Spis treści

1. Wybór danych

Dane jakie wybraliśmy to Energy Efficiency Dataset pochodzą z platformy Kaggle i dotyczą analizy energetycznej przy użyciu 12 różnych kształtów budynków symulowanych w Ecotect. Budynki różnią się między innymi powierzchnią przeszklenia, rozkładem powierzchni przeszklenia i orientacją. Symulujemy różne ustawienia jako funkcje wyżej wymienionych cech, aby uzyskać 768 kształtów budynków. Zbiór danych obejmuje 768 próbek i 8 cech, których celem jest przewidzenie dwóch rzeczywistych odpowiedzi. Jako zmienną zależną wybraliśmy ilość energii cieplnej wymaganej do ogrzania budynku. Zmienna objaśniające to: X1 - Względna zwartość, X2 - Powierzchnia, X3 - Powierzchnia Ścian, X4 - Powierzchnia Dachu, X5 - Całkowita Wysokość, X6 - Orientacja, X7 - Powierzchnia przeszklenia oraz X8 - Rozkład powierzchni przeszklenia.

X1	X2	X3	X4	X5	X6	X7	X8	Y1
0,98	514,5	294	110,25	7	2	0	0	15,55
0,98	514,5	294	110,25	7	3	0	0	15,55
0,98	514,5	294	110,25	7	4	0	0	15,55
0,98	514,5	294	110,25	7	5	0	0	15,55
0,9	563,5	318,5	122,5	7	2	0	0	20,84
0,9	563,5	318,5	122,5	7	3	0	0	21,46
0,9	563,5	318,5	122,5	7	4	0	0	20,71
0,9	563,5	318,5	122,5	7	5	0	0	19,68
0,86	588	294	147	7	2	0	0	19,5
0,86	588	294	147	7	3	0	0	19,95
0,86	588	294	147	7	4	0	0	19,34
0,86	588	294	147	7	5	0	0	18,31
0,82	612,5	318,5	147	7	2	0	0	17,05
0,82	612,5	318,5	147	7	3	0	0	17,41
0,82	612,5	318,5	147	7	4	0	0	16,95
0,82	612,5	318,5	147	7	5	0	0	15,98
0,79	637	343	147	7	2	0	0	28,52
0,79	637	343	147	7	3	0	0	29,9
0,79	637	343	147	7	4	0	0	29,63
0,79	637	343	147	7	5	0	0	28,75
0,76	661,5	416,5	122,5	7	2	0	0	24,77
0,76	661,5	416,5	122,5	7	3	0	0	23,93
0,76	661,5	416,5	122,5	7	4	0	0	24,77
0,76	661,5	416,5	122,5	7	5	0	0	23,93
0,74	686	245	220,5	3,5	2	0	0	6,07
0,74	686	245	220,5	3,5	3	0	0	6,05
0,74	686	245	220,5	3,5	4	0	0	6,01
0,74	686	245	220,5	3,5	5	0	0	6,04
0,71	710,5	269,5	220,5	3,5	2	0	0	6,37
0,71	710,5	269,5	220,5	3,5	3	0	0	6,4
0,71	710,5	269,5	220,5	3,5	4	0	0	6,37
0,71	710,5	269,5	220,5	3,5	5	0	0	6,4
0,69	735	294	220,5	3,5	2	0	0	6,85
0,69	735	294	220,5	3,5	3	0	0	6,79
0,69	735	294	220,5	3,5	4	0	0	6,77
0,69	735	294	220,5	3,5	5	0	0	6,81

Wczytanie danych do Excela z pliku CSV pobranego z Kaggle.

x1	x2	x3	x4	x5	x6	x7	x8	y
0,98	514,5	294	110,25	7	2	0	0	15,55
0,98	514,5	294	110,25	7	3	0	0	15,55
0,98	514,5	294	110,25	7	4	0	0	15,55
0,98	514,5	294	110,25	7	5	0	0	15,55
0,9	563,5	318,5	122,5	7	2	0	0	20,84
0,9	563,5	318,5	122,5	7	3	0	0	21,46
0,9	563,5	318,5	122,5	7	4	0	0	20,71
0,9	563,5	318,5	122,5	7	5	0	0	19,68
0,86	588	294	147	7	2	0	0	19,5
0,86	588	294	147	7	3	0	0	19,95
0,86	588	294	147	7	4	0	0	19,34
0,86	588	294	147	7	5	0	0	18,31
0,82	612,5	318,5	147	7	2	0	0	17,05
0,82	612,5	318,5	147	7	3	0	0	17,41
0,82	612,5	318,5	147	7	4	0	0	16,95
0,82	612,5	318,5	147	7	5	0	0	15,98
0,79	637	343	147	7	2	0	0	28,52
0,79	637	343	147	7	3	0	0	29,9
0,79	637	343	147	7	4	0	0	29,63
0,79	637	343	147	7	5	0	0	28,75
0,76	661,5	416,5	122,5	7	2	0	0	24,77
0,76	661,5	416,5	122,5	7	3	0	0	23,93
0,76	661,5	416,5	122,5	7	4	0	0	24,77
0,76	661,5	416,5	122,5	7	5	0	0	23,93
0,74	686	245	220,5	3,5	2	0	0	6,07
0,74	686	245	220,5	3,5	3	0	0	6,05
0,74	686	245	220,5	3,5	4	0	0	6,01
0,74	686	245	220,5	3,5	5	0	0	6,04
0,71	710,5	269,5	220,5	3,5	2	0	0	6,37
0,71	710,5	269,5	220,5	3,5	3	0	0	6,4
0,71	710,5	269,5	220,5	3,5	4	0	0	6,37
0,71	710,5	269,5	220,5	3,5	5	0	0	6,4
0,69	735	294	220,5	3,5	2	0	0	6,85
0,69	735	294	220,5	3,5	3	0	0	6,79
0,69	735	294	220,5	3,5	4	0	0	6,77
0,69	735	294	220,5	3,5	5	0	0	6,81

Fragment wczytanych danych w Excel

	x1	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*	*y*
x1	1	-0,9919	-0,20378	-0,86882	0,827747	0	1,28E-17	1,76E-17	0,622272
x2	-0,9919	1	0,195502	0,88072	-0,85815	0	1,32E-16	-3,6E-16	-0,65812
x3	-0,20378	0,195502	1	-0,29232	0,280976	0	-8E-19	0	0,455671
x4	-0,86882	0,88072	-0,29232	1	-0,97251	0	-1,4E-16	-1,1E-16	-0,86183
x5	0,827747	-0,85815	0,280976	-0,97251	1	0	1,86E-18	0	0,889431
x6	0	0	0	0	0	1	0	0	-0,00259
x7	1,28E-17	1,32E-16	-8E-19	-1,4E-16	1,86E-18	0	1	0,212964	0,269841
x8	1,76E-17	-3,6E-16	0	-1,1E-16	0	0	0,212964	1	0,087368
y	0,622272	-0,65812	0,455671	-0,86183	0,889431	-0,00259	0,269841	0,087368	1

Korelacja między zmiennymi.

