Michalina Szłapka

MIERNIKI ZALEŻNOŚCI DLA DANYCH PRZEDSTAWIONYCH W TABLICY KORELACYJNEJ

Wstęp

Badanie powiązań i zależności między zjawiskami jest ważnym elementem dla każdej z nauk, dostarcza bowiem informacji o funkcjonowaniu otaczającej nas rzeczywistości. Szczególnego znaczenia badanie to nabiera w naukach o zarządzaniu. Znajomość wzajemnych relacji zachodzących w otoczeniu ekonomicznym, politycznym i społeczno-gospodarczym pozwala przedsiębiorstwu na lepsze przystosowanie się do zachodzących w nim zmian oraz przewidywanie konsekwencji zachodzących zdarzeń.

W dziedzinie nauk ilościowych wnioskowanie o istniejących pomiędzy zmiennymi relacjach umożliwia wiele mierników statystycznych. Mierniki te powinny dostarczać w miarę jednoznacznej informacji o sile i kierunku zależności zmiennych. Jednak każdy z nich ma pewne wady. Do niektórych należy ograniczona możliwość stosowania, do innych pracochłonność wykonywanych obliczeń.

W literaturze można spotkać kilka miar stosowanych dla danych zestawionych w formie tablicy, zaczynając od miar opartych na wielkości χ^2, a na stosunkach korelacyjnych charakterystycznych dla skali przedziałowej kończąc.

W opracowaniu omówimy i porównamy mierniki korzystające z tablicy korelacyjnej (współczynnik korelacji, stosunki korelacji) oraz miary zależności określone dla skali nominalnej oparte na tablicy wielodzielczej (współczynnik Czuprowa, Cramera i Hellwiga).

Wartości tych współczynników wyznaczymy dla trzech przykładowych serii mierzalnych danych skali przedziałowej, charakteryzujących sprzedaż, produkcję, jakość i sytuację rynkową pewnego surowca. Dodatkowo rozpatrzymy związany z tworzeniem tablicy korelacyjnej problem tworzenia szeregów przedziałowych danych. Zaproponujemy sposób ustalania wymiarów tablicy, opierając się na dopasowaniu współczynnika korelacji szeregu wyliczającego danych do jego wartości wyznaczonej z tablicy.

1. Zastosowane miary

Teoria pomiaru wyróżnia cztery skale, dla których można przeprowadzać analizy ilościowe: najsłabszą skalę nominalną, porządkową, przedziałową, i najsilniejszą z nich skalę ilorazową. Dla każdej z tych skal istnieje wiele miar statystycznych pozwalających na wnioskowanie o zależności zmiennych [4; 5], przy czym dla skali wyższych można korzystać również z miar określonych dla skali niższych.

Statystyka odróżnia zależność stochastyczną i jej najczęściej badany rodzaj – zależność korelacyjną – od zależności funkcyjnej. Pierwszy typ zależności oznacza, że zmiana jednej zmiennej prowadzi do zmiany rozkładu prawdopodobieństwa drugiej z nich, podczas gdy w zależności funkcyjnej zmiana jednej wartości zmiennej prowadzi do ściśle określonej zmiany wartości drugiej zmiennej. Terminu „tablica korelacyjna" wolno używać tylko wtedy, gdy można badać zależność korelacyjną zmiennych. Ma to miejsce w przypadku, gdy zebrane dane są mierzalne i gdy określonym wartościom jednej zmiennej można przyporządkować pewne średnie z kilku wartości drugiej zmiennej. Użyte w opracowaniu dane spełniają ten warunek, a ponadto dotyczą skali przedziałowej. W prowadzonych badaniach mogą więc zostać wykorzystane zarówno mierniki tej skali, jak i miary niższej skali nominalnej.

1.1. Miary zależności dla skali nominalnej

Do pomiaru siły zależności na tej skali wykorzystuje się miary zależności statystycznej, obliczane na podstawie tablic wielodzielczych (kontyngencji), oparte na wielkości statystyki χ^2:

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} \tag{1}$$

gdzie:

n_{ij} – empiryczne liczebności tablicy kontyngencji,

$\hat{n}_{ij}$ – teoretyczne liczebności tablicy kontyngencji,

i – numer wiersza tej tablicy,

j – numer kolumny tablicy.

Teoretyczne liczebności tablicy kontyngencji wyznacza się przy założeniu niezależności stochastycznej zmiennych X i Y:

$$\hat{n}_{ij} = \frac{n_i \cdot n_j}{n} \tag{2}$$

gdzie:

n – całkowita liczebność tablicy kontyngencji,
n_i – liczba wystąpień wartości $X = X_i$,
n_j – liczba wystąpień wartości $Y = Y_j$.

Zmienne uznaje się za niezależne, gdy wielkość empiryczna χ^2 nie przekracza statystyki tej liczby odczytywanej z tablic rozkładu χ^2 dla zadanego poziomu istotności α i (r-1)(s-1) stopni swobody.

