Raport 1

Komputerowa analiza szeregów czasowych

Wykorzystanie poznanych metod służących do analizy zależności liniowej dla wybranych danych rzeczywistych

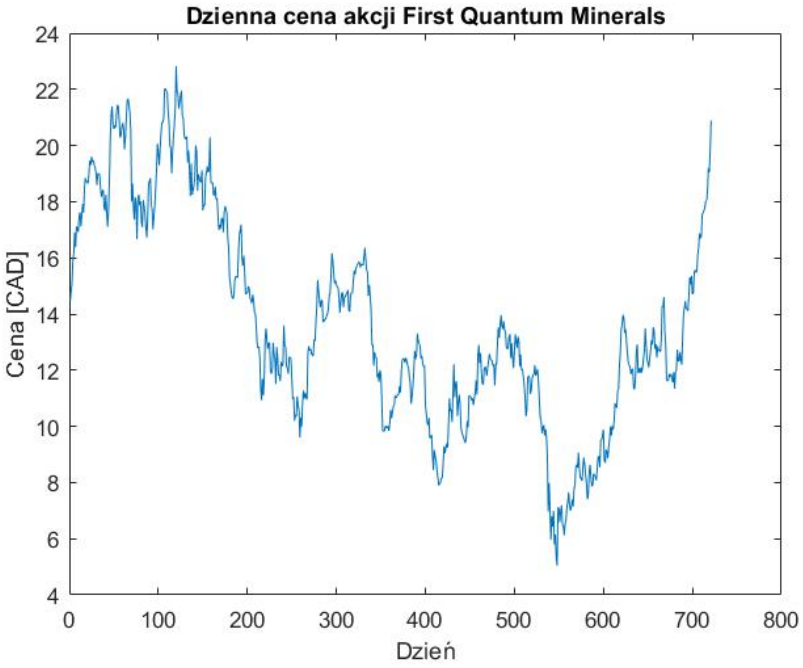
Patrycja Ozgowicz 249795
Martyna Boniatowska 249763

16.12.2020r.

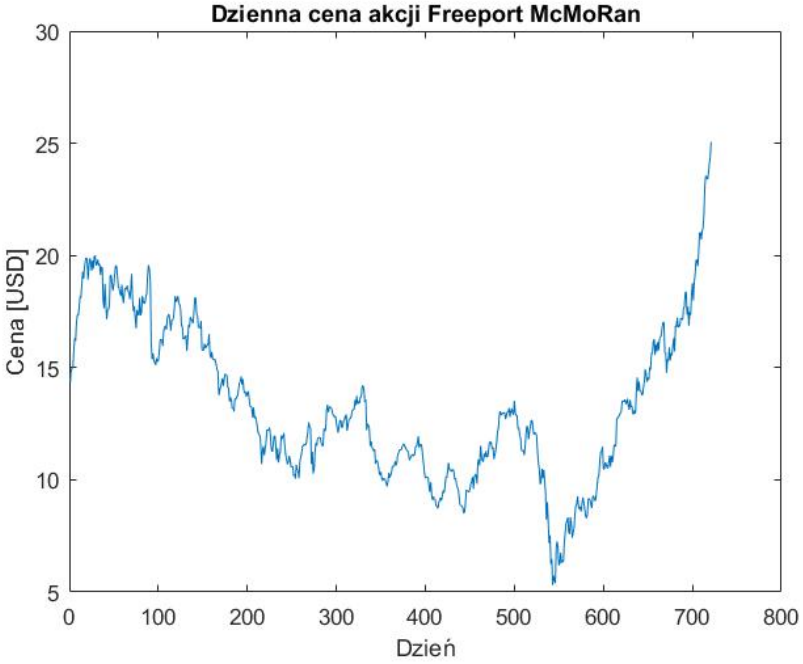
1 Przedstawienie i opis danych

Dane, które będą analizowane dotyczą dwóch firm First Quantum Minerals Ltd. oraz Freeport McMoRan. Pierwsza z nich jest kanadyjską spółką zajmującą się wydobyciem metali, której podstawową działalnością jest poszukiwanie, zagospodarowanie i wydobycie złóż. Jej głównym produktem jest miedź, która stanowi 80% przychodów od 2016 r. (właśnie w tym roku, wyprodukowano jej aż 539 458 ton). First Quantum Minerals założona w 1983 obecnie obsługuje kopalnie i projekty rozwojowe w Afryce, Australii, Finlandii, Hiszpanii, Turcji i Ameryce Łacińskiej. Druga z firm jest amerykańskim przedsiębiorstwem z siedzibą w Phoenix, operującym w branży wydobywczej miedzi, złota, molibdenu, kobaltu, ropy naftowej i gazu ziemnego. Freeport McMoRan założony w 1912 r. jest największym na świecie producentem molibdenu i największym publicznie notowanym producentem miedzi. Spółka angażuje się w wydobywanie surowców mineralnych w Ameryce Północnej, Ameryce Południowej i Indonezji. Dane, które będą analizowane to wartości kursu akcji tych firm odpowiednio na giełdach w Toronto (Toronto Stock Exchange) i Nowym Yorku (New York Stock Exchange). Rozpatrywany okres czasu to 3 lata, od 6 grudnia 2017 r. do 4 grudnia 2020 r. Wartości podano tylko dla dni roboczych, od poniedziałku do piątku. W każdej próbie jest ich 721, a rozpatrywany okres trwa 1094 dni. Ceny dotyczą teoretycznego kursu zamknięcia, ustalanego w fazie przed zamknięciem sesji. Są to ceny, po jakiej w sesji giełdowej inwestorzy zawarli ostatnią transakcję lub kilka ostatnich transakcji na danym instrumencie. Jednostką danych dotyczących firmy First Quantum Minerals jest dolar kanadyjski, natomiast firmy Freeport McMoRan dolar amerykański. Dane historyczne pochodzą ze strony www.kgmh.com[1], która jest oficjalną witryną spółki strategicznej KGHM Polska Miedź S.A. Strona ta, poświęcona jednej z największych polskich spółek skarbu państwa przedstawia m.in. zestawienia cen akcji kilku spółek zajmujących się wydobyciem i obróbką miedzi. Omawiane obserwacje pochodzą z jednego z takich zestawień. Obydwie grupy danych zostały przedstawione na poniższych wykresach 1,2:

[1]Link: https://kghm.com/pl/inwestorzy/akcje-i-obligacje/notowania



Rysunek 1: Wykres dziennych cen akcji firmy kanadyjskiej



Rysunek 2: Wykres dziennych cen akcji firmy amerykańskiej

Na wykresach 1, 2 przestawiono ceny akcji dwóch analizowanych firm od numeru obserwacji. Na osiach dni, wartości 1 odpowiada data 06.12.2017 r., natomiast wartości ostatniej - dzień 04.12.2020 r. Można zauważyć, że trajektorie obydwu wykresów są podobne. Tendencje wzrostowe lub spadkowe są utrzymywane w tych samych okresach. Wspólny dla obydwu wykresów jest spadek cen przypadający na końcówkę marca 2020 r., przy którym osiągana jest wartość minimalna. Był to okres, w którym państwa na całym świecie wprowadzały restrykcje mające zapobiec rozprzestrzenianiu się epidemii koronawirusa, które przyniosły poważne konsekwencje dla gospodarki. Jednocześnie załamanie cen ropy naftowej w wyniku rywalizacji Arabii Saudyjskiej z Rosją uderzyło mocno w sektor wydobywczy, a także w kurs dolara kanadyjskiego – który w tamtym czasie był najsłabszy od czterech lat. Można więc przypuszczać, że wykresy zachowują się w podobny sposób ze względu na wspólny obszar działalności spółek, głównie związany z wydobyciem i obróbką miedzi. Ceny akcji obydwu spółek zależą m.in. od cen miedzi na giełdzie czy światowego stanu gospodarki, stąd w ich zachowaniach można znaleźć podobieństwo. Przeprowadzona zostanie dokładniejsza analiza przedstawionych danych historycznych, z której zostaną wyciągnięte wnioski.

2 Informacje techniczne

W raporcie będą używane następujące oznazenia:

- MIN - wartość minimalna próby,
- MAX - wartość maksymalna próby,
- Q_i - i-ty kwartyl próby dla i=1,2,3,
- IQR - rozstęp międzykwartylowy,
- D - domiananta z próby,
- $\overline{X}$ - średnia arytmetyczna z próby,
- $\overline{X}_H$ - średnia harmoniczna z próby,
- $\overline{X}_G$ - średnia geometryczna z próby,
- n - ilość danych w zbiorach.

W raporcie będą używane nazwy:

- zbiór pierwszy - dane opisujące ceny akcji spółki First Quantum Minerals,
- zbiór drugi - zbiór danych opisujących ceny przedsiębiorstwa Freeport McMoRan.

Wszystkie przedstawiane wykresy zostały stworzone w programie Matlab.

3 Analiza jednomiarowa

3.1 Zbiór pierwszy - dane firmy First Quantum Minerals

Dla pierwszego zbioru danych tj. cen akcji spółki First Quantum Minerals obliczone zostały podstawowe statystyki miar położenia, rozproszenia, asymetrii i spłaszczenia. Uzyskane wyniki zostały przestawione w Tabeli 1 poniżej:

Charakterystyki	Wartości
Miary położenia	
Średnia arytmetyczna	13.7429
Średnia geometryczna	13.2011
Średnia harmoniczna	12.6470
Q_1	11.1275
Mediana	12.8800
Q_3	16.9200
MAX	22.8200
MIN	5.0400
Dominanta	11.7300
Miary rozproszenia	
Wariancja	14.6184
Odchylenie standardowe	3.8234
Rozstęp próby	17.7800
Rozstęp międzykwartylowy	5.7925
Współczynnik zmienności	0.2782
Miary asymetrii	
Współczynnik skośności	0.7898
Miary spłaszczenia	
Kutoza	2.3217

Tabela 1: Tabela z podstawowymi statystykami zbioru pierwszego

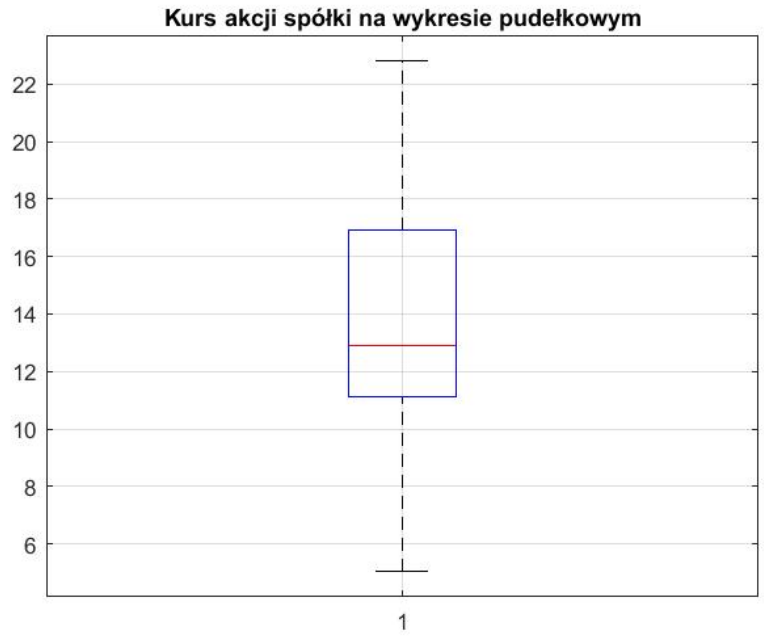
Z Tabeli 1 możemy odczytać, że średnia wartość cen akcji firmy First Quantum Minerals na przestrzeni ostatnich 3 lat wynosiła $\overline{X} = 13.7429$ CAD. Jest to wartość zbliżona do średniej geometrycznej i harmonicznej, wynoszących odpowiednio $\overline{X}_G = 13.2011$ CAD i $\overline{X}_h = 12.6470$ CAD. Średnia harmoniczna jest mniej wrażliwa na wartości odstające. Skoro jej wartość jest mniejsza od średniej arytmetycznej, to możemy przypuszczać, że maksimum jest bardziej oddalone od danych niż minimum. Jeżeli teraz spojrzymy na te wartości ekstremalne $MIN = 5.04$ i $MAX = 22.82$ możemy zauważyć, że średnia znajduje się

mniej więcej równo pomiędzy tymi wartościami, ale nieco bliżej minimum. Dodatkowo mediana dla danych jest równa 12.88 CAD, co jest wartością mniejszą od średniej arytmetycznej, choć zbliżoną.