2. Dobór zmiennych do modelu

- Metoda eliminacji Quasi

	0,64	784	343	220,5	3,5	4	0,4	5	18,16
	0,64	784	343	220,5	3,5	5	0,4	5	17,88
	0,62	808,5	367,5	220,5	3,5	2	0,4	5	16,54
	0,62	808,5	367,5	220,5	3,5	3	0,4	5	16,44
	0,62	808,5	367,5	220,5	3,5	4	0,4	5	16,48
	0,62	808,5	367,5	220,5	3,5	5	0,4	5	16,64
Średnia	0,764167	671,7083	318,5	176,6042	5,25	3,5	0,234375	2,8125	22,3072
Odchylenie stand.	0,105709	88,02875	43,59807	45,13654	1,75	1,118034	0,133134	1,54995	10,08362
V	13,83%	13,11%	13,69%	25,56%	33,33%	31,94%	56,80%	55,11%	45,20%
		Eliminacja x1,x2,x3,x4							

Metoda eliminacji Quasi – stałych z obraną wartością krytyczną równą 30%. Wyeliminowane zostały zmienne, których współczynnik zmienności jest mniejszy wartość krytyczna. Są nimi x1, x2, x4 oraz x4.

- Metoda Hellwiga (metodą wskaźników pojemności informacji)

	x5	*x6*	*x7*	*x8*	*y*
x5	1				
x6	0	1			
x7	1,86142E-18	0	1		
x8	0	0	0,212964	1	
y	0,889430674	-0,00259	0,269841	0,087368	1

Macierze: R_0 i R

	0,889430674		1,0000	0	1,861E-18	0
$R_{0=}$	-0,002586534	R=	0	1,0000	0	0
	0,269840996		1,86E-18	0	1,0000	0,212964
	0,087367594		0	0	0,2129642	1,0000

Nr kombinacji C	Zmienne występujące w danej kombinacji	Indywidualne pojemności nośników informacji **h**		Integralne pojemności nośników informacji **H**	
1	**{X5}**	h_{11}	0,791	H_1	**0,791**
2	**{X6}**	h_{22}	0,000	H_2	**0,000**
3	**{X7}**	h_{33}	0,073	H_3	**0,073**
4	**{X8}**	h_{44}	0,008	H4	**0,008**
5	**{X5, X6}**	h_{51}	0,791	**H5**	**0,791**
		h_{52}	0,000		
6	**{X5, X7}**	h_{61}	0,791	H6	**0,86**
		h_{63}	0,073		
7	**{X5, X8}**	h_{71}	0,791	H7	**0,799**
		h_{74}	0,008		
8	**{X6, X7}**	h_{82}	0,000	**H8**	**0,073**
		h_{83}	0,073		
9	**{X6, X8}**	h_{92}	0,000	H9	**0,008**
		h_{94}	0,008		
10	**{X7, X8}**	h_{103}	0,060	H10	**0,066**
		h_{104}	0,006		
11	**{X5, X6, X7}**	h_{111}	0,791	H11	**0,86**
		h_{112}	0,000		
		h_{113}	0,073		
12	**{X5, X7, X8}**	h_{121}	0,791	H12	**0,857**
		h_{123}	0,060		
		h_{124}	0,006		
13	**{X5, X6, X8}**	h_{131}	0,791	H13	**0,799**
		h_{132}	0,000		
		h_{134}	0,008		
14	**{X6, X7, X8}**	h_{142}	0,000	H14	**0,066**
		h_{143}	0,060		
		h_{144}	0,006		
15	**{X5, X6, X7, X8}**	h_{151}	0,791	H15	**0,857**
		h_{152}	0,000		
		h_{153}	0,060		
		h_{154}	0,006		

Wybieramy te zmienne dla których kombinacja daje maksymalną wartość Integralnego wskaźnika pojemności informacyjnej H_I. Kombinacje 6 oraz 11 dają podobny wskaźnik. Zdecydowaliśmy na wybór kombinacji 6 {X5, X7}, ponieważ występują w niej 2 zmienne.

Otrzymany model to: Y = ax5 + bx7 + c

Projekttest.R | najlepsze15 | tab

Filter

	x5	x6	x7	x8	hellwigP
1	0.7910869	6.690158e-06	0.07281416	0.000000000	8.639078e-01
2	0.7910869	0.000000e+00	0.07281416	0.000000000	8.639011e-01
3	0.7910869	6.690158e-06	0.06002993	0.006292928	8.574165e-01
4	0.7910869	0.000000e+00	0.06002993	0.006292928	8.574098e-01
5	0.7910869	6.690158e-06	0.00000000	0.007633096	7.987267e-01
6	0.7910869	0.000000e+00	0.00000000	0.007633096	7.987200e-01
7	0.7910869	6.690158e-06	0.00000000	0.000000000	7.910936e-01
8	0.7910869	0.000000e+00	0.00000000	0.000000000	7.910869e-01
9	0.0000000	6.690158e-06	0.07281416	0.000000000	7.282085e-02
10	0.0000000	0.000000e+00	0.07281416	0.000000000	7.281416e-02
11	0.0000000	6.690158e-06	0.06002993	0.006292928	6.632955e-02
12	0.0000000	0.000000e+00	0.06002993	0.006292928	6.632286e-02
13	0.0000000	6.690158e-06	0.00000000	0.007633096	7.639787e-03
14	0.0000000	0.000000e+00	0.00000000	0.007633096	7.633096e-03
15	0.0000000	6.690158e-06	0.00000000	0.000000000	6.690158e-06

Wyniki uzyskane w R za pomocą skryptu z zajęć

3.Estymacja

Beta	-9,38898685	c
	5,124962798	a
	20,43789945	b

Otrzymanie konkretnych wartości liczbowych i zastąpienie nimi parametrów równania

20,4379	5,124962798	-9,38898685
1,010241	0,076855548	0,486782796
0,863901	3,727295998	#N/D
2427,956	765	#N/D
67461,9	10627,94263	#N/D

Wyniki funkcji REGLINP

PODSUMOWANIE - WYJŚCIE

Statystyki regresji			
Wielokrotność	0,929462795		
R kwadrat	0,863901087	> 0,72	
Dopasowany R	0,863545273		Wyrazistość
Błąd standardo	3,727295998		17%
Obserwacje	768		

ANALIZA WARIANCJI						
	df	*SS*	*MS*	*F*	*Istotność F*	
Regresja	2	67461,89966	33730,95	2427,955958	0	< 0.05
Resztkowy	765	10627,94263	13,89274			
Razem	767	78089,84228				

