Politechnika Wrocławska

Wydział Matematyki

Analiza danych z wykorzystaniem regresji liniowej

Marcelina Kosiorowska, Maria Lubneuskaya

Spis treści

1. Wstęp

W niniejszym raporcie poddamy analizie dane przy użyciu regresji liniowej, tak aby odkryć zależności między nimi. Przed rozpoczęciem głównej części kluczowe jest zrozumienie podstawowych założeń tego modelu oraz poznanie niezbędnych definicji i wzorów.
Główne założenia modelu:
1. Istnieje liniowa zależność między zmienną niezależną a zmienną zależną.
2. Homoskedastyczność, czyli założenie, że błędy modelu mają stałą wariancję wzdłuż całego zakresu wartości przewidywanych.
3. Błędy modelu są nieskorelowane i mają rozkład normalny.

1.1. Definicje

* **Regresja liniowa** - statystyczna metoda analizy, której celem jest zrozumienie i modelowanie zależności pomiędzy zmiennymi. Zakłada ona, że zależność pomiędzy zmiennymi jest liniowa.
* **Zmienna objaśniania** - w kontekście regresji liniowej jest to zmienna, której wartości estymujemy poprzez model. Będziemy ją oznaczać jako y.
* **Zmienna objaśniająca** - w kontekście regresji liniowej jest to zmienna na której podstawie oblicza się zmienną objaśnianą. Będziemy ją oznaczać jako x.
* **Współczynnik determinacji** - ozn. R^2, jest to miara, która pozwala zbadać jakość dopasowania modelu do przybliżania danych. Im bliżej jest on wartości 1, tym lepiej dobrany jest model.
* **Wielkości deterministyczne** - to wielkości, które są w pełni przewidywalne i pozbawione elementu losowego.
* **Biały szum** - sekwencja niezależnych i identycznie rozkładających się (i.i.d.) zmiennych losowych o zerowej średniej i stałej wariancji (σ^2).

1.2. Wzory

1. Statystyki opisowe:

   — Średnia arytmetyczna (empiryczna wartość oczekiwana):

   $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{1}$$

   gdzie x_i to wartości danych, a n to liczba obserwacji.

   — Estymator wariancji:

   $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2}$$

   — Odchylenie standarowe:

   $$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3}$$

   — Współczynnik skośności (miara asymetrii):

   $$\alpha = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3 \tag{4}$$

   Dla: $\alpha > 0$ mamy asymetrię prawostronną (dodatnią), $\alpha < 0$ mamy asymetrię lewostronną, a dla $\alpha = 0$ mamy symetrię rozkładu.

— Kurtoza (miara spłaszczenia):

$$K = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2\right)^2} \tag{5}$$

Gdy $K > 0$ to oznacza, że w danych jest więcej skrajnych wartości odstających niż w przypadku rozkładu normalnego. Gdy $K < 0$ to w danych jest ich mniej niż w rozkładzie normalnym.

2. Wzory dotyczące modelu regresji liniowej:

— Współczynniki b_0 oraz b_1, używane do dobrania prostej regresji dla danych postaci $y = b_0 + b_1 x$:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{6}$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x} \tag{7}$$

— Równanie klasycznego modelu regresji liniowej:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{8}$$

gdzie x_i to wartości danych (wielkości deterministyczne), β_0, β_1 to parametry modelu (wielkości deterministyczne), a $\{\epsilon_i\}_{i=1}^{n}$ to biały szum.

— Współczynnik korelacji Pearsona:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \tag{9}$$

Im wartość ρ jest bliższa 1 tym zależność danych jest silniejsza i dodatnia (jeżeli x rośnie to y rośnie),
Im wartość ρ jest bliższa -1 tym zależność danych jest silniejsza i ujemna (jeżeli x rośnie to y maleje),
Gdy wartość ρ jest równa 0 to oznacza brak związku liniowego pomiędzy zmiennymi.

— Współczynnik determinacji:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{10}$$

Im wartość R^2 jest bliższa 1, tym lepsze dopasowanie prostej regresji do danych.

— Residua:

$$e_i = \hat{y}_i - y_i, \tag{11}$$

gdzie $\hat{y}_i$ - estymowana wartość y_i.

3. Przedziały ufności:

— Przedział ufności dla β_1 przy nieznanej wartości σ na poziomie ufności $1 - \alpha$:

$$\left[\hat{\beta}_1 - t_{n-2,1-\frac{\alpha}{2}}\frac{S}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \hat{\beta}_1 + t_{n-2,1-\frac{\alpha}{2}}\frac{S}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right], \tag{12}$$

gdzie $S = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y})^2$, a $t_{n-2,1-\frac{\alpha}{2}}$ oznacza kwartyl rzędu $1 - \frac{\alpha}{2}$ z rozkładu t-Studenta z $n-2$ stopniami swobody.

— Przedział ufności dla β_0 przy nieznanej wartości σ na poziomie ufności $1-\alpha$:

$$\left[\hat{\beta}_0 - t_{n-2,1-\frac{\alpha}{2}} S\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}, \hat{\beta}_0 + t_{n-2,1-\frac{\alpha}{2}} S\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}\right]. \quad (13)$$

— Przedział ufności dla prognozy przy nieznanej wartości σ na poziomie ufności $1-\alpha$:

$$\left[\hat{Y}(x_0) \pm t_{n-2,1-\alpha/2}\sqrt{\frac{s^2}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}\left(\hat{Y}(x_0) \pm t_{n-2,1-\alpha/2}\sqrt{\frac{s^2}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}\right)\right]. \quad (14)$$

gdzie $\hat{Y}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$ jest zmienną losową.

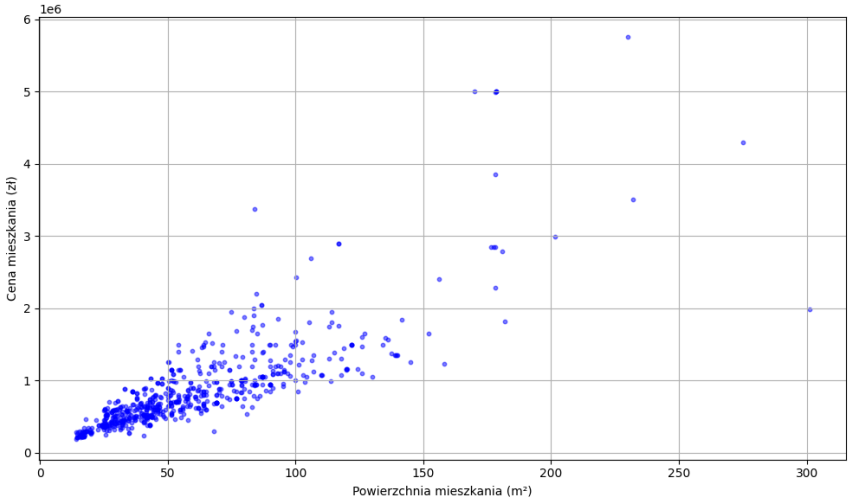
2. Opis danych

2.1. Tematyka danych

* W niniejszym raporcie przeprowadzimy analizę cen mieszkań w krakowskiej dzielnicy Stare Miasto w zależności od powierzchni lokalu.
* Użyte dane pochodzą z [2].
* Dane pochodzą z lutego 2021 roku i zostały zebrane z 674 obserwacji.
* Zmienną objaśnianą (y) będziemy definiować jako cenę mieszkania, natomiast zmienną objaśniającą (x) jako powierzchnię mieszkania.