Najczęściej stosowane miary zależności oparte na statystyce χ^2 to m.in. współczynniki[1]:

– Czuprowa:

$$T^2 = \frac{\chi^2}{n\sqrt{(r-1)(s-1)}} \tag{3}$$

– Cramera:

$$C^2 = \max\{\frac{\chi^2}{n \cdot (r-1)}; \frac{\chi^2}{n \cdot (s-1)}\} \tag{4}$$

– Hellwiga[2]:

$$\delta^2 = \frac{n - \sum_{i,j} \min(n_{ij}, \hat{n}_{ij})}{n \cdot (1 - \frac{1}{\min(r,s)})} \tag{5}$$

gdzie:

r – liczba wartości, jakie może przyjmować zmienna losowa X,
s – liczba wartości, jakie może przyjmować zmienna losowa Y.

Współczynnik Czuprowa wynosi 1 dla kwadratowej tablicy kontyngencji, w której każdej kategorii zmiennej odpowiada tylko jedna kategoria drugiej zmiennej. Miara Cramera osiąga górną granicę wtedy i tylko wtedy, gdy każdy wiersz lub każda kolumna zawiera tylko jeden element różny od zera.

[1] We wzorach podano kwadraty omawianych mierników.
[2] Wzór podano po zamianie oznaczeń charakteryzujących teoretyczne rozkłady zmiennej [1] na oznaczenia charakteryzujące rozkłady empiryczne.

Współczynnik Hellwiga jest równy 1, gdy niezerowe elementy tablicy kontyngencji są sobie równe i leżą na jednej z przekątnych macierzy.

Takie własności opisanych współczynników mogą doprowadzić do sytuacji, w której siła związku zmiennej samej ze sobą nie zostanie uznana za równą 1. Jest to ich wadą.

Zaletą opisanych miar nieparametrycznych jest ich uniwersalność. Mogą one być stosowane do pomiaru zależności cech mierzalnych, niemierzalnych, a także w sytuacjach, gdy jedna z nich jest mierzalna, a druga niemierzalna.

1.2. Miary zależności dla skali przedziałowej

Najpopularniejszą miarą zależności zmiennych jest współczynnik korelacji liniowej Pearsona. Wyznacza się go następująco:

$$r_{xy} = \frac{\operatorname{cov}(X,Y)}{S_X \cdot S_Y} \tag{6}$$

$$\operatorname{cov}(X,Y) = \frac{\sum_{i=1}^{r}\sum_{j=1}^{s}(x_i - \bar{x})(y_j - \bar{y}) \cdot n_{ij}}{n} \tag{7}$$

gdzie:

S_X, S_Y – odchylenia standardowe zmiennych,
$\bar{x}$, $\bar{y}$ – średnie wartości zmiennych.

Miara ta ma kilka ograniczeń, takich jak:
- brak skorelowania między zmiennymi nie oznacza, że są one niezależne,
- współczynnik ten można stosować tylko w przypadku regresji liniowej.

Współczynnik korelacji liniowej przyjmuje wartości od -1 do 1. Wartość dodatnia jest przyjmowana, gdy wraz ze wzrostem wartości zmiennej niezależnej rosną wartości zmiennej zależnej, natomiast ujemna – w przypadku ujemnej zależności zmiennych. Wartości r_{xy} zbliżone do 0 oznaczają słabą zależność korelacyjną liniową, bliskie jedności – silną.

Współczynnik korelacji o wartości 0 oznacza brak liniowego związku korelacyjnego między zmiennymi. Może mieć jednak wtedy miejsce nieliniowy związek korelacyjny, który można wykryć za pomocą stosunków (wskaźników) korelacyjnych.

Stosunki korelacyjne (zwane też wskaźnikami siły korelacji lub stosunkami współzależnościowymi) wprowadzone przez K. Pearsona są wyznaczane ze wzorów:

$$e_{xy} = \frac{S\bar{x}_{/j}}{S_X} \tag{8}$$

$$e_{yx} = \frac{S\bar{y}_{/i}}{S_Y} \tag{9}$$

Stosowanie tego miernika ma sens tylko wtedy, gdy dane są liczne i mogą być ujęte w formie tablicy korelacyjnej.

Podstawą tego miernika jest równość wariancyjna, która mówi, że ogólna wariancja zmiennej zależnej jest równa sumie wariancji średnich warunkowych tej zmiennej i średniej z jej wariancji warunkowych.

$$S_y^2 = S^2\bar{y}_{/i} + \overline{S_i^2 y} \tag{10}$$

Wariancja średnich warunkowych, nazywana też wariancją międzygrupową, mierzy zróżnicowanie wartości zmiennej zależnej (Y) spowodowane zmiennością zmiennej niezależnej (X), natomiast średnia z wariancji warunkowych (średnia wariancja wewnątrzgrupowa) określa tę część zmienności zmiennej zależnej, która jest spowodowana innymi czynnikami niż uwzględniona w badaniu zmienna niezależna.