< 0.05

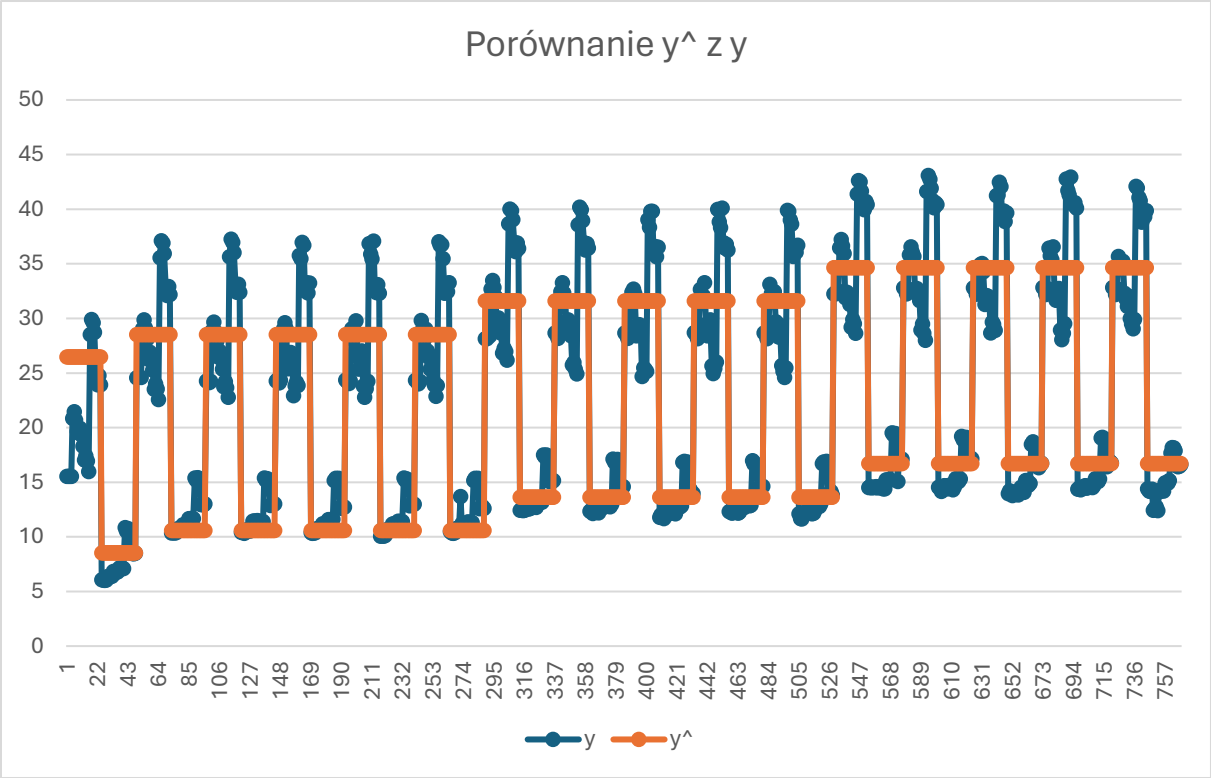
	Współczynniki	*Błąd standardowy*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*	*olne 95,0%*	*órne 95,0%*
Przecięcie	-9,38898685	0,486782796	-19,2878	7,42197E-68	-10,34457547	-8,4334	-10,3446	-8,4334
x5	5,124962798	0,076855548	66,68306	0	4,97408999	5,275836	4,97409	5,275836
x7	20,43789945	1,010240876	20,23072	3,16479E-73	18,45472607	22,42107	18,45473	22,42107

Zastosowanie Dane > Analiza danych > Regresja w Excelu

4. Weryfikacja i testy

t	x5	x7	y	y^	e	e^2		
1	7	0	15,55	26,48575	-10,9358	119,5907	14,22966	51,31617
2	7	0	15,55	26,48575	-10,9358	119,5907	14,22966	51,31617
3	7	0	15,55	26,48575	-10,9358	119,5907	14,22966	51,31617
4	7	0	15,55	26,48575	-10,9358	119,5907	14,22966	51,31617
5	7	0	20,84	26,48575	-5,64575	31,87452	2,303649	51,31617
6	7	0	21,46	26,48575	-5,02575	25,25819	4,570094	51,31617
7	7	0	20,71	26,48575	-5,77575	33,35932	1,925927	51,31617
8	7	0	19,68	26,48575	-6,80575	46,31827	0,128005	51,31617
9	7	0	19,5	26,48575	-6,98575	48,80074	0,031605	51,31617
10	7	0	19,95	26,48575	-6,53575	42,71606	0,394105	51,31617
11	7	0	19,34	26,48575	-7,14575	51,06178	0,000316	51,31617
12	7	0	18,31	26,48575	-8,17575	66,84293	1,024594	51,31617
13	7	0	17,05	26,48575	-9,43575	89,03343	5,162994	51,31617
14	7	0	17,41	26,48575	-9,07575	82,36929	3,656594	51,31617
15	7	0	16,95	26,48575	-9,53575	90,93058	5,627438	51,31617
16	7	0	15,98	26,48575	-10,5058	110,3708	11,17045	51,31617
17	7	0	28,52	26,48575	2,034247	4,138162	84,59912	51,31617
18	7	0	29,9	26,48575	3,414247	11,65708	111,8894	51,31617
19	7	0	29,63	26,48575	3,144247	9,886291	106,2503	51,31617
20	7	0	28,75	26,48575	2,264247	5,126816	88,88299	51,31617
21	7	0	24,77	26,48575	-1,71575	2,943807	29,67828	51,31617
22	7	0	23,93	26,48575	-2,55575	6,531872	21,23162	51,31617
23	7	0	24,77	26,48575	-1,71575	2,943807	29,67828	51,31617
24	7	0	23,93	26,48575	-2,55575	6,531872	21,23162	51,31617
25	3,5	0	6,07	8,548383	-2,47838	6,142382	175,6214	116,0756
26	3,5	0	6,05	8,548383	-2,49838	6,241917	176,1519	116,0756
27	3,5	0	6,01	8,548383	-2,53838	6,443388	177,2153	116,0756
28	3,5	0	6,04	8,548383	-2,50838	6,291985	176,4174	116,0756
29	3,5	0	6,37	8,548383	-2,17838	4,745352	167,7601	116,0756
30	3,5	0	6,4	8,548383	-2,14838	4,615549	166,9838	116,0756
31	3,5	0	6,37	8,548383	-2,17838	4,745352	167,7601	116,0756
32	3,5	0	6,4	8,548383	-2,14838	4,615549	166,9838	116,0756
33	3,5	0	6,85	8,548383	-1,69838	2,884505	155,5563	116,0756
34	3,5	0	6,79	8,548383	-1,75838	3,091911	157,0566	116,0756
35	3,5	0	6,77	8,548383	-1,77838	3,162646	157,5583	116,0756
36	3,5	0	6,81	8,548383	-1,73838	3,021975	156,5557	116,0756
37	3,5	0	7,18	8,548383	-1,36838	1,872472	147,4336	116,0756
38	3,5	0	7,1	8,548383	-1,44838	2,097813	149,3827	116,0756
39	3,5	0	7,1	8,548383	-1,44838	2,097813	149,3827	116,0756
40	3,5	0	7,1	8,548383	-1,44838	2,097813	149,3827	116,0756
41	3,5	0	10,85	8,548383	2,301617	5,297441	71,77855	116,0756
42	3,5	0	10,54	8,548383	1,991617	3,966539	77,12743	116,0756
766	3,5	0,4	16,44	16,72354	-0,28354	0,080396	8,307205	6,753135
767	3,5	0,4	16,48	16,72354	-0,24354	0,059313	8,078227	6,753135
768	3,5	0,4	16,64	16,72354	-0,08354	0,006979	7,194316	6,753135
						10627,94	84932,8	74304,85
						SSE	SST	SSR

Obliczenie miar dopasowania SSE, SST oraz SSR.



