Uniwersytet Warszawski
Wydział Matematyki, Informatyki i Mechaniki

Witold Chodor
Nr albumu: 291520

Prognozowanie możliwości sportowców w lekkoatletycznych dyscyplinach biegowych

Praca licencjacka
na kierunku MATEMATYKA

Praca wykonana pod kierunkiem
dr. inż. Przemysława Biecka
Instytut Matematyki Stosowanej i Mechaniki - Zakład Statystki Matematycznej

Sierpień 2013

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 przedstawione zostały przewidywane możliwości lekkoatletów w ośmiu dyscyplinach biegowych. Ich uzyskanie było możliwe dzięki zbudowaniu odpowiedniego modelu potęgowego, ukazującego zależność pomiędzy czasem a dystansem. Praca składa się z trzech części: teoretycznej, poświęconej funkcji `nls` oraz praktycznej. W części teoretycznej omówione zostały modele liniowe, test ilorazu wiarogodności oraz modele nieliniowe. W dalszej kolejności przedstawione zostało zastosowanie funkcji `nls` do modeli nieliniowych. Część praktyczna przedstawia kolejne etapy konieczne do uzyskania szukanych granic możliwości. W każdym z etapów wykorzystane są podane zagadnienia teoretyczne.

Słowa kluczowe

predykcja, model liniowy, test ilorazu wiarogodności, model zagnieżdżony, model nielinowy, `nls`, lekkoatletyka

Dziedzina pracy (kody wg programu Socrates-Erasmus)

11.2 Statystyka

Klasyfikacja tematyczna

62F03, 62P99, 62J02

Tytuł pracy w języku angielskim

Predicting athletes' abilities in athletic running events

Spis treści

Spis rysunków

Spis tabel

Przedmowa

Tematem prezentowanej pracy jest przewidywanie możliwości sportowców w wybranach lekkoatletycznych dyscyplinach biegowych. Przyglądając się występom lekkoatletów często zadajemy sobie pytanie, jak bardzo mogą oni poprawić swoje najlepsze osiągnięcia. Wydaje nam się bowiem, że musi istnieć jakaś fizjologiczna granica ludzkich możliwości. Również naukowcy, w tym statystycy, zadają sobie to pytanie i przy pomocy odpowiednich narzędzi statystycznych starają się jak najdokładniej wyznaczyć granice możliwości sportowców ([4], [5]).

Praca składa się z trzech rozdziałów. W pierwszym z nich przedstawione zostały zagadnienia teoretyczne, które zostały podzielone na trzy części. Najpierw omówiona została tematyka modeli linowych ze szczególnym naciskiem na estymację parametrów. W dalszej kolejności przedstawiona została idea testu ilorazu wiarogodności wraz z przykładem jego wykorzystania w modelach liniowych zagnieżdżonych. Na koniec części teoretycznej zaprezentowana została ogólnie tematyka modeli nieliniowych. Drugi rozdział stanowi krótki opis wykorzystania funkcji `nls` podczas tworzenia modeli nieliniowych. Z kolei trzeci rozdział to analiza danych rzeczywistych, w której modelowanym zagadnieniem jest przewidywanie możliwości lekkoatletów na podstawie rekordów, uzyskanych w ośmiu konkurencjach biegowych w ciągu 100 lat. Kolejne podrozdziały, nie wliczając opisu zbioru danych, są umieszczone tak, że odpowiedają kolejnym zagadnieniom teoretycznym omówionym w pierwszym rozdziale. Do pracy zostały również dołączone dwa dodatki. W pierwszym z nich znajduje się informacja na temat rozkładu QR macierzy. Z kolei drugą część dodatku stanowią kody programu R, które wykorzystałem do realizacji części praktycznej.

Podziękowania

Chciałbym w tym miejscu wyrazić moją ogromną wdzięczność dla osób, dzięki którym ta praca powstała. Przede wszystkim pragnę podziękować mojemu promotorowi dr.inż. Przemysławowi Bieckowi, którego cenne rady i wskazówki pozwoliły mi rozwiązać wszelkie trudności jakie napotykałem w trakcie pisania. Jestem również ogromnie zobowiązany Panu Andrzejowi Skibie, którego pasja i zangażowanie w ogromnej mierze przyczyniły się do wybrania przeze mnie studiów matematycznych. Na koniec chciałbym skierować moje podziękowania dla mojej mamy Anny Chodor za wspieranie mnie w wielu trudnych dla mnie momentach podczas studiów.

Rozdział 1

Wstęp teoretyczny

1.1. Modele liniowe

Modele liniowe zajmują szczególne miejsce w statystyce ze względu na liczne zastosowania w biologii, chemii, medycynie, ekonomii i wielu innych naukach doświadczalnych. Stanowią one jedną z najbardziej popularnych metod, pozwalających na opisanie zależności między zbiorem zmiennych objaśniających, a zmienną objaśnianą. Ich ogromną zaletą jest możliwość prostego wyznaczenia statystyk potrzebnych do estymacji i testowania. Dlatego też wykorzystywano je w praktyce zanim zaczęto korzystać z możliwości komputerów.

W poniższym rozdziale omówię ogólnie najważniejsze zagadnienia dotyczące modeli liniowych, które wykorzystane zostaną później w części praktycznej. Więcej szczegółów na ten temat możemy znaleźć, między innymi, w: rozdziałach pierwszym i drugim monografii [2], rozdziale ósmym skryptu [9] oraz rozdziale ósmym wykładów [11].

1.1.1. Opis modelu

Będzie nas interesowało opisanie zależności między zmienną objaśnianą a zbiorem zmiennych objaśniających. Zależność tę będziemy chcieli ocenić na podstawie wartości poszczególnych zmiennych dla zbioru n obiektów.

Zazwyczaj w roli y przyjmujemy pewną funkcję zmiennej objaśnianej. W celu uproszczenia opisu modelu przyjmijmy, że będzie nią przekształcenie tożsamościowe. Niech zatem y będzie zmienną objaśnianą. Wówczas każda z wartości tej zmiennej dla i-tego obiektu to liczba rzeczywista y_i. Zbiór s zmiennych objaśniających oznaczmy jako $V = \{V_1, \ldots, V_s\}$. V_j jest zmienną objaśniającą, która określa wartość każdego z n obiektów w sposób jakościowy bądź też ilościowy. Na początku będziemy zatem dysponowali n rzeczywistymi wartościami zmiennej y oraz $n \times s$ niekoniecznie ilościowymi wartościami zmiennej V.

W statystyce często bywa tak, że należy wybrać, które ze zmienych V_j będą użyteczne. Ponadto może się zdarzyć, że zmienne V_j będą jakościowe. Dlatego też musimy dokonać transformacji zmiennej V. Niech $X = \{X_1, \ldots, X_p\}$ będzie zbiorem zmiennych odpowiednio zakodowanych przy pomocy zmiennych V_j, z których każda przyjmuje wyłącznie wartości rzeczywiste. Po tym przekształceniu dysponujemy wciąż n rzeczywistymi wartościami zmiennej y, ale tym razem $n \times p$ rzeczywistymi wartościami zmiennej X.

Możemy zatem przejść do przedstawienia liniowej zależności pomiędzy zmiennymi y oraz

X przy pomocy modelu liniowego

$$\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}}_{\substack{\text{losowe,} \\ \text{obserwowane,} \\ n \times 1}} = \underbrace{\begin{matrix} X_1 & & X_p \end{matrix}}_{} \underbrace{\begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix}}_{\substack{\text{deterministyczne,} \\ \text{obserwowane,} \\ n \times p}} \cdot \underbrace{\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}}_{\substack{\text{szukane,} \\ p \times 1}} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{\substack{\text{losowe,} \\ \text{nieobserwowane,} \\ n \times 1}} .$$

W postaci macierzowej możemy zapisać ten model na dwa sposoby:

$$y = X\beta + \varepsilon \tag{1.1.1}$$

lub

$$y = X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon, \tag{1.1.2}$$

gdzie y to wektor kolumnowy z wartościami zmiennej objaśnianej dla kolejnych obiektów; X jest macierzą planu, w której kolejne kolumny to wartości zmiennych X_j dla każdego z n obiektów, tak więc element x_{ij} macierzy X oznaczać będzie wartość zmiennej X_j dla i-tego obiektu; β jest wektorem kolumnowym z nieznanymi parametrami modelu; ε to wektor kolumnowy ze składnikami losowymi.

Uwaga 1.1.1. Gdybyśmy chcieli rozważyć model z wyrazem wolnym wystarczyłoby, żeby, bez straty ogólności, pierwsza kolumna macierzy X była n-elemntowym wektorem złożonym z samych jedynek. Należy jednak podkreślić, że wtedy rozważalibyśmy $p-1$ odpowiednio zakodowanych zmiennych objaśniających X_j.

Uwaga 1.1.2. Pamiętajmy, że przymiotnik „liniowy" odnosi się tylko do liniowej zależności między y a X. Nie musi dotyczyć zależności pomiędzy zmienną objaśnianą y a zbiorem zmiennych objaśniających V. Spodziewając się, na przykład, potęgowej zależności między zmiennymi, możemy w modelu rozważyć liniową zależność między ich logarytmami. Wyraz „liniowy" wskazuje nam również, że parametry B_j wystepują wyłącznie w potędze pierwszej i są pojedynczo zależne od poszczególnych zmiennych X_j, co widzimy dokładnie w równaniu (1.1.2).

Uwaga 1.1.3. Zwróćmy również uwagę, że, w celu uproszczenia opisu, zapisujemy y małą literą ponieważ będziemy jej używać zamiennie jako zmienną losową bądź też wartość obserwacji. Z kontekstu będzie wiadomo, o którą z tych dwóch możliwości nam chodzi.

Dla tak sformułowanego modelu (1.1.1) przyjmiemy następujące założenia:

Założenie 1.1.1. Mamy $p < n$ oraz macierz X jest pełnego rzędu tzn. $\operatorname{rank}(X) = p$.

Wydaje się, że chemy mieć więcej niż p obiektów, w celu wyestymowania p szukanych parametrów. Ponadto chcielibyśmy, żeby poszczególne kolumny macierzy X były liniowo niezależne. W praktyce oznacza to, że odpowiednio zakodowane zmienne objaśniające X_j są niezależne.

Założenie 1.1.2. Wektor składników losowych ε ma rozkład $\mathcal{N}(0, \sigma^2 I_{n\times n})$.

W wyniku przekształcenia afinicznego otrzymujemy, że

$$y \backsim \mathcal{N}(X\beta, \sigma^2 I_{n\times n}). \tag{1.1.3}$$

Model z ograniczeniami

Warto również wspomnieć, że zdarzają się sytuacje podczas tworzenia modeli liniowych, kiedy to statystyk posiada pewne informacje na temat liniowych zależności pomiędzy parametrami β. Wtedy też model liniowy, wraz z uwzględnionymi w dodatkowym równaniu ograniczeniami na parametry, możemy przedstawić w następujący sposób

$$\begin{cases} y = X\beta + \varepsilon \\ A\beta = 0 \end{cases}, \tag{1.1.4}$$

gdzie y, X, β oraz ε wprowadziliśmy już w modelu (1.1.1) natomiast A jest macierzą o wymiarach $(p-q) \times p$. Przez $p-q$ oznaczyliśmy liczbę parametrów liniowo zależnych, więc q oznacza minimalną liczbę parametrów potrzebną do uzyskania pełnej informacji o wektorze β. Używając terminologii statystycznej powiemy, że q oznacza liczbę stopni swobody.

Przykład 1.1.1. Przyjmijmy, że $p = 5$, $p - q = 2$. Wtedy wektor p współczynników jest postaci

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}.$$

Załóżmy ponadto, że chcemy nałożyć następujące $p - q$ ograniczenia liniowe na jego współczynniki: $\beta_1 = \beta_3$, $\beta_2 = 2\beta_4$. Wówczas możemy zapisać je w następującej postaci

$$\underbrace{\begin{bmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & -2 & 0 \end{bmatrix}}_{A_{(p-q)\times p}} \beta = 0.$$

1.1.2. Estymacja parametrów

W poprzednim rozdziale przedstawiliśmy w klasycznej postaci model liniowy (1.1.1). Nieznanym parametrem w tym modelu jest oczywiście β. Zwróćmy również uwagę, że, zgodnie z przyjętym przez nas założeniem 1.1.2, nie wiemy jaka jest wartość parametru σ^2. Dlatego też, w poniższym rozdziale, przedstawię pokrótce teorię dotyczącą estymatorów tych parametrów.

Metoda najmniejszych kwadratów

Na początku zauważmy, że zgodnie z równaniem (1.1.3)

$$\mathbb{E}(y) = X\beta.$$

Po tej uwadze możemy wprowadzić następującą definicję.

Definicja 1.1.1. Estymator najmniejszych kwadratów parametru β to taka jego wartość, dla której kwadraty odległości euklidesowych przybliżanych danych do krzywej je przybliżających są najmniejsze,

$$\widehat{\beta} = \min_{\beta} ||y - X\beta||^2 = \min_{\beta} \mathrm{RSS}(\beta),$$

gdzie RSS (ang. *Residual Sum of Squares*) oznacza sumę kwadratów reszt, $\mathrm{RSS} = ||y - X\beta||^2$.

Jest kilka sposobów umożliwiających dokładne wyznaczenie $\widehat{\beta}$. Jedna z metod opiera się na rozkładzie QR (dodatek A) macierzy planu X i prowadzi do następującego oszacowania tego parametru.

Twierdzenie 1.1.1. *Estymator najmniejszych kwadratów wyraża się wzorem*

$$\widehat{\beta}_{mnk} = R_1^{-1} Q_1^T y,$$

gdzie R_1 oraz Q_1 pochodzą z wąskiego rozkładu QR macierzy X.

Dowód tego twierdzenia możemy znaleźć w rozdziale 8.2 pracy [11].

Innym sposobem na znalezienie estymatora mnk jest rozwiązanie zadanie BLUE (ang. *Best Linear Unbiased Estimator*). Opis tego rozwiązania znajdziemy w rozdziale 8.3 pracy [11]. W efekcie dostajemy estymator

$$\widehat{\beta}_{mnk} = (X^T X)^{-1} X^T y,$$

należący do klasy estymatorów BLUE. Dla pewności powinniśmy jeszcze sprawdzić czy te dwa estymatory najmniejszych kwadratów uzyskane dwiema różnymi metodami są sobie równe.

Stwierdzenie 1.1.1. *Estymator najmniejszych kwadratów jest równy liniowemu, nieobciążonemu estymatorowi o najmniejszej wariancji.*

Dowód. W poniższym dowodzie skorzystam z wąskiego rozkładu QR (dodatek A).

$$\begin{aligned}
\widehat{\beta} &= R_1^{-1} Q_1^T y \\
&= R_1^{-1} I_{p\times p} Q_1^T y \\
&= R_1^{-1} (R_1^T)^{-1} R_1^T Q_1^T y \\
&= (R_1^T R_1)^{-1} R_1^T Q_1^T y \\
&= (R_1^T I_{p\times p} R_1)^{-1} R_1^T Q_1^T y \\
&= (R_1^T Q_1^T Q_1 R_1)^{-1} R_1^T Q_1^T y \overset{(A.0.2)}{=} \\
&= (X^T X)^{-1} X^T y.
\end{aligned}$$

□

Metoda największej wiarogodności

Zanim przejdziemy do formalnych defincji związanych z tą metodą, warto podkreślić intuicję jaka stoi za tym sposobem wyznaczania estymatorów. Otóż estymator największej wiarogodności to taka wartość parametru, dla której prawdopodobieństwo zaobserwowania danych jest największe.

Uwaga 1.1.4. Załóżmy, że Z to zmienna losowa, którą będziemy traktować jako obserwację. Wtedy $z = Z(\omega)$ oznaczać będzie ustaloną wartość obserwacji Z.

Definicja 1.1.2. Wiarogodność jest to funkcja $\mathcal{L} : \Theta \to \mathbb{R}$ dana wzorem

$$\mathcal{L}(\theta(z)) = f_\theta(z),$$

gdzie $f_\theta(z)$ jest łączną gęstością rozkładu obserwacji Z.

Przykład 1.1.2 (**wiarogodność parametrów** β**,** σ^2)**.** Zgodnie z równaniem (1.1.3) wiemy, że zmienna $y \backsim \mathcal{N}(X\beta, \sigma^2 I_{n\times n})$. Przyjmijmy następujące oznaczenia:

$$\mu = \mathbb{E}(y) = X\beta, \tag{1.1.5}$$

$$\Sigma = \mathrm{Var}(y) = \sigma^2 I_{n\times n}. \tag{1.1.6}$$

Korzystając z twierdzenia o gęstości zmiennej o wielowymiarowym rozkładzie normalnym (dowód tego twierdzenia możemy, na przykład, znaleźć w rozdziale 5 pracy [10]) dostajemy, że funkcja wiarogodności jest postaci

$$\begin{aligned}
\mathcal{L}(\beta, \sigma^2) = f_{\beta,\sigma^2}(y) &= \frac{\sqrt{\det(\Sigma^{-1})}}{(2\pi)^{\frac{n}{2}}} \exp[-\frac{1}{2}\langle \Sigma^{-1}(y-\mu), y-\mu\rangle] \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}(\sigma^2)^{\frac{n}{2}}} \exp[-\frac{1}{2\sigma^2}\langle I_{n\times n}(y-\mu), y-\mu\rangle] \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}(\sigma^2)^{\frac{n}{2}}} \exp[-\frac{1}{2\sigma^2}\langle y-\mu, y-\mu\rangle] \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}(\sigma^2)^{\frac{n}{2}}} \exp[-\frac{1}{2\sigma^2}||y-\mu||^2] \overset{(1.1.5)}{=} \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}(\sigma^2)^{\frac{n}{2}}} \exp[-\frac{1}{2\sigma^2}||y-X\beta||^2]
\end{aligned} \tag{1.1.7}$$

Definicja 1.1.3. Powiemy, że $\widehat{\theta} = \widehat{\theta}(Z)$ jest **estymatorem największej wiarogodności** parametru θ, jeśli

$$\mathcal{L}(\widehat{\theta}(z)) = f_{\widehat{\theta}}(z) = \sup_{\theta\in\Theta} f_\theta(z) = \sup_{\theta\in\Theta} \mathcal{L}(\theta(z))$$

dla dowolnego z.

Uwaga 1.1.5. Często bywa tak, że funkcja wiarogodności jest przedstawiona w postaci iloczynu pewnych czynników. Żeby ułatwić sobie znajdowanie supremum takiej funkcji wystarczy, że ją zlogarytmujemy. Działanie to jest jak najbardziej uprawnione ponieważ logarytm jest funkcją ściśle rosnącą. Wprowadźmy zatem następujące oznaczenie na logarytm funkcji wiarogodności:

$$\ell(\theta(z)) = \log \mathcal{L}(\theta(z)).$$

Po zapoznaniu się z teorią możemy przejść do znalezienia estymatorów najmniejszej wiarogodności parametrów β oraz σ^2 modelu (1.1.1). Zauważmy, że wprowadzona w przykładzie 1.1.2 funkcja wiarogodności jest iloczynem wielu czynników. Dlatego też zamiast ją maksymalizować łatwiej nam będzie szukać maksimum jej logarytmu. Wybierzemy logarytm naturalny jako, że jednym z czynników jest funkcja exp. Po zlogarytmowaniu równania (1.1.7) otrzymujemy

$$\begin{aligned}
\ell(\beta, \sigma^2) = \ln(f_{\beta,\sigma^2}(y)) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}||y-X\beta||^2 \\
&= C - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}||y-X\beta||^2,
\end{aligned}$$

gdzie C jest stałą niezależną od szukanych parametrów. Zmaksymalizowanie funkcji $\ell(\beta, \sigma^2)$ jest równoważne minimalizacji funkcji $-2\ell(\beta, \sigma^2)$

$$-2\ell(\beta, \sigma^2) = C' + n\ln(\sigma^2) + \frac{1}{\sigma^2}||y-X\beta||^2, \tag{1.1.8}$$

Po ustaleniu parametru σ^2 okazuje się, że minimalizacja równania (1.1.8), ze względu na parmetr β, prowadzi do znanej już nam z definicji 1.1.1 minimalizacji składnika $||y - X\beta||^2$. Stąd też otrzymujemy, że estymator najmniejszej wiarogodności parametru β jest równy estymatorowi najmniejszych kwadratów

$$\widehat{\beta}_{mnw} = \widehat{\beta}_{mnk} = (X^T X)^{-1} X^T y = R_1^{-1} Q_1^T y.$$

Ponieważ $\widehat{\beta}$ nie zależy od σ^2 możemy w równaniu (1.1.8) ustalić $\beta = \widehat{\beta}$, otrzymując równanie z jedną niewiadomą σ^2

$$-2\ell(\widehat{\beta}, \sigma^2) = C' + n \ln(\sigma^2) + \frac{1}{\sigma^2} ||y - X\widehat{\beta}||^2. \tag{1.1.9}$$

Zatem znalezienie estymatora parametru σ^2 sprowadza się do minimalizacji prawej strony równania (1.1.9). Przyjmijmy oznaczenie $\tau = \sigma^2$, żeby nie pomylić się podczas różniczkowania. Wyznaczmy zatem pochodną cząstkową prawej strony równania (1.1.9) po parametrze τ i przyrównajmy ją do zera, żeby znaleźć punkt stacjonarny.

$$\frac{\partial(-2\ell(\widehat{\beta}, \tau))}{\partial(\tau)} = \frac{n}{\tau} - \frac{1}{\tau^2} ||y - X\widehat{\beta}||^2 = 0. \tag{1.1.10}$$

Mnożąc równanie (1.1.10) przez τ^2 otrzymujemy

$$n\tau - ||y - X\widehat{\beta}||^2 = 0,$$

skąd dostajemy, że

$$\widehat{\tau} = \widehat{\sigma^2} = \frac{||y - X\widehat{\beta}||^2}{n}.$$

Sprawdźmy jeszcze czy $\widehat{\sigma^2}$ jest argumentem dla którego prawa strona równania (1.1.9) jest minimalna

$$\begin{aligned} \frac{\partial^2(-2\ell(\widehat{\beta}, \tau))}{(\partial\tau)^2}\Big|_{\tau=\widehat{\tau}} &= -\frac{n}{\widehat{\tau}^2} + \frac{2||y - X\widehat{\beta}||^2}{\widehat{\tau}^3} \\ &= \frac{-n\widehat{\tau} + 2n\widehat{\tau}}{\widehat{\tau}^3} = \frac{n\widehat{\tau}}{\widehat{\tau}^3} = \frac{n}{\widehat{\tau}^2} > 0 \end{aligned}$$

Zatem faktycznie prawa strona równania (1.1.9) przyjmuje w punkcie $\widehat{\sigma^2}$ minimum.

