldsymbol{\gamma}_{jk} = \mathbf{h}_{jk} = h_{jk}(\mathbf{x}_{jk})$ dla wszystkich kombinacji j i k (4.1). Wtedy wynikiem będzie semi-parametryczna addytywna formuła GAMLSS:

$$g_k(\theta_k) = \eta_k = \mathbf{X}_k \boldsymbol{\beta}_k + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}), \quad (4.3)$$

w skróconej notacji z θ_k dla $k = 1, 2, 3, 4$ do reprezentowania wektora parametrów rozkładu μ, σ, ν, τ (opisana wcześniej przez (4.1)). Jeśli nie ma addytywnych składników dla wszystkich parametrów rozkładu, wówczas jest to prosty liniowy parametryczny model GAMLSS:

$$g_k(\theta_k) = \eta_k = \mathbf{X}_k \boldsymbol{\beta}_k. \quad (4.4)$$

Model (4.3) może być rozszerzony tak, aby zawierał nieliniowe parametryczne składniki dla μ, σ, ν, τ :

$$g_k(\theta_k) = \eta_k = h_k(\mathbf{X}_k, \boldsymbol{\beta}_k) + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}), \quad (4.5)$$

gdzie $h_k(\cdot)$ dla $k = 1, 2, 3, 4$ są nieliniowymi funkcjami i \mathbf{X}_k jest znaną macierzą $n \times J_k''$. Do modelu (4.5) można odnosić się jak do nieliniowego semi-parametrycznego addytywnego modelu GAMLSS. Jeżeli $J_k = 0$ dla $k = 1, 2, 3, 4$, czyli dla wszystkich parametrów rozkładu nie występują addytywne składniki, wtedy model (4.5) redukuje się do nieliniowego parametrycznego modelu GAMLSS:

$$g_k(\theta_k) = \eta_k = h_k(\mathbf{X}_k, \boldsymbol{\beta}_k). \quad (4.6)$$

Jeśli dodatkowo $h_k(\mathbf{X}_k, \boldsymbol{\beta}_k) = \mathbf{X}_k^T \boldsymbol{\beta}_k$ dla $i = 1, 2, \dots, n$ i $k = 1, 2, 3, 4$, wówczas (4.6) redukuje się do liniowego, parametrycznego modelu (4.4). Warto zauważyc, że niektóre składniki w każdym $h_k(\mathbf{X}_k, \boldsymbol{\beta}_k)$ mogą być liniowe, wtedy model GAMLSS jest kombinacją liniowych i nieliniowych parametrycznych składników. Z tego wynika, że do każdej kombinacji modeli (4.4) lub (4.6) można odnosić się jak do parametrycznego modelu GAMLSS.

4.3. Estymacja modelu

W pracy [18] zostało pokazane, że nieznane wektory współczynników β_k i wektory efektów losowych γ_{jk} , $j = 1, 2, \dots, J_k$ i $j = 1, 2, 3, 4$ występujące w ogólnej postaci modelu GAMLSS (4.2), są estymowane dla ustalonych λ_{jk} , zgodnie ze strukturą modelu GAMLSS przez maksymalizację funkcji wiarygodności z karą l_p daną przez:

$$l_p = l - \frac{1}{2} \sum_{k=1}^p \sum_{j=1}^{J_k} \boldsymbol{\lambda}_{jk} \boldsymbol{\gamma}_{jk}^T \mathbf{G}_{jk} \boldsymbol{\gamma}_{jk}, \quad (4.7)$$

gdzie $l = \sum_{i=1}^n \log(f(y_i|\boldsymbol{\theta}^i))$ jest logarymem funkcji wiarygodności danych $\boldsymbol{\theta}^i$ dla $i = 1, 2, \dots, n$. Zostało udowodnione [18], iż maksymalizacja l_p może być osiągnięta przez zastosowanie algorytmu wielokrotnego dopasowania. Maksymalizacja l_p prowadzi do uzyskania ściągającej (wygładzającej) macierzy \mathbf{S}_{jk} . Macierz ta jest wykorzystywana do aktualizowania estymatora predyktora $\mathbf{Z}_{jk} \boldsymbol{\gamma}_{jk}$ przy pomocy algorytmu wielokrotnego dopasowania. \mathbf{S}_{jk} dana jest przez:

$$\mathbf{S}_{jk} = \mathbf{Z}_{jk} (\mathbf{Z}_{jk}^T \mathbf{W}_{kk} \mathbf{Z}_{jk} + \mathbf{G}_{jk})^{-1} \mathbf{Z}_{jk}^T \mathbf{W}_{kk} \quad (4.8)$$

dla $j = 1, 2, \dots, J_k$ i $k = 1, 2, 3, 4$, gdzie \mathbf{W}_{kk} jest diagonalną macierzą wag. Różne postacie \mathbf{Z}_{jk} i \mathbf{G}_{jk} odpowiadają różnym typom addytywnych składników w liniowym predyktorze η_k . Dla składnika wygładzonych kubicznych funkcji klejanych $\gamma_{jk} = \mathbf{h}_{jk}$, $\mathbf{Z}_{jk} = \mathbf{I}_n$ i $\mathbf{G}_{jk} = \boldsymbol{\lambda}_{jk} \mathbf{K}_{jk}$, gdzie \mathbf{K}_{jk} jest macierzą kary. Szczegółowy opis tych macierzy został zawarty w pracy [18].

Algorytmy wyznaczania modeli GAMLSS

Maksymalizacja logarytmu funkcji wiarygodności z karą l_p (4.7) odbywa się z wykorzystaniem różnych metod obliczeniowych. Istnieją dwa podstawowe algorytmy wykorzystywane do wyznaczenia modelu GAMLSS i maksymalizowania funkcji l_p .

Pierwszy, *algorytm CG* jest uogólnieniem algorytmu Cole'a - Green'a (1992). Korzysta on z pierwszych, drugich i mieszanych pochodnych funkcji wiarygodności ze względu na parametry $\boldsymbol{\theta} = (\mu, \sigma, \nu, \tau)$. Dla wielu rozkładów z funkcją $f(y|\boldsymbol{\theta})$, parametry $\boldsymbol{\theta}$ są ortogonalne, czyli wartości oczekiwane mieszanych pochodnych $f(y|\boldsymbol{\theta})$ są zerowe. W takim przypadku lepszym rozwiązaniem jest zastosowanie *algorytmu RS*, który stanowi uogólnienie algorytmu zaproponowanego przez Rigby'ego i Stasinopoulos'a w 1996 roku. Algorytm RS nie korzysta z wartości oczekiwanych pochodnych mieszanych i może być z powodzeniem wykorzystywany dla wszystkich rozkładów zawartych w tabelach 4.2 i 4.3, chociaż czasami wolniej zbiega, [18].

Zasadniczo algorytm RS posiada zewnętrzny cykl, który maksymalizuje funkcję l_p ze względu na β_k i γ_{jk} dla $j = 1, \dots, J_k$ w modelu dla każdej wartości $\boldsymbol{\theta}_k$, $k = 1, 2, 3, 4$ po kolei. Do każdego kolejnego kroku algorytmu są używane obecne, aktualizowane wartości wszystkich wielkości. Należy również zauważyć, że algorytm RS nie jest szczególnym przypadkiem algorytmu CG, ponieważ w algorytmie RS diagonalna macierz wag \mathbf{W}_{kk} , występująca we wzorze (4.8), jest aktualizowana w procesie dopasowywania każdego parametru $\boldsymbol{\theta}_k$, natomiast w algorytmie CG wszystkie wagi macierzy \mathbf{W}_{ks} , $k = 1, 2, 3, 4$, $s = 1, 2, 3, 4$ są wyznaczone po dopasowaniu wszystkich $\boldsymbol{\theta}_k$, $k = 1, 2, 3, 4$. Szczegółowy opis obu algorytmów można znaleźć w dodatku B pracy [18].

Podsumowując, celem algorytmów RS i CG jest maksymalizacja funkcji l_p dla ustalonych współczynników λ . Dla modeli całkowicie parametrycznych (4.4) lub (4.6) algorytmy maksymalizują funkcję l . Wybrana metoda maksymalizacji funkcji wiarygodności jest podawana w funkcji `gamlss()` jako argument `method`. Dozwolona jest także kombinacja obu

algorytmów. Główne zalety algorytmów dostępnych w `gamlss` to modułowy charakter procedury estymacji, co umożliwia inną diagnostykę modelu dla każdego parametru rozkładu. Dzięki konstrukcji algorytmów RS i CG, w łatwy sposób można dodawać dodatkowe rozkłady i addytywne składniki, co zostało opisane w [17]. Łatwo także znajdowane są wartości początkowe, których wyznaczenie wymaga tylko wcześniejszego zadania wartości początkowych dla parametrów $\boldsymbol{\theta} = (\mu, \sigma, \nu, \tau)$, a nie dla parametrów β . Ogólnie, algorytmy RS i CG są uznawane za stabilne i szybkie, gdy mają dla parametru $\boldsymbol{\theta}$ zadane bardzo proste wartości początkowe (np. stałe wartości) [17].

4.4. Liniowy predyktor w GAMLSS

W ogólnym modelu GAMLSS (4.2) liniowy predyktor $\boldsymbol{\eta}_k$ obejmuje zarówno liniowy $\mathbf{X}_k\boldsymbol{\beta}_k$, jak i losowy $\mathbf{Z}_{jk}\boldsymbol{\gamma}_{jk}$ składnik, $j = 1, \dots, J_k$, $k = 1, 2, 3, 4$. Parametryczna część zawiera składniki liniowe, interakcje występujące pomiędzy zmiennymi objaśniającymi, składniki wielomianowe, wymierne wielomiany oraz funkcje kawalkami wielomianowe (ze ustalonymi węzłami) zmiennych objaśniających. Składniki $\mathbf{Z}_{jk}\boldsymbol{\gamma}_{jk}$ w modelu (4.2) mogą uwzględniać m.in. wygładzające składniki, efekty losowe, jak również składniki, które są przydatne w analizie szeregów czasowych. Wybrane funkcje programu R, które odpowiadają za dodawanie do modelu wyżej wymienionych składników zostały przedstawione w rozdziale 4.5.4.

4.5. Pakiet `gamlss`

Oprogramowanie odpowiadające za wyznaczanie modeli GAMLSS zostało zaimplementowane jako seria pakietów w języku R. Autorami oprogramowania są M. Stasinopoulos, B. Rigby i C. Akantziliotou. Wszystkie wersje pakietów są dostępne pod adresem:

<http://CRAN.R-project.org/>.

W tej pracy korzystam z wersji 4.0-3 udostępnionej we wrześniu 2010 roku³ i skupiam się głównie na dwóch pakietach:

- oryginalnym pakiecie `gamlss` dopasowującym model GAMLSS,
- `gamlss.dist` zawierającym wszystkie dostępne rodziny rozkładów.

Szczegółowa dokumentacja dla powyższych pakietów znajduje się na stronie:

<http://cran.r-project.org/web/packages/gamlss/index.html>.

Opis pakietów i dostępnych funkcji wraz z ciekawymi przykładami ich zastosowania można znaleźć na stronie <http://www.gamlss.org/> w sekcji *Quick links* w postaci kilku odnośników do dokumentów pdf.

4.5.1. Różne funkcje w pakiecie `gamlss`

Pakiet `gamlss` udostępnia wiele funkcji, które można następująco pogrupować według ich funkcjonalności:

- funkcje wykorzystywane do dopasowania lub zmiany modelu: `gamlss()`, `refit()`, `update()` i `histDist()`.

³Od maja 2011 roku jest dostępna nowsza wersja 4.0-8 serii pakietów do dopasowania modelu GAMLSS.

- funkcje wydobywające właściwości z obiektów klasy `gamlss`: `AIC()`, `GAIC()`, `coef()`, `deviance()`, `extractAIC()`, `fitted()`, `formula()`, `fv()`, `logLik()`, `lp()`, `lpred()`, `model.frame()`, `model.matrix()`, `predict()`, `print()`, `summary()`, `terms()`, `vcov()` i `residuals()`. Niektóre z funkcji zależą od parametru rozkładu, tzn. mają dodatkowy argument `what`, który określa wymaganą wartość parametru. Przykładowo `fitted(m1, what = "sigma")` wyświetla dopasowane wartości dla parametru σ z modelu `m1`.
- funkcje wykorzystywane do wyboru modelu: `addterm()`, `dropterm()`, `find.hyper()`, `gamlss.scope()`, `stepGAIC()`, `stepGAIC.CH()`, `stepGAIC.VR()` i `VGD()`.
- funkcje wykorzystywane do rysowania wykresów lub diagnostyki: `plot()`, `par.plot()`, `pdf.plot()`, `Q.stats()`, `prof.dev()`, `prof.term()`, `rqres.plot()`, `show.link()`, `wp()` i `term.plot()`.
- funkcje stosowane do estymacji krzywych centylowych w przypadku, gdy występuje tylko jedna zmienna objaśniająca: `centiles()`, `centiles.com()`, `centiles.split()`, `centiles.pred()`, `fitted.plot()`.

Więcej informacji o wymienionych funkcjach można odnaleźć w pomocy R do pakietu `gamlss` [23] lub w dostępnej literaturze [17], [20]. Przykłady zastosowania niektórych funkcji zawiera rozdział 5.

4.5.2. Funkcja `gamlss()`

Główną funkcją pakietu `gamlss` jest funkcja `gamlss()`, która jest używana do dopasowania modelu GAMLSS, i w konsekwencji odpowiada za utworzenie obiektu typu `gamlss`. Funkcja `gamlss` swoją budową przypomina funkcje `gam()` z pakietu `gam` i `mgcv`, ale jest znacznie bardziej elastyczna – może dopasowywać więcej rozkładów (nie tylko te należące do rodziny rozkładów wykładniczych) oraz modelować wszystkie parametry rozkładu za pomocą funkcji zmiennych objaśniających. Obecna implementacja `gamlss()` pozwala na modelowanie do czterech parametrów rozkładu zwyczajowo nazwanych `mu`, `sigma`, `nu` i `tau`.

Wywołanie funkcji:

```
gamlss(formula = formula(data), sigma.formula = ~1,
nu.formula = ~1, tau.formula = ~1, family = NO(),
data = sys.parent(), weights = NULL,
contrasts = NULL, method = RS(), start.from = NULL,
mu.start = NULL, sigma.start = NULL,
nu.start = NULL, tau.start = NULL,
mu.fix = FALSE, sigma.fix = FALSE, nu.fix = FALSE,
tau.fix = FALSE, control = gamlss.control(...),
i.control = glim.control(...), ...)
```

Wybrane argumenty funkcji `gamlss` zostały opisane w tabeli 4.1.

Formuły funkcji `gamlss` akceptują wszystkie typy formuł używanych w `glm` i większość `gam`, oraz dodatkowo kilka innych zawierających składniki omówione w rozdziale 4.5.5.

4.5.3. Dostępne rozkłady

W GAMLSS postać rozkładu $f(y_i|\mu_i, \sigma_i, \nu_i, \tau_i)$ przyjętego dla zmiennej odpowiedzi y może być bardzo ogólna. Jedynym ograniczeniem jakie istnieje w implementacji R dla GAMLSS jest różniczkowalność funkcji $\log f(y_i|\mu_i, \sigma_i, \nu_i, \tau_i)$ oraz istnienie jej pierwszej pochodnej (opcjonalnie drugiej i mieszanych pochodnych) ze względu na każdy z parametrów θ .

Tabela 4.1: Argumenty funkcji `gamlss()`, Źródło: [23]

formula	formuła modelu dla <code>mu</code> ze zmienną objaśnianą po lewej stronie operatora \sim i zmiennymi objaśniającymi z prawej strony oddzielonymi operatorem <code>+</code>
sigma.formula	formuła modelu dla parametru <code>sigma</code> ; w tej samej postaci można zapisywać formuły dla parametrów <code>nu</code> i <code>tau</code>
family	definicja rozkładu zmiennej odpowiedzi; dostępnymi rozkładami są rozkłady zawarte w tabelach 4.2, 4.3; domyślnie zmienna <code>family</code> przyjmuje rozkład normalny (<code>NO</code>)
data	dane zawierające zmienne występujące w modelu
weights	wektor wag
contrasts	lista kontrastów stosowana do niektórych lub wszystkich zmiennych z modelu
method	algorytmy dla GAMLSS np. <code>RS()</code> , <code>CG()</code> lub <code>mixed()</code> , rozdział 4.5.3
start.from	model GAMLSS, dla którego dopasowane wartości są uznane jako początkowe w bieżącym modelu
mu.start	wektor skalarów wartości początkowych dla parametru <code>mu</code> ; w analogiczny sposób można przekazać wartości początkowe dla parametrów <code>sigma</code> , <code>nu</code> i <code>tau</code>
mu.fix	określa czy parametr <code>mu</code> będzie przyjmował stałą wartość w procesie dopasowania modelu; analogicznie definiuje się określenie dla parametrów <code>sigma</code> , <code>nu</code> i <code>tau</code>
control	parametr kontrolujący iteracje algorytmu

Tabela 4.2 przedstawia różne jedno-, dwu-, trzy- i czteroparametrowe rodziny ciągłych rozkładów, które są dostępne w aktualnej wersji oprogramowania, natomiast tabela 4.3 zawiera rozkłady dyskretnie. Rozkłady obu tabel są wykorzystywane jako argumenty zmiennej `gamlss.family`. Szczegółowy opis wszystkich rozkładów można znaleźć m.in. w [17]. Niektóre z ciągłych rozkładów zawartych w tabeli 4.2 i wykorzystywanych w rozdziale 5, zostały omówione w dodatku B.4.