Wariancję międzygrupową zmiennych wyznacza się następująco:

$$S^2\bar{y}_{/i} = \frac{\sum_{i=1}^{r}(\bar{y}_{/i} - \bar{y})^2 \cdot n_i}{n} \tag{11}$$

$$S^2\bar{x}_{/j} = \frac{\sum_{j=1}^{s}(\bar{x}_{/j} - \bar{x})^2 \cdot n_j}{n} \tag{12}$$

$$\bar{y}_{/i} = \frac{\sum_{i=1}^{r} y_j \cdot n_{ij}}{n_i} \tag{13}$$

$$\bar{x}_{/j} = \frac{\sum_{j=1}^{s} x_i \cdot n_{ij}}{n_j} \tag{14}$$

gdzie:

$\bar{y}_{i/}, \bar{x}_{/j}$ – średnia warunkowa zmiennych,
i – numer wiersza tej tablicy,
j – numer kolumny tablicy,
r – liczba wartości, jakie może przyjmować zmienna losowa X,
s – liczba wartości, jakie może przyjmować zmienna losowa Y.

Stosunki korelacyjne przyjmują unormowane wartości z przedziału <0,1>. Wartość 0 świadczy o braku zróżnicowania średnich warunkowych zmiennych, a więc o ich niezależności korelacyjnej. Wartość 1 jest przyjmowana w sytuacji, gdy zróżnicowanie zmiennej zależnej jest całkowicie wyjaśniane przez zróżnicowanie zmiennej ją objaśniającej.

Stosunek korelacyjny pozwala na obliczenie zarówno korelacji prostoliniowej, jak i krzywoliniowej. Uwzględnia również to, że zależność zmiennych może nie być obustronna, tj. wpływ jednej zmiennej na drugą może mieć siłę różną od siły, z jaką druga zmienna kształtuje pierwszą.

Spośród opisanych poniżej mierników tylko współczynnik korelacji może być stosowany do wnioskowania o kierunku zależności zmiennych, przy założeniu liniowego charakteru tej zależności. Współczynniki T, C, δ oraz e_{xy} dopuszczają zależność krzywoliniową zmiennych i w związku z tym przyjmują tylko wartości dodatnie. Podczas gdy współczynniki: r_{xy}, T, C, δ zakładają symetryczność relacji dwóch zmiennych, stosunek korelacyjny e_{xy} jest miarą, która dopuszcza różne siły wzajemnego oddziaływania zmiennych na siebie.

2. Tworzenie tablicy korelacyjnej

Do pewnych ograniczeń zastosowanych mierników można zaliczyć wymaganie zestawienia danych w formie tablicy wielodzielczej. W celu uproszczenia obliczeń, a także lepszego odwzorowania charakteru zależności zmiennych, pożądany jest podział szeregów danych na przedziały. Problemem jest ustalenie rozpiętości i liczby tych przedziałów. Dla określonej liczebności zebranego materiału jest zalecana górna i dolna liczba klas przedziałów, co daje kilka możliwych rozwiązań.

W niniejszym opracowaniu problem wyboru wymiarów tablicy korelacyjnej został rozwiązany przez dobór takiego wariantu, dla którego otrzymana wartość współczynnika korelacji liniowej Pearsona jest jak najbardziej zbliżona do wartości tego współczynnika wyznaczonej dla danych indywidualnych. Wartość tego współczynnika można wyznaczać dla obydwu form prezentacji materiału liczbowego. Kryterium określenia wymiarów tablicy korelacyjnej jest więc jak najmniejsze zniekształcenie wyników, które zostałyby otrzymane, gdyby dane nie zostały przedstawione w formie przedziałowej tablicy korelacyjnej.

Minimalna liczba kolumn/wierszy została ustalona jako zaokrąglona w górę minimalna wartość spośród opisanych poniżej wskaźników, podczas gdy górny próg liczebności stanowi zaokrąglona w dół maksymalna wartość otrzymana z poniższych wzorów:

$$k = 1 + 3{,}322 \cdot \log n \tag{15}$$

$$k = \sqrt{n} \tag{16}$$

$$k = 5 \cdot \log n \tag{17}$$

gdzie:

k – minimalna liczba kolumn/wierszy.

Tak przedstawiony schemat postępowania został zautomatyzowany za pomocą programu Ms Excel dla prób obejmujących do 120 obserwacji. Dla najdokładniejszych dostępnych pomiarów statystycznych, które z reguły dotyczą miesięcy, liczba ta odzwierciedla ciąg obserwacji z 10 lat. Nie wydaje się być uzasadnione zbieranie pomiarów z dłuższego okresu, ponieważ tracą one wtedy swoją aktualność.

Rozpatrzono trzy możliwe warianty liczby kolumn oraz wierszy. Założono również, że liczba kolumn może być różna od liczby wierszy, co daje maksymalnie 9 możliwych wymiarów. Po podstawieniu maksymalnej założonej liczby obserwacji (120) do powyższych wzorów, można zauważyć, że dla tego zakresu liczebności próby nie ma większej liczby możliwych wariantów.

