Luku 6

Varianssianalyysi

6.1 Johdanto

Regressioanalyysissa tarkastellaan selitettävän muuttujan riippuvuutta yhdestä tai useammasta selittävästä muuttujasta. Varianssianalyysissa asetelma on samankaltainen kuin regressioanalyysissa, mutta varianssianalyysi eroaa tavallisesta regressioanalyysista ainakin kahdessa suhteessa:

1) Selittävät muuttujat ovat luokittelutason muuttujia (esimerkiksi sukupuoli, laji, tuotemerkki).
2) Vaikka selittäjät olisivatkin kvantitatiivisia, ei varianssianalyysissa oleteta mitään riippuvuuden luonteesta.

Varianssianalyysissa tutkitaan riippuvan muuttujan keskiarvoeroja selittävien muuttujien arvojen (tasojen) määrittämissä luokissa.

6.2 Yksisuuntainen varianssianalyysi

6.2.1 Varianssianalyysin malli

Oletetaan, että havainnot kuuluvat yhden selittävän muuttujan arvojen määrittämiin ryhmiin. Olkoon ryhmien lukumäärä a ja havaintojen lukumäärät ryhmissä $n_1, n_2, \ldots, n_a$.

Ryhmä	1	2	3	$\ldots$	a
Lukumäärät	n_1	n_2	n_3	$\ldots$	n_a
	y_{11}	y_{21}	y_{31}	$\cdots$	y_{a1}
	y_{12}	y_{22}	y_{32}	$\cdots$	y_{a2}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	y_{1n_1}	y_{2n_2}	y_{3n_3}	$\cdots$	y_{an_a}
Ryhmäkeskiarvot	$\bar{y}_{1.}$	$\bar{y}_{2.}$	$\bar{y}_{3.}$	$\cdots$	$\bar{y}_{a}$

Kaikkien havaintojen lukumäärä $N = n_1+n_2+\ldots+n_a$ ja havaintojen keskiarvo on $\bar{y}_{..}$. Jokaisen havainnon y_{ij} $(i = 1, 2, \ldots, a;\ j = 1, 2, \ldots, n_i)$ ajatellaan syntyvän systemaattisen ja satunnaisen osan summana siten, että

$$\begin{aligned} y_{ij} &= \mu_i + \varepsilon_{ij} \\ &= \mu + \alpha_i + \varepsilon_{ij}; \quad i = 1, 2, \ldots, a; \quad j = 1, 2, \ldots, n_i; \end{aligned} \tag{6.1}$$

missä

$$\begin{aligned} \mu_i &= i.\ \text{ryhmäkeskiarvo} \\ \mu &= \text{kaikkien havaintojen keskiarvo} \\ \alpha_i &= \mu_i - \mu. \end{aligned}$$

Komponentti ε_{ij} edustaa satunnaisvaihtelua. Oletetaan, että

$$E(\varepsilon_{ij}) = 0 \quad \text{ja} \quad \text{var}(\varepsilon_{ij}) = \sigma^2$$

kaikilla i:n ja j:n arvoilla ja

$$\text{cov}(\varepsilon_{ij}, \varepsilon_{ks}) = 0, \quad \text{kun } i \neq j \text{ tai } j \neq s.$$

Kaikilla satunnaistermeillä on siis sama varianssi ja ne ovat keskenään korreloimattomia. Lisäksi tilastollista testausta varten tehdään normaalisuusoletus:

$$\varepsilon_{ij} \sim N(0, \sigma^2).$$

6.2.2 Varianssianalyysin ja regressioanalyysin yhteys

Huomattakoon, että varianssianalyysin malli voidaan kirjoittaa samaan tapaan kuin regressioanalyysin malli. Oletetaan esimerkiksi, että selittäjien arvojen lukumäärä $k = 3$. Silloin

$$\begin{aligned} y_{1j} &= \mu_1 + \varepsilon_{1j} \\ &= \mu_1 \cdot 1 + \mu_2 \cdot 0 + \mu_3 \cdot 0 + \varepsilon_{1j}, \quad j = 1, 2, \ldots, n_1; \\ y_{2j} &= \mu_1 \cdot 0 + \mu_2 \cdot 1 + \mu_3 \cdot 0 + \varepsilon_{2j}, \quad j = 1, 2, \ldots, n_2; \\ y_{3j} &= \mu_1 \cdot 0 + \mu_2 \cdot 0 + \mu_3 \cdot 1 + \varepsilon_{3j}, \quad j = 1, 2, \ldots, n_3. \end{aligned}$$

Nyt voidaan ajatella, että μ_1, μ_2 ja μ_3 ovat regressiokertoimia ja selittävät muuttujat x_1, x_2 ja x_3 saavat joko arvon 0 tai 1 sen mukaan, mihin luokkaan havainto kuuluu. Siten esimerkiksi $x_1 = 1$, $x_2 = 0$ ja $x_3 = 0$, kun havainto on peräisin 1. ryhmästä.

