Aleksandra Górska, 268734,
Marcin Pałyga, 268768

Komputerowa Analiza Szeregów czasowych - Raport I

1. Wstęp i opis danych

Będziemy rozważać klasyczny model regresji liniowej oparty na metodzie najmniejszych kwadratów. Prosta regresji opisana jest wzorem: $\hat{y} = b_0 + b_1 x$.

Chcemy, aby wartość sumy kwadratów:

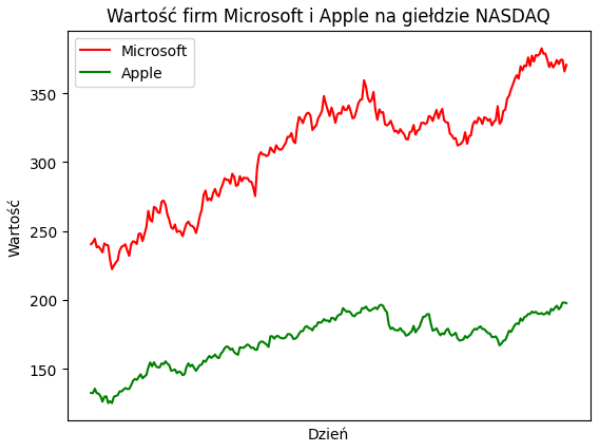
$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))$$

była minimalna. Naszą zmienną objaśniającą będzie wartość firmy Apple na giełdzie NASDAQ w dniach 19.12.2022-18.12.2023. r., za to zmienną objaśnianą będzie wartość firmy Microsoft w tym samym okresie. Długość prób wynosi 250 i zostały one pobrane z odpowiednio "https://finance.yahoo.com/quote/AAPL/history?p=AAPL" dla Apple oraz "https://finance.yahoo.com/quote/MSFT/history?p=MSFT" dla Microsoftu

2. Analiza jednowymiarowa zmiennych

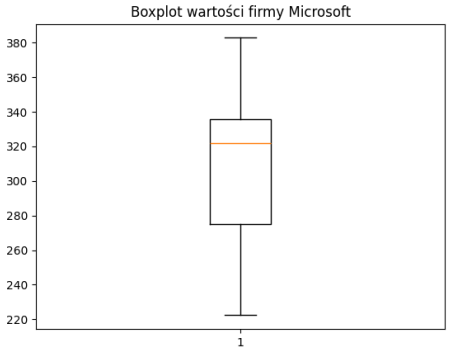
2.1. Wizualizacja danych

Na poniższym wykresie możemy zobaczyć porównanie wartości obu firm na giełdzie podczas okresu jednego roku.

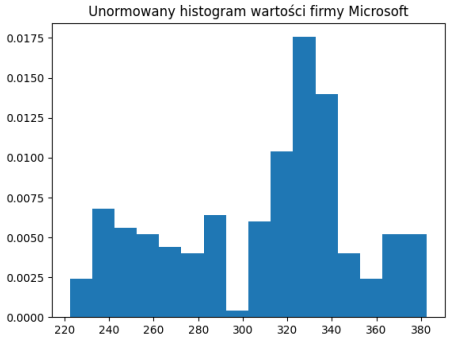


Rysunek 1: Wykres wartości firm.

Poniżej możemy zobaczyć wykres pudełkowy oraz histogram wartości firmy Microsoft oraz firmy Apple.