Dla określonej liczby klas ich rozpiętość została nieznacznie podwyższona (o 0,0001) w celu zapewnienia, że maksymalna wartość badanej zmiennej zmieści się w ostatnim prawostronnie otwartym przedziale.

3. Prezentacja przykładowych danych

Opisane współczynniki zostaną wyznaczone dla 3 przykładowych serii danych o różnej liczebności próby (100, 60 i 36 obserwacji) w celu ich porównania oraz określenia, które wymiary macierzy korelacyjnej były najczęściej wybierane przez zaprezentowany algorytm.

Zebrany materiał liczbowy[3] dotyczy sprzedaży (ogółem i na eksport), produkcji i parametrów jakościowych pewnego surowca w polskim przedsiębiorstwie wydobywczym i jego dwóch zakładach. Dodatkowo w przykładzie 3 wykorzystamy dane dotyczące produkcji materiału zużywającego surowiec w krajach jego największych europejskich producentów i cen wyrobów finalnych tworzonych z tego materiału.

Zebrane dane dotyczą miesięcznych okresów (lata 2000-2004) z wyjątkiem przykładu 1, który korzysta z danych dotyczących dziennych transakcji przedsiębiorstwa.

1. Pierwsza seria danych dotyczy całego przedsiębiorstwa:
   - sprzedaż surowca [t] – y,
   - wartość opałowa [Mcal] – X_1,
   - zawartość wilgoci [%] – X_2,
   - zawartość popiołu [%] – X_3,
   - zawartość siarki [%] – X_4.
2. Druga seria danych dotyczy jednego z zakładów przedsiębiorstwa:
   - produkcja surowca [t] – y,
   - zatrudnienie – X_1,
   - sprzedaż surowca [t] – X_2,
   - średnia cena sprzedaży [zł/t] – X_3,
   - koszt sprzedaży [zł/t] – X_4,
   - wartość opałowa [Mcal] – X_5,
   - zawartość wilgoci [%] – X_6,
   - zawartość popiołu [%] – X_7,
   - zawartość siarki [%] – X_8.
3. Trzecia seria danych dotyczy drugiego zakładu przedsiębiorstwa oraz danych rynkowych:
   - sprzedaż surowca [t] – y,
   - średnia cena sprzedaży [zł/t] – X_1,
   - koszt sprzedaży [zł/t] – X_2,
   - produkcja materiału zużywającego surowiec [tys. t]:

[3] Ze względu na tajemnicę handlową w opracowaniu nie podano nazwy zakładu ani produkowanego przez niego surowca.

  - w Niemczech – X_3,
  - w Rosji – X_4,
  - na Ukrainie – X_5,
- ceny produktów wykorzystujących ten materiał w Europie [USD/t]:
  - produkt I – X_6,
  - produkt II – X_7,
- wartość opałowa [Mcal] – X_8,
- zawartość wilgoci [%] – X_9,
- zawartość popiołu [%] – X_{10},
- zawartość siarki [%] – X_{11}.

4. Oczekiwane i otrzymane wielkości współczynników zależności

Dla użytych danych liczbowych otrzymano następujące wielkości mierników zależności zmiennych[4]:

Przykład 1

Spodziewany wpływ zmiennych X_1-X_4 na zmienną y powinien mieć charakter malejący dla zmiennych X_2-X_4 i stymulujący dla zmiennej X_1, która jest parametrem wymaganym przez odbiorców.

Dla zastosowanych danych otrzymano następujące wartości mierników zależności:

- Macierz współczynników Czuprowa: $M_T^T = [0{,}299 \quad 0{,}205 \quad 0{,}211 \quad 0{,}163]$.
- Macierz współczynników Cramera: $M_c^T = [0{,}309 \quad 0{,}212 \quad 0{,}218 \quad 0{,}173]$.
- Macierz współczynników Hellwiga: $M_\delta^T = [0{,}448 \quad 0{,}483 \quad 0{,}432 \quad 0{,}386]$.
- Macierz współczynników korelacyjnych: $M_r^T = [0{,}011 \quad 0{,}054 \quad 0{,}169 \quad 0{,}024]$.
- Macierz stosunków korelacyjnych: $M_e = [0{,}318 \quad 0{,}215 \quad 0{,}237 \quad 0{,}244]$.

Porównując otrzymane wartości z posiadaną wiedzą na temat zmiennych, można zauważyć, że wpływ zmiennych jakościowych na sprzedaż badanego surowca jest bardzo mały, co więcej, zmienne X_2-X_4 nie wykazały malejącego wpływu na zmienną. Może to wynikać z bardzo słabej korelacji tych zmiennych, której kierunek trudno ocenić.

[4] Wartości 1 na przekątnej macierzy zostały założone odgórnie, ponieważ nieuzasadnione jest badanie korelacji zmiennej samej ze sobą, a nie wszystkie rozpatrywane macierze były kwadratowe.