Taulukko 6.1 Koiranmuonan A myynti (pakkausten lukumäärä) hyllyn korkeuden mukaan

Hyllyn korkeus						
Polven taso		Vyötarön taso		Silmän taso		Yhteensä
y_{11}	77	y_{21}	88	y_{31}	85	
y_{12}	82	y_{22}	94	y_{32}	85	
y_{13}	86	y_{23}	93	y_{33}	87	
y_{14}	78	y_{24}	90	y_{34}	81	
y_{15}	81	y_{25}	91	y_{35}	80	
y_{16}	86	y_{26}	94	y_{36}	79	
y_{17}	77	y_{27}	90	y_{37}	87	
y_{18}	81	y_{28}	87	y_{38}	93	
$y_{1.} = 648$		$y_{2.} = 727$		$y_{3.} = 677$		$y_{..} = 2052$
$\bar{y}_{1.} = 81.0$		$\bar{y}_{2.} = 90.9$		$\bar{y}_{3.} = 84.6$		$\bar{y}_{..} = 85.5$

6.2.3 Varianssianalyysin neliösummahajotelma

Havaintojen y_{ij} poikkeama kokonaiskeskiarvosta $\bar{y}_{..}$ voidaan jakaa kahteen komponenttiin:

$$\begin{aligned} y_{ij} - \bar{y}_{..} =& (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.}) \\ =& \text{ ryhmäkeskiarvon poikkeama kaikkien keskiarvosta } + \\ & \text{ havainnon poikkeama ryhmäkeskiarvosta.} \end{aligned}$$

Laskemalla nähdään, että

$$\begin{aligned} \text{SST} &= \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2 \\ &= \sum_{i=1}^{a}\sum_{j=1}^{n_i}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 \\ &= \sum_{i=1}^{a} n_i(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 \\ &= \text{SSH} + \text{SSE}. \end{aligned} \tag{6.2}$$

Otoksesta lasketut ryhmäkeskiarvot $\bar{y}_{1.}$, ..., $\bar{y}_{a.}$ ovat pienimmän neliösumman mielessä parametrien μ_1, μ_2, ..., μ_a parhaita estimaatteja. Tämä tarkoittaa sitä, että

$$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \hat{\mu}_i)^2 \geq \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2,$$

valittiinpa estimaatit $\hat{\mu}_1, \ldots \hat{\mu}_a$ miten tahansa. Pienimmän neliösumman ratkaisuun liittyvä virheneliösumma eli kontrolloimaton vaihtelu

$$\text{SSE} = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$$

on pienempi kuin millään muulla ratkaisulla. Voimme laskea mallin selitysasteen samalla tavalla kuin regressioanalyysissa:

$$R^2 = \frac{\text{SSH}}{\text{SST}} = \frac{\text{SSH}}{\text{SSH} + \text{SSE}}.$$

R^2 osoittaa, paljonko ryhmäkeskiarvojen välinen vaihtelu selittää kokonaisvaihtelusta. Parametrien μ_1, μ_2, ..., μ_a pienimmän neliösumman estimaatit $\bar{y}_{1.}$, $\bar{y}_{2.}$, ..., $\bar{y}_{a.}$ ovat harhattomia eli

$$E(\bar{y}_{i.}) = \mu_i$$

kaikilla i:n arvoilla $i = 1, 2, \ldots, a$. Ryhmäkeskiarvojen poikkeamien $\alpha_i = \mu_i - \mu$ pienimmän neliösumman estimaatit ovat vastaavasti

$$\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}, \quad i = 1, 2, \ldots, a.$$

6.2.4 Virhevarianssin estimointi

Virheneliösummasta

$$\mathrm{SSE} = \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2$$

voidaan estimoida virhevarianssi $\mathrm{var}(\varepsilon_{ij}) = \sigma^2$ harhattomasti, sillä

$$E(\mathrm{SSE}) = \sigma^2(N - a).$$

Vapausasteillaan $N - a$ jaettu SSE on σ^2:n harhaton estimaatti:

$$s_e^2 = \frac{\mathrm{SSE}}{N - a} = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2}{N - a}. \tag{6.3}$$

Toinen tapa estimoida σ^2 olisi käyttää neliösummaa SSH. Lauseke

$$\frac{\mathrm{SSH}}{a - 1}$$

tuottaa yleensä liian suuria estimaatteja. Jos kuitenkin H_0: $\mu_1 = \mu_2 = \ldots = \mu_a$ on tosi, niin

$$E(\mathrm{SSH}) = (a - 1)\sigma^2$$

ja

$$\frac{\mathrm{SSH}}{a - 1}$$

on σ^2:n harhaton estimaatti. Tähän perustuu myös hypoteesin

$$H_0: \ \mu_1 = \mu_2 = \ldots = \mu_a$$

testaus.

