En Introduktion til Statistik

Bind 3B

Generaliserede lineære modeller

Poul Thyregod

LYNGBY 1998

IMM

Indhold

denne side blank

Afsnit 2

Likelihoodfunktion, generaliserede lineære modeller for endimensionale eksponentielle dispersions-parameterfamilier

Fil: glm1.tex 98-02-06

2.0 Indledning

I dette afsnit vil vi indføre de såkaldte generaliserede lineære modeller for eksponentielle dispersionsparametermodeller. Disse modeller kan opfattes som en generalisering af den generelle lineære model for normaltfordelte variable, som blev introduceret i Introduktion til Statistik, Bind 2. Den generelle lineære model for normaltfordelte variable sigter mod at beskrive

middelværdistrukturen i et observationssæt som en lineær funktion af en række forklarende variable. Klassen af generelle lineære modeller omfatter specielt de fra Introduktion til Statistik, Bind 1kendte lineære regressionsanalysemodeller og en- og tosidede variansanalyser. I disse normalfordelingsmodeller foretages estimationen ved mindste kvadraters metode og analysen af betydningen af de forklarende variable foretages ved at betragte en opspaltning af observationernes kvadratafvigelsessum i bidrag, der hver for sig beskriver indflydelsen af de forklarende variable.

De generaliserede lineære modeller udvider disse betragtninger til at omfatte modeller for middelværdistrukturen i observationssæt, hvis fordeling kan beskrives ved enhver af de naturlige fordelinger, der blev introduceret i Introduktion til Statistik, Bind 1. Det er imidlertid ikke altid formålstjenligt at betragte lineære modeller for middelværdistrukturen i disse fordelinger, da sådanne lineære modeller ofte vil indebære prædiktioner, der ligger uden for middelværdirummet i fordelingen. Imidlertid findes der for hver af disse fordelinger en naturlig afbildning, der afbilder middelværdirummet på hele den reelle talakse. Afbildningen fører middelværdien over i den såkaldte kanoniske parameter. De generaliserede lineære modeller omfatter lineære modeller for den kanoniske parameter, og endvidere omfatter de lineære modeller for vilkårlige funktioner af middelværdien.

I modsætning til normalfordelingen, hvor variansen ikke afhænger af middelværdien, gælder det for de fordelinger, vi her betragter, at variansen ændres med middelværdien. Det er således ikke tilfredsstillende blot at bruge kvadratafvigelsessummen som et mål for modeltilpasning. Vi indfører derfor et mål for modeltilpasning, den såkaldte devians, som er baseret på forskellen i værdien af likelihoodfunktionen svarende til observationen, og til den fittede værdi. Deviansen spiller den samme rolle ved analyse af generaliserede lineære modeller, som kvadratafvigelsessummen spiller ved analyse af generelle lineære modeller for normalfordelte størrelser. Således foregår vurderinger af modeltilpasningen ved betragtning af deviansen, og analysen af betydningen af de forklarende variable foretages ved at betragte en opspaltning af deviansen i bidrag, der hver for sig beskriver indflydelsen af de forklarende variable.

Da begrebsapparatet, der bruges i forbindelse med analyse af de generaliserede lineære modeller er nært knyttet til maksimum-likelihood estimation og kvotienttestprincipper, indleder vi i afsnit 2.1 med at resumere teorien omkring likelihoodfunktionen og maksimum-likelihood estimation.

I afsnit 2.2 introducerer vi de naturlige eksponentielle familier af fordelinger

og diskuterer parametriseringer af sådanne familier. Vi viser, hvorledes disse familier kan beskrives på en fælles standardform (den kanoniske form) med en tilhørende parametrisering ved den kanoniske parameter. Vi indfører en parametrisering ved middelværdien og viser, at en naturlig eksponentiel familie er karakteriseret ved sin variansfunktion, nemlig den funktion, der angiver hvorledes variansen afhænger af middelværdien. Imidlertid er en eksponentiel familie ikke altid rig nok til at kunne dække de praktiske anvendelser, og vi betragter udvidelsen med en såkaldt dispersionsparameter. Den herved fremkomne familie kaldes en eksponentiel dispersionsparameterfamilie.

Med udgangspunkt i likelihoodfunktionen indfører vi endelig et mål for afvigelse, deviansen mellem en observation og dens (evt estimerede) middelværdi.

Med henblik på at behandle lineære modeller for vilkårlige funktioner af observationernes middelværdi indfører vi begrebet linkfunktion, som er den funktion af middelværdien, som man ønsker at beskrive ved en lineær model.

De generaliserede lineære modeller indføres som lineære modeller for de ved link-funktionen transformerede middelværdier for de betragtede exponentielle dispersionsparameterfamilier.

Vi introducerer en generel matrix-formulering af lineære modeller, og viser, hvorledes de generaliserede lineære modeller (regressionsmodeller og variansanalysemodeller) kan udtrykkes som modeller for reduktion af affine rum og vi illustrerer estimation og test i sådanne modeller. Specielt diskuteres en hierarkisk modelreduktion.

Endelig introduceres en række mål for modeltilpasning, og for vurdering af observationers betydning, og vurdering af afvigende observationer.

2.1 Likelihoodfunktionen

Betragt en stokastisk variabel Y, der kan antage værdier i $\mathcal{Y} \subseteq \mathbb{R}^k$. I de situationer, vi betragter i statistikken, vil fordelingen af Y ikke være fuldstændig kendt (så var der jo ingen grund til at foretage den statistiske analyse). Sædvanligvis vil man imidlertid have fastlagt formen af fordelingen af Y, eller på anden måde have afgrænset samlingen af mulige fordelinger

for Y. Når man har afgrænset en en sådan samling $\mathfrak{P}$ af mulige fordelinger for Y siger man, at man har angivet en statistisk model for Y.

Sædvanligvis vil samlingen, $\mathfrak{P}$, af mulige sandsynlighedsfordelinger for Y være indekseret på en passende måde. En sådan indekseret mængde af fordelinger kaldes ofte en familie af fordelinger. Ved indekseringen har man parametriseret familien $\mathfrak{P}$ af fordelinger. En parameterfremstilling af $\mathfrak{P}$ med parametermængde Θ er en afbildning π, der afbilder parametermængden Θ ind i mængden af alle sandsynlighedsmål på $\mathcal{Y}$, sådan at værdimængden $\pi(\Theta) = \mathfrak{P}$. Ved afbildningen føres en parameterværdi $\theta \in \Theta$ over i en sandsynlighedsfordeling $\pi(\theta) = P_\theta \in \mathfrak{P}$.

En sådan familie kan eksempelvis være familien af binomialfordelinger med fastlagt antalsparameter n, indekseret ved sandsynlighedsparameteren p, $0 < p < 1$, eller samlingen af endimensionale normalfordelinger, indekseret ved forventningsværdi μ og varians σ^2, $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$. Som eksempel på en familie med en lidt mere struktureret indeksmængde kan man betragte en familie til beskrivelse af fordelingen af et sæt positive heltallige stokastiske variable Y_{ij}, $i = 1, \ldots, r$, $j = 1, \ldots, s$. Man kan i dette tilfælde eksempelvis betragte en familie af Poissonfordelinger $\mathrm{P}(\lambda_{ij})$, hvor λ_{ij} kan udtrykkes på formen $\lambda_{ij} = \eta_i \zeta_j$ med $\eta_i \in \mathbb{R}_+$, $i = 1, \ldots, r$ og $\zeta_j \in \mathbb{R}_+$, $j = 1, \ldots, s$

Da parametriseringen essentielt blot er udtryk for en indeksering af samlingen af fordelinger, er det klart, at en given familie af fordelinger kan parametriseres på mange måder. Man kan ikke nødvendigvis hævde, at en bestemt parametrisering altid vil være den bedste. Familien af binomialfordelinger $\mathrm{B}(n, p)$ med fastholdt antalsparameter n og med sandsynlighedsparameter $0 < p < 1$ er således parametriseret ved sandsynlighedsparameteren p med parametermængden $]0, 1[$. I andre sammenhænge kan det være mere relevant at parametrisere familien ved logaritmen til odds, $\vartheta = \ln[p/(1-p)]$ og parametermængden $\mathbb{R}$. De to parametriseringer beskriver samme familie af fordelinger, og er således ækvivalente.

Det vil ofte være naturligt at betragte en sandsynlighedsfordeling (eller et sandsynlighedsmål) beskrevet ved frekvensfunktionen, eller sandsynlighedstætheden. I overensstemmelse med den generelle mål- og integralteori vil vi i den følgende fremstilling benytte betegnelsen tæthed for en tæthed med hensyn til et vilkårligt mål $\nu\{\cdot\}$. I det følgende vil betegnelsen tæthed således omfatte såvel de sædvanlige frekvensfunktioner for kontinuerte variable (tætheder m.h.t. Lebesgue-målet), som frekvensfunktioner for diskrete

variable (tætheder m.h.t. f.eks. tællemålet). En familie af sandsynlighedsfordelinger $\mathfrak{P} = \{P_\theta\}_{\theta\in\Theta}$ vil således kunne repræsenteres ved en samling, $\{f(y;\theta)\}_{\theta\in\Theta}$ af sandsynlighedstætheder, hvor parameteren θ tilhører en given parametermængde Θ.

Den statistiske reduktion af data vil bestå i en indskrænkning af den familie af sandsynlighedsfordelinger, man vil benytte til beskrivelse af data. Ved modelreduktion vil man tilstræbe at benytte en mindre familie, og ved estimation af parametre vil man udpege en enkelt fordeling i familien.

Definition 2.1.1 ***Likelihoodfunktion, log likelihood***

Lad den statistiske model for observationssættet $Y_1, Y_2, \ldots, Y_n$ være givet ved familien af simultane tætheder

$$\{f(y_1, y_2, \ldots, y_n; \theta)\}_{\theta\in\Theta} \tag{2.1.1}$$

med hensyn til et mål ν_n på $\mathcal{Y}^n$

Ved likelihoodfunktionen for θ svarende til observationen $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$ forstår vi da funktionen

$$L(\theta; \mathbf{y}) \stackrel{\text{DEF}}{=} f(y_1, y_2, \ldots, y_n; \theta) \tag{2.1.2}$$

Likelihoodfunktionen angiver således den simultane tæthed for observationssættet som funktion af parameteren θ. Vi bemærker, at mens tætheden for observationssættet opfattes som en funktion af de mulige observationsværdier, $(y_1, y_2, \ldots, y_n)$ for en fastholdt værdi af parameteren θ, så opfattes likelihoodfunktionen som en funktion af parameteren θ for en fastholdt værdi (den observerede) af observationssættet $(y_1, y_2, \ldots, y_n)$. Vi har symboliseret denne forskel ved at lade det første argument i likelihoodfunktionen være θ, mens det første argument i den simultane tæthed er et muligt observationssæt.

Ofte er det mere bekvemt at betragte logaritmen til likelihoodfunktionen.

Vi indfører derfor log-likelihoodfunktionen svarende til observationssættet $\mathbf{y}$ som

$$l(\theta;\ \mathbf{y}) \stackrel{\text{DEF}}{=} \ln(L(\theta;\ \mathbf{y})) \tag{2.1.3}$$

hvor $L(\theta;\ \mathbf{y})$ er givet ved (2.1.2).

Vi vil almindeligvis udelade afhængigheden af observationen $\mathbf{y}$ ved angivelsen af likelihoodfunktionen. □

Når parameteren θ er flerdimensional, kan det undertiden være nyttigt at betragte grupper af parametre ad gangen.

Definition 2.1.2 ***Profil-likelihood***

Lad $(\theta^{(1)}, \theta^{(2)})$ angive en opdeling af parametervektoren θ for familien (2.1.1) af tætheder. Ved profil-likelihooden for $\theta^{(1)}$ vil vi forstå funktionen

$$\widetilde{L}(\theta^{(1)};\mathbf{y}) \stackrel{\text{DEF}}{=} \max_{\theta^{(2)}} L(\theta^{(1)}, \theta^{(2)};\mathbf{y}) \tag{2.1.4}$$

hvor $L(\theta^{(1)}, \theta^{(2)};\mathbf{y})$ angiver likelihoodfunktionen for $\theta = (\theta^{(1)}, \theta^{(2)})$

Tilsvarende defineres profil-log-likelihooden for $\theta^{(1)}$ som

$$\tilde{l}(\theta^{(1)};\mathbf{y}) \stackrel{\text{DEF}}{=} \max_{\theta^{(2)}} l(\theta^{(1)}, \theta^{(2)};\mathbf{y}) \tag{2.1.5}$$

hvor $l(\theta^{(1)}, \theta^{(2)};\mathbf{y})$ angiver log-likelihoodfunktionen for $\theta = (\theta^{(1)}, \theta^{(2)})$. □

Betegnelsen profil-likelihood skyldes, at i situationen, hvor parameteren θ er todimensional, da svarer funktionen $\widetilde{L}(\theta^{(1)})$ netop til profilen af grafen af $L(\theta^{(1)}, \theta^{(2)})$ når man ser ud ad $\theta^{(1)}$-aksen.

Definition 2.1.3 ***Likelihood-uafhængighed***

Betragt en situation, hvor parameteren θ er flerdimensional, og opdelt i komponenterne $(\theta^{(1)}, \theta^{(2)})$.

Såfremt likelihoodfunktionen for $(\theta^{(1)}, \theta^{(2)})$ kan faktoriseres i et produkt

$$L(\theta^{(1)}, \theta^{(2)}; \mathbf{y}) = L_1(\theta^{(1)}; \mathbf{y})\ L_2(\theta^{(2)}; \mathbf{y})$$

af to funktioner, hvor den ene alene afhænger af $\theta^{(1)}$, og den anden kun afhænger af $\theta^{(2)}$, siges komponenterne $\theta^{(1)}$ og $\theta^{(2)}$ at være likelihood-uafhængige.

Det er klart, at såfremt log-likelihoodfunktionen for $(\theta^{(1)}, \theta^{(2)})$ kan skrives som en sum

$$l(\theta^{(1)}, \theta^{(2)}; \mathbf{y}) = l_1(\theta^{(1)}; \mathbf{y}) + l_2(\theta^{(2)}; \mathbf{y})$$

af to funktioner, hvor den ene alene afhænger af $\theta^{(1)}$ og den anden kun afhænger af $\theta^{(2)}$, da er $\theta^{(1)}$ og $\theta^{(2)}$ likelihood-uafhængige. □

2.1.1 Sufficiens

Vi minder om definitionen på sufficiens fra Introduktion til Statistik, Bind 1.

Definition 2.1.4 *Sufficient stikprøvefunktion*

Betragt en statistisk model for observationssættet $Y_1, Y_2, \ldots, Y_n$; givet ved familien af tætheder (2.1.1), og betragt en afbildning, $T = T(y_1, y_2, \ldots, y_n)$ fra $\mathcal{Y}^n$ ind i $\mathfrak{T} \subset \mathbb{R}^p$.

Vi siger da, at stikprøvefunktionen T er sufficient for θ under modellen (2.1.1), såfremt den betingede fordeling af $Y_1, Y_2, \ldots, Y_n$ givet T ikke afhænger af θ. □

Bemærkning 1 *En sufficient stikprøvefunktion indeholder al stikprøveinformationen om parameteren θ*

Betingelsen siger jo netop, at når man kender værdien af T, så er resten af fordelingen af de enkelte Y'er blot støj, hvis fordeling ikke afhænger af θ, og hvis individuelle værdier derfor heller ikke kan sige noget om θ. □

Sætning 2.1.1 *Neyman's kriterium*

Lad den statistiske model for $Y_1, Y_2, \ldots, Y_n$ være givet ved familien af tætheder (2.1.1). Funktionen $T = T(y_1, y_2, \ldots, y_n)$ er sufficient for θ hvis og kun hvis likelihoodfunktionen (2.1.2) kan skrives som

$$f(y_1, y_2, \ldots, y_n; \theta) = h(T(y_1, y_2, \ldots, y_n); \theta)\ k(y_1, y_2, \ldots, y_n) \tag{2.1.6}$$

hvor $h(\cdot\,;\cdot)$ kun afhænger af $t = T(y_1, y_2, \ldots, y_n)$ og θ, og $k(\cdot, \cdot, \cdots)$ kun afhænger af $y_1, y_2, \ldots, y_n$.

Bevis:
Beviset overspringes, se f.eks. Dudewicz and Mishra (1988) □

Eksempel 2.1.1 ***Likelihoodfunktionen for uafhængige identisk fordelte obs.***

Såfremt observationerne $Y_1, Y_2, \ldots, Y_n$ er uafhængige og identisk fordelte med tætheder $f(\cdot;\theta)$ bliver likelihoodfunktionen

$$L(\theta;\ \mathbf{y}) = \prod_1^n f(y_i;\theta)$$

og tilsvarende fås log-likelihoodfunktionen

$$l(\theta;\ \mathbf{y}) = \sum_1^n \ln(f(y_i;\theta)\ )$$

□

Eksempel 2.1.2 ***Likelihoodfunktionen for uafhængige normalt fordelte obs.***

Såfremt observationerne $Y_1, Y_2, \ldots, Y_n$ er uafhængige og identisk fordelte med $Y_i \in \mathrm{N}(\mu, \sigma^2)$, da er likelihoodfunktionen for (μ, σ^2)

$$L(\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mu)^2\right] \tag{2.1.7}$$

Det følger af opspaltningen

$$\sum_{i=1}^n (y_i - \mu)^2 = \sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2\,,$$

at likelihoodfunktionen kan omformes til

$$L(\mu,\sigma^2) = (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i-\bar{y})^2 + n(\bar{y}-\mu)^2\right]\right\}.$$

Det gælder altså, at parret $\left(\overline{Y}, \sum_{i=1}^{n}(Y_i-\overline{Y})^2\right)$ er sufficient for parametrene (μ,σ^2).

(Sufficienskravet fastlægger ikke entydigt en stikprøvefunktion. Det gælder således også at eksempelvis parret $\left(\sum Y_i, \sum_{i=1}^{n}(Y_i-\overline{Y})^2/(n-1)\right)$ er sufficient for (μ,σ^2)).

Profillikelihoodfunktionen for σ^2 er

$$\begin{aligned}\widetilde{L}(\sigma^2) &= \max_{\mu} L(\mu,\sigma^2)\\ &= (2\pi\sigma^2)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y})^2\right]\ \max_{\mu}\left\{\exp\left[\frac{-n(\bar{y}-\mu)^2}{2\sigma^2}\right]\right\}\\ &= (2\pi\sigma^2)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y})^2\right]\end{aligned}$$

idet maksimum mht μ fås for $\mu=\bar{y}$. □

Definition 2.1.5 *Likelihood-sufficiens*

Betragt en statistisk model for observationssættet $Y_1, Y_2, \ldots, Y_n$; givet ved familien af tætheder (2.1.1), og betragt en afbildning, $T = T(y_1, y_2, \ldots, y_n)$ fra $\mathcal{Y}^n$ ind i $\mathfrak{T} \subset \mathbb{R}^p$, og antag at parameteren θ er flerdimensional og opdelt i komponenterne $(\theta^{(1)}, \theta^{(2)})$.

Såfremt profillikelihoodfunktionen for $\theta^{(1)}$ kun afhænger af observationssættet $y_1, \ldots, y_n$ gennem funktionen T, siges T at være likelihood-sufficient for $\theta^{(1)}$. □

Eksempel 2.1.3 *Likelihoodsufficiens for σ^2 ved uafhængige normalt fordelte obs.*

Betragt situationen som i eksempel 2.1.2.

Idet

$$\widetilde{L}(\sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \bar{y})^2\right]$$

ses det, at størrelsen $\sum_{i=1}^{n}(y_i - \bar{y})^2$ er likelihood-sufficient for σ^2. □

2.1.2 Scorefunktionen og Informationsmatricen

Definition 2.1.6 *Scorefunktionen*

Lad den statistiske model for observationssættet $Y_1, Y_2, \ldots, Y_n$ være givet ved familien af simultane tætheder (2.1.1), hvor parameterområdet Θ er en åben delmængde af $\mathbb{R}^k$. Såfremt log-likelihoodfunktionen er kontinuert differentiabel, betragter vi vektorfunktionen

$$l'_\theta(\theta;\, \mathbf{y}) \stackrel{\text{DEF}}{=} \frac{\partial}{\partial\theta}\, l(\theta;\, \mathbf{y}) = \begin{pmatrix} \frac{\partial}{\partial\theta_1}\, l(\theta;\, \mathbf{y}) \\ \vdots \\ \frac{\partial}{\partial\theta_k}\, l(\theta;\, \mathbf{y}) \end{pmatrix} \tag{2.1.8}$$

Funktionen $l'_\theta(\theta;\, \mathbf{y})$ benævnes scorefunktionen. □

Bemærkning 1 *Scorefunktionen for transformerede parametre*

Betragt en anden parametrisering givet ved $\theta = \theta(\beta) \in \mathbb{R}^k$ for $\beta \in \mathrm{B} \subset \mathbb{R}^m$ med $m \leq k$ og antag, at afbildningen er kontinuert differentiabel.

Da er scorefunktionen svarende til parametersættet β bestemt ved

$$l'_\beta(\beta;\, \mathbf{y}) = \mathbf{J}^T\, l'_\theta(\theta(\beta);\, \mathbf{y}) \tag{2.1.9}$$

hvor Jacobi-matricen $\mathbf{J}$ er givet ved

$$\mathbf{J} = \frac{\partial\theta}{\partial\beta}$$

med elementer

$$\mathbf{J}_{ij} = \frac{\partial \theta_i}{\partial \beta_j} \quad i = 1, 2, \ldots, k; \quad j = 1, 2, \ldots, m$$

□

Eksempel 2.1.4 *Log-likelihoodfunktionen svarende til den sædvanlige regressionsmodel*

Lad $Y_1, Y_2, \ldots, Y_n$ være en følge af indbyrdes uafhængige observationer med $Y_i \in \mathrm{N}(\alpha + \beta x_i, \sigma^2)$, hvor $x_1, x_2, \ldots, x_n$ er kendte størrelser.

For et givet observationssæt $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$ får vi da log-likelihoodfunktionen

$$l(\alpha, \beta, \sigma^2;\ \mathbf{y}) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_1^n (y_i - \alpha - \beta x_i)^2$$

således at scorefunktionen bliver

$$\frac{\partial}{\partial(\alpha, \beta, \sigma^2)} l(\alpha, \beta, \sigma^2;\ \mathbf{y}) = \begin{pmatrix} \frac{1}{\sigma^2} \sum_i (y_i - \alpha - \beta x_i) \\ \\ \frac{1}{\sigma^2} \sum_i x_i (y_i - \alpha - \beta x_i) \\ \\ -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_i (y_i - \alpha - \beta x_i)^2 \end{pmatrix}$$

□

Sætning 2.1.2 *Momenter af scorefunktionen*

Lad den statistiske model for observationssættet $Y_1, Y_2, \ldots, Y_n$ være givet ved den simultane tæthed (2.1.1), hvor parameterområdet Θ er en åben delmængde af $\mathbb{R}^k$. Såfremt log-likelihoodfunktionen er to gange kontinuert differentiabel og såfremt støtten for fordelingerne ikke afhænger af θ, da gælder

$$\mathsf{E}\left[\frac{\partial}{\partial\theta}\, l(\theta;\, \mathbf{Y})\right] = \mathbf{0}$$

$$\mathsf{E}\left[\frac{\partial^2}{\partial\theta\,\partial\theta^T}\, l(\theta;\, \mathbf{Y})\right] + \mathsf{E}\left[\frac{\partial}{\partial\theta}\, l(\theta;\, \mathbf{Y})\,(\frac{\partial}{\partial\theta}\, l(\theta;\, \mathbf{Y}))^T\right] = \mathbf{0}\,, \tag{2.1.10}$$

hvor

$$\mathsf{E}\left[\frac{\partial^2}{\partial\theta\,\partial\theta^T}\, l(\theta;\, \mathbf{Y})\right]$$

er matricen, hvis (i,j)'te element er givet ved

$$\mathsf{E}\left[\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\, l(\theta;\, \mathbf{Y})\right]$$

Bevis:
Beviset følger ved at differentiere udtrykket

$$\int f(\mathbf{y};\theta)\;\mu\{dx\} = 1$$

med hensyn til θ.

□

Definition 2.1.7 ***Observeret og forventet information***

Lad log-likelihoodfunktionen for en statistisk model være givet ved (2.1.3), hvor parameterområdet Θ er en åben delmængde af $\mathbb{R}^k$. Matricen

$$\mathbf{i}(\theta;\mathbf{y}) = -\,\frac{\partial^2}{\partial\theta\,\partial\theta^T}\, l(\theta;\, \mathbf{y}) \tag{2.1.11}$$

med elementer

$$\mathbf{i}(\theta;\mathbf{y})_{ij} = -\,\frac{\partial^2}{\partial\theta_i\partial\theta_j} l(\theta;\, \mathbf{y})$$

kaldes for den observerede information svarende til observationen $\mathbf{y}$ og parameterværdien θ.
Forventningsværdien af den observerede information,

$$\mathbf{i}(\theta) = \mathsf{E}\,[\,\mathbf{i}(\theta;\mathbf{Y})\,] \tag{2.1.12}$$

med $\mathbf{i}(\theta; \mathbf{Y})$ givet ved (2.1.11) kaldes for den forventede information, eller blot informationsmatricen svarende til parameterværdien θ. Den forventede information (2.1.12) kaldes undertiden Fisher informationen. □

Bemærkning 1 *Den forventede information er dispersionsmatricen for scorefunktionen*

Det følger af sætning 2.1.2, at under regularitetsbetingelserne for sætningen kan den forventede information $\mathbf{i}(\theta)$ udtrykkes som

$$\begin{aligned} \mathbf{i}(\theta) = -\mathsf{E}\left[\frac{\partial^2}{\partial\theta\,\partial\theta^T} l(\theta;\,\mathbf{Y})\right] &= \mathsf{E}\left[\frac{\partial}{\partial\theta} l(\theta;\,\mathbf{Y})\,(\frac{\partial}{\partial\theta} l(\theta;\,\mathbf{Y})\,)^T\right] \\ &= \mathbf{D}\,[l'_\theta(\theta;\,\mathbf{Y})\,] \end{aligned} \tag{2.1.13}$$

Den forventede information, der er defineret som forventningsværdien af minus den anden afledede af log-likelihoodfunktionen, er altså lig med dispersionsmatricen for scorefunktionen.

Da informationsmatricen således er en dispersionsmatrix, er den positiv semidefinit.

Den observerede information angiver den aktuelle krumning (med modsat fortegn) af log-likelihoodfunktionen svarende til observationen $\mathbf{y}$ og parameterværdien θ.

Informationsmatricen angiver tilsvarende den forventede krumning (med modsat fortegn) af log-likelihoodfunktionen svarende til parameterværdien θ. Informationsmatricen er således et mål for, hvor godt parameteren bestemmes. Hvis informationsmatricen er "stor", er parameteren godt bestemt. □

Sætning 2.1.3 *Den forventede information er additiv over uafhængige data*

Lad X og Y være stokastiske variable svarende til de statistiske problemer $\mathcal{X}, P_\theta, \theta \in \Theta$ og $\mathcal{Y}, Q_\theta, \theta \in \Theta$ og sådan at X og Y er uafhængige med hensyn til produktmålet $P_\theta \otimes Q_\theta$. Der gælder da, at

$$\mathbf{i}_{X\times Y}(\theta) = \mathbf{i}_X(\theta) + \mathbf{i}_Y(\theta)$$

Bevis:
Overspringes □

Eksempel 2.1.5 ***Den forventede information for uafhængige identisk fordelte observationer***

Såfremt specielt observationssættet $Y_1, Y_2, \ldots, Y_n$ er uafhængige og identisk fordelte med tætheder $f(y;\theta)$, bliver den forventede information

$$\begin{aligned}
\mathbf{i}_n(\theta) &= -\mathsf{E}\left[\frac{\partial^2}{\partial\theta\ \partial\theta^T}\, l(\theta;\ \mathbf{Y})\right] \\
&= \sum_1^n \mathsf{E}\left[\frac{\partial^2}{\partial\theta\ \partial\theta^T}\, \ln(f(Y_i;\theta)\ )\right] \\
&= n \times \mathsf{E}\left[\frac{\partial^2}{\partial\theta\ \partial\theta^T}\, \ln(f(Y_i;\theta)\ )\right] = n \times \mathbf{i}_1(\theta)\ .
\end{aligned}$$

Informationen svarende til n uafhængige observationer fra samme fordeling er således n gange så stor som informationen fra en enkelt observation. □

Sætning 2.1.4 ***Den forventede information er monoton***

Hvis $Z = t(Y)$, er informationen i Z ikke større, end informationen i Y Såfremt parameteren θ er endimensional gælder

$$\mathbf{i}_Y(\theta) \geq \mathbf{i}_Z(\theta)\ .$$

Såfremt θ er flerdimensional gælder, at matricen

$$\mathbf{i}_Y(\theta) - \mathbf{i}_Z(\theta)$$

er positiv semidefinit.
Bevis:
Overspringes □

Sætning 2.1.5 *Den forventede information ved sufficiente transformationer*

Hvis $T = t(Y)$ er sufficient for θ, er

$$\mathbf{i}_Y(\theta) = \mathbf{i}_T(\theta)$$

Bevis:
Overspringes □

Bemærkning 1 *Informationsmatricen for transformerede parametre*

Betragt en anden parametrisering givet ved $\theta = \theta(\beta) \in \mathbb{R}^k$ for $\beta \in \mathrm{B} \subset \mathbb{R}^m$ med $m \leq k$ og antag, at afbildningen er to gange kontinuert differentiabel.

Da er den forventede information med hensyn til β givet ved

$$\mathbf{i}(\beta) = -\mathsf{E}\left[\frac{\partial^2}{\partial\beta\partial\beta^T} l(\theta(\beta);\ \mathbf{Y})\right] = \mathbf{J}^T\ \mathbf{i}(\theta(\beta))\ \mathbf{J} \tag{2.1.14}$$

hvor Jacobi-matricen $\mathbf{J}$ er angivet i bemærkningen til definition 2.1.6.

Resultatet vises ved at bemærke, at den forventede information kan udtrykkes som dispersionsmatricen for scorefunktionen (2.1.9). □

2.1.3 Maksimum likelihood estimat

Definition 2.1.8 *Maksimum likelihood estimat*

Lad den statistiske model for observationssættet $Y_1, Y_2, \ldots, Y_n$ være givet ved familien af simultane tætheder (2.1.1), hvor parameteren $\theta \in \Theta$.

Ved et maksimum-likelihood estimat $\hat{\theta}$ for parameteren θ svarende til observationen $y_1, y_2, \ldots, y_n$ forstås den (eller de) værdi(er) af θ, der giver anledning til maksimum af likelihoodfunktionen (2.1.2).

□

Bemærkning 1 *Bestemmelse af maksimum-likelihood estimat*

Bestemmelse af maksimum-likelihood estimatet er forbundet med de sædvanlige problemer ved bestemmelse af størsteværdien af en funktion. Såfremt parameterområdet Θ er afsluttet (kompakt), findes maksimum enten som et lokalt ekstremumspunkt, eller på randen. Maksimum likelihood estimatet er ikke nødvendigvis entydigt. Det kan således ikke udelukkes, at flere værdier af θ giver anledning til samme maksimale værdi. For specielle familier af fordelinger er det imidlertid muligt at udtale sig mere præcist om eksistens og entydighed af maksimum-likelihood estimatet. Således gælder det for regulære og stejle eksponentielle familier af fordelinger, at maksimum-likelihood estimatoren i det væsentlige er entydig (se Oversigt over fordelinger med anvendelser i Statistik, IMM 1998). □

Bemærkning 2 *Maksimum-likelihood estimatoren er en stokastisk variabel*

Da likelihoodfunktionen afhænger af observationssættet $\mathbf{y}$, vil også maksimum likelihood estimatet afhænge af observationssættet, $\hat{\theta} = \hat{\theta}(\mathbf{y})$. Maksimum likelihood estimatoren er således den estimator, der til enhver observation $\mathbf{y}$ tilordner maksimum likelihood estimatet $\hat{\theta}(\mathbf{y})$. Maksimum likelihood estimatoren er således en stokastisk variabel, hvis fordeling afhænger af parameteren θ. □

Bemærkning 3 *Maksimum-likelihood estimatoren afhænger kun af observationerne igennem den sufficiente stikprøvefunktion*

Det følger af Neyman's kriterium (Sætning 2.1.1) at likelihoodfunktionen svarende til den sufficiente stikprøvefunktion kun adskiller sig ved en konstant faktor fra likelihoodfunktionen svarende til hele observationssættet. Ved maksimum-likelihood estimation er det således ligegyldigt om man betragter likelihoodfunktionen svarende til den sufficiente størrelse, eller likelihoodfunktionen svarende til hele observationssættet. □

Bemærkning 4 *Maksimum-likelihood metoden ved estimation af flere parametre under likelihood-sufficiens*

Såfremt den statistiske model for observationssættet $Y_1, Y_2, \ldots, Y_n$ er sådan at parameteren θ er flerdimensional, og at funktionen $T = T(y_1, y_2, \ldots, y_n)$ er likelihood-sufficient for komponenten $\theta^{(1)}$ af θ (jvf definition 2.1.5), da vil det være naturligt at bruge den marginale fordeling af T til estimation af $\theta^{(1)}$, dvs. at man bestemmer maksimum-likelihood estimatet for $\theta^{(1)}$ ved at maksimere likelihoodfunktionen svarende til den marginale fordeling af T.

□

Eksempel 2.1.6 ***Maksimum-likelihood estimation af σ^2 ved uafhængige normalt fordelte obs.***

Betragt situationen som i eksempel 2.1.2. I eksempel 2.1.3 så vi, at størrelsen $\sum_{i=1}^{n}(y_i - \bar{y})^2$ er likelihood-sufficient for σ^2.

Vi ved fra Introduktion til Statistik, Bind 1, at den marginale fordeling af $T = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$ er en $\sigma^2\ \chi^2(f)$-fordeling med $f = n - 1$, dvs en $\mathrm{G}\big((n-1)/2, 2\sigma^2\big)$-fordeling med tætheden

$$g(t|f, \sigma^2) = \exp\left\{(f/2 - 1)\ln(t) - t/(2\sigma^2) - (f/2)\ln(2\sigma^2) - \ln(\Gamma(f/2)\right\}$$

for $t \in \mathbb{R}_+$.

Logaritmen til likelihoodfunktionen for σ^2 er således på formen

$$l(\sigma^2; t) = C - t/(2\sigma^2) - (f/2)\ln(2\sigma^2)\ ,$$

der netop antager sit maksimum for $\sigma^2 = t/f$, hvorfor maksimum-likelihood-estimatoren for σ^2 er

$$\widehat{\sigma}^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2/(n-1)$$

altså netop den centrale estimator, S^2, som blev introduceret i Introduktion til Statistik, Bind 1.

Hvis man blot havde maksimeret den simultane likelihoodfunktion for μ og σ^2 (2.1.7) under ét, ville man have fundet at maksimum-likelihood estimatet for σ^2 (ved estimation i den simultane likelihoodfunktion) er af formen

$$\widehat{\sigma}^2 = \sum_{i=1}^{n}(y_i - \mu)^2 \Big/ n\ ,$$

hvor μ erstattes med maksimum-likelihood estimatet $\widehat{\mu}$. Men da maksimum-likelihood estimatet for μ netop vælges som den værdi, der minimerer

$$\sum_{i=1}^{n}(y_i - \mu)^2 ,$$

ville dette estimat for σ^2 jo systematisk blive lidt for lille.

Såfremt man - i en situation med flere parametre - kan identificere likelihood-sufficiens, bør man altså, som her, estimere i den marginale fordeling af den likelihood-sufficiente størrelse. □

2.2 Eksponentielle familier og dispersionsmodeller

fil glm2.tex 1998-02-07

De fleste af de fordelinger, der blev introduceret i Introduktion til Statistik, Bind 1, har nogle væsentlige egenskaber fælles. Således har vi set, at ved uafhængige, identisk fordelte gentagelser vil summen af observationerne, eller summen af en funktion af observationerne være sufficient. Det gælder endvidere, at ved en passende parametrisering vil affine hypoteser vedrørende parameteren modsvares af tilsvarende affine transformationer af observationerne.

Sådanne egenskaber deles af en stor gruppe af fordelinger, der under ét betegnes eksponentielle familier.

I Oversigt over fordelinger med anvendelser i Statistik, IMM 1998 er de eksponentielle familier introduceret. Vi skal her kun resumere nogle vigtige resultater for uafhængige observationer fra endimensionale eksponentielle familier.

Vi indfører standardformen (den kanoniske form) for en eksponentiel familie og den tilsvarende parametrisering ved den kanoniske parameter. Ønsker man direkte at relatere observationer til parameterværdier er det imidlertid mere bekvemt at betragte den såkaldte middelværdiparametrisering af familierne. For de endimensionale familier gælder specielt, at familien er karakteriseret ved den funktion, der beskriver variansen som funktion af middelværdien.

I afsnit 2.2.2 generaliserer vi den eksponentielle familie til at omfatte en ekstra parameter, den såkaldte dispersionsparameter. Herved kan fx normalfordelingen med ukendt middelværdi og spredning beskrives. Afsnit 2.2.3 giver en oversigt over de vigtigste størrelser i fortolkningen af de sædvanlige familier af fordelinger som eksponentielle dispersionsmodeller.

I afsnit 2.2.1 introduceres et likelihoodbaseret mål, deviansen, for en observations afvigelse fra dens (evt estimerede) middelværdi. Denne størrelse er fundamental for behandlingen af de generaliserede lineære modeller.

2.2.1 Naturlige eksponentielle familier af fordelinger

Definition 2.2.1 *Naturlig eksponentiel familie af fordelinger*

Betragt en endimensional stokastisk variabel, X, hvis fordeling beskrives ved en familie af tætheder (eller frekvensfunktioner) $f(\cdot;\vartheta)_{\vartheta\in D}$. Såfremt tæthederne kan skrives på formen

$$f(x;\vartheta) = c(x)\exp\{\vartheta x - \kappa(\vartheta)\} \quad \text{for } \vartheta \in D, \tag{2.2.1}$$

hvor

$$D = \left\{\vartheta \in \mathbb{R} : \int \exp(\vartheta x)\nu\{dx\} < \infty\right\},$$

kaldes familien for en naturlig eksponentiel familie.

Parameteren ϑ kaldes den kanoniske parameter, parameterområdet D kaldes det kanoniske parameterområde, og funktionen $\kappa(\cdot)$ kaldes kumulantfrembringeren for familien.

Støtten S for familien er mængden af $x \in \mathbb{R}$ for hvilke tætheden (2.2.1) er positiv, og den konvekse støtte er den mindste konvekse mængde (interval), der indeholder støtten S.

Når en eksponentiel familie er på formen (2.2.1) siges den at være på standardform eller kanonisk form. □

Eksempel 2.2.1 ***Familien af Bernoullifordelinger som eksponentiel familie.***

Lad $X \in \mathrm{B}(1,p)$ med $0 < p < 1$. Frekvensfunktionen for X er da

$$g(x;p) = p^x(1-p)^{1-x} \quad \text{for } x \in \{=0,1\}$$

Indfører vi

$$\vartheta = \ln\left(\frac{p}{1-p}\right) \tag{2.2.2}$$

ser vi, at frekvensfunktionen kan udtrykkes som

$$f(x;\vartheta) = \exp\{\vartheta\, x - \ln\left(1+\exp(\vartheta)\right)\}, \tag{2.2.3}$$

□

hvilket netop er på formen (2.2.1) med

$$\kappa(\vartheta) = \ln\left(\exp(\vartheta)\right) .$$

Det kanoniske parameterområde er $D = \mathbb{R}$.

Vi bemærker, at selv om det har mening at definere familien af $\mathrm{B}(1,p)$-fordelinger til også at omfatte $p = 1$ og $p = 0$, omfatter den naturlige eksponentielle familie ikke disse to udartede fordelinger. Disse fordelinger ville svare til værdierne $\vartheta = \pm\infty$.

Sætning 2.2.1 ***Momenterne udtrykt ved kumulantfrembringeren***

Lad fordelingen af X tilhøre en naturlig eksponentiel familie på formen (2.2.1). Såfremt ϑ er i det indre af D gælder

$$\mathsf{E}[X] = \kappa'(\vartheta) \tag{2.2.4}$$
$$\mathsf{V}[X] = \kappa''(\vartheta) , \tag{2.2.5}$$

□

Definition 2.2.2 ***Middelværdiafbildning*** eksponentiel familie

Lad fordelingen af X tilhøre en naturlig eksponentiel familie med kumulantfrembringer $\kappa(\cdot)$ og parameterområde D, og lad

$$\tau(\vartheta) \stackrel{\mathrm{DEF}}{=} \kappa'(\vartheta) \tag{2.2.6}$$

Afbildningen $\tau(\cdot)$ kaldes middelværdiafbildningen , og billedmængden

$$\mathcal{M} = \{\mu \in \mathbb{R} \,|\, \mu = \tau(\vartheta) \text{ for et } \vartheta \in D\} .$$

kaldes middelværdirummet for familien. □

Definition 2.2.3 *Middelværdiparametrisering af naturlig eksponentiel familie*

Middelværdiafbildningen $\tau(\cdot)$ er en monoton, injektiv afbildning fra det indre af D ind på M . I stedet for at parametrisere familien ved den kanoniske parameter $\vartheta \in D$, kan man altså lige så godt parametrisere familien ved middelværdien $\mu = \tau(\vartheta) \in \mathcal{M}$. Denne parametrisering kaldes middelværdiparametriseringen.

Den omvendte afbildning, $\tau^{-1}(\cdot) : \mathcal{M} \to D$ tilordner den kanoniske parameter ϑ til middelværdien. □

Bemærkning 1 *Middelværdirummet modsvarer rummet af observationer*

Vi bemærker, at mens den "naturlige parameter", ϑ ikke direkte er sammenlignelig med observationerne x, så er punkterne i middelværdirummet direkte sammenlignelige med observationerne. Det gælder, at middelværdirummet er indeholdt i (evt lig med) det indre af den konvekse støtte for familien. Således angives punkter i middelværdirummet netop i de samme (fysiske) enheder som observationerne, mens punkter i parameterrummet algebraisk set er repræsentanter for linearformer på observationerne (se Oversigt over fordelinger med anvendelser i Statistik, IMM 1998). □

Definition 2.2.4 *Variansfunktion for naturlig eksponentiel familie*

Lad fordelingen af X tilhøre en naturlig eksponentiel familie med kumulantfrembringer $\kappa(\cdot)$ og middelværdiafbildning $\tau(\cdot)$.

Funktionen

$$V(\mu) \stackrel{\text{DEF}}{=} \kappa''(\tau^{-1}(\mu)) \quad \text{for } \mu \in \mathcal{M} \tag{2.2.7}$$

kaldes variansfunktionen for familien. □

Det følger af (2.2.5), at såfremt fordelingen af X tilhører en naturlig eksponentiel familie med variansfunktion $V(\cdot)$, da er

$$\mathsf{V}[X] = V(\mu) ,$$

hvor $\mu = \mathsf{E}[X]$, dvs at variansfunktionen angiver variansen for X som en funktion af middelværdien for X. □

Bemærkning 1 *En naturlig eksponentiel familie er bestemt af sin variansfunktion*

Vi bemærker, at variansfunktionen fastlægger den naturlige eksponentielle familie, se Oversigt over fordelinger med anvendelser i Statistik, IMM 1998. Det er således nok at angive hvorledes variansen afhænger af middelværdien og at familien er en naturlig eksponentiel familie, så er hele familien fastlagt. □

Eksempel 2.2.2 *Middelværdiafbildning og variansfunktion for Bernoullifordelingen.*

Vi så i eksempel 2.2.1, at familien af $\mathrm{B}(1,p)$-fordelinger med $0 < p < 1$ er en naturlig eksponentiel familie med kanonisk parameter $\vartheta = \ln\big(p/(1-p)\big)$ og med kumulantfrembringer

$$\kappa(\vartheta) = \ln\big(\exp(\vartheta)\big) .$$

Man finder da middelværdiafbildningen

$$\tau(\vartheta) = \kappa'(\vartheta) = \frac{\exp(\vartheta)}{1+\exp(\vartheta)} .$$

Indsættes heri $\vartheta = \ln\big(p/(1-p)\big)$ får vi netop for $X \in \mathrm{B}(1,p)$, at $\mathsf{E}[X] = \tau(\vartheta) = p$. Den sædvanlige parametrisering er netop ved middelværdien, p.

Middelværdirummet for den naturlige eksponentielle familie er det åbne interval, $\mathcal{M} = \{p : 0 < p < 1\}$.

Den anden afledede af kumulantfrembringeren er

$$\kappa''(\vartheta) = \tau'(\vartheta) = \frac{\exp(\vartheta)}{\big(1+\exp(\vartheta)\big)^2} .$$

Indsætter vi heri $\vartheta = \tau^{-1}(p) = \ln\big(p/(1-p)\big)$ finder vi variansfunktionen

$$V(p) = p(1-p) .$$

□

Vi har tidligere bemærket (Bemærkning 1 på side 126), at der er en kvalitativ forskel på rummet D af naturlige parametre og rummet $\mathcal{M}$ af middelværdier. Hvis man vil vurdere observationers afvigelse fra en bestemt fordeling i familien, kan man ikke umiddelbart sammenligne en observation x med en postuleret (eller estimeret) værdi af den kanoniske parameter ϑ.

Her kan man imidlertid drage fordel af middelværdiparametriseringen af familien. Den konvekse støtte for familien er jo stort set (bortset fra randpunkterne) sammenfaldende med middelværdirummet og vi kan derfor umiddelbart sammenligne observationer med postulerede (eller estimerede) værdier af middelværdien.

I det følgende vil vi indføre en størrelse, deviansen, til vurdering af afvigelsen mellem en observation x og en postuleret forventningsværdi μ i fordelingen af X. I overensstemmelse med at vi bruger maksimum-likelihood metoden til estimation af parametre, vil vi vurdere afvigelsen mellem en observation og en postuleret forventningsværdi ved forskellen i log-likelihood mellem at bruge x og μ.

Vi indfører

Definition 2.2.5 *Devians (enhedsdevians)*

Lad x angive en observation af en stokastisk variabel X, hvis fordeling tilhører en naturlig eksponentiel familie med kumulantfrembringer $\kappa(\cdot)$, (dvs med tætheden (2.2.1)) og med den konvekse støtte C.

Størrelsen

$$d(x;\mu) \stackrel{\text{DEF}}{=} 2\{\sup_{\vartheta\in\Theta}[\vartheta x - \kappa(\vartheta)] - [x\tau^{-1}(\mu) - \kappa(\tau^{-1}(\mu))]\} \tag{2.2.8}$$

for $x \in C$ og $\mu \in \mathcal{M}$

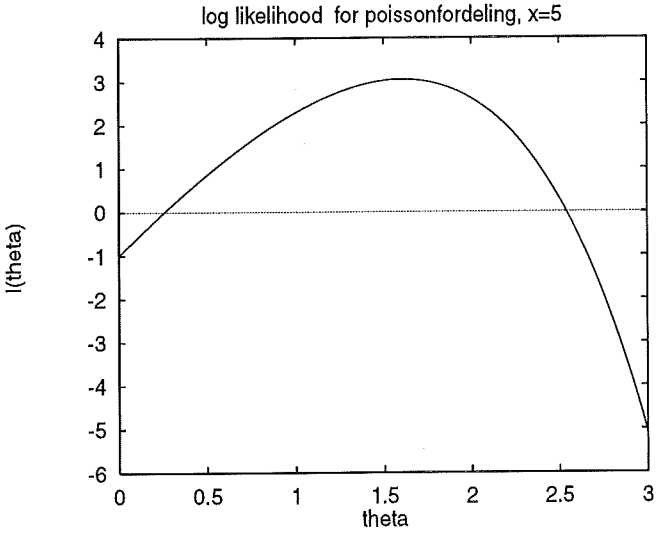
kaldes enhedsdeviansen (engelsk *unit deviance*) svarende til observationen x og middelværdien μ. □

I afsnit 2.5.6 vil vi betragte forskellige andre måder at vurdere afvigelsen mellem en observation og en postuleret (eller tilpasset) forventningsværdi.

Eksempel 2.2.3 *Enhedsdeviansen svarende til Poissonfordelte observationer*

Lad X være Poissonfordelt, $X \in \mathrm{P}(\lambda)$. Familien af Poissonfordelinger for $\lambda \in \mathbb{R}_+$ er en naturlig eksponentiel familie med den kanoniske parameter $\vartheta = \ln(\lambda)$, $\Omega = \mathbb{R}$ og med kumulantfrembringer $\kappa(\vartheta) = \exp(\vartheta)$, hvorfor $\tau(\vartheta) = \exp(\vartheta)$ og $\mathsf{V}[X] = \exp(\vartheta) = \mathsf{E}[X]$.

Nedenstående figur illustrerer forløbet af log-likelihood funktionen som funktion af den kanoniske parameter $\vartheta = \ln(\lambda)$ for familien af $\mathrm{P}(\lambda)$-fordelinger.

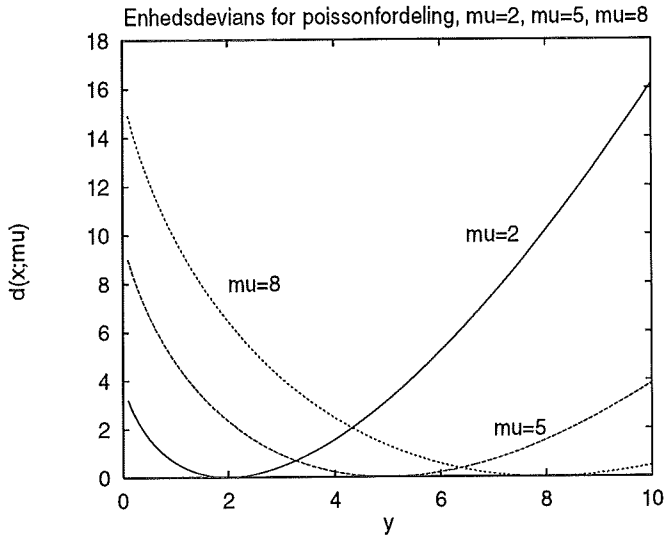


Den sædvanlige parametrisering af Poissonfordelingen er netop ved middelværdiparameteren, $\lambda = \mathsf{E}[X]$.

Man finder enhedsdeviansen

$$d(x;\mu) = 2 \times \{x \ln(x/\mu) - (x-\mu)\} \tag{2.2.9}$$

Nedenstående figur viser enhedsdeviansen for Poissonfordelingen som funktion af x for $\mu = 5$.



□

Bemærkning 1 ***Enhedsdeviansen svarende til observationer i middelværdirummet***

Vi bemærker, at første led i (2.2.8) netop udtrykker maksimum af

$$l(\vartheta; x) = \vartheta x - \kappa(\vartheta) \ .$$

Differentieres $l(\vartheta; x)$ med hensyn til μ finder man ligningen

$$x - \tau(\vartheta) = 0 \ .$$

Såfremt x ligger i middelværdirummet $\mathcal{M}$, har denne ligning løsningen $\widehat{\vartheta} = \tau^{-1}(x)$, og for $x \in \mathcal{M}$ har man derfor det alternative udtryk for enhedsdeviansen:

$$d(x; \mu) = 2\big[x\{\tau^{-1}(x) - \tau^{-1}(\mu)\} - [\kappa\{\tau^{-1}(x)\} - \kappa\{\tau^{-1}(\mu)\}]\big] \ , \tag{2.2.10}$$

dvs

$$d(x; \mu) = 2\{l_\mu(x; x) - l_\mu(\mu; x)\} \ ,$$

hvor

$$l_\mu(\mu; x) = \ln(f(x; \tau^{-1}(\mu)))$$

angiver log-likelihooden (for μ) svarende til observationen x. □

Deviansen mellem x og μ måler således forskellen mellem log-likelihood svarende til at bruge henholdsvis x og μ som middelværdi.

Bemærkning 2 ***Lokale egenskaber for enhedsdeviansen***

Man finder de første to afledede af enhedsdeviansen

$$\frac{\partial}{\partial \mu} d(x;\mu) = -2\frac{x-\mu}{V(\mu)} \tag{2.2.11}$$

$$\frac{\partial^2}{\partial \mu^2} d(x;\mu) = 2\left\{\frac{1}{V(\mu)} + (x-\mu)\frac{V'(\mu)}{V(\mu)^2}\right\} \tag{2.2.12}$$

Tilsvarende finder man den afledede af enhedsdeviansen med hensyn til observationen x som

$$\frac{\partial}{\partial x} d(x;\mu) = 2\{\tau^{-1}(x) - \tau^{-1}(\mu)\}$$

□

Bemærkning 3 ***Taylorudvikling af enhedsdeviansen***

Ved en Taylorudvikling af enhedsdeviansen omkring $x = \mu$ finder man jvf Oversigt over fordelinger med anvendelser i Statistik, IMM 1998side 40

$$d(x;\mu) \approx \frac{(x-\mu)^2}{V(\mu)} \tag{2.2.13}$$

□

Sætning 2.2.2 ***Relation mellem variansfunktion og enhedsdeviansen***

Betragt en naturlig eksponentiel familie med kumulantfrembringer $\kappa(\cdot)$, dvs med tætheden (2.2.1). Lad enhedsdeviansen $d(\cdot;\cdot)$ være givet ved (2.2.8) og variansfunktionen $V(\cdot)$ ved (2.2.7).

Der gælder da, at variansfunktionen kan udtrykkes ved enhedsdeviansen som

$$V(\mu) = 2 \Big/ \left[\frac{\partial^2}{\partial \mu^2} d(x;\mu) \right]_{x=\mu} \tag{2.2.14}$$

og omvendt kan enhedsdeviansen udtrykkes som

$$d(x;\mu) = 2 \int_{\mu}^{x} \frac{x-u}{V(u)} \, du \quad . \tag{2.2.15}$$

for $x \in \mathcal{M}$ og $\mu \in \mathcal{M}$

□

Vi anfører endelig

Sætning 2.2.3 ***Tætheden for en naturlig eksponentiel familie udtrykt ved enhedsdeviansen***

Betragt en naturlig eksponentiel familie med kumulantfrembringeren $\kappa(\cdot)$.

Tætheden (2.2.1) for fordelingen svarende til parameteren μ kan da udtrykkes ved hjælp af enhedsdeviansen som

$$f(x;\mu) = a(x) \, \exp\left\{ -\frac{1}{2} d(x;\mu) \right\} \tag{2.2.16}$$

hvor normeringsfaktoren $a(x)$ er bestemt ved

$$a(x) = \exp\left[\sup_{\vartheta \in D} \{\vartheta x - \kappa(\vartheta)\} \right]$$

Bevis:
Se fx Jørgensen (1997). □

Eksempel 2.2.4 *Enhedsdeviansen for normalfordelingen*

Lad $X \in \mathrm{N}(\mu, 1)$ med $\mu \in \mathbb{R}$. Den kanoniske parameter er netop $\vartheta \equiv \mu$, og som vi ved, er familien netop parametriseret ved sin middelværdi, μ.

Enhedsdeviansen er

$$d(x;\mu) = (x-\mu)^2$$

og det sædvanlige udtryk for tætheden for fordelingen af X

$$f(x;\mu) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{(x-\mu)^2}{2} \right\}$$

er jo netop på formen (2.2.16). □

2.2.2 Eksponentielle dispersionsmodeller

De fleste af de fordelingsfamilier, der blev behandlet i Introduktion til Statistik, Bind 1 kan formuleres som én- eller todimensionale eksponentielle familier (se Oversigt over fordelinger med anvendelser i Statistik, IMM 1998).

Det er imidlertid ikke altid, at de to "naturlige" parametre i en todimensional eksponentiel familie er helt så naturlige i relation til det statistiske problem, der skal behandles. Når man vil modellere middelværdistrukturer for et datasæt, kan det tit være praktisk at separere parametrene i en strukturel parameter knyttet til middelværdien, og en parameter (indeks- eller skalaparameter), der udtrykker antal addender, skala, vægtning el lign.

Vi vil derfor udvide den naturlige eksponentielle familie til at omfatte en sådan indeks- eller skalaparameter. Da denne parameter har betydning for hvorledes variansfunktionen skal korrigeres for at udtrykke variansen (dispersionen) for observationen, kaldes de udvidede familier for eksponentielle dispersionsmodeller.

Definition 2.2.6 *Additiv eksponentiel dispersionsmodel*

Betragt en naturlig eksponentiel familie af fordelinger, og antag at familien har tætheder på formen

$$f_z(z;\vartheta) = c(z)\exp\{\vartheta z - \kappa(\vartheta)\} \tag{2.2.17}$$

for $\vartheta \in D$.

Lad mængden $\Lambda \subset \mathbb{R}$ være mængden af reelle tal, λ, for hvilke der findes en ikke-negativ funktion c^* på $\mathbb{R}$ sådan at udtrykket

$$f^*(z;\vartheta,\lambda) = c^*(z;\lambda)\exp\{\vartheta z - \lambda\kappa(\vartheta)\} \tag{2.2.18}$$

er tætheden for fordelingen af en stokastisk variabel.

