ISSN 1230-3801 Zeszyt 130 nr 2/2014, 67-78

REGRESJA JAKO METODA PROCESU PREDYKCJI

Dariusz AMPUŁA
Wojskowy Instytut Techniczny Uzbrojenia

Streszczenie: W artykule autor we wstępie przedstawia krótki rys historyczny a następnie zapoznaje czytelnika z metodą analizy regresji. Na początku scharakteryzowano funkcję regresji I i II rodzaju po czym na przykładzie podjętych decyzji podiagnostycznych zapalników typu B-23U, przedstawiono sposób wyznaczenia tych funkcji. Scharakteryzowano klasyczny model regresji liniowej oraz za pomocą ww. wyników badań przestawiono postać graficzną i analityczną funkcji regresji łącznie z wyliczonymi przedziałami ufności. Zgodnie z procedurą wyznaczania linii regresji dokonano także estymacji parametrów modelu regresji oraz weryfikacji tego modelu. Zaprezentowano również metodę określania predykcji na podstawie modelu regresji liniowej dla analizowanych w artykule zapalników typu B-23U. Przedstawiono graficzną interpretację przedziału predykcji dla ww. zapalników. Ze względu na obszerność artykułu, nie omówiono testów sprawdzających podczas weryfikacji statystycznej zaprezentowanego modelu regresji. W artykule zastosowano uniwersalne narzędzie statystyczne jakim jest program Statistica. Dzięki niemu przedstawiono interpretację graficzną oraz arkusze wyników analizowanych danych statystycznych zapalników typu B-23U. Na końcu artykułu przedstawiono zwięzłe wnioski dotyczące analizy regresji liniowej.

Słowa kluczowe: predykcja, model, regresja, zmienna, estymacja.

THE REGRESS AS THE METHOD OF PREDICTION PROCESS

Dariusz AMPUŁA
Military Institute of Armament Technology

Abstract: In the introduction of the article the author presents a short historical outline and then acquaints the reader with the method of the regress analysis. The function of regress 1st and 2nd kind is shown at the beginning and on the example undertaken after diagnostic decisions fuses type B-23U, the way of marking this function was introduced. The classic model of the linear regress was characterized and using the results of the tests the graphic and analytic figure of the regress function together with enumerate trust range was introduced. According to the procedure of making regress line, the estimation of the parameters regress model and verification of this model was also executed. The method of defining prediction on the basis linear regress model for analysed fuses type B-23U in article was presented. The graphic interpretation of the prediction range for this fuses was discussed. Due to the extensiveness article, tests checking during the statistical verification of the presented regress model were not outlined. In the article, the universal statistical tool was applied i.e. the Statistica programme. Thanks to it, the graphic interpretation and the sheets of analysed statistical results of the fuses type B-23U were introduced. Concise conclusions relating to analyses linear of regress were introduced at the end of the article.

Keywords: prediction, model, regress, variable, estimation.

1. Wstęp

Większość zjawisk w otaczającym nas świecie występuje nie w odosobnieniu, ale w różnorodnych związkach. O powiązaniach między nimi mówią różnorodne prawa, formułując przeróżne zależności między występującymi tam zmiennymi.

Współzależność między zmiennymi, według [1], może być dwojakiego rodzaju: funkcyjna i stochastyczna (probabilistyczna). Istota zależności funkcyjnej polega tym, że zmiana wartości jednej zmiennej powoduje ściśle określoną zmianę drugiej zmiennej. Tak więc w zależności funkcyjnej określonej wartości jednej zmiennej X odpowiada jedna i tylko jedna wartość drugiej zmiennej Y**.** Symbolem X oznacza się zmienną niezależną (objaśniającą), natomiast symbolem Y zmienną zależną (objaśnianą). Ze związkami typu funkcyjnego spotykamy się w naukach przyrodniczych i w technice.

Zależność stochastyczna występuje wtedy, gdy wraz ze zmianą jednej zmiennej zmienia się rozkład prawdopodobieństwa drugiej zmiennej. Szczególnym przypadkiem zależności stochastycznej jest zależność korelacyjna (statystyczna). Polega ona na tym, że określonym wartościom jednej zmiennej odpowiadają ściśle określone średnie wartości drugiej zmiennej. Możemy zatem ustalić, jak zmieni się średnio biorąc wartość zmiennej zależnej Y w zależności od wartości zmiennej niezależnej X**.**

Jeśli między badanymi zmiennymi nie ma związku stochastycznego, to nie ma również związku korelacyjnego. Twierdzenie odwrotne nie jest prawdziwe. Wynika to z tego, że określonej liczbie identycznych wariantów zmiennej odpowiada zawsze ta sama średnia, ale daną średnią można uzyskać z różnej kombinacji takich wariantów.

