Raport nr 1

Model i Analiza Regresji Liniowej

Piotr Mikler 236895

Spis treści

1 Wstęp **2**

2 Przygotowanie danych **2**

3 Prosta Regresji **3**

4 Poprawność modelu **4**

4.1 Analiza estymatorów b_0 i b_1 4

4.2 Współczynnik determinacji 5

4.3 Analiza Reziduów 6

4.3.1 Rozkład 6

4.3.2 Niezmienna wariancja 7

4.3.3 Niezależność Reziduów 8

5 Prognoza **10**

5.1 Predykcja wartości średniej 10

5.2 Predykcja przyszłej wartości 11

6 Wnioski **13**

1 Wstęp

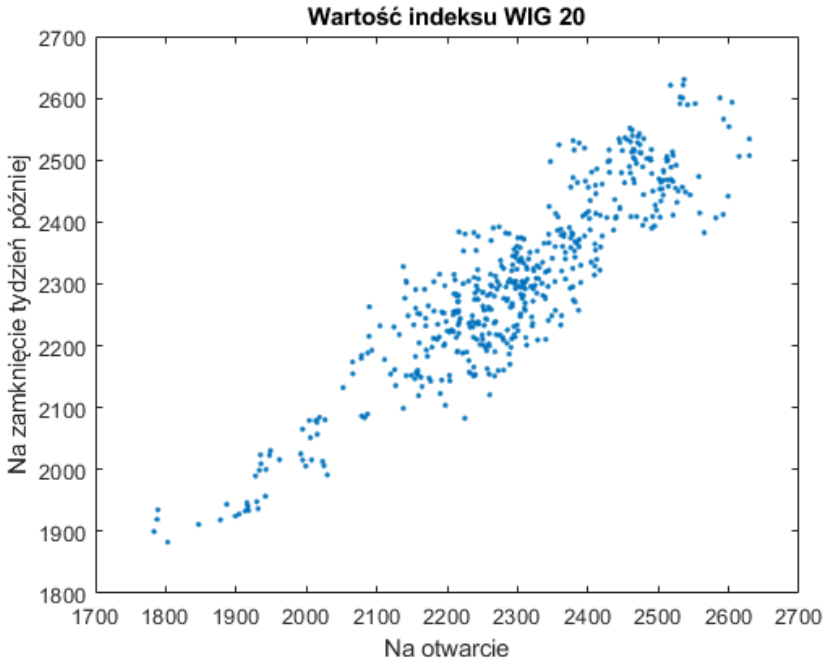
Niniejsza praca traktuje o badaniu istnienia i siły zależności liniowej pomiędzy kursem otwarcia indeksu giełdowego WIG 20, a jego kursem zamknięcia tydzień później w okresie od 2016-11-30 do 2018-11-30.

W kolejnych rozdziałach czytamy o przygotowaniu danych, dopasowaniu do nich prostej metodą najmniejszych kwadratów, a następnie o badaniu poprawności modelu. Na sam koniec, za pomocą modelu spróbujemy dokonać predykcji kursów zamknięcia dla kursów otwarcia z kilku następnych dni.

2 Przygotowanie danych

Dane pobrane zostały ze strony *https://stooq.pl* zawierającej wyczerpującą historyczną bazę wartości indeksów giełdowych. Pobrane dane zawierają wartości indeksu WIG20 (dalej nazywanego indeksem) na otwarcie sesji, zamknięcie sesji, wolumen oraz wartości: maksymalną i minimalną w dziennych interwałach (z pominięciem sobót i niedziel). W sumie, razem z datami jest to 500 obserwacji 6 zmiennych.

Dane zostały załadowane do programu Matlab, tam odcięte zostały niepotrzebne zmienne,



Rysunek 1: Wartość indeksu wartości na otwarcie przeciw wartości na zamknięcie po tygodniu

tak, aby zostały jedynie zmienna objaśniana *Zamknięcie*, oraz zmienna objaśniająca *Otwarcie*. Następnie wektory zostały odpowiednio przesunięte, więc obserwacji zmiennej objaśniającej odpowiada obserwacja zmiennej objaśnianej po tygodniu. Tak wstępnie przygotowane dane można zobaczyć na rysunku 1. Widzimy na pierwszy rzut oka, że próba dopasowania zależności liniowej do danych jest jak najbardziej uzasadniona. Punkty układają się wizualnie w prostą, z pewnymi wahaniami wokół niej. Poszukajmy jej.

3 Prosta Regresji

Prosta Regresji jest metodą poszukiwania zależności liniowej między dwoma zmiennymi. Wierzymy, że za zmiennością danych stoi pewna liniowa zależność $f(x_i) = \beta_1 x_i + \beta_0$, ale jej wartości zostały zniekształcone przez losowy składnik ϵ_i. Otrzymujemy zatem następującą postać modelu regresji liniowej:

$$Y_i = \beta_1 x_i + \beta_0 + \epsilon_i,$$

w której to β_0 i β_1 to rzeczywiste stałe deterministyczne, a dla $i = 1, ..., n$ składniki x_i to deterministyczne wartości zmiennej objaśniającej, a $\epsilon_i - iid \sim \mathcal{N}(0, \sigma^2)$.
Naszym zadaniem jest dopasować prostą regresji $\hat{y}_i = b_1 x_i + b_0$, w taki sposób, aby globalnie przechodziła jak najbliżej zadanych punktów. Zauważmy że przy ustalonych parametrach b_0 i b_1 estymowane przez nas wartości będą różnić się od zadanych o wartość $y_i - \hat{y}_i = e_i$. Taką wartość nazywamy rezyduum. Zatem można myśleć o dopasowywaniu prostej jako o minimalizacji sumy owych reziduów, lub ich kwadratów. To drugie podejście jest nazywane metodą najmniejszych kwadratów i prowadzi do następujących postaci estymatorów:

$$\begin{cases} b_1 = \hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i(x-\overline{x})}{\sum_{i=1}^{n}(x-\overline{x})^2} \approx 0.84 \\ b_0 = \hat{\beta}_0 = \overline{y} - b_1\overline{x} \approx 363.79 \end{cases}$$

Dopasowaną w ten sposób prostą można obejrzeć na rysunku (2). Widzimy tu, że prosta wygląda na dobrze dopasowaną, przecina chmurę punktów praktycznie na pół. Widzimy również, że niektóre punkty są blisko niej, a niektóre nieco dalej. Pojawia się zatem pytanie, czy być może istnieją w tym zbiorze wartości odstające, tzn niepasujące charakterem obserwacje, które psują zależność. Aby je zlokalizować, przeanalizuję wartości otrzymanych reziduów. Za odstające uznawać będziemy te obserwacje, których rezidua odbiegają od innych. Użyję kwantyli rozkładu reziduów, żeby sprawdzić które odstają o więcej niż 1.5 IQR od dolnego i górnego kwantyla.

Rysunek 2: Prosta Regresji MNK

Okazuje się, że taka obserwacja jest tylko jedna. Można wyrzucić ją ze zbioru i przeliczyć na nowo parametry modelu regresji. Wtedy otrzymamy $b_1 \approx 0.84$ oraz $b_0 \approx 363.44$. Widzimy zatem, że ta obserwacja nie jest wpływowa ponieważ praktycznie nie zmienia przebiegu prostej. Nie ma zatem konieczności usuwania jej.

Tak dopasowana prosta jest estymatorem zależności pomiędzy zmienną objaśnianą i objaśniającą. Warto zastanowić się teraz na ile dobre jest nasze dopasowanie.

4 Poprawność modelu

4.1 Analiza estymatorów b_0 i b_1

Jeśli zapiszemy bardzo ogólnie model regresji liniowej

$$Y_i = \beta_1 x_i + \beta_0 + \epsilon_i, \text{ gdzie } \epsilon_i \sim \mathcal{N}(0, \sigma^2),$$

to będziemy mogli rozważyć b_0 i b_1 jako zmienne losowe.

Zapiszmy najpierw zmienną losową $S^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{y}_i)^2}{n-2}$. Nazywamy ją błędem średniokwadratowym i jest ona bardzo użyteczna w modelu regresji. Jej wartość oczekiwana wynosi σ^2, więc jest ona naturalnym estymatorem wariancji, co będziemy wykorzystywać. Oznaczmy dodatkowo

błędy standardowe:

$$\begin{cases} SE_{b_1} = \left(\frac{S^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)^{0.5} \\ SE_{b_0} = S\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)^{0.5} \end{cases}$$

Wtedy można pokazać, że

$$\begin{cases} \frac{b_1-\beta_1}{SE_{b_1}} \sim t_{n-2} \\ \frac{b_0-\beta_0}{SE_{b_0}} \sim t_{n-2} \end{cases}$$

Z tych zależności dostajemy natychmiast przedziały ufności dla β_0 i β_1. Są one postaci:

$$\begin{cases} \beta_1 \in b_1 \pm t_{n-2,0.5\alpha} \cdot SE_{b_1} \\ \beta_0 \in b_0 \pm t_{n-2,0.5\alpha} \cdot SE_{b_0} \end{cases}$$

W przypadku naszych danych na poziomie ufności $\alpha = 0.05$ otrzymujemy, że $\beta_1 \in [0.81; 0.88]$, oraz, że $\beta_0 \in [283.94, 443.64]$. Widzimy, że pierwszy przedział ufności jest bardzo wąski, czego nie można powiedzieć o drugim. Według mnie wynika to z faktu, że widzimy stosunkowo małą zmienność wokół prostej, co pozwala dość dobrze określić β_1, natomiast wartości zmiennej objaśniającej są dalekie od zera, co powoduje, że nawet mała zmiana współczynnika kierunkowego zmienia znacząco wartość funkcji w zerze, a więc nasz estymator parametru β_0- ciężko jest o nim wnioskować bardzo precyzyjnie.

4.2 Współczynnik determinacji

Gdyby nie istniała żadna zależność między zmienną objaśnianą a objaśniającą, oraz przeprowadzilibyśmy superdokładne pomiary, to zmienna objaśniana byłaby zdegenerowana, tzn przyjmowałaby stałą wartość. Tak jednak się nie dzieje- wartości odbiegają od swojej średniej. Tę zmienność można podzielić na dwie części: pierwsza, jest związana z liniową zależnością. Zmieniamy wartość X, więc zmienia się Y. Widzimy jednak, że w naszym modelu mamy do czynienia z czymś więcej. Wartości zmiennej objaśnianej oscylują też wokół samej zależności. Jest to druga część zmienności, związana z losowością składnika ϵ.