Familien $\{f^*(\cdot;\vartheta,\lambda)\}_{(\vartheta,\lambda)\in D\times\Lambda}$ kaldes den additive eksponentielle dispersionsmodel frembragt af familien (2.2.17).

Parameteren ϑ kaldes den kanoniske parameter, parameteren λ kaldes indeksparameteren, og mængden Λ kaldes indeksmængden.

Mængden Λ er ikke tom; vi ved jo at $\lambda = 1$ tilhører mængden, idet (2.2.17) jo er en tæthed for en sandsynlighedsfordeling. □

Eksempel 2.2.5 ***Familien af binomialfordelinger som additiv eksponentiel dispersionsmodel***

Lad Z være Binomialfordelt, $Z \in \mathrm{B}(n,p)$. Fordelingen af Z har da tætheden

$$g(z;p,n) = \binom{n}{z}(1-p)^n[p/(1-p)]^z \quad \text{for} \quad z \in \{0,1,\ldots,n\}\,. \tag{2.2.19}$$

Sætter vi som i eksempel 2.2.1

$$\vartheta = \ln\{p/(1-p)\} \tag{2.2.20}$$

og

$$\kappa(\vartheta) = \ln(1+\exp(\vartheta))$$

ser vi, at tætheden (2.2.19) netop er på formen (2.2.18) med ϑ og $\kappa(\cdot)$ som ovenfor, og med indeksparameteren λ lig med antalsparameteren n, dvs. $\Lambda = \mathbb{N}$. Familien af $\mathrm{B}(n,p)$-fordelinger, $0 < p < 1$ og $n \in \mathbb{N}$ er således en additiv eksponentiel dispersionsmodel med kanonisk parameter $\vartheta = \ln(p/(1-p))$ og kumulantfrembringer $\kappa(\vartheta) = \ln(1+\exp(\vartheta))$.

Vi skal senere (i eksempel 2.2.8) diskutere familien af binomialfordelinger mere indgående. □

Definition 2.2.7 ***Reproduktiv eksponentiel dispersionsmodel***

Hvis vi - i definitionen af en additiv dispersionsmodel - havde formuleret kravet (2.2.18) udtrykt ved den variable $Y = Z/\lambda$ i stedet for Z, kalder vi den derved fremkomne familie af fordelinger for $Y = Z/\lambda$ for $(\vartheta, \lambda) \in \Theta \times \Lambda$ for den reproduktive eksponentielle dispersionsmodel frembragt af familien (2.2.17).

Tætheden for fordelingen af Y er på formen:

$$f(y;\vartheta,\lambda) = c(y;\lambda)\exp[\lambda\{\vartheta y - \kappa(\vartheta)\}]\,. \tag{2.2.21}$$

En reproduktiv eksponentiel dispersionsmodel parametriseres sædvanligvis ved parametrene μ og σ^2, hvor

$$\mu \overset{\text{DEF}}{=} \tau(\vartheta)\,, \quad \text{og} \quad \sigma^2 = 1/\lambda\,, \tag{2.2.22}$$

hvor, som sædvanligt, $\tau(\vartheta) = \kappa'(\vartheta)$.

For en reproduktiv eksponentiel dispersionsmodel kalder man parameteren μ for middelværdiparameteren, og $\sigma^2 = 1/\lambda$ kaldes dispersionsparameteren. □

Eksempel 2.2.6 ***Familien af Normalfordelinger som reproduktiv eksponentiel dispersionsmodel***

Lad Y være normalfordelt, $Y \in \mathrm{N}(\mu, \sigma^2)$.

Fordelingen af Y har da tætheden

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\,\sigma^2}}\,\exp\{-(y-\mu)^2/(2\sigma^2)\} \quad \text{for} \quad y \in \mathbb{R}\,.$$

Udtrykket for tætheden kan omskrives til

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\,\sigma^2}}\,\exp\Big\{-\frac{y^2}{2\sigma^2}\Big\}\,\exp\Big\{\frac{1}{\sigma^2}\Big(\mu y - \frac{\mu^2}{2}\Big)\Big\}\,, \tag{2.2.23}$$

med netop er på formen (2.2.21) med $\mu = \vartheta$ og $\sigma^2 = 1/\lambda$. og $\kappa(\vartheta) = \vartheta^2/2$. Idet alle ikke-negative værdier af σ^2 hører med til familien, har vi altså $\Lambda = \mathbb{R}_+$.

Familien af $\mathrm{N}(\mu, \sigma^2)$-fordelinger med $\mu \in \mathbb{R}$ og $1/\sigma^2 \in \mathbb{R}_+$ er altså en reproduktiv eksponentiel dispersionsmodel med kanonisk parameter $\vartheta = \mu$ og kumulantfrembringer $\kappa(\vartheta) = \vartheta^2/2$. Vi vil senere (i eksempel 2.2.9) behandle familien af normalfordelinger mere indgående. □

Eksempel 2.2.7 ***Fordelingen af empiriske varianser som reproduktiv eksponentiel dispersionsmodel***

Lad $X_1, X_2, \ldots, X_n$ være uafhængige med $X_i \in \mathrm{N}(\mu, \sigma^2)$ og betragt den empiriske varians,

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_\cdot)^2}{n-1}$$

med $\overline{X} = \sum_{i=1}^{n} X_i/n$.

Vi ved fra Introduktion til Statistik, Bind 1, at

$$S^2 \in \sigma^2\chi^2(f)/f$$

med $f = n-1$.

Men da denne fordeling er den samme som en $\mathrm{G}(f/2, \sigma^2/(f/2))$-fordeling (jvf. Introduktion til Statistik, Bind 1), har vi altså idet familien af gammafordelinger er en reproduktiv eksponentiel dispersionsmodel (se Oversigt over fordelinger med anvendelser i Statistik, IMM 1998), at

familien af fordelinger for S^2 er således en vægtet (reproduktiv) eksponentiel dispersionsmodel med middelværdiparameter

$$\mathsf{E}\,[S^2] = \sigma^2\ ,$$

variansfunktionen $V_G(\sigma^2) = (\sigma^2)^2$, dispersionsparameter 1, og med vægten $w = f/2$. Den kanoniske link svarende til familien er den reciprokke

$$\eta = 1/\sigma^2$$

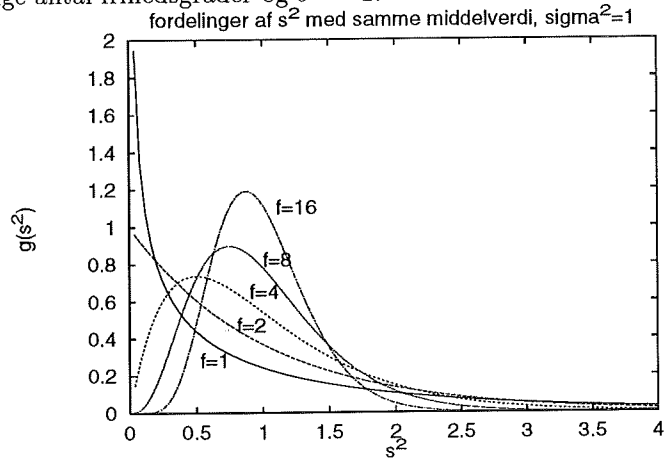
Egentlig kunne man vælge at fastsætte dispersionsparameteren til $2/f$, men sædvanligvis vælger man at benytte en vægtet model med dispersionsparameter 1, og vægten $f/2$.

Det følger af Oversigt over fordelinger med anvendelser i Statistik, IMM 1998, at der er en enentydig sammenhæng mellem variationskoefficienten i fordelingen og formparameteren, $f/2$, nemlig

$$\frac{2}{f} = \frac{\mathsf{V}\,[S^2]}{(\mathsf{E}\,[S^2])^2} = \mathsf{V}\,[S^2/\sigma^2] \qquad (2.2.24)$$

Fordelingen af den empiriske varians for normaltfordelte observationer, S^2, er således karakteriseret ved sin middelværdi (dvs variansen σ^2 i den underliggende normalfordeling af X'erne) og af variationskoefficienten i fordelingen af S^2.

Nedenstående figurer viser tæthederne svarende til fordelingen af S^2 for forskellige antal frihedsgrader og $\sigma^2 = 1$.



□

Bemærkning 1 ***Relation mellem additiv og reproduktiv dispersionsmodel***

Der er en stærk analogi mellem de additive og de reproduktive eksponentielle dispersionsmodeller.

Såfremt fordelingen af Z tilhører en additiv eksponentiel dispersionsmodel med kumulantfrembringeren κ og med parametrene ϑ og λ, da vil den tilsvarende fordeling af $Y = Z/\lambda$ tilhøre en reproduktiv eksponentiel dispersionsmodel, ligeledes med kumulantfrembringeren κ, og med parametrene $\mu = \tau(\vartheta)$ og $\sigma^2 = 1/\lambda$.

Det er dog ikke altid, der eksisterer en familie af fordelinger svarende til den reproduktive form af en additiv familie. Hvis fordelingen (i den additive form) af Z er diskret, er den tilsvarende reproduktive form uden mening (da

den jo skulle placere sandsynlighedsmassen på forskellige gitre afhængigt af værdien af indeksparameteren λ).

Tætheden for Z er af formen

$$f^*(z;\vartheta,\lambda) = c^*(z;\lambda)\exp\{\vartheta z - \lambda\kappa(\vartheta)\} \tag{2.2.25}$$

og tætheden for $Y = Z/\lambda$ er af formen

$$f(y;\vartheta,\lambda) = c(y;\lambda)\exp[\lambda\{\vartheta y - \kappa(\vartheta)\}] \tag{2.2.26}$$

Idet vi som vanligt indfører middelværdiafbildningen:

$$\tau(\vartheta) \stackrel{\text{DEF}}{=} \kappa'(\vartheta) \tag{2.2.27}$$

og enhedsvariansfunktionen

$$V(\mu) = \kappa''(\tau^{-1}(\mu)) = \tau'(\tau^{-1}(\mu)) \tag{2.2.28}$$

har vi, at såfremt fordelingen af Z kan beskrives ved en additiv eksponentiel dispersionsmodel med den kanoniske parameter ϑ og indeksparameter λ, da gælder

$$\mathsf{E}[Z] = \xi\,, \quad \text{og} \quad \mathsf{V}[Z] = \lambda V(\xi/\lambda) \tag{2.2.29}$$

med

$$\xi = \lambda\tau(\vartheta)\,, \tag{2.2.30}$$

hvor $V(\cdot)$ angiver enhedsvariansfunktionen, og $\tau(\cdot)$ angiver middelværdiafbildningen.

Sætter man i (2.2.29) $\mu = \tau(\vartheta)$, får man middelværdi og varians for Z udtrykt ved middelværdien μ af en "enhedsobservation",

$$\mathsf{E}[Z] = \lambda\mu\,, \quad \text{og} \quad \mathsf{V}[Z] = \lambda V(\mu) \tag{2.2.31}$$

med

$$\mu = \tau(\vartheta) \tag{2.2.32}$$

Tilsvarende har man, at såfremt fordelingen af Y kan beskrives ved en reproduktiv eksponentiel dispersionsmodel med kanonisk parameter ϑ og dispersionsparameter σ^2:

$$\mathsf{E}[Y] = \mu\,, \quad \text{og} \quad \mathsf{V}[Y] = \sigma^2 V(\mu) \tag{2.2.33}$$

med $\mu = \tau(\vartheta)$.

De to former, den additive og den reproduktive bestemmer samme eksponentielle dispersionsmodel. Det er imidlertid praktisk at operere med begge former. Den additive form er velegnet til repræsentation af diskrete fordelinger (man kan repræsentere deres tætheder med hensyn til tællemålet) og fordelinger, hvor summerne af de variable er af interesse (se sætning 2.2.4). Den reproduktive form er velegnet til behandling af fordelinger, hvor de vægtede gennemsnit er af størst interesse (se sætning 2.2.5).

Den reproduktive form parametriseres direkte ved middelværdien μ af den variable Y, men i den additive form er middelværdien af Z et multiplum, λ gange middelværdien μ af en "enhedsobsrvation". I den additive form indgår μ dog stadig som argument for enhedsvariansfunktionen. □

Bemærkning 2 ***Baggrund for betegnelsen "enhedsvariansfunktion"***

Grunden til at vi har indført betegnelsen "enhedsvariansfunktion" for den variansfunktionen, der er knyttet til kumulantfrembringeren, er, at variansen for Z (eller for Y) ikke er den samme funktion af middelværdien for forskellige værdier af indeksparameteren λ (eller σ^2). Den variansfunktion, der er knyttet til kumulantfrembringeren, spiller dog stadig en fundamental

rolle for bestemmelse af variansen som funktion af middelværdien, hvorfor det er praktisk at give den en speciel benævnelse. □

Eksempel 2.2.8 *Variansforholdene for familien af binomialfordelinger*

Lad $Z \in \mathrm{B}(n,p)$.

Vi så i eksempel 2.2.5, at familien af $\mathrm{B}(n,p)$-fordelinger, $0 < p < 1$ og $n \in \mathbb{N}$ er en additiv eksponentiel dispersionsmodel med kanonisk parameter $\vartheta = \ln(p/(1-p))$ kumulantfrembringer $\kappa(\vartheta) = \ln(1+\exp(\vartheta))$.

Middelværdiafbildningen er betemt ved

$$\tau(\vartheta) = \frac{\exp(\vartheta)}{1+\exp(\vartheta)} .$$

For $\vartheta = \ln(p/(1-p))$ finder vi netop

$$\tau(\ln[p/(1-p)]) = p .$$

Enhedsvariansfunktionen er givet ved

$$V(\mu) = \tau'(\tau^{-1}(\mu)) .$$

Idet

$$\tau'(\vartheta) = \frac{\exp(\vartheta)}{[1+\exp(\vartheta)]^2}$$

finder man enhedsvariansfunktionen

$$V(\mu) = \mu(1-\mu) ,$$

der for $\mu = p$ netop er det kendte udtryk, $p(1-p)$, for variansen i Bernoullifordelingen.

Vi får nu af (2.2.31) for $\lambda = n$, at $\mathsf{E}[Z] = \xi = np$ og

$$\mathsf{V}[Z] = nV(\xi/n) = nV(p) = np(1-p)$$

□

Eksempel 2.2.9 *Variansforholdene for familien af normalfordelinger*

Vi så i eksempel 2.2.6 at familien af $N(\mu, \sigma^2)$-fordelinger er en reproduktiv eksponentiel dispersionsmodel med kanonisk parameter $\vartheta = \mu$ og kumulantfrembringer $\kappa(\vartheta) = \vartheta^2/2$.

Middelværdifunktionen er netop

$$\tau(\vartheta) = \kappa'(\vartheta) = \vartheta$$

dvs middelværdien af Y er værdien af den kanoniske parameter, og variansfunktionen er

$$V(\mu) = \tau'(\tau^{-1}(\mu)) = 1$$

idet $\tau'(\vartheta) = 1$ for alle $\vartheta \in \mathbb{R}$.

Variansen for Y er (jvf. (2.2.33))

$$\mathsf{V}[Y] = \sigma^2 V(\mu) = \sigma^2$$

□

Sætning 2.2.4 ***Additionsegenskaber for additive eksponentielle dispersionsmodeller***

Lad $Z_1, Z_2, \ldots, Z_n$ være uafhængige og antag, at fordelingerne af Z_i kan beskrives ved fordelinger fra en additiv eksponentiel dispersionsmodel med samme værdi ϑ af den naturlige parameter, og med indeksparametre $\lambda_1, \lambda_2, \ldots, \lambda_n$ med $(\vartheta, \lambda_i) \in \Theta \times \Lambda$ for alle i.

Da vil fordelingen af

$$Z_+ = Z_1 + Z_2 + \cdots + Z_n$$

tilhøre samme familie med samme værdi ϑ af den naturlige parameter, og med indeksparameteren λ_+ givet ved

$$\lambda_+ = \lambda_1 + \lambda_2 + \cdots + \lambda_n$$

Middelværdien i fordelingen af Z_+ er $\xi_+ = \lambda_+ \tau(\vartheta)$ og variansen er

$$\mathsf{V}[Z_+] = \lambda_+ t V(\xi_+/\lambda_+)$$

Bevis:
Følger umiddelbart fra egenskaberne ved de tilsvarende naturlige eksponentielle familier. □

Sætning 2.2.5 ***Reproduktivitetsegenskaber for reproduktive eksponentielle dispersionsmodeller***

Lad $Y_1, Y_2, \ldots, Y_n$ være uafhængige og antag, at fordelingerne af Y_i kan beskrives ved fordelinger fra en reproduktiv eksponentiel dispersionsmodel med samme middelværdi μ og med dispersionsparametre $\sigma^2/w_1, \sigma^2/w_2, \ldots, \sigma^2/w_n$ med $(\mu, w_i/\sigma^2) \in \mathcal{M} \times \Lambda$ for alle i.

Da vil fordelingen af

$$\overline{Y}. = \sum_{i=1}^{n} w_i Y_i \Big/ \sum_{i=1}^{n} w_i$$

følge en reproduktiv eksponentiel dispersionsmodel med middelværdi μ og med dispersionsparameter

$$\sigma^2 \Big/ \sum_{i=1}^{n} w_i$$

Variansen i fordelingen af $\overline{Y}.$ er således

$$\mathsf{V}[\overline{Y}.] = \sigma^2 V(\mu) \Big/ \sum_{i=1}^{n} w_i$$

Bevis:
Følger af ovenstående sætning ved at benytte transformationen $Z_i = w_i Y_i/\sigma^2$. □

Definition 2.2.8 ***Enhedsdevians for en endimensional eksponentiel dispersionsmodel***

Ved enhedsdeviansen for en endimensional eksponentiel dispersionsmodel vil vi forstå enhedsdeviansen (definition 2.2.5) for den frembringende naturlige eksponentielle familie. Det vil sige, at hvis Y følger en reproduktiv eksponentiel dispersionsmodel med kumulantfrembringeren $\kappa(\cdot)$, da er enhedsdeviansen svarende til observationen y og en postuleret middelværdi, μ

$$d(y;\mu) \stackrel{\text{DEF}}{=} 2\{\sup_{\vartheta \in \Theta}[\vartheta y - \kappa(\vartheta)] - [y\tau^{-1}(\mu) - \kappa(\tau^{-1}(\mu))]\} \tag{2.2.34}$$

Såfremt specielt y ligger i middelværdirummet M, kan enhedsdeviansen (jvf (2.2.10) udtrykkes direkte som

$$d(y;\mu) = 2\left[y\{\tau^{-1}(y) - \tau^{-1}(\mu)\} - [\kappa\{\tau^{-1}(y)\} - \kappa\{\tau^{-1}(\mu)\}]\right]. \tag{2.2.35}$$

Når vi betragter observationer, z, svarende til additive dispersionsmodeller for fordelingen af Z, bruges den relative værdi, $y = z/\lambda$, som argument for enhedsdeviansen, svarende til at enhedsdeviansen måler afvigelsen imellem $y = z/\lambda$ og $\mu = \mathsf{E}[Z]/\lambda$. □

Resultaterne vedrørende devianser for naturlige eksponentielle familier, bemærkning 2 på side 131, samt sætning 2.2.2 gælder også for enhedsdeviansen svarende til en eksponentiel dispersionsmodel.

2.2.3 Oversigt over enhedsvariansfunktioner, dispersionsparametre og enhedsdevianser for sædvanlige eksponentielle dispersionsmodeller

Tabel 2.1 og 2.2 resumerer enhedsvariansfunktion og dispersionsparametre for eksponentielle dispersionsmodeller:

Tabel 2.1. **Enhedsvariansfunktionen $V(\mu)$ for reproduktive eksponentielle dispersionsmodeller**

Fordeling af Y	$\mathsf{E}[Y]$ μ	Variansfunktion Betegnelse	Dispersions-param. σ^2
$\mathrm{N}(\mu, \sigma^2)$	μ	$V_N(\mu) = 1$	σ^2
$\mathrm{G}(\alpha, \mu/\alpha)$	μ	$V_G(\mu) = \mu^2$	$1/\alpha$
$\mathrm{IG}(\mu, \lambda)$	μ	$V_{IG}(\mu) = \mu^3$	$1/\lambda$
$\mathrm{GHS}(\mu, \sigma^2)$	μ	$V_{GHS}(\mu) = 1 + \mu^2$	σ^2

Vi bemærker, at for disse familier gælder at variansen er et polynomium i højst tredie grad af middelværdien. Ses bort fra den inverse Gauss-fordeling, er variansen en (højst) kvadratisk funktion af middelværdien. De anførte seks familier med højst kvadratisk variansfunktion er de eneste endimensionale eksponentielle familier med denne egenskab. Morris (1982) resumerer en række egenskaber for disse seks familier.

For de sædvanlige modeller er enhedsdeviansen $d(y; \mu)$ angivet i tabel 2.3.

Tabel 2.2. **Enhedsvariansfunktionen $V(\mu)$ for additive eksponentielle dispersionsmodeller**

Fordeling af Z	Y	$\mathsf{E}[Y]$ μ	Variansfunktion Betegnelse	Indeks-param. λ
$\mathrm{P}(\mu)$	Z	μ	$V_P(\mu) = \mu$	†
$\mathrm{B}(n,p)$	Z/n	p	$V_{Bin}(\mu) = \mu(1-\mu)$	n
$\mathrm{NB}^*(\alpha,p)$	Z/α	$p/(1-p)$	$V_{NB}(\mu) = \mu(1+\mu)$	α
$\mathrm{G}(\alpha,\beta)$	Z/α	β	$V_G(\mu) = \mu^2$	α
$\mathrm{IG}(\lambda\mu,\lambda^2)$	Z/λ	μ	$V_{IG}(\mu) = \mu^3$	λ
† Middelværdi og indeksparameter kan ikke adskilles				

Tabel 2.3. **Enhedsdevians svarende til sædvanlige endimensionale fordelinger**

Fordeling af Y	$\mathsf{E}[Y]$ μ	Enhedsdevians $d(y;\mu)$
$\mathrm{N}(\mu,\sigma^2)$	μ	$(y-\mu)^2$
$\mathrm{P}(\mu)$	μ	$2 \times \{y\ln(y/\mu) - (y-\mu)\}$
$\mathrm{B}(n,p)/n$	p	$2 \times \{y\ln(y/\mu) + (1-y)\ln((1-y)/(1-\mu))\}$
$\mathrm{NB}^*(n,p)/n$	$p/(1-p)$	$2 \times \left\{y\ln\left(\frac{y(1+\mu)}{(1+y)\mu}\right) + \ln\frac{1+\mu}{1+y}\right\}$
$\mathrm{G}(\alpha,\mu/\alpha)$	μ	$2 \times \{y/\mu - \ln(y/\mu) - 1\}$
$\mathrm{IG}(\mu,\lambda)$	μ	$(y-\mu)^2/(y \times \mu^2)$
$\mathrm{GHS}(\mu,\sigma^2)$	μ	$2y\{\arctan(y) - \arctan(\mu)\} + \ln\left(\frac{1+\mu^2}{1+y^2}\right)$

Bemærk: Tabellinien svarende til binomialfordelingen og den negative binomialfordeling vedrører fordelingen af $Y = Z/n$, hvor $Z \in \mathrm{B}(n,p)$, hhv $Z \in \mathrm{NB}^*(n,p)$.

Bemærkning 1 *Momenter for enhedsdeviansen*

Ved benyttelse af bemærkning 2 på side 131 finder man, at såfremt fordelingen af Y tilhører en reproduktiv eksponentiel dispersionsmodel med enhedsdeviansen $d(y;\mu)$, variansfunktion $V(\mu)$ og dispersionsparameter σ^2 gælder

$$\mathsf{E}\left[\frac{\partial}{\partial\mu}\,d(Y;\mu)\right] = 0 \tag{2.2.36}$$

$$\mathsf{V}\left[\frac{\partial}{\partial\mu}\,d(Y;\mu)\right] = \frac{4}{\sigma^2 V(\mu)} \tag{2.2.37}$$

$$-\mathsf{E}\left[\frac{\partial^2}{\partial\mu^2}\,d(Y;\mu)\right] = \frac{2}{\sigma^2 V(\mu)} \tag{2.2.38}$$

Såfremt fordelingen af Z følger en additiv eksponentiel dispersionsmodel med indeksparameter λ gælder ovenstående udtryk med $\sigma^2 = 1$ for $Y = Z/\lambda$. Dog erstattes (2.2.37) af

$$\mathsf{V}\left[\frac{\partial}{\partial\mu}\,d(Z/\lambda;\mu)\right] = \frac{4}{\lambda V(\mu)} \tag{2.2.39}$$

□

2.2.4 Lidt om likelihoodfunktionen svarende til observationer fra eksponentielle dispersionsmodeller

Definition 2.2.9 *Vægtet model*

For reproduktive modeller kan man undertiden komme ud for at dispersionsparameteren svarende til den i'te observation er af formen $\sigma_i^2 = \sigma^2/w_i$, hvor størrelsen w_i er kendt, mens σ^2 er ukendt, dvs at variansen for Y_i er på formen:

$$\mathsf{V}[Y_i] = \frac{\sigma^2}{w_i}\,V(\mu_i) \tag{2.2.40}$$

Vi siger da, at vi har en vægtet model med middelværdi μ_i, variansfunktion $V(\cdot)$, dispersionsparameter σ^2 og vægten w_i.

Vægtningen er blot udtryk for at vi har trukket en kendt faktor ud af dispersionsparameteren. □

En sådan vægtning er netop i harmoni med reproduktivitetsegenskaberne for en reproduktiv familie jvf sætning 2.2.5.

Eksempel 2.2.10 *Vægtet normalfordelingsmodel*

Antag, at Y_i, $i = 1, 2, \ldots, k$ hver er gennemsnittet af n_i uafhængige, identisk fordelte størrelser, der hver for sig følger en $\mathrm{N}(\mu_i, \sigma^2)$-fordeling. Da vil fordelingen af Y_i være en $\mathrm{N}(\mu_i, \sigma^2/n_i)$-fordeling, dvs en vægtet model med middelværdi μ_i, variansfunktion $V(\mu) = 1$, dispersionsparameter σ^2 og vægt $w_i = n_i$. □

Det analoge resultat til sætning 2.2.3 er:

Sætning 2.2.6 *Tætheden for en eksponentiel dispersionsmodel udtrykt ved enhedsdeviansen*

Antage, at fordelingen af Y kan beskrives ved en reproduktiv eksponentiel dispersionsmodel parametriseret ved middelværdiparameteren μ og dispersionsparameteren σ^2 og med enhedsdeviansen $d(y;\mu)$. Da kan tætheden (2.2.26) udtrykkes som

$$f(y;\mu,\sigma^2) = a(y;\sigma^2)\exp\left\{-\frac{d(y;\mu)}{2\sigma^2}\right\}, \tag{2.2.41}$$

hvor

$$a(y;\sigma^2) = c(y;\sigma^{-2})\ \exp\left[\sigma^{-2}\sup_{\vartheta\in\Theta}\{\vartheta y - \kappa(\vartheta)\}\right]$$

Såfremt Y har vægten w, er tætheden

$$f(y;\mu,\sigma^2) = a(y;\sigma^2)\exp\left\{-\ w\ \frac{d(y;\mu)}{2\sigma^2}\right\},$$

Såfremt fordelingen af Z kan beskrives ved en additiv eksponentiel dispersionsmodel med middelværdien $\mathrm{E}[Z] = \xi$ og indeksparameter λ gælder

$$f(z;\xi,\lambda) = a^*(z;\lambda)\exp\left\{-\frac{\lambda}{2}\,d(z/\lambda;\xi/\lambda)\right\}, \tag{2.2.42}$$

hvor

$$a^*(z;\lambda) = c^*(z;\lambda)\ \exp\left[\lambda \sup_{\vartheta\in\Theta}\{\vartheta z - \kappa(\vartheta)\}\right]$$

Bevis:
Beviset for en reproduktiv familie følger af sætning 2.2.3

Beviset for en additiv familie følger dernæst ved at indføre parameteren $\xi = \lambda\mu$. □

Vi minder om, at indeksparameteren λ for en additiv familie kan tages som udtryk for antallet af observationer, og tilsvarende (jvf bemærkning 1 på side 137) at $\mu = \xi/\lambda$ er udtryk for "enhedsmiddelværdien", dvs middelværdien af $Y = Z/\lambda$.

Sætter man derfor $z/\lambda = y$ og $\xi/\lambda = \mu$ i (2.2.42), får man tætheden svarende til fordelingen af $Y = Z/\lambda$:

$$f^*(y;\mu,\lambda) = a^*(\lambda y;\lambda)\exp\left\{-\frac{\lambda}{2}\,d(y;\mu)\right\}. \qquad (2.2.43)$$

Bemærkning 1 ***En observations bidrag til log-likelihood funktionen***

Betragter vi nu en situation, hvor der foreligger et observationssæt $y_1,\ldots,y_k$, der kan opfattes som observationer af de uafhængige variable $Y_1,\ldots,Y_k$, hvor fordelingerne af $Y_1,\ldots,Y_k$ tilhører samme eksponentielle dispersionsmodel, men eventuelt med forskellige parametre $\mu_1,\ldots,\mu_k$.

Såfremt familien er en reproduktiv familie med dispersionsparameter σ^2, vil loglikelihoodfunktionen for $\mu_1,\ldots,\mu_k$ og σ^2 være en sum af bidrag af formen

$$\ell(y_i;\mu_i,\sigma^2) = c(y_i;\sigma^2) - \frac{1}{2\sigma^2}w_i d(y_i;\mu_i)\ , \qquad (2.2.44)$$

hvor w_i angiver den eventuelle vægt for den i'te observation, og $d(y;\mu)$ angiver enhedsdeviansen.

Såfremt familien er en additiv familie med indeksparameter λ, vil loglikelihoodfunktionen for $\mu_1,\ldots,\mu_k$ være en sum af bidrag af formen

$$\ell(y_i;\mu_i) = -\frac{1}{2}\,\lambda_i d(y_i;\mu_i) \qquad (2.2.45)$$

med $y = z/\lambda$ og $\mu = \mathsf{E}[Z]/\lambda$. Også her angiver $d(y;\mu)$ enhedsdeviansen. □

Definition 2.2.10 ***Devians og skaleret devians mellem observationssæt og model.***

Lad $y_1, y_2, \ldots, y_k$ være som i ovenstående bemærkning. Vi vil opfatte sættet af observationer opstillet som en søjlevektor (egentlig en søjlematrix af koordinater)

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{pmatrix} \tag{2.2.46}$$

og tilsvarende vil vi opfatte sættet af middelværdier $(\mu_1, \mu_2, \ldots, \mu_k)$ som en søjlevektor

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix} \tag{2.2.47}$$

Ved deviansen mellem observationerne $\mathbf{y}$ og modellen (karakteriseret ved middelværdierne $\boldsymbol{\mu}$) vil vi forstå udtrykket

$$D(\mathbf{y};\boldsymbol{\mu}) = \sum_{i=1}^{k} w_i d(y_i;\mu_i) \ , \tag{2.2.48}$$

hvor $d(y;\mu)$ angiver enhedsdeviansen svarende til den betragtede familie, og hvor w_i angiver vægten (for en vægtet reproduktiv familie), eller værdien af indeksparameteren λ (for en additiv familie).

Størrelsen

$$D^*(\mathbf{y};\boldsymbol{\mu}) = \frac{D(\mathbf{y};\boldsymbol{\mu})}{\sigma^2} \tag{2.2.49}$$

hvor σ^2 angiver værdien af dispersionsparameteren, kaldes den skalerede devians. □

Sætning 2.2.7 *Maksimum likelihood estimatet for $\mu_1, \ldots, \mu_k$ bestemmes ved at minimere deviansen*

Lad $\mathbf{y}$ være som i ovenstående bemærkning.

Maksimum likelihood estimatet for $\boldsymbol{\mu}$ fås da som det sæt af værdier, $\hat{\mu}_1, \ldots, \hat{\mu}_k$, der minimerer deviansen $D(\mathbf{y}; \boldsymbol{\mu})$

Bevis:
For en additiv familie fremgår det umiddelbart af udtrykket (2.2.45) for den i'te observations bidrag til loglikelihoodfunktionen, at likelihoodfunktionen for $\boldsymbol{\mu}$ netop er deviansen $D(\mathbf{y}; \boldsymbol{\mu})$.

For en reproduktiv familie med en eventuel ukendt dispersionsparameter, σ^2, vil loglikelihoodfunktionen for $\boldsymbol{\mu}$ og σ^2 også indeholde led af formen $c(y_i; \sigma^2)$. Disse led har imidlertid ingen indflydelse på bestemmelsen af maksimum med hensyn til μ'erne. Dette maksimum bestemmes ved at minimere $D(\mathbf{y}; \boldsymbol{\mu})$. □

Bemærkning 1 *Deviansen udtrykt ved variansfunktionen*

Ved at indsætte udtrykket (2.2.15) for enhedsdeviansen i (2.2.48) finder vi

$$D(\mathbf{y}; \boldsymbol{\mu}) = \sum_{i=1}^{k} w_i \int_{\mu_i}^{y_i} \frac{y_i - u}{V(u)}\, du \quad , \tag{2.2.50}$$

der i forbindelse med sætning 2.2.7 viser, at maksimum likelihood estimatet for $\mu_1, \ldots, \mu_k$ bestemmes ved at minimere den vægtede afvigelse mellem observationerne og modellen, vægtet med den reciproke varians, hvor vægtningen forløber glidende over hele intervallet mellem observation og model. □

Sætning 2.2.8 *Scorefunktionen for $\mu_1, \mu_2, \ldots, \mu_k$*

Lad $\mathbf{y}$ være som ovenfor.

Scorefunktionen (se def. 2.1.6 på side 114) for sættet $\mu_1, \mu_2, \ldots, \mu_k$ af middelværdier er

$$l'_{\mu}(\boldsymbol{\mu};\, \mathbf{y}) = \frac{1}{\sigma^2} \left(w_1 \frac{y_1 - \mu_1}{V(\mu_1)}, \ldots, w_k \frac{y_k - \mu_k}{V(\mu_k)} \right)^T \quad , \tag{2.2.51}$$

hvor w_i angiver de eventuelle vægte (eller værdier af indeksparameteren) og hvor faktoren $1/\sigma^2$ udelades for en additiv familie.

Bevis:
Beviset følger af udtrykket (2.2.11) for den afledede af enhedsdeviansen mht μ □

Sætning 2.2.9 *Den observerede information for* $\mu_1, \mu_2, \ldots, \mu_k$

Lad $\mathbf{y}$ være som ovenfor.

Den observerede information (se def. 2.1.7 side 116) svarende til observationen $\mathbf{y}$ og parametersættet $\boldsymbol{\mu}$ er

$$\mathbf{i}(\boldsymbol{\mu}; \mathbf{y}) = \operatorname{diag}\left\{ \frac{1}{V(\mu_i)} + (y_i - \mu_i)\frac{V'(\mu_i)}{V(\mu_i)^2} \right\}, \tag{2.2.52}$$

og den forventede information svarende til parametersættet $\boldsymbol{\mu}$ er

$$\mathbf{i}(\boldsymbol{\mu}) = \operatorname{diag}\left\{ \frac{1}{V(\mu_i)} \right\}, \tag{2.2.53}$$

hvor $\operatorname{diag}(a_i)$ betegner en diagonalmatrix med diagonalelementerne a_i.
Bevis:
Følger af (2.2.12). □

Bemærkning 1 *Maksimum likelihood estimation af dispersionsparameteren* σ^2

Vi bemærker, at en maksimum likelihood estimation af σ^2 kræver at man inddrager bidragene $c(y_i; \sigma^2)$ i (2.2.44) i maksimeringen. Vi vil senere diskutere alternative metoder til estimation af dispersionsparameteren. □

Definition 2.2.11 *Quasi-devians og quasi log-likelihood*

Det simple resultat i sætning 2.2.7 frister umiddelbart til at bruge en minimering af en "devians" som et almindeligt estimationsprincip.

Man kan konstruere en familie af fordelinger, hvor variansen er en fastlagt funktion $V(\cdot)$ af middelværdien og derefter i lighed med (2.2.50) definere en quasi-devians ved

$$d(y;\mu) = 2\int_{\mu}^{y} \frac{y-u}{V(u)}\,du \quad . \tag{2.2.54}$$

og derefter estimere $\boldsymbol{\mu}$ ved at minimere deviansen som i sætning 2.2.7.

En sådan størrelse kaldes en quasi-devians (*quasi=lige som*), og den tilsvarende størrelse $D(\mathbf{y};\boldsymbol{\mu})$ kaldes en quasi loglikelihood, da den jo ikke nødvendigvis indeholder al information om μ'erne. Med mindre familien er en naturlig eksponentiel familie vil den tilsvarende sandsynlighedstæthed jo indeholde en yderligere faktor, der fremkommer ved normering af quasilikelihood'en til en sandsynlighedstæthed, der integrerer til 1, og denne faktor vil i almindelighed afhænge af både μ og y. □

2.3 Linkfunktioner

Fil: glmlink.tex 1998-02-08

I behandlingen af de eksponentielle dispersionsmodeller har vi betragtet to forskellige parametriseringer af en familie af fordelinger, nemlig ved middelværdien og ved den kanoniske parameter.

Undertiden er man dog interesseret i at betragte en anden parametrisering, end disse to, nemlig når man vil beskrive indflydelsen fra forklarende variable som en lineær effekt. Det er ikke altid naturligt at beskrive middelværdien, eller den kanoniske parameter, som en lineær funktion af de forklarende variable.

I situationer, hvor interessevariablen er en sandsynlighed, p, hvis værdiområde er begrænset til intervallet $p \in]0,1[$, vil en lineær (egentlig affin) funktion $p = \beta_1 + \beta_2 x$ kunne føre til værdier af p, der ligger uden for intervallet $]0,1[$. En mulig transformation kunne her være den logistiske,

$$p = \frac{\exp(\beta_1 + \beta_2 x)}{1 + \exp(\beta_1 + \beta_2 x)} .$$

Såfremt observationerne kan beskrives ved $\mathrm{B}(n,p)$-fordelte variable, vil middelværdiparameteren μ netop være sandsynligheden p, og den logistiske transformation svarer da at modellere den kanoniske parameter $\vartheta = \ln(p/(1-p))$ ved en lineær funktion af x,

$$\vartheta = \beta_1 + \beta_2 x .$$

I visse (specielt medicinske og biologiske) sammenhænge er der imidlertid en tradition for at bruge den såkaldte probit transformation, dvs man modellerer p's afhængighed af den forklarende variable, x ved

$$p = \Phi(\beta_1 + \beta_2 x) ,$$

hvor $\Phi(\cdot)$ angiver den kumulerede fordelingsfunktion for den standardiserede normalfordeling.

For at kunne behandle også sådanne situationer indfører vi endnu et begreb, en linkfunktion.

Definition 2.3.1 ***Linkfunktion, prædiktor, kanonisk link***

Betragt en stokastisk variabel, Y, hvis fordeling kan beskrives ved en eksponentiel dispersionsmodel, parametriseret ved middelværdiparameteren μ.

Lad $\mathcal{M} = \tau(D) \subset \mathbb{R}$ angive middelværdirummet for familien.

En afbildning $g : \mathcal{M} \to \mathbb{R}$ med værdier

$$\eta = g(\mu) \tag{2.3.1}$$

kaldes en linkfunktion, værdierne η kaldes prædiktorer, og billedrummet $\mathcal{H} = g(\mathcal{M})$ kaldes prædiktorrummet.

Den omvendte funktion $g^{-1}(\cdot)$ til linkfunktionen beskriver hvorledes middelværdien afhænger af den lineære prædiktor

$$\mu = g^{-1}(\eta)$$

Såfremt linkfunktionen netop er afbildningen $\tau^{-1}(\cdot)$, der fører middelværdien over i den kanoniske parameter ϑ kaldes den for den kanoniske link.

□

Ved formuleringen af generaliserede lineære modeller er linkfunktionen den komponent af modellen, der specificerer hvilken funktion, $g(\cdot)$ af middelværdien, μ, man ønsker at modellere ved den lineære prædiktor $\eta_i = \mathbf{x}_i{}^T\boldsymbol{\beta}$.

Linkfunktionen $g(\cdot) : M \to \mathcal{H}$ angiver de komponentvise afbildninger

$$\eta_i = g(\mu_i)$$

Bemærkning 1 *Linkfunktioner og variansstabiliserende transformationer*

For god ordens skyld bemærker vi, at formålet med linkfunktionen er at angive den funktion af middelværdien, som man ønsker at modellere ved en lineær funktion.

Det er vigtigt at sondre mellem linkfunktionen, som er en funktion af middelværdien, og de såkaldte variansstabiliserende transformationer, som er

transformationer af observationerne. De variansstabiliserende transformationer har til formål at transformere observationerne til variable, der har nogenlunde samme varians, således at man kan bruge en simpel estimationsmetode, (oftest mindste kvadraters metode) til estimation af en lineær funktion af middelværdien af de transformerede observationer.

Linkfunktionen, derimod, ændrer ikke ved observationerne, men opererer alene paa middelværdien med henblik på at finde en passende lineær prædiktor. Observationernes forskellige varians bliver tilgodeset ved at maksimumlikelihood estimatet netop bestemmes ved at minimere deviansen, som jo afspejler observationernes varians. □

2.3.1 Sædvanlige linkfunktioner

Eksempel 2.3.1 *Almindeligt brugte linkfunktioner*

De almindeligt benyttede link-funktioner, $\eta = g(\mu)$, er angivet i nedenstående tabel

Benævnelse	link funktion $\eta = g(\mu)$	$\mu = g^{-1}(\eta)$
Identitet	μ	η
logaritme	$\ln(\mu)$	$\exp(\eta)$
logit	$\ln(\mu/(1-\mu))$	$\exp(\eta)/[1+\exp(\eta)]$
reciprok	$1/\mu$	$1/\eta$
potens	μ^k	$\eta^{1/k}$
kvadratrod	$\sqrt{\mu}$	η^2
probit	$\Phi^{-1}(\mu)$	$\Phi(\eta)$
log-log	$\ln(-\ln(\mu))$	$\exp(-\exp(\eta))$

□

Bemærkning 1 *Potenstransformationer og logaritmetransformation*

I nogle programsystemer optræder logaritmefunktionen som et specialtilfælde af potenstransformationen sådan at man fortolker

$$\eta = \mu^k = \begin{cases} \mu^k & \text{for } k \neq 0 \\ \log(\mu) & \text{for } k = 0 \end{cases} \tag{2.3.2}$$

Denne fortolkning er rent formel. Overgangen til udtrykket for $k = 0$ er ikke kontinuert i k. □

Eksempel 2.3.2 *Kanoniske linkfunktioner*

Den kanoniske (naturlige) linkfunktion er den linkfunktion, der fører middelværdien over i den kanoniske parameter. Når man vælger den kanoniske link i en generaliseret lineær model betyder det altså at man ønsker at konstruere lineære prædiktorer for de kanoniske parametre ϑ_i.

De kanoniske linkfunktioner svarende til de sædvanligt anvendte eksponentielle dispersionsmodeller er angivet i nedenstående tabel

Fordeling	Link: $\eta = g(\mu)$	Benævnelse
Normal	$\eta = \mu$	identitet
Poisson	$\eta = \ln(\mu)$	logaritme
Binomial	$\eta = \ln[\mu/(1-\mu)]$	logit
Gamma	$\eta = 1/\mu$	reciprok
Invers Gauss	$\eta = 1/\mu^2$	potens ($k = -2$)

Bemærk: I linien for binomialfordelingen svarer μ til forventningsværdien p af en enkelt elementarhændelse. □

Eksempel 2.3.3 *Linkfunktioner for binomialfordelte observationer*

Specielt for binomialt fordelte størrelser ser man en række forskellige linkfunktioner anvendt:

Lad $Z \in \mathrm{B}(n,p)$ med $p \in]0,1[$. Middelværdirummet for $Y = Z/n$ er netop intervallet $]0,1[$.

logit Den kanoniske link er logitfunktionen

$$\eta = \log[p/(1-p)]$$

probit I biologiske sammenhænge bruges undertiden probitfunktionen

$$\eta = u_p = \Phi^{-1}(p)$$

hvor $\Phi^{-1}(\cdot)$ angiver den inverse funktion til den kumulerede normalfordeling $p = \Phi(u)$.

log-log

$$\eta = \log[-\log(p)]$$

eller den komplementære log-log

$$\eta = \log[-\log(1-p)]$$

potenstransformationer

$$\eta = (p^k - 1)/k \quad \text{for } k \neq 0$$

med grænseværdien

$$\eta = \log(p) \quad \text{for } k \to 0$$

eller

$$\begin{aligned} \eta &= p^k \quad \text{for } k \neq 0 \\ \eta &= \log(p) \quad \text{for } k = 0 \end{aligned}$$

Selv om de to potenstransformationer ser næsten ens ud, har den første transformation, at overgangen til udtrykket for $k = 0$ er en kontinuert (i k) grænseovergang, i modsætning til den anden transformation (jvf. bemærkning 1 på side 154.

□

2.3.2 Illustration af afbildningerne ved forskellige link-funktioner

I de følgende afsnit vil vi illustrere afbildningerne ved nogle af disse link-funktioner.

Desuden illustreres effekten af at transformere den uafhængige variable, $x^* = h(x)$.

Betragt skemaet af link-funktioner i tabel 2.4:

Tabel 2.4. Oversigt over linkfunktioner, $\eta = g(\mu) = \alpha + \beta x^*$.

$g(\mu)$	transformation $x^* = h(x)$		
η	x	$1/x$	$\ln(x)$
μ	$\mu = \alpha + \beta x$	$\mu = \alpha + \beta/x$	$\mu = \alpha + \beta \ln(x)$
$1/\mu$	$1/\mu = \alpha + \beta x$	$1/\mu = \alpha + \beta/x$	$1/\mu = \alpha + \beta \ln(x)$
$\ln(\mu)$	$\ln(\mu) = \alpha + \beta x$	$\ln(\mu) = \alpha + \beta/x$	$\ln(\mu) = \alpha + \beta \ln(x)$
μ^k	$\mu^k = \alpha + \beta x$	$\mu^k = \alpha + \beta/x$	$\mu^k = \alpha + \beta \ln(x)$
$\ln\left(\frac{\mu}{1-\mu}\right)$	$\text{logit}(\mu) = \alpha + \beta x$	$\text{logit}(\mu) = \alpha + \beta/x$	$\text{logit}(\mu) = \alpha + \beta \ln(x)$

og de tilsvarende omvendte funktioner $\mu = g^{-1}(\alpha + \beta x^*)$

$g(\mu)$	transformation $x^* = h(x)$		
η	x	$1/x$	$\ln(x)$
μ	$\mu = \alpha + \beta x$	$\mu = \alpha + \beta/x$	$\mu = \alpha + \beta \ln(x)$
$1/\mu$	$\mu = \frac{1}{\beta(x + \alpha/\beta)}$	$\mu = \frac{1}{\alpha} - \frac{\beta}{\alpha(\alpha x + \beta)}$	$\mu = \frac{1}{\beta(\ln(x) + \alpha/\beta)}$
$\ln(\mu)$	$\mu = \gamma \exp(\beta x)$	$\mu = \gamma \exp(\beta/x)$	$\mu = \gamma x^\beta$
μ^k	$\mu = (\alpha + \beta x)^{1/k}$	$\mu = (\alpha + \beta/x)^{1/k}$	$\mu = [\alpha + \beta \ln(x)]^{1/k}$
$\ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\gamma \exp(\beta x)}{1 + \gamma \exp \beta x)}$	$\mu = \frac{\gamma \exp(\beta/x)}{1 + \gamma \exp(\beta/x)}$	$\mu = \frac{\gamma x^\beta}{1 + \gamma x^\beta}$

$\gamma = \exp(\alpha)$

Tabel 2.5. Transformationer af forklarende variable, og linkfunktioner

2.3.3 Hyperbelfunktioner

De tre funktioner

$$\begin{aligned} \mu &= \alpha + \beta/x \\ \mu &= \frac{1}{\beta(x + \alpha/\beta)} \\ \mu &= \frac{1}{\alpha} - \frac{\beta}{\alpha(\alpha x + \beta)} \end{aligned}$$

svarende til link= identitet og reciprok x, link = reciprok og lineær og reciprok x er eksempler på hyperbelfunktioner

$$(\mu - a)(x - b) = c \tag{2.3.3}$$

dvs

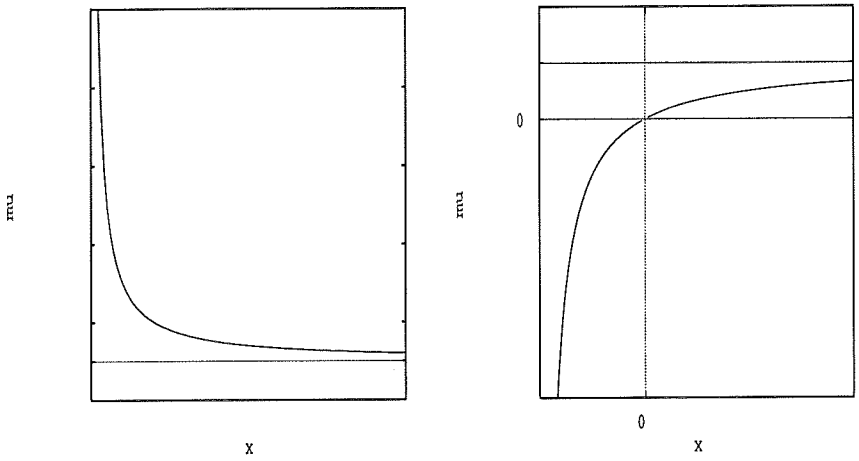
$$\mu = a + \frac{c}{x - b}$$

Ligningen (2.3.3) repræsenterer en hyperbel med asymptoterne $x = b$ og $\mu = a$ og krumningen c

De asymptotiske forhold for de tre funktioner er anført i nedenstående tabel:

Ligning	lodret asymptote	vandret asymptote	Krumning
$\mu = \alpha + \beta/x$	$x = 0$	$\mu = \alpha$	β
$1/\mu = \alpha + \beta x$	$x = -\frac{\alpha}{\beta}$	$\mu = 0$	$\frac{1}{\beta}$
$1/\mu = \alpha + \beta/x$	$x = -\frac{\beta}{\alpha}$	$\mu = \frac{1}{\alpha}$	$-\frac{\beta}{\alpha^2}$

Som illustration viser figur 2.1 to eksempler svarende til henholdsvis identitetslinken med reciprok x-transformation og reciprok link med reciprok x-transformation.



Figur 2.1. Eksempel på hyperbelfunktioner

$$\mu = \alpha + \beta/x, \text{ for } \beta > 0 \quad 1/\mu = \alpha + \beta/x, \text{ for } \beta > 0$$

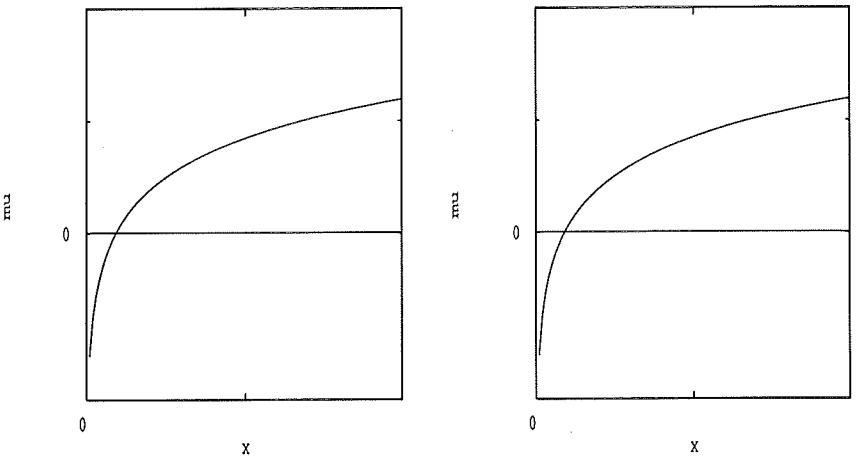
2.3.4 Logaritmefunktioner

Ved identitetslinken og logaritmisk x-transformation har man

$$\mu = \alpha + \beta \ln(x)$$

Kurven har μ-aksen som lodret asymptote, og kurven går gennem punktet $(1, \alpha)$. For $\beta > 0$ er kurven voksende mod ∞, og for $\beta < 0$ er kurven aftagende mod $-\infty$

Figur 2.2 viser eksempler på logaritmefunktionen



Figur 2.2. Eksempel på logaritmefunktionen med identitetslink

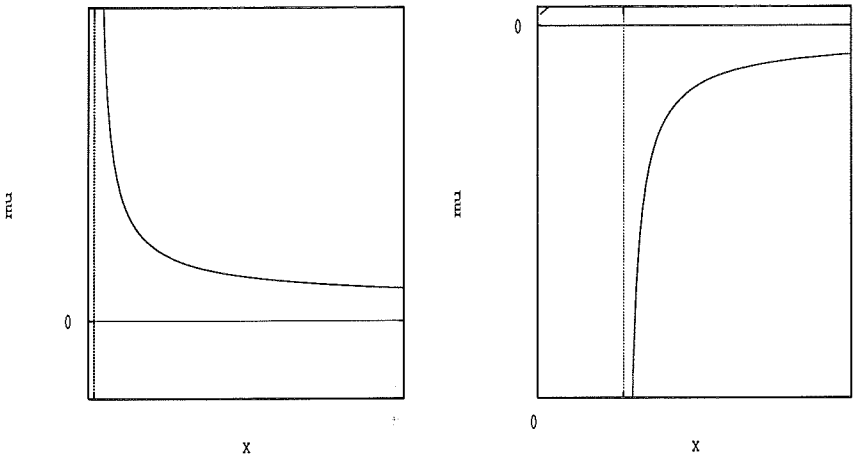
$$\mu = \alpha + \beta \ln(x), \text{ for } \beta > 0 \quad \mu = \alpha + \beta \ln(x), \text{ for } \beta < 0$$

Ved den reciproke link og logaritmisk x-transformation har man

$$\mu = \frac{1}{\alpha + \beta \ln(x)}$$

Kurven har linien $\mu = \exp(-\alpha/\beta)$ μ-aksen som lodret asymptote, og x-aksen som vandret asymptote. For $\beta > 0$ går kurven gennem punktet $(1, 1/\alpha)$ og går aftagende mod 0. For $\beta < 0$ er kurven voksende mod 0.

Figur 2.3 viser eksempler på logaritmefunktionen



Figur 2.3. Eksempel på logaritmefunktionen ved reciprok link

$$\mu = \frac{1}{\alpha + \beta \ln(x)}, \text{ for } \beta > 0 \quad \mu = \frac{1}{\alpha + \beta \ln(x)}, \text{ for } \beta < 0$$

2.3.5 Eksponentialfunktioner

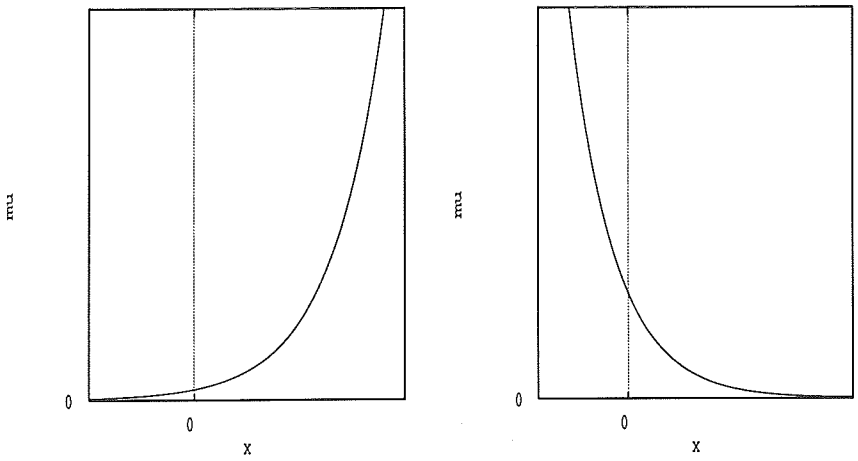
Ved logaritmisk link og lineær x transformation har man

$$\mu = \gamma \exp(\beta x) \ , \quad (-\infty < x < \infty)$$

med $\gamma = \exp(\alpha)$.

Kurven passerer gennem punktet $(0, \gamma)$ og har x-aksen som asymptote.