Trzy spośród badanych mierników wskazały na zmienną X_1 jako najistotniej kształtującą y spośród wybranych zmiennych. Poza tym współczynniki Hellwiga wskazują na największy wpływ drugiej zmiennej (wilgotność), a współczynniki korelacji na zmienną X_3 (zawartość popiołu). Prawie wszystkie zastosowane mierniki (z wyjątkiem stosunków korelacyjnych) wykazują najsłabszy wpływ zmiennej X_4 (zawartość siarki w surowcu).

Podczas gdy współczynniki Czuprowa, Cramera i stosunki korelacyjne wskazują na zbliżoną słabą siłę zależności, wszystkie współczynniki Pearsona wskazują na praktycznie zerową siłę zależności zmiennych. Przyczyną tej różnicy jest prawdopodobnie słaby krzywoliniowy charakter zależności zmiennych, którego współczynnik korelacji liniowej nie uwzględnia. Największą – umiarkowaną siłę korelacji zmiennych objaśniających z objaśnianą wskazuje współczynnik Hellwiga.

Przykład 2

Obliczone mierniki powinny wskazywać na silny dodatni wpływ zmiennych X_1 i X_2 na zmienną y oraz na silny ujemny wpływ zmiennych X_3 i X_4. Siłę wpływu zmiennych jakościowych na tę zmienną trudno określić, przypuszczalnie kierunek tej zależności będzie malejący, z wyjątkiem zmiennej X_1.

- Macierz współczynników Czuprowa:

$$M_T^T = [0{,}319 \quad 0{,}633 \quad 0{,}282 \quad 0{,}367 \quad 0{,}326 \quad 0{,}326 \quad 0{,}333 \quad 0{,}315]$$

- Macierz współczynników Cramera:

$$M_c^T = [0{,}319 \quad 0{,}633 \quad 0{,}282 \quad 0{,}381 \quad 0{,}326 \quad 0{,}326 \quad 0{,}346 \quad 0{,}315]$$

- Macierz współczynników Hellwiga:

$$M_\delta^T = [0{,}617 \quad 0{,}761 \quad 0{,}5 \quad 0{,}617 \quad 0{,}575 \quad 0{,}599 \quad 0{,}611 \quad 0{,}556]$$

- Macierz współczynników korelacji:

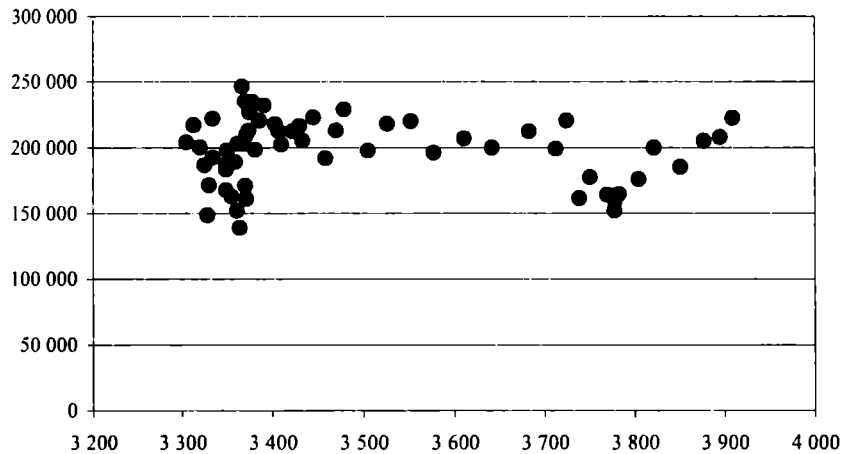
$$M_r^T = [-0{,}126 \quad 0{,}871 \quad -0{,}13 \quad -0{,}551 \quad 0{,}339 \quad 0{,}078 \quad -0{,}216 \quad -0{,}069]$$

- Macierz stosunków korelacyjnych:

$$M_e = [0{,}515 \quad 0{,}903 \quad 0{,}37 \quad 0{,}594 \quad 0{,}417 \quad 0{,}295 \quad 0{,}28 \quad 0{,}154]$$

Wyznaczone wartości potwierdzają początkowe przypuszczenia, jednoznacznie wskazując na najsilniejszy wpływ zmiennej X_2 (sprzedaż) na badaną wielkość (produkcja). Zgodne z oczekiwaniami wartości współczynników korelacji zmiennych X_3 i X_4 (cena i koszty sprzedaży) są ujemne, przy czym wszystkie miary zgodnie określają wpływ pierwszej z tych zmiennych jako słaby, a wpływ drugiej jako najsilniejszy po zmiennej X_2 w wybranym zbiorze

zmiennych. Zaskakujące są natomiast wskaźniki dotyczące wpływu zatrudnienia na zmienną objaśnianą. Tylko współczynnik Hellwiga oraz stosunki korelacyjne wskazują na jej duże znaczenie dla wielkości produkcji, przy czym pierwsza z wymienionych miar po raz kolejny wskazuje na znacznie większą i mniej więcej równą siłę zależności dla wszystkich badanych zmiennych. Dodatkowo, przy założeniu liniowej zależności, współczynnik korelacji określa jej kierunek jako ujemny, co jest zupełnie niezgodne z logiką. Wynika stąd, że łącząca zmienne zależność ma charakter krzywoliniowy i nie da się ocenić jej kierunku. Dowodzi tego duża wartość stosunku korelacji oraz sporządzony diagram korelacyjny tych zmiennych.