Analizując miary rozproszenia, warto zauważyć, że rozstęp z próby, będący różnicą wartości minimalnej i maksymalnej wynosi 17.78 CAD. Oznacza to, że na przestrzeni trzech lat, ceny akcji firmy różniły się od siebie maksymalnie o 17.78 dolarów kanadyjskich. To dość nieduża różnica sugerująca stabilność spółki giełdowej First Quantum Minerals. Z tabeli 1 możemy również odczytać, że odchylenie standardowe z próby to niecałe 4 CAD. Oznacza to, że ceny w większości przyjmowały wartości o tyle odchylone od średniej. Wariancja, będąca kwadratem odchylenia standardowego nie jest dużą wielkością. Pozwala to sądzić, że dane zachowują się w sposób stabilny. Współczynnik zmienności przyjmuje około 28%, co jest dość typową wartością dla danych rzeczywistych i nie świadczy o ich dużej zmienności. Zatem ceny akcji kanadyjskiego przedsiębiorstwa zachowują się raczej w sposób regularny z nielicznymi wartościami odbiegającymi.

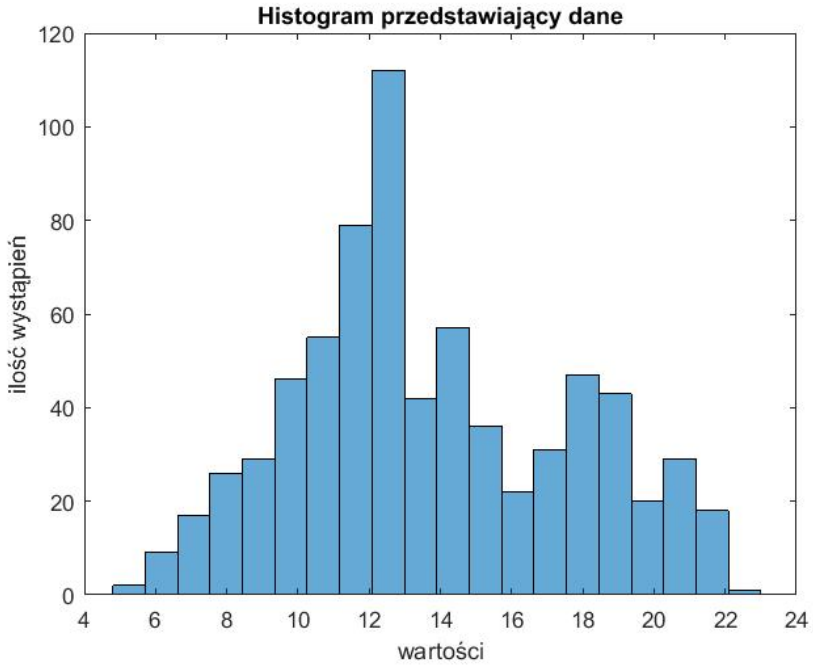
Zbadaną miarą asymetrii pierwszego zbioru danych jest współczynnik skośności. Jego wartość 0.7898 jest większa od zera. Oznacza to, że dane są prawostronnie skośne. Dodatkowo zachodzi: $D < Q_2 < \overline{X}$, co także świadczy o skośności dodatniej. Warto jednak zwrócić uwagę, że wszystkie te wartości są sobie bliskie, a współczynnik skośności nie jest znacząco większy od 0. Oznacza to, że większość wyników jest poniżej średniej, przy czym nie ma bardzo wielu elementów znacząco odbiegających od reszty.

Ostatnią zbadaną charakterystyką jest kurtoza, należąca do miar spłaszczenia. Jest to tak zwana miara koncentracji wyników, informująca o tym, jak bardzo obserwacje są skoncentrowane wokół średniej. Dla danych kurtoza wynosi 2.3217. Oznacza to, że rozkład jest mniej wysmukły niż normalny i posiada większe spłaszczenie (dane nieco mniej skoncentrowane niż przy rozkładzie normalnym). Dla lepszego zobrazowania rozkładu danych przedstawiono poniższy wykres pudełkowy:



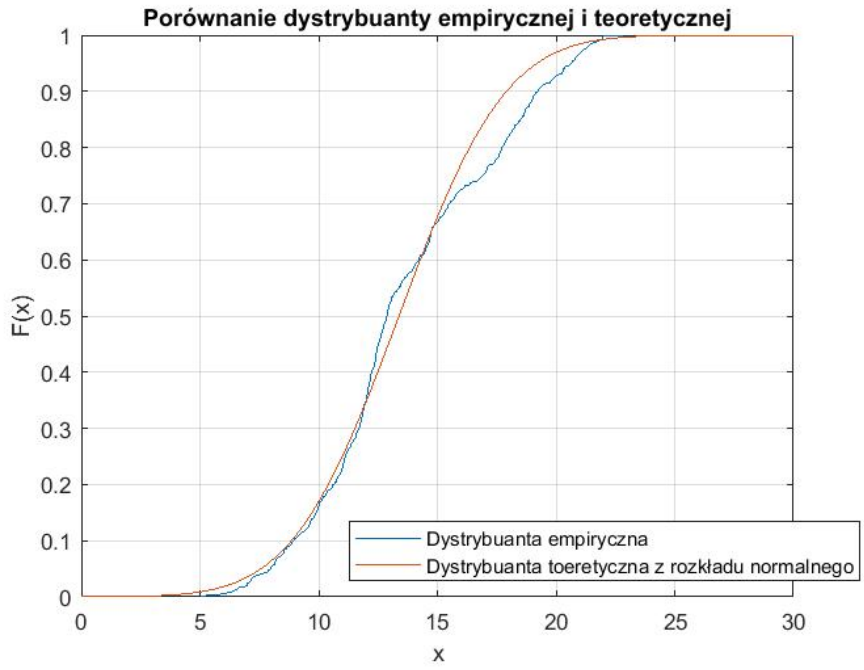
Rysunek 3: Wykres pudełkowy dziennych akcji zbioru pierwszego

Na wykresie 3 przedstawiono wykres skrzynkowy dla danych. Można z niego odczytać, że rozkład jest prawostronnie skośny, ponieważ pudełko nie jest równo podzielone przez medianę (czerwona linia na wykresie), a odległość MAX od mediany, jest większa niż odległość MIN od mediany. Ciekawym jest fakt, że na wykresie nie pojawiły się czerwone plusy, oznaczające wartości odstające. Są to dane, których wartości są poniżej dolnego wąsa lub powyżej górnego wąsa. Dolny wąs przyjmuje wartość $Q_1 - 1.5IQR$, natomiast góry $Q_3 + 1.5IQR$. Oznacza to, że w analizowanym zbiorze danych nie ma wartości wykraczających poza górny lub dolny wąs, które można by uznać za odstające. Dodatkowo długość pudełka, równa rozstępowi międzykwartylowemu (z Tabeli 1: 5.7925) mówi nam, że właśnie tam znalazło się 50% typowych danych. Poniżej przedstawiono histogram dla danych:



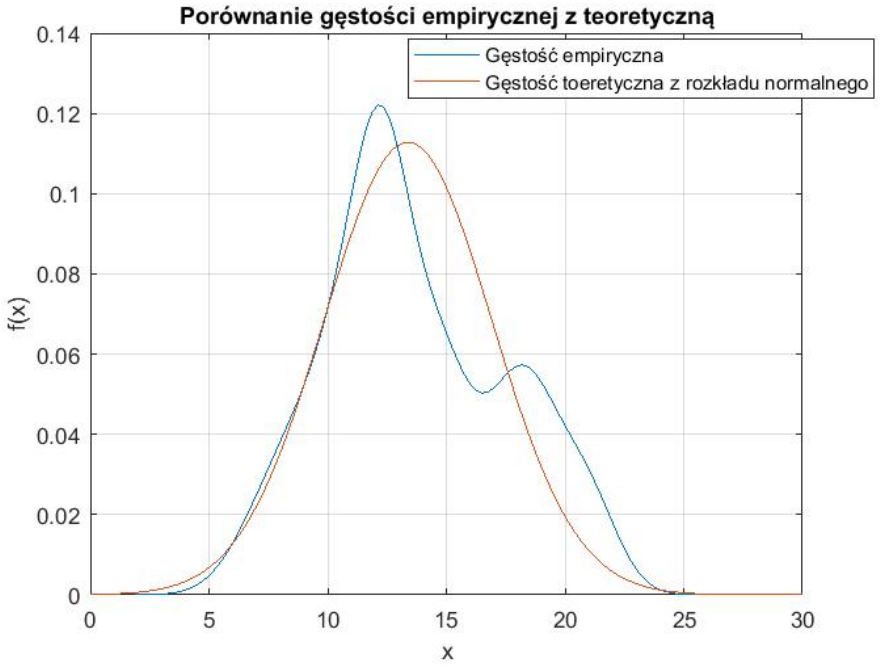
Rysunek 4: Histogram dla dziennych akcji zbioru pierwszego

Wykres 4 przedstawiający histogram dla danych potwierdza wcześniej przeprowadzoną analizę. Jest to histogram prawostronnie skośny. Dodatkowo najwyższy słupek pojawia się dla wartości w okolicach 12 CAD, a wyznaczona dominanta z Tabeli 1 przyjmowała wartość 11.73 CAD. Najwięcej danych znajduje się na lewo od wartości 12 CAD, czyli poniżej średniej, co również zauważono podczas analizy współczynnika skośności i wykresu pudełkowego. Na wykresie słupkowym nie pojawiło się dużo bardzo niskich słupków, co oznacza, że dane są dość mocno skoncentrowane i nie zawierają dużej ilości wartości skrajnych. Przeprowadzona analiza pozwala sugerować, że dane mogą pochodzić z rozkładu normalnego, mimo niewielkiej asymetrii i spłaszczenia. By sprawdzić te przypuszczenia, na kolejnych dwóch wykresach porównano gęstość i dystrybuantę empiryczną z gęstością i dystrybuantą rozkładu normalnego z parametrami średniej i odchylenia standardowego wyliczonymi z danych.



Rysunek 5: Wykres porównujący dystrybuanty dla danych ze zbioru pierwszego

Na wykresie 5 porównano dystrybuantę empiryczną z danych z dystrybuantą rozkładu normalnego z parametrami $\mu = 13.7429$ i $\sigma^2 = 14.6184$ - średnia i wariancja dla danych podana w Tabeli 1. Możemy zauważyć, że przebieg dystrybuant jest podobny. Dystrybuanta empiryczna przyjmuje wartości wokół tej pochodzącej z rozkładu normalnego. Ich kształt jest bliski, choć nie identyczny. Stąd przypuszczenie, że rozkład normalny mógłby być przybliżeniem dla analizowanych danych. W celu lepszego sprawdzenia tej teorii porównano również gęstości.



Rysunek 6: Wykres porównujący gęstości dla danych ze zbioru pierwszego

Wykres 6 przedstawia gęstość empiryczną porównaną z gęstością rozkładu normalnego, z parametrami średniej i wariancji z danych. Przebieg tych dwóch funkcji jest do siebie zbliżony. W początkowej fazie wykresy się prawie pokrywają. Gęstość z danych ma nieco smuklejszy kształt u góry i bardziej spłaszczony u dołu z prawej strony. Porównując funkcje z wykresu 6 można zauważyć większą rozbieżność, niż w przypadku dystrybuant. Stąd wniosek że dane zachowują się podobnie do rozkładu normalnego, ale z niego nie podchodzą.

