WPROWADZENIE DO ANALIZY KORELACJI I REGRESJI

dr Janusz Wątroba, StatSoft Polska Sp. z o.o.

Prezentowany artykuł poświęcony jest wybranym zagadnieniom analizy korelacji i regresji. Po przedstawieniu najważniejszych informacji potrzebnych do zrozumienia idei tych metod zaprezentowane zostaną przykłady ich zastosowań. Wszystkie potrzebne obliczenia oraz wykresy zostały wykonane za pomocą programu *STATISTICA*.

Wprowadzenie

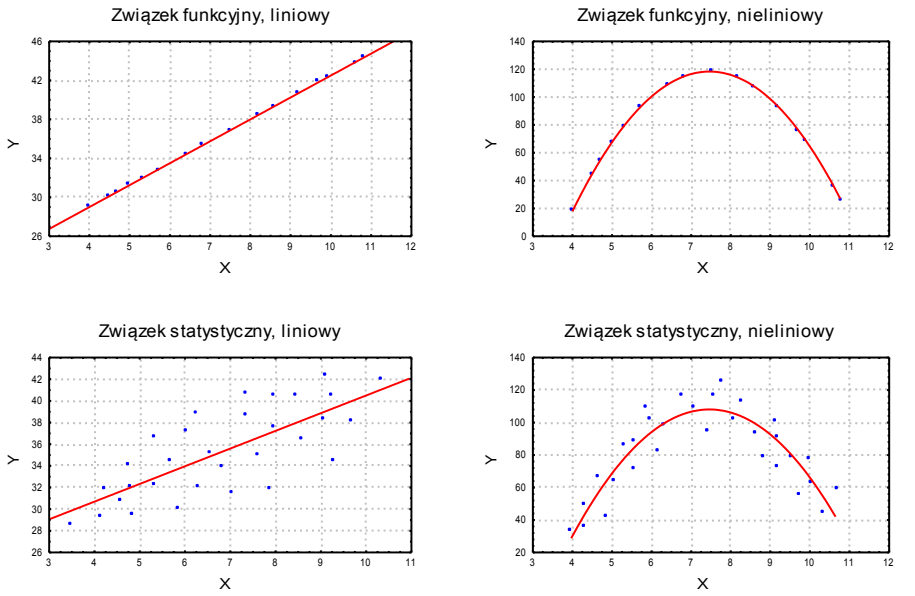
Poszczególne elementy wchodzące w skład badanej zbiorowości jednostek są zazwyczaj opisywane za pomocą więcej niż jednej cechy (zmiennej). W większości przypadków analizowane zmienne są w jakiś sposób powiązane ze sobą. W takich sytuacjach zachodzi zatem potrzeba ich łącznego badania. Celem analizy jest więc stwierdzenie, czy między badanymi zmiennymi zachodzą jakieś zależności, jaka jest ich **siła**, jaka jest ich **postać** i **kierunek**.

Współzależność między zmiennymi może być dwojakiego rodzaju: **funkcyjna** lub **stochastyczna** (**probabilistyczna**). Istota zależności funkcyjnej polega na tym, że zmiana wartości jednej zmiennej powoduje ściśle określoną zmianę wartości drugiej zmiennej. W przypadku zależności funkcyjnej określonej wartości jednej zmiennej (X) odpowiada jedna i tylko jedna wartość drugiej zmiennej (Y). Symbolem X oznaczamy zmienną niezależną (objaśniającą), natomiast symbolem Y - zmienną zależną (objaśnianą).

Zależność stochastyczna występuje wtedy, gdy wraz ze zmianą wartości 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.

Na zamieszczonym poniżej wykresie przedstawiono przykładowe postacie związków funkcyjnych i statystycznych.

Copyright © StatSoft Polska 2002



Związki typu statystycznego są możliwe do wykrycia oraz ilościowego opisu w przypadku, kiedy mamy do czynienia z wieloma obserwacjami, opisującymi badane obiekty, zjawiska czy też procesy.