Te rozważania podsumowuje krótko zależność:

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

Lewa strona równania jest nazywana Total Sum of Squares. Reprezentuje całkowitą zmienność w modelu. Po prawej stronie równania mamy natomiast kolejno Regression Sum of Squares

(zmienność wynikającą z liniowej zależności) oraz Error Sum of Squares (zmienność rezyduów wokół zera).

To prowadzi nas do definicji współczynnika R^2, który jest miarą dopasowania modelu liniowego do danych. Opisuje on stosunek zmienności wyjaśnianej przez model do zmienności danych. Wartości bliskie jedynce oznaczają zatem, że model wyjaśnia dobrze zmienność danych.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \approx 0.82$$

Wynika z tego, że nasz model odpowiada za 82% zachowania danych. Model liniowy jest zatem dość dobry do opisu naszych danych. Widać to było już wcześniej wizualnie, w postaci punktów układających się w prostą.

Jeżeli usunęlibyśmy wcześniej wspomnianą obserwację odstającą, współczynnik determinacji wzrasta o zaledwie 0.002. Potwierdza się więc, że nie jest ona wpływowa i nie warto jej usuwać.

4.3 Analiza Reziduów

Aby sprawdzić poprawność modelu należy przeanalizować własności powstałych z niego reziduów. Przy początkowych założeniach normalności i niezależności ϵ_i, oczekujemy, że rezidua będą przejawiały te same cechy.

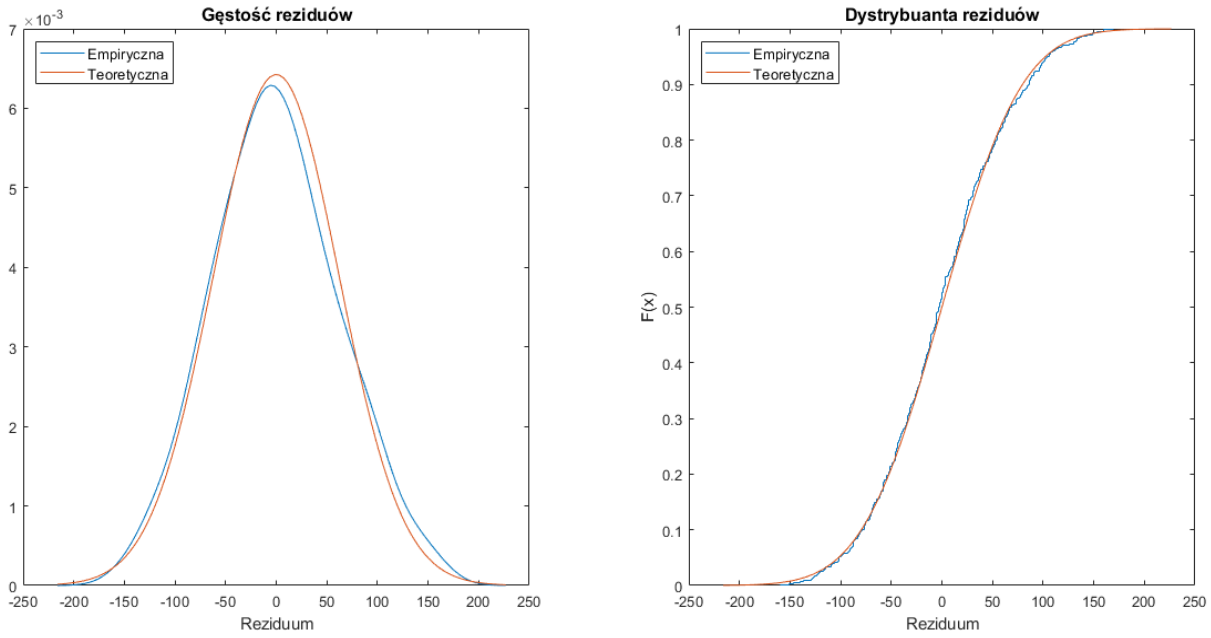
4.3.1 Rozkład

Rezidua powinny mieć rozkład normalny o średniej zero i pewnej nieznanej wariancji σ^2.

Nieznaną wariancję estymujemy jako wcześniej już wspomniany błąd średniokwadratowy, czyli

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

Testowanie normalności zaczniemy od sprawdzenia gęstości, oraz dystrybuanty rozkładu. Sporządziłem wykresy gęstości i dystrybuanty rozkładu reziduów, oraz gęstości i dystrybuanty rozkładu $\mathcal{N}(0, \hat{\sigma^2})$. (rysunek (3)). Widzimy, że nie pokrywają się idealnie, ale patrząc na ilość obserwacji jest całkiem dobrze. Widzimy, tu również, że otrzymujemy średnią w okolicy zera. Na poparcie tezy normalności mamy również akceptację testów: Kołmogorowa-Smirnoffa, z p-wartością na poziomie 0.67, oraz Lillieforse z p-wartością 0.25. Te testy badają czy istnieją powody na odrzucenie hipotezy zerowej jaką jest pochodzenie próbki z rodziny rozkładów normalnych. P-wartości ilustrują na ile "typowa" jest próba dla założeń hipotezy zerowej. Jest