3.2 Zbiór drugi - dane firmy Freeport McMoRan

Dla drugiego zbioru danych tj. cen akcji spółki amerykańskiej również obliczone zostały podstawowe statystyki miar położenia, rozproszenia, asymetrii i spłaszczenia. Otrzymane wyniki przestawiono w poniższej Tabeli 2:

Charakterystyki	Wartości
Miary położenia	
Średnia arytmetyczna	13.3814
Średnia geometryczna	12.9177
Średnia harmoniczna	12.4532
Q_1	10.8075
Mediana	12.8000
Q_3	16.1925
MAX	25.0600
MIN	5.3100
Dominanta	11.4900
Miary rozproszenia	
Wariancja	12.5075
Odchylenie standardowe	3.5366
Rozstęp próby	19.7500
Rozstęp międzykwartylowy	5.3850
Współczynnik zmienności	0.2643
Miary asymetrii	
Współczynnik skośności	0.5200
Miary spłaszczenia	
Kutoza	2.7282

Tabela 2: Tabela z podstawowymi statystykami zbioru drugiego

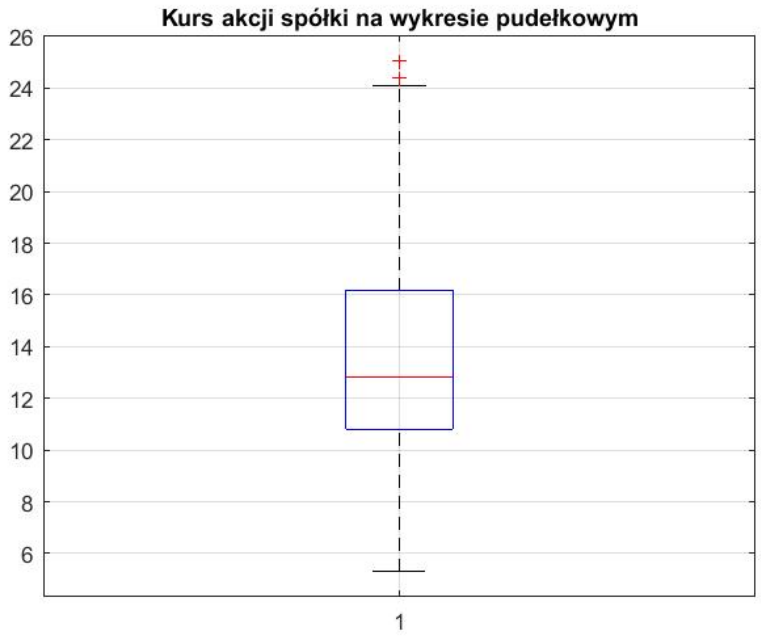
Na podstawie danych z tabeli 2 można ocenić, w jaki sposób rozkładają się dane ze zbioru drugiego. Wszystkie wyliczone średnie oscylują w okolicach 13 USD. $\overline{X} = 13.3814$ pokazuje jaką wartość powinny osiągać dane gdyby w obserwowanym okresie czasu przyjmowały równą wartość. Średnia ta jest niestety wrażliwa na obserwacje odstające. $\overline{X}_g = 12.9177$ nie jest jednak najbardziej wiarygodną statystyką, ponieważ najlepiej opisuje dane, zmieniające się w postępie geometrycznym, jako że poddawane analizie są dane rzeczywiste i nie sprawiają one wrażenia zachowujących się w sposób potęgowy, średnia ta nie będzie dobrym wyznacznikiem. Ostatnia średnia $\overline{X}_h = 12.4532$ jest mniej wrażliwa na wartości odstające niż $\overline{X}$, ponieważ daje ona równą wagę wszystkim danym. Widać, że średnia ta jest bardziej zbliżona do dominanty, czyli najczęściej występującej wartości w zbiorze. Największa osiągnięta przez dane wartość w badanym okresie wynosi 25.06 USD i patrząc na wykres 2, można zaobserwować, że została ona osiągnięta w początkowym okresie obserwacji. Minimalna wartość osiągnięta w zbiorze drugim to 5.31 USD, która zgodnie z wykresem 2 została osiągnięta w okolicach 550 obserwacji. Punktem, który rozdziela zbiór

cen akcji spółki FreePoint McMoRan na dwa równoliczne podzbiory jest mediana, która w tym przypadku wynosi 12.80 USD. Ostatnimi miarami położenia, zawartymi w tabeli 2 są kwartyle. Q_1 poniżej, którego znajduje się 25% obserwacji, jest znacząco oddalony od osiąganej wartości minimalnej. Natomiast Q_3 poniżej, którego znajduje się 75% obserwacji zachowuje się podobnie w stosunku do maksimum. Stąd wniosek, że w zbiorze drugim znajduje się kilka mocno odstających elementów.

Biorąc pod uwagę miary rozproszenia, zamieszczone w tabeli 2 można zauważyć, że wariancja zbioru, świadcząca o tym jak duże jest zróżnicowanie danych w zbiorze, przyjmuje wartość 12.5075. Nie jest to duża wartość, stąd wniosek o niewielkim zróżnicowaniu przyjmowanych wartości. Patrząc na dane, można wnioskować, że zróżnicowanie nie jest duże, ponieważ sam rozstęp próby wynosi ok. 20 USD. Rozstęp międzykwartylowy daje informację o długości przedziału, w którym znajduje się 50% typowych danych z pominięciem drugiej połowy skrajnych wartości. Otrzymana wartość 5.3850 jest nieduża, stąd wniosek, że dane są dość mocno skoncentrowane poza pewnymi wartościami odstającymi.

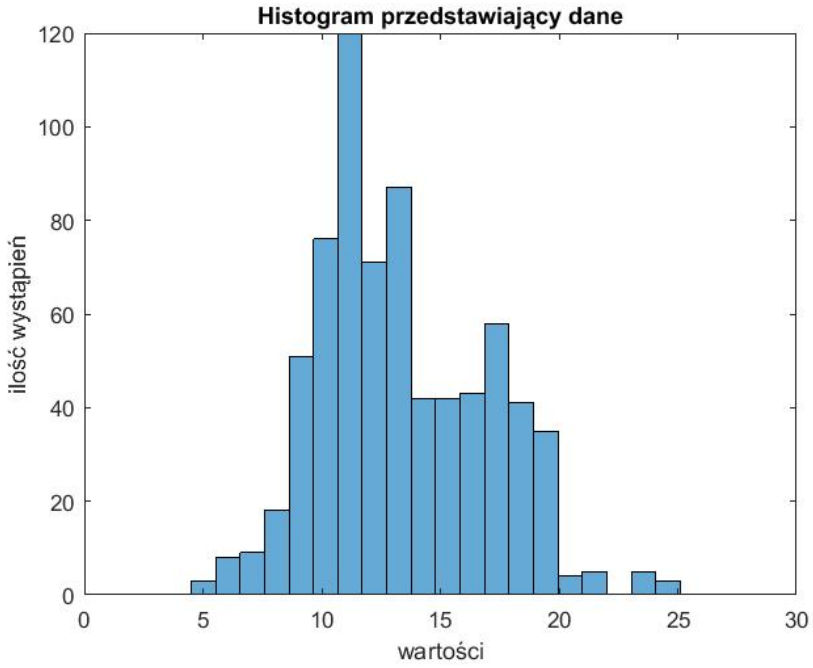
Kolejną miarą jest współczynnik skośności, który w tym przypadku wynosi 0.52. Zgodnie z jego interpretacją, otrzymana wartość, która jest większa od 0, pozwala wysnuć wniosek o prawostronnej asymetrii rozkładu.

Ostatnią opisaną w tabeli 2 miarą jest kurtoza, której wartość jest równa 2.7282. Zgodnie z definicją im wyższy współczynnik kurtozy, tym większa koncentracja danych wokół średniej. Otrzymany dla danych ze zbioru drugiego wynik jest mniejszy od 3, a co za tym idzie, ich rozkład jest mniej wysmukły niż rozkład normalny. Innymi słowy, następuje tu mniejsza koncentracja danych wokół średniej niż dla rozkładu normalnego.



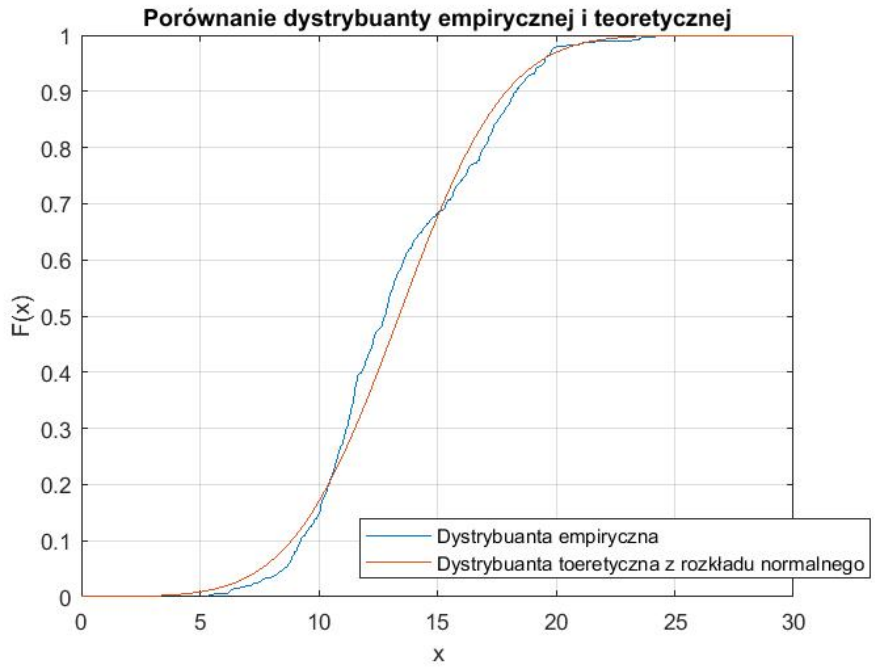
Rysunek 7: Wykres pudełkowy dziennych akcji zbioru drugiego

Na wykresie pudełkowym można łatwo zobrazować rozkład danych przy pomocy charakterystyk liczbowych przedstawionych w tabeli 2. Na wykresie 7 zobrazowany został rozkład dziennych cen akcji spółki FreePoint McMoRan. Jak widać, znajduje się na nim bardzo mało wartości odstających, czyli tzw. outliers'ów. Znajdują się one jedynie powyżej górnego wąsa wykresu. Czerwoną linią została oznaczona mediana, dzieli ona zbiór danych na dwa równoliczne pozdbiory. Z wykresu 2 można odczytać, że znajduje się ona bliżej Q_1, stąd więc pomiędzy nimi powinny się znajdować najpopularniejsze wartości. Z tabeli 2 można odczytać wartość dominanty, która należy właśnie do tego zbioru. Różnica pomiędzy MAX a Q_3, wydaje się być większa niż między MIN a Q_1 stąd wniosek, że znajduje się więcej obserwacji odstających powyżej średniej niż tych ekstremalnie małych.



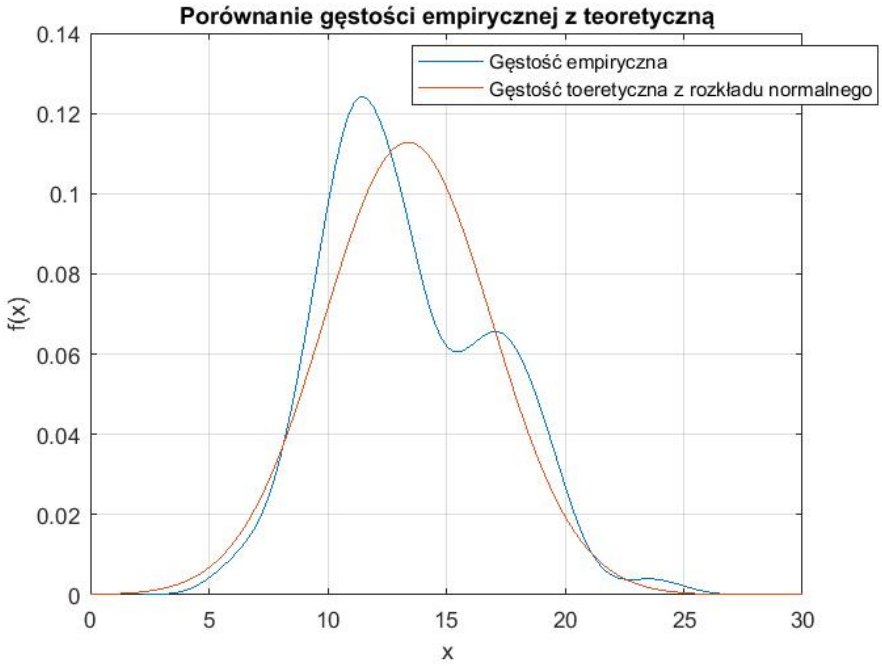
Rysunek 8: Histogram dla dziennych akcji zbioru drugiego

Analizując histogram przedstawiony na wykresie 8 widać, że wcześniej postawione założenia o rozkładzie wartości w zbiorze drugim są prawdziwe. Najwyższy słupek to ten dla 11 USD, a właśnie w okolicach tej wartości pojawiła się dominanta wyznaczona w tabeli 2. Największa koncentracja danych znajduję się pomiędzy 9 USD a 13 USD. Podczas analizy wykresu 7, czyli wykresu pudełkowego zauważone zostało skoncentrowanie danych między medianą a pierwszym kwartylem. Patrząc na wyznaczone tam wartości oraz to, co dzieje się na przedstawionym histogramie można powiedzieć, że ta tendencja jest prawdziwa. Widać, że na lewo od wartości 9 USD występuje już niewiele obserwacji, oraz że najmniejsza zaznaczona wartość to ok. 5, co pokrywa się z wyznaczonym w tabeli 2 minimum. Można zauważyć, że choć ekstremalnie wyższe wartości nie występują często to jest ich więcej niż tych ekstremalnie małych. Opisując dane ze zbioru drugiego, warto zbadać, z jakiego mogą być rozkładu. W tym celu wyrysowano gęstość i dystrybuantę.



