Kapitel 10

Lineære normale modeller

Vi har indtil nu ved præcisering af statistiske modeller benyttet formuleringer af typen "$y_1, \ldots, y_n$ er observationer af stokastiske variable $Y_1, \ldots, Y_n$ med fordelinger givet ved … ". Det bliver for tungt i længden. Vi vil fremover tillade os simpelthen at sige at observationerne (y'erne) selv er normalfordelte, Poissonfordelte eller hvad de nu er. Begrebsmæssigt er det noget sjusk, fordi observationer jo er givne tal og derfor aldeles ikke er "fordelte" på nogen som helst måde. Men det er i hvert fald nemmere, og det skulle ikke føre til misforståelser, når bare man husker hvad det i virkeligheden dækker over. Samtidig afskaffer vi betegnelsen Y_i og skriver fra nu af $\mathrm{E}y_i$, $\mathrm{var}(y_i)$ osv. i stedet for $\mathrm{E}Y_i$, $\mathrm{var}(Y_i)$ osv. Kampen for at opretholde distinktionen mellem Y'er og y'er har vi allerede tildels tabt, fordi vi for størrelser som $\hat{\beta}$ og $\mathrm{SSD}_{\mathrm{res}}$ ikke har indført en notation, der gør det muligt at se om vi tænker på dem som beregnede størrelser (funktioner af y'erne) eller afledte stokastiske variable (funktioner af Y'erne). I det følgende bruges betegnelsen $Y_1, \ldots, Y_n$ for de "observationsgenererende stokastiske variable" kun når vi helt explicit taler om fordelinger af stokastiske variable, som er afledt af observationerne.

10.1. Den generelle lineære normalfordelingsmodel.

Den simple regressionsmodel og den ensidede variansanalysemodel (og dermed også den simple model for identisk fordelte normale variable) er medlemmer af en meget større klasse af modeller, som kaldes de lineære normalfordelingsmodeller. Fælles for disse modeller er, at observationerne (som vi altid vil betegne med y'er, oftest $y_1, \ldots, y_n$) er uafhængige, normalfordelte med samme varians og en middelværdi der er specificeret som en lineær funktion af et sæt af ukendte parametre, kaldet *middelværdiparametrene.* Disse modeller omfatter for eksempel

Multiple regressionsmodeller, som er regressionsmodeller med to eller flere uafhængige variable. For to uafhængige variable x og z for eksempel

$$\mu_i = \mathrm{E}y_i = \alpha + \beta_x x_i + \beta_z z_i.$$

I Toyota Hiace eksemplet (eksempel 8.1) kunne man for eksempel udvide modellen ved foruden bilens alder x_i at inddrage antal kørte km som z_i.

Flersidede variansanalysemodeller, som er modeller hvori der kun indgår faktorer i udtrykket for middelværdien. *Tosidet variansanalyse* bruger man for eksempel i tilfældet hvor observationerne naturligt kan arrangeres i et tosidet skema, svarende til at man har to inddelingskriterier eller faktorer. Hvis vi her med y_{rsi} betegner den i'te observation i den celle af tabellen, som er bestemt ved rækkenummeret r og søjlenummeret s, vil man her normalt starte med modellen

$$\mu_{rsi} = \mathrm{E}y_{rsi} = \beta_{rs}$$

som siger at observationer i samme celle har samme middelværdi. Det er strengt taget bare en ensidet variansanalysemodel, svarende til opdelingen i tabellens celler, eller, i en mere indforstået statistikerjargon, den ensidede variansanalyse hvor grupperne er bestemt ved *krydsklassifikationen* givet ved række– og søjlefaktorerne, eller — kortere — *produktet* af disse to faktorer. Denne model kalder man også for modellen *med vekselvirkning* eller *interaktion* (engelsk *interaction*), idet man ved modellen *uden vekselvirkning,* også kaldet den *additive model,* forstår den simplere model hvor middelværdien er sum af *hovedvirkninger,*

$$\mu_{rsi} = \alpha_r + \beta_s.$$

Generelt specificerer man en lineær normalfordelingsmodel ved et udtryk for middelværdien som en sum af et antal led, der hver for sig kan antage en af følgende former:

β_f, effekten af en faktor med niveauer $f = 1, \ldots, F$.

β_{fg}, effekten af to faktorer med niveauer $f = 1, \ldots, F$ og $g = 1, \ldots, G$. Man taler her om *vekselvirkning af første orden* eller *to–faktor vekselvirkning*, for at præcisere modsætningen til en model hvor de to faktorer indgår additivt med hvert sit led α_f og β_g.

β_{fgh}, effekten af tre faktorer med *anden ordens* eller *tre–faktor vekselvirkning.*

Læseren kan selv fortsætte denne opremsning. Vekselvirkninger kan i princippet være af vilkårlig høj orden. Men vekselvirkninger af højere orden bliver hurtigt uhåndterlige i praksis, fordi de repræsenterer så mange parametre. Antallet af parametre af formen $\beta_{f_1 f_2 \ldots f_k}$ er jo produktet $F_1 F_2 \ldots F_k$ af antallene af niveauer for de faktorer der indgår.

βx_i, lineær effekt af en regressionsvariabel x.

Desuden forekommer "blandede" led af formen

$\beta_f x_i$, svarende til lineære effekter af regressionsvariable hvor hældningen afhænger af niveauet af en faktor. Undertiden bruger man også betegnelsen vekselvirkninger om sådanne led, idet man så udvider begrebet vekselvirkning til at omfatte vekselvirkninger af en faktor med en regressionsvariabel. Her kan faktoren igen være givet som et et produkt af flere faktorer, f.eks.

$\beta_{fg} x_i$ osv.

Terminologien i forbindelse med disse modeller er uoverskuelig og ikke altid lige logisk. En model hvor der indgår mange regressionsvariable, men ingen eller kun få faktorer, vil man oftest kalde en multipel regressionsmodel. Hvis der kun er faktorer kalder man den en variansanalysemodel. Hvis man i en variansanalysemodel introducerer én eller nogle få regressionsvariable taler man undertiden om *kovariansanalyse.*

Matematisk formuleres den generelle lineære model ved hjælp af en $n \times p$ matrix

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

som kaldes *modelmatricen* eller (især i variansanalysemodeller, hvor faktorniveauerne varieres efter en på forhånd fastlagt forsøgsplan) *designmatricen.* Den ved X bestemte lineære normalfordelingsmodel er en model for n uafhængige normalfordelte observationer med samme varians, og med middelværdier givet ved

$$\mu_i = \mathrm{E}y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}.$$

Modellen har altså, foruden den ukendte varians σ^2, p middelværdiparametre $\beta_1, \dots, \beta_p$.

