Komputerowa analiza szeregów czasowych - raport I

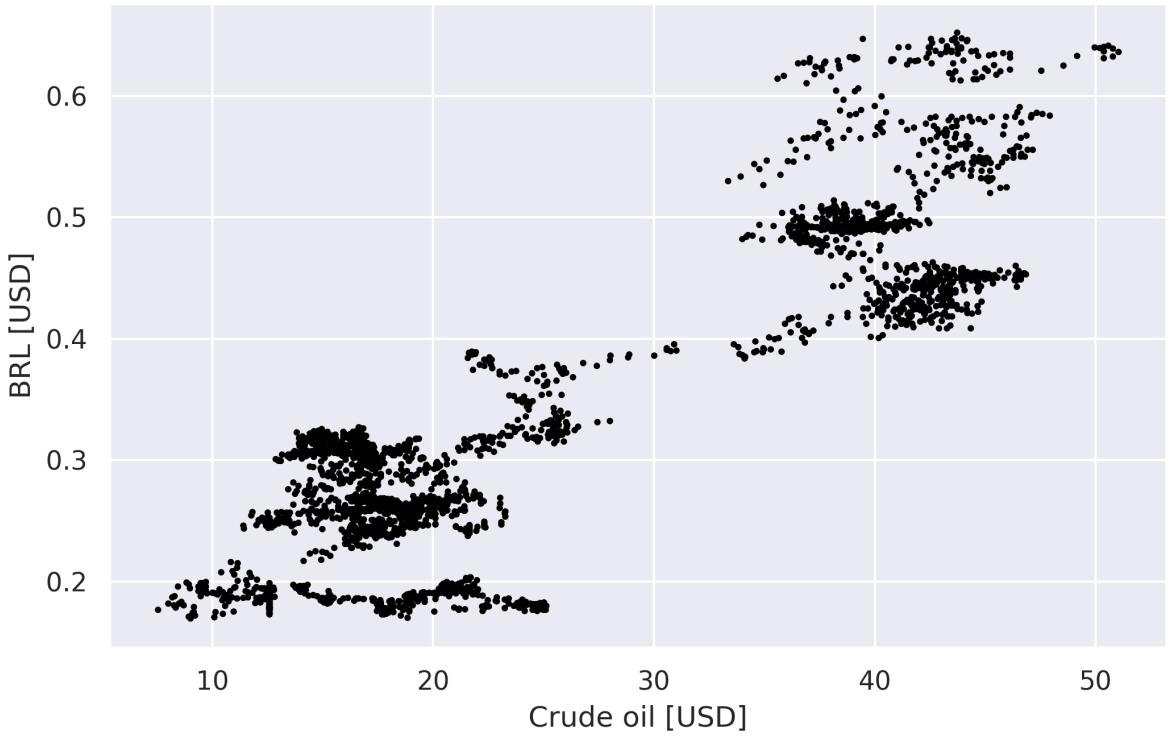
Analiza danych z wykorzystaniem regresji liniowej

Ada Majchrzak, Aleksander Jakóbczyk

22 grudnia 2021

1 Wstęp

Opracowywane przez nas dane pochodzą ze strony Yahoo Finance: iPath Pure Beta Crude Oil ETN, BRL/USD - pierwszy zbiór (zmienna objaśniająca) to ceny ropy naftowej, natomiast drugi, czyli zmienna objaśniana, dotyczy cen reala brazylijskiego (oba w USD). Analizowane obserwacje prowadzone były dziennie w okresie 20.04.2011r. - 10.12.2021r., z tym, że przy wstępnym przygotowaniu danych natrafiliśmy na pewne wartości brakujące oraz NaN, które usunęliśmy. Ostatecznie mamy więc po 2675 punktów w obu zbiorach danych. Na rysunku 1 prezentowany jest wykres rozrzutu:



Rysunek 1: Wykres danych rozproszonych

2 Statystyki opisowe

Będziemy oznaczać: X - zmienna objaśniająca, Y - zmienna objaśniana, A - dowolny zbiór danych (na potrzeby wypisania wzorów, aby nie powielać oznaczeń) oraz kolejno x_i, y_i, a_i - i-ta obserwacja ze zbioru. Ponadto n - długość próby.