Wszystkie rozkłady zawarte w tabelach (4.2) i (4.3) mają funkcje o nazwie danego rozkładu z dodanym **prefixem**: `d`, `p`, `q` i `r`. Funkcje o tych nazwach oznaczają odpowiednio: funkcję gęstości prawdopodobieństwa, dystrybuantę danego rozkładu, funkcję wyznaczającą wartości kwantyl (tj. odwrotność dystrybuanty) dla danego rozkładu w punktach podanych jako argument tej funkcji, oraz funkcję generującą losowe wartości z danego rozkładu. Przykładowo, dla rozkładu gamma (ozn. `GA`) są dostępne funkcje: `dGA`, `pGA`, `qGA` i `rGA`.

Dodatkowo każdy rozkład posiada tzw. funkcję „fitting”, czyli funkcję odpowiadającą za dopasowanie rozkładu, która pomaga w procedurze wyznaczania modelu poprzez dostarczenie funkcji wiążących, pierwszej i drugiej pochodnej, wartości początkowych, itp. Wszystkie funkcje „fitting” jako argumenty przyjmują funkcje wiiązające dla danego parametru rozkładu. Domyślne funkcje wiiązające dla wszystkich rozkładów `gamlss.family` są podane odpowiednio w kolumnach 3-6 tabeli (4.2) oraz w kolumnach 3-5 tabeli (4.3).

Tabela 4.2: Rozkłady ciągłe realizowane w ramach pakietu `gamlss` (z domyślnymi funkcjami wiążącymi). Źródło: [17].

nazwa rozkładu	nazwa w gamlss	μ	σ	ν	τ
beta	BE()	logit	logit	-	-
beta inflated (at 0)	BEOI()	logit	log	logit	-
beta inflated (at 1)	BEZI()	logit	log	logit	-
beta inflated (at 0 and 1)	BEINF()	logit	logit	log	log
Box-Cox Cole and Green	BCCG()	identity	log	identity	-
Box-Cox power exponential	BCPE()	identity	log	identity	log
Box-Cox-t	BCT()	identity	log	identity	log
exponential	EXP()	log	-	-	-
exponential Gaussowski	exGAUS()	identity	log	log	-
exponential gen. beta type 2	EGB2()	identity	identity	log	log
gamma	GA()	log	log	-	-
generalized beta type 1	GB1()	logit	logit	log	log
generalized beta type 2	GB2()	log	identity	log	log
generalized gamma	GG()	log	log	identity	-
generalized inverse Gaussian	GIG()	log	log	identity	-
generalized y	GT()	identity	log	log	log
Gumbel	GU()	identity	log	-	-
inverse Gaussian	IG()	log	log	-	-
Johnson's SU (μ the mean)	JSU()	identity	log	identity	log
Johnson's original SU	JSUo()	identity	log	identity	log
logistic	LO()	identity	log	-	-
log normal	LOGNO()	log	log	-	-
log normal (Box-Cox)	LNO()	log	log	fixed	-
NET	NET()	identity	log	fixed	fixed
normal	NO()	identity	log	-	-
normal family	NOF()	identity	log	identity	-
power exponential	PE()	identity	log	log	-
reverse Gumbel	RG()	identity	log	-	-
skew power exponential type 1	SEP1()	identity	log	identity	log
skew power exponential type 2	SEP2()	identity	log	identity	log
skew power exponential type 3	SEP3()	identity	log	log	log
skew power exponential type 4	SEP4()	identity	log	log	log
sinh-arcsinh	SHASH()	identity	log	log	log
skew t type 1	ST1()	identity	log	identity	log
skew t type 2	ST2()	identity	log	identity	log
skew t type 3	ST3()	identity	log	log	log
skew t type 4	ST4()	identity	log	log	log
skew t type 5	ST5()	identity	log	identity	log
t Family	TF()	identity	log	log	-
Weibull	WEI()	log	log	-	-
Weibull (PH)	WEI2()	log	log	-	-
Weibull (μ the mean)	WEI3()	log	log	-	-
zero adjusted IG	ZAIG()	log	log	logit	-

Tabela 4.3: Rozkłady dyskretne realizowane w ramach pakietu `gamlss` (z domyślnymi funkcjami wiążącymi). Źródło: [17].

nazwa rozkładu	nazwa w gamlss	μ	σ	ν
beta binomial	BB()	logit	log	-
binomial	BI()	logit	-	-
Delaporte	DEL()	log	log	logit
Negative Binomial type I	NBI()	log	log	-
Negative Binomial type II	NBII()	log	log	-
Poisson	PO()	log	-	-
Poisson inverse Gaussian	PIG()	log	log	-
Sichel	SI()	log	log	identity
zero inflated poisson	ZIP()	log	log	-

4.5.4. Addytywne składniki

Ogólna postać modelu GAMLSS (4.2) umożliwia modelowanie wszystkich parametrów rozkładu μ, σ, ν i τ jako liniowe parametryczne funkcje, nieliniowe parametryczne lub nieparametryczne wygładzone funkcje zmiennych wyjaśniających czy też składniki efektów losowych. Do modelowania funkcji liniowej może zostać użyta notacja stosowana w modelach liniowych `lm()` i uogólnionych modelach liniowy `glm()`. Dla dopasowania funkcji nieliniowych lub nieparametrycznych funkcji gładkich, czy efektów losowych należy wybrać addytywny składnik. Poniższa tabela 4.4 przedstawia addytywne składniki funkcji realizowanych w wersji 4.0-3 pakietu `gamlss`.

Tabela 4.4: Addytywne składniki udostępnione w pakiecie `gamlss`. Źródło: [17].

addytywny składnik	nazwa w gamlss
cubic splines	<code>cs()</code>
varying coefficient	<code>vc()</code>
penalized splines	<code>ps()</code>
loess	<code>lo()</code>
fractional polynomials	<code>fp()</code>
power polynomials	<code>pp()</code>
non-linear fit	<code>n1()</code>
random effects	<code>random()</code>
random effects	<code>ra()</code>

Najpopularniejszym i najczęściej stosowanym addytywnym składnikiem w GAMLSS jest składnik `cs()` wykorzystywany do jednowymiarowego wygładzania przy pomocy kubicznych funkcji sklejanych. Funkcja `cs()` opierają się na funkcji `smooth.spline()` opisanej w rozdziale 2.2. Ta metoda wymaga spełnienia założenia, że dla modelu (4.1) funkcje $h(t)$ są dwukrotnie różniczkowalne w sposób ciągły. Idea polega na zmaksymalizowaniu logarytmu funkcji wiarygodności l ze składnikiem kary postaci $\lambda \int_{\infty}^{\infty} [h''(t)]^2 dt$. Rozwiążanie maksymalizacji jest naturalną kubiczną funkcją sklejaną, a zatem może być wyrażone jako liniowa kombinacja naturalnych bazowych kubicznych funkcji sklejanych.

Innym składnikiem dostępnym w `gamlss` jest funkcja `vc()`, która została wprowadzona w celu włączenia do modelu szczególnego rodzaju interakcji pomiędzy dwoma zmiennymi wyjaśniającymi oznaczonymi `r` i `x`. Te interakcje przyjmują postać $\beta(r)x$, zatem liniowy

współczynnik zmiennej objaśniającej x zmienia się ze względu na inną zmienną r . W pewnych zastosowaniach r jest zmienną określającą czas. W ogólności r powinna być zmienną ciągłą, a x może być ciągłą lub kategoryczną zmienną. Kolejny składnik to `ps()`, czyli *penalized splines* (tzw. *P-splines*), które są funkcjami kawałkami wielomianowymi zdefiniowanymi przez bazowe funkcje sklejane (B-splajny) zmiennych objaśniających. Podejście zastosowane w `ps()` polega na tym, iż współczynniki bazowych funkcji zostały „ukarane” aby zagwarantować dostateczną gładkość. Kolejna funkcja ozn. `lo()` umożliwia korzystanie z metody wygładzania przy pomocy lokalnie wygładzanych wielomianów pierwszego lub drugiego stopnia, a funkcja `fp()` jest realizacją wymiernych wielomianów. Kombinacje liniowe funkcji potęgowych `pp()` w `gamlss` służą do dopasowania modeli postaci $b_0 + b_1x^{p_1} + b_2x^{p_2}$ z zadanimi wykładnikami p_1, p_2 . Funkcja `n1()` odpowiada za dopasowanie nieliniowego parametrycznego modelu umożliwiając dopasowanie nieliniowego składnika razem z liniowym lub wygładzonym składnikiem. Funkcje `random()` oraz `ra()` umożliwiają włączanie efektów losowych do modelu GAMLSS. Przykładowo funkcja `random()` pozwala by dopasowane wartości dla predyktora, który jest czynnikiem grupującym, były „ściągające” w kierunku całkowitej średniej. Wielkość „ściągania” zależy albo od parametru λ lub od liczby stopni swobody `df`. Obie funkcje `random()` i `ra()` dla danych parametrów λ są używane do oszacowania efektu losowego γ .

Funkcja `gamlss()` używa tego samego rodzaju addytywnego algorytmu wielokrotnego dopasowania, jaki został zaimplementowany w pakiecie `gam()`. Powodem korzystania algorytmu wielokrotnego dopasowania w `gamlss` jest to, że łatwo można rozszerzać ten algorytm o dodawanie nowych addytywnych składników. Każdy addytywny składnik w `gamlss()` to w rzeczywistości dwie funkcje. Pierwsza funkcja, która jest widziana przez użytkownika, definiuje addytywny składnik i określa dodatkowe macierze projektu wymagane dla definiowania liniowej części modelu. Przykładowo, składnik `cs(x)` definiujący kubiczne wygładzone funkcje sklejane dla ciągłej zmiennej objaśniającej x , jest używany podczas określania macierzy projektu dla odpowiedniego parametru rozkładu oraz dodawania liniowego składnika x w macierzy projektu. Druga funkcja odpowiada za faktyczne wykonanie algorytmu wielokrotnego dopasowania. Ta funkcja jest wywoływana przez `gamlss.name()`, gdzie `name` jest jedną z nazw funkcji drugiej kolumny tabeli 4.3. tj. aktualnie dostępnych addytywnych funkcji. Odwołując się do przykładu składnika `cs(x)` – funkcja `gamlss.cs()` wykonuje dopasowanie dla kubicznych funkcji sklejanych. Ogólną zasadą podczas wykonywania algorytmu wielokrotnego dopasowania w `gamlss` jest włączenie części liniowej addytywnych składników do odpowiedniego składnika macierzy projektu. Dla kubicznych splajnów `cs()` zmienna objaśniająca x jest umieszczana w liniowej macierzy z odpowiednim parametrem rozkładu i dopasowanie wygładzającej funkcji odbywa się jako odchylenie od tej części liniowej. Jest to równoważne dopasowaniu pewnej modyfikacji algorytmu wielokrotnego dopasowania. Stopnie swobody dla wygładzonych addytywnych składników uznaje się na początku dopasowania za dodatkowe stopnie swobody. Na przykład pojedyncza wygładzona kubiczna funkcja sklejana dla x z całkowitą liczbą stopni swobody równą 5 powinna zostać zapisana jako `cs(x, df = 3)`, gdyż 2 stopnie swobody wykorzystane zostaną do dopasowania stałego czynnika i liniowej części modelu zmiennej objaśniającej x . To jest różnica pomiędzy funkcją `s()` z pakietu `gam`, który używa formuły `s(x, df = 4)` zakładając, że tylko stały składnik został wybrany oddziennie.

Przykłady dopasowania modeli GAMLSS za pomocą funkcji `gamlss` z addytywnymi składnikami wykorzystującymi kubiczne funkcje sklejane `cs` zostały podane w rozdziale 5.

4.5.5. Kilka słów o budowie modelu

Ze względu na elastyczność i szeroki zakres możliwości GAMLSS, procedura wyznaczania „najlepszego” (w pewnym sensie) modelu nie jest prostym zadaniem. Ogólnie, reprezentację modelu GAMLSS można zapisać następująco: $\mathcal{M} = \{\mathcal{D}, \mathcal{G}, \mathcal{T}, \lambda\}$, gdzie odpowiednie składniki definiują: \mathcal{D} - rozkład zmiennej odpowiedzi, \mathcal{G} - zbiór funkcji wiążących (g_1, g_2, g_3, g_4) dla parametrów (μ, σ, ν, τ), \mathcal{T} - zbiór predyktorów ($t_\mu, t_\sigma, t_\nu, t_\tau$) dla ($\eta_\mu, \eta_\sigma, \eta_\nu, \eta_\tau$), λ - zbiór współczynników wygładzania.

Proces budowy modelu GAMLSS odbywa się przez porównywanie wielu różnych modeli dla których są sprawdzane różne kombinacje składników $\mathcal{M} = \{\mathcal{D}, \mathcal{G}, \mathcal{T}, \lambda\}$. Wnioskowanie o korzyściach jednego modelu nad innymi może polegać albo na wybraniu pojedynczego „ostatecznego” modelu albo poprzez uśrednienie wybranych modeli. W wyborze modelu GAMLSS mogą być wykorzystywane różne strategie szeroko opisywane w literaturze np. [17], [20]. Jednak ważniejsze jest sprawdzenie adekwatności modelu w odniesieniu do kwestii merytorycznych, a nie w oderwaniu od całości. Oznacza to, że różne problemy mogą wymagać różnych strategii wyboru modelu.

4.6. Podsumowanie GAMLSS

Modele GAMLSS stanowią bardzo ogólną klasę modeli będącą rozwiązaniem dla wielu problemów statystycznych napotykanych do tej pory w regresji z jednowymiarową zmienną odpowiedzi. Modele te wprowadzają jednolite ramy, uogólniają estymację oraz wnioskowanie dla różnych typów modeli, które zostały wprowadzone przez naukowców w ciągu ostatnich kilku dekad, [18]. Mogą być zatem bardzo pożytecznym narzędziem do celów edukacyjnych. Co więcej, autorzy GAMLSS udoskonaliли istniejące już typy modeli rozszerzając je o możliwość modelowania, poza średnią i wariancją – dalszych parametrów rozkładu. Umożliwiły korzystanie z bardzo szerokiej rodziny rozkładów dla zmiennej odpowiedzi, co automatycznie powoduje zmniejszenie niebezpieczeństwa niedopasowania rozkładu. Ważną zaletą GAMLSS jest możliwość określenia w elastyczny sposób rozkładu zmiennej odpowiedzi włączając w to wysoce skośne i/lub kurtyczne rozkłady. Modele te także pozwalają estymować wszystkie parametry rozkładu przy pomocy szerokiej gamy różnych addytywnych składników. Położenie, skala, asymetria i kurtoza – każdy z tych parametrów może być modelowany, gdy zachodzi taka potrzeba. Szczególnie zachęcającymi aspektami GAMLSS jest możliwość dodawania do modelu składników nieparametrycznych tj. wygładzeń, które mają wiele praktycznych możliwości w dopasowaniu relacji pomiędzy zmiennymi wyjaśniającymi i zmienną wyjaśnianą. Algorytm modelu GAMLSS jest szybki dla bardzo dużych i złożonych zestawów danych. Dzięki temu zanim zostanie dokonany ostateczny wybór danego modelu, lub zanim uwzględniona zostanie kombinacja różnych modeli, można badać i dopasowywać wiele alternatywnych modeli. Modułowy charakter algorytmu umożliwia użytkownikom włączanie nowych alternatywnych rozkładów oraz nowych addytywnych składników. Dla średnich i dużych rozmiarów danych modele GAMLSS stanowią wyjątkowo elastyczną metodą modelowania statystycznego, [18]. Ta elastyczność pozwala na bardziej realistyczne założenia dotyczące danych. Za wadę modelowania przy pomocy metody GAMLSS można uznać fakt, iż trudno jest wybrać najlepszy model. W modelowaniu z wykorzystaniem GAMLSS ta część wymaga największego wkładu pracy.

Rozdział 5

Analiza danych medycznych z wykorzystaniem gamlss

Celem niniejszego rozdziału jest zastosowanie metody GAMLSS do wyznaczania *wartości referencyjnych*¹ wysokości ciśnienia skurczowego. Analizy przeprowadzone w tym rozdziale zostaną wykonane na podstawie danych dla dzieci i młodzieży płci żeńskiej w wieku 6.5 – 18.5 lat, z wykluczeniem osób z nadwagą. Otrzymane wyniki posłużą do przeliczenia wartości ciśnienia tętniczego na adekwatne *poziomy centylowe*², które będą stanowić podstawę interpretacji diagnostycznej ciśnienia tętniczego u dzieci i młodzieży w Polsce.