På matrixform kan disse udtryk for middelværdien sammenfattes på formen

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$$

eller, med betegnelsen $\mu = (\mu_1, \dots, \mu_n) \in \mathbf{R}^n$ for *middelværdivektoren* og betegnelsen $\beta = (\beta_1, \dots, \beta_p) \in \mathbf{R}^p$ for *parametervektoren,*

$$\mu = X\beta.$$

Bemærk, at vi her følger den konvention at alle vektorer i observationsrummet $\mathbf{R}^n$ og middelværdiparameterrummet $\mathbf{R}^p$ opfattes som søjlevektorer, dvs. som $n \times 1$ eller $p \times 1$ matricer — også selv om vi skriver dem på sædvanlig måde som “vandrette” vektorer i teksten.

Eksempler. Modelmatricen for en simpel regression består af to søjler, med lutter ettaller i den første og den uafhængige variables værdier i den anden:

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} \alpha + \beta x_1 \\ \alpha + \beta x_2 \\ \vdots \\ \alpha + \beta x_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}.$$

Tilsvarende kan middelværdivektoren i en model med to regressionsvariable skrives

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} \alpha + \beta_x x_1 + \beta_z z_1 \\ \alpha + \beta_x x_2 + \beta_z z_2 \\ \vdots \\ \alpha + \beta_x x_n + \beta_z z_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & z_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_x \\ \beta_z \end{bmatrix}.$$

Modelmatricen for en ensidet variansanalysemodel består af lige så mange søjler som der er grupper, altså $p = G$. Den i'te værdi i den g'te søjle har værdien 1 hvis observation nr. i tilhører gruppe g, 0 ellers. Søjlerne er altså indikatorer for gruppetilhørsforhold, også kaldet "dummy'er". For eksempel kan middelværdivektoren i en ensidet variansanalysemodel for 6 observationer fordelt på 3 grupper af størrelse 2 (idet vi antager at observationerne i gruppe 1 kommer først osv.) skrives

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_1 \\ \beta_2 \\ \beta_2 \\ \beta_3 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

(idet vi her har valgt at kalde middelværdien i gruppe g for β_g i stedet for μ_g, for at undgå forveksling af middelværdivektoren $\mu \in \mathbf{R}^6$ med parametervektoren $\beta \in \mathbf{R}^3$).

Bemærk at det betyder, at en ensidet variansanalysemodel med G grupper kan beskrives som en multipel regressionsmodel med G regressionsvariable, hvor de G uafhængige variable kun antager værdierne 0 og 1, og hvor der for hver observation er netop én af dem som antager værdien 1.

Mere generelt kan ethvert led af formen β_g skrives som

$$\beta_g = \beta_1 x_{1i} + \cdots + \beta_G x_{Gi},$$

hvor

$$x_{gi} = \begin{cases} 1 \text{ hvis observation } i \text{ tilhører gruppe } g, \\ 0 \text{ ellers.} \end{cases}$$

På denne måde kan man rent formelt skrive en hvilken som helst lineær normalfordelingsmodel som en multipel regressionsmodel.

Som et sidste og lidt mere avanceret eksempel kan vi tage en model der involverer et "blandet" led af typen $\beta_f x_i$. Antag at vi har seks observationer $y_1, \ldots, y_6$. I modellen indgår en regressionsvariabel $x_1, \ldots, x_6$ og en faktor på to niveauer, der deler observationerne i de første to og de sidste fire observationer. Betragt modellen "to regressionslinier" givet ved

$$\mu_i = \alpha_f + \beta_f x_i$$

På matrixform kan denne model skrives

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \end{bmatrix} = \begin{bmatrix} \alpha_1 + \beta_1 x_1 \\ \alpha_1 + \beta_1 x_2 \\ \alpha_2 + \beta_2 x_3 \\ \alpha_2 + \beta_2 x_4 \\ \alpha_2 + \beta_2 x_5 \\ \alpha_2 + \beta_2 x_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & x_1 & 0 \\ 1 & 0 & x_2 & 0 \\ 0 & 1 & 0 & x_3 \\ 0 & 1 & 0 & x_4 \\ 0 & 1 & 0 & x_5 \\ 0 & 1 & 0 & x_6 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

Modellen kan kort beskrives ved at observationerne i de to grupper følger hver sin simple regressionsmodel, bortset fra at variansen er fælles. Som en umiddelbart interessant delmodel har den modellen "to parallelle regressionslinier" givet ved $\beta_1 = \beta_2 = \beta$, altså

$$\mu_i = \alpha_f + \beta x_i$$

som på matrixform kan skrives

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \end{bmatrix} = \begin{bmatrix} \alpha_1 + \beta x_1 \\ \alpha_1 + \beta x_2 \\ \alpha_2 + \beta x_3 \\ \alpha_2 + \beta x_4 \\ \alpha_2 + \beta x_5 \\ \alpha_2 + \beta x_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & x_1 \\ 1 & 0 & x_2 \\ 0 & 1 & x_3 \\ 0 & 1 & x_4 \\ 0 & 1 & x_5 \\ 0 & 1 & x_6 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{bmatrix}$$

10.2. Normalligningerne.

Likelihoodfunktionen for den lineære model bestemt ved $n \times p$ modelmatricen X for observationer $y_1, \ldots, y_n$ kan skrives

$$\begin{aligned} L(\beta_1, \ldots, \beta_p, \sigma^2) &= \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) \exp\left(-\frac{1}{2\sigma^2} (y_i - \mu_i)^2 \right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu_i)^2 \right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left(-\frac{1}{2\sigma^2} \|y - \mu\|^2 \right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left(-\frac{1}{2\sigma^2} \|y - X\beta\|^2 \right), \end{aligned}$$

og log–likelihooden bliver således, på nær en additiv konstant,

$$-\frac{1}{2} \left(n \log \sigma^2 + \frac{1}{\sigma^2} \|y - X\beta\|^2 \right).$$

