Kapitel 11

Successiv testning

Som nævnt i kapitel 10 har de fleste statistikpakker en facilitet, der gør det muligt at få foretaget de F–tests for modelreduktioner som svarer til at modelformlens led fjernes ét for ét bagfra. I dette kapitel vil vi gå mere i detaljer med hvordan regnestørrelserne fra disse tests normalt sammenfattes i et såkaldt *variansanalyseskema*, og hvordan sådan et skema skal læses.

Antag at vi har specificeret en model ved modelformlen

$$\texttt{Y=led(1)+led(2)+}\ \ldots\ \texttt{+led(k)}$$

hvor `led(1)` typisk (men ikke nødvendigvis) er `1`, medens de øvrige led er faktorer, regressionsvariable eller produkter af sådanne. Med

$$\mathrm{SSD}_{\mathrm{res}}\,(\texttt{modelformel})$$

betegnes i det følgende residualkvadratsummen i modellen specificeret ved en given modelformel. For kortheds skyld indfører vi midlertidigt notationen

$$\mathrm{SSD}^{j}_{\mathrm{res}} = \mathrm{SSD}_{\mathrm{res}}\big(\texttt{led(1)+led(2)+}\ \ldots\ \texttt{+led(j)}\big)$$

for residualkvadratsummen i modellen bestemt ved de første j led. Med f_j betegner vi antallet af frihedsgrader for denne residualkvadratsum (dvs. antallet af frihedsgrader i dens χ^2–fordeling), bestemt ved at $n - f_j$ er dimensionen af modellens middelværdiunderrum. Vi definerer naturligt $f_0 = n$ og $\mathrm{SSD}^{0}_{\mathrm{res}} = \sum y_i^2$, svarende til at modellen givet ved den "tomme modelformel" er den hvor alle observationer har middelværdi 0.

F–teststørrelsen for reduktion af modellen

$$\texttt{Y=led(1)+led(2)+}\ \ldots\ \texttt{+led(j)}$$

til

$$\texttt{Y=led(1)+led(2)+}\ \ldots\ \texttt{+led(j-1)}$$

er så ifølge sætning 10.3

$$F_j = \frac{(\mathrm{SSD}^{j-1}_{\mathrm{res}} - \mathrm{SSD}^{j}_{\mathrm{res}})/(f_{j-1} - f_j)}{\mathrm{SSD}^{j}_{\mathrm{res}}/f_j}.$$

11.1. Variansanalyseskemaet.

Resultaterne af disse tests og nogle af de regnestørrelser der indgår i dem plejer man at sammenfatte i et skema af følgende form:

Effect	S.S.	d.f.	M.S.	F
`led(1)`	$\mathrm{SSD}^0_{\mathrm{res}} - \mathrm{SSD}^1_{\mathrm{res}}$	$f_0 - f_1$	$\frac{\mathrm{SSD}^0_{\mathrm{res}} - \mathrm{SSD}^1_{\mathrm{res}}}{f_0 - f_1}$	F_1
$\vdots$			$\vdots$	
`led(j)`	$\mathrm{SSD}^{j-1}_{\mathrm{res}} - \mathrm{SSD}^j_{\mathrm{res}}$	$f_{j-1} - f_j$	$\frac{\mathrm{SSD}^{j-1}_{\mathrm{res}} - \mathrm{SSD}^j_{\mathrm{res}}}{f_{j-1} - f_j}$	F_j
$\vdots$			$\vdots$	
`led(k)`	$\mathrm{SSD}^{k-1}_{\mathrm{res}} - \mathrm{SSD}^k_{\mathrm{res}}$	$f_{k-1} - f_k$	$\frac{\mathrm{SSD}^{k-1}_{\mathrm{res}} - \mathrm{SSD}^k_{\mathrm{res}}}{f_{k-1} - f_k}$	F_k
`Residual`	$\mathrm{SSD}^k_{\mathrm{res}}$	f_k	$\hat{\sigma}^2$	
`Total`	$\sum y_i^2$	n	$\frac{\sum y_i^2}{n}$	