2.2. Graficzny opis - wykres rozproszenia

Dane na wykresie wyglądają następująco:



Rysunek 1: Wykres zależności danych

Na Rysunku 1 możemy zaobserować dodatnią zależność liniową pomiędzy zmiennymi.

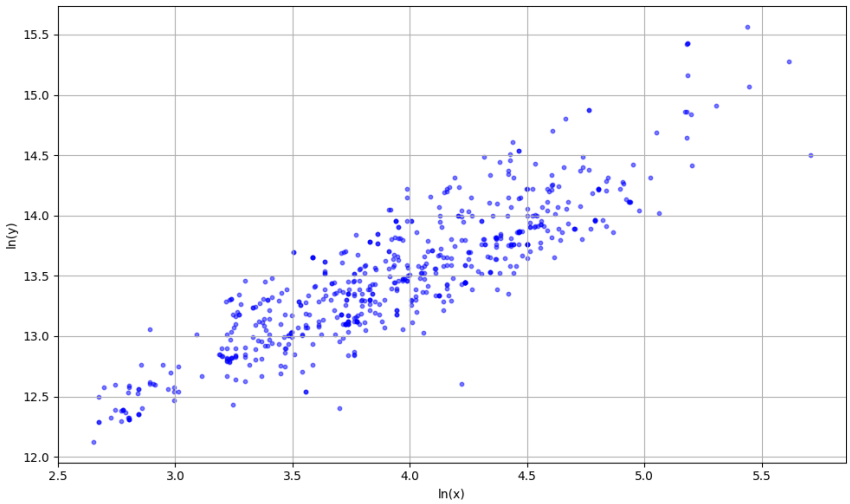
Dodatkowo korzystając ze wzoru (9) wyznaczamy współczynnik korelacji Pearsona: $\rho \approx 0.815$, który również implikuje silną dodatnią liniową zależność (wartość bliska 1).

Na wykresie zauważamy jednak, że wariancja cen mieszkań będzie rosnąć wraz ze wzrostem powierzchni, a więc nie będzie spełnione jedno z założeń modelu regresji liniowej (homoskedastyczność). W naszym przypadku heteroskedastyczność wynika najprawdopodobniej z dużej różnicy miedzy najmniejszą ceną a

największą. Nawet jeśli stosunek odchylenia standardowego cen do ich średniej byłby stały, to absolutne wartości odchylenia standardowego rosłyby ze wzrostem średniej cen.

Aby zniwelować heteroskedastyczność, zlogarytmujemy obie zmienne (logarytmowanie może pomóc w stabilizacji wariancji, skoro wariancja rośnie wraz ze wzrostem średniej cen, zastosowanie logarytmu może spłaszczyć rozkład danych, równoważąc wariancję w różnych obszarach).

Wówczas wykres naszych danych prezentuje się nastepująco:



Rysunek 2: Wykres zależności zlogarytmowanych danych

W dalszej części będziemy analizować oba przypadki - dla zmiennych zlogarytmowanych i niezlogarytmowanych.

3. Statystyki opisowe

3.1. Tabele z wartościami statystyk

Dla badanych zmiennych, tj. cena mieszkań (zm. objaśniana) i powierzchnia mieszkania (zm. objaśniająca), przeprowadziliśmy obliczenia kluczowych statystyk.

Wyniki przedstawiają poniższe tabele (są zaokrąglone do 3 miejsc po przecinku):

Nazwa statystyki	Powierzchnia lokum
Średnia	59.624
Wariancja	1368.173
Odchylenie standardowe	36.989
Pierwszy kwartyl (Q1)	34
Trzeci kwartyl (Q3)	77.548
Mediana	50.19
Wartość maksymalna	301
Wartość minimalna	14.03
Skośność	1.927
Kurtoza	6.105

Tabela 1: Statystyki opisowe dla powierzchni

Nazwa statystyki	Cena mieszkań
Średnia	$8.624 \cdot 10^5$
Wariancja	$4.226 \cdot 10^11$
Odchylenie standardowe	$6.501 \cdot 10^5$
Pierwszy kwartyl (Q1)	$4.955 \cdot 10^5$
Trzeci kwartyl (Q3)	$1.029 \cdot 10^6$
Mediana	$6.939 \cdot 10^5$
Wartość maksymalna	$5.75 \cdot 10^6$
Wartość minimalna	$1.845 \cdot 10^5$
Skośność	3.54
Kurtoza	17.98

Tabela 2: Statystyki opisowe dla ceny

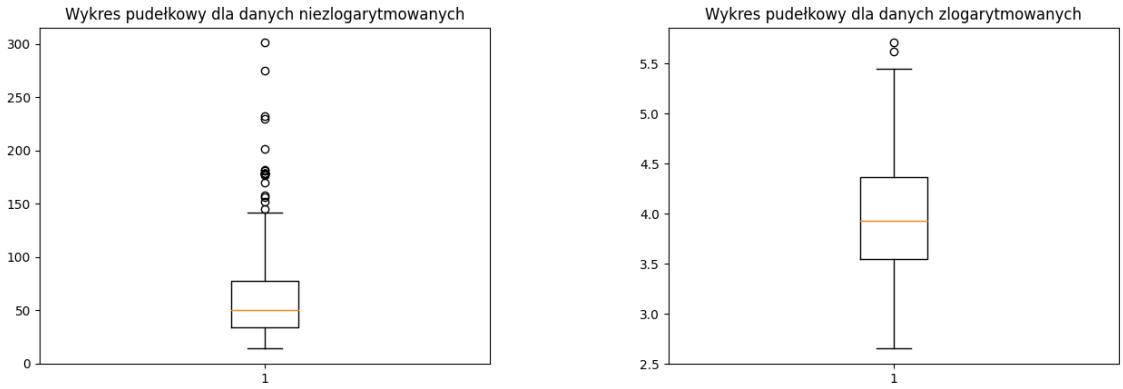
3.2. Analiza statystyk opisowych

Teraz przejdziemy do analizy najważniejszych (według rezultatu) statystyk dla obu zmiennych. Porównamy je ze statystykami dla danych zlogarytmowanych.

3.2.1. Dla powierzchni mieszkania

Widzimy, że średnia powierzchnia mieszkania wynosi około 59.6 m^2 (co ciekawe jest to wartość podobna do średniej powierzchni mieszkania w budynkach wielorodzinnych – 52,1 m^2 [3]).