Opisywane w niniejszym artykule postacie związków pomiędzy zmiennymi zawęzimy do związków liniowych. Ogólnie związki pomiędzy zmiennymi mogą przyjmować postać krzywej drugiego i wyższych stopni lub też inne postacie. Dlatego też, przy badaniu danych, ważnym krokiem jest sporządzenie wykresu. Jeśli okaże się, że badany związek nie jest liniowy, wówczas trzeba zastosować odpowiednie rozwiązanie nieliniowe. Odpowiednie narzędzia są dostępne w komercyjnych pakietach do analiz statystycznych, np. w pakiecie *STATISTICA*.

Wykres rozrzutu

Wykres rozrzutu pozwala w graficzny sposób zaprezentować postać związku pomiędzy zmiennymi. Przy interpretacji tego wykresu należy jednak zachować pewną ostrożność. Przede wszystkim należy pamiętać, że wykres nie pozwala stwierdzić związku przyczynowo-skutkowego. Na przykład może się okazać, że faza księżyca wykazuje wpływ na proces, w przypadku którego występuje cykl miesięczny.

Tworząc wykres rozrzutu, należy najpierw jasno zdefiniować zmienne. Następnie trzeba zebrać przynajmniej 30 par liczb (lepiej jest, jeśli danych jest więcej; 50 lub 100

 Copyright © StatSoft Polska 2002

przypadków). Powszechnie przyjęło się na osi poziomej umieszczać zmienną, która stanowi przyczynę, a na osi pionowej zmienną, która jest uważana za skutek.

Współczynnik korelacji liniowej

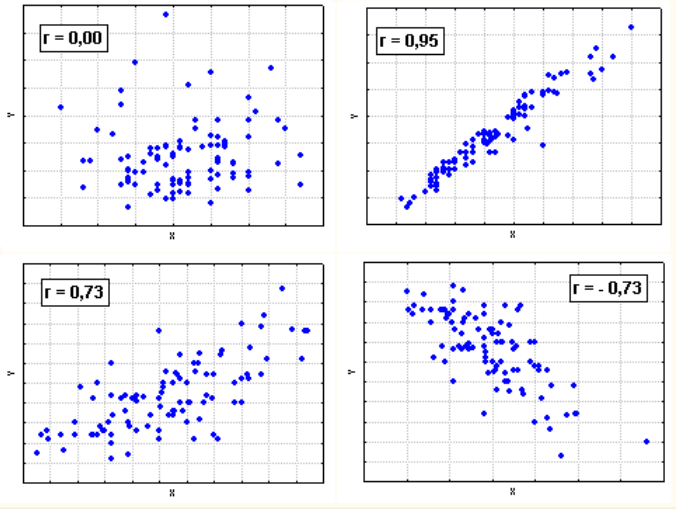
Statystyką, która opisuje siłę liniowego związku pomiędzy dwiema zmiennymi, jest współczynnik korelacji z próby (r). Przyjmuje on wartości z przedziału domkniętego <-1; 1>. Wartość –1 oznacza występowanie doskonałej korelacji ujemnej (to znaczy sytuację, w której punkty leżą dokładnie na prostej, skierowanej w dół), a wartość 1 oznacza doskonałą korelację dodatnią (punkty leżą dokładnie na prostej, skierowanej w górę). Wartość 0 oznacza brak korelacji liniowej.

Wzór, za pomocą którego obliczamy współczynnik korelacji, ma postać:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

gdzie x_i oraz y_i oznaczają odpowiednio wartości zmiennych x i y, a $\bar{x}$ oraz $\bar{y}$ oznaczają średnie wartości tych zmiennych.

Po obliczeniu wartości współczynnika korelacji zawsze zalecane jest utworzenie wykresu rozrzutu. Chodzi o to, aby wizualnie stwierdzić, czy badany związek rzeczywiście najlepiej opisuje funkcja liniowa. Może się bowiem okazać, że wyliczona wartość współczynnika korelacji jest zbliżona do zera, a mimo to pomiędzy korelowanymi zmiennymi występuje współzależność, tyle że nieliniowa.