O badaniu związku korelacyjnego można mówić tylko wtedy, gdy przynajmniej jedna zmienna jest mierzalna. W celu określenia stopnia zależności między badanymi zmiennymi można posłużyć się współczynnikiem korelacji lub funkcją regresji. Współczynnik korelacji jest miernikiem siły zależności miedzy badanymi zmiennymi. W wyniku analizy regresji można odpowiedzieć na pytanie, jakiej zmiany średniej wartości zmiennej zależnej należy oczekiwać przy zmianie wartości zmiennej niezależnej o jednostkę. Obie metody badania współzależności, tj. korelacja i regresja, odpowiadają zatem na różne, ale uzupełniające się nawzajem pytania**.**

Nazwa regresja według [2] i [5] jest nieco niefortunna, ponieważ ma niewiele wspólnego z cofaniem się jak sugerowałby to wyraz. Nazwa ta została po raz pierwszy użyta w 1885 roku przez Sir Francisa Galtona (ucznia Darwina) podczas badań nad zależnością wzrostu potomstwa od wzrostu rodziców. Obserwacje te dały początek teorii „regresji w kierunku przeciętności”. W istocie regresja jest o wiele starsza. Jak zauważono w literaturze, francuscy matematycy w XVIII wieku (Laplace) przeprowadzili analizy, które nazwalibyśmy regresją.

2. Regresja liniowa

Jeżeli przyjmiemy [2], że zbiorowość jest badana ze względu na dwie zmienne X i Y, gdzie oczywiście zmienna Y jest zmienną mierzalną, wartości tych zmiennych w n-elementowej próbie oznaczamy odpowiednio $x_1, x_2, \dots, x_n$ oraz $y_1, y_2, \dots, y_n$.

Narzędziem badania mechanizmu powiązań między tymi zmiennymi jest funkcja regresji. Funkcja regresji jest to funkcja matematyczna określonej postaci, która jest przybliżeniem faktycznej zależności między zmiennymi. Postać funkcji jest ustalana na podstawie zaobserwowanych wartości (x_i, y_i).

Dokładny obraz takiej zależności w populacji daje funkcja regresji I rodzaju, która pokazuje, jak zmieniają się wartości oczekiwane zmiennej zależnej w zależności od zmian wartości innych zmiennych (zmienne niezależne).

Dla omawianych zmiennych X i Y funkcję regresji I rodzaju możemy zapisać:

$$E\,(\,Y|X=\,x_i\,)=\,f\,(x_i\,) \qquad (1)$$

gdzie: $E(Y|X=x_i)$ - średnia z wartości zmiennej Y, które odpowiadają wartości zmiennej x_i .