Podsumujmy teraz ostatecznie oszacowania parametrów jakie uzyskaliśmy:

Wniosek 1.1.1. *Estymatory parametru β dla metody najmniejszych kwadratów i metody największej wiarogodności są równe*

$$\widehat{\beta}_{mnw} = \widehat{\beta}_{mnk} = (X^T X)^{-1} X^T y = R_1^{-1} Q_1^T y. \tag{1.1.11}$$

Wniosek 1.1.2. *Estymator parametru σ^2 dla metody największej wiarogodności jest równy z dokładnością do stałej*

$$\widehat{\sigma^2}_{mnw} = \frac{||y - X\widehat{\beta}||^2}{n}. \tag{1.1.12}$$

1.1.3. Własności estymatora parametru β

Po wyznaczeniu różnymi metodami estymatorów parametru β możemy omówić ich najważniejsze własności.

Wartość oczekiwana estymatora $\widehat{\beta}$

Okazuje się, że $\widehat{\beta}$ jest nieobciążonym estymatorem parametru β.

$$\mathbb{E}(\widehat{\beta}) \overset{(1.1.11)}{=} \mathbb{E}((X^TX)^{-1}X^Ty) \overset{(1.1.5)}{=} (X^TX)^{-1}X^TX\beta = \beta. \tag{1.1.13}$$

Wariancja estymatora $\widehat{\beta}$

Pamiętajmy, że $\widehat{\beta}$ jest wektorem dlatego też jego wariancja będzie macierzą. Będę korzystał z założenia 1.1.1, które zapewnia, że mogę odwracać macierze X, X^T jak również iloczyn XX^T.

$$\begin{aligned}
&\mathrm{Var}(\widehat{\beta}) = \mathbb{E}[(\widehat{\beta} - \mathbb{E}(\widehat{\beta}))(\widehat{\beta} - \mathbb{E}(\widehat{\beta}))^T] \overset{(1.1.11)}{=} \\
&= \mathbb{E}[((X^TX)^{-1}X^Ty - \mathbb{E}((X^TX)^{-1}X^Ty))((X^TX)^{-1}X^Ty - \mathbb{E}((X^TX)^{-1}X^Ty))^T] \\
&= \mathbb{E}[((X^TX)^{-1}X^Ty - (X^TX)^{-1}X^T\mathbb{E}(y))((X^TX)^{-1}X^Ty - (X^TX)^{-1}X^T\mathbb{E}(y))^T] \\
&= \mathbb{E}[((X^TX)^{-1}X^T(y - \mathbb{E}(y))((X^TX)^{-1}X^T(y - \mathbb{E}(y))^T] \\
&= \mathbb{E}[(X^TX)^{-1}X^T(y - \mathbb{E}(y))(y - \mathbb{E}(y))^T((X^TX)^{-1}X^T)^T] \\
&= (X^TX)^{-1}X^T\mathbb{E}[(y - \mathbb{E}(y))(y - \mathbb{E}(y))^T]((X^TX)^{-1}X^T)^T \\
&= (X^TX)^{-1}X^T\mathrm{Var}(y)((X^TX)^{-1}X^T)^T \overset{(1.1.6)}{=} (X^TX)^{-1}X^T\sigma^2I_{n\times n}((X^TX)^{-1}X^T)^T \\
&= \sigma^2(X^TX)^{-1}X^T((X^TX)^{-1}X^T)^T = \sigma^2(X^TX)^{-1}X^TX((X^TX)^{-1})^T \\
&= \sigma^2(X^TX)^{-1}X^TX(X^{-1}X^{-T})^T = \sigma^2(X^TX)^{-1}X^TXX^{-1}X^{-T} \\
&= \sigma^2(X^TX)^{-1}.
\end{aligned} \tag{1.1.14}$$

Rozkład estymatora parametru β

Stwierdzenie 1.1.2. *Przy założeniu 1.1.2 estymator parametru β ma rozkład*

$$\widehat{\beta} \backsim \mathcal{N}(\beta, \sigma^2(X^TX)^{-1}).$$

Dowód. Na podstawie wniosku 1.1.1 oraz równania 1.1.3 wiemy, że zmienna $\widehat{\beta} = (X^TX)^{-1}X^Ty$ ma rozkład normalny w $\mathbb{R}^p$, ponieważ otrzymaliśmy ją w wyniku przekształcenia liniowego ($(X^TX)^{-1}X^T$ jest macierzą tego przekształcenia) zmiennej y o rozkładzie normalnym w $\mathbb{R}^n$. Znając wartość oczekiwaną (1.1.13) oraz wariancję (1.1.14) mamy jednoznacznie wyznaczony wielowymiarowy rozkład normalny $\widehat{\beta}$. □

1.2. Test ilorazu wiarogodności

Podczas tworzenia modeli statystycznych często chcemy porównać dwa modele, które opisują nasz zbiór obserwacji. Okazuje się, że jednym ze sposobów, pozwalających na podjęcie decyzji, który z dwóch modeli wybrać, jest test ilorazu wiarogodności (ang. *likelihood ratio test*). Test ten, jak sama nazwa wskazuje, oparty jest na ilorazie funkcji wiarogodności dwóch modeli. Można więc nieformalnie powiedzieć, że chcemy sprawdzić, na którym z dwóch rozważanych modeli prawdopodobieństwo otrzymania danej obserwacji jest większe.

Przejdźmy teraz do formalnego opisu testowania hipotez przy pomocy testu ilorazu wiarogodności.

1.2.1. Ogólna postać testu

Rozważmy dwa modele statystyczne określone na tej samej przestrzeni obserwacji. W modelu pierwszym mamy do czynienia z rodziną rozkładów prawdopodobieństwa o gęstościach $f_0(\theta, z)$, gdzie $\theta \in \Theta_0$, natomiast w modelu drugim mamy gęstości $f_1(\theta, z)$, gdzie $\theta \in \Theta_1 = \Theta \setminus \Theta_0$. Będziemy chcieli zdecydować, który z dwóch modeli wybrać do opisu obserwacji Z. Przypuśćmy, że chcemy testować

$$H_0 : Z \backsim f_0(\theta, z) \text{ dla pewnego } \theta \in \Theta_0,$$

przeciw

$$H_1 : Z \backsim f_1(\theta, z) \text{ dla pewnego } \theta \in \Theta_1.$$

Za statystykę testową przyjmiemy

$$\lambda(z) = \frac{\sup_{\theta \in \Theta_1} f_1(\theta, z)}{sup_{\theta \in \Theta_0} f_0(\theta, z)} = \frac{f_1(\widehat{\theta}, z)}{f_0(\widetilde{\theta}, z)},$$

gdzie:

$\widehat{\theta}$ to estymator największej wiarogodnośći parametru $\theta \in \Theta_1$,
$\widetilde{\theta}$ to estymator największej wiarogodnośći parametru $\theta \in \Theta_0$.

Procedura testowa:

$$\delta : \text{ odrzucamy } H_0, \text{ jeśli } \lambda > c,$$

gdzie c jest pewną stałą.

1.2.2. Modele zagnieżdżone

Zdarza się, że określiwszy już model statystyczny z rodziną rozkładów prawdopodobieństwa o gęstościach $f_\theta(z)$, gdzie $\theta \in \Theta$, mamy informację wskazującą na pewną zależność między parametrami. W praktyce oznacza to, że możemy ograniczyć się do modelu z gęstościami $f_\theta(z)$, gdzie $\theta \in \Theta_0 \subseteq \Theta$. Mówimy wtedy, że nowy model jest zagnieżdżony w modelu pierwotnym.

Formalnie mamy do czynienia z następującym problemem:

$$\Theta \subseteq \mathbb{R}^p, \quad h : \mathbb{R}^p \to \mathbb{R}^{p-q}, \quad \Theta_0 = \{\theta \in \Theta : h(\theta) = 0\}, \quad p > q \geqslant 1, \quad p, q \in \mathbb{N}.$$

Jeśli h określa przekształcenie liniowe to możemy stwierdzić, że

$$h(\theta) = A\theta = 0,$$

gdzie A jest macierzą przekształcenia liniowego o wymiarach $(p-q) \times p$. Możemy zatem powiedzieć, że mamy $p-q$ równań liniowych z p parametrami. Stąd wniosek, że $\Theta_0 \subseteq \mathbb{R}^q$, a więc $\Theta_1 = \Theta \setminus \Theta_0$ jest zbiorem gęstym w przestrzeni $\mathbb{R}^p$. Wówczas w typowej sytuacji gdy f jest ciągła mamy

$$\sup_{\theta \in \Theta_1} f_\theta(z) = \sup_{\theta \in \Theta} f_\theta(z).$$

Zapis ten pozwala nam uprościć statystykę testową testu ilorazu wiarogodności. Otrzymujemy wówczas, że

$$\lambda(z) = \frac{\sup_{\theta \in \Theta} f_\theta(z)}{\sup_{\theta \in \Theta_0} f_\theta(z)} = \frac{f_{\widehat{\theta}(z)}(z)}{f_{\widetilde{\theta}(z)}(z)}. \tag{1.2.1}$$

1.2.3. Test ilorazu wiarogodności dla modeli liniowych zagnieżdżonych

W tym podrozdziale naszym głównym celem będzie rozważenie szczególnego przypadku testu ilorazu wiarogodności dla modeli linowych z ograniczeniami na parametry. Test pozwoli nam na podjęcie decyzji, czy model wprowadzony w równaniu (1.1.4) nie jest istotnie gorszy od modelu (1.1.1). Oczywiście założenia 1.1.1 oraz 1.1.2 pozostają w mocy.

W równaniu (1.1.7) wyprowadziliśmy wzór na funkcję wiarogodności zmiennej y. Ponieważ zależy ona od parametrów β oraz σ^2, przyjmijmy, że $\theta = (\beta, \sigma^2)$. Załóżmy również, że ograniczona przestrzeń parametrów to $\Theta_0 = \{\theta \in \Theta \ : A\beta = 0\}$. Będziemy chcieli testować

$$H_0 : y \backsim f_\theta(y) \text{ dla pewnego } \theta \in \Theta_0,$$

przeciw

$$H_1 : y \backsim f_\theta(y) \text{ dla pewnego } \theta \in \Theta_1 = \Theta \setminus \Theta_0$$

lub równoważnie

$$H_0 : A\beta = 0,$$

przeciw

$$H_1 : A\beta \neq 0.$$

Zgodnie ze wzorem (1.2.1) wyprowadzimy teraz statystykę testową w tym szczególnym przypadku.

Stwierdzenie 1.2.1. *Statystyka testowa testu ilorazu wiarogodności jest równa*

$$\lambda(y) = \left(\frac{\widetilde{\sigma^2}}{\widehat{\sigma^2}}\right)^{\frac{n}{2}} = \left(\frac{\|y - X\widetilde{\beta}\|^2}{\|y - X\widehat{\beta}\|^2}\right)^{\frac{n}{2}},$$

gdzie:

$\widehat{\theta} = (\widehat{\beta}, \widehat{\sigma^2})$ *estymatory największej wiarogodności w modelu bez ograniczeń (1.1.1) (na zbiorze* Θ*),*

$\widetilde{\theta} = (\widetilde{\beta}, \widetilde{\sigma^2})$ *estymatory największej wiarogodności w modelu z ograniczeniami (1.1.4) (na zbiorze* Θ_0*).*

Dowód. Będę korzystał z estymatora największej wiarogodności parametru σ^2 (równanie (1.1.12)). Otrzymujemy zatem, że:

$$\widetilde{\sigma^2} = \frac{||y - X\widetilde{\beta}||^2}{n}, \quad \widehat{\sigma^2} = \frac{||y - X\widehat{\beta}||^2}{n}. \tag{1.2.2}$$

$$\begin{aligned}
\lambda(y) = \frac{f_{\widehat{\theta}(y)}(y)}{f_{\widetilde{\theta}(y)}(y)} &\overset{(1.1.7)}{=} \frac{\frac{1}{(2\pi)^{\frac{n}{2}}(\widehat{\sigma^2})^{\frac{n}{2}}} \exp\left(-\frac{1}{2\widehat{\sigma^2}}||y - X\widehat{\beta}||^2\right)}{\frac{1}{(2\pi)^{\frac{n}{2}}(\widetilde{\sigma^2})^{\frac{n}{2}}} \exp\left(-\frac{1}{2\widetilde{\sigma^2}}||y - X\widetilde{\beta}||^2\right)} \overset{(1.2.2)}{=} \\
&= \frac{\frac{1}{(\widehat{\sigma^2})^{\frac{n}{2}}} \exp\left(-\frac{n\widehat{\sigma^2}}{2\widehat{\sigma^2}}\right)}{\frac{1}{(\widetilde{\sigma^2})^{\frac{n}{2}}} \exp\left(-\frac{n\widetilde{\sigma^2}}{2\widetilde{\sigma^2}}\right)} = \frac{\frac{1}{(\widehat{\sigma^2})^{\frac{n}{2}}} \exp\left(-\frac{n}{2}\right)}{\frac{1}{(\widetilde{\sigma^2})^{\frac{n}{2}}} \exp\left(-\frac{n}{2}\right)} \\
&= \left(\frac{\widetilde{\sigma^2}}{\widehat{\sigma^2}}\right)^{\frac{n}{2}} = \left(\frac{\frac{||y - X\widetilde{\beta}||^2}{n}}{\frac{||y - X\widehat{\beta}||^2}{n}}\right)^{\frac{n}{2}} = \left(\frac{||y - X\widetilde{\beta}||^2}{||y - X\widehat{\beta}||^2}\right)^{\frac{n}{2}}.
\end{aligned}$$

□

Z uwagi na praktyczne zastosowanie, statystyka wyprowadzona w stwierdzeniu 1.2.1 nie jest użyteczna. Dlatego też podam teraz stwierdzenie przedstawiające statystykę równoważną.

Stwierdzenie 1.2.2. *Statystyka testowa*

$$\lambda(y) = \left(\frac{\widetilde{\sigma^2}}{\widehat{\sigma^2}}\right)^{\frac{n}{2}}$$

jest równoważna statystyce

$$F(y) = \frac{(R_0 - R)/(p - q)}{R/(n - p)}, \tag{1.2.3}$$

gdzie $R_0 = ||y - X\widetilde{\beta}||^2$, $R = ||y - X\widehat{\beta}||^2$.

Dowód tego stwierdzenia możemy znaleźć w rozdziale 8.7.3 pracy [11].

Na koniec przedstawimy bardzo ważne twierdzenie o rozkładzie statystyki z równania (1.2.3).

Twierdzenie 1.2.1. *Statystyka F przy $p \ll n$ ma rozkład $\mathcal{F}$-Snedecora,*

$$F(y) = \frac{(R_0 - R)/(p - q)}{R/(n - p)} \backsim \mathcal{F}(p - q, n - p).$$

Dowód tego twierdzenia ze względu na swoją obszerność zostanie przeze mnie pominięty, ale można go znaleźć w rozdziale 8.7.3 pracy [11].

1.3. Modele nieliniowe

W podrozdziale 1.1 omówiłem pokrótce teorię związaną z modelami liniowymi. W uwadze 1.1.2 wspomniałem, że nie zawsze przedstawiają one liniową zależność między zmienną objaśnianą, a zmienną objaśniającą. Niemniej jednak odpowiednie transformacje tych zmiennych powodują, że w wielu przypadkach jesteśmy w stanie zbudować liniową zależność między odpowiednio przekształconymi zmiennymi. Niestety zdarzają się również sytuacje kiedy musimy założyć, że pomiędzy zmiennymi jest pewna zależność nieliniowa. Prowadzi to do stworzenia modelu nieliniowego, dla którego wyznaczenie chociażby oszacowań parametrów nie jest już tak proste jak w modelu liniowym. Jednak dzięki szybkiemu rozwojowi algorytmów numerycznych oraz geometrii różniczkowej (źródło:[13]) jesteśmy w stanie z coraz wiekszą precyzją dopasowywać modele nieliniowe do danych.

1.3.1. Opis modelu

W ogólności, podobnie jak w przypadku modelu liniowego, chcielibyśmy opisać zależność pomiędzy zmienną objaśnianą a zbiorem zmiennych objaśniających. W tej pracy ograniczymy się jednak do przypadku kiedy mamy do czynienia tylko z jedną zmienną objaśniającą. Nieliniową zależność pomiędzy zmienną objaśnianą a zmienną objaśniającą ocenimy na podstawie wartości tych zmiennych dla zbioru n obserwacji.

Przyjmijmy, że y oznaczać będzie zmienną objaśnianą natomiast x będzie zmienną objaśniającą. Dla uproszczenia opisu załóżmy również, że zarówno zmienna y jak i x są już odpowiednio zakodowanymi zmiennymi o wartościach rzeczywistych. Można powiedzieć, że dysponujemy zbiorem danych składającym się z n par liczb rzeczywistych: $(x_1, y_1), \ldots, (x_n, y_n)$. Dla każdej z tych par opisujemy zależność pomiędzy zmiennymi x_i oraz y_i przy pomocy modelu nieliniowego

$$\underbrace{y_i}_{\substack{\text{losowe,}\\ \text{obserwowane}}} = f(\underbrace{x_i}_{\substack{\text{deterministyczne,}\\ \text{obserwowane}}}, \overbrace{\beta}^{\text{szukane}}) + \underbrace{\varepsilon_i}_{\substack{\text{losowe,}\\ \text{nieobserwowane}}}, \tag{1.3.1}$$

gdzie y_i oznacza wartość zmiennej objaśnianej dla i-tej obserwacji; x_i jest wartością zmiennej objaśniającej dla i-tej obserwacji; $\beta = (\beta_1, \ldots, \beta_p)$ to wektor p nieznanych parametrów; f jest funkcją nieliniową ze względu na co najmniej jeden z parametrów $\beta_j, j \in (1, \ldots, p)$; ε_i jest losowym błędem pomiaru i-tej zmiennej y_i.

Uwaga 1.3.1. Zakłada się, że osoba analizująca dane dysponuje wiedzą, opartą na teoretycznych lub empirycznych rozważaniach na temat danego zjawiska, a w związku z tym mniej więcej wie jakiego rodzaju nieliniowej funkcji f należy użyć.

Dla tak sformułowanego modelu (1.3.1) przyjmijmy następujące założenia:

Założenie 1.3.1. Mamy $p < n$ oraz zmienne $x_1, \ldots, x_n$ są niezależne.

Rozpatrujemy analogicznie jak w przypadku modelu liniowego sytuację, kiedy mamy więcej obserwacji niż parametrów do westymowania. Ponadto chcemy, żeby zmienne objaśniające były niezależne.