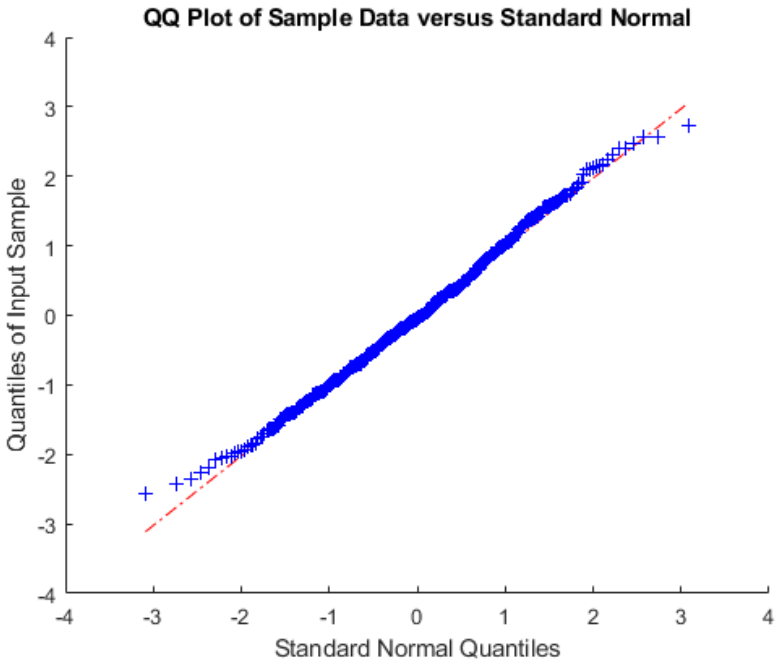
Rysunek 3: Gęstość i dystrybuanta rozkładu reziduów

więc dobrze.

Rysunek (4) przedstawia natomiast wykres kwantylowy unormowanych (podzielonych przez σ) reziduów. Jeśli próba jest z rozkładu normalnego, to kwantyle powinny układać się wzdłuż prostej, przynajmniej w centralnej części wykresu, gdzie jest ich najwięcej. Nasz wykres wygląda bardzo obiecująco. Dopiero ostatnie punkty się rozchodzą, ale mają do tego prawo, bo obserwacji jest mniej w ogonach próby.

4.3.2 Niezmienna wariancja

Rezidua powinny być zmiennymi losowymi o stałej wariancji. Jednak jeśli narysujemy wykres rozproszenia reziduów, możemy mieć co do tego wątpliwości (Rysunek (5) po lewej stronie). Wizualnie wygląda to jakby wariancja rosła w czasie. Jednym ze sposobów na sprawdzenie tego założenia jest szukanie punktu zmiany reżymu. Po prawej stronie rysunku (5) widzimy zakumulowane kwadraty reziduów. Jeżeli miałyby one stałą wariancję, to wykres powinien przypominać linię prostą. Widzimy jednak spore wahania, wskazujące na to, że wariancja wcale nie była stała. Jeśliby połączyć ostatni i pierwszy punkt wykresu linią, to cały znalazłby się pod nią. Wynika stąd, że wariancja rosła wraz ze wzrostem zmiennej objaśniającej. Jednocześnie należy zauważyć, że nie wygląda to na przypadkowe zmiany, można odnieść wrażenie, że wariancja postępuje



Rysunek 4: Wykres kwantylowy unormowanych reziduów

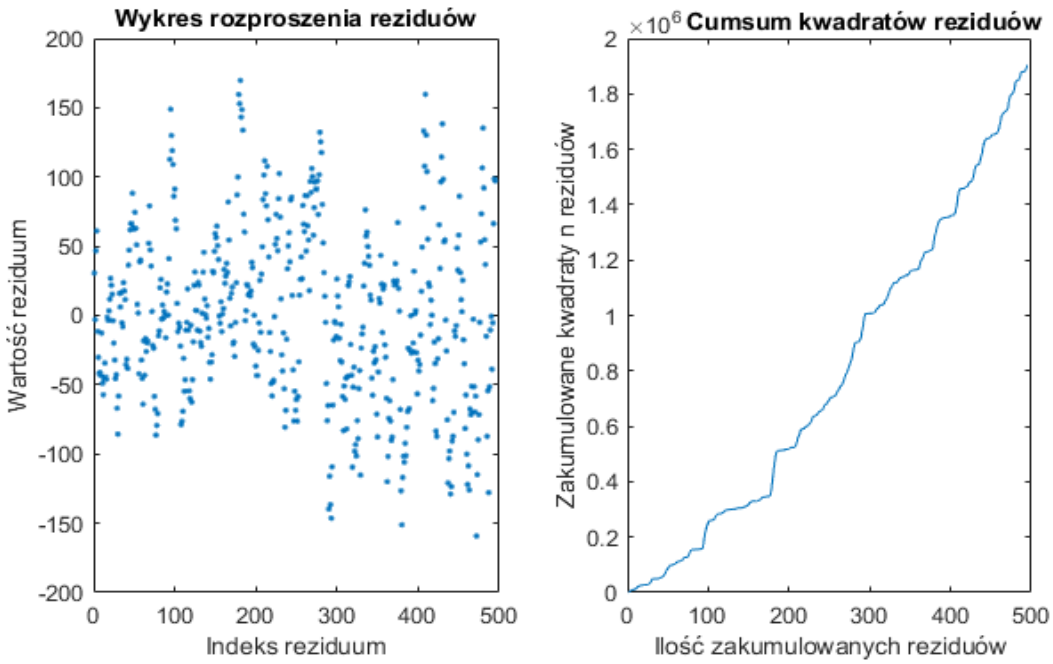
zgodnie z jakąś deterministyczną funkcją, np kwadratową. To z kolei mogłoby wskazywać na to, że rezidua nie są niezależne. W tej sytuacji możnaby wrócić do pierwszego rozdziału i powtórzyć rozważania używając Metody Najmniejszych Ważonych Kwadratów, która być może dałaby lepsze wyniki. Jednak osobiście nigdy tego nie robiłem, więc nie będę przy niej eksperymentował i przymrużę oko na to zachowanie reziduów, tym bardziej że jak na rzeczywiste dane to nie jest najgorzej.