Rysunek 9: Wykres porównujący dystrybuanty dla danych ze zbioru drugiego

Po narysowaniu wykresu empirycznej dystrybuanty nasunął się jeden wniosek, a mianowicie, że mogą pochodzić one z rozkładu normalnego. Dla sprawdzenia została narysowana dystrybuanta teoretyczna z wyznaczonymi wcześniej parametrami $\mu = \overline{X}$ oraz $\sigma^2 = Var(X)$. Wykres 9 przedstawia to porównanie. Widać, że dystrybuanta empiryczna w znacznym stopniu pokrywa się z tą teoretyczną o wspomnianych wyżej parametrach. Stąd wniosek, że do analizy jednowymiarowej można by spróbować przybliżyć dane rozkładem normalnym. Po porównaniu dystrybuant warto zobaczyć, jak będą zachowywały się gęstości.

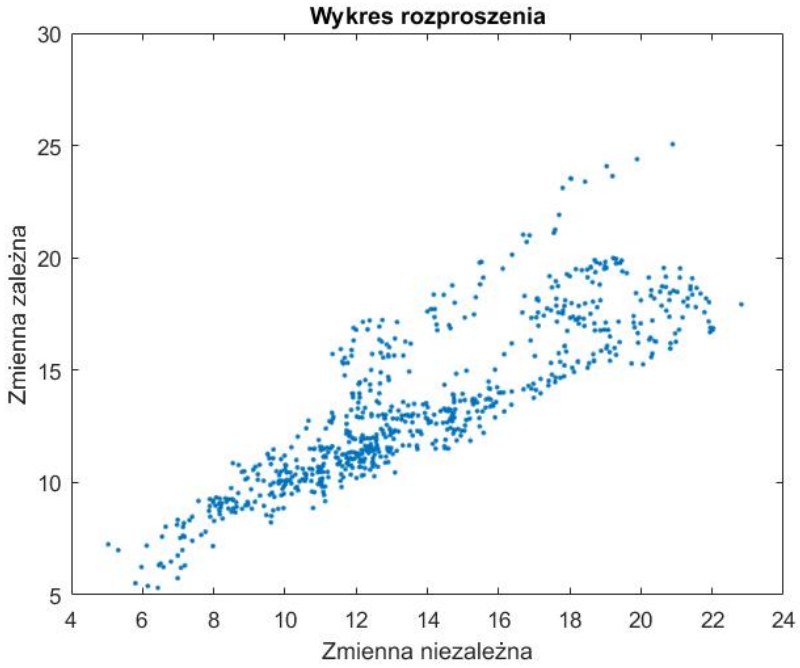


Rysunek 10: Wykres porównujący gęstości dla danych ze zbioru drugiego

Na wykresie 10 przedstawiona została gęstość empiryczna oraz teoretyczna z dokładnie takimi samymi parametrami jak w przypadku tworzenia teoretycznej dystrybuanty. Tutaj jednak funkcje nie są tak zbliżone jak w przypadku dystrybuanty. Widać, że gęstość empiryczna zachowuje się podobnie do tej teoretycznej, jednak różnice w przyjmowanych wartościach są trochę większe. W tabeli 2 wyznaczony został współczynnik kurtozy. Zgodnie z tym co zostało otrzymane, dane ze zbioru drugiego powinny mieć podobny rozkład wokół średniej jak w przypadku rozkładu normalnego. Dane rzeczywiste zachowują się podobnie do próby z rozkładu normalnego, jednak widoczne są różnice. Zatem nie można jednoznacznie stwierdzić, że rozkładem danych jest rozkład normalny.

4 Analiza zależności liniowej pomiędzy zbiorami danych

Niech dane opisujące ceny akcji spółki First Quantum Minerals będą zmienną niezależną, a dane opisujące ceny akcji przedsiębiorstwa Freeport McMoRan zmienną zależną. Wtedy wykres rozproszenia dla danych wygląda następująco:



Rysunek 11: Wykres rozproszenia danych, gdzie zbiór pierwszy jest zmienną niezależną, a zbiór drugi zmienną zależną

Na powyższym wykresie 11 można zauważyć pewną zależność między zbiorami danych. Wraz ze wzrostem zmiennej objaśniającej, rośnie zmienna objaśniana. Można powiedzieć, że wzrost ten ma w przybliżeniu charakter liniowy dodatni. Stąd przypuszczenie, że zmienne są zależne od siebie liniowo. Zatem rozpatrywanym modelem będzie klasyczny model regresji liniowej. Przez x_i będziemy oznaczać dane firmy First Quantum Minerals, natomiast y_i będzie oznaczać ceny spółki Freeport McMoRan dla $i = 1, 2, ..., n$, gdzie $n = 721$. Zależność liniowa wyraża się wzorem:

$$\hat{y_i} = b_0 + b_1 x_i \tag{1}$$

b_0, b_1 są współczynnikami, które można wyznaczać przy pomocy metody najmniejszych kwadratów. Wyrażają się wzorami:

$$b_1 = \frac{\sum_{i=1}^{n} x_i(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{2}$$

$$b_0 = \overline{y} - b_1 \overline{x} \tag{3}$$

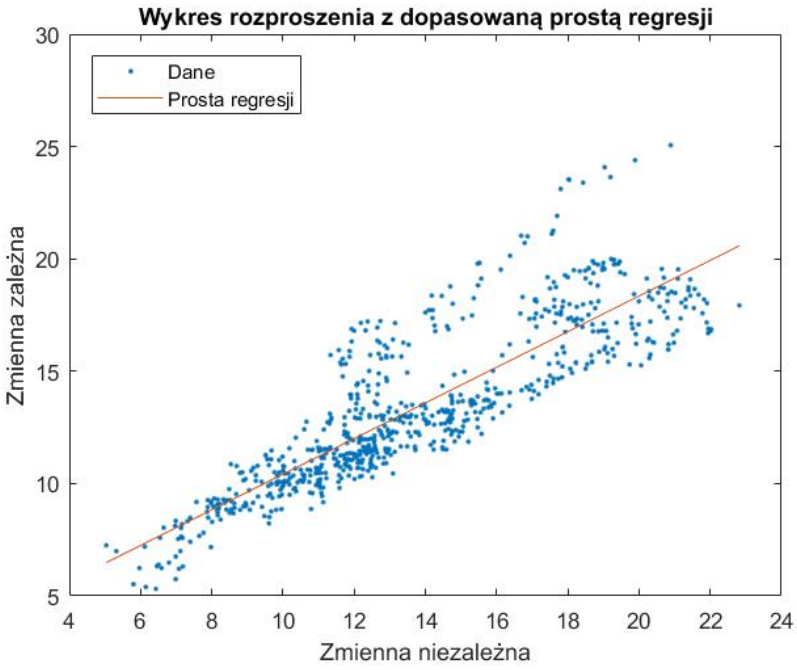
Dla analizowanych danych obliczono współczynniki z powyższych wzorów 2,3:

$$b_0 = 2.4800, \qquad b_1 = 0.7932$$

Zatem prosta regresji z wzoru 1 wygląda następująco:

$$\hat{y}_i = 2.48 + 0.7932 x_i$$

Wykres rozproszenia z naniesioną prostą regresji dla danych:



Rysunek 12: Wykres rozproszenia danych, gdzie zbiór pierwszy jest zmienną niezależną, a zbiór drugi zmienną zależną z dopasowaną prostą regresji

Wykres 12 przedstawia dane wraz z dopasowaną do nich prostą regresji wyznaczoną ze wzoru 1. Na wykresie widać, że prosta całkiem dobrze przybliża dane. Jednak jest część obserwacji, znajdujących się stosunkowo wysoko od prostej, które można by uznać za niepasujące do modelu. Stąd pomysł, by lepiej przygotować dane, wykorzystując poznane metody.

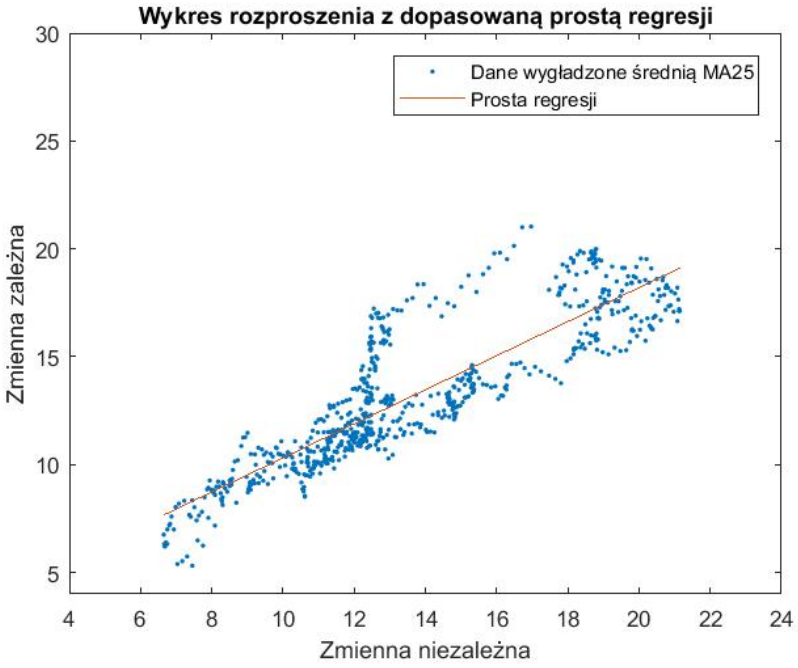
4.1 Metoda średniej ruchomej

Metoda średniej ruchomej polega na wygładzeniu danych poprzez ich średnią. Średnia ruchoma działa jak filtr, eliminując z szeregu wahania krótkookresowe poprzez zastąpienie pierwotnych wartości szeregu ciągiem średnich obliczanych z sąsiadujących ze sobą wyrazów szeregu czasowego. Jej obliczanie jest proste, dlatego często jest wykorzystywana na różnych rodzajach wykresów giełdowych.

Nieparzystą średnią ruchomą $(2p+1)$MA, $p \in \mathbb{N}$ wyrażamy wzorem:

$$\widetilde{y_k} = \frac{1}{2p+1} \sum_{j=-p}^{p} y_{k+j} \tag{4}$$

Średnią ruchomą MA25 ze wzoru 4 zastosowano do danych i otrzymano następujący wykres:



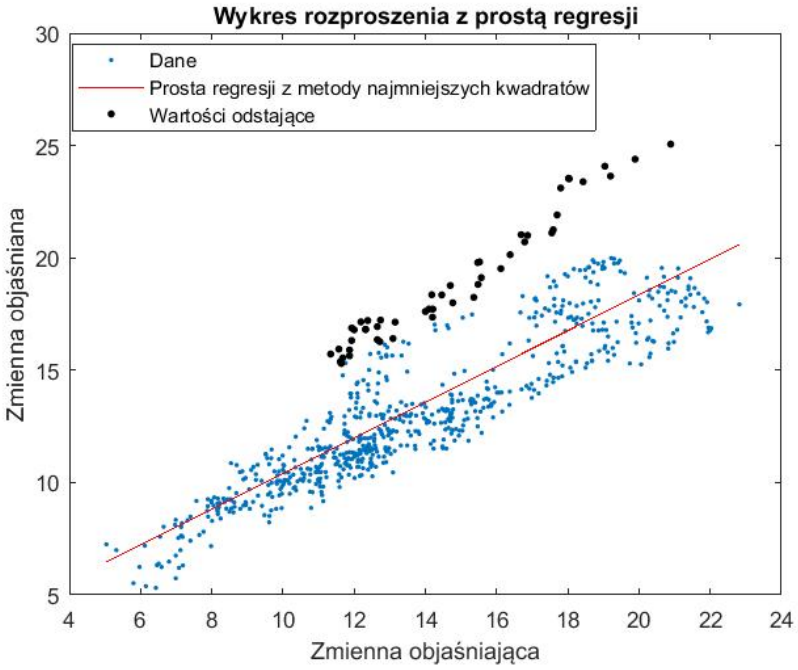
Rysunek 13: Wykres rozproszenia danych wygładzonych przy pomocy średniej ruchomej MA25 z nową prostą regresji dopasowaną do danych

Na wykresie 13 pokazano wykres rozproszenia wygładzonych danych oraz nową prostą regresji wyznaczoną ze wzoru 1 dla wygładzonych danych. Porównując ten wykres z rysunkiem 12, na którym znajdują się dane rzeczywiste, możemy zauważyć pewną różnicę. Mniej jest wartości bardzo dalekich od prostej (wcześniej obserwacje ponad prostą sięgały wysokości 25). Nowe dane wyglądają też na bardziej skoncentrowane. Można więc przypuszczać, że zależność liniowa będzie tu lepszym przybliżeniem.