Założenie 1.3.2. Błędy losowe $\varepsilon_1, \ldots, \varepsilon_n$ mają rozkład $\mathcal{N}(0, \sigma^2)$ i są niezależne.

1.3.2. Estymacja parametrów

Wyznaczywszy postać ogólną modelu nieliniowego możemy przejść do znalezienia oszacowań nieznanych parametrów. Oprócz współczynnika β, który jest jawnie przedstawiony w równaniu (1.3.1) musimy zgodnie z założeniem 1.3.2 wyznaczyć estymator parametru σ^2. W tym podrozdziale postaram się ogólnie przedstawić ideologię stojącą za znajdowaniem oszacowań tych parametrów. Więcej szczegółów na ten temat można znaleźć w pracy [1].

Parametr β

W celu wyznaczenia oszacowania parametru β wykorzystamy metodę najmniejszych kwadratów. Zgodnie z definicją 1.1.1

$$\widehat{\beta} = \min_{\beta} \mathrm{RSS}(\beta) = \min_{\beta} \sum_{i=1}^{n} (y_i - f(x_i, \beta))^2. \tag{1.3.2}$$

Zauważmy jednak, że szukanie powyższego minimum sprowadza się do rozwiązania układu p równań nieliniowych. Układ ten tworzą pochodne cząstkowe funkcji $\mathrm{RSS}(\beta)$ po każdym z p parametrów. Niestety nie istnieje ogólna analityczna metoda prowadząca do znalezienia rozwiązań tego typu układów równań. Dlatego też stosuje się numeryczne metody iteracyjne takie jak metoda Newtona (źródło: [8]), przy pomocy których możemy znaleźć przybliżone rozwiązania. W każdym kroku, w oparciu o zbiór danych, model oraz aktualne wartości parametrów, algorytm Newtona wyznacza nowe parametry. Procedura ta jest powtarzana aż do momentu uzyskania parametrów, które będą dostatecznie blisko minimum. Stosując metody numeryczne powinniśmy szczególnie uważać na następujące problemy:

- Jak wybrać dobre wartości początkowe?
- Jak zapewnić, że w wyniku działania procedury otrzymaliśmy globalne minimum, a nie lokalne?

Oczywiście te dwie rzeczy są ze sobą mocno powiązane. Jeśli bowiem wybierzemy parametry początkowe tak, żeby były dostatecznie blisko optymalnych, to algorytm będzie zbieżny po zaledwie kilku krokach. Jeśli natomiast źle wybierzemy parametry początkowe, to może się okazać, że uzyskamy jedynie minimum lokalne, a w konsekwencji model z nowymi parametrami nie będzie dobrze dopasowany do danych. Może się również okazać, że algorytm nie zbiega bez względu na to jakie parametry początkowe wybierzemy. Wtedy też jedynym słusznym rozwiązaniem jest wybranie nowego, prostszego modelu. Widzimy więc, że wybór parametów początkowych jest bardzo ważnym aspektem, któremu poświęcimy oddzielny podrozdział.

Parametr σ^2

W momencie kiedy mamy już odpowiednio dokładnie oszacowany parametr β możemy z łatwością wyznaczyć nieobciążony estymator wariancji σ^2,

$$\widehat{\sigma^2} = \frac{RSS(\widehat{\beta})}{n-p}. \tag{1.3.3}$$

Rozdział 2

Wykorzystanie funkcji `nls` w modelach nieliniowych

W ostatnim podzrozdziale wstępu teoretycznego przedstawiliśmy krótko teorię dotyczącą modeli nielinowych. Zaznaczyliśmy, że znajdowanie oszacowań parametrów jest niemożliwe przy pomocy metod analitycznych. Najczęściej wykorzystuje się iteracyjne metody numeryczne, a więc niezbędne jest wykorzystanie możliwości obliczeniowych komputera. W pakiecie R dostępna jest funkcja `nls`, która z powodzeniem może być wykorzystana do estymacji parametrów w modelu nieliniowym. W tym rozdziale postaram się omówić najważniejsze własności funkcji `nls`, które mogą być pomocne przy dopasowywaniu modeli nieliniowych do danych.

2.1. Wybór wartości początkowych

Przejdziemy teraz do dokładniejszej analizy wyboru wartości początkowych, która, jak już wspomniałem w podrozdziale 1.3.2, jest bardzo ważnym czynnikiem koniecznym do zbieżności algorytmu. Zanim zaczniemy wybierać wartości początkowe musimy, zgodnie z uwagą 1.3.1, mieć wiedzę, która pozwoli nam na wybranie odpowiedniej funkcji nieliniowej f. Od tej pory zakładamy zatem, że określiliśmy już funkcję f i do pełni szczęścia potrzebne nam są jedynie początkowe wartości współczynników.

Postaram się teraz przedstawić najpopularniejsze metody jakie stosuje się do wyboru wartości początkowych. Więcej szczegółów na ten temat wraz z dokładnie omówionymi przykładmi dla każdej z metod możemy znaleźć w pracy [12].

Wybieranie wartości początkowych *„na oko”*

Metoda ta, jak sama nazwa wskazuje, nie jest zbyt precyzyjnie określona. Wymaga ona sporych umiejętności i doświadczenia z poprzednich analiz. Jest ona, wbrew pozorom, najbardziej powszechną metodą wśród osób zajmujących się modelami nieliniowymi.

Graficzne poszukiwanie wartości początkowych

Zdarza się, że część z parametrów modelu ma pewną interpretację (biologiczną, chemiczną czy też fizyczną) i wtedy też łatwo możemy znaleźć ich początkowe wartości. Pozostałe parametry wybieramy metodą „na oko” i sprawdzamy jak dobrze nasz wykres przybliża dane. Jeśli okaże się, że wykres funkcji daje złe przybliżenie to próbujemy inny zetaw parametrów. W celu ułatwienia wyszukiwania odpowiednich parametrów możemy skorzystać z metody przeszukiwania po siatce.

Przeszukiwanie parametrów po siatce

Bardzo ciekawą metodą pozwalającą na znalezienie odpowiednich parametrów początkowych jest przeszukiwanie po siatce. Jest ona szczególnie przydatna kiedy wiemy mniej więcej z jakiego zakresu pochodzą nieznane parametry. Metoda ta polega na tym, że dla każdego z p parametrów podajemy zestaw wartości, które wydają nam się dobrymi wartościami początkowymi. Następnie, przy użyciu funkcji `expand.grid()` z pakietu R, tworzona jest tzw. siatka parametrów o p kolumnach i liczbie wierszy równej liczbie wszystkich możliwych kombinacji podanych wartości parametrów. Mówiąc ściślej, jeśli dysponujemy: n_1 wartościami parametru β_1,..., n_p wartościami parametru β_p to łącznie siatka będzie miała p kolumn oraz $n_1 \cdot \ldots \cdot n_p$ wierszy. Mając odpowiednio zaprojektowaną siatkę parametrów używamy funkcji `nls2` z pakietu R, żeby znaleźć zestaw p parametrów początkowych o najmniejszej wartości RSS(β).

Funkcje samostartujące

Jeśli nie mamy zupełnie pomysłu jakich parametrów początkowych użyć, to z pomocą może nam przyjść użycie funkcji samostartujących. Ich ogromną zaletą jest fakt, że wymagają one od nas podania jedynie zbioru danych z którego korzystamy, a funkcja sama wyznaczy wartości parametrów początkowych. Wadą może być to, że zbiór gotowych funkcji samostartujących jest mocno ograniczony. Ich przykłady możemy znaleźć w dodatku B pracy [12]. W momecie kiedy żadna z podanych funkcji nie odpowiada naszemu modelowi możemy sami stworzyć odpowiednią funkcji samostartującą, co potwierdza nieograniczony potencjał tej metody.

2.2. Kontrola algorytmu

Wiemy już jak wybrać dobre wartości początkowe, żeby algorytm był odpowiednio zbieżny. Warto również wspomnieć, jak możemy sami kontrolować tę zbieżność. Funkcja `nls` jest wyposażona w argument `control`, który pozwala na dokładne kontrolowanie procesu estymacji współczynników. Ponieważ zdarza się, że chcemy monitorować kilka rzeczy na raz, wygodnie jest użyć w tym celu funkcji `nls.control()`.

Postaram się teraz wymienić i opisać krótko najważniejsze argumenty funkcji `nls.control()`:

1. `maxiter` określa ile maksymalnie iteracji może zostać wykonanych przez algorytm, domyślna wartość to 50.

2. `tol` precyzuje, jaki jest poziom tolerancji zbieżności. Dokładniej można o tym kryterium poczytać w pracy ([1], str. 49-50), domyślna wartość to 0.00001.

3. `minFactor` to czynnik, który określa najmniejszą możliwą rożnicę pomiędzy wartościami parametrów uzyskanymi w kolejnych iteracjach, domyślnie wynosi on 1/1024.

4. `printEval` to argument logiczny wskazujący, czy chcemy, żeby informacje o pochodnych były raportowane w kolejnych iteracjach.

5. `warnOnly` to argument logiczny wskazujący, czy chcemy zobaczyć jak przebiegały kolejne kroki algorytmu do momentu wystąpienia jakiegoś błędu.

2.3. Podsumowanie algorytmu

Bardzo częstym sposobem służącym do porównywania różnych modeli na tym samym zbiorze danych jest wyznaczenie wartości RSS dla poszczególnych modeli i uszeregowanie ich według malejącej wartości. Ostateczną wartość RSS dla danego modelu możemy otrzymać, stosując metodę `deviance()` podając jako jej argument obiekt klasy `nls`.

2.4. Predykcja

W momencie kiedy mamy wyznaczone już oszacowania parametrów modelu (1.3.1) możemy z łatwością wyznaczyć predykcje zmiennej objaśnianej y. Funkcja `nls` z pakietu statystycznego R jest wyposażona w specjalne metody:

- `fitted` pozwala wyznaczyć predykcje, czyli wartości funkcji f dla zadanych w modelu zmiennych objaśniających x oraz wysetymowanych parametrów β,
- `predict` pozwala wyznaczyć predykcje, czyli wartości funkcji f na zbiorze zmiennych objaśniających wprowadzonych przez użytkownika oraz wyestymowanych parametrach β.

Może się zdarzyć, że naszą zmienną objaśniającą jest czas. Chcemy wówczas sprawdzić jak zachowują się funkcje nieliniowe z optymalnymi parametrami w odległym odstępie czasowym. Po uprzednim zdefiniowaniu naszej funkcji nieliniowej i znalezieniu oszacowań parametrów wystarczy, że użyjemy funkcji `curve()` na dowolnie wybranym przedziale czasu. Uzyskamy wówczas wykres tej funkcji i możemy próbować wyciągać wnioski na temat jej zachowania.

Rozdział 3

Analiza danych rzeczywistych

3.1. Opis zbioru danych

W niniejszym rozdziale dokonam analizy rozwoju rekordów świata w ośmiu lekkoatletycznych dyscyplinach biegowych na przestrzeni 100 lat.

Tabela 3.1 przedstawia rekordy odnotowane i zatwierdzone przez IAAF (ang. *International Association of Athletics Federation*) po mistrzostwach świata w Daegu w 2011 roku (źródło: [7]). Część danych, dotycząca biegów na 200 m od 1912 do 1924 roku, została pobrana z wikipedii ([17],[18],[19]). Każdy wiersz zawiera rok olimpijski oraz rekordowe czasy (w sekundach) na ośmiu różnych dystansach (w metrach): 100, 200, 400, 800, 1500, 5000, 10000 , 42195. W latach 1916, 1940 i 1944 olimpiady nie zostały rozegrane z powodu I i II wojny światowej. Mimo tego niektóre rekordy zostały poprawione. Przykładem może być bieg na 10000 m, kiedy to rekord z 1944 roku został poprawiony o ponad 17 s w stosunku do najlepszego wyniku z 1940 roku.

Większość czasów umieszczonych w tabeli 3.1 podanych jest z dokładnością do 0.1 s. Jest to efekt ułomności techniki, która przez długi czas nie pozwalała na dokładniejsze pomiary. Dopiero w 1968 roku na olimpiadzie w Meksyku odnotowany i zarazem ratyfikowany został przez IAAF pierwszy automatyczny pomiar czasu, który dawał dokładność do 0.01 s.
Oficjalnie elektroniczny pomiar czasu został zaakceptowany przez IAAF w 1975 roku. Zmiana dotyczyła początkowo tylko biegów do 400 m. Jednak już w 1981 roku zaakceptowanto go również dla biegów do 10000 m.

Warto również odnotować, że na dystansach 200, 400 i 800 m początkowe rekordy zostały pobite na nieco dłuższych dystansach. Było to odpowiednio 220 jardów (201.17 m), 440 jardów (402.34 m) i 880 jardów (804.68 m). Pomimo tej nieznacznej różnicy w długości dystansu czasy uzyskane przez sportowców zostały oficjalnie uznane przez IAAF za rekordowe.

Tabela 3.1: Rekordy świata w latach olimpijskich, źródło: [7],[17],[18],[19].

Rok	Rekordowe czasy (s) na dystansie:							
	100m	200m	400m	800m	1500m	5000m	10000m	42195m
1912	10.6	21.3	47.8[c]	111.9[d]	235.8	876.6	1858.8	9634.2
1916[a]	10.6	21.2[b]	47.4[c]	111.9[d]	235.8	876.6	1858.8	9366.6
1920	10.6	21.2[b]	47.4[c]	111.9[d]	234.7	876.6	1858.8	9155.8
1924	10.4	21.2[b]	47.4[c]	111.9[d]	232.6	868.2	1806.2	9155.8
1928	10.4	21.2[b]	47.0	110.6	231.0	868.2	1806.2	8941.8
1932	10.3	21.1	46.2	109.8	229.2	857.0	1806.2	8941.8
1936	10.2	20.7	46.1	109.7	227.8	857.0	1806.2	8802.0
1940[a]	10.2	20.7	46.0	106.6	227.8	848.8	1792.6	8802.0
1944[a]	10.2	20.7	46.0	106.6	223.0	838.2	1775.4	8802.0
1948	10.2	20.7	45.9	106.6	223.0	838.2	1775.4	8739.0
1952	10.2	20.6	45.8	106.6	223.0	838.2	1742.6	8442.2
1956	10.1	20.6	45.2	105.7	220.6	816.8	1722.8	8259.4
1960	10.0	20.5	44.9	105.7	215.6	815.0	1698.8	8116.2
1964	10.0	20.2[b]	44.9	104.3	215.6	815.0	1695.6	7931.2
1968	9.95	19.83	43.86	104.3	213.1	796.4	1659.4	7776.4
1972	9.95	19.83	43.86	104.3	213.1	793.0	1658.4	7713.6
1976	9.95	19.83	43.86	103.5	212.2	793.0	1650.8	7713.6
1980	9.95	19.72	43.86	102.33	211.36	788.4	1642.4	7713.6
1984	9.93	19.72	43.86	101.73	210.77	780.41	1633.81	7685.00
1988	9.92	19.72	43.29	101.73	209.46	778.39	1633.81	7610.00
1992	9.86	19.72	43.29	101.73	208.86	778.39	1628.23	7610.00
1996	9.84	19.32	43.29	101.73	207.37	764.39	1598.08	7610.00
2000	9.79	19.32	43.18	101.11	206.00	759.36	1582.75	7542.00
2004	9.79	19.32	43.18	101.11	206.00	757.35	1580.31	7495.00
2008	9.69	19.30	43.18	101.11	206.00	757.35	1577.53	7439.00
2012	9.58	19.19	43.18	101.01	206.00	757.35	1577.53	7382.00

[a] W 1916, 1940 i 1944 Olimpiady nie zostały rozegrane z powodu I i II wojny światowej.
[b] Biegi odbywały się na nieoficjalnym dystansie 220 jardów (201.17 m).
[c] Biegi odbywały się na nieoficjalnym dystansie 440 jardów (402.34 m).
[d] Biegi odbywały się na nieoficjalnym dystansie 880 jardów (804.68 m).

3.2. Modelowanie występów lekkoatletów

Głównym celem tego rozdziału jest znalezienie modelu, który będzie jak najdokładniej opisywał dane z tabeli 3.1, przedstawiające rekordy lekkoatletów w ośmiu różnych konkurencjach biegowych uzyskane w ciągu 100 lat.

3.2.1. Czas a dystans

Po wstępnym przeanalizowaniu danych z tabeli 3.1 nietrudno jest zauważyć, że dwukrotne pokonanie tego samego dystansu zajmuje biegaczom ponad dwa razy więcej czasu. Jest to efekt fizjologicznych ograniczeń możliwości wydolnościowych człowieka.

W celu potwierdzenia tej analizy, w tabeli 3.2, wyznaczona została zależność T/D pomiędzy czasem a dystansem dla każdych dwóch kolejnych badanych przeze mnie biegów. Ograniczyłem się do danych z lat 1912 i 2012, chcąc sprawdzić czy są jakieś istotne różnice dla poszczególnych zależności w tak dużym odstępie czasowym. Widzimy dokładnie, że wstępne przypuszczenia okazały się słuszne zarówno dla 1912 jak i 2012 roku. Co ciekawe różnice w wartościach poszczególnych zależności T/D różnią się niewiele dla tych lat.

Interesująca wydaje się być wartość współczynnika T/D porównująca 100 i 200 m. Okazuje się, że pokonanie dystansu 200 m zajmuje najlepszym biegaczom niemal dokładnie dwa razy więcej czasu niż pokonanie 100 m. Przyjrzyjmy się zatem dokładnie analizie rekordowych biegów Usaina Bolta na 100 i 200 m podczas Lekkoatletycznych Mistrzostw Świata w Berlinie w 2009 roku (źródło:[15], [16]). Reakcja startowa powoduje, że pierwsze 100 m w biegu na 200 m jest pokonywane w dłuższym czasie niż drugie. Niemniej, druga setka jest pokonywana szybciej niż sam bieg na 100 m. W efekcie średnia prędkość jest niemalże taka sama dla obu dystansów, a w konsekwencji wartość współcznnika T/D $\simeq 1$. Niestety przebiegnięcie kolejnych 200 m, ze średnią prędkością taką jak w przypadku 100 i 200 m, nie jest już możliwe ponieważ następuje efekt zmęczenia biegaczy. Stąd też współczynnik T/D jest większy niż jeden, gdy porównujemy dłuższe dystanse.

3.2.2. Model potęgowy

Analiza przeprowadzona w poprzednim rodziale pokazała, że rozsądne wydaje się być opisanie danych z tabeli 3.1, dotyczących danego roku, olimpijskiego przy pomocy modelu potęgowego.

Opis modelu

Przyjmijmy, że pierwsza kolumna tabeli 3.1 jest tylko opisem dla kolejnych wierszy. Niech i-ty wiersz oznacza dany rok olimpijski ($i \in \{1, ..., n\}$, n – liczba olimpiad). Z kolei j-ta kolumna (pomijając pierwszą) dotyczy kolejnych dystansów ($j \in \{1, ..., m\}$, m – liczba dystansów). Przyjmując następujące oznaczenia:

d_j – j-ty dystans,

t_{ij} – rekordowy czas uzyskany w i-tym roku olimpijskim podczas biegu na j-tym dystansie,

α_i, β_i – współczyniki modelu potęgowego dla i-tego roku olimpijskiego,

η_{ij} – składnik losowy o rozkładzie $\eta_{ij} \backsim \ln\mathcal{N}(0, \sigma^2)$, dla ustalonego i η_{ij} są niezależne,

dostaniemy nastepujące równanie modelu potęgowego dla i-tego roku olimpijskiego

$$t_{ij} = e^{\alpha_i} d_j^{\beta_i} \eta_{ij}. \tag{3.2.1}$$

Tabela 3.2: Zależność (T/D) pomiędzy stosunkiem dwóch kolejnych czasów, a stosunkiem dwóch kolejnych dystansów. Opracowanie własne.