Figur 2.4 viser eksempler på eksponentialfunktionen svarende til lineær x-transformation



Figur 2.4. Eksempel på lineær x-transformation med logaritmisk link

$$\mu = \gamma \exp(\beta x), \text{ for } \beta > 0 \quad \mu = \gamma \exp(\beta x), \text{ for } \beta < 0$$

Ved logaritmisk link og reciprok x transformation har man

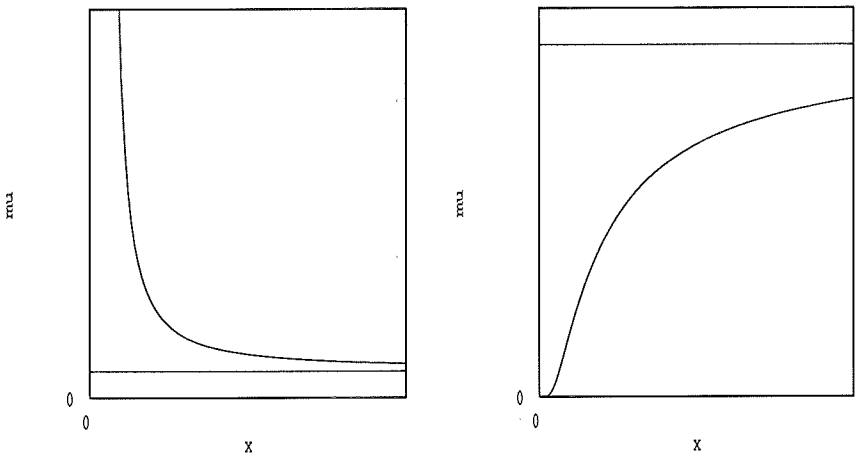
$$\mu = \gamma \exp(\beta / x) \ , \quad (0 < x < \infty)$$

med $\gamma = \exp(\alpha)$.

For $\beta > 0$ er kurven aftagende med μ-aksen og linien $\mu = \gamma$ som asymptoter.

For $\beta < 0$ er funktionen voksende. Grafen går gennem nulpunktet, har vendetangent i punktet $(\beta/2, \gamma \exp(-2))$ og har linien $\mu = \gamma$ som asymptote.

Figur 2.5 viser eksempler på eksponentialfunktionen svarende til reciprok x-transformation



Figur 2.5. Eksempel på reciprok x-transformation og logaritmisk link

$$\mu = \gamma \exp(\beta/x), \text{ for } \beta > 0 \quad \mu = \gamma \exp(\beta/x), \text{ for } \beta < 0$$

2.3.6 Potensfunktioner

Ved logaritmisk link og logaritmisk x transformation har man

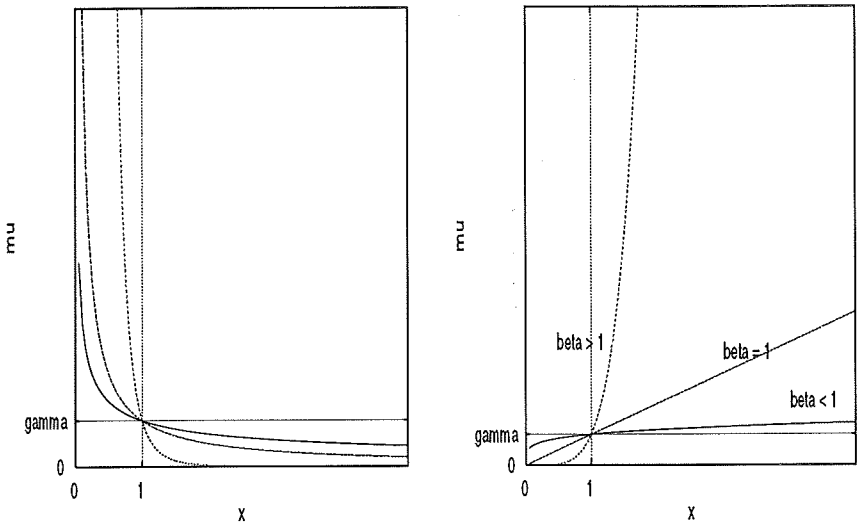
$$\mu = \gamma x^{\beta} , \quad (0 < x < \infty)$$

Kurven går igennem punktet $(x, \mu) = (1, \gamma)$.

For $\beta < 0$ har kurven μ−aksen som lodret asymptote for $x \to 0$ og x-aksen (linien $\mu = 0$) som asymptote for $x \to \infty$.

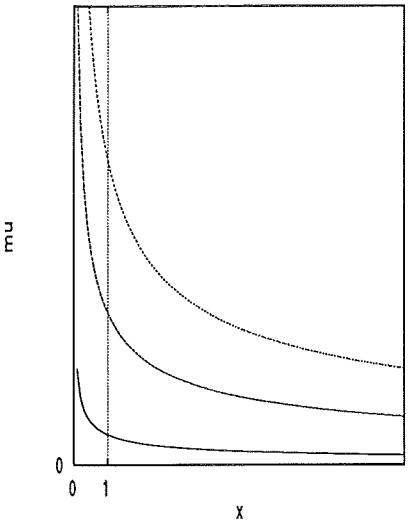
For $\beta > 0$ vil $\mu \to \infty$ for $x \to \infty$.

Figur 2.6 og 2.7 viser eksempler paa logaritmisk x-transformation sammen med en logaritmisk linkfunktion.



$\mu = \gamma x^{\beta}$ for $\beta < 0$ $\mu = \gamma x^{\beta}$ for $\beta > 0$

Figur 2.6. Eksempel på logaritmisk x-transformation med logaritmisk link for forskellige værdier af β



$\mu = \gamma x^{\beta}$ for $\beta < 0$

Figur 2.7. Eksempel på logaritmisk x-transformation med logaritmisk link for forskellige værdier af γ; $\beta < 0$

2.4 Generaliserede lineære modeller

Fil: glm2b.tex 98-02-15

2.4.0 Indledning

De følgende afsnit er viet til analysen af de såkaldte generaliserede lineære modeller.

Afsnittene er struktureret sådan at vi først introducerer modelklassen, derefter i afsnit 2.5 beskrives de generelle resultater vedrørende estimation, herunder iterative metoder til bestemmelse af estimaterne, den approximative fordeling af estimater, metoder til test af specifike værdier af enkelte parametrer samt en diskussion af forskellige mål (residualer) for enkeltobservationers afvigelse fra modellen.

I afsnit 2.6 beskrives test for modeltilpasning samt estimation af den såkaldte dispersionsparameter i de tilfælde, hvor en sådan parameter skal estimeres.

I afsnit 2.7 eksemplificeres den generelle teori ved at betragte homogenitetstest og simple regressionsmodeller.

De - lidt abstrakte - afsnit 2.8 og 2.9 beskriver forskellige varianter af parametrisering af disse modeller og afsnit 2.10 beskriver den såkaldte Wilkinson notation, som bruges i nyere programsystemer.

Endelig i afsnit 2.11 diskuteres test for modelreduktion. Selv om disse test er en vigtig bestanddel af enhver analyse har jeg fundet det praktisk først at beskrive modelstrukturerne i afsnittene 2.8 og 2.9 før indførelsen af disse test.

Afsnittene 2.12 og 2.13 beskriver generaliseringen af den tosidede variansanalysemodel fra Introduktion til Statistik, Bind 1til disse dispersionsmodeller og endelig sluttes af i afsnittene 2.14 og 2.15 med forskellige betragtninger om modelreduktion.

2.4.1 Definition af en generaliseret lineær model

Definition 2.4.1 *Generaliseret lineær model*

Antag, at fordelingen for $Y_1, Y_2, \ldots, Y_k$ kan beskrives ved at $Y_1, Y_2, \ldots, Y_k$ er uafhængige variable, hvis fordelinger kan beskrives ved en eksponentiel

dispersionsmodel med samme variansfunktion, $V(\mu)$. Såfremt fordelingen af Y_i beskrives ved en reproduktiv model antager vi desuden, at fordelingerne har samme dispersionsparameter σ^2.

En generaliseret lineær model for $Y_1, Y_2, \ldots, Y_k$ er en affin hypotese vedrørende værdierne $\eta_1, \eta_2, \ldots, \eta_k$ af en transformation,

$$\eta_i = g(\mu_i)$$

af middelværdierne $\mu_1, \mu_2, \ldots, \mu_k$. Hypotesen er af formen

$$H_0 \; : \; \boldsymbol{\eta} - \boldsymbol{\eta}_0 \in L \, , \tag{2.4.1}$$

hvor L er et lineært underrum af $\mathbb{R}^k$ af dimension m, og hvor $\boldsymbol{\eta}_0$ angiver en vektor af kendte offsetværdier. □

Definition 2.4.2 ***Fuld model svarende til generaliseret lineær model***

Betragt en generaliseret lineær model for $Y_1, \ldots, Y_k$. Den uindskrænkede model, der tillader alle μ_i at variere frit i deres værdiområde, kaldes den fulde, eller mættede model (eng: *saturated*). Dimensionen af den fulde model er antallet af observationer, k.

Under den fulde model er estimatet for μ_i blot observationen y_i.

□

De væsentligste egenskaber ved generaliserede lineære modeller afhænger ikke af den specifikke parametrisering af modellen, og vi har derfor valgt en koordinatuafhængig formulering i definitionen af en generaliseret lineær model.

Definition 2.4.3 ***Dimension af generaliseret lineær model***

Betragt den generaliserede lineære model givet ved (2.4.1). Dimensionen m af underrummet L kaldes for modellens dimension. □

Definition 2.4.4 *Modelmatrix for generaliseret lineær model*

Betragt en generaliseret lineær model med en hypotese af formen (2.4.1).

Såfremt der er valgt en basis for L, sådan at hypotesen kan udtrykkes på formen

$$\boldsymbol{\eta} - \boldsymbol{\eta}_0 = \mathbf{X}\boldsymbol{\beta} \quad \text{med } \boldsymbol{\beta} \in \mathbb{R}^m \,, \tag{2.4.2}$$

hvor $\mathbf{X}$ har fuld rang, kaldes matricen $\mathbf{X}$ for hypotesens modelmatrix.

Vektoren

$$\mathbf{x}_i^* = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{im} \end{pmatrix} \,, \tag{2.4.3}$$

hvis elementer er den i'te række i modelmatricen, kaldes modelvektoren for den i'te observation. □

Bemærkning 1 *Hypotesen udspænder et (side)underrum i* $\mathbb{R}^k$

Ved mange sædvanlige hypoteser er offset værdien $\mathbf{0}$, og hypotesen udtrykker, at prædiktoren $\boldsymbol{\eta}$ ligger i et m-dimensionalt underrum af $\mathbb{R}^k$, udspændt af søjlerne i $\mathbf{X}$.

Hvis offsetværdien $\boldsymbol{\eta}_0$ er forskellig fra nulvektoren, specificerer hypotesen, at prædiktoren $\boldsymbol{\eta}$ ligger i et m dimensionalt sideunderrum i $\mathbb{R}^k$. □

Bemærkning 2 *Formulering af hypotese ved lineære bånd.*

Vi bemærker, at en hypotese af formen (2.4.2) alternativt kunne specificeres ved

$$\mathbf{M}\boldsymbol{\eta} = \mathbf{M}\boldsymbol{\eta}_0 \tag{2.4.4}$$

hvor $\mathbf{M}$ er en $(k-m) \times m$-dimensional matrix af fuld rang.

Rummet udspændt af (2.4.2) er jo netop løsningsmængden svarende til et ligningssystem af formen (2.4.4). □

Bemærkning 3 *Geometrisk fortolkning*

Vi bemærker, at når parameteren β gennemløber parameterrummet $\beta \in B \subset \mathbb{R}^m$, da vil de tilsvarende værdier af den lineære prædiktor $\eta = h(\beta)$ beskrive en (evt parallelforskudt) m-dimensional hyperplan, $\eta(B)$ i $\mathbb{R}^m$. Hyperplanen er udspændt af søjlerne i $\mathbf{X}$. □

Bemærkning 4 *Komponenter i en generaliseret lineær model*

En generaliseret lineær model for observationerne $Y_1, Y_2, \ldots, Y_k$ er således karakteriseret ved

Variansfunktionen $V(\mu)$, der beskriver hvorledes variansen ændrer sig med middelværdien, μ,

En lineær komponent, den lineære prædiktor, der beskriver en lineær afhængighed af de forklarende variable

$$\eta = (\mathbf{x}^*)^T \beta$$

Link funktionen $g(\cdot)$, der beskriver hvilken funktion af forventningsværdien, der beskrives ved den lineære prædiktor

$$g(\mu) = (\mathbf{x}^*)^T \beta$$

Evt. vægte for en reproduktiv model, eller værdier af indeksparameteren for en additiv model.

Ad variansfunktion:
Udover variansfunktionen $V(\cdot)$ indgår evt dispersionsparameteren σ^2 og vægten w_i for den i'te observation til beskrivelse af variansforholdene. Dispersionsparameteren kan være kendt, eller den skal eventuelt estimeres fra data.

Ad linkfunktionen:

Linkfunktionen, $g(\cdot)$ sammenknytter den lineære prædiktor η_i med middelværdiparameteren $\mu_i = \mathrm{E}[Y_i]$ ved relationen

$$\eta_i = g(\mu_i) \tag{2.4.5}$$

Den omvendte afbildning $g^{-1}(\cdot)$ udtrykker forventningsværdien μ ved den lineære prædiktor η:

$$\mu = g^{-1}(\eta)$$

dvs

$$\mu_i = g^{-1}(\mathbf{x}_i{}^T\boldsymbol{\beta}) = g^{-1}(\sum_j x_{ij}\beta_j) \tag{2.4.6}$$

I afsnit 2.3 betragtede vi forskellige linkfunktioner. De almindeligt anvendte linkfunktioner var angivet i 2.3.1, og eksempel 2.3.2 gav en oversigt over de kanoniske linkfunktioner, og endelig i eksempel 2.3.3 gav vi en oversigt over linkfunktioner, der bruges i forbindelse med binomialt fordelte observationer.

Ad vægte:
For reproduktive modeller kan man undertiden komme ud for at dispersionsparameteren svarende til den i'te observation er af formen $\sigma_i^2 = \sigma^2/w_i$, hvor w_i er kendt, mens σ^2 er ukendt, se side 145. Vægtene spiller samme rolle som indeksparameteren i en additiv model. □

Definition 2.4.5 ***Lokal design matrix***

Betragt en generaliseret lineær model med modelmatricen $\mathbf{X}$ og variansfunktionen $V(\cdot)$.

De koordinatvise afbildninger

$$\mu_i = g^{-1}(\eta_i)$$

sammen med den lineære afbildning (2.4.2)

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

definerer $\boldsymbol{\mu}$ som en funktion af $\boldsymbol{\beta}$. Funktionen er en vektorfunktion $\mathbb{R}^m \to \mathbb{R}^k$.

Matricen

$$\mathbf{X}(\beta) \stackrel{\text{DEF}}{=} \frac{\partial \mu}{\partial \beta} = \left[\frac{\mathrm{d}\mu}{\mathrm{d}\eta}\right]^T \frac{\partial \eta}{\partial \beta} = \operatorname{diag}\left\{\frac{1}{g'(\mu_i)}\right\}\mathbf{X} \tag{2.4.7}$$

kaldes den lokale design matrix svarende til parameterværdien β

I udtrykket (2.4.7) angiver $\operatorname{diag}\{1/g'(\mu_i)\}$ en $k \times k$ dimensional diagonalmatrix, hvis i'te diagonalelement er $1/g'(\mu_i)$, hvor μ_i bestemt ved (2.4.6), dvs. som middelværdien af den i'te observation under hypotesen (2.4.2) svarende til parameterværdien β, dvs

$$\mu_i = \mu_i(\beta) \stackrel{\text{DEF}}{=} \mu(g^{-1}(\eta_i)) = \mu(g^{-1}(\mathbf{x}_i^*\beta)) \,, \tag{2.4.8}$$

hvor $\mathbf{x}_i^*$ er modelvektoren for den i'te observation givet ved (2.4.3)
For den kanoniske link er $g(\cdot) = \tau^{-1}(\cdot)$.

I tilfældet med den kanoniske link gælder derfor

$$g'(\mu) = \frac{1}{\tau'(\tau^{-1}(\mu))} = \frac{1}{V(\mu)} \tag{2.4.9}$$

□

Bemærkning 1 ***Fortolkning af lokal design matrix***

Den lokale design matrix $\mathbf{X}(\beta)$ udtrykker den lokale (for den aktuelle værdi af β) reskalering af modelmatricen, $\mathbf{X}$, der er nødvendig for at tilgodese ikke-lineariteten i transformationen fra den lineære prædiktor η til μ. □

Bemærkning 2 ***Den lokale design matrix ved kanonisk link***

Såfremt linkfunktionen netop er den kanoniske link, dvs $g(\mu) = \tau^{-1}(\mu)$ bliver

$$g'(\mu) = \frac{1}{V(\mu)} \,,$$

jvf (2.4.9), hvorfor vi har

$$\left[\frac{\mathrm{d}\mu}{\mathrm{d}\eta}\right] = \operatorname{diag}\left\{\frac{1}{g'(\mu_i)}\right\} = \operatorname{diag}\{V(\mu_i)\} \,,$$

således at den lokale designmatrix bliver

$$\mathbf{X}(\beta) = \operatorname{diag}\{V(\mu_i)\}\mathbf{X} \tag{2.4.10}$$

□

2.4.2 Eksempel på generaliserede lineære modeller

Eksempel 2.4.1 ***Fosterdødelighed hos mus***

Data fra C.J.Price, C.A.Kimmel, J.D.George and M.C.Marr. The developmental toxity of diethylene glycol dimethyl ether in mice. *Fund. Appl. Toxicol.* **8**, (1987), pp. 115-126.

I et eksperiment til vurdering af toxiteten af et industrielt opløsningsmiddel, diethylene glycol dimethyl æter, (diEGdiME) udvalgtes en række gravide mus, som i de ti første dage af graviditeten dagligt fik injiceret en dosis (diEGdiME). To dage senere undersøgte man fostrene.

Antallet af døde fostre som funktion af den indgivne koncentration er vist i nedenstående tabel:

Indeks i	Antal døde z_i	Antal fostre n_i	Indgiven concentration [mg/kg pr. dag] x_i
1	15	297	0
2	17	242	62.5
3	22	312	125
4	38	299	250
5	144	285	500

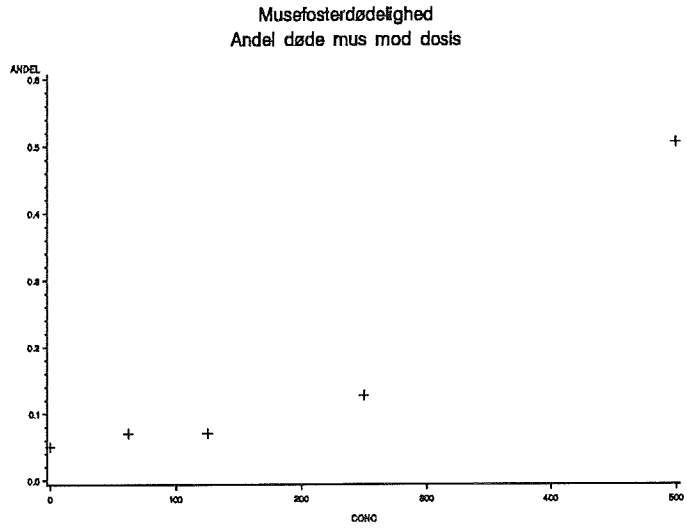
Vi modellerer andelen af døde fostre Y_i, svarende til det i'te forsøg ved $Y_i = Z_i/n_i$, hvor $Z_i \in \mathrm{B}(n_i, p_i)$, og vi antager at de enkelte forsøg er uafhængige.

Den fulde model svarende til observationssættet er modellen $Y_i = Z_i/n_i$, hvor $Z_i \in \mathrm{B}(n_i, p_i)$ og

$$H_F: \qquad p_i \in]0,1[\,, \; i = 1, 2, \ldots, 5$$

Modellen har dimensionen 5, idet modellen udsiger, at sættet $(p_1, \ldots, p_5)$ kan varierere frit i $]0,1[^5$.

Vi vil imidlertid reducere modellen ved at beskrive dødssandsynligheden p_i som funktion af den indgivne koncentration, x_i. Nedenstående figur viser de observerede andele døde mus tegnet op mod den indgivne koncentration. Det er klart, at en direkte lineær model, $p_i = \beta_1 + \beta_2 x_i$, ikke vil være særlig tilfredsstillende for dødssandsynligheder nær ved nul eller én.



Vi vil derfor vælge en link-funktion, og vi forsøger med den kanoniske link, dvs. vi betragter funktionen

$$g(p) = \ln\left(\frac{p}{1-p}\right)$$

og sætter

$$\eta_i = \ln\left(\frac{p_i}{1-p_i}\right)\ ,\ i = 1, 2, \ldots, 5\ .$$

Transformationen $\eta(\mu) = \ln\{\mu/(1-\mu)\}$ kaldes logittransformationen. Denne transformation fører netop middelværdien p i $\mathrm{B}(n,p)/n$ fordelingen over i den kanoniske parameter ϑ.

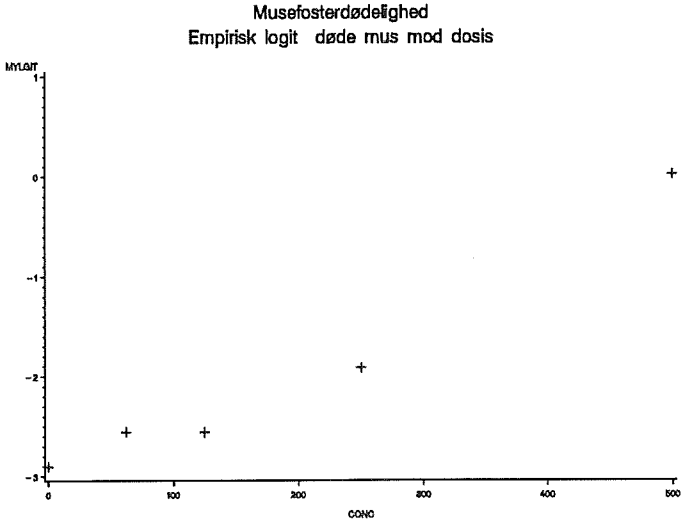
Som en indledende analyse til støtte for valget af linkfunktion kan man optegne de empiriske logit'er,

$$\ell_i = \ln\left(\frac{z_i + 1/2}{n_i - z_i + 1/2}\right)$$

mod x for at vurdere, om der kan tænkes at være en lineær sammenhæng.

Som anført, adderer man ofte 1/2 i brøkens tæller og nævner for at undgå problemer ved $z_i = 0$ og $z_i = n_i$.

Nedenstående figur viser de empiriske logit'er tegnet op mod den indgivne koncentration. Umiddelbart kan man ikke afvise at der er en lineær sammenhæng mellem de underliggende værdier af $\ln(p_i/(1-p_i))$ og den indgivne koncentration.



Vi formulerer derfor hypotesen

$$H_0: \qquad \eta_i = \beta_1 + \beta_2 x_i \ , \ i = 1, 2, \ldots, 5 \ , \tag{2.4.11}$$

hvor x_i angiver den indgivne koncentration i [mg/kg pr. dag].

Hypotesen (2.4.11) sammen med modelantagelserne $Y_i = Z_i/n_i$, hvor $Z_i \in \mathrm{B}(n_i, p_i)$, definerer en generaliseret lineær model med variansfunktionen $V(\mu) = \mu(1-\mu)$, indeksparameter (vægt), n_i, og linkfunktion den logistiske link. Modellen har dimensionen 2, idet hypotesen H_0 kan skrives på formen

$$\begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_5 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 62.5 \\ 1 & 125 \\ 1 & 250 \\ 1 & 500 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

dvs. netop af formen (2.4.2) med $\mathbf{X}$ givet ved den 5×2-dimensionale matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 62.5 \\ 1 & 125 \\ 1 & 250 \\ 1 & 500 \end{pmatrix}$$

der har rangen 2, dvs den udspænder et underrum L af dimension 2 i det 5-dimensionale rum $\mathbb{R}^5$

Modellen

$$H_0 : \; p_i = \exp(\alpha + \beta x_i)/\{1 + \exp(\alpha + \beta x_i)\}, \quad i = 1, 2, \ldots, k$$

kaldes den logistiske regressionsmodel. Estimation i denne model vil blive diskuteret i eksempel 2.5.2 og 2.6.1 samt i afsnit 3.2. □

Det følgende eksempel illustrerer formuleringen af en generaliseret lineær model svarende til den såkaldte tosidede variansanalysemodel i Introduktion til Statistik, Bind 1. Vi skal senere i afsnit 2.13 vende tilbage til en generisk (dvs for vilkårlige eksponentielle dispersionsmodeller) formulering af en sådan model.

Eksempel 2.4.2 ***Udplantning af blommestiklinger***

Vi betragter data fra et eksperiment, der havde til formål at vurdere forskellige faktorers betydning for blommestiklingers robusthed. (Hoblyn og Palmer 1934)

Vi betragter her et delforsøg, der havde til formål at vurdere hvorledes overlevelsesevnen afhænger af tykkelsen og længden af de udplantede stiklinger.

Ved forsøget undersøgte man to længder, henholdsvis 6 [cm] og 12 [cm] og tre tykkelsesgrader, 3-6 [mm], 6-9 [mm] og 9-12 [mm]. I oktober udplantede man 20 stiklinger af hver kombination af længde og tykkelse, og året efter vurderede man, hvor mange af disse, der stadig var i live.

Resultaterne er angivet i nedenstående tabel:

Obs nr	Længde	Tykkelse	Antal udplantede	Antal levende	Antal døde
1	LA	TYN	20	6	14
2	LA	MID	20	14	6
3	LA	TYK	20	18	2
4	KO	TYN	20	4	16
5	KO	MID	20	10	10
6	KO	TYK	20	11	9

hvor symbolerne LA og KO angiver hhv. lange og korte stiklinger, og TYN, MID og TYK angiver tynde, middel,og tykke stiklinger.

Der er åbenbart de to forklarende variable, længde og tykkelse. Vi vil opfatte begge variable som klassifikationsvariable (faktorvariable). Længden har de to niveauer {KO, LA} svarende til kort (6 [cm]) og lang (12 [cm]). Tykkelsen har de tre niveauer {MID, TYK, TYN} svarende til middel (6-9 [mm]), tyk (9-12 [mm]) og tynd (3-6 [mm]), .

Ved den parametriske repræsentation vil vi tilgodese den faktorielle struktur af de forklarende variable, og vi vil derfor parametrisere observationerne svarende til skemaet:

Antal overlevende/antal udplantede ved blommeudplantning

	tykkelse		
Længde	MID	TYK	TYN
KO	10/20	11/20	4/20
LA	14/20	18/20	6/20

dvs vi vil benævne observationerne Z_{ij}, $i = 1, 2$, $j = 1, 2, 3$, hvor i angiver indeks for længden, og j angiver indeks for tykkelsen

Vi vælger at modellere antallet, Z_{ij}, af overlevende stiklinger ved en $\mathrm{B}(n_{ij}, p_{ij})$-fordelt størrelse, hvor antallet af stiklinger, $n_{ij} = 20$.

Familierne af fordelinger af $Y_{ij} = Z_{ij}/n_{ij}$ er eksponentielle dispersionsmodeller med middelværdien p_{ij}, med den kanoniske link funktion

$$\eta = \ln\left(\frac{p}{1-p}\right)$$

variansfunktionen

$$V(p) = p(1-p)$$

og indeksparameter (vægt) $w_{ij} = n_{ij}$.

Vi vælger at betragte den kanoniske link, og formulerer hypotesen svarende til forsvindende vekselvirkning i analogi med modellen for den tosidede variansanalyse i Statistik 1.

$$H_0: \quad \eta_{ij} = \mu + \alpha_i + \gamma_j \tag{2.4.12}$$

Når vi vil formulere modellen ved en modelmatrix, skal vi sikre, at matricen har fuld rang. Der er derfor ikke plads til to længde-parametre og tre tykkelse-parametre.

Vælger vi en parametrisering med lange, tynde stiklinger som reference ($\alpha_2 = 0$, $\gamma_3 = 0$), finder vi modelmatricen svarende til modellen uden vekselvirkninger

$$\mathbf{X} = \begin{pmatrix} \text{ICPT} & \text{KORT} & \text{MID} & \text{TYK} \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix}$$

med den tilsvarende parametervektor

$$\boldsymbol{\beta} = \begin{pmatrix} \text{ICPT} \\ \text{KORT} \\ \text{MID} \\ \text{TYK} \end{pmatrix} = \begin{pmatrix} \mu \\ \alpha_1 \\ \gamma_1 \\ \gamma_2 \end{pmatrix}$$

□

De foregående eksempler svarede til, at man vælger den såkaldte kanoniske link som linkfunktion.

Det kan imidlertid forekomme, at problemstillingen lægger op til et andet valg af linkfunktion, end den kanoniske. Nedenstående anføres et eksempel på en sådan situation.

Eksempel 2.4.3 ***Bestemmelse af bakterietæthed fra fortyndingseksperiment***

I bakteriologien benyttes undertiden en såkaldt "fortyndings-opstilling" (eng.: *dilution assay*) til bestemmelse af koncentrationen af levedygtige bakterier i en given væske.

Metoden består primært i at tilberede et antal forskellige fortyndinger af den oprindelige væske. Fra hver fortynding udtages et antal prøver n_i, som sås i sterile glas med et vækstmedie.

Efter en passende tid (i varmeskab) undersøges for hver af prøverne, hvorvidt der var vækst, eller ej.

Lad x_i angive den kendte koncentration af den undersøgte væske i den i'te fortynding.

Under antagelse af at fordelingen af bakterier i den oprindelige væske er tilfældig og uden klumpning vil antallet af bakterier i et givet volumen kunne beskrives ved en Poisson-fordelt variabel. Det antages yderligere, at hvis der blot er én bakterie i den udsåede fortynding, da vil glasset vise vækst.

Under disse antagelser får man, at antallet af bakterier i et glas fra den i'te prøve kan beskrives ved en Poisson-fordelt variabel med middelværdi λx_i, hvor λ angiver den ukendte tæthed af bakterier i den oprindelige væske.

Sandsynligheden, p_i for at et glas fra den i'te prøve ikke vil vise vækst er sandsynligheden for at der ikke er nogle bakterier i prøven,

$$p_i = \mathrm{P}\left[\mathrm{P}(\lambda x_i) = 0\right] = \exp(-\lambda x_i) \; . \tag{2.4.13}$$

Lad nu Z_i betegne antallet af glas fra den i'te prøve, der ikke viste vækst. Der gælder da, at $Z_i \in \mathrm{B}(n_i, p_i)$, hvor p_i er givet ved (2.4.13).

Middelværdien af $Y_i = Z_i/n_i$ er p_i og modellen for p_i er

$$\ln p_i = -\lambda x_i \; , \tag{2.4.14}$$

hvor λ er den ukendte bakteriekoncentration. Modellen er en generaliseret lineær model for andelen Y_i af prøver uden vækst, hvor fordelingen af Y_i er $\mathrm{B}(n_i, p_i)/n_i$. Udtrykket (2.4.14) viser, at vi søger en lineær prædiktor for logaritmen til middelværdien. Linkfunktionen er således logaritmefunktionen.

Maksimum-likelihood estimatet for denne situation blev angivet af Fisher (1922).

Modellen finder anvendelse i en lang række andre sammenhænge, f.eks. når man måler fejl/ikke fejl i situationer med ukendt fejlintensitet, men kendte forhold mellem fejlintensiterne se f.eks. Hansen og Thyregod (1996). □

2.5 Estimation i generaliseret lineær model, fordeling af estimater

2.5.1 Maksimum likelihood estimat, observeret og forventet information

Sætning 2.5.1 ***Estimation i generaliserede lineære modeller***

Betragt den generaliserede lineære model (2.4.2) for observationerne $Y_1, \ldots Y_k$ og antag at $Y_1, \ldots Y_k$ er indbyrdes uafhængige med fordelinger, der kan beskrives ved en eksponentiel dispersionsmodel med variansfunktion $V(\cdot)$, dispersionsparameter σ^2 og eventuelt vægte (for en additiv model: indeksparametre) w_i.

Maksimum-likelihood estimatet $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ bestemmes som løsning til

$$[\mathbf{X}(\boldsymbol{\beta})]^T \mathbf{i}_\mu(\boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0} \,, \tag{2.5.1}$$

hvor $\mathbf{X}(\boldsymbol{\beta})$ angiver den lokale designmatrix (2.4.7), og $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta})$ angiver de fittede middelværdier svarende til parametersættet $\boldsymbol{\beta}$,

$$\mu_i(\boldsymbol{\beta}) = g^{-1}(\mathbf{x}_i^{*T}\boldsymbol{\beta}) \,, \tag{2.5.2}$$

og $\mathbf{i}_\mu(\boldsymbol{\mu})$ angiver den forventede information med hensyn til $\boldsymbol{\mu}$,

$$\mathbf{i}_\mu(\boldsymbol{\mu}) = \operatorname{diag}\left\{\frac{w_i}{V(\mu_i)}\right\} \,, \tag{2.5.3}$$

hvor w_i, $i = 1, 2, \ldots, k$ angiver den eventuelle vægt for den i'te observation. For en uvægtet model er $w_i = 1$.

Likelihoodligningen (2.5.1) må i almindelighed løses ved en iterativ procedure. I afsnit 2.5.4 skitserer vi nogle metoder, der kan benyttes.

Bevis:

Det følger af sætning 2.2.8, at scorefunktionen[1] med hensyn til middelværdiparameteren $\boldsymbol{\mu}$ er

$$l'_{\mu}(\boldsymbol{\mu};\mathbf{y}) = \operatorname{diag}\left\{\frac{w_i}{V(\mu_i)}\right\}(\mathbf{y}-\boldsymbol{\mu}) \tag{2.5.4}$$

Scorefunktionen med hensyn til $\boldsymbol{\beta}$ bliver da (jvf (2.1.9))

$$l'_{\beta}(\boldsymbol{\beta};\mathbf{y}) = \left[\frac{\partial\mu}{\partial\beta}\right]^T l'_{\mu}(\boldsymbol{\mu};\mathbf{y}) = [\mathbf{X}(\boldsymbol{\beta})]^T \operatorname{diag}\left\{\frac{w_i}{V(\mu_i)}\right\}(\mathbf{y}-\boldsymbol{\mu})\ . \tag{2.5.5}$$

Indsættes (2.5.3) i dette udtryk finder man

$$l'_{\beta}(\boldsymbol{\beta};\mathbf{y}) = \frac{\partial}{\partial\beta} l_{\beta}(\boldsymbol{\beta};\mathbf{y}) = [\mathbf{X}(\boldsymbol{\beta})]^T \mathbf{i}_{\mu}(\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})\ ,$$

der netop fører til (2.5.1). □

Bemærkning 1 ***For den kanoniske link fås estimatet ved løsning af middelværdiligningen***

For den kanoniske link reduceres den lokale designmatrix $\mathbf{X}(\boldsymbol{\beta})$ jvf bemærkning 2 på side 171 til $\operatorname{diag}\{V(\mu_i)\}\mathbf{X}$, hvorfor likelihoodligningen (2.5.1) bliver

$$\mathbf{X}^T \operatorname{diag}\{V(\mu_i)\} \operatorname{diag}\left\{\frac{w_i}{V(\mu_i)}\right\}(\mathbf{y}-\boldsymbol{\mu})\ ,$$

dvs

$$\mathbf{X}^T \operatorname{diag}\{w_i\}\,\mathbf{y} = \mathbf{X}^T \operatorname{diag}\{w_i\}\,\boldsymbol{\mu} \tag{2.5.6}$$

Ligningen (2.5.6) kaldes middelværdiligningen.

For en uvægtet model reduceres middelværdiligningen til den simple form

[1] Da vi kun betragter estimation af middelværdistrukturen, har vi set bort fra faktoren $1/\sigma^2$ i scorefunktion og informationsmatricer

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\boldsymbol{\mu} \tag{2.5.7}$$

Det kan vises (se f.eks. Barndorff-Nielsen (1978)), at hypotesen (2.4.2) definerer en m-dimensional eksponentiel familie med parametrene $\boldsymbol{\beta}$ og den sufficiente stikprøvefunktion $\mathbf{X}^T\text{diag}\{w_i\}\mathbf{y}$. I en sådan familie bestemmes maksimum-likelihood estimatet ved at sætte den observerede værdi af den (m-dimensionale) sufficiente stikprøvefunktion lig med sin middelværdi. □

Eksempel 2.5.1 ***Endimensional regression ved kanonisk link***

Lad $Y_1, \ldots Y_k$ være uafhængige variable, hvis fordeling kan beskrives ved en eksponentiel dispersionsmodel og betragt modellen specificeret ved parameterstrukturen

$$H_0: \quad \vartheta_i = \alpha + \beta x_i, \quad i = 1, 2, \ldots, k\,, \tag{2.5.8}$$

hvor $x_1, x_2, \ldots, x_n$ er et sæt af kendte kovariater, og hvor $\vartheta_i = \tau^{-1}(\mu_i)$ angiver værdierne af den kanoniske parameter.

Hypotesen H_0 er på formen

$$\begin{pmatrix} \vartheta_1 \\ \vartheta_2 \\ \vdots \\ \vartheta_k \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_k \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

dvs. netop af formen (2.4.2) med modelmatricen $\mathbf{X}$ givet ved den $k \times 2$-dimensionale matrix

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_k \end{pmatrix}$$

Idet

$$\mathbf{X}^T\ \text{diag}\{w_i\}\ \mathbf{y} = \begin{pmatrix} \sum_{i=1}^k w_i y_i \\ \\ \sum_{i=1}^k w_i x_i y_i \end{pmatrix}$$

og tilsvarende

$$\mathbf{X}^T \operatorname{diag}\{w_i\}\, \boldsymbol{\mu} = \begin{pmatrix} \sum_{i=1}^k w_i \tau(\alpha + \beta x_i) \\ \sum_{i=1}^k w_i x_i \tau(\alpha + \beta x_i) \end{pmatrix} ,$$

får man da for en uvægtet model ($w_i = 1$), at estimatet $\hat{\mu}_i$, $i = 1, 2, \ldots, n$ under modellen H_0 bestemmes ved

$$\hat{\mu}_i = \tau(\hat{\alpha} + \hat{\beta} x_i) , \tag{2.5.9}$$

hvor $\tau(\cdot)$ angiver middelværdifunktionen, og $\hat{\alpha}$ og $\hat{\beta}$ bestemmes som løsning til

$$\begin{aligned} \sum_{i=1}^k y_i &= \sum_{i=1}^k \tau(\alpha + \beta x_i) \\ \sum_{i=1}^k x_i y_i &= \sum_{i=1}^k x_i \tau(\alpha + \beta x_i) \end{aligned} \tag{2.5.10}$$

For en vægtet model med vægtene (for en additiv model: indeksparametrene) $w_1, \ldots, w_k$ bestemmes $\hat{\alpha}$ og $\hat{\beta}$ ved

$$\begin{aligned} \sum_{i=1}^k w_i y_i &= \sum_{i=1}^k w_i \tau(\alpha + \beta x_i) \\ \sum_{i=1}^k w_i x_i y_i &= \sum_{i=1}^k w_i x_i \tau(\alpha + \beta x_i) \end{aligned} \tag{2.5.11}$$

I afsnit 2.7 vil vi diskutere analysen af denne generiske regressionsmodel under forskellige dispersionsmodeller. □

2.5.2 Fittede værdier

Vi indfører først

Definition 2.5.1 ***Fittede værdier for generaliseret lineær model***

Betragt den generaliserede lineære model, der blev behandlet i sætning 2.5.1.

Lad $\widehat{\boldsymbol{\beta}}$ angive maksimum-likelihood estimatet for parameteren $\boldsymbol{\beta}$.

Ved de fittede værdier under hypotesen (2.4.2) vil vi forstå værdierne

$$\widehat{\boldsymbol{\mu}} \stackrel{\text{DEF}}{=} \boldsymbol{\mu}(\mathbf{X}\widehat{\boldsymbol{\beta}}) , \tag{2.5.12}$$

hvor den i'te værdi $\widehat{\mu}_i$ er givet ved

$$\widehat{\mu}_i = g^{-1}(\widehat{\eta}_i) \tag{2.5.13}$$

med den fittede værdi $\widehat{\eta}_i$ af den lineære prædiktor bestemt som

$$\widehat{\eta}_i = \sum_{j=1}^{m} x_{ij}\widehat{\beta}_j = (\mathbf{x}_i^*)^T \widehat{\boldsymbol{\beta}} \,, \tag{2.5.14}$$

hvor $\mathbf{x}_i^*$ angiver modelvektoren svarende til den i'te observation givet ved (2.4.3).

Såfremt specielt linkfunktionen er den kanoniske link, bruges den kanoniske parameter ϑ som lineær prædiktor, dvs (2.5.14) erstattes med

$$\widehat{\vartheta}_i = \sum_{j=1}^{m} x_{ij}\widehat{\beta}_j = (\mathbf{x}_i^*)^T \widehat{\boldsymbol{\beta}} \,, \tag{2.5.15}$$

□

2.5.3 Asymptotisk fordeling af maksimum likelihood estimatet

Med henblik på beskrivelsen af egenskaberne for maksimum likelihood estimatet anfører vi

Lemma 2.5.1 ***Den observerede og den forventede information med hensyn til $\boldsymbol{\beta}$***

Betragt den generaliserede lineære model, der blev behandlet i sætning 2.5.1.

Den observerede information $\mathbf{i}_\beta(\boldsymbol{\beta};\mathbf{y})$ med hensyn til $\boldsymbol{\beta}$ er

$$\mathbf{i}_\beta(\boldsymbol{\beta};\mathbf{y}) = -\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} l_\beta(\boldsymbol{\beta};\mathbf{y}) = \mathbf{i}_\beta(\boldsymbol{\beta}) - \frac{\partial^2\mu}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} l'_\mu(\boldsymbol{\mu};\mathbf{y}) \,, \tag{2.5.16}$$

hvor $\mathbf{i}_\beta(\boldsymbol{\beta})$ angiver den forventede information med hensyn til $\boldsymbol{\beta}$,

$$\mathbf{i}_\beta(\boldsymbol{\beta}) = [\mathbf{X}(\boldsymbol{\beta})]^T \mathbf{i}_\mu(\boldsymbol{\mu})\mathbf{X}(\boldsymbol{\beta}) \;= \mathbf{X}^T \operatorname{diag}\left\{\frac{w_i}{[g'(\mu_i)]^2 V(\mu_i)}\right\}\mathbf{X} \tag{2.5.17}$$

Bevis:
Vi har, at

$$\mathbf{i}_\beta(\boldsymbol{\beta};\mathbf{y}) = -\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} l_\beta(\boldsymbol{\beta};\mathbf{y}) = \left[\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\beta}}\right]^T \frac{\partial^2 l}{\partial\boldsymbol{\mu}\partial\boldsymbol{\mu}^T}\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\beta}} - \frac{\partial^2\mu}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} l'_\mu(\boldsymbol{\mu};\mathbf{y})$$

Første led er

$$\left[\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\beta}}\right]^T \frac{\partial^2 l}{\partial\boldsymbol{\mu}\partial\boldsymbol{\mu}^T}\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\beta}} = [\mathbf{X}(\boldsymbol{\beta})]^T \mathbf{i}_\mu(\boldsymbol{\mu};\mathbf{y})\mathbf{X}(\boldsymbol{\beta}) ,$$

hvorfor vi har

$$\mathbf{i}_\beta(\boldsymbol{\beta};\mathbf{y}) = [\mathbf{X}(\boldsymbol{\beta})]^T \mathbf{i}_\mu(\boldsymbol{\mu};\mathbf{y})\mathbf{X}(\boldsymbol{\beta}) - \mathbf{B} ,$$

hvor

$$\mathbf{B} = \frac{\partial^2\boldsymbol{\mu}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} l'_\mu(\boldsymbol{\mu};\mathbf{y}) ,$$

er det eneste led, der afhænger af observationen $\mathbf{y}$. Idet $\mathsf{E}\,[l'_\mu(\boldsymbol{\mu};\mathbf{y})] = \mathbf{0}$ (jvf. (2.1.10)) har vi, at $\mathsf{E}\,[\mathbf{B}] = \mathbf{0}$, hvorved vi får

$$\mathbf{i}_\beta(\boldsymbol{\beta}) = [\mathbf{X}(\boldsymbol{\beta})]^T \mathbf{i}_\mu(\boldsymbol{\mu})\mathbf{X}(\boldsymbol{\beta})$$

Ved indsættelse af udtrykket (2.5.3) for $\mathbf{i}_\mu(\boldsymbol{\mu})$ og udtrykket (2.4.7) for den lokale designmatrix $\mathbf{X}(\boldsymbol{\beta})$ får vi da (2.5.17).

□

Bemærkning 1 ***For den kanoniske link er den observerede information lig med den forventede***

Vi minder om, at såfremt linkfunktionen netop er den kanoniske link, reduceres scorefunktionen til

$$\frac{\partial}{\partial\beta} l_\beta(\beta;y) = \left[\frac{\partial\vartheta}{\partial\beta}\right]^T \frac{\partial}{\partial\vartheta} l(\vartheta(\beta);y) = \mathbf{X}^T(\mathbf{y}-\boldsymbol{\mu}) ,$$

hvor vi for nemheds skyld har set bort fra vægtningen.

Man får da udtrykket for den observerede information:

$$\mathbf{i}_\beta(\boldsymbol{\beta};\mathbf{y}) = \mathbf{X}^T \mathbf{i}_\mu(\boldsymbol{\mu})\mathbf{X}$$

Udtrykket afhænger ikke af $\mathbf{y}$, og i dette tilfælde er den observerede information altså lig med den forventede. □

Lemma 2.5.2 ***Differentiabilitet af maksimum likelihood estimatoren***

Betragt den generaliserede lineære model, der blev behandlet i sætning 2.5.1.

Såfremt linkfunktionen er således at afbildningen $\boldsymbol{\beta} \mapsto \vartheta(\boldsymbol{\beta})$ er en kontinuert bijektiv afbildning med kontinuert anden afledede, da definerer maksimum-likelihood ligningen (2.5.1) en funktion, $\widehat{\boldsymbol{\beta}} = b(\mathbf{y})$, hvor $b(\cdot)$ er kontinuert differentiabel og med den første afledede givet ved

$$\left[\frac{\partial}{\partial \mathbf{y}} b(\mathbf{y})\Big|_{y=\mu(\beta)}\right]^T = \left[\frac{\partial}{\partial \boldsymbol{\beta}^T}\vartheta\right] \mathbf{i}_\beta(\boldsymbol{\beta})^{-1} \tag{2.5.18}$$

hvor den forventede information med hensyn til $\boldsymbol{\beta}$, $\mathbf{i}_\beta(\boldsymbol{\beta})$ er givet ved (2.5.17).

Bevis:

Resultatet følger af standardresultaterne for implicit givne funktioner idet $b(\mathbf{y})$ er bestemt som løsning til maksimum-likelihood ligningen

$$\frac{\partial}{\partial \boldsymbol{\beta}} l_\beta(\boldsymbol{\beta}; \mathbf{y}) = \mathbf{0}$$

dvs $b(\mathbf{y})$ er givet som en implicit funktion af $\mathbf{y}$ ved relationen

$$h(b(\mathbf{y}), \mathbf{y}) = 0 \, .$$

Funktionalmatricen for $b(\cdot)$ er (jvf. Analyse 4, p. 305)

$$\mathbf{D}\mathbf{b} = \frac{\partial(b_1, b_2, \ldots, b_m)}{\partial y_1, y_2, \ldots, y_k}$$

Der gælder nu

$$\mathbf{D}_y \mathbf{h} + (\mathbf{D}_b \mathbf{h})\mathbf{D}_b = 0$$

dvs

$$\frac{\partial h}{\partial y} + \frac{\partial h}{\partial b}\,\frac{d\,b}{d\,y} = 0$$

Idet vi sætter

$$h(b(\mathbf{y}), \mathbf{y}) = [\mathbf{X}(\boldsymbol{\beta})]^T \mathbf{i}_\mu(\boldsymbol{\mu}, \mathbf{y})(\mathbf{y} - \boldsymbol{\mu})$$

har vi

$$\frac{\partial h}{\partial y} = [\mathbf{X}(\boldsymbol{\beta})]^T \mathbf{i}_\mu(\boldsymbol{\mu}, \mathbf{y})$$

og

$$\frac{\partial h}{\partial \beta} = -\mathbf{i}_\beta(\beta)$$

hvilket fører til

$$[\mathbf{X}(\beta)]^T \mathbf{i}_\mu(\mu, \mathbf{y}) - \mathbf{i}_\beta(\beta)\frac{\partial\, b}{\partial\, y} = 0$$

dvs

$$\frac{\partial\, b}{\partial\, y} = [\mathbf{i}_\beta(\beta)]^{-1}[\mathbf{X}(\beta)]^T \mathbf{i}_\mu(\mu, \mathbf{y})$$

Udnytter vi nu, at

$$[\mathbf{i}_\beta(\beta)]^{-1} = [[\mathbf{X}(\beta)^T \mathbf{i}_\mu(\mu, \mathbf{y})\mathbf{X}(\beta)]^{-1}$$

finder man (2.5.18). □

Sætning 2.5.2 ***Asymptotisk fordeling af maksimum-likelihood estimat***

Betragt den generaliserede lineære model, der blev behandlet i sætning 2.5.1.

Såfremt hypotesen $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ er sand, vil

$$\frac{\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}}{\sqrt{\sigma^2}} \overset{\text{as}}{\in} \mathrm{N}_m(\mathbf{0}, \boldsymbol{\Sigma}) \,, \tag{2.5.19}$$

hvor symbolet "$\overset{\text{as}}{\in}$" betyder "asymptotisk fordelt som" (når antallet af observationer, der indgår i de enkelte Y'er vokser mod uendeligt), og hvor σ^2 angiver dispersionsparameteren, og hvor dispersionsmatricen $\boldsymbol{\Sigma}$ for $\widehat{\boldsymbol{\beta}}$ er

$$\mathbf{D}\,[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\Sigma} = [\mathbf{X}^T \mathbf{W}(\boldsymbol{\beta})\mathbf{X}]^{-1} \tag{2.5.20}$$

med

$$\mathbf{W}(\boldsymbol{\beta}) = \mathrm{diag}\left\{\frac{w_i}{[g'(\mu_i)]^2 V(\mu_i)}\right\} \,, \tag{2.5.21}$$

Det i,j'te element, $\widehat{\sigma}_{ij}$, i $\boldsymbol{\Sigma}$ angiver kovariansen imellem $\widehat{\beta}_i$ og $\widehat{\beta}_j$.

Det j'te diagonalelement σ_{jj} i $\boldsymbol{\Sigma}$ angiver variansen for $\widehat{\beta}_j$.

Bevis:
Approximationen følger af lemma 2.5.2 ved at bemærke, at fordelingen af $\mathbf{Y}-\boldsymbol{\mu}$ asymptotisk er en $\mathrm{N}(\mathbf{0},\mathbf{V})$-fordeling, hvor dispersionsmatricen $\mathbf{V}$ for $\mathbf{Y}$ er givet ved

$$\mathbf{V} = \mathbf{D}\,[\mathbf{Y}] = \sigma^2 \mathrm{diag}(V(\mu_i)/w_i)$$

Den approximative dispersionsmatrix for den transformerede værdi $\widehat{\boldsymbol{\beta}} = b(\mathbf{Y})$ følger da ved at benytte lemma 2.5.2. □

Bemærkning 1 ***Estimation af dispersionsmatricen for*** $\widehat{\beta}$.

Dispersionsmatricen $\boldsymbol{\Sigma}$ (2.5.20) estimeres ved at indsætte de fittede værdier $\widehat{\boldsymbol{\mu}}$ i udtrykket, dvs

$$\widehat{\boldsymbol{\Sigma}} = [\mathbf{X}^T \mathbf{W}(\widehat{\boldsymbol{\beta}})\mathbf{X}]^{-1} \tag{2.5.22}$$

med

$$\mathbf{W}(\widehat{\boldsymbol{\beta}}) = \mathrm{diag}\left\{ \frac{w_i}{[g'(\widehat{\mu}_i)]^2 V(\widehat{\mu}_i)} \right\},$$

□

Bemærkning 2 ***Estimation af dispersionsmatricen ved kanonisk link***

Såfremt specielt linkfunktionen er den kanoniske link, reduceres "vægtmatricen" $\mathbf{W}(\boldsymbol{\beta})$ (2.5.21) til

$$\mathbf{W}(\boldsymbol{\beta}) = \mathrm{diag}\,\{w_i V(\mu_i)\} \tag{2.5.23}$$

□

Bemærkning 3 ***Fortolkning af "vægtmatricen"*** $\mathbf{W}(\boldsymbol{\beta})$

Vægtmatricen $\mathbf{W}(\boldsymbol{\beta})$ afspejler de vægte, hvormed de enkelte observationer indgår i estimationen, ud over den vægtning, der allerede er tilgodeset ved modelmatricen.

Således ser vi af (2.5.21), at de vægte, der er tilknyttet observationerne i en vægtet model indgår i vægtmatricen.

Endvidere gælder, at jo større værdi af

$$\mathsf{V}\,[g(Y)] \simeq [g'(\mu)]^2 \mathsf{V}\,[Y] = [g'(\mu)]^2 V(\mu)\ ,$$

desto mindre vægt tillægges observationen, da den transformerede observation (på η-skalaen) er mindre præcis.

Vi bemærker, at en stor varians $V(\mu)$ på den oprindelige Y-skala kan opvejes af en flad link-funktion (lille værdi af $g'(\mu)$). Hvis store ændringer i μ ikke indebærer væsentlige ændringer i η, betyder en stor usikkerhed på Y jo ikke så meget.

For den kanoniske link er vægtmatricen $\mathbf{W}(\boldsymbol{\beta})$ jvf (2.5.23)

$$\mathbf{W}(\boldsymbol{\beta}) = \mathrm{diag}\,\{w_i V(\mu_i)\}$$

Der gælder altså her, at jo større varians, desto større vægt tillægges den pågældende observation.

Dette tilsyneladende paradoks skyldes, at for den kanoniske link er linkfunktionen $g(\mu)$ bestemt af variansfunktionen, og der gælder

$$g'(\mu) = \frac{1}{V(\mu)}$$

dvs jo større varians, desto fladere linkfunktion. Betydningen af linkfunktionens fladhed er altså større end betydningen af usikkerheden på observationen. □

Vi nævner endvidere en anden metode til at vurdere betydningen af enkelte parametre ved at vurdere ændringen i maksimal likelihood svarende til at den pågældende parameter udelades af modellen:

Lemma 2.5.3 ***Fordeling af profillikelihood estimat***

Betragt en model H_0 parametriseret ved den r-dimensionale parametervektor $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_r)^T$.

Vi minder om, at, profil-loglikelihoodfunktionen $\tilde{l}(\beta_j; \mathbf{y})$ (jvf (2.1.5)) angiver maksimumværdien af log-likelihoodfunktionen for fastholdt β_j. Profilloglikelihoodfunktionen $\tilde{l}(\beta_j; \mathbf{y})$ er således en funktion af β_j.

Lad $\widehat{\beta}$ angive maksimum-likelihood estimatet for hele parametersættet under modellen H_0 (altså inklusive $\widehat{\beta}_j$). Betragt differensen

$$q(\beta_j; \mathbf{y}) = 2\{l(\widehat{\beta}; \mathbf{y}) - \tilde{l}(\beta_j; \mathbf{y})\}$$

mellem maksimum af likelihoodfunktionen og profillloglikelihood'en svarende til en fast værdi af parameteren β_j.

Såfremt den sande værdi af β_j er β_j^0, da vil $q(\beta_j^0; \mathbf{y})$ approximativt følge en $\chi^2(1)$-fordeling.

Bevis:

Følger ved at bemærke, at størrelsen $q(\beta_j^0; \mathbf{y})$ svarer til likelihood-kvotient teststørrelsen for hypotesen $H_1 : \beta_j = \beta_j^0$. Under H_1 vil $q(\beta_j^0; \mathbf{y})$ netop følge en $\chi^2(f)$-fordeling hvor antallet af frihedsgrader er forskellen i dimensionen af H_0 og H_1, altså netop $f = 1$. □

Sætning 2.5.3 ***Likelihoodkvotientbaseret konfidensinterval for enkelte parametre***

Et $100(1-\alpha)$ % konfidensinterval for β_j kan bestemmes som

$$\{\beta_j \ : \ \tilde{l}(\beta_j) \geq l_0 \}$$

hvor

$$l_0 = l(\widehat{\beta}) - 0.5\chi^2_{1-\alpha}(1)$$

Bevis:

Beviset følger af lemma 2.5.3. □

Intervallet må bestemmes ved iteration

Intervallet tilgodeser likelihoodfunktionens forløb. Således vil det sædvanligvis ikke være symmetrisk omkring maksimum-likelihood estimatet $\widehat{\beta}_j$.

Det sædvanlige (Wald) konfidensinterval er bare

$$\widehat{\beta}_j \pm u_{1-\alpha/2}\sqrt{\sigma^2 \widehat{\sigma}_{jj}}$$

2.5.4 Iterative metoder til estimation i generaliserede lineære modeller

Vi betragter atter den generaliserede lineære model (2.4.2) for observationerne $Y_1, \ldots Y_k$.

Ligningen til bestemmelse af maksimum-likelihood estimatet $\widehat{\beta}$ for β er anført i sætning 2.5.1 på side 179. Ligningen må sædvanligvis løses ved iteration, også selv om man bruger den kanoniske link.

Af hensyn til fortolkningen af udskrifter fra programpakker, der har moduler til behandling af generaliserede lineære modeller, anfører vi i dette afsnit de sædvanligt brugte iterationsmetoder.