Na zamieszczonym powyżej rysunku przedstawiono przykładowy wygląd wykresów przy określonych wartościach współczynnika korelacji. Wykres umieszczony z lewej strony u góry pokazuje sytuację braku skorelowania zmiennych. Z kolei wykres umieszczony

z prawej strony u góry przedstawia przypadek silnej korelacji dodatniej. Wykres umieszczony u dołu z lewej strony pokazuje nieco słabszą korelację dodatnią, a wykres umieszczony u dołu z prawej strony przedstawia przypadek korelacji ujemnej.

Badanie istotności korelacji

Współczynnik korelacji r (z próby) stanowi ocenę współczynnika korelacji ρ w zbiorowości generalnej i w związku z tym jest obciążony pewnym błędem. Współczynnik korelacji jest statystyką, w związku z czym powinien być traktowany jako zmienna losowa. Jeśli zatem N-elementowa próba została pobrana ze zbiorowości generalnej o dwuwymiarowym rozkładzie normalnym z parametrem $\rho = 0$, a więc gdy zmienne X i Y są nieskorelowane i zarazem niezależne, to zmienna losowa o postaci:

$$t = r\frac{\sqrt{N-2}}{\sqrt{1-r^2}}$$

ma rozkład t Studenta o N-2 stopniach swobody.

W praktyce oznacza to, że formułujemy hipotezę zerową: $H_0 : \rho = 0$, wobec hipotezy alternatywnej: $H_1 : \rho \neq 0$, a następnie porównujemy wartość graniczną t_α z wartością obliczoną t i podejmujemy odpowiednią decyzję odnośnie H_0.

Przykład analizy korelacji

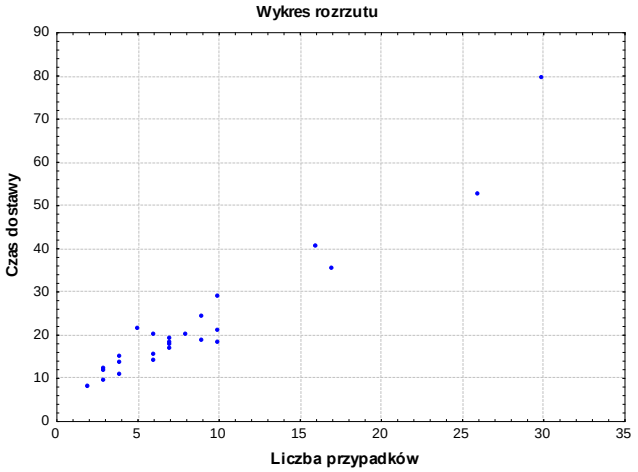
Prezentowany niżej przykład został zaczerpnięty z książki Breyfogle'a [1]. Obserwacji poddano czas dostawy (x) 25 napojów bezalkoholowych oraz wielkość dostawy (y). Fragment pliku danych przedstawia zamieszczony poniżej rysunek.

Data: Korelacja i regresja* (2v by 2...

	1 Liczba przypadków	2 Czas dostawy
1	7	16,68
2	3	11,50
3	3	12,03
4	4	14,88
5	6	13,75
6	7	18,11
7	2	8,00
8	7	17,83
9	30	79,24
10	5	21,50
11	16	40,33

 Copyright © StatSoft Polska 2002

Umieszczony poniżej wykres rozrzutu, utworzony dla tych danych, pokazuje, że pomiędzy analizowanymi zmiennymi najprawdopodobniej występuje dość mocny związek typu liniowego.