For fast σ^2 er maksimerering af likelihooden således ækvivalent med minimering af $\|y - X\beta\|^2$, som er kvadratet på afstanden fra observationsvektoren y til middelværdivektoren $X\beta$. Lad $L \subseteq \mathbf{R}^n$ betegne underrummet udspændt af søjlerne i modelmatricen X. Når β gennemløber middelværdiparameterrummet $\mathbf{R}^p$ vil $\mu = X\beta$ gennemløbe L. Modellen er således (idet vi ser bort fra valget af parametrisering, og kun interesserer os for hvilke fordelinger af observationssættet der er mulige under modellen) fuldstændigt beskrevet ved at middelværdivektoren μ tilhører underrummet L, som kaldes modellens *middelværdiunderrum.*

Maksimaliseringsestimatoren for middelværdivektoren μ er hermed karakteriseret som den vektor $\hat{\mu} \in L$ der (i sædvanlig Euklidisk afstand på $\mathbf{R}^n$) ligger nærmest ved observationsvektoren y. Det betyder naturligvis, hvis man tør stole på sin geometriske intuition når det gælder n–dimensionale rum, at $\hat{\mu}$ er *ortogonalprojektionen* af y på L.

Mere præcist gælder der følgende. Lad

$$P : \mathbf{R}^n \to \mathbf{R}^n$$

betegne den lineære afbildning, der til en vilkårlig vektor $y \in \mathbf{R}^n$ knytter dens ortogonalprojektion på L. For en given vektor y kan Py entydigt karakteriseres som den vektor i L der opfylder

$$y - Py \in L^{\perp}.$$

Ifølge en elementær udregning (eller Pytagoras' sætning, om man vil) har vi nu for $\mu \in L$

$$\|y - \mu\|^2 = \|y - Py\|^2 + \|Py - \mu\|^2,$$

hvoraf umiddelbart følger, at denne størrelse antager sit minimum for $\mu = Py$.

Matematisk set har vi hermed fundet estimatoren for middelværdivektoren μ. Beregningsmæssigt kan man ikke sige at vi har løst problemet, da løsningen involverer det abstrakte begreb "ortogonalprojektion". Men karakteriseringen af ortogonalprojektionen sætter os umiddelbart i stand til at opskrive de ligninger, der bestemmer maksimaliseringsestimatorerne for de oprindelige parametre $\beta_1, \ldots, \beta_p$. Ifølge ovenstående karakterisering af Py er løsningen (eller løsningerne) givet ved

$$(y - X\beta | u) = 0 \text{ for alle } u \in L,$$

eller

$$(y|u) = (X\beta|u) \text{ for alle } u \in L.$$

På matrixform kan denne betingelse skrives

$$u'y = u'X\beta \text{ for alle } u \in L,$$

og da denne ligning åbenbart er opfyldt for alle $u \in L$ hvis og kun hvis den er opfyldt for alle u i en mængde der udspænder L — f.eks. søjlerne i X — kan betingelsen også skrives

$$X'y = X'X\beta.$$

Dette lineære ligningssystem kaldes *normalligningerne.* Et sæt β af middelværdiparametre opfylder disse ligninger hvis og kun hvis likelihood'en i dette punkt (for fast σ^2) antager sin maksimale værdi.

Antag nu at parametriseringen af middelværdivektoren μ ved β er injektiv. Det betyder at den lineære afbildning $X : \mathbf{R}^p \to \mathbf{R}^n$ er injektiv, dvs. at matricen X har lineært uafhængige søjler. I så fald er $p \times p$–matricen $X'X$ regulær (faktisk positivt definit), hvoraf umiddelbart følger at normalligningerne har den entydige løsning

$$\hat{\beta} = (X'X)^{-1} X'y.$$

Hvis parametriseringen ikke er injektiv bliver løsningsmængden et underrum af $\mathbf{R}^p$. Men som vi senere skal se kan man altid reducere til tilfældet hvor X har lineært uafhængige søjler ved at fjerne nogle af søjlerne (svarende til at visse af de oprindelige parametre sættes til 0), og da det er sådan man håndterer problemet i praksis, behøver vi ikke her at interessere os mere for dette tilfælde.

Beregningsmæssigt har vi hermed reduceret problemet til noget der vedrører matrixregning, herunder inversion af $p \times p$–matricen $X'X$. Altså ikke noget man normalt vil kaste sig ud i at gøre i hånden for $p > 2$, men dog noget som er programmeringsmæssigt overkommeligt.

Bemærk at vi hermed også har vist at

$$Py = \hat{\mu} = X\hat{\beta} = X\,(X'X)^{-1} X'y,$$

hvoraf følger matrixrelationen

$$P = X\,(X'X)^{-1} X'$$

som viser hvordan man beregner matricen P for ortogonalprojektionen på underrummet udspændt af en given basis (søjlerne i X).

Om fordelingen af maksimaliseringsestimatoren gælder nu følgende:

SÆTNING 10.1. *Under antagelse af at X har lineært uafhængige søjler gælder at maksimaliseringsestimatoren*

$$\hat{\beta} = (X'X)^{-1} X'Y$$

er p–dimensionalt normalfordelt med middelværdi

$$\mathrm{E}\hat{\beta} = \beta$$

og kovariansmatrix

$$\operatorname{cov}(\hat\beta) = \sigma^2 \left(X'X\right)^{-1}.$$

Når vi her skriver $\mathrm{E}\hat\beta = \beta$ skal det naturligvis forstås sådan at vi tager middelværdien plads for plads, altså

$$\mathrm{E}\hat\beta_j = \beta_j \text{ for } j = 1, \ldots, p.$$

Når man på denne måde definerer middelværdier af stokastiske vektorer eller matricer Z ved at tage middelværdi plads for plads, gælder regneregler som

$$\mathrm{E}(AZ) = A(\mathrm{E}Z) \qquad \text{og} \qquad \mathrm{E}(ZA) = (\mathrm{E}Z)A,$$

der tillader ombytning af matrixmultiplikation med middelværdidannelse. Disse regler — som umiddelbart følger af matrixproduktets definition og simple regneregler for middelværdier — bruges flittigt i beviset for sætningen.