Bemærkning 1 ***Bestemmelse af maksimum-likelihood estimatet ved Newton-Raphson metoden***

Vi betragter scorefunktionen $l'_\beta(\beta; \mathbf{y})$ med hensyn til β givet i (2.5.5).

Idet

$$\frac{\partial^2}{\partial\beta\partial\beta^T} l_\beta(\beta_t; \mathbf{y}) = -\mathbf{i}_\beta(\beta_t; \mathbf{y})$$

finder vi Taylor udviklingen af $l'_\beta(\beta; \mathbf{y})$ omkring værdien $\beta = \beta_t$

$$l'_\beta(\beta; \mathbf{y}) = l'_\beta(\beta_t; \mathbf{y}) - \mathbf{i}_\beta(\beta_t; \mathbf{y})(\beta - \beta_t)$$

Indfører vi (2.5.5)

$$l'_\beta(\beta_t; \mathbf{y}) = [\mathbf{X}(\beta)]^T \mathbf{i}_\mu(\mu)\mathbf{r} \,,$$

hvor $\mathbf{r}$ udtrykker vektoren af responsresidualer

$$r_i = y_i - \mu_i \tag{2.5.24}$$

finder vi, at den $t+1$'te approximation β_{t+1} til $\widehat{\beta}$ bestemmes ved

$$\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t = [\mathbf{i}_\beta(\boldsymbol{\beta}_t; \mathbf{y})]^{-1}[\mathbf{X}(\boldsymbol{\beta})]^T[\mathrm{diag}\{V(\mu_i)\}]^{-1}\mathbf{r} \tag{2.5.25}$$

hvor $\mathbf{X}(\boldsymbol{\beta})$ er givet ved (2.4.7),

$$\mathbf{X}(\boldsymbol{\beta}) = \mathrm{diag}\left\{\frac{1}{g'(\mu_i)}\right\}\mathbf{X}$$

og hvor $\mathbf{i}_\beta(\boldsymbol{\beta}_t; \mathbf{y})$ er givet ved (2.5.16).

For at bestemme et operationelt udtryk for $\mathbf{i}_\beta(\boldsymbol{\beta}_t; \mathbf{y})$ kan man betragte

$$\mathbf{i}_\beta(\boldsymbol{\beta}; \mathbf{y}) = \mathbf{X}^T\mathbf{i}_\eta(\boldsymbol{\eta}; \mathbf{y})\mathbf{X}\,,$$

hvor $\mathbf{i}_\eta(\boldsymbol{\eta}; \mathbf{y})$ angiver den observerede information med hensyn til den lineære prædiktor $\boldsymbol{\eta}$.

Den $k \times k$ dimensionale matrix $\mathbf{i}_\eta(\boldsymbol{\eta}; \mathbf{y})$ er en diagonalmatrix med diagonalelementerne

$$-\frac{\partial^2}{\partial\eta_i^2}l_\eta(\eta; \mathbf{y}) = \frac{1}{[g'(\mu_i)]^2V(\mu_i)} + (y_i - \mu_i)\frac{V(\mu_i)g''(\mu_i) + V'(\mu_i)g'(\mu_i)}{[V(\mu_i)]^2[g'(\mu_i)]^3} \tag{2.5.26}$$

Vi har altså

$$\mathbf{i}_\beta(\boldsymbol{\beta}; \mathbf{y}) = \mathbf{X}^T\mathbf{W}_o\mathbf{X}\,, \tag{2.5.27}$$

hvor $\mathbf{W}_o$ er en diagonalmatrix hvis i'te diagonalelement er

$$\frac{1}{[g'(\mu_i)]^2V(\mu_i)} + (y_i - \mu_i)\frac{V(\mu_i)g''(\mu_i) + V'(\mu_i)g'(\mu_i)}{[V(\mu_i)]^2[g'(\mu_i)]^3} \tag{2.5.28}$$

For en vægtet model erstattes matricen $[\mathrm{diag}\{V(\mu_i)\}]^{-1}$ i (2.5.25) med $\mathrm{diag}\{w_i/V(\mu_i)\}$ og det i'te element i $\mathbf{W}_o$ (2.5.28) multipliceres med vægten w_i. □

Bemærkning 2 ***Fisher's scoringsmetode***

Fisher's scoringmetode fremkommer hvis man i (2.5.25) erstatter den observerede information $\mathbf{i}_\beta(\boldsymbol{\beta};\mathbf{y})$ med den forventede $\mathbf{i}_\beta(\boldsymbol{\beta})$

Det ses af (2.5.26), at den forventede information med hensyn til $\boldsymbol{\eta}$ blot er

$$\mathbf{i}_\eta(\boldsymbol{\eta}) = \text{diag}\left\{\frac{1}{[g'(\mu_i)]^2 V(\mu_i)}\right\} . \tag{2.5.29}$$

Den forventede information med hensyn til $\boldsymbol{\beta}$ fås derfor som

$$\mathbf{i}_\beta(\boldsymbol{\beta}) = \mathbf{X}^T \mathbf{W}_e \mathbf{X} , \tag{2.5.30}$$

hvor $\mathbf{W}_e$ er en diagonalmatrix hvis i'te diagonalelement er

$$\frac{1}{[g'(\mu_i)]^2 V(\mu_i)} . \tag{2.5.31}$$

For den kanoniske link er den observerede information lig den forventede (jvf. bemærkning 1 på side 185), og Fisher's scoringsmetode er derfor i dette tilfælde den samme som Newton-Raphson metoden. □

Bemærkning 3 ***Iterativ genvægtet mindste kvadraters metode***

Betragt Taylor-udviklingen af link-funktionen:

$$g(y) \simeq g(\mu) + (y-\mu)g'(\mu)$$

med højresiden

$$z = \eta + (y-\mu)\frac{\mathrm{d}\eta}{\mathrm{d}\mu} .$$

z kaldes den lokale afhængige variable. Der gælder

$$\mathsf{V}[Z] = \left(\frac{\mathrm{d}\eta}{\mathrm{d}\mu}\right)^2 V(\mu) = [g'(\mu)]^2 V(\mu) .$$

Sæt nu

$$z_i = \widehat{\eta}_i + \frac{\mathrm{d}\eta_i}{\mathrm{d}\mu_i}(y_i - \widehat{\mu}_i) \ .$$

z_i kaldes undertiden for "working response" og størrelsen

$$(y_i - \widehat{\mu}_i)\,\frac{\mathrm{d}\eta}{\mathrm{d}\mu}$$

kaldes "working residual" (jvf. definition 2.5.6 på side 205).

Betragt nu det vægtede mindste kvadraters problem:

Bestem $\boldsymbol{\beta}_{t+1}$ så man opnår minimum for udtrykket

$$S(\boldsymbol{\beta}) = (\mathbf{z} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{W} (\mathbf{z} - \mathbf{X}\boldsymbol{\beta})$$

hvor vægtmatricen

$$\mathbf{W} = \left[\mathrm{diag} \left\{ \left(\frac{\mathrm{d}\eta_i}{\mathrm{d}\mu_i} \right)^2 V(\widehat{\mu}_i) \right\} \right]^{-1}$$

og hvor $\mathbf{z}$ angiver vektoren af working responses.

Løsningen til dette problem er (jvf. Oversigt over fordelinger med anvendelser i Statistik, IMM 1998)

$$\boldsymbol{\beta}_{t+1} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} (\mathbf{X}\boldsymbol{\beta}_t + \mathrm{diag}\{g'(\mu_i)\}(\mathbf{y} - \boldsymbol{\mu})) \tag{2.5.32}$$

"Hatmatricen" svarende til denne løsning er

$$\mathbf{H} = \mathbf{X}[\mathbf{X}^T \mathbf{W} \mathbf{X}]^{-1} \mathbf{X}^T \mathbf{W} \tag{2.5.33}$$

□

Bemærkning 4 *Den estimerede dispersionsmatrix ved iterativ bestemmelse af maksimum-likelihood estimatet*

Når $\widehat{\boldsymbol{\beta}}$ bestemmes ved Newton-Raphson metoden eller ved Fishers scoringsmetode er den såkaldte Hessian-matrix matricen

$$\mathbf{M} = \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} l_\beta(\boldsymbol{\beta}_t; \mathbf{y}) = -\mathbf{i}_\beta(\boldsymbol{\beta}_t; \mathbf{y}) \ ,$$

og opdateringen er af formen (2.5.25)

$$\beta_{t+1} = \beta_t - \mathbf{M}^{-1}\mathbf{s}\,,$$

hvor s er score-vektoren

$$\mathbf{s} = l'_\beta(\beta;\mathbf{y}) = [\mathbf{X}(\beta)]^T[\mathrm{diag}\{V(\mu_i)\}]^{-1}(\mathbf{y}-\mu)$$

Når iterationen er afsluttet, kan man benytte Hessian-matricen M til at bestemme et estimat $\widehat{\mathbf{\Sigma}}$ for $\mathbf{D}\,[\widehat{\beta}]$

Man har nemlig

$$\mathbf{i}_\beta(\widehat{\beta};\mathbf{y}) \simeq -\mathbf{M}$$

hvorfor man kan sætte

$$\widehat{\mathbf{\Sigma}} = -\mathbf{M}^{-1}$$

Hvis man har benyttet den observerede information (2.5.28), svarer dette estimat $\widehat{\mathbf{\Sigma}}$ til den observerede informationsmatrix. Hvis man har benyttet den forventede information (2.5.31) (Fisher's scoringsmetode) svarer estimatet til den estimerede værdi af den forventede information.

For den kanoniske link er de to metoder ækvivalente

Hvis $\widehat{\beta}$ bestemmes ved den iterativt genvægtede mindste kvadraters metode vil vægtmatricen

$$\mathbf{W} = \left[\mathrm{diag}\left\{\left(\frac{\mathrm{d}\eta_i}{\mathrm{d}\mu_i}\right)^2 V(\widehat{\mu}_i)\right\}\right]^{-1}$$

svarende til at den sidste iteration netop angiver matricen $\mathbf{W}(\widehat{\beta})$ (2.5.21). □

2.5.5 Eksempler på estimation i generaliserede lineære modeller

Eksempel 2.5.2 ***Fosterdødelighed hos mus (fortsat)***

Vi betragter atter data fra eksempel 2.4.1 på side 172.

I eksemplet opstillede vi regressionsmodellen for den kanoniske parameter:

$$H_0:\ \vartheta_i = \alpha + \beta x_i, \quad i = 1, 2, \ldots, k$$

Hypotesen er således et specialtilfælde af eksempel 2.5.1 på side 181, og vi finder at middelværdiligningen (2.5.11) til bestemmelse af α og β bliver

$$\begin{pmatrix} \sum_{i=1}^{k} n_i \exp(\alpha+\beta x_i)/\{1+\exp(\alpha+\beta x_i)\} \\ \sum_{i=1}^{k} x_i n_i \exp(\alpha+\beta x_i)/\{1+\exp(\alpha+\beta x_i)\} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{k} z_i \\ \sum_{i=1}^{k} x_i z_i \end{pmatrix}, \tag{2.5.34}$$

idet vi har benyttet vægtene $w_i = n_i$ og udnyttet, at $w_i y_i = z_i$, når $y_i = z_i/n_i$.

Ligningerne må løses ved iteration. Som udgangspunkt for iterationen kan man bruge mindste kvadraters estimat for linien tilpasset de observerede logit'er, ℓ_i

dvs man kan benytte

$$\widetilde{\beta} = \frac{\sum(\ell_i - \bar{\ell})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \tag{2.5.35}$$

og

$$\widetilde{\alpha} = \bar{\ell} - \widetilde{\beta}\bar{x} \tag{2.5.36}$$

Man finder $\widetilde{\alpha} = -3.07497$ og $\widetilde{\beta} = 0.005827$. Disse værdier kan benyttes som udgangspunkt for en iterativ procedure, hvorved man finder de endelige estimater $\hat{\alpha} = -3.248$ og $\hat{\beta} = 0.006389$.

Som en vurdering af tilpasningen kan de fittede værdier $\widehat{\vartheta}_i$ (logit'er) og de tilsvarende fittede middelværdier $\hat{\mu}_i$ beregnes.

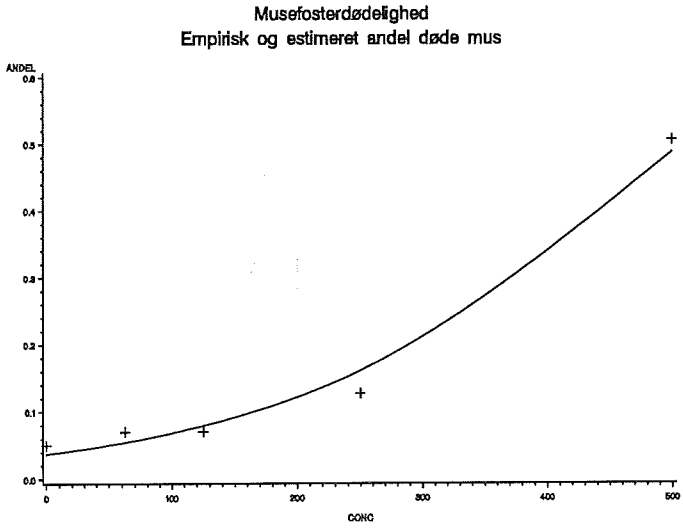
De fittede værdier af middelværdiparameteren p_i fås som

$$\widehat{p}_i = \frac{\exp(-3.248 + 0.006389 x_i)}{1 + \exp(-3.248 + 0.006389 x_i)}$$

Værdierne er anført i nedenstående tabel.

Obs nr. i	1	2	3	4	5
Koncentr. [mg/kg/dag]	0.0	62.5	125.0	250.0	500.0
$\widehat{\vartheta}_i$	-3.24800	-2.84869	-2.44938	-1.65075	-0.05350
$\widehat{p}_i$	0.03740	0.05475	0.07948	0.16101	0.48663

Nedenstående figurer viser de observerede og de fittede logit'er og tilsvarende de observerede og de fittede andele døde mus.



Der ses at være en rimelig god overensstemmelse mellem observationer og model. I afsnit 2.6 vil vi diskutere en metode til at vurdere tilpasningen.

Dispersionsmatricen for estimaterne fås af (2.5.22), hvor vægtmatricen $\mathbf{W}(\beta)$

bestemmes ved (2.5.23) idet vi jo har brugt den kanoniske link. Man får

$$\widehat{\mathbf{D}} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{pmatrix} 0.02486 & -0.00006 \\ -0.00006 & 0.0000002 \end{pmatrix}$$

Ved brug af SAS® PROC INSIGHT eller PROC GENMOD finder man det likelihoodkvotientbaserede 95 % konfidensinterval for parametrene α og β som $\alpha \in [-3.5673, -2.9486]$ og $\beta \in [0.00555, 0.00726]$. □

Eksempel 2.5.3 ***Udplantning af blommestiklinger (fortsat)***

Vi betragter atter situationen fra eksempel 2.4.2.

Estimationen foregår ved at løse middelværdiligningen (2.5.6)

$$\begin{aligned} \sum_{i=1}^{2} y_{ij} &= \sum_{i=1}^{2} \widehat{p}_{ij} \ , \quad j = 1,2,3 \\ \sum_{j=1}^{3} y_{ij} &= \sum_{j=1}^{3} \widehat{p}_{ij} \ , \quad i = 1,2 \end{aligned}$$

med

$$\widehat{p}_{ij} = \frac{\exp(\widehat{\mu} + \widehat{\alpha}_i + \widehat{\gamma}_j)}{1 + \exp(\widehat{\mu} + \widehat{\alpha}_i + \widehat{\gamma}_j)}$$

Ligningerne må løses ved iteration.

Som udgangspunkt for iterationen kan man betragte mindste kvadraters estimatet svarende til de empiriske logiter

$$\ell_{ij} = \ln\left(\frac{z_{ij} + 1/2}{n_{ij} - z_{ij} + 1/2}\right)$$

dvs

$$\widehat{\boldsymbol{\beta}}_1 = [\mathbf{X}^T\mathbf{X}]^{-1}\mathbf{X}^T\boldsymbol{\ell}$$

hvor

$$\boldsymbol{\ell} = \begin{pmatrix} -0.3485 \\ 0.3485 \\ 0.8692 \\ -0.5643 \\ 0.0000 \\ 0.0830 \end{pmatrix}$$

angiver de empiriske logit'er.

Denne første approximation er

$$\widehat{\boldsymbol{\beta}}_1 = \begin{pmatrix} -0.2313 \\ -0.4502 \\ 0.6306 \\ 0.9325 \end{pmatrix}$$

Man kan nu bestemme de fittede værdier $\widehat{\mathbf{p}}$ og den tilsvarende værdi af vægtmatricen $\mathbf{W}(\widehat{\mathbf{p}})$ (2.5.21) svarende til dette estimat,

Næste iteration kan da bestemmes ved

$$\widehat{\boldsymbol{\beta}}_2 = [\mathbf{X}^T\mathbf{W}(\widehat{\boldsymbol{\beta}})\mathbf{X}]^{-1}\mathbf{X}^T(\mathbf{y} - \widehat{\mathbf{p}})$$

Dette er netop Fisher's scoringsmetode, idet scorefunktionen (2.5.5) er

$$\mathbf{s} = l'_{\boldsymbol{\beta}}(\boldsymbol{\beta}; \mathbf{y}) = \mathbf{X}^T \mathrm{diag}\left\{\frac{1}{g'(p)V(p)}\right\}(\mathbf{y} - \mathbf{p}) ,$$

der i tilfældet med den kanoniske link reduceres til

$$l'_{\boldsymbol{\beta}}(\boldsymbol{\beta}; \mathbf{y}) = \mathbf{X}^T(\mathbf{y} - \mathbf{p}) ; .$$

Efter yderligere nogle iterationer finder man maksimum-likelihood estimatet

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} -0.6342 \\ -1.0735 \\ 1.6059 \\ 2.2058 \end{pmatrix} .$$

De fittede værdier, $\widehat{p}$, samt variansfunktionens værdi $V(\widehat{p}) = \widehat{p}(1 - \widehat{p})$ svarende til de fittede værdier er anført i nedenstående tabel.

Tabellen viser desuden kvadratroden af variansfunktionen $\sqrt{\widehat{p}(1 - \widehat{p})}$.

Obs nr	1	2	3	4	5	6
y_i	0.30	0.70	0.90	0.20	0.50	0.55
$\widehat{p}$	0.34655	0.72545	0.82800	0.15345	0.47455	0.62200
$V(\widehat{p})$	0.2265	0.1992	0.1424	0.1299	0.22494	0.2351
$\sqrt{V(\widehat{p})}$	0.4759	0.4463	0.3774	0.3604	0.4994	0.4849

Dispersionsmatricen estimeres jvf (2.5.22)ved

$$\widehat{\boldsymbol{\Sigma}} = [\mathbf{X}^T \mathbf{W}(\widehat{\boldsymbol{\beta}})\mathbf{X}]^{-1}$$

hvor (jvf. (2.5.23))

$$\mathbf{W}(\widehat{\boldsymbol{\beta}}) = \text{diag}\{w_i V(\widehat{p}_i)\} = \text{diag}\{w_i \widehat{p}_i(1-\widehat{p}_i)\}$$

$$= \begin{pmatrix} 4.5291 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3.9834 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2.8483 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2.5981 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4.9870 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4.7023 \end{pmatrix},$$

dvs

$$\widehat{\mathbf{D}}\,[\widehat{\boldsymbol{\beta}}] = [\mathbf{X}^T \text{diag}\{w_i V(\widehat{p}_i)\}\mathbf{X}]^{-1} = \begin{pmatrix} 23.6483 & 12.2874 & 8.9705 & 7.5506 \\ 12.2874 & 12.2874 & 4.9870 & 4.7023 \\ 8.9705 & 4.9870 & 8.9705 & 0 \\ 7.5506 & 4.7023 & 0 & 7.5506 \end{pmatrix}^{-1},$$

hvorfor vi endelig har

$$\widehat{\mathbf{D}}\,[\widehat{\boldsymbol{\beta}}] = \begin{pmatrix} 0.1639 & -0.0646 & -0.1279 & -0.1236 \\ -0.0646 & 0.1773 & -0.0339 & -0.0458 \\ -0.1279 & -0.0339 & 0.2583 & 0.1491 \\ -0.1236 & -0.0458 & 0.1491 & 0.2846 \end{pmatrix}.$$

De estimerede spredninger på de enkelte parameterværdier fås som kvadratroden af diagonalelementerne.

De enkelte parameterværdier kan nu vurderes ved (2.11.3):

Til illustration vises nedenfor udskriften svarende til SAS® PROC GENMOD og PROC INSIGHT

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi

```
INTERCEPT          1     -0.6342      0.4048      2.4548   0.1172
LENG       KO      1     -1.0735      0.4211      6.4992   0.0108
LENG       LA      0      0.0000      0.0000           .        .
TYK        MID     1      1.6059      0.5082      9.9849   0.0016
TYK        TYK     1      2.2058      0.5335     17.0977   0.0000
TYK        TYN     0      0.0000      0.0000           .        .
SCALE              0      1.0000      0.0000           .        .

NOTE:  The scale parameter was held fixed.
```

Størrelserne `Std Err` angiver den estimerede spredning $\sqrt{\widehat{\mathrm{V}}\,[\widehat{\beta}_j]} = \widehat{\sigma}_{jj}$ på den estimerede parameterværdi $\widehat{\beta}_j$, dvs kvadratroden af det tilsvarende diagonalelement i $\widehat{\mathbf{D}}\,[\widehat{\beta}]$.

Når de forklarende variable som her er klassifikationsvariable, vælger man i SAS-procedurerne at udskrive χ^2-teststørrelsen (2.11.4).

Vi ser, at alle koefficienter er signifikant forskellige fra nul.

I Programmeringssproget S-plus ville et kald af rutinen `glm` af formen

```
glm(cbind(lev,dod) ~ lengd + tyks, family = binomial(logit))
```

blandt andet bevirke udskriften

```
Coefficients:
                Value Std. Error   t value
(Intercept) -0.6342496  0.4047938 -1.566846
      lengd -1.0735260  0.4210604 -2.549577
    tyksmid  1.6058968  0.5081781  3.160106
    tykstyk  2.2058014  0.5333981  4.135375

(Dispersion Parameter for Binomial family taken to be 1 )
```

Overskriften `t-value` for teststørrelsen (2.11.3) indikerer ikke, at størrelsen nødvendigvis følger en t-fordeling, men blot, at størrelsen er fremkommet som forholdet mellem en estimeret middelværdi og den estimerede spredning for denne. □

2.5.6 Residualer

Betegnelsen residual bruges i almindelighed til at betegne en afvigelse mellem en observation og den tilsvarende fittede værdi. Sædvanligvis vil det gælde - i det mindste approximativt -, at residualer har middelværdien nul.

I det følgende vil vi introducere forskellige former for residualer, responsresidualer, deviansresidualer, Pearson-residualer og Wald-residualer. Disse former adskiller sig ved den måde, hvorpå man beregner afvigelsen mellem observeret og fittet værdi. De vigtigste af disse er responsresidualet, der måler afvigelsen mellem observation og tilpasset (fitted) direkte i samme enheder, som observationerne y og middelværdierne μ, og deviansresidualet, der måler afvigelsen udtrykt ved deviansen, dvs udtrykt som ændringer i log-likelihood.

Definition 2.5.2 ***Responsresidual***

Betragt den generaliserede lineære model (2.4.2) for observationerne $Y_1, \ldots Y_k$.

Ved responsresidualet (eller blot residualet) svarende til modellen vil vi forstå værdierne

$$r_i^r = r_R(y_i; \widehat{\mu}_i) \stackrel{\text{DEF}}{=} y_i - \widehat{\mu}_i \ , \quad i = 1, 2, \ldots, k \ , \tag{2.5.37}$$

hvor $\widehat{\mu}_i$ angiver den fittede værdi (2.5.13) svarende til den i'te observation.

□

Definition 2.5.3 ***Deviansresidual***

Betragt den generaliserede lineære model (2.4.2) for observationerne $Y_1, \ldots Y_k$.

Ved deviansresidualet svarende til modellen vil vi forstå værdierne

$$r_i^D = r_D(y_i; \widehat{\mu}_i) \stackrel{\text{DEF}}{=} \text{sign}(y_i - \widehat{\mu}_i)\sqrt{w_i d(y_i, \widehat{\mu}_i)} \tag{2.5.38}$$

hvor funktionen sign(x) er fortegnsfunktionen, sign(x) = 1 for $x > 0$ og sign(x) = -1 for $x < 0$, og hvor w_i angiver den eventuelle vægt, $d(y;\mu)$ angiver enhedsdeviansen og $\widehat{\mu}_i$ angiver den fittede værdi svarende til den i'te observation. □

Da deviansen er en ikke-negativ størrelse, viser den ikke, hvorvidt observationen y_i er større eller mindre end den fittede værdi $\widehat{\mu}_i$. Da fortegnet for observationens afvigelse fra den fittede værdi kan have interesse ved residualanalyser, har vi derfor introduceret dette fortegn i definitionen af deviansresidualet.

Vi skal senere (i afsnit 2.6.1) betragte kvadratsummen af deviansresidualerne som et mål for modellens tilpasning.

Definition 2.5.4 ***Pearson-residual***

Ved Pearson-residualet svarende til den generaliserede lineære model (2.4.2) vil vi forstå værdierne

$$r_i^P = r_P(y_i;\widehat{\mu}_i) \stackrel{\text{DEF}}{=} \frac{y_i - \widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i)/w_i}} \tag{2.5.39}$$

Pearson residualet fremkommer af responsresidualet ved at skalere med $\sqrt{\mathrm{V}\,[Y_i]}$. Pearson-residualet måler således responsresidualet i enheder af den estimerede standardafvigelse for observationen. □

Bemærkning 1 ***Kvadratet på Pearson residualet approximerer deviansen***

Vi så i bemærkning 3 på side 131, at Taylorudviklingen af enhedsdeviansen $d(y;\mu)$ omkring $y = \mu$ er

$$d(y;\mu) \approx \frac{(y-\mu)^2}{V(\mu)}\,, \tag{2.5.40}$$

altså netop kvadratet på det uvægtede Pearson-residual.

Idet kvadratet på deviansresidualet netop fremkommer ved at multiplicere enhedsdeviansen med den eventuelle vægtfaktor har vi altså

$$r_D(y;\widehat{\mu})^2 \approx \frac{w(y-\widehat{\mu})^2}{V(\widehat{\mu})} = r_P(y;\widehat{\mu})^2 \tag{2.5.41}$$

□

Definition 2.5.5 ***Wald-residual***

Betragt den generaliserede lineære model (2.4.2) for observationerne $Y_1, \ldots Y_k$ og lad middelværdiafbildningen (2.2.27) svarende til modellen være $\tau(\cdot)$ med den omvendte afbildning $\vartheta = \tau^{-1}(\mu)$.

Ved Wald-residualet svarende til den generaliserede lineære model (2.4.2) vil vi forstå værdierne

$$r_i^W = r_W(y_i;\widehat{\mu}_i) \stackrel{\text{DEF}}{=} \left\{\tau^{-1}(y_i) - \tau^{-1}(\widehat{\mu}_i)\right\}\sqrt{V(y_i)} \tag{2.5.42}$$

□

Bemærkning 1 ***Wald-residual måler afvigelser i rummet af kanoniske parametre***

Afbildningen $\tau^{-1}(\cdot)$ fører netop middelværdien μ over i den tilsvarende værdi af den kanoniske parameter ϑ. Vi kan altså opfatte Wald residualet som en standardiseret afvigelse mellem y og μ målt på aksen svarende til den kanoniske parameter.

Idet

$$\frac{\partial}{\partial\mu}\tau^{-1}(\mu) = \frac{1}{V(\mu)}$$

(jvf. Oversigt over fordelinger med anvendelser i Statistik, IMM 1998side 35), har man, at afbildningen $\tau^{-1}(\cdot)$ også kan udtrykkes som

$$\tau^{-1}(\cdot) = \int_{\mu_0}^{\mu} V^{-1}(u)du \tag{2.5.43}$$

hvor $V^{-1}(\cdot)$ angiver den omvendte afbildning til variansfunktionen og hvor μ_0 er en vilkårlig værdi i middelværdirummet $\mathcal{M}$. Definitionen afhænger ikke af valget af μ_0

Den approximative varians for Wald-residualet er dispersionsparameteren σ^2. □

Definition 2.5.6 ***Arbejdsresidual***

Undertiden møder man betegnelsen arbejdsresidual (engelsk *"working residual"*). Arbejdsresidualet fremkommer som afvigelsen mellem arbejdsresponset (engelsk *"working response"*)

$$z_i = \widehat{\eta}_i + \frac{\mathrm{d}\eta_i}{\mathrm{d}\mu_i}(y_i - \widehat{\mu}_i)$$

og den fittede η-værdi $\widehat{\eta}_i$. Arbejdsresponset bruges ofte som led i den iterative bestemmelse af parametrene (se afsnit 2.5.4).

Arbejdsresidualet er bestemt som

$$r_i^a = r_D(y_i;\widehat{\mu}_i) \stackrel{\text{DEF}}{=} (y_i - \widehat{\mu}_i)g'(\widehat{\mu}_i) \tag{2.5.44}$$

Arbejdsresidualet måler afvigelsen på $g(\mu)$-skalaen, nemlig den skala, hvor den lineære prædiktor η er udtrykt. □

Eksempel 2.5.4 ***Fosterdødelighed hos mus (fortsat)***

Vi betragter atter data fra eksempel 2.4.1 på side 172.

I eksempel 2.5.2 på side 195 bestemte vi estimaterne $\hat{\alpha} = -3.248$ og $\hat{\beta} = 0.006389$ for koefficienterne i den logistiske regressionsmodel $\vartheta_i = \alpha + \beta x_i$ og endvidere bestemte vi de fittede værdier svarende til denne model.

Deviansen svarende til binomialfordelte observationer fås af tabel 2.3. Man har enhedsdeviansen

$$d(y;p) = 2 \times \left\{ y \ln(y/p) + (1-y) \ln\left((1-y)/(1-p)\right) \right\}$$

og den vægtede devians bliver derfor

$$wd(y;p) = 2n \times \left\{ y \ln(y/p) + (1-y) \ln\left((1-y)/(1-p)\right) \right\}$$

I nedenstående tabel er for hver observation angivet de forskellige typer af residualer.

Obs nr. i	1	2	3	4	5
Koncentr. [mg/kg/dag]	0.0	62.5	125.0	250.0	500.0
Antal fostre w_i	297	242	312	299	285
Andel døde y_i	0.05051	0.07025	0.07051	0.12709	0.50526
$\tau^{-1}(y_i)$	-2.93386	-2.58289	-2.57884	-1.92693	0.02105
$\widehat{\vartheta}_i$	-3.24800	-2.84869	-2.44938	-1.65075	-0.05350
$\widehat{p}_i$	0.03740	0.05475	0.07948	0.16101	0.48663
$w_i d(y_i;\widehat{p}_i)$	1.28117	1.03557	0.35573	2.70902	0.39599
$r_r(y_i;\widehat{p}_i)$	0.01311	0.01550	-0.00897	-0.03392	0.01864
$r_D(y_i;\widehat{p}_i)$	1.13189	1.01763	-0.59643	-1.64591	0.62928
$r_P(y_i;\widehat{p}_i)$	1.19043	1.05984	-0.58585	-1.59571	0.62941
$r_W(y_i;\widehat{p}_i)$	1.18555	1.05673	-0.58544	-1.59065	0.62927

Man ser, at - bortset fra responsresidualet - er der ikke så stor forskel på de forskellige typer af residualer. □

Eksempel 2.5.5 *Udplantning af blommestiklinger (fortsat)*

Vi betragter atter situationen fra eksempel 2.4.2. I eksempel 2.5.3 bestemte vi estimaterne og de fittede værdier.

Nedenstående tabel viser deviansresidualerne svarende til disse værdier.

Obs nr	1	2	3	4	5	6
y_i	0.30	0.70	0.90	0.20	0.50	0.55
$\widehat{p}$	0.34655	0.72545	0.82800	0.15345	0.47455	0.62200
$n_i d(y;\widehat{p})$	0.1958	0.0638	0.8323	0.3097	0.0519	0.4318
$r_D(y;\widehat{p})$	-0.443	-0.253	0.912	0.556	0.228	-0.657

□

2.5.7 Fordeling af fittede værdier og residualer

Bemærkning 1 *Geometriske egenskaber for responsresidualer*

Udtrykkes maksimaliseringsligningen (2.5.1) ved vektoren af responsresidualer, $\mathbf{r} = (\mathbf{y} - \widehat{\boldsymbol{\mu}})$, ser vi, at maksimum-likelihood estimatet skal opfylde

$$[\mathbf{X}(\widehat{\boldsymbol{\beta}})]^T \, \mathbf{i}_\mu(\widehat{\boldsymbol{\mu}}) \, \mathbf{r} = \mathbf{0} \, ,$$

dvs ialt m ligninger af formen

$$\mathbf{u}_j^T \, \mathbf{i}_\mu(\widehat{\boldsymbol{\mu}}) \, \mathbf{r} = 0 \, ,$$

hvor vektorerne $\mathbf{u}_j$ er søjlerne i den lokale design matrix $\mathbf{X}(\widehat{\boldsymbol{\beta}})$.

Det gælder således, at vektoren af responsresidualer er ortogonal på rummet udspændt af søjlerne i den lokale design matrix $\mathbf{X}(\widehat{\boldsymbol{\beta}})$, hvor ortogonalitet måles med hensyn til det indre produkt defineret ved $\mathbf{i}_\mu(\widehat{\boldsymbol{\mu}})$ dvs

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T \mathbf{i}_\mu(\widehat{\boldsymbol{\mu}}) \mathbf{v} \, .$$

□

Ved en vurdering af residualer vil det være en fordel at kende variansen på disse residualer. Desuden kan det evet også være af interesse at kende størrelsen af korrelationen mellem de enkelte residualer.

Det er ikke overraskende, at en sådan korrelation eksisterer. Observationerne $\mathbf{y}$ kan variere frit i det k-dimensionale rum, men de fittede værdier ligger på en m-dimensional flade, da er der kun dimensionen $k - m$ tilbage til resten af variationen, dvs til residualerne.

Det følgende (ret tekniske) lemma diskuterer varians-kovariansforholdene for responsresidualet.

Lemma 2.5.4 ***Fordeling af fittede værdier og residualer***

Betragt observationssættet $Y_1, \ldots Y_k$ og antag, at $Y_1, \ldots Y_k$ er indbyrdes uafhængige, hvis fordelinger tilhører en eksponentiel dispersionsmodel med samme variansfunktion $V(\mu)$ og samme dispersionsparameter σ^2, og antag at link-funktionen er givet ved $\eta = g(\mu)$ og en eventuel vægtning (eller værdier af indeksparametre) er givet ved vægtene w_i.

Betragt en generaliseret lineær model af formen (2.4.2), og lad $\widehat{\boldsymbol{\beta}}$ angive maksimaliseringsestimatet for parameteren $\boldsymbol{\beta}$, og lad endvidere $\widehat{\boldsymbol{\mu}}$ angive de fittede værdier under denne hypotese.

$$\widehat{\boldsymbol{\mu}} \stackrel{\text{DEF}}{=} \boldsymbol{\mu}(\mathbf{X}\widehat{\boldsymbol{\beta}}) \, ,$$

Lad $\widetilde{\mathbf{y}}$ angive de standardiserede værdier af y

$$\widetilde{y}_i \stackrel{\text{DEF}}{=} \frac{1}{\sqrt{V(\mu_i)/w_i}}\, y_i \tag{2.5.45}$$

og $\widetilde{\boldsymbol{\mu}}$ de lokalt standardiserede fittede værdier er

$$\widetilde{\mu}_i \stackrel{\text{DEF}}{=} \frac{1}{\sqrt{V(\mu_i)/w_i}}\, \widehat{\mu}_i \tag{2.5.46}$$

og lad endelig $\widetilde{\mathbf{r}}$ angive de tilsvarende lokalt standardiserede responsresidualer

$$\widetilde{r}_i(\widehat{\mu}) \stackrel{\text{DEF}}{=} \frac{1}{\sqrt{V(\mu_i)/w_i}}\, (y_i - \widehat{\mu}_i) \tag{2.5.47}$$

Der gælder da:

$$\mathbf{D}\,[\widetilde{\mathbf{r}}] \simeq \sigma^2(\mathbf{I}_n - \mathbf{H}(\beta)) \tag{2.5.48}$$

og altså specielt

$$\mathrm{V}\,[\widetilde{r}_i] \simeq \sigma^2(1 - h_{ii}) \tag{2.5.49}$$

Endvidere gælder at varians-kovariansmatricen for de (standardiserede) observerede værdier $\widetilde{\mathbf{y}}$ og de (lokalt standardiserede) prædikterede værdier $\widetilde{\boldsymbol{\mu}}$ er

$$\mathbf{D}\begin{bmatrix} \widetilde{\mathbf{y}} \\ \widetilde{\boldsymbol{\mu}} \end{bmatrix} = \sigma^2 \begin{pmatrix} \mathbf{I}_n & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{pmatrix} \tag{2.5.50}$$

hvor $\mathbf{H}(\beta)$ er matricen:

$$\mathbf{H}(\boldsymbol{\beta}) = \mathbf{W}(\boldsymbol{\beta})^{1/2}\mathbf{X}[\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta})\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta})^{1/2} \tag{2.5.51}$$

med

$$\mathbf{W}(\boldsymbol{\beta}) = \text{diag}\left\{\frac{1}{(g'(\mu_i))^2 V(\mu_i)}\right\} \tag{2.5.52}$$

Såfremt linkfunktionen $g(\cdot)$ specielt er den kanoniske link er

$$\mathbf{W}(\boldsymbol{\beta}) = \text{diag}(V(\mu_i)) \tag{2.5.53}$$

Bevis:
Idet den forventede information med hensyn til $\boldsymbol{\mu}$ er (jvf. (2.5.3))

$$\mathbf{i}_\mu(\boldsymbol{\mu}) = \text{diag}\left\{\frac{w_i}{V(\mu_i)}\right\} ,$$

kan vi udtrykke de lokalt standardiserede værdier ved

$$\begin{aligned} \widetilde{\mathbf{y}} &= \mathbf{i}_\mu(\boldsymbol{\mu})^{1/2}\mathbf{y} \\ \widetilde{\boldsymbol{\mu}} &= \mathbf{i}_\mu(\boldsymbol{\mu})^{1/2}\widehat{\boldsymbol{\mu}} \\ \widetilde{\mathbf{r}}(\widehat{\boldsymbol{\mu}}) &= \mathbf{i}_\mu(\boldsymbol{\mu})^{1/2}\mathbf{r} \end{aligned}$$

hvor vektoren $\mathbf{r}$ af responsresidualer defineres som $(y_1 - \widehat{\mu}_1, \ldots, y_k - \widehat{\mu}_k)^T$

Den standardiserede lokale design matrix er

$$\widetilde{\mathbf{X}}(\boldsymbol{\beta}) \stackrel{\text{DEF}}{=} \mathbf{i}_\mu(\boldsymbol{\mu})^{1/2}\mathbf{X}(\boldsymbol{\beta}) \tag{2.5.54}$$

hvor den lokale designmatrix $\mathbf{X}(\boldsymbol{\beta})$ er defineret i (2.4.7)

$$\mathbf{X}(\boldsymbol{\beta}) = \text{diag}\left\{\frac{1}{g'(\mu_i)}\right\}\mathbf{X} .$$

Vi har da, idet vi udtrykker $y_i = \widehat{\mu}_i + (y_i - \widehat{\mu}_i)$

$$\mathbf{y} = \widehat{\boldsymbol{\mu}} + \mathbf{r} ,$$

at der tilsvarende gælder for de lokalt standardiserede variable

$$\widetilde{\mathbf{y}} = \widetilde{\boldsymbol{\mu}} + \widetilde{\mathbf{r}} . \tag{2.5.55}$$

Men da $\widetilde{\mathbf{y}}$ er de standardiserede værdier af de uafhængige variable $Y_1, \ldots, Y_k$ har vi

$$\mathbf{D}\,[\widetilde{\mathbf{y}}] = \sigma^2\, \mathbf{I}_k .$$

Betragt nu højre side af (2.5.55). Likelihoodligningen $l'_\beta(\boldsymbol{\beta}; \mathbf{y}) = \mathbf{0}$ (2.5.1) svarer til ligningen:

$$[\mathbf{X}(\boldsymbol{\beta})]^T \mathbf{i}_\mu(\boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0} \tag{2.5.56}$$

dvs

$$[\widetilde{\mathbf{X}}(\boldsymbol{\beta})]^T \widetilde{\mathbf{r}}(\widehat{\boldsymbol{\mu}}) = \mathbf{0} \tag{2.5.57}$$

I lighed med betragtningerne i bemærkning 1 på side 206 ser vi nu af (2.5.57), at vektoren af lokalt standardiserede residualer $\widetilde{\mathbf{r}}(\widehat{\boldsymbol{\mu}})$ er ortogonal på rummet udspændt af søjlerne i den standardiserede lokale design matrix $\widetilde{\mathbf{X}}(\widehat{\boldsymbol{\beta}})$, hvor ortogonalitet er med hensyn til det sædvanlige indre produkt i $\mathbb{R}^k$.

Det følger da af resultater fra normalfordelingsteorien, at $\widehat{\boldsymbol{\beta}}$ og $\widetilde{\mathbf{r}}$ er approximativt uafhængige, hvorfor også $\widetilde{\boldsymbol{\mu}}$ og $\widetilde{\mathbf{r}}$ er approximativt uafhængige. Vi har derfor

$$\mathbf{D}\,[\widetilde{\mathbf{y}}] = \mathbf{D}\,[\widetilde{\boldsymbol{\mu}} + \widetilde{\mathbf{r}}] \simeq \mathbf{D}\,[\widetilde{\boldsymbol{\mu}}] + \mathbf{D}\,[\widetilde{\mathbf{r}}] \tag{2.5.58}$$

Vi har fra sætning 2.5.2, at

$$\mathbf{D}\,[\widehat{\boldsymbol{\beta}}] \simeq \sigma^2\, [\mathbf{i}_\beta(\boldsymbol{\beta})]^{-1} = \sigma^2\, [\mathbf{X}(\boldsymbol{\beta})^T \mathbf{i}_\mu(\boldsymbol{\mu}) \mathbf{X}(\boldsymbol{\beta})]^{-1}$$

Idet $\mu_i = g^{-1}(\eta_i)$ og $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ har vi altså $\boldsymbol{\mu} = \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta})$.

Idet den afledede af $\boldsymbol{\mu}$ med hensyn til $\boldsymbol{\beta}$ er den lokale designmatrix $\mathbf{X}(\boldsymbol{\beta})$ (2.4.7) har vi, at Taylor-udviklingen af $\boldsymbol{\mu}(\mathbf{X}\widehat{\boldsymbol{\beta}})$ omkring den sande værdi $\boldsymbol{\beta}$ er

$$\boldsymbol{\mu}(\mathbf{X}\widehat{\boldsymbol{\beta}}) \simeq \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}) + \mathbf{X}(\boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

og derfor er

$$\mathbf{D}\,[\widehat{\boldsymbol{\mu}}] \simeq \mathbf{X}(\boldsymbol{\beta})\ \mathbf{D}\,[\widehat{\boldsymbol{\beta}}]\mathbf{X}(\boldsymbol{\beta})^T$$

Vi har derfor endelig, for de lokalt standardiserede prædikterede værdier $\widetilde{\boldsymbol{\mu}}$, at

$$\mathbf{D}\,[\widetilde{\boldsymbol{\mu}}] = \sigma^2\ \mathbf{i}_\mu(\boldsymbol{\mu})^{1/2}\ \mathbf{D}\,[\widehat{\boldsymbol{\mu}}][\mathbf{i}_\mu(\boldsymbol{\mu})^{1/2}]^T \simeq \sigma^2\ \mathbf{H}(\boldsymbol{\beta})\,, \tag{2.5.59}$$

hvor $\mathbf{H}(\boldsymbol{\beta})$ er matricen

$$\begin{aligned} \mathbf{H}(\boldsymbol{\beta}) &= \widetilde{\mathbf{X}}(\boldsymbol{\beta})[\widetilde{\mathbf{X}}(\boldsymbol{\beta})]^{-1}\widetilde{\mathbf{X}}(\boldsymbol{\beta})^T \qquad (2.5.60)\\ &= \mathbf{i}_\mu(\boldsymbol{\mu})^{1/2}\mathbf{X}(\boldsymbol{\beta})[\mathbf{X}(\boldsymbol{\beta})^T\mathbf{i}_\mu(\boldsymbol{\mu})\mathbf{X}(\boldsymbol{\beta})]^{-1}\mathbf{X}(\boldsymbol{\beta})^T\mathbf{i}_\mu(\boldsymbol{\mu})^{1/2} \end{aligned}$$

Dispersionsmatricerne har bestemt ikke fuld rang. Vi ved jo, at de lokalt standardiserede prædikterede værdier, $\widetilde{\boldsymbol{\mu}} \in \text{span}(\widetilde{\mathbf{X}}(\widehat{\boldsymbol{\beta}}))$ som har dimensionen m, og tilsvarende at de lokalt standardiserede residualer $\widetilde{\mathbf{r}} \in \text{span}(\widetilde{\mathbf{X}}(\widehat{\boldsymbol{\beta}}))^\perp$ som har dimensionen $k - m$.

Vi har nu af (2.5.58), idet $\mathbf{D}\,[\widetilde{\mathbf{y}}] = \sigma^2\ \mathbf{I}_k$, at

$$\mathbf{D}\,[\widetilde{\mathbf{r}}] \simeq \sigma^2\ (\mathbf{I}_n - \mathbf{H}(\boldsymbol{\beta}))$$

hvilket netop er (2.5.48).

For den kanoniske link er (jvf (2.4.10))

$$\mathbf{X}(\boldsymbol{\beta}) = \text{diag}\{V(\mu_i)\}\mathbf{X}$$

hvorved man får (2.5.53). □

Bemærkning 2 ***Geometriske egenskaber for de lokalt standardiserede residualer*** Vi så tidligere (bemærkning 1 på side 206), at vektoren af responsresidualer er ortogonal på rummet udspændt af søjlerne i den lokale design matrix $\mathbf{X}(\widehat{\boldsymbol{\beta}})$, hvor ortogonalitet måles med hensyn til det indre produkt defineret ved $\mathbf{i}_\mu(\widehat{\boldsymbol{\mu}})$.

Det fremgår af (2.5.57), at vektoren af lokalt standardiserede residualer $\tilde{\mathbf{r}}(\widehat{\boldsymbol{\mu}})$ er ortogonal på rummet udspændt af søjlerne i den standardiserede lokale design matrix $\widetilde{\mathbf{X}}(\widehat{\boldsymbol{\beta}})$, hvor ortogonalitet er med hensyn til det sædvanlige indre produkt i $\mathbb{R}^k$. Idet søjlerne i den standardiserede lokale design matrix udspænder tangentplanen til den middelværdiflade, der er fastlagt ved hypotesen H_0, har vi således, at vektoren af lokalt standardiserede residualer er vinkelret på denne tangentplan. □

Bemærkning 3 *Matricen* $\mathbf{H}(\boldsymbol{\beta})$ *kaldes "hat"-matricen*

Matricen $\mathbf{H}(\boldsymbol{\beta})$ bestemt ved (2.5.51) kaldes den lokale hat matrix.

Matricen, der kan opfattes som en lokal projektionsmatrix, flytter $\mathbf{y}$ ned i de tilsvarende fittede værdier $\widehat{\boldsymbol{\mu}}$. Da de fittede værdier er symboliseret ved en $\widehat{\ }$ (en "hat"), kaldes matricen ofte for hat-matricen og symboliseres ved $\mathbf{H}$. □

Bemærkning 4 *Variansen på den fittede værdi og residualvariansen afhænger af diagonalelementerne i "hat"-matricen*

Vi bemærker, at residualerne ikke har samme varians. Vi har

$$\mathrm{V}[r_i] \simeq \sigma^2 \, V(\mu_i) \, (1 - h_{ii})$$

hvor h_{ii} er det i'te diagonalelement i "hat"-matricen. Hvis specielt h_{ii} er nær ved 1, er der altså ikke stor varians på residualet.

Modsvarende har vi, at

$$\mathrm{V}[\widehat{\mu}_i] \simeq \sigma^2 \, V(\mu_i) \, h_{ii} \; ;$$

dvs, at hvis h_{ii} er lille, er der ikke stor varians på den fittede værdi.

Vi skal senere (afsnit 2.15.3) ses, at når h_{ii} er lille, kan dette tages som udtryk for at der ligger mange observationer til grund for den pågældende fittede værdi.

Det vil med andre ord sige, at hvis der ligger information fra mange observationer til grund for en bestemt fittet værdi, vil der være en stor varians på det tilsvarende residual. Hvis omvendt den fittede værdi stort set blot er bestemt på grundlag af en enkelt observation, vil det pågældende residual have en lille varians, dvs den fittede værdi vil ligge tæt på den pågældende observation.

Vi finder specielt, at korrelationen mellem en observation og den fittede værdi er

$$\rho(y_i, \widehat{\mu}_i) = \sqrt{h_{ii}}$$

hvilket er i overensstemmelse med ovenstående betragtninger.

Vi skal senere (afsnit 2.15.3) komme tilbage til betydningen af hat-matricen. □

2.5.8 Residualer, standardisering og studentisering

Vi så i det foregående afsnit, at variansen (dvs usikkerheden) på residualet landt andet afhang af det tilsvarende diagonalelement i hat-matricen. Det vil være naturligt at skalere residualerne med spredningen for at sikre en sammenlignelighed af residualerne. Man siger sædvanligvis at residualerne er standardiserede, hvis de er skaleret med en størrelse, der sikrer, at de har samme varians under modellen. Der er imidlertid ikke fuldstændig enighed om denne sprogbrug. Undertiden bruger man betegnelsen "standardiseret" til at angive, at residualerne er skaleret med deres standardafvigelse, sådan at de får spredningen 1.

En del af de generaliserede lineære modeller, vi betragter, har tilknyttet en ukendt dispersionsparameter, σ^2, som også skal estimeres fra data (se afsnit 2.6.2). Dette estimat vil sædvanligvis baseret på residualerne, og hvis man vil vurdere størrelsen af et residual i forhold til den estimerede spredning på dette residual, kan det være ønskeligt, at den estimerede spredning er uafhængig af netop den betragtede observation. Man siger, at et sæt residualer er studentiserede, hvis de er skaleret med et uafhængigt estimat for spredningen.

I det følgende vil vi introducere de standardiserede og studentiserede residualer, der sædvanligvis optræder i programmer til analyse af generaliserede lineære modeller, som feks programsystemerne SAS® og S-Plus.

I udtrykkene betegner h_{ii} det i'te diagonalelement i hat-matricen (2.5.51), $\widehat{\sigma}^2$ angiver estimatet for dispersionsparameteren σ^2 ved benyttelse af alle observationer, og $\widehat{\sigma}^2_{(i)}$ angiver estimatet, beregnet ved udeladelse af den i'te observation.

Standardiseret responsresidual

$$r_i^{rs} = \frac{r_i^r}{\sqrt{\widehat{\sigma}^2(1 - h_{ii})}}$$

Studentiseret responsresidual

$$r_i^{rt} = \frac{r_i^r}{\sqrt{\widehat{\sigma}^2_{(i)}(1-h_{ii})}}$$

Standardiseret deviansresidual

$$r_i^{Ds} = \frac{r_i^D}{\sqrt{\widehat{\sigma}^2(1-h_{ii})}}$$

Studentiseret deviansresidual

$$r_i^{Dt} = \frac{r_i^D}{\sqrt{\widehat{\sigma}^2_{(i)}(1-h_{ii})}}$$

Standardiseret Pearsonresidual

$$r_i^{Ps} = \frac{r_i^P}{\sqrt{\widehat{\sigma}^2(1-h_{ii})}}$$

Studentiseret Pearsonresidual

$$r_i^{Pt} = \frac{r_i^P}{\sqrt{\widehat{\sigma}^2_{(i)}(1-h_{ii})}}$$

I S-plus programsystemet fås de forskellige (ustandardiserede) residualer ved at bruge en *extractor* funktion residuals( ) på et glm-objekt. Funktionen har et type= argument, som kan være "deviance", "working", "pearson", eller "response". Såfremt man ikke specificerer nogen type fås deviansresidualerne.

Tabel 2.6 giver en oversigt over disse standardiserede og studentiserede residualer.

Størrelse	Udtryk	Fortolkning	side
Standardiserede residualer	$r/\left(\sqrt{\widehat{\sigma}^2\,(1-h)}\,\right)$	Residualer med varians $V(\mu_i)$	213
Studentiserede residualer	$r/\left(\sqrt{\widehat{\sigma}^2_{(i)}\,(1-h)}\,\right)$	Som ovenfor med $\widehat{\sigma^2}_{(i)}$ som dispersionsparameter	214
Standardiserede deviansresidualer	$r_d/\left(\sqrt{\widehat{\sigma}^2\,(1-h)}\,\right)$	Deviansresidualer med variansen 1	214
Studentiserede deviansresidualer	$r_d/\left(\sqrt{\widehat{\sigma}_{(i)^2}\,(1-h)}\,\right)$	Som ovenfor med $\widehat{\sigma^2}_{(i)}$ som dispersionsparameter	214
Standardiserede Pearsonresidualer	$r_P/\left(\sqrt{\widehat{\sigma}^2\,(1-h)}\,\right)$	Pearsonresidualer med variansen 1	214
Studentiserede Pearsonresidualer	$r_P/\left(\sqrt{\widehat{\sigma}_{(i)^2}\,(1-h)}\,\right)$	Som ovenfor med $\widehat{\sigma^2}_{(i)}$ som dispersionsparameter	214

Tabel 2.6. Oversigt over sædvanlige standardiserede og studentiserede residualer

2.5.9 Forudsigelse, prædiktion

Når man har tilpasset en model, kan det undertiden være af interesse at bruge modellen til at forudsige eller kontrollere værdier svarende til en bestemt kombination af de forklarende variable samt at bestemme usikkerheden på en sådan forudsigelse.

Den forudsagte værdi fås ved at indsætte værdierne af de forklarende variable i det parametriske udtryk for modellen

$$\widetilde{\eta} = \widehat{\beta}_1 \tilde{x}_1 + \cdots \widehat{\beta}_m \tilde{x}_m \tag{2.5.61}$$

Den forudsagte værdi på middelværdiskalaen er da

$$\widetilde{\mu} = g^{-1}(\widetilde{\eta}) \tag{2.5.62}$$

Sætning 2.5.4 ***Fordelingsforhold ved forudsigelser i generaliseret lineær model***

Betragt en generaliseret lineær model for observationerne $Y_1, \ldots Y_k$ med variansfunktion $V(\cdot)$, linkfunktion $g(\cdot)$ og den $k \times m$-dimensionale modelmatrix $\mathbf{X}$.

Betragt nu et sæt af p nye værdier af de forklarende variable

$$\widetilde{\mathbf{X}} = \begin{pmatrix} \tilde{x}_{11} & \tilde{x}_{12} & \cdots & \tilde{x}_{1m} \\ \tilde{x}_{21} & \tilde{x}_{22} & \cdots & \tilde{x}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{x}_{p1} & \tilde{x}_{p2} & \cdots & \tilde{x}_{pm} \end{pmatrix} \tag{2.5.63}$$

Det tilsvarende sæt af forudsagte værdier af den lineære prædiktor er

$$\widetilde{\boldsymbol{\eta}} = \widetilde{\mathbf{X}}\widehat{\boldsymbol{\beta}} \tag{2.5.64}$$

Dispersionsmatricen for sættet af forudsagte værdier af den lineære prædiktor er

$$\mathbf{D}\,[\widetilde{\eta}] = \widetilde{\mathbf{X}}[\mathbf{X}^T\mathbf{W}(\beta)\mathbf{X}]^{-1}\widetilde{\mathbf{X}}^T\,, \tag{2.5.65}$$

hvor $\mathbf{X}$ angiver den $k \times m$-dimensionale modelmatrix, der blev brugt ved estimation af $\widehat{\beta}$, og hvor $\mathbf{W}(\beta)$ er givet ved

$$\mathbf{W}(\beta) = \text{diag}\left\{\frac{w_i}{[g'(\mu_i)]^2 V(\mu_i)}\right\}\,.$$

Det tilsvarende sæt af forudsagte middelværdier fås som $\widetilde{\mu}_i = g^{-1}(\widehat{\eta}_i)$.

Dispersionsmatricen for sættet af forudsagte middelværdier er

$$\mathbf{D}\,[\widetilde{\mu}] \simeq \text{diag}\left\{\frac{1}{[g'(\widetilde{\mu}_i)]}\right\}\;\mathbf{D}\,[\widetilde{\eta}]\text{diag}\left\{\frac{1}{[g'(\widetilde{\mu}_i)]}\right\} \tag{2.5.66}$$

Bevis:
Det følger af Sætning 2.5.2 formel (2.5.20) at variansen for $\widehat{\beta}$ er

$$\mathbf{D}\,[\widehat{\beta}] = \mathbf{\Sigma} = [\mathbf{X}^T\mathbf{W}(\beta)\mathbf{X}]^{-1}$$

Idet forudsigelsen $\widetilde{\eta}$ blot er linearkombinationer af $\widehat{\beta}$ fås (2.5.65) ved de sædvanlige resultater for lineære transformationer og (2.5.66) følger da ved Taylorudvikling. □

Bemærkning 1 ***Forudsigelsesvarians ved en enkelt forudsigelse***

Betragt forudsigelsen af en enkelt ny observation.

