ISSN 1230-3801 Zeszyt 133 nr 1/2015, 81-102

# WARTOŚCI RESZTOWE W PROCESIE REGRESJI

Dariusz AMPUŁA
Wojskowy Instytut Techniczny Uzbrojenia

**Streszczenie:** W artykule autor we wstępie argumentuje potrzebę przeprowadzenia weryfikacji opracowanego modelu regresji za pomocą analizy wartości resztowych. Na początku scharakteryzowano założenia analizy reszt, zwracając uwagę na własności tych reszt czyli: normalność, stałość wariancji i brak autokorelacji. Do analizy wzięto wyniki badań zapalników artyleryjskich typu MD-7.Scharakteryzowano własność wartości resztowych, która głosi, że reszty modelu mają rozkład normalny. Następnie omówiono sposób wyznaczania autokorelacji wartości resztowych, uwzględniając test Durbina-Watsona, który służy do weryfikacji współczynnika autokorelacji. Ze względu na obszerność artykułu nie analizowano metody mnożników Lagrange'a. Omówiono własność stałości wariancji wartości resztowych, czyli założenie homoscedastyczności składnika losowego. W tym celu przedstawiono wykresy rozrzutu reszt względem wartości przewidywanych. Zaprezentowano sposób określania obserwacji nietypowych w analizie procesu regresji. Przedstawiono graficzną postać interpretacji obserwacji odstających za pomocą wykresu rozrzutu reszt względem reszt usuniętych. Na końcu artykułu przedstawiono zwięzłe wnioski dotyczące wartości resztowych w procesie regresji.

Słowa kluczowe: wartość resztowa, normalność, wariancja, autokorelacja, obserwacja odstająca.

## 1. Wstęp

Większość założeń procesu regresji dotyczy wartości resztowych, czyli tzw. reszt. Analizując reszty [1], możemy szybko i skutecznie wykryć wszystkie odstępstwa jakie mogły nastąpić od poprawnej analizy procesu regresji. Mimo, iż nie wszystkie założenia dotyczące procesu regresji możemy sprawdzić, to jednak największe odstępstwa możemy wykryć i ewentualnie wyeliminować. Analiza rozkładu wartości resztowych powinna być jednym z najważniejszych etapów weryfikacji modelu regresyjnego. Badając reszty modelu szybko wykryjemy również odstające obserwacje. Obserwacje takie mogą w poważny sposób zaburzyć równanie regresji poprzez „naciągnięcie" linii regresji w ich kierunku. Powoduje to zmiany współczynników regresji. Może się zdarzyć, że usunięcie takiego odstającego punktu danych obserwacji daje kompletnie różne wyniki analizy regresji, a co za tym idzie na tej podstawie przewidujemy inne wartości procesu predykcji.

Chcąc otrzymać w miarę poprawny model regresji, musimy zawsze po estymacji i weryfikacji tego modelu przeanalizować otrzymane wartości resztowe. Tylko poznanie ich wykresów i statystyk z nimi związanych gwarantuje szybkie wykrycie odstępstw i odpowiednią ich interpretację statystyczną. Regułą powinna się stać analiza wartości resztowych zawsze po oszacowaniu parametrów modelu regresji.

## 2. Założenia modelu

Analizę reszt według [1] należy rozpocząć od sprawy najważniejszej, tzn. sprawdzenia założeń klasycznej metody najmniejszych kwadratów. Poprawnie skonstruowany model to

taki, który charakteryzuje się pewnymi pożądanymi własnościami reszt (normalność, stałość wariancji, brak autokorelacji). Procedurę sprawdzenia założeń modelu stosuje się *ex post,* tzn. po oszacowaniu parametrów modelu klasyczną metodą najmniejszych kwadratów. Następnie, gdy okaże się, iż niektóre z tych warunków nie są spełnione, parametry tego modelu szacuje się ponownie, stosując inną metodę estymacji albo inną postać modelu.

Do analizy wartości resztowych weźmiemy wyniki badań diagnostycznych dla zapalników typu MD-7, jakie zostały uzyskane podczas badań laboratoryjnych do roku 2010 włącznie. Zapalniki tego typu były stosowane w amunicji artyleryjskiej różnych kalibrów z pociskami przeciwpancerno-smugowymi. W tabeli 1 przedstawione zostały ilości zbadanych partii w poszczególnych latach badawczych, decyzje negatywne i decyzje dodatnie uzyskane po badaniach laboratoryjnych oraz wartości frakcji decyzji dodatnich. Jako decyzje dodatnie przyjęto decyzje „B5" i „B3", natomiast, jako decyzje negatywne przyjęto wszystkie pozostałe decyzje.

## 3. Normalność wartości resztowych

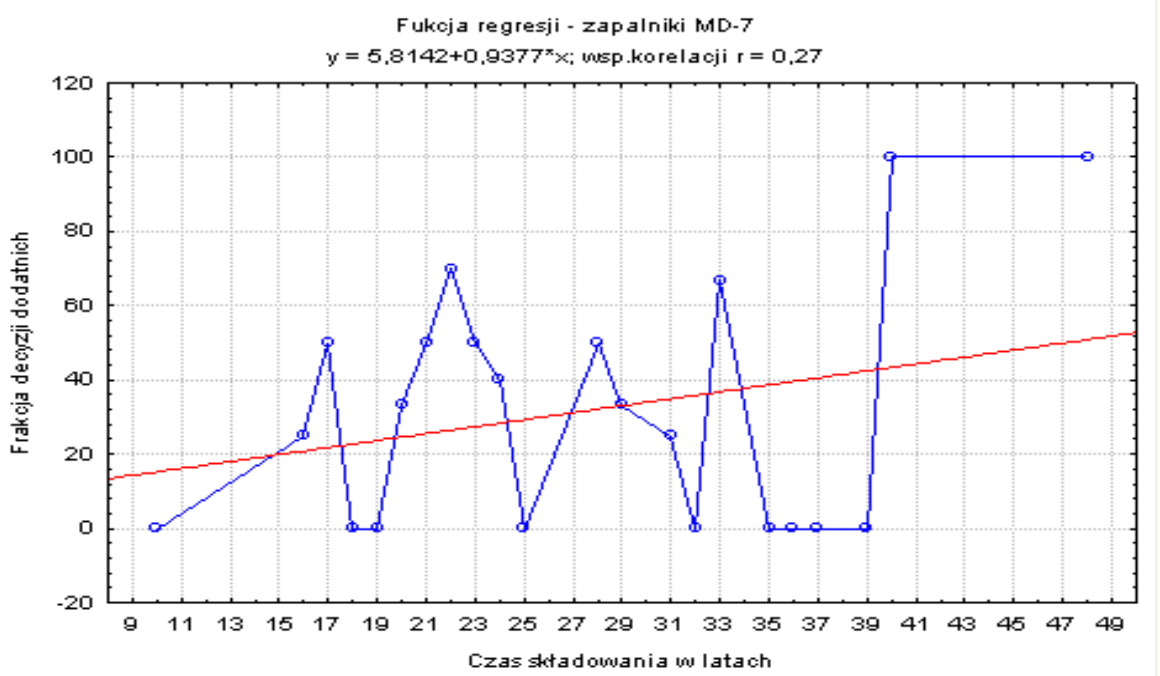
Pierwszym z założeń wartości resztowych jest założenie głoszące, że reszty modelu mają rozkład normalny. To założenie nie jest konieczne dla dopasowania modelu metodą najmniejszych kwadratów, ale dla weryfikacji istotności otrzymanych parametrów. Stosowane w tym celu testy ( t-Studenta, F ) są odporne na niewielkie odchylenia od normalności. Jeżeli założenie o normalności jest ewidentnie naruszone, to oceny istotności współczynników regresji mogą być zaburzone. Możemy wówczas dokonać transformacji zmiennej zależnej – przykładowo $logY$ lub $\sqrt{Y}$ – dla otrzymania rozkładu normalnego.

Tabela 1 – zapalniki typu MD-7

| Czas składowania | Liczba decyzji negatywnych | Liczba decyzji dodatnich | Frakcja decyzji dodatnich | Liczba zbadanych partii |
|---|---|---|---|---|
| 10 | 1 | 0 | 0,00 | 1 |
| 16 | 3 | 1 | 25,00 | 4 |
| 17 | 3 | 3 | 50,00 | 6 |
| 18 | 1 | 0 | 0,00 | 1 |
| 19 | 2 | 0 | 0,00 | 2 |
| 20 | 2 | 1 | 33,33 | 3 |
| 21 | 1 | 1 | 50,00 | 2 |
| 22 | 3 | 7 | 70,00 | 10 |
| 23 | 3 | 3 | 50,00 | 6 |
| 24 | 3 | 2 | 40,00 | 5 |
| 25 | 4 | 0 | 0,00 | 4 |
| 28 | 1 | 1 | 50,00 | 2 |
| 29 | 2 | 1 | 33,33 | 3 |
| 31 | 3 | 1 | 25,00 | 4 |
| 32 | 1 | 0 | 0,00 | 1 |
| 33 | 1 | 2 | 66,67 | 3 |
| 35 | 1 | 0 | 0,00 | 1 |
| 36 | 1 | 0 | 0,00 | 1 |
| 37 | 2 | 0 | 0,00 | 2 |
| 39 | 1 | 0 | 0,00 | 1 |
| 40 | 0 | 1 | 100,00 | 1 |
| 48 | 0 | 3 | 100,00 | 3 |
| Razem | 39 | 27 | 40,91 | 66 |

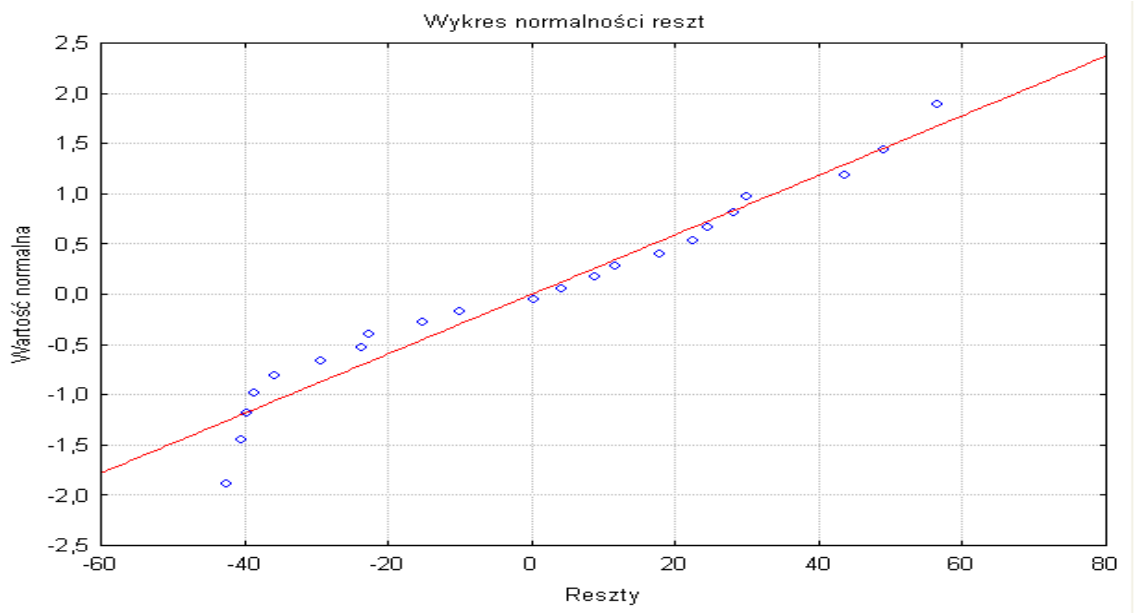
Analizując dane z tabeli 1 widzimy, że brak jest ciągłości wyników badań. Najstarsze badane zapalniki miały 48 lat, natomiast brak jest wyników badań dla zapalników o wieku od 41÷47 lat. Podobna sytuacja jest dla zapalników w wieku składowania od 11÷15 lat. Brak ciągłości wyników badań diagnostycznych może ewidentnie wpływać na zmianę rzeczywistych parametrów modelu regresji, co w konsekwencji prowadzi do możliwości wyciągnięcia błędnych wniosków dotyczących procesu predykcji dla tego typu zapalników.