Obliczona w programie *STATISTICA* wartość współczynnika korelacji liniowej Pearsona wyniosła tu *0,9646*. Po zbadaniu istotności współczynnika korelacji okazało się, że jego wartość istotnie różni się od zera na poziomie $p<0,01$. Tak więc możemy stwierdzić, że pomiędzy wielkością dostaw a czasem dostaw istnieje bardzo mocna, dodatnia zależność, a więc wraz ze wzrostem wielkości dostaw, średnio rzecz biorąc, wydłuża się czas dostawy.

Wybrane elementy analizy regresji prostej

Analiza korelacji jest stosowana do pomiaru stopnia powiązania pomiędzy zmiennymi, natomiast jej rozszerzenie, jeśli chodzi o metody badania współzależności, stanowi analiza regresji. Pozwala ona na ilościowy opis zachodzących powiązań. Pojęcie funkcji w zastosowaniu do badań empirycznych nie może być zazwyczaj stosowane bez pewnych zastrzeżeń. Elementarna matematyka żąda bowiem, aby jednej wartości zmiennej niezależnej (objaśniającej, predyktora) była przyporządkowana dokładnie jedna wartość zmiennej zależnej (objaśnianej). Badacz natomiast w praktyce ma zazwyczaj do czynienia z sytuacją, w której przy kilku powtórzeniach doświadczenia, zachowując za każdym razem te same wartości zmiennej niezależnej, otrzymuje inne wartości mierzonej zmiennej zależnej. Wartości te zwykle leżą blisko siebie, ale nie są na ogół identyczne. Tak więc rozsądek podpowiada, żeby pojęcie funkcji uczynić bardziej elastycznym, a terminy „zmienna niezależna" i „zmienna zależna" dostosować odpowiednio do nowych potrzeb. Dla tego celu w statystyce matematycznej wprowadzono pojęcie „regresji", oznaczające obliczenia wykorzystywane do ilościowego opisu zależności jednej zmiennej od drugiej.

Copyright © StatSoft Polska 2002

Model regresji liniowej prostej (tzn. takiej, w przypadku której występuje tylko jeden predyktor) przyjmuje postać:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

gdzie β_0 oznacza wyraz wolny, β_1 współczynnik kierunkowy, a ε błąd. Jak to zostało już wcześniej powiedziane, zazwyczaj nie wszystkie punkty układają się dokładnie na prostej modelu regresji. Źródłem błędu są wpływy innych, nieuwzględnionych w modelu zmiennych, takich jak np. błędy pomiarowe. Zakłada się przy tym, że błędy mają średnią wartość równą zero i nieznaną wariancję σ^2, oraz że błędny nie są nawzajem skorelowane. W sytuacji gdy wartość współczynnika determinacji R^2 (wielkość ta oznacza kwadrat współczynnika korelacji) jest duża, to oznacza to, że błędy dla tego modelu są stosunkowo małe i w związku z tym model jest dobrze dopasowany do rzeczywistych danych.

Jeśli model regresji liniowej zawiera tylko jedną zmienną niezależną, wtedy jest nazywany modelem regresji liniowej prostej. Natomiast jeśli model regresji zawiera większą liczbę zmiennych niezależnych, to wówczas odpowiedni model nazywamy modelem regresji wielorakiej.

Zasadniczy cel analizy regresji polega na ocenie nieznanych parametrów modelu regresji. Ocena ta jest dokonywana za pomocą *metody najmniejszych kwadratów* (*MNK*). Metoda ta sprowadza się do minimalizacji sum kwadratów odchyleń wartości teoretycznych od wartości rzeczywistych (czyli tzw. **reszt** modelu). Dopasowany model regresji prostej, który daje punktową ocenę średniej wartości *y* dla określonej wartości *x*, przyjmuje postać:

$$\hat{y} = b_0 + b_1 x$$

gdzie $\hat{y}$ oznacza teoretyczną wartość zmiennej zależnej, a b_o i b_1 odpowiednio oceny wyrazu wolnego i współczynnika kierunkowego, uzyskane na podstawie wyników z próby.