Geometrisk kan tallene i anden søjle, `S.S.`'erne, fortolkes som leddene i opspaltningen

$$\|y\|^2 = \|P_1 y\|^2 + \|P_2 y - P_1 y\|^2 + \cdots + \|P_k y - P_{k-1} y\|^2 + \|y - P_k y\|^2$$

hvor P_j betegner ortogonalprojektionen på middelværdiunderrummet svarende til modellen "`Y=led(1)+` ... `+led(j)`". Denne formel udtrykker blot at $\|y\|^2$ kan udregnes som summen af de kvadratiske normer af komponenterne i opspaltningen

$$y = P_1 y + (P_2 y - P_1 y) + \cdots + (P_k y - P_{k-1} y) + (y - P_k y)$$

svarende til fremstillingen

$$\mathbf{R}^n = L_1 \oplus (L_2 \cap L_1^\perp) \oplus \cdots \oplus (L_k \cap L_{k-1}^\perp) + L_k^\perp$$

af $\mathbf{R}^n$ som direkte ortogonal sum.

Vi har her gengivet variansanalyseskemaet som det ser ud i den matematisk enkleste version, som også er den man finder i output fra ISU. Filosofien bag denne variant af variansanalyseskemaet er, at et konstantled i en model skal behandles lige som alle andre led; dels for ikke at forplumre logikken, og dels for at give mulighed for specifikation af modeller uden konstantled, eller modeller med et konstantled placeret senere i en modelformel med henblik på at forsøge at teste det væk før visse andre led.

Men i output fra de fleste andre programpakker (f.eks. SAS, hvor dette variansanalyseskema går under betegnelsen "type I"), og også i den klassiske fremstilling af variansanalyseskemaet (som er meget ældre end de statistiske programpakker), benyttes en lidt anden konvention, som groft sagt bygger på den antagelse, at enhver model har et (oftest underforstået) konstantled der står som første led i modelformlen, og testet for fjernelse af dette led (altså reduktion fra "fuldstændig homogenitet" til "fælles middelværdi = 0") er uden interesse. Da testet i første linie af variansanalyseskemaet således er irrelevant, udelades denne linie. Testet i den modificerede tabels første linie (altså anden linie i skemaet ovenfor) bliver herefter et test for fuldstændig homogenitet. Idet sidste linie `Total` modificeres tilsvarende, så `S.S.`'er og `d.f.`'er stadig er summer af tallene ovenover, får skemaet nu følgende udseende:

Effect	S.S.	d.f.	M.S.	F
`led(2)`	$\mathrm{SSD}^1_{\mathrm{res}} - \mathrm{SSD}^2_{\mathrm{res}}$	$f_1 - f_2$	$\frac{\mathrm{SSD}^1_{\mathrm{res}} - \mathrm{SSD}^2_{\mathrm{res}}}{f_1 - f_2}$	F_2
$\vdots$			$\vdots$	
`led(j)`	$\mathrm{SSD}^{j-1}_{\mathrm{res}} - \mathrm{SSD}^j_{\mathrm{res}}$	$f_{j-1} - f_j$	$\frac{\mathrm{SSD}^{j-1}_{\mathrm{res}} - \mathrm{SSD}^j_{\mathrm{res}}}{f_{j-1} - f_j}$	F_j
$\vdots$			$\vdots$	
`led(k)`	$\mathrm{SSD}^{k-1}_{\mathrm{res}} - \mathrm{SSD}^k_{\mathrm{res}}$	$f_{k-1} - f_k$	$\frac{\mathrm{SSD}^{k-1}_{\mathrm{res}} - \mathrm{SSD}^k_{\mathrm{res}}}{f_{k-1} - f_k}$	F_k
`Residual`	$\mathrm{SSD}^k_{\mathrm{res}}$	f_k	$\hat{\sigma}^2$	
`Total`	$\sum(y_i - \bar{y})^2$	$n-1$	$\frac{\sum(y_i - \bar{y})^2}{n-1}$	

Geometrisk kan tallene i `S.S.`–søjlen her fortolkes som leddene i opspaltningen

$$\|y - P_1 y\|^2 = \|P_2 y - P_1 y\|^2 + \cdots + \|P_k y - P_{k-1} y\|^2 + \|y - P_k y\|^2$$

hvor (for `led(1)=1`) $\|y - P_1 y\|^2 = \mathrm{SSD}^1_{\mathrm{res}} = \sum(y_i - \bar{y})^2$.

Forskellen mellem de to varianter af variansanalyseskemaet kan siges at ligge i hvor langt ned man er interesseret i at teste. Hvis man som den mindste slutmodel vælger homogenitetsmodellen får man den klassiske version. Hvis man som slutmodel tager modellen hvor alle observationer har middelværdi 0 får man ISU–versionen. Begge varianter er specialtilfælde af en mere generel — temmelig indlysende — definition af et variansanalyseskema hørende til en vilkårlig aftagende følge af modeller.