1912 r					
i	dystans (m)	$D = \frac{\text{dystans}(i)}{\text{dystans}(i-1)}$	czas (s)	$T = \frac{\text{czas}(i)}{\text{czas}(i-1)}$	**T/D**
1	100		10.6		
2	200	2.000	21.2	2.000	**1.000**
3	400	2.000	47.4	2.236	**1.118**
4	800	2.000	111.9	2.361	**1.180**
5	1500	1.875	235.8	2.107	**1.124**
6	5000	3.333	876.6	3.718	**1.115**
7	10000	2.000	1858.8	2.120	**1.060**
8	42195	4.220	9366.6	5.039	**1.194**
2012 r					
i	dystans (m)	$D = \frac{\text{dystans}(i)}{\text{dystans}(i-1)}$	czas (s)	$T = \frac{\text{czas}(i)}{\text{czas}(i-1)}$	**T/D**
1	100		9.58		
2	200	2.000	19.19	2.003	**1.002**
3	400	2.000	43.18	2.250	**1.125**
4	800	2.000	101.01	2.339	**1.170**
5	1500	1.875	206.00	2.039	**1.088**
6	5000	3.333	757.35	3.676	**1.103**
7	10000	2.000	1577.53	2.083	**1.041**
8	42195	4.220	7382	4.679	**1.109**

Transformacja modelu

W celu uproszczenia struktury modelu zlogarytmujmy stronami równanie (3.2.1). W wyniku tego przekształcenia otrzymamy model liniowy

$$\ln(t_{ij}) = \alpha_i + \beta_i \ln(d_j) + \ln(\eta_{ij}).$$

Chcąc pozbyć się logarytmów w oznaczeniach, podstawmy

$$T_{ij} := \ln(t_{ij}), D_j := \ln(d_j), \ln(\eta_{ij}) := \xi_{ij}.$$

Dostaniemy wówczas

$$T_{ij} = \alpha_i + \beta_i D_j + \xi_{ij}. \tag{3.2.2}$$

W ten sposób, w wyniku prostych transformacji, otrzymujemy dla każdego i-tego roku olimpijskiego model regresji prostej, w którym:

- zmienną objaśniającą jest D_j, czyli zlogarytmowany j-ty dystans d_j,
- zmienna objaśniana to T_{ij}, czyli zlogarytmowany czas t_{ij},
- α_i, β_i współczynniki modelu regresji prostej dla i-tego roku olimpijskiego,
- ξ_{ij} - składnik losowy o rozkładzie $\xi_{ij} \backsim \mathcal{N}(0, \sigma^2)$, dla ustalonego i ξ_{ij} są niezależne.

Postać ogólna modelu

Zapisując równania (3.2.2) w kolejnych wierszach, dostajemy następujące równanie macierzowe

$$\begin{bmatrix} T_{1,1} \\ \vdots \\ T_{1,m} \\ T_{2,1} \\ \vdots \\ T_{2,m} \\ \vdots \\ T_{n,1} \\ \vdots \\ T_{n,m} \end{bmatrix} = \begin{bmatrix} 1 & D_1 & & & & & & \\ \vdots & \vdots & & & & & 0 & \\ 1 & D_m & & & & & & \\ & & 1 & D_1 & & & & \\ & & \vdots & \vdots & & & & \\ & & 1 & D_m & & & & \\ & & & & \ddots & & & \\ & & & & & 1 & D_1 & \\ & 0 & & & & \vdots & \vdots & \\ & & & & & 1 & D_m & \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \\ \vdots \\ \alpha_n \\ \beta_n \end{bmatrix} + \begin{bmatrix} \xi_{1,1} \\ \vdots \\ \xi_{1,m} \\ \xi_{2,1} \\ \vdots \\ \xi_{2,m} \\ \vdots \\ \xi_{n,1} \\ \vdots \\ \xi_{n,m} \end{bmatrix}$$

$$T = D\theta_1 + \xi, \tag{3.2.3}$$

gdzie T jest wektorem o długośći nm, macierz D jest wymiaru $(nm) \times (2n)$, θ_1 jest wektorem o długości $2n$, a $\xi \backsim \mathcal{N}(0, \sigma^2 I_{nm \times nm})$. Tym samym dostaliśmy model liniowy realizujący wszystkie nm obserwacji, w którym poszczególne podmodele (3.2.2) są wyznaczone przez kolejne bloki macierzy D.

Oszacowanie wpółczynników α_i oraz β_i

Do wyznaczenia ocen $\widehat{\alpha_i}, \widehat{\beta_i}$ użyłem oprogramowania statystycznego R. Opis wraz z listingiem umieszczone zostały w dodatku B.1. Najważniejsze szczegóły, przedstawiające wyniki jakie uzyskałem podczas wyznaczania ocen, przedstawione są w tabeli 3.3.

Zgodność dopasowania podmodeli

Jednym z kryteriów służących do oceny modelu jest współczynnik determinacji R^2. Analizując wartości tego współczynnika, umieszczone w tabeli 3.3 widzimy, że są one bliskie jeden. Najmniejszy z nich ma wartość 0.9995 i dotyczy roku 1972. Świadczy to o tym, że niemalże 100% zmienności zmiennej objaśnianej jest wyjaśnione przez zmienność zmiennej objaśniającej. Wynika stąd również, że zmienne D_j oraz T_{ij} są ze sobą mocno powiązane w każdym z podmodeli (3.2.2).

Możemy również potwierdzić dokładność dopasowania krzywych regresji w podmodelach (3.2.2), porównując zmienne objaśniane T_{ij} z ich predykcjami $\widehat{T_{ij}}$, wyznaczonymi przez proste trendu. Odpowiednie porównanie wykonałem dla roku 1972, o najgorszym współczynniku determinacji, a zarazem największej sumie kwadratów reszt (ang. *RSS*). Wyniki przedstawione zostały w tabeli 3.4.

Tabela 3.3: Oceny współczynników α i β w podmodelach regresji dla poszczególnych lat olimpijskich. W kolejnych kolumnach znajdują się: rok, ocena $\widehat{\alpha_i}$, błąd oceny $\widehat{\alpha_i}$, ocena $\widehat{\beta}_i$, błąd oceny $\widehat{\beta}_i$, współczynnik determinacji oraz suma kwadratów reszt. Opracowanie własne.

Rok	$\widehat{\alpha_i}$	$\sigma_{\widehat{\alpha_i}}$	$\widehat{\beta}_i$	$\sigma_{\widehat{\beta}_i}$	R^2	RSS
1912	-2.8883	0.0675	1.1335	0.00896	0.99971	0.01102
1916	-2.8751	0.0675	1.131	0.00896	0.99965	0.01326
1920	-2.8598	0.0675	1.1284	0.00896	0.99963	0.01403
1924	-2.864	0.0675	1.1279	0.00896	0.99967	0.01258
1928	-2.8552	0.0675	1.1258	0.00896	0.99966	0.01287
1932	-2.8738	0.0675	1.1273	0.00896	0.99967	0.0123
1936	-2.884	0.0675	1.1278	0.00896	0.99961	0.01472
1940	-2.8885	0.0675	1.1276	0.00896	0.99967	0.01259
1944	-2.8854	0.0675	1.1264	0.00896	0.99975	0.009393
1948	-2.8813	0.0675	1.1257	0.00896	0.99974	0.00998
1952	-2.8546	0.0675	1.121	0.00896	0.99964	0.01349
1956	-2.8478	0.0675	1.1183	0.00896	0.99966	0.0126
1960	-2.8467	0.0675	1.1168	0.00896	0.99967	0.01221
1964	-2.8425	0.0675	1.1153	0.00896	0.99959	0.01522
1968	-2.8441	0.0675	1.1134	0.00896	0.99951	0.018
1972	-2.8375	0.0675	1.1123	0.00896	0.9995	0.01857
1976	-2.8385	0.0675	1.1122	0.00896	0.99953	0.01727
1980	-2.8426	0.0675	1.1122	0.00896	0.99956	0.0161
1984	-2.8399	0.0675	1.1113	0.00896	0.9996	0.01485
1988	-2.8399	0.0675	1.1107	0.00896	0.99957	0.01594
1992	-2.8438	0.0675	1.1111	0.00896	0.99958	0.01534
1996	-2.849	0.0675	1.1106	0.00896	0.99957	0.01593
2000	-2.8458	0.0675	1.1094	0.00896	0.99959	0.015
2004	-2.8404	0.0675	1.1085	0.00896	0.99958	0.01539
2008	-2.8432	0.0675	1.1085	0.00896	0.99956	0.01596
2012	-2.8505	0.0675	1.1091	0.00896	0.99953	0.01732
$\frac{1}{n}\sum_{i=1}^{n}\widehat{\alpha_i}$	-2.8562	$\frac{1}{n}\sum_{i=1}^{n}\widehat{\beta}_i$	1.1185		$\sum_{i=1}^{n} RSS_i$	0.37195

Tabela 3.4: Porównanie obserwowanych i przewidywanych zlogarytmowanych czasow dla 1972 roku. Opracowanie własne.

bieg(m)	ln(czasu (s))		reszta
	obserwowanego	przewidywanego	
100	2.298	2.285	0.013
200	2.987	3.056	-0.069
400	3.781	3.827	-0.046
800	4.647	4.598	0.049
1500	5.362	5.297	0.065
5000	6.676	6.636	0.04
10000	7.414	7.407	0.0063
42195	8.951	9.009	-0.058

Interpretacja ocen $\widehat{\alpha_i}$ oraz $\widehat{\beta}_i$

Chcąc dokładniej przyjrzeć się wartościom ocen $\widehat{\alpha_i}$ oraz $\widehat{\beta}_i$ (tabela 3.3) wykonałem wykresy (rysunek 3.1a, 3.1b) przedstawiające zmiany oszacowań tych parametrów w kolejnych latach olimpijskich. Dodatkowo dla każdego i wyznaczyłem przedziały: $[\widehat{\alpha_i} - \sigma_{\widehat{\alpha_i}}, \widehat{\alpha_i} + \sigma_{\widehat{\alpha_i}}]$, $[\widehat{\beta}_i - \sigma_{\widehat{\beta}_i}, \widehat{\beta}_i + \sigma_{\widehat{\beta}_i}]$. Kod w programie R umożliwiający narysowanie tych wykresów umieszczony został w dodatku B.1.1.

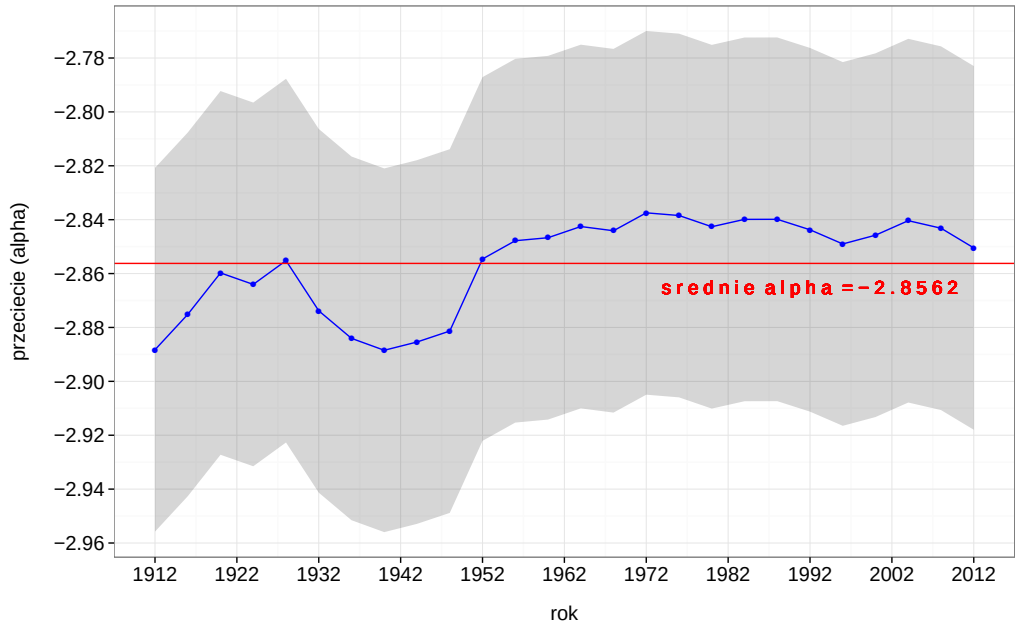
Wartości $\widehat{\alpha_i}$ wahają się pomiędzy -2.8885 a -2.8375 natomiast wartości $\widehat{\beta}_i$ należą do przedziału [1.1085, 1.1335]. Zauważmy, że, po ustaleniu współczynnika $\widehat{\beta}_i$, niewielkie zmiany współczynnika $\widehat{\alpha_i}$ nie wpłyną mocno na krzywą regresji. Natomiast gdybyśmy ustalili $\widehat{\alpha_i}$, to małe wahania współczynnika $\widehat{\beta}_i$ powodują, że prosta regresji zmieni się znacząco. Zwróćmy również uwagę, że średnia wartość ocen parametru α wynosi -2.8562 i całkowicie mieści się we wszystkich przedziałach $[\widehat{\alpha_i} - \sigma_{\widehat{\alpha_i}}, \widehat{\alpha_i} + \sigma_{\widehat{\alpha_i}}]$, w przeciwieństwie do średniego nachylenia. Dlatego też warto, być może, zastąpić oceny $\widehat{\alpha_i}$ przez ich średnią, a następnie estymować jedynie parametr nachylenia.

Na wykresie przedstawiającym zależność rok-nachylenie widoczne są dwie interesujące zmiany trendu malejącego:

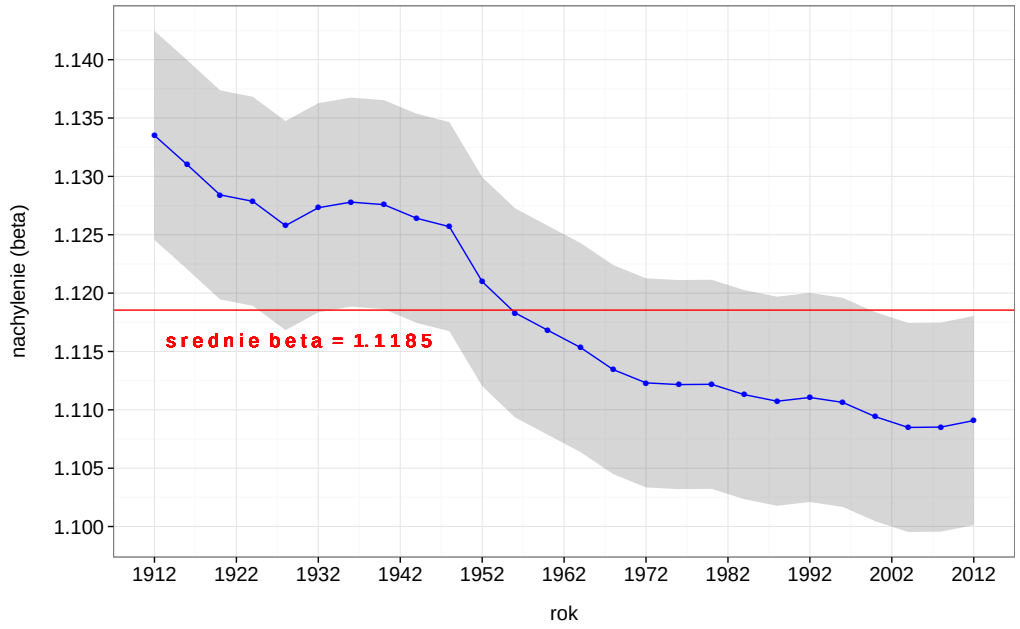
- Pierwsza z nich jest bardzo wyraźna i dotyczy lat 1928-1948. Widzimy, że w tym okresie poziom nachylenia nie spadł poniżej wartości 1.1258 z 1928 roku. Wskazuje to na stagnację w rozwoju rekordów świata na rozpatrywanych przeze mnie dystansach. Lata 1928-1948 obejmują okres II wojny światowej, kiedy nastąpił zdecydowany spadek w aktywności lekkoatletycznej. Jest to najprawdopodobniej główna przyczyna tego zastoju.

- Druga zmiana jest dosyć subtelna, ale jednak widoczna. Przypada ona na lata 1988-1996, kiedy ponownie nastąpiło zatrzymanie się rozwoju rekordów. W roku 1988 rozegrana została olimpiada w Seulu. Według zeznań ponad 100 świadków wiele spośród pobitych rekordów w lekkiej atletyce zostało uzyskanych przy pomocy środków dopingujących (źródło [14]). Dlatego też wiele z rekordów uzyskanych przez lekkoatletów podczas tej olimpiady nie zostało pobitych przez nastepne lata.

(a)



(b)



Rysunek 3.1: Wykresy przedstawiające oceny parametrów α oraz β dla kolejnych podmodeli regresji tzn. dla kolejnych lat olimpijskich: (3.1a) zależność pomiędzy rokiem a przecięciem α, (3.1b) zależność pomiędzy rokiem a nachyleniem β. Czerwona linia oznacza średnią wartość ocen danego parametru. Na szaro zaznaczone zostały przedziały: $[\widehat{\alpha_i} - \sigma_{\widehat{\alpha_i}}, \widehat{\alpha_i} + \sigma_{\widehat{\alpha_i}}]$, $[\widehat{\beta_i} - \sigma_{\widehat{\beta_i}}, \widehat{\beta_i} + \sigma_{\widehat{\beta_i}}]$. Opracowanie własne.

3.2.3. Model alternatywny

W tym rozdziale dokonam modyfikacjii modelu (3.2.3) opierając się na wniosku z poprzednich rozważań.

Opis modelu

Rysunek 3.1a pokazuje, że średnia wartość oceny parametru nachylenia mieści się całkowicie we wszystkich przedziałach $[\widehat{\alpha_i} - \sigma_{\widehat{\alpha_i}}, \widehat{\alpha_i} + \sigma_{\widehat{\alpha_i}}]$. Możemy zatem stwierdzić, że

$$\alpha_i := \frac{1}{n}\sum_{i=1}^{n} \widehat{\alpha_i} = -2.8562 = A \qquad \text{dla każdego } i \in \{1, ..., n\}.$$

Załóżmy również, że $\beta_i := \gamma_i$ dla każdego $i \in \{1, ..., n\}$. Wówczas model (3.2.2) przyjmie następującą postać

$$T_{ij} = A + \gamma_i D_j + \xi_{ij}. \tag{3.2.4}$$

Tym samym otrzymaliśmy i-ty model regresji prostej, gdzie nieznanym parametrem, który będzie oceniany jest jedynie współczynnik nachylenia γ_i.

Postać ogólna modelu

Zapisując równania (3.2.4) w postaci macierzowej dostajemy

$$\begin{bmatrix} T_{1,1} \\ \vdots \\ T_{1,m} \\ T_{2,1} \\ \vdots \\ T_{2,m} \\ \vdots \\ T_{n,1} \\ \vdots \\ T_{n,m} \end{bmatrix} = \begin{bmatrix} 1 & D_1 & & & & & & \\ \vdots & \vdots & & & & 0 & & \\ 1 & D_m & & & & & & \\ & & 1 & D_1 & & & & \\ & & \vdots & \vdots & & & & \\ & & 1 & D_m & & & & \\ & & & & \ddots & & & \\ & & & & & 1 & D_1 \\ & 0 & & & & \vdots & \vdots \\ & & & & & 1 & D_m \end{bmatrix} \begin{bmatrix} A \\ \gamma_1 \\ A \\ \gamma_2 \\ \vdots \\ A \\ \gamma_n \end{bmatrix} + \begin{bmatrix} \xi_{1,1} \\ \vdots \\ \xi_{1,m} \\ \xi_{2,1} \\ \vdots \\ \xi_{2,m} \\ \vdots \\ \xi_{n,1} \\ \vdots \\ \xi_{n,m} \end{bmatrix}$$

$$T = D\theta_2 + \xi, \tag{3.2.5}$$

gdzie macierz D jest wymiaru $(nm) \times (2n)$, θ_2 jest wektorem o długości $2n$, a $\xi \backsim \mathcal{N}(0, \sigma^2 I_{nm \times nm})$. Otrzymaliśmy zatem model liniowy realizujący wszystkie nm obserwacji. Poszczególne podmodele (3.2.4) są wyznaczone przez kolejne bloki macierzy D. Zauważmy, że wszystkie podmodele mają wspólny wyraz wolny A.

Oszacowanie współczynnika γ_i

Podobnie jak w poprzednim rozdziale, użyłem oprogramowania statystycznego R, żeby wyznaczyć oceny $\widehat{\gamma_i}$. Opis wraz z listingiem umieszczone zostały w dodatku B.2. Najważniejsze szczegóły, przedstawiające wyniki jakie uzyskałem, przedstawione są w tabeli 3.5.

Zgodność dopasowania podmodeli

Musimy teraz sprawdzić, czy modyfikacja jakiej dokonałem w modelu (3.2.4) nie spowoduje znaczących zmian w dopasowaniu i-tej krzywej trendu do danych.

Okazuje się, że współczynniki determinacji R^2 z tabeli 3.5 mają niemalże takie same wartości w porównaniu z ich odpowiednikamii z tabeli 3.3. Wnioskujemy zatem, że zmienne D_j oraz T_{ij} są ze sobą mocno skorelowane.

Sumy kwadratów reszt są rzędu 0.01, a w związku z tym dane dla poszczególnych modeli są bardzo dobrze opisane przez prostą trendu. Zwróćmy uwagę, że $\sum_{i=1}^{n} RSS_i$ wynoszą odpowiednio: 0.3719 dla modelu (3.2.3) oraz 0.376 dla modelu (3.2.5). Różnica jest niewielka, lecz wskazuje na to, że dane w poprzednim modelu byly minimalnie lepiej dopasowane. Dokładnym porównaniem obu modeli zajmiemy się jednak w następnym podrozdziale i użyjemy w tym celu bardziej specjalistycznego narzędzia.