Taulukko 6.2 Yksisuuntainen varianssianalyysi

Vaihtelun lähde	Vapaus-asteet	Neliösummat	MSE	F-testisuure
Malli	$a - 1$	$\mathrm{SSH} = \sum_i n_i(\bar{y}_{i.} - \bar{y}_{..})^2$	$\mathrm{MSH} = \frac{\mathrm{SSH}}{a-1}$	$F = \frac{\mathrm{MSH}}{\mathrm{MSE}}$
Jäännös	$N - a$	$\mathrm{SSE} = \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2$	$\mathrm{MSE} = \frac{\mathrm{SSE}}{N-a}$	
Kaikki	$N - 1$	$\mathrm{SST} = \sum_i \sum_j (y_{ij} - \bar{y}_{..})$		

6.2.5 Hypoteesien testaus

Hypoteesin

$$H_0: \ \mu_1 = \mu_2 = \ldots = \mu_a \tag{6.4}$$

testaus perustuu siihen huomioon, että vain H_0:n vallitessa MSH on σ^2:n harhaton estimaatti. Jos H_0 ei pidä paikkaansa, MSH antaa liian suuria estimaatteja. Sen sijaan MSE on σ^2:n harhaton estimaatti olipa H_0 tosi tai ei.

Jos H_0 on tosi, niin

$$F = \frac{\text{MSH}}{\text{MSE}} = \frac{\text{SSH}\ /(a-1)}{\text{SSE}\ /(N-a)} \tag{6.5}$$

saa 'pieniä' arvoja (ykkösen lähistöllä). Jos H_0 ei ole tosi, niin suurella todennäköisyydellä MSH $\gg$ MSE ja F saa 'liian' suuria arvoja. Sen perusteella voidaan epäillä H_0:n paikkansapitävyyttä.

6.3 Kaksisuuntainen varianssianalyysi

6.3.1 Kaksisuuntaisen varianssianalyysin malli

Kaksisuuntaisessa varianssianalyysissa tarkastellaan kahden tekijän A ja B vaikutusta selittävään muuttujaan y. Malli on muotoa

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \qquad \begin{aligned} i &= 1, \ldots, a \\ j &= 1, \ldots, b \\ k &= 1, \ldots, n. \end{aligned} \tag{6.6}$$

Virhetermien ε_{ijk} korreloimattomuutta ja varianssia koskevat oletukset ovat samanlaiset kuin yksisuuntaisessa varianssianalyysissa. Olkoon esimerkiksi $a = 2$ ja $b = 3$ ja $n = 3$. Havaintoaineisto ja ryhmäkeskiarvot on esitetty oheisessa taulukossa.

A \ B	1	2	3	
1	$y_{111}, y_{112}, y_{113}$	$y_{121}, y_{122}, y_{123}$	$y_{131}, y_{132}, y_{133}$	$\bar{y}_{1..}$
2	$y_{211}, y_{212}, y_{213}$	$y_{221}, y_{222}, y_{223}$	$y_{231}, y_{232}, y_{233}$	$\bar{y}_{2..}$
	$\bar{y}_{.1.}$	$\bar{y}_{.2.}$	$\bar{y}_{.3.}$	$\bar{y}_{...}$

Solukeskiarvot ovat muotoa

$$\bar{y}_{ij.} = \frac{1}{n} \sum_{k=1}^{n} y_{ijk}.$$

Vastaavasti mallin mukaiset teoreettiset ryhmäkeskiarvot ovat

$i \backslash j$	1	2	3	
1	μ_{11}	μ_{12}	μ_{13}	$\mu_{1.} = \frac{1}{3}(\mu_{11} + \mu_{12} + \mu_{13})$
2	μ_{21}	μ_{22}	μ_{23}	$\mu_{2.}$
	$\mu_{.1}$	$\mu_{.2} = \frac{1}{2}(\mu_{12}+\mu_{22}).$	$\mu_{.3}$	$\mu_{..}$

Kaksisuuntaisen varianssianalyysin malli voidaan kirjoittaa myös muodossa

$$
\begin{aligned}
y_{ijk} &= \mu_{..} + (\bar{\mu}_{i.} - \bar{\mu}_{..}) + (\bar{\mu}_{.j} - \bar{\mu}_{..}) + (\mu_{ij} - \bar{\mu}_{i.} - \bar{\mu}_{.j} + \bar{\mu}_{..}) + \varepsilon_{ijk} \\
&= \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij}, \qquad (6.7)
\end{aligned}
$$

missä

$$
\begin{aligned}
\mu &= \mu_{..} \\
\alpha_i &= \bar{\mu}_{i.} - \bar{\mu}_{..} \\
\beta_j &= \bar{\mu}_{.j} - \bar{\mu}_{..} \\
\gamma_{ij} &= \mu_{ij} - \bar{\mu}_{i.} - \bar{\mu}_{.j} + \bar{\mu}_{..}.
\end{aligned}
$$

Parametri α_i kuvaa rivivaikutusta, β_j sarakevaikutusta ja γ_{ij} yhdysvaikutusta.