Om begge typer af variansanalyseskemaer gælder følgende. I tabellens overskrifter står `S.S.` naturligvis for "sum of squares", `d.f.` for "degrees of freedom" og `M.S.` for "mean of squares". Den sidste søjle med F–teststørrelserne er i de fleste statistikpakker suppleret med en ekstra søjle, der indeholder de tilhørende P–værdier (højre halesandsynligheder i de relevante F–fordelinger).

Bemærk at F–størrelserne i tabellens sidste søjle kan udregnes ud fra `S.S.`'er og `d.f.`'er. Søjlen af `M.S.`'er, som er `S.S.`'er divideret med `d.f.`'er, kan forekomme lidt overflødig. Når man alligevel plejer at tage den med er det fordi man kan bruge den til at danne sig et o-verblik over hvor de væsentlige effekter i modellen ligger — ikke kun hvad angår signifikans, men også når det gælder faktisk størrelsesorden. Under hypotesen om fuldstændig homogenitet følger det let af tidligere fordelingsresultater at alle disse størrelser har middelværdi σ^2. Hvis et af modellens led har særligt stor effekt på responsernes variation vil derimod den pågældende `M.S.` (som jo er tælleren i det tilsvarende F–test) typisk være stor.

Ved brug af variansanalyseskemaet til successiv testning skal man være opmærksom på, at man kun i heldige tilfælde kan aflæse alle relevante tests. Man kan fjerne led nedefra indtil man støder på et der er signifikant; men om der så er et andet led ovenover der kan fjernes, kan man ikke umiddelbart se — det kræver et nyt skema, hvor det pågældende led skrives som det sidste af dem der er tilbage. Derfor bør man altid, ved specifikation af en modelformel, sørge for at de mindst vigtige led (dvs. dem man tror eller håber på at få fjernet, f.eks. vekselvirkninger af højere orden) kommer sidst.

EKSEMPEL. I en ensidet variansanalysemodel ser variansanalyseskemaet — med udeladelse af `M.S.` søjlen (fordi den ikke er særligt informativ, og fordi der ikke er plads til den) ud som følger. Vi gengiver begge versioner for en sikkerheds skyld:

ISU versionen:

`Effect`	`S.S.`	`d.f.`	`F`
`CONSTANT`	$n\bar{y}^2$	1	$\frac{n\bar{y}^2}{\sum\sum(y_{gi}-\bar{y})^2/(n-1)}$
`Mellem grupper`	$\sum n_g(\bar{y}_{g\cdot}-\bar{y})^2$	$G-1$	$\frac{\sum n_g(\bar{y}_{g\cdot}-\bar{y})^2/(G-1)}{\sum\sum(y_{gi}-\bar{y}_{g\cdot})^2/(n-G)}$
`Residual`	$\sum\sum(y_{gi}-\bar{y}_{g\cdot})^2$	$n-G$	
`Total`	$\sum\sum y_{gi}^2$	n	

Den traditionelle version:

`Effect`	`S.S.`	`d.f.`	`F`
`Mellem grupper`	$\sum n_g(\bar{y}_{g\cdot}-\bar{y})^2$	$G-1$	$\frac{\sum n_g(\bar{y}_{g\cdot}-\bar{y})^2/(G-1)}{\sum\sum(y_{gi}-\bar{y}_{g\cdot})^2/(n-G)}$
`Residual`	$\sum\sum(y_{gi}-\bar{y}_{g\cdot})^2$	$n-G$	
`Total`	$\sum\sum(y_{gi}-\bar{y})^2$	$n-1$	

I det første skema er den totale kvadratsum af observationerne opspaltet i tre bidrag, der kan fortolkes således:

`CONSTANT`: Totalgennemsnittets afstand fra 0.

`Mellem grupper`: Variationen mellem grupper.

`Residual`: Variationen indenfor grupper.

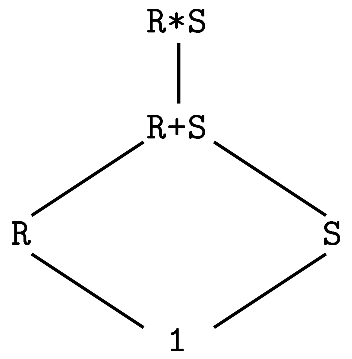
I det andet skema er det den totale kvadrat*afvigelses*sum omkring gennemsnittet, der er opspaltet i de to sidste bidrag.

