Analiza Statystyczna: Projekt 5

Prorok Kacper, Popkiewicz Szymon

2024-05-22

Celem naszego projektu jest oszacowanie modelu, który na podstawie zmiennych objaśniających takich jak:*gęstość zaludnienia, ludność powiatu, przystanki na 10 tys. mieszkańców, odsetek ludności powyżej 65 lat, średniej ceny za metr kwadratowy, liczba parków na 10 tys. mieszkańców, średni dochód* oszacuje liczbę samochodów o napędzie zielonym na 10,000 mieszkańców.

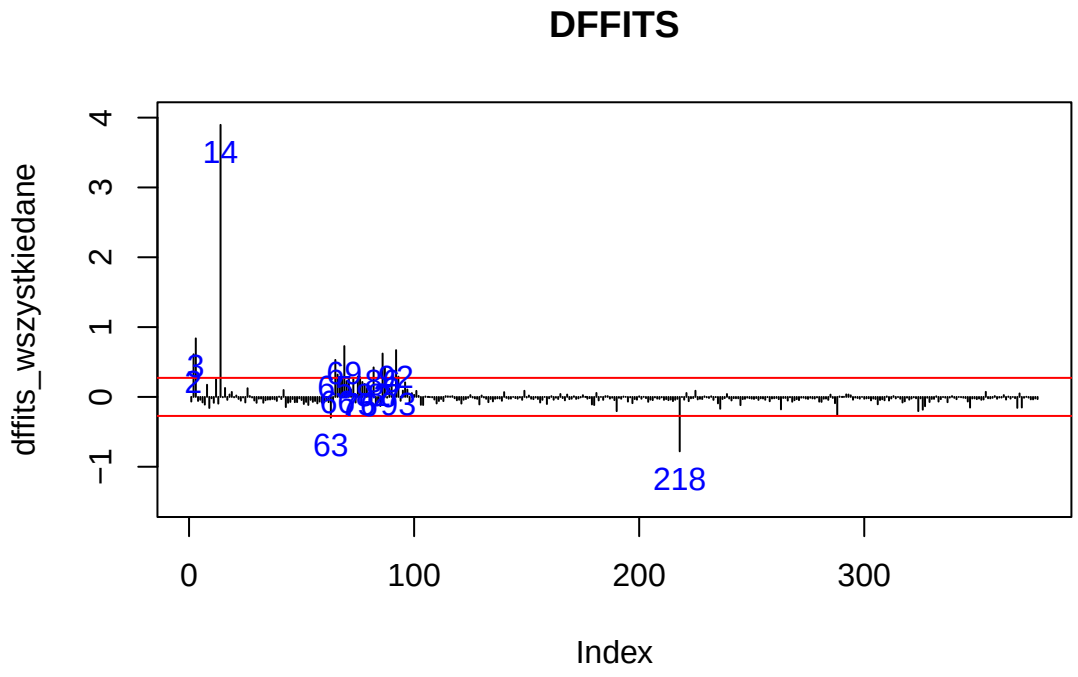
Analiza wszystkich danych

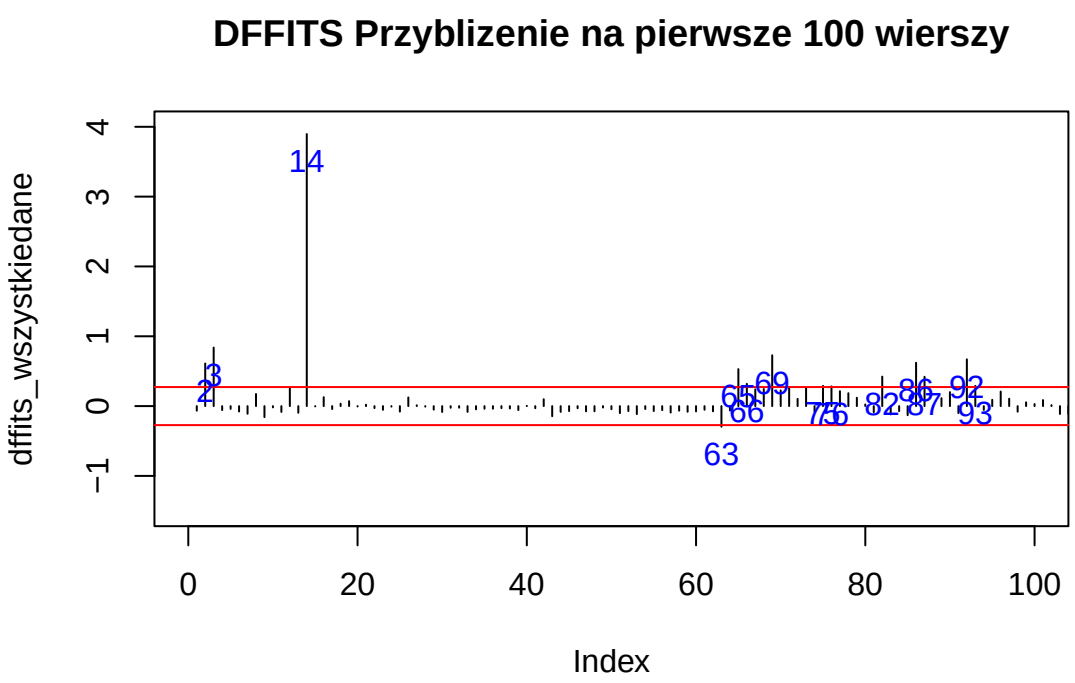
Analizę występowania obserwacji odstających

Pod uwagę weźmiemy odległość Cooka, dffits oraz dfbeats.

DFFITS

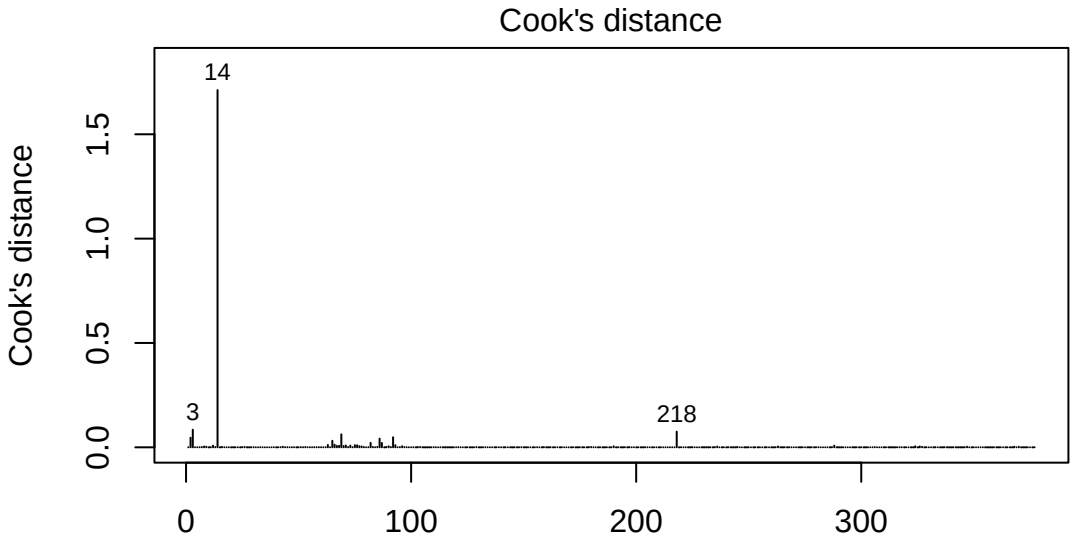
Kolorem czerwonym została zaznaczona wartość progowa - 2 * sqrt(k*/n).