Dla przedstawionych danych w tabeli 1 wykres zależności frakcji decyzji dodatnich od czasu składowania przedstawia rysunek 1. Na rysunku tym zaznaczona została na czerwono linia regresji.




**Rys. 1. Linia regresji dla zapalników MD-7**

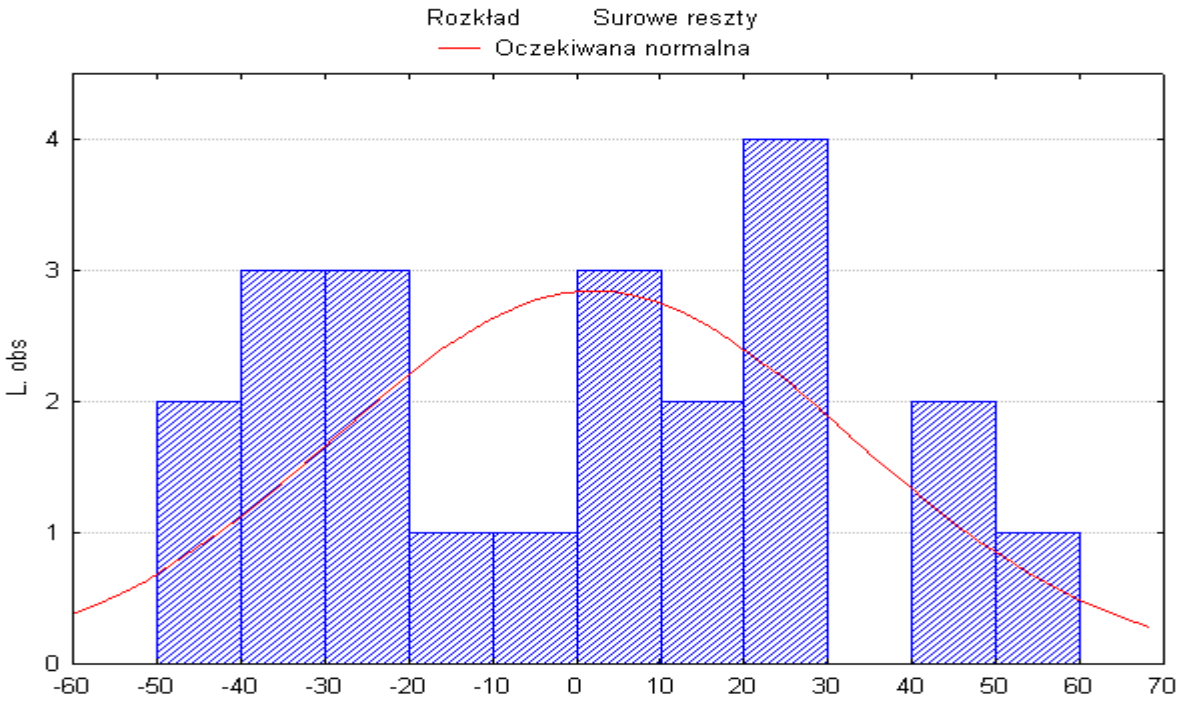
W celu uzyskania graficznego sprawdzenia normalności tego modelu regresji, wykorzystam oprogramowanie [4]. Wykres normalności reszt dla tego modelu przedstawia rysunek 2.




**Rys. 2. Wykres normalności wartości resztowych dla zapalników MD-7**

Otrzymany wykres umożliwia wzrokową ocenę zgodności reszt z rozkładem normalnym: jeśli reszty nie mają rozkładu normalnego, to punkty będą się odchylać od linii prostej, jeśli punkty tworzą wyraźny kształt wokół prostej, to sugeruje zastosowanie jakiejś transformacji. Przykładowo, jeśli punkty tworzą literę S, to możemy zastosować transformację logarytmiczną. Wykres normalności reszt może również ujawniać obserwacje odstające.

W naszym przypadku widzimy, że punkty układają się wzdłuż prostej, potwierdzając w ten sposób normalność rozkładu reszt. Można mieć obiekcję dotyczącą pierwszej obserwacji, ponieważ jest ona nieco oddalona od linii, ale oddalenie to nie wpływa znacząco na normalność wartości resztowych.




**Rys. 3. Histogram reszt dla zapalników MD-7**

Podobnych informacji jak wykres normalności prawdopodobieństwa dostarcza także histogram reszt. W sytuacji idealnej linia normalnej powinna przechodzić przez środki górnych krawędzi słupków. Niewielkie odchylenie od normalności nie jest niebezpieczne, zwłaszcza dla liczebnie dużych prób. W naszym przykładzie liczność próby wyniosła 66 obserwacji. Histogram reszt dla zapalników typu MD-7 przedstawia rysunek 3. Jak widać z rysunku, nie jest to sytuacja idealna, ponieważ linia odbiega nieco od wierzchołków słupków, nie można jednoznacznie przyjąć, że mamy do czynienia z normalnością wartości resztowych.

## 4. Autokorelacja wartości resztowych

Kolejnym założeniem dla wartości resztowych jest brak autokorelacji składnika losowego. Brzmi ono następująco: składniki losowe (reszty) są nieskorelowane, czyli $e_i$ oraz $e_j$ są ze sobą nieskorelowane dla wszystkich par $i, j$ gdzie $i, j = 1,2, \dots, n$ oraz $i \neq j$. Gdy założenie to nie jest spełnione mówimy o autokorelacji reszt.

Mówienie o autokorelacji ma sens tylko wtedy, gdy obserwacje w próbie są uporządkowane (najczęściej przez czas). Tak więc to założenie dotyczy przede wszystkim szeregów czasowych. W naszym przypadku kolejne obserwacje uzależnione są od czasu składowania i przeprowadzenia badań diagnostycznych, czyli mamy tu do czynienia z uporządkowanym szeregiem czasowym.

Wystąpienie autokorelacji powoduje, że estymatory otrzymane klasyczną metodą najmniejszych kwadratów przestają być najefektywniejszymi estymatorami. Mogą być obciążone dodatkowym błędem, a przez to mało precyzyjne, czyli mało przydatne. Najczęściej występuje autokorelacja pierwszego rzędu. Oznacza to, że składnik resztowy $e_t$ zależy od:

- składnika $e_{t-1}$,
- nowego składnika losowego $\varepsilon_t$.

Zachodzi zatem zależność:

$$e_t = \rho\, e_{t-1} + \varepsilon_t \qquad (1)$$

gdzie: $\rho$ - jest współczynnikiem autokorelacji. Przyjmuje on wartości z przedziału $< -1,1 >$. Jeżeli $\rho > 0$, mówimy o pozytywnej autokorelacji, w przeciwnym przypadku mówimy o negatywnej autokorelacji.

Problem jest, gdy występują sytuacje wątpliwe. Wówczas należy zbadać, czy w analizowanym modelu spełnione jest omawiane założenie. Należy zbadać, czy reszty modelu podlegają autokorelacji pierwszego rzędu. W tym celu musimy zweryfikować hipotezę mówiącą, że współczynnik autokorelacji jest równy zero. Do jej weryfikacji posłużyć może najbardziej popularny test Durbina-Watsona postaci:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} {e_t}^2} \qquad (2)$$

gdzie: $e_t$ - reszty modelu,

$n$ - liczność próby.

Aby poprawnie stosować ten test, muszą być spełnione następujące założenia:

- analizowany model musi mieć wyraz wolny,
- składniki resztowe mają rozkład normalny,
- w modelu nie występuje opóźniona zmienna zależna w charakterze zmiennej niezależnej,
- liczba obserwacji jest większa od 15.

W oprogramowaniu [4] wartość testu Durbina-Watsona dla naszego przypadku przedstawia rysunek 4.

d Durbina-Watsona (MD7czas%dod) i korelacja seryjna reszt

| | **d Durbin - Watsona** | **Seryjna - Kor.** |
|---|---|---|
| **Estymac.** | 1,417284 | 0,257851 |

**Rys. 4. Arkusz wyników z wartością testu Durbina-Watsona dla zapalników MD-7**

Z rysunku 4 widać, że wartość testu Durbina-Watsona wynosi $d = 1{,}417284$. Otrzymaliśmy również estymator $\hat{\rho}$ współczynnika autokorelacji ( seryjna – kor ). Wartości te są powiązane, zachodzi bowiem w przybliżeniu równość:

$$d = 2 * (1 - \hat{\rho}) \qquad (3)$$

Wartość statystyki $d$ należy do przedziału $< 0, 4 >$. Wartość $d$ zbliżona do zera wskazuje na istnienie autokorelacji dodatniej, z kolei wartość $d$ bliska 4 wskazuje na autokorelację ujemną. Wreszcie wartość $d$ bliska 2 wskazuje na brak autokorelacji.

Dla naszego przypadku trudno przyjąć, że opracowany model regresji posiada wartości resztowe o własności braku autokorelacji. Wartość testu Durbina-Watsona odbiega od wartości 2, gdyż wynosi 1,417284.

Podczas stosowania testu Durbina-Watsona zdarzają się sytuacje, w których test ten nie rozstrzyga o istnieniu autokorelacji pierwszego rzędu reszt modelu. Można wówczas zastosować test mnożników Lagrange'a, który jest skomplikowany i z racji swej obszerności nie będzie opisywany w tym artykule.