4.3.3 Niezależność Reziduów

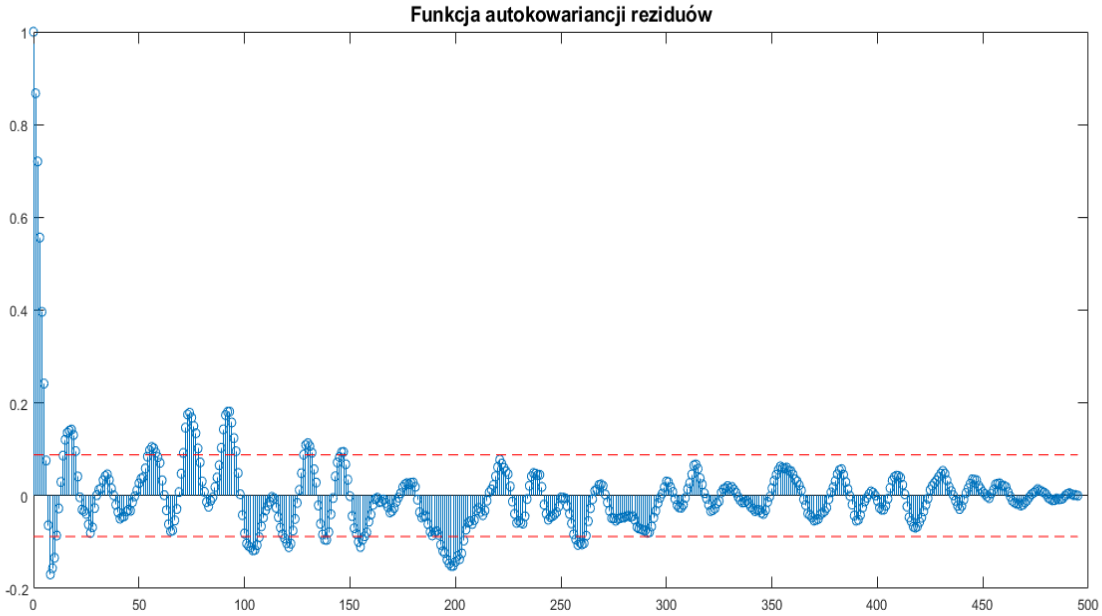
Ostatnim z założeń dotyczącym reziduów jest ich niezależność. Aby ją sprawdzić, użyję funkcji autokowariancji próbkowej. Dla próby losowej $y_1, y_2, y_3, ..., y_n$ taka funkcja dana jest wzorem

$$ACVF(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} (y_{t+|h|} - \overline{y})(y_t - \overline{y})$$

Mierzy ona zależność między wartością próby, a wartością o innym indeksie. Jeśli próba jest prosta, to funkcja powinna zamykać się (na danym poziomie ufność) pomiędzy narysowanymi na czerwono wartościami, oraz powinna wyglądać na losowy szum. Funkcję autokowariancji reziduów można obejrzeć na rysunku (6).



Rysunek 5: Analiza reziduów



Rysunek 6: Autokowariancja próbkowa reziduów

Funkcja autokowariancji zostawia nam wiele do życzenia. Widzimy, że dla małych wartości funkcja osiąga wartości porównywalne z jedynką, więc jest spora zależność reziduów o nieodległych od siebie indeksach. Poza tym, funkcja ewidentnie oscyluje wokół zera, wskazując na

istnienie okresowej, lecz nieregularnej zależności w danych. Gdyby była to regularna okresowa funkcja, możnaby przerobić dane tak, aby tę zależność zminimalizować. W takiej sytuacji, niestety nie jestem w stanie nic z tym zrobić.

5 Prognoza

Mając adekwatny i działający model możemy już powiedzieć coś o przewidywanych wartościach zmiennej objaśnianej, nawet poza znanymi wartościami zmiennej objaśniającej. Można tak zrobić, bo zakładając poprawność modelu regresji, możemy określić przedział ufności dla wartości próby przy danej wielkości zmiennej objaśniającej. Należy tylko pamiętać, że im dalej odsuwamy się od zadanego zbioru punktów, tym mniejszą pewność naszej predykcji dostaniemy. Dla punktów bliskich naszym danym predykcja powinna być jednak całkiem dokładna.
Wyróżniamy dwa rodzaje predykcji. Możemy próbować estymować średnią, lub konkretną wartość. Naturalnie druga opcja jest trudniejsza i mniej pewna, ponieważ aby wnioskować o konkretnej wartości, należy wcześniej wnioskować o średniej.
W moich rozważaniach przyjmowałem poziom ufności $\alpha = 0.05$.