Podrozdział 5.1 zawiera przeprowadzone analizy, które mają na celu skonstruowanie krzywych centylowych oraz wyznaczenie wartości centyl wzrostu dla danej populacji. W kolejnym podrozdziale przedstawiona jest analiza ciśnienia skurczowego w zależności od dwóch zmiennych objaśniających – wieku i wzrostu. Ostatni podrozdział stanowi podsumowanie dwóch poprzednich i zawiera wyniki analiz zebrane w postaci tabeli, która może stanowić podstawę diagnostyki choroby nadciśnieniowej.

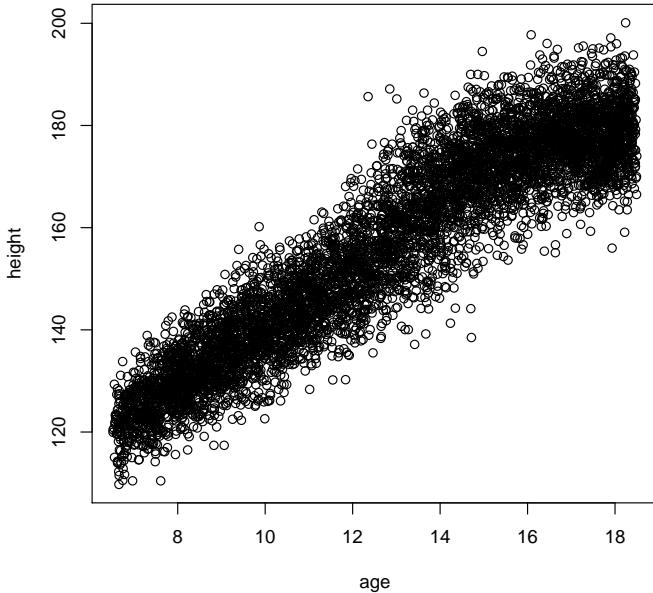
5.1. Krzywe centylowe wzrostu dla chłopców

Powołując się na opinie autorów metody GAMLSS zaktualizowaną w pracy [7] oraz na raport *WHO Child Growth Standards: Methods and Development* [24], metoda GAMLSS została uznana przez WHO (Światałową Organizację Zdrowia) za standardową metodę konstrukcji krzywych rozwoju dziecka. W tym podrozdziale zastosuję zatem GAMLSS do wyznaczenia centyl wzrostu oraz konstrukcji krzywych centylowych dla dzieci i młodzieży płci żeńskiej. Do analiz wykorzystam 6672 obserwacje z bazy danych `danemed` ze zmienną objaśnianą $y = \text{height}$ oraz zmienną objaśniającą $x = \text{age}$. Analizowane dane zostały przedstawione na rys. 5.1. Zarówno `height` jak i `age` są zmiennymi ciągłymi.

Dla znanego $X = x$, Y będzie modelowane przy pomocy rozkładu Box-Cox-Cole-Green'a, oznaczanego $BCCG(\mu, \sigma, \nu)$, który został zdefiniowany w dodatku B.5.3. Rozkład ten stanowi transformację modelu Box'a-Cox'a w wersji wykorzystanej przez Cole'a i Green'a. Metoda wyznaczania centyl za pomocą tego rozkładu w środowisku medycznym znana jest pod nazwą *metody LMS*, [15]. Parametry μ, σ i ν z rozkładu $BCCG$ (dla $Y \sim BCCG(\mu, \sigma, \nu)$) zostaną tutaj wyznaczone za pomocą pewnego przypadku modelu GAMLSS (4.1) jako gładkie

¹Wartości referencyjne/normy – wartości, które są opracowywane na podstawie analizy rozkładu statystycznego wysokości danej zmiennej w pewnej populacji uznanej za zdrową, [12]

²Poziom/wartość centylowa analizowanej cechy – liczba wskazująca odsetek osób w populacji o mniejszej wartości cechy.



Rysunek 5.1: Wzrost w centymetrach (`height`) dla chłopców ze względu na zmienną `age`.

nieparametryczne funkcje x :

$$\begin{aligned} g_1(\mu) &= h_1(x) \\ g_2(\sigma) &= h_2(x) \\ g_3(\nu) &= h_3(x) \end{aligned} \tag{5.1}$$

Zgodnie z oznaczeniami z rozdziału 4, dla $k = 1, 2, 3$ odpowiednie $g_k(\cdot)$ są znanimi monotonicznymi funkcjami wiążącymi, natomiast $h_k(x)$ to funkcje zmiennej x wygładzone za pomocą wybranych nieparametrycznych składników. Wykorzystywaną przeze mnie w tym rozdziale metodą wygładzającą są kubiczne funkcje sklejane, które zostały udostępnione w pakiecie `gamlss` w funkcji `cs()`, (rozdział 4.5.4).

Procedura modelowania składa się z wyboru funkcji wiążących $g_k(\cdot)$ oraz liczby efektywnych stopni swobody dla wygładzonych kubicznych funkcji sklejanych $h_k(x)$, $k = 1, 2, 3$, które zostaną oznaczone odpowiednio df_μ , df_σ , df_ν . W tym przykładzie wybieram domyślne funkcje wiążące dla rozkładu *BCCG*, czyli identycznościową funkcję wiążącą dla μ i ν oraz logarytmiczną funkcję wiążącą dla σ (aby zapewnić $\sigma > 0$). Do wyznaczenia parametrów df_μ , df_σ i df_ν wykorzystana została automatyczna procedura `find.hyper()` oparta na numerycznej optymalizacji funkcji `optim` w R, którą Rigby i Stasinopoulos zastosowali do zminimalizowania $GAIC(penalty) = -2\hat{l} + penalty \ df$, gdzie: \hat{l} – maksimum logarytmu funkcji wiarygodności, $penalty$ – kara, i df – całkowita liczba efektywnych stopni swobody w modelu. Korzystając z funkcji `find.hyper` należy być ostrożnym, gdyż $GAIC(penalty)$ może potencjalnie mieć wiele lokalnych minimum, zatem aby zapewnić znalezienie minimum globalnego procedurę warto uruchomić z różnymi wartościami początkowymi.

Poniższy kod R został użyty do znalezienia parametrów odpowiadających domyльнemu argumentowi funkcji `find.hyper()`, czyli `penalty=2`. W funkcji `cs()` argument `c.spar` zapewnia, że liczba stopni swobody jest w stanie przyjąć małe wartości. Parametry df_μ , df_σ , df_ν

w poniższym kodzie są reprezentowane przez składniki `p[1]`, `p[2]` i `p[3]`. Argument `par` określa wartości początkowe, natomiast `lower` zawiera dolne granice parametrów.

```
> mod1<-quote(gamlss(height~cs(age,df=p[1]),sigma.fo=~cs(age,df=p[2]),
+ nu.fo=~cs(age, df=p[3]),
+ family=BCCG, data=danemed,
+ control=gamlss.control(trace=FALSE)))
> op<-find.hyper(model=mod1, par=c(3,1,1),
+ lower=c(0.1,0.1,0.1), steps=c(0.05,0.05,0.05))
par 3 1 1 crit= 44149.17 with pen= 2
par 3.05 1 1 crit= 44146.33 with pen= 2
par 2.95 1 1 crit= 44152.21 with pen= 2
...
par 12.16354 8.447437 6.802487 crit= 44034.09 with pen= 2
```

Procedura trwała bardzo długo (ponad godzinę!). Wynikowe wartości dla parametrów df_μ , df_σ , df_ν oraz wartość `GAIC()` zostały zawarte odpowiednio w zmiennej `op$par` i `op$value`.

```
> op$par
[1] 12.16354 8.447437 6.802487
> op$value
[1] 44034.09
```

Otrzymane wartości stopni swobody nie uwzględniają stopni swobody dla stałych i liniowych składników. W celu uzyskania całkowitej liczby stopni swobody dla danego parametru rozkładu do każdej wartości należy dodać 2 (np. całkowita liczba stopni swobody dla modelu μ to $12.16 + 2 = 14.16$).

Ta sama procedura została uruchomiona z parametrem `penalty` równym 3 oraz $\log(n) = 8.8$ (ponieważ n to liczba obserwacji, zatem $\text{penalty} = \log(6672) = 8.8$). Tabela 5.1 przedstawia uzyskaną liczbę stopni swobody z optymalizacji wykonanej z wykorzystaniem funkcji `find.hyper` dla różnych wartości kary.

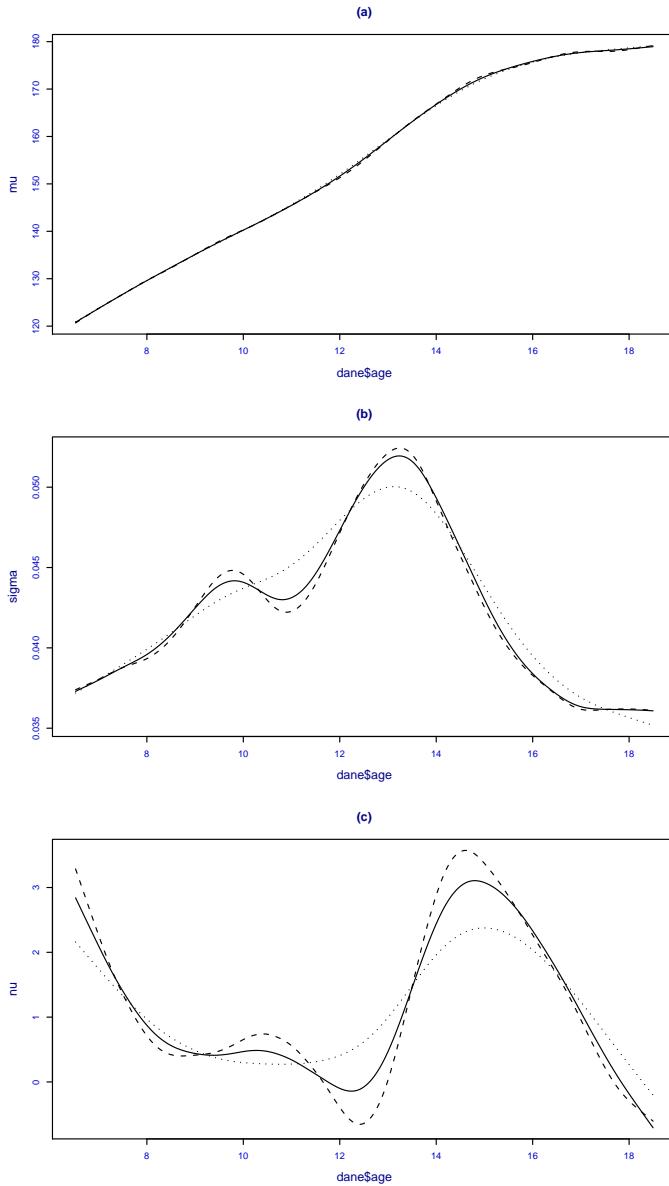
Tabela 5.1: Wartości stopni swobody uzyskane z optymalizacji w modelu *BCCG*.

penalty	df_μ	df_σ	df_ν
2	14.16	10.45	8.8
3	8.87	8.7	7
$\log(n)$	6.75	5.71	4.89

Poniższe trzy modele `m2`, `m3`, `m8` zostały utworzone w oparciu o uzyskane wartości stopni swobody zawarte w tabeli 5.1:

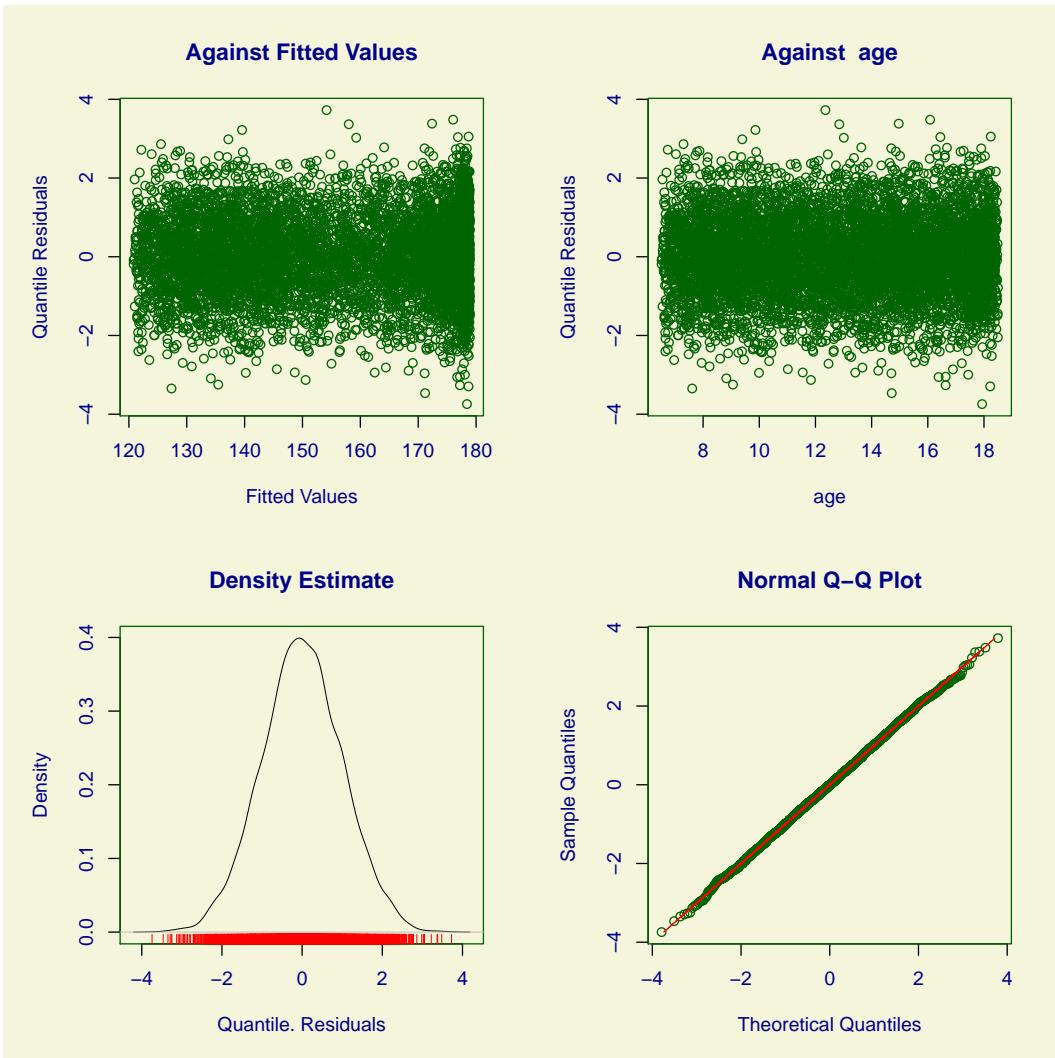
```
> m2<-gamlss(height~cs(age,df=12.16),sigma.fo=~cs(age,df=8.45),
+ nu.fo=~cs(age, df=6.8), family=BCCG,
+ data=danemed, control=gamlss.control(trace=FALSE,c.crit = 0.1))
> m3<-gamlss(height~cs(age,df=6.87),sigma.fo=~cs(age,df=6.7),
+ nu.fo=~cs(age, df=5), family=BCCG,
+ data=danemed, control=gamlss.control(trace=FALSE,c.crit = 0.1))
> m8<-gamlss(height~cs(age,df=4.75),sigma.fo=~cs(age,df=3.71),
+ nu.fo=~cs(age, df=2.89), family=BCCG,
+ data=danemed, control=gamlss.control(trace=FALSE,c.crit = 0.1))
```

Rysunek 5.2 przedstawia dopasowane wartości otrzymane z modeli `m2`, `m3` i `m8` dla parametrów rozkładu μ , σ , ν . Dopasowanie dla μ w przypadku trzech modeli jest bardzo podobne. Dla σ i ν , w zależności od wielkości kary, modele różnią się między sobą. Kryterium AIC posłużyło nam do wybrania modelu, który posiada dużą liczbę stopni swobody, co przełożyło



Rysunek 5.2: Dopasowane parametry ze względu na zmienną `age` dla „najlepszego” modelu *BCCG* z parametrami wybranymi w wyniku zastosowania kryterium AIC (---), GAIC(3) (—) oraz BIC (...): (a) μ , (b) σ , (c) ν .

się na znaczne dopasowanie do danych. Wynikowy model otrzymany za pomocą kryterium BIC jest bardzo wygładzony i posiada znacznie mniejszą liczbę stopni swobody w porównaniu z modelem `m2`. Ogólnie, można zaobserwować, że większa kara prowadzi do zmniejszenia liczby stopni swobody, a tym samym do prostszego modelu z bardziej wygładzonymi dopasowanymi wartościami dla μ , σ , ν , czego następstwem są m.in. bardziej wygładzone krzywe centylowe. Ze względu na mniejszą liczbę stopni swobody dla `m8` oraz wygładzony charakter krzywych na rysunku 5.2, można przypuszczać, że model otrzymany z wykorzystaniem kryterium BIC jest najlepszym wyborem. Powołując się na diagnostykę reszt dla modeli `m2`, `m3` i `m8` z wykorzystaniem funkcji `wp`, o której jest mowa w dalszej części rozdziału, do dalszych analiz wybieram model `m3`, czyli model *BCCG*(8.87, 8.7, 7).