- Test T współczynnika korelacji

Test T współczynnika korelacji		
t obl	69,68437	
t	1,96307	

- Błąd oszacowania parametrów

	Se^2=	13,89274	
Błąd oszacowania parametru			
Se * (XTX)^-1	0,236957	-0,03101	-0,2392
	-0,03101	0,005907	6,28E-15
	-0,2392	6,3E-15	1,020587
błędy	0,486783	0,076856	1,010241

S_e określa na ile dany parametr może się zmieniać w różnych badaniach tego samego zjawiska.

- Wyrazistość modelu

Wyrazistość modelu	
V	16,71%

Powinien być mniejszy niż 30%, informuje jaką część średniej wartości zmiennej prognozowanej y stanowi odchylenie standardowe reszt dla danego modelu.

- Współliniowość zmiennych objaśniających (VIF)

VIF								
PODSUMOWANIE - WYJŚCIE								
Statystyki regresji				VIF	1	Brak współliniowości predyjtorów		
Wielokro	1,39E-08							
R kwadra	1,93E-16							
Dopasow	-0,00131							
Błąd stan	1,752283							
Obserwac	768							
ANALIZA WARIANCJI								
	df	*SS*	*MS*	*F*	*Istotność F*			
Regresja	1	4,55E-13	4,55E-13	1,48E-13	1			
Resztkow	766	2352	3,070496					
Razem	767	2352						
	Współczynni	*d standard*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*	*olne 95,0%*	*órne 95,0%*
Przecięcie	5,25	0,128018	41,00978	2E-195	4,998692	5,501308	4,998692	5,501308
x7	7,16E-15	0,474936	1,51E-14	1	-0,93233	0,932331	-0,93233	0,932331

Jest to cecha nie pożądana, ponieważ może prowadzić do zaniżenia wartości statystyki t-Studenta w ocenie istotności parametrów. W naszym przypadku VIF jest równe 1 co oznacza brak współliniowości predyktorów.

- Koincydencja – niezgodność znaków

	x5	*x7*	*y*		
x5	1				
x7	1,86E-18	1			
y	0,889431	0,269840996	1		-9,38899
Brak koincydencji					5,124963
znaki korerlacji i znaki współczynników są takie same					20,4379

- MSE

Średni błąd kwadaratowy		
MSE	13,83847	

- Mierniki stopnia dopasowania modelu do danych

```
> mae <- mean(abs(Y_values$y - Y_values$`y^`))
> print(paste("MAE:", mae))
[1] "MAE: 2.90854633795797"
> # Root Mean Squared Error (RMSE)
> rmse <- sqrt(mean((Y_values$y - Y_values$`y^`)^2))
> print(paste("RMSE:", rmse))
[1] "RMSE: 3.72000899999296"
> # Mean Absolute Percentage Error (MAPE)
> mape <- mean(abs((Y_values$y - Y_values$`y^`) / Y_values$y)) * 100
> print(paste("MAPE:", mape))
[1] "MAPE: 13.0844642658981"
> # Root Mean Squared Percentage Error (RMSPE)
> rmspe <- sqrt(mean(((Y_values$y - Y_values$`y^`) / Y_values$y)^2)) * 100
> print(paste("RMSPE:", rmspe))
[1] "RMSPE: 16.2497060618352"
```

MAE, RMSE: o ile przeciętnie mylimy się, prognozując z modelu (w jednostkach pomiaru zmiennej)