5.1 Predykcja wartości średniej

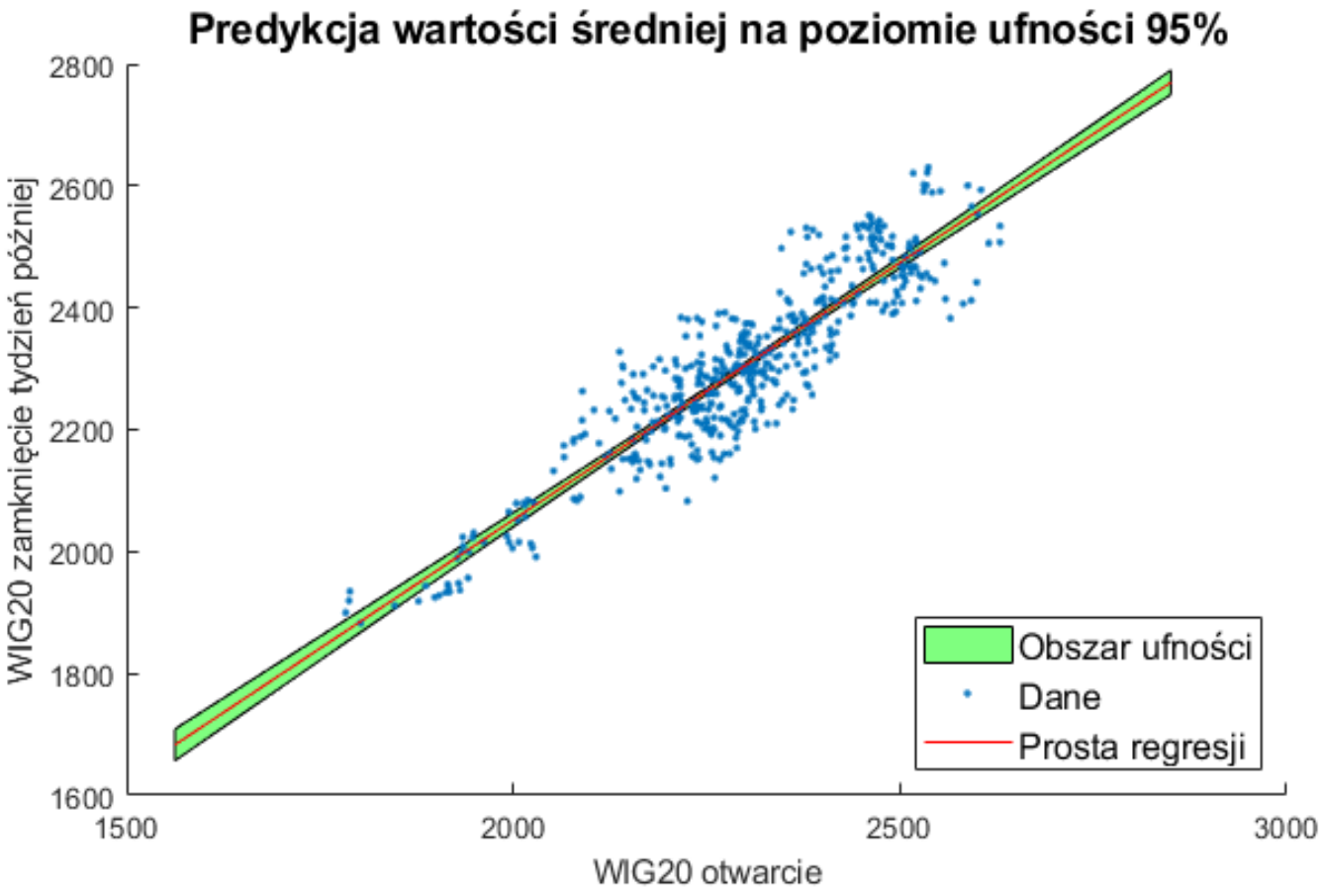
Zaczniemy zatem od predykcji wartości średniej. Aby tego dokonać skorzystamy z faktu, że studentyzowana średnia wartość $\hat{Y}$ w punkcie x_0 ma własność:

$$\frac{\hat{Y}(x_0) - \mu_{\hat{Y}(x_0)}}{SE_{\hat{Y}(x_0)}} \sim t_{n-2}$$

W powyższym wzorze mamy $\mu_{\hat{Y}(x_0)} = E[\hat{Y}(x_0)] = \beta_0 + \beta_1 x_0$, oraz $SE_{\hat{Y}(x_0)} = S\big(\frac{1}{n} + \frac{(x_0-\overline{x})^2}{\sum_{i=1}^{n}(x-\overline{x})^2}\big)^{0.5}$
Z tej własności otrzymujemy przedziały ufności dla wartości średniej w punkcie x_0.

$$\hat{Y}(x_0) \in \mu_{\hat{Y}(x_0)} \pm SE_{\hat{Y}(x_0)} \cdot t_{n-2,0.5\alpha}$$

Te przedziały zmieniają płynnie rozmiar w miarę zmieniania zmiennej objaśniającej. Rysując je wystarczająco gęsto dostajemy cały obszar w którym może znaleźć się średnia wartość. Ilustruje to rysunek (7).Należy zauważyć, że długość przedziału zależna jest od x_0. Zauważmy, że im bliżej jesteśmy średniej $\overline{x}$, tym krótszy jest przedział, co potwierdza wcześniejsze rozważania o tym że efektywnie przewidywać możemy jedynie blisko zadanej chmury punktów.



