Józef Czaja *, Edward Preweda *

ANALIZA STATYSTYCZNA ZMIENNEJ LOSOWEJ WIELOWYMIAROWEJ W ASPEKCIE KORELACJI I PREDYKCJI **

1. Zmienna losowa dwuwymiarowa

W badaniach statystycznych wielowymiarowej zmiennej losowej zakłada się, że jedna zmienna losowa stanowi zmienną objaśnianą (zależną), zaś pozostałe zmienne losowe mają charakter zmiennych objaśniających (niezależnych). Zależność zmiennej objaśnianej względem poszczególnych zmiennych objaśniających może być opisywana za pomocą jednego modelu wielorakiej regresji lub za pomocą kilku niezależnych modeli dwuwymiarowej regresji.

W analizie statystycznej rynku zmienną zależną jest cena, zaś cechy nieruchomości reprezentują zmienną niezależną. Związki między ceną i tymi cechami nieruchomości można opisywać za pomocą jednego modelu wielowymiarowej zmiennej losowej lub za pomocą kilku niezależnych modeli dwuwymiarowej zmiennej losowej.

W celu ustalenia modelu regresji dla dwóch zmiennych losowych, np.: X -zmiennej niezależnej reprezentującej atrybut i Y -zmiennej zależnej reprezentującej cenę nieruchomości, trzeba dla nich ustalić skale interwałowe. Graficzny obraz tych dwóch zmiennych w ortogonalnym układzie współrzędnych nosi nazwę diagramu korelacyjnego. Na podstawie tego obrazu można wstępnie wnioskować, jaka jest współzależność pomiędzy obiema zmiennymi.

Parametrem określającym wzajemną zależność zmiennych losowych w dwuwymiarowym rozkładzie jest mieszany moment centralny drugiego rzędu, czyli kowariancja pomiędzy zmienną X i zmienną Y, którą oznacza się przez cov[Y,X]. W zależności od wymiarów rozpatrywanych zmiennych losowych wielkość cov[Y, X] może przyjmować

* Akademia Górniczo-Hutnicza, Katedra Informacji o Terenie

** Praca stanowi wynik realizacji badań statutowych w AGH Nr 11.11.150.316

różne wartości. W związku z tym, wielkość cov[Y, X] można standaryzować za pomocą wartości odchyleń standardowych obu zmiennych losowych w rozkładach brzegowych, czyli $\sigma[X]$ i $\sigma[Y]$.

Standaryzowana cov[Y, X] stanowi miarę współzależności liniowej zmiennych losowych X i Y w dwuwymiarowym rozkładzie i nosi nazwę **współczynnika korelacji liniowej (Pearsona)**, czyli

$$r = \frac{\operatorname{cov}[X,Y]}{\sigma[X]\cdot\sigma[Y]} \tag{1.1}$$

Jeżeli wartości zmiennych losowych zaobserwowane w próbie (x_i, y_i), dla $i = 1, 2, ..., n$, będą posiadały równomierny rozkład prawdopodobieństwa

$$p_i = \frac{1}{n} \tag{1.2}$$

to estymatorem (oszacowaniem z próby) współczynnika korelacji liniowej będzie wartość współczynnika $\hat{r}$ liczona według następującego wzoru:

$$\hat{r} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_iy_i)-\bar{x}\cdot\bar{y}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i)^2-\bar{x}^2}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i)^2-\bar{y}^2}} \tag{1.3}$$

przy czym

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i \tag{1.4}$$

$$\sigma[x] = \sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}, \qquad \sigma[y] = \sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{1.5}$$

oznaczają wartość przeciętną oraz odchylenia standardowe rozpatrywanych zmiennych, określone na podstawie wyników z próby.

Siłę korelacji (współzależności) liniowej można określać na trzech poziomach, a mianowicie:

1) dla $|\hat{r}| \le 0{,}3$ – korelacja słaba,
2) dla $0{,}3 < |\hat{r}| \le 0{,}6$ – korelacja przeciętna,
3) dla $|\hat{r}| > 0{,}6$ – korelacja silna.

Jeżeli analizie statystycznej podlegają zmienne losowe, których opis jest utrudniony pod względem ilościowym, wtedy porównując elementy badanych zmiennych można uszeregować zaobserwowane wartości w ciągi niemalejące. Ustawienie elementów badanej zmiennej w ciąg niemalejący i przyporządkowanie im numeracji nosi nazwę nadawania rang. Numer miejsca, jakie element zajmuje w tym ciągu, nazywamy rangą

tego elementu. Dla próby losowej zawierającej n elementów rangami są liczby naturalne 1, 2, ..., n. Rangi są traktowane jako zaobserwowane wartości zmiennej losowej skokowej o równomiernym rozkładzie prawdopodobieństwa.

Jeżeli powyższe rozważania będą prowadzone w odniesieniu do dwóch cech (X, Y) i kolejnym obserwacjom każdej z cech z osobna nadano odpowiednie rangi, wtedy można wyznaczyć współczynnik korelacji rang (kolejności) tych cech.

Niech elementy (x_i, y_i) zaobserwowanej próby losowej będą uporządkowane według niemalejących wartości x_i, czyli

$$\begin{pmatrix} x_{(1)}, & x_{(2)}, & x_{(3)}, & ..., & x_{(n)} \\ y_1^* & y_2^* & y_3^* & ... & y_n^* \end{pmatrix}, \text{ gdzie: } x_{(1)} \le x_{(2)} \le x_{(3)} \le ... \le x_{(n)} \tag{1.6}$$

Dla zmiennej x_i rangami są liczby naturalne 1, 2, ..., n, zaś dla zmiennej y_i rangami są elementy $y_i^* \leftrightarrow s_i$.