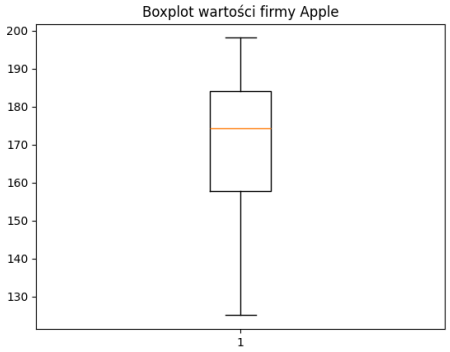


(a) Wykres pudełkowy

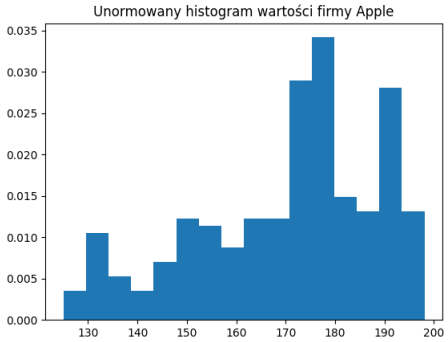


(b) Histogram

Rysunek 2: firma Microsoft



(a) Wykres pudełkowy



(b) Histogram

Rysunek 3: firma Apple

2.2. Wzory miar położenia, rozproszenia, skośności i spłaszczenia

Przedstawimy teraz wzory podstawowych miar statystycznych, które opisują nasze dane oraz tabelę z ich wartościami dla obu firm

— średnia arytmetyczna:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

— rozstęp międzykwartylowy:

$$IQR = Q_3 - Q_1$$

— wariancja:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

— odchylenie standardowe

$$S = \sqrt{S^2}$$

— współczynnik skośności:

$$\alpha = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} (\frac{X_i - \bar{X}}{S})^3$$

— kurtoza:

$$K = \frac{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^4}{(\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2)^2}$$

2.3. Tabela z wynikami

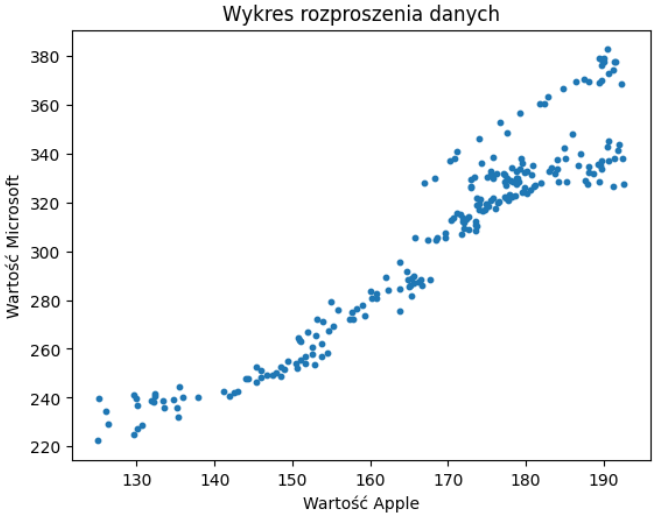
Podstawowe statystyki	Microsoft [USD]	Apple [USD]
średnia arytmetyczna	309.11	170.27
mediana	321.82	174.20
kwartyl Q1	275.27	157.69
kwartyl Q3	335.90	184.07
rozstęp międzykwartylowy	60.62	26.38
wariancja	1719.21	339.08
odchylenie standardowe	41.46	18.41
współczynnik skośności	-0.35	-0.68
kurtoza	-0.89	-0.38

Po przeanalizowaniu danych posortowaliśmy dane ze względu na wartość zmiennej objaśniającej oraz wydzieliliśmy pierwsze 230 danych jako dane treningowe oraz 20 ostatnich jako dane testowe. Następnie przeszliśmy do dopasowania modelu regresji liniowej.

3. Estymacja współczynników w klasycznym modelu regresji

3.1. Estymacja punktowa

Wykres rozproszenia naszych danych prezentuje się następująco:



Rysunek 4: Wykres rozproszenia danych.