Rys. 1. Diagram korelacyjny zmiennych produkcji pewnego zakładu i jego zatrudnienia

Jeśli chodzi o wpływ czynników jakościowych, zastosowane miary wskazują na nieco silniejszy ich wpływ na zmienną objaśnianą niż w przykładzie 1. Można z tego wnioskować, że jakość ma większy wpływ na ilość wyprodukowanego surowca niż na wielkość jego sprzedaży. Można jednak również przypuszczać, że do ogólnie mniejszych wartości współczynników zależności w przykładzie 1 przyczyniła się duża liczebność rozpatrywanej w nim próby dziennych danych, która utrudniła dostrzeżenie właściwych zależności zmiennych, podczas gdy przykład drugi opierał się na bardziej wygładzonych danych miesięcznych.

Współczynniki skal nominalnych wskazują w tym przypadku na najsilniejszy wpływ zmiennej X_7 (udział popiołu), podczas gdy miary skali przedziałowej jako najsilniejszy podają wpływ zmiennej X_5 (wartość opałowa

surowca). Również ujemne kierunki zależności w przykładzie zmiennych X_7 i X_8 odpowiadają oczekiwanym. Bardzo słabą zależność rosnącą zmiennych y i X_6 można wyjaśnić tak, jak w poprzednim przykładzie.

Przykład 3

Zależności zmiennych charakteryzujących sprzedaż, produkcję i jakość powinny się kształtować podobnie jak w poprzednich przykładach. Wpływ produkcji materiału w Europie – zmienne X_3-X_5 powinny być dość silną stymulantą dla sprzedaży, podczas gdy wpływ zmiennych X_6 i X_7 powinien mieć ujemny wpływ na zmienną y.

- Macierz współczynników Czuprowa:

$$M_T^T = [0{,}318 \quad 0{,}376 \quad 0{,}39 \quad 0{,}469 \quad 0{,}493 \quad 0{,}358 \quad 0{,}373 \quad 0{,}395 \quad 0{,}358 \quad 0{,}367 \quad 0{,}277]$$

- Macierz współczynników Cramera:

$$M_c^T = [0{,}318 \quad 0{,}376 \quad 0{,}39 \quad 0{,}491 \quad 0{,}493 \quad 0{,}358 \quad 0{,}373 \quad 0{,}414 \quad 0{,}375 \quad 0{,}384 \quad 0{,}277]$$

- Macierz współczynników Hellwiga:

$$M_\delta^T = [0{,}559 \quad 0{,}536 \quad 0{,}525 \quad 0{,}612 \quad 0{,}664 \quad 0{,}632 \quad 0{,}635 \quad 0{,}654 \quad 0{,}612 \quad 0{,}587 \quad 0{,}519]$$

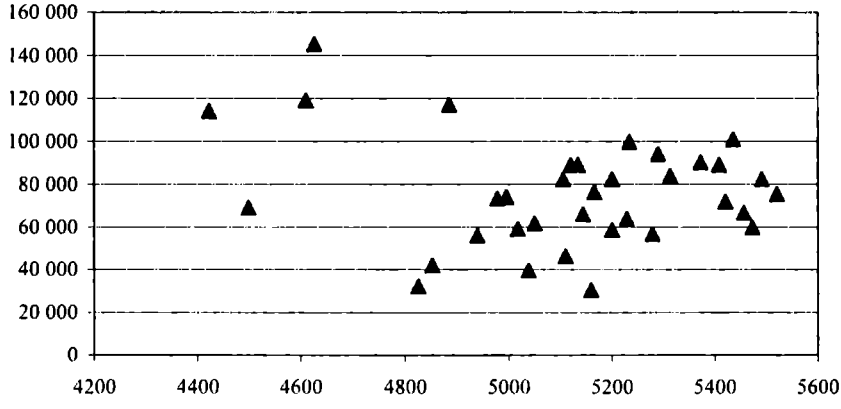
- Macierz współczynników korelacji:

$$M_r^T = [0{,}073 \quad -0{,}152 \quad -0{,}093 \quad -0{,}204 \quad -0{,}124 \quad 0{,}048 \quad -0{,}007 \quad -0{,}236 \quad 0{,}219 \quad 0{,}114 \quad 0{,}117]$$