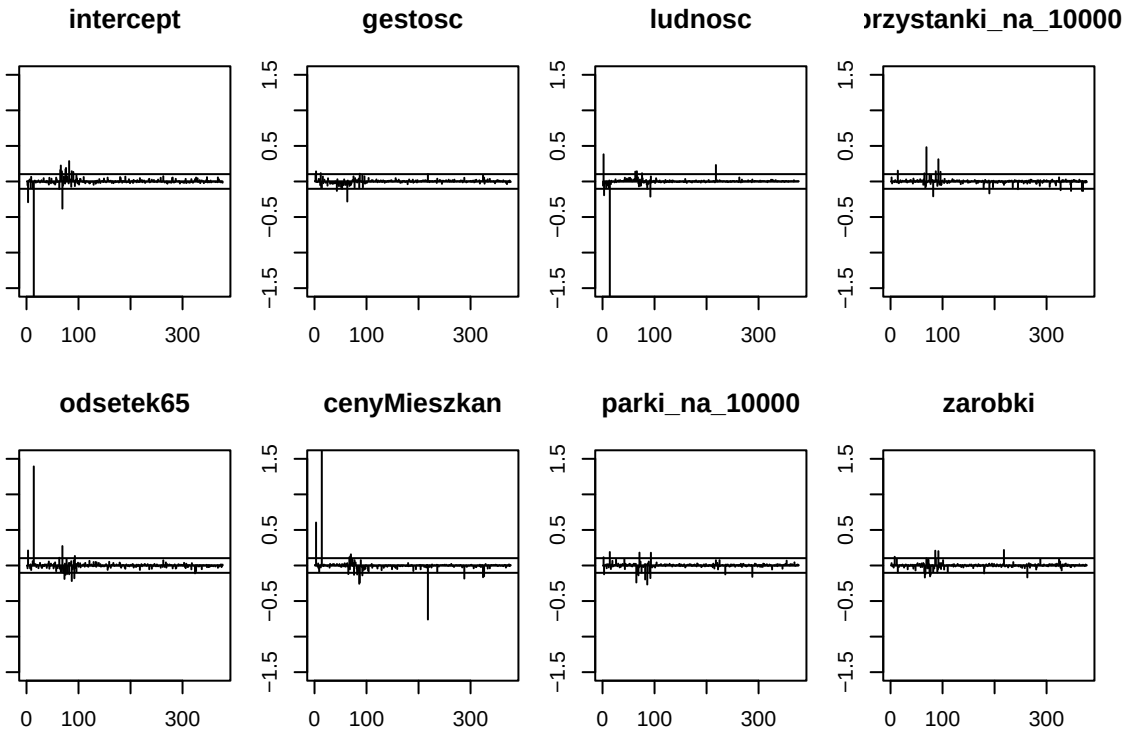


Na powyższym wykresie dla lepszej czytelności pokazujemy pierwsze 100 wierszy, ponieważ wystąpiło w nich najwięcej wartości odstających.

Odległość Cook'a



DFBETAS:



Po analizie powyższych wykresów decydujemy się na usunięcie tylko poniższych powiatów:

index	powiat
2	Powiat m. Wrocław
3	Powiat m. Gdynia
14	Powiat m. Sopot
63	Powiat m. Świętochłowice
65	Powiat jasielski
66	Powiat tarnowski
69	Powiat sandomierski
75	Powiat brzeski
76	Powiat krośnieński
82	Powiat dąbrowski
86	Powiat leski
87	Powiat sztumski
92	Powiat bieszczadzki
93	Powiat ząbkowicki
218	Powiat tatrzański

Powodem dlaczego nie usuwamy więcej zmiennych jest fakt, że mogłoby dojść do sytuacji gdzie usunelibyśmy większość powiatów grodzkich, a wtedy model straciłby sens.

Badanie współliniowości przy pomocy VIF

Inaczej określany jako współczynnik wariancji inflacji (Variance Inflation Factor) to wskaźnik, który pozwala określić czy między badanymi predyktorami występuje współliniowość.

	vif(model_wszystkie)
gestosc	2.114212
ludnosc	1.863545
przystanki_na_10000	1.351219
odsetek65	1.407667
cenyMieszkan	1.698685
parki_na_10000	1.067267
zarobki	1.375714

Wysokie wartości VIF, (np. VIF > 10) wskazują na występowanie współliniowości. Można spokojnie uznać, że w naszych danych nie występuje współliniowość.

Regresja krokowa

W metodzie krokowej model regresji jest stopniowo budowany przez dodawanie lub usuwanie zmiennych w kolejnych krokach na podstawie określonych kryteriów statystycznych.

Kryterium AIC

Kryterium Akaikego (AIC) jest miarą stosowaną do oceny jakości modeli statystycznych, z uwzględnieniem zarówno dopasowania modelu, jak i liczby parametrów. Niższa wartość AIC wskazuje na lepszy model, który dobrze dopasowuje się do danych przy jednoczesnym minimalizowaniu liczby parametrów.