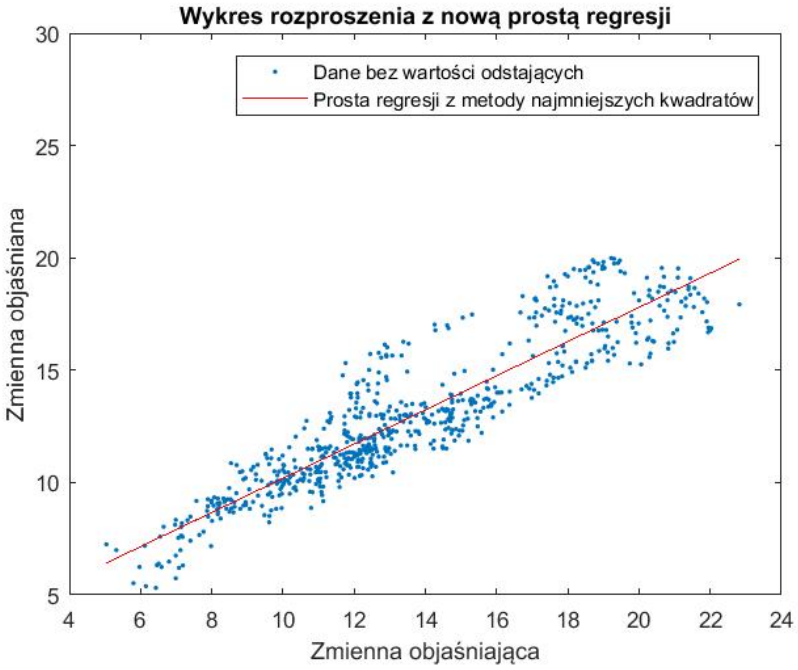
4.2 Metoda usuwania wartości odstających przy użyciu percentyli

W celu znalezienia wartości odstających stworzono wektor błędów, tzn. wektor różnic pomiędzy rzeczywistymi wartościami zmiennej objaśnianej, a wartościa-

mi na prostej regresji. Za wartości odstające uznano wartości znajdujące się poza przedziałem $(Q_1 - 1.5IQR,\ Q_3 + 1.5IQR)$, gdzie Q_1, Q_3 są kwartylami rzędu 0.25 i 0.75 wektora błędów, a IQR jest różnicą między nimi. Stworzono wykres z zaznaczonymi wartościami uznanymi za odstające oraz nowy wykres rozproszenia z dopasowaną prostą regresji dla nowych danych.



Rysunek 14: Wykres rozproszenia danych z dopasowaną prostą regresji wyznaczoną metodą najmniejszych kwadratów oraz zaznaczonymi wartościami odstającymi



Rysunek 15: Wykres rozproszenia danych bez wartości odstających z dopasowaną do nich prostą regresji wyznaczoną metodą najmniejszych kwadratów

Powyższe wykresy 14, 15 przedstawiają, jak zadziałała wybrana metoda usuwania wartości odstających. Na wykresie 14 zaznaczono obserwacje, które zostały uznane za ekstremalne, a następnie usunięte. Takich wartości było 49. Zatem liczność nowych danych po odrzuceniu outliersów wynosi 672. Wykres 15 wydaje się mieć najlepszą zależność liniową (tj. prosta regresji jest najlepszym przybliżeniem dla danych). Aby się o tym przekonać, wyliczone zostały pewne wskaźniki oceniające zależność liniową.

4.3 Ocena zależności liniowej

Wskaźnikami, które będą oceniać zależność liniową są:

- SSE - suma kwadratów błędów, czyli całkowita wartość błędu oszacowania. Wyraża się wzorem:
$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{5}$$
- SSR - suma regresyjna kwadratów. Stanowi miarę całkowitej poprawy dokładności przewidywań w przypadku stosowania regresji w porównaniu z sytuacją gdy nie uwzględniamy wartości zmiennej objaśniającej. Daje

informacje, jak bardzo wartość predykowana różni się od średniej z danych. Jest to tak zwana zmienność wyjaśniona przez model.

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 \tag{6}$$

- SST - całkowita suma kwadratów. Stanowi miarę całkowitej zmienności wartości y bez odniesienia do zmiennej objaśniającej. Mówi się, że SST jest funkcją wariancji zmiennej y. Można ją zapisać jako:

$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2 = SSE + SSR \tag{7}$$

- Współczynnik korelacji Pearsona - współczynnik określający poziom zależności liniowej między zmiennymi losowymi. Jego wartość mieści się w przedziale domkniętym [-1, 1]. Im większa jest jego wartość bezwzględna, tym silniejsza jest zależność liniowa między zmiennymi. 0 - oznacza brak liniowej zależność, 1 – oznacza zależność dodatnią, a -1 – oznacza zależność ujemną między cechami. Wyraża się wzorem:

$$R = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{8}$$

- Współczynnik determinacji R^2 - miara jakości dopasowania modelu. Mówi o tym, jaki procent jednej zmiennej wyjaśnia zmienność drugiej zmiennej. Przyjmuje on wartości od 0 do 1.

$$R^2 = \frac{SSR}{SST} \tag{9}$$

  Wartość maksymalna ostatniego współczynnika jest osiągana, gdy regresja idealnie pasuje do danych, co ma miejsce wtedy, gdy każdy z punktów danych leży dokładnie na oszacowanej linii regresji. Wówczas nie ma błędów oszacowania, a więc wartości resztowe (residua) wynoszą 0, a więc SSE=0, a wtedy SST = SSR, a R^2=1. Zatem model jest dobry, gdy wartości SST, SSE, SSR są bliskie 0, a R^2 jest blisko 1 i im większa wartość R^2, tym lepsze dopasowanie regresji do zbioru danych.

W celu porównania zależności liniowej danych rzeczywistych, danych wygładzonych średnią ruchoma MA25 i danych, z których usunięto wartości odstające, wyliczono wszystkie powyższe wskaźniki ze wzorów 5,6,7,8,9 i wyniki umieszczono w Tabeli 3:

Współczynnik	Dane rzeczywiste	Dane wygładzone MA25	Dane bez wartości odstających
SSE	2382.7	1858.5	1174.7
SSR	6622.7	5872.9	5868.2
SST	9005.4	7731.4	7042.9
R	0.8576	0.8716	0.9128
R^2	0.7354	0.7596	0.8332

Tabela 3: Tabela z porównaniem współczynników badających zależność liniową dla trzech zbiorów danych

Powyższa tabela 3 zestawia wskaźniki badające zależność liniową dla danych rzeczywistych, danych wygładzonych średnią ruchomą MA25 i danych, z których usunięto wartości odstające przy pomocy percentyli. Z tabeli można odczytać, że każdy ze wskaźników: SSE, SSR, SST jest najmniejszy dla ostatniego zbioru danych. Wtedy też wartości R i R^2 są największe w porównaniu z pozostałymi. Pozwala to jednoznacznie stwierdzić, że model regresji liniowej jest najlepiej dopasowany do trzeciego zbioru danych. Można też zauważyć, że dane wygładzone średnią ruchomą mają wskaźniki SSR, SST, SSE mniejsze od danych rzeczywistych, a współczynnik korelacji Pearsona i jego kwadrat są bliższe 1. Stąd wniosek, że przygotowanie danych zostało przeprowadzone poprawnie w obydwu przypadkach, jednak dla zbioru 3 regresja liniowa jest najlepszym przybliżeniem. Dane te są przedstawione na wykresie 15. Porównując ten wykres z wykresem 12 i 13, na których przedstawiono pozostałe zbiory danych, rzeczywiście można zauważyć, że tutaj obserwacje są najbliżej prostej. Zatem w dalszej części pracy analizowane będą dane, z których ucięto wartości odstające na podstawie metody z podrozdziału 4.2. Długość nowych danych to 672, a współczynniki prostej regresji wynoszą:

$$b_0 = 2.5830, \qquad b_1 = 0.7607$$

W dalszej części pracy zostanie przeprowadzona estymacja powyższych współczynników. Aby móc ją przeprowadzić, należy wprowadzić teoretyczny model regesji liniowej dla zmiennych losowych.

4.4 Teoretyczny model regresji liniowej

W teoretycznym modelu regresji pojawia się element losowości, czyli należy traktować niektóre elementy jako zmienne losowe. W modelu Y_i jest traktowany jako zmienna losowa. Według wzoru model ten prezentuje się następująco:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{10}$$

gdzie:

- Y_i zmienna losowa
- β_0 stała odpowiadająca wcześniej używanemu b_0

- β_1 stała odpowiadająca wcześniej używanemu b_1
- ϵ_i niezależne zmienne losowe, dla której $E(\epsilon_i) = 0$ oraz $Var(\epsilon_i) = \sigma^2$. Na zajęciach zawsze zakładano rozkład $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- x_i wartość deterministyczna
- $i = 1, 2, ..., 672$

Dla tak zdefiniowanego modelu regresji liniowej można wyznaczyć wartość średnią oraz wariancję:

- $E(Y_i) = \beta_0 + \beta_1 x_i$ - wartość tą można interpretować tak, że Y_i średnio dąży do modelu regresji liniowej.
- $Var(Y_i) = \sigma^2$ interpretacją tego wyniku może być fakt, że im mniejsza wariancja tym Y_i jest bliżej średniej wartości, czyli mniej się waha.

Posiadając tak zdefiniowany model można podjąć próbę estymacji parametrów stałych występujących w modelu. Wyznaczone estymatory będą już jednak zmiennymi losowymi.

Korzystając z Centralnego Twierdzenia Granicznego lub opierając się na funkcjach charakterystycznych, można wyznaczyć rozkłady estymatorów z modelu regresji liniowej, które prezentują się następująco:

$$\hat{\beta_1} \sim \mathcal{N}\left(\beta_1, \frac{\sigma}{\sqrt{\sum (x_i - \overline{x})^2}}\right)$$

$$\hat{\beta_0} \sim \mathcal{N}\left(\beta_0, \sigma\sqrt{\frac{1}{n}\frac{\overline{x}^2}{\sum (x_i - \overline{x})^2}}\right)$$

W dalszej części pracy założono, że analizowane dane pochodzą właśnie z takiego modelu regresji liniowej. Przyjęto, że $Y_i = y_i$. By sprawdzić, czy jest to poprawne założenie przeprowadzona zostanie poniższa analiza. Pozwoli ona wywnioskować, czy zastosowanie klasycznego modelu regresji liniowej jest odpowiednie.