MAPE,RMSPE: o ile procent przeciętnie mylimy się, prognozując z modelu

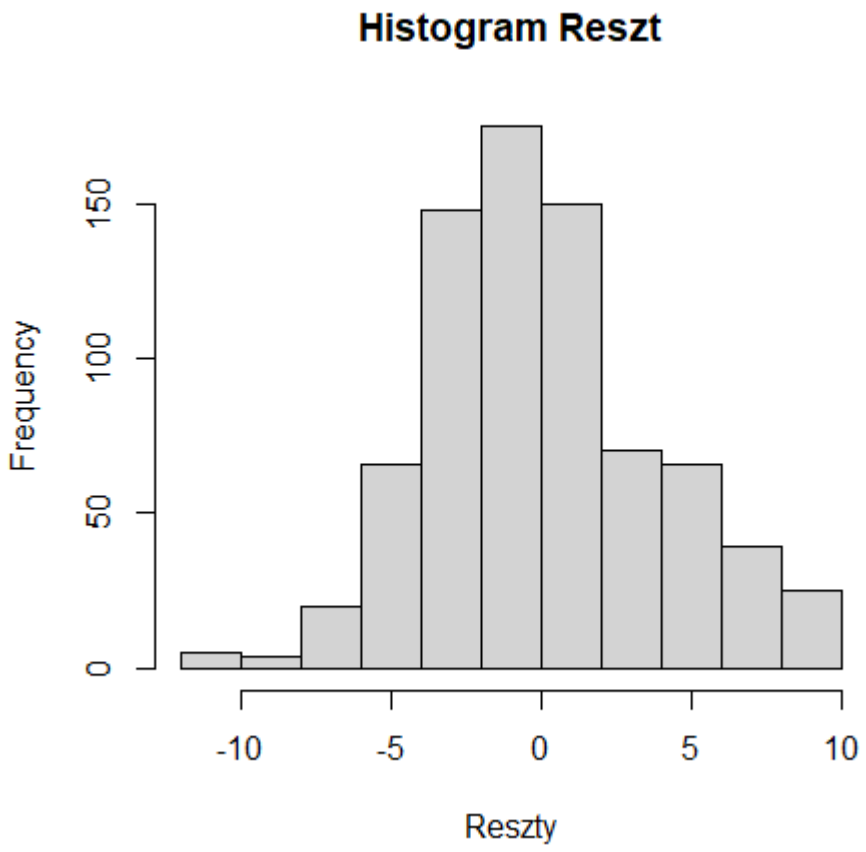
- AIC i AICC

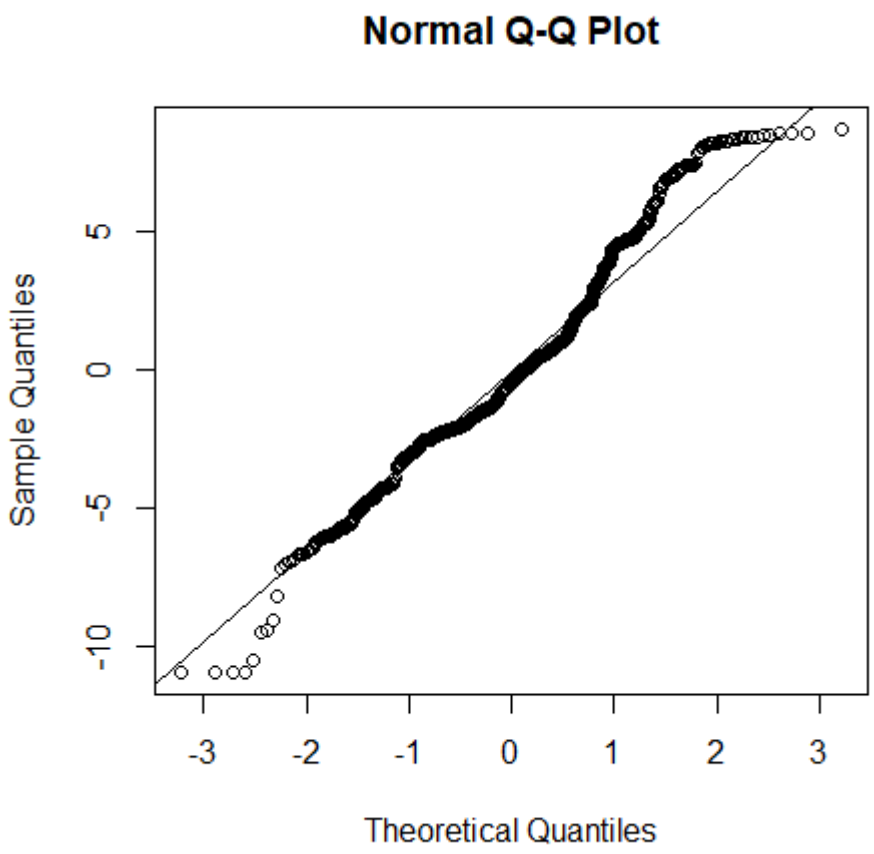
```
> aic_value <- AIC(model)
> aicc_value <- AIC(model) + 2 * (length(coef(model)) + 1) * (nobs(model) / (nobs(model) - length(coef(model)) - 1))
> print(paste("AIC:", aic_value))
[1] "AIC: 4205.37285760984"
> print(paste("AICC:", aicc_value))
[1] "AICC: 4213.41474242659"
```

- Badanie symetrii skł. Losowego

```
# Badanie symetrii skł. losowego

model <- lm(y ~ x5 + x7, data = Dane)
summary(model)
residuals <- residuals(model)
hist(residuals, main = "Histogram Reszt", xlab = "Reszty")
qqnorm(residuals)
qqline(residuals)
shapiro.test(residuals)
```

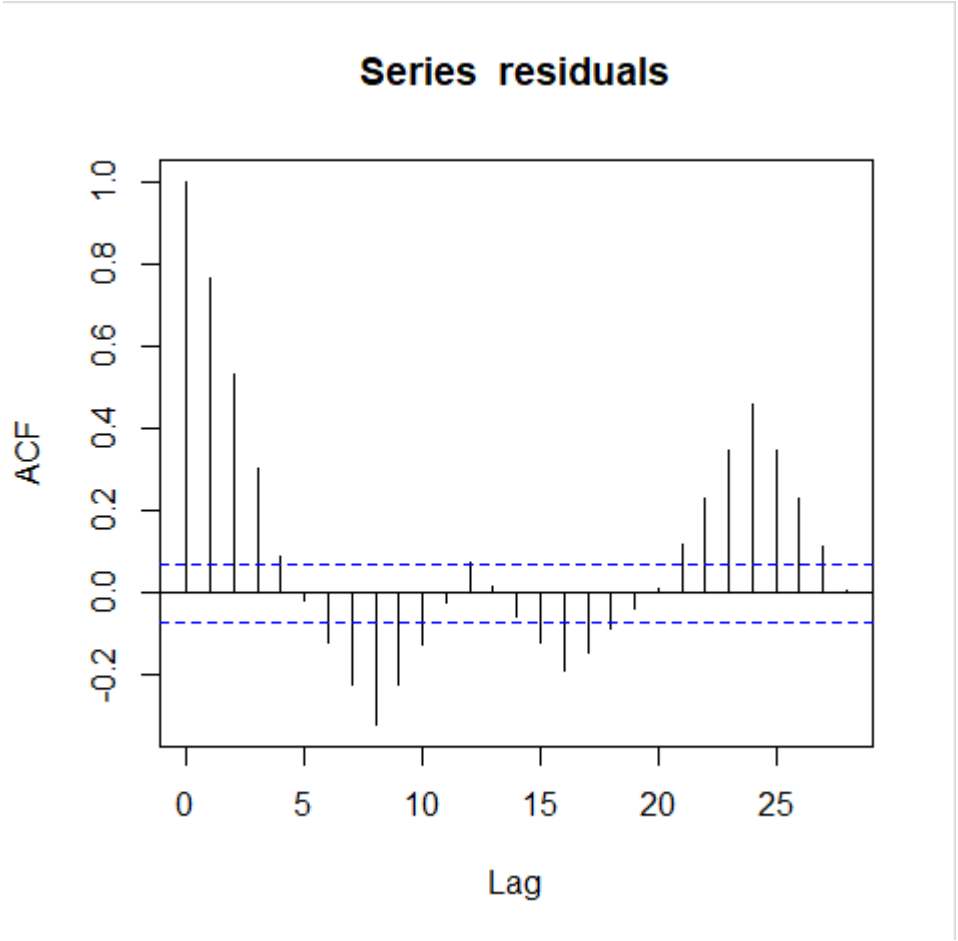
```
> model <- lm(y ~ x5 + x7, data = Dane)
> summary(model)

Call:
lm(formula = y ~ x5 + x7, data = Dane)

Residuals:
     Min       1Q   Median       3Q      Max
-10.9358  -2.2540  -0.4472   2.1403   8.7305

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.38899    0.48678  -19.29   <2e-16 ***
x5           5.12496    0.07686   66.68   <2e-16 ***
x7          20.43790    1.01024   20.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.727 on 765 degrees of freedom
Multiple R-squared:  0.8639,    Adjusted R-squared:  0.8635
F-statistic:  2428 on 2 and 765 DF,  p-value: < 2.2e-16
```


Histogram Reszt
Frequency
0
50
100
150
-10
-5
0
5
10
Reszty


Normal Q-Q Plot
Sample Quantiles
-10
-5
0
5
-3
-2
-1
0
1
2
3
Theoretical Quantiles