2.1 Miary położenia

2.1.1 Średnia arytmetyczna

$$\overline{a} = \frac{1}{n}\sum_{i=1}^{n} a_i$$

$$\overline{x} \approx 25.57$$

$$\overline{y} \approx 0.34$$

2.1.2 Średnia harmoniczna

$$\overline{a}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{a_i}}$$

$$\overline{x}_h \approx 20.92$$

$$\overline{y}_h \approx 0.30$$

2.1.3 Średnia geometryczna

$$\overline{a}_g = \sqrt[n]{\prod_{i=1}^{n} a_i}$$

$$\overline{x}_g \approx 23.06$$

$$\overline{y}_g \approx 0.32$$

2.1.4 Średnia ucinana

$$\overline{a}_t = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} a_i$$

Tutaj, chcąc wyznaczyć parametr k (liczba odrzuconych największych i najmniejszych obserwacji), skorzystamy z pierwszego i trzeciego kwartyla oraz rozstępu międzykwartylowego residuów - wzory na te miary podamy jednak w kolejnej części sprawozdania, powiemy też więcej o residuach. Teraz tylko wyznaczamy potrzebne statystyki i sprawdzamy, ile obserwacji wypada z przedziału $[Q1 - 1.5IQR,\ Q3 + 1.5IQR]$, gdzie:

$Q1$ - pierwszy kwartyl,

$Q3$ - trzeci kwartyl,

IQR - rozstęp międzykwartylowy.

W wyniku obliczeń otrzymujemy 11 obserwacji odstających. Nie sprawdzamy jednak, jak one się rozkładają (czy większość z nich jest powyżej, czy poniżej "normy"). Sortujemy próbę rosnąco i ucinamy po 11 elementów z każdej strony - przy tak dużym zbiorze danych nie powinno to znacząco wpłynąć na wynik. Liczymy zatem średnią ucinaną dla $k = 11$:

$$\overline{x}_t \approx 25.54$$

$$\overline{y}_t \approx 0.34$$

2.1.5 Mediana

Dla posortowanej próby:

$$a_{med} = \mathrm{Q2}_a = \begin{cases} a_{\frac{n+1}{2}}, & \text{gdy n nieparzyste} \\ \frac{1}{2}\left(a_{\frac{n}{2}} + a_{\frac{n}{2}+1}\right), & \text{gdy n parzyste} \end{cases}$$

$$x_{med} \approx 19.99$$

$$y_{med} \approx 0.31$$

2.1.6 Kwartyle

Pierwszy kwartyl Q1 to mediana grupy obserwacji mniejszych od Q2.
Trzeci kwartyl Q3 to mediana grupy obserwacji większych od Q2.

$$\mathrm{Q1}_x \approx 16.40$$

$$\mathrm{Q3}_x \approx 38.91$$

$$\mathrm{Q1}_y \approx 0.25$$

$$\mathrm{Q3}_y \approx 0.45$$

2.2 Miary rozproszenia

2.2.1 Rozstęp międzykwartylowy

$$\mathrm{IQR}_a = \mathrm{Q3}_a - \mathrm{Q1}_a$$

$$\mathrm{IQR}_x \approx 22.52$$

$$\mathrm{IQR}_y \approx 0.19$$

2.2.2 Rozstęp

Dla posortowanego zbioru danych:

$$R_a = a_n - a_1$$

$$R_x \approx 43.52$$

$$R_y \approx 0.48$$

2.2.3 Wariancja

$$S_a^2 = \frac{1}{n}\sum_{i=1}^{n}\left(a_i - \overline{a}\right)^2$$

$$S_x^2 \approx 138.05$$

$$S_y^2 \approx 0.02$$

2.2.4 Odchylenie standardowe

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

$$S_x \approx 11.75$$

$$S_y \approx 0.13$$

2.2.5 Współczynnik zmienności

$$V_a = \frac{S_a}{\overline{a}}$$

$$V_x \approx 0.46$$
$$V_y \approx 0.37$$

2.3 Miary asymetrii

2.3.1 Trzeci moment centralny

$$M_{3_a} = \frac{1}{n}\sum_{i=1}^{n}(a_i - \overline{a})^3$$

$$M_{3_x} \approx 901.10$$
$$M_{3_y} \approx 0.001$$

2.3.2 Współczynnik skośności (asymetrii)

$$\widetilde{M_{3_a}} = \frac{M_{3_a}}{S_a^3}$$

$$\widetilde{M_{3_x}} \approx 0.56$$
$$\widetilde{M_{3_y}} \approx 0.63$$

W obu przypadkach dodatni współczynnik skośności świadczy o prawostronnej asymetrii badanych danych.

