Politechnika Krakowska
im. Tadeusza Kościuszki
Wydział Informatyki i Telekomunikacji

Krzysztof Kubina

numer albumu: 138010

ESTYMACJA FUNKCJI REGRESJI W MODELU NIELINIOWYM WRAZ Z ZASTOSOWANIAMI DO DANYCH RZECZYWISTYCH

ESTIMATION OF REGRESSION FUNCTIONS IN A NONLINEAR MODEL WITH APPLICATIONS TO REAL DATA

praca inżynierska

na kierunku Matematyka stosowana

specjalność analiza danych

Praca przygotowana pod kierunkiem:
dr Elżbiety Gajeckiej-Mirek

Kraków, 2025

Spis treści

Wstęp

Estymacja funkcji regresji w modelach nieliniowych jest ważnym zagadnieniem w analizie danych, szczególnie w sytuacjach, gdy zależności między zmiennymi są złożone i trudno je opisać przy użyciu prostych modeli liniowych. W takich przypadkach, konieczne staje się zastosowanie zaawansowanych metod, które pozwalają na dokładniejsze dopasowanie modelu do danych i lepsze uchwycenie ich złożonych relacji.

Celem niniejszej pracy jest omówienie oraz porównanie wybranych zaawansowancyh metod estymacji funkcji regresji w modelu nieliniowym. Treść została podzielona na dwie główne części: teoretyczną i praktyczną.

W części teoretycznej, zostały przedstawione najważniejsze definicje, wykorzystywane w późniejszym etapie pracy. Zaprezentowano również charakterystykę modeli liniowych i nieliniowych, wraz z metodami estymacji ich parametrów. Szczegółowo opisano także metody zastosowane w pracy, takie jak aproksymacja Fouriera oraz metoda estymatorów jądrowych, uwzględniając ich teoretyczne podstawy oraz możliwości praktycznego wykorzystania.

Część praktyczna koncentruje się na implementacji opisanych metod. Zobrazowano wyniki uzyskane przy pomocy każdej z nich, a także omówiono rezultaty zastosowania ich różnych kombinacji.

1. Podtawowe definicje i pojęcia

W niniejszym rozdziale przedstawiono definicje oraz pojęcia, które wykorzystano w późniejszym etapie pracy. Korzystano z literatury: [4], [5] oraz [6].

1.1. Podstawowe definicje

Definicja 1. σ-algebra
Niech Ω będzie dowolnym zbiorem niepustym, zaś $\mathcal{P}(\Omega)$ oznacza rodzinę podzbiorów Ω. Rodzinę $\Sigma \subset \mathcal{P}(\Omega)$ nazywamy σ-algebrą, jeśli spełnione są następujące warunki:

1. $\Sigma \neq \emptyset$,
2. jeśli $A \in \Sigma$, to $A' = \Omega \setminus A \in \Sigma$,
3. jeśli $A_1, A_2, \ldots \in \Sigma$, to $\bigcup_{i=1}^{\infty} A_i \in \Sigma$.

Definicja 2. Miara probabilistyczna
Prawdopodobieństwem (miarą probabilistyczną) nazywamy dowolną funkcję $P : \Sigma \to [0, +\infty)$, która spełnia następujące aksjomaty:

1. $P(\Omega) = 1$ (warunek unormowania),
2. jeśli $A_1, A_2, \ldots \in \Sigma$ parami się wykluczają, to

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) \text{ (przeliczalna addytywność).}$$

Trójkę (Ω, Σ, P) nazywamy *przestrzenią probabilistyczną.*

Definicja 3. Zmienna losowa.
Niech (Ω, Σ, P) będzie przestrzenią probabilistyczną. Funkcję $X : \Omega \to \mathbb{R}^n$ nazywamy zmienną losową, jeśli jest funkcją mierzalną względem σ-algebry Σ, tzn.

$$\forall_{B \in \mathcal{B}(\mathbb{R}^n)} \quad X^{-1}(B) \in \Sigma.$$

Gdzie $\mathcal{B}(\mathbb{R}^n)$ oznacza σ-algebrę zbiorów borelowskich na przestrzeni metrycznej $\mathbb{R}$; $X^{-1}(B)$, to przeciwobraz zbioru B względem X dany wzorem:

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} = \{X \in B\}.$$

Uwaga: Jeśli $\Sigma = \mathcal{P}(\Omega)$, to każde odwzorowanie $X : \Omega \to \mathbb{R}$ jest zmienną losową.

Definicja 4. Rozkład prawdopodobieństwa zmiennej losowej.
Rozkładem prawdopodobieństwa zmiennej losowej $X : \Omega \to \mathbb{R}^n$ nazywamy rozkład P_X taki, że $P_X(B) = P(X^{-1}(B))$ dla $B \in \mathcal{B}(\mathbb{R}^n)$.

Definicja 5. Rozkład prawdopodobieństwa dyskretny.
Rozkład prawdopodobieństwa P nazywamy **dyskretnym**, jeżeli istnieje zbiór $K \in \mathcal{B}(\mathbb{R}^n)$ taki, że:

$$P(K) = 1 \quad \text{oraz} \quad P(\{x\}) > 0 \quad \text{dla każdego } x \in K.$$

Przykładami rozkładów dyskretnych są:

— **Rozkład Dwumianowy:** Rozkład P nazywamy rozkładem dwumianowym (o parametrach n, p), jeżeli istnieją liczby $p \in (0,1), n \in \mathbb{N}$, takie że:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n$$

— **Rozkład Poissona:** Rozkład P nazywamy rozkładem Poissona (z parametrem λ), jeżeli istnieje liczba $\lambda > 0$, taka że:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

Definicja 6. Rozkład prawdopodobieństwa absolutnie ciągły.
Rozkład prawdopodobieństwa P nazywamy **absolutnie ciągłym**, jeżeli jest absolutnie ciągły względem miary Lebesgue'a, tzn.:

$$m(A) = 0 \implies P(A) = 0 \quad \text{dla każdego } A \in \mathcal{B}(\mathbb{R}^n),$$

gdzie m oznacza miarę Lebesgue'a.

Definicja 7. Gęstość rozkładu prawdopodobieństwa.
Jeżeli zmienna losowa $X : \Omega \to \mathbb{R}^n$, gdzie $n = 1$, ma rozkład absolutnie ciągły, to istnieje funkcja $f : \mathbb{R} \to [0, \infty)$, nazywana **gęstością prawdopodobieństwa**, taka że:

$$P(a \leqslant X \leqslant b) = \int_a^b f(x)\, dx,$$

dla dowolnych $a, b \in \mathbb{R}$, gdzie $a \leqslant b$.

Funkcja f spełnia następujące warunki:
1. $f(x) \geqslant 0$ dla każdego $x \in \mathbb{R}$,
2. Warunek normalizacji:

$$\int_{-\infty}^{\infty} f(x)\, dx = 1.$$

Funkcja f opisuje sposób, w jaki prawdopodobieństwo jest rozkładane na zbiorze liczb rzeczywistych i jest podstawą do obliczania prawdopodobieństw dla przedziałów.

Przykład rozkładu ciągłego:
— **Rozkład Normalny:** Rozkład P nazywamy rozkładem normalnym o parametrach m, σ ($m > 0, \sigma \in \mathbb{R}$), jeżeli jego gęstość wyraża się wzorem:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

Jeżeli zmienna losowa X ma rozkład normalny o parametrach m i σ, to oznaczamy go jako: $X \sim N(m, \sigma)$. Standardowy rozkład normalny to N(0,1) wtedy:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, x \in \mathbb{R}$$

Definicja 8. Wartość oczekiwana.
Niech (Ω, Σ, P) będzie przestrzenią probabilistyczną, $X : \Omega \to \mathbb{R}$ będzie zmienną losową. Wartością oczekiwaną zmiennej losowej X o rozkładzie dyskretnym, nazywamy liczbę $\mathbb{E}X$ daną wzorem:

$$\mathbb{E}X = \sum_{i=1}^{n} x_i P(X = x_i).$$

Wartością oczekiwaną zmiennej losowej X o rozkładzie ciągłym oraz o gęstości f nazywamy liczbę $\mathbb{E}X$ daną wzorem:

$$\mathbb{E}X = \int_{-\infty}^{\infty} x f(x)\, dx,$$

o ile całka jest bezwzględnie zbieżna, tzn.

$$\int_{-\infty}^{\infty} |x| f(x)\, dx < \infty.$$

Definicja 9. Wariancja.
Niech (Ω, Σ, P) będzie przestrzenią probabilistyczną, $X : \Omega \to \mathbb{R}$ będzie zmienną losową. Jeśli $\mathbb{E}(X - \mathbb{E}X)^2 < \infty$, to liczbę tę nazywamy wariancją zmiennej losowej X i oznaczamy:

$$var(X) = \mathbb{E}(X - \mathbb{E}X)^2.$$

Definicja 10. Kowariancja.
Niech (Ω, Σ, P) będzie przestrzenią probabilistyczną, $X, Y : \Omega \to \mathbb{R}$ będą zmiennymi losowymi całkowalnymi, takimi, że $\mathbb{E}(|X - Y|) < \infty$. Kowariancją zmiennych losowych X i Y nazywamy liczbę:

$$\mathrm{cov}(X, Y) = \mathbb{E}((X - \mathbb{E}X)(Y - \mathbb{E}Y)).$$

Definicja 11. Autokorelacja.
Niech (Ω, Σ, P) będzie przestrzenią probabilistyczną, $X : \Omega \to \mathbb{R}$ będzie zmienną losową. Autokorelacją nazywamy miarę zależności pomiędzy wartościami tej samej zmiennej losowej X w różnych momentach czasu. Dla opóźnienia k autokorelację definiujemy jako:

$$\mathrm{corr}(X_t, X_{t-k}) = \frac{\mathrm{cov}(X_t, X_{t-k})}{\sqrt{\mathrm{var}(X_t)\,\mathrm{var}(X_{t-k})}},$$

Definicja 12. Korelacja.
Niech (Ω, Σ, P) będzie przestrzenią probabilistyczną, a $X, Y : \Omega \to \mathbb{R}$ będą zmiennymi losowymi. Korelację nazywamy miarą zależności pomiędzy dwiema zmiennymi losowymi X i Y. Korelację definiujemy jako:

$$\mathrm{corr}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)\,\mathrm{var}(Y)}},$$

Współczynnik korelacji mierzy stopień liniowej zależności pomiędzy dwiema zmiennymi losowymi. Przyjmuje wartości z przedziału $[-1, 1]$, gdzie:
— wartość bliższa 1 oznacza silną dodatnią korelację (wzrost jednej zmiennej powoduje wzrost drugiej),
— wartość bliższa -1 oznacza silną ujemną korelację (wzrost jednej zmiennej powoduje spadek drugiej),
— wartość bliska 0 wskazuje na brak liniowej zależności pomiędzy zmiennymi.

1.2. Regresja liniowa

Modele regresyjne służą do analizowania związków przyczynowo-skutkowych między jedną zmienną a inną zmienną lub grupą zmiennych. Zmienna, którą próbuje się przewidzieć lub wyjaśnić, to zmienna objaśniana lub zależna, a te, które są wykorzystywane do jej wyjaśnienia, nazywane są zmiennymi objaśniającymi lub niezależnymi. Gdy analizowana jest relacja z jedną zmienną objaśniającą, mowa jest o regresji prostej, natomiast w przypadku wielu takich zmiennych określa się to jako regresję wieloraką.
W przeciwieństwie do korelacji, która wskazuje jedynie na współzależność między zmiennymi, regresja umożliwia określenie, w jaki sposób zmienne objaśniane wpływają na zmienną objaśniającą.
Model w postaci macierzowej z jedną zmienną objaśniającą wygląda następująco:

$$Y = X\alpha + \xi.$$

lub, w postaci układu równań:

$$y_i = \alpha_0 + \alpha_1 x_i + \xi_i.$$

Regresję liniową z wieloma zmiennymi objaśniającymi można zapisać jako:

$$y_i = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_K x_{Ki} + \xi_i,$$

co odpowiada:

$$y_i = \alpha_0 + \sum_{k=1}^{K} \alpha_k x_{ki} + \xi_i,$$

gdzie
Y to wektor obserwacji zmiennej objaśnianej,
X to macierz obserwacji zmiennych objaśniających (pierwsza kolumna skada się z samych 1, co odpowiada parametrowi α_0)
K to liczba zmiennych objaśniających, k = 1,...,K. $K \in \mathbb{N}$,
n to liczba obserwacji, i=1,...n. $n \in \mathbb{N}$,
y_i to i-ta wartość realizacji zmiennej objaśnianej,
x_{ik} to regresory (realizacje zmiennych objaśniających),
ξ_i to i-ty składnik wektora losowego, $\xi_i \in \mathbb{R}$,
α_k to k-te współczynniki regresji, $\alpha_k \in \mathbb{R}$.

1.2.1. Założenia

Przedstawmy model w ogólnej postaci dla wielu zmiennych objaśniających:

$$y_i = f_i(x_{1i}, x_{2i}, ..., x_{Ki}, \xi_i).$$

W klasycznej analizie regresjii przyjmuje się następujące założenia, które są nazywane założeniami schematu Gaussa-Markowa i definiują tzw. standardowy model liniowy.

Założenie 1. *Model jest niezmienniczy ze względu na obserwacje:*

$$f_1 = f_2 = \cdots = f_i = f,$$

zatem:

$$y_i = f(x_{1i}, x_{2i}, \ldots, x_{Ki}, \xi_i).$$

Założenie 2. *Model jest liniowy względem parametrów:*

$$y = X\alpha + \xi.$$

Założenie 3. *Elementy macierzy X są nielosowe; są ustalone w powtarzalnych próbach, zatem:*

$$\mathbb{E}(y_i \mid x_{1i}, x_{2i}, \ldots, x_{Ki}) = \mathbb{E}(y_i),$$

oraz

$$var(y_i \mid x_{1i}, x_{2i}, \ldots, x_{Ki}) = var(y_i).$$

Jest to tzw. warunek identyfikacji.

Założenie 4. *Rząd macierzy X równy jest liczbie szacowanych parametrów:*

$$r(X) = K.$$

Oznacza to, że macierz X ma pełny rząd kolumnowy, a więc:
- *$n > K$, gdzie n oznacza liczbę obserwacji w danych lub wymiar macierzy X,*
- *żadnej z kolumn macierzy X nie można przedstawić jako kombinacji liniowej kolumn pozostałych; zmienne objaśniające nie są (dokładnie) współliniowe. Nie istnieje zatem taki wektor c_i, różny od wektora zerowego, $c_i \neq 0$, że:*

$$x_i c_i = 0.$$

gdzie

$$c_i = \begin{bmatrix} c_1 & c_2 & \dots & c_k \end{bmatrix}^T,$$

gdzie T oznacza transpozycję.
Zgodnie z tym założeniem zachodzi również:

$$r(X^T X) = K,$$

co oznacza, że macierz $X^T X$ jest nieosobliwa, a więc istnieje macierz do niej odwrotna.

Założenie 5. *Składnik losowy ξ ma n-wymiarowy rozkład normalny:*

$$\xi \sim N(\mathbb{E}(\xi), var(\xi)).$$

Założenie 6. *Wartość oczekiwana składnika losowego jest równa zeru:*

$$\mathbb{E}(\xi) = 0.$$

Założenie 7. *Składnik losowy jest sferyczny:*

$$var(\xi) = \sigma^2 I,$$

gdzie I oznacza macierz jednostkową o wymiarach $n \times n$.
Macierz wariancji-kowariancji składnika losowego $var(\xi)$ jest zatem diagonalna, o takich samych elementach na głównej przekątnej.
Własność tę można również wyrazić następująco: składnik losowy jest homoskedastyczny i nieskorelowany (nie występuje autokorelacja).

