

Tadeusz Drwięga, Zofia Graboś (Lublin)

ZASTOSOWANIE PROCEDURY JEDNOCZESNEGO TESTOWANIA HIPOTEZ
SANOVA W ANALIZIE REGRESJI WIELOKROTNEJ

1. Wstęp i streszczenie

W wielokrotnej analizie regresji testuje się hipotezy o nieistotności wektora współczynników regresji lub jego podwektora. W przypadku odrzucenia takiej hipotezy interesująca jest odpowiedź na pytanie, czy wszystkie składowe tego wektora należy uznać za niezerowe, czy tylko niektóre. Jedną z metod pozwalającą stwierdzić, które ze składowych podwektora wektora współczynników regresji należy uznać za istotne wraz z nim jest procedura jednoczesnego testowania hipotez SANOVA.

Ideę tej procedury, opartej na wielowymiarowym rozkładzie F , testującej jednocześnie układ hipotez liniowych przedstawił Ghosh (1955), a rozwinęli ją Ramachandran (1956) i Krishnaiah (1965), który udowodnił, że przedziały ufności związane z procedurą SANOVA są nie dłuższe niż odpowiednie przedziały ufności Scheffe'go. Zastosowanie tej procedury zostało omówione w pracach cytowanych powyżej, a także w publikacjach: Krishnaiah, Armitage (1970), Schuurmann, Krishnaiah, Chattopadhyay (1975), Drwięga, Graboś (1978). Zaletą procedury SANOVA jest to, że nie wymaga ona niezależności sum kwadratów dla poszczególnych hipotez liniowych. Pozwala to na stosowanie jej w wielokrotnej analizie regre-

sji dla kompozycyjnych układów czynnikowych. Możliwość takiego wykorzystania procedury SANOVA pokazano w niniejszej pracy. W paragrafie 2 podano definicję wielowymiarowego rozkładu F. Procedurę SANOVA przedstawiono w paragrafie 3. W paragrafie 4 wyprowadzono wzory umożliwiające stosowanie procedury SANOVA w analizie regresji. Zastosowanie tej procedury na przykładzie liczbowym przedstawiono w paragrafie 5.

2. Wielowymiarowy rozkład F

Podaną niżej definicję wielowymiarowego rozkładu F można znaleźć w pracy Krishnaiah i Armitage (1970).

Niech \underline{X} będzie $n \times k$ macierzą n niezależnych losowych wektorów wierszowych o tym samym k -wymiarowym rozkładzie normalnym $N(\underline{\mu}, \underline{\Sigma})$, z wektorem wartości oczekiwanych $\underline{\mu}$ i macierzą kowariancji $\underline{\Sigma} = (\sigma_{ij})$. Niech $\underline{S} = \underline{X}'\underline{X}$, a $\frac{s}{2}$, gdzie $E(s^2) = m\sigma^2$, niech będzie zmienną losową o rozkładzie χ^2 z m stopniami swobody, stochastycznie niezależną od elementów na przekątnej macierzy \underline{S} , tzn. od $s_{11}, s_{22}, \dots, s_{kk}$. Wspólny rozkład zmiennych F_1, F_2, \dots, F_k , gdzie $F_i = \frac{s_{ii}}{n\sigma_{ii}} : \frac{s}{2}$ nazywamy k -wymiarowym rozkładem F z (n, m) stopniami swobody i z $\underline{\Sigma}$ jako macierzą kowariancji "stowarzyszonego" k -wymiarowego rozkładu normalnego. Rozkład ten nazywamy centralnym lub niecentralnym rozkładem w zależności od tego, czy $\underline{\mu}$ jest wektorem zerowym czy niezerowym.

Tablice wielowymiarowego rozkładu F można znaleźć w publikacji Schuurmann, Krishnaiah i Chattopadhyay (1975).

Jeśli $n=1$ to wielowymiarowy rozkład F redukuje się do wielowymiarowego rozkładu t^2 . Tablice tego rozkładu znajdują się w pracy Krishnaiah i Armitage (1970).

3. Procedura testów jednoczesnych SANOVA

Rozważmy liniowy model stały

$$(3.1) \quad \underline{y} = \underline{A} \underline{\beta} + \underline{e}$$

gdzie \underline{y} jest wektorem N obserwacji, $\underline{\beta}$ jest wektorem p nieznanych parametrów, \underline{A} - znaną macierzą o wymiarach $N \times p$, a \underline{e} - wektorem błędów losowych o N -wymiarowym rozkładzie normalnym z wartością oczekiwaną $\underline{0}$ i macierzą kowariancji $\sigma^2 \underline{I}_N$ ($\underline{0}$ -wektor kolumnowy złożony z N zer, \underline{I}_N - macierz jednostkowa $N \times N$). Zakładamy, że σ^2 jest nieznanne.