Rysunek 5.3: Wykresy diagnostyczne dla reszt z modelu $BCCG(8.87, 8.7, 7)$. Kolejne wykresy przedstawiają reszty (a) ze względu na dopasowane wartości μ , (b) ze względu na zmienną `age`, (c) estymator gęstości jądra, (d) wykres kwantylowy dla rozkładu normalnego.

Diagnostykę dla wybranego modelu przeprowadzę wzorując się na podrozdziale 3.5 pracy [19]. Wykorzystam do tego zalecone do diagnostyki reszt funkcje pakietu `gamlss`: `plot()`, `wp()` oraz `Q.stats()`.

Następujące wywołanie:

```
> plot(m3)
*****
      Summary of the Quantile Residuals
      mean      =  0.000423349
      variance  =  1.000144
      coef. of skewness = -0.0007645265
      coef. of kurtosis =  2.977314
      Filliben correlation coefficient =  0.9999099
*****
```

odpowiada za wypisanie pewnych parametrów statystycznych dla reszt uzyskanych z modelu

`m3`, oraz tworzy zestaw wykresów, które zostały przedstawione na rysunku 5.3. Dla znormalizowanych kwantylowych reszt wyznaczona została: średnia, wariancja, współczynnik skośności i kurtozy oraz *współczynnik korelacji Filliben'a*³. Otrzymane wyniki dla tych parametrów utwierdzają w przekonaniu, iż reszty modelu *BCCG(8.87, 8.7, 7)* mają rozkład normalny.

Rysunek 5.3 przedstawia wykresy diagnostyczne dla znormalizowanych kwantylowych reszt. Odpowiednio rysunki: (a) i (b) przedstawiają reszty ze względu na dopasowane wartości dla μ i ze względu na zmienną `age`. Dla adekwatnego modelu reszty mają jednorodną wariancję i średnią równą zero niezależnie od wartości \hat{y}_i . Na tych wykresach widać, że wartość średnia reszt nie zależy od \hat{y}_i , jednak występują obserwacje dla których reszty > 2 . Rysunki (c) i (d) dostarczają estymator gęstości jądra i QQ-plot czyli tzw. wykres kwantylowy dla rozkładu normalnego. Jest to zależność pomiędzy wartościami zmiennej, a kwantylami rozkładu normalnego. W idealnym przypadku, jeśli rozkład zmiennej jest czysto normalny, wykres ten przedstawia linię prostą. QQ-plot nie wykazuje widocznych obserwacji odstających, można przypuszczać zatem, że model stanowi adekwatne dopasowanie do danych.

Kolejne kryterium wykorzystane do diagnozy modelu w znacznym stopniu zaważyło na wyborze przeze mnie modelu `m3` i wyższości tego modelu nad modelem `m8`.

Za pomocą następującego kodu R:

```
> wp.diagnostyka<-wp(m3, xvar = age, n.inter = 12, ylim.worm = 0.6,
+ cex = 0.3, pch = 20)
number of missing points from plot= 0 out of 557
number of missing points from plot= 0 out of 556
number of missing points from plot= 0 out of 555
number of missing points from plot= 0 out of 557
number of missing points from plot= 1 out of 555
number of missing points from plot= 1 out of 557
number of missing points from plot= 0 out of 555
number of missing points from plot= 0 out of 556
number of missing points from plot= 1 out of 556
number of missing points from plot= 0 out of 558
number of missing points from plot= 0 out of 554
number of missing points from plot= 1 out of 556
```

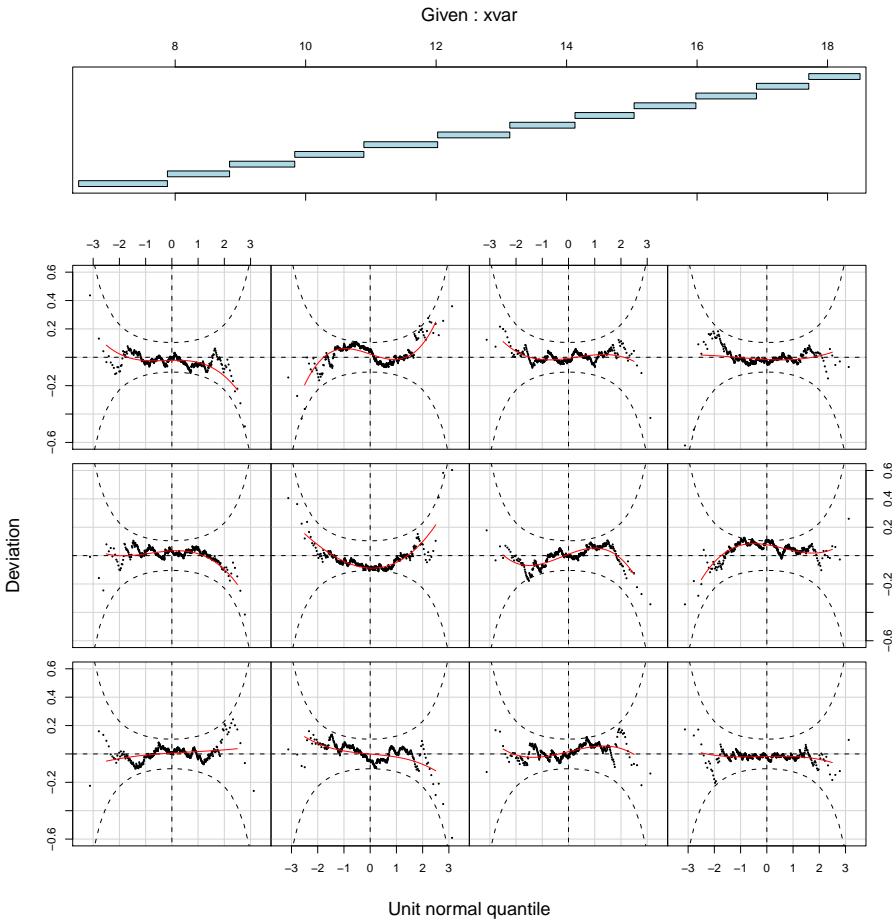
zostały utworzone wykresy przedstawione na rysunku 5.4. Rysunek ten obejmuje 12 wykresów, które zostały wyznaczone przez zmienną `n.inter = 12` dla podzielonej na 12 przedziałów zmiennej `age`, określonej przez `xvar=age`. Wykresy tego typu nazwane są często `worm plot`'ami lub *Detrended Normal QQ-plot* i dla danego modelu przedstawiają dokładną diagnozę reszt. Używana tutaj funkcja `wp` wprowadzona została przez van Buuren'a i Fredriks'a w 2001 roku. Wartości zmiennej `age` zostały podzielone na 12 sąsiadujących rozłącznych równiczych przedziałów wiekowych. Przedziały te zostały wyświetlane na górnym wykresie rysunku 5.4. Wykres `worm plot`, przedstawia różnice pomiędzy obserwowanymi, a oczekiwaniemi.

³ *Współczynnik korelacji Fillibena r* dla obserwacji $X_i, i = 1, 2, \dots, n$ zdefiniowany jest jako wartość korelacji pomiędzy uporządkowanymi niemalejąco obserwacjami X_i , a M_i , które są medianami statystyk pozycyjnych pochodzących ze standardowego rozkładu normalnego. Wartość współczynnika r wyznacza wzór:

$$r = Cor(X, M) = \frac{\sum_{i=1}^n (X_i - \bar{X})(M_i - \bar{M})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (M_i - \bar{M})^2}},$$

gdzie odpowiednio \bar{X} i \bar{M} oznaczają średnią z próby $X_i, M_i, i = 1, 2, \dots, n$.

Statystyka r mierzy zgodność z rozkładem normalnym próby X_i , dla $i = 1, 2, \dots, n$. Zgodność z rozkładem normalnym próby $X_i, i = 1, 2, \dots, n$ znajduje odzwierciedlenie w (prawie) jednostkowej wartości współczynnika r . Szczegóły dotyczące współczynnika korelacji opisane zostały przez J. Filliben'a w pracy pt. *The Probability Plot Correlation Coefficient Test for Normality*, [3].



Rysunek 5.4: Wykres `worm plot` reszt z modelu $BCCG(8.87, 8.7, 7)$.

nymi wartościami rozkładu normalnego w każdym z przedziałów wiekowych. Aby poprawnie analizować `worm plot` kolejne wykresy należy czytać wierszami, to znaczy od lewego dolnego rogu do prawego górnego. Każdy wykres odpowiada resztom pochodzących z obserwacji dla danego przedziału wiekowego. Wykresy pozwalają na wykrywanie nieprawidłowości w dopasowanym modelu wewnątrz wyznaczonych przedziałów zmiennej `age`. Prawidłowa postać wykresów typu `worm plot` jest wówczas, gdy obserwacje znajdują się wewnątrz przedziału ufności (95%) oznaczonego na rysunku przerywanymi liniami. Z wykresów `worm plot` dla modelu `m3` można odczytać poprawne dopasowanie dla większości danych z 12 przedziałów wiekowych, ze sporadycznymi niedoskonałościami.

Powyzsze wywołanie funkcji `wp` zwraca także wyniki sprawdzenia, czy wszystkie punkty odpowiadające 6672 obserwacjom zostały wyświetlane na wykresie `worm plot`. Jak widać 4 obserwacje nie zostały uwzględnione na wykresach z powodu znaczących odchyleń wykraczających poza zakres wykresu. Taki wynik w stosunku do całej analizowanej próby daje niewielki odsetek i jest akceptowalny.

Van Buuren i Fredriks w pracy z 2001 roku [1] do każdego z poszczególnych wykresów `worm plot` zaproponowali dodanie kubicznej krzywej (czerwona krzywa). Każda krzywa kubiczna zostaje utworzona przez wyznaczenie współczynników dla stałego, liniowego, kwadratowego i sześciennego składnika: $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ i $\hat{\beta}_3$. Współczynniki te odpowiednio oznaczają różnice między empirycznymi i pochodzącyymi z modelu następującymi momentami dla reszt: śred-

nią, wariancją, skośnością i kurtozą. Van Buuren i Fredriks podsumowali swoją interpretację charakteru danej krzywej kubicznej w tabeli II [1], przytoczonej przeze mnie jako rysunek 5.5. Dla diagnostyki modelu za niedopasowane wartości dla $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ i $\hat{\beta}_3$ Van Buuren i Fredriks uznali wartości przekraczające następujące progowe wielkości: 0.10, 0.10, 0.05 i 0.03 dla odpowiedniego $\hat{\beta}_i, i = 1, 2, 3, 4$.

Table II. Interpretation of various patterns in the worm plot.

Shape	Moment	If the	Then the
Intercept	Mean	worm passes above the origin, worm passes below the origin,	fitted mean is too small. fitted mean is too large.
Slope	Variance	worm has a positive slope, worm has a negative slope,	fitted variance is too small. fitted variance is too large.
Parabola	Skewness	worm has a U-shape, worm has an inverted U-shape,	fitted distribution is too skew to the left. fitted distribution is too skew to the right.
S-curve	Kurtosis	worm has an S-shape on the left bent down, worm has an S-shape on the left bent up,	tails of the fitted distribution are too light. tails of the fitted distribution are too heavy.

Rysunek 5.5: Interpretacja wykresów typu `worm plot`. Źródło: [1].

Wynikowa zmienna `wp.diagnostyka` zawiera w argumencie `$classes` przedziały wiekowe na jakie została podzielona zmienna `age`, natomiast `$coeff` podaje współczynniki $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ i $\hat{\beta}_3$.

```
> wp.diagnostyka
$classes
      [,1]      [,2]
[1,]  6.520192  7.880903
[2,]  7.880903  8.833676
[3,]  8.833676  9.832991
[4,]  9.832991 10.892539
[5,] 10.892539 12.023272
[6,] 12.023272 13.129363
[7,] 13.129363 14.128679
[8,] 14.128679 15.034908
[9,] 15.034908 15.982204
[10,] 15.982204 16.910335
[11,] 16.910335 17.712526
[12,] 17.712526 18.498289

$coeff
      [,1]      [,2]      [,3]      [,4]
[1,]  0.0065537605  0.011685876 -2.159400e-03  0.0009473888
[2,] -0.0007378092 -0.016640374  3.489517e-04 -0.0050508212
[3,]  0.0159047895  0.044782280  1.419344e-06 -0.0084140617
[4,] -0.0182686541  0.004073453 -1.317826e-03 -0.0028579223
[5,]  0.0334584425  0.008808407 -2.072855e-02 -0.0083243853
[6,] -0.0878996186 -0.007076363  4.379030e-02  0.0031402124
[7,]  0.0097652825  0.070706982 -1.179238e-02 -0.0157874850
[8,]  0.0764456476 -0.030294842 -2.227285e-02  0.0117147982
[9,] -0.0216233569  0.004241510 -8.162029e-03 -0.0108142856
[10,] 0.0223375436 -0.061897555  6.827984e-04  0.0240830073
[11,] -0.0076385095  0.024858745  7.804398e-03 -0.0084797025
[12,] -0.0160408224 -0.008540688  6.684386e-03  0.0020204253
```

Stosując kryterium diagnozy współczynników do tabeli `wp.diagnostyka$coef` otrzymuję, iż nie istnieją niedopasowania dla $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ i $\hat{\beta}_4$. Brak naruszeń może świadczyć o tym, że rozważany model jest adekwatny dla danych w odpowiednich przedziałach wiekowych.

Testowanie normalności reszt z uwzględnieniem grup wiekowych może być badane także przez obliczenie statystyki Q , [19]. Niech G będzie liczbą grup wiekowych i niech r_{gi} , $i = 1, 2, \dots, n_i$ będą resztami w grupie wiekowej g , ze średnią \bar{r}_g i odchyleniem standardowym s_g dla $g = 1, 2, \dots, G$. Statystyki $Z_{g1}, Z_{g2}, Z_{g3}, Z_{g4}$ wyznacza się dla reszt dla odpowiedniej grupy g i wykorzystuje do testu sprawdzającego, czy reszty w grupie g mają rozkład normalny $N(0, 1)$. Odpowiednie statystyki Z_{g1}, Z_{g2} to: $Z_{g1} = n_g^{1/2} \bar{r}_g$, $Z_{g2} = \{s_g^{2/3} - [1 - 2/(9n_g - 9)]\}/\{2/(9n_g - 9)\}^{1/2}$, natomiast Z_{g3} i Z_{g4} są statystykami testowymi dla skośności i kurtozy wyznaczonymi przez D'Agostino (1990) w pracy [22].

Statystykę Q , podaną przez Royston'a i Wright'a (2000) oblicza się ze wzoru $Q_j = \sum_{q=1}^G Z_{gj}^2$, $j = 1, 2, 3, 4$. W swojej pracy Royston i Wright przedyskutowali przybliżony rozkład statystyki Q na podstawie hipotezy zerowej mówiącej, że reszty mają rozkład normalny. Otrzymany poziom istotności powinien być uznany za wskaźnik nieadekwatności modelu, a nie formalny test. Istotność statystyk Q_1, Q_2, Q_3 lub Q_4 wskazuje możliwe wystąpienie nieprawidłowości odpowiednio dla parametrów modelu μ, σ, ν i τ , które mogą być pokonane poprzez zwiększenie liczby stopni swobody w modelu dla danego parametru. Wartość statystyki Q w pakiecie `gamlss` jest otrzymywana za pomocą funkcji `Q.stats()`.

	Z1	Z2	Z3	Z4	AgostinoK2	N
6.52019 to 7.96577	0.288	0.288	-0.351	0.310	0.220	607
7.96577 to 9.00342	-0.761	-0.912	0.692	-0.860	1.218	608
9.00342 to 10.1286	0.803	1.645	-0.397	0.087	0.165	605
10.1286 to 11.2676	-0.293	-1.679	-0.656	-2.283	5.641	606
11.2676 to 12.5325	-1.003	0.176	0.705	0.144	0.518	607
12.5325 to 13.7097	0.898	1.408	0.534	-0.883	1.064	606
13.7097 to 14.7364	-0.090	-0.688	-1.889	0.456	3.778	607
14.7364 to 15.7440	0.796	-0.265	0.449	-0.472	0.424	606
15.7440 to 16.7570	-0.301	0.181	0.248	2.009	4.097	607
16.7570 to 17.6276	0.307	-0.250	-0.351	-0.302	0.215	607
17.6276 to 18.4982	-0.526	0.011	0.540	0.344	0.410	606
TOTAL Q stats	4.310	9.092	6.221	11.530	17.751	6672
df for Q stats	2.129	6.149	4.001	11.000	15.001	0
p-val for Q stats	0.129	0.179	0.183	0.400	0.276	0

Kwadrat statystyki Z_{gj} stanowi wkład grupy wiekowej g do statystyki Q_j , a tym samym pomaga określić, która z grup wiekowych przyczynia się do istotności statystyki Q_j , a więc w których grupach wiekowych model może być nieadekwatny.