Wsteczna Po użyciu funkcji *step(model_wszystkie,direction="backward")*, dostajemy:

```
Step:  AIC=3485.55
samochody_na_10000 ~ gestosc + ludnosc + odsetek65 + cenyMieszkan

               Df Sum of Sq     RSS    AIC
<none>                      5349990 3485.5
- odsetek65     1     47076 5397066 3486.7
- gestosc       1     53863 5403852 3487.2
- ludnosc       1     74257 5424246 3488.5
- cenyMieszkan  1    280458 5630448 3502.0

Call:
lm(formula = samochody_na_10000 ~ gestosc + ludnosc + odsetek65 +
    cenyMieszkan, data = df)

Coefficients:
 (Intercept)       gestosc       ludnosc     odsetek65  cenyMieszkan
   1.369e+02     2.713e-02     1.567e-04    -5.391e+00     4.359e-04
```

W przód

```
## Start:  AIC=3487.77
## samochody_na_10000 ~ gestosc + ludnosc + przystanki_na_10000 +
##     odsetek65 + cenyMieszkan + parki_na_10000 + zarobki

##
## Call:
## lm(formula = samochody_na_10000 ~ gestosc + ludnosc + przystanki_na_10000 +
##     odsetek65 + cenyMieszkan + parki_na_10000 + zarobki, data = df)
##
## Coefficients:
##         (Intercept)             gestosc             ludnosc
##           81.5716501            0.0299000            0.0001310
## przystanki_na_10000            odsetek65         cenyMieszkan
##            0.1898727           -6.1599222            0.0004104
##      parki_na_10000              zarobki
##           -1.4891138            0.0134671
```

Wybór zmiennych

Dodatkowo zobaczymy jak wyglądają statystyki dla metody Hellwiga:

```
## $ludnosc_odsetek65_cenyMieszkan_parki_na_10000
## [1] 0.1783194
##
## $gestosc_ludnosc_cenyMieszkan_parki_na_10000
## [1] 0.1786377
##
## $ludnosc_cenyMieszkan
## [1] 0.1794827
##
## $ludnosc_cenyMieszkan_parki_na_10000
## [1] 0.1803378
```

Wyniki uległy poprawnie w stosunku do projektu 4, gdzie najwyższe kryterium informacyjne uzyskane metodą Hellwiga wyniosło: 0.0891. Oznacza to, że pozbycie się zmiennych odstających w tym projekcie przyniosło lepsze rezulaty niż w projekcie 4.

Biorąc pod uwagę wszystkie kryteria, dochodzimy do wniosku, by wybrać model składający się z gęstość, odsetków ludności powyzej 65 roku życia oraz ceny mieszkań.

Weryfikacja założeń Gaussa-Markowa

Heteroskedatyczność

~ Wariancja błędów modelu (reszt) nie jest stała dla wszystkich obserwacji. Oznacza to, że rozproszenie błędów różni się w zależności od wartości zmiennych niezależnych.

	Gold.Quand
p-value	0

Odrzucamy hiptezę zerową, na korzyść hipotezy alternatywnej mówiącej o heteroskedastyczności modelu - **model jest heteroskedastyczny.**

Autokorelacja

W przypadku danych, które nie są szeregiem czasowym, badanie autokorelacji wydaje się być zbędne.

Liniowość

Model liniowy zakłada, że zmiany w zmiennych niezależnych przekładają się na proporcjonalne zmiany w zmiennej zależnej, co sprawia, że relacja między nimi jest liniowa.

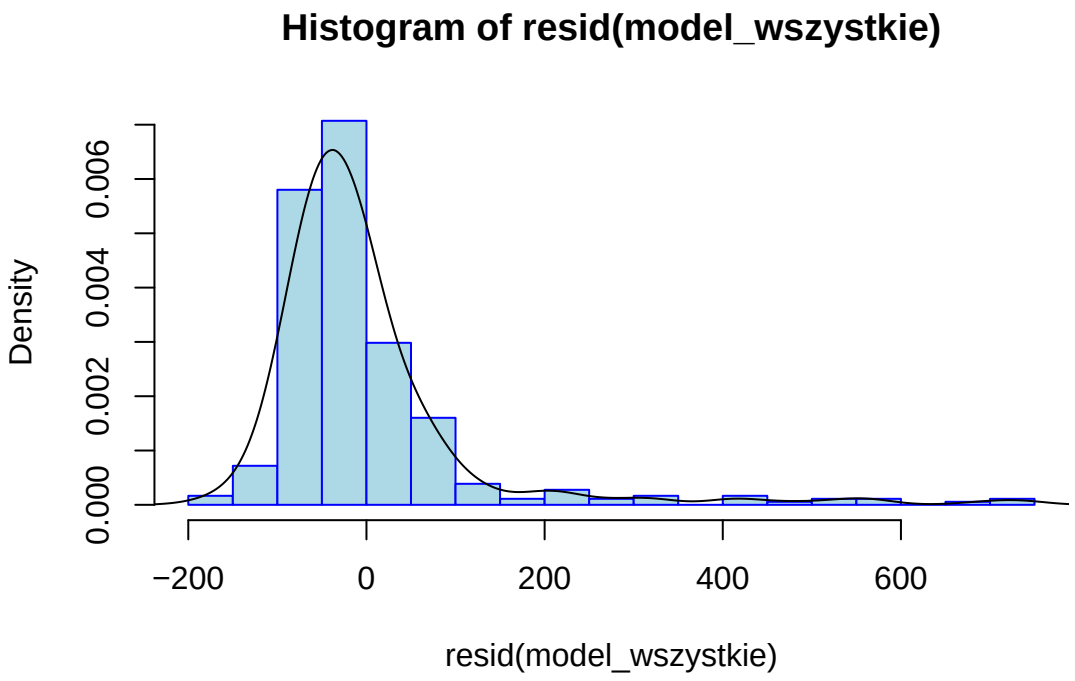
Wykonamy test RESET.

```
##
##  RESET test
##
## data:  model_wszystkie
## RESET = 2.5606, df1 = 2, df2 = 356, p-value = 0.07868
```

Ze względu na p-value większe niż 0.05, **możemy stwierdzić liniowość modelu.**

Normalność reszt modelu

Do sprawdzenia normalności wykonamy test Shapiro-Wilka oraz w Jarque-Bera. Poniżej znajduje się rozkład reszt:



Już po samym wykresie można stwierdzić, że rozkład reszt raczej nie będzie zgodny z rozkładem normalnym. Przeprowadzamy jeszcze testy statystyczne:

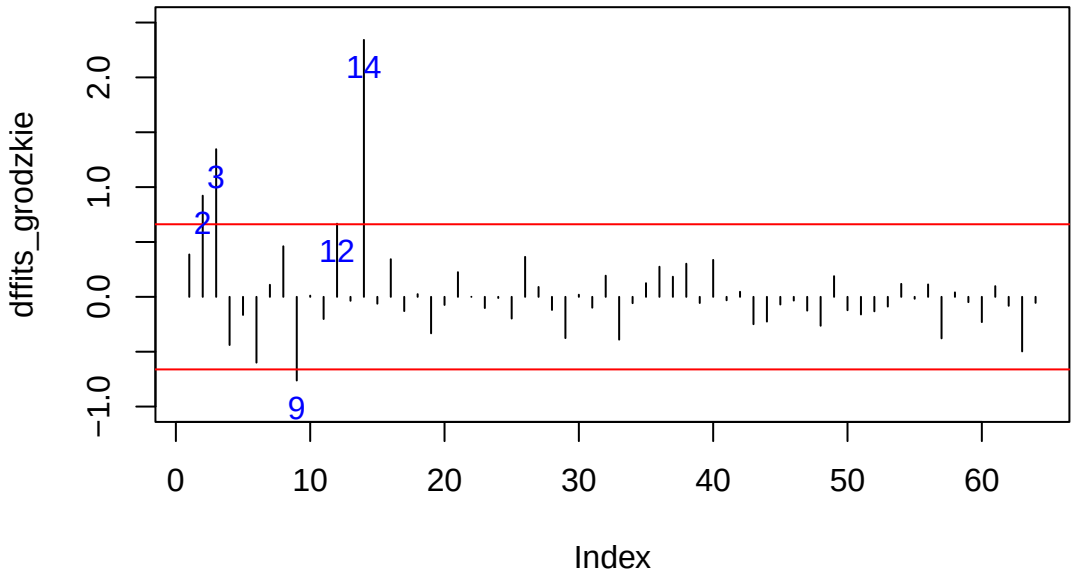
	shapiro	JB
p-value	0	0

Odrzucamy hipotezę zerową o normalności rozkładu reszt. **Reszty nie mają rozkładu normalnego.**

Dane grodzkie

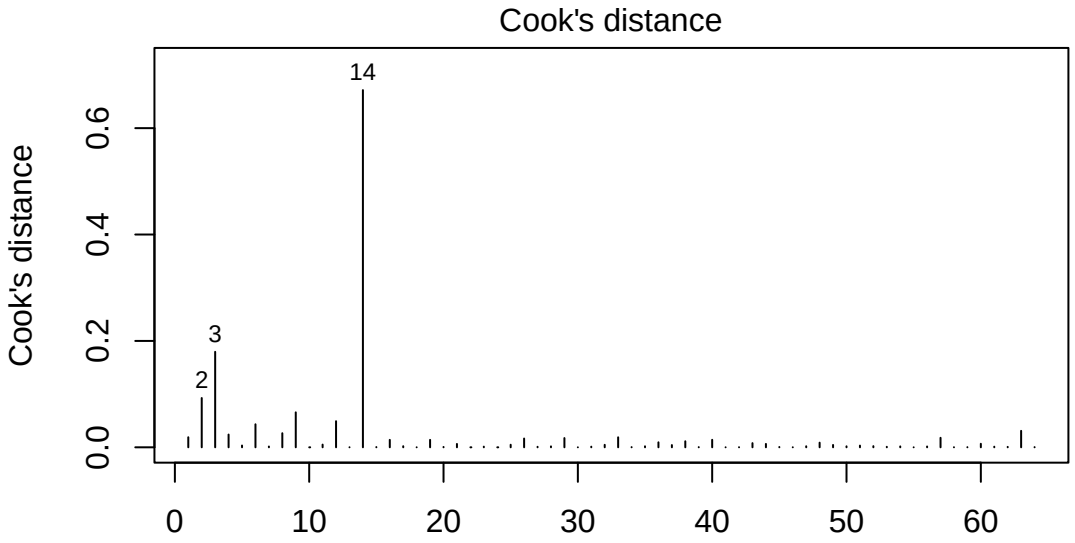
Analizę występowania obserwacji odstających

DFFITS

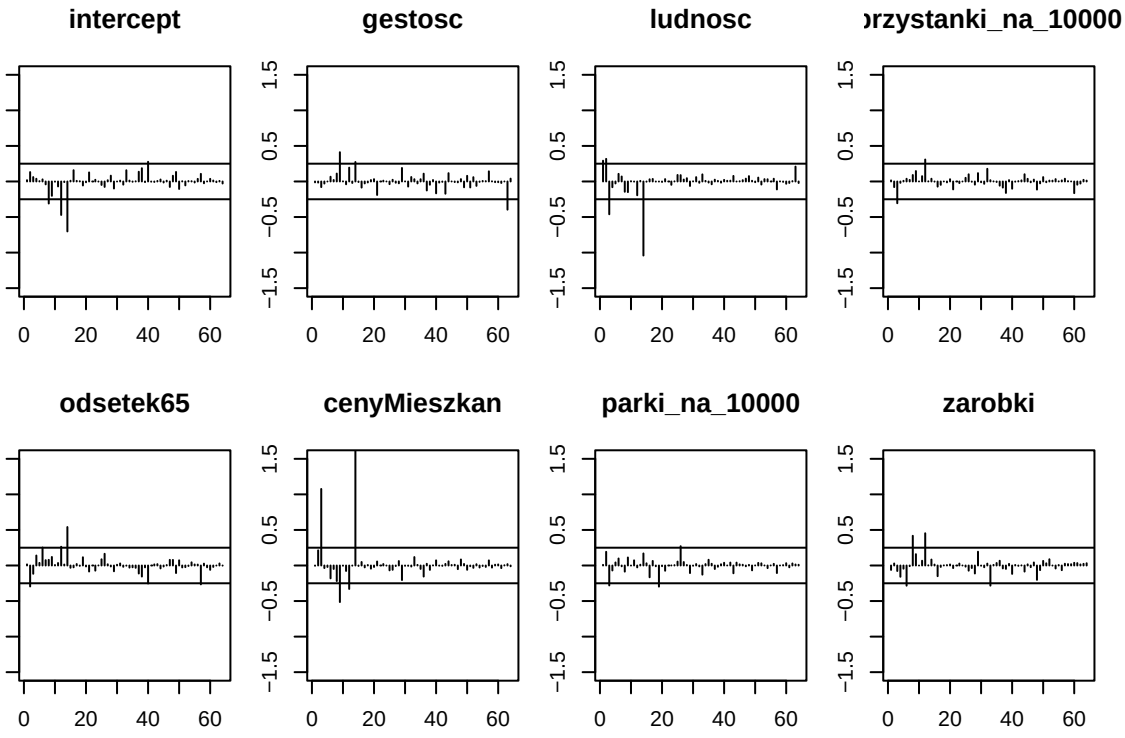


Kolorem czerwonym została zaznaczona wartość progowa - 2 * sqrt(k*/n).

Odległość Cook'a



DFBETAS



Z racji że danych o powiatach grodzkich jest dużo mniej niż o powiatach ziemskich usuniemy tylko najbardziej odstające obserwacje, aby mieć ich więcej.

