

UNIwersytet Ekonomiczny w Katowicach

Przedmiot:

Modele regresyjne

Temat:

Projekt zaliczeniowy – Modele regresyjne

Prowadzący:

dr Agnieszka Orwat - Acedańska

Anna Krzyżowska – 139503

Informatyka i Ekonometria,

Analityka Danych,

Rok 2, semestr 4

Spis treści

Regresja wieloraka	3
Regresja krokowa.....	7
Postępująca.....	7
Wsteczna	11
Regresje krzywoliniowe	14
Regresja logistyczna.....	17
Modele szeregów czasowych z analizą trendu	20
Diagnozowanie obserwacji odstających.....	28
Spis ilustracji	32
Bibliografia.....	33

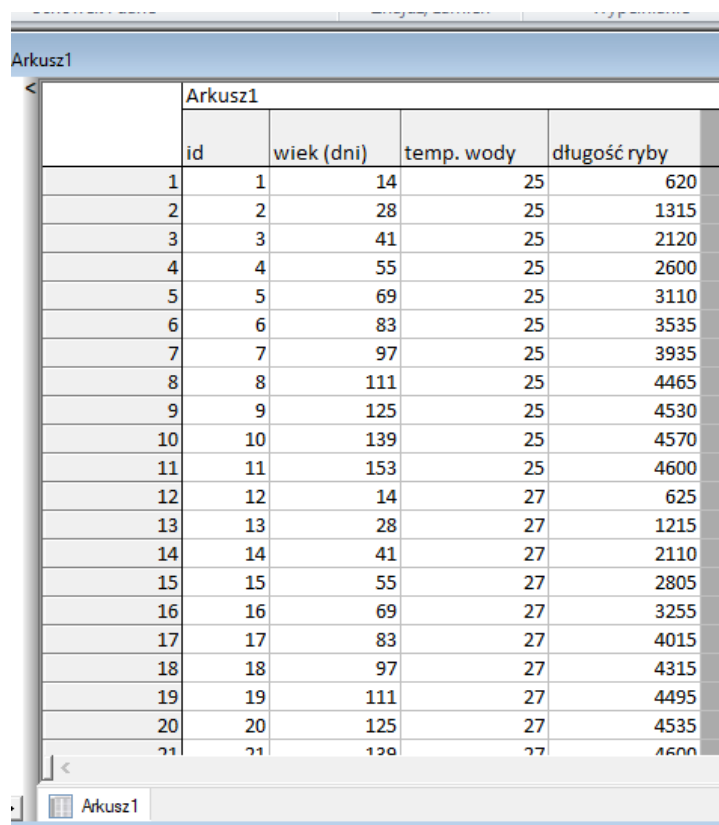
Regresja wieloraka

Opis danych

Bazę danych pobrałam z następującej strony:

<https://people.sc.fsu.edu/~jburkardt/datasets/regression/regression.html>

Długość danego gatunku ryb ma być reprezentowana jako funkcja wieku i temperatury wody. Ryby trzymane są w zbiornikach przy 25, 27, 29 i 31 stopniach Celsjusza. Po urodzeniu, próbka testowa jest wybierana losowo co 14 dni, a jego długość mierzona w [mm].



	id	wiek (dni)	temp. wody	długość ryby
1	1	14	25	620
2	2	28	25	1315
3	3	41	25	2120
4	4	55	25	2600
5	5	69	25	3110
6	6	83	25	3535
7	7	97	25	3935
8	8	111	25	4465
9	9	125	25	4530
10	10	139	25	4570
11	11	153	25	4600
12	12	14	27	625
13	13	28	27	1215
14	14	41	27	2110
15	15	55	27	2805
16	16	69	27	3255
17	17	83	27	4015
18	18	97	27	4315
19	19	111	27	4495
20	20	125	27	4535
21	21	139	27	4500

Rysunek 1 Fragment danych

Podsumowanie regresji zmiennej zależnej: długość ryby (Arkusz1 w dane do stat)						
R= ,89757907 R^2= ,80564818 Popraw. R2= ,79616761						
F(2,41)=84,979 p<,00000 Błąd std. estymacji: 600,00						
N=44	b*	Bł. std. z b*	b	Bł. std. z b	t(41)	p
W. wolny			3904,266	1149,044	3,39784	0,001522
wiek (dni)	0,879116	0,068850	26,241	2,055	12,76861	0,000000
temp. wody	-0,181118	0,068850	-106,414	40,452	-2,63063	0,011951

Rysunek 2 Podsumowanie regresji

Istotność regresji liniowej: Wartość $F=84,979$, $p=0,00$, czyli równanie regresji jest istotne. Współczynnik korelacji wynosi 0,90 i oznacza, że między zmiennymi istnieje zależność.

Dopasowanie modelu: $R^2=80\%$

Istotność częściowych współczynników regresji. Wszystkie zmienne są istotne statystycznie.

Długość ryby = $3904,266 + 26,241 \cdot \text{wiek} - 106,414 \cdot \text{temperatura}$

Wraz ze wzrostem starzenia się ryby o 1 dzień, przewidywana długość ryby wzrośnie o 26,24 mm przy założeniu, że pozostałe parametry nie ulegną zmianie;

Wraz ze wzrostem temperatury wody o 1 stopień Celsjusza, przewidywana długość ryby zmaleje o 106,41 mm przy założeniu, że pozostałe parametry nie ulegną zmianie.

Zmienna	Nadmiarowość zmiennych niezależnych; DV: długość ryby (Arkusz1 w dane do stat) kolumna R-kwadr. zawiera R-kwadrat odpowiedniej zmiennej ze wszystkimi innymi zmiennymi niezależnymi						
	Toleran.	R-kwadr.	Częstk. Korelac.	Semicz. Korelac.			
wiek (dni)	1,000000	0,00	0,893900	0,879116			
temp. wody	1,000000	0,00	-0,380014	-0,181118			

Rysunek 3 Nadmiarowość zmiennych niezależnych

Brak współliniowości (nadmiarowości) między zmiennymi niezależnymi. Tolerancja dla obu zmiennych jest bardzo wysoka=1, współczynnik przy R^2 jest bardzo niski, co świadczy o tym, że brak jest współliniowości między zmiennymi.

Autokorelacja

	d Durbin-Watsona (Arkusz1 w dane do stat) i korelacja seryjna reszt				
	d Durbin Watsona	Seryjna Kor.			
Estymac.	0,285132	0,874295			

Rysunek 4 Test Durbin-Watsona

Test Durbin-Watsona – badanie autokorelacji składnika losowego modelu.

$H_0: \rho_1=0$

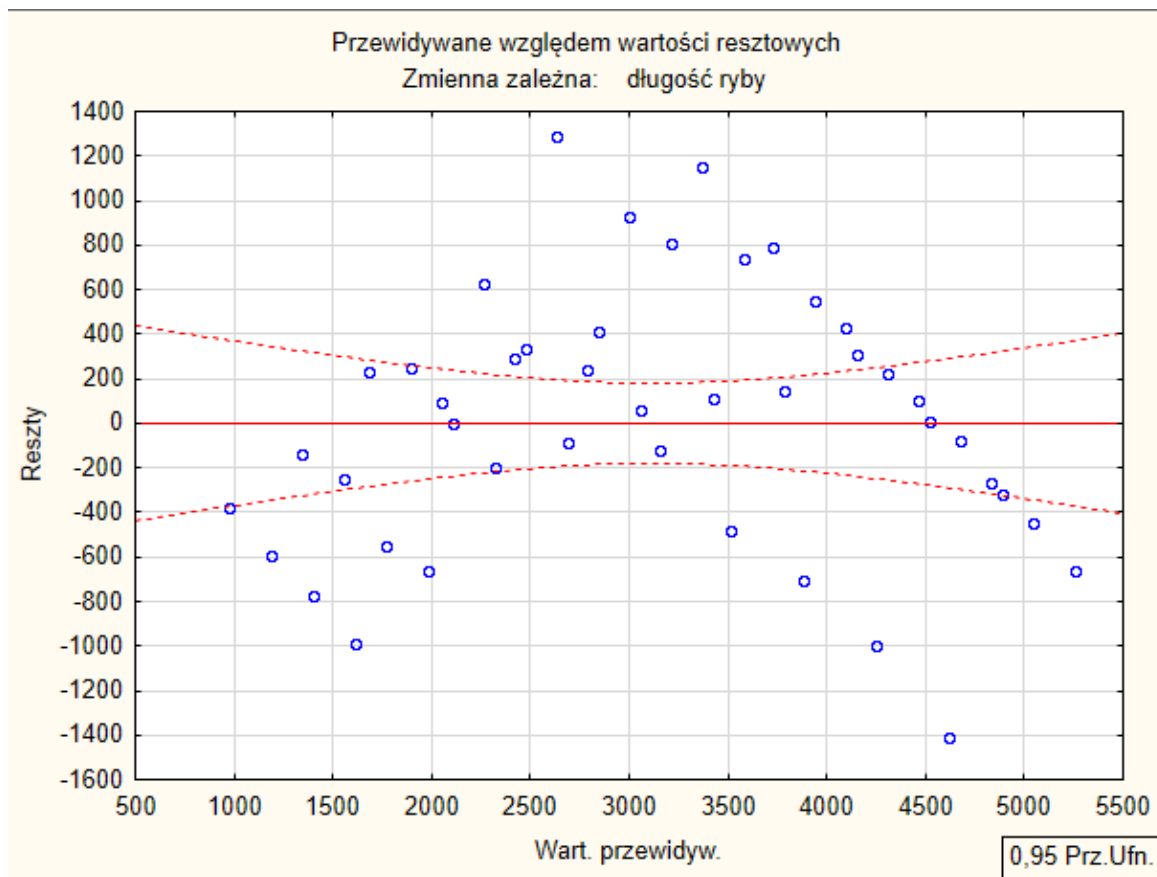
$H_1: \rho_1>0$

$dL = 1,4226$

$$dU = 1,6120$$

$d < dl$, występuje autokorelacja dodatnia, ponieważ nie zostały uwzględnione czynniki cykliczne.

Homoskedastyczność



Rysunek 5 Przewidywane względem wartości resztowych

Test White'a

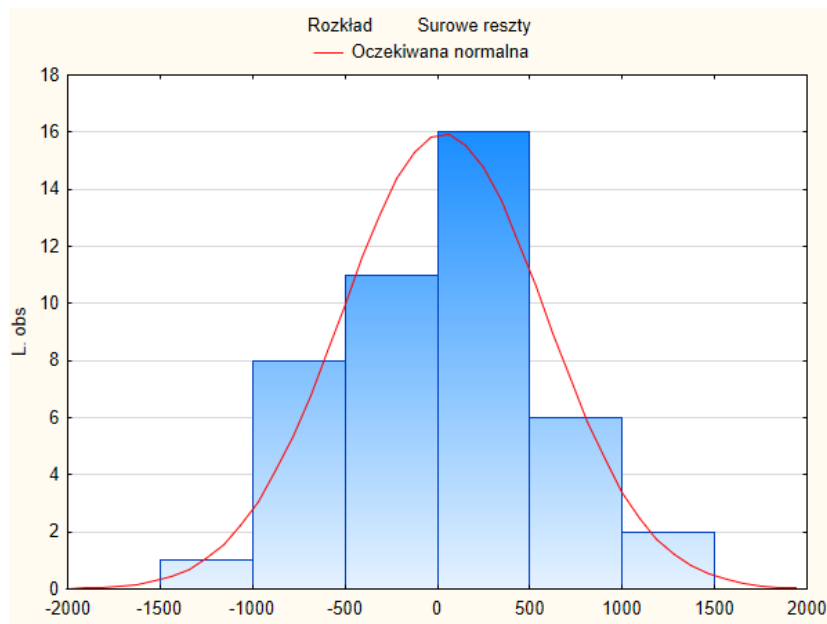
$$H_0: B_k = 0;$$

$$H_1: B_k \neq 0$$

$$LM = TR^2 = 10,767583, \text{ Krytyczna wart. } = 5,99146$$

$LM > \text{wartości krytycznej}$, odrzucamy H_0 , wariancja składnika losowego jest niejednorodna, wariancja reszt nie jest stała w czasie.