Lad de "nye" værdier af de forklarende variable være

$$\widetilde{\mathbf{x}} = \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \vdots \\ \tilde{x}_m \end{pmatrix}\,.$$

Den tilsvarende forudsagte værdi af den lineære prædiktor kan udtrykkes som

$$\widetilde{\eta} = \widetilde{\mathbf{x}}^T\widehat{\beta}\,.$$

Variansen på $\widetilde{\eta}$ er da

$$\mathsf{V}[\widetilde{\eta}] = \sigma^2 \, \widetilde{\mathbf{x}}[\mathbf{X}^T \mathbf{W}(\boldsymbol{\beta}) \mathbf{X}]^{-1} \widetilde{\mathbf{x}}^T \,, \tag{2.5.67}$$

hvor $\mathbf{W}(\boldsymbol{\beta})$ er givet ved

$$\mathbf{W}(\boldsymbol{\beta}) = \mathrm{diag}\left\{ \frac{w_i}{[g'(\mu_i)]^2 V(\mu_i)} \right\} \,,$$

og hvor $\mathbf{X}$ som før angiver den $k \times m$-dimensionale modelmatrix der blev brugt ved estimation af $\widehat{\boldsymbol{\beta}}$.

Variansen på den forudsagte middelværdi $\widetilde{\mu}$ er

$$\mathsf{V}[\widetilde{\mu}] \simeq \frac{\mathsf{V}[\widetilde{\eta}]}{[g'(\widetilde{\mu})]^2}$$

og variansen på forudsigelsen af observationen $\widetilde{Y}$ svarende til sættet $\widetilde{\mathbf{x}}$ af forklarende variable er

$$\mathsf{V}[\widetilde{Y}] \simeq V(\widetilde{\mu}) + \frac{\mathsf{V}[\widetilde{\eta}]}{[g'(\widetilde{\mu})]^2}$$

□

2.6 Test for modeltilpasning i generaliseret lineær model

Vi vil i dette afsnit diskutere test for modeltilpasning af en given generaliseret lineær model. Vi vil se, at likelihood-ratio testet bygger på devianserne mellem observationerne og de fittede værdier, og vi vil betragte forskellige andre tests for modeltilpasning.

Vi vil stadig betragte den generaliserede lineære model for observationssættet $Y_1, \ldots, Y_k$, hvor $Y_1, \ldots, Y_k$ er indbyrdes uafhængige, variable, hvis fordelinger tilhører en eksponentiel dispersionsmodel med samme variansfunktion $V(\mu)$ og samme dispersionsparameter σ^2.

Vi betragter den generaliserede lineære model givet ved (2.4.1), nemlig at

$$H_0 \ : \ \boldsymbol{\eta} - \boldsymbol{\eta}_0 \in L \ ,$$

hvor L er et lineært underrum af $\mathbb{R}^k$ af dimension m, og hvor $\boldsymbol{\eta}_0$ angiver en vektor af kendte offsetværdier og hvor transformationen $\eta = g(\mu)$ er specificeret ved linkfunktionen $g(\cdot)$..

Vi vil antage, at der valgt en basis for L sådan at hypotesen er på formen (2.4.2), dvs

$$H_0 : \quad \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} \ , \tag{2.6.1}$$

hvor modelmatricen $\mathbf{X}$ er en $k \times m$-dimensional matrix af fuld rang (m).

2.6.1 Residualdevians svarende til generaliseret lineær model

Definition 2.6.1 ***Residualdevians, skaleret residualdevians***

Betragt den generaliserede lineære model (2.6.1) for observationerne $Y_1, \ldots Y_k$.

Ved residualdeviansen $D(\mathbf{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))$ for hypotesen (2.6.1) vil vi forstå summen

$$D(\mathbf{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})) \stackrel{\text{DEF}}{=} \sum_{i=1}^{k} w_i d(y_i; \hat{\mu}_i) \tag{2.6.2}$$

hvor $d(y_i; \hat{\mu}_i)$ angiver deviansen svarende til observationen y_i og den fittede værdi $\widehat{\mu}_i$ (2.2.8), og hvor w_i angiver de eventuelle vægte.

Ved den skalerede residualdevians $D^*(\mathbf{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))$ for hypotesen (2.6.1) vil vi forstå summen

$$D^*(\mathbf{y};\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})) \stackrel{\text{DEF}}{=} \frac{D(\mathbf{y};\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))}{\sigma^2} \tag{2.6.3}$$

□

Bemærkning 1 ***Residualdeviansen er summen af de kvadrerede deviansresidualer***

Det følger af definitionen på deviansresidualet (def. 2.5.3 på side 202), at

$$r_D(y_i;\widehat{\mu}_i)^2 = w_i d(y_i,\widehat{\mu}_i)$$

□

Sætning 2.6.1 ***Kvotienttest for modeltilpasning i generaliserede lineære modeller***

Betragt den generaliserede lineære model (2.6.1) for observationerne $Y_1,\ldots Y_k$.

Kvotienttesstørrelsen for hypotesen

$$H_0 : \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} \tag{2.6.4}$$

imod den fulde model har teststørrelsen

$$G^2(H_0) = D^*(\mathbf{y};\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})) \tag{2.6.5}$$

hvor $D^*(\mathbf{y};\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))$ angiver den (evt skalerede) residualdevians (2.6.3).

Under hypotesen (2.6.1) vil teststørrelsen $G^2(H_0)$ approximativt følge en $\chi^2(k-m)$-fordeling, hvor k angiver dimensionen af den fulde model og hvor m angiver dimensionen af modellen svarende til H_0.

Hypotesen forkastes for store værdier af $G^2(H_0)$.

Bevis:
Kvotienttesstørrelsen er

$$G^2(H_0) = 2[\sup_{\mu\in M} l_\mu(\boldsymbol{\mu};\mathbf{y}) - \sup_{\mu=X\beta} l_\mu(\boldsymbol{\mu};\mathbf{y})] \tag{2.6.6}$$

hvor $l_\mu(\mu;\mathbf{y})$ angiver loglikelihoodfunktionen (2.1.3).

Det følger da af definitionen på enhedsdeviansen (side 142), at den (bortset fra vægtfaktoren w_i og dispersionsparameteren σ^2) netop måler bidraget fra den enkelte observation til kvotienttesteststørrelsen (2.6.6).

Beviset for fordelingsforholdene følger af overvejelser i lighed med lemma 2.5.4. □

Bemærkning 1 *Hvis dispersionsparameteren er 1, er kvotienttesteststørrelsen blot residualdeviansen* $D(\mathbf{y};\mu(\widehat{\beta}))$. □

Eksempel 2.6.1 *Fosterdødelighed hos mus*

Vi betragter atter data fra eksempel 2.4.1 på side 172.

I eksempel 2.5.4 på side 205 bestemte vi deviansresidualerne svarende til en logistisk regressionsmodel.

Som en vurdering af tilpasningen kan de fittede værdier $\hat{p}_i$ og de (vægtede) deviansbidrag $w_i d(y_i;\hat{p}_i)$ beregnes.

dosis	fost	andel	fit	devbidrag
0.0	297.0	0.05051	0.03740	1.28065
62.5	242.0	0.07025	0.05475	1.03504
125.0	312.0	0.07051	0.07949	0.35615
250.0	299.0	0.12709	0.16102	2.71072
500.0	285.0	0.50526	0.48665	0.39492
sum				5.7775

Residualdeviansen er $D(\mathbf{y};\widehat{\boldsymbol{p}}) = 5.78$, der skal sammenlignes med fraktilerne i en $\chi^2(3)$-fordeling. Idet $\chi^2(3)_{0.95} = 7.81$ er der således ingen grund til at afvise regressionsmodellen.

De enkelte afvigelser vurderes ved at betragte deviansresidualerne, eller Pearson-residualerne. Disse residualer er angivet i nedenstående skema.

dosis	Pears-res	dev-res
0.0	1.19	1.13

62.5	1.06	1.02
125.0	-0.59	-0.60
250.0	-1.60	-1.65
500.0	0.63	0.63

Ingen af residualerne overstiger ± 2, så der er ingen grund til at afvise modellen. □

Definition 2.6.2 ***Pearson-teststørrelse for modeltilpasning***

Betragt den generaliserede lineære model (2.6.1) for observationerne $Y_1, \ldots Y_k$.

Pearson-teststørrelsen, X^2, for test af hypotesen H_0 mod den fulde model defineres som kvadratsummen af Pearson-residualerne svarende til de fittede værdier $\widehat{\mu}$

$$X^2 \stackrel{\text{DEF}}{=} \sum_{i=1}^{k} r_P(y_i; \hat{\mu}_i)^2 \tag{2.6.7}$$

hvor Pearson-residualerne $r_P(y_i; \hat{\mu}_i)$ er givet ved (2.5.39)

Tilsvarende defineres den skalerede Pearson-teststørrelse, $(X^*)^2$ som

$$(X^*)^2 \stackrel{\text{DEF}}{=} \frac{X^2}{\sigma^2} \tag{2.6.8}$$

□

Bemærkning 1 ***Pearson-teststørrelsen er første led i Taylor approximationen til residualdeviansen***

Følger af bemærkning 1 på side 203 □

Sætning 2.6.2 ***Fordeling af Pearson-teststørrelsen for modeltilpasning***

Betragt den generaliserede lineære model (2.6.1) for observationerne $Y_1, \ldots Y_k$.

Under hypotesen (2.6.1) gælder, at den skalerede Pearson-teststørrelse $(X^*)^2$ givet ved (2.6.8) (svarende til de skalerede Pearson residualer) approximativt følger en $\chi^2(k-m)$-fordeling.

Hypotesen forkastes således for store værdier af $(X^*)^2$

Bevis:
Sætningen følger af ovenstående bemærkning i forbindelse med sætning 2.6.1. □

Definition 2.6.3 ***Wald-teststørrelse for modeltilpasning***

Betragt den generaliserede lineære model (2.6.1) for observationerne $Y_1, \ldots Y_k$.

Wald-teststørrelsen, X^2_W, for test af hypotesen H_0 mod den fulde model defineres som kvadratsummen af Wald-residualerne svarende til de fittede værdier $\widehat{\boldsymbol{\mu}}$

$$X^2_W \stackrel{\text{DEF}}{=} \sum_{i=1}^{k} r_W(y_i;\hat{\mu}_i)^2 \tag{2.6.9}$$

hvor Wald-residualerne $r_W(y_i;\hat{\mu}_i)$ er givet ved (2.5.42) Tilsvarende defineres den skalerede Wald-teststørrelse, $(X^*_W)^2$ som

$$(X^*_W)^2 \stackrel{\text{DEF}}{=} \frac{X^2}{\sigma^2}\,. \tag{2.6.10}$$

□

Sætning 2.6.3 ***Fordeling af Wald-teststørrelsen for modeltilpasning***

Betragt den generaliserede lineære model (2.6.1) for observationerne $Y_1, \ldots Y_k$. Under hypotesen (2.6.1) gælder, at den skalerede Wald-teststørrelse $(X^*_W)^2$ givet ved (2.6.10) approximativt følger en $\chi^2(k-m)$-fordeling.

Hypotesen forkastes således for store værdier af $(X^*_W)^2$
Bevis:
Sætningen følger af overvejelser i lighed med med lemma 2.5.4. □

Bemærkning 1 ***Brug af residualerne til modelkontrol***

De enkelte observationers bidrag $w_i d(y_i;\hat{\mu}_i)$ til $D^*(\mathbf{y};\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))$ kan tages som udtryk for denne observations afvigelse fra hypotesen. De enkelte observationers afvigelser kan vurderes ved at sammenligne de standardiserede residualer med fraktilerne i en standardiseret normalfordeling. □

Bemærkning 2 *Godhed af χ^2-approximationerne*

Godheden af χ^2-approximationen til fordelingen af de betragtede teststørrelser er nært knyttet til godheden af normalfordelingsapproximationen til fordelingen af de enkelte residualer. Der foreligger ikke mange universelt gyldige undersøgelser af disse approximationers godhed. McCullagh og Nelder (1989) har sammenlignet fordelingen af Pearson-residualerne med deviansresidualerne og konkluderet, at fordelingen af deviansresidualerne er nærmere en normalfordeling, end fordelingen af Pearson-residualerne. □

2.6.2 Estimation af dispersionsparameteren σ^2

Såfremt dispersionsparameteren, σ^2, er kendt, kan man umidddelbart benytte teststørrelsen for modeltilpasning til denne model, $G^2(H_0) = D^*(\mathbf{y}; \widehat{\boldsymbol{\mu}})$ til at vurdere, hvorvidt modellen kan antages at holde.

Såfremt dispersionsparameteren, σ^2, ikke er kendt, må man antage at modellen er dækkende, og man kan da benytte nedenstående fremgangsmåde til estimation af dispersionsparameteren

Sætning 2.6.4 *Estimation af dispersionsparameter*

Betragt den generaliserede lineære model (2.6.1) for observationerne $Y_1, \ldots Y_k$ og antag at dispersionsparameteren σ^2 er ukendt.

Såfremt modellen (2.6.1) er gyldig, kan man estimere σ^2 ud fra residualdeviansen ved

$$\widehat{\sigma^2} = \frac{D(\mathbf{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))}{k - m} \tag{2.6.11}$$

eller med udgangspunkt i Pearson residualerne

$$\widehat{\sigma^2} = \frac{\sum_1^k r_P(y_i; \hat{\mu}_i)^2}{k - m} \tag{2.6.12}$$

Bevis:

Estimationen består i at sætte den observerede værdi af teststørrelsen lig med den asymptotiske middelværdi. □

Såfremt likelihoodfunktionen med hensyn til σ^2 er et rimeligt simpelt udtryk, kan man også foretage en maksimum likelihood estimation ved at maksimere likelihoodfunktionen - også som funktion af σ^2.

I eksempel 2.7.2 vil vi illustrere en sådan maksimum-likelihood estimation.

2.7 Eksempler på regressions- og homogenitetsmodeller

2.7.1 Regressionsmodeller

Eksempel 2.7.1 ***Regression for normaltfordelte variable***

Lad observationerne $Y_1, Y_2, \ldots, Y_k n$ være indbyrdes uafhængige normalfordelte variable med samme varians σ^2.

Fordelingerne tilhører en (reproduktiv) eksponentiel dispersionsmodel, hvor den kanoniske parameter ϑ netop er middelværdien μ, enhedsvariansfunktionen er $V(\mu) = 1$, og hvor dispersionsparameteren er σ^2.

Enhedsdeviansen $d(y_i; \hat{\mu}_i)$ svarende til den fittede værdi $\hat{\mu}_i$ er jvf. tabel 2.3

$$d(y_i; \hat{\mu}_i) = (y_i - \hat{\mu}_i)^2 ,$$

hvor de fittede værdier $\hat{\mu}_i$ under modellen bestemmes ved $\hat{\mu}_i = \hat{\alpha} + \hat{\beta} x_i$.

Under regressionsmodellen bliver enhedsdeviansbidragene således

$$d(y_i; \hat{\mu}_i) = [y_i - (\hat{\alpha} + \hat{\beta} x_i)]^2$$

Residualdeviansen

$$D(\mathbf{y}; \widehat{\boldsymbol{\mu}}) = \sum_{i=1}^{k} d(y_i; \hat{\mu}_i)$$

er derfor kvadratafvigelsessummen

$$D(\mathbf{y}; \widehat{\boldsymbol{\mu}}) = \sum_{i=1}^{k} [y_i - (\hat{\alpha} + \hat{\beta} x_i)]^2 ,$$

Vi ved (f.eks. fra Introduktion til Statistik, Bind 1), at fordelingen af $D(\mathbf{y};\widehat{\boldsymbol{\mu}})$ er en $\sigma^2\chi^2(k-2)$-fordeling, dvs. at den skalerede residualdevians, $D^*(\mathbf{y};\widehat{\boldsymbol{\mu}})$, netop følger en $\chi^2(k-2)$-fordeling. Approximationen er altså eksakt i dette tilfælde.

Vi skal nedenfor (i eksempel 2.7.2) diskutere estimationen af σ^2.

Vi vil afslutte dette eksempel med at betragte matrixfremstillingen af modellen.

Enhedsdeviansen $d(y_i;\mu_i)$ svarende til en parameterværdi (α,β)

$$d(y_i;\mu_i(\alpha,\beta)) = [y_i - (\alpha+\beta x_i)]^2\,,$$

er den kvadratiske afvigelse mellem den observerede værdi y_i og den fittede værdi $\mu_i = \alpha+\beta x_i$ svarende til parameterværdien (α,β)

Maksimum-likelihood estimatet $\widehat{\boldsymbol{\beta}}$ er netop den værdi af $\boldsymbol{\beta}$, der minimerer kvadratafvigelsessummen $\sum[y_i-(\alpha+\beta x_i)]^2$. Udtrykt på matrixform har vi nemlig

$$\sum[y_i-(\alpha+\beta x_i)]^2 = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$$

Da $\tau(\vartheta)$ netop er identiteten, er middelværdiligningen (2.5.7) til bestemmelse af $\widehat{\boldsymbol{\beta}}$:

$$\mathbf{X}^T(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}) = 0\,. \tag{2.7.1}$$

Det er velkendt, at såfremt $\mathbf{X}$ har fuld rang, da er løsningen

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\,. \tag{2.7.2}$$

Betragt nu opspaltningen

$$(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = (\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})^T(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}) + (\mathbf{X}\widehat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{X}\widehat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta}) \tag{2.7.3}$$

af kvadratafvigelsessummen $\sum[y_i-(\alpha+\beta x_i)]^2$, hvor produktleddene forsvinder på grund af relationen (2.7.1)

Opspaltningen (2.7.3) udtrykker kvadratafvigelsessummen svarende til en vilkårlig værdi $\boldsymbol{\beta}$ som en sum af kvadratafvigelserne $(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})^T(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})$ svarende til afvigelserne omkring $\widehat{\boldsymbol{\beta}}$ og leddet

$$(\mathbf{X}\widehat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{X}\widehat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta}) = (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\mathbf{X}^T\mathbf{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})$$

svarende til afvigelserne mellem $\widehat{\boldsymbol{\beta}}$ og $\boldsymbol{\beta}$. Kvadratafvigelsessummen (2.7.3) antager derfor sin minimale værdi, netop for $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$.

Estimatet $\widehat{\boldsymbol{\beta}}$ bestemt ved (2.7.2) er altså den parameterværdi, der minimerer summen af kvadratiske afvigelser mellem observationerne $\mathbf{y}$ og de fittede værdier $\widehat{\boldsymbol{\mu}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$.

Opfatter vi observationssættet $\mathbf{y}$ som et punkt i det k-dimensionale rum, bemærker vi, at ligningen (2.7.1) udtrykker at residualerne

$$\mathbf{y} - \widehat{\boldsymbol{\mu}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$$

er ortogonale på den plan (2-dimensionale underrum) i $\mathbb{R}^k$, som er udspændt af søjlerne i modelmatricen $\mathbf{X}$. Estimatet $\widehat{\boldsymbol{\beta}}$ angiver koordinaterne (i denne plan) til projektionen af observationssættet $\mathbf{y}$ ned på denne plan, og de fittede værdier $\mathbf{X}\widehat{\boldsymbol{\beta}}$ angiver koordinaterne i det sædvanlige koordinatsystem i $\mathbb{R}^k$ til denne projektion.

Hatmatricen er her

$$\mathbf{H} = \mathbf{X}\left[\mathbf{X}^T\mathbf{X}\right]^{-1}\mathbf{X}^T$$

□

Eksempel 2.7.2 ***Regression for normaltfordelte variable, estimation af dispersionsparameteren***

Vi vil nu diskutere estimation af dispersionsparameteren σ^2 i det foregående eksempel ved maksimum likelihood metoden. Profillikelihood'en (2.1.4) for σ^2 kan udtrykkes som

$$\widetilde{L}(\sigma^2; \boldsymbol{y}) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} D(\mathbf{y}; \widehat{\boldsymbol{\mu}}) \, , \right]$$

hvilket viser, at størrelsen $T = D(\mathbf{y}; \widehat{\boldsymbol{\mu}})$ er likelihood-sufficient (def. 2.1.5) for σ^2.

Det følger da af bemærkning 4 på side 120, at man skal benytte likelihoodfunktionen svarende til den marginale fordeling for T.

I Introduktion til Statistik, Bind 1så vi, at fordelingen af $D(\mathbf{y}; \widehat{\boldsymbol{\mu}})$ er en $\sigma^2\chi^2(n-2)$-fordeling, dvs at $D(\mathbf{y}; \widehat{\boldsymbol{\mu}})$ følger en $\mathrm{G}((n-2)/2, 2\sigma^2)$-fordeling.

Det følger da i lighed med eksempel 2.2.7, at fordelingen af $T/f = D(\mathbf{y};\widehat{\boldsymbol{\mu}})/f$ med $f = n-2$ kan beskrives ved en (reproduktiv) eksponentiel dispersionsmodel med middelværdiparameter

$$\mathsf{E}\,[T/f] = \sigma^2\ ,$$

og da der kun er én observation af T/f har vi altså en fuld model, og det følger da af bemærkningen i sidste linie af definition 2.4.2 på side 167, at maksimum-likelihood estimatet for σ^2 fås ved at sætte observationen lig med sin middelværdi, dvs vi får (det marginale) maksimum-likelihood estimat

$$\widehat{\sigma}^2 = \frac{D(\mathbf{y};\widehat{\boldsymbol{\mu}})}{n-2}$$

altså netop det sædvanlige estimat, kendt fra Introduktion til Statistik, Bind 1.

Vi bemærker iøvrigt, at dette er det samme estimat, som vi ville have fået, hvis vi havde brugt sætning 2.6.4 på side 224. Normalt er dette ikke tilfældet. Sædvanligvis vil maksimum-likelihood estimatet for dispersionsparameteren skulle bestemmes ved iteration, og estimatet vil være forskelligt fra det simple devians-estimat bestemt ved (2.6.11) eller Pearson-estimat (2.6.12). □

Eksempel 2.7.3 ***Ramushøjder***

Nedenstående tabel viser målinger af højden af ramusknoglen hos en dreng. Målingerne er foretaget med halvårlige intervaller fra hans 8-års fødselsdag.

Alder [År]	højde [mm]
8	51.2
8 1/2	53.0
9	54.3
9 1/2	54.5

Antager vi nu, at ramushøjden Y_i ved alderen x_i kan beskrives ved en $\mathrm{N}(\mu_i, \sigma^2)$ fordelt variabel, har vi modellen

$$\mu_i = \alpha + \beta x_i\ , i = 1, 2, \ldots, 4$$

Vi vil illustrere brugen af modelmatricen til bestemmelse af løsningen.

For at simplificere beregningerne vælger vi at regne alderen x_i som alderen i år minus 8.75.

Denne transformation ændrer ikke analysen, men det er klart, at de fundne værdier af koefficienterne skal fortolkes i lys af dette valg af måleenhed.

Med dette valg har vi modelmatricen:

$$\mathbf{X} = \begin{pmatrix} 1.0 & -0.75 \\ 1.0 & -0.25 \\ 1.0 & 0.25 \\ 1.0 & 0.75 \end{pmatrix} \quad \text{med} \quad \mathbf{X}^T\mathbf{X} = \begin{pmatrix} 4.00 & 0.00 \\ 0.00 & 1.25 \end{pmatrix}$$

hvorfor

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 0.25 & 0.00 \\ 0.00 & 0.80 \end{pmatrix},$$

der fører til estimaterne $\widehat{\alpha}_1 = 53.25$, $\widehat{\beta} = 2.24$

De fittede værdier $\hat{\mu}_i$ bestemmes da som

$$\hat{\mu}_i = 53.25 + 2.24\, x_i$$

og enhedsdeviansen svarende til den i'te observation er

$$d(y_i, \hat{\mu}_i) = (y_i - 53.25 - 2.24\, x_i)^2$$

De fittede værdier og deviansbidragene er angivet i nedenstående tabel:

Alder [År]	x_i	højde [mm] y_i	fittet værdi $\hat{\mu}_i$	deviansbidrag $d(y_i, \hat{\mu}_i)$
8	-0.75	51.2	51.57	0.1369
8 1/2	-0.25	53.0	52.69	0.0961
9	0.25	54.3	53.81	0.2401
9 1/2	0.75	54.5	54.93	0.1849
sum			0.6580	

Da vi ikke kender dispersionsparameteren σ^2, kan vi ikke teste for modeltilpasning, men såfremt vi antager at den lineære model kan opretholdes, kan vi estimere dispersionsparameteren σ^2 ved

$$\widehat{\sigma}^2 = 0.6580/(4-2) = 0.3290$$

svarende til

$$\widehat{\sigma} = 0.5736 \text{ [mm]}$$

□

Eksempel 2.7.4 *Gamma regression*

Lad observationerne $Y_1, Y_2, \ldots, Y_k$ være indbyrdes uafhængige gammafordelte variable med $Y_i \in \mathrm{G}(\alpha_i, \mu_i/\alpha_i)$.

Modellen

$$H_0 : \mu_i = \gamma\, x_i, \quad i = 1, 2, \ldots, k\,, \tag{2.7.4}$$

hvor $\mu_i = \mathsf{E}\,[Y_i]$ og $x_1, x_2, \ldots, x_k$ er kendte kovariater, kaldes Gamma regressionsmodellen.

Idet den kanoniske linkfunktion for gammafordelingen er den reciproke $\vartheta = -1/\mu$, (jvf. eksempel 2.3.2), ser vi, at modellen (2.7.4) svarer til

$$H_0 : \vartheta_i = \frac{1}{\gamma x_i}\,,$$

dvs

$$H_0 : \vartheta_i = \frac{1}{\gamma}\, t_i$$

med $t_i = 1/x_i$. Modellen er således af formen (2.5.8), hvor interceptleddet α antages at være nul. Middelværdiligningen (2.5.11) reduceres her til den enkelte ligning

$$\sum_{i=1}^{k} \alpha_i y_i \frac{1}{x_i} = \sum_{i=1}^{k} \alpha_i \gamma\, x_i \frac{1}{x_i} \tag{2.7.5}$$

med løsningen

$$\widehat{\gamma} = \sum_{i=1}^{k} \alpha_i \frac{y_i}{x_i} \Big/ \sum_{i=1}^{k} \alpha_i \tag{2.7.6}$$

og de fittede værdier $\hat{\mu}_i$ under modellen (2.7.4) bliver

$$\hat{\mu}_i = \hat{\gamma} x_i \tag{2.7.7}$$

Såfremt formparameteren er den samme for alle observationer, $\alpha_i = \alpha$, bliver estimatet blot

$$\widehat{\gamma} = \frac{1}{k} \sum_{i=1}^{k} \frac{y_i}{x_i} \tag{2.7.8}$$

Vi bemærker, at gammafordelingens variansstruktur, $V(\mu) = \mu^2$, svarende til en konstant variationskoefficient $\sqrt{\mathsf{V}[Y_i]}/\mathsf{E}[Y_i]$, har den konsekvens, at estimatet for den fælles hældning netop bliver det simple aritmetiske gennemsnit af de individuelle hældninger. □

Eksempel 2.7.5 ***Regressionsmodel for empiriske varianser***

Ved et forsøg til bestemmelse af vaskeevnen for et vaskemiddel foretog man en ensartet tilsmudsning af en række lapper, hvorefter de blev vasket og renheden bestemt ved brug af et reflektometer.

Ved forsøget foretoges vask af hhv 3, 5 og 7 lapper samtidig. Nedenstående tabel viser de empiriske varianser for hvert af de tre forsøg.

Antal lapper x_i	3	5	7
SAK	2.5800	4.8920	4.9486
Frihedsgrader f_i	2	4	6
Empirisk varians s_i^2	1.2900	1.2230	0.8248

Man ønsker nu at vurdere, om variansen kan antages at aftage proportionalt med det reciprokke antal $1/x'$ af lapper (for antal imellem 3 og 7).

Idet det antages, at renheden efter vask kan beskrives ved en normalfordelt størrelse med en middelværdi og en varians, der kun afhænger af antallet af lapper følger det, af eksempel 2.2.7, at de empiriske varianser, $s_i^2 \in \mathrm{G}\big(f_i/2, \sigma_i^2/(f_i/2)\big)$ med $\mathsf{E}[s_i^2] = \sigma_i^2$.

Man ønsker at vurdere modellen

$$H_0 : \sigma_i^2 = \gamma/x_i', \quad i = 1, 2, 3,$$

dvs netop regressionsmodellen (2.7.4) med $x_i = 1/x_i'$.

Vi får derfor estimatet (2.7.5)

$$\widehat{\gamma} = \sum_{i=1}^{3}(f_i/2)s_i^2 x_i' \Big/ \sum_{i=1}^{3}(f_i/2) = 5.57$$

Nedenstående tabel viser de fittede værdier og deviansresidualerne, bestemt ved hjælp af enhedsdeviansen for Gammafordelingen (Tabel 2.3).

$$d(y;\mu) = 2 \times \{y/\mu - \ln(y/\mu) - 1\}$$

Antal lapper x_i	3	5	7
Empirisk varians s_i^2	1.2900	1.2230	0.8248
Fittet værdi	1.8566	1.1140	0.7957
$d(s_i^2; \widehat{sigma_i}^2)$	0.11785	0.00900	0.00130
$(f_i/2)d(s_i^2; \widehat{sigma_i}^2)$	0.11785	0.01800	0.00391
$r_D(s_i^2; \widehat{sigma_i}^2)$	-0.34329	0.13416	0.06254

Man finder goodness of fit teststørrelsen $D^*(\mathrm{s}^2; \widehat{\boldsymbol{\sigma}}^2) = 0.1398$, der skal sammenlignes med fraktilerne i en $\chi^2(2)$-fordeling. Der er således ingen grund til at afvise modellen.

Modellen kunne analyseres ved SAS® eller S-Plus ved at vælge en gammafordeling med vægtene $f_i/2$, kanonisk link og forklarende variabel x' (svarende til at $1/\sigma_i^2 = x'/\gamma$). Programmet ville da estimere størrelsen $1/\gamma$, mens de fittede værdier mv. ville være som i skemaet ovenfor. □

2.7.2 Homogenitetshypotesen, den minimale model

Sætning 2.7.1 ***Homogenitetstest for eksponentielle dispersionsmodeller***

Antag, at fordelingen for $Y_1, Y_2, \ldots, Y_k$ kan beskrives ved at $Y_1, Y_2, \ldots, Y_k$ er uafhængige variable, hvis fordelinger tilhører eksponentielle dispersionsmodeller med samme variansfunktion $V(\mu)$ og samme dispersionsparameter σ^2.

Betragt homogenitetshypotesen

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \tag{2.7.9}$$

mod alternativet (Den fulde model):

$$H_1 : \text{der findes mindst ét par} \quad (i,j)\text{så} \quad \mu_i \neq \mu_j$$

Hypotesen er den samme, hvilken link-funktion, vi end vælger, så vi kan lige så godt formulere hypotesen ved den kanoniske link, $g(\mu) = \tau^{-1}(\mu)$, dvs hypotesen bliver da

$$H_0 : \vartheta_1 = \vartheta_2 = \cdots = \vartheta_k \tag{2.7.10}$$

Modelmatricen er

$$\mathbf{X} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

Da hypotesen er formuleret ved den kanoniske link, kan vi udnytte bemærkning 1 på side 180.

Middelværdiligningerne (2.5.6) reduceres her til en enkelt ligning :

$$\sum_{i=1}^{k} y_i = \sum_{i=1}^{k} \tau(\alpha) \tag{2.7.11}$$

og estimatet for μ_i under hypotesen er da det fælles estimat:

$$\hat{\mu} = \sum_{i=1}^{k} y_i/k = \bar{y} \tag{2.7.12}$$

For en vægtet model finder man tilsvarende af (2.5.11) at den fælles middelværdi μ estimeres ved

$$\hat{\mu} = \sum_{j=1}^{k} w_j y_j / \sum_{j=1}^{k} w_j = \bar{y}_w \tag{2.7.13}$$

Kvotienttesstørrelsen $G^2(H_0)$ (2.6.5) følger en $\chi^2(k-1)$-fordeling.

Eksempel 2.7.6 ***Kvartalsvise uheldstal for motorkøretøjer***

Nedenstående tabel viser de kvartalsvise uheldstal for uheld med personskade motorkøretøjer for uheldskategorien "møde" i dagslys for ikke-spiritus-påvirkede førere for året 1987 i Danmark

Indeks i	Antal uheld y_i	kvartal
1	128	1
2	95	2
3	100	3
4	75	4

Vi ønsker at undersøge, om der er væsentlig forskel på ulykkestallene i de fire kvartaler, eller om de observerede variationer kan forklares som tilfældige variationer.

Vi opstiller modellen $Y_i \in \mathrm{P}(\lambda_i)$, og hypotesen om homogenitet er da $\lambda_1 = \ldots = \lambda_4$

Modellen svarer til ovenstående sætning 2.7.1.

Under den fulde model estimeres λ_i ved observationen y_i, og under hypotesen fås estimatet

$$\hat{\lambda}_i = \overline{y}_\cdot = \frac{128 + 95 + 100 + 75}{4} = \frac{398}{4} = 99.5 \text{ [ulykker/kvt]}$$

Deviansbidragene bliver jvf tabel 2.3

$$d_i(y_i) = 2(y_i \ln(y_i/\overline{y}_\cdot) - (y_i - \overline{y}_\cdot))$$

altså et udtryk for forskellen mellem observationens relative afvigelse fra gennemsnitsværdien $y_i/\overline{y}_\cdot$ og den absolutte afvigelse $y_i - \overline{y}_\cdot$.

Nedenstående tabel viser deviansbidragene:

Indeks i	Antal uheld y_i	devians-bidrag $d(y_i;\hat{\mu}_i)$
1	128	7.48
2	95	0.21
3	100	0.00
4	75	6.60
sum	398	14.28

Sammenligner vi med $\chi^2(3)$-tabellen ser vi, at $\chi^2_{0.995}(3) = 12.8$. Selv ved et test på 0.5 % niveauet vil vi altså afvise hypotesen om samme kvartalsvise uheldsfrekvens.

Vi må altså modellere de kvartalsvise uheldstal med hver sin parameter. En "naturlig" parametrisering kan være en additiv parametrisering af den kanoniske parameter $\vartheta_i = \ln \lambda_i$.

Parametriseringen

$$\vartheta_i = \vartheta_0 + \alpha_i;\ i = 1, 2, 3, 4$$

svarende til

$$\lambda_i = \lambda_0 \times \rho_i;\ i = 1, 2, 3, 4$$

er ikke entydig med mindre vi lægger et bånd på parametrene α_i, $i = 1, 2, 3, 4$.

Båndet $\sum \alpha_i = 0$ fastsætter $\vartheta_0 = \sum_i \vartheta_i/4$, altså

$$\lambda_0 = \{\prod_{i=1}^{4} \lambda_i\}^{1/4}$$

altså det geometriske gennemsnit af de enkelte kvartalers hyppighed.

Man kunne også vælge et af kvartalerne, f.eks. januar kvt. som referencekvartal, d.v.s. $\vartheta_0 = \vartheta_1$, svarende til $\alpha_1 = 0$ og til $\lambda_0 = \lambda_1$ med $\rho_1 = 1$. Under den sidste parametrisering vil ρ_i, $i = 2, 3, 4$ således angive en "indeksfaktor" for det i'te kvartal i relation til januar kvartal.

Vi skal senere vende tilbage til en diskussion af forskellige parametriseringer af en sådan model. □

Eksempel 2.7.7 ***Model for Poisson- "regression", offset-værdi***

Lad $Z_1, Z_2, \ldots, Z_k$ være uafhængige observationer, hvor $Z_i \in \mathrm{P}(\lambda_i)$. Betragt hypotesen

$$\lambda_i = x_i \lambda, \quad i = 1, 2, \ldots, k \tag{2.7.14}$$

hvor x_i angiver kendte "regressionskonstanter", og hvor λ angiver en ukendt fælles parameter.

Hypotesen svarer til en hypotese af formen

$$\mathrm{E}\,[Z_i/x_i] = \lambda^* \tag{2.7.15}$$

Fordelingen af $Y_i = Z_i/x_i$ er en $\mathrm{P}(\lambda x_i)/x_i$-fordeling, svarende til en en additiv dispersionsmodel for Z_i med indeksparameter x_i og variansfunktion $V(\mu) = \mu$. Vi kan opfatte indeksparameteren som vægte.

Hypotesen (2.7.14) bliver da en homogenitetshypotese i denne model.

Vi har derfor, at maksimaliseringsestimatoren $\hat{\lambda}^*$ for den fælles værdi af λ^* under hypotesen fås af (2.7.13)

$$\hat{\lambda}^* = \sum_i z_i / \sum_i x_i$$

Vi bemærker, at vi også direkte kunne have formuleret en hypotese vedrørende de kanoniske parametre for fordelingen af Z_i ved at formulere hypotesen som

$$\vartheta_i = \ln x_i + \beta$$

svarende til en offset værdi $\vartheta_{0i} = \ln x_i$. □

Eksempel 2.7.8 ***Poisson-regression***

Nedenstående eksempel viser antallet af diagnosticerede lungecancertilfælde i årene 1968-1971 for mandlige indbyggere i aldersgruppen 55-59 år for hver af byerne Fredericia, Horsens, Kolding og Vejle. (Kilde Clemmesen et. al. 1974).

Antal diagnosticerede tilfælde af lungecancer blandt 55-59 årige mænd i årene 1968-1971 i fire provinsbyer.

		By				
		Fredericia	Horsens	Kolding	Vejle	Ialt
Antal tilfælde	z_i	11	6	8	7	32
Antal mænd i aldersgruppen	x_i	800	1083	1050	878	3811
Rate	y_i	13.75	5.54	7.62	7.97	8.40

Det er af interesse at vurdere, hvorvidt risikoen for lungecancer kan antages at være den samme for indbyggerne i de fire byer.

Vi indleder med at opstille en stokastisk model for data. Den naturlige model er at beskrive antallet af diagnosticerede cancertilfælde Z_i i den i'te by ved en $B(x_i, p_i)$-fordelt variabel, hvor x_i angiver antallet af mænd i byen i den pågældende aldersgruppe.

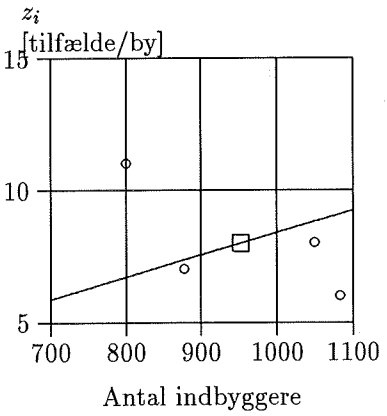
For at illustrere Poisson "regressionsmodellen" vælger vi imidlertid en Poisson model, dvs vi antager at antallet af diagnosticerede cancertilfælde Z_i

i den i'te by kan beskrives ved uafhængige variable $Z_i \in \mathrm{P}(\lambda_i)$, hvor λ_i angiver det underliggende forventede antal tilfælde i den i'te by. Enheden for λ_i er jo [antal tilfælde/by].

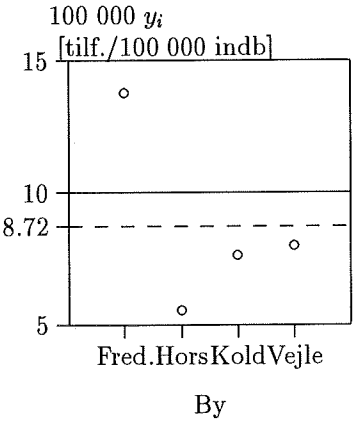
En umiddelbar sammenligning af de rå antal diagnosticerede tilfælde (dvs λ_i) synes ikke rimelig, da indbyggerantallene i den pågældende aldersgruppe ikke er ens i de fire byer. Man har derfor udregnet raten af diagnosticerede cancertilfælde pr 100 000 indbyggere ved normering af det observerede antal tilfælde med antallet af indbyggere (og multiplikation med 100 000). Det vil være naturligt at benytte den underliggende forventede værdi af denne rate som udtryk for aldersgruppens cancerrisiko i den pågældende by.

Nedenstaaende figur viser antallet af diagnosticerede tilfælde tegnet op mod indbyggerantallet samt de observerede rater:

Antal tilfælde pr by mod indbyggerantal



Antal tilfælde pr 100 000 indb. mod by



□ angiver gennemsnittet pr. by
Den indtegnede linie på figuren til venstre angiver linien igennem origo og stikprøvegennemsnittet, dvs. linien med hældning $\sum z_i / \sum x_i = 8.40$
Linien på figuren til højre angiver (det uvægtede) gennemsnit af de fire rater, $\sum(z_i/x_i)/4 = 8.72$.

I overensstemmelse med den valgte Poisson-model vil vi derfor vælge at parametrisere modellen ved $Z_i \in \mathrm{P}(x_i\beta_i)$, hvor x_i angiver den eksponerede population, dvs. indbyggertallet i den betragtede aldersgruppe. Parameteren $\beta_i = \lambda_i/x_i$ med enheden [antal tilfælde/indbygger] er netop udtryk for den forventede værdi af den observerede rate $Y_i = Z_i/x_i$.

Det er vigtigt at notere sig, at vi ikke bare kunne modellere de observerede rater ved Poisson-fordelte størrelser. Dels er Poissonfordelingen begrænset til heltallige værdier, mens de mulige værdier for de observerede rater er en række brudne tal mellem 0 og 1 (eller en række tal mellem 0 og 100 000). Dels vil usikkerheden på den observerede rate afhænge af nævneren, (indbyggertallet). Jo større indbyggertal, desto mindre vil den observerede rate afvige fra den underliggende forventede værdi. Under modellen $Z_i \in \mathrm{P}(x_i\beta_i)$ har vi jo

$$\mathrm{V}[Z_i] = x_i\beta_i$$

(nemlig lig forventningsværdien), mens

$$\mathrm{V}[Y_i] = \frac{\mathrm{V}[Z_i]}{x_i^2} = \frac{\beta_i}{x_i}$$

Ved normeringen har vi opnået at få nogle størrelser med sammenlignelige middelværdier, men varianserne er ikke ens. Disse forskelle i variansen udtrykkes ved vægtene w_i

Vi opstiller homogenitetshypotesen for raterne:

$$H_0 : \frac{\lambda_1}{x_1} = \frac{\lambda_2}{x_2} = \frac{\lambda_3}{x_3} = \frac{\lambda_4}{x_4} = \beta$$

svarende til (2.7.15)

Estimaterne under hypotesen fås som:

$$\widehat{\beta} = \frac{\sum z_i}{\sum x_i} = \frac{32}{3811} = \times 0.00840 \text{ [tilfælde/person]}$$

svarende til

$$\hat{\lambda}_i = x_i \times 0.00840 \text{ [tilfælde/by]}$$

Ved bestemmelsen af devianserne skal vi benytte enhedsdeviansen

$$d(y;\mu) = 2 \times \{y \ln(y/\mu) - (y - \mu)\}$$

		By				Ialt
		Fredericia	Horsens	Kolding	Vejle	
Antal tilfælde	z_i	11	6	8	7	32
Antal mænd i aldersgruppen	x_i	800	1083	1050	878	3811
Estimat under hypotesen	$\hat{\lambda}_i = \hat{\beta} x_i$	6.72	9.09	8.82	7.32	
devians-bidrag	$d(y_i, \hat{\beta})$	2.29	1.20	0.08	0.02	3.58
devians-residual	$r_D(y_i, \hat{\beta})$	1.51	-1.09	-0.28	-0.14	

Vi bemærker at hypotesen

$$H_0 : \frac{\lambda_1}{x_1} = \frac{\lambda_2}{x_2} = \frac{\lambda_3}{x_3} = \frac{\lambda_4}{x_4}$$

netop svarer til en regressionshypotese om en sammenhæng $\mathsf{E}\,[Z_i] = \beta x_i$, dvs $\beta_1 = \beta_2 = \beta_3 = \beta_4$,

og at estimationen under hypotesen ved

$$\hat{\lambda}_i = x_i \frac{\sum z_i}{\sum x_i}$$

udtrykker at den fælles værdi β estimeres ved

$$\widehat{\beta} = \frac{\sum z_i}{\sum x_i} = \frac{\sum y_i\, x_i}{\sum x_i}$$

altså et vægtet gennemsnit af de individuelle hældninger $\widehat{\beta}_i = y_i$ med vægtene x_i.

Til sammenligning kan anføres, at mindste kvadraters estimatet over den fælles hældning, β, under en normalfordelingsmodel, $Z_i \in \mathrm{N}(\beta x_i, \sigma^2)$ er

$$\widehat{\beta} = \frac{\sum z_i x_i}{\sum x_i^2} = \frac{\sum y_i\, x_i^2}{\sum x_i^2}$$

svarende til et vægtet gennemsnit af de individuelle hældninger $\widehat{\beta}_i = y_i = z_i/x_i$ med vægtene x_i^2, mens regressionen for gammafordelte variable i eksempel 2.7.4 førte til et simpelt (uvægtet) gennemsnit af de individuelle hældninger. □

Eksempel 2.7.9 ***Pearson teststørrelsen ved homogenitetstest i Poissonfordelingen***

Betragt atter situationen i eksempel 2.7.6

I eksemplet bestemte vi devianstesteststørrelsen svarende til homogenitetstestet for k Poissonfordelinger. Vi vil her betragte Pearsonteststørrelsen og Pearson-residualerne.

Indledningsvis vil vi illustrere det generelle resultat i bemærkning 1 på side 203 ved at udlede det direkte for denne situation.

Vi betragter Taylor-udviklingen af $y\ln(y)$ for $y \to 1$

$$y\ln(y) = (y-1) + \frac{(y-1)^2}{2} + \mathrm{O}((y-1)^2)$$

Under hypotesen vil $y_i/\overline{y}.$ være i nærheden af 1, hvorfor

$$y_i \ln(y_i/\overline{y}.) \simeq \overline{y}. \left(\frac{y_i}{\overline{y}.} - 1 + \frac{1}{2}\left(\frac{y_i}{\overline{y}.} - 1\right)^2 \right) ,$$

således at vi ved indsættelse i udtrykket

$$d_i(y_i) = 2(y_i \ln(y_i/\overline{y}.) - (y_i - \overline{y}.))$$

for enhedsdeviansen svarende til Poissonfordelingen har

$$D(\mathbf{y}; \widehat{\boldsymbol{\mu}}) \simeq \sum_{i=1}^{k} \frac{(y_i - \overline{y}.)^2}{\overline{y}.}$$

Hvilket netop er Pearson-teststørrelsen X^2.

Nedenstående tabel viser de to typer af residualer:

Indeks	Antal uheld	Fittet værdi	devians-bidrag	devians residual	Pearson residual	kvadr.
i	y_i	$\widehat{\mu}_i$	$d(y_i;\widehat{\mu}_i)$	$r_d(y_i)$	$r_P(y_i)$	r_P^2
1	128	99.5	7.48	2.74	2.85	8.16
2	95	99.5	0.21	-0.45	- 0.44	0.20
3	100	99.5	0.00	0.00	0.00	0.00
4	75	99.5	6.60	-2.56	-2.45	6.03
sum	398	99.5	14.28		0	14.39

Der er således en rimelig god overensstemmelse mellem de to teststørrelser. Begge teststørrelser leder til afvisning af homogenitetshypotesen ved test på et 0.5 % niveau.

□

Eksempel 2.7.10 *Test for varianshomogenitet for normalfordelte data*

Betragt k grupper af uafhængige normalfordelte stokastiske variable

$$X_{ij} \quad j = 1, 2 \ldots, n_i \; ; \; i = 1, 2 \ldots, k$$

med

$$\begin{aligned} \mathsf{E}\,[X_{ij}] &= \mu_i \\ \mathsf{V}\,[X_{ij}] &= \sigma_i^2 \end{aligned}$$

Vi ønsker at teste

$$H_0 \; : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 \tag{2.7.16}$$

mod alternativet

$$H_1 \; : \exists\, i, j \text{ så } \sigma_i^2 \neq \sigma_j^2$$

Lad

$$\begin{aligned} \overline{X}_{i\cdot} &= \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \\ SAK_i &= \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i\cdot})^2 \\ f_i &= n_i - 1 \\ s_i^2 &= \frac{1}{f_i} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i\cdot})^2 \end{aligned} \tag{2.7.17}$$

Vi så i eksempel 2.2.7, at familierne af fordelinger for S_i^2 kan beskrives ved reproduktive eksponentielle dispersionsmodelr med forventningsværdi

$$\mathsf{E}\,[s_i^2] = \sigma_i^2$$

variansfunktionen $V_G(\sigma^2) = (\sigma^2)^2$, kanonisk link $\vartheta = -1/\sigma^2$, og med vægten $w = f/2$.

Hypotesen (2.7.16) er netop en hypotese om middelværdihomogenitet for disse familier.

Lad σ^2 betegne den fælles varians under hypotesen. Under hypotesen har man jvf (2.7.13) estimatet

$$s^2 = \frac{\sum_{i=1}^{k}(f_i/2)s_i^2}{\sum_{i=1}^{k}(f_i/2)} \tag{2.7.18}$$

Man finder enhedsdeviansen svarende til Gamma-fordelingen i tabel 2.3 på side 144

$$d(s^2;\sigma^2) = 2 \times \{(s^2-\sigma^2)/\sigma^2 - \ln(s^2/\sigma^2)\}\ , \tag{2.7.19}$$

hvorfor det vægtede deviansbidrag er

$$w_i d(s_i^2;s^2) = (f_i/2)d(s_i^2;s^2) = f_i\{(s_i^2-s^2)/s^2 - \ln(s_i^2/s^2)\} \tag{2.7.20}$$

Man får således udtrykket for residualdeviansen

$$D(\mathbf{s}^2;\widehat{\sigma}^2) = \sum_{i=1}^{k}\frac{f_i(s_i^2-s^2)}{s^2} - \sum_{i=1}^{k} f_i \ln(s_i^2/s^2)$$

Den første sum er nul pga relationen (2.7.18) og teststørrelsen er derfor

$$G^2(H_0) = D(\mathbf{s}^2;\widehat{\sigma}^2) = f\ln(s^2) - \sum_{i=1}^{k} f_i \ln(s_i^2) \tag{2.7.21}$$

med

$$f = \sum_{i=1}^{k} f_i$$

Under hypotesen er $D(\mathbf{s^2};\widehat{\sigma}^2)$ approximativt fordelt som $\chi^2(k-1)$. Testet forkaster for store værdier af $D(\mathbf{s^2};\widehat{\sigma}^2)$

Testet er anført som Sætning 11 i Introduktion til Statistik, Bind 1, afsnit 5.3.

Testet kaldes også Bartlett's test efter den engelske statistiker M.S.Bartlett. Bartlett (1937) angav ikke blot teststørrelsen, men udledte desuden en korrektionsfaktor,

$$c = 1 + \frac{1}{3(k-1)} \left(\sum_{i=1}^{k} \frac{1}{f_i} - \frac{1}{f} \right), \tag{2.7.22}$$

til teststørrelsen $D(\mathbf{s^2};\widehat{\sigma}^2)$, sådan at fordelingen af

$$\frac{1}{c} D(\mathbf{s^2};\widehat{\sigma}^2)$$

svarer bedre til $\chi^2(k-1)$-fordelingen. Princippet for bestemmelsen af denne korrektionsfaktor kan generaliseres til den generelle kvotienteststørrelse (2.6.5) i sætning 2.6.1. Korrektionen kaldes "Bartlett-korrektionen".

Vi bemærker, at teststørrelsen (2.7.21) kan udtrykkes som

$$D(\mathbf{s^2};\widehat{\sigma}^2) = \ln\left(\left[\frac{s^2}{s_g^2} \right]^f \right)$$

hvor

$$s_g^2 = \left(\prod (s_i^2)^{f_i} \right)^{1/f}$$

angiver det geometriske gennemsnit af de empiriske varianser.

Testet måler således forholdet mellem det aritmetiske, s^2, og det geometriske gennemsnit, s_g^2, af de empiriske varianser. Testet forkaster, hvis det aritmetiske gennemsnit er markant større end det geometriske. □

Eksempel 2.7.11 ***Analyseusikkerhed ved forskellige kuvettelængder***

Nedenstående tabel viser resultaterne fra et kalibreringseksperiment med henblik på bestemmelsen af extinktionskoefficienten for cement ved forskellige titaniumindhold. (Efter Hald (1948))

Der blev undersøgt fire forskellige titaniumindhold, 0.1 %, 0.25 %, 0.50 % og 1.00 %. For hver titaniumværdi, i blev der udført 5 uafhængige bestemmelser af ekstinktionskoefficienten z_{ij}; $i = 1, \ldots, 4$; $j = 1, \ldots, 5$. Da ekstinktionskoefficienten er proportional med det absolutte titaniumindhold blev der brugt forskellige kuvettelængder for de forskellige titaniumindhold.

De registrerede størrelser z_{ij} er angivet som [ekstinktionskoefficient]

Den anvendte cuvettelængde er ligeledes anført i tabellen.

Titanium indhold	0.1 %	0.25 %	0.50 %	1.0 %	
Kuvette-længde l [cm]	5 [cm]	3 [cm]	2 [cm]	1 [cm]	
		Ext.koeff.			Sum
	237	267	308	304	
	235	246	310	300	
	247	260	330	311	
	222	251	304	301	
	245	273	316	322	
SAK	393	467	411	333	1604
f_i	4	4	4	4	16
s_i^2	98.20	116.80	102.80	83.30	
$f_i \ln(s_i^2)$	18.348	19.042	18.531	17.690	73.611

Man finder det fælles skøn

$$s^2 = \frac{1604}{16} = 100.25$$

med

$$f \ln(s^2) = 16 \ln(100.25) = 16 \times 4.608 = 73.723$$

Teststørrelsen (2.7.21) bliver altså

$$D(\mathbf{s^2}; \hat{\sigma}^2) = 73.723 - 73.611 = 0.112$$

Idet $\chi^2_{0.95}(3) = 6.25$ er der ingen grund til at afvise hypotesen. Den fundne størrelse svarer til ca 1 %-fraktilen så der er endog særdeles god overensstemmelse med hypotesen. □

2.8 Parametrisk repræsentation af modeller

Fil: glm3b.tex 98-02-16

2.8.1 Introduktion

Almindeligvis vil et observationssæt bestå af en række værdier $y_1, y_2, \ldots, y_k$ af interessevariablen samt tilhørende værdier af en række andre variable, de såkaldte kovariable, eller forklarende variable. For enhver observation, y_i, $i = 1, 2, \ldots, k$, af interessevariablen foreligger der et sæt værdier $\{x_{i1}, \ldots, x_{im}\}$ af de kovariable, der beskriver de omstændigheder, der er knyttet til observationen y_i.

Formålet med analysen ved en generaliseret lineær model er at vurdere relationen mellem

$$\eta_i = g(\mu_i) = g(\mathsf{E}\,[Y_i])$$

og sættet af kovariable $x_{i1}, x_{i2}, \ldots, x_{im}$

Den lineære (dimensionsreducerende) komponent af en generaliseret lineær model er karakteriseret ved at sættet $\eta_1, \ldots, \eta_k$ er beliggende i et lineært (evt.affint) underrum af $\mathbb{R}^k$, fastlagt af værdierne af de forklarende variable. Et sådant underrum kan imidlertid specificeres på en række forskellige måder.

I den foregående fremstilling har vi hovedsageligt betragtet en karakterisering af underrummet ved et sæt af basisvektorer, samlet i modelmatricen $\mathbf{X}$. Vi har endvidere betragtet hypoteser, formuleret på en parametrisk form, som f.eks.

$$\eta_i = \beta_1 + \beta_2 x_i$$

I dette og det følgende afsnit vil vi diskutere forskellige former for repræsentation af generaliserede lineære modeller og delhypoteser i sådanne modeller.

Vi vil betragte følgende repræsentationer:

datarepræsentation: En repræsentation, der afspejler hvorledes data er organiseret

repræsentation ved modelmatrix: En repræsentation ved et sæt af vektorer, der udspænder det lineære underrum

Parametrisk repræsentation: En repræsentation ved et regneudtryk (byggende på koordinater i et lineært underrum).