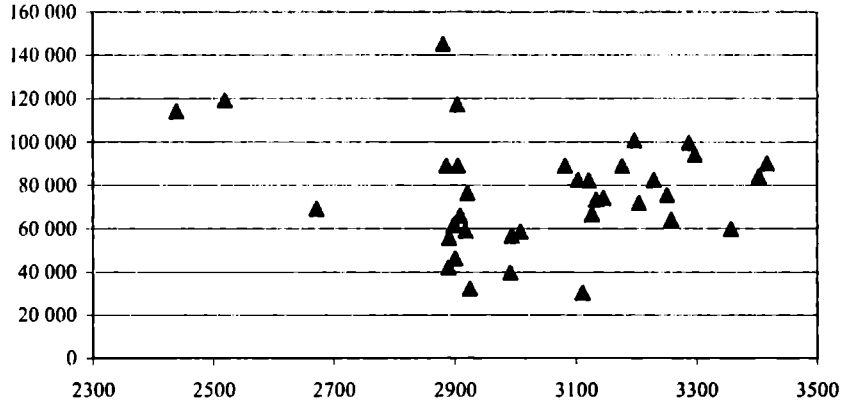
- Macierz stosunków korelacji:

$$M_e = [0{,}237 \quad 0{,}289 \quad 0{,}318 \quad 0{,}59 \quad 0{,}439 \quad 0{,}311 \quad 0{,}407 \quad 0{,}447 \quad 0{,}414 \quad 0{,}223 \quad 0{,}137]$$

Wyznaczone mierniki wskazują, że największy wpływ na badaną zmienną (sprzedaż surowca na eksport) miały zmienne X_4 i X_5.



Rys. 2. Diagram korelacyjny sprzedaży na eksport surowca i produkcji zużywającego go materiału w Rosji



Rys. 3. Diagram korelacyjny sprzedaży na eksport surowca i produkcji zużywającego go materiału na Ukrainie

Cena i koszt sprzedaży wykazują niewielki wpływ na zmienną objaśnianą, przy czym wpływ ceny sprzedaży na podstawie współczynnika korelacji Pearsona jest dodatni. Zjawisko to można wyjaśnić bardzo słabą i trudno rozpoznawalną zależnością zmiennych lub pominięciem w badaniu innych czynników istotnych dla kształtowania się relacji tych zmiennych.

Wpływ cen wyrobów końcowych na zmienną objaśnianą można określić jako umiarkowany na podstawie współczynników T,C, e_{yx}, bardzo słaby (r_{xy}) lub silny (δ). Wobec tak sprzecznych informacji trudno wyciągnąć wnioski z otrzymanych wielkości.

W grupie zmiennych jakościowych jednoznacznie została wskazana zmienna X_1 (wartość opałowa surowca) jako najistotniej wpływająca na y. Siła tego wpływu w większości przypadków jest określona jako umiarkowana (z wyjątkiem współczynnika Hellwiga). Po raz kolejny otrzymane znaki zależności zmiennych jakościowych ze zmienną objaśnianą są niezgodne z wiedzą na temat tej zależności. Sporządzone diagramy korelacyjne potwierdziły tylko słabą siłę zależności tych zmiennych. Przyczyną może być faktyczny brak zależności zmiennych jakościowych i sprzedaży na eksport lub nieuwzględnienie w badaniu dodatkowych czynników.

5. Otrzymane wymiary macierzy

W rozdziale 2 opisano algorytm doboru wymiarów macierzy korelacyjnej pod względem najlepszego dopasowania wartości współczynnika korelacji. Tabela 1 prezentuje procentowy udział określonych wymiarów macierzy uzyskanych w poszczególnych przykładach.

Tabela 1

Udziały procentowe wykorzystanych w badaniach wariantów wymiarów macierzy korelacyjnej

Wymiary	axa	axb	axc	bxa	bxb	bxc	cxa	cxb	cxc
Przykład 1	30,00%	10,00%	0,00%	10,00%	0,00%	20,00%	30,00%	0,00%	0,00%
Przykład 2	22,22%	27,78%	-	22,22%	27,78%	-	-	-	-
Przykład 3	24,24%	21,21%	-	27,27%	27,27%	-	-	-	-

gdzie:

- a – najmniejszy wymiar macierzy korelacji, który wynosił:
  - 8 – przykład 1, 100 obserwacji,
  - 7 – przykład 2, 60 obserwacji,
  - 6 – przykład 3, 36 obserwacji,
- b – następny po a wymiar macierzy, b = a+1 przyjmowany tylko w sytuacji, gdy nie przekracza ustalonej górnej granicy wymiarów macierzy[5], wynosił:
  - 9 – przykład 1, 100 obserwacji,
  - 8 – przykład 2, 60 obserwacji,
  - 7 – przykład 3, 36 obserwacji,
- c – następny po b wymiar macierzy, c = b+1 przyjmowany tylko w sytuacji, gdy nie przekracza ustalonej górnej granicy wymiarów macierzy, wynosił:
  - 10 – przykład 1, 100 obserwacji,
  - w przykładach 2 i 3 c przekroczyło górną dopuszczalną liczbę wierszy/kolumn i było pominięte.