Zakładając, że liczba grup G jest wystarczająco duża w stosunku do stopni swobody dla danego parametru modelu, wartości Z_{gj} powinny mieć w przybliżeniu standardowy rozkład normalny pod warunkiem hipotezy zerowej, że reszty mają rozkład $N(0, 1)$. Zgodnie z pracą [17] do interpretacji wyników zwracanych przez funkcję `Q.stats` wykorzystam stwierdzenie, że wartość $|Z_{gj}|$ większa niż 2 wskazuje na występowanie niedoskonałości w modelu.

Wartości zawarte w tablicy zwracanej przez funkcję `Q.stats` sugerują, że w wybranym modelu mogą występować niewielkie nieprawidłowości (szczególnie dla ν i τ). Reszty w przedziałach wiekowych (10.13, 11.27) i (15.74, 16.76) są odpowiednio platykuryczne i leptokurytyczne. Statystyka $|Z_{gj}|$ przekracza tam nieznacznie wartość 2, co wskazuje możliwe niedoskonałości modelowanego τ w tym zakresie lub występowanie obserwacji odstających zmiennej `height`. P-wartości dla odpowiednich Q-statystyk są odpowiednio równe 0.129, 0.179, 0.183

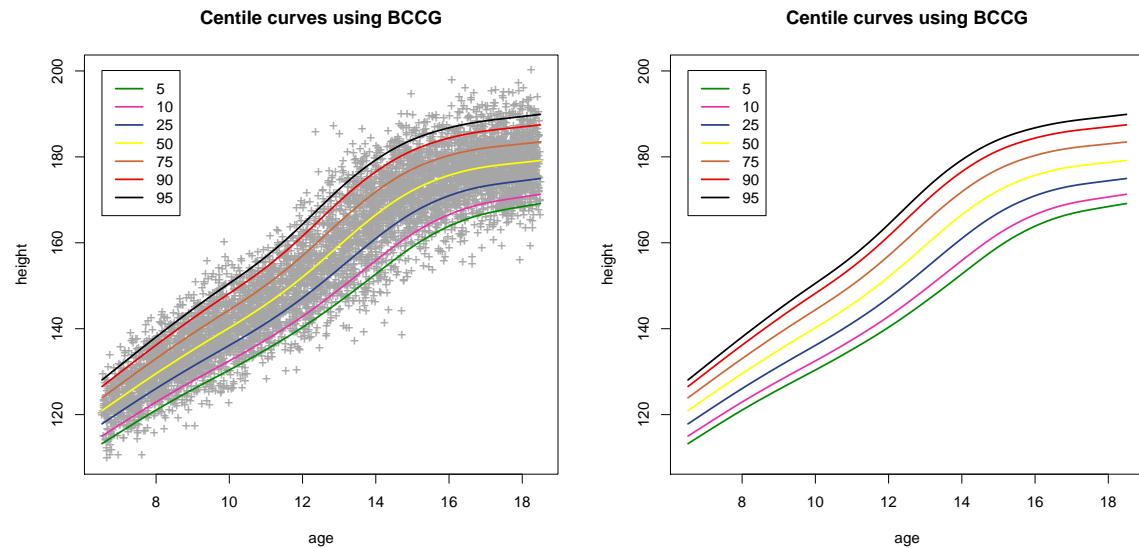
i 0.4. Pozwala to zatem przypuszczać, że nie ma podstaw do odrzucenia hipotezy zerowej o normalności rozkładu reszt.

Korzystając z funkcji `centiles` i wybranego modelu $BCCG(8.87, 8.7, 7)$ rysuję krzywe centylowe wysokości. Funkcja `centiles` oprócz wykresu 5.6 zwraca również, jaki procent obserwacji wyznaczonych na podstawie modelu znajduje się poniżej/powyżej oczekiwanej centyla podanego w argumencie `cent`. Rozkład ten został podany w tabeli 5.2. Jak widać poniższe wyniki potwierdzają dobre dopasowanie modelu.

```
> (centiles(m3, xvar=age, points=TRUE, cent=c(5,10,25,50,75,90,95),
+ lwd.cent=c(2,2,2,2,2,2,2),
+ ylab = "height", xlab = "age"))
```

Tabela 5.2: Porównanie % obserwacji oczekiwanych i otrzymanych z modelu `m3`.

% obserwacji oczekiwanych	5	10	25	50	75	90	95
% obserwacji otrzymanych	4.95	9.94	24.49	50.24	75.15	89.78	94.89



Rysunek 5.6: Krzywe centylowe wysokości dla dzieci i młodzieży płci męskiej utworzone na podstawie modelu $BCCG(8.87, 8.7, 7)$.

Wykres 5.6 przedstawia siedem krzywych centylowych z modelu $BCCG(8.87, 8.7, 7)$, które zostały uzyskane ze wzoru (5.2) z uwzględnionymi następującymi centylami α : 5, 10, 25, 50, 75, 90, 95. Dla każdego α krzywą centylową y_α ze względu na x otrzymuje się poprzez znalezienie dopasowanych wartości $(\hat{\mu}, \hat{\sigma}, \hat{\nu})$ dla każdego x , i podstawienie wartości do wzoru:

$$y_\alpha = \begin{cases} \mu[1 + \sigma\nu z_\alpha]^{1/\nu} & \text{jeżeli } \nu \neq 0 \\ \mu \exp[\sigma z_\alpha] & \text{jeżeli } \nu = 0 \end{cases} \quad (5.2)$$

gdzie z_α jest α -tym kwantylem ze standardowego rozkładu normalnego.

Korzystając z funkcji `centiles.pred` wyznaczam następujące wartości centylów wysokości: 5, 10, 25, 50, 75, 90, 95 dla dzieci i młodzieży płci męskiej w wieku 7-18:

```
> centyle <- centiles.pred(m3, xname="age", xvalues=seq(7,18,1),
+ data=danemed, cent=c(5,10,25,50,75,90,95))
```

Tabela 5.3: Centyle wysokości dla chłopców (bez nadwagi).

wiek	Centyl						
	5	10	25	50	75	90	95
7	115.72	117.54	120.52	123.73	126.86	129.61	131.22
8	121.20	123.05	126.15	129.61	133.07	136.20	138.08
9	125.85	127.86	131.27	135.11	139.01	142.57	144.73
10	130.28	132.44	136.11	140.25	144.45	148.29	150.61
11	135.39	137.58	141.28	145.47	149.74	153.66	156.03
12	140.32	142.74	146.88	151.64	156.56	161.14	163.96
13	145.97	148.84	153.70	159.20	164.81	169.94	173.05
14	152.34	155.69	161.07	166.76	172.18	176.85	179.57
15	159.31	162.43	167.38	172.54	177.41	181.56	183.95
16	164.22	166.88	171.19	175.83	180.31	184.21	186.50
17	167.06	169.42	173.35	177.71	182.06	185.97	188.31
18	168.24	170.44	174.20	178.49	182.90	186.99	189.48

Wykonując porównanie otrzymanych wyników dla chłopców bez nadwagi, z wynikami opublikowanymi w [11], które zostały wyznaczone dla danych `danemed` z wyłączeniem dzieci z nadwagą, zauważam, iż wartości centyl wzrostu dla całej populacji są nieznacznie wyższe. Obserwacja ta może stanowić ciekawy wniosek, że w Polsce dzieci z nadwagą i otyłością charakteryzują się wyższą wysokością ciała.

5.2. Model ciśnienia skurczowego dla chłopców bez nadwagi

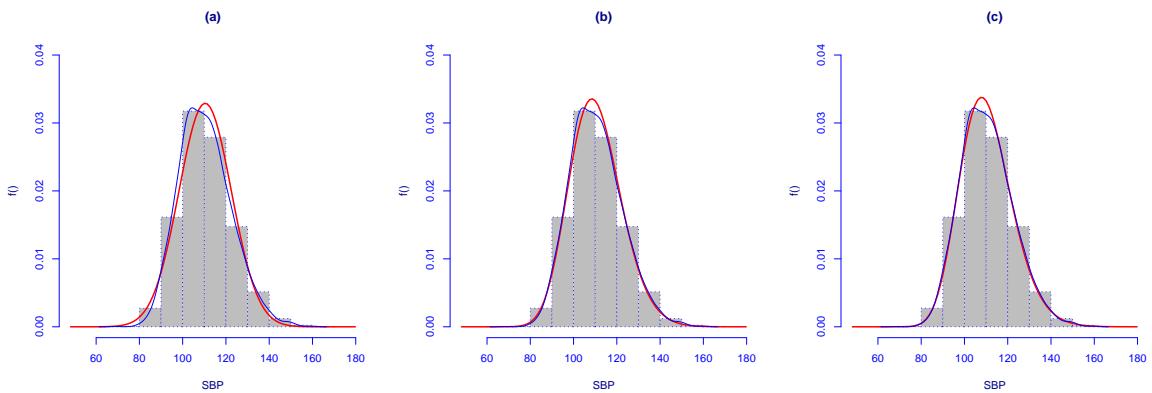
W tym podrozdziale ciśnienie skurczowe SBP dla dzieci i młodzieży płci męskiej zostało analizowane z uwzględnieniem dwóch zmiennych objaśniających: `age` i `height`. Model, który został wybrany w wyniku poniższych analiz, posłuży w następnym rozdziale do wyznaczenia norm ciśnienia dla dzieci i młodzieży płci męskiej. Do analiz zostały wybrane dane ze zbioru `danemed` dla chłopców bez nadwagi w wieku 6.5–18.5, selekcjonowane na podstawie zmiennej `bp.ref=1`, będącej wskaźnikiem dla obserwacji, które należy uwzględnić do obliczeń centylu ciśnienia. Zbiór ten liczy 6627 obserwacji.

Wykresy przedstawione na rysunku 5.7 zostały utworzone z wykorzystaniem poniższych wywołań funkcji `histDist`. Przedstawiają histogramy, dopasowane rozkłady do wartości zmiennej SBP z próby (niebieska krzywa) oraz dopasowany parametryczny rozkład wskazany przez argument `family`.

```
> mNO <- histDist(SBP, family="NO", density = TRUE, main = "(a)",  
+ ylim=c(0,0.04))  
> mLNO <- histDist(SBP, family="LNO", density = TRUE, main = "(b)",  
+ ylim=c(0,0.04))  
> mBCCG <- histDist(SBP, family="BCCG", density = TRUE, main = "(c)",  
+ ylim=c(0,0.04))
```

Porównując rozkład *NO* z rozkładem *LNO* można sprawdzić, czy dane są dodatnio skośne. Natomiast rozkład *BCCG* umożliwia modelowanie także rozkładów ujemnie skośnych⁴. Na podstawie wykresów można zauważyc, że do modelowania rozkładu zmiennej SBP warto użyć 3-parametrowego rozkładu log-normalnego $LNO(\mu, \sigma, \nu)$, dla którego kolejne parametry

⁴Wszystkie wymienione tutaj rozkłady zostały opisane w dodatku B.4, natomiast parametry i domyślne funkcje związane dla odpowiednich rozkładów znajdują się w tabeli 4.2.



Rysunek 5.7: Dopasowany rozkład dla zmiennej SBP (niebieska krzywa) oraz parametryczny rozkład dla odpowiedniej rodziny rozkładów: (a) *NO*, (b) *LNO*, (c) *BCCG*, (czerwona krzywa).

oznaczają medianę, wariancję i skośność. Aby potwierdzić to przypuszczenie zostały utworzone trzy modele, w których parametr położenia μ został opisany przez addytywną formułę zmiennych `age` i `height` modelowanych z wykorzystaniem kubicznych funkcji sklejonymi z rozkładami: normalnym, log-normalnym oraz rozkładem Box-Cox-Cole-Green'a. Pozostałe parametry rozkładu dla wszystkich modeli są stałe i nie zależą od zmiennych objaśniających.

```
> con <- gamlss.control ( trace = FALSE )
> mod.NO<-gamlss(SBP~cs(age)+cs(height),
+ sigma.fo=~1, data=danemed, family=NO, control =con)
> mod.LNO<-gamlss(SBP~cs(age)+cs(height),
+ sigma.fo=~1, nu.fo=~1,data=danemed, family=LNO, control =con)
> mod.BCCG<-gamlss(SBP~cs(age)+cs(height),
+ sigma.fo=~1, nu.fo=~1,data=danemed, family=BCCG, control =con)
```

Skorzystałem z wartości oryginalnego kryterium AIC z `penalty=2`, kryterium BIC z karą $\log(6627) = 8.798907 \approx 8.8$ oraz w celu sprawdzenia wrażliwości modeli także `penalty=4` i `penalty=8`⁵. Opierając się na wynikach kryterium GAIC zawartych w tabeli 5.1 do modelowania zmiennej $y = \text{SBP}$ został wybrany rozkład log-normalny.

Tabela 5.4: Wyniki kryterium GAIC.

model	AIC (penalty=2)	penalty=4	penalty=8	BIC (penalty=8.8)
mod.NO	49024.74	49044.74	49084.74	49092.73
mod.LNO	48831.74	48851.74	48891.74	48899.73
mod.BCCG	48829.82	48851.82	48895.82	48904.61

Po wybraniu rodziny rozkładów kolejnym zadaniem było jak najlepiej dopasować addytywne składniki modelu GAMLSS, które opiszą parametry zmiennej y . Dla zmiennych objaśniających `age` i `height` oznaczonych odpowiednio przez x_1 i x_2 oraz zmiennej objaśnianej $y = \text{SBP}$ z rozkładem *LNO* zagadnienie jest następujące:

$$\begin{cases} \log(\mu) = h_1(x_1) + h_2(x_2) \\ \log(\sigma) = h_3(x_1) + h_4(x_2) \\ \nu = c \end{cases} \quad (5.3)$$

⁵Wybór wartości $k = 8$ do wyboru najlepszego modelu został zasugerowany w artykule [15].

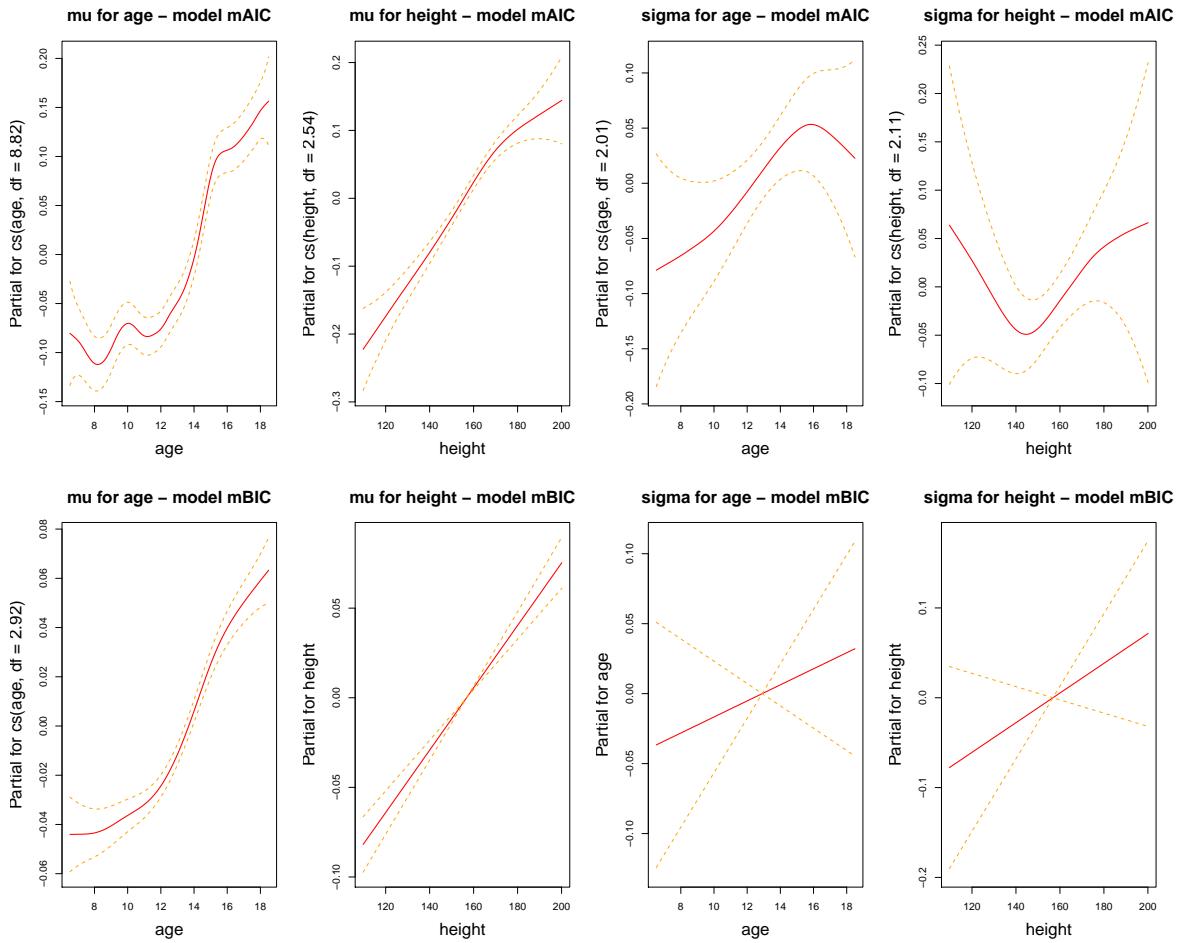
gdzie c – stała. Parametry μ i σ rozkładu LNO z logarytmiczną funkcją wiążącą, wstępnie chcę modelować nieparametrycznie jako funkcje zmiennych objaśniających, natomiast współczynnik skośności, czyli ν dla rozkładu $LNO(\mu, \sigma, \nu)$ jest ustalony. Do opisania formuły dla μ i σ korzystam z funkcji `cs()` postępując zgodnie z metodyką wygładzania kubicznymi funkcjami sklejonymi. Złożoność każdej krzywej kubicznej jest zdefiniowana w kategoriach efektywnych stopni swobody, gdzie w konwencji R stopień swobody równy 0 odpowiada liniowemu składnikowi, podczas gdy większa liczba stopni swobody wskazuje na coraz bardziej złożoną krzywą. Efektywne stopnie swobody wyznaczam z użyciem procedury `optim()` oraz funkcji `GAIC()` stosując kryterium AIC ($k = 2$) jak i BIC ($k \approx 8.8$).

```
> fn1 <- function(p) AIC(gamlss(SBP ~ cs(age, df = p[1]) +
+ cs(height, df=p[2]),
+ sigma.fo = ~cs(age, df = p[3]) + cs(height, df = p[4]),
+ nu.fix=TRUE, nu.start=p[5],
+ data = danemed,
+ family=LNO, control = gamlss.control(trace=FALSE, c.crit = 0.1)),
+ k=2)
> opAIC <- optim(par = c(5,5,5,5,0), fn1, method = "L-BFGS-B",
+ lower = c(0.1,0.1,0.1,0.1,-2),
+ upper = c(13,13,13,13,2))
> opAIC$par
[1] 8.8174258 2.5393365 2.0143224 2.1175887 0.2173152
> fn2 <- function(p) AIC(gamlss(SBP ~ cs(age, df = p[1]) +
+ cs(height, df=p[2]),
+ sigma.fo = ~cs(age, df = p[3]) + cs(height, df = p[4]),
+ nu.fix=TRUE, nu.start=p[5],
+ data = danemed,
+ family=LNO, control = gamlss.control(trace=FALSE, c.crit = 0.1)),
+ k=log(length(age)))
> opBIC <- optim(par = c(3,3,3,3,3,0), fn2, method = "L-BFGS-B",
+ lower = c(0.1,0.1,0.1,0.1,-2),
+ upper = c(7,7,7,7,2))
> opBIC$par
[1] 2.9007598 0.1000000 0.1756228 0.1000000 0.2082036
```