Normalność



Rysunek 6 Histogram

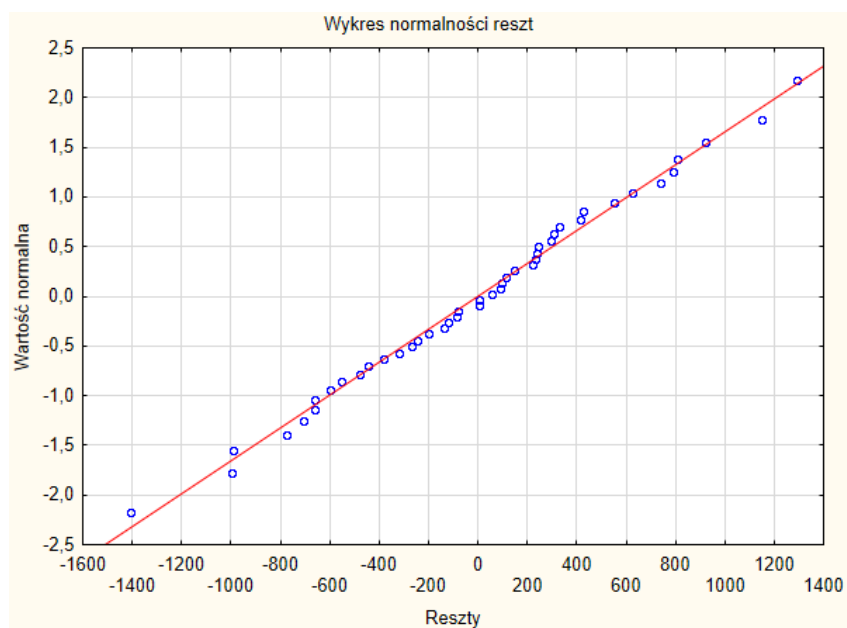
Test Jarque'a-Bera

H0: składnik losowy ma rozkład normalny

H1: składnik losowy nie ma rozkładu normalnego

JB= 0,0708305, Krytyczna wart. = 5,99146

JB<wartości krytycznej, składnik losowy modelu ma rozkład normalny.



Rysunek 7 Wykres normalności reszt

Regresja krokowa

Dane zostały użyte te same co do regresji wielorakiej.

Postępująca

Regresja krokowa postępująca zakłada kolejne dołączanie do listy zmiennych objaśniających, uwzględnionych w modelu tych zmiennych, które mają najistotniejszy wpływ na zmienną zależną.

Podsumowanie regresji zmiennej zależnej: długość ryby (Arkusz1 w dane do stat) R= ,89757907 R^2= ,80564818 Popraw. R2= ,79616761 F(2,41)=84,979 p<,00000 Błąd std. estymacji: 600,00						
N=44	b*	Bł. std. z b*	b	Bł. std. z b	t(41)	p
W. wolny			3904,266	1149,044	3,39784	0,001522
wiek (dni)	0,879116	0,068850	26,241	2,055	12,76861	0,000000
temp. wody	-0,181118	0,068850	-106,414	40,452	-2,63063	0,011951

Rysunek 8 Podsumowanie regresji zmiennej zależnej

Długość ryby = $3904,266 + 26,241 \cdot \text{wiek} - 106,414 \cdot \text{temperatura}$

statystyka	Stat. podsum.; Zmn. Wartość
R wielorakie	0,897579068
Wielorakie R2	0,805648183
Skorygowane R2	0,796167607
F(2,41)	84,9788185
p	2,60655077E-15
Błąd std. estymacji	599,997517

Rysunek 9 Statystyki podsumowujące

Istotność regresji linowej Test F =84,979, p<0,000, czyli liniowość jest istotna.

Dopasowanie modelu: $R^2=80\%$

Istotność częściowych współczynników regresji Dla zmiennych wiek i temperatura współczynniki są istotne (p<0,05).

Brak nadmiarowości między zmiennymi niezależnymi

Zmienna	Nadmiarowość zmiennych niezależnych; DV: długość ryby (Arkusz1 w rybki) kolumna R-kwadr. zawiera R-kwadrat odpowiedniej zmiennej ze wszystkimi innymi zmiennymi niezależnymi					
	Toleran.	R-kwadr.	Cząstk. Korelac.	Semicz. Korelac.		
wiek (dni)	1,000000	0,00	0,893900	0,879116		
temp. wody	1,000000	0,00	-0,380014	-0,181118		

Rysunek 10 Nadmiarowość zmiennych niezależnych

Należy sprawdzić, czy żadna ze zmiennych niezależnych nie jest kombinacją liniową innych zmiennych niezależnych, czyli czy brak jest współliniowości. Tolerancja dla obu zmiennych jest bardzo wysoka=1, współczynnik przy R^2 jest bardzo niski, co świadczy o tym, że brak jest współliniowości między zmiennymi.

Autokorelacja

	d Durbina-Watsona (Arkusz1 w rybki) i korelacja seryjna reszt		
	d Durbin Watsona	Seryjna Kor.	
Estymac.	0,261213	0,930625	

Rysunek 11 Test Durbina Watsona

Test Durbina-Watsona – badanie autokorelacji składnika losowego modelu.

$H_0: \rho_1 = 0$

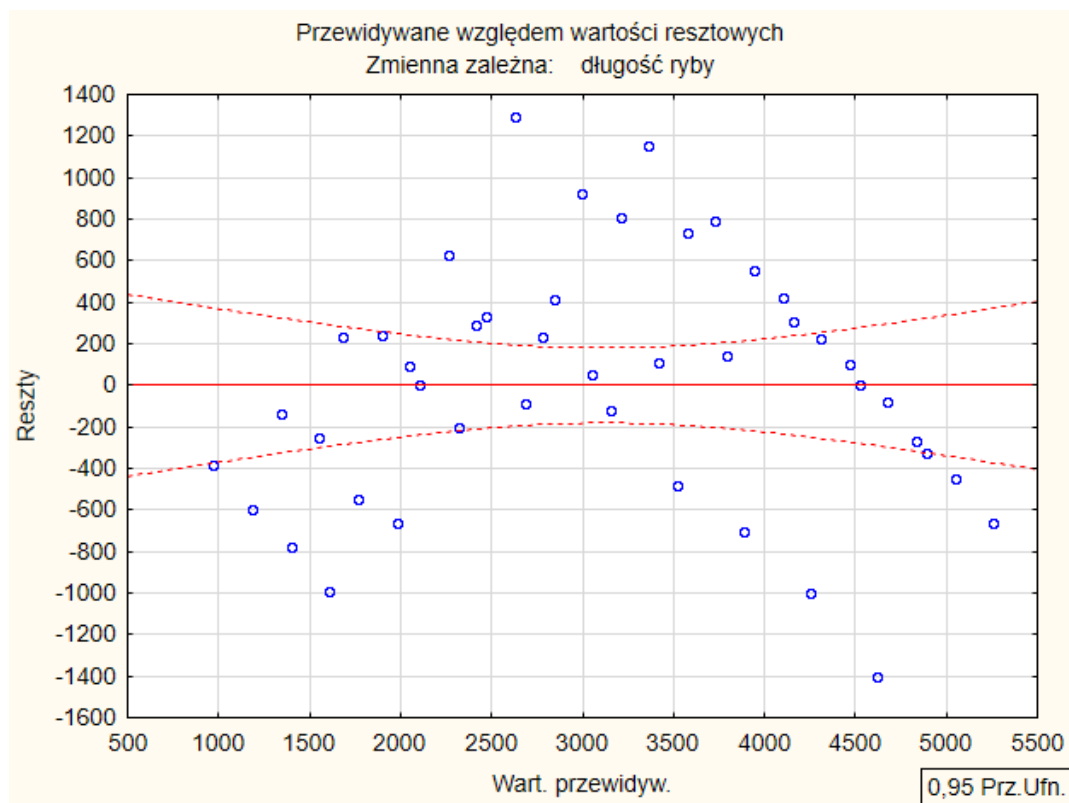
$H_1: \rho_1 > 0$

$d_L = 1,469$

$d_U = 1,56$

$d < d_L$, występuje autokorelacja dodatnia, ponieważ nadal nie zostały uwzględnione czynniki cykliczne.

Homoskedastyczność



Rysunek 12 Przewidywane względem wartości resztowych

Test White'a

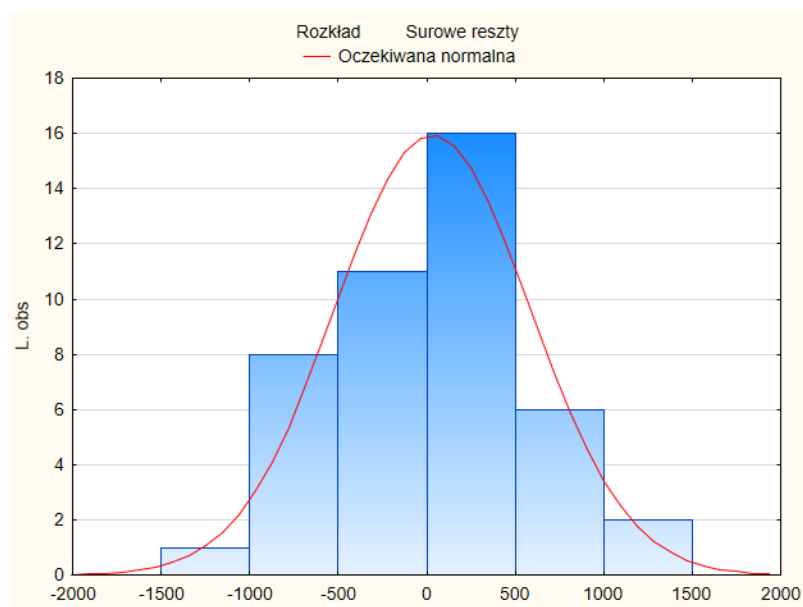
$H_0: B_k = 0;$

$H_1: B_k \neq 0$

$LM = TR^2 = 10,767583$, Krytyczna wart. = 5,99146

Wariancja składnika losowego jest niejednorodna, wariancja reszt nie jest stała w czasie.

Normalność



Rysunek 13 Histogram

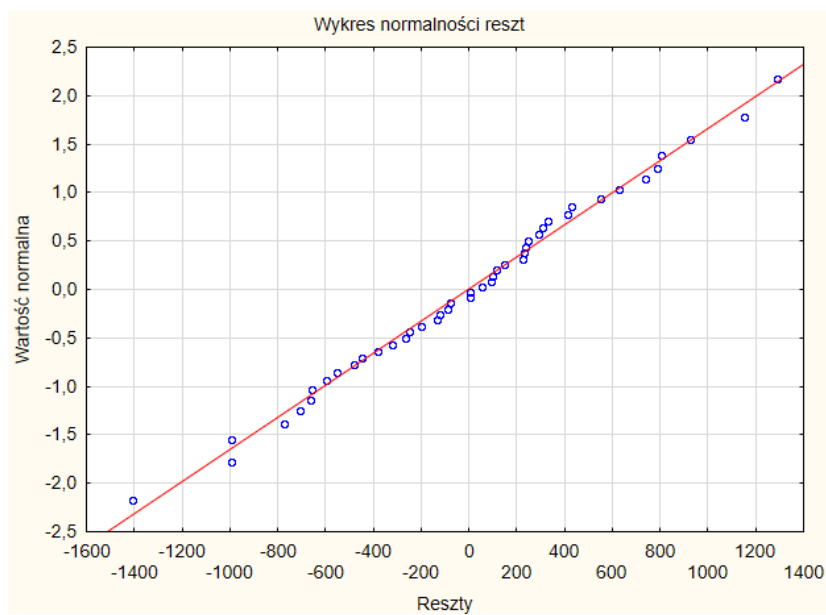
Test Jarque'a-Bera

H_0 : składnik losowy ma rozkład normalny

H_1 : składnik losowy nie ma rozkładu normalnego

$JB = 0,0708305$, Krytyczna wart. = 5,99146

Składnik losowy modelu ma rozkład normalny.