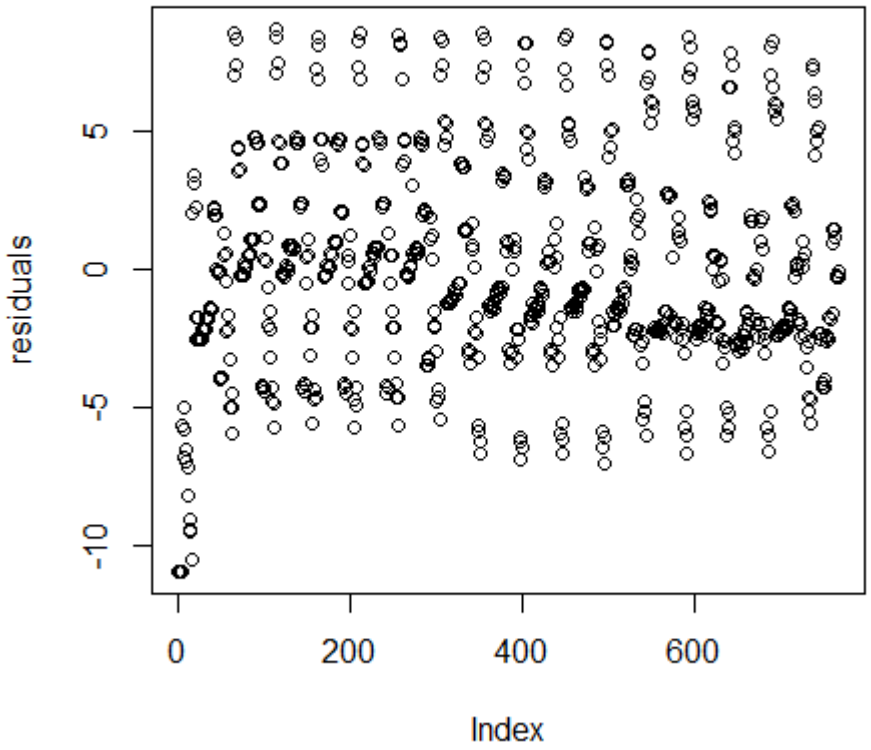
- Badanie losowości skł. Losowego

```
# Badanie losowości skł. losowego

acf(residuals)
Box.test(residuals, lag = 20, type = "Ljung-Box")
plot(residuals)

ks.test(residuals, "pnorm")
```



Na wykresie, ACF jest dodatni dla pierwszych kilku opóźnień. Sugeruje to, że w resztach modelu występuje dodatnia autokorelacja.

```
> Box.test(residuals, lag = 20, type = "Ljung-Box")

        Box-Ljung test

data:  residuals
X-squared = 1001.1, df = 20, p-value < 2.2e-16
```

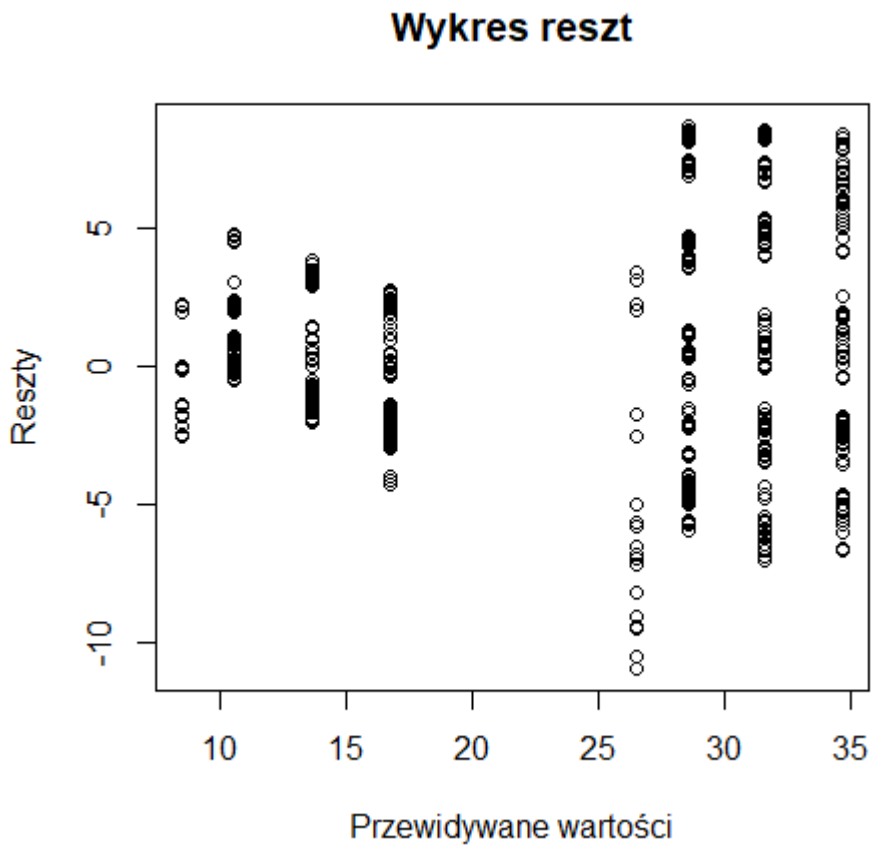


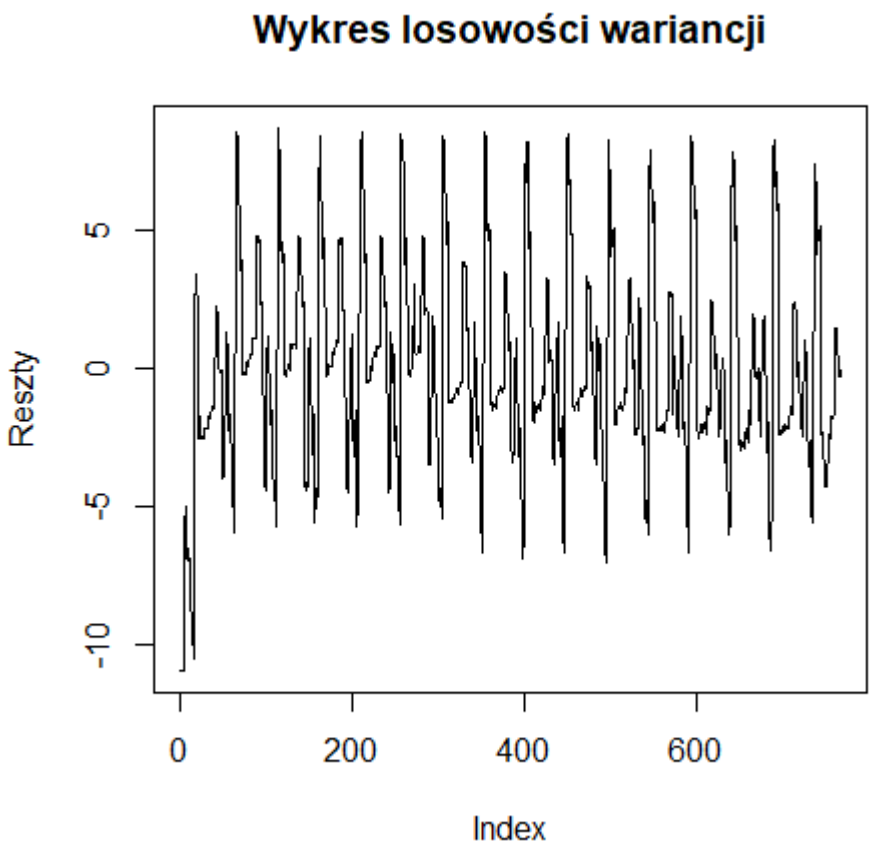

```
> ks.test(residuals, "pnorm")

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  residuals
D = 0.33332, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Wyniki te sugerują, że składniki losowe nie mają rozkładu normalnego.

- Badanie heteroskedastyczności/homoskedastyczności