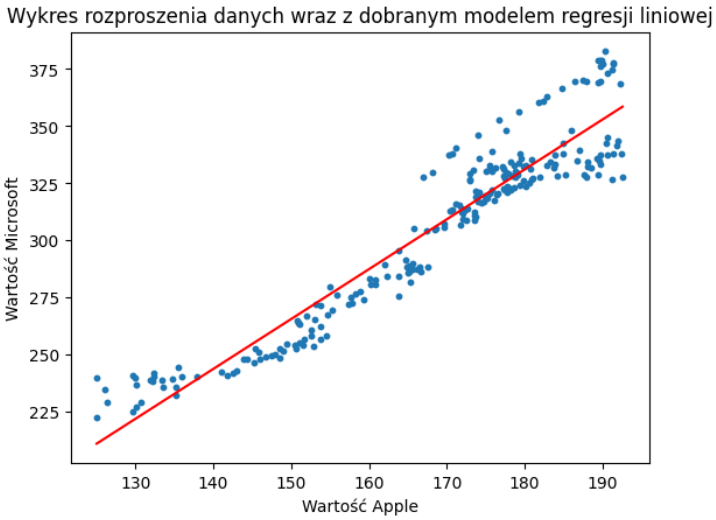
Podczas estymacji punktowej korzystaliśmy z następujących wzorów, dzięki którym obliczyliśmy nieobciążone estymatory parametrów prostej regresji.

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} = 2.1845$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = -62.1691$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = -62.17 + 2.18 X_i$$

Poniżej możemy zobaczyć wygenerowaną prostą regresji w porównaniu do wykresu rozproszenia.



Rysunek 5: Dobrany model regresji liniowej.

Aby sprawdzić dopasowanie modelu do danych treningowych, użyliśmy następujących miar oceny jakości:

— SST - całkowita suma kwadratów

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

— SSR - regresyjna suma kwadratów

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

— SSE - suma kwadratów z błędów

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

— błąd średniokwadratowy S^2 - estymator wariancji σ^2:

$$MSE = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2}{n-2}$$

— średni błąd bezwzględny:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

współczynnik determinacji

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

Ich wartości możemy zobaczyć w poniższej tabeli.

Miary jakości modelu	Wartość
SST	379617.38
SSR	341689.74
SSE	37927.64
współczynnik korelacji Pearsona	0.94
współczynnik determinacji	0.90
MSE	164.90
MAE	10.15

Mimo dużego SST oraz MSE, współczynniki korelacji Pearsona oraz determinacji są na wysokim poziomie co może świadczyć o dobrym dopasowaniu modelu. Potwierdza nam to także wykres modelu oraz rozproszenia, na których możemy zobaczyć, że prosta regresji dobrze przybliża dane.

3.2. Estymacja przedziałowa

Następnie przeszliśmy do estymacji przedziałów ufności estymatorów $\hat{\beta}_0$ oraz $\hat{\beta}_1$ na poziomie istotności $\alpha = 0.05$. Korzystając z odpowiednich wzorów wyznaczyliśmy ich wartości.

$$S = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2}{n-2}}$$

Przedział ufności dla estymatora $\hat{\beta}_1$ wynosi:

$$\left[\hat{\beta}_1 - t_{n-2,1-\frac{\alpha}{2}} \frac{S}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}, \hat{\beta}_1 + t_{n-2,1-\frac{\alpha}{2}} \frac{S}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}\right] = [2.1841; 2.1848]$$

,gdzie $t_{1-\frac{\alpha}{2},n-2}$ jest kwantylem rzędu $1-\frac{\alpha}{2}$ rozkładu t-studenta dla parametru $n-2$.

Przedział ufności dla estymatora $\hat{\beta}_0$ wynosi:

$$\left[\hat{\beta}_0 - t_{n-2,1-\frac{\alpha}{2}} S \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}, \hat{\beta}_0 + t_{n-2,1-\frac{\alpha}{2}} S \frac{1}{n} + \frac{\bar{X}^2}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}\right] = [-78.0588; -46.2794]$$

Jak widać wcześniej wyznaczone wartości estymatorów wpadają w wyznaczone na ich podstawie przedziały ufności.

4. Predykcja dla danych testowych

Używając następującego wzoru wyznaczyliśmy przedział ufności na poziomie istotności $\alpha = 0.05$ dla każdego indeksu danych testowych.