Bevis. Det er klart at $\hat\beta$ er normalfordelt, da $\hat\beta$ jo er fremkommet ved lineær transformation af observationsvektoren Y. Middelværdien bliver

$$\mathrm{E}\hat\beta = \mathrm{E}\big((X'X)^{-1}X'Y\big) = (X'X)^{-1}X'\mathrm{E}Y = (X'X)^{-1}X'X\beta = \beta,$$

og kovariansmatricen

$$\begin{aligned}
\operatorname{cov}(\hat\beta) &= \mathrm{E}\big((\hat\beta-\beta)(\hat\beta-\beta)'\big) \\
&= \mathrm{E}\big(((X'X)^{-1}X'(Y-\mu))((X'X)^{-1}X'(Y-\mu))'\big) \\
&= \mathrm{E}\big((X'X)^{-1}X'(Y-\mu)(Y-\mu)'X(X'X)'^{-1}\big) \\
&= (X'X)^{-1}X'\mathrm{E}\big((Y-\mu)(Y-\mu)'\big)X(X'X)^{-1} \\
&= (X'X)^{-1}X'(\sigma^2 I)X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}.
\end{aligned}$$

Af sætningen følger specielt at estimaterne for middelværdiparametrene er centrale, og at variansen på maksimaliseringsestimatet $\hat\beta_j$ for den j' te parameter β_j kan udregnes som σ^2 ganget med det j'te diagonalelement i matricen $(X'X)^{-1}$.

Opgave 10.2.1. I en simpel regressionsmodel er $X'X$ en 2×2–matrix, som er nem at regne ud. Benyt dette — samt en velkendt formel for den inverse til en 2×2–matrix — til at udlede formlerne for maksimaliseringsestimatorerne $\hat\alpha$ og $\hat\beta$ fra kapitel 8 og disses varianser.

Opgave 10.2.2. I en ensidet variansanalysemodel bliver $X'X$ en diagonalmatrix, der som bekendt er nem at invertere. Benyt dette til at udlede formlen for maksimaliseringsestimatorerne $\hat\mu_g$ fra kapitel 9 og disses varianser.

10.3. Estimation af variansen.

Med $\mathrm{SSD}_{\mathrm{res}}$ betegner vi som sædvanlig kvadratsummen af observationernes afvigelser fra deres estimerede middelværdier, altså

$$\mathrm{SSD}_{\mathrm{res}} = \sum (y_i - \hat{\mu}_i)^2 = \|y - \hat{\mu}\|^2 = \|y - Py\|^2.$$

Den delvis maksimerede log–likelihood har som sædvanlig udseendet

$$l(\hat{\beta}_1, \ldots, \hat{\beta}_p, \sigma^2) = -\frac{1}{2}\left(n \log \sigma^2 + \frac{1}{\sigma^2}\mathrm{SSD}_{\mathrm{res}}\right),$$

og antager sin maksimale værdi for

$$\sigma^2 = \sigma^2_{ML} = \frac{\mathrm{SSD}_{\mathrm{res}}}{n}.$$

Som skøn for variansen benytter man imidlertid det korrigerede estimat

$$\hat{\sigma}^2 = \frac{\mathrm{SSD}_{\mathrm{res}}}{n-p}.$$

Vi er nu endelig i den lykkelige situation, at vi kan give et generelt bevis for det resultat, der som konsekvens har at denne estimator er central, samt en række andre påstande som vi har henvist til under gennemgangen af den simple regressionsmodel og den ensidede variansanalysemodel:

Sætning 10.2. *Residualkvadratsummen* $\mathrm{SSD}_{\mathrm{res}}$ *er* χ^2*–fordelt med* $n-p$ *frihedsgrader og skalaparameter* σ^2*. Desuden er residualvektoren* $Y - PY = Y - \hat{\mu}$ *(og dermed også* $\mathrm{SSD}_{\mathrm{res}} = \|Y - PY\|^2$*) stokastisk uafhængig af vektoren af fittede værdier* $\hat{\mu} = PY$ *(og dermed af* $\hat{\beta}$*).*

Bemærk at når vi taler om $\hat{\beta}$ som en veldefineret størrelse (her: stokastisk variabel) er det underforstået at parametriseringen er injektiv, altså at X har lineært uafhængige søjler. Resten af sætningen gælder sådan set uden denne antagelse (idet p så skal betegne antallet af parametre i en injektiv parametrisering, eller dimensionen af L).

Bevis. Lad $e_1, \ldots, e_n$ være en ortonormal basis for $\mathbf{R}^n$, som er valgt på en sådan måde at de første p enhedsvektorer $e_1, \ldots, e_p$ netop udspænder middelværdiunderrummet L. Det betyder at $e_1, \ldots, e_p$ er en ortogonalbasis for L (for eksempel konstrueret ved Gram–Schmidt ortogonalisering af søjlerne i X), medens $e_{p+1}, \ldots, e_n$ er en ortogonalbasis for $L^{\perp}$. Sæt

$$U_i = \frac{Y_i - \mu_i}{\sigma}.$$

Så er $U_1, \ldots, U_n$ stokastisk uafhængige, normeret normalfordelte. Lad U betegne den n–dimensionale stokastiske variabel $(U_1, \ldots, U_n)$. Af sætning 7.2.1 i Ssr. følger nu, at de stokastiske variable $V_1, \ldots, V_n$ givet ved

$$V_i = (e_i|U)$$

igen er uafhængige, normeret normalfordelte.