W przypadku stwierdzenia autokorelacji w opracowanym modelu regresji, można postąpić na kilka sposobów:

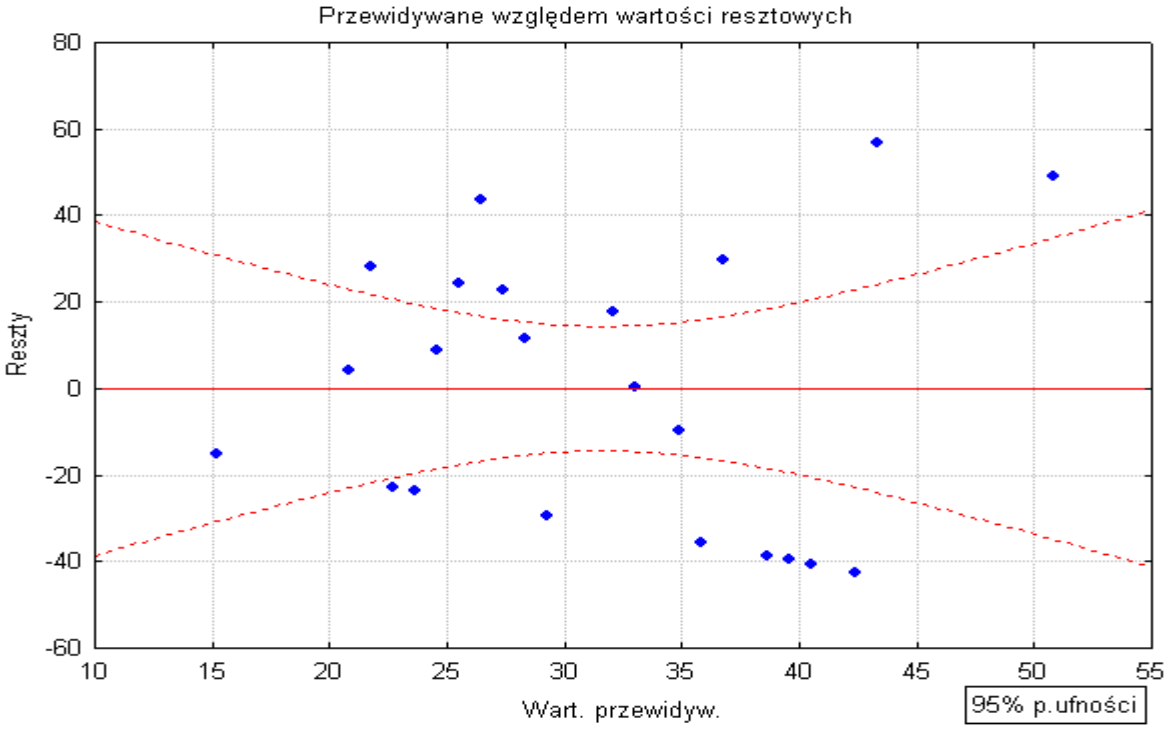
- przeanalizować raz jeszcze strukturę modelu – często autokorelację powoduje niezgodna postać funkcyjna modelu lub pominięcie ważnej zmiennej niezależnej,
- zastosować uogólnioną metodę najmniejszych kwadratów – jest to najczęściej stosowane wyjście, ale prowadzi często do utraty dokładności estymatora,
- zastosować inne bardziej złożone metody (interakcyjne, różniczki zupełnej, itp.), które znaleźć można w podręcznikach [2] i [3]. Dokładny ich opis nie będzie omawiany w tym artykule,
- nie robić nic - zostawić parametry modelu oszacowane metodą najmniejszych kwadratów pamiętając, że w tym przypadku nie są to estymatory efektywne.

## 5. Stałość wariancji wartości resztowych

Kolejną pożądaną własnością wartości resztowych jest założenie o homoscedastyczności składnika losowego. Założenie to brzmi następująco: wariancja składnika losowego (reszty $e_i$) jest taka sama dla wszystkich obserwacji ( war $(e_i) = \sigma^2$ dla wszystkich $i = 1, 2, \ldots, n$ ), czyli mówimy o stałości wariancji składnika losowego.

Naruszenie warunku homoscedastyczności nosi nazwę heteroscedastyczność. Najlepszym sposobem na sprawdzenie, czy heteroscedastyczność jest obecna, jest utworzenie odpowiednich wykresów rozrzutu. Jeśli spodziewamy się różnych $\sigma_i^2$ (wariancji składnika losowego) dla różnych $E(y_i)$, najlepiej sporządzić wykres rozrzutu reszt ( które są estymatorami składników losowych) względem wartości przewidywanych (które są estymatorami $E(y_i)$ ). Można też sporządzić wykres rozrzutu wartości przewidywanych względem kwadratów reszt, który w pewnych sytuacjach może okazać się bardziej przydatny.

Wykres rozrzutu reszt względem wartości przewidywanych dla zapalników MD-7 przedstawia rysunek 5.




**Rys. 5. Wykres rozrzutu reszt względem wartości przewidywanych dla zapalników MD-7**

Z rysunku 5 widać, że wartości reszt są nieznacznie bardziej zróżnicowane (rozrzucone) tylko dla niektórych wartości przewidywanych. Fakt ten nie wpływa na precyzję oszacowania. Widoczna jest chmurka punktów bez wyraźnej tendencji wzrostu (lub spadku) wariancji reszt przy wzroście wartości przewidywanej reszt. Można przyjąć, że założenie o stałości wariancji składnika losowego jest spełnione, czyli zjawisko homoscedastyczności jest zachowane.

Czasami wartość wariancji różni się dla jednej lub więcej zmiennych niezależnych. Jeżeli w opracowanym modelu regresji istnieje kilka zmiennych niezależnych, które mogą mieć wpływ na wariancję, to wówczas należy sporządzić wykresy rozrzutu reszt dla wszystkich zmiennych niezależnych tego modelu. Tylko pełny obraz rozrzutu reszt względem wartości przewidywanych dla wszystkich zmiennych niezależnych, daje możliwość stwierdzeniao stałości wariancji wartości resztowych dla opracowanego modelu regresji.

Czasami wykres reszt względem wartości przewidywanych może nasuwać pewne wątpliwości. Musimy wówczas zweryfikować hipotezę o stałości wariancji składnika losowego. Można posłużyć się wieloma testami sprawdzającymi. Zainteresowanych czytelników odsyłam do [ 2 ] lub [ 3 ].

W przypadku stwierdzenia heteroscedastyczności składnika losowego, można na kilka sposobów szukać estymatora bardziej efektywnego:

- zastosować uogólnioną metodę najmniejszych kwadratów,
- w przypadku, gdy wariancja zakłóceń ma tendencję wzrostu, można zastosować ważoną metodę najmniejszych kwadratów,
- zastosować odpowiednią transformację zmiennej zależnej np. logarytmiczną, pierwiastkową, odwrotnościową, kwadratową, itp.,
- nie robić nic, czyli zostać przy parametrach oszacowanych klasyczną metodą najmniejszych kwadratów, ale należy pamiętać, że nie jest to estymator efektywny.

## 6. Obserwacje nietypowe w analizie regresji

Po dopasowaniu równania regresji na podstawie wyników obserwacji należy zawsze przeanalizować wartości przewidywane i wartości reszt. W analizie regresji ważne jest to, aby opracowany model nie był nadmiernie uwarunkowany przez pojedyncze obserwacje o wartościach mocno różniących się od typowych dla danej próby. Takie odstające wartości mogą nam istotnie zakłócić wyniki obliczeń i prowadzić do błędnych wniosków. Czasami tylko usunięcie takiej jednej obserwacji wybawia nas z opresji. Z drugiej strony obserwacje niepasujące do modelu mogą wskazywać na braki w modelu lub złą postać algebraiczną modelu.

Dzięki oprogramowaniu [4] możemy wykryć takie obserwacje odstające. Wykorzystam w tym w celu reszty modelu. Należy przejrzeć wartości reszt i statystyki z nimi związane. Arkusz wyników z wyliczonymi wartościami reszt oraz innymi statystykami dla rozpatrywanego w artykule przykładu przedstawiono na rysunku 6.

Wartości przewidywane i reszty (MD7czas%dod) Zmienna zależna: Frakcja dec.dod.

| Nr obserwacji | Obserw. - Wartość | Przewidyw. - Wartość | Reszta | Standard - Przewid. | Standard - Reszta | Bł. std. - W.przew. | Mahaln. - Odległ. | Usunięte - Reszta | Cooka - Odległ. |
|---|---|---|---|---|---|---|---|---|---|
| | I | II | III | IV | V | VI | VII | VIII | IX |
| 1 | 0,0000 | 15,19100 | -15,1910 | -1,84422 | -0,46969 | 14,72966 | 3,401135 | -19,1664 | 0,036420 |
| 2 | 25,0000 | 20,81708 | 4,1829 | -1,20861 | 0,12933 | 10,96855 | 1,460743 | 4,7265 | 0,001228 |
| 3 | 50,0000 | 21,75476 | 28,2452 | -1,10268 | 0,87332 | 10,39774 | 1,215898 | 31,5010 | 0,049023 |
| 4 | 0,0000 | 22,69244 | -22,6924 | -0,99674 | -0,70163 | 9,85063 | 0,993498 | -25,0127 | 0,027741 |
| 5 | 0,0000 | 23,63012 | -23,6301 | -0,89081 | -0,73062 | 9,33139 | 0,793542 | -25,7757 | 0,026436 |
| 6 | 33,3333 | 24,56780 | 8,7655 | -0,78488 | 0,27102 | 8,84492 | 0,616030 | 9,4741 | 0,003209 |
| 7 | 50,0000 | 25,50548 | 24,4945 | -0,67894 | 0,75735 | 8,39693 | 0,460961 | 26,2649 | 0,022226 |