Założenie 8. *Informacje zawarte w próbie są jedynymi, na podstawie których estymuje się parametry strukturalne modelu.*

1.2.2. Estymacja parametrów

Jedną z najczęściej stosowanych metod w analizie danych jest MNK (metoda najmniejszych kwadratów). Dzięki niej otrzymuje się estymatory nieobciążone współczynników regresji.
W poniższym akapicie zostało przedstawione MNK dla modelu z jedną zmienną objaśniającą:

$$y_i = \alpha_0 + \alpha_1 x_i + \xi_i,$$

gdzie i=1,..n.
Postać szacowana modelu regresji to:

$$\hat{y_i} = \hat{a_0} + \hat{a_1} x_i, \tag{1.1}$$

gdzie $\hat{y_i}$ - szacowana wartość y_i.
MNK ma na celu oszacowanie parametrów linii regresji w taki sposób aby jak najdokładniej odzwierciedlała dane empiryczne, czyli sprowadza się do znalezienia minimum sumy kwadratów reszt.

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \to \min.$$

Korzystając ze wzoru (1.1), można zredagować powyższą sumę:

$$\sum_{i=1}^{n} (y_i - \hat{a_0} + \hat{a_1} x_i)^2 \to \min.$$

Dana jest zatem funkcja zależna od dwóch parametrów $M(\hat{a_0}, \hat{a_1})$.
W celu znalezienia ekstremum funkcji dwóch zmiennych, pierwszym krokiem jest wyznaczenie gradientu.

$$\nabla f = (\frac{\partial f}{\partial \hat{a_0}}, \frac{\partial f}{\partial \hat{a_1}}),$$

$$\begin{cases} \frac{\partial f}{\partial \hat{a_0}} = -2 \sum\limits_{i=1}^{n} (y_i - \hat{a_0} - \hat{a_1} x_i) \\ \frac{\partial f}{\partial \hat{a_1}} = -2 \sum\limits_{i=1}^{n} (y_i - \hat{a_0} - \hat{a_1} x_i) \end{cases}$$

Punkt dla którego gradient jest równy zero, może być ekstremum tej funkcji.

$$\begin{cases} -2 \sum\limits_{i=1}^{n} (y_i - \hat{a_0} - \hat{a_1} x_i) = 0 \\ -2 \sum\limits_{i=1}^{n} (y_i - \hat{a_0} - \hat{a_1} x_i) x_i = 0 \end{cases}$$

$$\begin{cases} \sum\limits_{i=1}^{n} (y_i) - \sum\limits_{i=1}^{n} (\hat{a_0}) - \sum\limits_{i=1}^{n} (\hat{a_1} x_i) = 0 \\ \sum\limits_{i=1}^{n} (x_i y_i) - \sum\limits_{i=1}^{n} (\hat{a_0} x_i) - \sum\limits_{i=1}^{n} (\hat{a_1} x_i^2) = 0 \end{cases}$$

Punkty ekstremalne są wyznaczane poprzez obliczenie $\hat{a_0}$ i $\hat{a_1}$ z powyższego układu równań. Zauważa się, że:

$$\sum_{i=1}^{n} (y_i) = \bar{y}.$$

Korzystając z tej zależności otrzymano:

$$\begin{cases} n\bar{y} - n\hat{a_0} - n\hat{a_1}\bar{x} = 0 \\ \sum\limits_{i=1}^{n} (x_i y_i) - n\hat{a_0}\bar{x} - \hat{a_1} \sum\limits_{i=1}^{n} (x_i^2) = 0 \end{cases}$$

$$\begin{cases} \hat{a_0} = \bar{y} - \hat{a_1}\bar{x} \\ \sum\limits_{i=1}^{n} (x_i y_i) - n\bar{y}\bar{x} + n\hat{a_1}\bar{x}^2 - \hat{a_1} \sum\limits_{i=1}^{n} (x_i^2) = 0 \end{cases}$$

W drugim równaniu również zachodzą pewne zależności:

$$\sum_{i=1}^{n} (x_i y_i) - n\bar{y}\bar{x} = n\mathrm{cov}(x, y),$$

$$n\hat{a_1}\bar{x}^2 - \hat{a_1} \sum_{i=1}^{n} (x_i^2) = n\hat{a_1}\mathrm{var}(x).$$

Zatem w wyniku wychodzą następujące punkty, które mogą być ekstremum funkcji:

$$\begin{cases} \hat{a_0} = \bar{y} - \hat{a_1}\bar{x} \\ \hat{a_1} = \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)} \end{cases}$$

$$\nabla f = (\bar{y} - \hat{a_1}\bar{x}, \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)}).$$

W celu sprawdzenia, czy dany punkt jest minimum, należy obliczyć Hesjan, czyli macierz pochodnych drugiego rzędu:

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial \hat{a_0}^2} & \frac{\partial^2 f}{\partial \hat{a_0} \partial \hat{a_1}} \\ \frac{\partial^2 f}{\partial \hat{a_1} \partial \hat{a_0}} & \frac{\partial^2 f}{\partial \hat{a_1}^2} \end{bmatrix}$$

$$\frac{\partial^2 f}{\partial \hat{a_0}^2} = 2n$$

$$\frac{\partial^2 f}{\partial \hat{a_0} \partial \hat{a_1}} = \frac{\partial^2 f}{\partial \hat{a_1} \partial \hat{a_0}} = 2\sum_{i=1}^{n}(x_i)$$

$$\frac{\partial^2 f}{\partial \hat{a_1}^2} = 2\sum_{i=1}^{n}(x_i^2)$$

$$H = \begin{bmatrix} 2n & 2\sum\limits_{i=1}^{n}(x_i) \\ 2\sum\limits_{i=1}^{n}(x_i^2) & 2n \end{bmatrix}$$

W kolejnym kroku obliczony został wyznacznik Hesjanu:

$$\det(H) = \left(\frac{\partial^2 f}{\partial \hat{a_0}^2}\right)\left(\frac{\partial^2 f}{\partial \hat{a_1}^2}\right) - \left(\frac{\partial^2 f}{\partial \hat{a_0} \partial \hat{a_1}}\right)^2,$$

$$\det(H) = 4n\sum_{i=1}^{n}(x_i^2) - 4(\sum_{i=1}^{n}(x_i^2))^2.$$

Można zauważyć pewną zależność z wariancją:

$$\mathrm{var}(x) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right).$$

Zatem ten wyznacznik jest równy:

$$\det(H) = 4n\mathrm{var}(x)$$

n jak i var(x) są większe od zera, dodatkowo $2n > 0$

Funkcja jest wypukła i ekstremum lokalne jest minimum globalnym funkcji regresji liniowej. Otrzymano więc za pomocą MNK dwa estymatory, przedstawione wzorami:

$$\hat{a_0} = \bar{y} - \hat{a_1}\bar{x},$$

$$\hat{a_1} = \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)}$$

Twierdzenie 1. *Przy spełnionych założeniach z rozdziału 1.2.1 estymatory MNK $\hat{a_0}$ i $\hat{a_1}$ są najlepsze wśród wszystkich liniowych i nieobciążonych estymatorów, ponieważ mają najmniejszą wariancję.*

Uchylenie tych założeń powoduje, że estymatory uzyskane za pomocą MNK przestają być nieobciażone (czyli wartość oczekiwana estymatora różni się od estymowanego parametru).

1.3. Uogólnione modele regresji liniowej

Uogólnione modele regresji liniowej są szerszą wersją klasycznych modeli liniowych. Główna różnica między nimi widoczna jest w założeniach. Uogólnione modele regresji liniowej nie muszą spełniać klasycznych założeń, między innymi założenia 4 o normalności rozkładu zmiennej zależnej. Pozwalają one na korzystanie z różnych rozkładów prawdopodobieństwa. np.:

— **Rozkład Dwumianowy**,

— **Rozkład Poissona**.

Można to zaliczyć do jednej z trzech głównych cech uogólnionych modeli liniowych. Kolejnymi są funkcja łącząca oraz metody estymacji. Uogólnione modele liniowe korzystają z funkcji łączącej, która przekształca zmienne zależne, w taki sposób, żeby były liniowo związane z zmiennymi niezależnymi. Dodatkowo w ramach estymacji korzysta się z metody największej wiarygodności.

1.3.1. Przykład

Przykładem uogólnionych modeli regresji liniowych jest regresja logistyczna. Jest to metoda statystyczna, która wykorzystywana jest w modelu, sprawdzającym zależności między zmienną objaśnianą, która przyjmuje wartości binarne (0 lub 1), a jedną lub większą liczbą zmiennych objaśniających. Model logistyczny opiera się na funkcji logistycznej (sigmoidalnej), która pozwala przekształcić wartość prognozowaną do przedziału $[0, 1]$.
Funkcja logistyczna definiowana jest jako:

$$p_i = \frac{1}{1 + e^{-z_i}},$$

gdzie $p_i \in [0, 1]$, $z_i = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_k x_{Ki}$, $x_{1i}, x_{2i}, \ldots, x_{Ki}$, i = 1,...,n, są to obserwacje zmiennych objaśniających. $\alpha_0, \alpha_1, \ldots, \alpha_K$ to parametry modelu.
Regresja logistyczna jest często wyrażana w formie funkcji logit:

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1 - p_i}\right) = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_k x_{Ki}.$$

Parametry $\alpha_0, \alpha_1, \ldots, \alpha_K$ są estymowane za pomocą metody maksymalnego prawdopodobieństwa, która polega na maksymalizacji funkcji wiarygodności:

$$L(\alpha) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i},$$

gdzie y_i to obserwacje zmiennej zależnej, p_i to przewidywane prawdopodobieństwo dla obserwacji i.

1.4. Regresja nieliniowa

Regresja nieliniowa to metoda pozwalająca na modelowanie nieliniowych zależności pomiędzy zmienną zależną a zestawem zmiennych niezależnych. W przeciwieństwie do klasycznej regresji liniowej, ograniczonej do modeli liniowych, regresja nieliniowa umożliwia estymację modeli o bardziej złożonych relacjach między zmiennymi. Proces ten opiera się na zastosowaniu iteracyjnych algorytmów estymacji. Taki model (z jedną zmienną objaśniającą) w postaci układu równań można wyrazić jako:

$$y_i = h(x_{Ki}, \boldsymbol{\alpha}) + \xi_i, \tag{1.2}$$

gdzie
h to funkcja różnowartościowa, ciągła i dwukrotnie różniczkowalna,
x_{Ki} to wektor realizacji zmiennych niezależnych,
$\boldsymbol{\alpha}$ to wektor parametrów modelu,
ξ_i to składnik losowy opisujący błąd modelu.

1.4.1. Rodzaje regresji nieliniowej

Regresję nieliniową można podzielić na trzy rodzaje.

— **Rodzaj pierwszy: Modele liniowe względem parametrów**
Pierwszym rodzajem są modele, które są nieliniowe względem zmiennych niezależnych, a liniowe względem parametrów. Te modele można łatwo przekształcić do formy liniowej za pomocą odpowiednich podstawień.
Przykład 1. Przykładowym modelem może być:

$$y = \alpha_0 + \alpha_1\sqrt{x_1} + \alpha_2 x_2^2 + \frac{\alpha_3}{x_3} + \xi$$

Aby go zlinearyzować, wprowadzono następujące podstawienia:

$$x_1' = \sqrt{x_1}, \quad x_2' = x_2^2, \quad x_3' = \frac{1}{x_3}$$

Otrzymano wtedy liniowy model:

$$y = \alpha_0 + \alpha_1 x_1' + \alpha_2 x_2' + \alpha_3 x_3' + \xi$$

— **Rodzaj drugi: Modele transformowalne do postaci liniowej**
Drugim rodzajem są modele nieliniowe względem zmiennych niezależnych i parametrów, które można sprowadzić do postaci liniowej dzięki odpowiednim transformacjom.
Przykład 2. Przykładem jest model logarytmiczny:

$$y = a_0 \cdot x_1^{a_1} \cdot e^{a_2 x_2} + \xi$$

Logarytmując obustronnie, otrzymano:

$$\ln y = \ln \alpha_0 + \alpha_1 \ln x_1 + \alpha_2 x_2 + \xi$$

W kolejnym kroku wprowadzono nowe zmienne:

$$Y' = \ln y, \quad X_1' = \ln x_1, \quad X_2' = x_2$$

Finalnie otrzymano postać liniową:

$$Y' = \ln \alpha_0 + \alpha_1 X_1' + \alpha_2 X_2' + \xi$$

— **Rodzaj trzeci: Modele nieliniowe niesprowadzalne do postaci liniowej**
Trzeci rodzaj obejmuje modele, które są nieliniowe zarówno względem zmiennych niezależnych, jak i parametrów, których dodatkowo nie można sprowadzić do postaci liniowej poprzez transformację. Tym różnią się od modeli z rodzaju drugiego.
Przykład 3. Przykładem może być model sigmoidalny:

$$y = \frac{\alpha_0}{1 + e^{-\alpha_1(x_1 - \alpha_2)}} + \xi$$

Tego typu modelu nie da się sprowadzić do postaci liniowej, ponieważ zależności między zmiennymi i parametrami mają charakter nieliniowy, niezależny od transformacji.

2. Model nieliniowy

2.1. Aproksymacja

Materiał przedstawiony w tym rozdziale został przygotowany na podstawie [1],[2], [3] oraz [5]

2.1.1. Nieliniowa metoda najmniejszych kwadratów - metody iteracyjne

Nieliniowe estymatory metody najmniejszych kwadratów (NMNK, ang. *nonlinear least squares*) można wyznaczyć analogicznie do estymatorów liniowych, minimalizując sumę kwadratów błędów. Po podstawieniu (1.2) otrzymano:

$$S(\boldsymbol{\alpha}_k) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - h(\mathbf{x}_{ki}, \boldsymbol{\alpha}_k)\right]^2 .$$

Alternatywnie, w zapisie macierzowym:

$$S(\boldsymbol{\alpha}_k) = \mathbf{e}^\top \mathbf{e} = \left[\mathbf{y} - \mathbf{h}\right]^\top \left[\mathbf{y} - \mathbf{h}\right].$$

gdzie

$$\mathbf{y} = [y_1, \ldots, y_n]^\top, \quad \mathbf{e} = [e_1, \ldots, e_n]^\top,$$
$$\mathbf{h} = [h(x_{kn}, \alpha_k), \ldots, h(x_{kn}, \alpha_k)]^\top.$$

x_{ki} oznacza i-tą obserwację wektora cech zmiennych objaśniających,
n to liczba obserwacji, $n \in \mathbb{N}$,
i = 1, . . . ,n. $i \in \mathbb{N}$,
k = 1, . . . ,n. $k \in \mathbb{N}$.
Warunkiem koniecznym dla ekstremum funkcji jest równość zeru jej pochodnej:

$$\frac{\partial S(\boldsymbol{\alpha}_k)}{\partial \boldsymbol{\alpha}_k} = 0.$$

Z powyższego wynika:

$$-2 \sum_{i=1}^{n} \left[y_i - h(x_{ki}, \alpha_k)\right] \frac{\partial h(x_{ki}, \alpha_k)}{\partial \alpha_k} = 0.$$

Rozwiązanie tego układu równań (zazwyczaj nieliniowych) wymaga stosowania metod iteracyjnych, ponieważ brakuje analitycznego rozwiązania. Proces iteracyjny rozpoczyna się od wektora początkowego α_k^0 i dąży do wyznaczenia ciągu ocen $\alpha^0, \alpha^1, \ldots, \alpha^n, \ldots$, które spełniają warunek:

$$S(\alpha_k^{n+1}) < S(\alpha_k^n).$$

Poprawa przybliżenia ocen parametrów odbywa się przez dodanie pewnego wektora $\mathbf{v}_k^n$ (krok):

$$\alpha_k^{n+1} = \alpha_k^n + \mathbf{v}_k^n. \tag{2.1}$$

Zatem

$$S(\alpha_k^n + \mathbf{v}_k^n) < S(\alpha_k^n). \tag{2.2}$$

Wektor $\mathbf{v}_k$ można wyrazić jako iloczyn kierunku $\mathbf{d}_k$ oraz długości kroku l (Długością kroku l może być dowolna stała):

$$\mathbf{v}_k = l \cdot \mathbf{d}_k,$$

Wstawiając to do wzorów (2.2) i (2.1) otrzymano:

$$S(\alpha_k + l \cdot \mathbf{d}_k) < S(\alpha_k).$$

Poszukiwany jest kierunek $\mathbf{d}_k$, dla którego funkcja $S(\alpha_k + l \cdot \mathbf{d}_k)$ maleje wzdłuż l, dla l dostatecznie bliskiego 0. Rozważając pochodną względem l:

$$\frac{\partial S(\alpha_k + l \cdot \mathbf{d}_k)}{\partial l} = \left[\frac{\partial S}{\partial \alpha_k}\right]^\top \frac{\partial(\alpha_k + l \cdot \mathbf{d}_k)}{\partial l} < 0.$$