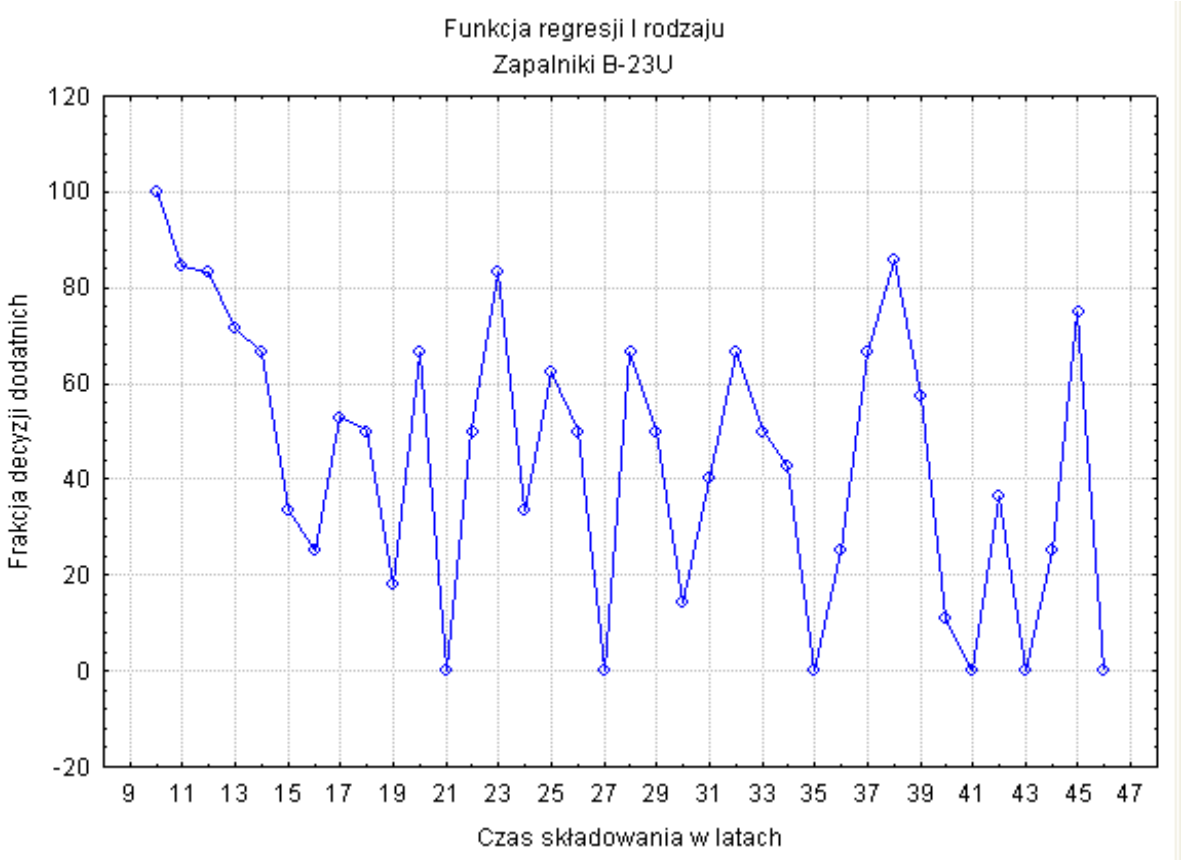
Tabela 1 - zapalniki typu B-23U

Czas składowania	**Liczba decyzji negatywnych**	**Liczba decyzji dodatnich**	**Frakcja decyzji dodatnich**	**Liczba zbadanych partii**
10	0	2	100,00	2
11	4	22	84,62	26
12	2	10	83,33	12
13	2	5	71,43	7
14	3	6	66,67	9
15	2	1	33,33	3
16	12	4	25,00	16
17	8	9	52,94	17
18	3	3	50,00	6
19	9	2	18,18	11
20	1	2	66,67	3
21	4	0	0,00	4
22	4	4	50,00	8
23	1	5	83,33	6
24	4	2	33,33	6
25	3	5	62,50	8
26	3	3	50,00	6
27	3	0	0,00	3
28	2	4	66,67	6
29	1	1	50,00	2
30	6	1	14,29	7
31	3	2	40,00	5
32	3	6	66,67	9
33	3	3	50,00	6
34	4	3	42,86	7
35	5	0	0,00	5
36	6	2	25,00	8
37	1	2	66,67	3
38	1	6	85,71	7
39	3	4	57,14	7
40	8	1	11,11	9
41	5	0	0,00	5
42	7	4	36,36	11
43	9	0	0,00	9
44	3	1	25,00	4
45	1	3	75,00	4
46	1	0	0,00	1
Razem	140	128	47,76	268

Dla przedstawionych danych w tabeli 1, wykres zależności frakcji decyzji dodatnich od czasu składowania przedstawia rysunek 1.

Analityczna postać funkcji $f\,(x_i\,)$ jest najczęściej nieznana. Tym samym wyznaczenie regresji I rodzaju jest trudne. W praktyce postępuje się w ten sposób, że na podstawie zaobserwowanych wyników z próby możemy graficznie przedstawić tzw. empiryczną linię regresji. Przy jej pomocy wyznaczamy najbardziej odpowiednią postać analityczną funkcji opisującą powiązanie między zmiennymi. Funkcja ta nosi nazwę funkcji regresji II rodzaju. Funkcja regresji II rodzaju zastępuje nam, jeśli uzyskane przybliżenie możemy uznać za wystarczające, nieznaną krzywą regresji I rodzaju.

Jako przykład wyznaczenia funkcji regresji I rodzaju posłużymy się wynikami badań diagnostycznych zapalników typu B-23U. Są to zapalniki stosowane w 25 mm nabojach morskich. Przedstawiona tabela 1 zawiera wszystkie decyzje podjęte po badaniach laboratoryjnych dla tego typu zapalników do roku 2010 włącznie.



Rys. 1. Empiryczna linia regresji I rodzaju dla zapalników B-23U

Wykres empirycznej linii regresji podpowiada nam typ funkcji opisującej powiązanie pomiędzy badanymi zmiennymi. Tę funkcję nazywamy funkcją regresji II rodzaju. Funkcja ta jest więc przybliżeniem funkcji regresji I rodzaju. Empiryczny wykres regresji pozwala na wskazanie analitycznej postaci funkcji regresji II rodzaju. Dla jej dokładniejszego wyznaczenia korzystamy także ze źródeł pozastatystycznych, np. z poprzednich badań doświadczalnych, opinii doświadczonych badaczy.

Ogólnie według [2] proces wyznaczania modelu funkcji regresji II rodzaju możemy podzielić na cztery etapy:

- specyfikacja modelu;
- estymacja parametrów modelu;
- weryfikacja modelu;
- użycie modelu do procesu predykcji.

Specyfikacja to sformułowanie modelu w postaci nadającej się do dalszej analizy i weryfikacji empirycznej. Punktem wyjścia do sformułowania modelu jest istniejąca teoria oraz postawiona przez nas hipoteza naukowa, którą chcielibyśmy zweryfikować w oparciu o dane liczbowe, którymi dysponujemy.

Estymacja to zastosowanie odpowiednich metod statystycznych w celu otrzymania jak najlepszych ocen występujących w modelu parametrów w oparciu o dane liczbowe. Dane te

to najczęściej wyniki badań przeprowadzonych w losowo wybranej próbce. Na tym etapie następuje nadanie parametrom modelu wartości liczbowej.

Weryfikacja to sprawdzenie, czy otrzymane oszacowanie modelu wytrzymuje konfrontację z teorią oraz czy dane potwierdzają poprawność utworzonego modelu. Wykorzystując metody statystyczne, szacujemy istotność otrzymanych parametrów. Jeśli model nie spełnia stawianych mu wymagań, możemy sformułować nowy model (zmienić postać funkcji, zebrać nowe dane, wykorzystać inną teorię itd.).