Jak to zostało wyżej powiedziane, różnica pomiędzy wartością obserwowaną y_i i odpowiadającą jej wartością teoretyczną (dopasowaną) $\hat{y}_i$ jest nazywana resztą. Tak więc dla danej obserwacji *i* można ją zapisać jako:

$$e_i = y_i - \hat{y}_i$$

Analiza reszt stanowi ważny element oceny adekwatności uzyskanego modelu regresji oraz oceny odchyleń występujących w stosunku do przyjmowanych założeń.

Przy testowaniu istotności współczynników regresji korzystamy z rozkładu *t* Studenta, a przy przeprowadzaniu analizy wariancji (przeprowadzanej w celu oceny liniowości modelu regresji) z rozkładu *F*. W pierwszym przypadku jedna hipoteza zerowa zakłada, że β_0 ma wartość stałą (przeciw alternatywnej, zakładającej, że β_0 nie jest wartością stałą), a druga przyjmuje, że ocena β_1 wynosi zero (przeciw alternatywnej, zakładającej, że ocena β_1 różni się od zera).

 Copyright © StatSoft Polska 2002

W przypadku tabeli analizy wariancji, całkowita zmienność jest dzielona na dwie części, opisywane przez odpowiednie sumy kwadratów (SK):

$$SK_{\text{cała}} = SK_{\text{modelu}} = SK_{\text{błędu}}$$

gdzie

$$SK_{\text{cała}} = \sum(y_i - \bar{y})^2$$

$$SK_{\text{modelu}} = \sum(\hat{y}_i - \bar{y})^2$$

$$SK_{\text{błędu}} = \sum(y_i - \hat{y}_i)^2$$

Każda z określonych wyżej sum kwadratów jest powiązana z odpowiednimi liczbami stopni swobody, które wynoszą kolejno: *n*-1, 1 i *n*-2.

Po podzieleniu przez podane liczby stopni swobody odpowiednie sumy kwadratów stanowią dobre oceny odpowiednich źródeł zmienności. Uzyskane w ten sposób wielkości są nazywane w skrócie średnimi kwadratami (ŚK). Są one wykorzystywane do oceny liniowości regresji. Wartość statystyki testowej jest wyliczana ze wzoru:

$$F_0 = \frac{\acute{S}K_{\text{modelu}}}{\acute{S}K_{\text{błędu}}}$$

Następnie jest ona porównywana z wartością teoretyczną $F_{\alpha,1,n-2}$, po czym hipoteza o liniowości regresji jest odpowiednio weryfikowana.

Kolejnym przeprowadzanym testem jest ocena adekwatności modelu. W tym celu sprawdzamy, czy ocena współczynnika determinacji

$$R^2 = \frac{SK_{\text{modelu}}}{SK_{\text{cała}}}$$

istotnie różni się od zera. Po przemnożeniu tej wartości przez 100 otrzymujemy ocenę udziału wariancji wyjaśnianej przez oszacowany model regresji.

Analiza reszt

Jak to zostało już wcześniej wspomniane, analiza reszt odgrywa ważną rolę przy badaniu adekwatności dopasowanego modelu oraz ocenie prawdziwości przyjmowanych założeń. Zazwyczaj obejmuje ona następujące elementy:

- sprawdzenie założenia normalności rozkładu reszt, które jest przeprowadzane za pomocą oceny wykresu normalności reszt lub oceny histogramu rozkładu reszt,
- ocenę skorelowania reszt poprzez wykreślenie reszt w funkcji numeru obserwacji,

- ocenę poprawności modelu przez wykreślenie wartości reszt względem wartości dopasowanych.

Przykład analizy regresji prostej

Przykład zostanie przeprowadzony w oparciu o te same dane, które były wykorzystywane poprzednio przy analizie korelacji. Tak jak poprzednio, wszystkie potrzebne obliczenia i wykresy zostaną wykonane przy pomocy programu *STATISTICA*.

W pierwszej kolejności przeprowadzimy ocenę wyrazu wolnego i współczynnika regresji. Odpowiednie wyniki zawiera zamieszczona poniżej tabela.