Gradient funkcji kryterium jest oznaczany jako:

$$\mathbf{g}_k = \frac{\partial S}{\partial \alpha_k},$$

co prowadzi do nierówności:

$$\mathbf{g}_k^\top \mathbf{d}_k < 0.$$

Kierunek kroku jest przyjmowany jako:

$$\mathbf{d}_k = -\mathbf{P}\mathbf{g}_k,$$

gdzie $\mathbf{P}$ jest dodatnio określoną macierzą kwadratową stopnia K, co zapewnia, że funkcja kryterium zmniejsza się w kierunku gradientu w każdym kroku iteracyjnym.
Iteracyjny wzór obliczeń przyjmuje postać:

$$\alpha_k^{n+1} = \alpha_k^n - l^n \mathbf{P}^n \mathbf{g}_k,$$

Kryteria zakończenia procesu iteracyjnego to:
1. Stabilizacja wartości funkcji kryterium:

$$|S(\alpha_k^{n+1}) - S(\alpha_k^n)| < \epsilon,$$

   gdzie ϵ jest złożonym poziomem dokładności, nazywanym niekiedy wsółczynnikiem dokładnosci. Ozacza to, że długość kroku powinna być tak dobrana aby

$$S(\alpha_k^{n+1} - l^n \mathbf{P}^n \mathbf{g}_k) < S(\alpha_k^n),$$

2. Stabilizacja wartości parametrów:

$$|\alpha_k^{n+1} - \alpha_k^n| < \xi, \quad k = 1, \ldots, K.$$

Przykładem jest metoda najszybszego spadku, gdzie $\mathbf{P}^n$ to macierz jednostkowa:

$$\mathbf{P}^n = \mathbf{I}. \tag{2.3}$$

2.1.2. Metody rozwinięć w szereg Fouriera

Definicja 13. Szeregiem trygonometrycznym lub szeregiem Fouriera nazywamy szereg funkcyjny postaci:

$$\frac{a_0}{2} + \sum_{n=1}^{n} (a_n cosnx + b_n sinnx),$$

gdzie $(a_n)_{n\geqslant 0}$ i $(b_n)_{n\geqslant 0}$ są danymi ciągami liczbowymi.

Twierdzenie 2. *Jeśli całkowalna funkcja okresowa $f : \mathbb{R} \to \mathbb{R}$ jest sumą jednostajnie zbieżnego szeregu trygonometrycznego*

$$\frac{a_0}{2} + \sum_{n=1}^{n} (a_n cosnx + b_n sinnx)$$

to współczynniki tego szeregu dane są wzorami

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\,dx,$$
$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cdot\cos nx\,dx \quad dla\ n = 1,2,\ldots,$$
$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cdot\sin nx\,dx \quad dla\ n = 1,2,\ldots.$$

W praktyce, ze względu na ograniczoną liczbę obserwacji $y_1, y_2, \ldots, y_n$, nie jest możliwe dokładne oszacowanie nieskończonej liczby współczynników Fouriera. Dlatego proces budowy estymatora funkcji regresji przebiega w następujący sposób:

1. Wyznaczenie parametru l_n, czyli punktu odcięcia szeregu Fouriera. Parametr l_n jest wybierany tak, aby zminimalizować błąd średniokwadratowy estymatora. Na dużych próbach l_n przyjmuje wartość w przybliżeniu proporcjonalną do $\sqrt{n}$:

   $$l_n \approx c\sqrt{n},$$

   gdzie stała c zależy od właściwości funkcji f i rozkładu błędu. Jej wartość można określić za pomocą wzoru:

   $$c = \frac{|f(\pi) - f(-\pi)|}{2\pi\sigma},$$

   gdzie:
   $f(\pi)$ i $f(-\pi)$ to wartości funkcji na końcach przedziału, opisujące jej zgodność lub różnicę na brzegach,
   σ^2 oznacza wariancję błędu pomiarowego.
2. Wyznaczenie współczynników Fouriera $\hat{a}_j$ i $\hat{b}_j$ na podstawie próby $y_1, \ldots, y_n$, zgodnie ze wzorami:

   $$\hat{a}_j = \frac{1}{\pi}\sum_{i=1}^{n} y_i \cos jx, \quad \hat{b}_j = \frac{1}{\pi}\sum_{i=1}^{n} y_i \sin jx, \quad j = 1,\ldots,l_n.$$

3. Konstrukcja estymatora funkcji regresji w postaci obciętego szeregu Fouriera:

   $$f_n(x) = \frac{\hat{a}_0}{2} + \sum_{j=1}^{l_n}\left(\hat{a}_j \cos jx + \hat{b}_j \sin jx\right), \quad t \in \langle -\pi, \pi\rangle.$$

Estymator $f_n(x)$, zwany *estymatorem Fouriera funkcji regresji*, jest przybliżeniem funkcji $f(x)$. Właściwości estymatora zależą od wybranego l_n oraz od spełnienia warunku zgodności wartości funkcji na końcach przedziału, tj. $f(-\pi) = f(\pi)$. W przypadku dużych odchyleń wariancja estymatora może być redukowana poprzez dodatkowe pomiary na brzegach przedziału.

2.1.3. Metoda estymatorów jądrowych

Estymacja jądrowa stosowana jest między innymi do estymacji gęstości rozkładu oraz funkcji regresji w sytuacjach, gdy brak jest możliwości założenia konkretnego kształtu modelu analitycznego. Metoda ta opiera się na funkcji jądra $K(u)$, która realizuje lokalne wygładzanie danych, oraz parametrze szerokości pasma h, kontrolującym poziom wygładzenia.

Definicja 14. Estymator jądrowy
Podstawowy estymator jądrowy funkcji gęstości $f(x)$ definiuje się jako:

$$f_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

gdzie
n to liczba obserwacji,

$h > 0$ to szerokość pasma (parametr wygładzający),
$K(u)$ to funkcja jądra spełniająca warunek normalizacji:

$$\int_{-\infty}^{\infty} K(u)\,du = 1.$$

Przykłady Funkcji Jądra Najczęściej stosowane funkcje jądra $K(u)$ to:

— **Jądro Gaussowskie**:

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right),$$

cechujące się ciągłością i nieskończonym zakresem wsparcia.

— **Jądro Epanechnikowa**:

$$K(u) = \frac{3}{4}(1 - u^2), \quad \text{dla } |u| \leqslant 1,$$

optymalne w sensie minimalizacji wariancji błędu estymacji.

— **Jądro Prostokątne**:

$$K(u) = \begin{cases} \frac{1}{2}, & \text{dla } |u| \leqslant 1, \\ 0, & \text{dla } |u| > 1, \end{cases}$$

które jest proste w implementacji, ale mniej efektywne w praktyce.

2.1.4. Estymatory jądrowe funkcji regresji

W celu oszacowania funkcji regresji $f(t)$, dla zbioru obserwacji $(x_1, y_1), \ldots, (x_n, y_n)$, przyjmuje się model:

$$y_i = f(x_i) + \xi_i, \quad i = 1, \ldots, n,$$

gdzie ξ_i to niezależne błędy o wartości oczekiwanej równej zero i skończonej wariancji.

2.1.4.1. Podstawowe założenia

Estymatory jądrowe zakładają, że większy wpływ na oszacowanie funkcji regresji w punkcie x mają obserwacje znajdujące się bliżej tego punktu. Do opisu wpływu tych obserwacji wykorzystuje się funkcję wagową $K(u)$, zwaną jądrem estymatora. Parametr $h > 0$, określany jako szerokość pasma (ang. *bandwidth*), kontroluje zakres wpływu punktów odległych od x.
Wyróżnia się dwa główne estymatory jądrowe funkcji regresji:

Definicja 15. Estymator Gassera-Müllera

$$f_n(t) = \frac{1}{h} \sum_{i=1}^{n} y_i \int_{\xi_i}^{\xi_{i+1}} K\left(\frac{t-u}{h}\right) du,$$

gdzie $\xi_0 = a, \xi_{n+1} = b$, zaś $\xi_i = \frac{x_i + x_{i+1}}{2}$ dla $i = 1, 2, \ldots, n-1$. Jądro $K(u)$ jest funkcją symetryczną, a $h > 0$ to szerokość pasma.

Definicja 16. Estymator Nadayai-Watsona

$$f_n(t) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{t-x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{t-x_i}{h}\right)},$$

gdzie licznik odpowiada ważonej sumie wartości Y_i, a mianownik jest sumą wag. Wartość $f_n(t)$ w punkcie $t = x_i$ jest równa wartości Y_i, co sprawia, że estymator jest zgodny z obserwacjami.

2.1.4.2. Błąd estymatora i dobór szerokości pasma

Jakość estymacji funkcji regresji przy użyciu estymatorów jądrowych jest oceniana za pomocą błędu średniokwadratowego:

$$R(f_n) = \frac{1}{4}h^4 C_1^2 \int (f''(t))^2 dt + \frac{\sigma^2 C_2}{nh} \int \frac{1}{g(x)} dx + o\left(h^4 + \frac{1}{nh}\right),$$

gdzie
$C_1 = \int u^2 K(u) du$,
$C_2 = \int K(u)^2 du$,
σ^2 to wariancja błędu estymacji,
$g(x)$ to funkcja gęstości rozkładu zmiennej losowej x.

Dobór szerokości pasma h jest kluczowy dla jakości estymatora:
— Małe wartości h prowadzą do niestabilności estymatora, ponieważ wynik jest nadmiernie dopasowany do danych (overfitting).
— Duże wartości h powodują zbyt mocne wygładzenie, co może prowadzić do utraty szczegółowych informacji o funkcji (underfitting).
Optymalna szerokość pasma h może być obliczona jako:

$$h = c \cdot n^{-1/5},$$

gdzie stała c zależy od funkcji jądra K, szerokości przedziału $g(x)$ oraz innych parametrów danych. W szczególnych przypadkach, takich jak rozkład normalny danych, wzór ten odpowiada tzw. regule Silvermana, w której stała c jest równa $1.06 \cdot \sigma$, gdzie σ to odchylenie standardowe zmiennych.

Estymatory Gassera-Müllera i Nadayai-Watsona mają różne zalety i ograniczenia:
— Estymator Nadayai-Watsona jest łatwiejszy w implementacji, ale może mieć wysokie błędy na brzegach przedziału.
— Estymator Gassera-Müllera jest bardziej skomplikowany, ale lepiej radzi sobie z oszacowaniem funkcji na brzegach, dzięki zastosowaniu modyfikacji jądra.
Estymatory jądrowe są skutecznym narzędziem do estymacji funkcji regresji, zwłaszcza gdy nie zakłada się konkretnego kształtu funkcji. Ostateczna jakość estymacji zależy od doboru jądra K oraz szerokości pasma h. W praktyce szerokość pasma h jest dobierana iteracyjnie lub na podstawie minimalizacji błędu średniokwadratowego.

2.2. Klasyfikacja oceny modelu

2.2.1. Współczynnik determinacji R^2

Definicja 17. W analizie regresji, suma całkowita kwadratów (*Total Sum of Squares*) to:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Można również zapisać ją w postaci rozbicia na składniki:

$$SST = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = SSE + SSR,$$

gdzie
— SSE to suma kwadratów błędów (*Error Sum of Squares*),
— SSR to suma kwadratów regresji (*Regression Sum of Squares*).

Definicja 18. Suma kwadratów regresji SSR może być interpretowana jako indeks zmienności wartości przewidywanych $\hat{y}_i$ wokół średniej wartości $\hat{y}$, natomiast suma kwadratów błędów SSE reprezentuje zmienność reszt wokół wartości średniej równej 0.
Dzięki temu można wyprowadzić współczynnik determinacji R^2, który opisuje proporcję zmienności wyjaśnionej przez model:

$$R^2 = \frac{SSR}{SST},$$

co odpowiada stosunkowi:

$$R^2 = \frac{\text{zmienność wyjaśniona przez model}}{\text{zmienność całkowita}}.$$

Stwierdzenie to oznacza, że wartość R^2 jest ściśle związana z wartością współczynnika korelacji próbki.

2.2.2. MSE i RMSE

Definicja 19. MSE (średni błąd kwadratowy) to średnia z kwadratów różnic między wartościami rzeczywistymi a przewidywanymi przez model. Wyraża, jak daleko przewidywania modelu są od rzeczywistych wyników.

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

gdzie $n \in \mathbb{N}$ to liczba obserwacji, $y_i \in \mathbb{R}$ to obserwowana wartość zmiennej objaśnianej, $\hat{y}_i \in \mathbb{R}$ to przewidywana wartość zmiennej objaśnianej przez model, $\text{MSE} \in \mathbb{R}$.

Definicja 20. RMSE (pierwiastek średnich kwadratów błędów) jest pierwiastkiem kwadratowym z MSE, co pozwala interpretować wyniki w jednostkach oryginalnych danych.

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

3. Zastosowanie

W pracy wykorzystano dane meteorologiczne występujące we Włoszech na przestrzeni 70 lat, od 1950 roku do 2022 roku. Zebrane zostały z Portalu Wiedzy o Zmianach Klimatu (Climate Change Knowledge Portal), jest to platforma stworzona przez Bank Światowy (World Bank Group), która umożliwia dostęp do danych i analiz dotyczących zmian klimatycznych na świecie ([7]). Wykorzystując je przedstawiono i przeanalizowano kilka opcji estymacji funkcji regresji w modelu nieliniowym, opisanych w poprzednich rozdziałach. Do analizy wykorzystano program R-studio, w którym skorzystano z pakietów: „readxl",„tidyverse",„ggplot2", „GGally", „KernSmooth", „np" i „lmtest".