Interpretacja oceny $\widehat{\gamma}_i$

Analizę wartości poszczególnych ocen $\widehat{\gamma}_i$ ułatwi nam rysunek 3.2. Przedstawione są na nim zmiany oszacowań tego parametru w latach kolejnych olimpiad. Kod w programie R pozwalający na narysowanie tego wykresu umieściłem w dodatku B.2.1.

Zauważmy, że wartości oszacowań parametru γ tworzą ciąg ściśle malejący. Na rysunku 3.2 widoczne są interesujące zmiany:

- Pierwsza z nich dotyczy okresu 1916-1920, kiedy to tempo obniżania się wartości współczynnika γ spadło. Jest to oczywiście związane z brakiem aktywności lekkoatletów w czasie I wojny światowej.

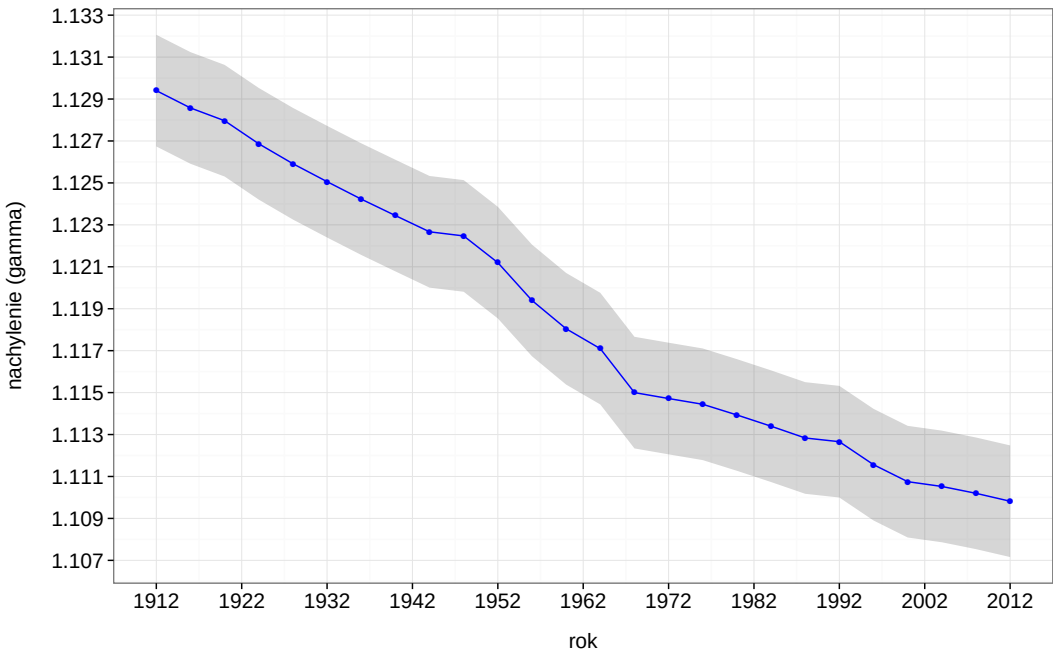
- Druga z nich przypada na okres 1944-1948. Widzimy, że wartości współczynnika γ niemalże ustabilizowały się na poziomie z roku 1944. Wynika to z niewielkiej aktywności lekkoatletów podczas rozgrywania II wojny światowej.

- Kolejna zmiana dotyczy lat 1964-1968. Tym razem łatwo możemy zauważyć, że współczynnik gamma malał zdecydowanie szybciej. Jest to najprawdopodobniej spowodowane tym, że w roku 1968 wprowadzono po raz pierwszy elektroniczny pomiar czasu, a w związku z tym czasy stały się bardziej dokładne. Ponadto w tym samym roku została rozegrana olimpiada w Meksyku, który jest położony 2240 m n.p.m. Na tej wysokości powietrze jest około pięć razy mniej gęste niż na poziomie morza (źródło [5]). Dlatego też biegającym na krótkich dystansach łatwiej było uzyskać lepsze wyniki. Spoglądając na rekordy pobite w tym roku okazuje się, że faktycznie w Meksyku pobite zostały rekordy świata na dystansach krótkich: 100, 200 i 400 m.

- Ostatnia zmiana dotyczy lat 1988-1992, kiedy to ponownie nastąpiła stagnacja w rozwoju rekordów. Jak już wcześniej wspomniałem, 1998 rok to Olimpiada w Seulu, kiedy to według zeznań świadków duża część sportowców użyła środków dopingujących. W konsekwencji wielu rekordów nie udało się pobić przez najbliższe kilka lat.

Tabela 3.5: Oceny współczynników γ w podmodelach regresji dla poszczególnych lat olimpijskich. W kolejnych kolumnach znajdują się: rok, ocena $\widehat{\gamma_i}$, błąd oceny $\widehat{\gamma_i}$, współczynnik determinacji oraz suma kwadratów reszt - opracowanie własne.

Rok	$\widehat{\gamma_i}$	$\sigma_{\widehat{\gamma_i}}$	R^2	RSS
1912	1.1294	0.00266	0.9997	0.0116
1916	1.1286	0.00266	0.99965	0.0134
1920	1.128	0.00266	0.99963	0.014
1924	1.1269	0.00266	0.99967	0.0126
1928	1.1259	0.00266	0.99966	0.0129
1932	1.1251	0.00266	0.99967	0.0125
1936	1.1242	0.00266	0.9996	0.0151
1940	1.1234	0.00266	0.99965	0.0131
1944	1.1227	0.00266	0.99974	0.00984
1948	1.1225	0.00266	0.99973	0.0103
1952	1.1212	0.00266	0.99964	0.0135
1956	1.1194	0.00266	0.99966	0.0126
1960	1.118	0.00266	0.99967	0.0123
1964	1.1171	0.00266	0.99959	0.0153
1968	1.115	0.00266	0.99951	0.0181
1972	1.1147	0.00266	0.99949	0.0188
1976	1.1144	0.00266	0.99953	0.0174
1980	1.1139	0.00266	0.99956	0.0162
1984	1.1134	0.00266	0.99959	0.015
1988	1.1128	0.00266	0.99956	0.0161
1992	1.1127	0.00266	0.99958	0.0154
1996	1.1116	0.00266	0.99956	0.016
2000	1.1108	0.00266	0.99959	0.0151
2004	1.1105	0.00266	0.99958	0.0155
2008	1.1102	0.00266	0.99956	0.0161
2012	1.1098	0.00266	0.99953	0.0173
			$\sum_{i=1}^{n} RSS_i$	0.37599



Rysunek 3.2: Wykres przedstawiający oceny parametru γ dla kolejnych podmodeli regresji tzn. dla kolejnych lat olimpijskich. Na szaro zaznaczone zostały przedziały $[\widehat{\gamma}_i - \sigma_{\widehat{\gamma}_i}, \widehat{\gamma}_i + \sigma_{\widehat{\gamma}_i}]$ dla kolejnych ocen parametru γ. Opracowanie własne.

3.3. Porówanie modeli

W celu porównania modeli (3.2.3) i (3.2.5) wykonam test ilorazu wiarogodności. Najpierw jednak postaram się wytłumaczyć, jak porównać oba modele zgodnie z teorią testowania hipotez.

Model z ograniczeniami

Będziemy chcieli sprawdzić, czy jeśli dla każdego $i \in \{1, ..., n\}$ zastąpimy w modelu (3.2.3) parametr α_i przez $A = \frac{1}{n}\sum_{i=1}^{n} \alpha_i$, to model (3.2.5) nie będzie istotnie gorszy. Oczywiście równoważne są stwierdzenia:

(a) $\forall_{k\in\{1,...,n\}} \quad \alpha_k = \frac{1}{n}\sum_{i=1}^{n} \alpha_i$.

(b) $\alpha_1 = \alpha_2 = \cdots = \alpha_n$.

Co więcej $n-1$ równości stwierdzenia (b) możemy zapisać w postaci macierzowej

$$\begin{bmatrix} 1 & 0 & -1 & & & & & & \\ & & 1 & 0 & -1 & & & 0 & \\ & 0 & & & & \ddots & & & \\ & & & & & & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \\ \vdots \\ \alpha_n \\ \beta_n \end{bmatrix} = 0$$

$$P\theta_1 = 0,$$

gdzie macierz P jest wymiaru $(n-1) \times (2n)$.

W ten sposób widzimy, że porównanie tych dwóch modeli sprowadza się do porównania (3.2.3) z modelem z ograniczeniami na parametry

$$\begin{cases} T = D\theta_1 + \xi \\ P\theta_1 = 0 \end{cases},$$

gdzie $T \backsim \mathcal{N}(D\theta_1, \sigma^2 I_{nm\times nm})$.

Test ilorazu wiarogodności

Krok 1. Będziemy chcieli testować czy dodatkowe oganiczenia na parametry wpłyną na poprawę modelu (3.2.3). Badamy zatem następującą hipotezę zerową

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_n,$$

równoważnie

$$H_0 : P\theta_1 = 0,$$

przeciw hipotezie alternatywnej

$$H_1 : \exists_{k,l} \quad \alpha_k \neq \alpha_l.$$

Krok 2. Zgodnie ze stwierdzeniem 1.2.1 statystyka testu ilorazu wiarogodności jest równa:

$$\lambda(T) = \left(\frac{\widetilde{\sigma^2}}{\widehat{\sigma^2}}\right)^{\frac{nm}{2}} = \left(\frac{\|T - D\widetilde{\theta}_2\|^2}{\|T - D\widehat{\theta}_1\|^2}\right)^{\frac{nm}{2}}, \tag{3.3.1}$$

gdzie:

$\widehat{\theta} = (\widehat{\theta}_1, \widehat{\sigma^2})$ estymatory najwiekszej wiarogodności dla modelu bez ograniczeń (3.2.3),

$\widetilde{\theta} = (\widetilde{\theta}_2, \widetilde{\sigma^2})$ estymatory najwiekszej wiarogodności dla modelu z ograniczeniami (3.2.5).

Ponadto korzystając ze stwierdzenia 1.2.2 statystyka (3.3.1) jest równoważna statystyce:

$$F(T) = \frac{(R_0 - R)/(n-1)}{R/(nm-2n)},$$

gdzie $R_0 = \|T - D\widetilde{\theta}_2\|^2$, $R = \|T - D\widehat{\theta}_1\|^2$. Dodatkowo wiemy również, że statystyka F ma rozkład $\mathcal{F}$-Snedecora (twierdzenie 1.2.1):

$$F \backsim \mathcal{F}(n-1, nm-2n).$$

Krok 3. Musimy teraz ustalić jak będzie wyglądała procedura testowa. Spójrzmy najpierw na statystykę testową (3.3.1). Jeśli hipoteza zerowa nie byłaby prawdziwa, to co najmniej dwa spośród α_i byłyby różne. To z kolei oznaczałoby, że $\|T - D\widetilde{\theta}_2\|^2$ byłaby większa niż w przypadku gdyby H_0 była prawdziwa. Oczywiście $\|T - D\widehat{\theta}_1\|^2$ pozostaje bez zmian, bez względu na to czy H_0 jest prawdziwa czy nie. Niech zatem procedura testowa δ będzie następująca:

$$\delta: \text{ Odrzucam } H_0 \text{ jeśli } F(T) \geqslant c.$$

Krok 4. Zauważmy, że w naszym przypadku wartości R_0 oraz R to $\sum_{i=1}^{n} RSS_i$ odpowiednio dla modelu (3.2.5) oraz modelu (3.2.3). Zatem $R_0 = 0.37599, R = 0.37195$. Przypomnijmy jednocześnie, że $n = 26$ oznacza liczbę olimpiad, z kolei $m = 8$ to liczba dystansów. Otrzymujemy zatem ostatecznie, że wartość obserwowana statystyki testowej wynosi:

$$F(T_{obs}) = f = \frac{(0.37599 - 0.37195)/(26-1)}{0.37195/(208-52)} = 0.068,$$

a odpowiadająca tej statystyce p-wartość:

$$p = \mathbb{P}(F \geqslant f) = 1 - 8.823841 \times 10^{-11} \approx 1.$$

Wnioskujemy stąd, że gdybyśmy przeprowadzili test na każdym poziomie istotności $\alpha_0 < p$, to nie odrzucilibyśmy hipotezy zerowej H_0.

Wniosek 3.3.1. *Wynik testu ilorazu wiarogodności pozwala nam stwierdzić, że model alternatywny (3.2.5) nie jest istotnie gorszy od modelu (3.2.3), dlatego też w dalszych rozważaniach będę wykorzystywał model (3.2.5).*

3.4. Predykcja granic możliwości lekkoatletów

W podrozdziale 3.2 udało nam się znaleźć zagnieżdżony model liniowy (3.2.5), który dobrze opisuje dane dotyczące występów lekkoatletów na ośmiu różnych dystansach na przestrzeni 100 lat. Nieznanym parametrem w tym modelu był wektor $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \ldots, \gamma_n]^T$, którego kolejne współrzędne oznaczają współczynniki nachylenia γ_i modeli (3.2.4). Przy pomocy oprogramowania statystycznego R udało nam się oszacować wartości tych n parametrów. Jednak, żeby znaleźć granice możliwości lekkoatletów, chcielibyśmy wiedzieć jakie byłyby wartości tych parametrów w kolejnych latach olimpijskich . Dlatego też w tej części pracy postaram się znaleźć krzywą regresji nieliniowej jak najlepiej dopasowaną do ocen parametrów γ_i. Następnie znajdę jej poziomą asymptotę, będącą zarazem granicą $\lim\limits_{n\to\infty} \gamma_n = \gamma_\infty$. Na koniec, wykorzystując wartość tej asymptoty, podam przewidywane granice możliwości lekkoatletów na ośmiu rozważanych dystansach.

3.4.1. Istnienie asymptoty γ_∞

Jak już wcześniej zauważyliśmy oceny parametrów γ_i tworzą malejący ciąg (rysunek 3.2). Ponadto, z tabeli 3.2, możemy wywnioskować, że nachylenie γ_i, głównie w wyniku ograniczeń wydolnościowych sportowców, nie powinno spaść poniżej wartości jeden. Trudno bowiem sobie wyobrazić w realnym świecie, żeby sportowcy byli w stanie przebiec z tą samą średnią prędkością zarówno 100 m jak i maraton. W szczególności nachylenie γ_i musi być większe od zera. Zatem oceny parametrów γ_i tworzą ciąg malejący, a w dodatku ograniczony z dołu. W związku z tym musi istnieć granica tego ciągu, którą możemy wyznaczyć poprzez znalezienie asymptoty krzywej regresji nieliniowej.

3.4.2. Modele nieliniowe

Oznaczenia

Dokonamy najpierw pewnej transformacji zmiennej „rok". Będziemy chcieli patrzeć na kolejne lata olimpijskie jak na kolejne jednostki czasu. Oznacza to, że okres czterech lat bedzie odpowiadał jednostce czasu. Nazwijmy ją „numer olimpiady". Przyjmijmy, że zaczynamy liczyć czas od pierwszej olimpiady, która dotyczy naszych danych, czyli 1912 roku.

Wprowadźmy następujące oznaczenia:

x_i – zmienna objaśniająca, czyli numer i-tej olimpiady, x_i są niezależne,
y_i – zmienna objaśniana, czyli wartość oceny parametru γ_i dla i-tej olimpiady,
$\boldsymbol{\varphi}$ – wektor nieznanych parametrów $\boldsymbol{\varphi} = (\varphi_1, ..., \varphi_p)$,
f – funkcja nieliniowa ze względu na co najmniej jedej parametr φ_i,
ε_i – składnik losowy o rozkładzie $\varepsilon_i \backsim \mathcal{N}(0, \sigma^2)$, ε_i są niezależne.

Wówczas otrzymujemy model nieliniowy nastepującej postaci

$$y_i = f(x_i, \boldsymbol{\varphi}) + \varepsilon_i. \tag{3.4.1}$$

W dalszej części pracy będę dla uproszczenia określał modele przy pomocy jedynie funkcji f, mając jednak w pamięci właściwą postać modelu (3.4.1) ze składnikiem losowym ε.

Wybór funkcji

Jak już wcześniej wspomniałem będziemy się teraz starali znaleźć krzywą regresji nieliniowej, która będzie jak najlepiej dopasowana do danych. W moich badaniach będę korzystał z doświadczeń Davida C. Blesta, który w swojej pracy *Lower bounds for athletic performance* (źródło: [4]) zajmował się tym zagadnieniem. Idąc za Blestem, wykorzystam siedem różnych modeli, które następnie uszeregują według malejącej wartości sumy kwadratów reszt. Informacje o postaci funkcyjnej tych modeli wraz z ich nazwami własnymi znajdują sie w pierwszych dwóch kolumnach tabeli 3.6.

Wybór wartości parametrów początkowych

W zadaniu znajdowania optymalnej krzywej regresji nieliniowej będziemy szukali takich wartości parametrów $\boldsymbol{\varphi}$, które będą minimalizowały sumy kwadratów reszt,

$$RSS(\boldsymbol{\varphi}) = \sum_{i=1}^{n} (y_i - f(x_i, \boldsymbol{\varphi}))^2. \tag{3.4.2}$$

Jak wspomniałem w podrozdziale 1.3.2, szukanie optymalnych parametrów sprowadza się do rozwiązania układu p równań nieliniowych, których nie jesteśmy w stanie rozwiązać analitycznie. Dlatego też użyjemy iteracyjnej metody numerycznej, która wymaga od nas podania wartości początkowych parametrów.

W naszym zadaniu użyję jako wartości początkowych ostatecznych ocen parametrów, które uzyskał w swojej pracy Blest. Na rysunkach: 3.3, 3.4, 3.5, 3.6, 3.7, 3.8, 3.9 zaznaczone zostały czarną przerywaną linią wykresy kolejnych rozważanych przeze mnie modeli dla zadanych parametrów początkowych. Wartości tych parametrów przedstawione zostały w trzeciej kolumnie tabeli 3.6. Widzimy, że wszystkie te krzywe znajdują się stosunkowo blisko danych, dlatego też powinny być dobrym przybliżeniem początkowym rozwiązania.

Oszacowanie współczynników modeli nieliniowych

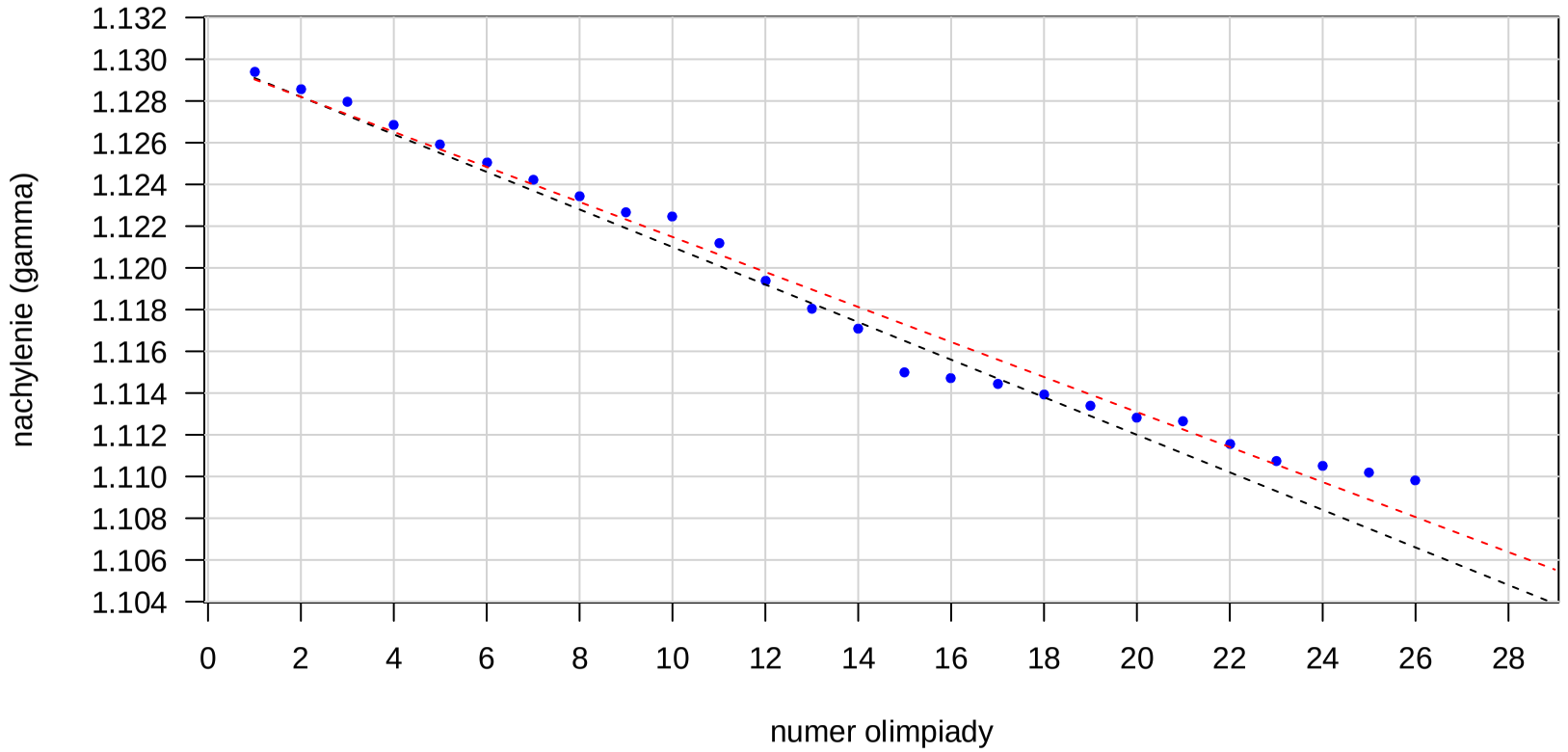
Po określeniu wartości parametrów początkowych dla wybranych modeli nieliniowych możemy przejść do znalezienia optymalnych wartosci tych parametrów. W tym celu wykorzystam funkcję `nls` z pakietu R. Kod wraz z opisem znajdują się w dodatku B.3. Szczegóły wyników jakie uzyskałem umieszczone zostały w czterech ostatnich kolumnach tabeli 3.6. Dodatkowo na rysunkach: 3.3, 3.4, 3.5, 3.6, 3.7, 3.8, 3.9 zaznaczone zostały czerwoną przerywaną linią wykresy kolejnych funkcji dla optymalnych parametrów uzyskanych przy pomocy funkcji `nls`.