4.5 Estymacja punktowa

Do estymacji parametrów β_1, β_2 w modelu regresji liniowej użyto metody najmniejszych kwadratów, która pozwala wyznaczyć estymatory nieobciążone o najmniejszej wariancji. Wzory pozwalające wyznaczyć konkretne wartości estymowanych parametrów prezentują się następująco:

1. Estymator współczynnika β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})Y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{11}$$

2. Estymator współczynnika β_0

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{x} \tag{12}$$

Dla analizowanych danych estymatory wyliczone ze wzorów 11,12 osiągają poniższe wartości:

$$\hat{\beta}_1 = 0.760653950792117$$

$$\hat{\beta}_0 = 2.583032769497956$$

W celu uzyskania lepszej informacji o dokładności szacunków warto wyznaczyć błędy standardowe dla wyznaczonych estymatorów. Zanim jednak będzie można przejść do szacowania błędów, należy wyjaśnić jak wygląda wzór na estymator parametru σ^2, ponieważ analizowane dane są rzeczywiste, w związku z czym σ jest nieznana. Estymator ten będzie przydatny zarówno podczas wyznaczania błędów standardowych jak i w późniejszym paragrafie do konstrukcji przedziałów ufności.

Estymator parametru σ^2:

$$S^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} \tag{13}$$

gdzie:

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
Y_i - zmienne losowe ze wzoru 10, przyjęte jako y_i

Posiadając tak zdefiniowany estymator σ^2 można już wyznaczyć błędy standardowe:

$$std(\hat{\beta}_1) = S\sqrt{\frac{1}{n}\frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = 0.0069$$

$$std(\hat{\beta}_0) = S\sqrt{\frac{1}{n}\frac{\overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = 5.0720 \cdot 10^{-4}$$

Otrzymane błędy standardowe przyjmują bardzo małe wartości stąd wniosek,

o niewielkim popełnionym błędzie estymacji i bardzo dobrym dopasowaniu danych do modelu.

4.6 Estymacja przedziałowa

Kolejnym sposobem na estymację parametrów w modelu regresji liniowej będzie estymacja przedziałowa, w tym wypadku, jak mówi sama nazwa wyznaczony zostaje przedział, do którego z zadanym prawdopodobieństwem będzie należeć estymowany parametr. Do wyznaczenia przedziału ufności potrzebny jest parametr α, który wyznacza to prawdopodobieństwo jako $1 - \alpha$. Jako że poddawane analizie dane są rzeczywiste, zostanie rozważony przypadek z nieznaną wariancją, dla którego stosowane poniżej kwantyle muszą pochodzić z rozkładu t-student z $n-2$ stopniami swobody. Do wyznaczenia przedziału potrzebne jest zdefiniowanie statystyki:

Dla estymatora β_1 statystyka T_1 o rozkładzie t-studenta o n-2 stopniach swobody ma postać:

$$T_1 = \frac{\hat{\beta}_1 - \beta_1}{S} \sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Natomiast skonstruowany przy jej pomocy przedział:

$$\left[\hat{\beta}_1 - \frac{St_{n-2,1-\frac{\alpha}{2}}}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}; \hat{\beta}_1 + \frac{St_{n-2,1-\frac{\alpha}{2}}}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} \right]$$

gdzie $t_{1-\frac{\alpha}{2},n-2}$ to kwantyl z rozkładu t-studenta z n-2 stopniami swobody. Dla analizowanych zbiorów danych przedział ufności estymatora β_1 wygląda tak:

$$[0.7348, 0.7865]$$

Wyznaczony wyżej parametr β_1 mieści się w wyznaczonym przedziale ufności na dla $\alpha = 0.05$

Dla parametru β_0 postepowanie jest analogiczne, zaczynając od zdefiniowania statystki:

$$T_0 = \frac{\hat{\beta}_0 - \beta_0}{S} \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right)^{-\frac{1}{2}}$$

Skonstruowany przy pomocy tej statystyki przedział ufności dla $\alpha = 0.05$ przdstawia się tak:

$$\left[\hat{\beta}_0 - St_{n-2,1-\frac{\alpha}{2}}\sqrt{\frac{1}{n} + \frac{\overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}; \hat{\beta}_0 + St_{n-2,1-\frac{\alpha}{2}}\sqrt{\frac{1}{n} + \frac{\overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}\right]$$

gdzie $t_{1-\frac{\alpha}{2},n-2}$ to kwantyl z rozkładu t-studenta z n-2 stopniami swobody.

Po podstawieniu wartości przedział przyjmuje poniższą postać:

$$[2.2162, 2.9499]$$

W tym przypadku również estymowany parametr mieści się w zadanym przedziale.

4.7 Predykcja i przedziały ufności dla pewnej części danych

W celu przeprowadzenia predykcji przyjmowanych wartości na podstawie modelu regresji liniowej i pewnej części danych, zgromadzone dane zostały posortowane, następnie odcięto 20 (czyli ok 3% danych) największych wartości, więc prosta regresji liniowej została wyznaczona na podstawie pozostałych 652 obserwacji. Przy ich pomocy zostały również wyznaczone estymatory.

W celu wyznaczenia przedziału ufności dla danych predykowanych należy wprowadzić pewne oznaczenia:

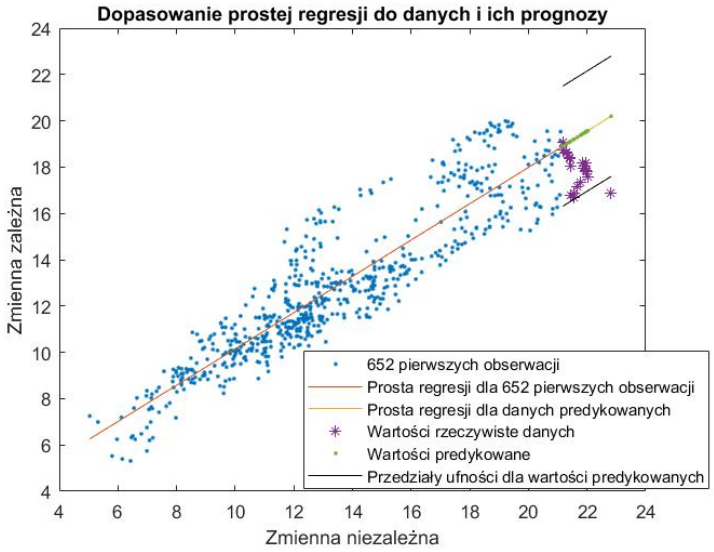
- x_0 - obserwacja odcięta, która będzie predykowana
- $Y(x_0)$ - wartość zmiennej zależnej dla odciętej obserwacji
- $\hat{Y}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$ - wartość estymowanej zmiennej zależnej dla odciętej obserwacji

Dla takiego modelu przedział ufności przyjmuje następującą postać:

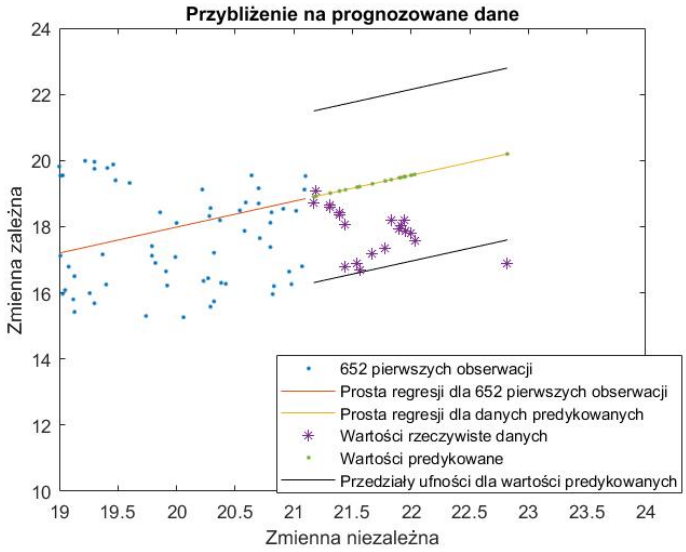
$$\left[\hat{Y}(x_0) - t_{1-\frac{\alpha}{2},n-2}S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}, \hat{Y}(x_0) + t_{1-\frac{\alpha}{2},n-2}S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}\right]$$

gdzie $t_{1-\frac{\alpha}{2},n-2}$ to kwantyl z rozkładu t-studenta z n-2 stopniami swobody.

W tym przypadku nie można podać konkretnego przedziału, ponieważ 20 obserwacji zostało odciętych, a każdą z nich traktuje się jako kolejny x_0 w związku z czym dla każdego z nich zostanie skonstruowany osobny przedział. Najwygodniej więc zobrazować przedziały oraz predykowane wartości na wykresach, które znajdują się poniżej. Przyjęto $\alpha = 0.05$.



Rysunek 16: Wykres rozproszenia dla 652 pierwszych danych z dopasowaną prostą regresji oraz wyznaczonymi wartościami predykowanymi i ich prostą regresji i przedziałami ufności



Rysunek 17: Przybliżenie na predykowane dane z wykresu 16

Na wykresie 16 i 17 przedstawiono graficznie przeprowadzoną predykcję danych. Zielonymi kropkami oznaczono dane predykowane, których wartości były wyliczane na podstawie prostej regresji dopasowanej do pierwszych 652 danych. Czarne linie są natomiast przedziałami ufności dla tych predykowanych danych. Dodatkowo fioletowymi gwiazdkami zostały oznaczone rzeczywiste obserwacje, które wcześniej zostały odcięte. Warto zauważyć, że prawie wszystkie te wartości mieszczą się w przedziałach ufności skonstruowanych dla danych predykowanych. Oznacza to, że model został dobrany prawidłowo i rzeczywiste dane nie odbiegają znacznie od danych predykowanych, które znalazły się na prostej regresji.

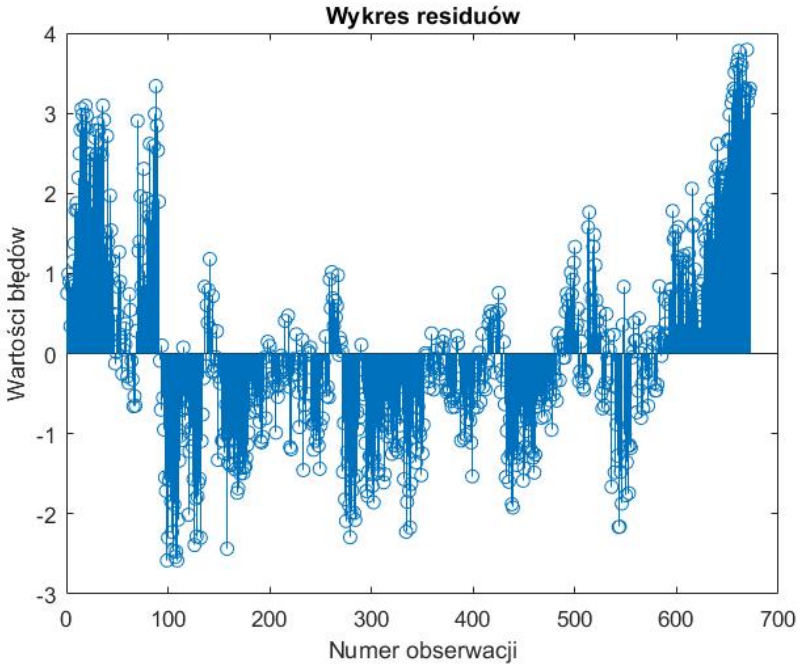
5 Analiza residuów

Residua są to wartości resztowe otrzymane w procesie regresji. Mówią one o tym, o ile estymowana wartość różni się od wartości rzeczywistej zmiennej losowej. Wartość tą można wyrazić przy pomocy poniższego wzoru:

$$e_i = Y_i - \hat{Y}_i = Y_i - (\beta_0 + \beta_1 x_i)$$

Residua są więc błędami popełnianymi podczas przybliżania danych. Aby sprawdzić poprawność zastosowanego modelu regresji liniowej danego wzorem 10 należy zbadać zachowanie residuów. Zgodnie z teorią wartości resztowe oznaczane w modelu jako ϵ_i powinny mieć rozkład normalny. Dodatkowo ich średnia powinna wynosić zero, wariancja być stała, a residua powinny być niezależne od

siebie. Przeprowadzona zostanie analiza błędów, sprawdzająca te warunki, która zweryfikuje poprawność zastosowanego modelu. Początkowo wygenerowano wykres przedstawiający residua.



Rysunek 18: Wykres przedstawiający residua

Powyższy wykres 18 wizualizuje wartości residuów od numeru obswerwacji. Przyjmują one na zmianę wartości dodatnie i ujemne. Można zauważyć, że residua są największe na krańcach badanego okresu i najmniejsze w środku jego trwania. Ich własności zbadano poniżej.