Ciąg dalszy tabeli

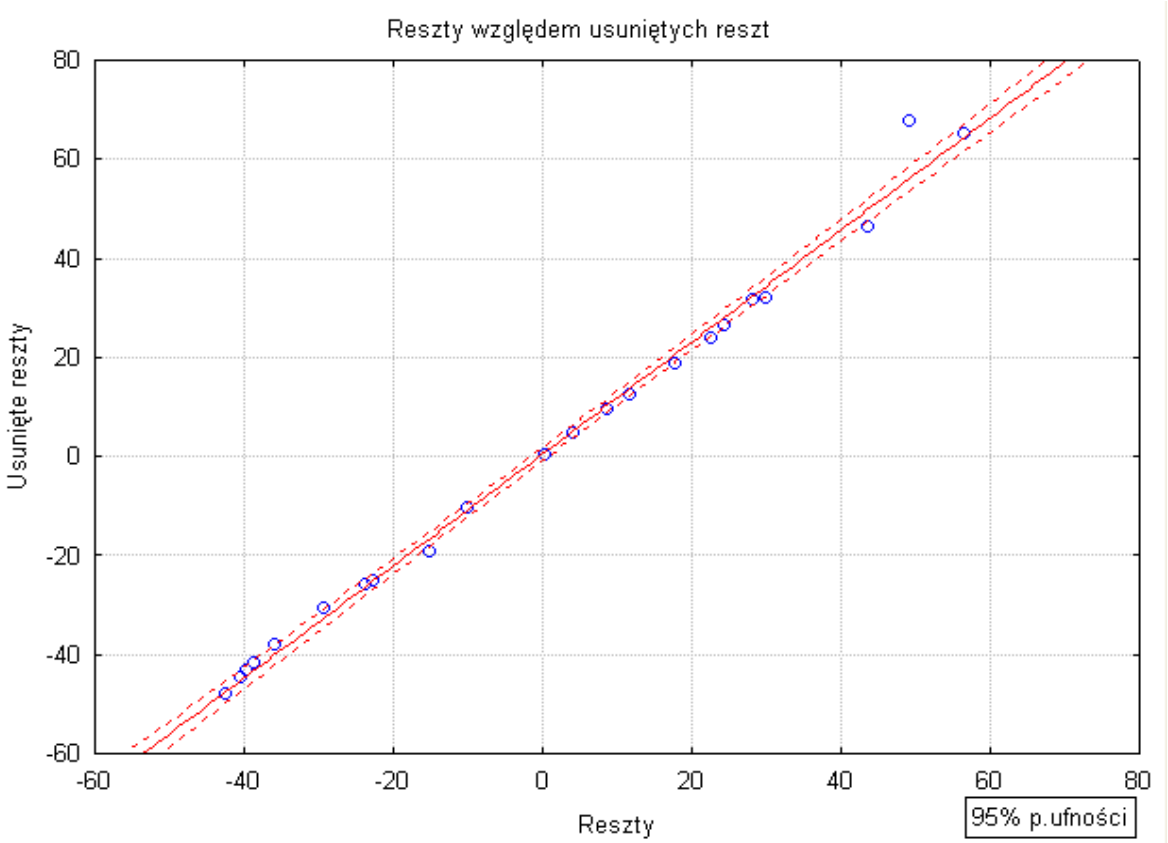
| | I | II | III | IV | V | VI | VII | VIII | IX |
|---|---|---|---|---|---|---|---|---|---|
| 8 | 70,0000 | 26,44316 | 43,5568 | -0,57301 | 1,34674 | 7,99389 | 0,328337 | 46,3908 | 0,062843 |
| 9 | 50,0000 | 27,38084 | 22,6192 | -0,46707 | 0,69936 | 7,64291 | 0,218157 | 23,9570 | 0,015320 |
| 10 | 40,0000 | 28,31852 | 11,6815 | -0,36114 | 0,36118 | 7,35145 | 0,130421 | 12,3179 | 0,003747 |
| 11 | 0,0000 | 29,25620 | -29,2562 | -0,25520 | -0,90457 | 7,12682 | 0,065130 | -30,7493 | 0,021945 |
| 12 | 50,0000 | 32,06924 | 17,9308 | 0,06260 | 0,55440 | 6,90960 | 0,003918 | 18,7883 | 0,007701 |
| 13 | 33,3333 | 33,00692 | 0,3264 | 0,16853 | 0,01009 | 6,99729 | 0,028403 | 0,3424 | 0,000003 |
| 14 | 25,0000 | 34,88227 | -9,8823 | 0,38040 | -0,30555 | 7,39968 | 0,144704 | -10,4281 | 0,002721 |
| 15 | 0,0000 | 35,81995 | -35,8199 | 0,48633 | -1,10752 | 7,70252 | 0,236521 | -37,9737 | 0,039093 |
| 16 | 66,6667 | 36,75763 | 29,9090 | 0,59227 | 0,92476 | 8,06351 | 0,350781 | 31,8913 | 0,030218 |
| 17 | 0,0000 | 38,63299 | -38,6330 | 0,80414 | -1,19449 | 8,93069 | 0,646635 | -41,8218 | 0,063745 |
| 18 | 0,0000 | 39,57067 | -39,5707 | 0,91007 | -1,22349 | 9,42351 | 0,828228 | -43,2416 | 0,075876 |
| 19 | 0,0000 | 40,50835 | -40,5084 | 1,01600 | -1,25248 | 9,94816 | 1,032265 | -44,7413 | 0,090527 |
| 20 | 0,0000 | 42,38371 | -42,3837 | 1,22787 | -1,31046 | 11,07459 | 1,507671 | -48,0132 | 0,129196 |
| 21 | 100,0000 | 43,32139 | 56,6786 | 1,33381 | 1,75245 | 11,66893 | 1,779040 | 65,1606 | 0,264183 |
| 22 | 100,0000 | 50,82283 | 49,1772 | 2,18128 | 1,52051 | 16,86859 | 4,757981 | 67,5534 | 0,593368 |
| Minimum | 0,0000 | 15,19100 | -42,3837 | -1,84422 | -1,31046 | 6,90960 | 0,003918 | -48,0132 | 0,000003 |
| Maksimum | 100,0000 | 50,82283 | 56,6786 | 2,18128 | 1,75245 | 16,86859 | 4,757981 | 67,5534 | 0,593368 |
| Średnia | 31,5151 | 31,51515 | -0,0000 | -0,00000 | -0,00000 | 9,43736 | 0,954545 | 0,5202 | 0,071217 |
| Mediana | 29,1667 | 30,66272 | 2,2547 | -0,09630 | 0,06971 | 8,88780 | 0,631332 | 2,5345 | 0,028980 |

**Rys. 6. Arkusz wyników z wartościami przewidywanymi i resztami dla zapalników MD-7**

W tabeli na rysunku 6 obliczone są poszczególne wartości statystyk dla rozpatrywanych zapalników typu MD-7. Poszczególne kolumny oznaczają:

- I – wartości obserwowane zmiennej zależnej,
- II – wartości przewidywane wyliczone z równania regresji,
- III – wartości reszt, czyli różnica między wartościami obserwowanymi a przewidywanymi,
- IV – standaryzowane wartości przewidywane,
- V – standaryzowane wartości resztowe,
- VI – błędy standardowe niestandaryzowanej wartości przewidywanej,
- VII – odległości Mahalanobisa. Odległość Mahalanobisa to odległość danej obserwacji od określonego centrum rozkładu lub środka ciężkości rozkładu. Jest ona uogólnieniem odległości euklidesowej, uwzględniającej skorelowane zmienne. Miara ta pozwala ustalić, czy dana obserwacja może być zaliczana do odstających,
- VIII – usunięte wartości resztowe
- IX – odległości Cooka. Miara ta jest miarą wpływu i-tego przypadku na współczynniki regresji ( mierzy stopień zmiany współczynników regresji, gdyby dany przypadek pominąć w obliczeniach współczynników ). O ile odległości Mahalanobisa mierzą odległość przypadku od środka ciężkości, wyznaczonego przez zmienne niezależne, a standaryzowane reszty od linii regresji, to odległość Cooka łączy te dwie odległości i przez to jest łączną miarą wpływu poszczególnych obserwacji na linię regresji.

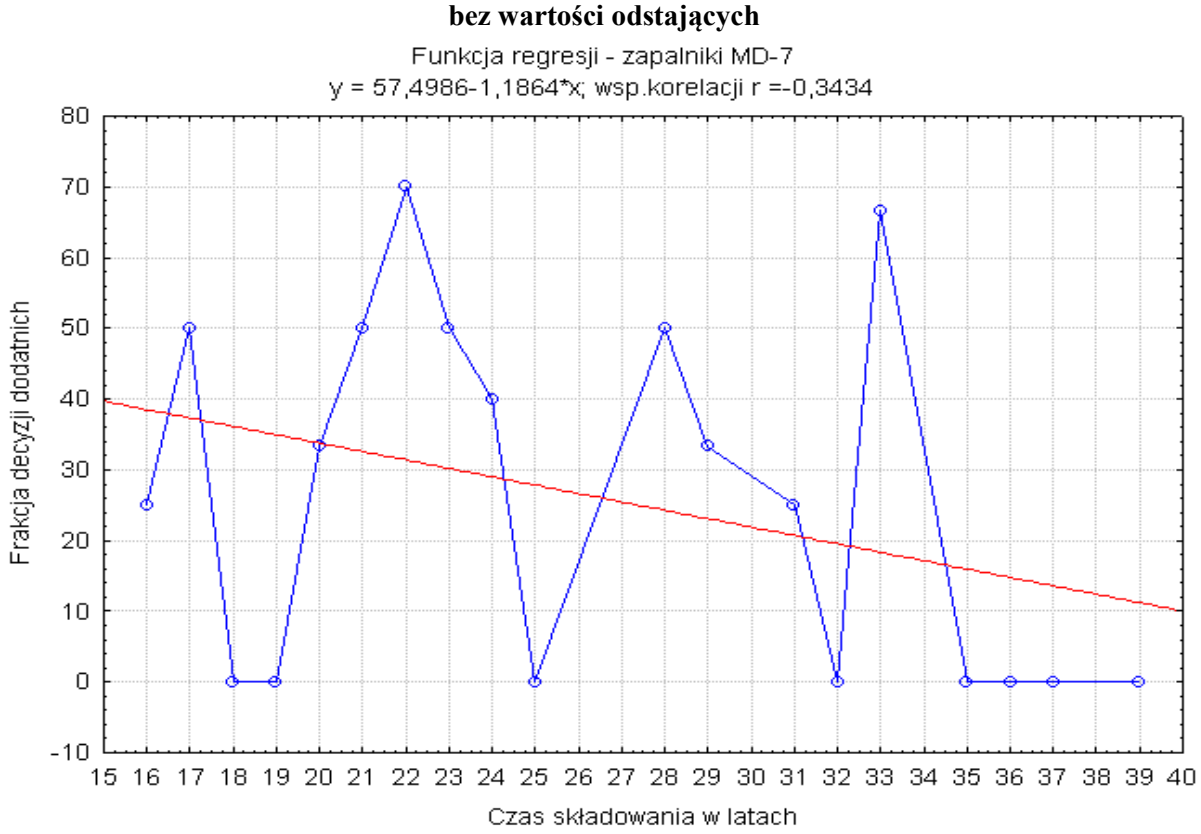
Analizując tabelę przedstawioną na rysunku 6, szukając wartości odstających można sytuację tą zinterpretować graficznie na wykresie rozrzutu reszt względem reszt usuniętych. Wykres ten przedstawiony został na rysunku 7.




**Rys. 7. Wykres rozrzutu reszt względem reszt usuniętych dla zapalników MD-7**

Z wykresu tego widać, że jedna obserwacja jest trochę odstająca. Jest to obserwacja 22. Należy zmienić model regresji i wyeliminować tę obserwację z modelu. Wówczas otrzymujemy trochę inną linię regresji, jednakże wykres rozrzutu reszt względem reszt usuniętych wykazuje kolejne dwie obserwacje, które nie leżą na linii. Są to obserwacje 1 i 21. Po ich usunięciu wykres rozrzutu reszt względem reszt usuniętych przestawiony został na rysunku 8.


**Rys. 8. Wykres rozrzutu reszt względem reszt usuniętych dla zapalników MD-7**




**Rys. 9. Linia regresji dla zapalników MD-7 bez obserwacji odstających**

Jak widać z wykresu, wszystkie analizowane obserwacje leżą na linii co oznacza, że model regresji został prawidłowo oszacowany. Linia regresji dla tego modelu została przedstawiona na rysunku 9. Po wyeliminowaniu obserwacji odstających, otrzymane współczynniki linii regresji zostały oszacowane prawidłowo.