Rysunek 7: Predykcja wartości średniej modelu

Trzeba jednak przyznać, że obserwując rozstrzał chmury punktów, obszar ufności średniej na wysokim poziomie ufności jest względnie mały. Możemy zatem dość dokładnie estymować wartość średnią.

5.2 Predykcja przyszłej wartości

W poprzednim podrozdziale wyznaczaliśmy przedział ufności dla średniej wartości zmiennej. Można jednak iść o krok dalej i podać przedział ufności dla wartości zmiennej objaśnianej dla konkretnego argumentu x_0. Tutaj niepewność będzie większa, ponieważ ten rodzaj estymacji, to najpierw wyestymowanie średniej, a do tego jeszcze rozrzutu konkretnego punktu. Analitycznie, da się to jednak wyznaczyć, korzystając ze zmiennej losowej $\hat{Y}(x_0) - Y(x_0)$. Zauważmy, że

$$E[\hat{Y}(x_0) - Y(x_0)] = 0$$

$$\sigma^2_{\hat{Y}(x_0)-Y(x_0)} = \sigma^2_{\hat{Y}(x_0)} + \sigma^2 = \sigma^2\left(1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^{n}(x-\overline{x})^2}\right)$$

Podobnie jak wcześniej, będziemy studentyzować naszą zmienną. Do tego potrzebny nam jest znów jej błąd standardowy, który wynika w naturalny sposób z ostatniej równości.

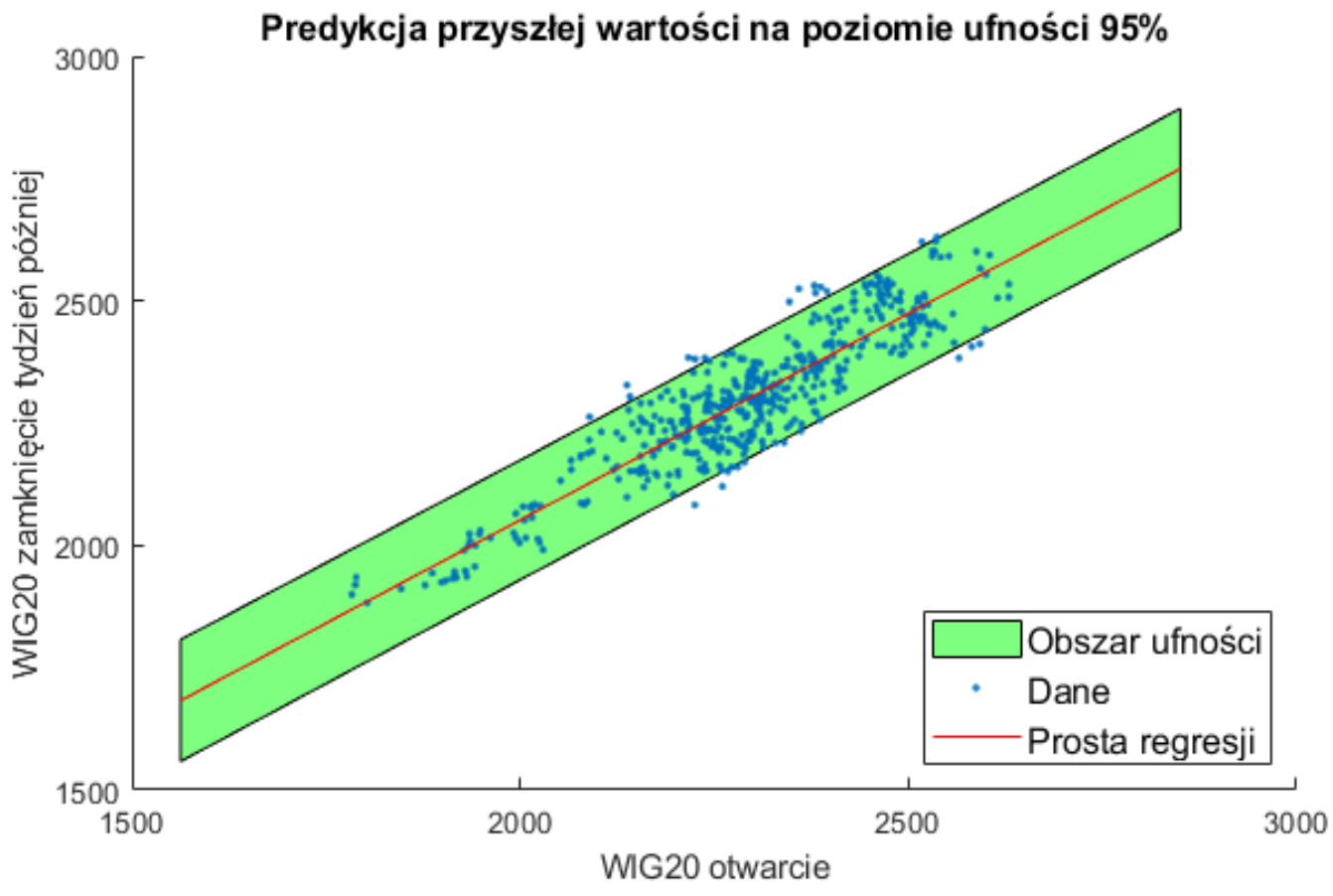
$$SE_{\hat{Y}(x_0)-Y(x_0)} = S\left(1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^{n}(x-\overline{x})^2}\right)^{0.5}$$