Porównanie modeli

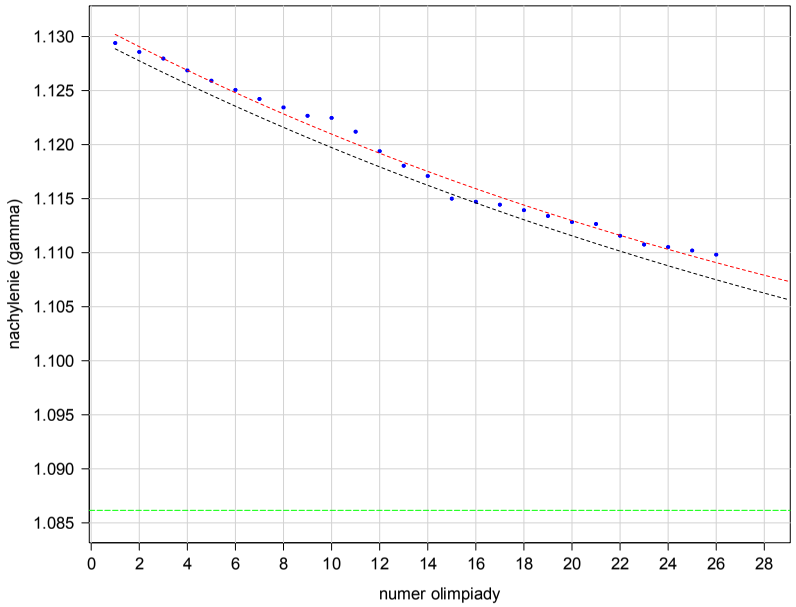
Wartości RSS dla optymalnych parametrów w rozważanych modelach nieliniowych są rzędu 10^{-5} lub 10^{-6}. Okazuje się, że najdokładniej, z rozpatrywanych przez mnie modeli, dane opisuje model wykładniczy antysymetryczny. Jego RSS jest najmniejszy i wynosi $4,91 \times 10^{-6}$.

Wzór	Model	Parametry początkowe	Oceny parametrów	$\sigma_{\widehat{\varphi}_i}$	$RSS \times 10^3$	γ_∞
$y_i = \varphi_1 + \varphi_2 x_i$	Liniowy	$\varphi_1 = 1.129$	$\widehat{\varphi}_1 = 1.12987$	0.000376	0.02076	nie istnieje
		$\varphi_2 = -0.0009$	$\widehat{\varphi}_2 = -0.00084$	0.0000243		
$y_i = \varphi_1 + \varphi_2 \exp(-\varphi_3 x_i)$	Wykładniczy	$\varphi_1 = 1.08$	$\widehat{\varphi}_1 = 1.08616$	0.00841	0.01214	1.08616
		$\varphi_2 = 0.05$	$\widehat{\varphi}_2 = 0.0452$	0.00806		
		$\varphi_3 = 0.023$	$\widehat{\varphi}_3 = 0.02612$	0.00681		
$y_i = \varphi_4 - \varphi_1[1 - \exp(-\varphi_2 x_i)]^{\varphi_3}$	Rozszerzony	$\varphi_1 = 0.029$	$\widehat{\varphi}_1 = 0.02313$	0.00173	0.00674	1.105806
	Chapmana-Richardsa	$\varphi_2 = 0.058$	$\widehat{\varphi}_2 = 0.09972$	0.01596		
		$\varphi_3 = 1.426$	$\widehat{\varphi}_3 = 2.392$	0.428		
		$\varphi_4 = 1.13$	$\widehat{\varphi}_4 = 1.12893$	0.00044		
$y_i = \varphi_4 + \varphi_1 \exp\{-\exp[\varphi_3(x_i - \varphi_2)]\}$	Zreparametryzowany	$\varphi_1 = 0.025$	$\widehat{\varphi}_1 = 0.02788$	0.0031	0.00598	1.109818
	Gompertza	$\varphi_2 = 10.23$	$\widehat{\varphi}_2 = 11.305$	0.904		
		$\varphi_3 = 0.115$	$\widehat{\varphi}_3 = 0.1001$	0.0137		
		$\varphi_4 = 1.111$	$\widehat{\varphi}_4 = 1.10982$	0.00053		
$y_i = \varphi_4 - \varphi_1 \exp[-\exp(\varphi_2 - \varphi_3 x_i)]$	4-parametrowy	$\varphi_1 = 0.0274$	$\widehat{\varphi}_1 = 0.02294$	0.00172	0.00595	1.107169
	Gompertza	$\varphi_2 = 0.792$	$\widehat{\varphi}_2 = 1.2371$	0.173		
		$\varphi_3 = 0.102$	$\widehat{\varphi}_3 = 0.1277$	0.014		
		$\varphi_4 = 1.133$	$\widehat{\varphi}_4 = 1.13011$	0.00092		
$y_i = \varphi_4 - \varphi_1[1 + \exp(\varphi_2 - \varphi_3 x_i)]^{-1}$	Logistyczny	$\varphi_1 = 0.027$	$\widehat{\varphi}_1 = 0.02431$	0.002	0.00551	1.108497
		$\varphi_2 = 1.417$	$\widehat{\varphi}_2 = 1.8933$	0.292		
		$\varphi_3 = 0.156$	$\widehat{\varphi}_3 = 0.1765$	0.0203		
		$\varphi_4 = 1.135$	$\widehat{\varphi}_4 = 1.1328$	0.0014		
$y_i = \varphi_4 + \varphi_1 \exp[-\varphi_2(x_i - \varphi_3)]$ dla $x_i \geqslant \varphi_3$	Wykładniczy	$\varphi_1 = 0.016$	$\widehat{\varphi}_1 = 0.01449$	0.00148	0.00491	1.105911
$y_i = \varphi_4 + \varphi_1\{2 - \exp[\varphi_2(x_i - \varphi_3)]\}$ dla $x_i < \varphi_3$	antysymetryczny	$\varphi_2 = 0.069$	$\widehat{\varphi}_2 = 0.0873$	0.0133		
		$\varphi_3 = 10.267$	$\widehat{\varphi}_3 = 11.039$	0.437		
		$\varphi_4 = 1.1037$	$\widehat{\varphi}_4 = 1.10591$	0.00132		

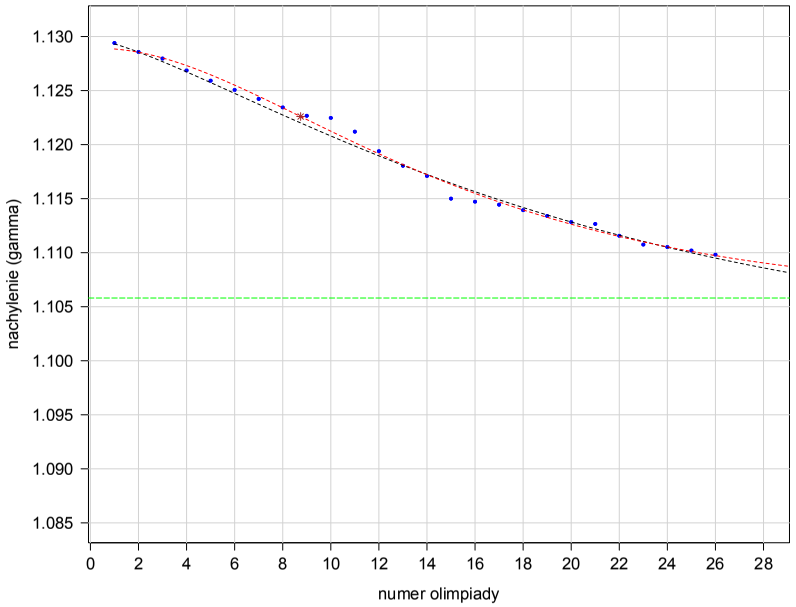
Tabela 3.6: Szczegóły dopasowania krzywych regresji nieliniowych do danych: y - zmienna objaśniana odpowiadająca za parametr γ_i, x - zmienna objaśniająca czyli numer olimpiady. W kolejnych kolumnach dane są: wzór modelu nieliniowego, nazwa własna modelu, wartości początkowe parametrów modelu, oceny parametrów modelu, błędy ocen parametrów, resztowa suma kwadratów, asymptota pozioma regresji nieliniowej. Modele zostały uporządkowane według malącej wartośći RSS. Opracowanie własne.



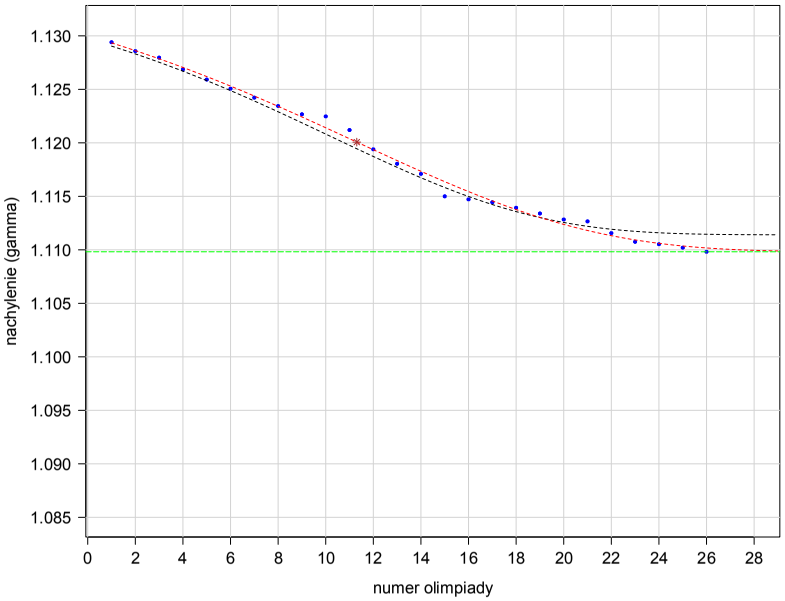
Rysunek 3.3: Oceny parametru γ dla kolejnych olimpiad wraz z krzywymi modelu liniowego $y = \varphi_1 + \varphi_2 x$. Linia czarna przerywana oznacza model z parametrami początkowymi $\varphi_1 = 1.13$, $\varphi_2 = 0.0009$. Linia czerwona przerywana to model najlepiej dopasowany do danych. Opracowanie własne.



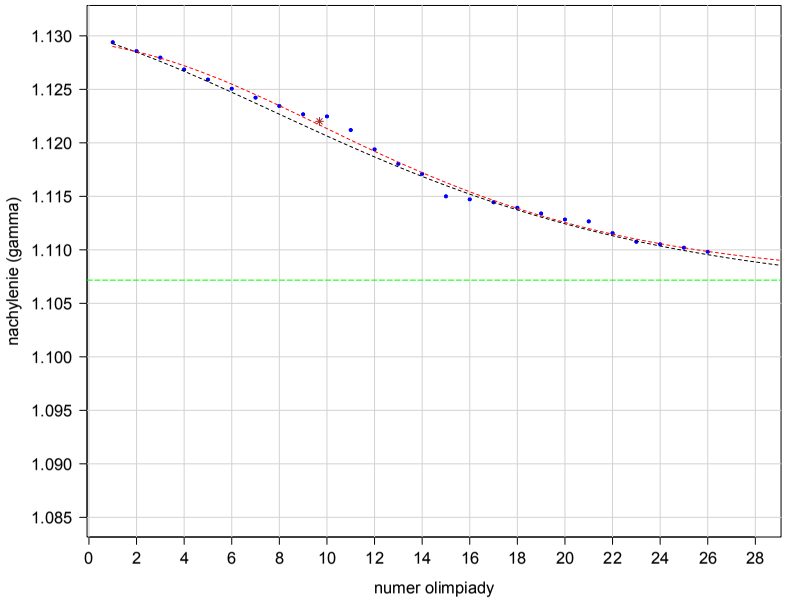
Rysunek 3.4: Oceny parametru γ dla kolejnych olimpiad wraz z krzywymi modelu wykładniczego $y = \varphi_1 + \varphi_2 \exp(-\varphi_3 x)$. Linia czarna przerywana oznacza model z parametrami początkowymi $\varphi_1 = 1.08$, $\varphi_2 = 0.05$, $\varphi_3 = 0.023$. Linia czerwona przerywana to model najlepiej dopasowany do danych. Linią zieloną przerywaną oznaczona została γ_∞. Opracowanie własne.



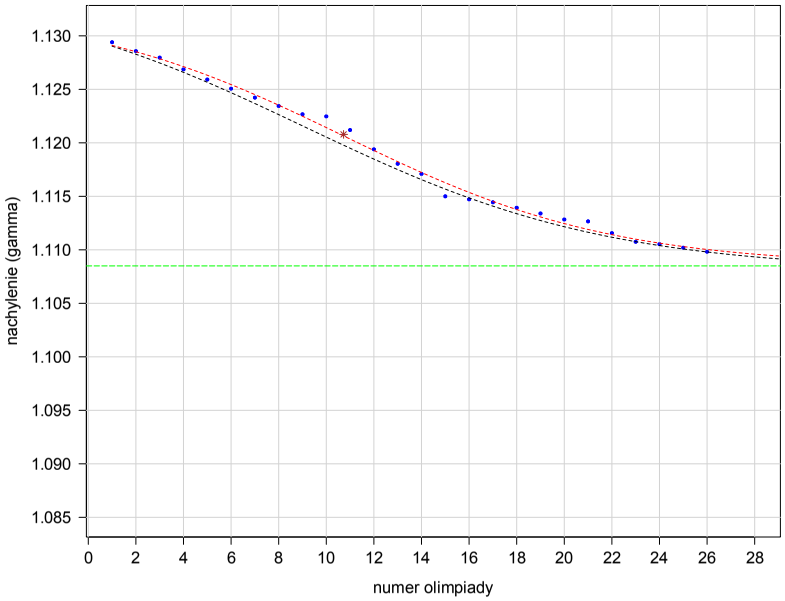
Rysunek 3.5: Oceny parametru γ dla kolejnych olimpiad wraz z krzywymi rozszerzonego modelu Chapmana-Richardsa $y = \varphi_4 - \varphi_1[1 - \exp(-\varphi_2 x)]^{\varphi_3}$. Linia czarna przerywana oznacza model z parametrami początkowymi $\varphi_1 = 0.029$, $\varphi_2 = 0.058$, $\varphi_3 = 1.4257$, $\varphi_4 = 1.1298$. Linia czerwona przerywana to model najlepiej dopasowany do danych. Linią zieloną przerywaną oznaczona została γ_∞. Brązowa gwiazdka oznacza punkt przegięcia optymalnej krzywej. Opracowanie własne.



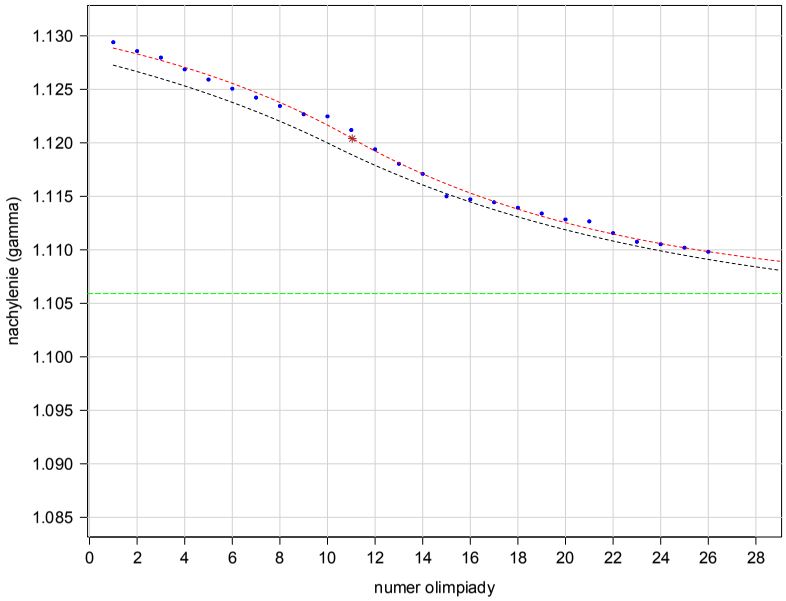
Rysunek 3.6: Oceny parametru γ dla kolejnych olimpiad wraz z krzywymi zreparametryzowanego modelu Gompertza $y = \varphi_4 + \varphi_1 \exp\{-\exp[\varphi_3(x - \varphi_2)]\}$. Linia czarna przerywana oznacza model z parametrami początkowymi $\varphi_1 = 0.02495, \varphi_2 = 10.227, \varphi_3 = 0.1148, \varphi_4 = 1.111402$. Linia czerwona przerywana to model najlepiej dopasowany do danych. Linią zieloną przerywaną oznaczona została γ_∞. Brązowa gwiazdka oznacza punkt przegięcia optymalnej krzywej. Opracowanie własne.



Rysunek 3.7: Oceny parametru γ dla kolejnych olimpiad wraz z krzywymi 4-parametrowego modelu Gompertza $y = \varphi_4 - \varphi_1 \exp[-\exp(\varphi_2 - \varphi_3 x)]$. Linia czarna przerywana oznacza model z parametrami początkowymi $\varphi_1 = 0.0274$, $\varphi_2 = 0.792$, $\varphi_3 = 0.102$, $\varphi_4 = 1.133$. Linia czerwona przerywana to model najlepiej dopasowany do danych. Linią zieloną przerywaną oznaczona została γ_∞. Brązowa gwiazdka oznacza punkt przegięcia optymalnej krzywej. Opracowanie własne.



Rysunek 3.8: Oceny parametru γ dla kolejnych olimpiad wraz z krzywymi modelu logistycznego $y = \varphi_4 - \varphi_1[1 + exp(\varphi_2 - \varphi_3 x)]^{-1}$. Linia czarna przerywana oznacza model z parametrami początkowymi $\varphi_1 = 0.027$, $\varphi_2 = 1.417$, $\varphi_3 = 0.156$, $\varphi_4 = 1.135$. Linia czerwona przerywana to model najlepiej dopasowany do danych. Linią zieloną przerywaną oznaczona została γ_∞. Brązowa gwiazdka oznacza punkt przegięcia optymalnej krzywej. Opracowanie własne.



Rysunek 3.9: Oceny parametru γ dla kolejnych olimpiad wraz z krzywymi modelu wykładniczego antysymetrycznego $y = \varphi_4 + \varphi_1 \exp[-\varphi_2(x - \varphi_3)]$ dla $x \geqslant \varphi_3$, $y = \varphi_4 + \varphi_1\{2 - \exp[\varphi_2(x - \varphi_3)]\}$ dla $x < \varphi_3$. Linia czarna przerywana oznacza model z parametrami początkowymi $\varphi_1 = 0.016$, $\varphi_2 = 0.069$, $\varphi_3 = 10.267$, $\varphi_4 = 1.1037$. Linia czerwona przerywana to model najlepiej dopasowany do danych. Linią zieloną przerywaną oznaczona została γ_∞. Brązowa gwiazdka oznacza punkt przegięcia optymalnej krzywej. Opracowanie własne.