Huomattakoon, että

$$
\sum_i \alpha_i = \sum_j \beta_j = \sum_i \sum_j \gamma_{ij} = 0.
$$

Parametrien pienimmän neliösumman estimaatit ovat

$$
\begin{aligned}
\hat{\mu} &= \bar{y}_{..} \\
\hat{\alpha}_i &= \bar{y}_{i.} - \bar{y}_{..} \\
\hat{\beta}_j &= \bar{y}_{.j} - \bar{y}_{..}
\end{aligned}
$$

ja

$$
\hat{\gamma}_{ij} = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}.
$$

Kaksisuuntaisessa varianssianalyysissa voidaan testata hypoteeseja

$$
\begin{aligned}
&H_A: \ \alpha_1 = \alpha_2 = 0 && A\text{:lla ei vaikutusta} \\
&H_B: \ \beta_1 = \beta_2 = \beta_3 = 0 && B\text{:llä ei vaikutusta} \\
&H_{AB}: \ \gamma_{11} = \gamma_{21} = \gamma_{12} && \\
&\qquad = \gamma_{22} = \gamma_{13} = \gamma_{23} = 0 && \text{ei yhdysvaikutusta.} \qquad (6.8)
\end{aligned}
$$

Testauksessa oletetaan, että havainnot noudattavat normaalijakaumaa.

Taulukko 6.3 Kaksisuuntainen varianssianalyysi

Vaihtelun lähde	Neliösummat	Vapaus-asteet	Keskineliö-virhe	F-testisuure
A	$\mathrm{SSA} = nb\sum_i(\bar{y}_{i..} - \bar{y}_{...})^2$	$a - 1$	$\mathrm{MSA} = \frac{\mathrm{SSA}}{a-1}$	$F = \frac{\mathrm{MSA}}{\mathrm{MSE}}$
B	$\mathrm{SSB} = na\sum_j(\bar{y}_{.j.} - \bar{y}_{...})^2$	$b - 1$	$\mathrm{MSB} = \frac{\mathrm{SSB}}{b-1}$	$F = \frac{\mathrm{MSB}}{\mathrm{MSE}}$
AB	$\mathrm{SSAB} = n\sum_{ij}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{...})^2$	$(a-1)(b-1)$	$\mathrm{MSAB} = \frac{\mathrm{SSAB}}{(a-1)(b-1)}$	$F = \frac{\mathrm{MSAB}}{\mathrm{MSE}}$
Virhe	$\mathrm{SSE} = \sum_{ijk}(y_{ijk} - \bar{y}_{ij.})^2$	$ab(n - 1)$		
Kaikki	$\mathrm{SST} = \sum_{ijk}(y_{ijk} - \bar{y}_{...})^2$	$abn - 1$		

Kokonaisvaihtelu $\mathrm{SST} = \sum_{ijk}(y_{ijk} - \bar{y}_{...})^2$ voidaan esittää neliösummana

$$\mathrm{SST} = \mathrm{SSA} + \mathrm{SSB} + \mathrm{SSAB} + \mathrm{SSE}, \tag{6.9}$$

jossa oikean puolen neliösummat ovat toisistaan riippumattomia. Keskineliövirheiden lausekkeet MSA, MSB, MSAB ja MSE muodostetaan F-testiä varten jakamalla neliösumma vapausasteillaan. Esimerkiksi hypoteesi

$$H_A: \ \alpha_1 = \alpha_2 = \ldots = \alpha_a = 0$$

testataan F-suureella $F = \mathrm{MSA}/\mathrm{MSE}$, joka noudattaa F–jakaumaa $F_{a-1,ab(n-1)}$ vapausastein $a - 1$ ja $ab(n - 1)$.