2.4 Miary spłaszczenia

2.4.1 Czwarty moment centralny

$$M_4 = \frac{1}{n}\sum_{i=1}^{n}(a_i - \overline{a})^4$$

$$M_{4_x} \approx 31751.81$$
$$M_{4_y} \approx 0.0006$$

2.4.2 Kurtoza

$$K_a = \frac{M_{4_a}}{S_a^4}$$

$$K_x \approx 1.67$$
$$K_y \approx 2.37$$

W obu przypadkach kurtoza mniejsza od 3 świadczy o mniejszym skupieniu wartości wokół średniej w porównaniu z rozkładem $\mathcal{N}(0,1)$.

2.5 Miary zależności

2.5.1 Kowariancja

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n}$$
$$cov(X,Y) \approx 1.30$$

2.5.2 Współczynnik korelacji liniowej Pearsona

Współczynnik korelacji Pearsona określa miarę zależności liniowej między zmiennymi losowymi. Wartość współczynnika korelacji mieści się w przedziale [-1, 1]. Im większa jest jego wartość bezwzględna, tym silniejsza jest zależność liniowa między zmiennymi.

$$r_{xy} \approx \frac{cov(X,Y)}{S_x S_y}$$
$$r_{xy} \approx 0.88$$

Możemy zatem stwierdzić silną liniową zależność panującą między danymi.

3 Prosta regresji i współczynnik determinacji

3.1 Parametry modelu

Z Rysunku 1 widzimy dodatnią zależność między danymi, przypominającą liniową - możemy więc zastosować model regresji liniowej. Aby wyznaczyć współczynniki prostej $y = \beta_1 x + \beta_0$, korzystamy z następujących wzorów:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$
$$\hat{\beta}_0 = \overline{y} - \beta_1 \overline{x},$$

gdzie:

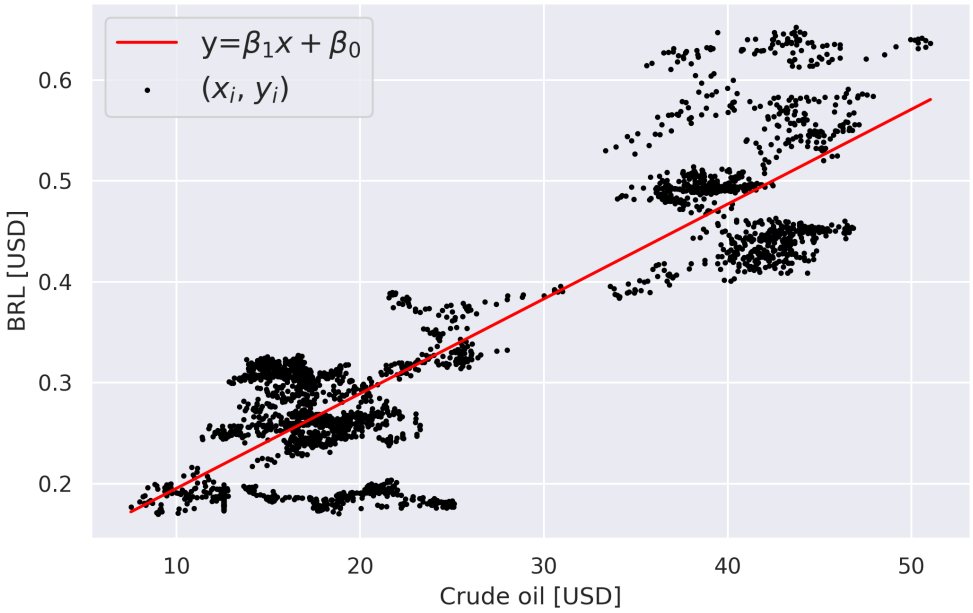
n - ilość obserwacji,

x - ceny ropy,

y - ceny reala brazylijskiego.