Repræsentation ved modelformel: En repræsentation ved symboler for de variable, der indgår i modellen

Ved behandlingen af de forklarende variable er det praktisk at sondre mellem:

kontinuerte kovariable som er variable, hvis værdier udtrykkes på en måleskala, en såkaldt intervalskala, så som længde, vægt, hastighed, koncentration, antal etc.
En model af formen

$$\eta_i = \sum_{j=1}^{m} \beta_j x_{ij} \ , \quad i = 1, 2, \ldots, k \ , \tag{2.8.1}$$

hvor der alene indgår kontinuerte kovariable x_{ij} kaldes ofte en lineær regressionsmodel for η.

kvalitative kovariable som er variable, hvis værdier udtrykkes på en nominal skala. Kvalitative variable udtrykker resultatet af en klassifikation. Sådanne kovariable betegnes ofte faktorvariable. En faktor svarer til en klassifikation af indeksmængden i disjunkte grupper. En faktor kan kun have et begrænset antal værdier, de såkaldte niveauer (eng. *levels*). For en faktor med r niveauer kan vi altid repræsentere niveauerne ved heltallene $1, 2, \ldots, r$, de formelle niveauer. I praksis vil der ofte være knyttet navne, eller evt en numerisk kodning til de formelle niveauer.

intercept Sædvanligvis vil man inddrage et såkaldt intercept led i modelleringen. Leddet bruges specielt til at beskrive den minimale model, nemlig en model svarende til fuldstændig homogenitet.

Som eksempler på kvalitative kovariable (klassifikationer) kan man forestille sig en klassifikation af trafikuheld efter uheldskvartal med niveauerne { januar, april, juli, oktober }, eller efter uheldstype med niveauerne {Eneuheld, Flerpartsuheld }, etc.

For en kontinuert kovariabel er der et naturligt afstandsmål, (nemlig differensen) mellem to værdier af den kovariable, og vi kan derfor relatere forskelle mellem y-værdierne til forskellene mellem de tilsvarende værdier af den kovariable x, sådan at vi kan forsøge at udtrykke de observerede forskelle i responsparameteren η ved ændringen, β, pr enhed hvormed x ændres, dvs en affin model $\eta = \alpha + \beta x$.

For kvalitative kovariable stiller sagen sig anderledes, idet der ikke er nogen fælles måleenhed for den kovariable. Der er ikke noget kvartal mellem kvartalerne { januar, april, juli, oktober }. I trafikuheldssammenhæng har det således ikke mening at opfatte juli kvartal som liggende to kvartaler senere end januar og et kvartal tidligere end oktober.

Ved klassifikation af et halvfabrikata efter leverandør { leverandør A, leverandør B, ... } er der ikke umiddelbart nogen naturlig måleskala, hvorpå man kan indordne leverandørerne.

Ved kvalitative kovariable kan vi derfor kun relatere forskelle mellem y-værdierne til de kvalitative forskelle mellem de tilhørende værdier af de kovariable. Bidraget svarende til en sådan klassifikation må derfor modelleres anderledes end bidraget svarende til en kontinuert kovariabel. Principielt kan vi kun beskrive forskelle, såkaldte kontraster mellem responserne svarende til forskellige niveauer af de betragtede kovariable.

For en ordens skyld gør vi opmærksom på at afgørelsen af, hvorvidt en kovariabel skal opfattes som kvantitativ (regressionsvariabel), eller som kvalitativ (faktorvariabel) ikke altid har et entydigt svar. Valget mellem de to fortolkninger styres af den aktuelle situation.

Datarepræsentation og parametrisk repræsentation

Datarepræsentationen er den måde, hvorpå data er organiseret, på papir, eller i et edb-program.

Den parametriske repræsentation af en hypotese angiver et regneudtryk, hvorved man kan bestemme den lineære prædiktor som funktion af et antal parametre. Den parametriske repræsentation knytter sig ofte til en bestemt datarepræsentation.

2.8.2 Kontinuerte kovariable

Betragt et observationssæt med de observerede værdier $y_1, y_2, \ldots, y_k$ af interessevariablen.

Udtrykket

$$\eta_i = \alpha + x_i\beta \,,$$

hvor $\alpha \in \mathbb{R}$ og $\beta \in \mathbb{R}$ er parametre, og $x_1, x_2, \ldots, x_k$ er værdierne af en kovariabel, angiver en parametrisk repræsentation af en hypotese.

Denne repræsentation af hypotesen knytter sig til en repræsentation af observationerne af Y og de tilsvarende værdier x_i af den forklarende variable som to k-dimensionale søjlevektorer.

Undertiden, når der ikke synes at herske tvivl om indices og om hvilke symboler, der angiver kendte kovariater, og hvilke, der angiver parametre, udelader man indeks og skriver blot

$$\eta = \alpha + x\beta$$

For en hypotese med m kontinuerte kovariable vil datarepræsentationen af interessevariablen Y være en k-dimensional søjlevektor $\mathbf{y}$, og datarepræsentationen af værdierne af de m kovariable vil tilsvarende være k-dimensionale søjlevektorer, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ med

$$\mathbf{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{kj} \end{pmatrix} \tag{2.8.2}$$

Søjlevektoren $\mathbf{x}_j$ (2.8.2) kaldes også modelmatricen svarende til den j'te kovariable.

Til en bestemt repræsentation af de forklarende variable hører der en parametrisk repræsentation. Hvis vi ændrer repræsentationen af de forklarende variable (f.eks. ændrer nulpunkt og/eller skala, eller danner nye forklarende variable som linearkombinationer af de gamle) får vi en anden parametrisk repræsentation, selv om hypotesen essentielt er uændret.

Parametrene i den parametriske repræsentation af en hypotese kan opfattes som "koordinater" i det lineære underrum, L, der svarer til den pågældende hypotese.

Enhver repræsentation af de forklarende variable ved en $k \times m$ matrix

$$\mathbf{X} = (\mathbf{x}_1|\mathbf{x}_2|\cdots|\mathbf{x}_m)$$

modsvares af en parametrisering af underrummet $L = \text{span}(\mathbf{X})$ sådan at ethvert $\eta \in L$ kan udtrykkes som en linearkombination

$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} \ , \ i = 1, 2, \ldots, k \ . \tag{2.8.3}$$

Matricen $\mathbf{X}$ kaldes modelmatricen svarende til de m kovariable.

Det kan forekomme, at man har valgt at overparametrisere hypotesen, svarende til at matricen $\mathbf{X}$ ikke har fuld rang m, dvs man har ikke brug for alle vektorerne i matricen $\mathbf{X}$ for at danne en basis for L. Vi skal senere (afsnit 2.9.8) vende tilbage til den eventuelle overbestemthed, der indføres ved en parametrisk repræsentation.

Når man i denne sammenhæng betegner en model som lineær, mener man således modeller af formen (2.8.3), der udtrykker prædiktoren $\eta = g(\mu)$ ved en lineær (egtl. affin) funktion af de kovariable. En model med de kovariable tryk$\approx \mathbf{x}_1$, ln(tryk)$\approx \mathbf{x}_2$ og kvadratet på værdien af trykket,$\approx \mathbf{x}_3$ vil således i denne sammenhæng også blive betegnet som en lineær model, idet den udtrykker prædiktoren η_i ved en linearkombination af de tilsvarende værdier af de kovariable, også selv om trykket indgår kvadratisk.

2.8.3 Intercept led

Sædvanligvis vil man i modelleringen inddrage et såkaldt intercept led, β_0, der specificerer den minimale model.

For en model med kontinuerte kovariable erstattes den parametriske repræsentation (2.8.3) da med udtrykket

$$\eta_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} \ , i = 1, 2, \ldots, k \ .$$

Den tilsvarende modelmatrix for den generaliserede lineære model er

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1m} \\ 1 & x_{21} & x_{22} & \ldots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{k1} & x_{k2} & \ldots & x_{km} \end{pmatrix} , \tag{2.8.4}$$

for den generaliserede lineære model, idet modelmatricen svarende til den minimale model er en søjlevektor bestående af k ettaller.

Den minimale model fremkommer netop for

$$\beta_1 = \beta_2 = \cdots = \beta_k = 0$$

2.8.4 Kvalitative kovariable, faktorvariable

I forbindelse med kvalitative kovariable, benytter de fleste programsystemer en speciel repræsentation af data.

Værdimængden for en faktor med r niveauer kan repræsenteres ved en liste $\{1, 2, \ldots, r\}$ med r elementer (de formelle niveauer). Ofte er der til mængden af formelle niveauer knyttet endnu en liste, $\{l_1, l_2, \ldots, l_r\}$, af "labels" for de r niveauer.

Nogle statistiksprog indeholder yderligere et objekt "en ordnet faktor". Et sådant objekt adskiller sig kun fra det sædvanlige faktorbegreb ved at mængden af faktorniveauer opfattes som en ordnet mængde, dvs $l_1 < l_2 < \cdots < l_r$. Ordningen har betydning ved automatiske valg af kontraster og ved analyser af træ-strukturer.

Når interessevariablen Y er repræsenteret som en k-dimensional søjlevektor, er den tilsvarende datarepræsentation af en faktorvariabel et ordnet sæt af k værdier i en rækkefølge svarende til observationsvektoren. Såfremt faktoren har r niveauer, vil datarepræsentationen således være en k-dimensional vektor, hvor værdien af det enkelte element er et af de formelle niveauer $\{1, 2, \ldots, r\}$, eller af de aktuelle niveauer $\{l_1, l_2, \ldots, l_r\}$.

Hvis $a(i)$ angiver de formelle faktorniveauer for faktoren A svarende til den i'te observation, $i = 1, 2, \ldots, k$, vil datarepræsentationen af faktorvariabel A være tallene $a(1), a(2), \ldots, a(k)$ organiseret som en søjlevektor.

Den parametriske repræsentation af en model med én kvalitativ kovariabel er af formen

$$\eta_i = \alpha_{a(i)}\,, \quad i = 1, 2, \ldots, k\,, \tag{2.8.5}$$

eller eventuelt med interceptleddet β_0 svarende til den minimale model:

$$\eta_i = \beta_0 + \alpha_{a(i)}\,, \quad i = 1, 2, \ldots, k\,, \tag{2.8.6}$$

hvor parametrene α_1, α_p i fremstillingen (2.8.6) må underkastes nogle bånd for at tilgodese overparametriseringen (se afsnit 2.9.3 og 2.9.8)

Hvis der er flere gentagelser for hvert faktorniveau kan man opstille data på tabelform. Således kan man - i situationen med en enkelt faktor - opstille data i et skema med r rækker (svarende til de r faktorniveauer) og på hver af de r linier anføre de observerede værdier svarende til dette faktorniveau. Den i'te observation bliver da placeret i p'te række, hvor $p = a(i)$ angiver det formelle niveau af faktoren for den i'te observation. Observationerne indiceres da som $y_{p,j}$, hvor $j = 1, 2, \ldots, n_p$ angiver gentagelsesnummeret ved den pågældende faktorværdi.

Når data repræsenteres på denne form er den tilsvarende parametriske model på formen

$$\eta_p = \alpha_p \ , p = 1, 2, \ldots, r$$

eller, hvis man inddrager et interceptled:

$$\eta_p = \beta_0 + \alpha_p \ , p = 1, 2, \ldots, r$$

Eksempel 2.8.1 ***Repræsentation af værdier af faktorvariabel svarende til et enkelt klassifikationskriterium***

I et forsøg til bestemmelse af effekten af de tre forskellige fødeadditiver {spinat, jordnød, kryptonit} foretoges eksperimenter på fem forsøgsenheder.

De fem forsøgsenheder blev nummereret ved tallene fra 1 til 5, hvor enhed nr. 1 fik spinat, nr 2 og 3 fik jordnød og nr. 4 og 5 fik kryptonit. De tilsvarende værdier af interessevariablen (et mål for styrke) var 25, 5, 7 , 14, 17.

Man har således en faktorvariabel Additiv med de tre formelle niveauer 1,2 og 3, hvor f.eks. niveau 1 angiver spinattilstætning, niveau 2 angiver jordnøddetilsætning og niveau 3 angiver kryptonittilsætning.

Udtrykt ved de formelle niveauer er datarepræsentationen for Additiv

$$\begin{pmatrix} 1 \\ 2 \\ 2 \\ 3 \\ 3 \end{pmatrix}$$

Hvis eksempelvis de tre niveauer er kodet med symbolerne "s', "j" og "k", bliver datarepræsentationen

$$\begin{pmatrix} \texttt{s} \\ \texttt{j} \\ \texttt{j} \\ \texttt{k} \\ \texttt{k} \end{pmatrix}$$

Hvis faktorniveauerne er ordnede, f.eks. i rækkefølgen $j < s < k$ ændrer det ikke på denne datarepræsentation.

Hvis faktorniveauerne har de tilknyttede labels {spinat,jordnød,kryptonit}, vil datarepræsentationen ikke ændres, men en udskrivning af faktorvariablen Additiv vil resultere i udskriften

$$\begin{pmatrix} \texttt{spinat} \\ \texttt{jordnød} \\ \texttt{jordnød} \\ \texttt{kryptonit} \\ \texttt{kryptonit} \end{pmatrix}$$

Tabelformen af data er:

Additiv	Styrke
spinat	25
jordnød	5, 7
Kryptonit	14,17

□

Hvis værdierne af en variabel er klart kvalitative (nominelle), dvs. f.eks. karaktervariable, vil de fleste programsystemer behandle den pågældende variabel som en faktorvariabel.

I modsat fald vil den som hovedregel blive behandlet som en kontinuert kovariabel (intervalskala), med mindre den eksplicit er erklæret som en faktorvariabel

I datavinduet i SAS® Insight proceduren er der en rubrik over navnet på den variable, der markerer, hvorvidt den pågældende variable fortolkes som målt på intervalskala eller på nominal skala. Ved at klikke i rubrikken kan man ændre denne fortolkning. I de fleste SAS®-procedurer kan man ændre en variabel fra intervalskala til nominal skala ved sætningen CLASS *navn* ;

Når der optræder flere faktorvariable vil man ofte opstille data på tabelform.

Således vil man i en situation med to inddelingskriterier opstille observationerne i et tosidet skema (tabel), organiseret i rækker (det ene inddelingskriterium), og søjler (det andet inddelingskriterium). Tabellen er netop indiceret ved de formelle faktorniveauer. Sammenhængen mellem observationer og værdier af de kovariable er her beskrevet ved observationens position i tabellen.

Antag således at de k observationer $y_1, \ldots, y_k$ organiseres i et tosidet skema med r rækker (svarende til faktoren A) og s søjler (svarende til faktoren B). Den i'te observation, $i = 1, 2, \ldots, k$ bliver da placeret i p'te række og q'te søjle, hvor $p = a(i)$ angiver det formelle niveau af faktor A for den pågældende observation, og $q = b(i)$ tilsvarende angiver det formelle niveau af faktor B.

Den parametriske repræsentation på tabelform fremkommer da af den sædvanlige parametriske repræsentation ved at indicere observationer og parametre ved deres position i skemaet. For en tosidet inddeling med én observation i hver celle kan man eksempelvis have en model

$$\begin{aligned} \mathsf{E}\,[Y_{p,q}] &= \mu_{p,q} \\ \eta_{p,q} &= g(\mu_{p,q}) \\ \eta_{p,q} &= \gamma_{p,q}\,; \qquad p = 1, 2, \ldots, r\,;\ \ q = 1, 2, \ldots, s\,; \end{aligned} \tag{2.8.7}$$

i analogi med eksempel 2.4.2.

Ofte vælger man at skrive (2.8.7) på formen

$$\eta_{p,q} = \alpha_p + \beta_q + \gamma^*_{p,q}\,, \tag{2.8.8}$$

eller med et interceptled, β_0:

$$\eta_{p,q} = \beta_0 + \alpha_p + \beta_q + \gamma^*_{p,q}\,, \tag{2.8.9}$$

hvor parametrene α_p, β_q og $\gamma^*_{p,q}$ må underkastes nogle bånd for at tilgodese overparametriseringen.

I en sådan fremstilling kaldes parameteren $\gamma^*_{p,q}$ for vekselvirkningsleddet. Vi vil senere (afsnit 2.12) nærmere diskutere de såkaldte vekselvirkninger.

I nogle fremstillinger skrives den parametriske model (2.8.8) på formen

$$\eta_{p,q} = \alpha_p + \beta_q + (\alpha\beta)_{p,q} \,, \tag{2.8.10}$$

dvs man bruger det sammensatte symbol $(\alpha\beta)_{p,q}$ i stedet for $\gamma^*_{p,q}$. Specielt hvis man har mange faktorvariable, kan det være formålstjenligt på denne måde at indikere, hvilke faktorer den pågældende vekselvirkning refererer til.

Det skal imidlertid understreges, at der ikke er forskel på modellerne, der er repræsenteret ved (2.8.7), (2.8.8) og (2.8.10). Det er kun parameterfremstillingerne, der er forskellige.

Eksempel 2.8.2 ***Kvartalsvise uheld for motordrevne køretøjer***

På basis af politiets indberetninger opgøres for hvert kvartal antallet af registrerede uheld med motordrevne køretøjer. For hvert enkelt uheld registreres en række oplysninger, blandt andet klassificeres uheldet i en af 10 hovedsituationer, endvidere efter uheldstidspunkt, samt efter hvorvidt en af parterne var spirituspåvirket.

Tabel 2.7 viser de kvartalsvise uheldstal for motorkøretøjer for uheldskategorierne "single" og "møde" i dagslys for ikke-spirituspåvirkede førere for årene 1987 - 1990.

Vi har valgt at klassificere de 24 observationer i $3 \times 4 \times 2$ kategorier, nemlig år, kvartal og uheldsituation.

Vi siger, at vi har de tre faktorer (eller inddelingskriterier) år, kvartal og uheldssituation med henholdsvis 3, 4 og 2 niveauer.

Tabellen nedenfor viser for hver af de tre faktorer det tilhørende antal niveauer, samt betegnelsen for hvert niveau.

Tabel 2.7. Kvartalsvise uheldsantal med motorkøretøjer, opdelt på uheldskategorier
(ikke-spiritusuheld, dagslys, kun single- og mødeuheld)

År	Kvartal	Hovedsituation	
		single	møde
1987	1	127	128
	2	149	95
	3	157	100
	4	110	75
	Ialt	543	398
1988	1	107	94
	2	181	85
	3	145	119
	4	111	71
	Ialt	544	369
1989	1	97	82
	2	158	81
	3	133	98
	4	100	72
	Ialt	488	333

Faktor	antal niveauer	niveau nummer	værdi
år	3	1	1987
		2	1988
		3	1989
Kvartal	4	1	JAN
		2	APR
		3	JUL
		4	OKT
Situation	2	1	single
		2	møde

Ved at krydsklassificere data har vi således foretaget en afbildning af indeksmængden $I = \{1, 2, \ldots, 24\}$ af de 24 observationer ind i mængden

$$\{1, 2, 3\} \times \{1, 2, 3, 4\} \times \{1, 2\}$$

således at en observation kan repræsenteres som $Y_{i,j,k}$; $i = 1,2,3\,; j = 1,2,3,4\,; k = 1,2$. □

I afsnit 2.9 vil vi yderligere diskutere modeller med kvalitative kovariable

2.8.5 Parametrisk repræsentation af blandede led

Betragt en model med den parametriske fremstilling

$$\eta_i = \alpha_{a(i)} + \beta x_i\,, \quad i = 1,2,\ldots,k \tag{2.8.11}$$

Grafen for denne model i et (x,η)-koordinatsystem er nogle parallelle rette linier, en linie for hvert af niveauerne af faktoren A.

Linierne har hældningen β, og linien svarende til niveau p af faktor A har afskæringen α_p på η-aksen.

De såkaldte "blandede led" i den parametriske fremstilling fremkommer ved en udvidelse af denne model til en model af formen

$$\eta_i = \alpha_{a(i)} + \gamma_{a(i)} x_i\,,$$

der angiver, at linierne svarende til de forskellige niveauer af faktoren A ikke nødvendigvis er parallelle, men at hældningerne kan være forskellige.

2.9 Modelmatrix, kontraster

I det følgende vil vi give en formel beskrivelse af konstruktionen af modelmatricen svarende til en generaliseret lineær model.

Største delen af afsnittet drejer sig om opstilling af modelmatricen for én eller flere kvalitative kovariable, da disse situationer kræver lidt ekstra omhu.

Ved brug af statistiske programsystemer som S-plus, SAS® mv. behøver man i almindelighed ikke selv opstille modelmatricerne. Ved formulering af hierarkiske hypotesekæder og test af sådanne hypoteser er det imidlertid en fordel at have en fornemmelse af relationen mellem parametriseringen og de underrum, der udspændes af de forskellige hypoteser.

2.9.1 Modelmatrix for kontinuerte kovariable

Som anført i afsnit 2.8.2 fremkommer modelmatricen svarende til et sæt af kontinuerte kovariable ved at danne matricen, hvis enkelte søjler er modelmatricen (søjlevektoren) svarende til datarepræsentationen af de enkelte kovariable.

Såfremt man ønsker et intercept led (svarende til den minimale model) tilføjer man yderligere en søjle med ettaller til modelmatricen.

2.9.2 Incidensmatrix for faktorvariabel

Vi indfører en lidt abstrakt definition

Definition 2.9.1 ***Klassifikation***

Betragt en indiceret mængde $\{y_i\}_{i\in I}$ af observationer.

En klassifikation af indeksmængden er en afbildning $a : I \to A$ af indeksmængden $I = \{1, 2 \ldots, k\}$ ind i en endelig mængde, de formelle faktorniveauer $A = \{1, 2, \ldots, r\}$. □

En sådan klassifikation betegnes undertiden også en faktor.

Klassifikationen bestemmer en disjunkt opdeling af indeksmængden I i r grupper

$$I = \bigcup_{f\in A} a^{-1}(f) = \bigcup_{f\in A} \{i \in I | a(i) = f\}$$

Afbildningen a bestemmer en lineær afbildning $a^{\square} : \mathbb{R}^A \to \mathbb{R}^I$ ved

$$a^{\square} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_r \end{pmatrix} = \begin{pmatrix} \beta_{a(1)} \\ \beta_{a(2)} \\ \vdots \\ \beta_{a(k)} \end{pmatrix}$$

Afbildningen fører det r-dimensionale talsæt $(\beta_1, \ldots, \beta_r)^T$ over i et k-dimensionalt talsæt, hvor β_f optræder på de pladser i I, der svarer til faktorniveauet f, $f = 1, 2, \ldots, r$.

Billedmængden for afbildningen er

$$L_A = a^{\square}(\mathbb{R}^A) = \{(\beta_{a(i)})_{i\in I} \in \mathbb{R}^I | (\beta_f)_{f\in A} \in \mathbb{R}^A\}$$

altså et r-dimensionalt underrum $L_A \subset \mathbb{R}^k$.

Lad nu

$$\delta_{p,q} = \begin{cases} 1 & \text{for } p = q \\ 0 & \text{ellers} \end{cases} \tag{2.9.1}$$

være den sædvanlige Kronecker's delta.

Idet

$$\eta_{a(i)} = \sum_{f\in A} \delta_{a(i),f}\beta_f$$

for $i \in I$ finder vi ved valg af de sædvanlige baser i $\mathbb{R}^A$ og $\mathbb{R}^I$, at matricen for $a^{\square}$ er

$$\mathbf{U}_A = (\delta_{a(i),j})_{(i,j)\in I\times A} \tag{2.9.2}$$

Søjlevektorerne $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$ i $\mathbf{U}_A$ udspænder underrummet $L_A \subset \mathbb{R}^I$ svarende til faktoren A.

En søjle $\mathbf{u}_j$ kaldes incidensvektoren eller indikatorvektoren for det j'te niveau

Matricen $\mathbf{U}_A$ kaldes incidensmatricen svarende til faktoren A.

Incidensmatricen er en $n \times r$ dimensional matrix med nuller og ettaller, hvor det (i,j)'te element er bestemt ved

$$u_{ij} = \begin{cases} 1 & \text{hvis } i\text{'te enhed er på niveau } j \\ 0 & \text{ellers} \end{cases}$$

I hver række er der netop ét ettal. Elementet i i'te række og j'te søjle er et ettal, såfremt $a(i) = j$, dvs. såfremt den i'te observation svarer til klassen j.

Vi bemærker, at incidensvektorerne $\mathbf{u}_j$ er underkastet det lineære bånd

$$\mathbf{u}_1 + \mathbf{u}_2 + \cdots + \mathbf{u}_r \equiv \mathbf{1}$$

uanset den aktuelle allokering af faktorniveauer til forsøgsenheder. Båndet udtrykker at hver forsøgsenhed skal være tilknyttet netop ét af de r faktorniveauer.

Vi har således indført en række "dummy" variable, $\mathbf{u}_j$, $j = 1, 2 \ldots, r$, én for hvert niveau af den pågældende faktor, til at beskrive hvilke observationer, der repræsenterer det pågældende niveau af denne faktor.

Såfremt der er foretaget mindst én observation for hver klasseværdi, vil afbildningen a være surjektiv og $\dim(L_A) = r$. Matricen $\mathbf{U}_A$ har da fuld rang, r.

Incidensmatricen $\mathbf{U}_A$ er matrix for indlejringen af underrummet $\mathbb{R}^A$ ind i $\mathbb{R}^I$. Såfremt matricen har fuld rang finder man, at den surjektive ortogonalprojektion $p_U : \mathbb{R}^I \to \mathbb{R}^A$ (med hensyn til det sædvanlige indre produkt på $\mathbb{R}^I$) har matricen

$$\mathbf{P}_A = (\mathbf{U}_A{}^T\mathbf{U}_A)^{-1}\mathbf{U}_A{}^T$$

og ortogonalprojektionen på $\mathbb{R}^I$ har matricen (hat-matricen)

$$\mathbf{H}_A = \mathbf{U}_A\mathbf{P}_A = \mathbf{U}_A(\mathbf{U}_A{}^T\mathbf{U}_A)^{-1}\mathbf{U}_A{}^T$$

Matricen $\mathbf{U}_A{}^T\mathbf{U}_A$ er en $r \times r$ diagonalmatrix, hvis f'te diagonalement er antallet af observationer svarende til faktorniveauet f, $f = 1, 2, \ldots, r$.

En klassifikation, der kun antager én værdi, kaldes den konstante faktor. Den konstante faktor svarer således til en model med fuld homogenitet (den minimale model). Det tilhørende endimensionale underrum betegnes med L_M. Undertiden bruger man dog symbolet O for den konstante faktor, og L_O for det tilsvarende underrum.

Underrummet svarende til den identiske afbildning af I på sig selv betegnes med $L_I = \mathbb{R}^I$.

For en faktorvariabel A med r værdier $\beta_1, \ldots, \beta_r$ kan vi udtrykke bidraget $\eta_{a(i)}$ svarende til den i'te observation som

$$\eta_{a(i)} = \sum_{j=1}^{r} u_{ij}\beta_j = (\mathbf{u}_i^*)^T\beta$$

hvor $(\mathbf{u}_i^*)$ er vektoren bestående af i'te række i $\mathbf{U}_A$.

Eksempel 2.9.1 *Incidensmatrix svarende til et enkelt klassifikationskriterium*

Vi betragter atter undersøgelsen af fødeadditiver fra eksempel 2.8.1

Konstruktionen af modelmatricen fremgår af nedenstående tabel:

Forsøgsenhed	Additiv	niveau	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$
1	spinat	1	1	0	0
2	jordnød	2	0	1	0
3	jordnød	2	0	1	0
4	kryptonit	3	0	0	1
5	kryptonit	3	0	0	1

□

2.9.3 Parametrisering af faktormodel ved kontraster

En faktor med r niveauer inducerer således et lineært underrum af dimension r i $\mathbb{R}^k$, udspændt af søjlerne i incidensmatricen $\mathbf{U}_A$.

Imidlertid vil det ofte være relevant at indføre et interceptled, β_0 og parametrisere en faktormodel i analogi med (2.8.6) dvs

$$\eta_{a(i)} = \beta_0 + \sum_{j=1}^{r} u_{ij}\beta_j \tag{2.9.3}$$

hvor β_0 angiver interceptleddet. Værdien af β_0 kan også fortolkes som et referenceniveau.

Repræsentationen (2.9.3) giver imidlertid ikke mulighed for at identificere værdierne af parametrene β_j, da matricen

$$\mathbf{X}_A = \{\mathbf{1}, \mathbf{u}_1, \ldots, \mathbf{u}_r\}$$

svarende til repræsentationen (2.9.3) kun har rangen r. Én af vektorerne $\mathbf{u}_1, \ldots, \mathbf{u}_r$ kan udtrykkes som en linearkombination af de øvrige og vektoren $\mathbf{1}$. Vi ved jo netop, at rækkesummen af $\mathbf{U}_A$ er 1.

Man må derfor indføre et lineært bånd mellem parametrene $\beta_1, \ldots, \beta_r$, eller reparametrisere problemet.

Såfremt man ønsker at operere med en modelmatrix af fuld rang, kan man vælge at reparametrisere problemet sådan at man opnår en $k \times r$ dimensional matrix

$$\mathbf{X}_A = \{\mathbf{1}, \mathbf{c}_1, \ldots, \mathbf{c}_{r-1}\}$$

af fuld rang, der udspænder L_A.

Vektorerne $\mathbf{c}_j$ definerer en række såkaldte estimable kontraster imellem de r størrelser $\beta_1, \ldots, \beta_r$

Der findes en række forskellige muligheder for at definere kontraster mellem r niveauer af en faktor. I nedenstående eksempel skal vi illustrere nogle af disse muligheder i tilfældet $r = 4$.

Eksempel 2.9.2 *Kontraster mellem* $\mathbf{r = 4}$ *niveauer*

Helmert-transformation
Parametrisering ved den såkaldte Helmert-transformation svarer til modellen

$$\boldsymbol{\eta} = \mathbf{X}_H \begin{pmatrix} \beta_0 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix}$$

hvor modelmatricen $\mathbf{X}_H$ er

$$\mathbf{X}_H = \begin{pmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 0 & 2 & -1 \\ 1 & 0 & 0 & 3 \end{pmatrix},$$

Parametriseringen svarer til at første kontrast angiver forskellen $\beta_2 - \beta_1$ mellem værdien svarende til niveau 2 og niveau 1; anden kontrast angiver forskellen $\beta_3 - (\beta_1 + \beta_2)/2$ mellem værdien svarende til niveau 3 og gennemsnittet mellem niveau 1 og niveau 2. Den j'te kontrast angiver forskellen $\beta_{j+1} - (\beta_1 + \cdots \beta_j)/j$ mellem niveau $j+1$ og gennemsnittet af niveauerne 1 til j.

Helmert-transformationen sikrer, at de enkelte søjler $\mathbf{c}_j$ er indbyrdes ortogonale, og endvidere ortogonale på vektoren $\mathbf{1}$.

Sum-kodning
En anden mulighed er den såkaldte "sum"-kodning

$$\boldsymbol{\eta} = \mathbf{X}_S \begin{pmatrix} \beta_0 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix}$$

hvor modelmatricen $\mathbf{X}_S$ er

$$\mathbf{X}_H = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 \end{pmatrix},$$

svarende til at første kontrast angiver forskellen mellem β_1 og β_r, anden kontrast angiver forskellen mellem β_2 og β_r, etc. Også disse kontraster er indbyrdes ortogonale og ortogonale på $\mathbf{1}$.

Treatment-kodning
Ofte benytter man en kodning svarende til repræsentationen

$$\boldsymbol{\eta} = \mathbf{X}_T \begin{pmatrix} \beta_0 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix}$$

hvor modelmatricen $\mathbf{X}_T$ er

$$\mathbf{X}_T = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix},$$

Kodningen kaldes undertiden "treatment"-kodning.

Reparametriseringen svarer til at benytte det første niveau β_1 som referenceniveau. Parametrene $\gamma_1, \ldots, \gamma_{r-1}$ svarende til denne kodning er ikke kontraster, da søjlesummerne ikke er nul, og søjlerne derfor ikke er ortogonale på vektoren $\mathbf{1}$. □

2.9.4 Modelmatrix svarende til blandede led

I afsnit 2.8.5 diskuterede vi den parametriske repræsentation af modeller med såkaldt "blandede led". Vi vil her beskrive konstruktionen af modelmatricen svarende til sådanne led.

Lad faktoren A have r værdier og den $k \times r$-dimensionale incidensmatrix

$$\mathbf{U} = (\mathbf{u}_1|\mathbf{u}_2|\ldots|\mathbf{u}_r),$$

og lad den kvantitative kovariate have modelvektoren $\mathbf{x}$.

Modelmatricen $\mathbf{W}$ svarende til de blandede led for A og x er da den $k \times r$-dimensionale matrix

$$\mathbf{W} = (\mathbf{w}_1|\mathbf{w}_2|\ldots|\mathbf{w}_r)$$

hvor den n-dimensionale søjlevektor $\mathbf{w}_i$ fremkommer ved elementvis multiplikation af $\mathbf{x}$ og $\mathbf{u}_i$.

Eksempel 2.9.3 *Modelmatrix svarende til et enkelt klassifikationskriterium*

Betragt atter den konstruerede situation i eksempel 2.9.1 og antag, at der blev anvendt forskellige mængder x_i af fødeadditivet i de fem eksperimenter.

Modelmatricen $\mathbf{W}$ svarende til de blandede led er angivet i nedenstående tabel:

Forsøgsenhed	Additiv	niveau	mængde $\mathbf{x}$	$\mathbf{w}_1$	$\mathbf{w}_2$	$\mathbf{w}_3$
1	spinat	1	1	1	0	0
2	jordnød	2	3	0	3	0
3	jordnød	2	5	0	5	0
4	kryptonit	3	7	0	0	7
5	kryptonit	3	9	0	0	9

Det anvendte design tjener til at illustrere konstruktion af modelmatricen. Det er næppe egnet til at give information om sammenhængen mellem indtaget mængde og responsvariabel.
Vi bemærker at der gælder

$$\mathbf{w}_1 + \mathbf{w}_2 + \cdots + \mathbf{w}_r = \mathbf{x}$$

uanset allokeringen af niveauer til forsøgsenheder. □

2.9.5 Incidensmatrix svarende til to klassifikationskriterier

Vi vil nu vende tilbage til den betragtningsmåde, der blev benyttet i afsnit 2.9.2 og beskrive konstruktionen af incidensmatricen svarende til to klasifikationskriterier.

Betragt to klassifikationer $a : I \to A = \{1, 2 \ldots, r\}$ og $b : I \to B = \{1, 2 \ldots, s\}$ med incidensmatricerne

$$\mathbf{U} = (\mathbf{u}_1 | \mathbf{u}_2 | \cdots \mathbf{u}_r)$$

og

$$\mathbf{V} = (\mathbf{v}_1 | \mathbf{v}_2 | \cdots \mathbf{v}_s)$$

De to klassifikationer bestemmer den nye klassifikation $a.b : I \to A \times B$ ved

$$(a.b)(i) = (a(i), b(i)) \in A \times B$$

Matricen svarende til afbildningen

$$(a.b)^{\square} : \mathbb{R}^{A \times B} \to \mathbb{R}^{I}$$

er en $k \times (r \times s)$ dimensional matrix, $\mathbf{W}$, hvis søjler $\mathbf{w}_{a.b(i)}$ fremkommer ved elementvis multiplikation af $\mathbf{u}_{a(i)}$ og $\mathbf{v}_{b(i)}$

I lighed med incidensmatricen for én faktor gælder at rækkesummen i $\mathbf{W}$ er 1.

Endvidere gælder

$$\sum_{\nu : b(\nu) = j} \mathbf{w}_{a.b(\nu)} = \mathbf{v}_j, \; j = 1, 2 \ldots, s$$

og

$$\sum_{\nu : a(\nu) = i} \mathbf{w}_{a.b(\nu)} = \mathbf{u}_i, \; i = 1, 2 \ldots, r$$

For en tosidet inddeling efter to kriterier A og B med henholdsvis r og s niveauer er den parametriske fremstilling af formen :

$$\eta_i = \gamma_{a(i), b(i)}$$

med de $r \times s$ værdier

$$\gamma_{i,j} \; ; \; i = 1, 2, \ldots, r \; ; \; j = 1, 2, \ldots, s$$

Parametriseringen afspejler blot, at vi har afbildet den endimensionale indeksmængde I ind i produktmængden $A \times B$. Hvis der er flere observationer for hver kombination (p, q) af de to klassifikationer, er afbildningen dimensionsreducerende; ellers er den bare udtryk for en reparametrisering.

Definition 2.9.2 *Additiv model ved to klassifikationer*

Betragt to klassifikationer $a : I \to A = \{1, 2 \ldots, r\}$ og $b : I \to B = \{1, 2 \ldots, s\}$

Såfremt den parametriske fremstilling er af formen

$$\eta_i = \alpha_{a(i)} + \beta_{b(i)} \tag{2.9.4}$$

siges modellen at være additiv i de to kriterier A og B.

Såfremt indeksmængden $I = A \times B$, dvs. såfremt der kun er ét element svarende til hver kombination (p, q) af de to klassifikationer, udtrykkes modellen (2.9.4) ofte som

$$\eta_{p,q} = \alpha_p + \beta_q \tag{2.9.5}$$

□

Bemærkning 1 *En additiv model svarer til en sum af underrummene for de enkelte klassifikationer*

Lad incidensmatricerne for de to klassifikationer være

$$\mathbf{U}_A = (\mathbf{u}_1 | \mathbf{u}_2 | \cdots \mathbf{u}_r)$$

og

$$\mathbf{V}_B = (\mathbf{v}_1 | \mathbf{v}_2 | \cdots \mathbf{v}_s) \ ,$$

og antag, at incidensmatricerne har fuld rang.

Som nævnt i afsnit 2.9.2, kan incidensmatricen $\mathbf{U}_A$ opfattes som matricen for en lineær afbildning $\mathbb{R}^A \to \mathbb{R}^I$. Billedet $L_A \subset \mathbb{R}^I$ af denne afbildning er mængden af afbildninger $I \to \mathbb{R}$, der er konstante på de ækvivalensklasser i I, der er induceret af klassifikationen A. Tilsvarende er billedet $L_B \subset \mathbb{R}^I$ mængden af afbildninger $I \to \mathbb{R}$, der er konstante på ækvivalensklasserne induceret af klassifikationen B

Projektionerne p_A og p_B på $\mathbb{R}^A$ og $\mathbb{R}^B$ med matricerne

$$\mathbf{P}_A = (\mathbf{U}_A{}^T \mathbf{U}_A)^{-1} \mathbf{U}_A{}^T$$

og

$$\mathbf{P}_B = (\mathbf{V}_B{}^T \mathbf{V}_B)^{-1} \mathbf{V}_B{}^T$$

tilordner netop enhver af disse afbildninger til sin ækvivalensklasse (indiceret ved $\mathbb{R}^A$ og $\mathbb{R}^B$), og matricerne $\mathbf{H}_A = \mathbf{U}_A\mathbf{P}_A$ og $\mathbf{H}_B = \mathbf{V}_B\mathbf{P}_B$ er netop hatmatricerne for de tilsvarende projektioner ned på $L_A \subset \mathbb{R}^I$ og $L_B \subset \mathbb{R}^I$.

Rummet $L_A + L_B \subset \mathbb{R}^I$ er således netop mængden af $\boldsymbol{\eta} \in \mathbb{R}^I$, der kan udtrykkes på formen (2.9.4).

Incidensmatricen $\mathbf{X}$ svarende til modellen (2.9.4) er da den $k \times (r+s)$ dimensionale matrix, $\mathbf{X} = (\mathbf{U}_A|\mathbf{V}_B)$, der fremkommer ved at stille matricerne $\mathbf{U}_A$ og $\mathbf{V}_B$ ved siden af hinanden. □

Definition 2.9.3 *Ortogonale klassifikationer af indeksmængden*

Betragt to klassifikationer $a : I \to A = \{1, 2 \ldots, r\}$ og $b : I \to B = \{1, 2 \ldots, s\}$ med incidensmatricerne

$$\mathbf{U}_A = (\mathbf{u}_1|\mathbf{u}_2|\cdots\mathbf{u}_r)$$

og

$$\mathbf{V}_B = (\mathbf{v}_1|\mathbf{v}_2|\cdots\mathbf{v}_s) ,$$

og antag at incidensmatricerne har fuld rang.

Klassifikationerne siges at være ortogonale, såfremt der gælder

$$\mathbf{H}_A\mathbf{H}_B = \mathbf{H}_B\mathbf{H}_A ,$$

hvor $\mathbf{H}_A$ og $\mathbf{H}_B$ er projektionsmatricerne

$$\mathbf{H}_A = \mathbf{U}_A(\mathbf{U}_A{}^T\mathbf{U}_A)^{-1}\mathbf{U}_A{}^T$$

og

$$\mathbf{H}_B = \mathbf{V}_B(\mathbf{V}_B{}^T\mathbf{V}_B)^{-1}\mathbf{V}_B{}^T$$

□

Bemærkning 1 *Ortogonalitet af klassifikationer svarer til geometrisk ortogonalitet af de tilsvarende underrum*

Lad L_A og L_B være som i bemærkning 1 til definition 2.9.2.

Da projektionerne h_A og h_B på $\mathbb{R}^I$ og $\mathbb{R}^I$ givet ved matricerne $\mathbf{H}_A$ og $\mathbf{H}_B$ netop er projektionen ned på L_A og L_B, ser vi at da definitionen udtrykker, at disse projektioner kommuterer, svarer det til at sige, at underrummene L_A og L_B er geometrisk ortogonale. □

Bemærkning 2 *Ortogonalitet af klassifikationer er ensbetydende med proportionalitet mellem antallet af observationer i cellerne*

Matricen $\mathbf{U}_A{}^T\mathbf{U}_A$ er er en diagonalmatrix, hvis i'te element netop er antallet $n_p(A)$ af elementer i indeksmængden I, der svarer til niveauet p for klassifikationen A, $p = 1, 2 \ldots, r$.

Vi finder således specielt, at to faktorer er ortogonale, hvis og kun hvis der gælder

$$(\mathbf{W}^T\mathbf{W})_{p,q} = (\mathbf{U}_A{}^T\mathbf{U}_A)_{p,p}(\mathbf{V}_B{}^T\mathbf{V}_B)_{q,q} \ , \tag{2.9.6}$$

hvor matricen $\mathbf{W}$ er den $k \times (r \times s)$ dimensionale incidensmatrix svarende til produktklassifikationen $a.b : I \to A \times B$.

To faktorer er således ortogonale, hvis celleantallene $n_{p,q}(A \times B) = n_p(A)n_q(B)$. □

2.9.6 Klassifikationer med hierarkisk ordnet indeksmængde

En hierarkisk klassifikation optræder når niveauerne for én klassifikation, f.eks. B, er underordnet niveauerne for en anden klassifikation, f.eks. A, dvs. svarende til ethvert niveau, p, af A er der en klassifikation, B_p, af mængden

$$a^{-1}(p) = \{i \in I | a(i) = p\}$$

Der er således en klassifikation svarende til hvert niveau af A, men klassifikationen har kun mening indenfor dette niveau af A. Vi siger at klassifikationen B er underordnet klassifikationen A. (På engelsk siges B at være "nested" indenfor A). I praksis optræder en sådan hierarkisk klassifikation eksempelvis i situationer, hvor der optræder gentagelser for hvert niveau af faktor A. Indeks for gentagelsen vil være underordnet faktoren A. Gentagelse nummer ν på niveau p i A har intet til fælles med gentagelse nummer ν på niveau q i A, selv om vi har valgt at kode dem med det samme nummer.

2.9.7 Partiel ordning af klassifikationer

Vi bemærker, at der ikke er grund til at opfatte klassifikationer som forskellige, hvis de svarer til den samme inddeling af indeksmængden. Vi vil derfor

sige, at to klassifikationer er ækvivalente, hvis de inducerer den samme opdeling af indeksmængden. Mængden $\mathfrak{F}(I)$ af ækvivalensklasser er endelig. Det er nok, at betragte repræsentanter for ækvivalensklasser af klassifikationer af indeksmængden I.

En klassifikation A siges at være finere end en klassifikation B, hvis opdelingen af I svarende til A er en underopdeling af opdelingen svarende til B. Lader vi $\phi_B : I \to B$ og $\phi_A : I \to A$ angive de to klassifikationer har vi, at A er finere end B hvis og kun hvis der findes en afbildning $\phi_{B|A} : A \to B$ så

$$\phi_B = \phi_{B|A} \circ \phi_A \ .$$

Hvis klassifikationen A er finere end klassifikationen B skriver vi $A \leq B$.

Med ordningen $\leq$ kan $\mathfrak{F}(I)$ organiseres som en partielt ordnet mængde (engelsk: *poset*).

Den trivielle klassifikation O svarer til den konstante afbildning $\phi_O : I \to O$, hvor O angiver en vilkårlig mængde med kun ét element. Ét-faktoren I svarer til identiteten på indeksmængden I, $\phi_I : I \to I$.

Såfremt der foreligger to klassifikationer A og B af indeksmængden, er der to mulige faktoriseringer af afbildningen $\phi_{O|A\times B}$ fra klassifikationen

$$A.B : I \to A \times B \tag{2.9.7}$$

til en mængde O med kun ét element, nemlig en faktorisering

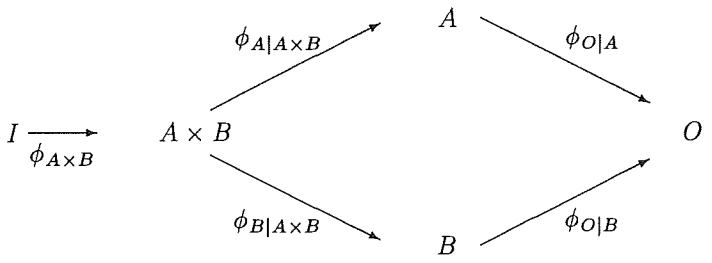
$$\phi_{O|A\times B} = \phi_{O|A} \circ \phi_{A|A\times B}$$

igennem A, og en faktorisering

$$\phi_{O|A\times B} = \phi_{O|B} \circ \phi_{B|A\times B}$$

igennem B.

De to faktoriseringer er illustreret i nedenstående figur:

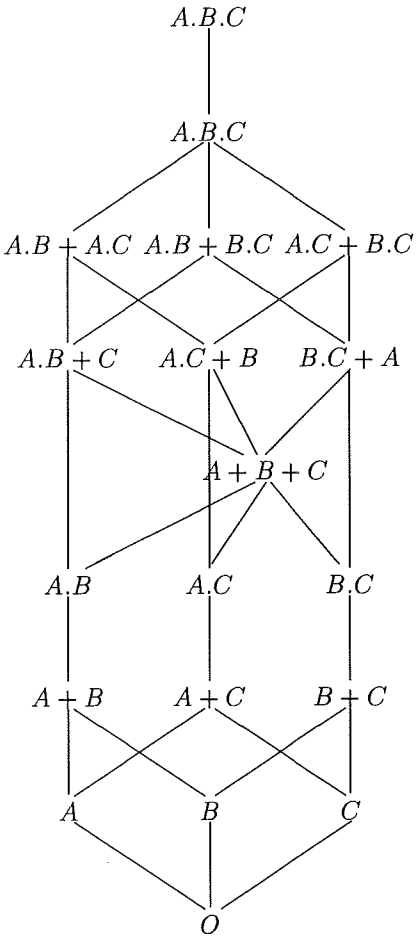


Den tilsvarende partielle ordning af klassifikationerne I, $A.B$, A, B, og O er illustreret i nedenstående diagram:

$$I \leq A.B \begin{array}{c} \leq \\ \leq \end{array} \begin{array}{c} A \\ \\ B \end{array} \begin{array}{c} \leq \\ \leq \end{array} O$$

Eksempel 2.9.4 ***Delmodeller af trefaktormodel***

Nedenstående inklusionsdiagram viser modelformlerne (se afsnit 2.10) svarende til delmodellerne af den fulde model for tre faktorer A, B og C.



□

2.9.8 Aliasrelationer mellem parametre, marginalitet

Betragt en model specificeret ved matricen $\mathbf{X}$ (svarende til et vilkårligt sæt af kontinuerte og kvalitative kovariable og til eventuelle blandede led).

En søjlevektor $\mathbf{x}_j$, $j = 1, 2, \ldots, m$ i $\mathbf{X}$ kan opfattes som en vektor i $\mathbb{R}^k$. Sættet af søjlevektorerne $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ udspænder et underrum af $\mathbb{R}^k$. Vi

siger at underrummet er udspændt af $\mathbf{X}$. Rummet symboliseres undertiden span($\mathbf{X}$).

Såfremt $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ er lineært uafhængige, har underrummet span($\mathbf{X}$) maksimal dimension,
dim(span($\mathbf{X}$)) $= m$.

Hvis der findes p uafhængige lineære bånd imellem $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ er dim(span($\mathbf{X}$)) $m - p$.

Betragt nu en matrix $\mathbf{X} = (\mathbf{U}|\mathbf{W})$, hvor dim($\mathbf{U}$) $= k \times p$ og dim($\mathbf{W}$) $= k \times q$ Vi kan uden indskrænkning antage at $q \leq p$.

Matricen svarer til at opfatte delrummet $L_X \subset \mathbb{R}^I$ som en sum, $L_X = L_U + L_W$.

Der er tre mulige relationer mellem span($\mathbf{U}$) og span($\mathbf{W}$), svarende til tre former for opspaltning af L_X:

a) Alle $p + q$ vektorer, der udspænder span($\mathbf{U}$) og span($\mathbf{W}$) er lineært uafhængige. I dette tilfælde er dim(span($\mathbf{X}$)) $= p + q$. Rummet L_X kan udtrykkes som en direkte sum af rummene L_U og L_W.

b) Alle vektorer i $\mathbf{W}$ kan udtrykkes som linearkombination af vektorer i $\mathbf{U}$, dvs. span($\mathbf{W}$) $\subset$ span($\mathbf{U}$) og $L_W \subset L_U$. I dette tilfælde, hvor span($\mathbf{W}$) er et underrum af span($\mathbf{U}$), gælder, at span($\mathbf{X}$) = span($\mathbf{U}$), hvorfor dim(span($\mathbf{X}$)) $= p$ og $L_X = L_U$. Vi siger at modellen svarende til $\mathbf{W}$ er en delmodel af modellen svarende til $\mathbf{U}$

c) r af de q vektorer i $\mathbf{W}$ kan udtrykkes som linearkombination af vektorer i $\mathbf{U}$ svarende til at fællesmængden for span($\mathbf{U}$) og span($\mathbf{W}$) har dimensionen r, svarende til dim$(L_U \cap L_W) = r$.

I en parametrisk repræsentation

$$\eta_i = \alpha_1 u_{i1} + \cdots + \alpha_p u_{ip} + \beta_1 w_{i1} + \cdots + \beta_q w_{iq}, \; i = 1, 2, \cdots, k$$

vil overlappende underrum som i ovenstående tilfælde b) og c) indebære, at ikke alle parametre er identificerbare. Man siger, at der er en aliasing mellem α-parametrene og β-parametrene.

I tilfælde b) vil man principielt kunne undvære parametrene $\alpha_1, \ldots, \alpha_p$. I tilfælde c) vil man principielt kunne undvære r af α-parametrene, eller af β-parametrene. I en række tilfælde vælger man alligevel at bevare overparametriseringen (modelmatricen $\mathbf{X}$), og evt. indføre nogle lineære bånd for at sikre entydigheden.

Aliasing ved faktorvariable

Betragt en model, der alene indeholder et interceptled for niveau og en enkelt faktor A.

En parametrisk repræsentation (på faktorform) af modellen er

$$\eta_{p\nu} = \kappa + \alpha_p$$

hvor p indikerer niveauerne for A, og ν indikerer gentagelserne inden for gruppen. Vi har allerede tidligere bemærket at vektorerne i modelmatricen for A adderer til vektoren for den konstante faktor. Der er således en aliasrelation, der sammenknytter κ med $\sum_p \alpha_p$. Dette er en iboende aliasrelation, da den gælder, uanset hvordan vi allokerer enheder til A.

Relationen mellem κ og α_p er ikke symmetrisk, da modelvektoren $\mathbf{1}$ svarende til den konstante faktor er fuldstændig indeholdt i rummet udspændt af modelmatricen for A svarende til tilfælde b). Vi siger, at parameteren κ er marginal i forhold til α'erne.

Det er netop på grund af denne marginalitet, at vi har opskrevet den parametriske model med κ-leddet først. En konsekvens af marginaliteten er, at det ikke har mening at undersøge en hypotese om κ, f.eks. $\kappa = 0$, såfremt værdierne af α ikke er specificerede.

Overparametriseringen er uden betydning for bestemmelsen af den lineære prædiktor $\eta_{p\nu}$. Værdien af den lineære prædiktor ændres således ikke, selv om vi adderer en konstant c til κ, og subtraherer den samme konstant fra hvert α_p. En sådan addition og subtraktion ændrer ikke størrelsen $\kappa + \alpha_i$, ligesom den heller ikke ændrer kontraster af formen $\sum_p \lambda_p \alpha_p$ med $\sum_p \lambda_p = 0$. Størrelser, der ikke ændres ved sådanne additioner og subtraktioner af konstanter til parameterværdier, kaldes estimable.

Når man imidlertid betragter estimater $\hat{\kappa}$ og $\hat{\alpha}_p$ af de indgående parametre, er det nødvendigt, at estimaterne er entydigt identificerbare. Man indfører sædvanligvis nogle lineære bånd imellem parameterestimaterne for at opnå at de kan bestemmes entydigt. For en ordens skyld skal vi påpege, at sådanne bånd er ikke en del af modelspecifikationen. De påvirker ikke vurderingen af modeltilpasningen. De er blot en bekvem måde til løsning af problemer, der skyldes overparametriseringen.

I ovenstående model vil man sædvanligvis benytte en af de følgende muligheder

1) Fastsæt $\kappa = 0$, sådan at α_p direkte udtrykker gruppemiddelværdierne

2) Fastsæt $\alpha_1 = 0$, sådan at middelværdien i den første gruppe tillægges værdien $\hat{\kappa}$. Parametrene α_p udtrykker således forskellene mellem middelværdien i den p'te gruppe og den første, $p = 2, \ldots, r$

3) Fastsæt $\sum \alpha_p = 0$, sådan at parameterestimatet κ udtrykker gennemsnittet af gruppemiddelværdierne, og α_p udtrykker forskellen mellem værdien i den p'te gruppe og dette gennemsnit.

Vi så i afsnit 2.9.3, at indførelsen af sådanne lineære bånd imellem parametrene kan opfattes som en "kodning" af faktoreffekterne, $\boldsymbol{\alpha} = \mathbf{C}\boldsymbol{\gamma}$, hvor $\mathbf{C}$ er en $r \times (r-1)$ matrix af fuld rang $r-1$. Modelmatricen $\mathbf{U}_A^*$ svarende til denne kodning bliver den $k \times (r-1)$ dimensionale matrix

$$\mathbf{U}_A^* = \mathbf{U}_A\mathbf{C}$$

I eksempel 2.9.2 har vi set forskellige eksempler på en sådan kodning. Vi vil her blot supplere med at nævne en parametrisering ved ortogonale polynomier.

Såfremt niveauerne $\{1, 2, \ldots, r\}$ for en faktor kan opfattes som repræsentererende en underliggende numerisk variabel, hvor værdierne svarende til de r faktorniveauer har samme indbyrdes afstand,kan man vælge at kode koefficienterne α_i som koefficienter γ_i i $r-1$ ortogonale polynomier af grad 1 til $r-1$.

For $r = 4$ får man

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} = \begin{pmatrix} -0.6708204 & 0.5 & -0.2236068 \\ -0.2236068 & -0.5 & 0.67082404 \\ 0.2236068 & -0.5 & -0.6708204 \\ 0.6708204 & 0.5 & 0.2236068 \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix}$$

Eksempel 2.9.5 ***Aliasrelationer ved tosidet klassifikation***

Vi betragter her et tosidigt sæt af observationer,

$$\begin{pmatrix} Y_{11} & Y_{12} & \ldots & Y_{1s} \\ Y_{21} & Y_{22} & \ldots & Y_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{r1} & Y_{r2} & \ldots & Y_{rs} \end{pmatrix} \tag{2.9.8}$$

organiseret i r rækker og k søjler.

Data svarer til en klassifikation af indeksmængden $I = \{1, 2, \ldots, r \cdots\}$ efter to kriterier A (rækker) og B (søjler).

Såfremt vi i en situation med en tosidet klassifikation med faktorerne A og B havde formuleret den parametriske model

$$\eta_{p,q} = \kappa + \alpha_p + \beta_q + \gamma_{p,q} \tag{2.9.9}$$

ville vi klart have en overparametrisering.

Denne overparametriserede repræsentation kan illustrere de forskellige aliasrelationer i modellen.

Parametriseringen har følgende aliasrelationer, hidrørende fra de lineære bånd $\mathbf{u}_1 + \mathbf{u}_2 + \cdot = \mathbf{1}$ imellem incidensvektorerne:

$$\begin{array}{ll} \sum_p \alpha_p \equiv \kappa & \sum_p \alpha_p \equiv \kappa \\ \sum_p \gamma_{p,q} \equiv \alpha_p & \sum_q \gamma_{p,q} \equiv \beta_q \ , \end{array}$$

hvor symbolet $\equiv$ her betyder "er det samme som".