6.4 Monimuuttujainen varianssianalyysi (MANOVA)

6.4.1 Yksisuuntainen MANOVA

MANOVAssa on selitettäviä muuttujia eli vastemuuttujia useampia. Tutkitaan esimerkiksi uralla etenemisen vaikutusta työtä koskeviin asenteisiin. Selitettäviä muuttujia voisivat olla tyytyväisyys työhön (y_1) ja tyytyväisyys työympäristöön (y_2). Tutkitaan tilannetta uran vaiheissa A_1, A_2 ja A_3. Selitettävänä on vektoriarvoinen muuttuja

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}.$$

Silloin havaintoaineisto on muotoa

A_1	A_2	A_3
$\mathbf{y}_{11}, \mathbf{y}_{12}, \ldots, \mathbf{y}_{1n_1}$	$\mathbf{y}_{21}, \mathbf{y}_{22}, \ldots, \mathbf{y}_{2n_2}$	$\mathbf{y}_{31}, \mathbf{y}_{32}, \ldots, \mathbf{y}_{3n_3}$

Yksittäinen havaintovektori on

$$\mathbf{y}_{ij} = \begin{pmatrix} y_{ij1} \\ y_{ij2} \end{pmatrix}.$$

Yleisessä tilanteessa selitettävänä on samanaikaisesti p muuttujaa.Vastemuuttuja on vektori $\mathbf{y} = (y_1, y_2, \ldots, y_p)'$. Havaintojen $\mathbf{y}_{ij} = (y_{ij1}, y_{ij2}, \ldots y_{ijp})'$ malli on

$$\begin{aligned} \mathbf{y}_{ij} &= \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_{ij} \\ &= \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_{ij}; \quad i = 1, 2, \ldots, a, \quad j = 1, 2, \ldots, n_i. \end{aligned} \tag{6.10}$$

Jokaisen yksittäisen muuttujan y_r $(r = 1, \ldots, p)$ osalta malli on

$$y_{ijr} = \mu_r + \alpha_{ir} + \varepsilon_{ijr} = \mu_{ir} + \varepsilon_{ijr}.$$

Edellä esitetyssä esimerkissä $p = 2$ ja $a = 3$.

Voidaan ajatella, että meillä on a populaatiota, joista jokaisesta on tehty otos. Havaintojen kokonaismäärä $N = n_1 + n_2 + \ldots + n_a$.

Otos 1: $\mathbf{y}_{11}, \mathbf{y}_{12}, \ldots \mathbf{y}_{1n_1}$ otos populaatiosta $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$
Otos 2: $\mathbf{y}_{21}, \mathbf{y}_{22}, \ldots \mathbf{y}_{2n_2}$ otos populaatiosta $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$
$\vdots$
Otos a: $\mathbf{y}_{a1}, \mathbf{y}_{a2}, \ldots \mathbf{y}_{an_a}$ otos populaatiosta $N_p(\boldsymbol{\mu}_a, \boldsymbol{\Sigma})$.

Ryhmäkeskiarvot ovat muotoa

$$\bar{\mathbf{y}}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{y}_{ij}, \quad i = 1, 2, \ldots, a$$

ja koko aineiston keskiarvo

$$\bar{\mathbf{y}}_{..} = \frac{1}{N} \sum_{i=1}^{a} \sum_{j=1}^{n_i} \mathbf{y}_{ij}.$$

Havaintovektoreiden $\mathbf{y}_{ij}$ poikkeama kokonaiskeskiarvosta $\bar{\mathbf{y}}_{..}$ voidaan jakaa "käsittelyefektiin" ja "residuaaliin" vastaavalla tavalla kuin yhden muuttujan tapauksessa:

$$\mathbf{y}_{ij} - \bar{\mathbf{y}}_{..} = (\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..}) + (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i.}).$$

Ristitulo $(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{..})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{..})'$ voidaan nyt kirjoittaa muodossa

$$\begin{aligned} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{..})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{..})' &= [(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..}) + (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i.})][(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..}) + (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i.})]' \\ &= (\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})' + (\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i.})' \\ &+ (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i.})(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})' + (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i.})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i.})'. \end{aligned}$$

Jos yllä olevat ristitulot summataan puolittain indeksin j suhteen, niin keskimmäiset termit ovat nollamatriiseja, koska $\sum_{j=1}^{n_i}(\mathbf{y}_{ij}-\bar{\mathbf{y}}_{i.}) = \mathbf{0}$. Kun lausekkeet summataan yli indeksien i ja j, saadaan siis

$$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(\mathbf{y}_{ij}-\bar{\mathbf{y}}_{..})(\mathbf{y}_{ij}-\bar{\mathbf{y}}_{..})' = \sum_{i=1}^{a} n_i(\bar{\mathbf{y}}_{i.}-\bar{\mathbf{y}}_{..})(\bar{\mathbf{y}}_{i.}-\bar{\mathbf{y}}_{..})'+\sum_{i=1}^{a}\sum_{j=1}^{n_i}(\mathbf{y}_{ij}-\bar{\mathbf{y}}_{i.})(\mathbf{y}_{ij}-\bar{\mathbf{y}}_{i.})'.$$