Arkusz wyników testu Durbina-Watsona dla nowego modelu regresji przedstawiono na rysunku 10. Wartość ta wynosi $d$ = 1,897226, czyli w tym przypadku możemy powiedzieć o braku autokorelacji wartości resztowych dla naszego modelu regresji. Spełniona jest także zależność (3), różnica wynosi zaledwie 0,029252.

d Durbina-Watsona (MD7czas%dod) i korelacja seryjna reszt

| | d Durbin - Watsona | Seryjna - Kor. |
|---|---|---|
| **Estymac.** | 1,897226 | 0,036761 |

**Rys. 10. Arkusz wyników z wartością testu Durbina-Watsona bez obserwacji odstających**

## 7. Wnioski

Przedstawiona analiza wartości resztowych, po oszacowaniu modelu regresji jak widać jest elementem koniecznym. W trakcie jej opracowywania uzyskujemy bardzo ważną informację dotyczącą prawidłowości oszacowania współczynników regresji. Zauważone obserwacje odstające, jakie mogą wystąpić w opracowywanym modelu regresji, mają jak widać z artykułu, znaczący wpływ na cały proces regresji.

Z wykresu 9 widać, że w miarę upływu lat składowania, frakcja decyzji dodatnich maleje, czyli stan jakościowy zbioru przechowywanych zapalników typu MD-7 ulega pogorszeniu. Świadczy o tym ujemna wartość współczynnika korelacji liniowej wynosząca r=-0,3434.

Pierwsza uzyskana linia regresji pokazuje nam całkiem inny trend opracowywanego modelu. Z wykresu ( rys. 1 ) można by wysnuć wniosek, że stan jakościowy analizowanego zbioru zapalników typu MD-7 ulega poprawie w miarę lat składowania. Oczywiście wniosek taki jest błędny, co pokazuje analiza wartości resztowych przedstawiona w artykule.

Wspomniana na początku w artykule uwaga, dotycząca braku ciągłości wyników badań, okazała się trafna. Opracowany model regresji dla zapalników typu MD-7 wykazał, że aby proces regresji mógł być prawidłowy musi być ciągłość wyników badań.

Reasumując, opracowywany model regresji dla każdego analizowanego zbioru obserwacji wymaga, aby analizę tego modelu przeprowadzić łącznie z analizą wartości resztowych tego modelu. Nie do końca oszacowany model regresji prowadzi do błędnie opracowanych wniosków, a co za tym idzie wpływa bardzo znacząco na opracowywany proces predykcji, dotyczący analizowanego zbioru danych.

## Literatura

[1] A. Stanisz – *Przystępny kurs statystyki* – Statsoft Polska, Kraków 2007 r..

[2] J. Gajda – *Ekonometria praktyczna* – Przedsiębiorstwo Specjalistyczne „Absolwent" Sp. z o.o., Łódź 2002 r..

[3] A. Welfe – *Ekonometria* – Polskie Wydawnictwo Ekonomiczne, Warszawa 2003 r..

[4] *Statistica 9* – Statsoft Polska 2009 r. – oprogramowanie komputerowe.

[5] M. Sobczyk – *Prognozowanie, teoria, przykłady, zadania* – Wydawnictwo Placet, Warszawa 2008 r..

[6] M. Gruszczyński, M. Podgórska – *Ekonometria* – Szkoła Główna Handlowa w Warszawie, Warszawa 2004 r.

# RESIDUUM VALUES IN THE REGRESS PROCESS

Dariusz AMPUŁA
Military Institute of Armament Technology

**Abstract:** In the introduction of the article the author presents the need of verifying a developed regress model by using the analysis of residuum values. On the beginning the presumptions of the residuum analysis are characterized, paying the attention to properties of these residuum i.e.: normality, constancy of variance and lack of autocorrelation. Tests results of artillery fuses type MD-7 were taken to analysis. A property of residuum values showing that residuum of a model have a normal distribution was characterized. Then a way of determining the autocorrelation of residuum values was described, taking into account Durbin-Watson's test, which is used to verify the autocorrelation coefficient. Due to the extensiveness of the article, the method of Lagrange multipliers is not analyzed.
A property of variance residuum values constancy i.e. presumption on homoscedasticity of random component is described. The residuum scatter diagrams in relation to forecast values are presented to prove it. A way defining atypical observations in the analysis of regress process was outlined. A graphic figure of interpretation for atypical observations by using residuum scatter diagram in relation to deleted residuum is discussed. Concise conclusions relating to the residuum values in the regress process are presented at the end of the article.

Keywords: residuum value, normality, variance, autocorrelation, coming off observation

## 1. Introduction

The majority of the presumptions of the regress process relates to residuum values typically called the residuum. Analyzing the residuum [1], we can quickly and effectively detect all departures which could happen from correct analysis of regress process. Although, we cannot check all presumptions relating to the regress process, however we can detect the largest departures and alternatively eliminate them. The analysis of distribution of residuum values should be one of the most important stages verifying the regress model. Studying the model residuum, we can also detect coming off observations quickly. Such observations can in the serious way disturb the regress equation through “straining” the regress line towards their direction. This causes a change of the regress coefficients. It may happen that removing such coming off observation data point gives completely various results of regress analysis and basing on this we foresee another values of prediction process.

If we want to receive a relatively correct regress model, we always have to analyze received residuum values after estimation and verification of this model. Only recognition their graphs and statistics connected with them, guarantees a quick detection of the departures and suitable statistical interpretation. The analysis of the residuum values should always become the rule after the estimation of regress model parameters.

## 2. Presumptions of the model

The analysis of residuum according to [1], should begin from the most important matter i.e. checking presumptions of the classic method of the smallest squares. This is because the correctly constructed model is characterized by certain desirable properties of the residuum (such as normality, constancy of variance, lack of the autocorrelation). The procedure of checking presumptions of the model is applied *ex post*, i.e. after estimation of model parameters by the classic method of smallest squares. Then, when it turns out, that some of these conditions are not fulfilled, the parameters of this model need to be estimated again, applying another method of estimation or another figure of the model.

We will take the results of MD-7 fuses diagnostic tests which were obtained during laboratory tests until 2010 to the analysis of residuum values. The fuses of this type were applied for artillery ammunition with anti-tank-tracer projectiles of various calibers. The quantities of tested lots in different years, negative decisions and positive decisions obtained after laboratory tests and values of the fraction for positive decisions were introduced in the table 1. Decisions "B5" and "B3" were accepted as positive decisions whereas all remaining decisions were accepted as negative decisions.

## 3. Normality of residuum values

The first among presumptions of residuum values is one saying that the residuum of the model has the normal distribution. This presumption is not necessary for adapting the model of the smallest squares, but for the verification of received parameters significance. Tests applied in this aim (t-distribution, F) are resistant against small deviations from normality. If presumption about normality is evidently infringed, the evaluations of regress coefficients significance can be disturbed. Then we can execute transformation of dependent variable – for example $logY$ or $\sqrt{Y}$ – to receive the normal distribution.

Analyzing data from table 1, we can see that there is a lack of continuity of tests results. The oldest tested fuses were 48 years old, but there is a lack of tests results for fuses aged between 41÷47 years. The similar situation is for fuses which have been stored for 11÷15 years. The lack of continuity in diagnostic results can evidently change the real parameters of regress model, what in consequence leads to a possibility of drawing some incorrect conclusions relating to prediction process for this type fuses.

Table 1 – MD-7 fuses

| Storage time | The quantity of negative decisions | The quantity of positive decisions | Fraction of positive decisions | Quantity of tested parties (lots) |
|---|---|---|---|---|
| 10 | 1 | 0 | 0,00 | 1 |
| 16 | 3 | 1 | 25,00 | 4 |
| 17 | 3 | 3 | 50,00 | 6 |
| 18 | 1 | 0 | 0,00 | 1 |
| 19 | 2 | 0 | 0,00 | 2 |
| 20 | 2 | 1 | 33,33 | 3 |
| 21 | 1 | 1 | 50,00 | 2 |
| 22 | 3 | 7 | 70,00 | 10 |
| 23 | 3 | 3 | 50,00 | 6 |
| 24 | 3 | 2 | 40,00 | 5 |
| 25 | 4 | 0 | 0,00 | 4 |
| 28 | 1 | 1 | 50,00 | 2 |
| 29 | 2 | 1 | 33,33 | 3 |
| 31 | 3 | 1 | 25,00 | 4 |
| 32 | 1 | 0 | 0,00 | 1 |
| 33 | 1 | 2 | 66,67 | 3 |
| 35 | 1 | 0 | 0,00 | 1 |
| 36 | 1 | 0 | 0,00 | 1 |
| 37 | 2 | 0 | 0,00 | 2 |
| 39 | 1 | 0 | 0,00 | 1 |
| 40 | 0 | 1 | 100,00 | 1 |
| 48 | 0 | 3 | 100,00 | 3 |
| Total | 39 | 27 | 40,91 | 66 |

Figure 1 represents the positive decision fraction dependence on storage time for data displayed in table 1. The regress line is marked by red in this figure

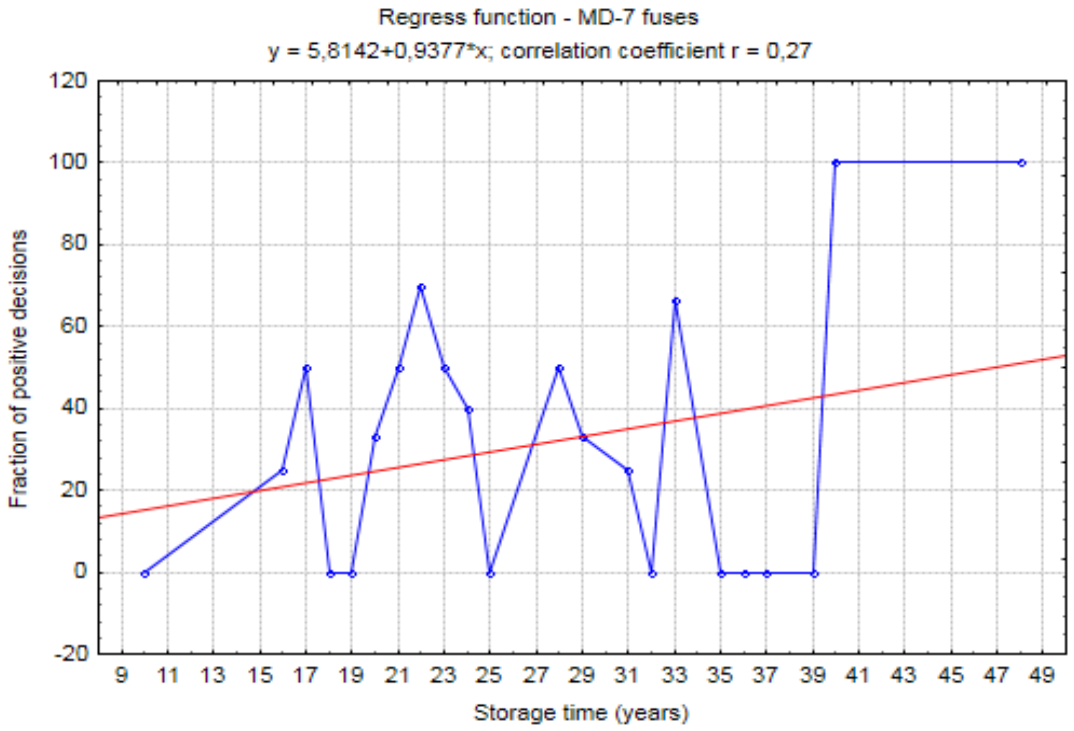



**Fig. 1. Regress line for MD-7 fuses**

In order to obtain a graphic normality checking of this regress model, the software [4] will be used. The graph of the residuum normality for this model is represented in figure 2.