Disse identiteter indebærer i sig selv, at summen af incidensvektorerne svarende til klassifikationen $A \times B$ er $\mathbf{1}$, svarende til

$$\sum_{p,q} \gamma_{p,q} \equiv \kappa$$

Der gælder således, at

κ er marginal i forhold til α_p, β_q og $\gamma_{p,q}$,
α_p er marginal i forhold til $\gamma_{p,q}$,
β_q er marginal i forhold til $\gamma_{p,q}$

Det vil sige, at vi har en kæde, hvor κ er marginal i forhold til (α_p, β_q), som igen er marginal i forhold til $\gamma_{p,q}$

De estimable parameterkombinationer er dels den lineære prædiktor η_{ipq} (2.9.9) og kontraster af formen

$$\begin{array}{lll} \sum l_p(\alpha_p + \bar{\gamma}_{p\cdot}) & \text{med} & \sum l_p = 0 \\ \sum l_q(\beta_q + \bar{\gamma}_{\cdot q}) & \text{med} & \sum l_q = 0 \\ \sum l_{p,q}\gamma_{p,q} & \text{med} & \sum_p l_{p,q} = \sum_q l_{p,q} = 0, \end{array}$$

hvor $\bar{\gamma}_{p\cdot}$ angiver gennemsnittene af $\gamma_{p,q}$ i den p'te række, og $\bar{\gamma}_{\cdot q}$ tilsvarende angiver gennemsnittene af $\gamma_{p,q}$ i den q'te søjle.

Ofte benyttes de lineære bånd

$$\sum \hat{\alpha}_p = 0 \qquad \sum \hat{\beta}_q = 0$$
$$\sum_q \hat{\gamma}_{p,q} = 0 \,, p = 1, \ldots, r \qquad \sum_p \hat{\gamma}_{p,q} = 0 \,, q = 1, \ldots, s$$

Imidlertid er det kun $r + s - 1$ af de sidste $r + s$ bånd, der er lineært uafhængige, så der er essentielt kun $r + s + 1$ lineære bånd. Indføres disse bånd kan vi løse de $r \cdot s$ middelværdiligninger for en kanonisk linkfunktion. Under den identiske linkfunktion får man

$$\begin{aligned} \hat{\kappa} &= \bar{y}_{\cdot\cdot} \\ \hat{\alpha}_p &= \bar{y}_{p\cdot} - \bar{y}_{\cdot\cdot} \\ \hat{\beta}_q &= \bar{y}_{\cdot q} - \bar{y}_{\cdot\cdot} \\ \hat{\gamma}_{p,q} &= y_{p,q} - \bar{y}_{p\cdot} - \bar{y}_{\cdot q} + \bar{y}_{\cdot\cdot} \end{aligned}$$

Man kunne imidlertid også have valgt andre bånd, f.eks. $\hat{\alpha}_1 = \hat{\beta}_1 = \hat{\gamma}_{i,1} = \hat{\gamma}_{1,q} = 0$ jvf. eksempel 2.9.2. □

Aliasing frembragt af de valgte variable

En analog form for overparametrisering optræder, hvis de valgte variable har en iboende lineær afhængighed.

Antag således, at man betragter en samling rektangulære objekter og registrerer de variable x_1 = logaritmen til længden, x_2 = logaritmen til bredden, og x_3 = logaritmen til arealet. Såfremt man bruger en model med x_1,x_2 og x_3 som kovariater, vil det lineære bånd $x_3 = x_1 + x_2$ indebære, at udtrykket

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

reduceres til

$$\eta = \beta_0 + (\beta_1 + \beta_3) x_1 + (\beta_2 + \beta_3) x_2$$

Vi kan således kun gøre os håb om at bestemme koefficienterne β_0, $(\beta_1 + \beta_3)$ og $(\beta_2 + \beta_3)$.

En sådan overparametrisering vil komme til udtryk ved at modelmatricen $\mathbf{X}$ ikke har rangen 4 (svarende til de fire parametre β_0, β_1, β_2, β_3 og β_4), men kun rangen 3 (svarende til parametrene β_0, $(\beta_1 + \beta_3)$ og $(\beta_2 + \beta_3)$.

Hvis der er en næsten lineær relation mellem nogle af de forklarende variable, vil man være tæt på en sådan situation. I et sådant tilfælde siger man, at der er kollinearitet mellem de forklarende variable.

Når der optræder en sådan kollinearitet vil estimationen af de berørte parametre være ganske usikker, og man kan risikere, at afrundingsfejl øver en væsentlig indflydelse på resultatet.

Omfanget af kollinearitet kan undersøges ved at betragte egenværdierne af modelmatricen $\mathbf{X}$. Ved en singulær værdi dekomposition af modelmatricen kan denne udtrykkes som

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}$$

hvor $\mathbf{U}$ er en $n \times m$ matrix, $\mathbf{V}$ en $m \times m$-matrix og $\mathbf{\Lambda}$ en $m \times m$ diagonalmatrix, og hvor søjlerne i $\mathbf{U}$ og $\mathbf{V}$ er ortogonale, dvs $\mathbf{U}^T\mathbf{U} = \mathbf{I}_n$ og $\mathbf{V}^T\mathbf{V} = \mathbf{I}_m$. Søjlerne i $\mathbf{U}$ er egenvektorer for $\mathbf{X}\mathbf{X}^T$; søjlerne i $\mathbf{V}$ er egenvektorer for $\mathbf{X}^T\mathbf{X}$ og egenværdierne for $\mathbf{X}^T\mathbf{X}$ er kvadratet på diagonalelementerne i $\mathbf{\Lambda}$. Indsættes singulær værdi dekompositionen af $\mathbf{X}$ i udtrykket

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

får man

$$\boldsymbol{\eta} = \mathbf{Z}\boldsymbol{\gamma}$$

hvor $\mathbf{Z} = \mathbf{U}\mathbf{\Lambda} = \mathbf{X}\mathbf{V}^T$ og $\boldsymbol{\gamma} = \mathbf{V}\boldsymbol{\beta}$ svarende til, at de m forklarende variable $\mathbf{x}_1, \ldots, \mathbf{x}_m$ er erstattet af de m principalkomponenter $\mathbf{z}_1, \ldots, \mathbf{z}_m$, hvor

$$\mathbf{z}_j = v_{j1}\mathbf{x}_1 + v_{j2}\mathbf{x}_2 + \cdots + v_{jm}\mathbf{x}_m$$

De m søjler i $\mathbf{Z}$ er ortogonale, da

$$\mathbf{Z}^T\mathbf{Z} = (\mathbf{U}\mathbf{\Lambda})^T(\mathbf{U}\mathbf{\Lambda}) = \mathbf{\Lambda}^2$$

Kvadratsummen af elementerne i den j'te søjle af $\mathbf{Z}$ er netop kvadratet på den$j$'te egenværdi:

$$\sum_{i=1}^{n} z_{ij}^2 = \lambda_j^2$$

Størrelsen λ_j^2 udtrykker således variationen i den retning i $\mathbb{R}^m$, der bestemmes af $\mathbf{w}_j$. Hvis λ_j er lille i forhold til de øvrige λ'er, indikerer det således, at der ikke er stor variation i den pågældende retning, hvorfor den ikke har så stor indflydelse på fastlæggelse af planen.

Aliasing frembragt af de valgte værdier

Endelig kan det forekomme, at de værdier, der er valgt for kovariaterne, er valgt så uheldigt, at der reelt er lineære relationer mellem værdierne af forskellige kovariater.

Også i dette tilfælde vil ubestemtheden komme til udtryk i at modelmatricen $\mathbf{X}$ ikke har fuld rang.

2.10 Modelformler

I mange programsystemer benyttes en kompakt modelformel til specifikation af lineære modeller.

Vi vil her skitsere en symbolsk notation, der er foreslået af Wilkinson og Rogers (1978). Notationen bruges i det store og hele i de sædvanlige programsystemer, f.eks. BMDP, Genstat, SAS og S-plus, hvor man kan referere til variable ved symbolske variabelnavne.

Vi vil lade A og B betegne faktorvariable (klassifikationer af indeksmængden), og X angive en kontinuert kovariabel. Lad incidensmatricerne for A og B være hhv $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r)$ for A, og $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_s)$ for B, og lad X have modelvektoren $\mathbf{x}$

En modelformel er en symbolsk beskrivelse af en lineær model ved en kombination af symboler for de variable og operatorer.

Vi vil nedenfor beskrive disse operatorer ved hjælp af de tilknyttede operationer på de tilsvarende modelmatricer og de led, der genereres i den parametriske repræsentation.

Interceptled

I de fleste systemer indgår der implicit et såkaldt interceptled som det første led i enhver modelformel.

Interceptleddet symboliseres undertiden ved et ettal. Interceptleddet svarer til en faktor, der kun har ét niveau.

Der gælder derfor i almindelighed for det første led i en modelformel

$$A \equiv 1 + A$$

eller

$$X \equiv 1 + X$$

Sædvanligvis er det en klog strategi at tillade et sådant interceptled, med mindre man har meget gode grunde til at udelade det.

Prik- eller punktoperatoren

Operatoren skrives ofte som et punktum "." eller kolon ":". Operatoren genererer led i den parametriske repræsentation svarende til alle kombinationer af de to variable, der "prikkes".

For to faktorvariable A og B dannes den $n \times (r \cdot s)$-dimensionale incidensmatrix svarende til $A.B$ ved elementvis multiplikation af alle søjlepar $\mathbf{u}_p$ og $\mathbf{w}_q$ svarene til konstruktionen i afsnit 2.9.5. Leddene i den parametriske repræsentation bliver de $r \cdot s$ led af formen $\gamma_{p,q}$, $p = 1, 2, \ldots, r$; $q = 1, 2, \ldots, s$.

For en faktorvariabel A og en kontinuert variabel X dannes den $n \times r$-dimensionale modelmatrix svarende til $A.B$ ved elementvis multiplikation af $\mathbf{x}$ med hver af søjlerne $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r$, som beskrevet ved konstruktionen af de blandede led i afsnit 2.8.5. Leddene i den parametriske repræsentation på faktorform bliver de r led af formen $\alpha_p x$, $p = 1, 2, \ldots, r$.

Der gælder

$$A.A = A$$

svarende til at incidensmatricen for $A.A$ har r søjler lig med søjlerne for A, og de resterende $r(r-1)$ søjler er nulvektorer.

Almindeligvis vil $X.X \neq X$. Venstre side kan fortolkes som en vektor med komponenter x_i^2. De fleste programsystemer tillader dog ikke denne konstruktion. Hvis man ønsker en forklarende variabel svarende til X^2 bør denne konstrueres eksplicit.

Prikoperatoren er kommutativ. Der gælder

$$A.B \equiv B.A$$

Den er desuden associativ, idet

$$(A.B).C \equiv A.(B.C)$$

Man kan derfor blot skrive $A.B.C$ etc.

Plusoperatoren

Plusoperatoren i modelformler udtrykker addition af parametre i den parametriske repræsentation som f.eks. i den additive model (2.9.4). Matricen svarende til plusoperationen fremkommer ved at sætte addendernes matricer ved siden af hinanden, dvs. matricen svarende til $A+B$ fås som $(\mathbf{U}|\mathbf{W})$ og matricen svarende til $A+X$ fås som $(\mathbf{U}|\mathbf{x})$.

Vi vil i afsnit 2.9.8 se, at matricen udspænder et underrum af $\mathbb{R}^n$. Det rum, der udspændes af incidensmatricen svarende til $A+B$ er således summen af rummene udspændt af modelmatricen svarende til A og incidensmatricen svarende til B.

Der gælder

$$A + A \equiv A$$

For at slippe for at skrive for mange parenteser bruger man almindeligvis den samme konvention som ved de sædvanlige plus og gange operationer, at "+" har lavere prioritet end ".", dvs.

$$A.B + C \equiv (A.B) + C$$

Prikoperatoren "." er distributiv med hensyn til plusoperatoren "+", dvs

$$A.(B + C) \equiv A.B + A.C$$

Krydsede faktorer

Krydsningsoperatoren "$*$" bruges hovedsagelig til at give en kompakt beskrivelse af modeller med flere faktorvariable. Operatoren "$*$" defineres som

$$A * B \stackrel{\text{DEF}}{=} A + B + A.B$$

$$A * B * C \stackrel{\text{DEF}}{=} A + B + C + A.B + A.C + B.C + A.B.C$$

etc. I den parametriske repræsentation svarer modelformlen $A * B$ således til led af formen $\alpha_p + \beta_q + \gamma_{p,q}$, dvs. modellering af hovedvirkningerne af A og B samt af vekselvirkningen mellem A og B.

Mange programsystemer tillader yderligere, at de variable A, B etc. i ovenstående udtryk erstattes med modelformler.

Krydsningsoperatoren "*" har højere prioritet end "+", men lavere prioritet end ".", dvs

$$\begin{aligned} A*B+C &\equiv A+B+C+A.B \\ A*B.C &\equiv A+B.C+A.B.C \end{aligned}$$

Krydsningsoperatoren "*" er associativ og distributiv med hensyn til "+". Der gælder nemlig

$$\begin{aligned} A*(B+C) &\equiv A+(B+C)+A.(B+C) \\ &\equiv A+B+C+A.B+A.C \\ &\equiv A+B+A.B+A+C+A.C \\ &\equiv A*B+A*C \end{aligned}$$

Fjernelse af led i modelformel

Operatoren "-" defineres som den modsatte operation af "+". Operatoren bruges til at udtrykke fjernelse af led i en modelformel.

Man har således udtrykket for en tofaktormodel uden vekselvirkning, dvs. alene med hovedvirkninger

$$A*B-A.B \equiv A+B$$

og tilsvarende udtrykker

$$A*B*C-A.B.C \equiv A+B+C+A.B+A.C+B.C$$

en model med tre faktorer, der indeholder alle tre hovedvirkninger og alle tre to-faktorvekselvirkninger.

2.10.1 Hierarkisk organiseret indeksmængde, underordnede faktorer

Vi nævner kort den særlige struktur, der er knyttet til en hierarkisk klassifikation af indeksmængden.

En hierarkisk klassifikation optræder når niveauerne for én klassifikation, f.eks. B, er underordnet niveauerne for en anden klassifikation, f.eks. A, se afsnit 2.9.6.

I modelformlen udtrykker man underordning ved symbolet "/". Således udtrykker

$$A/B \stackrel{\text{DEF}}{=} A + A.B$$

at B er underordnet A. Fortolkningen af $A.B$ er her effekter af B indenfor A.[2]

Vi bemærker, at A/B adskiller sig fra $A * B$ ved at der ikke er noget led, der udtrykker hovedvirkning af B. Et sådant led er jo uden mening, da niveauerne af faktor B jo relaterer til det pågældende niveau af faktor A.

Såfremt man tillader symbolerne i modelformlen selv at symbolisere modelformler, defineres "/" ved

$$A/B \stackrel{\text{DEF}}{=} A + \text{pl}(A).B$$

hvor vi fortolker pl(A) som "produktleddet" (ved brug af prikoperatoren) af alle elementerne i A.

Således har vi eksempelvis

$$(A * B)/C \equiv A * B + A.B.C$$

Underordningsoperatoren er associativ. Der gælder således

$$A/(B/C) \equiv (A/B)/C$$

Underordningsoperatoren er distributiv med hensyn til "+" idet der gælder

$$A/(B + C) = A + A.(B + C) = A + A.B + A + A.C = A/B + A/C$$

I lighed med krydsningsoperatoren har underordningsoperatoren en prioritet imellem "." og "+". Man vælger ofte konventionen at give den prioritet over "*".

[2] I programsystemet SAS® bruges en parentes til at angive underordning af faktorer. I SAS® symboliserer leddet $B(A)$, at faktoren B er underordnet A. Angivelse af leddet $B(A)$ i en modelformel indebærer sædvanligvis, at alle vekselvirkningsleddene mellem B og A indgår i modellen. Såfremt man ønsker hovedeffekterne af A modelleret (af hensyn til deviansopspaltningen), skal man eksplicit angive det i modelformlen ved at skrive $A + B(A)$.

2.11 Test for modelreduktion

Fil: glm3a.tex 97-03-16

2.11.1 Indledning, strategier for modeltilpasning

Vi vil i dette afsnit kort introducere principperne for test for modelreduktion.

Ved modellering af et sæt observationer $y_1, \ldots, y_k$ fra en eksponentiel dispersionsparameterfamilie vil man tage udgangspunkt i den fulde model, der tillader alle observationer at variere frit,

$$H_F: \quad \boldsymbol{\mu} \in \Omega_F \subset \mathbb{R}^k . \tag{2.11.1}$$

Når man vil modellere observationssættet ved en generaliseret lineær model vil man sædvanligvis begynde med en rimeligt omfattende model, udgangshypotesen H_0, som indeholder alle de forklarende variable, man kan forestille sig, der kan komme på tale.

Såfremt dispersionsparameteren, σ^2, er kendt, kan man umidddelbart benytte teststørrelsen for modeltilpasning til denne model, $G^2(H_0) = D^*(\mathbf{y}; \widehat{\boldsymbol{\mu}})$ til at vurdere, hvorvidt denne model kan antages at holde. Hvis hypotesen H_0 må afvises, kan man overveje, hvorvidt det er muligt at inddrage flere forklarende variable i modellen (feks efter en analyse af residualerne med henblik på at finde systematiske effekter), eller om der er tale om overdispersion (dvs større tilfældig variation end den, der er tilgodeset ved den pågældende fordeling og dens variansfunktion). Vælger man at fortolke den ringe modeltilpasning som et udtryk for overdispersion må man estimere dispersionsparameteren σ^2. Evt kan man benytte betragtningerne i afsnit 6 til at fortolke overdispersionen.

Såfremt dispersionsparameteren, σ^2, ikke er kendt, er man sædvanligvis nødt til at antage, at modellen er dækkende, og man kan da som anført i sætning 2.6.4 bruge residualdeviansen $D(\mathbf{y}; \widehat{\boldsymbol{\mu}})$ til estimation af skaleringsfaktoren σ^2 ved (2.6.11)

$$\widehat{\sigma}^2 = \frac{D(\mathbf{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))}{f} ,$$

hvor f angiver antallet af frihedsgrader knyttet til residualdeviansen.

Resultatet af tilpasningen til H_0 resumeres i SAS® proceduren INSIGHT[3] i en oversigtstabel af formen:

```
                    Summary of Fit
Mean of Response    0.5250  Deviance          1.8854
SCALE               1.0000  Deviance / DF     0.9427
                        .   Scaled Dev        1.8854

                    Summary of Fit
Pearson Chi-Sq                                1.8107
Pearson Chi-Sq / DF                           0.9054
Scaled Chi-Sq                                 1.8107
```

hvor rubrikken `Deviance` angiver residualdeviansen $D(\mathbf{y};\widehat{\boldsymbol{\mu}})$, størrelsen `SCALE` angiver skaleringsfaktoren $1/\sigma^2$ og rubrikken `Scaled Dev` angiver den skalerede residualdevians $D^*(\mathbf{y};\widehat{\boldsymbol{\mu}})$.

Endvidere udskrives størrelsen `Deviance / DF` , nemlig $D(\mathbf{y};\widehat{\boldsymbol{\mu}})/f$. Såfremt modellen er sand og dispersionsparameteren er 1, vil denne størrelse følge en $\chi^2(f)/f$-fordeling Hvis der indgår en ukendt dispersionsparameter i modellen, kan man i stedet bruge denne størrelse til at estimere σ^2. . (Udskriften beskriver resultatet af tilpasningen af modellen $LEV = LENG + TYKS$ med en binomial responsfordeling og logit-link til data i eksempel 2.4.2, hvor dispersionsparameteren σ^2 netop var valgt til at være 1).

Såfremt teststørrelsen $G^2(H_0) = D^*(\mathbf{y};\widehat{\boldsymbol{\mu}})$ ikke giver anledning til afvisning af hypotesen H_0, og såfremt en analyse af residualerne heller ikke giver anledning til tvivl om hypotesens validitet, vil man anse modellen H_0 for at være tilstrækkelig til at beskrive data.

En analyse er imidlertid ikke fuldstændig, før man har vurderet betydningen af de forklarende variable, og herunder har vurderet om alle indgående variable også er nødvendige, eller om der er basis for at reducere modellen.

En sådan vurdering kan tage udgangspunkt dels i en vurdering af estimaterne af koefficienterne svarende til de forklarende variable (afsnit 2.11.2), og dels i en analyse af deviansforøgelsen hvis de pågældende variable fjernes fra modellen (afsnit 2.11.3).

[3] ved brug af menuen `Fit` under `Analyze` menuvalg

2.11.2 Test af enkelte parametre

Såfremt man blot vil undersøge enkelte koefficienter i modellen, kan man benytte resultatet i nedenstående sætning til at vurdere hvorvidt koefficienten kan antages at have en specificeret værdi (sædvanligvis værdien 0 svarende til at den pågældene variable ikke har nogen betydning).

Sætning 2.11.1 *Test af hypoteser vedrørende enkelte værdier β_j*

I tilfældet, hvor dispersionsparameteren σ^2 er kendt, kan hypoteser $\beta_j = \beta_j^0$ vedrørende specifike værdier for de enkelte parametre testes ved at betragte teststørrelsen

$$u_j = \frac{\hat{\beta}_j - \beta_j^0}{\sqrt{\sigma^2 \widehat{\sigma}_{jj}}}, \tag{2.11.2}$$

hvor $\widehat{\sigma}_{jj}$ angiver det$j$'te diagonalelement i $\widehat{\boldsymbol{\Sigma}}$, og sammenligne med fraktiler i en $\mathrm{N}(0,1)$ fordeling.

Testet forkaster for ekstreme værdier af u_j. Ved et tosidet test på niveau α har testet har testet således det kritiske område

$$|u_j| > u_{1-\alpha/2}$$

Specielt bliver teststørrelsen for hypotesen $\beta_j = 0$

$$u_j = \frac{\hat{\beta}_j}{\sqrt{\sigma^2 \widehat{\sigma}_{jj}}} \tag{2.11.3}$$

Et ækvivalent test fås ved at betragte teststørrelsen

$$z_j = u_j^2 \tag{2.11.4}$$

og forkaste for $z_j > \chi^2_{1-\alpha}(1)$.

Såfremt dispersionsparameteren σ^2 ikke er kendt, men er estimeret ved (2.6.11) eller (2.6.12), benyttes i stedet teststørrelserne

$$t_j = \frac{\hat{\beta}_j - \beta_j^0}{\sqrt{\widehat{\sigma}^2 \widehat{\sigma}_{jj}}} \tag{2.11.5}$$

Under hypotesen $\beta_j = \beta_j^0$ vil t_j approximativt følge en $t(k-m)$ fordeling. Det kritiske område er derfor

$$|t_j| > t_{1-\alpha/2}(k-m)$$

Bevis:
Følger af sætning 2.5.1. □

Eksempel 2.11.1 ***Udplantning af blommestiklinger, test af hypotese vedrørende enkelte koefficienter***

Vi betragter atter situationen i eksempel 2.4.2.

Vi vil her undersøge hypotesen om effekten af tykkelse "middel" adskiller sig fra referenceværdien, tykkelse "tynd".

Parameteren γ_1 svarende til tykkelse "middel" blev estimeret til $\widehat{\gamma}_1 = 1.6059$ med den estimerede spredning $\sqrt{\sigma_{33}} = 0.5082$.

Teststørrelsen for hypotesen $\gamma_1 = 0$ er da

$$u_3 = \frac{1.6059 - 0}{0.5082} = 3.16$$

Idet $\mathsf{P}\ [\mathrm{N}(0,1) \geq 3.16] = 0.0079$ ses det, at hypotesen må afvises ved (tosidet) test på ethvert niveau større end 1.6 %. □

Eksempel 2.11.2 ***Udplantning af blommestiklinger, bestemmelse af konfidensinterval for enkelte koefficienter***

Vi betragter atter situationen i eksempel 2.4.2. I det foregående eksempel (eksempel 2.11.1) betragtede vi det sædvanlige test (u-test) for hvorvidt det kunne antages, at overlevelsen svarende til tykkelse "middel" kunne tænkes at være den samme som overlevelsen svarende til referenceniveauet, tykkelse "tynd", dvs om koefficienten svarende til tykkelse "middel" kunne tænkes at være 0.

Vi vil her illustrere, hvorledes denne vurdering tilsvarende kunne foretages ved brug af et konfidensinterval.

Parametrene til det sædvanlige Wald-konfidensinterval fås af eksempel 2.11.1. Parameteren γ_1 svarende til tykkelse "middel" blev estimeret til $\widehat{\gamma}_1 =$

1.6059 med den estimerede spredning $\sqrt{\sigma_{33}} = 0.5082$. Wald-konfidensintervallet svarende til 95 % konfidenssandsynlighed fås da som

$$\widehat{\gamma}_1 \pm 1.96\,\sqrt{\sigma_{33}}\,,$$

idet $u_{0.975} = 1.96$. Man får således intervallet

$$1.6059 \pm 1.96 \times 0.5082 = \begin{cases} 0.6098 \\ 2.6020 \end{cases}$$

Intervallet omfatter ikke værdien nul.

Det likelihoodkvotientbaserede konfidensinterval må bestemmes ved iteration.

Logaritmen til likelihoodfunktionen i maksimumspunktet er

$$l(\widehat{\beta}; \mathbf{y}) = -69.5127$$

og idet $\chi^2_{0.95}(1) = 3.841$, skal man altså bestemme de to værdier af γ_1 (henholdsvis mindre og større end $\widehat{\gamma}_1 = 1.6059$) sådan at logaritmen til profillikehooden $\widetilde{l}(\gamma_1; \mathbf{y})$ netop er

$$\widetilde{l}(\gamma_1; \mathbf{y}) = -69.5127 - 0.5 \times 3.841 = -71.4334$$

I SAS® Insight-proceduren under menuen Fit kan man klikke på Output knappen, hvorved der fremkommer en menu, der giver mulighed for valg af output-tabeller. Her kan man specielt vælge hhv. 95 % Wald-konfidensintervaller eller 95 % LR (likelihood-kvotient) konfidensintervaller for de estimerede parametre.

Herved fås konfidensintervaller for samtlige parametre, som vist nedenfor:

95% C.I. (Wald) for Parameters

Variable	LENG	TYK	Estimate	Lower	Upper
INTERCEPT			-0.6342	-1.4277	0.1592
LENG	KO		-1.0735	-1.8989	-0.2482
	LA		0.0000	.	.
TYK		MID	1.6059	0.6098	2.6020
		TYK	2.2058	1.1602	3.2514
		TYN	0.0000	.	.

95% C.I. (LR) for Parameters

Variable	LENG	TYK	Estimate	Lower	Upper
INTERCEPT			-0.6342	-1.4672	0.1378
LENG	KO		-1.0735	-1.9267	-0.2659
	LA		0.0000	.	.
TYK		MID	1.6059	0.6393	2.6439
		TYK	2.2058	1.2012	3.3050
		TYN	0.0000	.	.

Vi ser, at det likelihood-kvotient baserede 95 % konfidensinterval for koefficienten svarende til tykkelse "middel" er (0.6393; 2.6439). Intervallet er ikke symmetrisk omkring maksimum-likelihood estimatet $\widehat{\gamma}_1 = 1.6059$, men intervallet afviger dog ikke særlig meget fra det simplere Wald-konfidensinterval.

Hvis man bruger SAS®-proceduren GENMOD, kan man bruge optionerne WALDCI og LRCI i MODEL sætningen til at specificere, at man ønsker hhv Wald-konfidensintervaller eller likelihood-kvotient baserede konfidensintervaller for parametrene.

Nedenfor er vist udskriften fra PROC GENMOD

Likelihood Ratio Based Confidence Intervals For Parameters

Two-Sided Confidence Coefficient: 0.9500

Parameter		Confidence Limits	Parameter Values			
			PRM1	PRM2	PRM4	PRM5
PRM1	Lower	-1.4672	-1.4672	-0.7866	2.2815	2.8648
PRM1	Upper	0.1378	0.1378	-1.4112	1.0248	1.6507
PRM2	Lower	-1.9267	-0.3698	-1.9267	1.8377	2.5161
PRM2	Upper	-0.2659	-0.9701	-0.2659	1.5103	2.0764
PRM4	Lower	0.6393	-0.1552	-0.9838	0.6393	1.6712
PRM4	Upper	2.6439	-1.1515	-1.2497	2.6439	2.8344
PRM5	Lower	1.2012	-0.1839	-0.9472	1.0858	1.2012
PRM5	Upper	3.3050	-1.0978	-1.2886	2.1903	3.3050

Koefficienterne er her blot nummereret i den rækkefølge, de optræder i modelformlen, og referenceværdierne, længde "lang" og tykkelse "tynd" er udeladt, dvs koefficienten γ_1 skal her findes i linien PRM4. Udskriften viser

desuden estimaterne af de øvrige koefficienter svarende til denne værdi af γ_1.

Identifikationen af parametrene er i overensstemmelse med eksempel 2.4.2 på side 175:

PRM1	PRM2	PRM4	PRM5
μ	α_1	γ_1	γ_2
Intcpt	Kort	Mid	Tyk

Såfremt man yderligere ønsker at følge iterationerne kan man i SAS®-proceduren GENMOD bruge optionen ITPRINT i MODEL-sætningen. □

2.11.3 Test af delhypotese

Ved en vurdering af hvilke koefficienter, der er nødvendige til beskrivelse af interessevariablen, kan man naturligvis godt vurdere én koefficient ad gangen, som illustreret i det foregående afsnit. Man skal dog være forsigtig med at bruge en sådan fremgangsmåde til at fjerne mere end én koefficient uden at reestimere modellen. Den korrelation, der er mellem estimaterne af de enkelte koefficienter, kan bevirke, at selv om et antal af koefficienterne hver for sig kan antages at være nul, kan de ikke alle antages at være nul samtidig.

En anden indvending, man kan rette mod denne fremgangsmåde er, at selv om det enkelte test udføres, fx. på et 5 % niveau, vil niveauet svarende til den enkeltvise vurdering af et antal koefficienter være noget mindre end de formelle 5 %. (Hvis alle koefficienter virkelig var nul, ville sandsynligheden for at mindst én af r koefficienter blev dømt som værende forskellig fra nul ved r uafhængige tests af koefficienterne være $1-(0.95)^r$, altså væsentligt større end 5 %).

Man vil derfor sædvanligvis følge den fremgangsmåde, at man forsøger at vurdere et led i modelformlen ad gangen. Såfremt der er belæg for at fjerne det pågældende led, reestimeres modellen uden dette led og man vurderer, hvorvidt der er nogle de resterende led, der kan fjernes, etc.

Vi skal i det følgende introducere det begrebsapparat, der benyttes ved en sådan successiv testning af delhypoteser. Det vigtigste resultat er angivet i sætning 2.11.2, der angiver den fundamentale opspaltning af residualdevi-ansen $D(\mathbf{y};\widehat{\boldsymbol{\mu}})$ i additive bidrag, der måler afvigelsen mellem delmodeller.

Betragt den fulde model

$$H_F : \boldsymbol{\mu} \in \Omega_F \subset \mathbb{R}^k$$

for $Y_1, \ldots, Y_k$ og betragt modellen (den generaliserede lineære model)

$$H_0 : \quad \boldsymbol{\eta} \in L \subset \mathbb{R}^m$$

svarende til en $k \times m$-dimensional modelmatrix $\mathbf{X}_0$ af fuld rang m:

$$H_0 : \quad \boldsymbol{\eta} = \mathbf{X}_0\boldsymbol{\beta} , \quad \boldsymbol{\beta} \in B \subset \mathbb{R}^m , \tag{2.11.6}$$

Ved anvendelse af sætning 2.6.1 kan man teste, om modellen kan opretholdes.

Såfremt testet ikke fører til afvisning af hypotesen, kan man altså antage, at man har inddraget tilstrækkelig mange forklarende variable i modellen.

Lad H_1 angive en lineær delhypotese af H_0, dvs

$$H_1 : \quad \boldsymbol{\eta} \in L_1 \subset L$$

hvor $L_1 \subset \mathbb{R}^r$ med $r < m$.

Antag, at H_1 er udtrykt på den parametriske form

$$H_1 : \quad \boldsymbol{\beta} = \mathbf{G}_1\boldsymbol{\alpha} , \quad \boldsymbol{\alpha} \in A \subset \mathbb{R}^r ,$$

hvor den $m \times r$-dimensionale matrix $\mathbf{G}$ antages at have fuld rang, r.

Udtrykt ved den lineære prædiktor $\boldsymbol{\eta}$ er hypotesen:

$$H_1 : \quad \boldsymbol{\eta} = \mathbf{X}_1\boldsymbol{\alpha} , \quad \boldsymbol{\alpha} \in A \subset \mathbb{R}^r , \tag{2.11.7}$$

hvor modelmatricen $\mathbf{X}_1$ svarende til H_1 tilfredsstiller

$$\mathbf{X}_1 = \mathbf{X}_0\mathbf{G}$$

Lad $\boldsymbol{\mu}(\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}}))$ angive de fittede værdier svarende til maksimum-likelihood estimatet af $\boldsymbol{\alpha}$ under hypotesen H_1

Man kan da benytte sætning 2.6.1 til at udføre et test for modeltilpasning til hypotesen H_1. Kvotientteststørrelsen for modeltilpasning af H_1 (imod den fulde model) er den skalerede residualdevians, $G^2(H_1) = D^*\big(\mathbf{y}; \boldsymbol{\mu}(\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}}))\big)$,

hvor residualdeviansen $D(\mathbf{y};\boldsymbol{\mu}(\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}})))$ måler afvigelsen mellem H_F og H_1, dvs. modeltilpasningen til H_1.

Under H_1 følger $G^2(H_1)$ approximativt en $\chi^2(k-r)$ fordeling.

Teststørrelsen tager imidlertid ikke hensyn til, at vi allerede har accepteret H_0. En del af variationen i de observerede data består i afvigelsen mellem H_F og H_0. Ved testet for modeltilpasning til H_0 har vi allerede vurderet disse afvigelser som værende tilfældige. Når vi derfor nu ønsker et test for H_1, bør denne afvigelse ikke indgå i den benyttede teststørrelse.

Udtrykt på en anden måde: De to størrelser $G^2(H_0) = D^*(\mathbf{y};\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))$, og $G^2(H_1) = D^*(\mathbf{y};\boldsymbol{\mu}(\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}})))$ er ikke stokastisk uafhængige.

Da H_1 er en delhypotese af H_0 gælder åbenbart

$$\frac{\sup\limits_{\mu\in\Omega_1} L_y(\mu)}{\sup\limits_{\mu\in\Omega_F} L_y(\mu)} \leq \frac{\sup\limits_{\mu\in\Omega_0} L_y(\mu)}{\sup\limits_{\mu\in\Omega_F} L_y(\mu)},$$

hvor

$$\Omega_0 = \{\mu_1,\ldots,\mu_k \in \mathbb{R}^k \mid g(\mu_1),\ldots,g(\mu_k) \in L\}$$

og

$$\Omega_1 = \{\mu_1,\ldots,\mu_k \in \mathbb{R}^k \mid g(\mu_1),\ldots,g(\mu_k) \in L_1\}$$

angiver middelværdirummene under de respektive hypoteser H_0 og H_1.

Men $G^2(H_1)$ fås netop som -2 gange logaritmen til venstre side, og $G^2(H_0)$ er -2 gange logaritmen til højre side. Der gælder derfor, at

$$G^2(H_0) \leq G^2(H_1)$$

Forskellen $G^2(H_1) - G^2(H_0)$ kan imidlertid tillægges en selvstændig fortolkningsom anført i nedenstående sætning:

Sætning 2.11.2 ***Kvotienttest for delhypotese af generaliseret lineær model (kendt dispersionsparameter)***

Betragt den generaliserede lineære model

$$H_0: \quad \boldsymbol{\eta} \in L \subset \mathbb{R}^m$$

svarende til matrixfremstillingen $\boldsymbol{\eta} = \mathbf{X}_0\boldsymbol{\beta}$ (2.11.6), og betragt delhypotesen

$$H_1: \quad \boldsymbol{\eta} \in L_1 \subset \mathbb{R}^r$$

med matrixfremstillingen $\boldsymbol{\eta} = \mathbf{X}_1\boldsymbol{\alpha}$ (2.11.7).

Antag, at dispersionsparameteren σ^2 er kendt.

Kvotienttestet for H_1 under antagelse af at modellen H_0 gælder, har da teststørrelsen

$$\begin{aligned} G^2(H_1|H_0) &= G^2(H_1) - G^2(H_0) \\ &= D^*\left(\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}); \boldsymbol{\mu}(\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}}))\right) \end{aligned} \tag{2.11.8}$$

Under hypotesen H_1 vil teststørrelsen $D^*\left(\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}); \boldsymbol{\mu}(\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}}))\right)$ asymptotisk følge en $\chi^2(m-r)$ fordeling. Endvidere vil fordelingen af $D^*\left(\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}); \boldsymbol{\mu}(\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}}))\right)$ være asymptotisk uafhængig af $G^2(H_0) = D^*\left(\mathbf{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})\right)$.

Bevis:

Da H_1 er en delhypotese af H_0, kan man faktorisere kvotienttteststørrelsen $q_{F,1}$ for test af H_1 mod H_F i et produkt af kvotienttteststørrelsen $q_{F,0}(\mathbf{y})$ for test af H_0 mod den fulde model, og kvotienttteststørrelsen $q_{0,1}(\mathbf{y})$ for test af H_1 under H_0:

$$q_{F,1}(\mathbf{y}) = \frac{\sup\limits_{\mu\in\Omega_1} L_y(\mu)}{\sup\limits_{\mu\in\Omega_F} L_y(\mu)} = \frac{\sup\limits_{\mu\in\Omega_0} L_y(\mu)}{\sup\limits_{\mu\in\Omega_F} L_y(\mu)} \frac{\sup\limits_{\mu\in\Omega_1} L_y(\mu)}{\sup\limits_{\mu\in\Omega_0} L_y(\mu)} = q_{F,0}(\mathbf{y})q_{0,1}(\mathbf{y})$$

svarende til at logaritmen til likelihoodkvotienten spalter op i en sum

$$G^2(H_1) = G^2(H_0) + G^2(H_1|H_0)\,, \tag{2.11.9}$$

der udtrykker teststørrelsen for tilpasning til hypotesen H_1 som en sum af teststørrelsen for tilpasning til H_0 og teststørrelsen for det betingede test for tilpasning til H_1 (givet H_0 er sand).

Udtrykt ved residualdevianserne har man

$$D(\mathbf{y};\boldsymbol{\mu}(\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}}))) = D(\mathbf{y};\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})) + D(\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}});\boldsymbol{\mu}(\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}}))) \ . \tag{2.11.10}$$

Fordelingsforholdene følger af de generelle asymptotiske resultater for kvotienttestet. □

Bemærkning 1 ***Deviansanalyse***

Spaltningen (2.11.10) kaldes en deviansanalyse, fordi den spalter deviansen svarende til H_1 i to relevante komponenter. Spaltningen er analog til opspaltningen af kvadratafvigelsessummer for normalt fordelte observationer. □

Bemærkning 2 ***Fortolkning af opspaltningen***

Deviansopspaltningen (2.11.9) svarer til en opspaltning af residualdeviansen svarende til H_1 i i et bidrag, $D(\mathbf{y};\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))$, der beskriver afvigelsen fra H_0, og et bidrag, $D(\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}});\boldsymbol{\mu}(\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}})))$, der beskriver afvigelsen af H_1 fra H_0.

□

Bemærkning 3 ***Bestemmelse af devianserne***

Residualdeviansen $D(\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}});\boldsymbol{\mu}(\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}})))$ kan udtrykkes ved enhedsdevianserne $d(\cdot;\cdot)$ (2.2.8) som

$$D(\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}});\boldsymbol{\mu}(\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}}))) = \sum_{i=1}^{k} w_i \, d(\mu_i(\widehat{\boldsymbol{\beta}});\mu_i(\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}}))) \ , \tag{2.11.11}$$

hvor vi har brugt betegnelsen $\mu_i(\widehat{\boldsymbol{\beta}})$ for de fittede værdier under H_0, og $\mu_i(\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}}))$ for de fittede værdier under H_1.

Når vi har bestemt de fittede værdier $\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})$ og $\boldsymbol{\mu}(\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}}))$ under de to modeller H_0 og H_1, kan deviansbidragene $w_i \, d(\mu_i(\widehat{\boldsymbol{\beta}});\mu_i(\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}}))$ altså bestemmes ved hjælp af tabel 2.3. □

Tabel 2.8. **Eksempel på deviansanalyseskema omkring interceptled**

Variationskilde	f	Devians	middeldevians	Goodness of fit fortolkning
Model H_0	$m-1$	$D(\mu(\widehat{\beta});\widehat{\mu}_M)$	$\frac{D(\mu(\widehat{\beta});\widehat{\mu}_M)}{m-1}$	$G^2(H_M\|H_0)$
Residual (Error)	$k-m$	$D(\mathbf{y};\mu(\widehat{\beta}))$	$\frac{D(\mathbf{y};\mu(\widehat{\beta}))}{k-m}$	$G^2(H_0)$
Korrigeret Total	$k-1$	$D(\mathbf{y};\widehat{\mu}_M)$		$G^2(H_M)$

Note: $\widehat{\mu}_M$ angiver estimatet for den fælles middelværdi.

Bemærkning 4 ***Deviansanalyseskema med reference til minimal model***

Man ser undertiden deviansanalyseskemaer svarende til en model H_0 udformet svarende til en deviansopspaltning for H_0 og den minimale model, H_M, nemlig en model, der kun indeholder et konstantled, det såkaldte interceptled. Modelmatricen for en model, der kun indeholder interceptled-det, er en søjle med lutter ettaller.

Et sådant deviansanalyseskema ser ud som vist i tabel 2.8

□

Eksempel 2.11.3 ***Devians svarende til minimal model***
Udplantning af blommestiklinger, fortsat

Vi betragter atter situationen i eksempel 2.4.2.

Den minimale model er modellen svarende til fuldstændig homogenitet, $p_{i,j} = p$. Under denne model finder man estimatet for overlevelsessandsyn-

ligheden

$$\widehat{\widehat{p}} = \frac{10+11+4+14+18+6}{120} = \frac{63}{120} = 0.5250$$

Deviansbidragene svarende til denne model er anført i nedenstående tabel

Obs nr	1	2	3	4	5	6
y_i	0.30	0.70	0.90	0.20	0.50	0.55
$\widehat{\widehat{p}}$	0.5250	0.5250	0.5250	0.5250	0.5250	0.5250
$w_i d(y_i;\widehat{\widehat{p}})$	0.0501	0.0502	8.9609	2.5407	13.1713	4.1420

Man finder

$$G^2(H_M) = D(\mathbf{y};\widehat{\widehat{\mathbf{p}}}) = \sum_{i=1}^{2}\sum_{j=1}^{3} w_{i,j}\, d(y_{i,j};0.5250) = 28.9152$$

Fra tabellen i eksempel 2.4.2 finder man

$$G^2(H_0) = D(\mathbf{y};\widehat{\mathbf{p}}) = \sum_{i=1}^{2}\sum_{j=1}^{3} w_{i,j}\, d(y_{i,j};\hat{p}_{i,j}) = 1.8854$$

sådan at man har

$$G^2(H_M|H_0) = G^2(H_M) - G^2(H_0) = 28.9152 - 1.8854 = 27.0298$$

I SAS® proceduren INSIGHT bevirker optionen `summary of fit table` under menuen `Fit` netop at der udskrives en tabel, der viser opspaltningen af residualdeviansen, $D(\mathbf{y};\widehat{\mu}_M)$, svarende til den minimale model i en sum af residualdeviansen, $D(\mathbf{y};\widehat{\mu})$, svarende til tilpasningen til H_0, og deviansen, $D(\widehat{\mu};\widehat{\mu}_M)$, svarende til afvigelsen mellem H_0 og H_M. Desuden udskrives middelresidualdeviansen og de skalerede residualdevianser $D^*(\mathbf{y};\widehat{\mu})$ og $D^*(\widehat{\mu};\widehat{\mu}_M)$, samt testsandsynligheden, $\mathrm{P}\left[\chi^2(m-1) > D^*(\widehat{\mu};\widehat{\mu}_M)\right]$ svarende til test af den minimale model, H_M, under antagelse af at udgangsmodellen H_0 er sand.

Analysis of Deviance

Source	DF	Deviance	Deviance / DF	Scaled Dev	Pr > Scaled Dev
Model	3.0	27.0298	9.0099	27.0298	0.0001
Error	2.0	1.8854	0.9427	1.8854	.
C Total	5.0	28.9152	.	.	.

hvor værdierne i søjlen `Deviance` og `Deviance / DF` fås som

Source	Deviance	Deviance/DF
Model	$D(\widehat{\mu};\widehat{\mu}_M)$	$D(\widehat{\mu};\widehat{\mu}_M)\big/(m-1)$
Error	$D(\mathbf{y};\widehat{\mu})$	$D(\mathbf{y};\widehat{\mu})\big/(k-m)$
C Total	$D(\mathbf{y};\widehat{\mu}_M)$	

hvor $\widehat{\mu}_M$ angiver middelværdiestimatet under den minimale model H_M.

Hvis man har specificeret en model uden intercept-led, er den minimale model en model, der tillægger alle observationer en middelværdi svarende til at den lineære prædiktor $\eta_M = 0$. Rubrikken `C Total` (Corrected total) erstattes da af en rubrik med betegnelsen `U Total` (Uncorrected total), der angiver $D(\mathbf{y};(\eta = 0))$.

Programsystemet S-plus udskriver tilsvarende størrelserne

```
    Null Deviance: 28.9152 on 5 degrees of freedom

Residual Deviance: 1.885398 on 2 degrees of freedom
```

svarende til $D(\mathbf{y};\widehat{\mu}_M)$ og $D(\mathbf{y};\widehat{\mu})$. □

Eksempel 2.11.4 ***Fosterdødelighed hos mus, (deviansanalyseskema)***

Vi betragter atter situationen fra eksempel 2.4.1.

I eksemplet forelå der $k = 5$ observationer af andelen $y = z/n$ af døde fostre ved forskellige koncentrationer, x [mg/kg pr. dag] af (diEGdiME).

Vi opstillede en generaliseret lineær model for den forventede andel p_i af døde fostre ved koncentrationen x_i. Vi modellerede antallet Z_i af døde fostre ved uafhængige $\mathrm{B}(n_i, p_i)$-fordelte størrelser, og betragtede den logistiske linkfunktion:

$$\eta = \ln\left(\frac{p}{1-p}\right)$$

med den lineære prædiktor

$$H_0: \quad \eta_i = \alpha + \beta x_i$$

I eksempel 2.5.2 på side 195 fandt vi estimaterne $\widehat{\alpha} = -3.248$ og $\widehat{\beta} = 0.006389$.

I eksempel 2.6.1 på side 221 fandt vi residualdeviansen svarende til den logistiske regressionsmodel, $D(\mathbf{y};\widehat{\mathbf{p}}) = 5.77$

Residualdeviansen skal sammenlignes med fraktilerne i en $\chi^2(3)$-fordeling. Idet $\chi^2_{0.95}(3) = 7.81$, er der ingen grund til at afvise hypotesen ved test på et 5 %-niveau.

Den minimale model:

$$H_M: \quad \eta_i = \alpha$$

svarende til fuldstændig homogenitet udtrykker, at andelen af døde fostre ikke afhænger af koncentrationen af tilsætningsstoffet.

Estimatet under H_M findes af middelværdiligningen (2.7.13):

$$\widehat{\widehat{p}} = \frac{\sum n_i y_i}{\sum n_i} = 0.164$$

Man kunne naturligvis også have estimeret parameteren α i den logistiske regressionsmodel ved middelværdiligningen:

$$\frac{\sum n_i y_i}{\sum n_i} = \frac{\exp(\alpha)}{1+\exp(\alpha)} ,$$

der fører til

$$\widehat{\widehat{\alpha}} = \ln(\frac{0.164}{0.836}) = -1.625 ,$$

hvilket også fører til $\widehat{\widehat{p}} = 0.164$.

Deviansbidragene svarende til denne hypotese er anført i nedenstående tabel:

Dosis x_i	Ant. fostre n_i	Ant. døde z_i	fitted $n_i\widehat{\widehat{p}}$	dev.bidr $n_i d(y_i;\widehat{\widehat{p}})$
0.0	297.0	15.00	48.86	36.72
62.5	242.0	17.00	39.81	19.17
125.0	312.0	22.00	51.33	24.56
250.0	299.0	38.00	49.19	3.26
500.0	285.0	144.00	46.89	175.39
Sum				259.11

Tabel 2.9. **Deviansanalyseskema for musefostre**

Variationskilde	f	Devians
Dosis	$2-1$	253.33
Omkring linie	$5-2$	5.78
Total	$5-1$	259.11

Deviansanalyseskemaet svarende til den logistiske regressionsmodel er vist i tabel 2.9.

Teststørrelsen for effekt af dosis, under antagelse af linearitet (af logit'erne) er da $D(\widehat{\mathbf{p}};\widehat{\widehat{p}}) = 253.33$, der skal sammenlignes med en $\chi^2(1)$-fordeling.

Idet $\chi^2_{0.9995}(1) = 12.1$, må den minimale model H_M klart afvises, dvs leddet svarende til afhængighed af dosis er nødvendigt. □

Definition 2.11.1 ***Wald-teststørrelse***

Betragt en generaliseret lineær model med en udgangshypotese H_0, givet ved (2.11.6), hvor modelmatricen svarende til H_0 er den $k \times m$-dimensionale matrix $\mathbf{X}_0$ med fuld rang (m), dvs

$$H_0 \quad \boldsymbol{\eta} = \mathbf{X}_0\boldsymbol{\beta}$$

Lad den r-dimensionale delhypotese H_1 være udtrykt som

$$H_1 \; : \quad \mathbf{L}^T\boldsymbol{\beta} = \mathbf{0} \tag{2.11.12}$$

hvor $\mathbf{L}^T$ angiver en $(m-r) \times m$-dimensional matrix af fuld rang $(m-r)$.

Størrelsen

$$Z = (\mathbf{L}^T\widehat{\boldsymbol{\beta}})^T(\mathbf{L}^T\widehat{\boldsymbol{\Sigma}}\mathbf{L})^{-1}(\mathbf{L}^T\widehat{\boldsymbol{\beta}}) \,, \tag{2.11.13}$$

hvor $\widehat{\beta}$ angiver maksimaliseringsestimatoren for β under H_0, og hvor $\widehat{\Sigma}$ angiver den estimerede dispersionsmatrix for $\widehat{\beta}$, kaldes Wald-teststørrelsen for test af H_1 under H_0.

Den estimerede dispersionsmatrix $\widehat{\Sigma}$ er bestemt ved (2.5.22) som matricen

$$\widehat{\Sigma} = [\mathbf{X}_0{}^T \mathbf{W}(\widehat{\beta}) \mathbf{X}_0]^{-1} \tag{2.11.14}$$

med

$$\mathbf{W}(\widehat{\beta}) = \text{diag}\left\{ \frac{w_i}{[g'(\widehat{\mu}_i)]^2 V(\widehat{\mu}_i)} \right\} , \tag{2.11.15}$$

□

Sætning 2.11.3 ***Fordeling af Wald's teststørrelse for modelreduktion***

Betragt hypotesen H_1 (2.11.12) i definition 2.11.1.

Under H_1 vil den skalerede Wald-teststørrelse

$$Z^* = \frac{Z}{\sigma^2} ,$$

hvor Z er givet ved (2.11.13), approximativt følge en $\chi^2(m-r)$-fordeling,

Hypotesen forkastes for store værdier af Z^*.

Bevis:

Følger af sætning 2.5.1 □

Bemærkning 1 ***Wald's teststørrelse måler direkte afvigelsen i β-rummet***

Teststørrelsen (2.11.13) er netop den generaliserede (og standardiserede) kvadratsum af de $m-r$ kontraster mellem β-værdierne. □

Sætning 2.11.4 ***Estimation og test af hypotesen H_1 under antagelse af H_0 for ukendt dispersionsparameter***

Såfremt dispersionsparameteren σ^2 ikke er kendt, kan man ikke teste for tilpasning til modellen H_0.

Dispersionsparameteren σ^2 kan estimeres som anført i sætning 2.6.4 ved (2.6.11), eller med udgangspunkt i Pearson residualerne (2.6.12), eller eventuelt ved maksimum-likelihood metoden.

Såfremt H_0 kan antages at være opfyldt, er teststørrelsen for modelreduktion til H_1 givet ved

$$Z_1 = \frac{G^2(H_1|H_0)/(m-r)}{G^2(H_0)/(k-m)} = \frac{D(\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}); \boldsymbol{\mu}(\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}})))/(m-r)}{D(\mathbf{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))/(k-m)} \quad (2.11.16)$$

Under H_1 vil Z_1 approximativt være fordelt som $\mathrm{F}(m-r, k-m)$. Hypotesen forkastes for store værdier af Z_1.

Bevis:
Approximationen hviler på det resultat, at $D^*(\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}); \boldsymbol{\mu}(\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}})))$ og $D^*(\mathbf{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))$ er approximativt uafhængige, og approximativt χ^2-fordelte. Da vil forholdet mellem de skalerede middeldevianser approximativt følge en F-fordeling. Men da dispersionsparameteren σ^2 forkortes ud ved bestemmelsen af forholdet mellem de skalerede middeldevianser, gælder resultatet altså også for forholdet mellem de uskalerede middeldevianser. □

Bemærkning 1 ***Deviansanalyseskema ved ukendt dispersionsparameter***

Tabel 2.10 illustrerer deviansanalyseskemaet svarende til disse tests. □

Variationskilde	f	Devians	middeldevians	Test
Mellem H_1 og H_0	$m-r$	$D(\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}});\boldsymbol{\mu}(\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}})))$	$\dfrac{D(\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}});\boldsymbol{\mu}(\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}})))}{m-r}$	$\dfrac{D(\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}});\boldsymbol{\mu}(\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}})))/(m-r)}{D(\mathbf{y};\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))/(k-m)}$
Afvigelse fra H_0	$k-m$	$D(\mathbf{y};\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))$	$\dfrac{D(\mathbf{y};\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))}{k-m}$	
Total	$k-r$	$D(\mathbf{y};\boldsymbol{\mu}(\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}})))$	$\dfrac{D(\mathbf{y};\boldsymbol{\mu}(\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}})))}{k-r}$	

Tabel 2.10. ***Eksempel på deviansanalyseskema for ukendt dispersionsparameter***

2.11.4 Modelreduktion ved successiv testning i hierarkiske hypoteser

Såfremt udgangsmodellen H_0 kan accepteres, kan det som nævnt være af interesse at forsøge at foretage en yderligere simplifikation af modellen.

I dette afsnit vil vi diskutere tests for en sådan yderligere reduktion. Vi vil her kun betragte hierarkisk organiserede lineære hypoteser for den lineære prædiktor, dvs. hypotesekæder af formen:

$$H_M \subset \ldots \subset H_2 \subset H_1 \subset H_0 \subset H_F \tag{2.11.17}$$

dvs.

$$H_M \Rightarrow \ldots \Rightarrow H_2 \Rightarrow H_1 \Rightarrow H_0 \Rightarrow H_F$$

hvor H_i er en lineær hypotese, og hvor H_F angiver den fulde model. Bunden af kæden, H_M, angiver den minimale model, der overhovedet kan komme i betragtning.

En hypotese H_i specificerer et underrum $L_i \subset \mathbb{R}^k$. Vi har tidligere set, at sådanne hypoteser kan repræsenteres ved en modelmatrix.

Modelmatricen svarende til den fulde model H_F er den k-dimensionale enhedsmatrix $\mathbf{I}_k$. Vi vil lade $\mathbf{X}_i$ angive en modelmatrix svarende til H_i. Den minimale model, H_M, vil sædvanligvis være en model svarende til fuldstændig homogenitet, dvs en model med blot et interceptled. Modelmatricen for denne minimale model er en søjle med lutter ettaller.

Modelhierarkiet svarende til hypoteserne kan da udtrykkes ved følgen af lineære underrum

$$\mathbb{R} \subseteq L_M \ldots \subset L_2 \subset L_1 \subset L_0 \subset \mathbb{R}^k\,, \tag{2.11.18}$$

svarende til

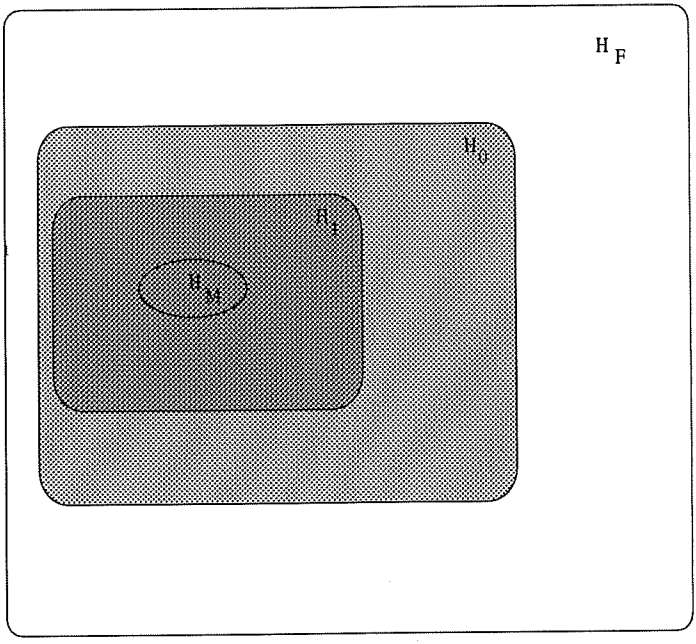
$$\mathbb{R} \subset \ldots \subset \text{span}(\mathbf{X}_2) \subset \text{span}(\mathbf{X}_1) \subset \text{span}(\mathbf{X}_0) \subset \mathbb{R}^k\,, \tag{2.11.19}$$

hvor $\text{span}(\mathbf{X})$ angiver underrummet udspændt af søjlerne i matricen $\mathbf{X}$.