Po zastąpieniu par elementów $(x_{(i)}, y_i^*)$ odpowiadającymi im rangami zbiór (1.6) zastępujemy zbiorem następującej postaci

$$\begin{pmatrix} 1, & 2, & 3, & ..., & n \\ s_1 & s_2 & s_3 & ... & s_n \end{pmatrix} \tag{1.7}$$

Z porównania (1.6) i (1.7) widać, że zbiór wyników obserwacji (x_i, y_i) został zastąpiony nowym zbiorem (i, s_i), którego elementy zawierają tylko numery kolejności uporządkowania wartości (x_i, y_i).

Współczynnik korelacji rang (kolejności uporządkowania) rozpatrywanych cech można zdefiniować na podstawie wzoru (1.3). Do realizacji wzoru (1.3) należy najpierw wyznaczyć wartości przeciętne rang cechy X i cechy Y. Ponieważ w jednym i drugim przypadku mamy do czynienia ze zmienną losową, która z prawdopodobieństwem równym $1/n$ przyjmuje wartości naturalne 1, 2, 3, ..., n, zatem wartości przeciętne rang cechy X i cechy Y będą identyczne i równe średniej arytmetycznej rang. Wykorzystując wzór na sumę kolejnych liczb naturalnych można zapisać wartości przeciętne rang, które wyrażają się następującą zależnością

$$\bar{x} = \bar{y} = 1 + 2 + 3 + ... + n = \frac{1}{2} n(n+1) \tag{1.8}$$

Po uwzględnieniu wartości rang według oznaczeń (1.7) i wartości (1.8) wzór (1.3) przyjmuje następującą postać

$$\hat{r}_s = \frac{\sum_{i=1}^{n} \left[i - \frac{1}{2}(n+1) \right] \cdot \left[s_i - \frac{1}{2}(n+1) \right]}{\sqrt{\sum_{i=1}^{n} \left[i - \frac{1}{2}(n+1) \right]^2} \cdot \sqrt{\sum_{i=1}^{n} \left[s_i - \frac{1}{2}(n+1) \right]^2}} \tag{1.9}$$

Sumy kwadratów wyrażeń występujących w mianowniku są identyczne, gdyż różnią się tylko kolejnością składników (liczb naturalnych), stąd ich wartość zostanie wyznaczona na podstawie wzoru na sumę kwadratów kolejnych liczb naturalnych, czyli

$$1^2+2^2+3^2+\ldots+n^2=\frac{1}{6}n\cdot(n+1)(2n+1) \tag{1.10}$$

Zatem

$$\sum_{i=1}^{n}\left[i-\frac{1}{2}(n+1)\right]^2=\frac{1}{n}\sum_{i=1}^{n}i^2-\frac{1}{4}(n+1)^2=\frac{(n+1)(2n+1)}{6}-\frac{(n+1)^2}{4}=\frac{n}{12}(n^2-1) \tag{1.11}$$

$$\sum_{i=1}^{n}\left[s_i-\frac{1}{2}(n+1)\right]^2=\frac{1}{n}\sum_{i=1}^{n}s_i^2-\frac{1}{4}(n+1)^2=\frac{(n+1)(2n+1)}{6}-\frac{(n+1)^2}{4}=\frac{n}{12}(n^2-1) \tag{1.12}$$

W celu uproszczenia wyrażenia występującego w liczniku wzoru (1.9) wykonamy następujące przekształcenie

$$\begin{aligned}\sum_{i=1}^{n}(i-s_i)^2&=\sum_{i=1}^{n}\left\{\left[i-\frac{1}{2}(n+1)\right]\left[s_i-\frac{1}{2}(n+1)\right]\right\}^2=\sum_{1=1}^{n}\left[i-\frac{1}{2}(n+1)\right]^2+\\&+\sum_{1=1}^{n}\left[s_i-\frac{1}{2}(n+1)\right]^2+2\sum_{i=1}^{n}\left[i-\frac{1}{2}(n+1)\right]\left[s_i-\frac{1}{2}(n+1)\right]=\\&=\frac{n}{6}(n^2-1)+2\sum_{i=1}^{n}\left[i-\frac{1}{2}(n+1)\right]\left[s_i-\frac{1}{2}(n+1)\right]\end{aligned} \tag{1.13}$$

Z tego przekształcenia wynika, że

$$\sum_{i=1}^{n}\left[i-\frac{1}{2}(n+1)\right]\left[s_i-\frac{1}{2}(n+1)\right]=\frac{1}{12}(n^3-n)\cdot\frac{1}{2}\sum_{n=1}^{n}(i-s_i)^2 \tag{1.14}$$

Po uwzględnieniu (1.11), (1.12) i (1.14) wzór (1.9) na współczynnik korelacji rang przyjmuje postać

$$\hat{r}_s=\frac{\frac{1}{12}(n^3-n)-\frac{1}{2}\sum_{i=1}^{n}(i-s_i)^2}{\frac{1}{12}(n^3-n)}=1-\frac{6\sum_{i=1}^{n}(i-s_i)^2}{n^3-n} \tag{1.15}$$

Wielkość $\hat{r}_s$ określona wzorem

$$\hat{r}_s=1-\frac{6\sum_{i=1}^{n}(i-s_i)^2}{n^3-n} \tag{1.16}$$

nosi nazwę **współczynnika korelacji rang (kolejności) Spearmana** – stąd wskaźnik s.

Oprócz współczynnika korelacji rang Spearmana można również, na bazie zbioru rang (1.7), konstruować współczynnik korelacji rang Kendalla. W tym celu, dla każdej z rang cechy Y ($y_i^* \leftrightarrow s_i$) formułujemy pary z wszystkimi występującymi po niej rangami. Jeżeli w tak utworzonych parach następnik jest większy niż poprzednik, to dla takich par przyporządkowujemy notę (+1), natomiast gdy następnik jest mniejszy niż poprzednik, to dla takich par przyporządkowujemy notę (–1).