W końcowym etapie, kiedy model uznajemy za poprawny, możemy go wykorzystać albo do obliczenia nieznanych wartości zmiennej zależnej (mówimy wówczas o procesie predykcji), albo do wyznaczenia wartości zmiennych niezależnych dla uzyskania odpowiedniego poziomu zmiennej zależnej (mówimy wówczas o procesie sterowania).

W podanym przykładzie 1, wyspecyfikowaną krzywą, która najlepiej opisuje wzajemne powiązanie zmiennych, jest wykres funkcji liniowej. Tą krzywą mogą być funkcje logarytmiczne, wykładnicze, potęgowe, wielomianowe itp. Funkcja liniowa jest więc najprostszą, a zarazem najczęściej spotykaną zależnością opisującą wyspecyfikowaną krzywą.

3. Klasyczny model regresji liniowej

Przy rozpatrywaniu regresji liniowej, regresja I rodzaju opisująca zależność zmiennej Y od X, jest więc w takiej sytuacji postaci:

$$Y = E\,(\,Y|X = x\,) \;=\; \beta_0 + \beta_1\, x \;+ \varepsilon \qquad (\,2\,)$$

gdzie: $E\,(\,Y \mid X = x\,) = E\,(\,Y \mid X\,)$ - oznacza wartość zmiennej Y oczekiwana przy warunku że zmienna przyjmuje wartość x;

ε - składnik losowy.

Oznacza to, że znajomość jednej zmiennej (X) pozwala nam na uściślenie jakich wartości należy oczekiwać od drugiej zmiennej. Składnik losowy reprezentuje losowe zakłócenia funkcyjnego powiązania między wartościami zmiennej zależnej a wartościami zmiennej niezależnej. Składnik ten wyraża wpływ wszystkich czynników, które obok X wpływać mogą na zmienną objaśnianą Y oraz związany jest z brakiem pełnego dopasowania analitycznej postaci funkcji regresji do rzeczywistego powiązania między analizowanymi zmiennymi. Składnik losowy jest więc losową zmienną, która pozwala na obliczenie dokładności szacunku parametrów liniowej funkcji regresji. Równania lub układy równań opisujące powiązania między zmiennymi statystycy nazywają modelami. Tak więc równanie (2) wraz z dodatkowymi założeniami nosi nazwę klasycznego modelu regresji liniowej.

Należy pamiętać, że w rzeczywistości nie są znane parametry $\beta_0\; i\; \beta_1$. Możemy je jedynie oszacować na podstawie n-elementowej próbki, składającej się z par obserwacji $(\,x_i, y_i)$ dla $i = 1, 2, \ldots, n$. Oszacowana funkcja regresji przyjmuje wówczas następującą postać:

$$y_i = \; b_0 + \; b_1 x_i + \; e_i = \hat{y}_i + \; e_i \qquad (\,3\,)$$

gdzie: $i = 1, 2, \ldots, n$ - kolejne numery elementów obserwacji;

e_i - reszty (zmienna losowa) definiowane jako $e_i = \; y_i \; - \; \hat{y}_i$.

Dla statystyków przy wyznaczaniu dopasowanej linii regresji, punktem wyjścia są reszty, a właściwie suma kwadratów reszt, opisująca rozbieżność pomiędzy wartościami empirycznymi zmiennej zależnej a jej wartościami teoretycznymi, obliczonymi na podstawie wybranej funkcji. Oszacowania $b_0\; i\; b_1$ dobieramy tak, aby suma kwadratów reszt osiągnęła minimum. Ta najbardziej znana i stosowana metoda szacowania parametrów linii regresji nosi nazwę metody najmniejszych kwadratów. Wszystkich czytelników chcących dokładnie poznać zasady tworzenia i stosowania tej metody odsyłam do wcześniejszego mojego

artykułu nt. *Predykcja za pomocą metody najmniejszych kwadratów*, który ukazał się w *Problemach Techniki Uzbrojenia nr 1/2014*.

Ogólnie metoda najmniejszych kwadratów polega na takim oszacowaniu parametrów funkcji (2), by dla danych z próbki spełniony był warunek:

wyrażenie $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - b_1 x_i - b_0)^2$ ma osiągnąć minimum, (4)

gdzie: y_i oznaczają wartości empirycznej zmiennej Y, a $\hat{y}_i$ wartości teoretyczne wyznaczone na podstawie równania (3).

Wykorzystując więc (4) otrzymujemy wzory:

$$b_0 = \bar{y} - b_1 \bar{x} \quad (5)$$

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

gdzie: $\bar{x}$ - średnia wartość X;

$\bar{y}$ - średnia wartość Y.

Oceny parametrów b_0 *i* b_1 noszą nazwę współczynników regresji.

Licznik ułamka określającego wartość b_1 stanowi podstawowy element miary współzmienności dwóch zmiennych pozostający w związku liniowym. Nosi on nazwę kowariancji
i wyraża ją wzór postaci:

$$cov\,(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \quad (7)$$

Szacując parametry powyższych wzorów, otrzymujemy liniową funkcję regresji postaci:

$$y = 75{,}304 - 1{,}103 * x \quad (8)$$

Wykres regresji liniowej dla rozpatrywanego przykładu zapalników B-23U tzn. zależności frakcji decyzji dodatniej od czasu składowania przedstawia rysunek 2.