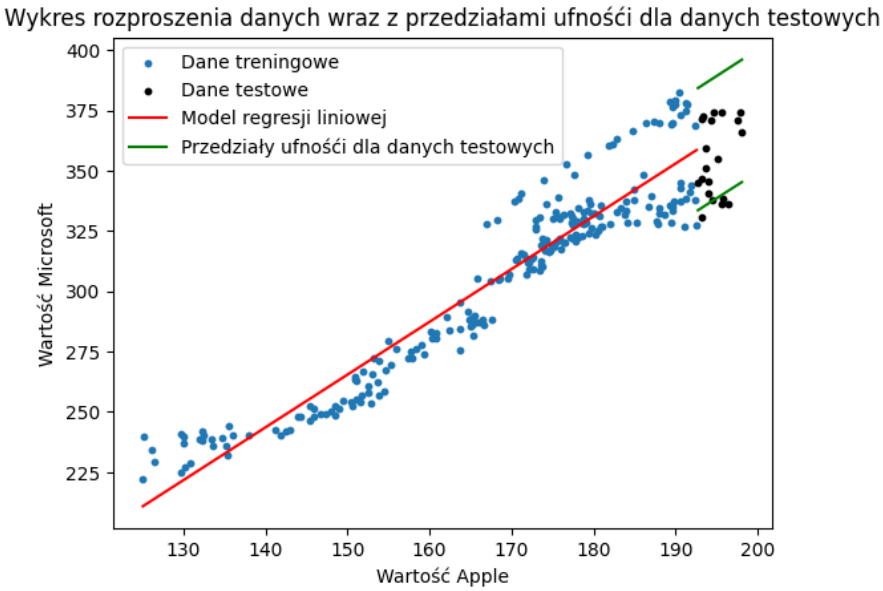
$$\left[\hat{\beta}_0 + \hat{\beta}_1 X_0 - t_{n-2,1-\frac{\alpha}{2}} S \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}, \hat{\beta}_0 + \hat{\beta}_1 X_0 + t_{n-2,1-\frac{\alpha}{2}} S \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}\right]$$

Przedstawiając to na wykresie wraz z wcześniej dopasowaną prostą regresji możemy zauważyć, że niektóre z wartości testowych wystają niewiele poza wcześniej wyliczony przedział.

Aby lepiej sprawdzić jakość predykcji danych testowych użyliśmy następujących miar estymacji punktowej:

— MSE = 292.11
— MAE = 14.38
— $R^2 = 0.24$

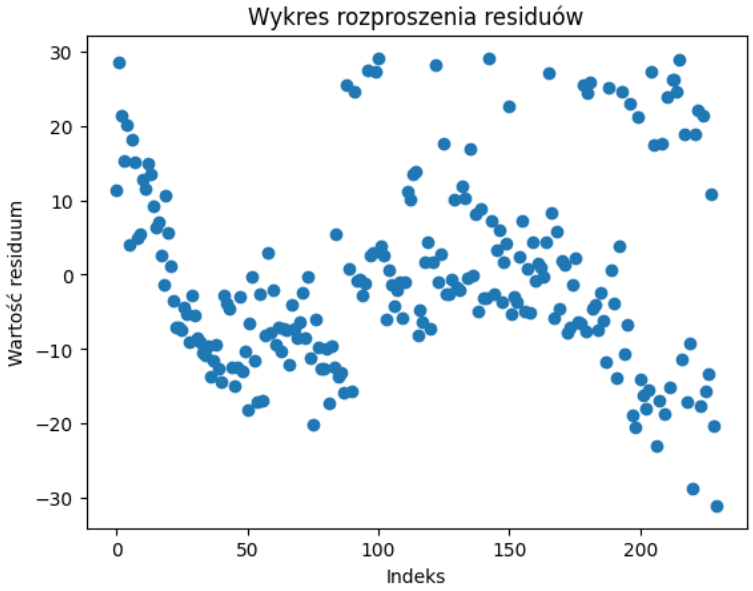
Wysoka wartość błędu średniokwadratowego oraz niski współczynnik determinacji mogą o na przykład świadczyć słabej zależności liniowej danych czy złym dopasowaniu modelu.