Rysunek 14 Wykres normalności reszt

Wsteczna

Regresja krokowa wsteczna zakłada kolejne usuwanie z modelu zbudowanego ze wszystkich potencjalnych zmiennych tych spośród nich, które w danym kroku mają najmniej istotny wpływ na zmienną zależną, aż do uzyskania „najlepszego” modelu.

Podsumowanie regresji zmiennej zależnej: długość ryby (Arkusz1 w dane do stat)						
R= ,87911573 R^2= ,77284446 Popraw. R2= ,76743600						
F(1,42)=142,90 p<,00000 Błąd std. estymacji: 640,89						
N=44	b*	Bł. std. z b*	b	Bł. std. z b	t(42)	p
W. wolny			924,6842	206,5837	4,47608	0,000057
wiek (dni)	0,879116	0,073542	26,2407	2,1952	11,95388	0,000000

Rysunek 15 Podsumowanie regresji zmiennej zależnej

Jak widać, na powyższym screenie, Statistica nie wzięła do modelu temperatury. Model prezentuje się teraz tak:

$$\text{Długość ryby} = 924,6842 + 26,24 \cdot \text{wiek}$$

Istotność regresji liniowej Test F=142,9, p<,0000, liniowość jest istotna statystycznie.

Dopasowanie modelu: R^2=77%

Istotność cząstkowych współczynników regresji Dla zmiennej wiek wartość p<0,05; współczynnik jest istotny.

statystyka	Stat. podsum.; Zmn. zal.: długość ryby (Arkusz1 w dane do stat)				
	Wartość				
R wielorakie	0,879115728				
Wielorakie R2	0,772844464				
Skorygowane R2	0,767435999				
F(1,42)	142,89534				
p	4,20356575E-15				
Błąd std. estymacji	640,890888				

Rysunek 16 Statystyki podsumowujące

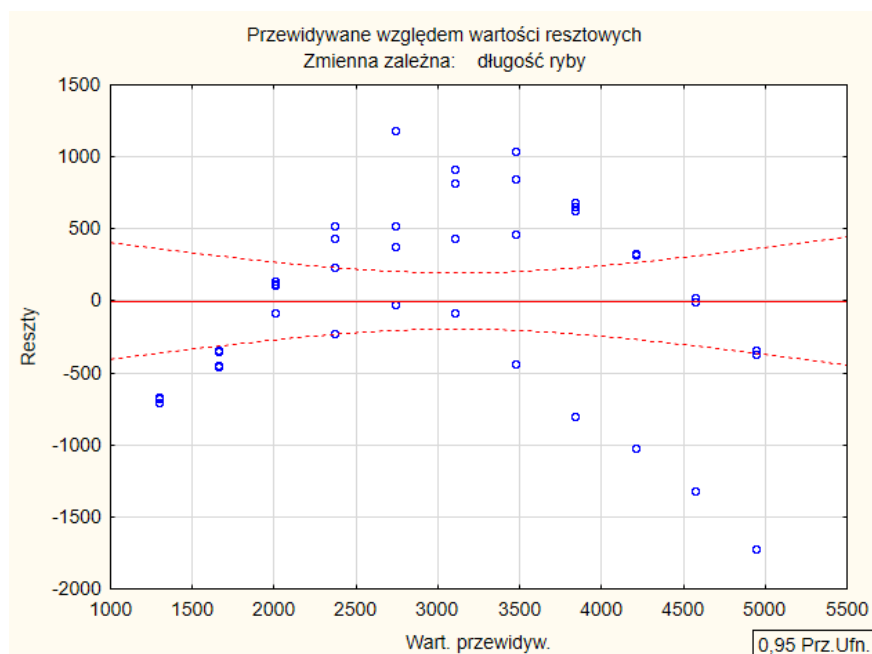
Brak nadmiarowości między zmiennymi niezależnymi

Nadmiarowość zmiennych niezależnych; DV: długość ryby (Arkusz1 w rybki)					
kolumna R-kwadr. zawiera R-kwadrat odpowiedniej zmiennej ze wszystkimi innymi zmiennymi niezależnymi					
Zmienna	Toleran.	R-kwadr.	Cząstk. Korelac.	Semicz. Korelac.	
wiek (dni)	1,000000	0,00	0,879116	0,879116	
temp. wody	1,000000	0,00	-0,380014	-0,181118	

Rysunek 17 Nadmiarowość zmiennych niezależnych

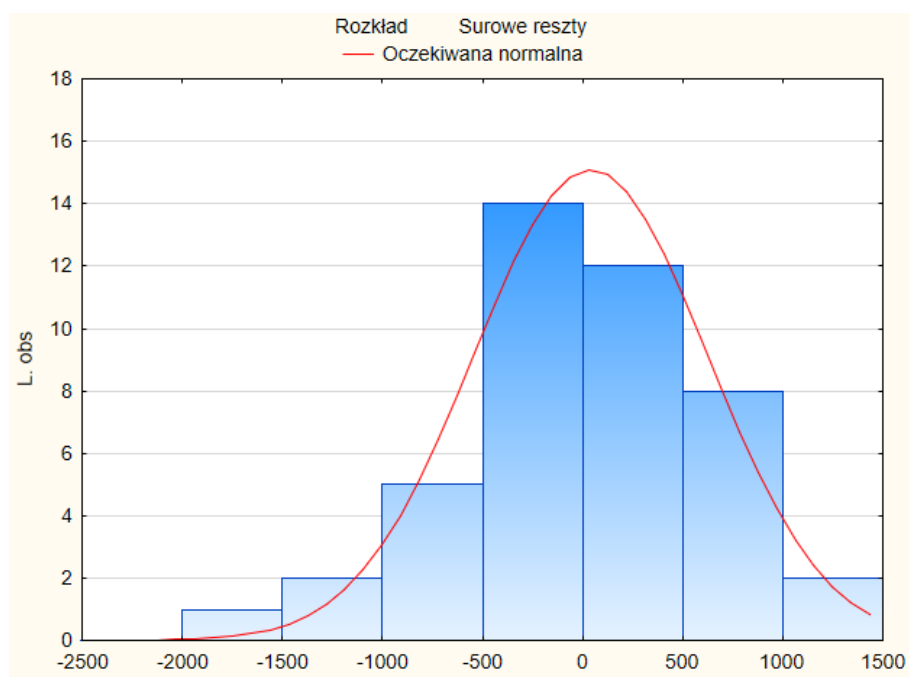
Korelacje smicząsteczkowe są bardzo małe dla temperatury i możliwe, że to było powodem odrzucenia zmiennej w tej korelacji. Świadczy to o słabej korelacji ze zmianą zależną.

Założenie homoskedastyczności



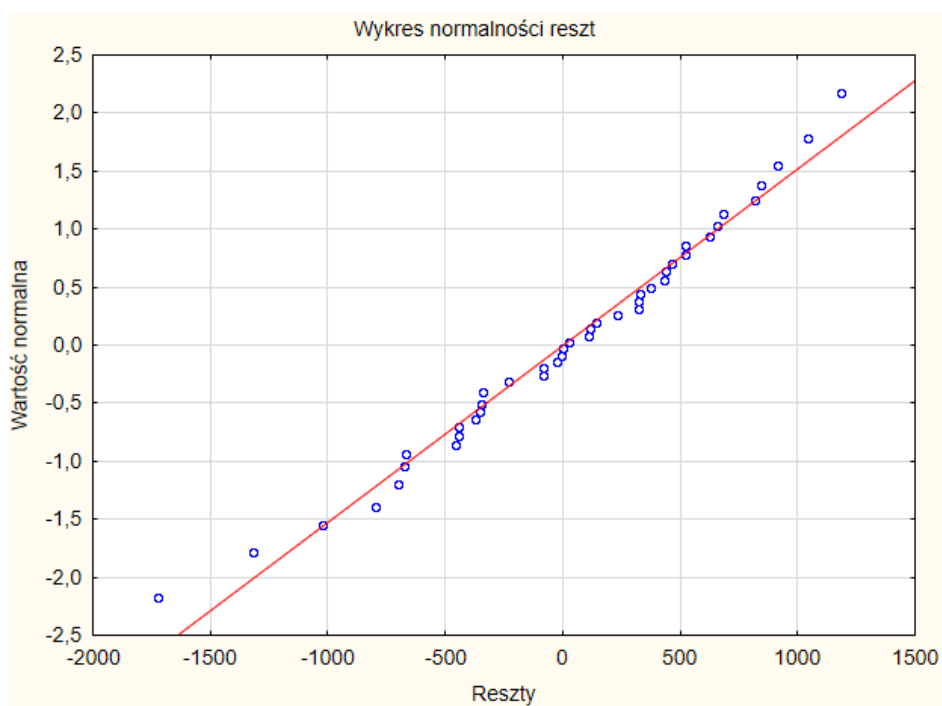
Rysunek 18 Przewidywane względem wartości resztowych

Na podstawie wykresu widać że, wariancja składnika losowego jest niejednorodna, wariancja reszt nie jest stała w czasie.



Rysunek 19 Histogram

Na podstawie histogramu można zauważyć, że składnik losowy modelu ma rozkład normalny.



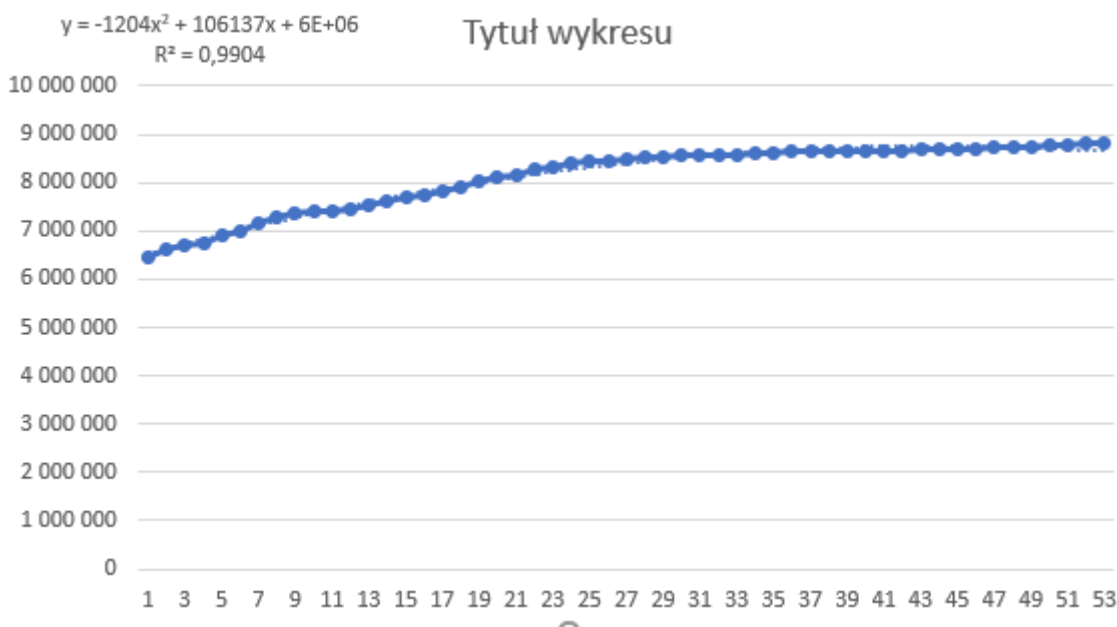
Rysunek 20 Wykres normalności reszt

Punkty oprócz początkowych i końcowych obserwacji przylegają do wykresu, co jest potwierdzeniem normalności rozkładu.