Dodatkowo, podczas wyznaczania funkcji regresji został wyliczony współczynnik korelacji liniowej Pearsona, który wynosi r = -0,4114.

Podczas wyznaczania linii regresji, oprócz kwestii jej specyfikacji, należy wziąć pod uwagę cztery bardzo ważne problemy:

- problem zastosowania metody najmniejszych kwadratów (spełnienie odpowiednich założeń);
- problem własności estymatorów otrzymanych metodą najmniejszych kwadratów;
- problem istotności ocen parametrów;
- problem dokładności dopasowania linii regresji do wartości empirycznych.

Wyznaczanie przedziałów ufności jest bardzo popularne, gdyż nie posiadają one szeregu wad jakimi obarczone są estymatory. Przedziały ufności określają nam prawdopodobny zasięg naszych wyliczeń od wartości rzeczywistej. Wyznaczenie arytmetyczne tego przedziału jest skomplikowane i wymaga zastosowania specjalnych wzorów, których postać zależy od liczności próbki oraz od pewnych założeń dotyczących badanej cechy. Dzięki pomocy techniki komputerowej, większość programów statystycznych wylicza przedziały precyzyjnie
i bez problemu.

Interpretacja przedziału ufności jest oczywista: im mniejszy przedział ufności, tym dokładniej obliczony przez nas estymator przybliża wartość rzeczywistą dla całej populacji. Odwrotnie, szeroki przedział ufności oznacza możliwość dużych odchyleń wartości z próby od wartości z populacji, czyli małą wiarygodność naszych wyników obliczeń.

Rys. 2. Wykres rozrzutu, linia regresji wraz z 95% przedziałem ufności

.

4. Estymacja parametrów modelu funkcji regresji

Na początku należy wyznaczyć oceny parametrów modelu metodą najmniejszych kwadratów. W tym celu użyjemy programu Statistica i otrzymujemy arkusz wyników przedstawiony na rysunku 3.

Podsumowanie regresji zmiennej zależnej: Frakcja decyzji dodatnich (Regresja B-23U) R= ,41140461 R^2= ,16925375 Skoryg. R2= ,14551815 F(1,35)=7,1308 p<,01142 Błąd std. estymacji: 26,819

	b*	Bł. std. - z b*	b	Bł. std. - z b	t(35)	p
W. wolny			75,30350	12,37471	6,08527	0,000001
Czas składowania	-0,411405	0,154064	-1,10272	0,41295	-2,67035	0,011415

Rys. 3. Arkusz wyników danych regresji dla zapalników B-23U

Współczynniki regresji to kolumna „b”. Poszukiwany model regresji ma więc postać równania (8). Badacz nigdy nie dysponuje pełną informacją o całej populacji. To co ma, to tylko funkcja regresji wyliczona na podstawie danych z badanej próbki. Ta funkcja regresji, jest aproksymacją regresji w całej populacji. Wiąże się z tym problem oceny rozbieżności między wartościami zmiennej zależnej y_i, a wartościami $\hat{y}_i$ wyliczonymi z modelu. Różnice $e_i = y_i - \hat{y}_i$, opisujące tę rozbieżność, jak wspomniałem wcześniej są to reszty. Im reszty będą mniejsze, tym bliżej wartości empirycznej y_i będą wartości $\hat{y}_i$ przewidywane przez model. Najlepiej byłoby, gdyby reszty były równe zero.

W statystyce precyzję estymatora mierzy jego wariancja. Wielkość dana wzorem (9), zwana błędem standardowym estymacji, informuje o przeciętnej wielkości odchyleń empirycznych wartości zmiennej zależnej od wartości wyliczonych z modelu.

$$S_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}} \qquad (9)$$

Im S_e mniejsze, tym lepiej jest dopasowany model. Wartość ta zawarta jest w arkuszu wyników danych jako „Błąd std. estymacji". W przypadku analizowanych zapalników B-23U wynosi on $S_e = 26{,}819$. Oznacza to, że przewidywane wartości zmiennej zależnej różnią się od wartości empirycznych średnio o 26,819%. Można więc równanie (8) zapisać jako:

$$y = -1{,}103 * x + 75{,}304 \pm 26{,}819 \qquad (10)$$

Wyliczone współczynniki regresji są oszacowaniami współczynników regresji dla całej populacji. Jakim więc błędem są one obarczone? Odpowiedzi na to pytanie udziela średni błąd szacunku parametru. Jest on oszacowaniem średniej rozbieżności między parametrami modelu a jego możliwymi ocenami. Wartości tych błędów są zawarte w arkuszu wyników danych jako „Bł. std.- z b". Dla zapalników B-23U mamy:

- oceny parametru b_1 odchylają się od tego parametru o $S_{b_1} = 0{,}41295$,
- oceny parametru b_0 odchylają się od tego parametru o $S_{b_0} = 12{,}37471$.