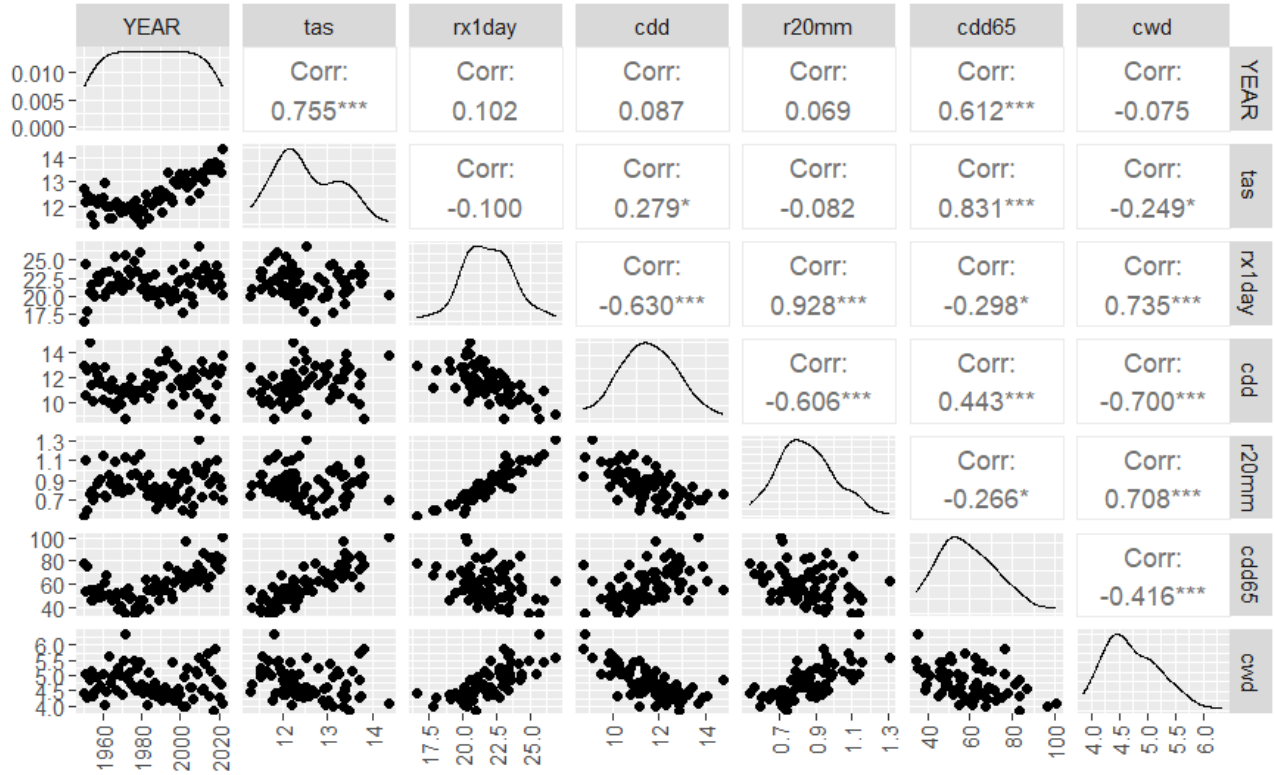
Zebrane zostały następujące dane:
— „tas" to średnia temperatura powietrza na poziomie powierzchni ziemi, zazwyczaj w stopniach Celsjusza.
— „rx1day" to średnia największa ilość opadów, jakie miały miejsce w ciągu jednego dnia w danym miesiącu. Zmienna ta wskazuje, jak ekstremalne były opady deszczu w poszczególnych dniach.
— „r20mm" to liczba dni w miesiącu, w których opady przekraczały 20 mm.
— „tr29" to liczba tropikalnych nocy, czyli dni, w których minimalna temperatura nie spadła poniżej 29°C.
— „cdd" to maksymalna liczba kolejnych dni suchych w danym miesiącu.
— „cdd65" to wskaźnik, który mierzy ilość energii potrzebnej do chłodzenia pomieszczeń w ciągu roku, oparty na temperaturze przekraczającej 65°F.
— „cwd" to maksymalna liczba kolejnych dni mokrych (z opadami) w danym miesiącu.
— „hd30" to liczba gorących dni, czyli dni, w których maksymalna temperatura przekracza 30°C.
— „YEAR" to dany rok dla danej analizy (od 1950 roku do 2020 roku)

W poniższej analizie, zmienną objaśnianą była zmienna „tas". Zastosowane poniżej modele przedstawiają jaki wpływ na średnią temperaturę powietrza miały inne czynniki lub jak zmienna „tas" zmieniała się na przestrzeni lat („YEAR").

3.1. Aproksymacja Fouriera

Jako pierwszą metodę estymacji zastosowana została aproksymacja danych za pomocą szeregu Fouriera, wykorzystując nieliniową metodę najmniejszych kwadratów. Aby móc z tej metody skorzystać zaimplementowana została funkcja fouriera z której skorzystano w modelu. Kolejnym potrzebnym krokiem było znormalizowanie zmiennych objaśniających do zakresu $[-\pi, \pi]$. Dodatkowo ustalono liczbę harmonicznych (częstotliwość) zgodnie z metodą opisaną w 2.1.2. Po ustaleniu liczby harmonicznych i znormalizowaniu danych pozostało jedynie dokonanie wyboru zmiennych objaśniających.

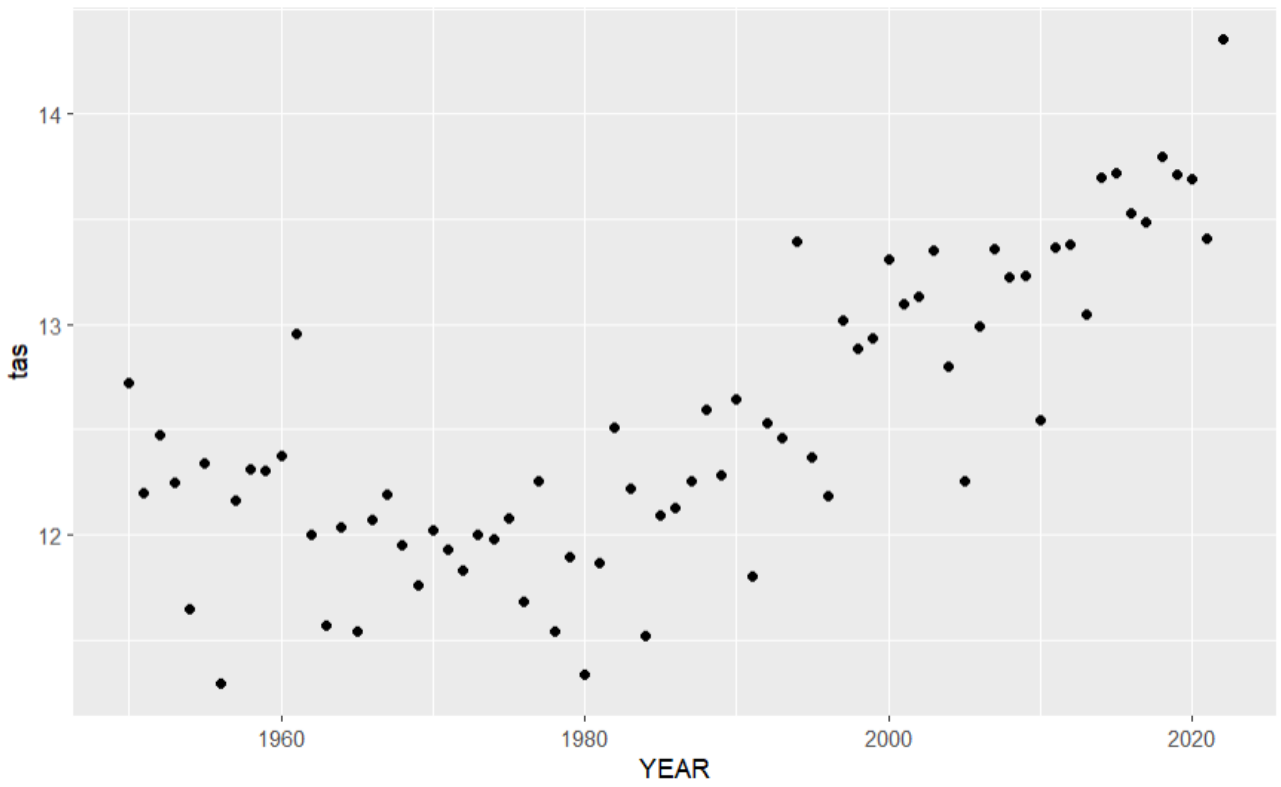
Podstawowym założeniem koniecznym do zastosowania opisanych wcześniej metod estymacji jest obecność nieliniowej zależności w danych. Pierwszym krokiem w procesie wyboru zmiennych było zwizualizowanie danych, analiza korelacji między nimi oraz ocena rozkładu za pomocą wykresów punktowych. Za pomocą biblioteki „GGally" wszystkie te elementy zostały połączone w jeden wykres (rysunek 3.1)



Rysunek 3.1. Wykresy i korelacja między zmiennymi. Źródło: opracowanie własne.

3.1.1. Model pierwszy

Najwyższą korelacje ze zmienną „tas" ma zmienna „cdd65" (0,831) z trzema gwiazdkami (***) – oznacza to, że związek między nimi jest istotny statystycznie ($p \leqslant 0.001$). Świadczy to o małym prawdopodobieństwie, aby tak wysoka korelacja wystąpiła losowo. Wykres punktowy pokazuje zależność liniową między nimi. Zatem odrzucamy zmienną „cdd65".
Kolejną zmienną z największą korelacją jest „YEAR". Korelacja między nimi wynosi 0,755 z trzema gwiazdkami (***). Wykres punktowy przedstawiony na rysunku 3.2 jednak sugeruje na związek nieliniowy.



Rysunek 3.2. Wykres punktowy między „tas” a „YEAR”. Źródło: opracowanie własne.

Zatem nie wykazano przeciwwskazań do odrzucenia tej zmiennej. Wykorzystano więc tą zmienną jako zmienną objaśniającą w pierwszym modelu.
Po dostosowaniu zmiennej „YEAR” do modelu w podsumowaniu widocznym na rysunku 3.3 otrzymano:

```
Formula: tas ~ fourier_multi(x_list2, params, n_harmonics)

Parameters:
        Estimate Std. Error t value Pr(>|t|)
params1 12.52488    0.04720 265.330  < 2e-16 ***
params2  0.65390    0.06717   9.734 3.01e-14 ***
params3 -0.38486    0.06634  -5.801 2.21e-07 ***
params4 -0.14487    0.06717  -2.157  0.03479 *
params5  0.04928    0.06634   0.743  0.46033
params6  0.20192    0.06717   3.006  0.00378 **
params7 -0.08378    0.06634  -1.263  0.21121
params8 -0.17638    0.06717  -2.626  0.01080 *
params9  0.01339    0.06634   0.202  0.84071
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.403 on 64 degrees of freedom

Algorithm "port", convergence message: both X-convergence and relative convergence (5)
```

Rysunek 3.3. Podsumowanie modelu pierwszego. Źródło: opracowanie własne.

Korzystając z rysunku 3.3 a dokładniej z podanych parametrów, można zauważyć, jak wyglądały parametry funkcji aproksymacji Fouriera:

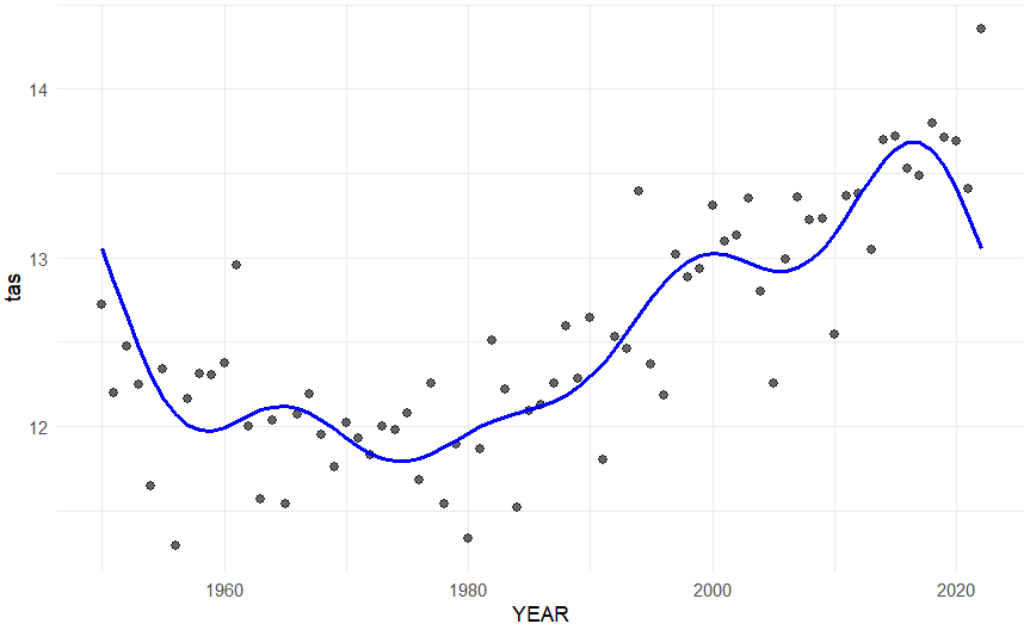
$$a_0 = 12,52488$$

$$a_1 = 0,65390, \quad b_1 = -0,38486$$
$$a_2 = -0,14487, \quad b_2 = 0,04928$$
$$a_3 = 0,20192, \quad b_3 = -0,08378$$
$$a_4 = -0,17638, \quad b_4 = 0,01339$$

Z tego wynika, że liczba harmonicznych (częstotliwość) wyniosła 4. Wzór funkcji aproksymacji Fouriera po podstawieniu wartości parametrów w modelu pierwszym wygląda zatem następująco:

$$\text{tas}(x) = 12,52488 + 0,65390 \cdot \sin(x) - 0,38486 \cdot \cos(x)$$
$$-0,14487 \cdot \sin(2x) + 0,04928 \cdot \cos(2x)$$
$$+0,20192 \cdot \sin(3x) - 0,08378 \cdot \cos(3x)$$
$$-0,17638 \cdot \sin(4x) + 0,01339 \cdot \cos(4x)$$

gdzie x to zmienna „YEAR".



Rysunek 3.4. Wykes modelu pierwszego. Źródło: opracowanie własne.

Analizując rysunek 3.4, można stwierdzić, że metoda estymacji danych przy użyciu szeregu Fouriera wykazuje dobre dopasowanie do danych. Linia funkcji regresji wydaje się złapać właściwe rozłożenie danych. Aby jednak potwierdzić tę obserwację, przeprowadzono dodatkową ocenę modelu za pomocą odpowiednich miar klasyfikacyjnych.

```
"MSE: 0.1424"
"RMSE: 0.3774"
"R^2: 0.7037"
```

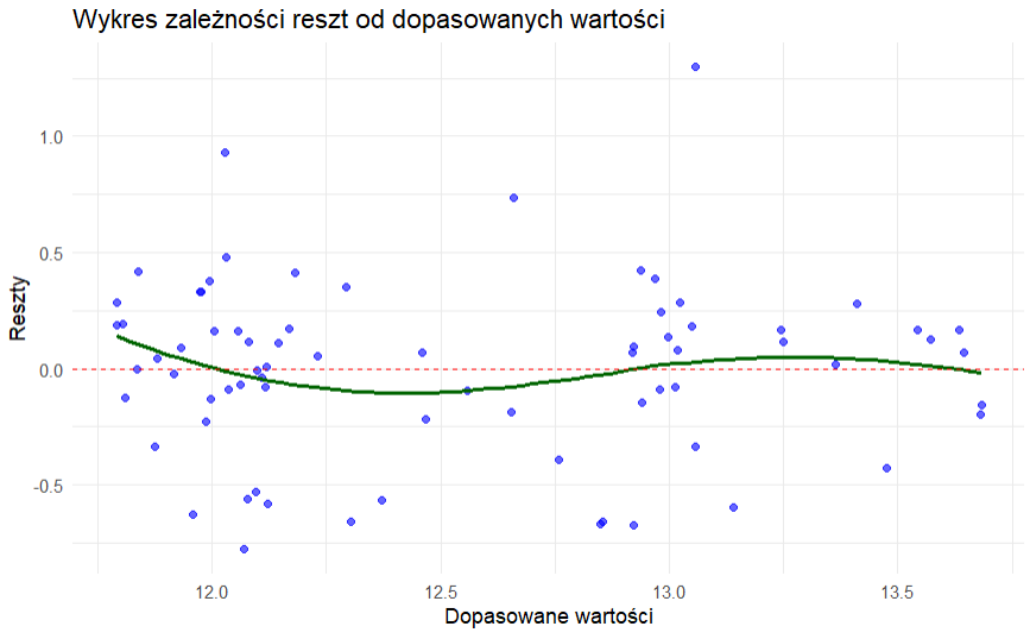
Rysunek 3.5. Ocena modelu pierwszego. Źródło: opracowanie własne.