Zajmiemy się weryfikacją testowalnej hipotezy liniowej

$$(3.2) \quad H : \underline{K}' \underline{\beta} = \underline{0}$$

gdzie \underline{K} jest macierzą o wymiarach $p \times q$, rzędu r .

Hipotezę (3.2) możemy przedstawić jako koniunkcję hipotez

$$(3.3) \quad H_i : \underline{v}_i' \underline{\beta} = 0 \quad (i = 1, 2, \dots, t; t \geq r)$$

gdzie wektory $\underline{v}_1, \underline{v}_2, \dots, \underline{v}_t$ są liniowymi kombinacjami kolumn macierzy \underline{K} , przy czym dokładnie r spośród nich jest liniowo niezależnych. W szczególności może być $t=q$ i $\underline{v}_1, \underline{v}_2, \dots, \underline{v}_q$

mogą być kolumnami macierzy \underline{K} . To, że hipoteza H jest koniunkcją hipotez H_i ($i = 1, 2, \dots, t$) będziemy zapisywali $H = \bigcap_{i=1}^t H_i$.

Każdej hipotezie H_i ($i = 1, 2, \dots, t$) można przyporządkować zmienną losową

$$(3.4) \quad F_i = \frac{SS_i}{V_e}$$

gdzie SS_i jest sumą kwadratów dla hipotezy H_i , a V_e - średnim kwadratem dla błędu. Jeśli hipoteza $H = \bigcap_{i=1}^t H_i$ jest prawdziwa, to łączny rozkład statystyk F_1, F_2, \dots, F_t jest t -wymiarowym rozkładem F z $(1, v_e)$ - stopniami swobody, gdzie v_e jest liczbą stopni swobody dla błędu. Hipotezę H_i ($i = 1, 2, \dots, t$) odrzucamy na poziomie istotności α , jeśli $F_i > F_{\alpha}^t$, przy czym wartość F_{α}^t jest tak dobrana, że

$$P[F_i \leq F_{\alpha}^t; i = 1, 2, \dots, t | \bigcap_{i=1}^t H_i] = 1 - \alpha$$

Hipotezę H odrzucamy, jeśli odrzucimy przynajmniej jedną hipotezę H_i (patrz Schuurmann, Krishnaiah i Chattopadhyay (1975)).

Z definicji wielowymiarowego rozkładu F podanej w poprzednim paragrafie wynika, że jest on scharakteryzowany przez liczby stopni swobody i macierz kowariancji "stowarzyszonego" wielowymiarowego rozkładu normalnego.

Suma kwadratów SS_i dla hipotezy H_i ($i = 1, 2, \dots, t$) postaci (3.3) jest równa

$$SS_i = \underline{Y}' \underline{A} \underline{G} \underline{v}_i (\underline{v}_i' \underline{G} \underline{v}_i)^{-1} \underline{v}_i' \underline{G} \underline{A}' \underline{Y} = [(\underline{v}_i' \underline{G} \underline{v}_i)^{-\frac{1}{2}} \underline{v}_i' \underline{G} \underline{A}' \underline{Y}]^2$$

(patrz Searle, 1971), gdzie \underline{G} oznacza uogólnioną macierz odwrotną do macierzy $\underline{A}'\underline{A}$, tzn. macierz spełniająca warunek $\underline{A}'\underline{A} \underline{G} \underline{A}'\underline{A} = \underline{A}'\underline{A}$. Macierz \underline{X} t-wymiarowego rozkładu normalnego będzie zatem wektorem wierszowym postaci

$$(3.5) \quad \underline{X} = [(\underline{v}'_1 \underline{G} \underline{v}_1)^{-\frac{1}{2}} \underline{v}'_1 \underline{G} \underline{A}' \underline{Y}, (\underline{v}'_2 \underline{G} \underline{v}_2)^{-\frac{1}{2}} \underline{v}'_2 \underline{G} \underline{A}' \underline{Y}, \dots, (\underline{v}'_t \underline{G} \underline{v}_t)^{-\frac{1}{2}} \underline{v}'_t \underline{G} \underline{A}' \underline{Y}]$$