Regresje krzywoliniowe

Bank Danych Polska (BDP) jest bazą gromadzącą historyczne dane dotyczące kraju ogółem pochodzące z systemu polskiej statystyki publicznej. Szeregi czasowe, w zależności od wybranej kategorii, rozpoczynają się w 1946 r. a kończą w 1999 r. Informacje są prezentowane w postaci tablic i wykresów, które mogą być eksportowane do powszechnie używanych formatów. Dane dotyczą zmieniającej się powierzchni Lasów Skarbu Państwa na przestrzeni lat 1946, a 1999. Dane zostały pobrane ze strony: <https://bdl.stat.gov.pl/>

Zastosowałam regresję wielomianową, funkcja trendu prezentuje się następująco:



Rysunek 21 Funkcja trendu wielomianowa

Podsumowanie regresji zmiennej zależnej: powierzchnia (Arkusz1 w regresja wielomian)							
R= ,99520830 R^2= ,99043955 Popraw. R2= ,99005713							
F(2,50)=2589,9 p<0,0000 Błąd std. estymacji: 68536,							
N=53	b*	Bł. std. z b*	b	Bł. std. z b	t(50)	p	
W. wolny			6418145	29342,55	218,7317	0,000000	
rok	2,38478	0,056328	106137	2506,92	42,3375	0,000000	
V2**2	-1,50701	0,056328	-1204	45,00	-26,7542	0,000000	

Rysunek 22 Podsumowanie regresji zmiennej zależnej

$$y = -1204x^2 + 106137x + 6418145$$

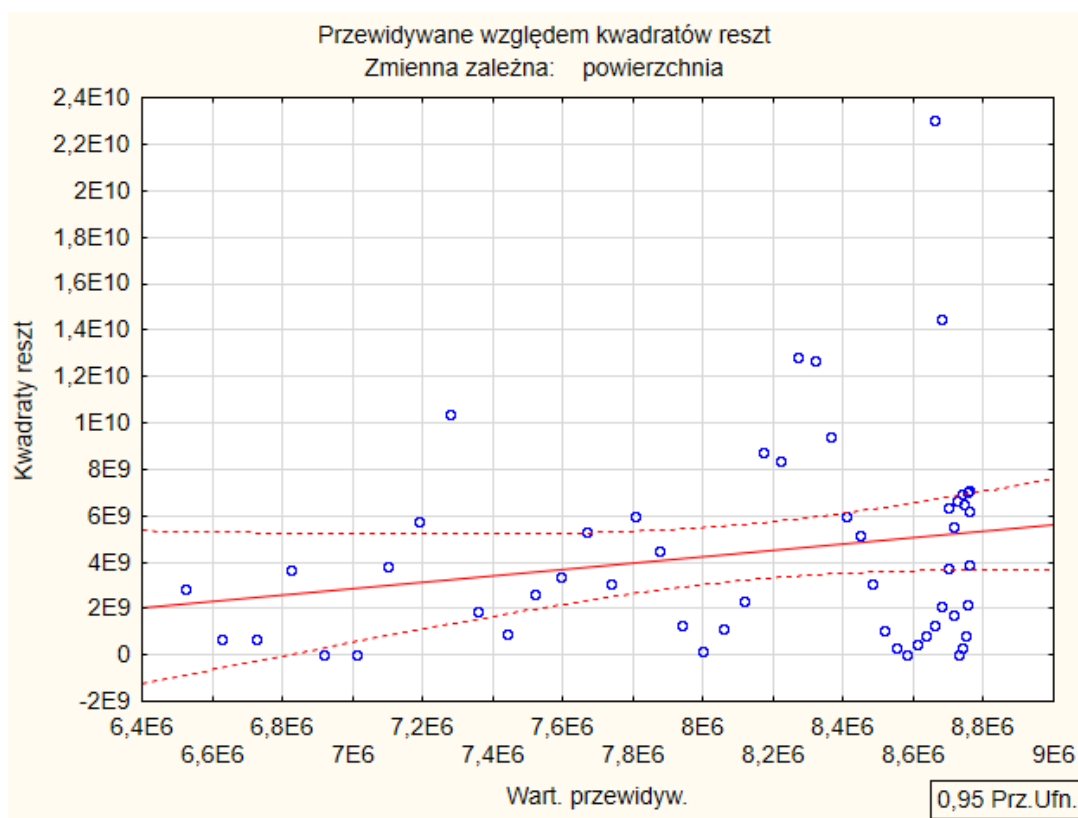
Z testu F($p < 0,0000$) wynika, że równanie regresji jest istotne. Wszystkie współczynniki są istotne, co oznacza, że model jest poprawny. Współczynnik korelacji, równy w przybliżeniu 0,99 świadczy o silnej korelacji między zmiennymi, a współczynnik determinacji w 99%

wyjaśnia zmienność pomiędzy zmienną czasową a powierzchnią lasów. Błąd standardowy estymacji jest bardzo duży, ale wynika to z zastosowania dużych liczb w regresji.

statystyka	Stat. podsum.; Zm
	Wartość
R wielorakie	0,995208296
Wielorakie R2	0,990439551
Skorygowane R2	0,990057134
F(2,50)	2589,94008
p	0
Błąd std. estymacji	68535,5578

Rysunek 23 Statystyki podsumowujące

Założenie homoskedastyczności



Rysunek 24 Przewidywane względem kwadratów reszt

Z wykresu można odczytać, że wariancja składnika losowego jest niejednorodna, wariancja reszt nie jest stała w czasie.

Autokorelacja

	d Durbina-Watsona (Arkusz i korelacja seryjna reszt)	
	d Durbin Watsona	Seryjna Kor.
Estymac.	0,157982	0,960130

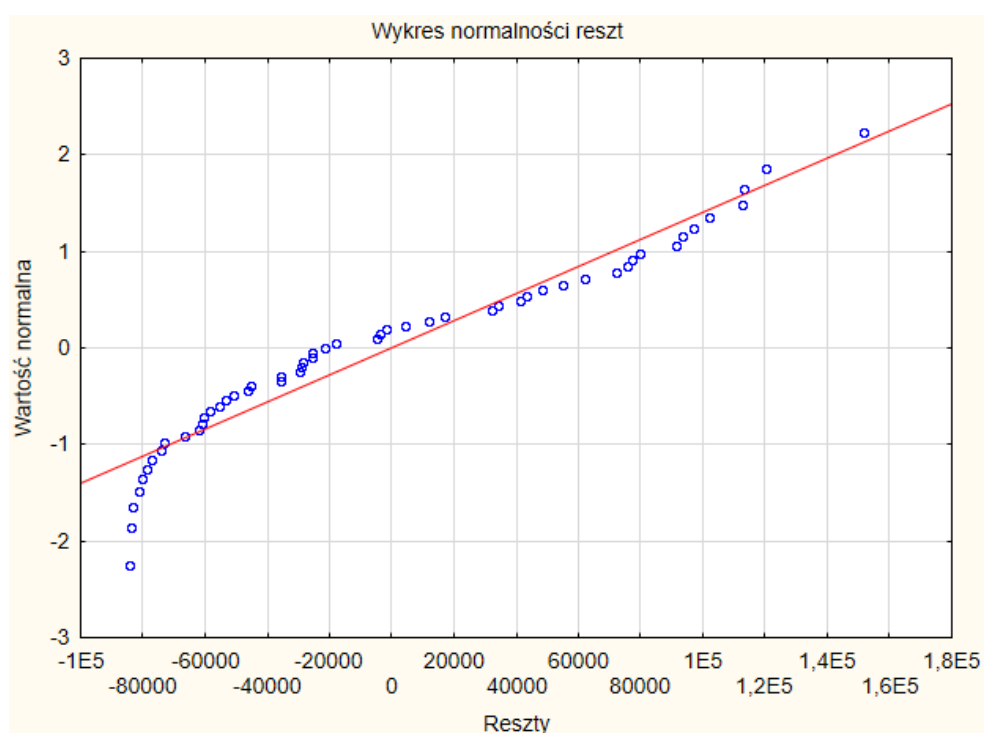
Rysunek 25 Test Durbina Watsona

Test Durbina-Watsona

$dL=1,59505$

$dU=1,47967$

$d < dU$, występuje autokorelacja, jest istotna statystycznie.



Rysunek 26 Wykres normalności reszt

Z wykresu normalności reszt wynika, że punkty układają się jak sinusoida. W tym przypadku widzimy, że punkty układają się wzdłuż prostej, potwierdzając w ten sposób normalność rozkładu. Można mieć wątpliwości co do pierwszej obserwacji, ponieważ jest nieco oddalona od linii, ale oddalenie nie wpływa znacząco na normalność wartości resztkowych.

Regresja logistyczna

Zostali zbadani pacjenci z zakażeniami krwi z punktem wyjścia w jamie brzusznej. Wyniki zostały powiązane z wiekiem i czynnikami takimi jak alkoholizm, zawały narządów oraz złe odżywianie. Dane zostały pobrane ze strony: <https://www.statsoft.pl/zasoby/do-pobrania/dane-do-ksiazek/>

Zmienne są następujące: PRZEŻYĆ – zmienna zależna opisująca przeżycie pacjenta (1-zgon, 0-przeżycie), WSTRZĄS – zmienna niezależna jakościowa (1 – objawy wystąpiły, 0 – brak objawów), ALKOHOL – zmienna jakościowa opisująca alkoholizm (1 – objawy wystąpiły, 0-brak objawów), WIEK – zmienna niezależna ilościowa opisująca wystąpienie zawału narządów (1 – objawy wystąpiły, 0 – brak objawów), ODŻYW – zmienna jakościowa opisująca złe odżywianie pacjenta (1 – objawy wystąpiły, 0 – brak objawów), ZAWAŁ – zmienna niezależna jakościowa opisująca wystąpienie zawału narządów (1 – objawy wystąpiły, 0 – brak objawów).

	1	2	3	4	5	6
	PRZEŻYC	WSTRZĄS	ODŻYW	ALKOHOL	WIEK	ZAWAŁ
1	0	0	0	0	56	0
2	0	0	0	0	80	0
3	0	0	0	0	61	0
4	0	0	0	0	26	0
5	0	0	0	0	53	0
6	1	0	1	0	87	0
7	0	0	0	0	21	0
8	1	0	0	1	69	0
9	0	0	0	0	57	0
10	0	0	1	0	76	0
11	1	0	0	1	66	1
12	0	0	0	0	48	0
13	0	0	0	0	18	0
14	0	0	0	0	46	0
15	0	0	1	0	22	0
16	0	0	1	0	33	0
17	0	0	0	0	38	0
18	0	0	0	0	27	0
19	1	1	1	0	60	1
20	0	0	0	0	31	0
21	0	0	0	0	59	1
22	0	0	0	0	29	0
23	0	1	0	0	60	0

Rysunek 27 Fragment danych

Do regresji pod uwagę wzięłam zmienną ALKOHOL i WIEK.

N=106	Stała B0	ALKOHOL	WIEK
Ocena	-5,666749	1,902917	0,06254818
Błąd standard.	1,340868	0,6094838	0,01879863
t(103)	-4,226179	3,122178	3,327273
p	0,00005150478	0,002330474	0,00121673
-95%CL	-8,326045	0,6941496	0,02526552
+95%CL	-3,007453	3,111685	0,09983083
Chi-kwadrat Walda	17,86059	9,747995	11,07074
p	0,00002383514	0,001796783	0,0008779934
Iloraz szans z jedn.	0,003459093	6,705426	1,064546
-95%CL	0,0002421278	2,002006	1,025587
+95%CL	0,04941739	22,45885	1,104984
Iloraz szans zakr.		6,705426	123,4961
-95%CL		2,002006	6,996746
+95%CL		22,45885	2179,769

Rysunek 28 Podsumowanie

Wartość p dla statystyki chi-kwadrat jest bardzo istotna. Można wyciągnąć wniosek, że alkohol i wiek ma istotny wpływ na przeżycie pacjenta.