Można więc powiedzieć, że szacując współczynnik kierunkowy na poziomie -1,103 mylimy się średnio o 0,413. Analogicznie, szacując wyraz wolny na poziomie 75,304, mylimy się średnio biorąc o 12,375. Nasuwa się pytanie czy to dużo czy też nie? Dla parametru b_1 błąd szacunku stanowi około 37%, natomiast dla parametru b_0 błąd ten stanowi około 16%. Uzyskane błędy szacowania są akceptowalne, dopiero wartości większe od 50% powinny wzmóc naszą czujność.

Wielu statystyków uważa, że średnie błędy szacunku są niewygodne w użyciu. Dużo łatwiej jest interpretować ilorazy t ($t = b_i/S_{b_i}$). Wartości te zawarte są w arkuszu wyników danych jako „t(35)". Mamy więc:

$$t_{b_1} = 6{,}08527\,,\ t_{b_0} = -2{,}67035 \qquad (11)$$

Ocena pierwszego parametru jest około 6 razy większa od błędu szacunku, natomiast dla drugiego 2,67 razy większa od błędu szacunku. Źle jest gdy błąd szacunku jest większy od oceny parametru ($|t| < 1$).

Najbardziej jednak popularną miarą dopasowania jest współczynnik determinacji oznaczany jako R^2. Punktem wyjścia do zbudowania takiej miary jest badanie sumy kwadratów odchyleń poszczególnych obserwacji y_i od średniej $\bar{y}$. Współczynnik determinacji mierzy nam, jaka część ogólnej zmienności zmiennej zależnej jest wyjaśniona przez regresję liniową. Tej miary dopasowania używamy tylko dla regresji liniowej. Symbol R^2wziął się z tego, że w modelu liniowym współczynnik determinacji jest równy kwadratowi współczynnika korelacji.

Dla zapalników B-23U wartość współczynnika determinacji wynosi $R^2 = 0{,}16925375$. Jak widać obliczony model wyjaśnia bardzo niski procent zaobserwowanej zmienności, a nie wyjaśnia wysokiego procentu zmienności.

Oczywiście im większe R^2 tym lepiej, dołączenie bowiem nowej zmiennej do istniejącego modelu zawsze spowoduje zwiększenie R^2. Pamiętajmy, że celem oceny istniejącego modelu nie jest dążenie do jak największego R^2, a znalezienie związku między rozpatrywanymi zmiennymi X i Y z rzetelnymi ocenami parametrów.

W praktyce, statystycy używają raczej poprawionego lub skorygowanego współczynnika determinacji. W naszym przypadku wynosi on (rys. 3) „Skoryg. R2 = 0,14551815". Współczynnik ten mówi, że R^2 jest obliczony z próby i jest trochę „za dobry" jeśli uogólniamy nasze wyniki na populację. Skorygowane R^2 mówi nam, jak dobrze dopasowane byłoby nasze równanie regresji do innej próby z tej samej populacji.

5. Weryfikacja modelu regresji

Każdy model regresji [2] powinien zostać poddany weryfikacji. Najważniejsze etapy to weryfikacja merytoryczna i statystyczna. W trakcie weryfikacji merytorycznej sprawdzamy, czy model spełnia nasze oczekiwania, czy poczynione założenia modelu są spełnione oraz czy model jest zgodny z założeniami teorii, która posłużyła do jego budowy. Weryfikacja merytoryczna musi być połączona z weryfikacją statystyczną. Jeśli ta ostatnia wypadnie źle, to model jest najprawdopodobniej zły i jego ocena merytoryczna budzi dużą wątpliwość.

Do weryfikacji modelu warto wykorzystać różnorodne dodatkowe informacje o równaniu regresji. Należą do nich przede wszystkim:

- współczynnik zmiennej zależnej względem zmiennej niezależnej – beta,
- postulat o koincydencji,
- elastyczność.

Obliczonych wartości współczynników regresji nie możemy porównywać ze względu na różne jednostki miary. Wprowadzono więc, współczynnik beta β, który dla zapalników B-23U wynosi β = -0,411405 (rys. 3). Można to zinterpretować, że zmiana zmiennej niezależnej o jedno odchylenie standardowe powoduje spadek wartości zmiennej zależnej o 0,411 jej odchylenia standardowego. Zaletą tej interpretacji jest jej niezależność od jednostek miary.

Postulat o koincydencji jest w tym przypadku spełniony, ponieważ w modelu z jedną zmienną postulat ten jest zawsze spełniony.

W trakcie weryfikacji wyliczamy także wartość elastyczności Y względem X:

$$elastyczność = b_1 * \frac{\bar{X}}{\bar{Y}} \qquad (12)$$

Elastyczność ta pokazuje, o ile procent zmienia się wartość Y, gdy wartość zmiennej X wzrasta o 1%. Dla zapalników B-23U wartość ta wynosi -0,695%. Oznacza to, że w otoczeniu wartości średnich zmiennych X i Y, wzrost o 1% zmiennej X powoduje około 0,7-procentowy spadek Y.

Na weryfikację statystyczną składa się cały szereg testów sprawdzających:

- istotność parametrów modelu,
- istotność całego modelu,
- założenia metody najmniejszych kwadratów.

Ze względu na obszerność artykułu testy te nie zostały omówione w tym artykule.