Data: Regression Summary for Dependent Variable: Czas dostawy (...

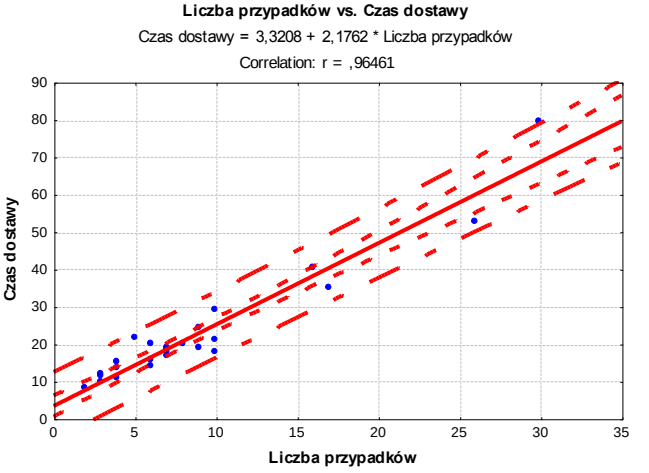
Regression Summary for Dependent Variable: Czas dostawy (
R= ,96461459 R2= ,93048131 Adjusted R2= ,92745876
F(1,23)=307,85 p<,00000 Std.Error of estimate: 4,1814

N=25	Beta	Std.Err. of Beta	B	Std.Err. of B	t(23)	p-level
Intercept			3,320780	1,371074	2,42203	0,023721
Liczba przypadków	0,964615	0,054978	2,176167	0,124030	17,54555	0,000000

Tak więc oszacowany model regresji przyjął postać:

$$y = 3{,}321 + 2{,}176\ x$$

Na podstawie wyników zawartych w tabeli możemy stwierdzić, że obydwie oceny istotnie różnią się od zera. Świadczą o tym wartości prawdopodobieństw testowych odpowiadające obydwu ocenom, które są niższe od 0,05.

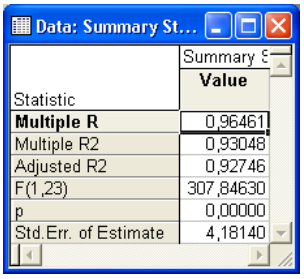


Na wykresie widzimy oszacowany model w postaci graficznej wraz z 95% granicami przedziału ufności i przedziału predykcji. Granice ufności odzwierciedlają szerokość prze-

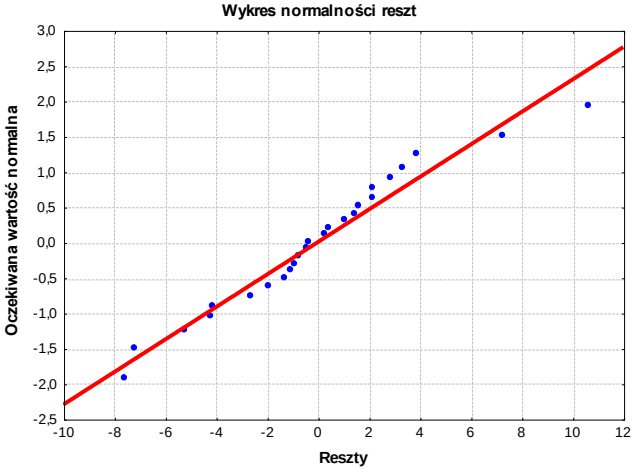
 Copyright © StatSoft Polska 2002

działu ufności dla ocen parametrów modelu, a granice predykcji odzwierciedlają szerokość przedziału ufności dla wartości zmiennej zależnej przy danym poziomie zmiennej niezależnej.