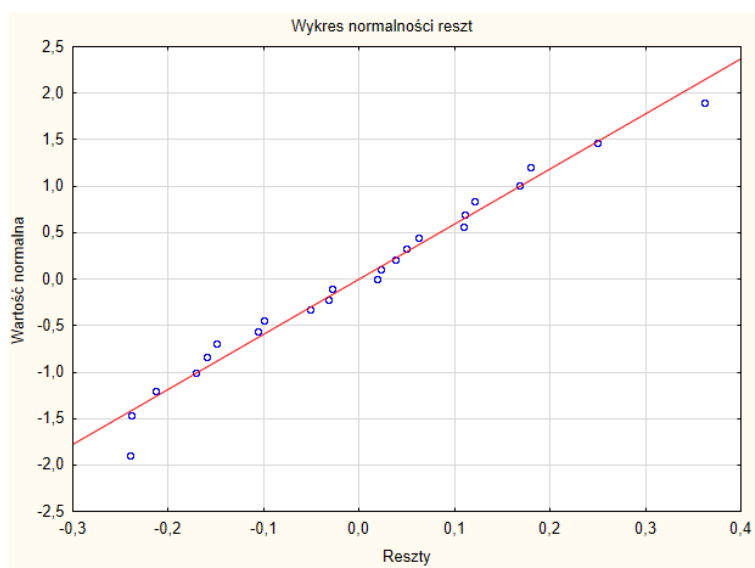
Model ma postać:

$$\text{Logit } P = -5,667 + 1,9 \cdot \text{alkohol} + 0,0625 \cdot \text{wiek}$$

Istotność funkcji regresji Poziom p dla testu Chi-kwadrat jest istotny, a to oznacza, że oszacowany model stanowi istotnie lepsze dopasowanie do danych niż model zerowy zawierający tylko wyraz wolny.

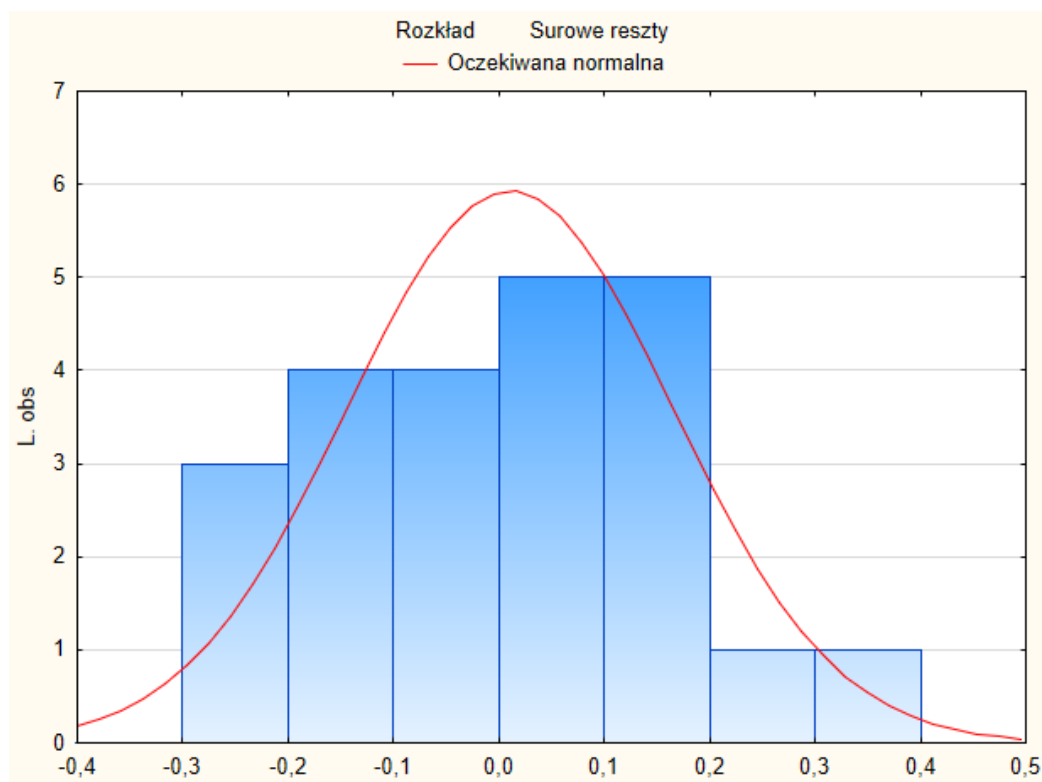
Istotność współczynników regresji Dla wszystkich zmiennych poziom $p < 0,05$, a to oznacza, że parametry regresji są istotne statystycznie.

Normalność rozkładu reszt



Rysunek 29 Wykres normalności reszt

Praktycznie wszystkie punkty (reszty) na wykresie normalności reszt leżą bardzo blisko linii, co oznacza, że reszty podlegają rozkładowi normalnemu.



Rysunek 30 Histogram

Ten wykres potwierdza, że reszty podlegają rozkładowi normalnemu.

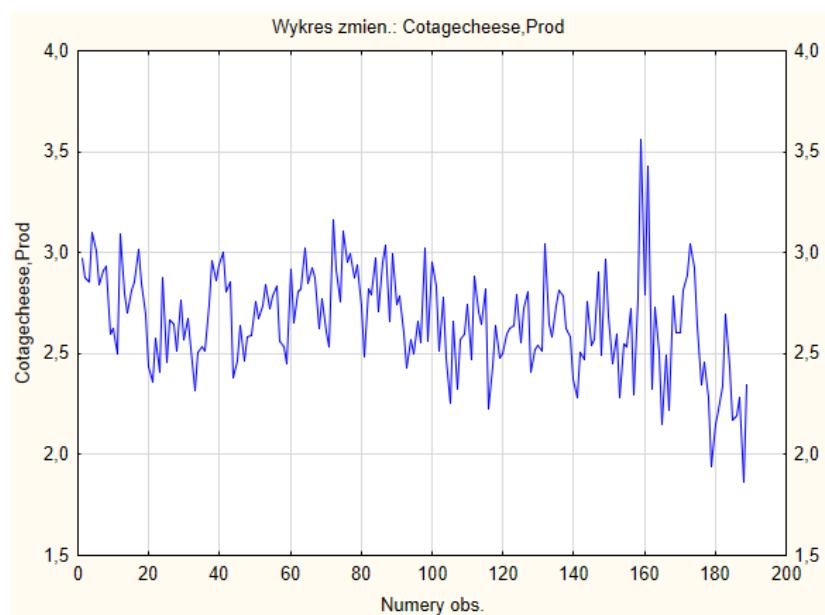
Dodatkowo szacunki współczynników odpowiadające zmiennym wskazują, że wzrost tych wielkości powoduje zwiększenie szansy na zgon pacjenta. Jednostkowy iloraz szans pokazuje, że szansa na zgon w przypadku, gdy człowiek był alkoholikiem to funkcja logit wzrosła o 1,9 w porównaniu, gdy człowiek nie spożywał w nadmiernych ilościach alkoholu, przy takich samych wartościach pozostałych zmiennych. Podobnie, jeśli wiek pacjenta wzrośnie o 1 rok, to funkcja logit. Błędy standardowe nie są duże, ale mogą powodować błędne wyniki. Przyczyną tego może być mała ilość danych. Prawidłowy model wymaga większej ilości danych.

Modele szeregów czasowych z analizą trendu

Baza pochodzi ze strony [kaggle.com](https://www.kaggle.com) dotyczy produkowania sera twarogowego od września 1998 do grudnia 2013. Baza zawiera 2 zmienne (lata, ilość produkowanego sera twarogowego) oraz 189 obserwacji.

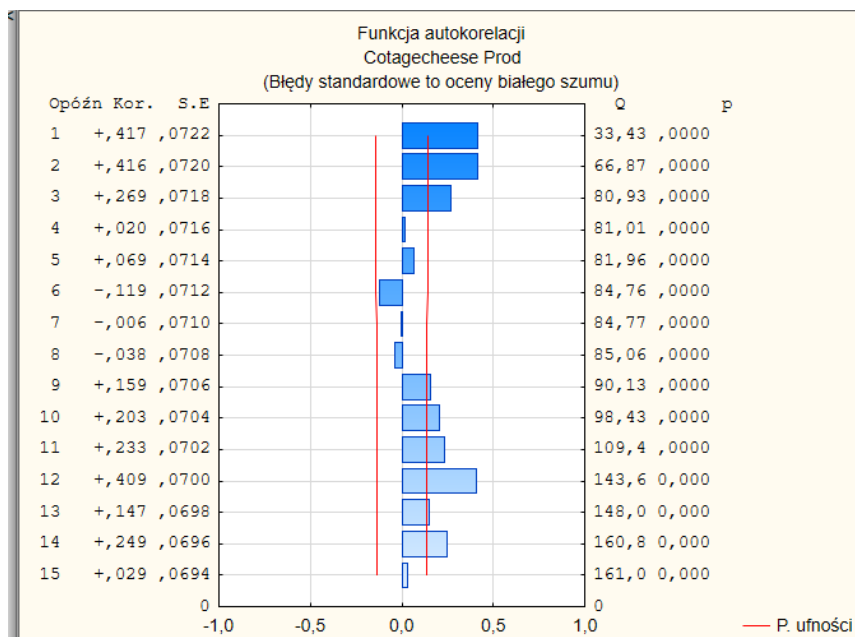
Year	Cottagecheese Prod
1998	2,971
1998	2,876
1998	2,853
1998	3,098
1998	3,007
1998	2,843
1998	2,909
1998	2,931
1998	2,597
1999	2,628
1999	2,496
1999	3,093
1999	2,801
1999	2,702
1999	2,813
1999	2,853
1999	3,018
1999	2,850
1999	2,692
1999	2,437
1999	2,363
2000	2,574
2000	2,409

Rysunek 31 Fragment danych



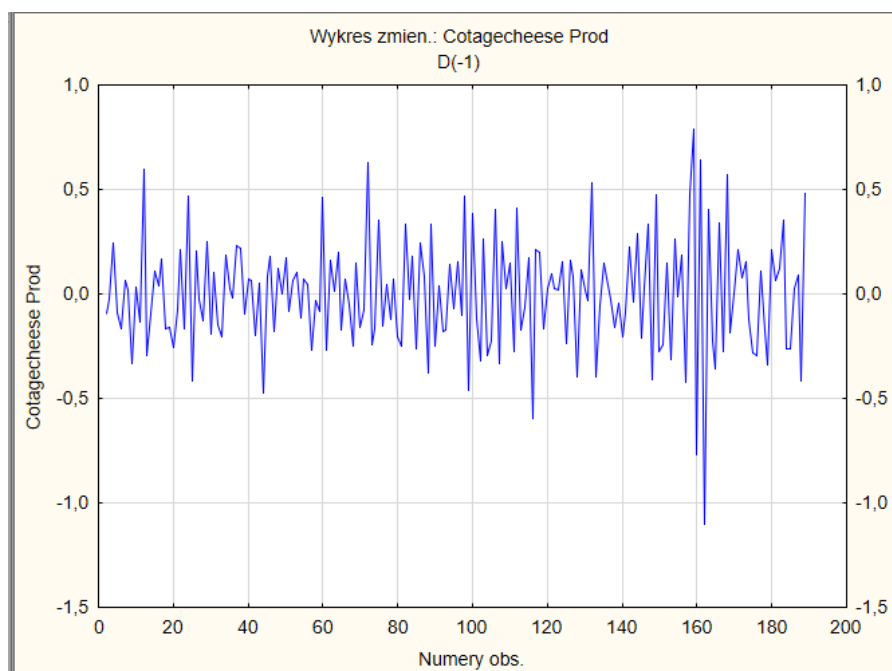
Rysunek 32 Wykres zmiennych

Na wykresie można zauważyć, że amplituda zmian sezonowych nie rośnie w czasie, więc możemy uznać, że sezonowość jest addytywna.