Z wyników przedstawionych na rysunku 3.5 można potwierdzić, że model aproksymacji Fouriera jest dobrze dopasowany do danych. Niskie wartości MSE i RMSE wskazują na dobrą precyzję przewidywań, a wysokie R^2 sugeruje, że model skutecznie wyjaśnia większość zmiennej „tas". MSE, RMSE

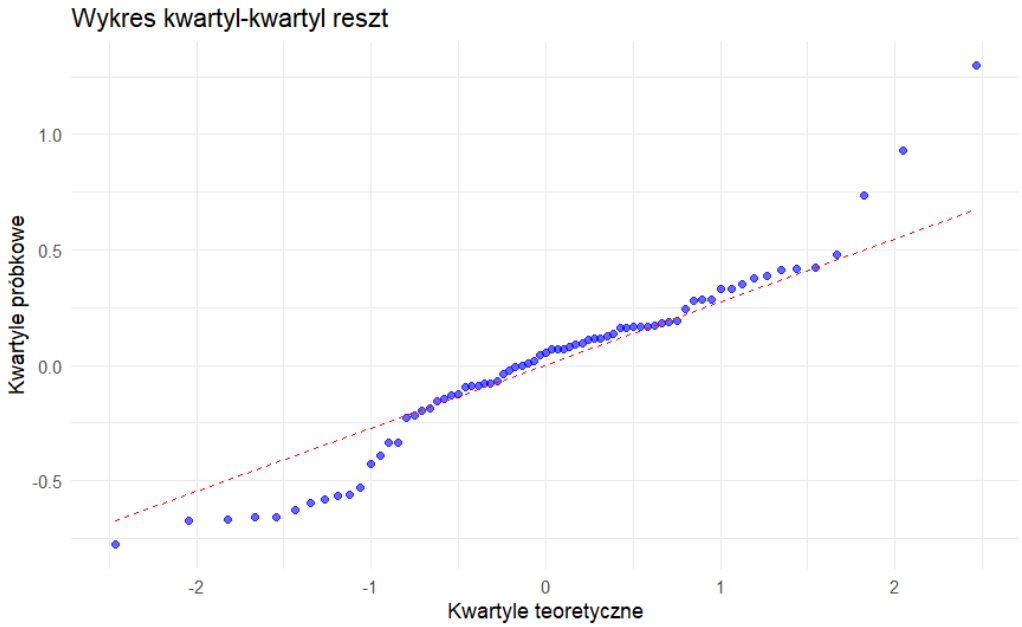
i R^2 zapewniają ogólną ocenę dopasowania modelu nieliniowego, jednak dokładna analiza reszt jest ważna, aby sprawdzić, czy model poprawnie opisuje nieliniowe zależności. Wykres zależności reszt od dopasowanych wartości, wykres kwartyl-kwartyl reszt (QQplot) oraz testy umożliwiają ocenę niezależności reszt, ich losowości rozkładu, normalności rozkładu i zgodności z założeniami modelu.

Wykres zależności reszt od dopasowanych wartości przedstawia zależność reszt od dopasowanych wartości, co pozwala na ocenę, czy model dobrze pasuje do danych. Na takim wykresie punkty powinny być rozmieszczone losowo względem linii $y = 0$. Zielona linia jest wygładzonym odzwierciedleniem średnich reszt w zależności od dopasowanych wartości.
Na poniższym wykresie, widocznym na rysunku 3.6 większość punktów jest rozłożona w sposób losowy względem linii $y = 0$. Potwierdza to zielona linia, która również jest bardzo zbliżona do prostej $y = 0$ Można więc wstępnie stwierdzić, że model dobrze pasuje do danych i nie zawiera systematycznych błędów.

Z kolei wykres kwartyl-kwartyl reszt służy do oceny, czy reszty mają rozkład normalny. Aby reszty spełniały to założenie, punkty na wykresie powinny układać się wzdłuż czerwonej linii odniesienia. Na wykresie, widocznym na rysunku 3.7 można zauważyć spore odchylenia punktów, szczególnie na końcach przedziału. Może to sugerować brak normalności reszt.



Rysunek 3.6. Wykres zależności reszt od dopasowanych wartości modelu pierwszego. Źródło: opracowanie własne.



Rysunek 3.7. Wykres kwartyl-kwartyl reszt modelu pierwszego. Źródło: opracowanie własne.

Aby potwierdzić tezy wyciągnięte z wykresów wykonano dodatkowo dwa testy: test Shapiro-Wilka oraz test Durbin-Watsona.

```
        Shapiro-Wilk normality test

data:  ITA_TABELA$residuals1
W = 0.95347, p-value = 0.008945


        Durbin-Watson test

data:  tas ~ taspred
DW = 1.6794, p-value = 0.06612
```

Rysunek 3.8. Testy dla modelu pierwszego. Źródło: opracowanie własne.

Wynik $DW = 2$ testu Durbin-Watsona wskazuje na brak korelacji reszt. W tym przypadku wartość $DW = 1.6794, p > 0.05$ testu Durbin-Watsona wskazuje, że reszty są bliskie braku korelacji, co oznacza, że model w dużym stopniu poprawnie odzwierciedla zależności w danych.
W teście Shapiro-Wilka ze względu na $p < 0.05$, hipoteza zerowa dotycząca normalności reszt została odrzucona, co wskazuje, że reszty nie mają rozkładu normalnego. W przypadku estymacji funkcji regresji brak normalności reszt nie stanowi koniecznie problemu, dodatkowo podsumowując wszystkie przeprowadzone oceny, można wywnioskować, że model poprawnie estymuje dane.
Jednak możliwe jest ulepszenie modelu poprzez na przykład uwzględnienie dodatkowych zmiennych objaśniających.

3.1.2. Model drugi

W modelu drugim wykorzystano trzy zmienne objaśniające: („rx1day”,„YEAR”,„cdd”). Jak już wyżej zostało wspomniane, głównym założeniem było, aby model był nieliniowy. Przy wyborze tych zmiennych usunięte zostały zmienne dla których istnieje liniowa zależnoś. Korzystając z rysunku 3.1 można taką zalezność zauważyć pomiędzy „rx1day” i „r20mm”(korelacja = 0,928). pomiędzy „rx1day” i „cwd”(korelacja = 0,735) oraz pomiędzy „r20mm” i „cwd”(korelacja = 0,708). Zatem te trzy zmienne razem nie wnoszą nic więcej, tak samo z osobna nie będą znacząco zmieniać modelu. Odrzucono więc dwie z nich(„r20mm”,„cwd”). Tak jak wspomniano wyżej, zależność liniowa była również pomiędzy

„tas” i „cdd65”, więc po usunięciu „cdd65” zostały nam podane wyżej trzy zmienne. Tak jak w modelu pierwszym po dostosowaniu zmiennych do modelu otrzymano w podsumowaniu (rysunek 3.9) następujące parametry:

```
Formula: tas ~ fourier_multi(x_list, params, n_harmonics)

Parameters:
           Estimate Std. Error t value Pr(>|t|)
params1  12.5322409  0.0811317 154.468  < 2e-16 ***
params2   0.5610990  0.0762357   7.360 2.07e-09 ***
params3  -0.4700496  0.0734012  -6.404 6.04e-08 ***
params4  -0.2541447  0.0772326  -3.291 0.001878 **
params5  -0.0401332  0.0686468  -0.585 0.561533
params6   0.1846716  0.0741764   2.490 0.016305 *
params7  -0.1030917  0.0692692  -1.488 0.143219
params8  -0.2502972  0.0695696  -3.598 0.000757 ***
params9   0.0261549  0.0667087   0.392 0.696737
params10 -0.0382869  0.1216141  -0.315 0.754261
params11  0.0811462  0.1134014   0.716 0.477726
params12 -0.0002665  0.1021531  -0.003 0.997929
params13 -0.1456790  0.1019608  -1.429 0.159545
params14  0.0511538  0.1101766   0.464 0.644540
params15  0.0283644  0.0894676   0.317 0.752592
params16 -0.0025989  0.1027213  -0.025 0.979920
params17 -0.1007140  0.0815165  -1.236 0.222656
params18  0.3299888  0.1123403   2.937 0.005072 **
params19 -0.0052562  0.1055076  -0.050 0.960474
params20 -0.1857521  0.1103760  -1.683 0.098888 .
params21 -0.0439163  0.0807355  -0.544 0.588992
params22  0.0088895  0.1189815   0.075 0.940753
params23  0.1622003  0.0835013   1.942 0.057957 .
params24  0.1238915  0.0879040   1.409 0.165164
params25 -0.0345902  0.0728936  -0.475 0.637274
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3706 on 48 degrees of freedom

Algorithm "port", convergence message: relative convergence (4)
```

Rysunek 3.9. Podsumowanie modelu drugiego. Źródło: opracowanie własne.

Liczba harmonicznych wyniosła 11 a wzór funkcji po podstawieniu parametrów wygląda tak:

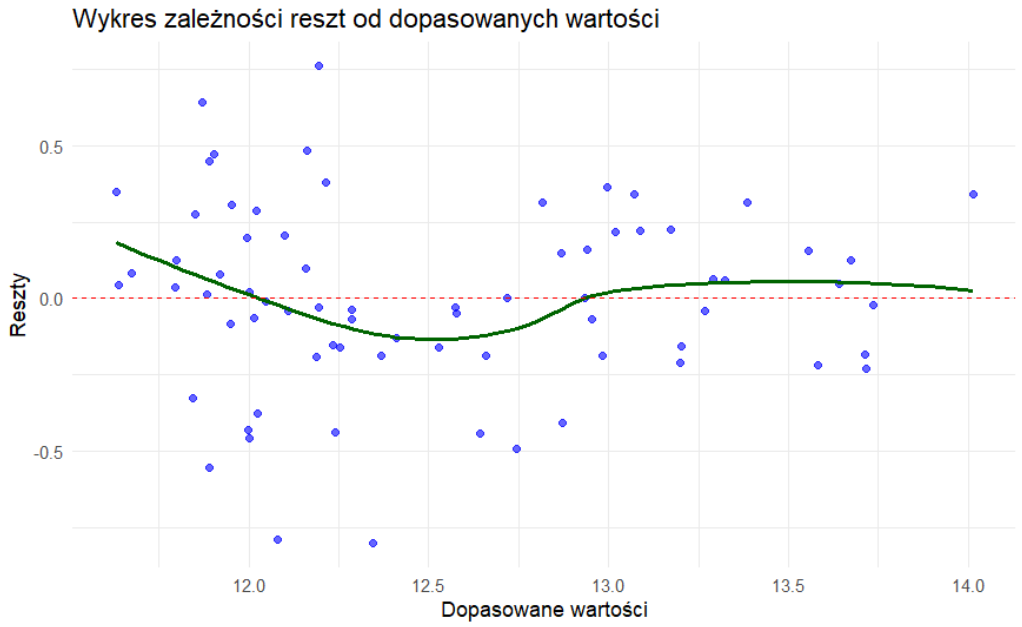
$$\begin{aligned}\text{tas}(x) = {} & 12,5322409 + 0,5610990 \cdot \sin(x) - 0,4700496 \cdot \cos(x) \\ & -0,2541447 \cdot \sin(2x) - 0,0401332 \cdot \cos(2x) - 0,1846716 \cdot \sin(3x) - 0,1300917 \cdot \cos(3x) \\ & -0,2502972 \cdot \sin(4x) + 0,0261549 \cdot \cos(4x) - 0,0382869 \cdot \sin(5x) + 0,0811462 \cdot \cos(5x) \\ & -0,0002665 \cdot \sin(6x) + 0,0025899 \cdot \cos(6x) - 0,1007140 \cdot \sin(7x) + 0,3299888 \cdot \cos(7x) \\ & -0,0052561 \cdot \sin(8x) - 0,1857521 \cdot \cos(8x) - 0,0439163 \cdot \sin(9x) + 0,0088895 \cdot \cos(9x) \\ & +0,1622003 \cdot \sin(10x) + 0,1238951 \cdot \cos(10x) - 0,0345902 \cdot \sin(11x) + 0,0728936 \cdot \cos(11x)\end{aligned}$$

gdzie x to zmienna „YEAR”.
Aby ocenic model potrzebne było wykonanie oceny modelu, jak przy modelu pierwszym.

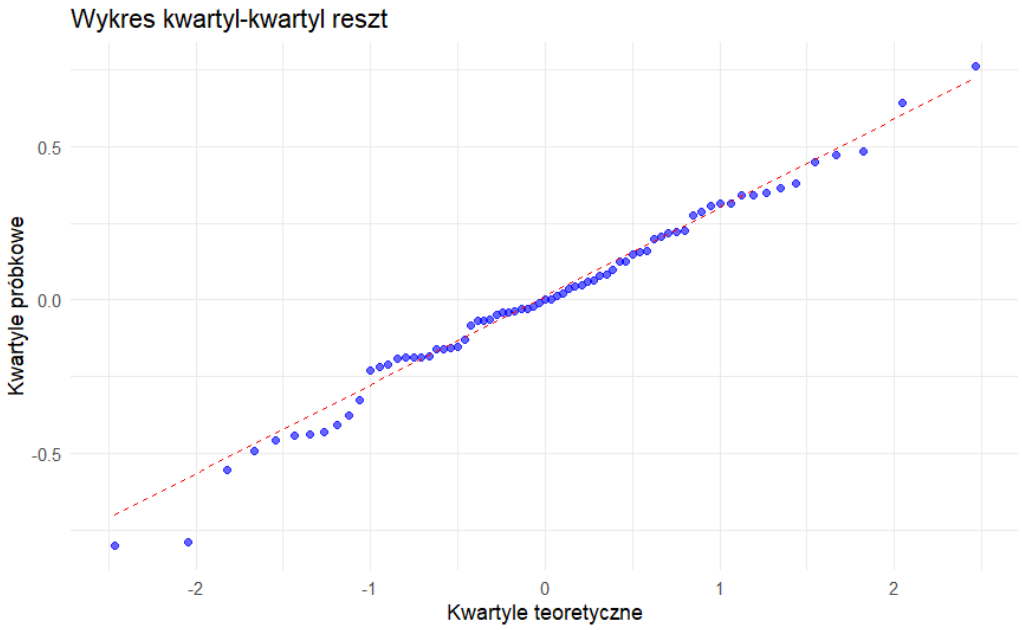
```
"MSE: 0.0903"
"RMSE: 0.3005"
"R^2: 0.8122"
```

Rysunek 3.10. Ocena modelu drugiego. Źródło: opracowanie własne.

Z podanych ocen przedstawionych na rysunku 3.10 wyszły niskie wartości błędów przewidywań (MSE i RMSE), co oznacza, że różnice między wartościami rzeczywistymi a przewidywanymi są niewielkie. Wartość współczynnika determinacji R^2 wskazuje, że model wyjaśnia około 80% zmienności danych, więc po tych ocenach, możemy wywnioskować, że model jest odpowiedni do analizy.
Z rysunku 3.11 na wykresie zależności reszt od dopasowanych wartości widać, że reszty są rozłożone losowo wokół osi $y = 0$.Jednak można zwrócić uwagę na drobne wybrzuszenia zielonej linii. Aby mieć pewność potrzebne było wykonanie testu Durbin-Watsona.
Wykres kwartyl-kwartyl (rysunek 3.12) wskazuje, że większość punktów układa się wzdłuż linii odniesienia, można spodziewać się zatem normalności reszt.



Rysunek 3.11. Wykres zależności reszt od dopasowanych wartości modelu drugiego. Źródło: opracowanie własne.



Rysunek 3.12. Wykres kwartyl-kwartyl reszt modelu drugiego. Źródło: opracowanie własne.

```
        Shapiro-Wilk normality test

data:  ITA_TABELA$residuals
W = 0.9868, p-value = 0.6473


        Durbin-Watson test

data:  tas ~ tas_pred
DW = 2.0377, p-value = 0.522
```

Rysunek 3.13. Testy dla modelu pierwszego. Źródło: opracowanie własne.