Opgave 11.1.1. Vi er ikke helt færdige med eksempel 9.1 (side 82, se også opgave 10.7.1 på side 112). Fortolk nedenstående to variansanalyseskemaer i relation til tegningerne på side 112. `PLAC` er faktoren Nord/Syd, `M2` er boligarealet og responsen er månedlig bruttoydelse.

Effect	D.F.	S.S.	M.S.	F	P
CONSTANT	1	7587571912	7587571912	1049.1793	0.000000
M2	1	159198455	159198455	32.2958	0.000000
PLAC	1	36837248	36837248	8.2995	0.005367
PLAC*M2	1	4750842	4750842	1.0715	0.304491
RESIDUAL	64	283751543	4433618		
TOTAL	68	8072110000			

Effect	D.F.	S.S.	M.S.	F	P
CONSTANT	1	7587571912	7587571912	1049.1793	0.000000
M2	1	159198455	159198455	32.2958	0.000000
PLAC*M2	1	41487398	41487398	9.5003	0.003013
PLAC	1	100692	100692	0.0227	0.880685
RESIDUAL	64	283751543	4433618		
TOTAL	68	8072110000			

11.2. Tosidet variansanalyse i det balancerede tilfælde.



Lad der være givet observationer

$$y_{rsi}\ ,\quad r=1,\ldots,R\ ,\quad s=1,\ldots,S\ ,\quad i=1,\ldots,n_{rs}$$

svarende til at data foreligger i form af en tosidet $R\times S$–tabel med n_{rs} observationer i celle (r,s). Vi forudsætter naturligvis at inddelingerne i rækker og søjler har en eller anden fortolkning, som gør det rimeligt at lade de tilsvarende faktorer indgå i modellen, og at der ikke er andre faktorer eller regressionsvariable at tage hensyn til.

Diagrammet på foregående side viser (med betegnelserne `R` og `S` for de to faktorer) de modeller, der er af potentiel interesse i en sådan situation. Ved at specificere modellen på formen `Y=1+R+S+R*S` kan vi få udført de tests, der svarer til at følge diagrammets pile fra oven og ned langs venstre side. De to tests der mangler (reduktionen fra `R+S` til `S` og fra `S` til `1` kan vi få udført ved at at bytte om på de to hovedvirkningsled `R` og `S` i modelformlen.

Der er i almindelighed ikke nogen simpel algebraisk sammenhæng mellem de tests for fjernelse af hovedvirkningsleddene `R` og `S`, som hører til de to forskellige rækkefølger. Undtagelsen herfra er tilfældet, hvor den tosidede tabel er *balanceret*, dvs. hvor alle celletallene n_{rs} er ens. I dette tilfælde gælder følgende resultat:

SÆTNING 11.1. *Hvis $n_{rs} = k$ for alle (r,s) er maksimaliseringsestimatorerne for de estimerede middelværdier under additivitetsmodellen* `R+S` *givet ved*

$$\hat{\mu}_{rsi} = \bar{y}_{r\cdot\cdot} + \bar{y}_{\cdot s\cdot} - \bar{y}_{\cdot\cdot\cdot} \quad ,$$

og mellem residualkvadratsummerne i modellerne gælder relationen

$$\mathrm{SSD}_{\mathrm{res}}(\mathtt{R+S}) = \mathrm{SSD}_{\mathrm{res}}(\mathtt{R}) + \mathrm{SSD}_{\mathrm{res}}(\mathtt{S}) - \mathrm{SSD}_{\mathrm{res}}(\mathtt{1}) \quad .$$

Sætningen gælder i øvrigt også under den svagere betingelse at celletallene er *proportionale*, dvs. $n_{rs} = n_{r\cdot}n_{\cdot s}/n_{\cdot\cdot}$, og det er ikke så svært at generalisere beviset så det dækker dette tilfælde også. Det vil vi overlade til læseren. Men vi udskyder beviset lidt, det er vigtigere at forstå sætningens konsekvenser.

Sætningen gælder naturligvis også for $k = 1$, men her må man tage det forbehold, at da modellen `R*S` er udartet, i den forstand at hver observation har sin egen frit varierende middelværdi, kan man i denne model ikke estimere variansen, og man kan derfor heller ikke bruge denne model som udgangspunkt for test af den additive model. I et tosidet skema med én observation per celle er man nødt til, som grundmodel, at tage den additive model.