Wykonując obliczenia w Pythonie otrzymujemy:

$$y \approx 0.0094x + 0.1015$$



Rysunek 2: Dane rozproszone i prosta modelu regresji liniowej

Rysunek 2 wyraźnie pokazuje, że nasz model dobrze opisuje zależność między zbiorami danych.

3.2 Współczynnik determinacji

Mając już wyznaczone współczynniki prostej regresji, możemy przejść do obliczenia współczynnika determinacji. W tym celu wyznaczamy najpierw wartość teoretyczną zmiennej objaśnianej:

$$\widehat{y}_i = \hat{\beta}_1 x_i + \hat{\beta}_0,$$

a następnie korzystamy ze wzoru:

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(yi - \overline{y})^2},$$

Wykonując potrzebne obliczenia otrzymujemy

$$R^2 \approx 0.78,$$

co potwierdza poprawność dobranego modelu.

3.3 Przedziały ufności parametrów modelu

W modelu regresji liniowej estymatory naszych parametrów β_1 i β_0 mają rozkład:

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right), \quad \hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{(\overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)\right).$$

Na podstawie powyższych estymatorów skonstruujmy przedziały ufności dla dwustronnej hipotezy, na poziomie istotności α i nieznanej σ:

$$\mathbb{P}\left(\hat{\beta}_1 - \mathrm{t}_{1-\frac{\alpha}{2},n-2} S\sqrt{\frac{1}{n} + \frac{(\overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}} \leqslant \beta_1 \leqslant \hat{\beta}_1 + \mathrm{t}_{1-\frac{\alpha}{2},n-2} S\sqrt{\frac{1}{n} + \frac{(\overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}\right) = 1 - \alpha,$$

$$\mathbb{P}\left(\hat{\beta}_0 - \mathrm{t}_{1-\frac{\alpha}{2},n-2} \frac{S}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} \leqslant \beta_0 \leqslant \hat{\beta}_0 - \mathrm{t}_{1-\frac{\alpha}{2},n-2} \frac{S}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}\right) = 1 - \alpha.$$

gdzie:

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

$\mathrm{t}_{1-\frac{\alpha}{2},n-2}$ - kwantyl rzędu $1 - \frac{\alpha}{2}$ z rozkładu t-studenta z $n-2$ stopniami swobody.

Skonstruujmy zatem przedziały ufności na poziomie istotności $\alpha = 0.05$, dla parametru β_1:

$$p_{\beta_1} \approx [0.0092,\ 0.0096]$$

oraz dla parametru β_0:

$$p_{\beta_0} \approx [0.0960,\ 0.1068].$$

3.4 Analiza residuów

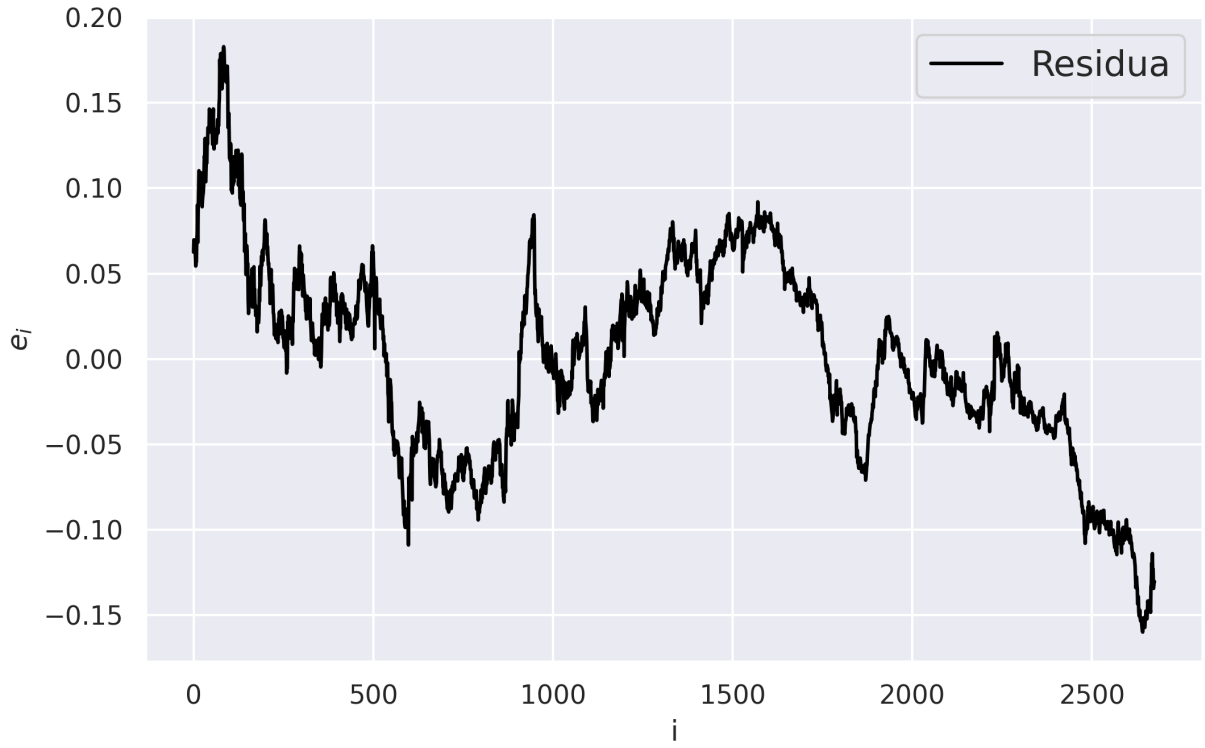
Zajmijmy się zatem analizą residuów. W modelu regresji oczekiwane residua powinny pochodzić z rozkładu normalnego, a same ich wartości wyrażamy za pomocą wzoru:

$$e_i = y_i - \hat{y}_i.$$

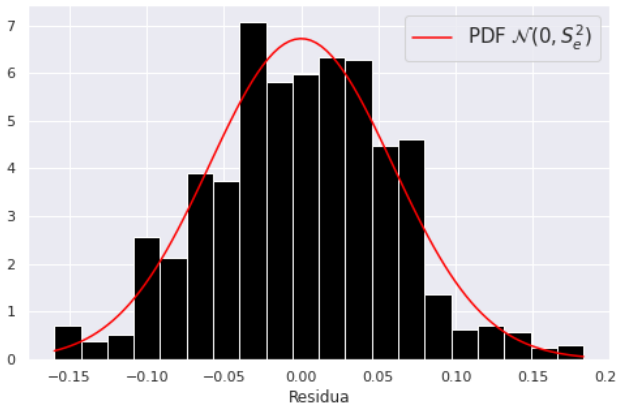
Obliczmy kilka podstawowych statystyk:

$\overline{e}$	S_e	$\mathrm{Q1}_e$	$\mathrm{Q3}_e$	IQR_e	e_{med}	R_e	$\widetilde{M_e}$	K_e
9.29e−18	0.0593	0.0370	0.0400	0.0771	0.0020	0.3428	0.0051	3.0110

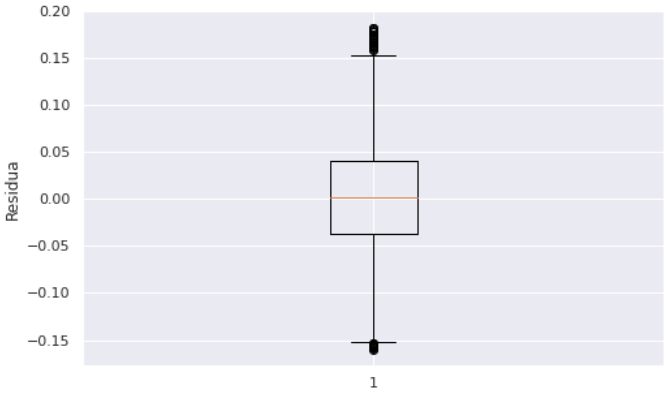
Tabela 1: Podstawce statystyk residuów



Rysunek 3: Wykres wartości residuów



Rysunek 4: Histogram prawdopodobieństwa residuów



Rysunek 5: Box-plot residuów

Na podstawie Rysunku 4, 5 i Tabeli 1 widzimy że rozkład residuów upodabnia się do rozkładu $\mathcal{N}(0, S_e^2)$.