5.1 Analiza średniej

Zgodnie z teorią wartość średnia residuów, powinna wynosić 0. W rozważanym przypadku jest to

$$\overline{e} = 7.4464 \cdot 10^{-15}$$

Wartość średniej jest bardzo mała i bliska 0. Dodatkowo, jeżeli przyjrzymy się wykresowi 18 przedstawiającemu wartości residuów możemy dostrzec, że znajdują się one powyżej i poniżej zera. Można więc przypuszczać, że wartości te się znoszą, a średnia wynosi 0. By to potwierdzić, przeprowadzono t test. Jest to test, który zwraca decyzję testową dla hipotezy zerowej, że dane pochodzą z rozkładu normalnego o średniej równej zero i nieznanej wariancji. Alternatywna hipoteza głosi, że rozkład populacji nie ma średniej równej zero. Wynik h

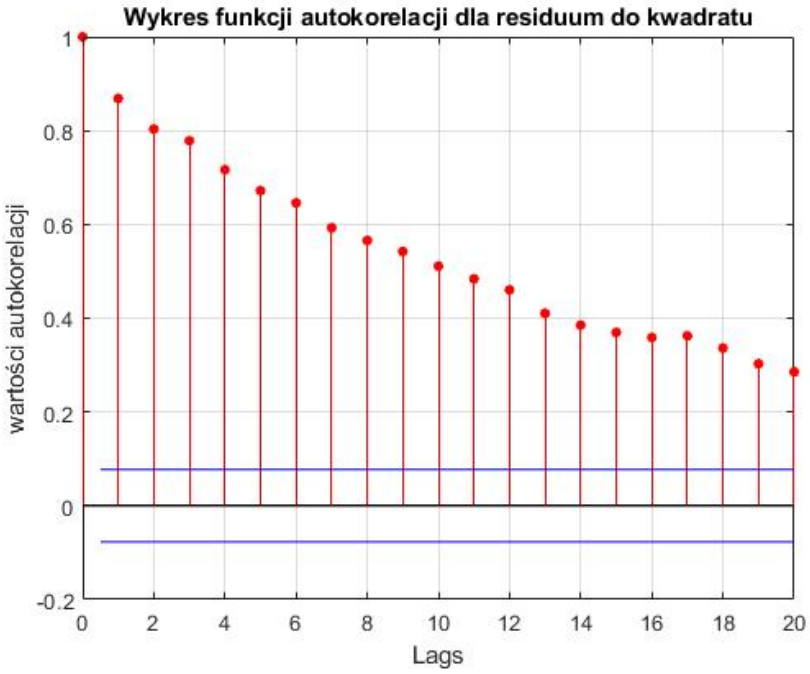
wynosi 1, jeśli test odrzuca hipotezę zerową na poziomie istotności 5%, a 0 w innym przypadku. Wynik:

$$h = 0, p = 1$$

Oznacza to, że hipoteza zerowa została przyjęta na poziomie istotności 0.05,a p-wartość wynosi 1, czyli nie ma podstaw do odrzucenia hipotezy zerowej. Stąd wniosek, że średnia wartości resztowych wynosi 0 i jest zgodna z założeniami.

5.2 Analiza wariancji

Według teoretycznych założeń wartość wariancji residuów powinna być stała dla wszystkich wartości. W celu sprawdzenia tego założenia zostanie narysowany wykres autokorelacji z e^2. Jeśli będzie zachowywał się tak jak zachowują się dane niezależne można będzie stwierdzić, że wariancja jest stała, ponieważ nie waha się znacząco. Oprócz tego sprawdzenie założenia o stałej wariancji będzie wykonane przy pomocy Arch testu, który w hipotezie zerowej zakłada, brak heteroskedastyczności, czyli tego, że przynajmniej jedna wartość będzie różniła się wariancją od pozostałych. Jeśli hipoteza zerowa zostanie przyjęta, będzie można wnioskować, że wariancja wartości resztowych jest stała, w przypadku przyjęcia hipotezy alternatywnej, głoszącej, że w wektorze znajduje się przynajmniej jedna wartość, dla której wariancja odstaje od pozostałych, należy przyjąć brak stałej wariancji.



Rysunek 19: Wykres autokorelacji residuów do kwadratu

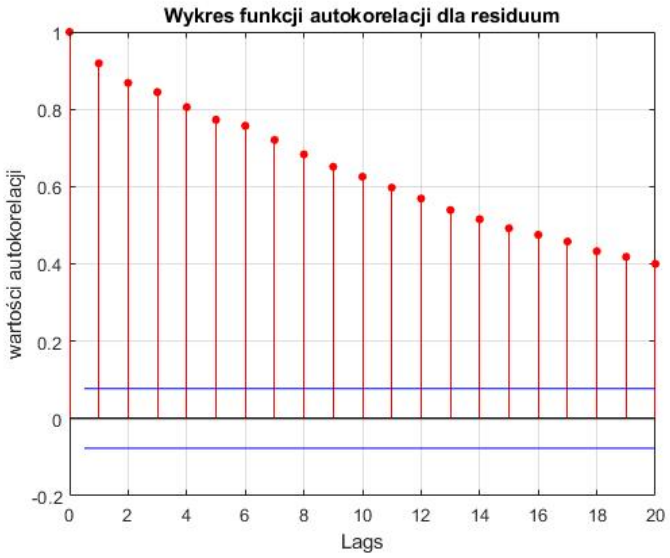
Przyglądając się wykresowi 19, od razu można zauważyć, że różni się on znacząco od wyników, które byłyby otrzymane dla wartości niezależnych. W związku, z czym można przypuszczać że wariancja nie jest stała, ponieważ widoczne na wykresie 19 wartości autokorelacji dla różnych lagów są bardzo duże, ewidentnie nie mieszczą się w zaznaczonych na niebiesko przedziałach ufności. W celu upewnienia się, czy założenie to jest poprawne, warto spojrzeć na wyniki wspomnianego wyżej Arch testu, który już formalnie daje odpowiedź, czy wariancja jest stała, czy też nie. Wynik:

$$h = 1, p = 0$$

Wartość $h = 1$ oznacza, że należy odrzucić hipotezę zerową na rzecz hipotezy alternatywnej. Z kolei $p = 0$ odpowiada wartości p-value, która zgodnie z definicją mówi, że im większe p, tym bardziej można się skłaniać do przyjąć hipotezy zerowej. Dla otrzymanego wyniku wniosek jest prosty, w 100% odrzucamy hipotezę zerową. W związku z czym w tym przypadku wariancja wartości resztowych nie jest stała.

5.3 Analiza niezależności

Otrzymane residua zgodnie z teorią powinny być od siebie niezależne. Pierwszym sprawdzeniem może być spojrzenie na wykres 18 przedstawiający wykres residuów. Dla wartości niezależnych, chmura powinna być równomiernie rozłożona, jednak na wykresie widać, że na początku i na końcu okresu obserwacji residua przyjmują dodatnią wartość, w środku natomiast ujemną. Może to poddawać wątpliwości niezależność wartości resztowych. Jednak przeprowadzając dokładniejszą analizę w sposób nieformalny, można narysować wykres funkcji autokorelacji dla residuów, jeśli wartości będą bliskie 0 dla całej próby, i tylko w zerze otrzymana zostanie wartość 1 można mówić o niezależności. Bardziej formalnym sposobem sprawdzenia będzie wykonanie testu Ljung'a-Box'a, który używany jest do oceny zależności między danymi. Za hipotezę zerową test ten przyjmuje, że korelacja między obserwacjami równa jest 0. Hipoteza alternatywna natomiast mówi, że są one zależne.



Rysunek 20: Wykres autokorelacji residuów

Aby dane rzeczywiste były zgodne z teorią, wartości residuów dla konkretnych lagów powinny oscylować w okolicach zera, pomiędzy niebieskimi liniami oznaczającymi przedziały ufności. Na wykresie 20 widać, że jest zupełnie inaczej. Wartości znacząco przewyższają zaznaczony poziom ufności, dlatego też można wnioskować, że wartości resztowe nie są niezależne. W celu formalnego sprawdzenia należy zobaczyć jaki wynik zwraca lbqtest. Wyniki testu:

$$h = 1, p = 0$$

Otrzymany wynik należy interpretować analogicznie jak w przypadku wariancji. Otrzymane $h = 1$, odpowiada odrzuceniu hipotezy zerowej na rzecz hipotezy alternatywnej. Z kolei $p = 0$ daje całkowitą pewność tego, że należy odrzucić hipotezę zerową. Stąd wniosek, że wartości resztowe z modelu nie są niezależne.

5.4 Analiza rozkładu

Zgodnie z założeniami, residua powinny mieć rozkład normalny. Aby to sprawdzić, przeprowadzono dwa rodzaje testów. Sprawdzenie mniej formalne, na które składało się wyliczenie charakterystyk, gęstości, dystrybuanty i kwantyli oraz sprawdzenie formalne, które zawiera testy statystyczne. Poniżej zaprezentowano wyniki przeprowadzonej analizy

SPRAWDZENIE NIEFORMALNE

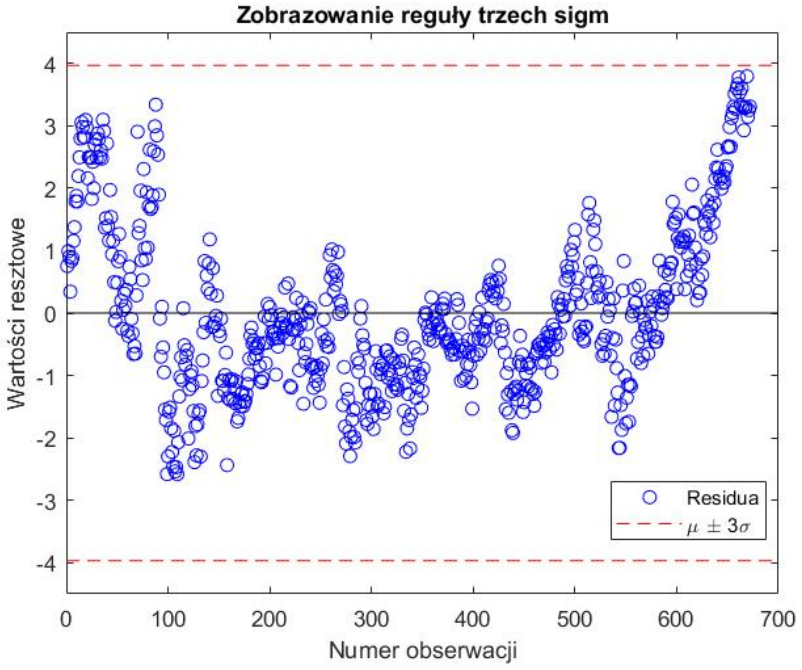
- Statystyki residuów:

Statystyka	Wartość dla residuów	Wartość dla N(0,σ)
Kutrtoza	3.1986	3
Współczynnik skośności	0.1977	0

Tabela 4: Tabela z porównaniem kurtozy i współczynnika skośności dla residuów i teoretycznego rozkładu normalnego N(0,σ)

Powyższa tabela 4 porównuje skośność i spłaszczenie rozkładu residuów z rozkładem normalnym. Można z niej odczytać, że wartość kurtozy 3.1986 dla wartości resztowych jest bliska wartości 3 dla rozkładu normalnego. Podobnie ze współczynnikiem skośności, którego wartość 0.1977 jest bliska 0.