Da kvadratsummerne $\mathrm{SSD}_{\mathrm{res}}(\mathtt{R})$, $\mathrm{SSD}_{\mathrm{res}}(\mathtt{S})$ og $\mathrm{SSD}_{\mathrm{res}}(\mathtt{1})$ er lette at udregne (det drejer sig jo om residualkvadratsummer i ensidede variansanalysemodeller, for den sidstes vedkommende endda en residualkvadratsum i en simpel homogenitetsmodel, jvf. kapitel 7), følger det af sætningen at alle udregningerne i forbindelse med tosidet variansanalyse i det balancerede tilfælde kan udføres relativt enkelt ved hjælp af en lommeregner. De fittede værdier — og dermed passende estimater af række– og søjleparametre — er jo også nemme at beregne, ifølge formlen for $\hat{\mu}_{rsi}$. Specielt ser man at ML–estimater for række– eller søjlekontraster har den simple form

$$\hat{\alpha}_{r'} - \hat{\alpha}_{r''} = \bar{y}_{r'\cdot\cdot} - \bar{y}_{r''\cdot\cdot}\ , \qquad \hat{\beta}_{s'} - \hat{\beta}_{s''} = \bar{y}_{\cdot s'\cdot} - \bar{y}_{\cdot s''\cdot}\ .$$

Det er i øvrigt forholdsvis let direkte at indse, at disse differenser mellem række– eller søjlegennemsnit netop har de tilsvarende differenser mellem række– eller søjleparametre som middelværdier hvis og kun hvis celletallene er proportionale.

Af sætningen følger (ved en simpel omskrivning af sætningens sidste relation, som overlades til læseren) at den totale kvadratafvigelsessum i et balanceret tosidet skema kan skrives på formen

$$\begin{aligned}
&\sum_{r,s,i}(y_{rsi} - \bar{y}_{\cdots})^2 \\
&= \sum_{r,s,i}(y_{rsi} - \bar{y}_{rs\cdot})^2 \\
&+ \sum_{r,s,i}(\bar{y}_{rs\cdot} - \bar{y}_{r\cdot\cdot} - \bar{y}_{\cdot s\cdot} + \bar{y}_{\cdots})^2 \\
&+ \sum_{r,s,i}(\bar{y}_{r\cdot\cdot} - \bar{y}_{\cdots})^2 \\
&+ \sum_{r,s,i}(\bar{y}_{\cdot s\cdot} - \bar{y}_{\cdots})^2
\end{aligned}$$

De fire led på højre side er netop dem man finder i variansanalyseskemaet, og de fortolkes naturligt således:

Total kvadratafvigelsesum

= variationen indenfor celler

\+ vekselvirkningen

\+ variationen mellem rækker

\+ variationen mellem søjler.

For k–sidede balancerede tabeller gælder et tilsvarende resultat, der siger at den totale kvadratafvigelsessum splitter op som en sum af bidrag der stammer fra variationen indenfor celler, hovedvirkningerne af de k faktorer og alle mulige vekselvirkninger imellem dem af ordener 1, 2, …, $k-1$. Historisk set er det denne opspaltning, snarere end tankegangen bag successiv testning, der ligger bag det klassiske variansanalyseskema. Det er måske til syvende og sidst disse algebraiske relationer der er skyld i, at vi den dag idag sammenfatter resultaterne af en hel række tests i et enkelt skema.

De beregningsmæssige konsekvenser var af afgørende betydning indtil omkring 1960–70. Før den tid fremstillede man oftest variansanalysen som analysen af balancerede data, hvor muligheden for analyse af ikke–balancerede tabeller ved hjælp af tidskrævende matrixberegninger indgik som en kuriositet, der snarere hørte hjemme under overskriften "multipel regression".