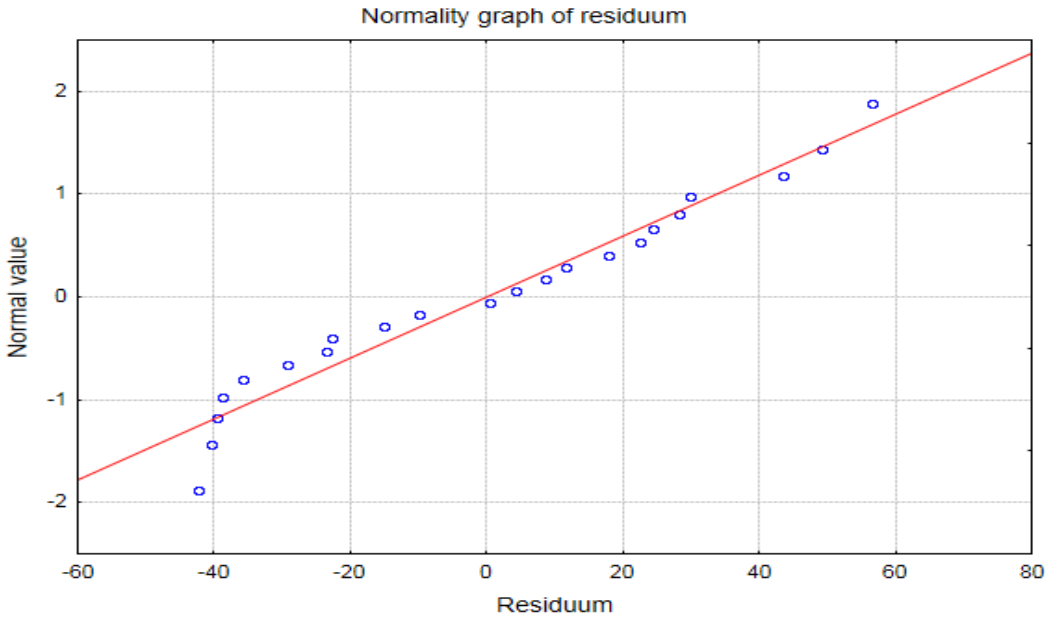



**Fig. 2. Normality graph of residuum values for MD-7 fuses**

The received graph enables a visual examination of residuum compliance with normal distribution - if the residuum has no normal distribution, then points deviate from the straight line, if points create a shape around the line, then it suggests use of some transformation. For example, if points create the S letter shape, we can apply logarithmic transformation. The graph of the residuum normality can also disclose coming off observations.

We can see in this case, that points are situated along the straight line, confirming the normality of residuum distribution. We can have some objections relating to the first observation, because it is a bit off from the line, but this distance has not influenced significantly the normality of residuum values.

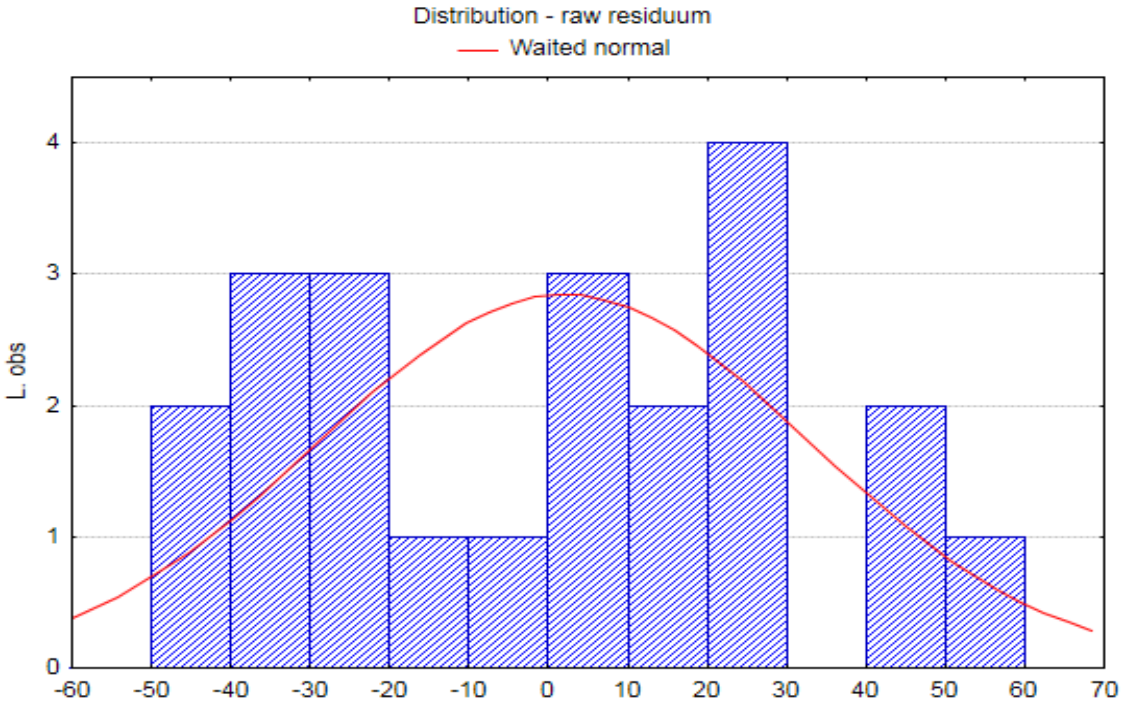



**Fig. 3. Residuum histogram for MD-7 fuses**

The similar information as showed on the graph of probability normality also delivers the residuum histogram. The normal line should cross in perfect situation the column upper edge centers. The small aberration from normality is not dangerous, especially for large number populations. The size of the population in this example was 66 observations. The figure 3 represents residuum histogram for MD-7 type fuses. It can be seen from the figure that this is not a perfect situation because the line deviates from upper column edges and it cannot be straightly accepted that there is a case of normality of residuum values.

## 4. Autocorrelation of residuum values

The lack of random component autocorrelation is the next presumption for residuum values. It is as follow: random components (residuum) are not correlated, that is $e_i$ and $e_j$ are not correlated with themselves for all couples $i, j$ where $i, j = 1,2, \dots, n$ and $i \neq j$. When this presumption is not fulfilled then there is a residuum autocorrelation.

The autocorrelation may happen only in cases when observations in the population (sample) are ordered (usually by the time). This presumption relates first of all to time series. In our case the consecutive observations depend on the storage time and carried out diagnostic tests what means that it is an ordered time series.

The existence of the autocorrelation makes the estimators received by classic method of the smallest squares stop to be the most effective estimators. They can be burdened with an additional error and for this become less precise and not so much useful. The autocorrelation of the first grade happens mostly. It means that residuum component $e_t$ depends on:

- component $e_{t-1}$,
- a new random component $\varepsilon_t$.

So dependence arises:

$$e_t = \rho\, e_{t-1} + \varepsilon_t \qquad (1)$$

where: $\rho$ – the coefficient of autocorrelation. It accepts values from range $< -1,1 >$.
If $\rho > 0$, we can talk about the positive autocorrelation, in the opposite case we talk about negative autocorrelation.

A problem occurs in doubtful situations. Then it should be checked if the discussed presumption is fulfilled in the analyzed model. It should be tested if residuum model is the subject of the first grade autocorrelation. To do so it is necessary to verify the hypothesis stating that autocorrelation coefficient is zero. This verification can be done by using the most popular Durbin-Watson's test in the form:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} {e_t}^2} \quad (2)$$

where: $e_t$ – residuum model,
$n$ - quantity of sample.

To apply this test correctly, following presumptions have to be fulfilled:
- the analyzed model has to have a free term,
- the residuum components have the normal distribution,
- in the model there is no delayed dependent variable in the character of independent variable,
- number of observations is greater than 15.

The value of Durbin-Watson's test in software [4] is presented for the considered case in figure 4.

d Durbin-Watson (MD7time%positive) and serial residuum correlation

| | d Durbin - Watson | Serial - Cor. |
|---|---|---|
| **Estimation** | 1,417284 | 0,257851 |

**Fig. 4. Results sheet with value of Durbin-Watson's test for MD-7 fuses**

From figure 4 it can be seen that value of Durbin-Watson's test is $d = 1{,}417284$. We also received estimator $\hat{\rho}$ autocorrelation coefficient ( serial – cor. ). These values are related as there is an approximate equation:

$$d = 2 * (1 - \hat{\rho}) \quad (3)$$

The value of statistic $d$ belongs to the range $< 0, 4 >$. The $d$ value approximates to zero then there is a positive autocorrelation, value $d$ close 4 shows on the negative autocorrelation. At last value $d$ close 2 shows on the lack of autocorrelation.

It is hard to accept in considered case, that elaborated regress model has residuum values characterized by the lack of autocorrelation. The value of Durbin-Watson's test somewhat deviates from value 2, because it is $1{,}417284$.

While applying Durbin-Watson test some situations may happen in which this test does not decide about existence of the first grade autocorrelation model residuum. Then we can apply Lagrange's multipliers test which is a bit complicated and because of its extensiveness it is not considered in this article.

In the case of confirmed autocorrelation in elaborated regress model there are possibilities to:
- analyze again the model structure – often autocorrelation is caused discordant figure of functional model or omission of important independent variable,
- apply the generalized method of smallest squares – this is the most applied solution but it often leads to losing the precision of estimator,
- apply another more complicated methods (interactive, total differential, etc.) which can be found in the literature [ 2 ] and [ 3 ]. Their exact description is not included in this article,
- do nothing – leaving model parameters estimated by the method of the smallest squares and remembering that in this case they are not the effective estimators.

## 5. Stability of residuum values variance

The next desirable property of residuum values is a presumption about homoscedasticity of random component. This presumption is as follows: variance of random component (residuum $e_i$) is the same for all observations ( var. $(e_i) = \sigma^2$ for all $i = 1, 2, \ldots, n$ ), so we can talk about the variance constancy of random component.

The infringement of homoscedasticity condition is called as heteroscedasticity. The best way of examining if there is heteroscedasticity is to create suitable distribution graphs. If we expect various $\sigma_i^2$ ( variance of random component ) for various $E(y_i)$, the best way is to prepare the graph of residuum distribution ( which are estimators of random components ) in relation to foreseen values ( which are estimators of $E(y_i)$ ). We can also prepare the graph of foreseen values distribution in relations to the squares of residuum, which in certain situations may be more useful. The graph of residuum distribution in relation to foreseen values for MD-7 fuses is presented on figure 5.

We can see from the figure 5 that residuum values are a bit diverse (scattered) only for some foreseen values. This fact does not influence on estimation precision. The cloud of points is visible without clear growth tendency ( or fall ) of residuum variance at the growth of foreseen residuum value. We can accept, that presumption about variance constancy of random component is fulfilled i.e. the homoscedasticity phenomenon is observed.