- Reguła trzech sigm mówi, że jeżeli zmienna ma rozkład normalny bądź zbliżony do rozkładu normalnego to $99,7\%$ obserwacji znajduje się w zakresie pomiędzy ± 3 odchylenia standardowe od średniej. Dla residuów wykres obrazujący tę regułę wygląda następująco:

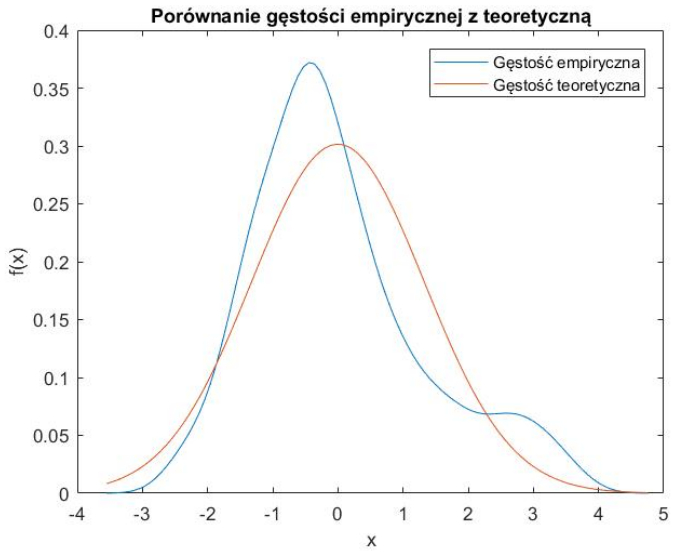


Rysunek 21: Wykres obrazujący regułę 3 sigm dla rozkładu residuów

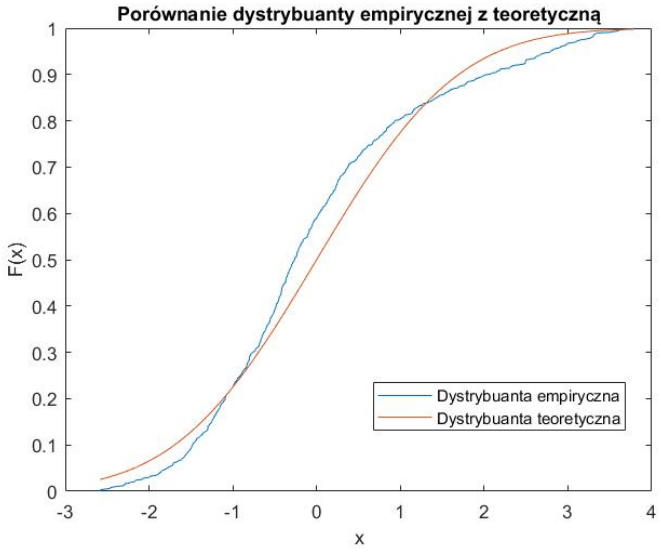
Na wykresie 21 przedstawiono wartości resztowe od numeru obserwacji.

Dodatkowo czerwoną przerywaną linią zaznaczono wartości $\mu \pm 3\sigma$, gdzie μ jest średnią z próby, a σ odchyleniem standardowym z próby. Na wykresie możemy zobaczyć, że żadna z wartości resztowych nie znalazła się poza przedziałem. Stąd wniosek, że reguła trzech sigm jest spełniona, ponieważ $99,7\% \approx 1$, tak jak wyszlo dla danych empirycznych.

- Porównanie gęstości i dystrybuanty empirycznej z teoretyczną



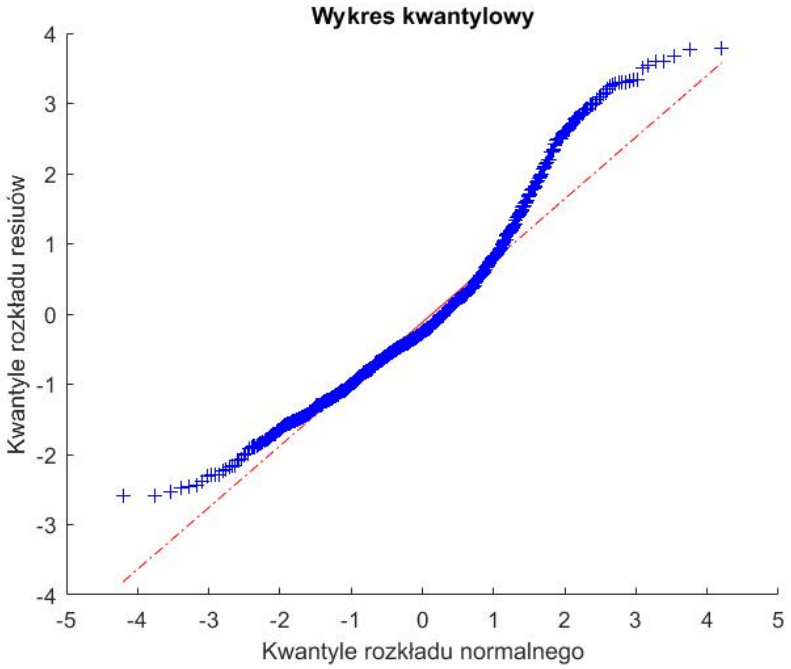
Rysunek 22: Wykres porównujący gęstość empiryczną residuów z dystrybuantą rozkładu normalnego $\mathcal{N}(0, 1.3231)$



Rysunek 23: Wykres porównujący dystrybuantę empiryczną residuów z dystrybuantą rozkładu normalnego $\mathcal{N}(0, 1.3231)$

Na powyższych wykresach 22, 23 porównano gęstości i dystrybuanty rozkładu wartości resztowych z rozkładem normalnym z średnią 0 i wariancją wyliczoną z próby $\mathcal{N}(0, 1.3231)$. Na pierwszym z nich widać, że gęstości się różnią. Ich kształt jest podobny, jednak gęstość empiryczna jest bardziej smukła u góry i bardziej spłaszczona u dołu (zwłaszcza z prawej strony). Funkcje z wykresu 23 mają podobny przebieg, dystrybuanta empiryczna przyjmuje wartości wokół tej teoretycznej, ale raczej nie pokrywają się na żadnym odcinku.

- Wykres kwantylowy



Rysunek 24: Wykres porównujący kwantyle residuów z kwantylami rozkładu normalnego $\mathcal{N}(0, 1.3231)$

Wykres 24 jest wykresem kwantylowym, który porównuje kwantyle rozkładu residuów z kwantylami rozkładu normalnego $\mathcal{N}(0, 1.3231)$. Z wykresu można odczytać, że kwantyle się nie pokrywają. Niebieskie plusy oznaczające kwantyle wartości resztowych mocno odbiegają od prostej na końcach przedziału. Jest to wyraźnie widoczne i nie pozostawia wątpliwości, że rozkład residuów nie może pochodzić z rozkładu normalnego $\mathcal{N}(0, 1.3231)$. Mimo iż statystyki tych rozkładów z tabeli 4 przyjmowały podobne wartości i zachowana była zasada trzech sigm i dystrybuanty oraz gęstości miały podobny kształt, to wykres kwantylowy nie pozostawia wątpliwości, że rozkładem residuów nie jest rozważany rozkład teoretyczny. Poniższe formalne sprawdzenie zależności między tymi rozkładami pozwoli na podjęcie ostatecznej decyzji, czy rozkład residuów jest rozkładem normalnym.

SPRAWDZENIE FORMALNE

- Test Kołmogorowa - Smirnowa to najczęściej stosowany test statystyczny sprawdzający, czy próba pochodzi z rozkładu normalnego. Bazuje on na różnicy pomiędzy dystrybuantą teoretyczną a empiryczną. Hipoteza zerowa zakłada właśnie należenie próby do rozkładu normalnego, natomiast

alternatywna odrzuca to założenie. Zastosowano modyfikację powyższego testu i analizowane dane przyrównano do rozkładu normalnego ze średnią zero i wariancją z danych.

- Test Anderson-Darling jest podobny do K-S testu, ponieważ również bazuje na różnicy pomiędzy dystrybuantą empiryczną a teoretyczną. Jest jednak bardziej wrażliwy na różnice w ogonach. Test ten również za hipotezę zerową przyjmuje, że wektor pochodzi z rozkładu normalnego, a hipoteza alternatywna odrzuca to założenie.

- Test Jarque-Bera bazuje na empirycznych parametrach rozkładu, takich jak skośność i kurtoza. Dzięki niemu można ocenić, czy próba pochodzi z rozkładu normalnego o nieznanej średniej i wariancji. Pochodzenie z takiego rozkładu zakłada hipoteza zerowa, natomiast alternatywna odrzuca to założenie.

Test	h	p
Kołmogorow - Smirnow	1	$9.0392 \cdot 10^{-6}$
Anderson-Darling	1	$5 \cdot 10^{-4}$
Jarque-Bera	1	$1 \cdot 10^{-3}$

Tabela 5: Tabela z wynikami testów normalności

Jeżeli test przyjmuje wartość $h = 1$, świadczy to o odrzuceniu jego hipotezy zerowej na rzecz hipotezy alternatywnej. Wartość p mówi o tym, jak bardzo bliska odrzucenia jest hipoteza zerowa. Im mniejsze p tym większe prawdopodobieństwo odrzucenia hipotezy zerowej. Jak widać w tabeli 5 wszystkie testy odrzuciły hipotezę zerową, w związku z czym należy wnioskować, że dane nie pochodzą z rozkładu normalnego. Analizując wartość p, za każdym razem widać, że jest ono bardzo małe, czyli szansa na odrzucenie hipotezy zerowej jest bardzo duża.

6 Podsumowanie i wnioski

W pracy analizowane były ceny akcji dwóch spółek giełdowych w okresie trzech lat, od 6 grudnia 2017 r. do 4 grudnia 2020 r. Pierwsza z nich to First Quantum Minerals Ltd., która jest kanadyjską firmą zajmującą się poszukiwaniem, zagospodarowaniem i wydobyciem złóż, głównie miedzi. Drugim omawianym przedsiębiorstwem było amerykańskie Freeport McMoRan również działające w branży wydobywczej. W pierwszej części raportu, dane obydwu firm były analizowane osobno. Omawiano ich rozkład, analizując położenie, rozproszenie, asymetrię i skośność. W obydwu przypadkach rozkład dla danych był podobny do rozkładu normalnego, choć występowały pewne różnice. Oznacza to, że dane

były skoncentrowane wokół swoich średnich, ale w badanym okresie pojawiały się pewne zaburzenia, które generowały obserwacje ekstremalne, np. pandemia koronawirusa. Stąd rozkłady danych nie były dokładnie rozkładem normalnym, ale zawierały wiele jego cech. W kolejnej części pracy przyjęto za zmienną niezależną, dane dotyczące firmy First Quantum Minerals Ltd, natomiast za zmienną zależną ceny akcji drugiej spółki. Dla tak dobranych zmiennych przedstawiono wykres rozproszenia danych (wykres 11). Na jego podstawie stwierdzono dodatnią zależność liniową między danymi i dobrano model regresji liniowej. W kolejnym etapie przygotowano dane, pozbywając się obserwacji najbardziej oddalonych od prostej regresji, by model jeszcze lepiej przybliżał dane. Warto dodać, że usunięte obserwacje uznano za odstające, ponieważ znajdowały się najdalej od wyznaczonej prostej. Nie oznacza to jednak, że były one nietypowe dla analizowanych zbiorów. Przybliżenie to pozwoliło na lepsze dopasowanie prostej, ale niekoniecznie miało sens w odniesieniu do ich rzeczywistego przebiegu i sytuacji rynkowej. Wyliczono wskaźniki badające zależność liniową, a ich wyniki pozwalały sądzić, że dobrany model jest prawidłowy. Wyestymowano parametry dla modelu regresji i ich przedziały ufności. Przeprowadzono predykcję dwudziestu obserwacji i zbadano przedziały ufności dla danych predykowanych. Zauważono, że dane rzeczywiste nie różniły się znacznie od tych predykowanych i w znaczej większości wpadły w wyznaczony przedział. Pozwoliło to przypuszczać, że zależność liniowa prawiłowo opisuje związek między zbiorami. Ostatni etap pracy, polegał na analizie residuów. Zgodnie z przyjętym modelem regresji liniowej, powinny one pochodzić z rozkładu normalnego, o średniej zero i stałej wariancji. Dodatkowo powinny one być niezależne od siebie. Zbadano więc te założenia, które okazały się być niespełnione. Wykryto zależność pomiędzy kolejnymi wartościami resztowymi na podstawie wykresu autokorelacji (wykres 21). Przeprowadzono 3 testy statystyczne (Test Kołmogorow - Smirnow, Anderson-Darling, Jarque-Bera), które zgodnie odrzuciły hipotezę, że rozkład residuów jest rozkładem normalnym. Zatem założenia z teoretycznego modelu regresji liniowej nie zostały spełnione. Mogło to na przykład wynikać z faktu, że część informacji ze zmiennej objaśnianej znalazła się w analizowanym błędzie. Oznaczałoby to, że residua przestały być losowe tak, jak to zakładano. To tylko jeden z wielu możliwych powodów, dla których wybrany model się nie sprawdził. Ostatecznie jednak można wywnioskować, że dobrany model regresji liniowej jest niewystarczający. Stąd istnieje realne ryzyko, że przeprowadzone analizy i przedstawione wyniki nie są poprawne dla omawianych danych.