Macierz kowariancji $\underline{\Sigma}$ tego wektora ma wyrazy równe

$$(3.6) \quad \sigma_{ij} = \sigma^2 \frac{\underline{v}'_i \underline{G} \underline{v}_j}{\sqrt{\underline{v}'_i \underline{G} \underline{v}_i} \sqrt{\underline{v}'_j \underline{G} \underline{v}_j}} \quad \text{dla } i, j = 1, 2, \dots, t$$

zaś elementy macierzy korelacji $\underline{\Omega} = (\rho_{ij})$ mają postać

$$(3.7) \quad \rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} = \frac{\underline{v}'_i \underline{G} \underline{v}_j}{\sqrt{\underline{v}'_i \underline{G} \underline{v}_i} \sqrt{\underline{v}'_j \underline{G} \underline{v}_j}}$$

Tablice wielowymiarowego rozkładu F skonstruowane są jedynie dla przypadku, gdy wszystkie $\rho_{ij} = \rho$ dla $i \neq j$. Jeżeli sumy kwadratów SS_i dla hipotez (3.3) są stochastycznie niezależne i suma ich jest równa sumie kwadratów dla hipotezy (3.2), to wówczas współczynniki korelacji (3.7) są równe zero dla $i \neq j$ (patrz Drwięga, 1976).

4. Zastosowanie procedury SANOVA w modelach regresji dla doświadczeń czynnikowych

Niech równanie (3.1) będzie równaniem regresji wielokrotnej. Wtedy macierz \underline{A} jest macierzą pełnego rzędu kolumnowego i każda hipoteza postaci (3.2) jest testowalna. W praktyce często weryfikuje się hipotezę, że jakiś podwektor $\underline{\beta}_1$ wektora parametrów $\underline{\beta}$ jest wektorem zerowym. Hipoteza (3.2) przybiera wówczas postać

$$(4.1) \quad H : \underline{\beta}_1 = \underline{0}$$

Hipoteza (4.1) jest koniunkcją hipotez

$$(4.2) \quad H_i : \beta_{1i} = 0 \quad i = 1, 2, \dots, t$$

gdzie t jest liczbą składowych wektora $\underline{\beta}_1$. Hipotezy postaci (4.2) mówią o nielstotności poszczególnych składowych wektora $\underline{\beta}_1$.

Niech $\sigma^2 \underline{D}$ oznacza macierz kowariancji estymatora najmniejszych kwadratów $\hat{\underline{\beta}}_1 = [\hat{\beta}_{11}, \hat{\beta}_{12}, \dots, \hat{\beta}_{1t}]'$ wektora $\underline{\beta}_1$. Jest ona oczywiście podmacierzą macierzy $\sigma^2 (\underline{A}'\underline{A})^{-1}$. Suma kwadratów SS_i dla hipotezy H_i jest równa

$$(4.3) \quad SS_i = \hat{\beta}_{1i} d_{ii}^{-1} \hat{\beta}_{1i} = (d_{ii}^{-\frac{1}{2}} \hat{\beta}_{1i})^2 \quad \text{dla } i=1, 2, \dots, t$$

(por. Oktaba, 1972), gdzie d_{ii} jest odpowiednim elementem macierzy \underline{D} .

Macierz \underline{X} t -wymiarowego "stowarzyszonego" rozkładu normalnego będzie zatem wektorem wierszowym postaci

$$\underline{X} = [d_{11}^{-\frac{1}{2}\hat{\beta}}, d_{22}^{-\frac{1}{2}\hat{\beta}}, \dots, d_{tt}^{-\frac{1}{2}\hat{\beta}}]$$

Współczynnik korelacji pomiędzy i-tą a j-tą składową wektora \underline{X} jest równy

$$(4.4) \quad \rho_{ij} = \frac{d_{ij}}{\sqrt{d_{ii}d_{jj}}}$$

Tablice wielowymiarowego rozkładu F skonstruowane są jedynie dla przypadku, gdy

$$(4.5) \quad \rho_{ij} = \rho \quad \text{dla } i \neq j$$

Warunek ten jest spełniony dla układu hipotez (4.2) np. wtedy, gdy macierz \underline{D} ma jednakowe elementy na przekątnej i jednakowe elementy poza przekątną. W modelu regresji drugiego stopnia dla doświadczeń kompozycyjnych macierz $(\underline{A}'\underline{A})^{-1}$ ma na ogół postać

$$(\underline{A}'\underline{A})^{-1} = \begin{bmatrix} \underline{s} & \underline{0} & \underline{0} & \underline{f} \\ \underline{0} & \underline{P} & \underline{0} & \underline{0} \\ \underline{0} & \underline{0} & \underline{Q} & \underline{0} \\ \underline{f} & \underline{0} & \underline{0} & \underline{R} \end{bmatrix}$$