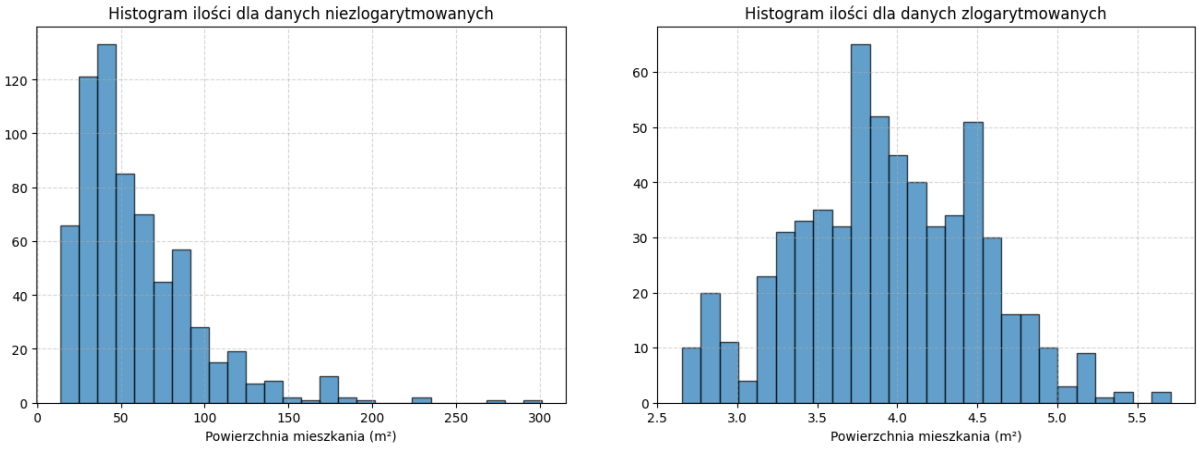
Zauważamy również, że wartości powierzchni mieszkań cechują się dużą rozpiętością, co pokazuje różnica między wartością maksymalną i minimalną oraz potwierdza poniższy wykres po lewej:



Rysunek 3: Porównanie wykresów pudełkowych dla danych powierzchni

Analizując powyższy Rysunek 6 zauważamy, że dla dla powierzchni niezlogarytmowanej istnieje znacznie więcej wartości odstających (wartość mediany jest mniejsza niż wartość średniej) niż dla danych zlogarytmowanych.

Teraz przeanalizujemy histogramy ilościowe:



Rysunek 4: Porównanie histogramów ilości dla danych powierzchni

Skośność w przypadku niezlogarytmowanym wynosi $\alpha \approx 1.927$, co świadczy o prawostronnej asymetrii (wydłużone są prawe ramiona rozkładu) i potwierdza to powyższy histogram po lewej. Natomiast sko-

śność w przypadku zlogarytmowanym wynosi $\alpha^{'} \approx 0.035$, co oznacza że takie dane są mniej asymetryczne.

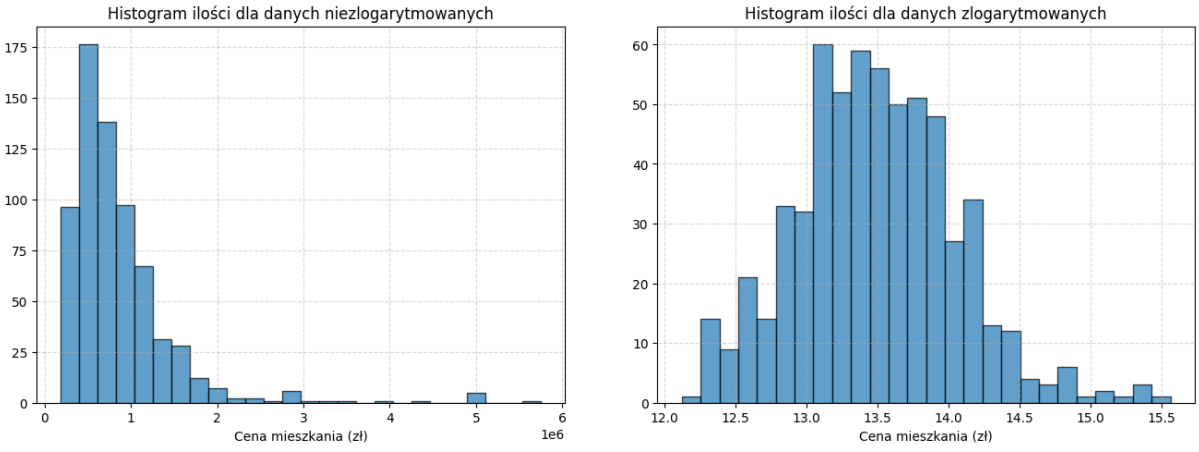
Kurtoza w niezlogarytmowanym przypadku wynosi $K \approx 6.105$ (rozkład ciężkoogonowy), natomiast w zlogarytmowanym $K^{'} \approx -0.259$ (rozkład lekkoogonowy). Oznacza to, że w 1 przypadku w danych jest więcej skrajnych wartości odstających niż w przypadku rozkładu normalnego, a w 2 przypadku mniej.

3.2.2. Dla ceny mieszkania

W przypadku cen, dane również cechują się dużą rozpiętością, co pokazuje różnica między wartością maksymalną i minimalną oraz potwierdza poniższy wykres po lewej :

Rysunek 5: Porównanie wykresów pudełkowych dla danych cen

Dla cen także istnieją wartości odstające, ponieważ wartość mediany jest mniejsza niż wartość średniej. Ponownie zlogarytmowany przypadek cechuje się mniejszą ilością wartości odstających.



Rysunek 6: Porównanie histogramów ilości dla danych cen

Tutaj skośność wynosi odpowiednio $\alpha \approx 3.54$ oraz $\alpha^{'} \approx 0.32$, co świadczy o prawostronnej asymetrii w obu przypadkach i potwierdzają to powyższe histogramy.

Ponadto kurtoza w obu przypadkach jest dodatnia, zatem w jest więcej skrajnych wartości odstających niż w przypadku rozkładu normalnego (rozkłady są ciężkoogonowe).

Jednak w przypadku zlogarytmowanym kurtoza wynosi $K^{'} \approx 0.5$ co jest wartością znacznie niższą niż kurtoza w 1 przypadku $K \approx 17.98$, zatem wartości odstających jest znacznie mniej.

Powyższe informacje podsumowuje fakt, że dane zlogarytmowane posiadają mniej wartości odstających od danych niezlogarytmowanych zatem ich dobór pozwoli nam na lepsze dopasowanie prostej regresji liniowej.

4. Prosta regresji dla danych

Naszym celem jest dopasowanie prostej regresji do danych (wiemy, że wariancja nie jest stała, jednak aby zastosować model regresji liniowej założymy, że jest ona stała).

4.1. Podział danych

Z racji iż w 7 podpunkcie będziemy prognozować przyszłe wartości cen, to podzielimy zbiór danych na dwie części w następujący sposób:

* Dane treningowe: pierwsze 1982 danych (ok. 90% zbioru) - zbiór danych, do którego będzie dobierana prosta regresji liniowej.
* Dane testowe: pozostałe 67 danych (ok. 10 % zbioru) - zbiór danych, które posluzą nam do prognozy

Ponieważ dane nie są posortowane ani ze względu na cenę, ani ze względu na powierzchnię, to możemy przyjąć powyższy podział, a nie wybierać dane w sposób losowy.

4.2. Zastosowanie modelu regresji liniowej

Przyjmujemy nastepujący model (wzór (8)) :

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, 2, ..., n.$$

gdzie jako x_i przyjmujemy powierzchnię mieszkania, ϵ_i są niezależnymi zmiennymi losowymi z zerową średnią i skończoną wariancją, a $n = 607$ to liczba danych.

Dane dotyczące ceny mieszkania oznaczymy jako y_i i będziemy je traktować, jako realizacje zmiennych losowych Y_i.