Korzystając z tego, otrzymujemy po raz kolejny znany rozkład t-Studenta.

$$\frac{\left(\hat{Y}(x_0) - Y(x_0)\right) - 0}{SE_{\hat{Y}(x_0)-Y(x_0)}} \sim t_{n-2}$$

A stąd otrzymamy wzór na przedział ufności na poziomie α dla $Y(x_0)$, która jest niczym innym jak wartością zmiennej objaśnianej w punkcie x_0. Wygląda on następująco:

$$Y(x_0) \in \hat{Y}(x_0) \pm t_{\alpha/2,n-2} \cdot SE_{\hat{Y}(x_0)-Y(x_0)}$$



Rysunek 8: Predykcja wartości przyszłej

Wyliczając takie przedziały ufności dla kolejnych punktów i łącząc je płynnie ze sobą otrzymujemy obszar ufności (na rysunku (8) zaznaczony na zielono). W tym obszarze znajdzie się

statystycznie 95% obserwacji. Dla naszych danych się to zgadza, wewnątrz jest 94.75% punktów. Pasmo to, podobnie jak na poprzedzającym rysunku rozszerza się tym bardziej im dalej od chmury odejdzemy. Jednakże przy tej skali rysunku nie jest to widoczne.
Używając tej techniki sprawdzimy czy nowe obserwacje będą zgadzać się z naszymi przewidywaniami. Dla okresu od 2018-12-03 do 2018-12-06 sporządziłem tabelę wartości indeksu WIG20 na otwarcie, obszarów ufności oraz wartości na zamknięcie. Na zielono zaznaczyłem przedziały które zgadzają się z realną wartością, a na czerwono te, które nie zawierają realizacji.

Data	Otwarcie	Predykcja średniej	Predykcja wartości	Zamknięcie po tygodniu
03.12.18	2310.75	[2308.77; 2319.78]	[2192.33; 2436.23]	2242.10
04.12.18	2322.24	[2318.42; 2329.53]	[2202.02; 2445.93]	2221.37
05.12.18	2319.32	[2315.97; 2327.05]	[2199.56; 2443.46]	2275.15
06.12.18	2335.28	[2329.33; 2340.64]	[2213.02; 2456.94]	2310.66

Widzimy że obserwacje wpadają do swoich przedziałów predykcji. Nie mieszczą się w przedziałach predykcji dla średniej, ale tego nie można od nich wymagać. W te przedziały powinna wpaść średnia wartość indeksu przy zadanym otwarciu, nie wartość sama w sobie.

6 Wnioski

Badając zależność liniową między cenami WIG20 na otwarcie i na zamknięcie po tygodniu metodą regresji liniowej doszliśmy do wniosku, że modelowanie zależnością liniową jest uzasadnione, trafne, lecz nie idealne.
Dopasowana prosta regresji o równaniu $\hat{y} = 0.84x + 363.79$ wyjaśnia około 82% zmienności w danych ($R^2 \approx 0.82$). Jest to dobrze rokująca, wysoka wartość. Dane nie mają w sobie potencjalnie niebezpiecznych dla modelu odstających obserwacji.
Rezidua modelu pasują do rozkładu normalnego, potwierdzają to metody graficzne (gęstość i dystrybuanta), oraz testy statystyczne (K-S i Lillieforse'a). Ich wariancja rośnie stopniowo wraz ze zwiększaniem się wartości zmiennej objaśniającej, oraz dodatkowo nie są one niezależne. Powoduje to możliwe niedokładności w próbach predykcji. Ukryta zależność między obserwacjami nie powinna nas jednak dziwić, ponieważ są to dane giełdowe, więc wartości indeksu zależą od decyzji uczestników giełdy, które zależą od wartości indeksu. Stąd pojawia się pewna nieregularna zależność między wartościami, co odzwierciedla się w zależności reziduów modelu.

Predykcja jednak okazała się trafna, wszystkie wartości znalazły się w swoich przedziałach ufności. Model można zatem efektywnie wykorzystywać do prób analizy zachowania indeksu WIG20. Widzimy na przykład, że mamy $\hat{y}(x) > x$ dla $x < 2330.9$. Dla inwestora oznacza to, że krótkoterminowe (tygodniowe) inwestycje w indeks będą opłacalne wtedy, gdy wartość indeksu będzie mniejsza od 2330.9.

Dalsza analiza danych potencjalnie mogłaby prowadzić do ciekawych wniosków dotyczących zachowania indeksu. Cieszy jednak fakt, że tak prosty model jak liniowa regresja jest w stanie dobrze opisać dane i wskazać nam interesujące obserwacje.