gdzie $\underline{0}$ oznacza macierz zerową, s jest elementem odpowiadającym wyrazowi wolnemu wielomianu regresyjnego, \underline{P} - macierzą odpowiadającą liniowym współczynnikom regresji, \underline{Q} - macierzą odpowiadającą

wiadającą mieszanym współczynnikom regresji, \underline{R} - macierzą odpowiadającą kwadratowym współczynnikom regresji, a $\sigma^2_{\underline{f}}$ jest wektorem kowariancji między estymatorami współczynników kwadratowych, a estymatorem wyrazu wolnego. W wielu przypadkach macierze \underline{P} i \underline{Q} są diagonalne, a macierz \underline{R} ma jednakowe elementy przekątne i jednakowe elementy poza przekątną (por. np. Zieliński (1974), Mańczak (1976)). Wektor $\underline{\beta}_1$ w hipotezie (4.1) może zatem składać się z efektów liniowych i mieszanych lub efektów kwadratowych.

5. Przykład

Zastosowanie przedstawionej w poprzednich paragrafach metody zilustrujemy na przykładzie skonstruowanym w oparciu o pracę Michałowskiego (1973). Badano wpływ trzech składników nawożenia: azotu N, fosforu P, potasu K na plon żyta ozimego. Przypuszczano, że zależność plonu od badanych czynników można opisać wielomianem drugiego stopnia, tzn.

$$E_y = \beta_0 + \beta_1 N + \beta_2 P + \beta_3 K + \beta_{12} NP + \beta_{13} NK + \beta_{23} PK + \beta_{11} N^2 + \beta_{22} P^2 + \beta_{33} K^2$$

Plan doświadczenia oraz obserwacje przedstawiono w tabeli 1.

Tabela 1

N	P	K	Y_i
-1	-1	-1	15.8
-1	-1	1	16.9
-1	1	-1	17.0
-1	1	1	18.4
1	-1	-1	20.8
1	-1	1	21.7
1	1	-1	21.7
1	1	1	23.1
1	0	0	21.9
-1	0	0	17.1
0	1	0	20.1
0	-1	0	18.8
0	0	1	20.0
0	0	-1	18.8
0	0	0	19.5

przy czym poziomy: -1, 0, 1 odpowiadają dawkom:

dla azotu - 20,40 i 60 kg/ha, dla fosforu - 16,32 i 48 kg/ha

oraz dla potasu - 24,48 i 72 kg/ha. Macierz planu \underline{A} dla takiego

doświadczenia ma postać:

$$\underline{A} = \begin{bmatrix} 1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Przypuśćmy, że interesuje nas hipoteza o nieistotności efektów kwadratowych, tzn. hipoteza

$$(5.1) \quad H : [\beta_{11}, \beta_{22}, \beta_{33}]' = \underline{0}$$

W przypadku odrzucenia tej hipotezy nasuwa się pytanie, czy każdy z efektów kwadratowych jest różny od zera, czy tylko niektóre.

Odpowiedź na nie można uzyskać traktując hipotezę (5.1) jako koniunkcję hipotez

$$(5.2) \quad H_i : \beta_{ii} = 0 \quad (i = 1, 2, 3)$$

i stosując procedurę testów jednoczesnych SANOVA. W tym celu należy wyznaczyć metodą analizy wariancji sumy kwadratów (4.3) dla hipotez (5.2) oraz średni kwadrat dla błędów. Dla danych z tabeli 1

wielkości te są odpowiednio równe 0.00, 0.0030, 0.0120, 0.0014. Wartości statystyki F wynoszą zatem 0.00, 2.14, 8.57. Macierz \underline{D} , której elementy są potrzebne do wyznaczenia współczynników korelacji (4.3) i sprawdzenia warunku (4.4) jest podmacierzą macierzy

$$(\underline{A}'\underline{A})^{-1} = \begin{bmatrix} \frac{13}{45} & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{9} & -\frac{1}{9} & -\frac{1}{9} \\ 0 & \frac{1}{10} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{10} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{10} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{8} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{8} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{8} & 0 & 0 & 0 \\ -\frac{1}{9} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{7}{18} & -\frac{1}{9} & -\frac{1}{9} \\ -\frac{1}{9} & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{9} & \frac{7}{18} & -\frac{1}{9} \\ -\frac{1}{9} & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{9} & -\frac{1}{9} & \frac{7}{18} \end{bmatrix}$$