Interpretacja krzywych modeli nieliniowych

Na każdym z rysunków poza 3.3 oraz 3.4 możemy dostrzec wyraźny punkt przegięcia krzywej najlepiej dopasowanej do danych. Punkty te znajdują się pomiędzy $x = 8.7$ a $x = 11.3$, czyli mniej więcej pomiędzy 1944 a 1960 rokiem. Widać zatem wyraźnie, że okres II wojny miał znaczący wpływ na spadek tempa rozwoju rekordów świata. Z drugiej strony koncentracja punktów przegięcia w tym przedziale może świadczyć o tym, że ciągłe polepszanie rezultatów sportowców poprzez: stosowanie odpowiedniej diety, dostosowanie indywidualnego programu treningowego, projektowanie jeszcze lepszego obuwia, przestanie być w pewnym momencie pomocne lekkoatletom w poprawie rezultatów. Stąd też na każdym z rysunków poza 3.3 widzimy spadek tempa rozwoju współczynnika γ aż do osiągnięcia wartości granicznej γ_∞.

3.4.3. Diagnostyka modelu wykładniczego antysymetrycznego

Zanim przejdziemy do jakichkolwiek konkluzji, powinniśmy jeszcze sprawdzić czy założenia 1.3.1 oraz 1.3.2 są spełnione. Wiemy, że zmienne x_i oznaczające numer kolejnej olimpiady są niezależne, a ponadto liczba parametrów naszego modelu (4) jest znacznie mniejsza niż liczba obserwacji (26). Dlatego też będzie nas wyłącznie interesowało sprawdzenie drugiego z założeń. Diagnostykę przeprowadzę dla modelu, który po analizie RSS, wydawał się najdokładniej opisywać dane. W dodatku B.4 umieściłem kody do wykresów i testów diagnostycznych, które wykonałem.

Wielokrotnie będę używał wyrażenia reszta (ang. *residual*) dlatego też podkreślę, że resztą nazywamy różnicę pomiędzy wartością zmiennej objaśnianej, a wartością dopasowaną przez model (w naszym przypadku nieliniowy)

$$\widehat{r}_i = y_i - f(x_i, \widehat{\beta}) = y_i - \widehat{y}_i.$$

Wykorzystam również temin wystandaryzowana reszta, który oznacza

$$\frac{\widehat{r}_i}{\sqrt{\widehat{\sigma^2}}},$$

gdzie $\widehat{\sigma^2}$ zgodnie z (1.3.3) oznacza nieobciążony estymator wariancji σ^2.

Funkcja nieliniowa f

Wydaje się, że wykres modelu wykładniczego antysymetrycznego z optymalnymi parametrami (rysunek 3.9) dosyć dokładnie opisuje zbiór danych. Jednak dla pewności sprawdzimy jak zmieniają się reszty w rozważanym modelu w zależności od predykcji wyznaczonych przez krzywą nieliniową.

Analizując rysunek 3.10a o nagłówku „Residuals vs. Fittted" widzimy, że wartości reszt zależą "sinusoidalnie" od wartości $\widehat{y}_i$. Niestety nie jest to dobra informacja, ponieważ świadczy to o tym, że wybrany model nie jest aż tak odpowiedni do opisu danych, jak mogło się wydawać po analizie jego wykresu.

Homoskedastyczność

Kolejnym założeniem, które powinniśmy sprawdzić jest ocena czy wariancje reszt są jednorodne.

Przyglądając się rysunkowi 3.10b o nagłówku „Scale location" nie jesteśmy w stanie stwierdzić jakiegokolwiek trendu, żadnej szczególnej zależności funkcyjnej. Możemy zatem wnioskować, że założenie o jednorodności wariancji reszt jest spełnione.

Rozkład normalny reszt

W dalszej kolejności powinniśmy sprawdzić, czy reszty mają rozkład normalny. W tym celu użyjemy dwóch narzędzi. Najpierw dokonamy analizy wykresu kwantylowego, a następnie wykonamy jeszcze test diagnostyczny Shapiro-Wilka.

Punkty na rysunku 3.10c układają się wzdłuż lini prostej przerywanej. Zauważmy jednak, że nie jest to prosta $y = x$. Możemy zatem wnioskować, że reszty mają rozkład normalny, który jest liniową transformacją standardowego rozkładu normalnego.

Żeby potwierdzić nasze przypuszczenia, przeprowadzimy dodatkowo test Shapiro-Wilka.

```
> shapiro.test(stdRes)

    Shapiro-Wilk normality test

data:  stdRes
W = 0.9711, p-value = 0.6529
```

Ponieważ p-wartość wynosi 0.6529 możemy stwierdzić, że zakładając poziom istotności 0.05 nie możemy odrzucić hipotezy zerowej o normalności rozkładu reszt.

Niezależność reszt

W celu zbadania czy reszty są niezależne dokonam analizy rysunku 3.10d, a następnie wykonam test serii (ang. *runs test*).

Na rysunku o nagłówku „Autocorrelation" widzimy, że zgromadzone punkty układają się w trend liniowy. Ponieważ są one rozproszone, nie powinniśmy być pewni tej zależności dlatego też przeprowadzimy dodatkowo test serii.

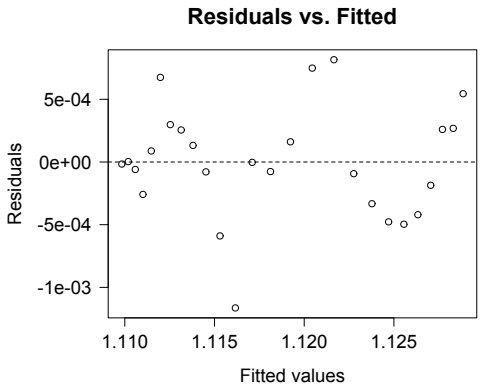
```
> runs.test(run)

    Runs Test

data:  run
Standard Normal = -2.3858, p-value = 0.01704
alternative hypothesis: two.sided
```

P-wartość wynosi 0.01704. Ustalając zatem poziom istotności równy 0.05 możemy odrzucić hipotezę zerową o losowości reszt modelu wykładniczego antysymetrycznego.

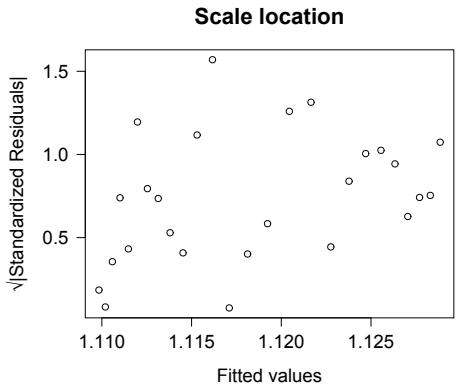
Wnioski z diagnostyki

Podsumowując wszystkie badania jakie wykonałem, możemy stwierdzić, że reszty mają rozkład normalny o tej samej wariancji, ale niestety nie są między sobą niezależne. Co więcej, możemy powiedzieć, że zależą one "sinusoidalnie" od wartości dopasowanych przez model. Mimo tego, że wykres modelu wykładniczego antysymetrycznego z optymalnymi parametrami wskazuje na dobre dopasowanie do danych, to jak się okazuje, nie spełnia on wszystkich wymaganych założeń.

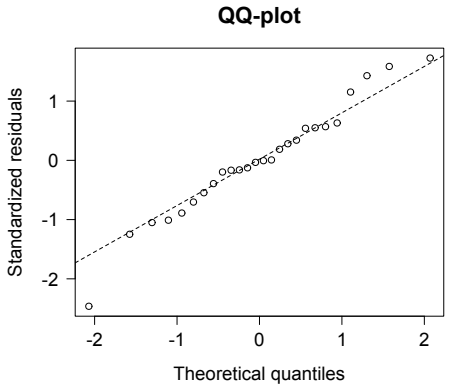
(a)



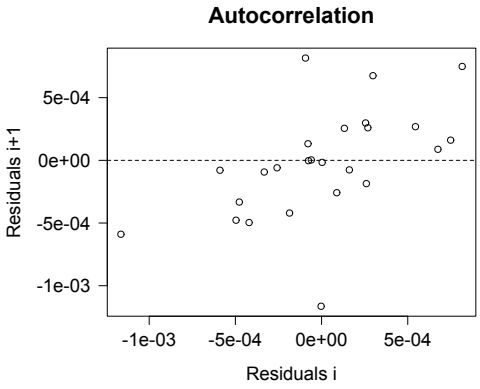
(b)



(c)



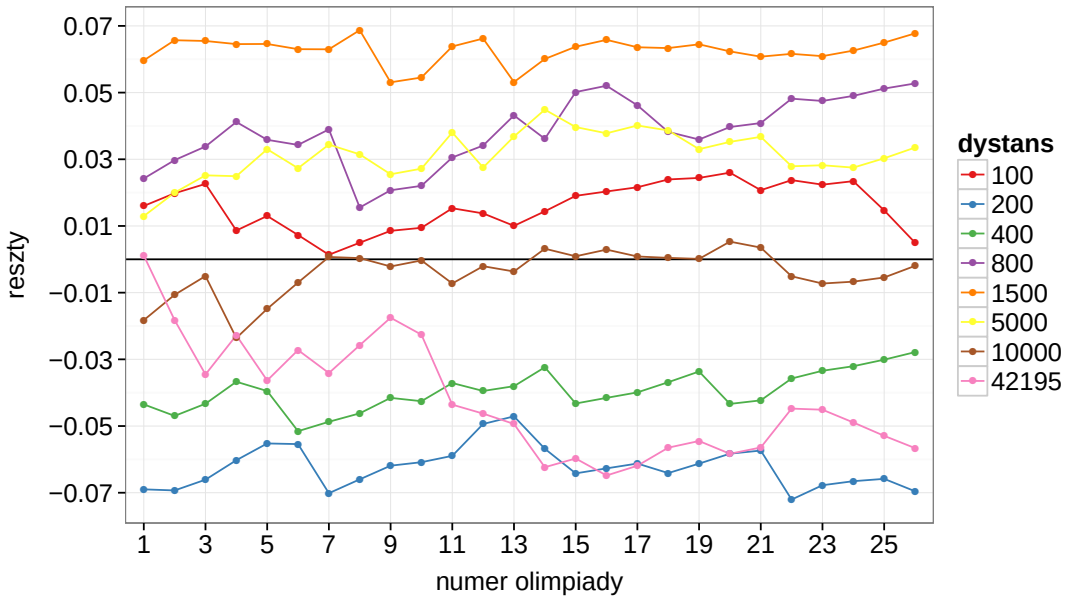
(d)



Rysunek 3.10: Wykresy diagnostyczne modelu wykładniczego antysymetrycznego: (3.10a) wykres reszt w zależności od predykcji $\widehat{y}_i$ modelu; (3.10b) wykres pierwiastków z modułów wystandaryzowanych reszt w zależności od predykcj $\widehat{y}_i$ modelu; (3.10c) wykres zależności pomiędzy wystandaryzowanymi resztami, czyli kwantylami próbkowymi a kwantylami teoretycznymi rozkładu normalnego. Przerywaną linią oznaczona została prosta łącząca I i III kwartyl; (3.10d) wykres przedstawiający zależność pomiędzy $(i+1)$-szą a i-tą resztą. Opracowanie własne.

3.4.4. Usprawnienie modelu

Zanim zaprezentujemy ostateczne wyniki, warto przyjrzeć się, jak wyglądają reszty w modelu (3.2.5). Rysunek 3.11 należy rozumieć w następujący sposób. Dla i-tej olimpiady ($i \in \{1, \ldots, 26\}$) oraz j-tego dystansu ($j \in \{1, \ldots, 8\}$) zaznaczona została wartość reszty ($\widehat{r}_{i,j} = T_{i,j} - \widehat{T}_{i,j}$).



Rysunek 3.11: Wartości reszt w zagnieżdżonym modelu liniowym (3.2.5) dla poszczególnych dystansów w kolejnych latach olimpijskich. Opracowanie własne.

Widzimy wyraźnie, że cztery dystanse: 200 m, 400 m, 10000 m oraz maraton mają ujemne reszty. Świadczy to o tym, że wartości czasów dopasowane przez model (3.2.5) dla tych dystansów są większe niż czasy przez nas zaobserwowane. Oczywiście pomijając przypadek idelanego dopasowania krzywej do danych, obserwacje zawsze będą znajdowały się po obu stronach najlepiej dopasowanych do niej krzywej. Zwróćmy jednak uwagę, że reszty dla dystansów 200 m, 400 m oraz maratonu są ujemne przez wszystkie 26 olimpiad. Postaramy się zatem wykorzystać tę stabilność reszt i dokonać korekty dla tych dystansów, dla których nasz model daje czasy gorsze niż wynoszą rekordy dla poszczególnych olimpiad. Podkreślmy najpierw, że w celu znalezienia minimalnego czasu będziemy korzystali z równania (3.2.4). Dokonamy zatem następującej korekty parmetru γ_∞ dla dystansów: 200 m, 400 m, 10000 m oraz maratonu,

$$\widetilde{\gamma}_\infty = \widehat{\gamma}_\infty - (\max_{i \in \{1,\ldots,26\}} |\widehat{r}_{i,j}|)/D_j, \tag{3.4.3}$$

dzięki czemu otrzymamy następujący wzór na obliczenie czasów granicznych

$$\widehat{t}_{\infty,j} = \exp(\widehat{T}_{\infty,j}),$$

gdzie

$$\widehat{T}_{\infty,j} = \begin{cases} A + \widetilde{\gamma}_\infty D_j = A + \widehat{\gamma}_\infty D_j \ - \max_{i \in \{1,\ldots,26\}} |\widehat{r}_{i,j}| & \text{dla } j = 2,3,7,8 \\ A + \widehat{\gamma}_\infty D_j & \text{dla } j = 1,4,5,6 \end{cases}.$$

3.4.5. Granice możliwości lekkoatletów

Znając wartości asymptoty poziomej γ_∞ oraz wprowadziwszy korektę możemy przejść do zaprezentowania szukanych granic możliwości lekkoatletów na rozpatrywanych dystansach.

W tabeli 3.7 przedstawione zostały czasy, które uzyskano po korekcie (3.4.3) dla trzech, najlepszych według wartości RSS, rozważanych przeze mnie modeli nieliniowych. Dla porównania umieszczone również zostały w drugiej kolumnie aktualne rekordy świata. Dodatkowo w trzeciej kolumnie przedstawione zostały przewidywane wyniki w 2012 roku, gdzie γ_{26} została wyznaczona przez model antysymetryczny wykładniczy dla 26 olimpiady czyli 2012 roku. Wyniki z trzeciej kolumny zostały umieszczone, żeby podkreślić znaczenie korekty (3.4.3) jakiej dokonano w poprzednim podrozdziale.

Dystans (m)	Aktualny rekord (s)	Przewidywane czasy (s) dla modelu:			
		Antysymetryczny wykładniczy (2012)	Granice możliwości		
			Antysymetryczny wykładniczy	4-parametrowy Gompertza	Logistyczny
100	9.58	9.53	9.36	9.42	9.47
200	19.19	20.57	18.75	18.88	19.01
400	43.18	44.4	41.19	41.5	41.84
800	101.01	95.83	93.35	94.14	94.98
1500	206	192.53	187.08	188.81	190.65
5000	757.35	732.49	708.41	716.04	724.19
10000	1577.53	1580.88	1489.19	1506.54	1525.09
42195	7382	7813.32	7023.04	7117.72	7219.15

Tabela 3.7: Przewidywane granice możliwości lekkoatletów. W kolejnych kolumnach dane są: badany dystans; rekordy świata z 2012 roku na wybranych dystansach; przewidywane czasy na 2012 rok z gammą uzyskaną przy pomocy modelu antysymetrycznego wykładniczego; przewidywane granice możliwości lekkoatletów ze skorygowaną gammą uzyskaną kolejno przy pomocy modeli: antysymetrycznego wykładniczego, 4-parametrowego Gompertza i Logistycznego.

Podsumowanie

Predykcje, które umieszczone zostały w tabeli 3.7 dotyczą nie bez powodu trzech różnych modeli. Należy bowiem zaznaczyć, że model antysymetryczny wykładniczy nie spełnia wszystkich wymaganych założeń. Dlatego też umieszczono pozostałe dwa modele, żeby pokazać pewien przedział granic możliwości lekkoatletów. Warto również zaznaczyć, że korekta czasów jakiej dokonano z pewnością nie jest idealnym rozwiązaniem i pozwala przyjąć pewien margines błędu przewidywanych czasów.

Zakończenie

W pracy tej przedstawiona została analiza danych rzeczywistych dotyczących rekordów świata lekkoatletów w ośmiu dyscyplinach biegowych na przestrzeni 100 lat. Analiza ta posłużyła do znalezienia granic możliwości sportowców na rozważanych dystansach.

Modelowanie oparte zostało na trzech zagadanieniach teoretycznych: modele liniowe, test ilorazu wiarogodności oraz modele nieliniowe, które pokrótce omówiono w rozdziale teoretycznym. W części praktycznej najpierw użyto modelu potęgowego do opisu zbioru danych. W dalszej kolejności przekształcono model potęgowy w model liniowy w celu ułatwienia znalezienia ocen parametrów. Następne podrozdziały zawierały opis modelu zagnieżdżonego, który mógł zostać wykorzystany w kolejnym etapie pracy dzięki pozytywnemu wynikowi testu ilorazu wiarogodności. W ostatnim podrozdziale wykorzystano zestaw modeli nielinowych w celu przybliżenia zachowania się współczynnika γ w kolejnych latach olimpijskich. Ostatecznie wybrano trzy modele nieliniowe, najlepiej opisujące zachowanie współczynika nachylenia modelu zagnieżdżonego i znaleziono ich asymptoty poziome. Dzięki temu możliwe było wyznaczenie granic możliwości lekkoatletów po uwzględnieniu odpowiedniej korekty.

Dodatek A

Rozkład QR macierzy

W podrozdziale 1.1.2 wielokrotnie będę używał dekompozycji QR macierzy planu X. Dlatego też warto przypomnieć czym jest rozkład QR. Wszystkie oznaczenia z rozdziału 1.1 pozostają w mocy.

Szeroki rozkład QR: Każdą rzeczywistą macierz pełnego rzędu X wymiaru $n \times p$ $(p < n)$ można zapisać jako iloczyn macierzy ortogonalnej Q wymiaru $n \times n$ $(Q^T Q = I_{n \times n})$ oraz macierzy górnotrójkątnej R wymiaru $n \times p$:

$$X = QR = \underbrace{\begin{bmatrix} | & & | \\ q_1 & & q_n \\ | & & | \end{bmatrix}}_{n \times n} \underbrace{\begin{bmatrix} * & * \\ 0 & * \\ \hline 0 & \end{bmatrix}}_{n \times p}, \tag{A.0.1}$$

gdzie $q_1, \ldots, q_n$ oznaczają ortogonalne kolumny macierzy Q

Wąski rozkład QR: Ponieważ $(n - p)$ dolnych wierszy macierzy R jest zerowa to możemy odpowiednio skróci zapis:

$$\begin{aligned} X = QR &= Q \begin{bmatrix} R_1 \\ \hline 0 \end{bmatrix} = \left[\begin{array}{c:c} Q_1 & Q_2 \end{array}\right] \begin{bmatrix} R_1 \\ \hline 0 \end{bmatrix} \\ &= Q_1 R_1 = \underbrace{\begin{bmatrix} | & | \\ q_1 & q_p \\ | & | \end{bmatrix}}_{n \times p} \underbrace{\begin{bmatrix} * & * \\ 0 & * \end{bmatrix}}_{p \times p}, \end{aligned} \tag{A.0.2}$$

gdzie Q_1 jest macierzą wymiaru $n \times p$ o ortogonalnych kolumnach $q_1, \ldots, q_p$, a R_1 jest macierzą górnotrójkątną wymiaru $p \times p$.

Warto również podkreślić, że kolumny macierzy Q_1 otrzymywane są w wyniku ortogonalizacji Gramma-Schmidta kolumn macierzy X. Dlatego też Q_1 jest macierzą ortogonalną $(Q_1^T Q_1 = I_{p \times p})$, a jej kolumny rozpinają tę samą przestrzeń co kolumny macierzy X.