Det beregningsmæssige aspekt af balancerthed er næsten uden betydning idag, og man kan med en vis ret hævde, at mange lærebøger — også nogle af de nyere — lægger uforholdsmæssigt stor vægt på udregningerne i det balancerede tilfælde. Men der er én konsekvens af ovenstående sætning og dens analoger for flersidede variansanalysemodeller, som det er væsentligt at være klar over, også selv om man aldrig kommer til at udføre en større variansanalyse ved hjælp af en lommeregner:

F–testet for fjernelse af $\mathtt{R}$ fra modellen $\mathtt{1+R+S}$ — altså testet for manglende rækkeeffekt i den additive model — bliver jo

$$F(R-1, RSk-(R+S-1)) = \frac{(\mathrm{SSD}_{\mathrm{res}}(\mathtt{S}) - \mathrm{SSD}_{\mathrm{res}}(\mathtt{R+S}))/(R-1)}{\mathrm{SSD}_{\mathrm{res}}(\mathtt{R+S})/(RSk-(R+S-1))},$$

medens testet for fjernelse af $\mathtt{R}$ fra modellen $\mathtt{1+R}$ — altså testet for homogenitet i den ensidede variansanalysemodel bestemt ved faktoren $\mathtt{R}$ — er baseret på F–størrelsen

$$F(R-1, RSk-R) = \frac{(\mathrm{SSD}_{\mathrm{res}}(\mathtt{1}) - \mathrm{SSD}_{\mathrm{res}}(\mathtt{R}))/(R-1)}{\mathrm{SSD}_{\mathrm{res}}(\mathtt{R})/(RSk-R)}.$$

Men af sætningen følger at

$$\mathrm{SSD}_{\mathrm{res}}(\mathtt{S}) - \mathrm{SSD}_{\mathrm{res}}(\mathtt{R+S}) = \mathrm{SSD}_{\mathrm{res}}(\mathtt{1}) - \mathrm{SSD}_{\mathrm{res}}(\mathtt{R}),$$

dvs. at de to F–teststørrelser har samme tæller. Det betyder, at de to kvadratsummer, der i variansanalyseskemaet er benævnt $\mathtt{R}$ og $\mathtt{S}$, *ikke* afhænger af hvilken af de to mulige testrækkefølger man vælger. Den eneste forskel på de to skemaer, hvad angår søjlerne `S.S.`, `d.f.` og `M.S.`, er at disse to linier er byttet om.

Heraf følger ganske vist ikke at de to tests for $\mathtt{R}$–effekt giver samme konklusion. Nævnerne i de to F–teststørrelser er jo ikke ens, og de skal heller ikke vurderes i den samme fordeling. Men det at de to tællere er ens betyder alligevel, at de to tests i almindelighed vil føre til omtrent samme konklusion i det balancerede tilfælde; forstået på den måde, at hvis F–størrelsen for reduktion af $\mathtt{R+S}$ til $\mathtt{R}$ er insignifikant, så vil de to mulige tests for fjernelse af $\mathtt{R}$ føre til omtrent samme konklusion. Hvis man —

som SAS gør i sine såkaldte "type I"–variansanalyseskemaer — benytter F–tests, der som nævner hele vejen igennem har variansestimatet i den oprindelige model (i stedet for at "poole" variansestimatet, dvs. opdatere med den yderligere information fra de hypoteser man tidligere har godkendt), får man *præcis* samme test for R–effekt, uanset rækkefølgen man tester i. Formelt er det forkert at gøre det på den måde, men man kan argumentere for, at denne fejl ikke betyder så meget i praksis. Forskellen fra det egentlige kvotienttest ligger i, at man benytter et andet (lidt dårligere, dvs. baseret på færre frihedsgrader) centralt estimat for variansen som nævner i F–størrelsen.

Bevis for sætning 11.1. Vi benytter i det følgende betegnelserne

$$L_{\mathtt{m.f.}} \subseteq \mathbf{R}^n \quad \text{og} \quad P_{\mathtt{m.f.}} : \mathbf{R}^n \to \mathbf{R}^n$$

for henholdsvis middelværdiunderrummet og ortogonalprojektionen på dette for modellen givet ved en modelformel **m.f.**.

Den afgørende egenskab ved det balancerede tilfælde er, at når vi ser på de to ortogonale opspaltninger

$$L_{\mathtt{R}} = L_{\mathtt{1}} \oplus (L_{\mathtt{R}} \cap L_{\mathtt{1}}^{\perp}) \qquad \text{og} \qquad L_{\mathtt{S}} = L_{\mathtt{1}} \oplus (L_{\mathtt{S}} \cap L_{\mathtt{1}}^{\perp})$$

(som altid gælder, uanset om vi har balancerthed eller ej) så har vi ikke alene ortogonalitet mellem underrummene indenfor hver af dem (det ligger jo i notationen "$\oplus$"), men der er også ortogonalitet "på tværs af de to faktorer", i den forstand at der gælder

$$(L_{\mathtt{R}} \cap L_{\mathtt{1}}^{\perp}) \perp L_{\mathtt{S}} \qquad \text{og} \qquad (L_{\mathtt{S}} \cap L_{\mathtt{1}}^{\perp}) \perp L_{\mathtt{R}}$$