Rysunek 6: Przedział ufności predykcji.

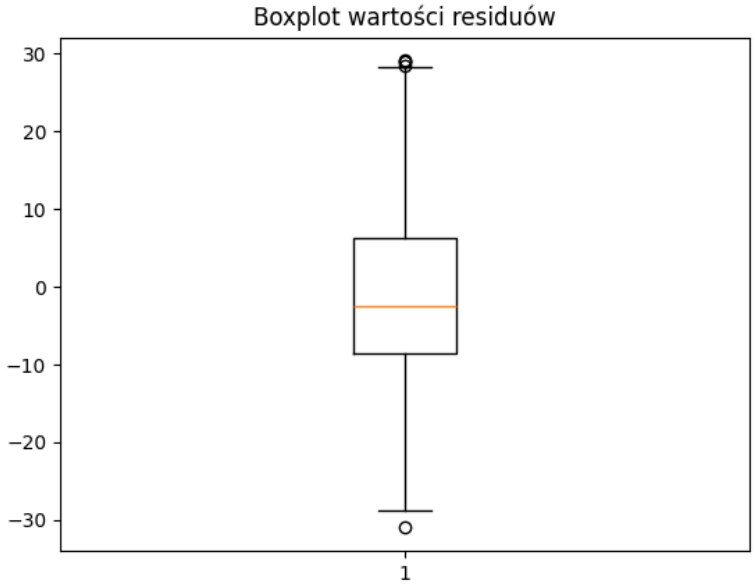
5. Analiza residuów

Analizując residua modelu określone jako $\epsilon_i = y_i - \hat{y}_i$ zajmiemy się sprawdzeniem dotyczących ich założeń w naszym modelu. Wykres rozproszenia residuów możemy zobaczyć poniżej:



Rysunek 7: Wykres rozproszenia residuów.

Na wykresie możemy zobaczyć, że residua układają się w określony sposób i nie są one mocno rozproszone po całym wykresie, co może świadczyć o zależności między nimi. Po wykresie pudełkowym także widać, że w rozkładzie residuów znajdują się wartości odstające, zwłaszcza powyżej górnej granicy rozstępu międzykwartylowego.



Rysunek 8: Wykres pudełkowy residuów.

5.1. Średnia równa zero

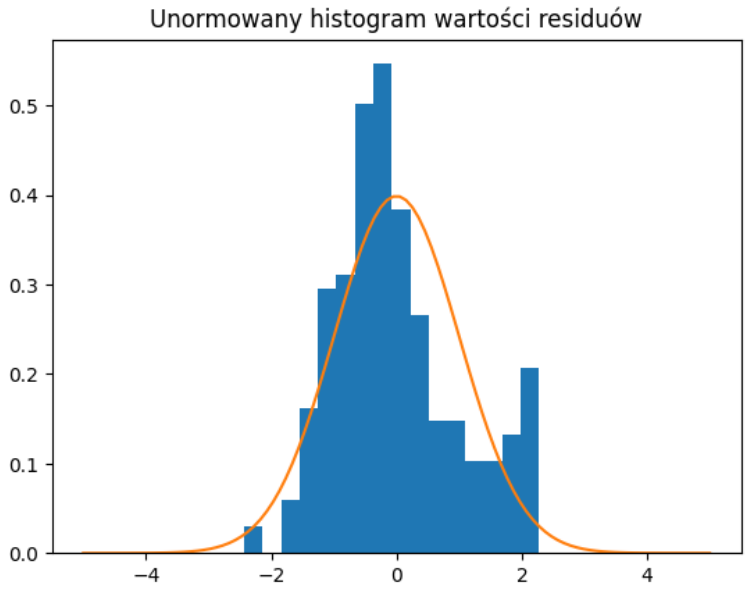
Pierwszym założeniem w naszym modelu dotyczącym residuów, które sprawdzimy jest średnia równa zero. Wartość wyliczona wynosi $-1.11 \cdot 10^{-14}$ co jest bardzo bliskie zeru, więc możemy stwierdzić, że nasze residua spełniają to założenie.