Potwierdzeniem powyższych rozważań są testy przedstwione na rysunku 3.13. Test Durbin-Watsona wykazał brak istotnej autokorelacji reszt $DW = 2,0377, p > 0.05$. Czyli model właściwie odzwierciedla zależności w danych. Test Shapiro-Wilka wykazał, że nie ma podstaw do odrzucenia hipotezy zerowej o normalności reszt $p > 0.05$. Oznacza to, że reszty charakteryzują się rozkładem normalnym.

3.1.3. Porównanie

W celu porównania przygotowano tabelę, w której zaprezentowano wyniki oceny obu modeli.

Tabela 3.1. Porównanie oceny modeli. Źródło: opracowanie własne.

Ocena modelu	**Model pierwszy**	**Model drugi**
MSE	0.09029483	0.1424166
RMSE	0.30049098	0.3773812
R^2	0.81215136	0.7037177

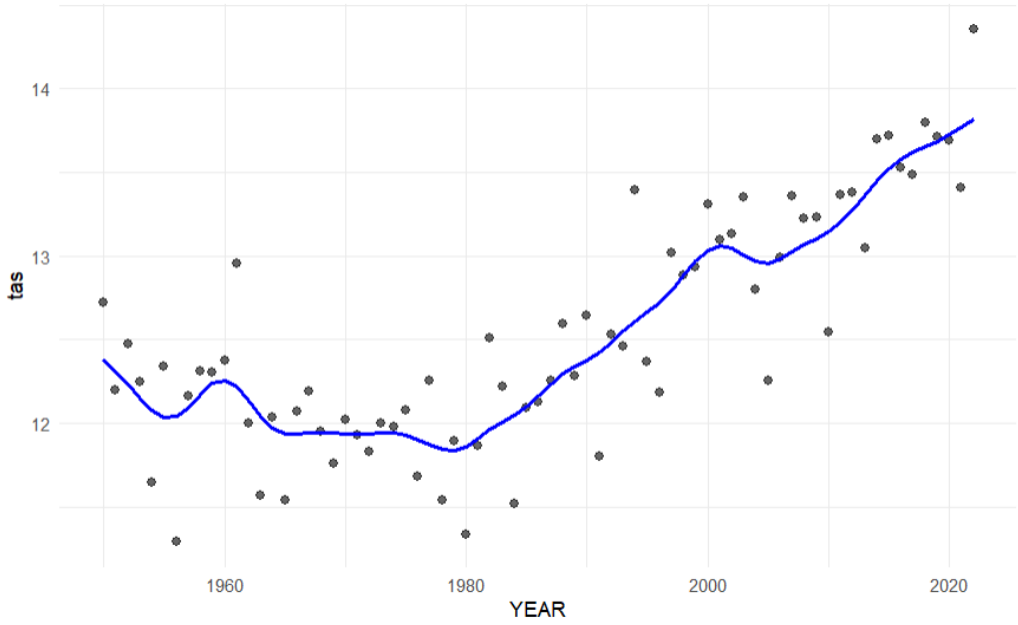
Porównując obie metody z tabeli 3.1, można stwierdzić, że model drugi, oparty na odpowiednio dobranych zmiennych objaśniających, lepiej dopasowuje się do danych. Wybór kluczowych zmiennych objaśniających przyniósł korzyści, co przełożyło się na lepszą jakość modelu. Niemniej jednak, model pierwszy, uwzględniający tylko jedną zmienną objaśniającą, również wykazuje dobre wyniki i nie należy go uznawać za nieodpowiedni. Trzeba jednak pamiętać o braku normalności reszt w tym modelu. Podsumowując, można wywnioskować, że estymacja funkcji regresji za pomocą aproksymacji Fouriera stanowi skuteczne rozwiązanie dla danych, w których występuje nieliniowa zależność.

3.2. Metoda estymatorów jądrowych

W ramach kolejnej metody estymacji zastosowano wygładzanie za pomocą estymatorów jądrowych. W tym podrozdziale przeanalizowano dwie strategie wyznaczania optymalnej szerokości jądra, z wykorzystaniem Gaussowskiego estymatora jądrowego. Wyniki tych analiz zostały zaprezentowane na danych zawierających jedną zmienną objaśniającą – „YEAR".

3.2.1. Model pierwszy

Jako pierwsze skorzystano z wbudowanej funkcji w programie R „dpill", z biblioteki „KernSmooth". Metode „dpill" wykorzystano do estymacji optymalcej szerokości (bandwidth) jądra, bazuje ona na minimalizacji średniego zintegrowanego błędu kwadratowego, podobnie jak reguła Silvermana. Do implementacji modelu regresji jądrowej wykorzystano funkcję „ksmooth".
Po zaimplementowaniu wszystkich niezbędnych parametrów i dodaniu wygładzonych wartości do tabeli, uzyskano wykres (rysunek 3.14) dopasowania modelu do danych.



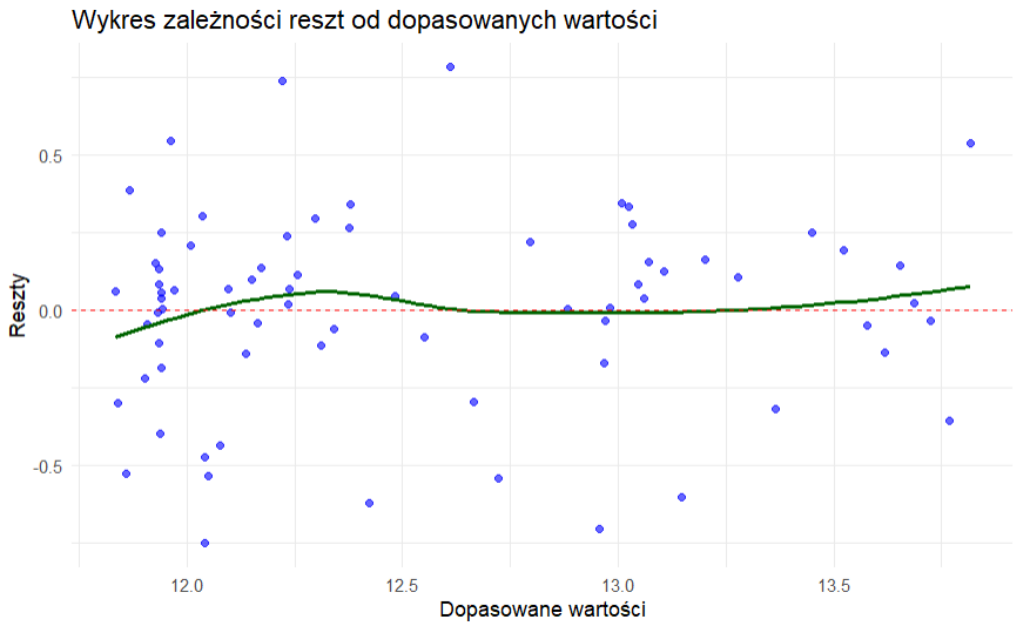
Rysunek 3.14. Wykres modelu pierwszego (Metoda estymatorów jądrowych). Źródło: opracowanie własne.

Szerokość jądra wyniosła dokładnie 6,093962. Po samym wykresie widocznym na rysunku 3.14, można zauważyć, że metoda estymacji za pomocą wygładzania jądrowego z estymacją szerokości pasma za pomocą „dpill" precyzyjnie odwzorowuje dane. Potwierdzeniem są wyniki MSE, RMSE i R^2.

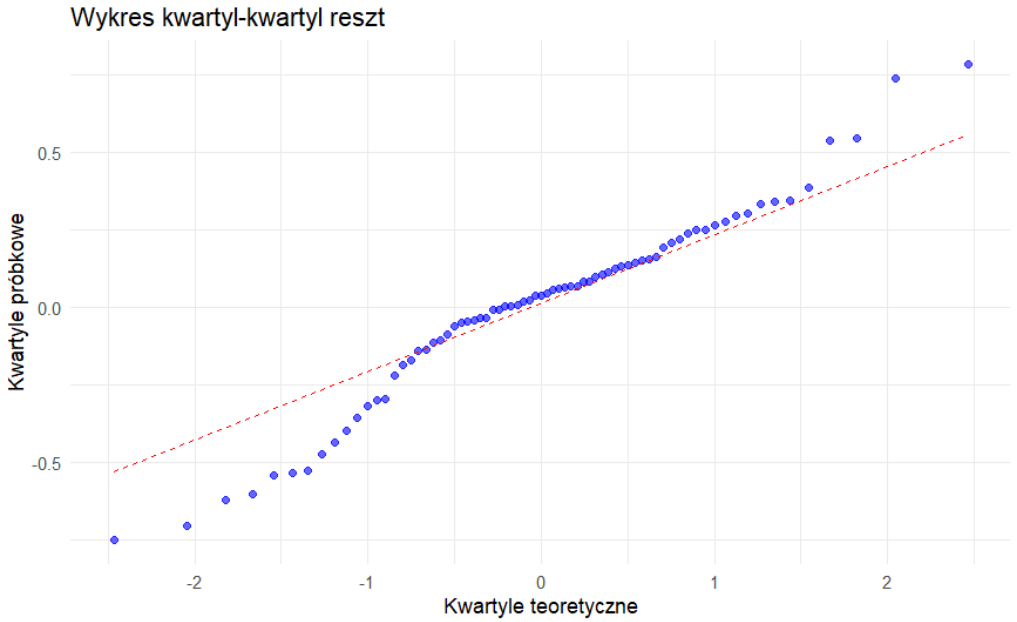
```
"MSE: 0.0952"
"RMSE: 0.3085"
"R^2: 0.802"
```

Rysunek 3.15. Ocena modelu pierwszego (Metoda estymatorów jądrowych). Źródło: opracowanie własne.

W wyniku uzyskano niskie wartości MSE (0.0952) jak i RMSE (0.3085). Wskazuje to na wysoką jakość modelu wygładzania jądrowego, który wyjaśnia 80.2% zmienności zmiennej zależnej ($R^2 = 0.802$). Dokładniejsza analiza wykresu zależności reszt od dopasowanych wartości (rysunek 3.16) oraz wykresu kwartyl-kwartyl reszt (rysunek 3.17) potwierdziła powyższe wnioski. Z wykresu widocznego na rysunku 3.17 oprócz bardzo drobnego wybrzuszenia dla niskich wartości przewidywanych zielona linia jest przy samej prostej y = 0. Z kolei z wykresu przedstawionego na rysunku 3.17, wynika, że większość danych punktów układa się wzdłuż linii odniesienia w szczególności w środkowej części. Jedynie krańce wykresu pokazują odchylenia od normalności. Zatem można wnioskować, że rozkład reszt może spełniać założenie normalności. Aby to potwierdzić wykonano testy (rysunek 3.18).



Rysunek 3.16. Wykres zależności reszt od dopasowanych wartości modelu pierwszego(Metoda estymatorów jądrowych). Źródło: opracowanie własne.



Rysunek 3.17. Wykres kwartyl-kwartyl reszt modelu pierwszego (Metoda estymatorów jądrowych). Źródło: opracowanie własne.

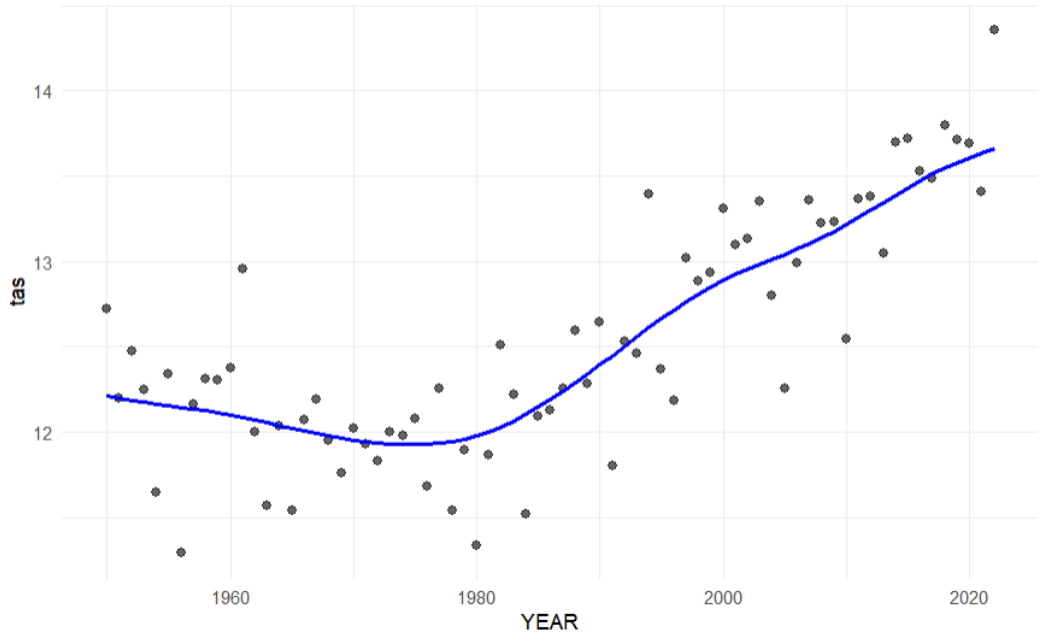
```
        Shapiro-Wilk normality test

data:  ITA_TABELA$residuals2
W = 0.96701, p-value = 0.05279


        Durbin-Watson test

data:  tas ~ tas_smooth
DW = 2.3905, p-value = 0.9417
```

Rysunek 3.18. Testy dla modelu pierwszego (Metoda estymatorów jądrowych). Źródło: opracowanie własne.

Ponieważ $p > 0.05$ testu Shapiro-Wilka, nie ma podstaw do odrzucenia hipotezy zerowej o normalności reszt, co wzmacnia wiarygodność modelu. Dodatkowo, wartość testu Durbin-Watsona powyżej 2 oraz $p > 0.05$ wskazuje na brak autokorelacji. Czyli analiza wykresów została potwierdzona.
Metoda wygładzania jądrowego z automatycznym doborem szerokości pasma („dpill") okazała się bardzo skuteczna w dopasowaniu modelu nieliniowego. Model dokładnie opisuje zmienność „tas" i może być prawidłową opcją do estymacji funkcji regresji w przypadku danych z nieliniowymi zależnościami.

3.2.2. Model drugi

Drugą metodą wykorzystaną dla estymacji optymalnej szerokości jądra jest użycie walidacji krzyżowej. Jest to technika oceny modeli, która polega na podziale danych na kilka części (tzw. foldów), z których każda pełni naprzemiennie rolę zbioru testowego, a pozostałe służą do trenowania modelu. W niniejszej pracy walidacja krzyżowa została zastosowana do automatycznego wyznaczenia szerokości pasma jądra, minimalizując błąd predykcji.
W tym celu użyto funkcji „npregbw" z pakietu „np", która na podstawie walidacji krzyżowej wyznacza szerokość jądra najlepiej dopasowaną do danych. Po wy-estymowaniu optymalnej szerokości, dopasowaniu modelu oraz po dodaniu wygładzonych wartości do tabeli, w wyniku otrzymano kolejną dobrze estymującą funkcję regresji metodę. Potwierdzeniem jest poniższy wykres (rysunek 3.19) oraz poniższe wyniki (rysunek 3.20).