Ostatecznie decydujemy się na usunięcie obserwacji pokazanych poniżej.

index	powiat
2	Powiat m. Wrocław
3	Powiat m. Gdynia
9	Powiat m. Szczecin
14	Powiat m. Sopot
63	Powiat m. Świętochłowice
12	Powiat m. Tarnów
16	Powiat m. Elbląg
8	Powiat m. Katowice
26	Powiat m. Kalisz

Badanie współliniowości przy pomocy VIF

	vif(model_grodzki)
gestosc	1.930833
ludnosc	4.244552
przystanki_na_10000	1.407447
odsetek65	1.126000
cenyMieszkan	3.159483
parki_na_10000	1.046499
zarobki	2.175126

Widzimy, że dla zmiennej 'ludnosc' oraz 'cenyMieszkan' występują wysokie wartości współczynnika VIF. Usuwamy zmienną 'ludnosc' i przedstawiamy ponownie wyniki:

	vif(lm(samochody_na_10000 ~ gestosc + przystanki_na_10000 + odsetek65 + cenyMieszkan + parki_na_10000 + zarobki, data = grodzkie))
gestosc	1.527923
przystanki_na_10000	1.402462
odsetek65	1.111280
cenyMieszkan	2.016630
parki_na_10000	1.043485
zarobki	1.851146

Wniosek: Usuwając ludność, pozbylibyśmy się współliniowości.

Regresja krokowa

Kryterium AIC

Kryterium Akaikego (AIC) jest miarą stosowaną do oceny jakości modeli statystycznych, z uwzględnieniem zarówno dopasowania modelu, jak i liczby parametrów. Niższa wartość AIC wskazuje na lepszy model, który dobrze dopasowuje się do danych przy jednoczesnym minimalizowaniu liczby parametrów.

Wsteczna .

```
Step:  AIC=446.19
samochody_na_10000 ~ gestosc + przystanki_na_10000 + cenyMieszkan +
    zarobki

                     Df Sum of Sq    RSS    AIC
<none>                             152971 446.19
- zarobki             1      8197 161167 447.06
- gestosc             1      8461 161432 447.15
- przystanki_na_10000 1     12330 165301 448.45
- cenyMieszkan        1    206459 359430 491.17

Call:
lm(formula = samochody_na_10000 ~ gestosc + przystanki_na_10000 +
    cenyMieszkan + zarobki, data = grodzkie)

Coefficients:
       (Intercept)             gestosc  przystanki_na_10000         cenyMieszkan
        -3.188e+02           2.345e-02           1.530e+00           9.544e-04
           zarobki
         2.322e-02
```

W przód

```
## Start:  AIC=451.52
## samochody_na_10000 ~ gestosc + ludnosc + przystanki_na_10000 +
##     odsetek65 + cenyMieszkan + parki_na_10000 + zarobki

##
## Call:
## lm(formula = samochody_na_10000 ~ gestosc + ludnosc + przystanki_na_10000 +
##     odsetek65 + cenyMieszkan + parki_na_10000 + zarobki, data = grodzkie)
##
## Coefficients:
##         (Intercept)              gestosc              ludnosc
##          -2.490e+02            1.875e-02            2.590e-05
## przystanki_na_10000            odsetek65         cenyMieszkan
##           1.442e+00           -1.364e+00            9.044e-04
##      parki_na_10000              zarobki
##          -5.845e-01            2.081e-02
```

Wybór zmiennych

Dla przypomnienia pokażemy kryteria pojemności informacyjnej uzyskanych metodą Hellwiga:

```
## $cenyMieszkan_parki_na_10000
## [1] 0.7078229
##
## $ludnosc_cenyMieszkan_parki_na_10000
## [1] 0.7113471
##
## $ludnosc_cenyMieszkan
## [1] 0.738465
##
## $cenyMieszkan
## [1] 0.7675052
```

W przypadku danych grodzkich wyniki uzyskane metodą Hellwiga uległy lekkiej poprawie - najwyższa wartość uzyskana w projekcie 4 wyniosła: 0.7071838, a teraz 0.7675.

Biorąc pod uwagę wszystkie kryteria, dochodzimy do wniosku, by wybrać model składający się z cen mieszkań,zarobków oraz przystanków na 10000 mieszkańców.

Weryfikacja założeń Gaussa-Markowa

Heteroskedatyczność

	Gold.Quand
p-value	0.1387

Brak podstaw do odrzucenia hipotezy H0. **Model jest homoskedastyczny**

Liniowość

Model liniowy zakłada, że zmiany w zmiennych niezależnych przekładają się na proporcjonalne zmiany w zmiennej zależnej, co sprawia, że relacja między nimi jest liniowa.

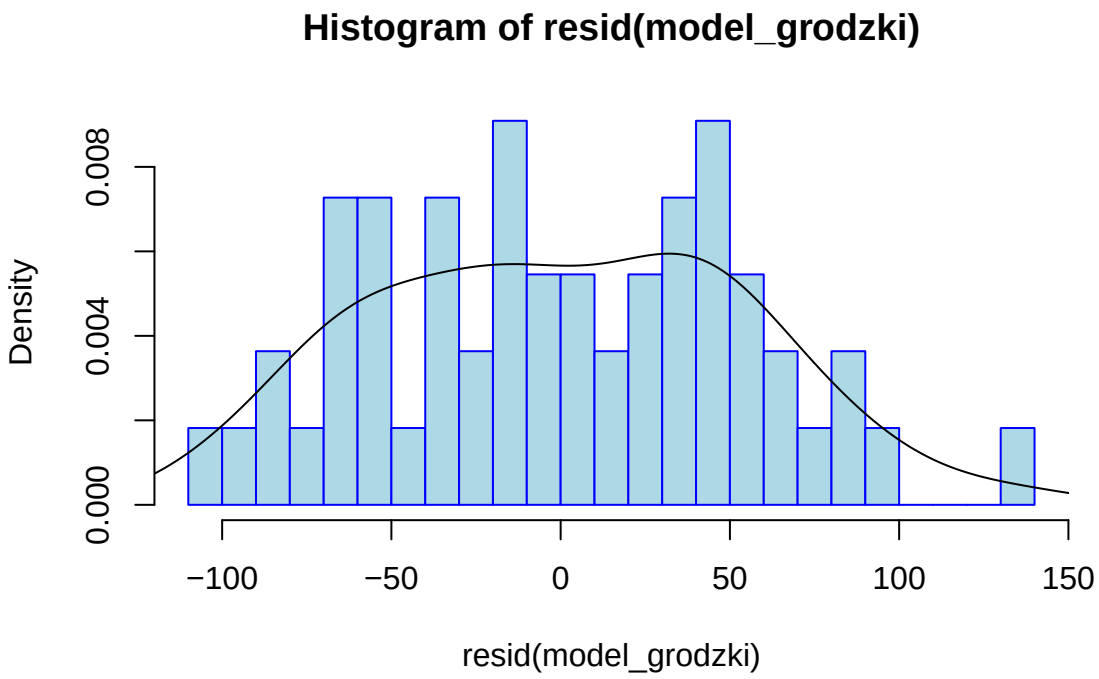
Wykonamy test RESET.

```
##
##  RESET test
##
## data:  model_grodzki
## RESET = 0, df1 = 3, df2 = 48, p-value = 1
```