Nu er imidlertid

$$PY = P(\mu + \sigma U) = P(\mu + \sigma(V_1 e_1 + \cdots + V_n e_n))$$

$$= P(\mu) + \sigma P(V_1 e_1 + \cdots + V_n e_n) = \mu + \sigma(V_1 e_1 + \cdots + V_p e_p),$$

og dermed

$$Y - PY = (\mu + \sigma(V_1 e_1 + \cdots + V_n e_n)) - (\mu + \sigma(V_1 e_1 + \cdots + V_p e_p))$$

$$= \sigma(V_{p+1} e_{p+1} + \cdots + V_n e_n).$$

Heraf følger, da kvadratet på den Euklidiske norm af en vektor jo kan udregnes som kvadratsummen af vektorens koordinater m.h.t. en hvilken som helst ortonormal basis, at

$$\mathrm{SSD}_{\mathrm{res}} = \|Y - PY\|^2 = \sigma^2(V_{p+1}^2 + \cdots + V_n^2),$$

hvoraf sætningens første påstand umiddelbart følger (jvf. χ^2–fordelingens definition Ssr. side 125–26). Den anden påstand følger af at $\hat{\mu} = PY$ (og dermed også $\hat{\beta}_1, \ldots, \hat{\beta}_p$) kan skrives som funktion af $V_1, \ldots, V_p$, medens residualvektoren $Y - PY$ (og dermed $\mathrm{SSD}_{\mathrm{res}}$) er en funktion af $V_{p+1}, \ldots, V_n$.

10.4. F–test for modelreduktion.

Betragt nu, ud over modellen bestemt ved $n \times p$–matricen X, en reduceret model bestemt ved en $n \times p_0$–matrix X_0, $p_0 < p$. Lad L og L_0 betegne de to modellers middelværdiunderrum, som er af dimensioner p og p_0 (idet vi stadig antager at parametriseringerne er injektive, altså at X og X_0 har lineært uafhængige søjler). Vi antager at $L_0 \subset L$, dvs. at modellen bestemt ved X_0 er en delmodel af modellen bestemt ved X. Med $\mathrm{SSD}_{\mathrm{res}}$ og $\mathrm{SSD}_{\mathrm{res}}^0$ betegnes residualkvadratsummerne i de to modeller. Om kvotienttestet for den tilsvarende modelreduktion har vi nu følgende resultat:

SÆTNING 10.3. *Kvotientteststørrelsen for hypotesen $\mu \in L_0$ i den ved X bestemte model er en monotont voksende funktion af*

$$F = \frac{(\mathrm{SSD}_{\mathrm{res}}^0 - \mathrm{SSD}_{\mathrm{res}})/(p - p_0)}{\mathrm{SSD}_{\mathrm{res}}/(n - p)}$$

som under hypotesen er F*–fordelt med $(p - p_0, n - p)$ frihedsgrader.*

BEVIS. Lad $e_1, \ldots, e_n$ være en ortonormal basis for $\mathbf{R}^n$ som er valgt på en sådan måde at de første p_0 basisvektorer udspænder L_0 og de første p basisvektorer udspænder L. En sådan basis kan umiddelbart konstrueres ved at vi først vælger $e_1, \ldots, e_{p_0}$ som en ortonormal basis for L_0, dernæst $e_{p_0+1}, \ldots, e_p$ som en ortonormal basis for $L \cap L_0^\perp$, og endelig $e_{p+1}, \ldots, e_n$ som en ortonormal basis for $L^\perp$. Som i beviset for forrige sætning definerer vi uafhængige, normeret normalfordelte stokastiske variable

$$U_i = \frac{Y_i - \mu_i}{\sigma}$$

og får ifølge sætning 7.2.1 i Ssr. at de stokastiske variable $V_1, \ldots, V_n$ defineret ved

$$V_i = (e_i|U)$$

ligeledes er uafhængige, normeret normalfordelte. Så er

$$\mathrm{SSD_{res}} = \sigma^2(V_{p+1}^2 + \cdots + V_n^2),$$

og tilsvarende for den reducerede model

$$\mathrm{SSD_{res}^0} = \sigma^2(V_{p_0+1}^2 + \cdots + V_n^2),$$

Log–likelihoodens maksimale værdi under den store model er

$$-\frac{1}{2}\left(n \log \frac{\mathrm{SSD_{res}}}{n} + n\right)$$

og under den reducerede model tilsvarende

$$-\frac{1}{2}\left(n \log \frac{\mathrm{SSD_{res}^0}}{n} + n\right)$$

Vi får således

$$\begin{aligned} -2\log Q &= 2\left(\frac{1}{2}\left(n \log \frac{\mathrm{SSD_{res}^0}}{n} + n\right) - \frac{1}{2}\left(n \log \frac{\mathrm{SSD_{res}}}{n} + n\right)\right) \\ &= n \log \frac{\mathrm{SSD_{res}^0}}{\mathrm{SSD_{res}}} = n \log\left(1 + \frac{\mathrm{SSD_{res}^0} - \mathrm{SSD_{res}}}{\mathrm{SSD_{res}}}\right). \end{aligned}$$

Heraf følger umiddelbart sætningens første påstand, og den sidste følger af omskrivningen

$$F = \frac{(V_{p_0+1}^2 + \cdots + V_p^2)/(p - p_0)}{(V_{p+1}^2 + \cdots + V_n^2)/(n-p)}$$

i forbindelse med definitionen af F–fordelingen (Ssr. side 127).

10.5. T–test for modelreduktion.

I tilfældet $p = p_0 + 1$ kan hypotesen $\mu \in L_0$ altid fortolkes som en hypotese om at en eller anden linearkombination

$$\vartheta = a_1\beta_1 + \cdots + a_p\beta_p$$

af middelværdiparametrene er lig med 0. Hvis X har lineært uafhængige søjler vil koefficienterne $a_1, \ldots, a_p$ være entydigt bestemt ved X, L og L_0 på nær proportionalitet. Dette følger for eksempel af at to ikke–proportionale betingelser af formen $a_1\beta_1 + \cdots + a_p\beta_p = 0$ bestemmer et $p-2$–dimensionalt underrum af $\mathbf{R}^p$, som ved den injektive parametrisering af middelværdivektoren ville blive afbildet over i et $p-2$–dimensionalt underrum af L.