6. Predykcja na podstawie modelu regresji liniowej

Podczas budowy modelu funkcji regresji, bierzemy pod uwagę możliwość predykcji wartości zmiennych tzn. jakie wartości przyjmie zmienna zależna przy różnych wartościach zmiennej niezależnej. Finalnym więc etapem analizy regresji jest wykorzystanie zweryfikowanego modelu regresji do predykcji zmiennej zależnej. Dzięki predykcji możemy ustrzec się skutków przyszłych zdarzeń, a także wpływać na ich bieg lub im zapobiegać.

Predykcję dzielimy na:

- ex post - kiedy wartości zmiennych niezależnych są znane, sama predykcja może być porównana z wartościami zaobserwowanymi,

– ex ante - kiedy nie znamy wartości zmiennej niezależnej. Zależą one od charakteru i wartości tych zmiennych we wcześniejszych okresach. Na ogół nie wiemy z całą pewnością, jak ukształtuje się zmienna niezależna w przyszłym okresie T + s. Znamy jedynie jej wartość w okresie T. Najczęściej dane te wyznaczamy prognostycznie z pewnym prawdopodobieństwem równocześnie dla obu zmiennych X i Y. Taka predykcja występująca przy szeregach czasowych nosi nazwę predykcji warunkowej.

Wykorzystując model zbudowany dla zapalników B-23U możemy wyliczyć predykcję frakcji decyzji dodatnich dla żądanego czasu składowania. I tak np. dla czasu składowania równego 22 lata wyliczona predykcja zmiennej zależnej została przedstawiona w arkuszu wyników danych na rysunku 4. Przedstawiony został również 95% przedział ufności.

Obliczanie wartości (Regresja B-23U) zmiennej: Frakcja decyzji dodatnich			
	Wagi b	**Wartość**	**Wagi b - *Wartość**
Czas składowania	-1,10272	22,00000	-24,2599
W. wolny			75,3035
Przewidyw.			51,0436
-95,0%GU			40,7762
+95,0%GU			61,3111

Rys. 4. Arkusz wyników danych z przedziałem ufności

Równanie predykcji frakcji decyzji dodatnich dla czasu składowania 22 lata ma postać wzoru (13). Przedział ufności ma postać (40,7762; 61,3111).

$$y_p = -1{,}10272 * 22 + 75{,}3035 = 51{,}0436 \qquad (13)$$

Natomiast równanie predykcji frakcji decyzji dodatnich dla czasu składowania 50 lat dla zapalników B-23U będzie miało postać:

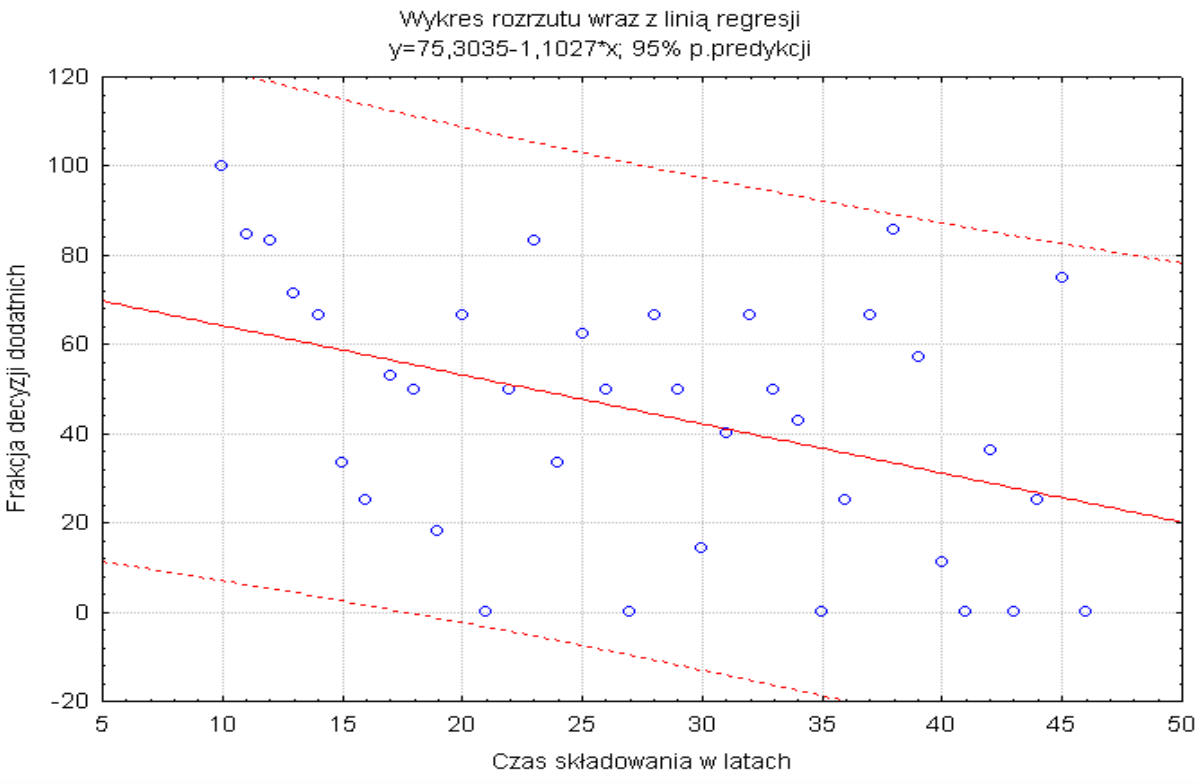
$$y_p = -1{,}10272 * 50 + 75{,}3035 = 20{,}1674 \qquad (14)$$

Arkusz wyników danych dla tego równania łącznie z przewidywanym przedziałem ufności przedstawiono na rysunku 5.

