

Den subtraktive model

“Den subtraktive model” er det navn jeg her har brugt for en lineær model, der beskriver observationerne i et kvadratisk tosidet skema som normalfordelte med middelværdier der er differenser mellem række- og søjleparametre, under den yderligere antagelse at række og søjleparametre er ens. Vi illustrerer med et gennemgående eksempel, som er lånt fra projekt 1, hvor vi blot (for at kunne opskrive de relevante matricer osv.) antager at der kun er én sæson hvor der spilles alle mod alle både på hjemme- og udebane mellem tre hold. De seks kampe antages at finde sted i følgende rækkefølge:

i	h	u
1	1	2
2	1	3
3	2	1
4	2	3
5	3	1
6	3	2

Her betegner i kampens nummer, h hjemmeholdet og u udeholdet. Responsen er

$y_i =$ differensen mellem hjemme- og udeholdets målscorer.

Vi kan naturligt opstille observationerne i et tosidet skema

udehold:		1	2	3
hj.hold:	1		y_1	y_2
	2	y_3		y_4
	3	y_5	y_6	

Model 1. Antag at y_i 'erne er uafhængige, normalfordelte med samme varians og middelværdier

$$E y_i = \alpha + \beta_h + \gamma_u ,$$

på skemaform

udehold:		1	2	3
hj.hold:	1		$\alpha + \beta_1 + \gamma_2$	$\alpha + \beta_1 + \gamma_3$
	2	$\alpha + \beta_2 + \gamma_1$		$\alpha + \beta_2 + \gamma_3$
	3	$\alpha + \beta_3 + \gamma_1$	$\alpha + \beta_3 + \gamma_2$	

Dette er en ganske almindelig additiv normalfordelingsmodel (modelformel $Y=1+H+U$) i det tosidede $H \times U$ -skema. Det eneste usædvanlige er, at der ikke er observationer i diagonalen. Modellen er overparametriseret på sædvanlig måde. En injektiv parametrisering får man ved at sætte $\beta_3 = \gamma_3 = 0$ (hvilket f.eks. er hvad ISUW eller SAS ville gøre), eller ved at udelade konstantleddet α og sætte $\gamma_3 = 0$ (hvilket er hvad ISUW gør hvis man udelader 1-tallet i modelformlen, og SAS hvis man specificerer at modellen skal være uden intercept). Fortolkning af parametrene: β_h er hold h 's styrke på hjemmebane, γ_u er hold u 's styrke på udebane **med modsat fortegn**. På matrixform ser parametriseringen af modellen sådan ud:

$$\begin{bmatrix} \alpha + \beta_1 + \gamma_2 \\ \alpha + \beta_1 + \gamma_3 \\ \alpha + \beta_2 + \gamma_1 \\ \alpha + \beta_2 + \gamma_3 \\ \alpha + \beta_3 + \gamma_1 \\ \alpha + \beta_3 + \gamma_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix}$$

Model 2 (Den subtraktive model). Denne model fremkommer som hypotesen $\gamma_h = -\beta_h$ i model 1. Med betegnelsen β_h for den fælles værdi af $-\gamma_h$ og β_h får vi så

$$E y_i = \alpha + \beta_h - \beta_u,$$

på skemaform

udehold:	1	2	3
hj.hold:	1	$\alpha + \beta_1 - \beta_2$	$\alpha + \beta_1 - \beta_3$
	2	$\alpha + \beta_2 - \beta_1$	$\alpha + \beta_2 - \beta_3$
	3	$\alpha + \beta_3 - \beta_1$	$\alpha + \beta_3 - \beta_2$

Dette er ikke en konstruktion, som er indbygget i sædvanlig modelformel-syntaks. Man er nødt til at konstruere modelmatricens søjler direkte, og benytte dem som forklarende variable i en multipel regression. Fortolkningen af parametrene er, at β_h er hold h 's styrke på såvel hjemme- som udebane. Parameteren α , som i model 1 var ubestemt og derfor kunne sættes lig med 0, får her en meget præcis betydning som en "hjemmebanefordelsparameter" (den forventede forskel mellem hjemmeholdets og udeholdets målscorer, når to lige gode hold spiller mod hinanden). Modellen er dog stadig overparametriseret, idet β 'erne kun er givet på nær en additiv konstant (kun differenser mellem β 'er indgår i midelværdien). På matrixform ser parametriseringen af modellen sådan

ud:

$$\begin{bmatrix} \alpha + \beta_1 - \beta_2 \\ \alpha + \beta_1 - \beta_3 \\ \alpha + \beta_2 - \beta_1 \\ \alpha + \beta_2 - \beta_3 \\ \alpha + \beta_3 - \beta_1 \\ \alpha + \beta_3 - \beta_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & -1 & 0 \\ 1 & 1 & 0 & -1 \\ 1 & -1 & 1 & 0 \\ 1 & 0 & 1 & -1 \\ 1 & -1 & 0 & 1 \\ 1 & 0 & -1 & 1 \end{bmatrix} \times \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

De tre sidste søjler d_1 , d_2 og d_3 i modelmatricen, som svarer til parametrene β_1 , β_2 og β_3 (og som hver for sig er differens mellem to af søjlerne i modelmatricen for model 2) ses at være givet som en slags “fortegnsgummier” på formen

$$d_{hi} = \begin{cases} 1 & \text{hvis hold } h \text{ er med i kamp } i \text{ som hjemmebanehold,} \\ -1 & \text{hvis hold } h \text{ er med i kamp } i \text{ som udebanehold,} \\ 0 & \text{hvis hold } h \text{ ikke er med i kamp } i. \end{cases}$$

Dette skulle forklare konstruktionen af vektorerne $\mathbf{d1}, \dots, \mathbf{d30}$ i indlæringsprogrammet READ.ISU, og den brug vi senere gør af disse vektorer som forklarende variable i lineære modeller. Samt i en del generaliserede lineære modeller, som vi skal se på senere.