Neliösummamatriisit ovat $p \times p$–matriiseja ja niitä merkitään seuraavasti:

$$\begin{aligned}
\mathbf{T} &= \sum_{i=1}^{a}\sum_{j=1}^{n_i}(\mathbf{y}_{ij}-\bar{\mathbf{y}}_{..})(\mathbf{y}_{ij}-\bar{\mathbf{y}}_{..})' \\
\mathbf{H} &= \sum_{i=1}^{a} n_i(\bar{\mathbf{y}}_{i.}-\bar{\mathbf{y}}_{..})(\bar{\mathbf{y}}_{i.}-\bar{\mathbf{y}}_{..})' \\
\mathbf{E} &= \sum_{i=1}^{a}\sum_{j=1}^{n_i}(\mathbf{y}_{ij}-\bar{\mathbf{y}}_{i.})(\mathbf{y}_{ij}-\bar{\mathbf{y}}_{i.})'.
\end{aligned}$$

Testin muodostaminen perustuu neliösummamatriiseihin $\mathbf{T}$, $\mathbf{H}$ ja $\mathbf{E}$.
Esimerkki 6.1 Olkoon havaintomatriisi $\mathbf{Y}$ seuraava:

$$\mathbf{Y}' = \begin{pmatrix} 9 & 6 & 9 & 0 & 2 & 3 & 1 & 2 \\ 3 & 2 & 7 & 4 & 0 & 8 & 9 & 7 \end{pmatrix}.$$

Ensimmäiset 3 havaintoa kuuluvat 1. ryhmään, 2 seuraavaa 2. ryhmään ja 3. viimeistä 3. ryhmään. Ryhmien keskiarvovektorit ja koko aineiston keskiarvo ovat

$$\bar{\mathbf{y}}_{1.} = \begin{pmatrix} 8 \\ 4 \end{pmatrix},\ \bar{\mathbf{y}}_{2.} = \begin{pmatrix} 1 \\ 2 \end{pmatrix},\ \bar{\mathbf{y}}_{3.} = \begin{pmatrix} 2 \\ 8 \end{pmatrix},\ \bar{\mathbf{y}}_{..} = \begin{pmatrix} 4 \\ 5 \end{pmatrix}.$$

Tästä aineistosta laskettu nelisummahajotelma $\mathbf{T} = \mathbf{H} + \mathbf{E}$ on

$$\begin{pmatrix} 88 & -11 \\ -11 & 72 \end{pmatrix} = \begin{pmatrix} 78 & -12 \\ -12 & 48 \end{pmatrix} + \begin{pmatrix} 10 & 1 \\ 1 & 24 \end{pmatrix}.$$

Nelisummiin liittyvät vapausateet ovat vastaavasti $n-1 = 8-1 = 7,\ 3-1 = 2$ ja $3+2+3-3 = 5$. Nyt $|\mathbf{H}| = 10 \cdot 24 - 1^2 = 239$ ja $|\mathbf{T}| = 88 \cdot 72 - (-11)^2 = 6215$, joten $\Lambda = 239/6215 = 0.0385$.

Testataan hypoteesia

$$H_0 :\ \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_a \quad \text{vs.} \quad H_1 :\ \text{ainakin kaksi } \boldsymbol{\mu}_i\text{:tä poikkeaa toisistaan.}$$

Hypoteesimatriisin $\mathbf{H}$ diagonaalilla on jokaisen muuttujan (p kappaletta) keskiarvojen välinen vaihtelu SSH . Virhematriisin $\mathbf{E}$ diagonaalilla on kunkin muuttujan virhevaihtelu SSE. Hypoteesiin H_0 liittyviä vapausasteita merkitään ν_H

ja virheneliösummaan liittyviä vapausasteita ν_E. Yksisuuntaisessa tapauksessa $\nu_H = a - 1$. Huomattakoon, että $\mathbf{S}_{pl} = \mathbf{E}/(N - a)$ on $\boldsymbol{\Sigma}$:n harhaton estimaatti

$$E\Big(\frac{\mathbf{E}}{N - a}\Big) = \boldsymbol{\Sigma}.$$

6.4.2 Wilksin testisuure

Hypoteesin $H_0: \ \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_a$ suurimman uskottavuuden testisuure

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}|} \tag{6.11}$$

on ns. Wilksin Λ. H_0 hylätään, kun $\Lambda \leq \Lambda_{\alpha,p,\nu_H,\nu_E}$. H_0 hylätään siis pienillä Λ:n arvoilla. Wilksin Λ:n jakauman parametrit ovat

p = muuttujien lukumäärä
ν_H = hypoteesiin liittyvät vapausasteet
ν_E = virheisiin liittyvät vapausasteet
α = riskitaso.