Aby dopasować prostą regresji, skorzystamy z metody najmniejszych kwadratów. Owa metoda polega na minimalizacji sumy kwadratów różnic pomiędzy wartościami obserwowanymi (y_i) a wartościami przewidywanymi przez model.

$$S(\beta_1, \beta_0) = \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2$$

szukamy takich współczynników β_0, β_1 które zminimalizują powyższą sumę. W tym celu rozwiązujemy równania:

$$\frac{dS}{d\beta_0} = 0$$
$$\frac{dS}{d\beta_1} = 0$$

Rozwiązaniem powyższych równań jest para estymatorów (ze wzorów (6).):

$$\begin{cases} \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x} \end{cases}$$

Po podstawieniu danych otrzymujemy nastepujące wartości:

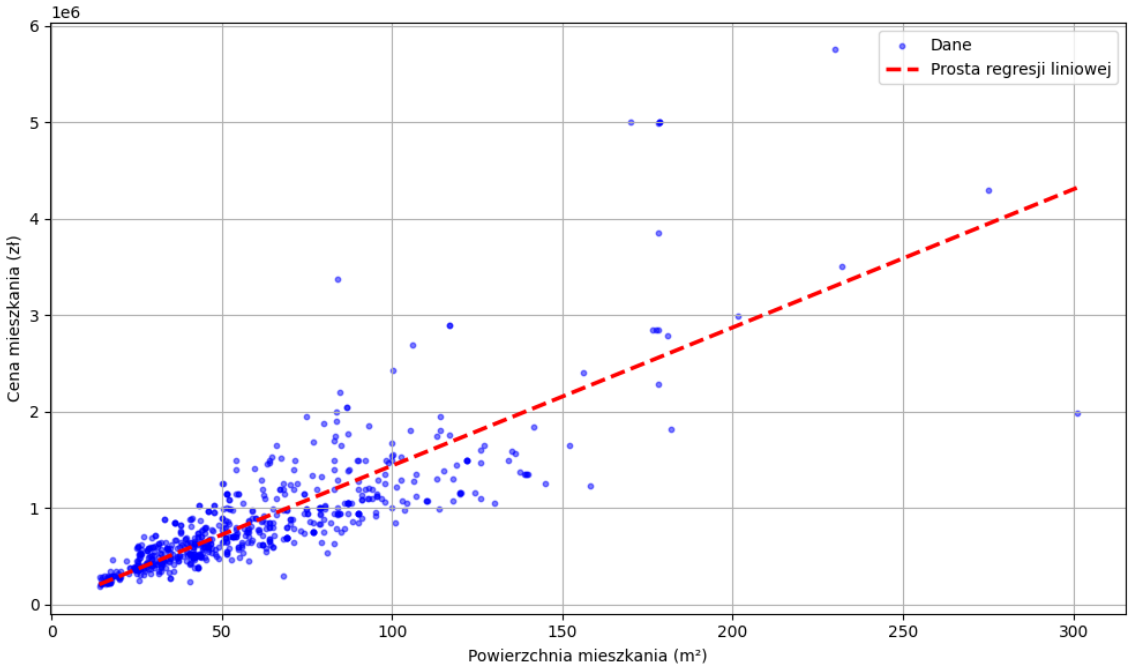
$$\begin{cases} \hat{\beta}_1 = 13639.55488020045 \approx 1.36 \cdot 10^4 \\ \hat{\beta}_0 = 38221.96091954433 \approx 3.82 \cdot 10^4 \end{cases} \tag{15}$$

Zatem wyestymowane wartości Y_i (cen mieszkań) będą miały postać:

$$\hat{y}_i = \hat{\beta}_1 x_i + \hat{\beta}_0 \approx (1.36 \cdot 10^4) x_i + 3.82 \cdot 10^4$$

4.3. Graficzne rozwiązanie (2 przypadki)

Dane z prostą regresji wyglądają następująco:

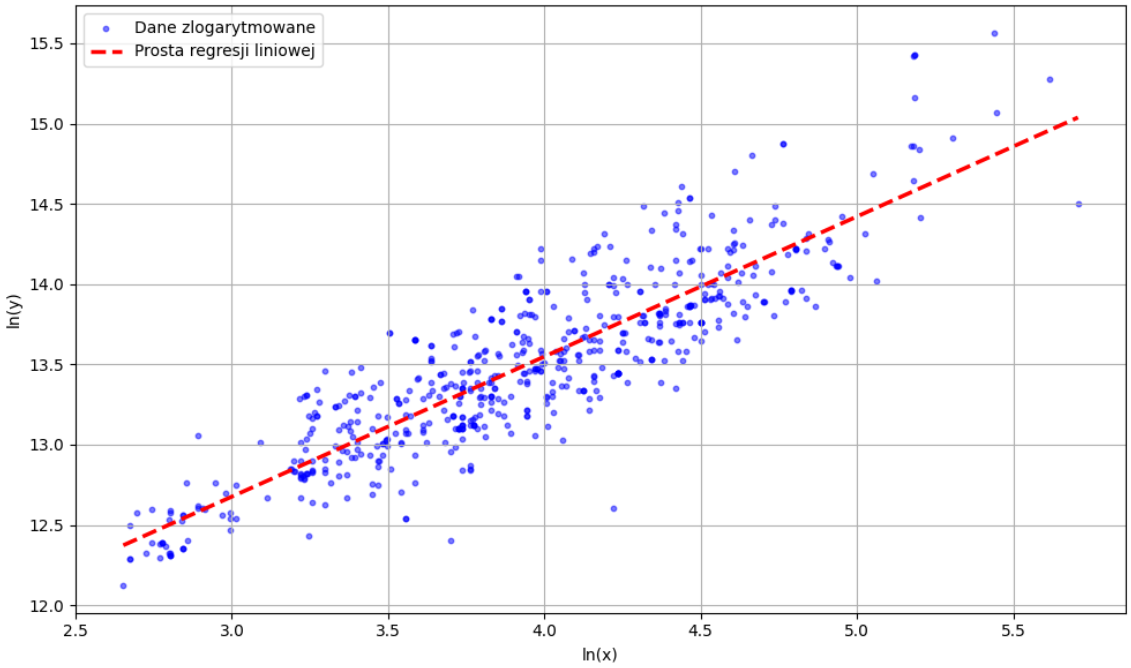


Rysunek 7: Wykres danych wraz z prostą regresji

Aby sprawdzić dokładniej jakość dopasowania naszego modelu obliczymy współczynnik determinacji (ze wzoru (10)):

$$R^2 \approx 0.674$$

Przeanalizujemy teraz drugi przypadek, w którym dane są zlogarytmowane. Wówczas wartości parametrów to $\hat{\beta}_1' \approx 0.87$ oraz $\hat{\beta}_0' \approx 10.06$, co oznacza że prosta dana jest równaniem $\hat{y}_i = 0.87x_i + 10.6$. Wykres danych zlogarytmowanych prezentuje się następująco:



Rysunek 8: Wykres zlogarytmowanych danych wraz z prostą regresji

W tym przypadku współczynnik determinacji wynosi:

$$R^2 \approx 0.776$$

Analizując powyższe informacje dochodzimy do następujących wniosków:

* Stopień dopasowania modelu regresji do danych jest dość dobry, ponieważ wartości w obu przypadkach (niezlogarytmowanym i zlogarytmowanym) są bliskie 0.7 (wartość w praktyce uznawana za dość dobrą).
* Prosta regresji jest lepiej dopasowana do danych zlogarytmowanych (współczynnik determinacji ma wyższą wartość).

5. Przedziały ufności dla parametrów β_0 i β_1

W obu przypadkach przyjmiemy poziom istotności $\alpha = 0.05$.

Wartością n jest ilość danych treningowych, zatem $n = 607$.

5.1. Dla danych niezlogarytmowanych

Będziemy konstruować przedziały ufności dla szukanych parametrów β_0 i β_1.