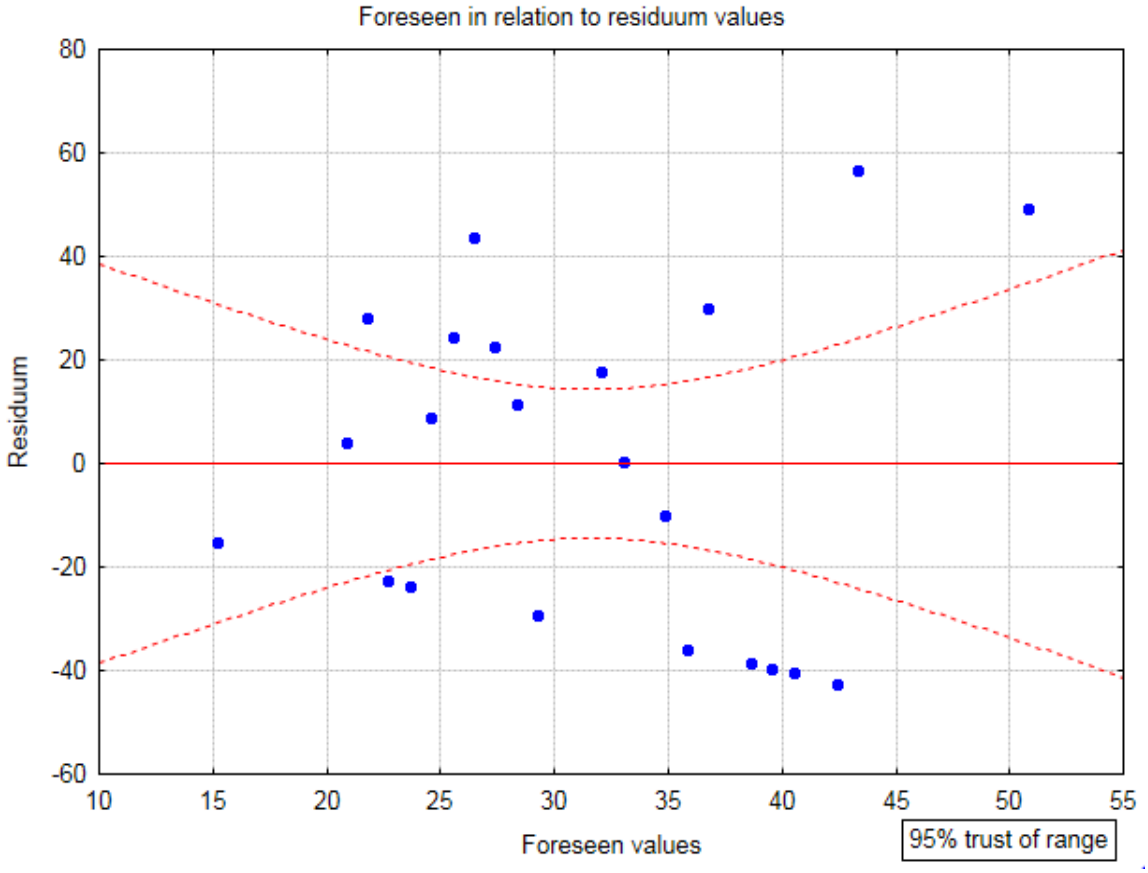



**Fig. 5. Graph of residuum distribution in relation to foreseen values for MD-7 fuses**

The variance value sometimes differs for one or more independent variables. If in elaborated regress model there are some several independent variables which can influence on variance, then it should be prepared the graphs of residuum distribution for all independent variables of this model. Only a full picture of residuum distribution in relation to foreseen

values for all independent variables gives a possibility of confirming the variance constancy of residuum values for elaborated regress model.

The residuum graph in relation to foreseen values can sometimes suggest certain doubts. Then we have to verify the hypothesis about the variance constancy of random component. Many checking tests can be used. More information for interested readers is in [ 2 ] or [ 3 ].

In case of stating the heteroscedasticity of random component we can look for a more effective estimator in several ways by:

- applying the generalized method of the smallest squares,
- applying the weighed method of the smallest squares in the case when disturbances variance has a growing tendency,
- applying a suitable transformation of dependent variable e.g. logarithmic, square root, reverse, square, etc.,
- doing nothing i.e. staying with parameters estimated by the classic method of the smallest squares and having in mind that this estimator is not effective.

## 6. Atypical observation in regress analysis

After adapting equation of regression on the basis of observation results the foreseen values and residuum values have to be always analyzed. It is important in the analysis of regress that the elaborated model is not excessively conditioned by single observations possessing values differing significantly from typical for a given sample. Such coming off values can indeed disturb the results of calculations and lead to incorrect conclusions. Sometimes this one observation has to be deleted to prevent such case. From the other side observations that do not match to this model can show its drawbacks or wrong algebraic form.

Thanks to software [ 4 ] we can detect such coming off observations. The residuum of model can be used for this reason. The residuum values and statistics connected with them have to be reviewed. The sheet of results with calculated residuum values and different statistics for the example considered in the article was introduced on figure 6.

Specific values of statistics for considered fuses MD-7 are calculated and presented in the table on the figure 6. The individual columns signify:

- I – observed values of dependent variable,
- II – foreseen values calculated from the regress equation,
- III – residuum values i.e. a difference between observed and foreseen values,
- IV – standardised foreseen values,
- V – standardised residuum values,
- VI – standard errors for no-standardised foreseen value,
- VII – the Mahalanobis distances. The distance of Mahalanobis is the distance of a given observation to a specific distribution centre or its gravity centre. It is a generalization of Euclid's distance complying correlated variables. This gauge evaluates if given observation can be ranked into coming off,
- VIII – expelled residuum values,
- IX – the Cook's distances. This gauge is the gauge of „i" case influence on regress coefficients ( it measures degree of changes for regress coefficients if a given case is skipped in calculations of coefficients ). Whereas the Mahalanobis distances measure the distance of the case from the gravity centre, determined through independent variables, and the standardized residuum from the regress line, the Cook distance bonds these two distances and thanks to this it is a total gauge showing an influence of individual observations on the regress line.

Foreseen values and residuum (MD7time%pos) Dependent variable: Frakcion pos.dec.

*Residuum Values in the Regress Process*

| No. of the observat. | Observ. - Value | Foreseen - Value | Resid. | Standard - Foreseen | Standard – Resid. | Std.error V.foresee | Mahaln. - distance | Expeled – Resid. | Cook's - distance |
|---|---|---|---|---|---|---|---|---|---|
| | I | II | III | IV | V | VI | VII | VIII | IX |
| 1 | 0,0000 | 15,19100 | -15,1910 | -1,84422 | -0,46969 | 14,72966 | 3,401135 | -19,1664 | 0,036420 |
| 2 | 25,0000 | 20,81708 | 4,1829 | -1,20861 | 0,12933 | 10,96855 | 1,460743 | 4,7265 | 0,001228 |
| 3 | 50,0000 | 21,75476 | 28,2452 | -1,10268 | 0,87332 | 10,39774 | 1,215898 | 31,5010 | 0,049023 |
| 4 | 0,0000 | 22,69244 | -22,6924 | -0,99674 | -0,70163 | 9,85063 | 0,993498 | -25,0127 | 0,027741 |
| 5 | 0,0000 | 23,63012 | -23,6301 | -0,89081 | -0,73062 | 9,33139 | 0,793542 | -25,7757 | 0,026436 |
| 6 | 33,3333 | 24,56780 | 8,7655 | -0,78488 | 0,27102 | 8,84492 | 0,616030 | 9,4741 | 0,003209 |
| 7 | 50,0000 | 25,50548 | 24,4945 | -0,67894 | 0,75735 | 8,39693 | 0,460961 | 26,2649 | 0,022226 |
| 8 | 70,0000 | 26,44316 | 43,5568 | -0,57301 | 1,34674 | 7,99389 | 0,328337 | 46,3908 | 0,062843 |
| 9 | 50,0000 | 27,38084 | 22,6192 | -0,46707 | 0,69936 | 7,64291 | 0,218157 | 23,9570 | 0,015320 |
| 10 | 40,0000 | 28,31852 | 11,6815 | -0,36114 | 0,36118 | 7,35145 | 0,130421 | 12,3179 | 0,003747 |
| 11 | 0,0000 | 29,25620 | -29,2562 | -0,25520 | -0,90457 | 7,12682 | 0,065130 | -30,7493 | 0,021945 |
| 12 | 50,0000 | 32,06924 | 17,9308 | 0,06260 | 0,55440 | 6,90960 | 0,003918 | 18,7883 | 0,007701 |
| 13 | 33,3333 | 33,00692 | 0,3264 | 0,16853 | 0,01009 | 6,99729 | 0,028403 | 0,3424 | 0,000003 |
| 14 | 25,0000 | 34,88227 | -9,8823 | 0,38040 | -0,30555 | 7,39968 | 0,144704 | -10,4281 | 0,002721 |
| 15 | 0,0000 | 35,81995 | -35,8199 | 0,48633 | -1,10752 | 7,70252 | 0,236521 | -37,9737 | 0,039093 |
| 16 | 66,6667 | 36,75763 | 29,9090 | 0,59227 | 0,92476 | 8,06351 | 0,350781 | 31,8913 | 0,030218 |
| 17 | 0,0000 | 38,63299 | -38,6330 | 0,80414 | -1,19449 | 8,93069 | 0,646635 | -41,8218 | 0,063745 |
| 18 | 0,0000 | 39,57067 | -39,5707 | 0,91007 | -1,22349 | 9,42351 | 0,828228 | -43,2416 | 0,075876 |
| 19 | 0,0000 | 40,50835 | -40,5084 | 1,01600 | -1,25248 | 9,94816 | 1,032265 | -44,7413 | 0,090527 |
| 20 | 0,0000 | 42,38371 | -42,3837 | 1,22787 | -1,31046 | 11,07459 | 1,507671 | -48,0132 | 0,129196 |
| 21 | 100,0000 | 43,32139 | 56,6786 | 1,33381 | 1,75245 | 11,66893 | 1,779040 | 65,1606 | 0,264183 |
| 22 | 100,0000 | 50,82283 | 49,1772 | 2,18128 | 1,52051 | 16,86859 | 4,757981 | 67,5534 | 0,593368 |
| Minimum | 0,0000 | 15,19100 | -42,3837 | -1,84422 | -1,31046 | 6,90960 | 0,003918 | -48,0132 | 0,000003 |
| Maximum | 100,0000 | 50,82283 | 56,6786 | 2,18128 | 1,75245 | 16,86859 | 4,757981 | 67,5534 | 0,593368 |
| Average | 31,5151 | 31,51515 | -0,0000 | -0,00000 | -0,00000 | 9,43736 | 0,954545 | 0,5202 | 0,071217 |
| Median | 29,1667 | 30,66272 | 2,2547 | -0,09630 | 0,06971 | 8,88780 | 0,631332 | 2,5345 | 0,028980 |

**Fig. 6. The sheet of results with foreseen values and residuum for MD-7 fuses**

Analyzing table introduced on the figure 6 to find out coming off values a graph may be prepared that shows the residuum distribution in relation to expelled residuum. The graph is presented in figure 7.