Współczynnik korelacji rang Kendalla ($\hat{r}_k$) oblicza się według wzoru

$$\hat{r}_k = \frac{\textit{Suma przyporządkowanych not}}{\textit{Liczba wszystkich par}} = \frac{2U}{n(n-1)} \tag{1.17}$$

przy czym U oznacza sumę uzyskanych not (+1) i (–1) dla wszystkich utworzonych par z rang (s_i).

Na podstawie wykonanej analizy formuł określających współczynnik korelacji rang Spearmana i współczynnik korelacji rang Kendalla można wyrazić opinię, że współczynniki korelacji rang powinny być stosowane do opisywania współzależności takich cech, dla których są ograniczone możliwości ich kwantyfikowania. Opisywanie zmiennych za pomocą rang ma charakter uszeregowania tych zmiennych bez uwzględniania różnic występujących pomiędzy wartościami sąsiednich zmiennych. W przypadku gdy w rozpatrywanej próbie wartości cech powtarzają się, wtedy ich uszeregowanie (rangowanie) może być wykonane na wiele sposobów, a to dowodzi, że wartości współczynników korelacji rang mogą przyjmować różne wartości.

W statystycznych analizach ilościowych powinno się używać współczynnika korelacji Pearsona, który posiada geometryczną interpretację oraz jest odniesiony do parametrów rozkładu normalnego.

Jeżeli założymy, że badane zmienne (cechy) mają dwuwymiarowy rozkład normalny, to przedział ufności dla oszacowanego współczynnika korelacji Pearsona można wyznaczyć na podstawie statystyki Fishera następującej postaci

$$Z = \frac{1}{2} \ln \frac{1+r}{1-r} \tag{1.18}$$

która ma w przybliżeniu rozkład normalny o wartości przeciętnej

$$E[Z] \cong \frac{1}{2} \ln \frac{1+r}{1-r} + \frac{r}{2(n-1)} \tag{1.19}$$

oraz wariancji

$$\sigma^2[Z] \cong \frac{1}{n-3} \tag{1.20}$$

Dla $n \geq 10$, w praktycznych rozważaniach, drugi składnik wartości oczekiwanej można pominąć, stąd po dokonaniu standaryzacji rozważanej statystyki otrzymamy nową statystykę postaci

$$U = \left(Z - \frac{1}{2} \ln \frac{1+r}{1-r} \right) \sqrt{n-3} \tag{1.21}$$

która ma rozkład $N(0, 1)$.

Przedział ufności dla wartości przeciętnej

$$E[Z] = \frac{1}{2} \ln \frac{1+r}{1-r} \tag{1.22}$$

oraz dla poziomu istotności $(1 - \alpha)$ określa się według nierówności

$$\hat{z} - \frac{1}{\sqrt{1-3}} \cdot u\left(1 - \frac{\alpha}{2}\right) < \frac{1}{2} \ln \frac{1+r}{1-r} < \hat{z} + \frac{1}{\sqrt{1-3}} \cdot u\left(1 - \frac{\alpha}{2}\right) \tag{1.23}$$

gdzie $\hat{z} = \frac{1}{2} \ln \frac{1+\hat{r}}{1-\hat{r}}$ jest wyliczane z próby.

Jeżeli nierówność (1.23) zapiszemy przy następujących oznaczeniach

$$\hat{z}_1 < \frac{1}{2} \ln \frac{1+r}{1-r} < \hat{z}_2 \tag{1.24}$$

to realizację przedziału ufności dla r, w postaci $r_1 < r < r_2$, wyznacza się według następujących zależności

$$r_1 = \frac{e^{2\hat{z}_1} - 1}{e^{2\hat{z}_1} + 1}, \qquad r_2 = \frac{e^{2\hat{z}_2} - 1}{e^{2\hat{z}_2} + 1} \tag{1.25}$$

lub z odpowiednich tablic.

Na podstawie analizy wariancji można rozszerzyć interpretację współczynnika korelacji Pearsona na zakres wyjaśniania zmiennej zależnej przez zmienną niezależną.

Ustalenie modelu regresji polega na wybraniu takiej funkcji

$$y = g(x) \tag{1.26}$$

która będzie reprezentowała (zastępowała) zbiór wartości obu zmiennych. Estymacja parametrów tej funkcji będzie wykonywana metodą najmniejszych kwadratów (MNK), czyli według zasady

$$\Psi(x_i) = \sum_i [y_i - g(x_i)]^2 = \min \tag{1.27}$$

Dla funkcji postaci liniowej

$$y = A + Bx \tag{1.28}$$

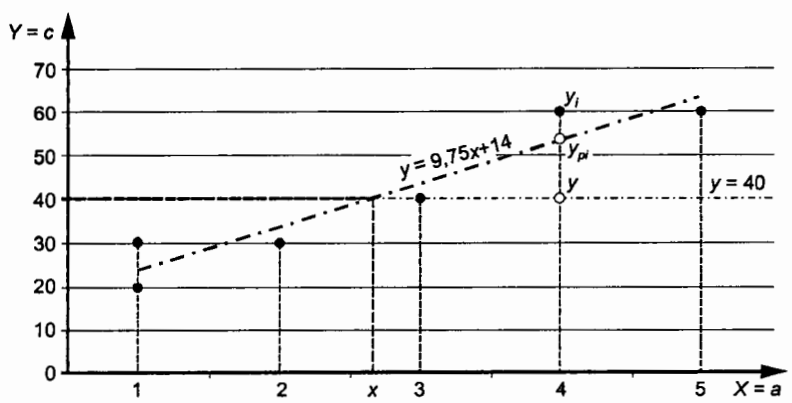
warunek (1.27) prowadzi do modelu regresji liniowej, który można wyrazić następującą zależnością

$$\frac{y-\bar{y}}{\sigma(y)} = r\frac{x-\bar{x}}{\sigma(x)} \tag{1.29}$$

przy czym wielkości występujące w tej zależności stanowią wartości przeciętne i odchylenia zmiennych losowych określonych na podstawie wyników z próby. Wielkości te są zdefiniowane wzorami od (1.3) do (1.5), a to oznacza, że w tym wzorze może występować tylko współczynnik korelacji liniowej (Pearsona) zmiennej X względem zmiennej Y.