Według kryterium AIC dla najlepszego modelu dla μ liczba stopni swobody wygładzeń stanowi sumę $8.82 \approx 9$ oraz $2.54 \approx 3$. Należy podkreślić, że 12 stopni swobody tego dopasowania odnosi się do dodatkowych stopni swobody, tzn. pomniejszonych o liniowe i stałe składniki. Po dopasowaniu stałej i liniowej części modelu całkowita liczba stopni swobody dla μ jest równa $12 + 2 + 2 = 16$.

Uwzględniając kryterium BIC, dla parametru rozkładu μ najlepszy model dla składnika wygładzenia posiada $2.9 \approx 3$ stopnie swobody (całkowita liczba stopni swobody to $3 + 2 + 2 = 7$). Wartości `df` bliskie 0 wskazują, że wpływ odpowiedniej zmiennej na zmienną y prawdopodobnie jest liniowy. Zatem w powyższym przypadku można zastąpić składnik `cs(height)` składnikiem `height`. Po uwzględnieniu liniowego wpływu zmiennej `height` na μ i σ zmiennej objaśnianej, wywołanie funkcji `optim` zostało powtórzone. Kierując się kryteriami BIC i AIC zostały otrzymane następujące modele `mBIC`, `mAIC`:

```
> mBIC <- gamlss(SBP ~ cs(age, df = 2.89)+height,
+ sigma.fo = ~cs(age, df=0.72)+height, nu.fix=TRUE, nu.start=0.2077279,
+ family=LNO, data = danemed)
> mAIC <- gamlss(SBP ~ cs(age, df = 8.82)+cs(height, df = 2.54),
+ sigma.fo = ~cs(age, df=2.01) + cs(height, df = 2.11),
+ nu.fix=TRUE, nu.start=0.2173152, family=LNO, data = danemed)
```



Rysunek 5.8: Dopasowane wartości dla addytywnych składników dla modeli mAIC i mBIC.

Na podstawie wykresów 5.8, które zostały utworzone za pomocą funkcji `term.plot`⁶ można ocenić, w jaki sposób poszczególne składniki regresji dla odpowiedniej zmiennej wpływają na zmienną objaśnianą. Model mAIC dla μ wydaje się być nadmiernie dopasowany do danych – jest to typowe zachowanie AIC. Model mBIC otrzymany w wyniku zastosowania kryterium BIC dla parametru μ jest całkiem dobrze dopasowany. W obu przypadkach modeli dla σ przedziały ufności są bardzo szerokie. Ta obserwacja sugeruje, że korzystniej byłoby przybliżać wariancję za pomocą stałego lub liniowego składnika, zamiast za pomocą addytywnego składnika kubicznych funkcji sklejanych zmiennych `age` i `height`.

W następnej kolejności konstruuję modele:⁷

```
> m1<-gamlss(SBP~cs(age, df = 3) + height,
+ sigma.fo = ~1,
+ nu.fix=TRUE, nu.start=-0.1787564,
+ data = danemed, family=LNO)
> m2<-gamlss(SBP~cs(age, df = 3) + height,
+ sigma.fo = ~age,
+ nu.fix=TRUE, nu.start=0.1766871,
```

⁶Funkcja `term.plot` w pakiecie `gamlss` wykonuje to samo, co funkcja `plot` z pakietu `mgcv`, czyli rysuje addytywne wygładzone składniki dla dowolnego parametru rozkładu zmiennej objaśnianej.

⁷W tych modelach stały składnik dla `nu` został wyznaczony przy pomocy procedury `optim()`.

```

+ data = danemed, family=LNO)
> m3<-gamlss(SBP~cs(age, df = 3) + height,
+ sigma.fo = ~height,
+ nu.fix=TRUE, nu.start=0.1648639,
+ data = danemed, family=LNO)
> m5<-gamlss(SBP~cs(age, df = 3) + height,
+ sigma.fo = ~cs(age,df=3),
+ nu.fix=TRUE, nu.start=0.2248048 ,
+ data = danemed, family=LNO)
> m6<-gamlss(SBP~cs(age, df = 3) + height,
+ sigma.fo = ~cs(height,df=3),
+ nu.fix=TRUE, nu.start= 0.2121624,
+ data = danemed, family=LNO)
> AIC(m1,m2,m3,m5,m6)
      df      AIC
m6 10.999978 48794.71
m5 11.001508 48794.74
m3  8.000753 48803.10
m2  8.000753 48804.08
m1  7.000754 48832.22
> AIC(m1,m2,m3,m5,m6,k=8)
      df      AIC
m3  8.000753 48851.11
m2  8.000753 48852.09
m6 10.999978 48860.71
m5 11.001508 48860.75
m1  7.000754 48874.23
> AIC(m1,m2,m3,m5,m6,k=log(length(age)))
      df      AIC
m3  8.000753 48857.50
m2  8.000753 48858.48
m6 10.999978 48869.50
m5 11.001508 48869.54
m1  7.000754 48879.82

```

Kierując się kryterium BIC oraz $GAIC(8)$ ostatecznie wybieram model m3:

```

> SBP.boys<-gamlss(SBP~cs(age, df = 3) + height,
+ sigma.fo = ~height,
+ nu.fix=TRUE, nu.start=0.1648639,
+ data = danemed, family=LNO)

```

Z wykorzystaniem funkcji `prof.dev` wywołanej dla modelu `SBP.boys` z następującymi parametrami:

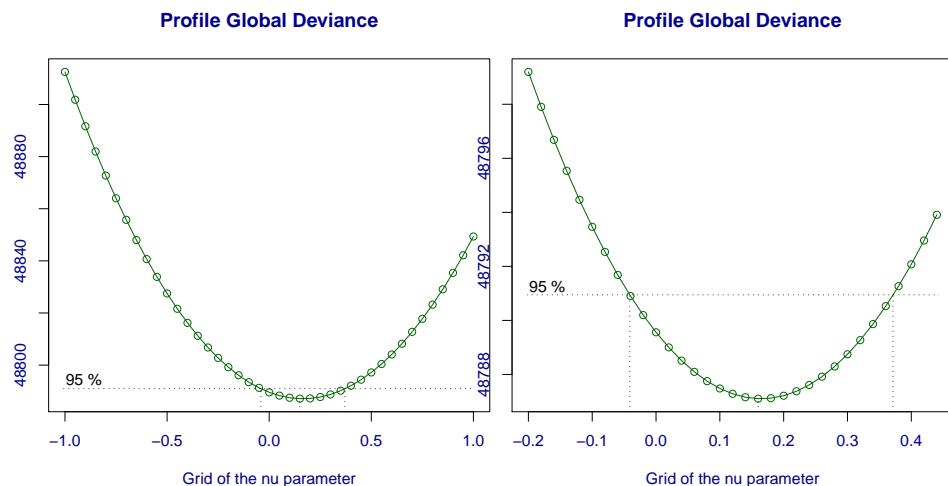
```

> prof.dev(SBP.boys,"nu",min=-1,max=1,step=0.05)
...
*****
Best estimate of the fixed parameter is  0.15
with a Global Deviance equal to  48787.12 at position  24
A 95 % Confidence interval is: ( -0.04064889 ,  0.3705185 )
*****
> prof.dev(SBP.boys,"nu",min=-0.2,max=0.45,step=0.02)
...
*****
Best estimate of the fixed parameter is  0.16
with a Global Deviance equal to  48787.1 at position  19

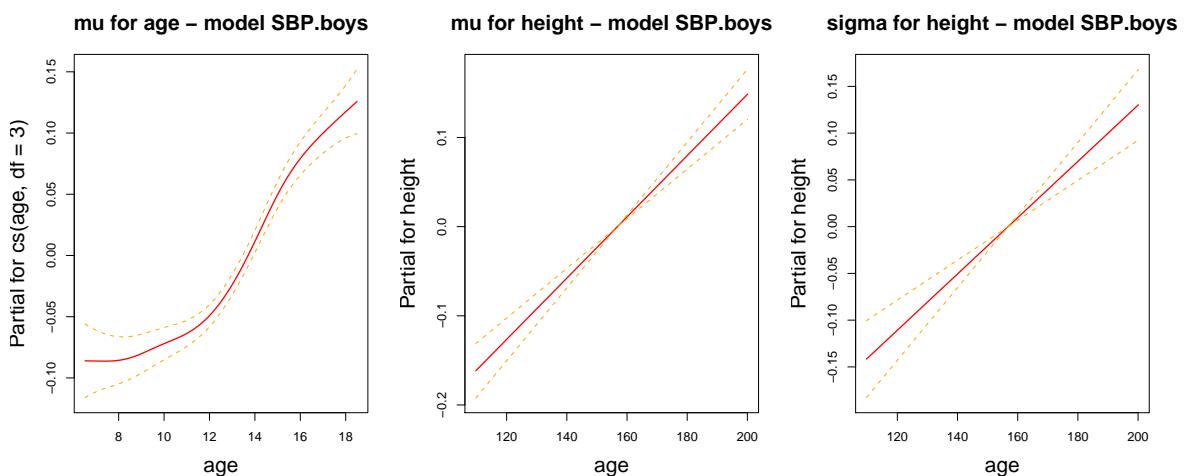
```

```
A 95 % Confidence interval is: ( -0.04105424 , 0.3712613 )
*****
```

został utworzony wykres profilowanej globalnej dewiancji (rys. 5.9). Funkcja ta wyznaczyła także 95% przedział ufności dla stałego parametru ν , który wyniósł $(-0.041, 0.37)$.



Rysunek 5.9: Wykres profilowanej dewiancji dla ν dla dopasowanego modelu SBP.boys.



Rysunek 5.10: Dopasowane wartości dla addytywnych składników modelu SBP.boys.

```
> summary(SBP.boys)
*****
Family:  c("LNO", "Box-Cox")
Call:
gamlss(formula = SBP ~ cs(age, df=3) + height, sigma.formula=~height,
       family = LNO, data = dane.ch, nu.start = 0.1648639, nu.fix = TRUE)
```

```

Fitting method: RS()

-----
Mu link function: identity
Mu Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.284043  0.0296527 211.92 0.000e+00
cs(age, df = 3) 0.021236  0.0018709 11.35 1.376e-29
height       0.003435  0.0003253 10.56 7.414e-26

-----
Sigma link function: log
Sigma Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.133576  0.0690567 -30.896 6.275e-196
height       0.003010  0.0004369   6.889 6.147e-12

-----
Nu parameter is fixed
Nu = 0.1648639

No. of observations in the fit: 6627
Degrees of Freedom for the fit: 8.000753
      Residual Deg. of Freedom: 6618.999
                           at cycle: 4

Global Deviance: 48787.1
      AIC: 48803.1
      BIC: 48857.5
*****

```

Podsumowując, ostatecznie wybrany model jest postaci:

$$\begin{cases} \log(\mu) = 6.284043 + 0.021236 * cs(age, df = 4) + 0.003435 * height \\ \log(\sigma) = -2.133576 + 0.003010 * height \\ \nu = 0.1648639. \end{cases} \quad (5.4)$$

Jest to model semi-parametryczny zmiennej `age`, dla której zostały użyte nieparametryczne metody wygładzania kubicznymi funkcjami sklejonymi z 3 efektywnymi stopniami swobody. Liczba efektywnych stopni swobody, które zostały wykorzystane w powyższym modelu do modelowania μ z użyciem zmiennej `age`, jest równa 5 (jeden dla stałego, jeden dla liniowego składnika, oraz 3 dla wygładzenia). Warto tutaj zwrócić uwagę, że notacja `gamlss()` różni się od tej z `gam()`, gdzie równoważna formuła modelu dla μ byłby zapisana ze składnikiem `s(age, 4)+height`.

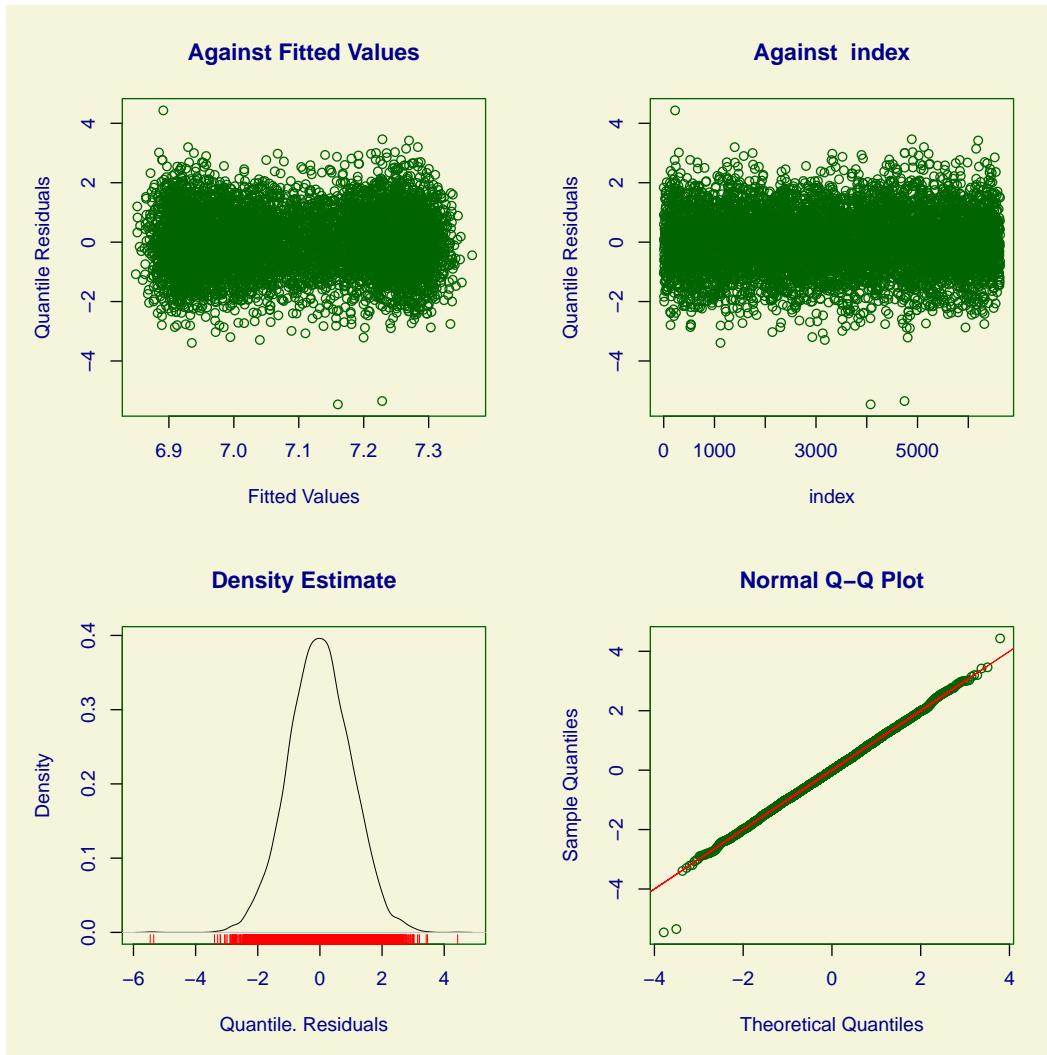
Dopasowanie modelu sprawdzam przy pomocy wykresów dla reszt:

```

> plot(SBP.boys)
*****
Summary of the Quantile Residuals
      mean     = -1.676737e-05
      variance = 1.000151
      coef. of skewness = -0.006101541
      coef. of kurtosis = 3.234984
Filliben correlation coefficient = 0.9993987
*****