From this graph we can see that there is one coming off observation enumerated as twenty second. We have incorporated a change into the regress model and eliminated this observation from model. Then we received a bit different regress line, however the graph of residuum distribution in relation to expelled residuum shows that there are next two observations which

not lie on the line. These are the first and twenty first observations. After their expelling the graph of residuum distribution in relation to expelled residuum was presented on the figure 8.

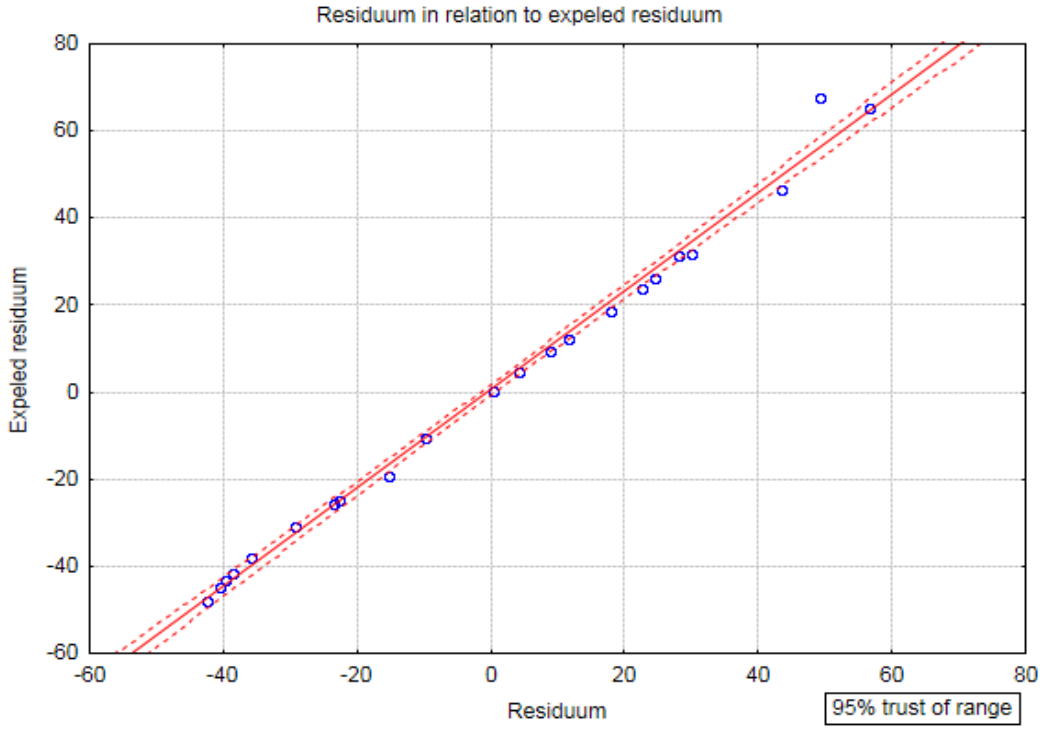



**Fig. 7. The graph of residuum distribution in relation to expelled residuum for MD-7 fuses**

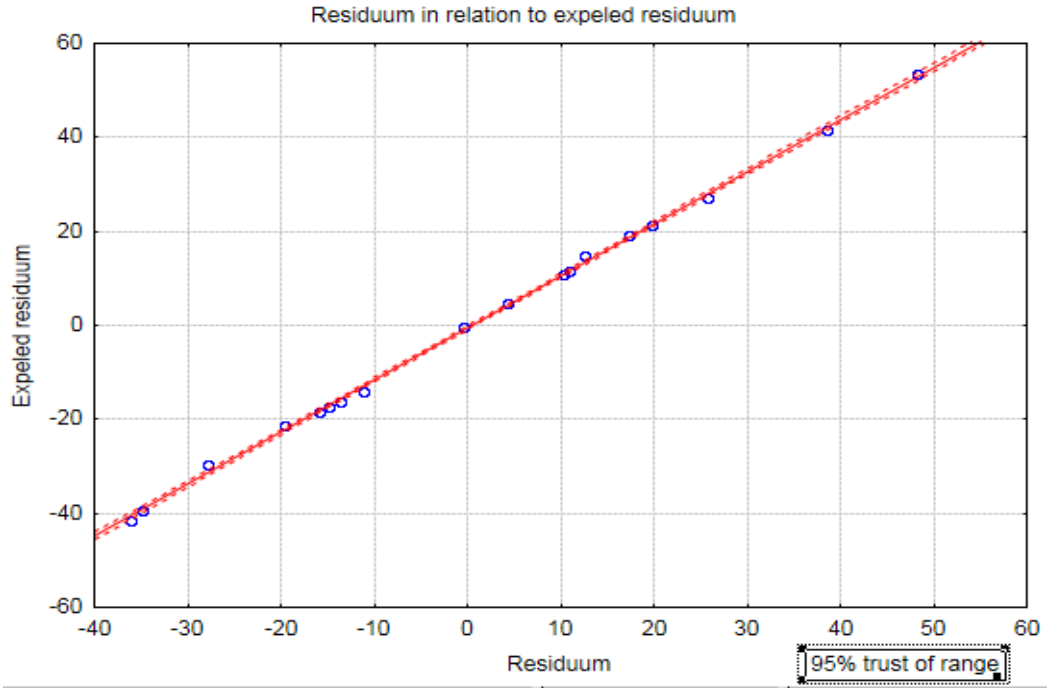



**Fig. 8. The graph of residuum distribution in relation to expelled residuum for MD-7 fuses without coming off values**

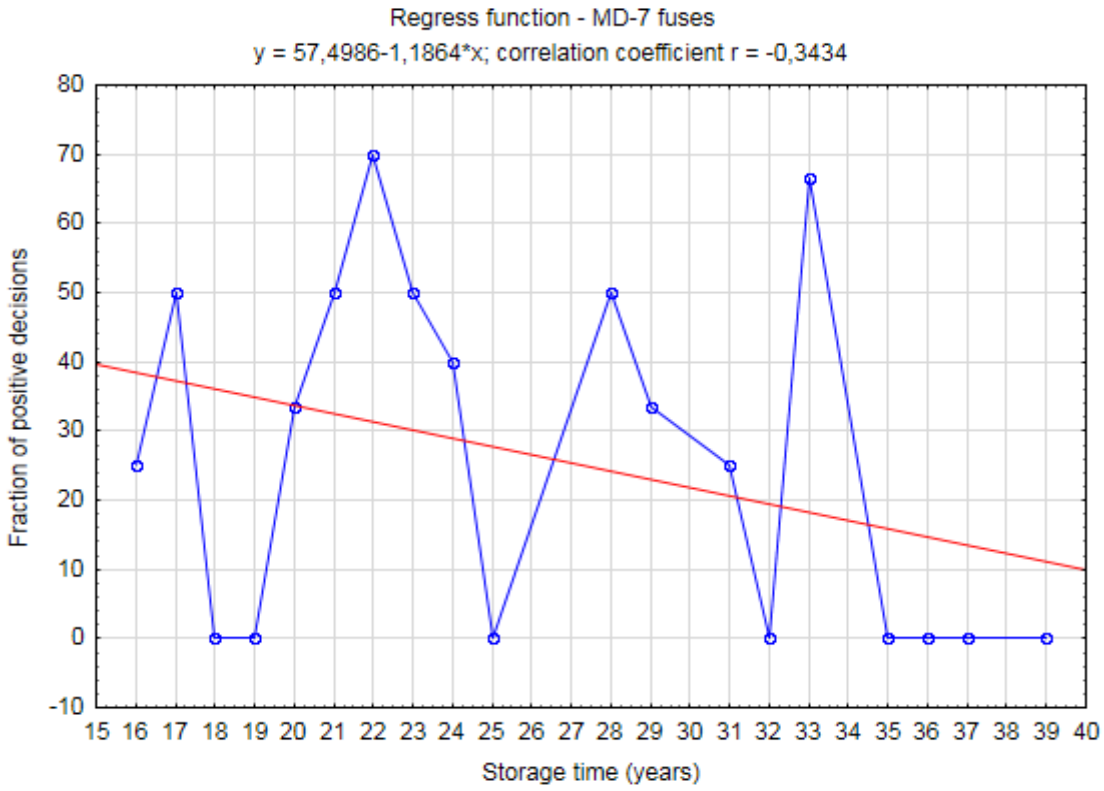



**Fig. 9. Regress line for MD-7 fuses without coming off observations**

As we see from the graph, all analyzed observations lie on the line what means, that regress model was estimated properly. The regress line for this model was introduced on the figure 9. After elimination coming off observations the received coefficients of regress line were properly estimated.

The results sheet of Durbin-Watson test for the new regress model was introduced on the figure 10. This value is $d = 1{,}897226$, so in this case we can say about lack of residuum values autocorrelation for our regress model. The dependence (3) is also fulfilled and the difference is only $0{,}029252$.

d Durbin-Watson (MD7time%positive) and serial residuum correlation

| | **d Durbin - Watson** | **Serial - Cor.** |
|---|---|---|
| **Estimation** | 1,897226 | 0,036761 |

**Fig. 10. Results sheet with value of Durbin-Watson test without coming off observations**

## Conclusions

The introduced analysis of residuum values, after estimation of regress model as we see is a necessary element. We can obtain very important information connected with the regularities of estimation of regress coefficients. The noticed coming off observations which can happen in elaborated regress model, as we could see from the article have a significant influence on whole regress process.

We can see from graph 9 that with the passing storage time the fraction of positive decisions diminish what means that qualitative condition of the set of stored fuses MD-7 type deteriorates. The negative value of linear correlation coefficient is $r = -0{,}3434$ and testifies this. The first obtained regress line shows us quite a different trend of elaborated model. The conclusion can be inferred from the graph ( fig. 1 ) that a qualitative condition of the set of

analyzed fuses MD-7 improves with the passing of storing time. Obviously such conclusion is incorrect what is showed by residuum values analysis introduced in this article.

The comment mentioned on the beginning of the article relating the lack of test results continuity has turned out to be justified. Elaborated regress model for fuses MD-7 type has showed that to get a correct regress process the continuity of test results is required.

Recapitulating, elaborated regress model for every analyzed observation set requires conducting the analysis of this model together with the analysis of its residuum values. The regress model that is not completely estimated leads to wrong conclusions and influences very significantly on elaborated prediction process that relates to the analyzed data set.

## Literature

[ 1 ] A. Stanisz – *Przystępny kurs statystyki* – Statsoft Polska, Kraków 2007 r..

[ 2 ] J. Gajda – *Ekonometria praktyczna* – Przedsiębiorstwo Specjalistyczne „Absolwent" Sp. z o.o., Łódź 2002 r..

[ 3 ] A. Welfe – *Ekonometria* – Polskie Wydawnictwo Ekonomiczne, Warszawa 2003 r..

[ 4 ] *Statistica 9* – Statsoft Polska 2009 r. – oprogramowanie komputerowe.

[ 5 ] M. Sobczyk – *Prognozowanie, teoria, przykłady, zadania* – Wydawnictwo Placet, Warszawa 2008 r..

[ 6 ] M. Gruszczyński, M. Podgórska – *Ekonometria* – Szkoła Główna Handlowa w Warszawie, Warszawa 2004 r..