Ved at omparametrisere modellen på formen

$$\mu = \vartheta_1 e_1 + \cdots + \vartheta_p e_p$$

med nye parametre $\vartheta_1, \ldots, \vartheta_p$, svarende til en modelmatrix der som søjler har de første p enhedsvektorer i en ortonormalbasis $e_1, \ldots, e_n$ der er valgt som i beviset for sætning 10.3, får vi en parametrisering hvor hypotesen $\mu \in L_0$ er ækvivalent med $\vartheta_p = 0$. Maksimaliseringsestimatoren for ϑ_p er under denne parametrisering simpelthen

$$\hat{\vartheta}_p = (Y|e_p) = \vartheta_p + \sigma V_p$$

(= p'te koordinat af ortogonalprojektionen af Y på middelværdiunderrummet, når $e_1, \ldots, e_n$ bruges som basis), og variansen på dette estimat er

$$\mathrm{var}(\hat{\vartheta}_p) = \sigma^2.$$

Den til F–teststørrelsen svarende T–størrelse (se Ssr. side 127) kan derfor skrives som

$$\frac{V_p}{\sqrt{(V_{p+1}^2 + \cdots + V_n^2)/(n-p)}} = \frac{\hat{\vartheta}_p - \vartheta_p}{\hat{\sigma}}$$

$$= \frac{\text{ML–estimatet for den parameter som ifølge hypotesen er 0}}{\text{den estimerede standardafvigelse for denne}} \; .$$

Men på grund af ovenstående bemærkninger om entydighed af den linearkombination, der ifølge hypotesen er lig med 0, holder den sidste fortolkning også i den oprindelige parametrisering. Parameteren ϑ_p er jo nødvendigvis, som funktion af $\beta_1, \ldots, \beta_p$, netop denne linearkombination. Det betyder at kvotienttestet for en hypotese af formen $\vartheta = 0$, hvor ϑ er en af parametrene $\beta_1, \ldots, \beta_p$ eller en linearkombination af disse, altid kan foretages som et T–test, hvor estimatet for ϑ divideret med sin estimerede standardafvigelse vurderes tosidet i en T–fordeling med $n-p$ frihedsgrader.

10.6. Konventioner i forbindelse med overparametrisering.

Antag at vi har data i form af en tosidet tabel med 2 rækker og 3 søjler, hvor hver celle indholder én observation y_{rs}, $r = 1, 2$, $s = 1, 2, 3$. Den additive model kan her skrives

$$\begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \\ \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{bmatrix} = \begin{bmatrix} \alpha_1 + \beta_1 \\ \alpha_1 + \beta_2 \\ \alpha_1 + \beta_3 \\ \alpha_2 + \beta_1 \\ \alpha_2 + \beta_2 \\ \alpha_2 + \beta_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

Denne parametrisering er ikke injektiv. Modelmatricens fem søjler er lineært afhængige, hvilket for eksempel følger af at summen af de første to er lig med summen af de sidste tre. Man ser da også umiddelbart at man kan ændre på de fem parametre uden at ændre på middelværdivektoren, ved at addere et tal til de to rækkeparametre og trække det samme tal fra de tre søjleparametre.

Man kalder en parameter eller parameterfunktion *identificerbar*, hvis den er entydigt bestemt ved den tilhørende sandsynlighedsfordeling. I dette tilfælde er ingen af de fem parametre α_1, α_2, β_1, β_2 og β_3 identificerbare. Differensen mellem de to rækkeparametre, eller enhver af de tre mulige differenser mellem søjleparametre, er derimod identificerbare. For eksempel er jo $\alpha_1 - \alpha_2 = \mu_{11} - \mu_{21}$.

Generelt er det ikke muligt at opskrive udtrykket for middelværdien i en model hvor to eller flere faktorer indgår additivt på en naturlig måde, således at parametriseringen er injektiv. Dette er en af begrundelserne for at man i praksis er nødt til at arbejde med modelmatricer, der ikke har lineært uafhængige søjler.

En anden begrundelse for dette er, at man ofte — af andre grunde, som har noget at gøre med hvilke estimater man vil have udregnet og hvilke tests man vil have udført af en statistisk programpakke — med vilje opskriver modellerne med overflødige led. En ensidet variansanalysemodel opskrives for eksempel normalt (når man snakker teori) på formen

$$\mu_{gi} = \beta_g,$$

men der er ikke noget i vejen for tilføje et overflødigt konstantled og i stedet skrive den på formen

$$\mu_{gi} = \gamma + \beta_g.$$

Det svarer til at tilføje en søjle bestående af lutter ettaller som den første søjle i modelmatricen. Herved bliver søjlerne naturligvis lineært afhængige, fordi den tilføjede søjle er summen af dem der stod der i forvejen. Formålet med at gøre det kan være at tvinge den statistiske programpakke man bruger til at udføre testet for homogenitet. En standard facilitet, som mange programpakker har, går nemlig ud på at man kan få udført alle de tests, der svarer til fjernelse af leddene i formlen for middelværdien ét for ét, startende med det sidste.

Lad os tage et mere kompliceret eksempel: En tosidet variansanalysemodel med vekselvirkninger vil man normalt skrive

$$\mu_{rsi} = \mathrm{E}y_{rsi} = \delta_{rs}.$$

Men ved i stedet for at skrive den på formen

$$\mu_{rsi} = \gamma + \alpha_r + \beta_s + \delta_{rs}$$

opnår man, at der blandt de tests for modelreduktioner som svarer til fjernelse af led bagfra optræder følgende:

Test for additivitet: Reduktionen fra

$$\mu_{rsi} = \gamma + \alpha_r + \beta_s + \delta_{rs} \qquad \text{til} \qquad \mu_{rsi} = \gamma + \alpha_r + \beta_s$$

Test for forsvindende søjlevirkning i den additive model: Reduktionen fra

$$\mu_{rsi} = \gamma + \alpha_r + \beta_s \qquad \text{til} \qquad \mu_{rsi} = \gamma + \alpha_r$$

Test for homogenitet i den ensidede variansanalysemodel svarende til opdelingen efter rækker: Reduktionen fra

$$\mu_{rsi} = \gamma + \alpha_r \qquad \text{til} \qquad \mu_{rsi} = \gamma$$

Bemærk at vi ikke får udført alle de tests som kunne være relevante. De to der mangler kunne vi få udført ved at bytte om på hovedvirkningsleddene α_r og β_s. Når man specificerer en model i en statistisk programpakke er det derfor normalt en god idé at skrive de led man først vil forsøge at teste væk som de sidste.