```
# Badanie heteroskedastyczności/homoskedastyczności
fitted_values <- fitted(model)
plot(fitted_values, residuals, main = "Wykres reszt", xlab = "Przewidywane wartości", ylab = "Reszty")
plot(residuals, type = "l", main = "Wykres losowości wariancji", ylab = "Reszty")

library(lmtest)

# Test Breuscha-Pagana
bptest(model)

# Test Goldfelda-Quandta
gqtest(model)
```


Wykres reszt
Reszty
5
0
-5
-10
10
15
20
25
30
35
Przewidywane wartości




```
> bptest(model)

        studentized Breusch-Pagan test

data:  model
BP = 196.56, df = 2, p-value < 2.2e-16

> # Test Goldfelda-Quandta
> gqtest(model)

        Goldfeld-Quandt test

data:  model
GQ = 0.88644, df1 = 381, df2 = 381, p-value = 0.8801
alternative hypothesis: variance increases from segment 1 to 2
```

- Badanie autokorelacji skł. Losowego

```
# Badanie autokorelacji skł. losowego

acf(residuals)
Box.test(residuals, lag = 20, type = "Ljung-Box")
plot(residuals)
cor(residuals[-length(residuals)], residuals[-1])

library(lmtest)

# Test Breuscha-Godfreya
bgtest(model)

# Statystyka Durbin-Watsona
durbinWatsonTest(model)

# Eliminacja autokorelacji poprzez dodanie lagów do modelu
model_with_lags <- lm(y ~ x5 + x7 + lag(residuals, 1), data = Dane)

#Badanie autokorelacji skł. losowego

# Test Breuscha-Godfreya
bgtest(model_with_lags)

# Statystyka Durbin-Watsona
durbinWatsonTest(model_with_lags)
```

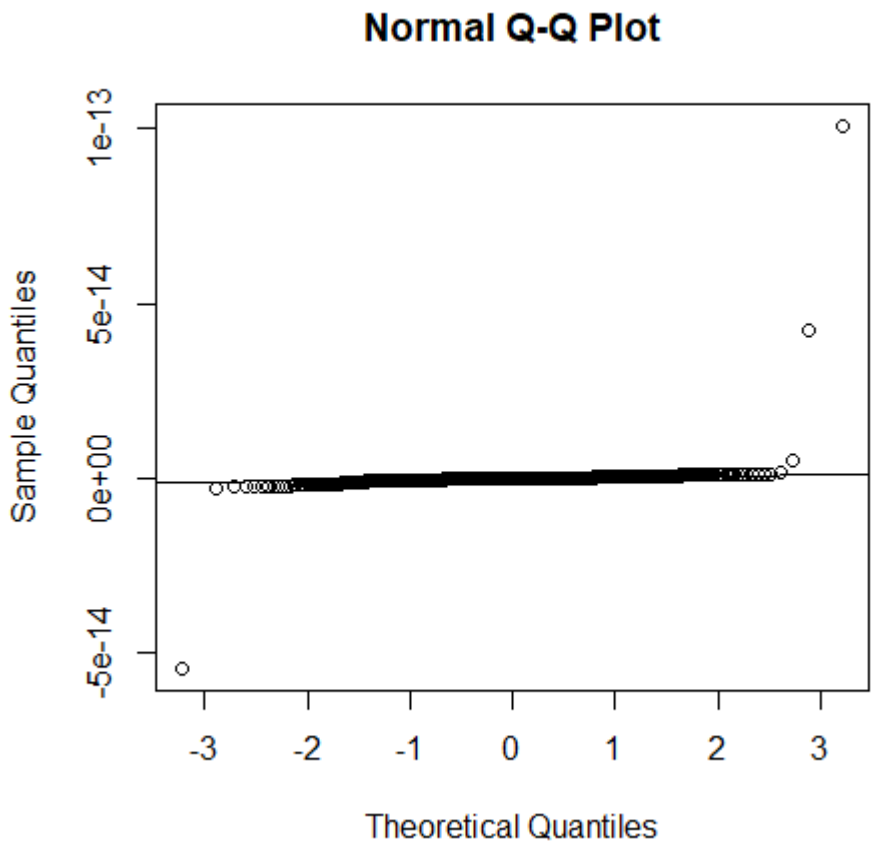
```
> cor(residuals[-length(residuals)], residuals[-1])
[1] 0.7706075

> bgtest(model)