5.2. Stała wariancja

Wykres residuów, może wskazywać brak jednorodność wariancji w czasie. Sprawdziliśmy to więc dwoma testami statystycznymi korzystając z biblioteki scipy, mianowicie testu Levene'a oraz Bartletta. Oba testy zwróciły wartość rzędu 10^{-6} co skłania nas do odrzucenia hipotezy zerowej i stwierdzenia, że nasze residua mają zmienną wariancję w czasie.

5.3. Rozkład normalny

Do sprawdzenia normalności rozkładu residuów skorzystaliśmy z porównania gęstości teoretycznej rozkładu normalnego o średniej równej 0 i wariancji równej 1. Po unormowaniu naszego rozkładu otrzymaliśmy następujący wykres.

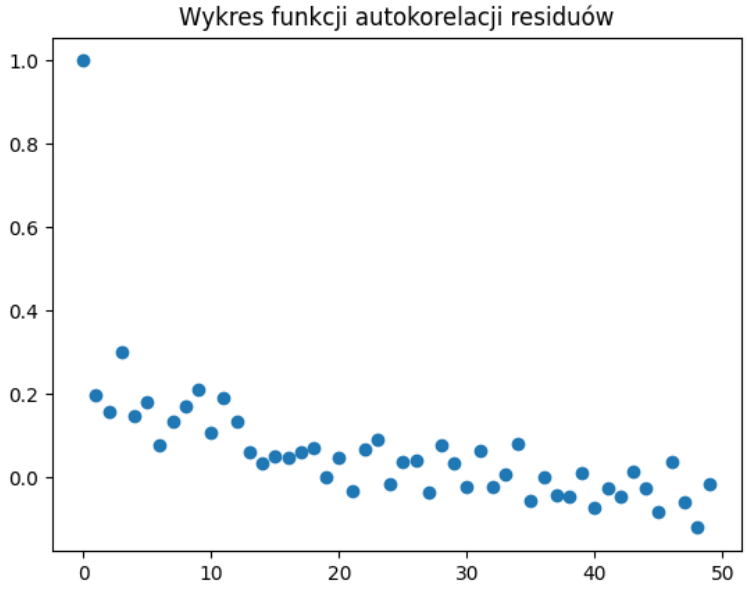


Rysunek 9: Histogram rozkładu residuów.

Na wykresie nie widać znaczącego podobieństwa między gęstościami, co potwierdzają także testy statystyczne Kołmogorowa-Smirnowa oraz Jarque-Bera. Oba testy zwróciły p-value około 0.01 co prowadzi do odrzucenia hipotezy, że rozkład residuów pochodzi z rozkładu normalnego.

5.4. Funkcja autokorelacji

Wykres funkcji autokorelacji dla 50 lagów pokazuje nam, że wartości tej funkcji są za wysokie aby stwierdzić nieskorelowanie, co także nie zgadza się z jednym z założeń klasycznego modelu regresji liniowej.



Rysunek 10: Wykres funkcji autokorelacji residuów.

6. Podsumowanie

Pomimo obiecujących początkowych wyników, po analizie regresji oraz jakości predykcji danych testowych możemy stwierdzić, że nasz model nie spełnia wszystkich założeń go dotyczących. Korzystanie z naszego modelu do predykcji przyszłych wartości firm na giełdzie byłoby ryzykowne. Jednak początkowe dopasowanie prostej świadczy o znacznej możliwości poprawy po ewentualnym usunięciu wartości odstających oraz wydłużeniu długości próbki.