3.5 Testy Normalności

3.5.1 Test Kołmogorowa-Smirnowa

Przeprowadzając test Kołmogorowa-Smirnowa na poziomie istotności $\alpha = 0.05$, dla naszych residuów otrzymujemy statystykę testową $D_n \approx 0.44$, a p-wartość $= 0.0284$. Jednak sam KS-test nie jest uważany za najlepszy możliwy sposób do testowania normalności, skorzystamy zatem z innej opcji.

3.5.2 Test Jarque-Bera

W przypadku JB-testu nasza statystyka testowa wynosi w przybliżeniu $JB \approx 0.0252$, a p-wartość ≈ 0.9875.

Na podstawie powyższego testu możemy stwierdzić, że otrzymane residua pochodzą z rozkładu normalnego.

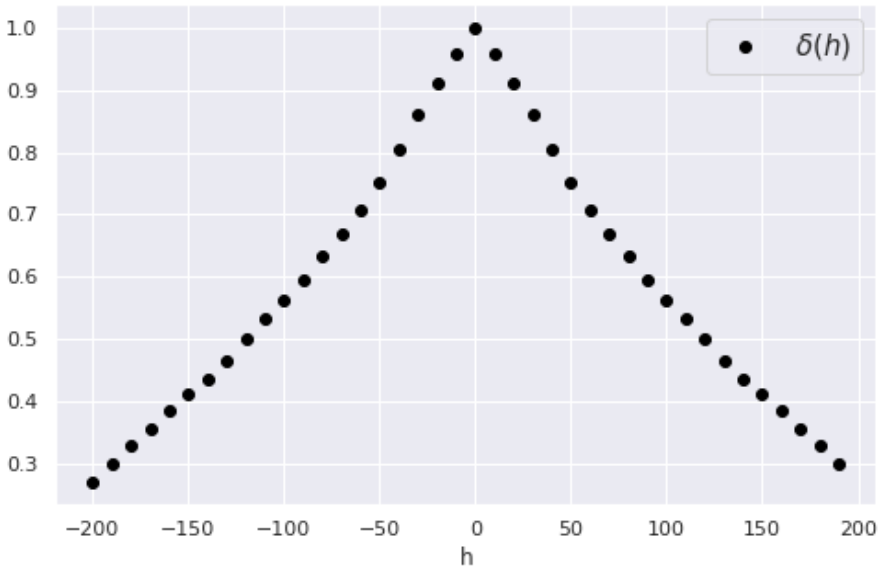
3.6 Korelacja

Analizę korelacji przeprowadzimy na podstawie wykresu funkcji autokorelacji:

$$\delta(h) = \frac{\gamma(h)}{\gamma(0)}$$

gdzie $\gamma(h)$ jest estymatorem funkcji autokowariancji danym wzorem:

$$\gamma(h) = \frac{1}{n} \sum_{i=1}^{n-|h|} (e_i - \overline{e})(e_{i+|h|} - \overline{e})$$



Rysunek 6: Funkcja Autokorelacji

Na podstawie Rysunku 6 możemy stwierdzić, że istnieje pewna korelacja między badanymi residuami, zatem nie są one nieskorelowane. Wynik ten nie jest zaskakujący - z Rysunku 1 widzimy, że nasze dane mają tendencję do grupowania się przy niektórych wartościach.

3.7 Wartości odstające

Jako wartości odstające będziemy zaliczać wszystkie residua niezawierające się w przedziale:

$$[\mathrm{Q1}_e - 1.5\mathrm{IQR}_e,\ \mathrm{Q3}_e + 1.5\mathrm{IQR}_e] \approx [-0.153,\ 0.156],$$

W naszym przypadku po przefiltrowaniu odrzucamy 34 obserwacji odstających.