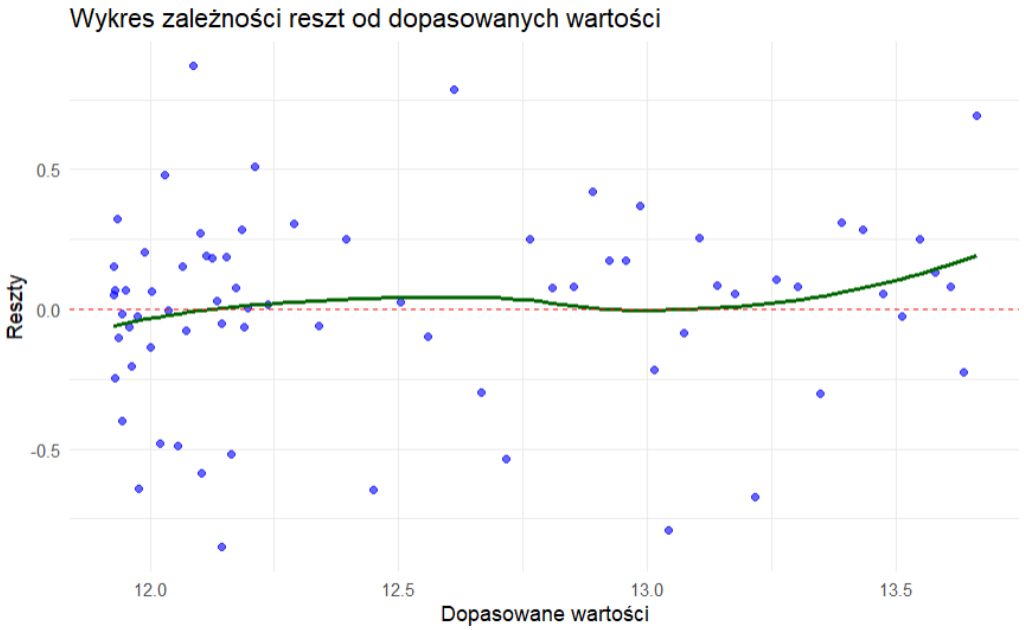


Rysunek 3.19. Wykres modelu drugiego (Metoda estymatorów jądrowych). Źródło: opracowanie własne.

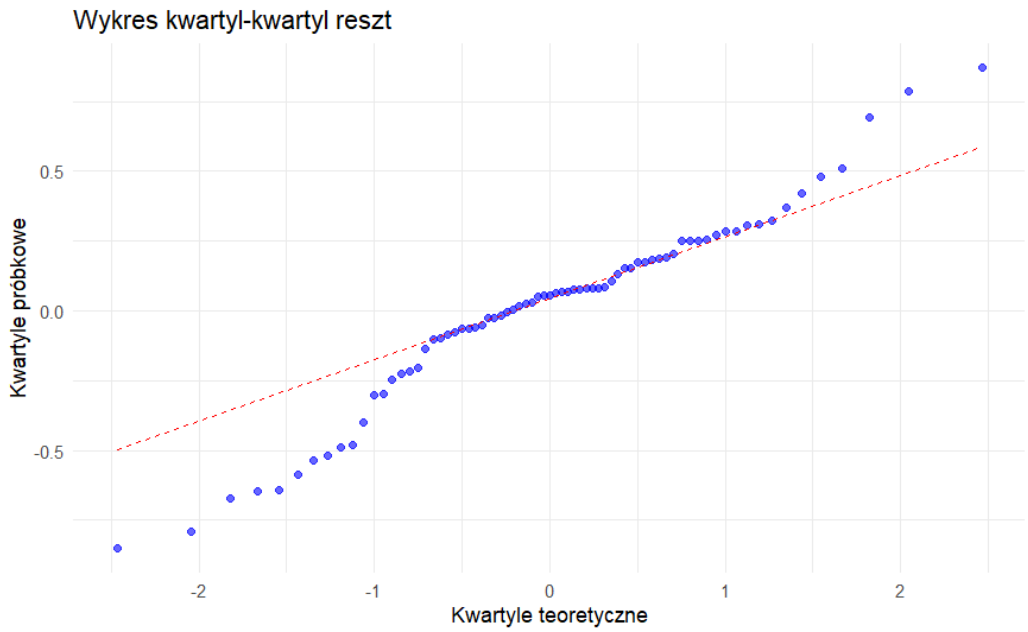
```
"MSE: 0.1151"
"RMSE: 0.3393"
"R^2: 0.7605"
```

Rysunek 3.20. Ocena modelu drugiego (Metoda estymatorów jądrowych). Źródło: opracowanie własne.

W tej opcji estymacji szerokość, wyszła mniejsza od opcji poprzedniej, a dokładnie 5.574122. Teoretycznie mniejsza szerokość powinna skutkować bardziej elastycznym modelem i lepszym dopasowaniem do danych, jednak w tym przypadku efekt jest odwrotny. Linia wygładzania jądrowego jest mniej elastyczna i bardziej wygładzona w stosunku do modelu pierwszego, co pokazuje, że szerokość pasma powinna być oceniana w kontekście konkretnej metody.
Chociaż model drugi, mimo mniejszej szerokości pasma, wydaje się mniej dokładny w dopasowaniu, wciąż uzyskuje bardzo dobre wyniki. Wynika to z R^2, które wynosi 76% oraz niskich MSE (0.1151) i RMSE (0.3393).



Rysunek 3.21. Wykres zależności reszt od dopasowanych wartości modelu drugiego (Metoda estymatorów jądrowych). Źródło: opracowanie własne.



Rysunek 3.22. Wykres kwartyl-kwartyl reszt modelu drugiego (Metoda estymatorów jądrowych). Źródło: opracowanie własne.

```
        Shapiro-Wilk normality test

data:  ITA_TABELA$residuals3
W = 0.96118, p-value = 0.02426


        Durbin-Watson test

data:  tas ~ tas_smoothcv
DW = 2.0624, p-value = 0.558
```

Rysunek 3.23. Testy dla modelu drugiego (Metoda estymatorów jądrowych). Źródło: opracowanie własne.

Wykres zależności reszt od dopasowanych wartości wygląda prawie idealnie. Różnice między zieloną linią a prostą y = 0 jest znikoma. Test Durbin-Watsona jest potwierdzeniem, DW = 2,0624 ($p > 0.05$). Problem pojawił się przy teście Shapiro-Wilka, ponieważ wynik wykazał $p < 0.05$, co sugeruje na brak normalności reszt. Po samym wykresie widocznym na rysunku 3.22 można zauważyć spore odchylenia od linii odniesienia oraz drobne odchylenia w środkowej części.
Podsumowując, model bardzo dobrze dopasował się do danych, co pokazuje, że ta metoda również stanowi skuteczne podejście do estymacji danych, jednak trzeba pamiętać o braku normalności reszt. W ramach potrzeby dalszej analizy, aby poprawić normalność, można rozważyć uwzględnienie dodatkowych zmiennych objaśniających, które mogą lepiej uchwycić strukturę danych.

3.2.3. Porównanie

Tabela 3.2. Porównanie oceny modeli. Źródło: opracowanie własne.

Ocena modelu	**Model pierwszy**	**Model drugi**
MSE	0.09515897	0.1151187
RMSE	0.30847848	0.3392914
R^2	0.80203204	0.7605081

Korzystając z powyższej tabeli, można stwierdzić, że oba modele wykazały się wysoką jakością dopasowania. Skutecznie dopasowują się do danych oraz objaśniają je, o czym świadczą pozytywne wyniki ocen modelu. Drobną przewagę ma jednak model pierwszy, objaśnia o około 4% więcej danych oraz oceny MSE oraz RMSE również wyszły minimalnie lepsze czyli mniejsze. Dodatkowo założenia dla modelu pierwszego zostały spełnione. Dlatego w kolejnym etapie skorzystano z właśnie modelu pierwszego.

3.3. Połączenie dwóch metod

W tej metodzie połączono dwa podejścia estymacji funkcji regresji w modelu nieliniowym, opisane we wcześniejszych sekcjach. Celem tej analizy jest ocena, czy zastosowanie tych metod w połączeniu przynosi lepsze rezultaty niż ich stosowanie oddzielnie. Rozważono dwa warianty: najpierw zastosowanie wygładzania jądrowego, a następnie aproksymacji Fouriera do wygładzonych danych, oraz odwrotną kolejność, czyli dopasowanie aproksymacji Fouriera do danych, a następnie wygładzenie uzyskanej funkcji.

3.3.1. Model pierwszy - Wygładzanie jądra jako preprocesowanie przed aproksymacją Fouriera

Jako pierwsza przedstawiona została metoda, która łączy wygładzanie jądrowe oraz aproksymację Fouriera. Podejście to zakłada, że na wstępnym etapie dane są wygładzane za pomocą jądrowego oszacowania gęstości, a następnie na tak przygotowanych danych przeprowadza się aproksymację w szereg Fouriera. Po zaimplementowaniu w podsumowaniu modelu otrzymaliśmy:

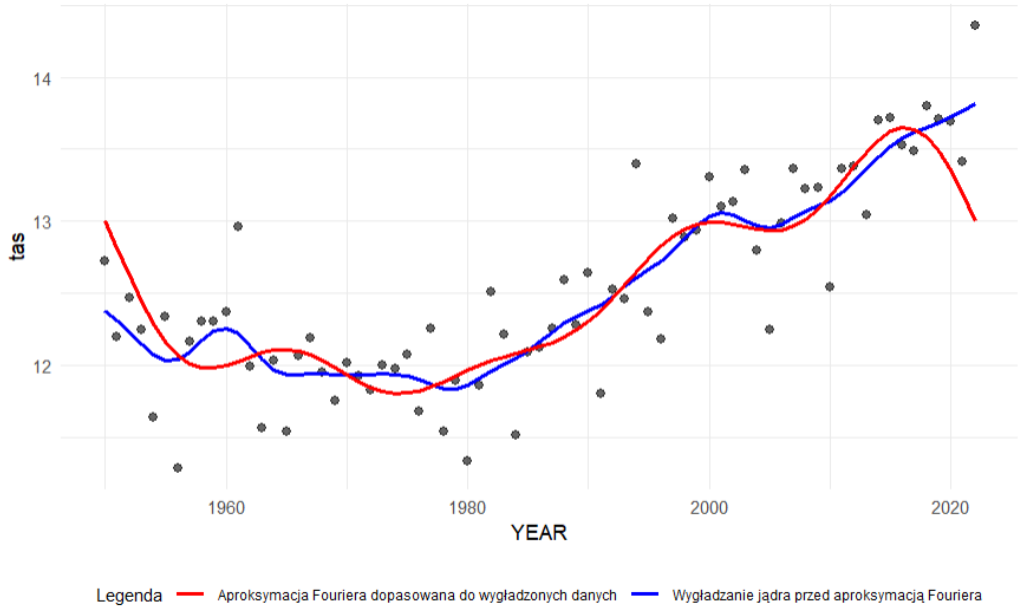
```
Formula: tas_smooth ~ fourier(YEAR_scaled, params, n_harmonics)

Parameters:
         Estimate Std. Error t value Pr(>|t|)
params1 12.521776   0.024025 521.208  < 2e-16 ***
params2  0.650346   0.034187  19.023  < 2e-16 ***
params3 -0.371489   0.033763 -11.003  < 2e-16 ***
params4 -0.151541   0.034187  -4.433 3.73e-05 ***
params5  0.040324   0.033763   1.194   0.2368
params6  0.194460   0.034187   5.688 3.43e-07 ***
params7 -0.066142   0.033763  -1.959   0.0545 .
params8 -0.160226   0.034187  -4.687 1.50e-05 ***
params9  0.006928   0.033763   0.205   0.8381
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2051 on 64 degrees of freedom

Algorithm "port", convergence message: both X-convergence and relative convergence (5)
```

Rysunek 3.24. Podsumowanie modelu pierwszego (połączenie dwóch metod). Źródło: opracowanie własne.

Z rysunku 3.24 można zapisać wzór funkcji Fouriera dopasowanej do danych, wygląda on następująco:

$$\begin{aligned}\text{tas}(x) = 12,521776 + 0,650346 \cdot \sin(x) - 0,371489 \cdot \cos(x) \\ -0,151541 \cdot \sin(2 \cdot x) + 0,040324 \cdot \cos(2 \cdot x) \\ +0,194460 \cdot \sin(3 \cdot x) - 0,066142 \cdot \cos(3 \cdot x) \\ -0,160226 \cdot \sin(4 \cdot x) + 0,006928 \cdot \cos(4 \cdot x)\end{aligned}$$

gdzie x to zmienna „YEAR_scaled”.

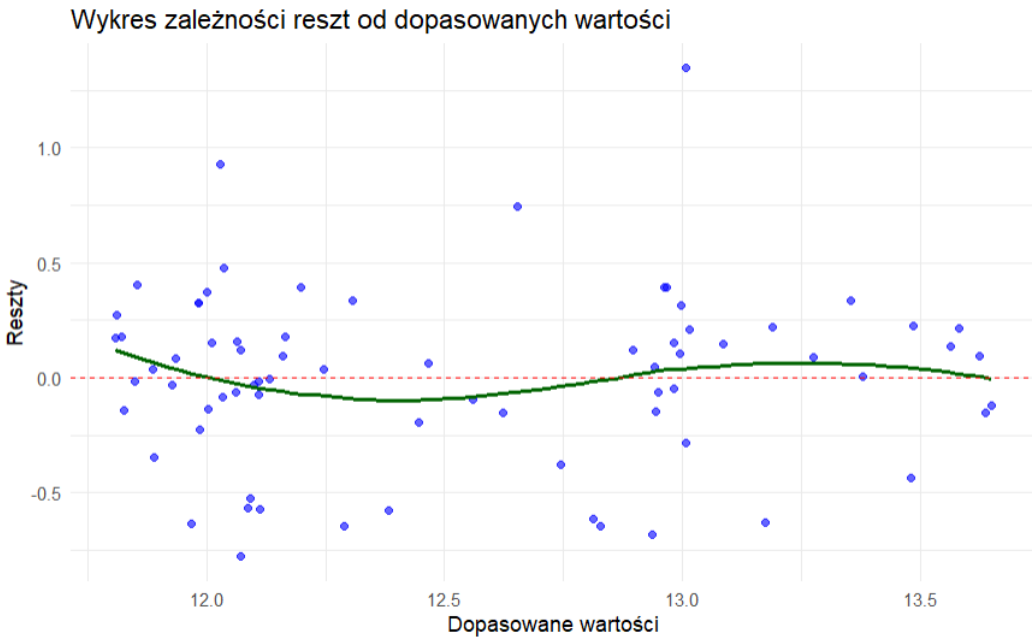


Rysunek 3.25. Wykres modelu pierwszego (Połączenie dwóch metod). Źródło: opracowanie własne.

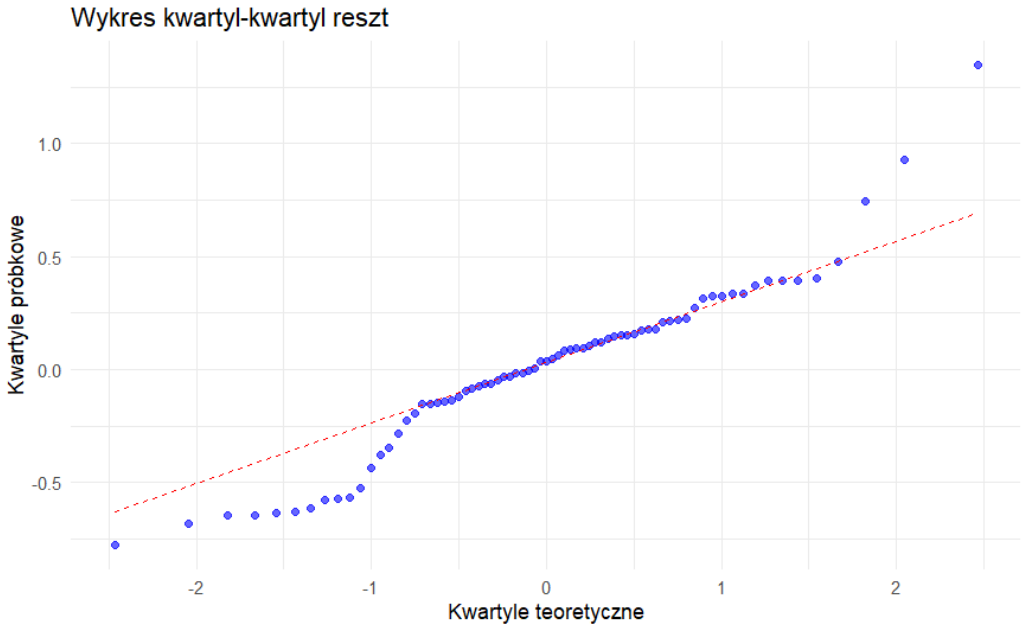
```
"MSE: 0.1429"
"RMSE: 0.3781"
"R^2: 0.7026"
```

Rysunek 3.26. Ocena modelu pierwszego (Połączenie dwóch metod). Źródło: opracowanie własne.