        Breusch-Godfrey test for serial correlation of order up to 1

data:  model
LM test = 451.56, df = 1, p-value < 2.2e-16

> # Statystyka Durbin-Watsona
> durbinWatsonTest(model)
 lag Autocorrelation D-W Statistic p-value
   1       0.7662537     0.4562394       0
 Alternative hypothesis: rho != 0
```

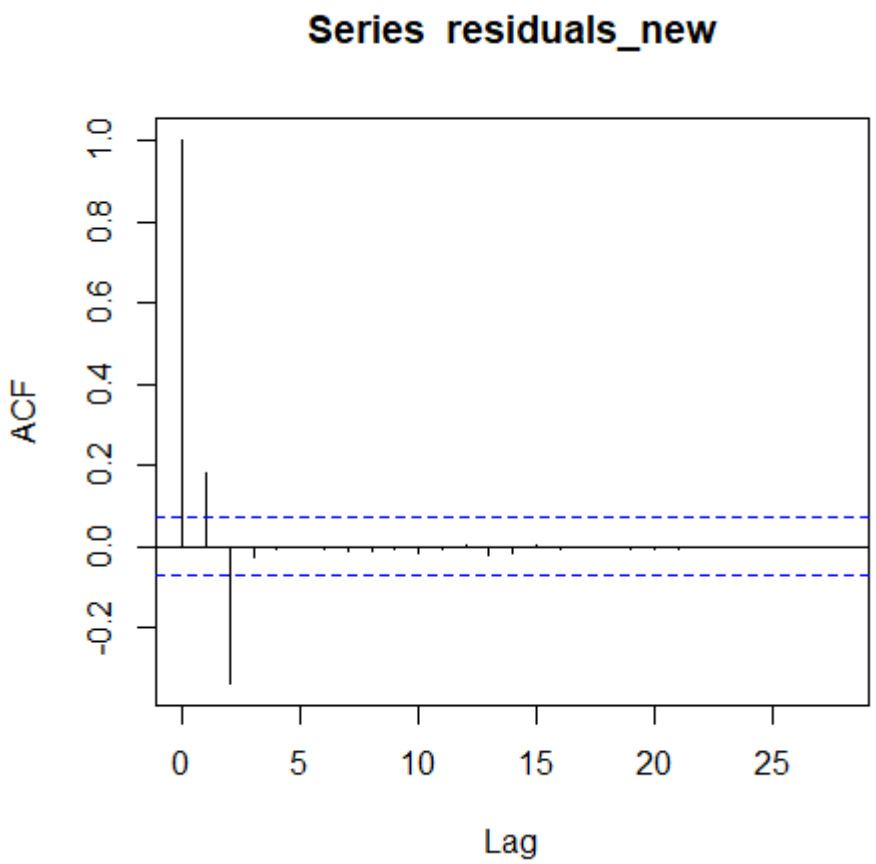
Wyniki testu wskazują na wystąpienie autokorelacji w modelu

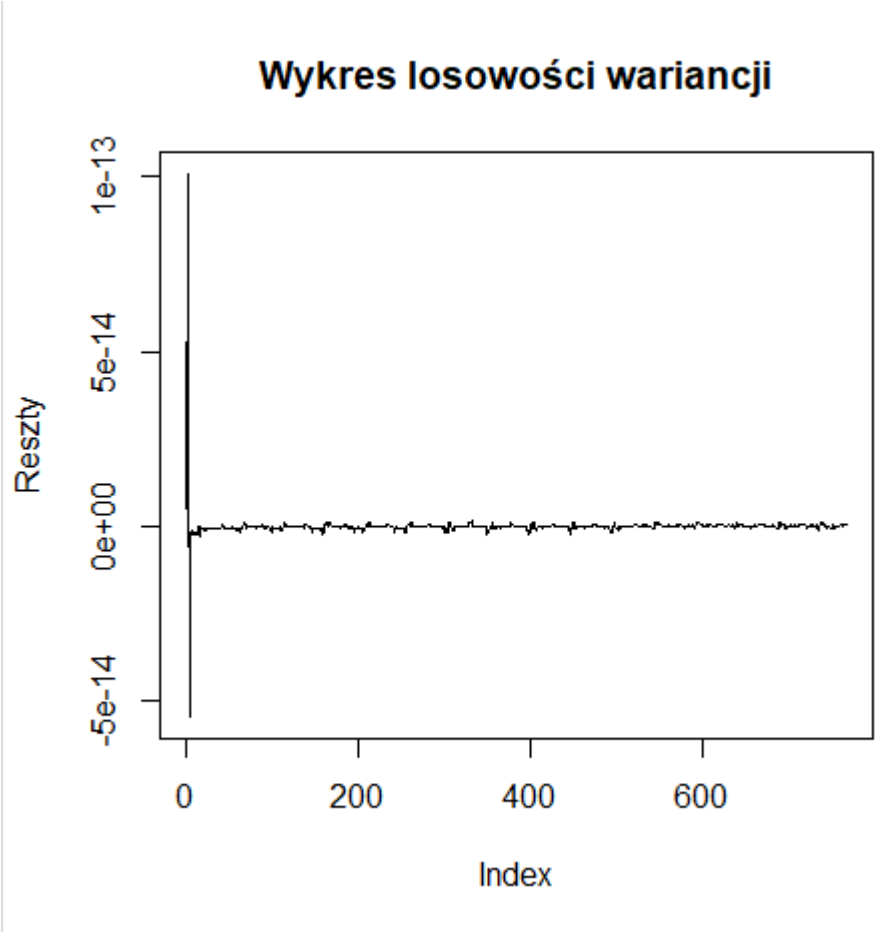
Eliminacja autokorelacji z modelu poprzez zmianę opóźnień

```
> model_with_lags <- lm(y ~ x5 + x7 + lag(residuals, 1), data = Dane)
> # Test Breuscha-Godfreya
> bgtest(model_with_lags)

        Breusch-Godfrey test for serial correlation of order up to 1

data:  model_with_lags
LM test = 24.947, df = 1, p-value = 5.894e-07

> # Statystyka Durbin-Watsona
> durbinWatsonTest(model_with_lags)
 lag Autocorrelation D-W Statistic p-value
   1       0.1800225      1.638365     0.3
 Alternative hypothesis: rho != 0
```

Wykresy dla reszt po eliminacji autokorelacji:


Normal Q-Q Plot
Sample Quantiles
1e-13
5e-14
0e+00
-5e-14
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Series residuals_new
ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0
5
10
15
20
25
Lag



- Badanie normalności skł. Losowego

```r
# Badanie normalności skł. losowego
result <- shapiro.test(residuals)
result
p_value <- result$p.value

if (p_value < 0.05) {

  cat("Odrzucamy hipotezę zerową - dane nie pochodzą z rozkładu normalnego.")

} else {

  cat("Nie ma podstaw do odrzucenia hipotezy zerowej - dane mogą pochodzić z rozkładu normalnego."

}
```

```
> # Badanie normalności skł. losowego
> result <- shapiro.test(residuals)
> result

        Shapiro-Wilk normality test

data:  residuals
W = 0.97469, p-value = 2.876e-10

> p_value <- result$p.value
> if (p_value < 0.05) {
+
+   cat("Odrzucamy hipotezę zerową - dane nie pochodzą z rozkładu normalnego.")
+
+ } else {
+
+   cat("Nie ma podstaw do odrzucenia hipotezy zerowej - dane mogą pochodzić z rozkładu normalnego.")
+
+ }
Odrzucamy hipotezę zerową - dane nie pochodzą z rozkładu normalnego.
```

Dane nie pochodzą z rozkładu normalnego

5. Wnioski

Na podstawie danych z Energy Efficiency Dataset, przeprowadziliśmy analizę energetyczną, koncentrując się na 12 różnych kształtach budynków symulowanych w programie Ecotect. Zbiór danych obejmuje 768 próbek, z których każda opisuje unikalne kombinacje cech. Zmienna objaśniające to: X1 - Względna zwartość, X2 - Powierzchnia, X3 - Powierzchnia Ścian, X4 - Powierzchnia Dachu, X5 - Całkowita Wysokość, X6 - Orientacja, X7 - Powierzchnia przeszklenia oraz X8 - Rozkład powierzchni przeszklenia.

Współczynniki regresji dla zmiennych objaśniających x5 i x7 są statystycznie istotne, co sugeruje, że obie te zmienne mają istotny wpływ na zmienną zależną y. Reszty modelu wydają się być dobrze rozłożone wokół zera na wykresach rozrzutu reszt. Brak widocznych wzorców w resztach może świadczyć o tym, że model jest adekwatny. Wnioski te sugerują, że zbudowany model regresji może być użyteczny do prognozowania zmiennych zależnych na podstawie wartości zmiennych objaśniających.

Dzięki tym wynikom możemy wnioskować, że efektywność energetyczna budynków może być istotnie związana z ich całkowitą wysokością i powierzchnią przeszklenia. To z kolei może być przydatne dla projektantów i architektów, ponieważ mogą oni dostosować te cechy w celu poprawy efektywności energetycznej planowanych budynków.

Analiza dodatkowych zmiennych lub zastosowanie różnych modeli mogą wpłynąć na lepsze zrozumienie związków między cechami budynków a ich efektywnością energetyczną.