Kun n on suuri, niin silloin

$$-(n - 1 - (p + a)/2) \ln \Lambda$$

noudattaa likimain χ^2-jakaumaan vapausastein $p(a-1)$. Tietyissä erikoistapuksissa Wilksin Λ:n tarkka jakauma saadaan F-jakauman avulla.

Wilksin Λ vertailee keskiarvovektoreiden sisäistä virhevaihtelua $\mathbf{E}$ kokonaisvaihteluun $\mathbf{E} + \mathbf{H}$. Matriisien $\mathbf{E}$ ja $\mathbf{H}$ sisältämä monimuuttujainen informaatio tiivistetään yksiulotteiselle asteikolle. Tällä asteikolla päätetään, poikkeavatko keskiarvovektorit merkittävästi toisistaan.

Joitakin Wilksin Λ:n ominaisuuksia:

1. Jotta $|\mathbf{E}| > 0$, on epäyhtälön $\nu_E \geq p$ oltava voimassa.

2. MANOVAssa ν_H ja ν_E ovat samat kuin vastaavassa yhden muuttujan analyysissa. Yksisuuntaisessa mallissa esimerkiksi $\nu_H = a - 1$ ja $\nu_E = N - a$.

3. $0 \leq \Lambda \leq 1$.

4. Wilksin Λ voidaan lausua matriisin $\mathbf{E}^{-1}\mathbf{H}$ ominaisarvojen $\lambda_1, \lambda_2, \ldots, \lambda_s$ avulla seuraavasti:

$$\Lambda = \prod_{i=1}^{s} \frac{1}{1 + \lambda_i},$$

missä $s = \min(p, \nu_H)$ on matriisin $\mathbf{E}^{-1}\mathbf{H}$ nollasta poikkeavien ominaisarvojen lukumäärä.

6.4.3 Monimuuttujainen kaksisuuntainen MANOVA

Kaksisuuntainen p:n riippuvan muuttujan MANOVA-malli saadaan sijoittamalla vastaavat $p \times 1$–vektorit yhden muuttujan malliin (6.7). Esimerkiksi $\mathbf{y}_{ijk} = (y_{ijk1}, y_{ijk2}, \ldots, y_{ijkp})$ ja $\gamma_{ij} = (\gamma_{ij1}, \gamma_{ij2}, \ldots, \gamma_{ijp})$. Malli on

$$\mathbf{y}_{ijk} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j + \boldsymbol{\gamma}_{ij} + \boldsymbol{\varepsilon}_{ijk} = \boldsymbol{\mu}_{ij} + \boldsymbol{\varepsilon}_{ijk}, \tag{6.12}$$
$$i = 1, 2, \ldots, a; \ j = 1, 2, \ldots, b; \ k = 1, 2, \ldots, n,$$

missä α_i on A:n i. tason vaikutus jokaisen $\mathbf{y}_{ijk}$:n muuttujaan, $\boldsymbol{\beta}_j$ on B:n j. tason vaikutus ja $\boldsymbol{\gamma}_{ij}$ on AB vuorovaikutus. Oletamme, että

$$\sum_i \boldsymbol{\alpha}_i = \sum_j \boldsymbol{\beta}_j = \sum_i \boldsymbol{\gamma}_{ij} = \sum_j \boldsymbol{\gamma}_{ij} = \mathbf{0},$$

$\boldsymbol{\varepsilon}_{ijk}$ ovat toisistaan riippumattomia ja

$$\boldsymbol{\varepsilon}_{ijk} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}), \quad \boldsymbol{\beta}_j = \bar{\boldsymbol{\mu}}_{.j} - \bar{\boldsymbol{\mu}}_{..} \quad \text{ja} \quad \boldsymbol{\gamma}_{ij} = \bar{\boldsymbol{\mu}}_{ij} - \bar{\boldsymbol{\mu}}_{i.} - \bar{\boldsymbol{\mu}}_{i.} - \bar{\boldsymbol{\mu}}_{..}.$$

Annettujen reunaehtojen vallitessa $\boldsymbol{\alpha}_i = \bar{\boldsymbol{\mu}}_{i.} - \bar{\boldsymbol{\mu}}_{..}$, missä $\bar{\boldsymbol{\mu}}_{i.} = \sum_j \boldsymbol{\mu}_{ij}/b$, $\bar{\boldsymbol{\mu}}_{.j} = \sum_i \boldsymbol{\mu}_{ij}/a$ ja $\bar{\boldsymbol{\mu}}_{..} = \sum_{ij} \boldsymbol{\mu}_{ij}/ab$. Määritelmät ovat vastaavat kuin yhden muuttujan mallissa.