3.8 Predykcje

Spróbujmy wyznaczyć prognozowane przedziały ufności dla naszego modelu regresji liniowej. Z modelu regresji wiemy, że dla zadanego poziomu istotności α przyjmują postać:

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

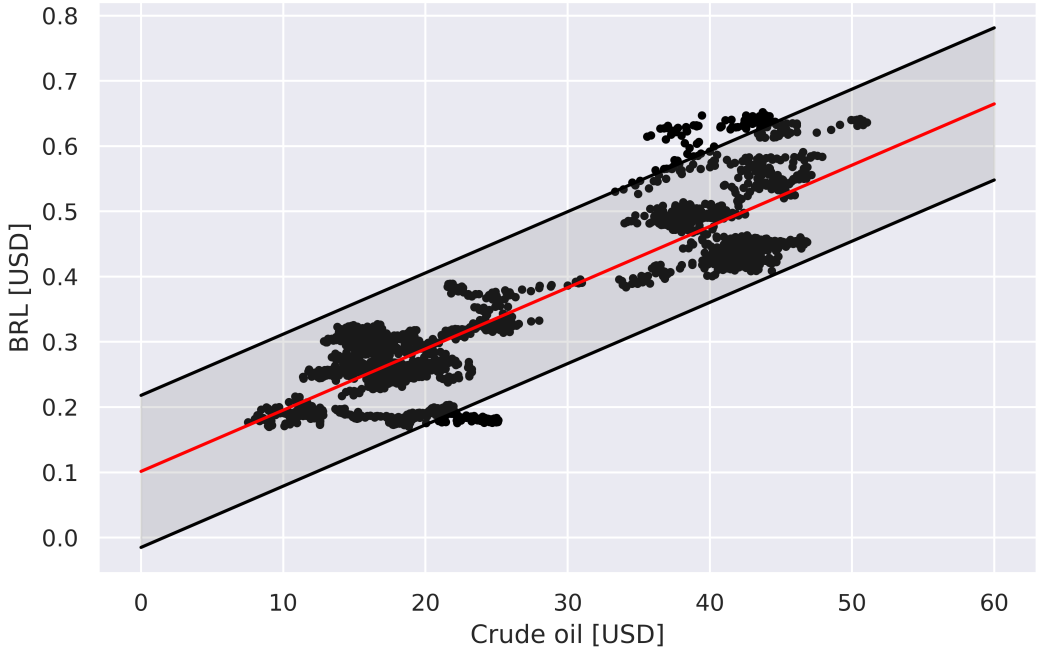
gdzie:

$\hat{y}(x_0) = \hat{B}_0 + \hat{B}_1 x_0$,

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

$\mathrm{t}_{1-\frac{\alpha}{2},n-2}$ - kwantyl rzędu $1-\frac{\alpha}{2}$ z rozkładu t-studenta z $n-2$ stopniami swobody.

Na tej podstawie skonstruujmy nasze prognozowane przedziały ufności:



Rysunek 7: Model predykcji liniowej

Na Rysunku 7 czerwoną linią zaznaczona jest prosta regresji liniowej, natomiast czarne linie to odpowiednio górna i dolna granica naszych przedziałów predykcji.

4 Wnioski

Po przeprowadzeniu analizy widzimy, że model regresji liniowej dobrze dopasowuje się do naszych danych. Świadczy o tym nie tylko wysoka wartość współczynnika determinacji, ale także możemy łatwo odczytać to z wykresu. Ponadto widzimy, że średnia cena ropy naftowej w badanym okresie to około 25.5 dolara, natomiast średnia cena reala brazylijskiego wynosiła 0.34 dolara. Mediana z kolei dla obu zbiorów wyszła nieznacznie niższa niż średnia, co nie dziwi nas ze względu na prawostronną skośność danych. Otrzymany współczynnik zmienności mówi nam o stosunkowo dużym rozproszeniu danych wokół średniej, z kolei wyliczone miary zależności pokazują dodatnią liniową zależność między nimi. Poza tym okazuje się, że tak jak oczekiwaliśmy, residua pochodzą z rozkładu normalnego, co zweryfikowaliśmy przy pomocy przeprowadzonych testów normalności, sprawdzając liczbę obserwacji odstających oraz generując odpowiednie wykresy.