```

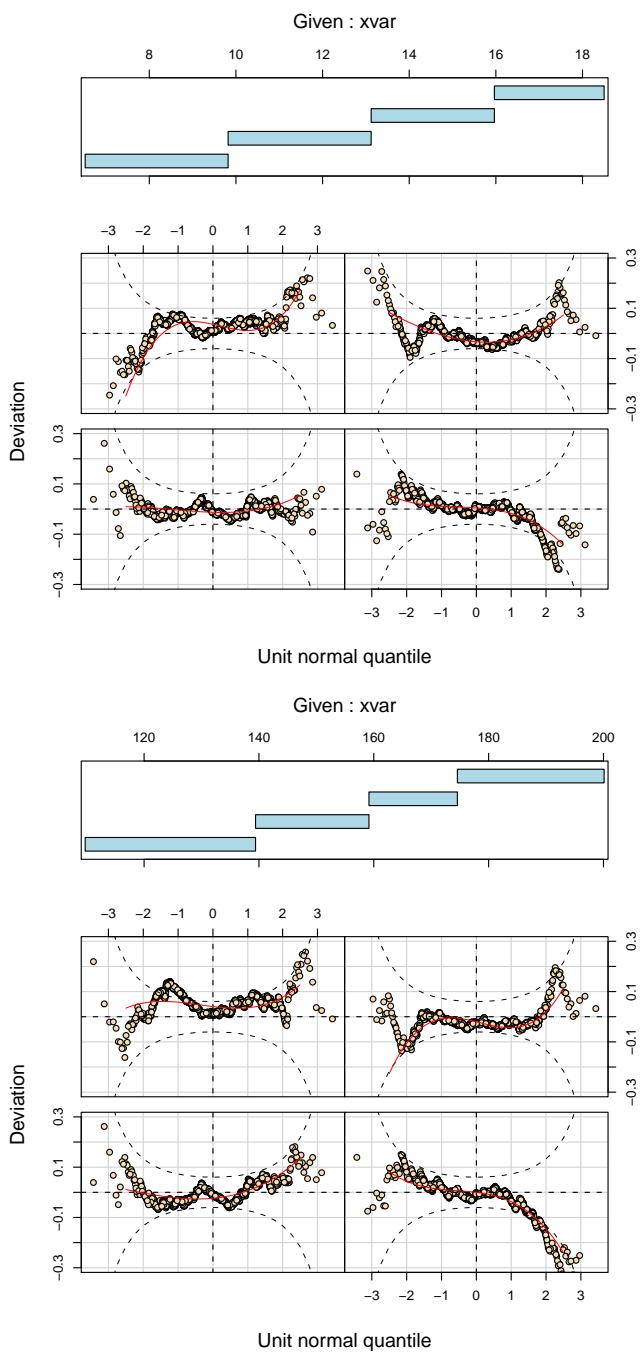
Wykres 5.11 przedstawia diagnostykę reszt. Z wykresów można odczytać, iż występują 3 znacznie odstające obserwacje. Poza tym nie widać dużych odstępstw od normalności.



Rysunek 5.11: Wykresy diagnostyczne dla reszt z modelu `SBP.boys`, przedstawiające reszty (a) ze względu na dopasowane wartości μ , (b) ze względu na indeks obserwacji, (c) estymator gęstości jądra, (d) wykres kwantylowy dla rozkładu normalnego.

Reszty z modelu sprawdzam także korzystając z wykresów `worm plot`. Rysunki 5.12 odpowiednio pokazują różnice pomiędzy obserwowanymi, a oczekiwanyymi wartościami rozkładu normalnego w każdej z 4 grup wiekowych (górny wykres) i w 4 grupach wzrostowych (dolny wykres).

Dodatkowo sprawdzam jaki procent obserwacji z próby mieści się pod następującymi wygładzonymi centylami: 50, 90, 95, 99 uwzględniając wszystkie grupy wiekowe razem oraz poszczególne kategorie wiekowe osobno. Dla całej próby wyniki zostały zawarte w tabeli 5.5. Procentowy rozkład przedstawiony w tejże tabeli wygląda zadowalająco, ponieważ różnica w dopasowaniu nie przekracza 0.4. Wykresy rysunku 5.13 przedstawiają wyniki odchyleń od oczekiwanych centylów (oznaczonych poziomą linią) dla poszczególnych kategorii wiekowych (oznaczonych kropkami). W tym podejściu można zauważać większe różnice (np. dla grupy 10 latków poniżej 50 centyla znajduje się tylko 45 procent obserwowanych osób). Analizując

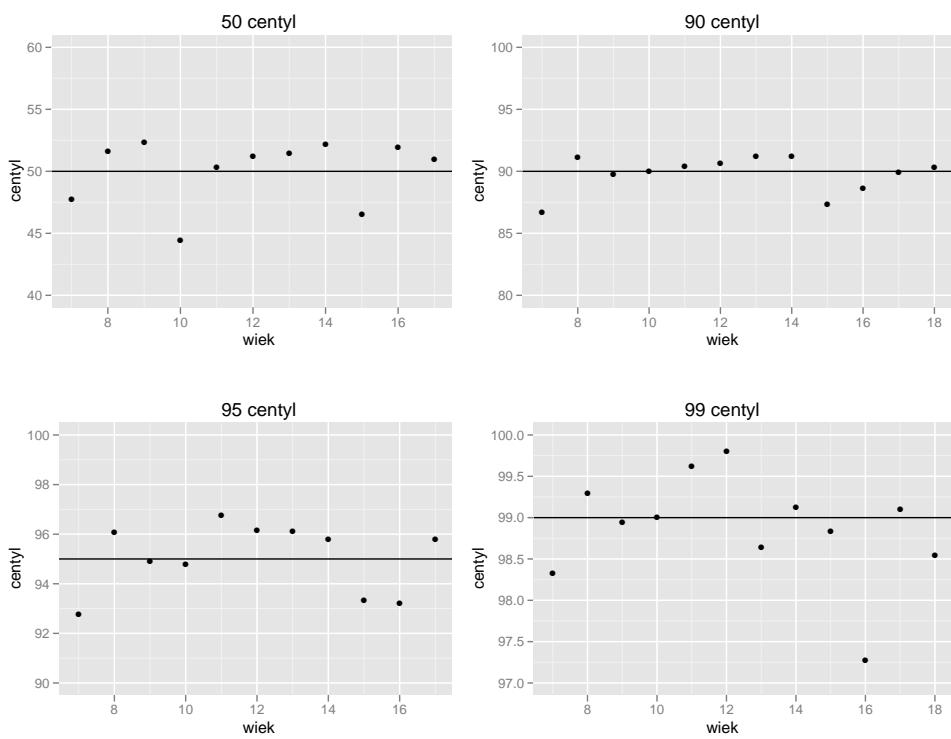


Rysunek 5.12: Wykresy `worm` plot reszt z modelu `SBP.boys`; kolejne wykresy odpowiadają (a) 4 grupom wiekowym, (b) 4 przedziałom wzrostowym.

wykresy należy pamiętać, że w poszczególnych podgrupach odchylenia są rzeczą naturalną.

Tabela 5.5: Porównanie % obserwacji oczekiwanych i otrzymanych z modelu `SBP.boys`.

% obserwacji oczekiwanych	50	90	95	99
% obserwacji otrzymanych	50.35	89.84	95.15	98.86



Rysunek 5.13: Dopasowanie modelu **SBP.boys** w poszczególnych kategoriach wiekowych.

Korzystając z wyników uzyskanych podczas poprzedniego testu (wyznaczyłem liczbę obserwacji, dla których procent obserwacji z próby mieści się pod wygładzonymi centylami 50, 90, 95, 99) dodatkowo sprawdzam dopasowanie modelu testem **p.adjust** [23] do złożonej hipotezy:

```
> p1<-prop.test(3337,6627,0.50)
> p1$p.val
[1] 0.5720287
> p2<-prop.test(5954,6627,0.90)
> p2$p.val
[1] 0.6882148
> p3<-prop.test(6306,6627,0.95)
> p3$p.val
[1] 0.5787741
> p4<-prop.test(6552,6627,p=0.99)
> p4$p.val
[1] 0.3095959
> p.adjust(p=c(p1$p.val,p2$p.val,p3$p.val,p4$p.val),
+ method = "bonferroni")
[1] 1 1 1 1
```

Powyższy test nie wykazuje istotnych statystycznych różnic.

Opierając się na powyższej diagnostyce i testach wykonanych dla modelu **SBP.boys** stwierdzam, że wybrany model dobrze opisuje ciśnienie skurczowe dla analizowanej grupy osób.

5.3. Podsumowanie

W tym podrozdziale korzystając z modelu `SBP.boys` oraz z wyznaczonych w podrozdziale 5.1 centyli wysokości wykonałam niezbędne obliczenia w celu wyznaczenia centyli ciśnienia skurczowego w zależności od dwóch zmiennych: wiek i wzrost. Do wykonania tych obliczeń wykorzystałam funkcje `predict.gamlss` oraz `qLNO`, [23]. Otrzymane wyniki przedstawiłam w tabeli 5.6. Kolumna `wzrost` w tych tabelach zawiera centyle wysokości dla danej grupy wiekowej (są to kolejno centyle: 5, 10, 25, 50, 75, 90, 95). Wyniki zawarte w tabeli 5.6 mogą stanowić zakres referencyjny wysokości ciśnienia w sprawowaniu opieki profilaktycznej, diagnostyce oraz leczeniu dzieci i młodzieży płci męskiej w wieku 7-18 lat.

Wartości referencyjne ciśnienia skurczowego i rozkurczowego dla dzieci i młodzieży płci żeńskiej i męskiej (z wykluczeniem osób z nadwagą i otyłością), które wyznaczyłam wykonując podobne analizy w oparciu o bazę danych `danemed` powiększoną o przypadki osób w wieku 6 i 19 lat można będzie niebawem znaleźć w artykule *Oscillometric blood pressure percentiles for Polish normal-weight school-aged children and adolescents*, [13].

Tabela 5.6: Centyle ciśnienia skurczowego według wieku i wysokości ciała dla dzieci i młodzieży płci męskiej w Polsce (bez nadwagi).

wiek	wzrost	50 centyl	90 centyl	95 centyl	99 centyl
7	116	99	109	112	118
	118	99	110	113	119
	121	100	110	113	120
	124	100	111	114	120
	127	101	112	115	121
	130	101	112	115	122
	131	101	112	116	122
8	121	100	110	114	120
	123	100	111	114	120
	126	101	111	115	121
	130	101	112	115	122
	133	102	113	116	123
	136	102	114	117	124
	138	102	114	117	124
9	126	101	112	115	121
	128	101	112	115	122
	131	102	113	116	123
	135	102	114	117	124
	139	103	114	118	125
	143	103	115	119	126
	145	104	116	119	126
10	130	102	113	116	123
	132	102	113	117	123
	136	103	114	118	124
	140	103	115	119	125
	144	104	116	120	127
	148	105	117	121	128
	151	105	117	121	128
11	135	103	115	118	125
	138	103	115	118	125
	141	104	116	119	126
	145	105	117	120	127
	150	105	118	121	128
	154	106	119	122	130
	156	107	119	123	130
12	140	105	116	120	127
	143	105	117	120	127
	147	106	118	121	129
	152	107	119	123	130
	157	107	120	124	131
	161	108	121	125	133
	164	109	122	126	133

wiek	wzrost	50 centyl	90 centyl	95 centyl	99 centyl
13	146	107	119	123	130
	149	107	120	123	131
	154	108	121	125	132
	159	109	122	126	133
	165	110	123	127	135
	170	111	125	129	137
	173	111	125	129	137
14	152	110	122	126	134
	156	110	123	127	135
	161	111	124	128	136
	167	112	126	130	138
	172	113	127	131	139
	177	114	128	132	141
	180	114	129	133	142
15	159	113	126	130	138
	162	113	127	131	139
	167	114	128	132	140
	173	115	129	134	142
	177	116	131	135	143
	182	117	132	136	145
	184	117	132	137	145
16	164	115	129	133	141
	167	116	130	134	142
	171	117	131	135	143
	176	117	132	136	145
	180	118	133	137	146
	184	119	134	138	147
	186	119	135	139	148
17	167	117	131	135	143
	169	117	131	136	144
	173	118	132	137	145
	178	119	134	138	147
	182	120	135	139	148
	186	120	136	140	149
	188	121	136	141	150
18	168	118	132	137	145
	170	118	133	137	146
	174	119	134	138	147
	178	120	135	139	148
	183	121	136	140	149
	187	122	137	142	151
	189	122	138	142	152

Dodatek A

Opis danych danemed

W niniejszej pracy do analiz została wykorzystana baza danych z udziałem ponad 17500 dzieci i młodzieży w wieku 6.5 - 18.5 lat. Dane te pochodzą z projektu OLAF „Opracowanie norm ciśnienia tętniczego dla populacji dzieci i młodzieży w Polsce-PL0080” prowadzonego przez zespół badaczy z Instytutu „Pomnik – Centrum Zdrowia Dziecka” w Warszawie, we współpracy z badaczami z całego kraju. Na potrzeby projektu OLAF zostały one zebrane w latach 2007-2009 na terenie całego kraju w 416 szkołach podstawowych, gimnazjach i szkołach ponadgimnazjalnych, zatem stanowią reprezentatywną próbę dla całej krajowej populacji [12].

Dane o nazwie **danemed**, z których korzystam w niniejszej pracy zostały odpowiednio przygotowana przez doktora Zbigniewa Kułągę poprzez usunięcie rekordów zawierających dane dzieci poniżej 6.5 i powyżej 18.5 roku życia, oraz rekordy tych dzieci, które zostały uznane za chore. Kryteria wykluczeń z analiz zostały szczegółowo opisane w [13]. Każda obserwacja danych **danemed** dotyczy jednego ucznia i zawiera informacje np. o wysokości ciśnienia skurczowego (SBP), rozkurczowego (DBP), wieku (**age**) lub wysokości ciała (**height**). Wszystkie zmienne bazy **danemed** wraz z opisem zostały przedstawione w tabeli A.1.

Tabela A.1: Zmienne bazy danych **danemed**

nazwa zmiennej	opis
ID	unikalny identyfikator danej osoby biorącej udział w badaniu
sex	1 – chłopcy, 2 – dziewczęta
age	wiek wyznaczony jako różnica daty pomiaru ciśnienia i daty urodzenia
wiek.kat	grupa wiekowa do której zalicza się dana osoba
bp.ref	1 – rekordy do wyliczenia centyli ciśnienia, 9 – rekordy, które nie należy uwzględniać w wyznaczeniu centyli ciśnienia
height	wysokość ciała (cm)
weight	waga (kg)
SBP	wysokość ciśnienia skurczowego (mmHg) wyznaczona jako średnia z drugiego i trzeciego pomiaru, które były pobierane w 15-minutowych odstępach czasu
BMI	wartość BMI (<i>body mass index</i>)
DBP	wysokość ciśnienia rozkurczowego (mmHg) wyznaczona jako średnia z drugiego i trzeciego pomiaru, które były pobierane w 15-minutowych odstępach czasu
nadwaga	0 – bez nadwagi, 1 – nadwaga

Dodatek B

Wybrane definicje

B.1. Kryteria informacyjne – AIC i BIC

W kontekście klasy modeli szacowanych za pomocą metody największej wiarygodności definiuje się *kryteria informacyjne* pozwalające porównywać jakość różnych modeli dla zmiennej zależnej. Konwencja, która została przyjęta dla tych modeli uznaje za najlepszy ten model, dla którego wartość kryterium informacyjnego jest najniższa. Najpopularniejszymi kryteriami informacyjnymi jest *kryterium informacyjne Akaike'go*, AIC (ang. *Akaike Information Criterion*) oraz *Bayesowskie kryterium informacyjne*, BIC (ang. *Bayes Information Criterion*).