Af beregningsmæssige og præsentationsmæssige grunde er det imidlertid nødvendigt at ændre parametriseringen af modellen så den bliver injektiv. Den simpleste metode hertil — som også, med visse modifikationer, er den der benyttes af de fleste statistiske programpakker — består i at udtynde modelmatricens søjler forfra, så dem der bliver tilbage er lineært uafhængige. Princippet er altså at man gennemgår søjlerne forfra én for én. Hver gang man støder på en, der kan skrives som linearkombination af de foregående, fjerner man den. De tilsvarende parametre udgår, eller sættes lig med 0 om man vil. Det er klart at man på denne måde ender med et sæt af søjler, som er lineært uafhængige og udspænder det samme underrum som de oprindelige — dvs. definerer den samme statistiske model, bare i en anden parametrisering.

EKSEMPLER. En ensidet variansanalysemodel med 6 observationer delt i 3 grupper af størrelse 2 kan, når vi tager et konstantled med, skrives på formen

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \end{bmatrix} = \begin{bmatrix} \gamma+\beta_1 \\ \gamma+\beta_1 \\ \gamma+\beta_2 \\ \gamma+\beta_2 \\ \gamma+\beta_3 \\ \gamma+\beta_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \gamma \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

Udtynding af modelmatricen til lineær uafhængighed fører åbenbart til at den sidste søjle slettes, så vi får den injektive parametrisering givet ved

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \end{bmatrix} = \begin{bmatrix} \gamma+\beta_1 \\ \gamma+\beta_1 \\ \gamma+\beta_2 \\ \gamma+\beta_2 \\ \gamma \\ \gamma \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \gamma \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

Denne parametrisering er ikke specielt naturlig, hvilket betyder at man skal holde hovedet koldt når man læser og fortolker de parameterestimater, der kommer ud af statistiske programpakker. Bemærk at fortolkningen af "konstantleddet" γ er, at det er middelværdien i gruppe 3 (!), medens β_1 og β_2 er differenser mellem middelværdierne i grupperne 1 og 2 og middelværdien i gruppe 3.

I en tosidet model svarende til en tabel med 2 rækker og 3 søjler og 2 observationer pr. celle bliver middelværdivektoren i modellen med vekselvirkning, opskrevet med konstantled og begge hovedvirkninger,

$$\begin{bmatrix} \mu_{111} \\ \mu_{112} \\ \mu_{121} \\ \mu_{122} \\ \mu_{131} \\ \mu_{132} \\ \mu_{211} \\ \mu_{212} \\ \mu_{221} \\ \mu_{222} \\ \mu_{231} \\ \mu_{232} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \gamma \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \delta_{11} \\ \delta_{12} \\ \delta_{13} \\ \delta_{21} \\ \delta_{22} \\ \delta_{23} \end{bmatrix}$$

Vi har her udbygget det tidligere eksempel på en tosidet model — en 2×3 tabel med 1 observation pr. celle — ved at fordoble antallet af observationer. Det er strengt taget ikke nødvendigt for at forklare de ting der vedrører parametriseringen. Men vekselvirkningsmodellen i en tabel med 1 observation pr. celle er lidt uinteressant derved at middelværdiunderrummet udgør hele observationsrummet, så man kan ikke estimere variansen (man får $\mathrm{SSD}_{\mathrm{res}} = 0$, $\hat{\mu} = y$ osv.). Man vil derfor normalt for en tosidet tabel med én observation per celle være tvunget til at vælge den additive model som sin grundmodel.

Udtynding af modelmatricen til lineær uafhængighed fører her til at søjlerne 3, 6, 9, 10, 11 og 12 fjernes, svarende til α_2, β_3, δ_{13}, δ_{21}, δ_{22} og δ_{23}. Med andre ord, alle parametre der har et "maksimalt indeks" sættes til 0. Herefter får vi den injektive parametrisering

$$\begin{bmatrix} \mu_{111} \\ \mu_{112} \\ \mu_{121} \\ \mu_{122} \\ \mu_{131} \\ \mu_{132} \\ \mu_{211} \\ \mu_{212} \\ \mu_{221} \\ \mu_{222} \\ \mu_{231} \\ \mu_{232} \end{bmatrix} = \begin{bmatrix} \gamma + \alpha_1 + \beta_1 + \delta_{11} \\ \gamma + \alpha_1 + \beta_1 + \delta_{11} \\ \gamma + \alpha_1 + \beta_2 + \delta_{12} \\ \gamma + \alpha_1 + \beta_2 + \delta_{12} \\ \gamma + \alpha_1 \\ \gamma + \alpha_1 \\ \gamma + \beta_1 \\ \gamma + \beta_1 \\ \gamma + \beta_2 \\ \gamma + \beta_2 \\ \gamma \\ \gamma \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \gamma \\ \alpha_1 \\ \beta_1 \\ \beta_2 \\ \delta_{11} \\ \delta_{12} \end{bmatrix}$$

Noget kompliceret, må man sige, i betragtning af at det i virkeligheden drejer sig om en ensidet variansanalysemodel med 6 grupper af størrelse 2. Men i praksis er der jo ikke noget i vejen for, at man estimerer denne model både ved at angive den fulde modelformel som her, for at få udført en række tests, og — hvis man ikke kan få den reduceret til noget simplere — på formen $\mu_{rsi} = \delta_{rs}$, for at få udskrevet parameterestimaterne på en mere læselig form.

10.7. Modelformler.

Vi kan ikke inden for rammerne af dette kursus introducere store statistiske programpakker som SAS og GENSTAT. Det kræver for lang tid bare at komme igang med dem. For ikke at lade læseren helt i stikken når det gælder statistiske beregninger i praksis vil vi i kapitel 12 og 13 gennemgå håndteringen af en række modeller i ISU (Interactive StatUnit), som er forfatterens egen, meget mindre og meget lettere tilgængelige, statistikpakke. Det vi vil sige her om den symbolske notation til specifikation af lineære modeller gælder imidlertid generelt for statistiske programpakker. Der er kun nogle nuanceforskelle. I SAS skal man for eksempel i hvert enkelt tilfælde præcisere hvilke variable der opfattes som faktorer ved et særligt `CLASS` direktiv. I GENSTAT er syntaksen i forbindelse med modelformler mere indviklet end antydet her, fordi der skelnes mellem `R.S` (som er en “ren vekselvirkning”) og `R*S` (som er ækvivalent med `R+S+R.S`). Hverken i SAS eller GENSTAT specificeres modellens konstantled (men man kan, ved en særlig indsats, undgå at få det med). Men disse små afvigelser er ikke svære at lære, og de er under alle omstændigheder umulige at lære uden at forstå de generelle principper som forklares her.