Posłużymy się estymatorami $\hat{\beta}_0$ oraz $\hat{\beta}_1$ policzonymi w poprzednim podpunkcie (15):

$$\begin{cases} \hat{\beta}_1 = 13639.55488020045 \approx 1.36 \cdot 10^4 \\ \hat{\beta}_0 = 38221.96091954433 \approx 3.82 \cdot 10^4 \end{cases}$$

Skoro wariancja nie jest znana, to do obliczenia przedziału ufności skorzystamy ze wzorów (13) i (12):

1. Dla β_1:

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

Implementując powyższy wzór otrzymujemy nastepujący przedział ufności β_1:

$$[12882.180; 14396.930]$$

2. Dla β_0:

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

Stosując powyższy wzór otrzymujemy przedział ufności β_0:

$$[-15410.121; 91854.043]$$

5.2. Dla danych zlogarytmowanych

Tak jak w pierwszym przypadku stosujemy te same wzory, z tą różnicą iż przyjmujemy wartości $\hat{\beta}_1^{'} = 0.87$ i $\hat{\beta}_0^{'} = 10.06$ (wyznaczone w poprzednim punkcie).

W taki sposób otrzymujemy nastepujące przedziały ufności:

1. Dla $\beta_1^{'}$: $[0.835; 0.910]$
2. Dla $\beta_0^{'}$: $[9.916; 10.209]$

6. Analiza residuów

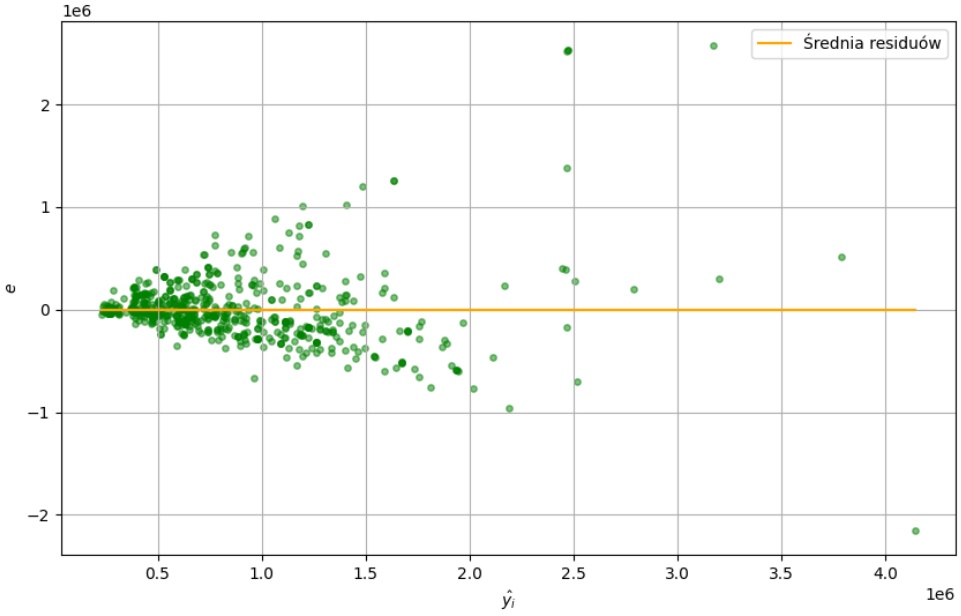
W tym punkcie zajmiemy się analizą residuów e_i, które traktujemy jako realizacje zmiennych losowych ϵ_i w modelu teoretycznym.

Analiza polega na zbadaniu poniższych założeń (będziemy sprawdzać 1,2,3):

1. $E(\epsilon_i) = 0$,
2. $Var(\epsilon_i) < \infty$,
3. $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
4. $\epsilon_1, ..., \epsilon_n$ są niezależne

6.1. Dla wartości niezlogarytmowanych

Ze wzoru (11) obliczamy residua e_i. Wyniki przedstawimy na wykresie residuów względem dopasowania (ang. "Residual vs fitted plot"), który jest powszechnie używany do wykrywania nieliniowości, nierówności wariancji błędów oraz wartości odstających.



Rysunek 9: Wykres residuów e_i względem dopasowania

6.1.1. Analiza średniej i wariancji

Jak widać na Rysunku 9, residua e_i układają się wokół wartości 0, co jest odpowiednią zależnością gdy model regresji liniowej jest odpowiedni dla zbioru danych.

Dodatkowo obliczyłyśmy średnia arytmetyczną owych residuów i wynosła ona: $\bar{e} \approx -6.291 \cdot 10^{-11}$, co jest wartością bardzo bliską 0.

Z rysunku odczytujemy także, że wariancja residuów rośnie ze wzrostem estymowanej ceny $\widehat{Y}$. Oznacza to, że residua nie mają stałej wariancji σ^2, zatem nie mogą mieć rozkładu normalnego. Jeśli jednak wyestymujemy wariancję ze wzoru (2), to jej wartość będzie wynosiła: $s^2 \approx 124266840695.495$.

6.1.2. Normalność residuów

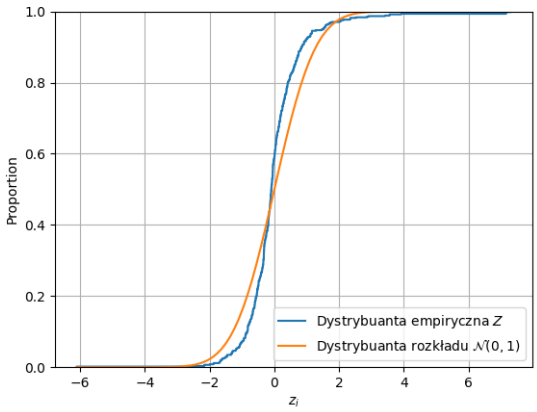
Sprawdzimy teraz czy standaryzowane residua będą mieć rokład $\mathcal{N}(0,1)$. Zrobimy to za pomocą testów statystycznych Kołmogorowa-Smirnowa oraz Jarque-Bera.

Przyjmijmy statystykę $Z = \frac{e_i - \bar{e}}{s}$ - jako standaryzowane residua.

Niech hipoteza zerowa to $H_0 : Z \sim \mathcal{N}(0,1)$, a poziom istotności to $\alpha = 0.05$. Wówczas:

* p-wartość dla testu Kołmogorowa-Smirnowa wynosi $\approx 7.690 \cdot 10^{-11} < 0.05 = \alpha$, co prowadzi do odrzucenia H_0.
* p-wartość dla testu Jarque-Bera wynosi $0 < 0.05 = \alpha$, co również prowadzi do odrzucenia H_0.

Ponadto, brak rozkładu normalnego residuów potwierdza wykres porównawczy dystrybuant na Rysunku 10:

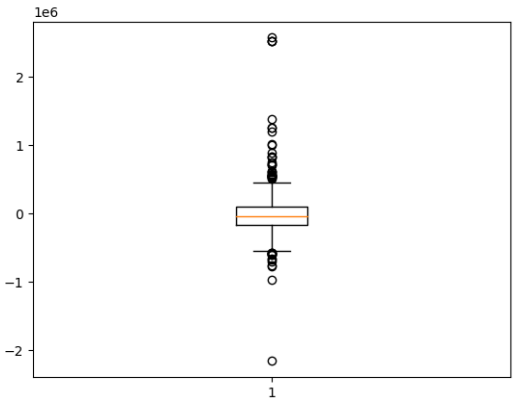


Rysunek 10: Wykres porównania dystrybuanty empirycznej Z z dystrybuantą rozkładu $\mathcal{N}(0,1)$

Z Rysunku 10 i wyników testów statystycznych wnioskujemy, że residua e_i nie pochodzą z rokładu normalnego.