Przyjmujemy hipotezę zerową, **stwierdzamy liniowość modelu.**

Normalność reszt modelu

Do sprawdzenia normalności wykonamy test Shapiro-Wilka oraz w Jarque-Bera. Poniżej znajduje się rozkład reszt:

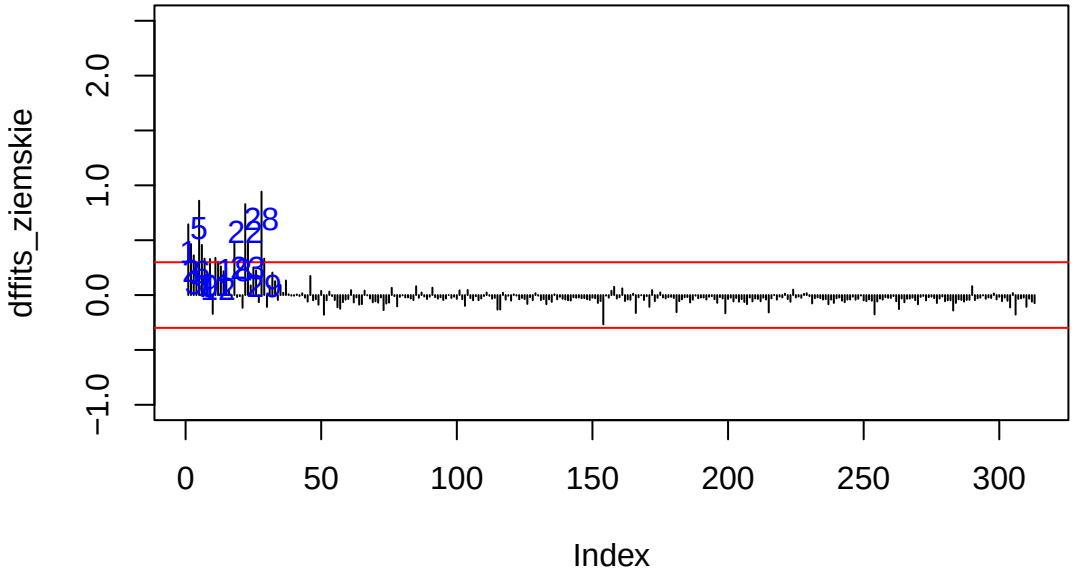


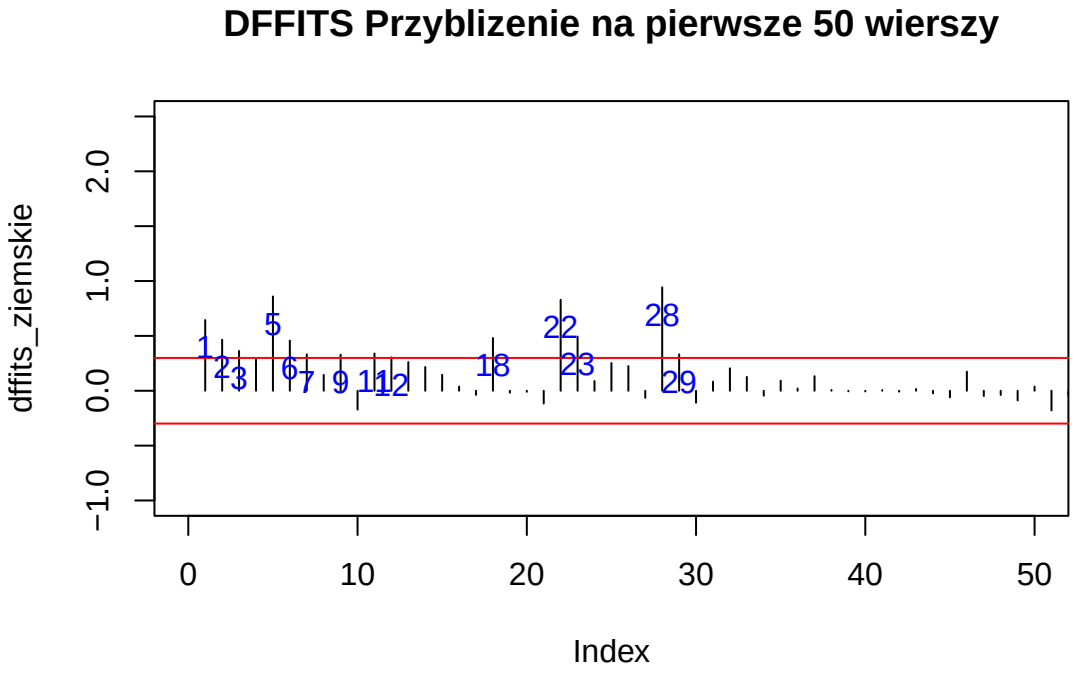
	shapiro	JB
p-value	0.6597873	0.5581442

Przyjmujemy hipotezę zerową o normalności rozkładu reszt.**Reszty modelu mają rozkład normalny**