Zatem

$$\underline{D} = \begin{bmatrix} \frac{7}{18} & -\frac{1}{9} & -\frac{1}{9} \\ -\frac{1}{9} & \frac{7}{18} & -\frac{1}{9} \\ -\frac{1}{9} & -\frac{1}{9} & \frac{7}{18} \end{bmatrix}$$

Jak łatwo zauważyć $\rho_{ij} = -0.3$ ($i \neq j$; $i, j = 1, 2, 3$). Wartość tablicowa trójwymiarowego rozkładu F z (1,5) stopniami swobody

dla $\rho = 0.3$ przy $\alpha = 0.05$ jest równa 11.3 (por.Krishnaiah i Armitage, 1970). Porównując wartości statystyk F z wartością tablicową otrzymujemy jako wniosek brak podstaw do odrzucenia którejkolwiek z hipotez (5.2) a w konsekwencji do odrzucenia hipotezy (5.1).

Ponieważ w naszym przykładzie macierz kowariancji wektora estymatorów efektów liniowych i mieszanych jest macierzą diagonalną

$$\begin{bmatrix} \frac{1}{10} & 0 & 0 & \vdots & & & \\ 0 & \frac{1}{10} & 0 & \vdots & & 0 & \\ 0 & 0 & \frac{1}{10} & \vdots & & & \\ \dots & \dots & \dots & \vdots & \dots & \dots & \\ & & & \vdots & \frac{1}{8} & 0 & 0 \\ & 0 & & \vdots & 0 & \frac{1}{8} & 0 \\ & & & \vdots & 0 & 0 & \frac{1}{8} \\ & & & & & & \vdots \end{bmatrix}$$

więc procedurę SANOVA można stosować dla hipotezy o nieistotności dowolnego podwektora wektora efektów liniowych i mieszanych. Np. hipotezę

$$(5.3) \quad H : [\beta_{12}, \beta_{13}, \beta_{23}]' = 0$$

o nieistotności wektora efektów mieszanych można przedstawić jako koniunkcję hipotez

$$(5.4) \quad H_{ij} : \beta_{ij} = 0 \quad (i \neq j; i, j = 1, 2, 3)$$

o nieistotności poszczególnych efektów mieszanych. Ponieważ sumy kwadratów (4.3) dla hipotez (5.4) są dla danych z tabeli 1 odpowiednio równe 0.02, 0.005, 0.08 więc statystyka F przyjmie wartości 14.29, 3.57, 57.00. Wartość tablicowa trójwymiarowego rozkładu F z (1,5) stopniami swobody dla $\alpha = 0.05$ przy $\rho = 0$ jest równa 11.54. Porównując wartości statystyk F z wartością tablicową otrzymujemy jako wniosek, że istotne są współczynniki β_{12} i β_{23} , a współczynnik β_{13} można uznać jako nieistotny. Konsekwencją tych wniosków jest odrzucenie hipotezy (5.3).

Literatura cytowana

- Drwięga, T., 1976: Testowanie hipotez liniowo-estymowalnych, VI Colloquium Metodologiczne z Agro-Biometrii, PAN, Warszawa, 120-130.
- Drwięga, T., Graboś, Z., 1978: Zastosowanie wielowymiarowego rozkładu F do jednoczesnego testowania hipotez, VIII Colloquium Metodologiczne z Agro-Biometrii, PAN, Warszawa.
- Ghosh, M.N., 1955: Simultaneous tests of linear hypotheses, *Biometrika* 42, 441-449.
- Krishnaiah, P.R., 1965: On the simultaneous ANOVA and MANOVA tests, *The Annals of Mathematical Statistics*, 17, 35-53.
- Krishnaiah, P.R., Armitage, J.V., 1970: On a multivariate F distribution, *Essays in Probability and Statistics*, University of North Carolina Press, Chapel Hill, North Carolina.
- Mańczak, K., 1976: Technika planowania eksperymentu, Wydawnictwa Naukowo-Techniczne, Warszawa.

- Michałowski, K., 1973: Efektywność nawożenia mineralnego w świetle doświadczeń ekstremalnych, Państwowe Wydawnictwa Rolnicze i Leśne, Warszawa.
- Oktaba, W., 1972: Metody statystyki matematycznej w doświadczeniach; PWN, Warszawa.
- Ramachandran, K.V., 1956: On the simultaneous analysis of variance tests. *The Annals of Mathematical Statistics*, 27, 521-528.
- Schuurmann, R.J., Krishnaiah, P.R., Chattopadhyay, A.K., 1975: Tables for a multivariate F distribution, *Sankhya*, B 37, 308-331.
- Zieliński, R., 1974: Wybrane zagadnienia optymalizacji statystycznej, PWN, Warszawa.