Når man skal fortælle en statistisk programpakke at man ønsker at få analyseret en tosidet model med vekselvirkning gør man det naturligvis ikke ved at skrive

$$\mathrm{E}y_i = \gamma + \alpha_r + \beta_s + \delta_{rs}.$$

Man gør det (med det nævnte forbehold vedrørende konstantleddet) ved at skrive

```
Y=1+R+S+R*S
```

Her forudsættes at `R` og `S` er navnene på de faktorer — dvs. vektorer eller tabeller hvis værdier er faktorniveauer — som for hver enkelt observation fortæller hvilken gruppe observationen hører til.

Generelt bruger man til specifikation af lineære modeller en symbolsk notation, hvor modellens led angives ved udtryk der er adskilt af plusser eller (i SAS) mellemrum. Normalt skriver man så responsen først, efterfulgt af et lighedstegn. Det enkelte led på højre side af lighedstegnet kan være

— Navnet på en faktor, svarende til et led af formen α_f.

— Navnet på en kvantitativ (regressions–)variabel af samme længde som responsvektoren, svarende til et led af formen βx_i.

— Navnene på flere faktorer adskilt med gangetegn ($*$), svarende til et vekselvirknings– eller produktled af typen α_{fg}, α_{fgh} etc.

— Navnet på en eller flere faktorer og en kvantitativ variabel, adskilt med gangetegn, svarende til led af formen $\beta_f x_i$, $\beta_{fg} x_i$ etc.

En naturlig forlængelse af denne notation går ud på, at hvis der i et produkt optræder flere regressionsvariable, så skal deres værdier simpelthen ganges sammen.

Den nemmeste måde at forstå disse regler på er måske via følgende beskrivelse af den algoritme, der genererer modelmatricen (før udtynding) ud fra modelformlen. Hvert led i modellen omsættes til et antal søjler i matricen. Et konstantled `1` omsættes til en søjle af ettaller. Et led som består af navnet på én kvantitativ variabel omsættes til en søjle der indeholder dennes værdier. Et produkt af kvantitative variable genererer en søjle bestående af produkterne af disses værdier. Kun hvis der optræder faktorer i et led genereres mere end én søjle. Antallet af søjler som genereres af et led er produktet af niveauantallene for de faktorer der forekommer i leddet. Hvis et led udelukkende indeholder faktorer bliver søjlerne "dummy'erne" for grupperne i den tilsvarende krydsklassifikation. Hvis der også forekommer kvantitative variable ganges alle søjlerne plads for plads med disses værdier.

Eksempler. Lad `F` og `G` være betegnelser for faktorer, medens `X` og `Z` er betegnelser for regressionsvariable, og `Y` er navnet på den vektor der indeholder responserne. Den ensidede variansanalysemodel svarende til `G` er så givet ved modelformlen

`Y=G` (minimal, injektiv parametrisering)

eller

`Y=1+G` (med konstantled, med henblik på at få udført homogenitetstestet).

Den tosidede model med vekselvirkning kan skrives

`Y=1+F+G+F*G`

hvor de tre første led kan udelades efter behag, afhængigt af hvad man vil have testet og hvilken parametrisering man ønsker.

En simpel regressionsmodel skrives typisk

`Y=1+X`

men hvis man af en eller anden grund vil have testet for afskæring=0 kan man skrive

`Y=X+1`

(dette gælder kun for ISU, i SAS og GENSTAT er det mere indviklet)

En regressionsmodel med to uafhængige variable kunne se sådan ud:

`Y=1+X+Z`

og hvis man af en eller anden grund ville forsøge at tilpasse et andengradspolynomium i de to variable kunne man skrive

`Y=1+X+Z+X*X+Z*Z+X*Z`

En model, der beskriver observationerne i et grupperet talmateriale ved en regressionsmodel, hvis parametre varierer fra gruppe til gruppe, kan skrives

`Y=F+F*X`

Hvis regressionslinierne skal være parallelle skriver man

`Y=F+X`

og hvis man ønsker at teste for parallellitet i den større model kan man derfor fra starten skrive

`Y=F+X+F*X`

således at fjernelse af sidste led netop kommer til at svare til hypotesen om parallellitet. Parametriseringen af den oprindelige model bliver så til gengæld ret kompliceret.

OPGAVE 10.7.1. Responserne y_i i eksempel 9.1 (side 82) er fremkommet ved at man for hver enkelt villa har divideret den månedlige bruttoydelse b_i med beboelsesarealet a_i:

$$y_i = \frac{b_i}{a_i}$$

En lineær normalfordelingsmodel, som er (mindst) lige så naturlig er følgende, hvor bruttoydelserne b_i opfattes som responser med arealet a_i og villaens placering $p \in \{\text{Nord,Syd}\}$ som forklarende variable:

$$\mathrm{E}b_i = \alpha_p + \beta_p a_i.$$

Modellen svarer til at man lægger en regressionslinie ind på hver af de to tegninger nedenfor, hvor punkterne (a_i, b_i) er indtegnet. Modellen forekommer umiddelbart rimelig, når man ser på tegningerne hver for sig, og selv om spredningen omkring linien måske er lidt større for område Nord end for område Syd, ser det ud til at være en brugbar model.

Residualkvadratsummen i denne model samt i to reducerede modeller fremgår af nedenstående skema:

Modelformel	$\mathrm{E}b_i$	Residualkvadratsum
`B=1+P+A+P*A`	$\alpha_p + \beta_p a_i$	283751543
`B=1+P+A`	$\alpha_p + \beta a_i$	288502385
`B=1+A`	$\alpha + \beta a_i$	325339633

Hvad kan man konkludere af dette? Fortolk resultatet økonomisk. Til orientering kan oplyses at parameterestimaterne i modellen $\mathrm{E}b_i = \alpha_p + \beta a_i$ blev

$$\hat{\alpha}_N = 6151$$
$$\hat{\alpha}_S = 4632$$
$$\hat{\beta} = 31.27$$

Hvad er det for en (forkert) antagelse i den ensidede variansanalysemodel fra eksempel 9.1, der gør at konklusionerne bliver så forskellige?