Odpowiednie wzory dla kryteriów są następujące:

$$AIC = 2K - 2l,$$

$$BIC = K \ln(N) - 2l,$$

gdzie l jest logarymem funkcji wiarygodności dla oszacowanego wektora parametrów, K jest liczbą parametrów w modelu, a N liczbą obserwacji.

Na ogół model o większej liczbie predyktorów daje dokładniejsze przewidywania, jednak ma też większą skłonność do przeuczenia. W przypadku AIC czynnik $2K$ odgrywa rolę kary jaką trzeba ponieść za dołączenie nowych zmiennych do modelu i ma na celu zrównoważenie oczywistego zmniejszania się w takim przypadku składnika l . Dysponując kilkoma modelami, z reguły wybiera się ten, dla którego wartość kryterium AIC jest najmniejsza lub ten, który charakteryzuje się „dostatecznie” małą wartością AIC i zarazem jest opisywany przez stosunkowo niewielki zestaw parametrów.

Różnica między kryterium AIC i BIC polega na innym ważeniu jakości dopasowania i prostoty modelu. Pierwszy składnik we wzorze na kryteria informacyjne mierzy prostotę modelu. W obu przypadkach element ten rośnie wraz ze wzrostem liczby parametrów i wzrost ten jest tym większy im mniejsza jest liczba obserwacji. Takie zdefiniowanie kryterium informacyjnego związane jest z faktem, że prostota modelu jest szczególnie ważna w przypadku modeli szacowanych na małych próbach. Jakkolwiek asymptotycznie oba kryteria wybierać będą jako prawidłowy model, model prawdziwy, to jednak w małych próbach ich wskazania mogą się znacznie różnić. W literaturze sugeruje się, że kryterium AIC ma tendencję do wybierania modelu o zbyt dużej liczbie parametrów.

B.2. Skośność

Miara asymetryczności, nazywana trzecim momentem standaryzowanym, który dla zmiennej losowej X zdefiniowany jest wzorem:

$$\nu = E \left[\left(\frac{X - \mu}{\sqrt{\sigma}} \right)^3 \right] = \frac{\mu_3}{(\sqrt{\sigma})^3},$$

gdzie: μ - wartość oczekiwana X , $\sqrt{\sigma}$ - odchylenie standardowe X , [14].

Dla próby x_i dla $i = 1, 2, \dots, n$ współczynnik skośność wyznacza wzór:

$$\nu = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{3/2}},$$

gdzie \bar{x} to próbowa średnia.

Warto zauważyć, że dla rozkładu symetrycznego (np. rozkładu normalnego) $\nu = 0$. Dwa górne wykresy rys. B.1 przedstawiają przykładowe funkcje gęstości prawdopodobieństwa dla ujemnie i dodatnio skośnych rozkładów.

B.3. Kurtoza

Miara koncentracji (spłaszczenia), nazywana czwartym momentem standaryzowanym, która dla zmiennej losowej X zdefiniowana jest wzorem:

$$\tau = E \left[\left(\frac{X - \mu}{\sqrt{\sigma}} \right)^4 \right] - 3 = \frac{\mu_4}{\sigma^2} - 3,$$

gdzie: μ_4 to czwarty moment centralny X^1 , $\sqrt{\sigma}$ to odchylenie standardowe² X , [14].

Dla próby x_i dla $i = 1, 2, \dots, n$ współczynnik kurtozy wyznacza wzór:

$$\tau = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^2} - 3,$$

gdzie \bar{x} to próbowa średnia.

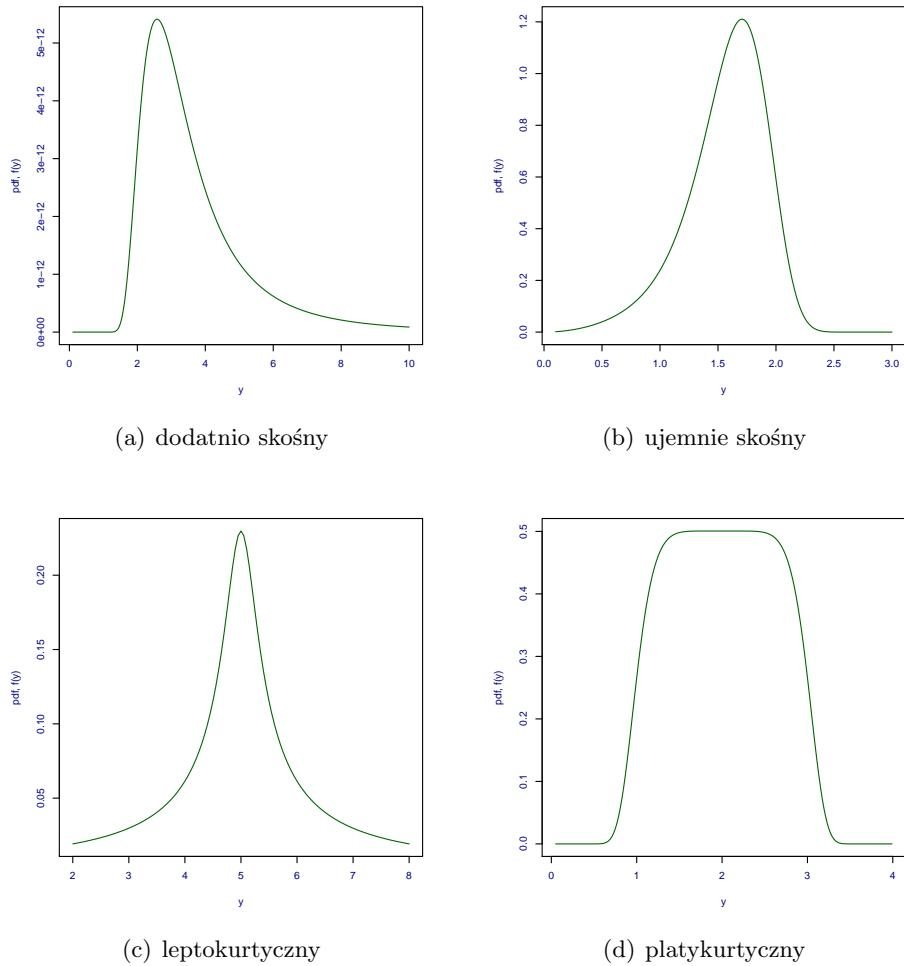
Warto zauważyć, iż:

- dla rozkładu normalnego $\tau = 0$, wykres rozkładu prawdopodobieństwa ma kształt normalny, tzw. *mezokurtyczny*,
- dla $\tau > 0$ rozkład jest bardziej wysmukły niż normalny, tzw. *leptokurtyczny*, który posiada większe skupienie wartości wokół średniej,
- dla $\tau < 0$ rozkład jest mniej wysmukły niż normalny, tzw. *platykurtyczny*, który posiada większe spłaszczenie rozkładu.

Dwa dolne wykresy rys. B.1 przedstawiają przykładowe funkcje gęstości prawdopodobieństwa dla platokurtycznego i leptokurtycznego rozkładu.

¹Czwarty moment centralny jest wyznaczany za pomocą wzoru: $\mu_4 = E[X - E(X)]^4$

²W niektórych pracach, (szczególnie starszych) można spotkać się ze wzorem dla współczynnika kurtozy, w którym nie jest odejmowana liczba 3 od powyższego ułamka. Nowa definicja kurtozy jest jednak bardziej wygodna, gdyż np. kurtoza rozkładu normalnego wynosi 0.



Rysunek B.1: Przykłady wykresów funkcji gęstości prawdopodobieństwa dla rozkładów skośnych i kurtycznych. Źródło: opracowanie własne.

B.4. Wybrane rozkłady

B.4.1. Rozkład normalny (NO)

Domyślnym rozkładem w funkcji `gamlss()` jest dwuparametryczny ciągły rozkład normalny. Parametryzacja wykorzystywana dla normalnej (gaussowskiej) funkcji gęstości prawdopodobieństwa ozn. ($NO(\mu, \sigma)$) jest wyrażona wzorem:

$$f_Y(y|\mu, \sigma) = \frac{1}{\sqrt{2\Pi}\sigma} \exp \left[-\frac{(y-\mu)^2}{2\sigma^2} \right] \quad (\text{B.1})$$

gdzie $-\infty < y < \infty$, $-\infty < \mu < \infty$ i $\sigma > 0$. Średnia i wariancja Y są dane odpowiednio przez $E(Y) = \mu$, $Var(Y) = \sigma^2$, zatem σ to odchylenie standardowe zmiennej losowej Y , [20].

B.4.2. Rozkład log-normalny (LOGNO, LNO)

Rozkład log-normalny (LOGNO)

Rozkład log-normalny jest odpowiedni dla dodatnio skośnych danych. Funkcja gęstości prawdopodobieństwa dla rozkładu log-normalnego ozn. $LOGNO(\mu, \sigma)$ dla $y > 0$ jest zdefiniowana:

$$f_Y(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma^2} \frac{1}{y} \exp \left[-\frac{(\log(y) - \mu)^2}{2\sigma^2} \right], \quad (\text{B.2})$$

gdzie $\mu > 0$ i $\sigma > 0$. Tutaj $E(Y) = \omega^{1/2}e^\mu$ i $Var(Y) = \omega(\omega - 1)e^{2\mu}$, gdzie $\omega = \exp(\sigma^2)$.

Rodzina rozkładów Log-normalnych (LNO)

W pakiecie gamlss funkcja $LNO(\mu, \sigma, \nu)$ umożliwia zastosowanie wykładowiczej transformacji Box'a-Cox'a, gdzie przekształcenie $Y(\nu)$ jest wprowadzone w celu usunięcia skośności, oraz gdzie $Z = (Y^\nu - 1)/\nu$ (jeśli $\nu \neq 0$) + $\log(Y)$ (jeśli $\nu = 0$). O przekształconej zmiennej Z zakłada się wówczas, że ma rozkład normalny $NO(\mu, \sigma)$, zaś o wynikowym rozkładzie Y można powiedzieć że ma rozkład $LNO(\mu, \sigma, \nu)$. Jeśli $\nu = 0$, rozkład będzie postaci (B.2). Dla wartości $\nu \neq 0$ oraz dla $y > 0$ wynikowy rozkład będzie 3-parametrowym rozkładem:

$$f_Y(y|\mu, \sigma, \nu) = \frac{y^{\nu-1}}{\sqrt{2\pi}\sigma^2} \exp \left[-\frac{(z - \mu)^2}{2\sigma^2} \right], \quad (\text{B.3})$$

gdzie $\mu > 0$, $\sigma > 0$, $-\infty < \nu < \infty$ oraz gdzie $z = (y^\nu - 1)/\nu$ (jeśli $\nu \neq 0$) + $\log(y)$ (jeśli $\nu = 0$). Alternatywą dla rozkładu (B.3) może być bardziej ortogonalna parametryzacja – rozkład $BCCG$, [20].

B.4.3. Rozkład Box'a-Cox'a-Cole'a-Green'a (BCCG)

Rozkład Box'a-Cox'a-Cole'a-Green'a ($BCCG$) jest dobry dla dodatnio lub ujemnie skośnych danych. Niech $Y > 0$ będzie dodatnią zmienną losową o rozkładzie Box'a-Cox'a-Cole'a-Green'a oznaczoną $BCCG(\mu, \sigma, \nu)$ i zdefiniowaną przez przekształcenie zmiennej losowej Z podanej przez:

$$Z = \begin{cases} \frac{1}{\sigma\nu} \left[\left(\frac{Y}{\mu} \right)^\nu - 1 \right] & \text{jeżeli } \nu \neq 0 \\ \frac{1}{\sigma} \log \left(\frac{Y}{\mu} \right) & \text{jeżeli } \nu = 0 \end{cases} \quad (\text{B.4})$$

dla $0 < Y < \infty$, gdzie $\mu > 0$, $-\infty < \nu < \infty$, oraz gdzie Z ma ucięty rozkład normalny. Warunek $0 < Y < \infty$ (wymagany do tego aby Y^ν była rzeczywistą liczbą $\forall \nu$) prowadzi do warunku $-\frac{1}{\sigma\nu} < Z < \infty$ jeżeli $\nu > 0$ oraz $-\infty < Z < -\frac{1}{\sigma\nu}$ dla $\nu < 0$, co wymaga uciętego rozkładu normalnego dla Z .

Stąd parametryzacja gęstości zmiennej Y będzie dana przez:

$$f_Y(y) = \frac{y^{\nu-1} \exp(-\frac{1}{2}z^2)}{\mu^\nu \sigma \sqrt{2\pi} \Phi(\frac{1}{\sigma|\nu|})} \quad (\text{B.5})$$

gdzie z jest dane wzorem (B.4), a $\Phi()$ to dystrybuanta standardowego rozkładu normalnego.

Jeśli prawdopodobieństwo obcięcia $\Phi(-\frac{1}{\sigma|\nu|})$ jest nieistotne, zmienna Y ma medianę μ . Parametryzacja (B.4) była używana przez Cole'a i Green'a (1992), którzy zakładali standaryzowany rozkład normalny dla Z i nieistotność obcięcia, [20].

Bibliografia

- [1] van Buuren S., Fredriks M., *Worm plot: a simple diagnostic device for modelling growth reference curves*, Statistics in Medicine, 20, str. 1259–1277, 2001.
- [2] Cole T. J., Stanojevic S., Stocks J., Coates A. L., Hankinson J. L., Wade A. M. *Age- and size-related reference ranges: A case study of spirometry through childhood and adulthood*, Stat Med., 28, str. 880–898, 2009.
- [3] Filliben, J. J., *The Probability Plot Correlation Coefficient Test for Normality*, <http://www.jstor.org/pss/1268008>, 2011.
- [4] Fox J., *Nonparametric Regression, Appendix to An R and S-PLUS Companion to Applied Regression*, <http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-nonparametric-regression.pdf>, 2002.
- [5] Gatnar E., *Nieparametryczna metoda dyskryminacji i regresji*, PWN, str. 13-19, 2001.
- [6] Gramacki A., Gramacki J., *Estymacja nieparametryczna wybranych parametrów bloku gazowo-parowego*, <http://lord.uz.zgora.pl:7777/skep/docs/F5828/GramackiKNWS092.pdf>, 2009.
- [7] Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning*, Springer, str. 139-158, 186-189, 295-304, 2001.
- [8] Jakubczyk K., *Interpolacja funkcjami sklejonymi*, <http://www.kaj.pr.radom.pl/prace/Splines.pdf>, 2008.
- [9] Kincaid D., Cheney W., *Analiza numeryczna*, ISBN, str. 328-361, Warszawa 2006.
- [10] Koronacki J., Ćwik J., *Statystyczne systemy uczące się*, ISBN, str. 168-179, Warszawa 2005.
- [11] Kułaga Z., Litwin M. i in., *Polish 2010 growth references for school-aged children and adolescents*, 170(5), str. 599-609, European journal of pediatrics, 2011.
- [12] Kułaga Z., Litwin M., Januszewicz A., Prejbisz A., *Nadciśnienie tętnicze u młodzieży i młodych dorosłych*, Medycyna Praktyczna, str. 67-98, 2011.
- [13] Kułaga Z., Litwin M., Lis A. i in., *Oscillometric blood pressure percentiles for Polish normal-weight school-aged children and adolescents*, w przygotowaniu.
- [14] Magiera R., *Modele i metody statystyki matematycznej, Cz. I, Rozkłady i symulacja stochastyczna*, str. 28, 180, GiS, 2005.

- [15] Neuhauser H. K., Thamm M., Ellert U., Werner Hense H., Schaffrath Rosario A., *Blood Pressure Percentiles by Age and Height From Nonoverweight Children and Adolescents in Germany*, Pediatrics, e978, 2011.
- [16] Rodriguez G., *Smoothing and Non-Parametric Regression*,
<http://data.princeton.edu/eco572/smoothing.pdf>, 2001.
- [17] Stasinopoulos D., Rigby B., *A flexible regression approach using GAMlSS in R*,
<http://www.jstatsoft.org/v23/i07/paper>, 2009.
- [18] Stasinopoulos D., Rigby B., *Generalized Additive Models for Location Scale and Shape*, Applied Statistics, 507–554, 2005.
- [19] Stasinopoulos D., Rigby B., *Generalized Additive Models for Location Scale and Shape (GAMlSS) in R*,
<http://www.jstatsoft.org/v23/i07/paper>, 2007.
- [20] Stasinopoulos D., Rigby B., Akantziliotou C., *Instructions on how to use the gamlss package in R*,
<http://www.gamlss.org/manual.pdf>, 2008.
- [21] Stasinopoulos D., Rigby B., Akantziliotou C., *Instructions on how to use the gamlss package in R – Second Edition*,
<http://studweb.north.londonmet.ac.uk/stasinom/papers/gamlss-manual.pdf>, 2008.
- [22] Stasinopoulos D., Rigby B., Akantziliotou C., *Package „gamlss” – Generalized Additive Models for Location Scale and Shape*, 2011.
- [23] Strona główna projektu R,
<http://www.r-project.org/>.
- [24] WHO Multicentre Growth Reference Study Group, *WHO Child Growth Standards: Methods and Development*, World Health Organization, Geneva, Switzerland, str.6, str. 209-212, 2006.