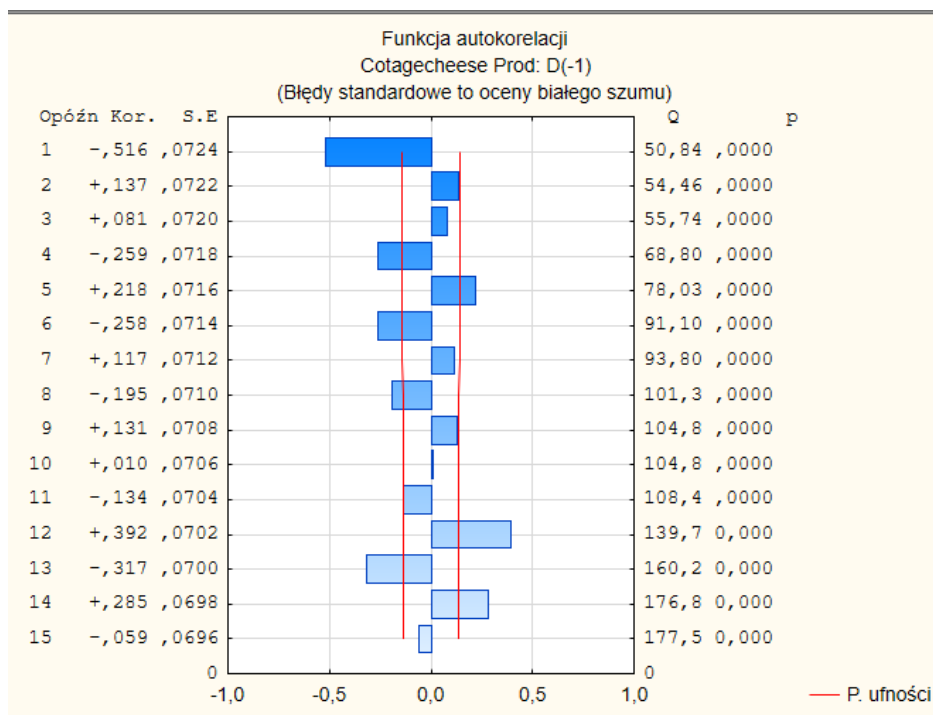
Rysunek 33 Funkcja autokorelacji

Na wykresie funkcji autokorelacji widać, że występuje autokorelacja, szereg zostanie poddany niesezonowemu różnicowaniu z opóźnieniem równym 1.



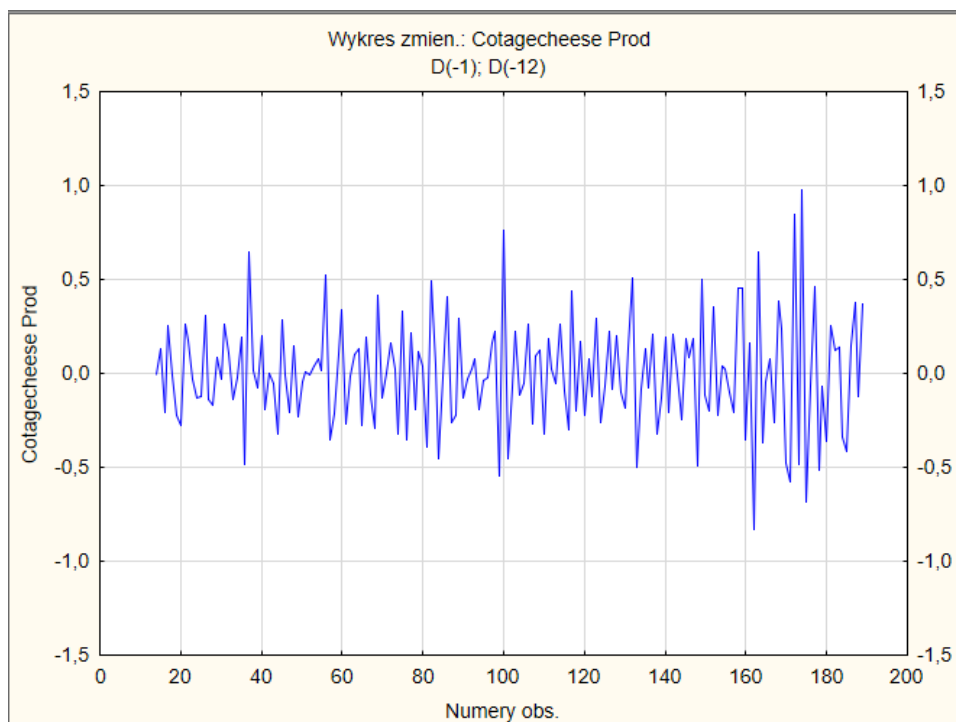
Rysunek 34 Wykres zmiennych

Szereg jest teraz krótszy o 1, czyli liczbę elementów opóźnionych, ponieważ pierwszego elementu nie można różnicować. Każdy element przekształconego szeregu reprezentuje różnicę między jego pierwotną wartością oraz jego początkową wartością sąsiadującego z nim elementu.

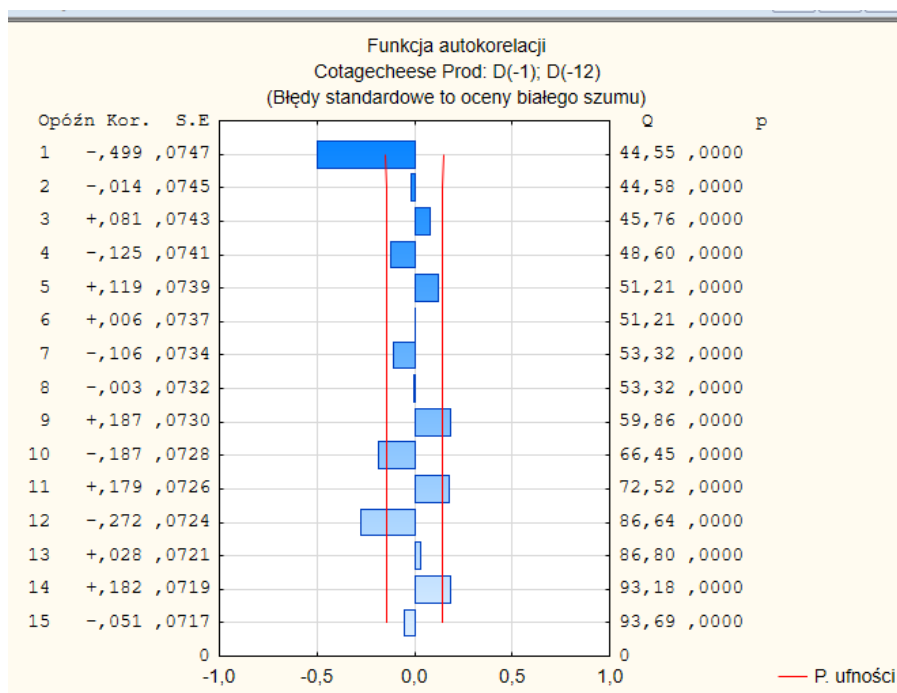


Rysunek 35 Funkcja autokorelacji

Z powyższego autokorelogramu wynika, że wciąż występują autokorelacje i dane należy jeszcze raz zróżnicować, tym razem ustawiłam opóźnienie na wartość 12.

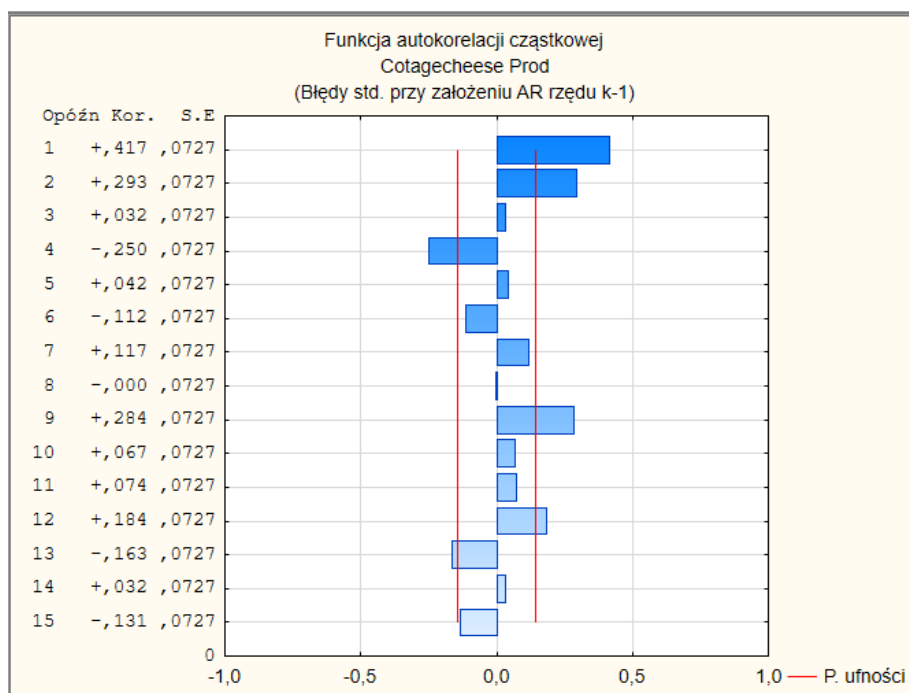


Rysunek 36 Wykres zmiennych



Rysunek 37 Funkcja autokorelacji

Z powyższego wykresu wynika, że większość silnych autokorelacji została usunięta. Mimo, że pozostało kilka autokorelacji, to na tym zakończę różnicowanie szeregu. Decyzja ta została podjęta, mając na uwadze ostrożność, aby za bardzo nie zróżnicować szeregu, czego konsekwencją byłoby zniesienie efektu parametrów średniej ruchomej.



Rysunek 38 Funkcja autokorelacji cząstkowej

Korelogram funkcji cząstkowej wygląda dobrze i jest gotowy do analizy ARIMA.

Model ARIMA dla pojedynczego szeregu: Arkusz1 w dane do arimy

OK (Rozpocznij estymację parametrów)

Zabezpiecz	Zmienna	Długa nazwa zmiennej (szeregu)
L	Cotagecheese,Prod	
	Cotagecheese,Prod	D(-1)

Liczba zapamiętanych przekształt. dla 1 zm.: 3

Zapisz zmienne Usun

Podstawowe Więcej Autokorelacje Przegląd szeregu

Parametry modelu ARIMA

☐ Szacuj stałą Opóźnienie sezonowe: 12

p - autoregresyjne: 0 P - sezonowe: 0

q - średnia ruchoma: 1 Q - sezonowe: 1

Przekształć zmienną (szereg) przed analizą

☐ Logarytm naturalny ☐ Potęgowanie: 2,0

☒ Różnice 1. opóźnienie: 1 N przeb.: 1

2. opóźnienie: 12 N przeb.: 1

Inne przekształcenia i wykresy

Metoda estymacji

☒ Przybliżona (McLeoda i Salesa)

Szacuj obs. poprzedzające: 0

☐ Dokładna (Melarda)

Opcje estymacji

Maksymalna liczba iteracji: 50

Kryterium zbieżności (wymagana dokładność): .0001

Maks. liczba iteracji dla szacowania obs. poprzedz.: 10

☐ Wartości początkowe użyt.

Rysunek 39 Wprowadzanie ustawień

Jak widać na powyższym oknie, ustawiłam następujące wartości dla parametrów: opóźnienie sezonowe równe 12, q, czyli średnia ruchoma=1 oraz Q - sezonowe zmieniłam na wartość 1. Następnie w części „Przekształć zmienną (szereg) przed analizą” zaznaczyłam opcję Różnice. W celu określenia różnicowania sezonowego w polu 1.opóźnienie dałam wartość 1, tak samo zrobiłam dla obu pól N przeb., w polu 2. opóźnienie nadałam wartość 12.

Wyniki analizy ARIMA pojedynczego szeregu: Arkusz1 w dane do arimy

Zmienna: Cotagech

Przekształcenia: D(1),D(12)

Model: (0,1,1)(0,1,1) Opóźn. sez.: 12

Liczba obs.:176 Wstępne SS=14,718 Końc. SS=6,5846 (44,74%) MS=.03784

Parametry (p/ps-autoregresyjne, q/Qs-średniej ruch.) podświetl: p<.05

q(1) Qs(1)

Ocena: .77518 .77749

Błąd std: .04457 .05151

Podstawowe Więcej Przegląd i reszty Rozkład reszt Autokorelacje

Podsum.: Oceny parametrów

Drukuj wyniki

Po kliknięciu przycisku Podsumowanie, reszty i przekształcenia szeregu będą dołączone do zmiennych w pamięci.

Podsum.