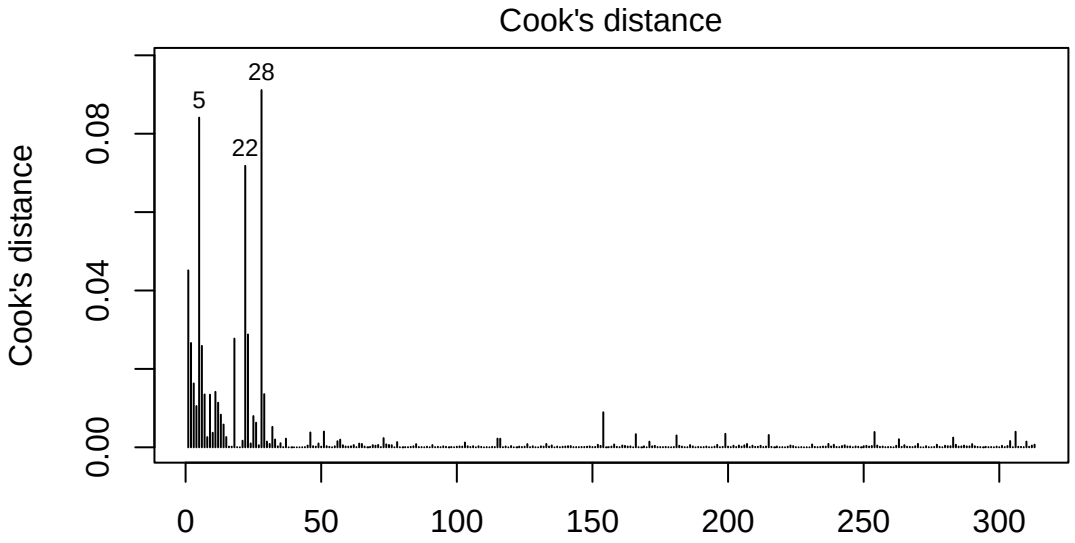
Dane ziemskie

Analiza występowania obserwacji odstających

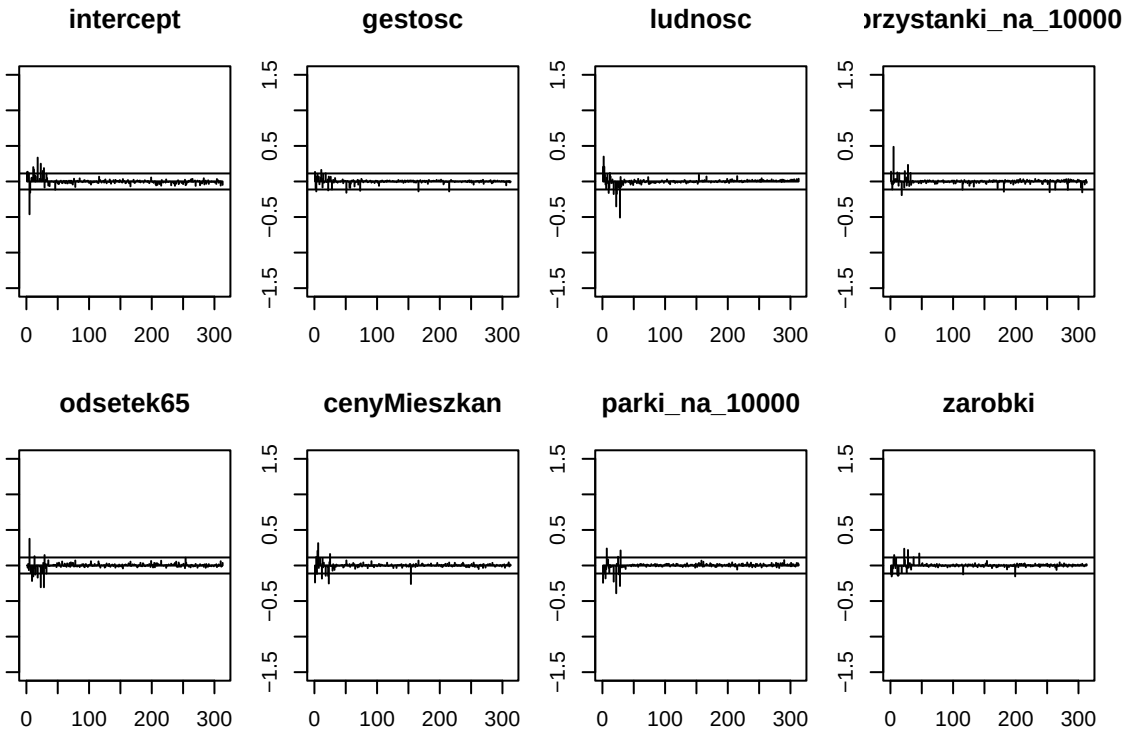




Odległość Cook'a



DFBETAS



Wśród otrzymanych rekordów mamy m.in:

index	powiat
1	Powiat jasielski
5	Powiat sandomierski
11	Powiat brzeski
18	Powiat dąbrowski
22	Powiat leski
23	Powiat sztumski
28	Powiat bieszczadzki
46	Powiat lubiński
51	Powiat wodzisławski
215	Powiat bieruńsko-lędziński
154	Powiat tatrzański

Jako,że danych grodzkich jest dość sporo, usuniemy około 100 wierszy

Zmienne odstające wybraliśmy na podstawie wyboru zmiennych, których dystans Cook'a wyniósł więcej niż 0.01. Operację tą powtórzyliśmy 5 razy, za każdym razem tworząc model na nowych zmiennych.

Badanie współliniowości przy pomocy VIF

	vif(model_ziemski)
gestosc	1.842439
ludnosc	1.906340
przystanki_na_10000	1.168457
odsetek65	1.151934
cenyMieszkan	1.204182
parki_na_10000	1.127211
zarobki	1.112423

Wartości nie przekraczają wartości 2. Możemy stwierdzić **brak współliniowości.**

Regresja krokowa

Kryterium AIC

Wsteczna .

```
Step:  AIC=1204.62
samochody_na_10000 ~ przystanki_na_10000 + odsetek65 + cenyMieszkan +
    parki_na_10000 + zarobki

                     Df Sum of Sq   RSS    AIC
<none>                             72271 1204.6
- odsetek65           1    1321.6 73593 1206.3
- parki_na_10000      1    1673.1 73944 1207.3
- przystanki_na_10000 1    8171.6 80443 1224.4
- cenyMieszkan        1    9294.9 81566 1227.2
- zarobki             1   12034.0 84305 1233.9

Call:
lm(formula = samochody_na_10000 ~ przystanki_na_10000 + odsetek65 +
    cenyMieszkan + parki_na_10000 + zarobki, data = ziemskie)

Coefficients:
       (Intercept)  przystanki_na_10000           odsetek65         cenyMieszkan       parki_na_10000             zarobki
         -1.514e+01            2.983e-01           -1.418e+00            1.171e-04           -5.433e-01            1.974e-02
```

W przód

```
## Start:  AIC=1207.36
## samochody_na_10000 ~ gestosc + ludnosc + przystanki_na_10000 +
##     odsetek65 + cenyMieszkan + parki_na_10000 + zarobki
```

```
##
## Call:
## lm(formula = samochody_na_10000 ~ gestosc + ludnosc + przystanki_na_10000 +
##     odsetek65 + cenyMieszkan + parki_na_10000 + zarobki, data = ziemskie)
##
## Coefficients:
##         (Intercept)             gestosc             ludnosc
##          -1.630e+01           7.319e-04           4.951e-05
## przystanki_na_10000           odsetek65        cenyMieszkan
##           3.025e-01          -1.338e+00           1.085e-04
##      parki_na_10000             zarobki
##          -4.809e-01           1.926e-02
```

Wybór zmiennych

Hellwig:

```
## $przystanki_na_10000_cenyMieszkan_parki_na_10000_zarobki
## [1] 0.2243831
##
## $ludnosc_przystanki_na_10000_cenyMieszkan_parki_na_10000_zarobki
## [1] 0.2264111
##
## $przystanki_na_10000_cenyMieszkan_zarobki
## [1] 0.2314234
##
## $ludnosc_przystanki_na_10000_cenyMieszkan_zarobki
## [1] 0.234629
```

W porównaniu do projektu 4 wyniki uległy poprawie - najwyższa wartość w projekcie 4 wyniosła: 0.02832488.