Dodatek B

Kody programu R użyte w pracy

B.1. Oszacowanie współczynników α_i oraz β_i

Przedstawiony program znajduje oceny $\widehat{\alpha_i}, \widehat{\beta_i}$ $(i \in \{1, ..., n\})$ modelu (3.2.3)

$$T = D\theta_1 + \xi.$$

Algorytm podzieliłem na dwa etapy:

- Wczytywanie danych z tabeli 3.1,
- Wyznaczenie ocen $\widehat{\alpha_i}, \widehat{\beta_i}$.

Wczytywanie danych

W pierwszym kroku wczytałem przy pomocy funkcji `read.table()` dane z pliku `dane.csv`, przechowującego informacje zgromadzone w tabeli 3.1. Następnie korzystając z polecenia `as.matrix()` zapisałem wczytane dane w formacie macierzowym.

```
setwd("C:/Users/Witek/Desktop/Praca licencjacka/analiza danych/dane by me/R")
dane<-as.matrix(read.table("dane.csv",sep=";",dec=",",header=FALSE))
```

Listing B.1: Wczytywanie danych.

Wyznaczanie ocen

W kolejnym kroku zająłem się wyznaczeniem ocen parametrów $\widehat{\alpha_i}, \widehat{\beta_i}$. Dodatkowo wyznaczyłem również współczynniki determinacji R^2 oraz sumy kwadratów reszt (ang. RSS) dla i-tego podmodelu (3.2.2).

```
# Postać ogólna modelu
dane_m1 <- dane[,2:ncol(dane)] # dane bez kolumny "Rok"
rownames(dane_m1) <- dane[,1]
n<-nrow(dane_m1) # liczba wierszy macierzy dane_m1
m<-ncol(dane_m1) # liczba kolumn macierzy dane_m1
dane_m1_po <- as.data.frame(as.table(dane_m1))

# dodaję do ramki danych wszystkie zmienne objaśniane Log(T)
dane_m1_po$logT <- log(dane_m1_po[,3])

# dodaję do ramki danych wszystkie zmiennn objaśniające Log(D)
dane_m1_po$logD <- log(as.numeric(as.character(dane_m1_po[,2])))
```

```
# Oceny współczynników modelu wraz z ich błędami
alpha<-c() # wektor ocen parametru alpha
beta<-c() # wektor ocen parametru beta
er_alpha<-c() # wektor błędów ocen parametru alpha
er_beta<-c() # wektor błędów ocen parametru beta
model1 <- summary(lm(logT~logD:Var1 + Var1-1, data=dane_m1_po)) # podsumowanie modelu
alpha<-coef(model1)[1:n,1]
er_alpha<-coef(model1)[1:n,2]
beta<-coef(model1)[(n+1):(2*n),1]
er_beta<-coef(model1)[(n+1):(2*n),2]

# Znajduję średnie ze zmiennych objaśnianych Log(T_ij) dla i-tego roku  olimpijskiego
M<-vector("integer",n) # wektor średnich zmiennych objaśnianych
t<-dane_m1_po[,4] # wektor 208 zmiennych objaśnianych dla wszystkich lat olimpijskich
for (i in 1:n){
  for (j in 1:m){
    M[i]<-M[i]+t[i+n*(j-1)]
  }
  M[i]<-M[i]/m
}

# Petla znajdujaca szczegoly (R^2, RSS) dotyczace i-tego podmodelu regresji
res<-residuals(model1)
TSS<-vector("integer",n) # inicjalizacja wektora TSS
RSS<-vector("integer",n) # inicjalizacja wektora RSS
Rsq<-vector("integer",n) # inicjalizacja wektora RSQ
for (i in 1:n){
  for (j in 1:m){
    RSS[i]<-RSS[i]+res[i+n*(j-1)]^2
    TSS[i]<-TSS[i]+(t[i+n*(j-1)]-M[i])^2
  }
  Rsq[i]<-1-RSS[i]/TSS[i]
}
```

Listing B.2: Wyznaczanie ocen $\widehat{\alpha_i}$ oraz $\widehat{\beta_i}$.

B.1.1. Wykresy ocen współczynników $\widehat{\alpha_i}$ oraz $\widehat{\beta_i}$ w zależności od roku

Program służy do narysowania wykresów przedstawiających zależności pomiędzy rokiem a estymatorami $\widehat{\alpha_i}$ oraz $\widehat{\beta_i}$. W tym celu użyłem pakietu `ggplot2` (źródło: [3],[6]), przy pomocy którego można w bardzo prosty sposób dodawać kolejne warstwy wykresu. Najważniejszą funkcją tego pakietu, którą wykorzystałem do stworzenia moich wykresów jest `ggplot()`.

```
inv<-function(x){
  d<-length(x)
  temp<-c()
  for(i in 1:d){
    temp[i]<-x[d+1-i]
  }
  inv<-temp
}
```

Listing B.3: Funkcja inv odwracająca kolejność elementów w wektorze.

```
library(ggplot2) # BIBLIOTEKA PAKIETU GGPLOT2
rok<-dane[1:n,1] # wektor kolejnych lat olimpijskich
M_alpha<-mean(alpha) # Średnie przecięcie
M_beta<-mean(beta) # Średnie nachylenie

# int_a_dane - punkty w których znajdują się końce przedziałów ufności alphy
int_a_dane<-data.frame(int_a_x=c(rok,inv(rok)),
                       int_a_y=c(alpha+er_alpha,inv(alpha-er_alpha)))
# a_dane -  punkty (rok,alpha)
a_dane<-data.frame(r=rok,przeciecie=alpha)
```

```
plot_a<-ggplot(data=a_dane,aes(x=r,y=przeciecie))+ # FUNKCJA GGPLOT()
  scale_x_continuous(breaks=seq(min(a_dane$r),max(a_dane$r),by=10))+
  scale_y_continuous(breaks=round(seq(min(int_a_dane$int_a_y),
                                      max(int_a_dane$int_a_y),by=0.02),2))+
  labs(x="rok",y="przeciecie (alpha)")+
  geom_polygon(data=int_a_dane,aes(x=int_a_x,y=int_a_y),col=gray,alpha=0.2)+
  geom_point(col="blue")+
  geom_line(col="blue")+
  geom_hline(aes(yintercept=M_alpha),col="red")+
  geom_text(x=1992,y=-2.865,
            label="s r e d n i e  a l p h a  = - 2 . 8 5 6 2",col="red")
plot_a<-plot_a+theme_bw()
plot_a+theme(axis.title=element_text(size=15),axis.text=element_text(size=15))
```

Listing B.4: Kod do wykresu rok - przecięcie ($\widehat{\alpha_i}$).

```
# int_b_dane - punkty w których znajdują się końce przedziałów ufności bety
int_b_dane<-data.frame(int_b_x=c(rok,inv(rok)),
                       int_b_y=c(beta+er_beta,inv(beta-er_beta)))
# b_dane -  punkty (rok,beta)
b_dane<-data.frame(r=rok,nachylenie=beta)

plot_b<-ggplot(data=b_dane,aes(x=r,y=nachylenie))+ # FUNKCJA GGPLOT()
  scale_x_continuous(breaks=seq(min(b_dane$r),max(b_dane$r),by=10))+
  scale_y_continuous(breaks=round(seq(min(int_b_dane$int_b_y),
                                      max(int_b_dane$int_b_y),by=0.005),3))+
  labs(x="rok",y="nachylenie (beta)")+
  geom_polygon(data=int_b_dane,aes(x=int_b_x,y=int_b_y),col=gray,alpha=0.2)+
  geom_point(col="blue")+
  geom_line(col="blue")+
  geom_hline(aes(yintercept=M_beta),col="red")+
  geom_text(x=1930,y=1.116,label="s r e d n i e  b e t a  =  1. 1 1 8 5 ",col="red")
plot_b<-plot_b+theme_bw()
plot_b+theme(axis.title=element_text(size=15),axis.text=element_text(size=15))
```

Listing B.5: Kod do wykresu rok - nachylenie ($\widehat{\beta_i}$).

B.2. Model alternatywny - oszacowanie współczynnika γ

Poniższy kod ma na celu znalezienie ocen $\widehat{\gamma_i}$ modelu (3.2.5)

$$T = D\theta_2 + \xi.$$

Dodatkowo wyznaczyłem również współczynniki determinacji R^2 oraz sumy kwadratów reszt (ang. *RSS*) dla i-tego podmodelu (3.2.4).

Wszystkie oznaczenia z Dodatku B.1 pozostają w mocy.

```
# Oceny współczynników modelu
gamma_ma<-c() # wektor ocen parametru gamma
er_gamma_ma<-c() # wektor bledow ocen parametru gamma
model2<-summary(lm(logT~logD:Var1, data=dane_m1_po)) # podsumowanie modelu

# Pierwszy współczynnik to przeciecie, które jest
# takie samo dla wszystkich podmodeli regresji i wynosi A=-2.8562
gamma_ma<-coef(model2)[2:(n+1),1]
er_gamma_ma<-coef(model2)[2:(n+1),2]

# Pętla znajdująca szczegóły (R^2, RSS) dotyczące i-tego podmodelu regresji
res_ma<-residuals(model2)
RSS_ma<-vector("integer",n) # inicjalizacja wektora RSS_ma
Rsq_ma<-vector("integer",n) # inicjalizacja wektora RSQ_ma
```

```
for (i in 1:n){
  for (j in 1:m){
    RSS_ma[i]<-RSS_ma[i]+res_ma[i+n*(j-1)]^2
  }
  Rsq_ma[i]<-1-RSS_ma[i]/TSS[i]
}
```

Listing B.6: Wyznaczanie ocen $\widehat{\gamma}_i$.

B.2.1. Wykres ocen współczynnika $\widehat{\gamma}_i$ w zależności od roku

W wyniku działania poniższego kodu narysowane zostały wykresy przedstawiające zależności pomiędzy rokiem a estymatorem $\widehat{\gamma}_i$. Do ich narysowania użyłem ponownie pakietu `ggplot2`.

```
# Przypomnijmy, że funkcja inv odwraca kolejność elementów w wektorze

# int_g_dane - punkty w których znajdują się końce przedziałów ufności gammy
int_g_dane<-data.frame(int_g_x=c(rok,inv(rok)),
                       int_g_y=c(gamma_ma+er_gamma_ma,inv(gamma_ma-er_gamma_ma)))

# g_dane -  punkty (rok,gamma)
g_dane<-data.frame(r=rok,nachylenie_g=gamma_ma)

plot_g<-ggplot(data=g_dane,aes(x=r,y=nachylenie_g))+
  scale_x_continuous(breaks=seq(min(g_dane$r),max(g_dane$r),by=10))+
  scale_y_continuous(breaks=round(seq(min(int_g_dane$int_g_y),
                                      max(int_g_dane$int_g_y)+0.002,by=0.002),3))+
  labs(x="rok",y="nachylenie (gamma)")+
  geom_polygon(data=int_g_dane,aes(x=int_g_x,y=int_g_y),col=gray,alpha=0.2)+
  geom_point(col="blue")+
  geom_line(col="blue")
plot_g<-plot_g+theme_bw()
plot_g+theme(axis.title=element_text(size=15),axis.text=element_text(size=15))
```

Listing B.7: Kod do wykresu rok - nachylenie($\widehat{\gamma}_i$).

B.3. Modele nieliniowe - oszacowanie współczynników

Zamieszczony w tym podrozdziale kod znajduje oceny parametrów modelu wykładniczego antysymetrycznego. Podkreślę, że kody do pozostałych modeli, które zostały wymienione w tabeli 3.6 różnią się w niewielkim stopniu. Dlatego postanowiłem umieścić jedynie kod do modelu o najmiejszej wartości RSS.

Warto również wspomnieć, że poniższy kod zakończył się po sześciu iteracjach, przy domyślnym poziomie kontroli zbieżności `1e-5`. Niewielka liczba iteracji potrzebnych do uzyskania żądanej zbieżności wskazuje, że dobrze zostały wybrane parametry początkowe, a minimalna wartość RSS, która wyniosła w tym przypadku `4.91e-6` wskazuje, że krzywa z optymalnymi parametrami bardzo dokładnie opisuje zbiór danych.

Wszystkie oznaczenia z dodatków B.1 i B.2 pozostają w mocy.

```
n_olimpiady<-(rok-1908)/4 # zmienna rok po transformacji na "numer olimpiady"
g_dane_pred<-data.frame(x=n_olimpiady,y=gamma_ma) # zbiór danych do opracowania modelu
     nieliniowego

# Definiuję funkcję wykorzystywaną w modelu wykładniczym antysymetrycznym
fun4<-function(x,phi1,phi2,phi3,phi4){
  ifelse(x>=phi3,phi4+phi1*exp(-phi2*(x-phi3)),
         phi4+phi1*(2-exp(phi2*(x-phi3))))
}
```

```
# Chcemy kontrolować nasze rozwiązania
control<-nls.control(tol=1e-5,warnOnly=TRUE,printEval=TRUE)


# stosuję funkcję nls
model.4<-nls(y~fun4(x,phi1,phi2,phi3,phi4), data=g_dane_pred,
             start=list(phi1=0.016,phi2=0.069,phi3=10.267,phi4=1.1037),control, trace=
                 TRUE)
deviance(model.4) # metoda do wyznaczenia minimalnego RSS dla parametrów
mod4<-summary(model.4)
coef(mod4) # metoda do wyznaczenia estymatorów parametrów i błędów estymatorów
phi1.4<-coef(mod4)[1,1] # estymator phi1.4
er_phi1.4<-coef(mod4)[1,2] # błąd estymatora phi1.4
phi2.4<-coef(mod4)[2,1] # estymator phi2.4
er_phi2.4<-coef(mod4)[2,2] # błąd estymatora phi2.4
phi3.4<-coef(mod4)[3,1] # estymator phi3.4
er_phi3.4<-coef(mod4)[3,2] # błąd estymatora phi3.4
phi4.4<-coef(mod4)[4,1] # estymator phi4.4
er_phi4.4<-coef(mod4)[4,2] # błąd estymatora phi4.4
```

B.3.1. Wykres modelu wykładniczego antysymetrycznego

W wyniku działania poniższego kodu narysowany został wykres ocen parametru γ dla kolejnych olimpiad wraz z krzywymi modelu wykładniczego antysymetrycznego. Dodatkowo narysowana została pozioma asymptota γ_∞ dla krzywej z optymalnymi parametrami.

```
par(lab=c(14,10,10),mar=c(5,6,4,2)) # ustawienia okna graficznego

plot(y~x,data=g_dane_pred,las=1,xlab="numer olimpiady",ylab="",
     xlim=c(min(x),max(x)+2),
     ylim=c(1.085,1.131),col="blue",pch=16,cex=0.7)
mtext("nachylenie (gamma)",side=2,line=4.5)
grid(lty=1,lwd=1)

# Zaznaczam dane, czyli oceny parametru gamma w poszczególnych olimpiadach
points(y~x,data=g_dane_pred,col="blue",pch=16,cex=0.7)

# Model nieliniowy z parametrami początkowymi
curve(fun4(x,theta1=0.016,theta2=0.069,theta3=10.267,theta4=1.1037),
      add=TRUE,lty=2,from=1,to=29)

# Model nieliniowy z najlepszymi parametrami
curve(fun4(x,theta1.4,theta2.4,theta3.4,theta4.4),add=TRUE,lty=2,
      from=1,to=29,col="red")

# Punkt przegięcia
x_p=theta3.4
y_p=fun4(x_p,theta1.4,theta2.4,theta3.4,theta4.4)
points(x=x_p,y=y_p,col="brown",pch=8)

# Pozioma asymptota gamma inf
abline(h=theta4.4,lty=5,col="green")
```

B.4. Diagnostyka modeli

W poniższym rozdziale zamieszczone są kody do wykresów i testów diagnostycznych, które znajdują się w podrozdziale 3.4.3.

```
# 1.Wykres residuals vs fitted
# Sprawdzamy czy funkcja nieliniowa jest dobrze dopasowana.
par(mar=c(5,6,4,2))
plot(residuals(model.4)~fitted(model.4),las=1,xlab="Fitted values",
     ylab="",main="Residuals vs Fitted")
```

```
mtext("Residuals",side=2,line=4.5,cex=1)
abline(h=0,lty=2)



# 2. Wykres pierwiastków modułów wystandaryzowanych reszt.
# Sprawdzamy w ten sposób czy wariancja jest jednorodna.
plot(sqrt(abs(residuals(model.4)/summary(model.4)$sigma))~
       fitted(model.4),las=1,xlab="Fitted values",
     ylab="",main="Scale location")
mtext("sqrt(|Standardized Residuals|)",side=2,line=3.5, cex=1)
stdRes<-residuals(model.4)/summary(model.4)$sigma
qqnorm(stdRes,las=1,xlab="Theoretical quantiles",
       ylab="",main="QQ-plot")
mtext("Standardized residuals",side=2,line=3, cex=0.8)
qqline(stdRes,lty=2)

# 3. Wykres kwantylowy dla rozkładu normalnego
# Sprawdzamy czy kwantyle empiryczne mają faktycznie rozkład normalny
stdRes<-residuals(model.4)/summary(model.4)$sigma
qqnorm(stdRes,las=1,xlab="Theoretical quantiles",
       ylab="",main="QQ-plot")
mtext("Standardized residuals",side=2,line=3, cex=0.8)
qqline(stdRes,lty=2)

# Test Shapiro-Wilka (normalność reszt)
shapiro.test(stdRes)

# 4. Wykres zależności pomiędzy kolejnymi resztami.
# Sprawdzamy w ten sposób czy założenie o niezależności reszt jest spełnione
plot(residuals(model.4),c(residuals(model.4)[-1], NA),las=1,
     xlab="Residuals i",ylab="",main="Autocorrelation")
mtext("Residuals i+1",side=2,line=4.5)
abline(h=0,lty=2)

# Test serii (losowość reszt)
require(tseries)
# kodujemy odpowiednio serię
run<-as.factor(residuals(model.4)>mean(residuals(model.4)))
runs.test(run)
```

Bibliografia

[1] D. M. Bates, D.G. Watts. *Nonlinear Regression Analysis and Its Applications*. Wiley, 1988. 39-52.

[2] P. Biecek. *Analiza danych z programem R. Modele liniowe z efektami stałymi, losowymi i mieszanymi*. Wydawnictwo naukowe PWN, 2011.

[3] P. Biecek. *Przewodnik po pakiecie R*. Oficyna Wydawnicza GiS, 2011.

[4] D. C. Blest. *Lower bounds for athletic performance*. Journal of the Royal Statistical Society. Series D (The Statistician),45(2):243-253,1996.

[5] The Economist *Faster, higher, no longer*, `http://www.economist.com/node/21559903` [Zacytowano dania: 14-05-2013].

[6] Strona internetowa poświęcona pakietowi ggplot2, `http://docs.ggplot2.org`.

[7] 13th IAAF World Championships in Athletics: IAAF Statistics Handbook. Daegu 2011.

[8] D. Kincaid, W.Cheney. *Numerical Analysis. Mathematics of Scientific Computing*. Tłumaczenie na polski: S.Paszkowski. *Analiza Numeryczna*. Wydawnictwo Naukowo-Techniczne. 71-72.

[9] W. Niemiro. *Statystyka*, 2011.

[10] A. Osękowski. *Wykład z rachunku prawdopodobieństwa I*.

[11] P. Pokarowski, A.Prochenka. *Statystyka II wykłady*, 2012.

[12] C. Ritz, J. C. Streibig. *Nonlinear Regression with R*. Springer, 2011.1-54.

[13] G. A. F. Seber, C. J. Wild. *Nonlinear regression*. Wiley-intescience, 2003. Preface V-VII.

[14] R. Stefani. *Overcoming the doping legacy: Can London's winners outperform the drugs of 1988*. Significance, 9(2):4-8, 2012.

[15] Video Analysis of Sports `http://videosportsanalysis.blogspot.com/2009/08/biomechanical-analysis-of-usain-bolts.htmll` [Zacytowano dnia: 21-06-2013].

[16] Video Analysis of Sports `http://videosportsanalysis.blogspot.com/2009/08/analyzing-usain-bolts-1919-second-200m.html` [Zacytowano dnia: 21-06-2013].

[17] Internetowa encyklopedia *Wikipedia*, `http://en.wikipedia.org/wiki/Athletics_at_the_1912_Summer_Olympics_%E2%80%93_Men%27s_200_metres` [Zacytowano dnia: 11-05-2013].

[18] Internetowa encyklopedia *Wikipedia*, http://en.wikipedia.org/wiki/Athletics_at_the_1920_Summer_Olympics_%E2%80%93_Men%27s_200_metres [Zacytowano dnia: 11-05-2013].

[19] Internetowa encyklopedia *Wikipedia*, http://en.wikipedia.org/wiki/Athletics_at_the_1924_Summer_Olympics_%E2%80%93_Men%27s_200_metres [Zacytowano dnia: 11-05-2013].