Na podstawie modelu regresji (1.29), przy ustalonej wartości jednej zmiennej, można prognozować (wyliczać) wartość drugiej zmiennej. W przypadku prognozy wartości nieruchomości ustalana będzie wartość atrybutu ($x_p = a_p$), zaś wyliczana będzie prognozowana cena jednostkowa nieruchomości, czyli ($y_p = c_p$), przy czym $y_{pi} = g(x_{pi})$.

Stosując analizę wariancji można wykazać, że współczynnik r ma ścisły związek z zakresem wyjaśniania zmiennej zależnej przez zmienną niezależną. Aby to uzasadnić, rozpatrzmy zmienną losową dwuwymiarową reprezentowaną przez próbę zawierającą 5 par wartości (x_i, y_i), dla której diagram korelacyjny i linię regresji wraz z jej parametrami przedstawiono na rysunku 1.



x	1	1	2	3	4	5
y	20	30	30	40	60	60

$x = 2{,}66(6)$; $\sigma(x) = 1{,}491$; $A = 14{,}00$; $r = 0{,}9515$
$y = 40{,}00$; $\sigma(y) = 15{,}275$; $B = 9{,}75$;

Rys. 1. Wartości zmiennej losowej w próbie, diagram korelacyjny, linia regresji

Miarą całkowitego rozproszenia wartości zmiennej losowej (y_i) względem jej wartości przeciętnej ($\bar{y}$) jest suma kwadratów ich różnic, zwana całkowitą sumą kwadratów (CSK), czyli

$$\mathrm{CSK} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{1.30}$$

Jeżeli zaobserwowane wartości zmiennej losowej będą aproksymowane linią regresji, to różnice ($y_i - \bar{y}$) można zastąpić dwoma składnikami (por. rys. 1)

$$y_i - \bar{y} = (y_i - y_{pi}) + (y_{pi} - \bar{y}) \tag{1.31}$$

Pierwszy składnik dotyczy części wyjaśnianej zmienności (y_i) przez linię regresji, zaś drugi składnik dotyczy części niewyjaśnianej tej zmiennej.

W aspekcie formuły (1.30) wynika, że suma kwadratów różnic ($y_i - \bar{y}$) stanowi wyjaśnioną część z całej sumy kwadratów, czyli

$$\mathrm{WSK} = \sum_{i=1}^{n} [(y_i - g(x_i)]^2 = \sum_{i=1}^{n} (y_i - y_{pi})^2 \tag{1.32}$$

Niewyjaśniona część sumy kwadratów zmiennej (y_i) stanowi dopełnienie do WSK względem całkowitej sumy kwadratów, czyli

$$\mathrm{NSK} = \mathrm{CSK} - \mathrm{WSK} \tag{1.33}$$

Celem przedstawianej analizy wariancji jest wykazanie, że stosunek wyjaśnianej sumy kwadratów do całkowitej sumy kwadratów jest równy kwadratowi współczynnika korelacji liniowej (Pearsona) określonego z próby, czyli

$$\frac{\mathrm{WSK}}{\mathrm{CSK}} = \frac{(y_{pi} - \bar{y})^2}{(y_i - \bar{y})^2} = \hat{r}^2 \tag{1.34}$$

Na podstawie linii regresji postaci (1.29) dla wartości x_i można obliczyć prognozowaną wartość zmiennej ($y_{pi} - \bar{y}$)

$$(y_{pi} - \bar{y}) = \hat{r} \frac{\sigma[y]}{\sigma[x]} (x_i - \bar{x}) \tag{1.35}$$

Po uwzględnieniu (1.35) i definicji na odchylenie standardowe zależność (1.34) można zapisać w następującej postaci

$$\frac{\mathrm{WSK}}{\mathrm{CSK}} = \frac{\sum_{i=1}^{n} \left[\hat{r} \frac{\sigma[y]}{\sigma[x]} (x_i - \bar{x}) \right]^2}{n \cdot \sigma^2[y]} = \frac{\hat{r}^2 \frac{\sigma^2[y]}{\sigma^2[x]} \cdot n \cdot \sigma^2[x]}{n \cdot \sigma^2[y]} = \hat{r}^2 \qquad \text{c.b.d.w.} \tag{1.36}$$

Zależność (1.36) dowodzi, że $\hat{r}^2$ określa miarę dopasowania linii prostej regresji postaci (1.29) do zbioru punktów reprezentujących dwuwymiarową zmienną losową (np. atrybutu i ceny poszczególnych nieruchomości). Wielkość $\hat{r}^2$ może stanowić miarę wyjaśniania zmiennej y przez model liniowej regresji zmiennej x, stąd jej wartość można przyjmować za wagę dokładności prognozy y_{pi} opartej na formule (1.29).

2. Zmienna losowa wielowymiarowa

Dla wielu zmiennych niezależnych $X_1, X_2, ..., X_k$ model liniowy regresji wielorakiej definiuje się za pomocą hiperpłaszczyzny w przestrzeni $(k+1)$ wymiarowej, czyli

$$Y = a_0 + a_1X_1 + a_2X_2 + ... + a_kX_k = G(X) \tag{2.1}$$

przy czym Y stanowi zmienną losową zależną (prognozowaną), wielkości $X_1, X_2, ..., X_k$ stanowią zmienne losowe niezależne (objaśniające), zaś $a_0, a_1, ..., a_k$ nazywamy współczynnikami regresji zmiennej Y, względem zmiennych X_i. W przypadku rozpatrywania tylko dwóch zmiennych niezależnych (X_i, X_j) równanie to będzie przedstawiało płaszczyznę regresji.