Biorąc pod uwagę wszystkie kryteria, dochodzimy do wniosku, by wybrać model składający się z: przystanki_na_10000, cenyMieszkan, zarobki.

Weryfikacja założeń Gaussa-Markowa

Heteroskedatyczność

	Gold.Quand
p-value	0.888616

Przyjmujemy hiptezę zerową. **Model jest homoskedastyczny**

Liniowość

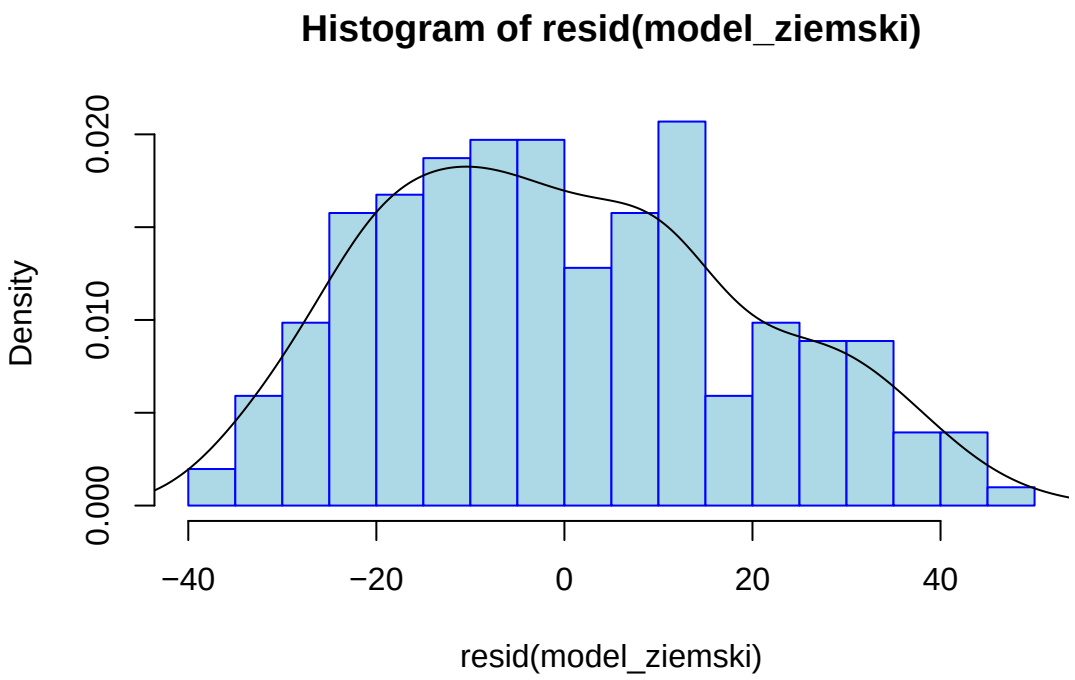
Test RESET.

```
##
##  RESET test
##
## data:  model_ziemski
## RESET = 1.0692, df1 = 2, df2 = 197, p-value = 0.3453
```

Przyjmujemy hipotezę zerową, stwierdzając liniowość modelu. **Model jest liniowy**

Normalność reszt modelu

Do sprawdzenia normalności wykonamy test Shapiro-Wilka oraz w Jarque-Bera. Poniżej znajduje się rozkład reszt:



	shapiro	JB
p-value	0.0026762	0.0269317

Odrzucamy hipotezę zerową. **Reszty modelu nie mają rozkładu normalnego**

Próba poprawy modelu

Aby poprawić sytuację naszego modelu przeprowadzimy transformacje danych.Testowane przez nas metody transformacji obejmują:

- Logarytmowanie: Używanie logarytmu zmiennej objaśnianej i/lub zmiennych objaśniających.
- Pierwiastkowanie: Używanie pierwiastka kwadratowego lub innego stopnia zmiennej.

Logarytm zmiennej objaśnianej

Zobaczmy na wyniki testów statystycznych, gdy 'zlogarytmujemy' zmienną objaśnianą:

Table 14: log(y)

	grodzkie	ziemskie	wszystkie
Gold.Quand	0.1855820	0.0070932	0.0000000
RESET_TEST	0.0761829	0.7039811	0.0968058
shapiro	0.1207080	0.0536207	0.0000000

Dla wszystkich powiatów oraz dla grodzkich nie widać poprawy. W przypadku powiatów ziemskich model stał się heteroskedastyczny, ale za to reszty mają rozkład normalny.

Pierwiastkowanie zmiennej objaśnianej

Spróbujmy spierwastkować zmienną objaśnianą.

Table 15: sqrt(y)

	grodzkie	ziemskie	wszystkie
Gold.Quand	0.9462192	0.0525598	0.000000
RESET_TEST	0.4737318	0.7932179	0.091421
shapiro	0.3642724	0.2558609	0.000000

Duża poprawa modelu dla powiatów ziemskich - tym razem wszystkie założenia Gaussa-Markova zostałyby spełnione.

Ważona metoda najmniejszych kwadratów

Na koniec spróbujemy spróbujemy skorzystać z bardziej zaawansowanej metody -'Weighted Least Squares Regression',to technika statystyczna używana do estymacji parametrów modelu regresji, w której różne obserwacje mają różne wagi.

Wagi obserwacji przypisujemy w taki sposób, by zrównoważyć obserwacje bardziej oraz mniej odstające.

Table 16: WLS

	grodzkie	ziemskie	wszystkie
Gold.Quand	0.1387000	0.2159493	0.0000000
RESET_TEST	0.8722356	0.8642258	0.0786789
shapiro	0.6668795	0.1993464	0.0000000

w tym przypadku również widzimy ewidentną poprawę modelu dla powiatów ziemskich - ponownie model spełniłby założenia Gaussa-Markova.

Podsumowanie

Na podstawie analizy przeprowadzonej w tym projekcie wybraliśmy najlepsze zestawy zmiennych objaśniających dla każdego z modeli. Ostatecznie **model dla danych grodzkich** okazuje się być najbardziej wiarygodnym modelem spośród wszystkich - ma on najwyższe wartości AIC oraz kryteriów informacyjnych Hellwiga oraz spełnione wszystkie założenia Gaussa-Markova. Model dla powiatów ziemskich po spierwiastkowaniu zmiennej objaśnianej również uległ poprawie i spełnia założenia Gaussa-Markova.