Powyższe wyniki pozwalają zaobserwować, że najlepsze dopasowanie współczynnika korelacji liniowej było uzyskiwane często zarówno dla macierzy kwadratowych (axa, bxb, cxc), jak i dla macierzy niekwadratowych. Dla najliczniejszej próby macierz kwadratowa była wykorzystywana rzadziej niż prostokątna (razem w 40% przypadków), dla 60 obserwacji w równo 50% przypadków. Dla najmniejszej liczby obserwacji (36) przeważały macierze kwadratowe (51,51% przypadków).

[5] Rozdział 2.

We wszystkich przykładach uzyskano dość duży rozrzut wykorzystanych w badaniach wymiarów tablicy korelacyjnej wśród określonych wariantów. Dowodzi to, że nie ma pewnej reguły określania wymiarów tej macierzy, która mogłaby zapewnić zgodność uzyskanych z niej wyników z wynikami uzyskanymi dla pierwotnych szeregów zmiennych.

Podsumowanie

Przeprowadzone badania wskazały na częstą rozbieżność wyników otrzymanych dla mierników skali nominalnej oraz przedziałowej. Jak można zauważyć, współczynniki skali nominalnej (a zwłaszcza współczynnik Hellwiga) znacznie zawyżają siłę zależności zmiennych w stosunku do współczynnika korelacji liniowej Pearsona. Dodatkowo miary te, niezależnie od liczebności próby, zacierają różnice pomiędzy siłą zależności zmiennych, określając ją jako albo średnią czy słabą (T, C), albo silną (δ) praktycznie dla wszystkich badanych zmiennych. Chociaż miary te można stosować dla wyższych skal, wskazane wydaje się być używanie mierników określonych dla możliwie najwyższej skali. Najlepszym miernikiem krzywoliniowej zależności zmiennych jest w tej sytuacji stosunek korelacyjny.

Stosunek korelacyjny jest jednocześnie jedyną znaną miarą zależności dwóch zmiennych, która pozwala na odróżnienie ich wzajemnego wpływu. Jednak ze względu na swoją konstrukcję, zawsze wskazuje on na silniejszą zależność zmiennych niż współczynnik korelacji, tym samym nie dopuszczając możliwości badania nieprzechodniej relacji dwóch zmiennych w sytuacji, gdy wpływowi jednej zmiennej nie towarzyszy oddziaływanie ze strony drugiej zmiennej.

Świadomość ograniczeń istniejących mierników dowodzi konieczności formułowania wniosków o charakterze zjawisk i łączących je relacji, opierając się na więcej niż jednej mierze statystycznej. Ma to szczególne znaczenie dla zmiennych, których zależność ma charakter krzywoliniowy i może zostać niedostrzeżona, jeśli jedyną stosowaną w badaniu miarą będzie współczynnik korelacji liniowej.

Analizując otrzymane wyniki, można niejednokrotnie dostrzec pewne punkty wspólne, ale także różnice w odpowiedzi na pytanie o zmienne najistotniej wpływające na badaną wielkość. Przeprowadzone badania wydają się więc potwierdzać, że nie opracowano do tej pory idealnego miernika, którego interpretacja sama w sobie stanowiłaby wystarczającą podstawę do wnioskowania o zależności zmiennych.

Literatura

1. Hellwig Z.: *Elementy rachunku prawdopodobieństwa i statystyki matematycznej*. Wydawnictwo Naukowe PWN, Warszawa 1995.
2. Kaczmarczyk S.: *Badania marketingowe. Metody i techniki*. PWE, Warszawa 2003.
3. Kassyk-Rokicka H.: *Statystyka nie jest trudna. Mierniki statystyczne*. PWE, Warszawa 1997.
4. *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*. Red. E. Gatnar, M. Walesiak. AE, Wrocław 2004.
5. Walesiak M.: *Metody analizy danych marketingowych*. Wydawnictwo Naukowe PWN, Warszawa 1996.
6. Zając K.: *Zarys metod statystycznych*. PWE, Warszawa 1994.

MEASURES OF THE VARIABLES' RELATIONSHIP FOR THE DATA PRESENTED IN THE CORRELATION ARRAY

Summary

Article includes description of a few commonly used measures of variables' relationship. The description concerns measures based on the observations of variables presented as an array. These measures apply to the nominal scale (the coefficients of Czuprow, Cramer, Hellwig) or to a correlation relationship in the interval scale (the Pearson's coefficient of linear correlation and the correlation ratio). The mentioned coefficients are used to measure the relationship between the selected explanatory variables and three explained variables. Statistical data, used in the research, concerns production, sale and quality of some material. Dimensions of the constructed arrays are established, respecting the formal requirements, in order to maximize the fitting of correlation coefficients to its values obtained for individual data. A comparison of the measures, used in research, has been made on the basis of the obtained results. The frequency, in which some variants of array's dimensions gave the best precision of correlation coefficient value, has been defined.