Wszystkie współczynniki regresji i korelacji mogą być zdefiniowane na podstawie macierzy korelacyjnej $\{K\}$, której elementy stanowią współczynniki korelacji zupełnej (Pearsona) r_{ij} pomiędzy poszczególnymi zmiennymi Y, X_i, X_j, przy czym wskaźnik „0" będzie odpowiadał zmiennej zależnej Y. Uwzględniając powyższe oznaczenia, macierz korelacyjna przyjmuje następującą postać

$$\{K\} = \begin{Bmatrix} 1 & r_{01} & r_{02} & \cdots & r_{0k} \\ r_{10} & 1 & r_{12} & \cdots & r_{1k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ r_{k0} & r_{k1} & r_{k2} & \cdots & 1 \end{Bmatrix} \tag{2.2}$$

Wartości współczynników korelacji zupełnej r_{ij} dla poszczególnych par zmiennych niezależnych (np. atrybutów), określają natężenie współzależności poszczególnych zmiennych (atrybutów). Pary atrybutów, które wykazują współczynnik korelacji wysoki ($|r_{ij}| > 0{,}6$) mają podobny wpływ na kształtowanie się cen nieruchomości, a tym samym oba atrybuty wyjaśniają analogiczną część zmienności zmiennej Y (cen). W takich parach powinno się zrezygnować w modelu regresji (2.1) z jednego atrybutu. Eliminacja zmiennych (atrybutów), które w parach wykazują wysoki współczynnik korelacji, ma na celu poprawienie wiarygodności modelu regresji, a także zapewnienia stabilności estymacji współczynników tego modelu.

Na podstawie macierzy korelacyjnej (2.2) można zdefiniować **współczynniki korelacji cząstkowej (częściowej)**, czyli współzależność dwóch ustalonych zmien-

nych X_i, X_j, ale przy uwzględnieniu wpływu pozostałych zmiennych losowych. Współczynniki korelacji cząstkowej najwygodniej określać za pomocą algebraicznych dopełnień elementów macierzy $\{K\}$, czyli

$$r_{ij\cdot 1,2,\ldots,k} = \frac{-Ad[K_{ij}]}{\sqrt{Ad[K_{ii}]\cdot Ad[K_{jj}]}} \tag{2.3}$$

gdzie:

$Ad[K_{ij}]$ – oznacza algebraiczne dopełnienie elementu k_{ij} macierzy $\{K\}$,

$Ad[K_{ii}]$ – oznacza algebraiczne dopełnienie elementu k_{ii} macierzy $\{K\}$,

$Ad[K_{jj}]$ – oznacza algebraiczne dopełnienie elementu k_{jj} macierzy $\{K\}$.

Rzędem współczynnika korelacji cząstkowej nazywamy liczbę wskaźników występujących po kropce w oznaczeniu $r_{ij\cdot 1,2,\ldots,k}$ postaci (2.3). Dla zmiennej losowej określonej w przestrzeni $(k+1)$ najwyższy rząd korelacji cząstkowej wynosi $(k+1)-2$, czyli dla dwuwymiarowej zmiennej losowej rząd korelacji cząstkowej jest równy zeru, stąd ich nazwa korelacji zupełnej (całkowitej).

Jeżeli rozpatrzymy trzy zmienne losowe (X_i, X_j, X_k), to badanie siły związku między zmiennymi (X_i, X_j) przy uwzględnieniu wpływu trzeciej zmiennej (korygującej) (X_k) odbywa się przy rozważaniu **współczynnika korelacji cząstkowej**, czyli

$$r_{ij\cdot k} = \frac{r_{ij} - (r_{ik})(r_{jk})}{\sqrt{1-r_{ik}^2}\sqrt{1-r_{jk}^2}} \tag{2.4}$$

Współczynnik korelacji cząstkowej stanowi miarę współzależności odchyłek losowych jakie otrzymuje się z predykcji zmiennych X_i, X_k na podstawie wartości zmiennych X_k, czyli odchyłek liczonych na podstawie równania postaci (1.29)

$$\delta_{is} = x_{is} - \hat{x}_i - r_{ik}\frac{\sigma_i}{\sigma_k}(x_{ks} - \hat{x}_k) \tag{2.5}$$

$$\delta_{js} = x_{js} - \hat{x}_j - r_{jk}\frac{\sigma_j}{\sigma_k}(x_{ks} - \hat{x}_k) \tag{2.6}$$

Wykorzystując formułę (1.29) można zapisać wzór na współczynnik korelacji cząstkowej za pomocą odchyłek (2.5) i (2.6)

$$\hat{r}[\delta_i,\delta_j] = \frac{\frac{1}{n}\sum_{s=1}^{n}(\delta_{is}\cdot\delta_{js}) - \bar{\delta}_i\bar{\delta}_j}{\sqrt{\frac{1}{n}\sum_{s=1}^{n}\delta_{is}^2 - \bar{\delta}_i^2}\cdot\sqrt{\frac{1}{n}\sum_{s=1}^{n}\delta_{js}^2 - \bar{\delta}_j^2}} \tag{2.7}$$

Porównując wartości współczynników korelacji zupełnej z odpowiadającymi współczynnikami korelacji cząstkowej można wnioskować, że zmienne występujące po kropce

w korelacji cząstkowej mają istotny wpływ na współzależność dwóch rozpatrywanych zmiennych w korelacji zupełnej.