6.1.3. Wartości odstające

Sprawdzimy teraz, czy są wśród residuów e_i wartości odstające.



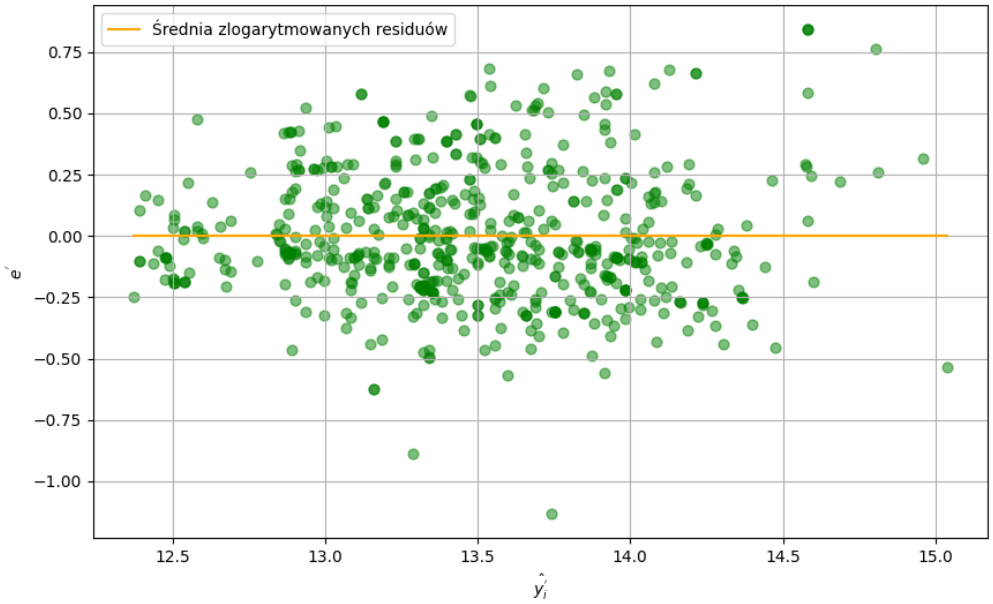
Rysunek 11: Wykres pudełkowy residuów e_i

Jak widać na wykresie pudełkowym na Rysunku 11 wśród residuów występuje dużo wartości odstających, przy czym jest ich więcej z góry, niż z dołu.

6.2. Dla wartości zlogarytmowanych

Ze wzoru (11) obliczyłyśmy residua $e_i^{'}$ dla zlogarytmowanych zmiennych.

Analogicznie jak w pierwszym przypadku wyniki prezentujemy na wykresie względem dopasowania. Przedstawiony jest on na Rysunku 12:



Rysunek 12: Wykres zlogarytmowanych residuów $e_i^{'}$ względem dopasowania

6.2.1. Analiza średniej i wariancji

Na Rysunku 12 obserwujemy, że residua e_i' układają się losowo wokół zera.

Ich średnia arytmetyczna wynosi $\bar{e}' \approx -4.887 \cdot 10^{-16}$, co jest wartością bardzo bliską 0.

Analizując powyższy wykres i zauważając, jak wartości są skoncentrowane i jak rozkładają się na wykresie, możemy stwierdzić, że dane posiadają stałą wariancję, którą estymujemy ze wzoru (2). Wynosi ona $s'^2 \approx 0.072$.

6.2.2. Normalność residuów

Sprawdzimy teraz, czy residua e' mają rozkład normalny.

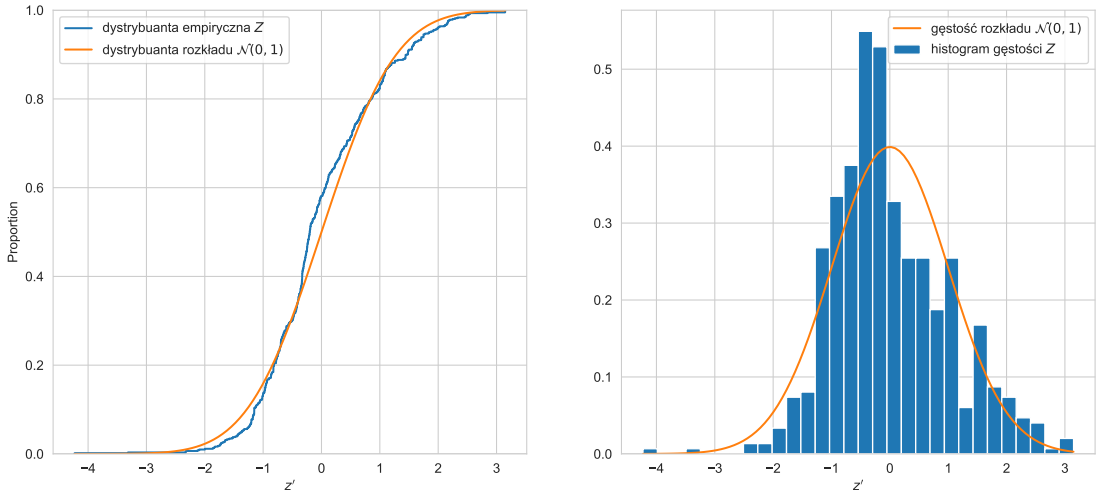
Tak jak w poprzednim przypadku, sprawdzimy, czy standaryzowane residua $Z' = \frac{e'-\bar{e}'}{s'}$ mają rozkład normalny $\mathcal{N}(0,1)$.

Analogicznie przyjmiemy poziom istotności $\alpha = 0.05$ i hipotezę zerową: $H_0 : Z^{'} \sim \mathcal{N}(0,1)$.

Wówczas:

* p-wartość dla testu Kołmogorowa-Smirnowa wynosi $\approx 6.900 \cdot 10^{-5} < 0.05 = \alpha$, co na przyjętym poziomie istotności α prowadzi do odrzucenia H_0.
* p-wartość dla testu Jarque-Bera wynosi $\approx 6.547 \cdot 10^{-6} < 0.05 = \alpha$, co prowadzi do odrzucenia H_0.

Teraz zobaczymy jak prezentuje się dystrybuanta empiryczna i histogram gęstości Z' w porównaniu do dystrybuanty i gęstości teoretycznych $\mathcal{N}(0, 1)$.



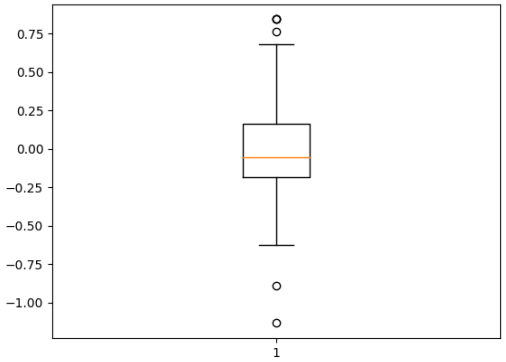
Rysunek 13: Wykres porównania dystrybuanty empirycznej i histogramu gęstości Z' do dystrybuanty i gęstości teoretycznych dla rozkładu $\mathcal{N}(0, 1)$

Z Rysunku 13 i wyników testów statystycznych wnioskujemy, że residua nie mają rokładu normalnego.