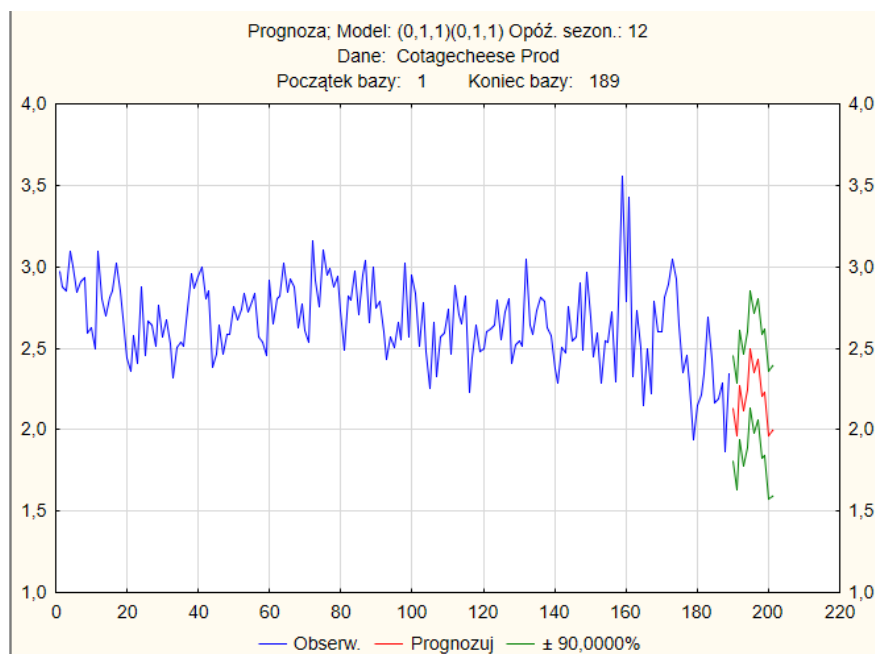
Anuluj

Opcje

Grupami

Rysunek 40 Wyniki analizy ARIMA

Zarówno parametr $q(1)$, jak i $Qs(1)$ jest oznaczony na czerwono, co oznacza, że są istotne statystycznie.



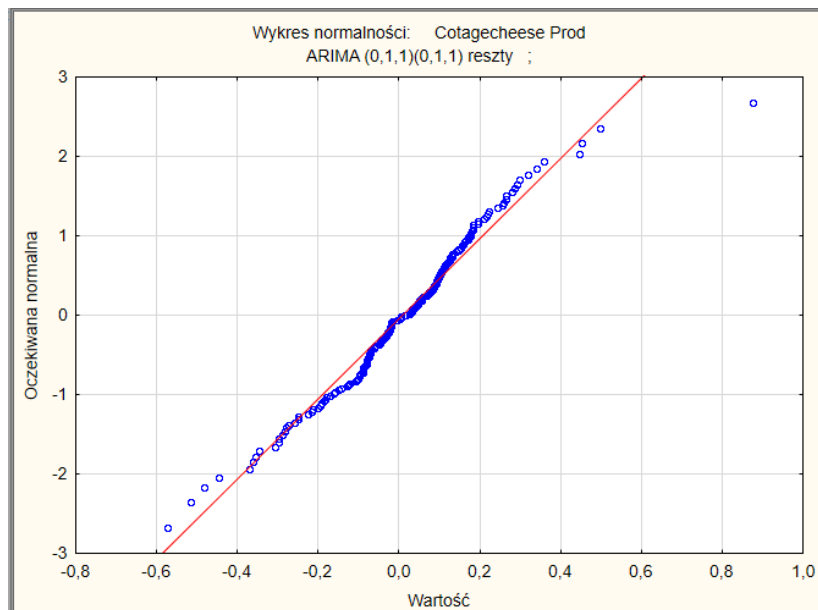
Rysunek 41 Prognoza

Proгноза; Model: (0,1,1)(0,1,1) Opóź. sezon.: 12 (A Dane: Cotagecheese Prod Początek bazy: 1 Koniec bazy: 189				
Nr obs.	Prognozu	Dolne 90,0000%	Górne 90,0000%	Błąd std
190	2,131671	1,809983	2,453359	0,194531
191	1,959376	1,629658	2,289094	0,199387
192	2,273545	1,935989	2,611101	0,204127
193	2,120035	1,774818	2,465251	0,208759
194	2,245738	1,893027	2,598448	0,213291
195	2,492976	2,132927	2,853025	0,217729
196	2,348334	1,981094	2,715575	0,222078
197	2,432248	2,057954	2,806542	0,226343
198	2,203077	1,821860	2,584294	0,230529
199	2,228011	1,839994	2,616027	0,234641
200	1,965500	1,570802	2,360199	0,238682
201	1,990898	1,589628	2,392168	0,242656

Rysunek 42 Tabela prognoz

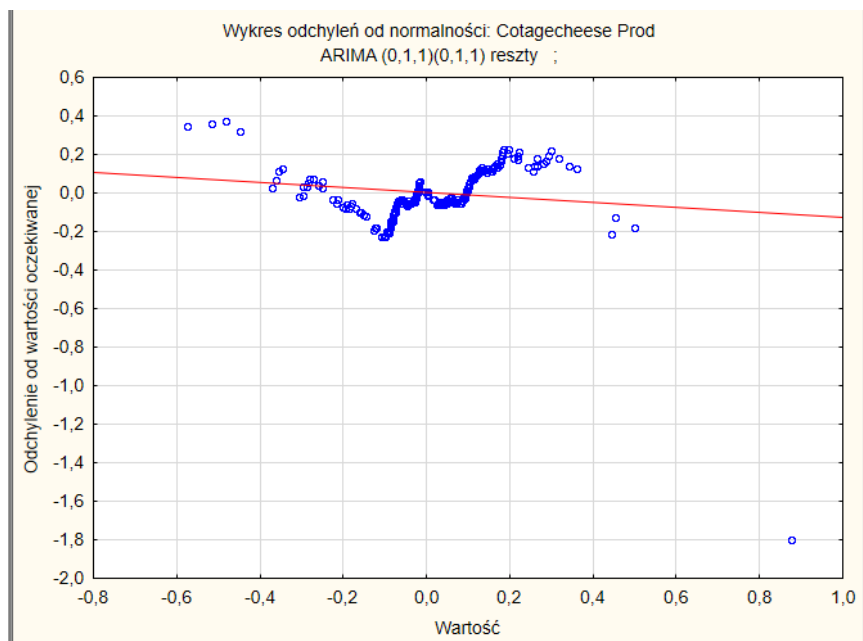
Dodałam do modelu 12 przewidywanych przypadków dla tego szeregu. Zarówno na wykresie jak i tabeli prognoz widać, że mają one niski poziom błędu oraz mieszczą się w przedziale ufności wartości przewidywanych. Prognozy tworzą logiczne przedłużenie wykresu.

Aby model ARIMA był dopasowany musi spełniać dwa założenia: reszty muszą mieć rozkład normalny oraz nie może występować autokorelacja reszt.

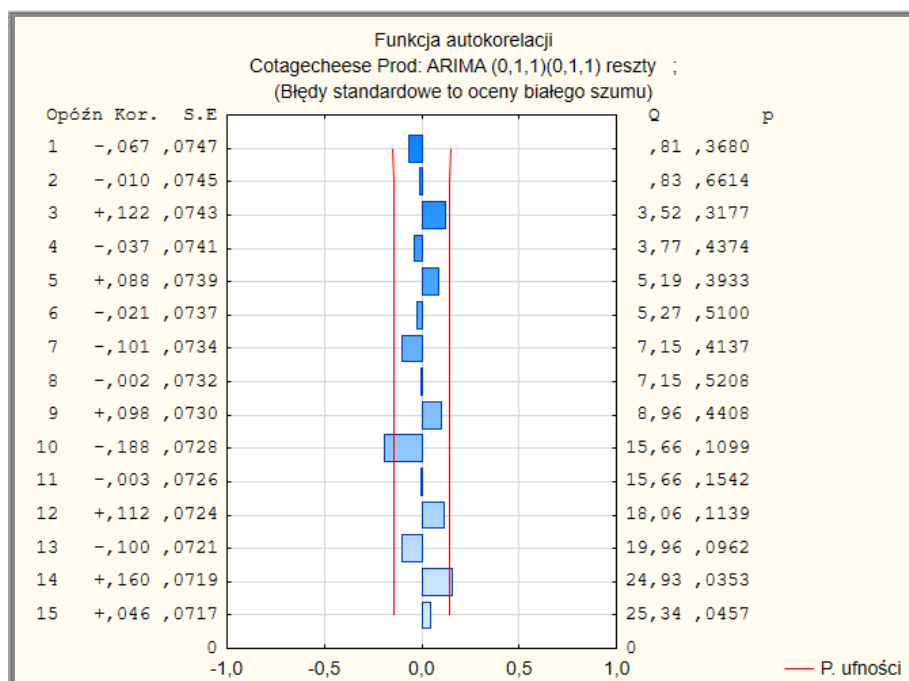


Rysunek 43 Wykres normalności

Na wykresie można zauważyć, że reszty układają się wokół linii prostej. Można więc uznać, że istnieje rozkład normalny.



Rysunek 44 Wykres odchylen od normalności

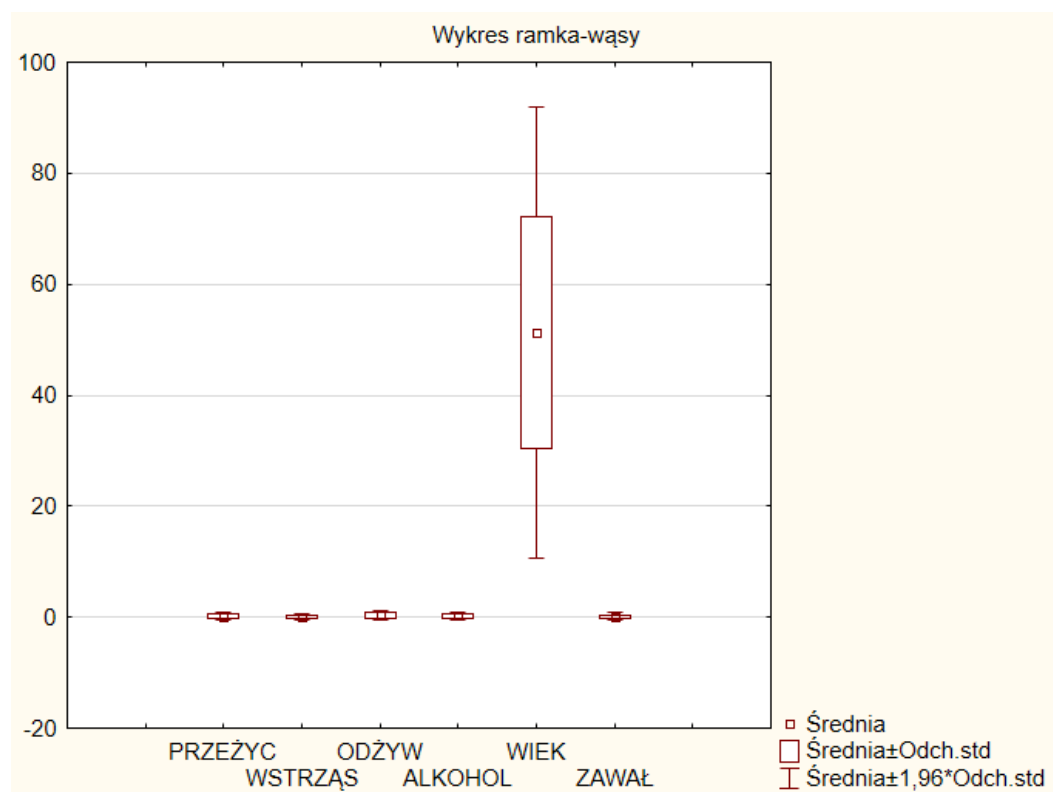


Rysunek 45 Funkcja autokorelacji

Zarówno wykres odchyień od normalności dla zmiennej Cotagechesse Prod jak i wykres funkcji autokorelacji wykazują, że praktycznie nie ma żadnych autokorelacji dla reszt, a więc możemy przyjąć, że warunek jest spełniony.