Stopień dopasowania hiperpłaszczyzny, wyrażonej formułą (2.1) do zbioru punktów reprezentujących ceny i atrybuty poszczególnych nieruchomości będzie określony przez **współczynnik liniowej korelacji wielorakiej *R*.**

Kwadrat współczynnika liniowej korelacji wielorakiej można również obliczyć w oparciu o macierz korelacyjną (1.2)

$$R^2 = 1 - \frac{\det\{K\}}{\det\{K_0\}} \tag{2.8}$$

gdzie:

$\det\{K\}$ – oznacza wyznacznik macierzy $\{K\}$,

$\det\{K_0\}$ – oznacza wyznacznik podmacierzy $\{K_0\}$, która powstaje ze skreślenia pierwszego wiersza i pierwszej kolumny macierzy $\{K\}$, czyli elementów macierzy dotyczących współczynników korelacji dla zmiennych niezależnych.

Równanie hiperpłaszczyzny regresji, określane na podstawie wyników próby losowej, można również zapisać w następującej postaci

$$y - \bar{y} = a_1(x_1 - \bar{x}_1) + a_2(x_2 - \bar{x}_2) + \cdots + a_k(x_k - \bar{x}_k) \tag{2.9}$$

gdzie:

$\bar{y}, \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$ – wartości przeciętne zmiennej zależnej i zmiennych niezależnych liczone z próby,

a_i – współczynniki regresji zmiennej Y względem zmiennej X_i.

Współczynniki regresji w modelu (2.9) można wyznaczyć na podstawie elementów macierzy korelacyjnej (2.1), czyli

$$a_i = \frac{Ad[K_{0i}]}{Ad[K_{00}]} \frac{\sigma[x_i]}{\sigma[y]} \tag{2.10}$$

gdzie:

$Ad[K_{0i}], Ad[K_{00}]$ – algebraiczne dopełnienia odpowiednich elementów macierzy $\{K\}$,

$\sigma[y], \sigma[x_i]$ – odchylenia standardowe dla zmiennej zależnej i zmiennych niezależnych liczone z próby.

Jeżeli współczynniki regresji poddamy standaryzacji, czyli podzielimy przez odchylenie standardowe zmiennej zależnej i pomnożymy przez odchylenie standardowe zmiennej niezależnej, to parametry te stanowią nowe wielkości, które często występują

pod nazwą wag beta (β). Standaryzowane współczynniki regresji (wagi β) oblicza się według następującego wzoru

$$\beta_i = \frac{Ad[K_{0i}]}{Ad[K_{00}]} \tag{2.10a}$$

Dla trzech zmiennych o wskaźnikach i, j, k wagę beta można wyrazić za pomocą współczynników korelacji zupełnej (Pearsona) według następującego wzoru

$$\beta_{ij \cdot k} = \frac{r_{ij} - r_{ik} r_{jk}}{1 - r_{jk}^2} \tag{2.11}$$

Współczynnik $\beta_{ij,k}$ oznacza, że prognozujemy zmienną i na podstawie zmiennych j, k, oraz że jest to współczynnik stojący przy j zmiennej.

Wartość współczynnika korelacji wielorakiej R można również określić na podstawie analizy wariancji, przy funkcji regresji (2.1)

$$R^2 = 1 - \frac{\sum_{i=1}^{n} [y_i - G(x_i)]^2}{\sum_{i=1}^{n} [y_i - \bar{y}]^2} = 1 - \frac{\sum_{i=1}^{n} \delta_i^2}{\sum_{i=1}^{n} \Delta_i^2} \tag{2.12}$$

gdzie:

$\sum_{i=1}^{n} [y_i - G(x_i)]^2$ – suma kwadratów odchyłek pomiędzy wartościami zmiennej zależnej z próby a jej wartościami modelowymi – czyli suma ta opisuje część zmienności y_i, która nie została wyjaśniona modelem regresji postaci (2.1),

$\sum_{i=1}^{n} [y_i - \bar{y}]^2$ – wyraża całkowite rozproszenie zmiennej y_i od jej wartości przeciętnej $\bar{y}$.

Jeżeli funkcja zmiennych losowych $X_1, X_2, ..., X_n$ ma postać liniową

$$F(X_1, X_2, ..., X_n) = b_1 X_1 + b_2 X_2 + ... + b_n X_n \tag{2.13}$$

w której zmienne są względem siebie zależne, czyli istnieje macierz korelacyjna $\{K_0\}$, to wariancja tej funkcji wyraża się następującą zależnością

$$\sigma^2[F] = \begin{Bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{Bmatrix}^T \cdot \begin{bmatrix} \sigma[X_1] & 0 & 0 & 0 \\ 0 & \sigma[X_2] & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \sigma[X_n] \end{bmatrix} \cdot \{K_0\} \cdot \begin{bmatrix} \sigma[X_1] & 0 & 0 & 0 \\ 0 & \sigma[X_2] & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \sigma[X_n] \end{bmatrix} \cdot \begin{Bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{Bmatrix} \tag{2.14}$$

Każdą funkcję wielkości losowych $X_1, X_2, ..., X_n$ postaci nieliniowej

$$G = g(X_1, X_2, ..., X_n) \tag{2.15}$$

można sprowadzić do postaci liniowej za pomocą rozwinięcia w szereg Taylora. Uwzględniając pierwsze pochodne tego rozwinięcia otrzymuje się następującą zależność

$$G = G_0 + dG = G_0 + \frac{\partial g}{\partial X_1} dX_1 + \frac{\partial g}{\partial X_2} dX_2 + ... + \frac{\partial g}{\partial X_n} dX_n \tag{2.16}$$

gdzie:

G_0 – stała wartość funkcji G, określona za pomocą przybliżonych wartości zmiennych losowych $X_{10}, X_{20}, ..., X_{n0}$;

$\frac{\partial g}{\partial X_i}$ – pochodna cząstkowa funkcji G względem zmiennej losowej X_i, której wartość jest liczona dla przybliżonych wartości zmiennych losowych $X_{10}, X_{20}, ..., X_{n0}$;

dX_i – przyrost zmiennej X_i.