Jednak trzeba zauważyć, że, jak widać na 13, dystrybuanta residuów e'_i lepiej pokrywa się z dystrybuantą rozkładu normalnego $\mathcal{N}(0, 1)$, niż dystrybuanta residuów e_i na Rysunku 10. Zatem zlogarytmowanie naszych danych możemy uznać za zasadne.

6.2.3. Wartości odstające

Sprawdzimy teraz, czy występują wśród residuów e' wartości odstające.



Rysunek 14: Wykres pudełkowy residuów e'_i

Z Rysunku 14 wynika, że w tym przypadku również występują wartości odstające, ale jest ich znacznie mniej, niż w przypadku zmiennych niezlogarytmowanych.

Świadczy to o tym, że prosta regresji liniowej w tym przypadku jest lepiej dopasowana do danych, a więc prognoza przyszłych wartości w przypadku zmiennych zlogarytmowanych prawdopodobnie będzie dokładniejsza.

7. Prognoza przyszłej wartości $Y(x_0)$

Teraz przechodzimy do ostatniego punktu naszego raportu. Zajmiemy się prognozą przyszłych wartości Y (cen mieszkań). Przy użyciu modelu regresji, który został zbudowany na podstawie dostępnych danych będziemy przewidywać wartości zmiennej zależnej Y dla nowych, nieobserwowanych punktów danych (powierzchnia mieszkania). Do prognozy skorzystamy z danych testowych (wyznaczonych w podpunkcie 4.1.).

7.1. Dla danych niezlogarytmowanych

Jak pokazałyśmy wyżej, residua nie mają w tym przypadku stałej wariancji ani rokładu normalnego.

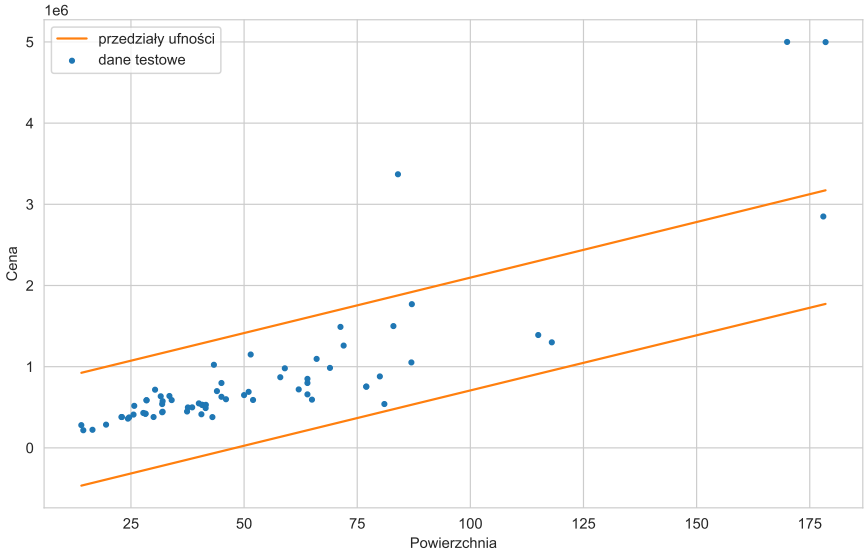
Mimo to, aby skonstruować przedział ufności dla wartości przyszłej $Y(x_0)$, założymy, że możemy wyestymować wariancję s^2 residuów ze wzoru (2) oraz że residua mają rozkład normalny $\mathcal{N}(0, s^2)$.

Dzięki powyższym założeniom możemy skorzystać ze wzoru na przedział ufności dla prognozy (14):

$$Y(x_0) \in \left[\widehat{Y}(x_0) - t_{n-2,1-\frac{\alpha}{2}} s\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i'-\bar{x}')^2}}; \widehat{Y}(x_0) + t_{n-2,1-\frac{\alpha}{2}} s\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i'-\bar{x}')^2}}\right],$$

gdzie $\widehat{Y}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$, a x_0 to nowa, nieobserwowana wcześniej wartość powierzchni.

Po podstawieniu danych i przyjęciu poziomu istotności $\alpha = 0.05$ i poziomu ufności $1-\alpha = 1-0.05 = 0.95$, otrzymujemy przedziały ufności dla wszystkich cen ze zbioru testowego (Rysunek 15).



Rysunek 15: Przedziały ufności dla cen ze zbioru testowego na poziomie ufności 0.95

Na Rysunku 15 możemy zauważyć, że większość danych testowych mieści się w zakresach przewidywań, co oznacza że poprawnie skonstruowałyśmy przedziały predykcji.

7.2. Dla danych zlogarytmowanych

Teraz skonstruujemy przedziały ufności dla zlogarytmowanych wartości nowych cen.

W przypadku zmiennych zlogarytmowanych residua mają stałą wariancję σ'^2, którą estymujemy jako

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

.

Tak jak w poprzednim przypadku, zakładamy, że residua mają rozkład normalny $\mathcal{N}(0, \sigma'^2)$.

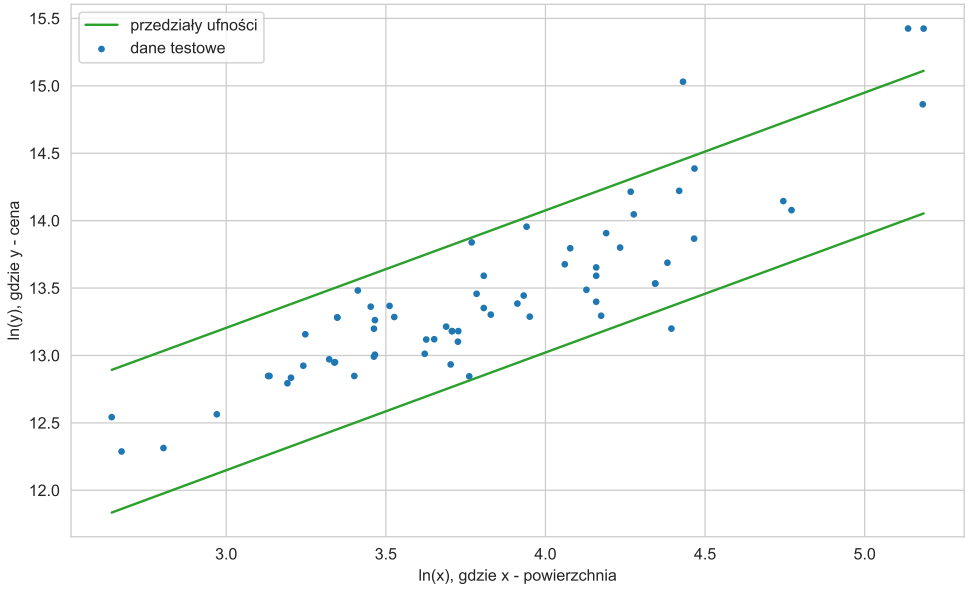
Przyszłe zlogarytmowane wartości cen mieszkań oznaczymy jako $Y'(x_0) = ln(Y(x_0))$.

Dzięki powyższym założeniom możemy skonstruować przedziały według wzoru (14):

$$Y'(x_0) \in \left[\widehat{Y}'(x_0) - A; \widehat{Y}'(x_0) + A\right],$$

gdzie $A = t_{n-2,1-\frac{1}{\alpha}} s' \sqrt{1 + \frac{1}{n} + \frac{(x'_0 - \bar{x}')^2}{\sum_{i=1}^{n}(x'_i - \bar{x}')^2}}$.