At det forholder sig sådan følger umiddelbart, hvis vi kan indse at der for vilkårlige vektorer $v \in L_{\mathtt{R}} \cap L_{\mathtt{1}}^{\perp}$ og $u \in L_{\mathtt{S}}$ gælder $(u|v) = 0$. At $u \in L_{\mathtt{S}}$ betyder jo, at vi (med en lidt løs notation) har $u_{rsi} = u_s$. At $v \in L_{\mathtt{R}} \cap L_{\mathtt{1}}^{\perp}$ betyder tilsvarende at $v_{rsi} = v_r$, samt at v er vinkelret på det 1–dimensionale underrum $L_{\mathtt{1}}$ (= "diagonalen" i $\mathbf{R}^n$), hvilket åbenbart er det samme som at v har koordinatsum 0:

$$\sum_r \sum_s \sum_i v_{rsi} = 0,$$

eller (på grund af v' specielle form)

$$\sum_r \sum_s \sum_i v_r = 0.$$

Men netop på grund af balancertheden ses dette umiddelbart at være ækvivalent med

$$\sum_r v_r = 0.$$

Herefter får vi, igen ved at udnytte balancertheden,

$$(u|v) = \sum_r \sum_s \sum_i u_s v_r = \sum_s \sum_r \sum_i u_s v_r = \sum_s u_s \left(\sum_r \sum_i v_r \right) = 0.$$

Af dette resultat følger at middelværdiunderrummet for den additive model splitter op som en direkte ortogonal sum på formen

$$L_{\mathtt{R+S}} = L_{\mathtt{R}} + L_{\mathtt{S}} = L_{\mathtt{1}} \oplus (L_{\mathtt{R}} \cap L_{\mathtt{1}}^{\perp}) \oplus (L_{\mathtt{S}} \cap L_{\mathtt{1}}^{\perp}).$$

Heraf følger at ortogonalprojektionen på dette underrum kan skrives

$$P_{\mathtt{R+S}} = (P_{\mathtt{R}} - P_{\mathtt{1}}) + (P_{\mathtt{S}} - P_{\mathtt{1}}) + P_{\mathtt{1}} = P_{\mathtt{R}} + P_{\mathtt{S}} - P_{\mathtt{1}}.$$

Fra den ensidede variansanalyse ved vi hvordan $P_{\mathtt{R}}$ og $P_{\mathtt{S}}$ transformerer en vektor koordinatvis, ved at erstatte den enkelte koordinat med gennemsnittet i den tilsvarende gruppe. Ved anvendelse af dette til koordinatvis udregning af $\hat{\mu} = P_{\mathtt{R+S}}y$ får man umiddelbart sætningens første påstand.

Den anden påstand kan (ved gentagne anvendelser af “Pytagoras’ sætning”) bevises sådan:

$$\begin{aligned}
\mathrm{SSD}_{\mathrm{res}}(\mathtt{R+S}) &= \|y - P_{\mathtt{R+S}}y\|^2 = \|y\|^2 - \|P_{\mathtt{R+S}}y\|^2 \\
&= \|y\|^2 - \left(\|P_{\mathtt{1}}y\|^2 + \|(P_{\mathtt{R}} - P_{\mathtt{1}})y\|^2 + \|(P_{\mathtt{S}} - P_{\mathtt{1}})y\|^2\right) \\
&= \|y\|^2 - \left(\|P_{\mathtt{1}}y\|^2 + \|P_{\mathtt{R}}y\|^2 - \|P_{\mathtt{1}}y\|^2 + \|P_{\mathtt{S}}y\|^2 - \|P_{\mathtt{1}}y\|^2\right) \\
&= \left(\|y\|^2 - \|P_{\mathtt{R}}y\|^2\right) + \left(\|y\|^2 - \|P_{\mathtt{S}}y\|^2\right) - \left(\|y\|^2 - \|P_{\mathtt{1}}y\|^2\right) \\
&= \mathrm{SSD}_{\mathrm{res}}(\mathtt{R}) + \mathrm{SSD}_{\mathrm{res}}(\mathtt{S}) - \mathrm{SSD}_{\mathrm{res}}(\mathtt{1}) \ .
\end{aligned}$$