Jeżeli zmienne losowe definiujące funkcję (2.13) będą względem siebie zależne, czyli będą posiadały macierz korelacyjną $\{K_0\}$, to wariancję tej funkcji można wyrazić iloczynem macierzowym następującej postaci

$$\sigma^2[G] = \begin{Bmatrix} \frac{\partial g}{\partial X_1} \\ \frac{\partial g}{\partial X_2} \\ \vdots \\ \frac{\partial g}{\partial X_n} \end{Bmatrix}^T \cdot \begin{Bmatrix} \sigma[X_1] & 0 & 0 \cdots & 0 \\ 0 & \sigma[X_2] & 0 \cdots & 0 \\ \vdots & \vdots & \vdots \cdots & \vdots \\ 0 & 0 & 0 \cdots & \sigma[X_n] \end{Bmatrix} \cdot \{K_0\} \cdot \begin{Bmatrix} \sigma[X_1] & 0 & 0 \cdots & 0 \\ 0 & \sigma[X_2] & 0 \cdots & 0 \\ \vdots & \vdots & \vdots \cdots & \vdots \\ 0 & 0 & 0 \cdots & \sigma[X_n] \end{Bmatrix} \cdot \begin{Bmatrix} \frac{\partial g}{\partial X_1} \\ \frac{\partial g}{\partial X_2} \\ \vdots \\ \frac{\partial g}{\partial X_n} \end{Bmatrix} \tag{2.17}$$

Wzór (2.17) ujmuje zasadę sumowania (narastania) wariancji dla nieliniowych funkcji zmiennych losowych zależnych. Jeżeli macierz $\{K_0\}$ będzie macierzą jednostkową to oznacza, że rozpatrywane zmienne losowe są względem siebie niezależne.

W zależności (2.17) iloczyn trzech macierzy środkowych określa macierz współczynników kowariancji zmiennych losowych $(X_1, X_2, ..., X_n)$, czyli

$$\operatorname{cov}\{X_1, X_2, ..., X_n\} = \begin{Bmatrix} \sigma[X_1] & 0 & 0 \cdots & 0 \\ 0 & \sigma[X_2] & 0 \cdots & 0 \\ \vdots & \vdots & \vdots \cdots & \vdots \\ 0 & 0 & 0 \cdots & \sigma[X_n] \end{Bmatrix} \cdot \{K_0\} \cdot \begin{Bmatrix} \sigma[X_1] & 0 & 0 \cdots & 0 \\ 0 & \sigma[X_2] & 0 \cdots & 0 \\ \vdots & \vdots & \vdots \cdots & \vdots \\ 0 & 0 & 0 \cdots & \sigma[X_n] \end{Bmatrix} \tag{2.18}$$

Zatem wzór (2.17) można zapisać w następującej postaci

$$\sigma^2[G] = \begin{Bmatrix} \frac{\partial g}{\partial X_1} \\ \frac{\partial g}{\partial X_2} \\ \vdots \\ \frac{\partial g}{\partial X_n} \end{Bmatrix}^T \cdot \operatorname{cov}\{X_1, X_2, ..., X_n\} \cdot \begin{Bmatrix} \frac{\partial g}{\partial X_1} \\ \frac{\partial g}{\partial X_2} \\ \vdots \\ \frac{\partial g}{\partial X_n} \end{Bmatrix} \tag{2.19}$$

Zależność (2.19) stanowi uogólnioną formułę określania wariancji dla zmiennych losowych zależnych.

Współczynniki liniowego modelu wielorakiej regresji (2.1) można również wyznaczyć na podstawie rozwiązania układu równań następującej postaci macierzowej

$$\{X\} \cdot \{a\} = \{Y\} \tag{2.20}$$

gdzie:

$$\{X\} = \begin{Bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{Bmatrix} \tag{2.21}$$

oznacza macierz zawierającą jedynki i zmienne niezależne (atrybuty),

$$\{a\} = \begin{Bmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{Bmatrix} \tag{2.22}$$

oznacza wektor (macierz jednokolumnową) współczynników regresji wielorakiej,

$$\{Y\} = \begin{Bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{Bmatrix} \tag{2.23}$$

oznacza wektor (macierz jednokolumnową) zmiennej zależnej (cen nieruchomości).

Rozwiązanie układu równań (2.20) według metody najmniejszej sumy kwadratów prowadzi do następujących zależności

$$\{\hat{a}\}=\left(\{X\}^T\cdot\{X\}\right)^{-1}\cdot\{X\}^T\cdot\{Y\} \tag{2.24}$$

Wariancję resztową określa się według wzoru

$$\hat{\sigma}_0^2=\frac{\{Y\}^T\{Y\}-\{\hat{a}\}^T\cdot\{X\}^T\cdot\{Y\}}{n-k-1} \tag{2.25}$$

gdzie n oznacza liczbę porównywanych nieruchomości, zaś k oznacza liczbę rozpatrywanych zmiennych niezależnych (atrybutów).

Prognozę (predykcję) jednostkowej ceny rynkowej ($\hat{y}=\hat{w}$) wycenianej nieruchomości wyznacza się według następującego iloczynu macierzowego

$$\hat{w}=\{\hat{x}_1\ \hat{x}_2\ \ldots\ \hat{x}_k\}^T\cdot\{\hat{a}\} \tag{2.26}$$

przy czym $\hat{x}_i$ oznacza wartość i-tego atrybutu wycenianej nieruchomości.