Monimuuttujaiset keskiarvovektorit $\bar{\mathbf{y}}_{i..}$, $\bar{\mathbf{y}}_{.j.}$, $\bar{\mathbf{y}}_{ij.}$ ja $\bar{\mathbf{y}}_{...}$ määritellään samalla tavalla kuin yhden muuttujan mallissa. Testauksessa tarvittavat neliösummamatriisit on annettu oheisessa taulukossa.

Taulukko 6.4 Kaksisuuntainen MANOVA

Vaihtelun lähde	Neliö- ja tulosummamatriisit	Vapaus asteet	Testisuure
A	$\mathbf{H}_A = nb\sum_i(\bar{\mathbf{y}}_{i..} - \bar{\mathbf{y}}_{...})\times(\bar{\mathbf{y}}_{i..} - \bar{\mathbf{y}}_{...})'$	$a-1$	$\Lambda_A = \frac{\lvert\mathbf{E}\rvert}{\lvert\mathbf{E}+\mathbf{H}_A\rvert} \sim \Lambda_{p,a-1,ab(n-1)}$
B	$\mathbf{H}_B = na\sum_j(\bar{\mathbf{y}}_{.j.} - \bar{\mathbf{y}}_{...})\times(\bar{\mathbf{y}}_{.j.} - \bar{\mathbf{y}}_{...})'$	$b-1$	$\Lambda_B = \frac{\lvert\mathbf{E}\rvert}{\lvert\mathbf{E}+\mathbf{H}_B\rvert} \sim \Lambda_{p,b-1,ab(n-1)}$
AB	$\mathbf{H}_{AB} = n\sum_{ij}(\bar{\mathbf{y}}_{ij.} - \bar{\mathbf{y}}_{i..} - \bar{\mathbf{y}}_{.j.} - \bar{\mathbf{y}}_{...})\times(\bar{\mathbf{y}}_{ij.} - \bar{\mathbf{y}}_{i..} - \bar{\mathbf{y}}_{.j.} - \bar{\mathbf{y}}_{...})'$	$(a-1)(b-1)$	$\Lambda_{AB} = \frac{\lvert\mathbf{E}\rvert}{\lvert\mathbf{E}+\mathbf{H}_{AB}\rvert} \sim \Lambda_{p,(a-1)(b-1),ab(n-1)}$
Virhe	$\mathbf{E} = \sum_{ijk}(\mathbf{y}_{ijk} - \bar{\mathbf{y}}_{ij.})\times(\mathbf{y}_{ijk} - \bar{\mathbf{y}}_{ij.})'$	$ab(n-1)$	
Kaikki	$\mathbf{T} = \sum_{ijk}(\mathbf{y}_{ijk} - \bar{\mathbf{y}}_{...})\times(\mathbf{y}_{ijk} - \bar{\mathbf{y}}_{...})'$	$abn-1$	

Kaksisuuntaisessa MANOVAssa kokonaisneliö- ja tulosummamatriisi $\mathbf{T}$ voidaan osittaa seuraavasti:

$$\mathbf{T} = \mathbf{H}_A + \mathbf{H}_B + \mathbf{H}_{AB} + \mathbf{E}.$$

Esimerkiksi matriisin $\mathbf{H}_A$ diagonaalialkio h_{Arr}, $r = 1, 2, \ldots, r$, on

$$h_{Arr} = nb\sum_{i=1}^{a}(\bar{y}_{i..r} - \bar{y}_{...r})^2,$$

missä $\bar{y}_{i..r}$ on vektorin $\bar{\mathbf{y}}_{i..}$ r. alkio ja $\bar{y}_{...r}$ on vektorin $\bar{\mathbf{y}}_{...}$ r. alkio.

Esimerkiksi hypoteesin

$$\begin{aligned} H_{0AB} &: \ \boldsymbol{\gamma}_{11} = \ldots = \boldsymbol{\gamma}_{a1} = \ldots = \boldsymbol{\gamma}_{ab} = \mathbf{0} \qquad \text{vs.} \\ H_{1AB} &: \ \text{ainakin yksi } \boldsymbol{\gamma}_{ij} \neq \mathbf{0} \end{aligned}$$

testaamiseksi käytetään testisuuretta

$$\Lambda_{AB} = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}_{AB}|} \sim \Lambda_{p,(a-1)(b-1),ab(n-1)},$$

joka noudattaa jakaumaa $\Lambda_p, (a-1)(b-1), ab(n-1)$, kun H_{0AB} on tosi. H_{0AB} hylätään pienillä Λ_{AB}:n arvoilla. Vastaavat testisuureet hypoteesien $H_A : \ \boldsymbol{\alpha}_1 = \ldots = \boldsymbol{\alpha}_a = \mathbf{0}$ ja $H_B : \ \boldsymbol{\beta}_1 = \ldots = \boldsymbol{\beta}_b$ testaamiseksi on esitetty oheisessa taulukossa.