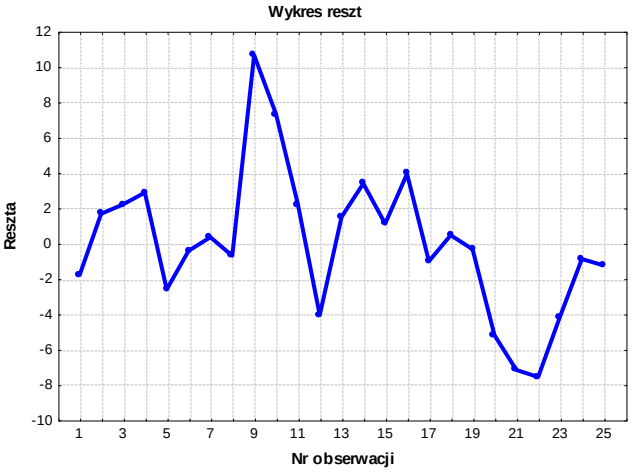
Kolejne wyniki analizy, pokazane na poniższym rysunku, mówią o tym, że oszacowany model wyjaśnia około 93% oryginalnej zmienności analizowanej zmiennej zależnej. Standardowy błąd estymacji wynosi 4,18. Jego wartość informuje nas o tym, jaki przeciętnie błąd popełnimy, przewidując czas dostawy na podstawie jej wielkości.



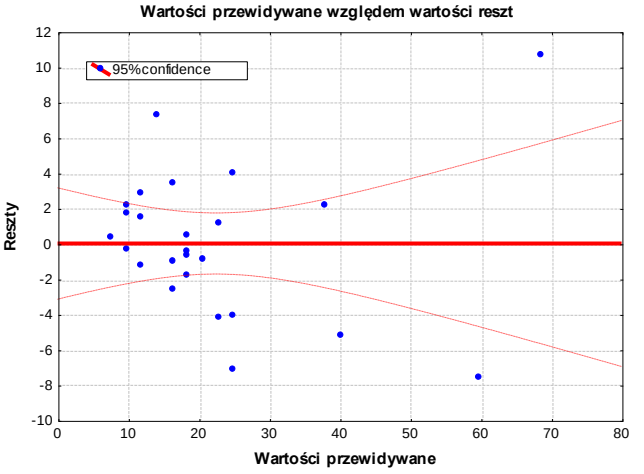
W tym miejscu przeprowadzimy jeszcze analizę reszt. Po pierwsze ocenimy, czy rozkład reszt jest rozkładem normalnym. W tym celu utworzymy wykres normalności reszt. Wykres ten został przedstawiony poniżej.



Następnie sprawdzimy, czy reszty są ze sobą skorelowane. Jak widać na poniższym wykresie, sytuacja taka raczej nie występuje.



Ostatni z zamieszczonych wykresów przedstawia wartości reszt względem wartości prognozowanych.



Podsumowanie

W tym momencie powinniśmy jeszcze raz powrócić do rzeczywistego celu przeprowadzanej analizy. Model pozwala dość dokładnie oddać przebieg prawdziwych danych. Być może dalsza analiza (np. budowa modelu opisującego związek odległości dostawy i liczby przypadków) pozwoliłaby znaleźć lepszy model. Można byłoby również spróbować zastosować model regresji wielorakiej, tzn. model uwzględniający większą liczbę zmiennych niezależnych (np. dzień tygodnia, w którym była realizowana dostawa, operator lub stan pogody). Z drugiej strony trzeba zdać sobie sprawę, że proces gromadzenia i analizy danych jest zazwyczaj czasochłonny, i w związku z tym należy rozważyć ewentualne zyski i straty.

 Copyright © StatSoft Polska 2002

Literatura

1. Breyfogle III F. W., 1999, Implementing Six Sigma. Smarter Solutions Using Statistical Methods. John Wiley & Sons, Inc.
2. Czermiński J. B., Iwasiewicz A., Paszek Z., Sikorski A., 1992, Metody statystyczne dla chemików, wyd. II, PWN Warszawa.
3. Montgomery D. C., 1997, Introduction to Statistical Quality Control, wyd. III, John Wiley & Sons,Inc.

Copyright © StatSoft Polska 2002