Aby taka prognoza była wiarygodna i możliwa do wykorzystania w szacowaniu nieruchomości, należy przestrzegać następującej zasady: wartości atrybutów ($\hat{x}_i$) wycenianej nieruchomości muszą zawierać się w przedziałach zmienności odpowiadających atrybutów nieruchomości wybranych do porównania. W każdym innym przypadku prognozowana wartość nieruchomości będzie miała małą wiarygodność, gdyż będzie ona określana na podstawie ekstrapolacji modelu regresji.

Ocenę dokładności prognozy wartości rynkowej określa się na podstawie wektora atrybutów występujących w zależności (2.26), macierzy kowariancji $(\{X\}^T\{X\})^{-1}$ występującej w zależności (2.24) oraz wariancji resztowej $\hat{\sigma}_0^2$ określonej wzorem (2.26)

$$\sigma^2[\hat{w}]=\hat{\sigma}_0^2\begin{Bmatrix}\hat{x}_1\\ \hat{x}_2\\ \vdots\\ \hat{x}_k\end{Bmatrix}^T\left(\{X\}^T\{X\}\right)^{-1}\begin{Bmatrix}\hat{x}_1\\ \hat{x}_2\\ \vdots\\ \hat{x}_k\end{Bmatrix}=\begin{Bmatrix}\hat{x}_1\\ \hat{x}_2\\ \vdots\\ \hat{x}_k\end{Bmatrix}^T\operatorname{cov}\{\hat{a}\}\begin{Bmatrix}\hat{x}_1\\ \hat{x}_2\\ \vdots\\ \hat{x}_k\end{Bmatrix} \tag{2.27}$$

Formuła (2.27) odpowiada formule (2.17), która jest zapisana za pomocą macierzy korelacyjnej.

3. Uwagi końcowe

W ostatnim okresie ukazuje się wiele publikacji dotyczących analizy statystycznej rynku nieruchomości mających na celu wnioskowanie, które atrybuty (cechy) nieruchomości wyjaśniają istotną część zmienności cen rynkowych tych nieruchomości.

Przedstawiona analiza różnych współczynników korelacji ma na celu wskazanie, w jakich obszarach badań można stosować odpowiedni rodzaj współczynnika korelacji oraz jaka jest ich interpretacja analityczna i geometryczna. Z analizy tej można wnioskować, że dla zmiennych losowych kwantyfikowanych powinien być zawsze stosowany współczynnik korelacji Pearsona, gdyż jego wartość jest związana z parametrami rozkładu zmiennej losowej oraz posiada ścisłą interpretację geometryczną. Wartość tego współczynnika ustala poziom istotności dla wielkości prognozowanych z modelu liniowej regresji. Współczynniki korelacji Pearsona definiują również elementy macierzy korelacyjnej dla zmiennej losowej wielowymiarowej, która jest podstawą do wszelkich analiz statystycznych tych zmiennych. Na podstawie elementów macierzy korelacyjnej można obliczyć współczynnik korelacji wielorakiej, współczynniki korelacji cząstkowej oraz współczynniki regresji. Macierz korelacyjna może być również wykorzystana do analizy dokładności nowych statystyk (parametrów) opartych na rozpatrywanych zmiennych losowych, a to prowadzi do określenia macierzy kowariancji dla tych parametrów.

W następnej publikacji z tego zakresu zostanie przedstawiona, na przykładzie liczbowym, analiza ilościowa rozważanych współczynników korelacji oraz praktyczne wnioski, które będzie można wykorzystać do analizy rynku nieruchomości.

Literatura

[1] Cegielski P.: *Przykład alternatywnej techniki analizy statystycznej rynku – część 2.* Warszawa, Rzeczoznawca Majątkowy, nr 2, 1999

[2] Czaja J.: *Estymacja uogólnionych modeli liniowych.* Warszawa, PAN Geodezja i Kartografia, t. XLIII, z. 3, 1994

[3] Czaja J. i in.: *Wycena nieruchomości majątkowych metodą cenowo-porównawczą.* Warszawa, Przegląd Geodezyjny, 12, 1996

[4] Czaja J.: *Zamiast techniki porównywania parami – technika interpolacyjna – wyceny nieruchomości.* Katowice, Kwartalnik Nieruchomość, nr 3, 1996

[5] Czaja J.: *Modele statystyczne w informacji o terenie.* Kraków, Wydawnictwa AGH 1997

[6] Czaja J.: *Interval Estimation of Generalized Linear Models.* Warszawa, PAN, Geodezja i Kartografia, t. XLVI, z. 1, 1997

[7] Czaja J.: *Praktyczna weryfikacja różnych technik wyceny w podejściu porównawczym.* Kraków, Rzeczoznawca Małopolski, nr 2, 1997

[8] Czaja J.: *Metody i systemy szacowania nieruchomości.* Kraków, podręcznik dla rzeczoznawców majątkowych, 1999

[9] Czaja J.: *Podejście porównawcze wyceny nieruchomości w aspekcie standardów*. Warszawa, Rzeczoznawca Majątkowy, nr 4, 1999
[10] Czaja J.: *Metody i systemy określania wartości nieruchomości*. Kraków, UWND AGH 1999
[11] Mazurkiewicz E.: *Uwagi do procedur określania wartości rynkowej nieruchomości przy zastosowaniu analizy statystycznej rynku*. Warszawa, Rzeczoznawca Majątkowy, nr 2, 1999
[12] Prystupa M.: *Wycena nieruchomości metodą cenowo-porównawczą*. Warszawa, Biblioteczka rzeczoznawcy majątkowego 1997
[13] Prystupa M.: *Jaki standard dla podejścia porównawczego?* Warszawa, Rzeczoznawca Majątkowy, nr 4, 1999
[14] Rao C.R.: *Modele liniowe statystyki matematycznej*. Warszawa, PWN 1982

Recenzent
prof. dr hab. inż. Stanisław Latoś