Obliczanie wartości (Regresja B-23U) zmiennej: Frakcja decyzji dodatnich			
	Wagi b	**Wartość**	**Wagi b - *Wartość**
Czas składowania	-1,10272	50,00000	-55,1361
W. wolny			75,3035
Przewidyw.			20,1674
-95,0%GU			-0,3332
+95,0%GU			40,6680

Rys. 5. Arkusz wyników danych z przedziałem ufności

Możemy również przedstawić graficznie wykres rozrzutu frakcji decyzji dodatnich w zależności od czasu składowania z naniesionym przewidywanym 95% przedziałem predykcji. Granice przedziału predykcji zostały przedstawione na rysunku 6 linią przerywaną.



Rys. 6. Wykres rozrzutu, linia regresji wraz z 95% przedziałem predykcji

Można także obliczyć analitycznie przykładowo dla czasu składowania wynoszącego 50 lat, przewidywany przedział predykcji. Arkusz wyników danych przedstawiono na rysunku 7.

Obliczanie wartości (Regresja B-23U) zmiennej: Frakcja decyzji dodatnich

	Wagi b	**Wartość**	**Wagi b - *Wartość**
Czas składowania	-1,10272	50,00000	-55,1361
W. wolny			75,3035
Przewidyw.			20,1674
-95,0%GP			-38,0106
+95,0%GP			78,3455

Rys. 7. Arkusz wyników danych z przedziałem predykcji

8. Wnioski

Klasyczny model regresji liniowej dzięki estymacji parametrów, czyli dzięki metodzie szacowania parametrów zmiennych losowych w całej populacji, pozwala nam przy ustalonym z góry prawdopodobieństwie (zwanym poziomem ufności) utworzyć dla nieznanego parametru populacji oszacowanie zwane przedziałem ufności.

Z każdym przedziałem ufności jest związany poziom ufności określony liczbą (oznaczany jako $1-\propto$). Wydawać by się mogło, że przyjęcie wysokiego poziomu ufności rozwiąże wszystkie problemy. Niestety tak nie jest. Zwiększenie poziomu ufności powoduje zwiększenie szerokości przedziału ufności, czyli zmniejszenie precyzji estymacji. Poprawa precyzji jest możliwa pod warunkiem zwiększenia liczności próbki, co powoduje zwiększenie kosztów badań.

W artykule przeprowadzono analizę regresji dla regresji liniowej z jedną zmienną niezależną dla zapalników B-23U, przyjmując jako zmienną niezależną czas składowania natomiast jako zmienną zależną frakcję decyzji dodatnich jakie zostały podjęte po przeprowadzonych badaniach diagnostycznych. Zaproponowany model regresji liniowej oraz przeprowadzona analiza tego modelu, mówi nam, że w miarę upływu lat składowania, stan jakościowy populacji zapalników typu B-23U ulega pogorszeniu. Świadczy o tym wyliczony współczynnik korelacji liniowej $r = -0{,}4114$ oraz współczynnik regresji $b_1 = -1{,}10272$.

W wyniku analizy regresji można więc odpowiedzieć na pytanie, jakiej zmiany średniej wartości zmiennej zależnej należy oczekiwać przy zmianie wartości zmiennej niezależnej o jednostkę. Podczas wykonywania analizy regresji, bardzo ważnym elementem jest wyznaczenie reszt. Analizując reszty, możemy bowiem szybko i skutecznie wykryć wszystkie odstępstwa od poprawnej analizy regresji. Mimo, że nie wszystkie założenia dotyczące regresji możemy sprawdzić, to jednak największe odstępstwa możemy wykryć i ewentualnie wyeliminować. Regułą powinna stać się analiza wartości resztowych zawsze po oszacowaniu parametrów modelu regresji.

Literatura

[1] M. Sobczyk – *Statystyka* – Wydawnictwo PWN, Warszawa 2002r.
[2] A. Stanisz – *Przystępny kurs statystyki* – Statsoft Polska, Kraków 2007r.
[3] *Statistica 9* – Statsoft Polska 2009 r. – oprogramowanie komputerowe
[4] S. Kot, J. Jakubowski, A. Sokołowski – *Statystyka* – Wydawnictwo Difin, Warszawa 2011r.
[5] B. R. Górecki – *Ekonometria, podstawy teorii i praktyki* – Wydawnictwo Key Text, Warszawa 2010r.
[6] M. Rabiej – *Statystyka z programem Statistica* – Wydawnictwo Helion, Gliwice 2012r.
[7] J. Paradysz – *Statystyka* – Wydawnictwo Akademii Ekonomicznej w Poznaniu, Poznań 2005r.
[8] *Wikipedia* – wolna encyklopedia