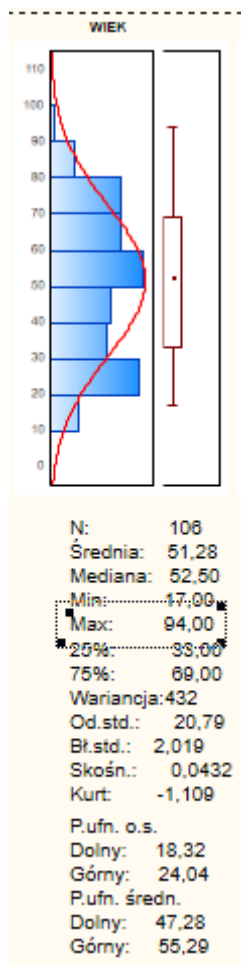
Diagnozowanie obserwacji odstających

Stosując bazę dotyczącą przeżyć, wszystkie zmienne są binarne oprócz wieku, dlatego sprawdzimy czy zmienna wiek posiada wartości odstające. Podstawowy wykres wygląda tak:



Rysunek 46 Diagnozowanie obserwacji odstających

Wiek ma największy zakres zmienności jako jedyna z pozostałych. Ta zmienna może zawierać obserwacje nietypowe, mając charakter nietypowości i wpływowości. Pomimo wybrania odstające i ekstremalne nic się nie dzieje. Nie ma obserwacji odstających.



Rysunek 47 Podsumowanie zmiennej

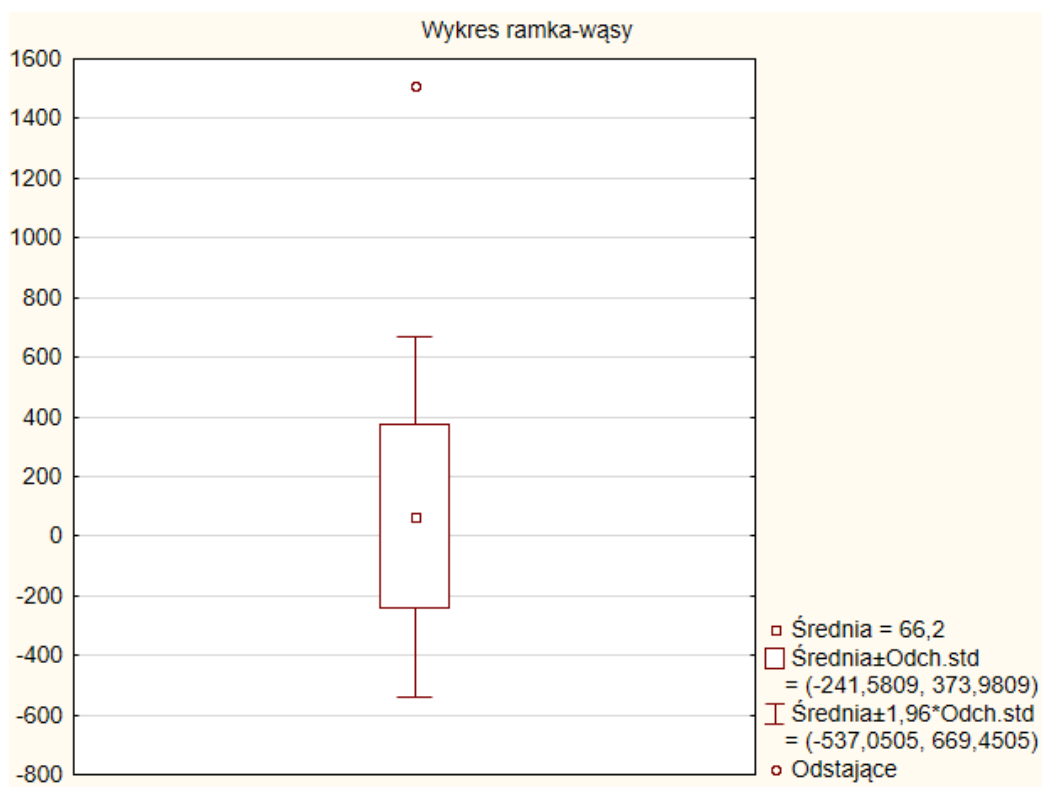
Widać, że dla zmiennej WIEK największą wartością jest 94. Nie ma obserwacji odstających.

Jako że nie ma obserwacji odstających, postanowiłam skorzystać z bazy, w której zawarte jest średnie roczne stężenie cezu 137 Bq/m² w opadzie całkowitym od 1986 roku. W niej na pewno będzie znajdować się obserwacja odstająca spowodowana wybuchem w Czarnobylu. Dane zostały pobrane ze strony GUS, Ochrona Środowiska, 2010.

rok	stężenie cezu 137
1986	1511
1987	22
1988	12
1989	8
1990	7,6
1991	5,3
1992	3,8
1993	3,8
1994	2,2
1995	2,1
1996	1,3
1997	1,5
1998	1
1999	0,8

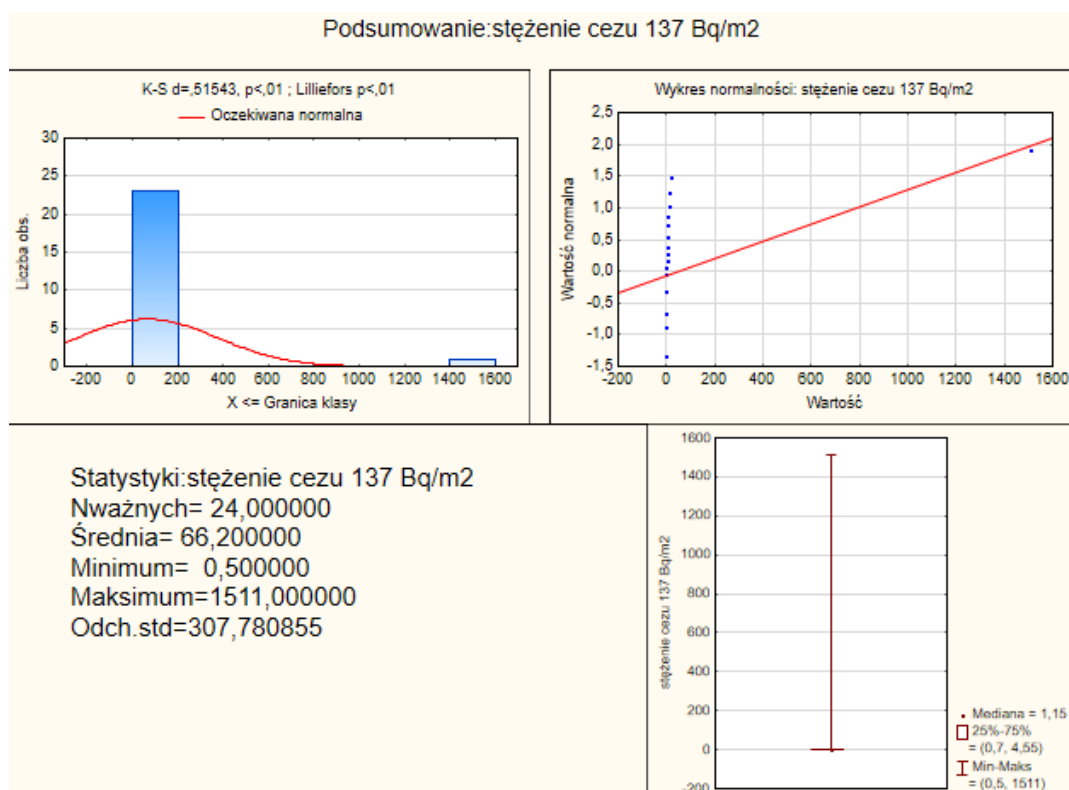
Rysunek 48 Fragment tabeli

Tworzę wykres boxplot ramka-wąsy dla stężenia cezu 137 Bq/m². Zaznaczam opcję, żeby wykres uwzględnił wartości odstające. Prezentuje się to następująco:



Rysunek 49 Wykres ramka-wąsy

Pojawia się kropczka na górze, która informuje, że znajduje się obserwacja nietypowa.



Rysunek 50 Podsumowanie zmiennej

Na powyższym wykresie widać, że ta zmienna nie ma rozkładu normalnego, co jest spowodowane wartością odstającą. Maksimum wynosi 1511 – co jest bardzo dużym odstępstwem od normy.

Zmienna	Statystyki opisowe (Arkusz1 w odstająca)						
	Nważnych	Średnia	Grubbsa statystyka	poziom p	Minimum	Maksimum	Odch.std
stężenie cezu 137 Bq/m ²	24	66,20000	4,694249	0,00	0,500000	1511,000	307,7809

Rysunek 51 Statystyki opisowe

Na powyższym screenie wykonany jest test Grubbsa:

H₀: Nie ma odchyień w zbiorze danych

H₁: Istnieje przynajmniej jedno odchylenie w zbiorze danych

Wartość $p < 0,05$, odrzucamy H₀ na korzyść H₁ – istnieje przynajmniej jedno odchylenie, które potwierdza wykres boxplot.

Spis ilustracji

Rysunek 1 Fragment danych	3
Rysunek 2 Podsumowanie regresji.....	3
Rysunek 3 Nadmiarowość zmiennych niezależnych	4
Rysunek 4 Test Durbina-Watsona	4
Rysunek 5 Przewidywane względem wartości resztowych.....	5
Rysunek 6 Histogram.....	6
Rysunek 7 Wykres normalności reszt.....	6
Rysunek 8 Podsumowanie regresji zmiennej zależnej	7
Rysunek 9 Statystyki podsumowujące	7
Rysunek 10 Nadmiarowość zmiennych niezależnych	8
Rysunek 11 Test Durbina Watsona	8
Rysunek 12 Przewidywane względem wartości resztowych.....	9
Rysunek 13 Histogram.....	10
Rysunek 14 Wykres normalności reszt.....	10
Rysunek 15 Podsumowanie regresji zmiennej zależnej	11
Rysunek 16 Statystyki podsumowujące	11
Rysunek 17 Nadmiarowość zmiennych niezależnych	11
Rysunek 18 Przewidywane względem wartości resztowych.....	12
Rysunek 19 Histogram.....	12
Rysunek 20 Wykres normalności reszt.....	13
Rysunek 21 Funkcja trendu wielomianowa.....	14
Rysunek 22 Podsumowanie regresji zmiennej zależnej	14
Rysunek 23 Statystyki podsumowujące	15
Rysunek 24 Przewidywane względem kwadratów reszt.....	15
Rysunek 25 Test Durbina Watsona	16
Rysunek 26 Wykres normalności reszt.....	16
Rysunek 27 Fragment danych	17
Rysunek 28 Podsumowanie.....	18
Rysunek 29 Wykres normalności reszt.....	18
Rysunek 30 Histogram.....	19
Rysunek 31 Fragment danych	20
Rysunek 32 Wykres zmiennych	20
Rysunek 33 Funkcja autokorelacji	21
Rysunek 34 Wykres zmiennych	21
Rysunek 35 Funkcja autokorelacji	22
Rysunek 36 Wykres zmiennych	22
Rysunek 37 Funkcja autokorelacji	23
Rysunek 38 Funkcja autokorelacji częściowej	23
Rysunek 39 Wprowadzanie ustawień	24
Rysunek 40 Wyniki analizy ARIMA	24
Rysunek 41 Prognoza	25
Rysunek 42 Tabela prognoz.....	25
Rysunek 43 Wykres normalności.....	26
Rysunek 44 Wykres odchyżeń od normalności.....	26
Rysunek 45 Funkcja autokorelacji	27
Rysunek 46 Diagnozowanie obserwacji odstających	28
Rysunek 47 Podsumowanie zmiennej	29

Rysunek 48 Fragment tabeli	30
Rysunek 49 Wykres ramka-wąsy	30
Rysunek 50 Podsumowanie zmiennej	31
Rysunek 51 Statystyki opisowe	31

Bibliografia

- M. Rabiej, *Statystyka z programem Statistica*, 2012 Helion;
- A. Stanisławski *Przystępny kurs statystyki w wykorzystaniu programu STATISTICA PL na przykładach z medycyny Tom II*, 2000 StatSoft;
- M. Maliński Wybrane zagadnienia statystyki matematycznej w Excelu i pakiecie Statistica, 2015 Wydawnictwo Politechniki Śląskiej;