Na wykresie przedstawionym na rysunku 3.25 niebieska linia pokazuje wygładzanie jądrowe przed aproksymacją Fouriera. Po tej linii można stwierdzić, że wygładzenie sprawnie usuwa szum z danych, umożliwiając lepsze dopasowanie aproksymacji.Czerwona linia przedstawia aproksymację Fouriera dopasowaną do tych wygładzonych danych. Model ten (linia czerwona) skutecznie uchwycił ogólny kształt trendu w danych, ale widoczne są pewne odchylenia na końcach przedziału (lata około 1960 i 2020), co może wynikać z ograniczeń funkcji aproksymacji Fouriera. Oceny modelu (rysunek 3.26) potwierdziły dobre dopasowanie, choć nie idealne. Współczynnik determinacji ($R^2 = 0{,}7026$) objaśnia 70% zmiennej tas.



Rysunek 3.27. Wykres zależności reszt od dopasowanych wartości modelu pierwszego (Połączenie dwóch metod). Źródło: opracowanie własne.



Rysunek 3.28. Wykres kwartyl-kwartyl reszt modelu pierwszego (Połączenie dwóch metod). Źródło: opracowanie własne.

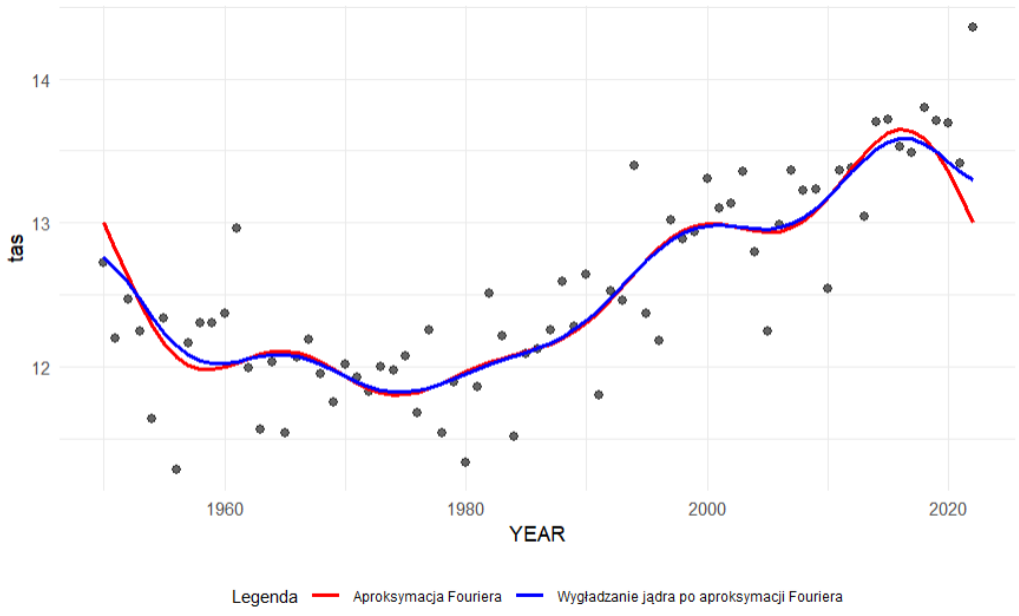
```
        Shapiro-Wilk normality test

data:  ITA_TABELA$residuals4
W = 0.94818, p-value = 0.004624


        Durbin-Watson test

data:  tas ~ tas_fourier
DW = 1.6737, p-value = 0.06298
```

Rysunek 3.29. Testy dla modelu pierwszego (Połączenie dwóch metod). Źródło: opracowanie własne.

Wykres widoczny na rysunku 3.27 ukazuje drobne odchylenia zielonej linii od prostej y = 0, jednak są to nadal drobne odchylenia, także nie powinno być problemu z założeniem homoskedastyczności, ponieważ rozrzut reszt wydaje się być stały w całym zakresie. Test Durbin-Watsona wykazał również brak problemów z rozkładem losowym reszt i homoskedastycznością ($DW = 1.6737, p > 0,05$). Połączenie modeli nie przyniosło jednak efektów jeżeli chodzi o normalność reszt ($p < 0,05$). Sprawdzona została zatem opcja druga połączenia tych metod.

3.3.2. Model drugi - Wygładzanie jądra jako korekta po aproksymacji Fouriera

W tym modelu początkowo została dopasowana aproksymacja fouriera do danych a następnie zastosowane zostało wygładzenie jądra funkcji. Zatem skorzystano z modelu opisanego w 3.1.1.



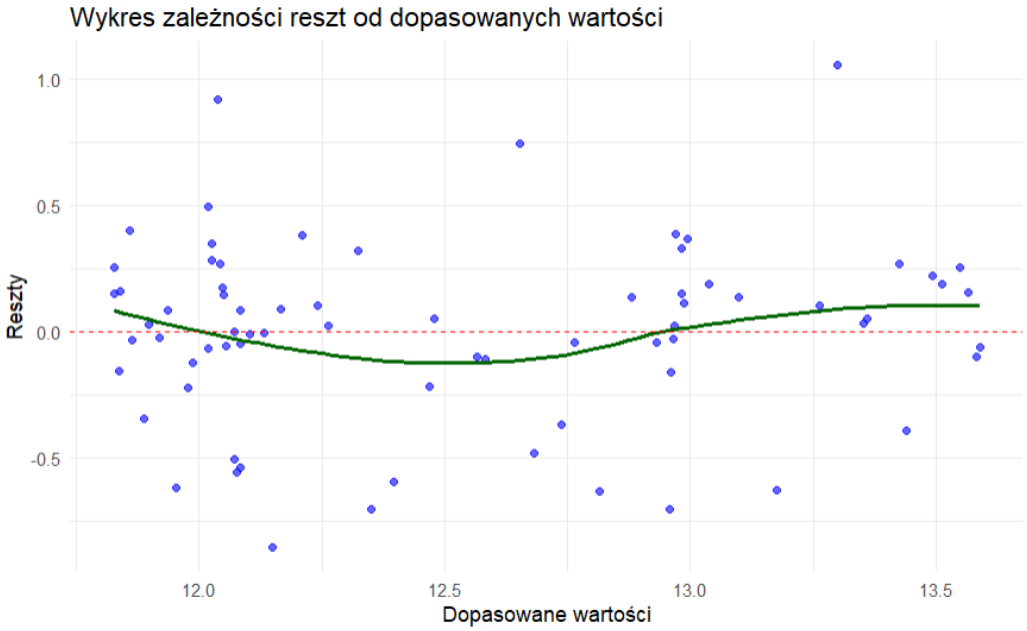
Rysunek 3.30. Wykres modelu drugiego (Połączenie dwóch metod). Źródło: opracowanie własne.

```
"MSE: 0.1284"
"RMSE: 0.3584"
"R^2: 0.7328"
```

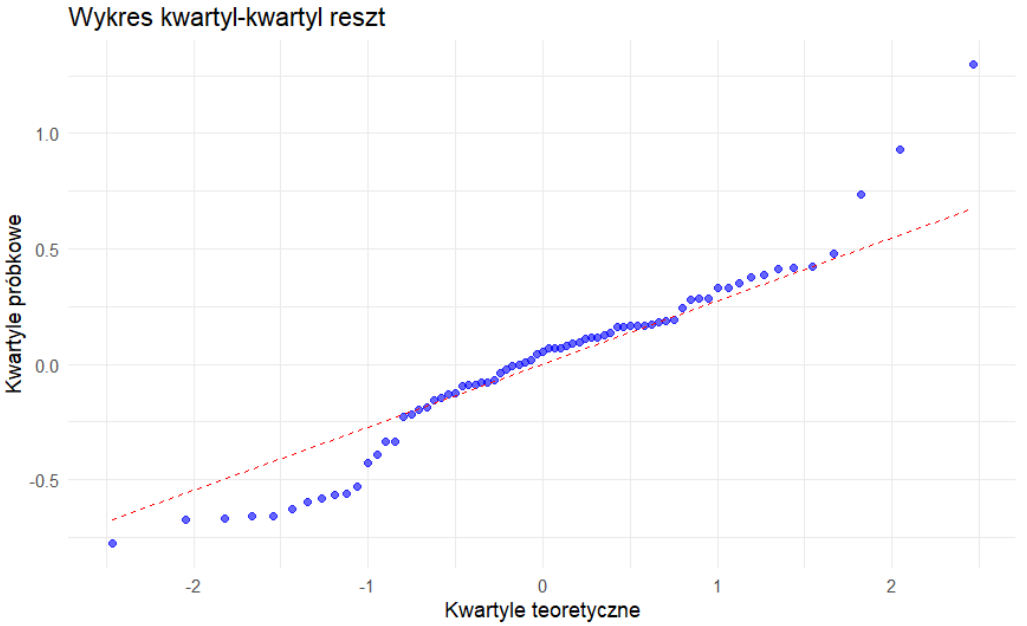
Rysunek 3.31. Ocena modelu drugiego (Połączenie dwóch metod). Źródło: opracowanie własne.

Na wykresie widocznym na rysunku 3.30 można zauważyć, że wygładzenie dodatkowo funkcji Fouriera, nie zmieniło znacząco dopasowania do danych. Pomogło przy końcach przedziału, w czym był problem w modelu pierwszym. Patrząc po samym wykresie można stwierdzić, że prawdopodobnie model ten

minimalnie lepiej dopasował się do danych ze względu właśnie na końce wykresu, co może świadczyć o bardzo dobrym dopasowaniu się do danych. Oceny modelu są tego potwierdzeniem. Model nie znakomicie, lecz dobrze dopasował się do danych. Mse oraz RMSE wyszło niskie a $R^2 = 0,7328$, pokazuje, że model objaśnia około 73% danych, co jest dobrym wynikiem. W kolejnym kroku, tak jak przy poprzednich modelach sprawdzono wykres zależności reszt od dopasowanych wartości modelu drugiego,Wykres kwartyl-kwartyl reszt oraz testy.



Rysunek 3.32. Wykres zależności reszt od dopasowanych wartości modelu drugiego (Połączenie dwóch metod). Źródło: opracowanie własne.



Rysunek 3.33. Wykres kwartyl-kwartyl reszt modelu drugiego (Połączenie dwóch metod). Źródło: opracowanie własne.

```
        Shapiro-Wilk normality test

data:  ITA_TABELA$residuals5
W = 0.95579, p-value = 0.01202


        Durbin-Watson test

data:  tas ~ tas_smooth_fourier
DW = 1.8488, p-value = 0.2204
```

Rysunek 3.34. Testy dla modelu drugiego (Połączenie dwóch metod). Źródło: opracowanie własne.

Test Shapiro-Wilka niestety również wykazał brak normalności reszt ($p < 0,05$), co można zauważyć na wykresie zależności reszt od dopasowanych wartości (rysunek 3.32). Widać na nim duże odchylenia w dolnym i górnym zakresie, oraz zbliżone lecz nie dokładne linii odniesienia punkty przy centralnym zakresie. Z kolei po wykresie kwartyl-kwartyl reszt (rysunek 3.33) można wywnioskować brak znaczących odchyleń w resztach, co potwierdza test Durbin-Watsona ($DW = 1.8488, p > 0,05$)

3.3.3. Porównanie

Tabela 3.3. Porównanie oceny modeli (Połączenie dwóch metod). Źródło: opracowanie własne.

Ocena modelu	**Model pierwszy**	**Model drugi**
MSE	0.1429455	0.1284176
RMSE	0.3780813	0.3583541
R^2	0.7026174	0.7328410

Tak jak zostało wspomniane przy modelu drugim, już po samym wykresie można było zauważyć, że model drugi minimalnie lepiej dopasował się do danych. Potwierdza to powyższa tabela 3.3. Model drugi, w którym początkowo została wykonana aproksymacja Fouriera a dopiero później wygładzenie jej za pomocą wbudowanej funkcji dpill objaśnia o około 3% więcej danych oraz MSE i RMSE wyszły minimalnie niższe czyli lepsze. Każda z metod wyszła dość dobrze, jednak nie idealnie. Są one jednak warte rozważenia przy próbie estymacji funkcji regresji w modelu nieliniowym, pamiętać jednak trzeba o problemie z normalnością reszt, który wyszedł.

3.4. Porównanie modeli

Poniżej przedstawiono porównanie wszystkich modeli użytych w pracy.

Tabela 3.4. Porównanie oceny wszystkich modeli. Źródło: opracowanie własne.

Model	**MSE**	**RMSE**	**R^2**
Model pierwszy - Aproksymacja Fouriera	0.1424166	0.3773812	0.7037177
Model drugi - Aproksymacja Fouriera	0.09029483	0.30049098	0.81215136
Model pierwszy - Metoda estymatorów jądrowych	0.09515897	0.30847848	0.80203204
Model drugi - Metoda estymatorów jądrowych	0.1151187	0.3392914	0.7605081
Model pierwszy - Połączenie modeli	0.1429455	0.3780813	0.7026174
Model drugi - Połączenie modeli	0.1284176	0.3583541	0.7328410

Z wszystkich badanych metod estymacji, najlepiej wypadł model drugi aproksymacji Fouriera oraz model pierwszy wykonany za pomocą metody estymatorów jądrowych. Z kolei najgorzej wypadły dwa modele. Połączenie tych dwóch metod w modelu pierwszym połączenia, czyli początkowe wygładzenie a następnie dopasowanie aproksymacji Fouriera oraz Model pierwszy aproksymacji Fouriera.

Podsumowanie

Celem pracy było przedstawienie różnych metod estymacji funkcji regresji w modelu nieliniowym wraz z przykładami, które zostały przeprowadzone na danych rzeczywistych. W pierwszym rozdziale zostały opisane niezbędne zagadnienia teoretyczne, m.in. regresja liniowa oraz nieliniowa, liniowa i nieliniowa metoda najmniejszych kwadratów, metoda estymatorów jądrowych, aproksymacja w szereg Fouriera oraz klasyfikacja oceny modelu. W drugim rozdziale zostało opisane praktyczne zastosowanie opisanych w rozdziale pierwszym metod estymacji. Przedstawione zostało wykorzystanie ich z osobna jak i połączonych. Sprawdzono wyniki ocen każdego z modeli. Dodatkowo wykonany został wykres zależności reszt od dopasowanych wartości oraz wykres kwartyl-kwartyl reszt, jak i testy sprawdzające założenia (test Durbin-Watsona i test Shapiro-Wilka). Po każdej z opcji estymacji przedstawione zostało podsumowanie danych metod. Wykonano to również na końcu pracy pokazując oceny wszystkich modeli.

Bibliografia

[1] R.L. Eubank, J.D. Hart, P. Speckman. Welfe, *Trigonometric series regression estimators with an application to partially linear models*,Journal of Multivariate Analysis, 1990.

[2] L. Gajek, M. Kałuszka, *Wnioskowanie statystyczne*, WNT, Warszawa 2000.

[3] T. Gasser, A. Kneip, W. Kohler, *A flexible and fast method for automatic smoothing*, Journal of Multivariate Analysis, 1991.

[4] J. Jakubowski, R. Sztencel, *Wstęp do teorii prawdopodobieństwa*, SCRIPT, Warszawa 2004.

[5] A. Welfe, *Ekonometria. Metody i ich zastosowanie*, Polskie Wydawnictwo Ekonomiczne SA, Warszawa 2009.

[6] M. Wiciak, *Elementy Probabilistyki w zadaniach*, Wydawnictwo Politechniki Krakowskiej, Kraków 2008.

[7] https://climateknowledgeportal.worldbank.org/download-data#htab-1497, dostęp-11.10.2024.