Po podstawieniu danych i przyjęciu poziomu ufności 0.95, otrzymujemy przedziały dla logarytmów wszytkich cen ze zbioru testowego (przedstawione na Rysunku 16):



Rysunek 16: Przedziały ufności dla logarytmów cen ze zbioru testowego na poziomie ufności 0.95

W tym przypadku ponownie większość danych testowych wpada do przedziałów prognozy, co oznacza poprawnie skonstruowane przedziały predykcji.

7.3. Porównanie

Przedstawione na Rysunku 16 przedziały ufności prognozują logarytmy cen.

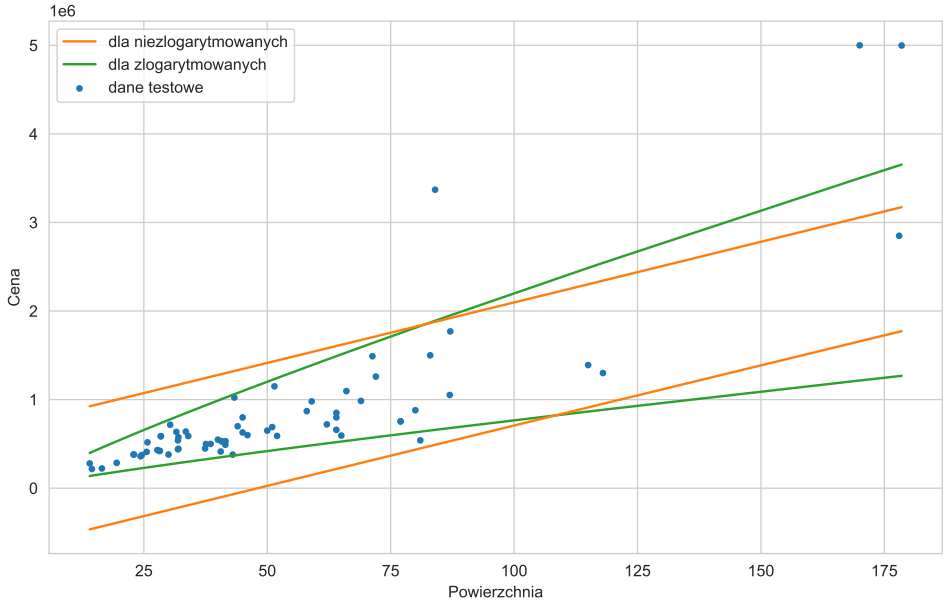
Użyjemy ich teraz tak, by prognozowały ceny, a następnie porównamy wyniki tej metody z klasyczną prognozą wykonaną w podpunkcie 7.1.

Rozpoczynamy od wyliczenia eksponenty z kresu dolnego i górnego dla przedziałów ufności wyznaczonych w podpunkcie 7.2.

Skoro wiemy, iż $Y'(x_0) \in \left[\widehat{Y}'(x_0) - A; \widehat{Y}'(x_0) + A\right]$, to $Y(x_0) \in \left[e^{\widehat{Y}'(x_0)-A}; e^{\widehat{Y}'(x_0)+A}\right]$

gdzie $A = t_{n-2,1-\frac{1}{\alpha}} s' \sqrt{1 + \frac{1}{n} + \frac{(x'_0 - \bar{x}')^2}{\sum_{i=1}^{n}(x'_i - \bar{x}')^2}}$.

Wówczas porównanie przedziałów ufności prezentuje się następująco:



Rysunek 17: Porównanie przedziałów ufności prognozy (2 sposoby)

Jak widać na powyższym Rysunku 17, przedziały ufności wyliczone dla zlogarytmowanych cen rozszerzają się ze wzrostem powierzchni mieszkań.

Im mniejsza powierzchnia mieszkania, tym węższy jest przedział ufności dla jego ceny i tym dokładniejsza jest prognoza. Wzrost szerokości przedziałów ufności ma sens w naszym przypadku, ponieważ im większa jest powierzchnia mieszkań tym większe jest standardowe odchylenie cen, a więc predykcja musi być mniej dokładna.

Średnia szerokość przedziałów ufności wyliczonych dla zmiennych zlogarytmowanych po przeliczeniu wynosi 818280.141, szerokość przedziałów ufności dla wyliczonych dla niezlogarytmowanych zmiennych wynosi 1389798.790.

8. Wnioski

Celem tego raporu było ustalić liniową zależność między cenami i powierzchniami mieszkań.

Jedno z założeń regresji liniowej - homoskedastyczność, nie było spełnione dla wybranych przez nas danych , zatem zastosowałyśmy transformację logarytmiczną i w ten sposób otrzymałyśmy dane o stałej wariancji.

Przeprowadziłyśmy analogiczną analizę dla obu przypadków: danych zlogarytmowanych i niezlogarytmowanych, i wysunęłyśmy następujące wnioski:

* W obu przypadkach proste regresji liniowej są dość dobrze dopasowane do danych, jednak w przypadku danych zlogarytmowanych współczynnik determinacji jest wyższy, co łączy się z bardziej efektywnym dopasowaniem.
* W przypadku danych zlogarytmowanych szerokość przedziałów ufności jest tym większa, im wyższe są ceny, co dla danych o zmiennej wariancji jest zasadne. Jednocześnie średnia szerokość przedziału ufności dla danych zlogarytmowanych jest mniejsza niż szerokość przedziałów dla danych niezlogarytmowanych, co oznacza, że prognoza jest dokładniejsza niż w nietrasformowanym przypadku.
* W przypadku danych niezlogarytmowanych wśród residuów jest dużo wartości odstających, co może pogarszać dopasowanie prostej do danych i zmniejszać dokładność prognozy.

Ogólnie można powiedzieć, że w przypadku danych zlogarytmowanych otrzymujemy lepsze wyniki, niż w przypadku danych niezlogarytmowanych.

Warto zaznaczyć, że ceny mieszkań zależą nie tylko od ich powierzchni, ale także od wielu innych czynników takich jak na przykład: rok budowy, liczba pokoi czy numer piętra. Z pewnością wzięcie ich pod uwagę pozwoliłoby nam na lepszą prognozę przyszłej wartości ceny.

Fakt, iz operowałyśmy na danych rzeczywistych spowodował, że nie zostały spełnione pewne założenia modelu regresji liniowej. Mimo to udało nam sie znaleźć prostą regresji, która dość dobrze dopasowała nasze dane, a ponadto przeprowadziłyśmy efektywną prognozę.

Co więcej wiedza o sposobie zniwelowania heteroskedastyczności pozwoliła nam na polepszenie dopasowania prostej regresji i predykcji.

Literatura

[1] Dr hab. inż. Agnieszka Wyłomańska, prof. uczelni, *Komputerowa Analiza Szeregów Czasowych: Wykłady*, semestr zimowy 2023/2024.

[2] Cegielski, D. (2021). House prices in Poland. Retrieved from `https://www.kaggle.com/datasets/dawidcegielski/house-prices-in-poland/`

[3] Central Statistical Office of Poland. (2022). Construction in 2022. Retrieved from `https://stat.gov.pl/download/gfx/portalinformacyjny/pl/defaultaktualnosci/5478/13/17/1/budownictwo_w_2022_r..pdf`