Endelig, hvis man udtrykker den i'te hypotese ved middelværdimængden, Ω_i, svarende til hypotesen H_i, har man

$$\Omega_M \subset \ldots \subset \Omega_2 \subset \Omega_1 \subset \Omega_0 \subset \Omega_F$$

Modelhierarkiet er illustreret i nedenstående figur



Det følger ved gentagen anvendelse af sætning 2.11.2, at

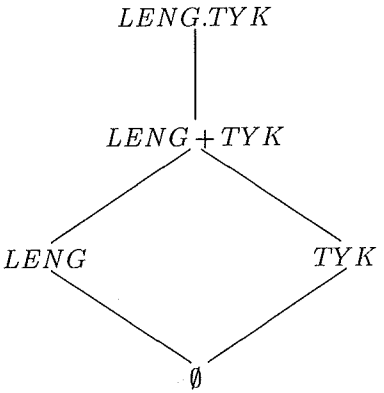
> såfremt en kæde af hypoteser er organiseret hierarkisk, kan man spalte residualdeviansen $D(\mathbf{y};\widehat{\boldsymbol{\mu}}_M)$ omkring den minimale model i en sum af bidrag, der hver for sig kan tillægges en selvstændig fortolkning som en kvotienttestørrelse.

For enhver model, H_0, kan man formulere mindst én kæde af hierarkisk organiserede hypoteser med den minimale model H_M som den laveste model i hierarkiet, og med H_0 som den højeste model. Som vi så af inklusionsdiagrammet i eksempel 2.9.4, vil man imidlertid i almindelighed kunne organisere en given model som en sådan kæde på flere forskellige måder.

Dog vil vekselvirkningsled være højere end de tilsvarende hovedeffekter.

Nedenstående inklusionsdiagram figur 2.8 illustrerer de to kæder, der kan konstrueres for et forsøg med to faktorer (jvf den partielle ordning af klassifikationerne side 269).

Figur 2.8. Inklusionsdiagram svarende til tofaktormodeller for blomme-forsøg



Det fremgår imidlertid af sætning 2.11.2, at

teststørrelsen for fjernelse af et bestemt led i modelformlen afhænger af, hvilke andre led, der indgår i modellen på det pågældende trin.

I praksis vil man derfor sikre sig at alle relevante led tilgodeses, når man vurderer, hvorvidt et bestemt modelled kan udelades af modellen.

Såfremt man i situationen svarende til figur 2.8 har accepteret hypotesen om forsvindende vekselvirkning, dvs at man har godtaget modellen svarende til modelformlen `LENG + TYK`, da vil testet svarende til den venstre gren i diagrammet modsvare at man undersøger, om der er påviselig effekt af tykkelsen, når man tilgodeser at stiklingerne har forskellig længde og korrigerer

herfor. Tilsvarende vil testet svarende til den højre gren i diagrammet modsvare at man undersøger, om der er påviselig effekt af længden, når man tilgodeser at stiklingerne har forskellig tykkelse og korrigerer herfor.

2.11.5 Modelreduktion ved partielle tests

Vi vil i dette afsnit beskrive proceduren for modelreduktion ved successiv fjernelse af led i modelformlen.

På et givet trin i modellen skal man vælge, ad hvilken gren man vil fortsætte reduktionen. Hertil benyttes sædvanligvis et partielt test:

Definition 2.11.2 *Partielt kvotienttest*

Betragt den hierarkisk organiserede kæde af hypoteser (2.11.17), og antag at hypotesen H_i har de forskellige delhypoteser $H_{i+1}^A \subset H_i$, $H_{i+1}^B \subset H_i$, ... $H_{i+1}^K \subset H_i$.

Ved det partielle kvotienttest for H_{i+1}^J under H_i forstås testet med teststørrelsen $G^2(H_{i+1}^J|H_i)$, dvs. testet, der måler, hvorvidt observationerne strider imod hypotesen H_{i+1}^J under antagelse af at hypotesen H_i kan opretholdes.

Det følger af sætning 2.11.2, at kvotienttesteststørrelsen for dette partielle test er

$$G^2(H_{i+1}^J|H_i) = G^2(H_{i+1}^J) - G^2(H_i) \tag{2.11.20}$$

og at kvotienttesteststørrelsen kan bestemmes som den skalerede devians mellem estimaterne $\widehat{\boldsymbol{\mu}}$ under H_i og estimaterne $\widehat{\widehat{\boldsymbol{\mu}}}$ under H_{i+1}^J

$$G^2(H_{i+1}^J|H_i) = D^*(\widehat{\boldsymbol{\mu}}; \widehat{\widehat{\boldsymbol{\mu}}}) \; . \tag{2.11.21}$$

□

Bemærkning 1 *Det partielle test betegnes undertiden et type III test*

Det partielle test af H_{i+1}^J under H_i kaldes undertiden det marginale kvotienttest af H_{i+1}^J (under H_i)

I SAS® procedurerne kaldes testet ofte et Type III test. □

Bemærkning 2 *Det partielle test "korrigerer" for de effekter, der er i modellen på det pågældende trin*

Ved tolkning af teststørrelserne svarende til et givet trin i modelhierarkiet siger man, at man vurderer effekten af det givne led korrigeret for de øvrige effekter, der er i modellen på det pågældende trin. □

Bemærkning 3 *Partielt Wald-test*

Wald-teststørrelsen, der blev defineret i definition 2.11.1 er netop teststørrelsen for det partielle Wald-test for H_1 under H_0. □

Eksempel 2.11.5 *Partielle kvotienttest for modelreduktion Udplantning af blommestiklinger, fortsat*

Vi betragter atter situationen i eksempel 2.4.2

Vi har udgangsmodellen

$$H_0: \quad \eta_{i,j} = \kappa + \alpha_i + \gamma_j$$

med $i = 1, 2$, og $j = 1, 2, 3$, og med $\alpha_2 = 0$ og $\gamma_3 = 0$.

Vi vil nu vurdere delhypoteserne

$$H_1^A: \quad \alpha_1 = \alpha_2 = 0 \Leftrightarrow \eta_{i,j} = \kappa + \gamma_j$$

dvs stiklingens længde er uden betydning for overlevelsesevnen, og

$$H_1^B: \quad \gamma_1 = \gamma_2 = \gamma_3 = 0 \Leftrightarrow \eta_{i,j} = \kappa + \alpha_i$$

dvs stiklingens tykkelse er uden betydning for overlevelsesevnen.

Idet dispersionsparameteren svarende til binomialfordelingen er 1, er teststørrelserne de uskalerede residualdevianser:

$$\begin{aligned} G^2(H_0) &= 1.8854 \quad f = 2 \\ G^2(H_1^A) &= 8.7412 \quad f = 3 \\ G^2(H_1^B) &= 23.2221 \quad f = 4 \end{aligned}$$

Vi får derfor teststørrelserne for de partielle tests ved at beregne differenserne mellem teststørrelserne for modeltilpasning:

Effekt	Teststørrelse	Frihedsgrader
Længde	$G^2(H_1^A \mid H_0) = G^2(H_1^A) - G^2(H_0)$	3-2
Tykkelse	$G^2(H_1^B \mid H_0) = G^2(H_1^B) - G^2(H_0)$	4-2

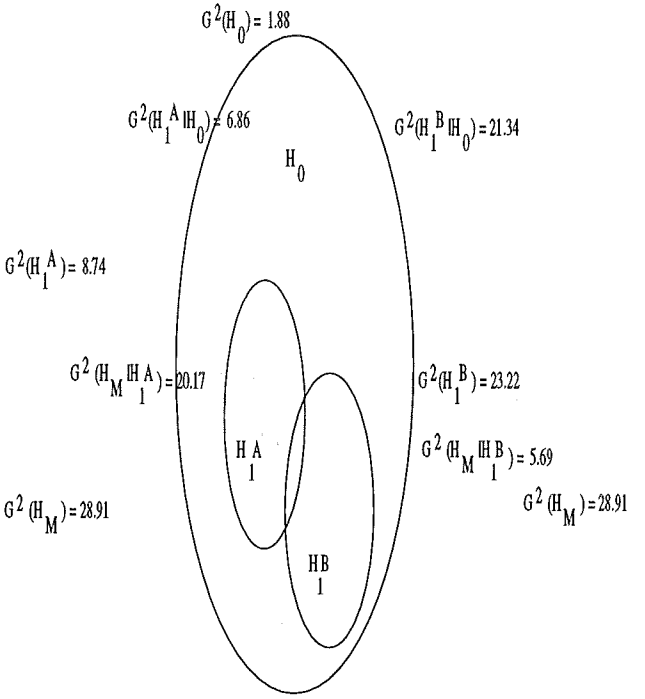
Havde man benyttet menuen `Fit` under SAS® proceduren INSIGHT, havde man fået tabellen:

```
                Type III (LR) Tests
Source              DF       Chi-Sq  Pr > Chi-Sq
LENG            1.0000       6.8558       0.0088
TYK             2.0000      21.3367       0.0001
```

altså netop teststørrelserne for henholdsvis længde, korrigeret for tykkelse (tabel 2.14), og tykkkelse korrigeret for længde (tabel 2.11)

Da begge størrelser er signifikante ved test på et 5 % niveau, kan ingen af leddene fjernes. Begge led er nødvendige til bestemmelse af overlevelsesevnen.

De forskellige test er illustreret i nedenstående figur



□

Eksempel 2.11.6 ***Partielle Wald-test for modelreduktion Udplantning af blommestiklinger, fortsat***

Havde man i det foregående eksempel i stedet ønsket at benytte Wald's teststørrelser, ville man formulere de tilsvarende kontraster:

Kontrasterne svarende til forsvindende længdeeffekt

$$H_1^A: \quad \alpha_1 = \alpha_2 = 0$$

er ganske enkelt vektoren, der udvælger α_1:

$$L_{lengd} = \begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix}.$$

Idet $\widehat{\boldsymbol{\Sigma}}$ er givet ved (2.5.3) har vi

$$\mathbf{L}_{lengd}{}^T \widehat{\boldsymbol{\Sigma}}\, \mathbf{L}_{lengd} = 0.1773$$

og

$$\mathbf{L}_{lengd}{}^T \widehat{\boldsymbol{\beta}} = \widehat{\alpha}_1 = -1.0735 ,$$

sådan at Walds teststørrelse bliver

$$Z = \frac{(-1.0735)^2}{0.1773} = 6.499$$

Da H_1 blot udtrykker et krav til en enkelt parameter, α_1, kunne vi lige så godt have benyttet et test for $\alpha_1 = 0$ svarende til analysen af parameterestimater på side 201. Vi får da også samme værdi af teststørrelsen som på side 201.

Kontrasterne svarende til forsvindende tykkelseseffekt er bestemt ved matricen

$$\mathbf{L}_{tyk} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} ,$$

der udvælger β_1 og β_2. Vi har

$$\mathbf{L}_{tyk}{}^T \widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 1.6059 \\ 2.22058 \end{pmatrix}$$

Den estimerede dispersionsmatrix svarende til denne effekt er

$$\mathbf{L}_{tyk}{}^T \widehat{\boldsymbol{\Sigma}} \mathbf{L}_{tyk} = \begin{pmatrix} 0.2583 & 0.1491 \\ 0.1491 & 0.2846 \end{pmatrix} = \begin{pmatrix} 5.5497 & -2,9072 \\ -2.9072 & 5.0370 \end{pmatrix}^{-1}$$

således at man finder Wald's teststørrelse

$$Z = (\mathbf{L}_{tyk}{}^T \widehat{\boldsymbol{\beta}})^T (\mathbf{L}_{tyk}{}^T \widehat{\boldsymbol{\Sigma}} \mathbf{L}_{tyk}) - 1 \mathbf{L}_{tyk}{}^T \widehat{\boldsymbol{\beta}} = 18.2235 ,$$

der skal sammenlignes med fraktilerne i en $\chi^2(2)$ fordeling.

Wald-teststørrelsen er således i god overensstemmelse med likelihood-kvotient teststørrelsen.

Havde man benyttet menuen `Fit` under SAS® proceduren INSIGHT og specificeret Wald-test, havde man fået tabellen

Analyse af Blomme planter
Type III (Wald) Tests

Source	DF	Chi-Sq	Pr > Chi-Sq
LENG	1.0000	6.4992	0.0108
TYK	2.0000	18.2235	0.0001

□

Der er næppe nogen universelt anvendelig strategi for tilpasning af en passende model til beskrivelse af et givet sæt data $\mathbf{y}$.

Valget af model er sædvanligvis en iterativ proces, hvor forskellige reduktioner af den fulde model vurderes i relation til den behandlede problemstilling ved betragtning af teststørrelsen for modelreduktion, ved vurdering af residualer m.v..

En god kontrol af en model er en validering af modellen på et nyt, uafhængigt datasæt. datasæt.

2.11.6 Total deviansopspaltning svarende til successiv tilføjelse eller fjernelse af led

De fleste af de programsystemer, der benytter sig af modelformler, angiver en total opspaltning af residualdeviansen svarende til den rækkefølge, hvori leddene er anført i modelformlen, begyndende med residualdeviansen svarende til en model alene med et intercept-led, efterfulgt af residualdevianserne svarende til de modeller, der fremkommer ved successiv tilføjelse af leddene i modelformlen.

Lad H_ν^* betegne modellen svarende til det ν'te trin i denne procedure og lad $\widehat{\boldsymbol{\mu}}_\nu$ angive middelværdiestimaterne svarende til denne model. Lad tilsvarende $H_{\nu+1}^*$ med estimaterne $\widehat{\boldsymbol{\mu}}_{\nu+1}$ betegne modellen svarende til det følgende trin, dvs svarende til tilføjelse af leddet $C_{\nu+1}$ i modelformlen.

$$H_M \subset \cdots \subset H_\nu \subset H_{\nu+1} \subset \cdots \subset H_0$$

Under hypotesen $H_{\nu+1}$ er kvotienteteststørrelsen for den hypotese, at koefficienterne til leddet $C_{\nu+1}$ alle er nul, dvs for hypotesen H_ν

$$\begin{aligned} G^2(H_\nu|H_{\nu+1}) &= G^2(H_\nu) - G^2(H_{\nu+1}) \\ &= D^*(\mathbf{y};\widehat{\boldsymbol{\mu}}_\nu) - D^*(\mathbf{y};\widehat{\boldsymbol{\mu}}_{\nu+1}) \\ &= D^*(\widehat{\boldsymbol{\mu}}_{\nu+1};\widehat{\boldsymbol{\mu}}_\nu) \end{aligned} \qquad (2.11.22)$$

Under H_ν vil teststørrelsen approximativt følge en χ^2-fordeling med et antal frihedsgrader, f, der er forskellen mellem dimensionen af underrummet svarende til $H_{\nu+1}^*$ og underrummet svarende til H_ν^*

Størrelsen (2.11.22) måler forøgelsen i modeldeviansen når leddet $C_{\nu+1}$ inddrages i modellen.

I SAS® procedurerne kaldes en sådan additiv opspaltning af deviansen svarende til den minimale model for en sekventiel Type I analyse, og de tilsvarende test, der svarer til en successiv tilføjelse af led i modelformlen kaldes for Type I test.

Eksempel 2.11.7 ***Type-I analyse i SAS og S-plus***

Udplantning af blommestiklinger, fortsat

Hvis man under`Fit`-menuen i SAS® proceduren INSIGHT har specificeret modellen (for binomialfordeling med antalsparameter `ANT` og logit-link):

```
LEV    =   LENG  TYK
```

og man vælger outputoptionen `Type I Tests`, udskrives en tabel af form som i tabel 2.11, hvor leddene i tabellen svarer til øgningen i modeldevians

Tabel 2.11. Eksempel på udskrift af Type I-test, SAS® procedure INSIGHT Leddene tilføjes i rækkefølgen `LENG` og `TYK`.

	Type I (LR) Tests		
Source	DF	Chi-Sq	Pr > Chi-Sq
LENG	1.0000	5.6931	0.0170
TYK	2.0000	21.3367	0.0001

når man bevæger sig op igennem den venstre gren af inklusionsdiagrammet fig 2.8. Rækkefølgen svarer til den rækkefølge, hvori leddene er angivet i modelformlen. Leddet `TYK` i tabellen svarer derfor her til effekten af tykkelse, når der er korrigeret for den eventuelle effekt af længden.

I SAS® proceduren GENMOD fås udskriften, der er vist i tabel 2.12.

Kolonnen `Deviance` angiver deviansen $D(\mathbf{y};\widehat{\boldsymbol{\mu}}_{\nu+1})$ svarende til en model, der indeholder alle led svarende til de foregående linier i tabellen, inklusive leddet $C_{\nu+1}$ svarende til den aktuelle tabellinie.

Tabel 2.12. Eksempel på udskrift af Type I-test, SAS® procedure GENMOD

The GENMOD Procedure

LR Statistics For Type 1 Analysis

Source	Deviance	DF	ChiSquare	Pr>Chi
INTERCEPT	28.9152	0	.	.
LENG	23.2221	1	5.6931	0.0170
TYK	1.8854	2	21.3367	0.0000

Kolonnen `ChiSquare` indeholder teststørrelserne $G^2(H_\nu|H_{\nu+1})$ (2.11.22) svarende til tilføjelse af den effekt, der er angivet i den pågældende linie.

I programsystemet S-plus vil kommandoen

```
< blomme <- glm(cbind(lev,dod) ~ leng + tyk,
   family = binomial(   logit))
```

efterfulgt af kaldet

```
<  anova(blomme, test=  "Chisq")
```

bevirke udskrift af en deviansanalysetabel i analogi med udskriften fra GENMOD. Udskriften er vist i tabel 2.13.

Kolonnen `Deviance Resid.` indeholder netop teststørrelserne $G^2(H^*_\nu|H^*_{\nu+1})$ (2.11.22) og kolonnen `Pr(Chi)` angiver den sandssynlighedsmasse, der ligger til højre for den observerede værdi af teststørrelsen.

Endelig angiver kolonnen `Resid. Dev` goodness of fit-teststørrelsen $G^2(H_{\nu+1})$ svarende til en model, der indeholder alle led svarende til de linier i tabellen, inklusive leddet $C_{\nu+1}$ svarende til den aktuelle tabellinie. I det aktuelle tilfælde angiver bidraget svarende til `LENG` indflydelsen af stiklingens længde, dvs et test for hvorvidt koefficienterne α_i, $i = 1, 2$ i modellen

$$H_1^*: \quad \eta_{i,j} = \kappa + \alpha_i \tag{2.11.23}$$

Tabel 2.13. Eksempel på udskrift af Type I-test, S-plus objekt glm

```
Analysis of Deviance Table

Binomial Model

Response: cbind(lev,dod)

Terms added sequentially (first to last)

     Df Deviance Resid. Df  Resid. Dev     Pr(Chi)
NULL                    5     28.9152
leng  1   5.6931         4     23.2221   0.0170320
 tyk  2  21.3367         2      1.8854   0.0000233
```

kan antages at være nul. Bidraget svarende til `TYK` er teststørrelsen for, hvorvidt koefficienterne γ_j, $j = 1, 2, 3$ i modellen

$$H_2^* : \quad \eta_{i,j} = \kappa + \alpha_i + \gamma_j$$

kan antages at være nul.

Havde vi i stedet angivet modelformlen som

```
LEV     /       ANT   =   TYK  LENG
```

dvs med effekten af tykkelse først, da ville opspaltningen modsvare hierarkiet af hypoteser svarende til den højre gren i inklusionsdiagrammet fig 2.8.

I menuen `Fit` under SAS® proceduren INSIGHT ville type I testene svarende til denne variant af modelformlen blive som anført i tabel 2.14, og vi ville sige, at teststørrelsen svarende til linien `LENG` måler effekten af længde, korrigeret for tykkelse.

Vi bemærker, at teststørrelsen svarende til effekten af længde, korrigeret for tykkelse i tabel 2.14 er anderledes end teststørrelsen svarende til effekten af længde i tabel 2.11.

Tabel 2.14. Eksempel på udskrift af Type I-test, SAS® procedure INSIGHT
Leddene tilføjes i rækkefølgen TYK og LENG

	Type I (LR) Tests		
Source	DF	Chi-Sq	Pr > Chi-Sq
TYK	2.0000	20.1740	0.0000
LENG	1.0000	6.8558	0.0088

Teststørrelsen for længde, korrigeret for tykkelse, i tabel 2.14 svarer til et test af, hvorvidt koefficienterne α_i, $i = 1, 2$ i modellen

$$H_2^{\square}: \quad \eta_{i,j} = \kappa + \gamma_j + \alpha_i$$

kan antages at være nul.
Teststørrelsen for hvorvidt tykkelsen har betydning for overlevelseschancen afhænger således af, om vi ved testet ønsker at tilgodese den eventuelle effekt af længde eller ej. □

I deviansopspaltningen svarende til en type 1 analyse er der ved testet for en given effekt (led i modelformlen) korrigeret for de eventuelle effekter svarende til de foregående led.

Når man fjerner led i modelformlen, skal leddene derfor fjernes "fra neden og opefter".

2.11.7 Successiv testning ved estimation af dispersionsparameter

I situationer, hvor det er nødvendigt at estimere dispersionsparameteren ved brug af residualdeviansen jvf sætning 2.6.4, må man tage udgangspunkt i udgangsmodellen H_0 og antage, at denne model er sand.

Ved en successiv fjernelse af led må man da benytte F-test for de skalerede residualdevianser i analogi med sætning 2.11.4. Sædvanligvis vil man da reestimere dispersionsparameteren efter hver reduktion af modellen.

fil glm5.tex 1998-02-08

2.12 Vekselvirkning

Betragt en model med to forklarende faktorvariable, og antag at den additive model (2.9.5)

$$\eta_{p,q} = \alpha_p + \beta_q \tag{2.12.1}$$

gælder.

Det følger da, at kontraster mellem niveauer af den ene faktor ikke afhænger af niveauet af den anden faktor. Man siger, at der ikke er vekselvirkning imellem de to inddelinger.

Omvendt, hvis kontraster mellem niveauer af den ene faktor afhænger af niveauet af den anden faktor siger man, at der er vekselvirkning.

Vekselvirkning kaldes undertiden interaction eller matrixeffekt.

Betragter vi den differentielle effekt

$$\Delta^A_{p,p';q} \stackrel{\text{DEF}}{=} \eta_{p,q} - \eta_{p',q} \tag{2.12.2}$$

mellem niveau p og p' af faktoren A, ser vi, at under den additive model (2.12.1) gælder:

$$\Delta^A_{p,p';q} = \alpha_p + \beta_q - (\alpha_{p'} + \beta_q) = \alpha_p - \alpha_{p'} \ ,$$

der ikke afhænger af q.

Sætning 2.12.1 ***Konstant differentiel effekt indebærer additiv model uden vekselvirkning***

Betragt en model med to faktorer. Hvis den differentielle effekt $\Delta^A_{p,p';q}$ ikke afhænger af niveauet q, da gælder den additive model (2.12.1)

Bevis:

Hvis den differentielle effekt $\Delta^A_{p,p';q}$ ikke afhænger af q, da kan den udtrykkes som

$$\Delta^A_{p,p';q} = \eta_{p,q} - \eta_{p',q} = \delta(p,p') \tag{2.12.3}$$

hvor $\delta(p, p')$ ikke afhænger af q.

Sætter vi $p' = 1$ har vi af (2.12.3), at

$$\eta_{p,q} = \eta_{1,q} + \delta(p, 1)$$

altså netop på formen (2.13.2). □

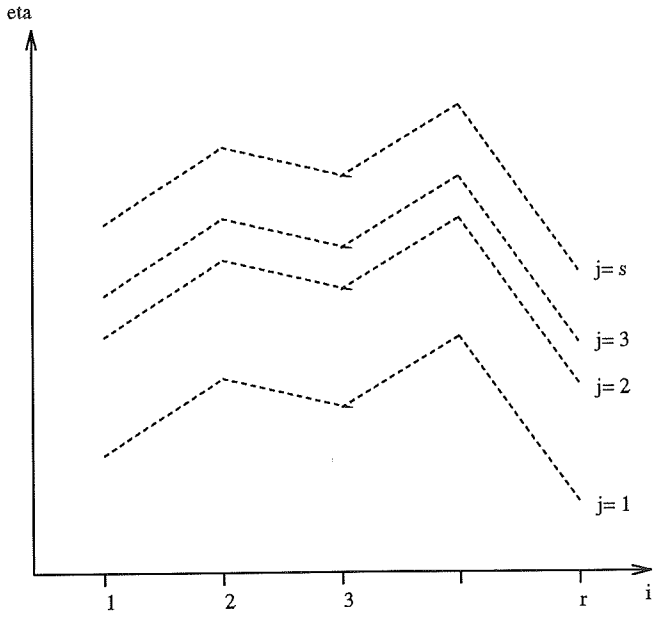
Grafisk kontrol, profilplot

Antagelsen om forsvindende vekselvirkning kan kontrolleres grafisk ved et såkaldt profilplot, dvs en grafisk afbildning af de transformerede værdier $g(y_{i,q})$ mod f.eks. rækkeindeks i. Profilplottet afspejler den egenskab, at kontrasten mellem to niveauer af én faktor ikke afhænger af niveauet af den anden faktor. Profilen af faktor A svarende til niveau j for faktor B er den stykkevis lineære kurve, der forbinder punkterne $g(y_{1,q}), g(y_{1,q}), \ldots, g(y_{r,q})$. Under den additive model vil profilerne af faktor A svarende til forskellige niveauer af faktor B være parallelle.

Modellen er symmetrisk i i og j, og profilplottet kan derfor lige så godt tegnes med $\eta_{i,j}$ som funktion af j. De to varianter af profilplottet indeholder samme information, og det er derfor i princippet ligegyldigt, hvilken af de to man vælger. Hvis den ene faktor repræsenterer tid eller en anden variabel med en klar ordning, vil det ofte være naturligt at tegne denne faktor ud ad den vandrette akse.

Modeller uden vekselvirkning er interessante, fordi række- og søjlesummerne indeholder al informationen. I en model uden vekselvirkning kan man meningsfuldt udtale sig om forskelle mellem række- eller søjleniveauer - uden at være nødt ti at specificere en bestemt celle i tabellen.

Hvis der var vekselvirkning, vil et udsagn om en forskel på eksempelvis to rækker afhænge af den specifike vægtning af søjlerne.



Figur 2.9. Eksempel på profilplot

$$\eta_{i,j} = \alpha_i + \beta_j$$

Eksempel 2.12.1 *Temperatur i termosikring*

Antag at man af hensyn til design af en termosikring, der sidder i et bestemt elektrisk apparat, er interesseret i driftstemperaturen i apparatet.

Man afprøver derfor apparatet, blandt andet ved forskellige spændinger og forskellige omgivelsestemperaturer.

Antag, at man får resultatet:

Spænding	Omgivelsestemperatur	
	25°C	45°C
220 Volt	94.5	109.9
210 Volt	99.3	120.1

Man ser, at hvis spændingen er lavere end standardspændingen 220 Volt, forøges driftstemperaturen. Men forøgelsen afhænger af omgivelsestemperaturen. Ved 25°C er forøgelsen 5°C, men ved 45°C er forøgelsen ca 10°C.

Der er således vekselvirkning mellem effekten på driftstemperaturen af spænding og omgivelsestemperatur. □

Eksempel 2.12.2 ***Kvartalsvise uheldstal for motorkøretøjer***

Nedenstående tabel viser de kvartalsvise uheldstal for uheld med personskade med motorkøretøjer for uheldskategorien "møde" i dagslys for ikke-spirituspåvirkede førere for årene 1987-89 i Danmark

Indeks i	Antal uheld y_i	år z_{i1}	kvartal z_{i2}
1	128	1987	1
2	95	1987	2
3	100	1987	3
4	75	1987	4
5	94	1988	1
6	85	1988	2
7	119	1988	3
8	71	1988	4
9	82	1989	1
10	81	1989	2
11	98	1989	3
12	72	1989	4

Vi opstiller tabellen som en todimensional tabel

Antal uheld med personskade i årene 1987-1989
Kategori "møde"

	Kvartal				
År	1	2	3	4	Ialt
1987	128	95	100	75	398
1988	94	85	119	71	369
1989	82	81	98	72	333
Ialt	304	261	317	218	1100

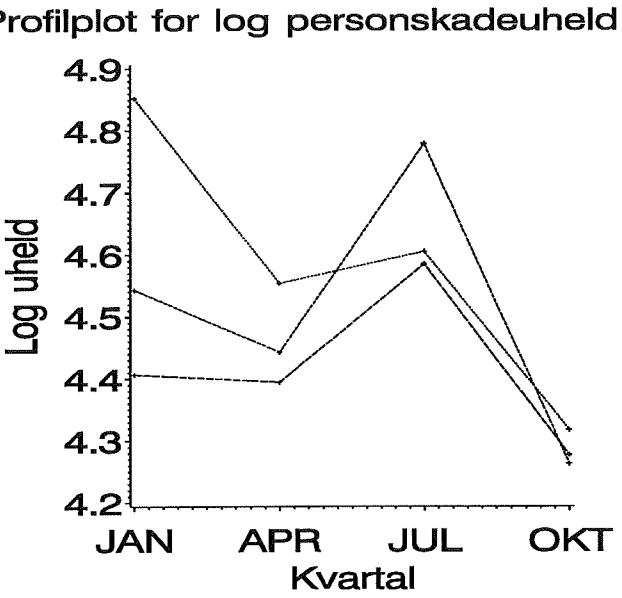
Man kunne forestille sig at modellere uheldstallene ved Poisson-fordelte variable med middelværdi μ_{pq}. Den kanoniske link er $\eta_{pq} = \log(\lambda_{pq})$.

For at vurdere, hvorvidt det vil være rimeligt at antage en additiv model

$$\eta_{pq} = \alpha_p + \beta_q$$

for den valgte linkfunktion, tegner man et profilplot for $\log(y_{pq})$.

Profilplottet er vist i figur 2.10.



Figur 2.10. Profilplot for logaritmen til kvartalsvise uheldstal i årene 1987-1989

Uheldstal for samme år er forbundet med rette linier

Liniestykkerne er med god tilnærmelse parallelle.

Den additive model for η_{pq} modsvares af en multiplikativ model for $\mu_{pq} = \mathsf{E}[Y_{pq}]$, dvs en model af formen

$$\mu_{pq} = \alpha_p^* \beta_q^* \tag{2.12.4}$$

Modellen uden vekselvirkning for den kanoniske parameter svarer altså til at der i alle årene er det samme forhold mellem uheldstallene i to kvartaler. (Den differentielle effekt $\Delta^A_{p,p';q}$ for η_{pq} oversættes jo i en kvotient for μ_{pq}).

Et udsagn "Der er 50% flere ulykker i juli kvartal, end i oktober kvartal" er således gyldigt såvel for alle årene under ét, som for de enkelte år.

I eksempel 2.13.2 vil vi vende tilbage til analysen af denne model. □

Eksempel 2.12.3 ***Overnatninger på campingpladser***

Nedenstående tabel viser antallet af overnatninger (i 1000 overnatninger) på danske campingpladser i 1994 for Bornholms og Fyns amt fordelt på udvalgte nationaliteter.

	Nationalitet			Ialt
Amt	Danmark	Sverige	Tyskland.	
Bornholm	134	62	108	304
Fyn	767	8	186	961
Ialt	901	70	294	1265

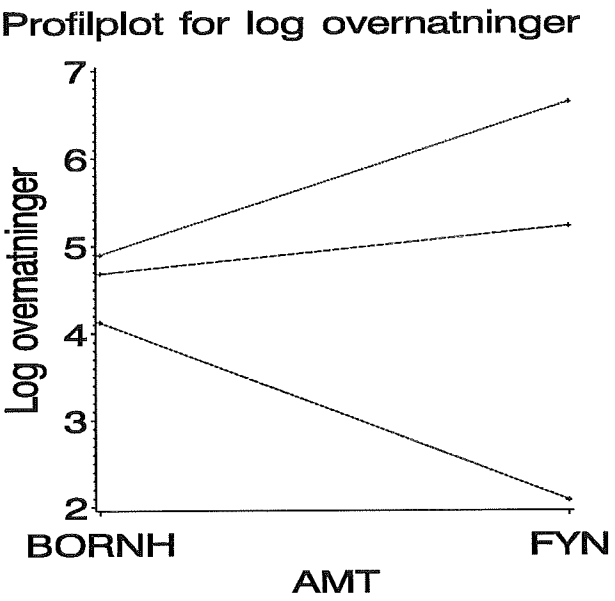
Man kunne forestille sig at modellere antallet af overnatninger ved Poisson-fordelte variable med middelværdi μ_{pq}, hvor $p = 1, 2$ angiver amtet, og $q = 1, 2, 3$ angiver nationaliteten.

Den kanoniske link er $\eta_{pq} = \log(\mu_{pq})$ og for at vurdere en eventuel vekselvirkning i den kanoniske parameter tegner vi derfor et profilplot for $\log(y_{pq})$.

Profilplottet er vist i figur 2.11.

Linierne er langt fra at være parallelle og man kan derfor ikke opretholde en hypotese om forsvindende vekselvirkning for den kanoniske parameter. Der er altså ikke multiplikativitet i middelværdiparameteren.

Udsagnet "der er tre gange så mange danske overnatninger som tyske" har således kun mening for de to amter under ét. I Bornholms amt er der næsten lige mange danske og tyske overnatninger, mens der i Fyns amt er næsten fire gange så mange danske som tyske overnatninger. □



Figur 2.11. Profilplot for logaritmen til antallet af overnatninger af forskellige nationaliteter for Fyns og Bornholms amter

Overnatninger for samme nationalitet er forbundet med rette linier

2.13 Tosidig inddeling

Som et specialtilfælde af den generelle teori vil vi nu diskutere analysen af en generaliseret lineær model med to forklarende variable, der begge er kvalitative.

Vi vil således betragte et sæt af observationer svarende til en klassifikation af indeksmængden $I = \{1, 2, \ldots, r, \ldots\}$ efter to kriterier, A (rækker) og B (søjler).

Et sådant observationssæt organiseres ofte på tabelform

$$\begin{pmatrix} Y_{11} & Y_{12} & \dots & Y_{1s} \\ Y_{21} & Y_{22} & \dots & Y_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{r1} & Y_{r2} & \dots & Y_{rs} \end{pmatrix} \tag{2.13.1}$$

med r rækker og s søjler.

Der gælder da

Sætning 2.13.1 ***Estimation og test ved tosidig inddeling***

Betragt observationssættet (2.13.1) og antag, at $Y_{11}, \dots, Y_{r,s}$ er indbyrdes uafhængige, hvis fordelinger tilhører en eksponentiel dispersionsparameterfamilie med samme variansfunktion $V(\mu)$ og samme dispersionsparameter σ^2 og antag at link-funktionen er givet ved $\eta = g(\mu)$.

Betragt den additive model

$$H_0: \quad \eta_{p,q} = \alpha_p + \gamma_q; \;\; p = 1, \dots, r; \;\; q = 1, \dots, s \tag{2.13.2}$$

svarende til at der ikke er vekselvirkning imellem række- og søjleeffekterne.

Betragt endvidere delhypotesen

$$H_1^A: \eta_{p,q} = \gamma_q; \;\; p = 1, \dots, r; \;\; q = 1, \dots, s \tag{2.13.3}$$

svarende til forsvindende rækkevirkning ($\alpha_1 = \cdots = \alpha_r = 0$)

og betragt endelig den minimale model

$$H_M: \eta_{p,q} = \kappa; \;\; p = 1, \dots, r; \;\; q = 1, \dots, s \tag{2.13.4}$$

svarende til fuldstændig homogenitet

Hypotesekæden $H_M \subset H_1^A \subset H_0 \subset H_F$ svarer til deviansopspaltningen

Variationskilde	f	Devians	Goodness of fit fortolkning
Mellem søjler	$s-1$	$D\big(\boldsymbol{\mu}(\widehat{\widehat{\gamma}});\boldsymbol{\mu}(\widehat{\kappa})\big)$	$G^2(H_M\|H_1^A)$
Mellem rækker korrigeret for søjler	$r-1$	$D\big(\boldsymbol{\mu}(\widehat{\alpha},\widehat{\gamma});\boldsymbol{\mu}(\widehat{\widehat{\gamma}})\big)$	$G^2(H_1^A\|H_0)$
Vekselvirkning	$(r-1)(s-1)$	$D\big(\mathbf{y};\boldsymbol{\mu}(\widehat{\alpha};\widehat{\gamma})\big)$	$G^2(H_0)$
Total	$rs-1$	$D^*(\mathbf{y};\boldsymbol{\mu}(\widehat{\kappa}))$	$G^2(H_M)$

hvor $\boldsymbol{\mu}(\widehat{\alpha},\widehat{\gamma})$ angiver de fittede værdier under additivitetshypotesen H_0, $\boldsymbol{\mu}(\widehat{\widehat{\gamma}})$ angiver de fittede værdier under H_1^A, og endelig $\boldsymbol{\mu}(\widehat{\kappa})$ angiver de fittede værdier under homogenitetshypotesen H_M.

Såfremt dispersionsparameteren σ^2 er kendt, kan hypotesen om forsvindende vekselvirkning testes ved den (evt skalerede) residualdevians

$$G^2(H_0) = D^*\big(\mathbf{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})\big) \ .$$

Under H_0 følger $D^*\big(\mathbf{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})\big)$ approximativt en $\chi^2((r-1)(s-1))$-fordeling. Hypotesen forkastes for store værdier af residualdeviansen.

Under antagelse af additivitet kan hypotesen H_1^A om forsvindende rækkevirkning testes ved den (evt skalerede) residualdevians

$$G^2(H_1^A|H_0) = D^*\big(\boldsymbol{\mu}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}); \boldsymbol{\mu}(\widehat{\widehat{\boldsymbol{\gamma}}})\big) \ .$$

Under H_0 følger $D^*\big(\boldsymbol{\mu}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}); \boldsymbol{\mu}(\widehat{\widehat{\boldsymbol{\gamma}}})\big)$ approximativt en $\chi^2(r-1)$-fordeling. Hypotesen forkastes for store værdier af $D^*\big(\boldsymbol{\mu}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}); \boldsymbol{\mu}(\widehat{\widehat{\boldsymbol{\gamma}}})\big)$.

Under forudsætning af forsvindende rækkevirkning kan Hypotesen H_M om total homogenitet testes ved residualdeviansen

$$G^2(H_M|H_1^A) = D^*\big(\boldsymbol{\mu}(\widehat{\widehat{\boldsymbol{\gamma}}}); \boldsymbol{\mu}(\widehat{\boldsymbol{\kappa}})\big) \ .$$

Under H_M følger $D^*\big(\boldsymbol{\mu}(\widehat{\widehat{\boldsymbol{\gamma}}}); \boldsymbol{\mu}(\widehat{\boldsymbol{\kappa}})\big)$ approximativt en $\chi^2(s-1)$-fordeling. Hypotesen forkastes for store værdier af $D^*\big(\boldsymbol{\mu}(\widehat{\widehat{\boldsymbol{\gamma}}}); \boldsymbol{\mu}(\widehat{\boldsymbol{\kappa}})\big)$

Se bemærkning 1 på side 326 for test af søjlevirkning før rækkevirkning

Såfremt dispersionsparameteren σ^2 ikke er kendt, og H_0 kan antages opfyldt, kan man estimere dispersionsparameteren σ^2 ved

$$\widehat{\sigma}^2 = \frac{D\big(\mathbf{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})\big)}{(r-1)(s-1)} \tag{2.13.5}$$

Hypotesen H_1^A kan da testes ved teststørrelsen

$$F_1 = \frac{D\big(\boldsymbol{\mu}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}); \boldsymbol{\mu}(\widehat{\widehat{\boldsymbol{\gamma}}})\big)/(r-1)}{D\big(\mathbf{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})\big)/[(r-1)(s-1)]} \tag{2.13.6}$$

Under H_1^A følger F_1 approximativt en $\mathrm{F}\big(r-1, (r-1)(s-1)\big)$-fordeling. Hypotesen forkastes for store værdier af F_1.

Såfremt H_1^A er sand, kan man teste hypotesen H_M om total homogenitet ved teststørrelsen

$$F_2 = \frac{D(\boldsymbol{\mu}(\widehat{\widehat{\gamma}}); \boldsymbol{\mu}(\widehat{\kappa}))/(s-1)}{D(\mathbf{y}; \boldsymbol{\mu}(\widehat{\alpha}, \widehat{\gamma}))/[(r-1)(s-1)]} \tag{2.13.7}$$

Under H_M følger F_2 approximativt en $\mathrm{F}\big(s-1, (r-1)(s-1)\big)$-fordeling.

Parametrene må estimeres ved den generelle ligning (2.5.1).

Såfremt linkfunktionen $g(\cdot)$ er den kanoniske link, kan de fittede værdier dog bestemmes ved middelværdiligningen (2.5.6).

Estimaterne under H_0 bestemmes af

$$\begin{aligned} y_{p\cdot} = \sum_{q=1}^{s} w_{p,q} y_{p,q} &= \sum_{q=1}^{s} w_{p,q} \mu(\alpha_p + \gamma_q); \quad p = 1, 2, \ldots, r \\ y_{\cdot q} = \sum_{p=1}^{r} w_{p,q} y_{p,q} &= \sum_{p=1}^{r} w_{p,q} \mu(\alpha_p + \gamma_q); \quad q = 1, 2, \ldots, s \end{aligned} \tag{2.13.8}$$

Estimaterne under H_1^A er blot søjlegennemsnittene

$$\mu(\widehat{\widehat{\gamma}}_q) = \bar{y}_{\cdot q} = \sum_{p=1}^{r} w_{p,q} y_{p,q} \Big/ \sum_{p=1}^{r} w_{p,q}\,, \quad q = 1, 2, \ldots, s \tag{2.13.9}$$

Endelig fås estimaterne under homogenitetshypotesen som det fælles gennemsnit

$$\mu(\widehat{\kappa}) = \bar{y}_{\cdot\cdot} = \sum_{p=1}^{r}\sum_{q=1}^{s} w_{p,q} y_{p,q} \Big/ \sum_{p=1}^{r}\sum_{q=1}^{s} w_{p,q}\,, \tag{2.13.10}$$

hvor $w_{p,q}$ angiver de eventuelle vægte. For en uvægtet model er $w_{p,q} = 1$.

Bevis:
Hypotesen H_0 definerer en $r+s-1$ dimensional delmodel af den fulde model. Parametriseringen (2.13.2) er nemlig ikke injektiv. Hvis sættet

$$\{\alpha_p, \gamma_q\}, \quad p = 1, \ldots, r;\ q = 1, \ldots, s$$

tilfredsstiller hypotesen, da vil også sættet

$$\{\alpha_p + c, \gamma_q - c\}, \quad p = 1, \ldots, r;\ q = 1, \ldots, s$$

med et vilkårligt c tilfredsstille hypotesen, svarende til at modelmatricen for den additive model kun udspænder et $r+s-1$ dimensionalt underrum i $\mathbb{R}^{rs}$

Selv om denne overparametrisering naturligvis er af betydning ved en eventuel estimation af de individuelle værdier α_p og γ_q, betyder den ikke ikke noget for bestemmelsen af de fittede værdier, da alle kombinationer $\{\alpha_p + c, \gamma_q - c\}, p = 1, \ldots, r;\ q = 1, \ldots, s$ vil give anledning til samme værdi $\eta_{p,q}$ efter (2.13.2).

Da modellen H_1^A er en delmodel af H_0, har vi jvf. sætning 2.11.2 deviansopspaltningen

$$D(\mathbf{y}; \boldsymbol{\mu}(\widehat{\widehat{\gamma}})) = D(\mathbf{y}; \boldsymbol{\mu}(\widehat{\alpha}, \widehat{\gamma})) + D(\boldsymbol{\mu}(\widehat{\alpha}, \widehat{\gamma}); \boldsymbol{\mu}(\widehat{\widehat{\gamma}}))$$

svarende til deviansanalyseskemaet

Variationskilde	f	Devians	Goodness of fit fortolkning
Mellem rækker korrigeret for søjler	$r-1$	$D(\boldsymbol{\mu}(\widehat{\alpha}, \widehat{\gamma}); \boldsymbol{\mu}(\widehat{\widehat{\gamma}}))$	$G^2(H_1^A \mid H_0)$
Vekselvirkning	$(r-1)(s-1)$	$D(\mathbf{y}; \boldsymbol{\mu}(\widehat{\alpha}, \widehat{\gamma}))$	$G^2(H_0)$
Total, omkring søjlemidler	$s(r-1)$	$D(\mathbf{y}; \boldsymbol{\mu}(\widehat{\widehat{\gamma}}))$	$G^2(H_1^A)$

Resultaterne følger da af sætning 2.11.2. □

Bemærkning 1 ***Deviansopspaltning ved test af søjlevirkning før rækkevirkning***

Ved ombytning af søjler og række ri ovensstående sætning finder man, at deviansanalyseskemaet svarende til test af hypotesekæden $H_M \subset H_1^B \subset H_0 \subset H_F$ med

$$H_1^B \; : \; \eta_{p,q} = \alpha_p \; ; \quad p = 1, \ldots, r; \; q = 1, \ldots, s$$

dvs til test af forsvindende søjlevirkning før test af fuldstændig homogenitet er

Variationskilde	f	Devians	Goodness of fit fortolkning
Mellem rækker	$r-1$	$D(\boldsymbol{\mu}(\widehat{\widehat{\boldsymbol{\alpha}}}); \boldsymbol{\mu}(\widehat{\kappa}))$	$G^2(H_M \mid H_1^B)$
Mellem søjler korrigeret for rækker	$s-1$	$D(\boldsymbol{\mu}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}); \boldsymbol{\mu}(\widehat{\widehat{\boldsymbol{\alpha}}}))$	$G^2(H_1^B \mid H_0)$
Vekselvirkning	$(r-1)(s-1)$	$D(\mathbf{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}))$	$G^2(H_0)$
Total	$rs-1$	$D(\mathbf{y}; \boldsymbol{\mu}(\widehat{\kappa}))$	$G^2(H_M)$

hvor $\boldsymbol{\mu}(\widehat{\widehat{\boldsymbol{\alpha}}})$ angiver de fittede værdier under H_1^B

For den kanoniske link bliver estimaterne under H_1^B netop rækkegennemsnittene

$$\mu(\widehat{\widehat{\alpha}}_p) = \bar{y}_{\cdot p} = \sum_{q=1}^{s} w_{p,q} y_{p,q} \Big/ \sum_{q=1}^{s} w_{p,q} \; , \qquad p = 1, 2, \ldots, r \tag{2.13.11}$$

Forskellen på de to fremgangsmåder er illustreret i nedenstående inklusionsdiagram, der illustrerer, at den additive model H_0 deles i to kæder af

successivt underordnede modeller.

$$H_M \subset \begin{matrix} H_1^A \\ \\ H_1^B \end{matrix} \subset H_0 \subset H_F$$

□

I eksemplerne 2.4.2 og 2.11.3 behandlede vi et eksempel på tosidig inddeling ved binomialfordelte observationer og kanonisk link.

Vi vil nu diskutere den tilsvarende situation for Poisson-fordelingen

Eksempel 2.13.1 ***Tosidig inddeling af Poisson-fordelte observationer ved kanonisk link***

Som et eksempel på anvendelsen af ovenstående sætning vil vi angive løsningen svarende til situationen $Y_{p,q} \in \mathrm{P}(\mu_{p,q})$ med den kanoniske link, $\eta = g(\mu) = \ln(\mu)$

Vi får da under hypotesen

$$H_0 : \eta_{p,q} = \alpha_p + \gamma_q ,$$

at $\mu_{p,q} = \exp(\alpha_p + \gamma_q)$, dvs modellen er en multiplikativ model i middelværdiparameteren μ.

Under H_0 får man af (2.13.8), at de fittede værdier bestemmes ved

$$\hat{\mu}_{p,q} = \frac{y_{p\cdot} y_{\cdot q}}{y_{\cdot\cdot}} \tag{2.13.12}$$

med

$$y_{\cdot\cdot} = \sum_{p=1}^{r} \sum_{q=1}^{s} y_{p,q}$$

Under

$$H_1^A : \mu_{p,q} = \exp(\gamma_q)$$

finder vi af (2.13.9)

$$\hat{\mu}_{p,q}^A = \bar{y}_{\cdot q} = \sum_{p=1}^{r} y_{p,q}/r \tag{2.13.13}$$

og under

$$H_1^B : \mu_{p,q} = \exp(\alpha_p)$$

finder vi af (2.13.11), at

$$\widehat{\mu}_{p,q}^B = \bar{y}_{p\cdot} = \sum_{q=1}^{s} y_{p,q}/s \tag{2.13.14}$$

og endelig under

$$H_M : \mu_{p,q} = \exp(\kappa)$$

fås af (2.13.10), at

$$\widehat{\mu}_{p,q}^0 = \bar{y}_{\cdot\cdot} = \sum_{p=1}^{r}\sum_{q=1}^{s} y_{p,q}/(rs) \tag{2.13.15}$$

□

Eksempel 2.13.2 ***Kvartalsvise uheldstal for motorkøretøjer Poisson tosidet inddeling***

Vi betragter atter data fra eksempel 2.12.2.

Vi havde den todimensionale tabel

Antal uheld med personskade i årene 1987-1989
Kategori "møde"

	Kvartal				
År	1	2	3	4	Ialt
1987	128	95	100	75	398
1988	94	85	119	71	369
1989	82	81	98	72	333
Ialt	304	261	317	218	1100

I eksempel 2.12.2 foretog vi en grafisk vurdering af den multiplikative hypotese H_0. Vi vil nu foretage et numerisk test for denne hypotese i overensstemmelse med sætning 2.13.1.

De fittede værdier under H_0 fås af (2.13.12). Værdierne er angivet i nedenstående skema

Estimerede kvartalsvise uheldstal $\widehat{\mu}_{p,q}$ med personskade under den multiplikative Poisson-model
Kategori "møde", årene 1987-1989

År	Kvartal				Ialt
	1	2	3	4	
1987	109.99	94.43	114.70	78.88	398
1988	101.98	87.55	106.34	73.13	369
1989	92.03	79.01	95.96	65.99	333
Ialt	304	261	317	218	1100

De tilsvarende deviansresidualer $r_D(y_{p,q};\widehat{\mu}_{p,q})$ fås ved at benytte tabellen over enhedsdevianser, tabel 2.3 (side 144. Deviansresidualerne er angivet i nedenstående tabel

Deviansresidualer $r_D(y_{p,q};\widehat{\mu}_{p,q})$ for kvartalsvise uheldstal svarende til den multiplikative Poisson-model
Kategori "møde", årene 1987-1989

År	Kvartal			
	1	2	3	4
1987	1.673	0.058	-1.403	-0.440
1988	-0.801	-0.274	1.205	-0.250
1989	-1.065	0.223	0.207	0.728

Til trods for enkelte store residualer finder man, at teststørrelsen for den multiplikative model imod den fulde model, $G^2(H_0) = \sum r_D(y_p;\widehat{\mu}_p)^2 = 8.953$ ikke er ekstrem ved sammenligning med fraktilerne i en $\chi^2(6)$-fordeling. Man har således $\chi^2(6)_{0.85} = 9.45$

Der er således ikke nogen grund til at afvise den multiplikative model.

Vi formulerer nu hypotesen $H_1^A : \alpha_1 = \alpha_2 = \alpha_3 = 0$.

Under H_1^A fås jvf (2.13.13) de fittede værdier

$$\widehat{\mu}^A_{p,q} = \bar{y}_{\cdot q} = \sum_{p=1}^{3} y_{p,q}/3 \ , p = 1,2,3;\ q = 1,2,3,4$$

nemlig de gennemsnitlige kvartalsværdier (midlet over år).

Deviansresidualerne svarende til H_1^A er angivet i nedenstående skema:

Deviansresidualer $r_D(y_{p,q}; \widehat{\mu}^A_{p,q})$ for kvartalsvise uheldstal med personskade svarende til modellen alene med kvartalseffekt
Kategori "møde", årene 1987-1989

	Kvartal			
År	1	2	3	4
1987	2.544	0.845	-0.556	0.272
1988	-0.738	-0.215	1.271	-0.196
1989	-1.987	-0.651	-0.755	-0.078

Man finder residualdeviansen svarende til H_1^A som kvadratsummen af deviansresidualerne, $G^2(H_1^A) = D(\mathbf{y}; \widehat{\boldsymbol{\mu}}^A) = 14.76$ med $f = 8$.

Afvigelsen mellem H_0 og H_1^A har da residualdeviansen

$$G^2(H_1^A|H_0) = D(\widehat{\boldsymbol{\mu}}; \widehat{\boldsymbol{\mu}}^A) = 14.76 - 8.95 = 5.81$$

Størrelsen skal sammenlignes med $\chi^2(2)$-fordelingen. Da $\chi^2(2)_{0.95} = 5.99$ overstiger vores teststørrelse ikke 95 %-fraktilen, og hypotesen kan således ikke afvises ved test på et 5 % niveau.

Vi bemærker, at vi lige så godt kunne have udregnet $G^2(H_1|H_0)$ direkte ved at bestemme devianserne $d(\widehat{\mu}_{i,q}; \widehat{\mu}^A_{i,q})$ mellem $\widehat{\mu}$ og $\widehat{\mu}^A$.

Vi opstiller nu til slut hypotesen H_M om total homogenitet

Under H_M finder vi jvf (2.13.15) estimatet for den fælles middelværdi:

$$\widehat{\mu}^0 = \sum_{p=1}^{3}\sum_{q=1}^{4} y_{p,q}/12 = 91.67$$

Deviansresidualerne svarende til H_M er angivet i nedenstående skema:

Deviansresidualer $r_D(y_{p,q}; \widehat{\mu}^0_{p,q})$ for kvartalsvise uheldstal med personskade svarende til modellen med fuldstændig homogenitet
Kategori "møde", årene 1987-1989

	Kvartal			
År	1	2	3	4
1987	3.578	0.346	0.858	-1.798
1988	0.243	-0.705	2.728	-2.2480
1989	-1.028	-1.137	0.654	-2.135

Man finder deviansen $G^2(H_M) = D(\mathbf{y}; \widehat{\boldsymbol{\mu}}^0) = 37.28$ med $f = 11$ frihedsgrader.

Da de undersøgte modeller danner en kæde af underordnede modeller, kan vi bestemme teststørrelsen $G^2(H_M|H_1^A)$ ved subtraktion sådan at vi får deviansanalyseskemaet

Variationskilde	f	Devians	middeldevians
Kvartal	3	22.52	7.507
År, kontrolleret for kvt	2	5.81	2.905
Residual	6	8.953	1.492
Total	11	37.28	3.389

Da $\chi^2(3)_{0.95} = 7.81$ må hypotesen om total homogenitet klart afvises. Der er således forskel på uheldsraterne i de forskellige kvartaler, men uheldsraterne ændrer sig ikke i de betragtede år.

Som før kunne vi have valgt at bestemme teststørrelsen $G^2(H_M|H_1^A)$ direkte ved at betragte de tilsvarende deviansbidrag $d(\widehat{\mu}_{i,q}^A; \widehat{\mu}_{i,q}^0)$. Disse bidrag ville kunne belyse, hvilke kvartaler, der eventuelt var særligt ekstreme. □

Bemærkning 1 Uafhængighed mellem estimater ved Poisson-fordeling og kanonisk link

I almindelighed vil estimaterne $\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_m$ være indbyrdes afhængige, også selv om søjlerne i modelmatricen $\mathbf{X}$ er ortogonale. Dette skyldes den reskalering af de enkelte observationer svarende til $\mathbf{W}(\boldsymbol{\beta})^{1/2}$, der principielt foretages under estimationen (jvf lemma 2.5.4 på side 207).

I normalfordelingstilfældet med den kanoniske link bliver $\mathbf{W}(\boldsymbol{\beta})$ dog netop enhedsmatricen, og ortogonalitetsegenskaber ved modelmatricen overføres til uafhængighedsegenskaber ved fordelingen af estimatorerne (2.5.20).

For Poisson-fordelte observationer gælder dog, at for modeller, der alene modellerer hovedeffekter af faktorvariable ved den kanoniske linkfunktion, vil kovarianserne mellem de estimerede koefficienter være nul på de samme pladser (dvs mellem de samme koefficienter) som i $[\mathbf{X}^T\mathbf{X}]^{-1}$ matricen.

Dette skyldes, at den multiplikative struktur af de fittede værdier (som i Poisson-tilfældet jo netop også er $V(\mu)$) netop svarer til ortogonalitetskravet til klassifikationer (Bemærkning 1 til definition 2.9.3). □

2.14 Forklaringsgrad R^2

Ved en søgning efter passende modeller kan man supplere de udførte test, først og fremmest med en vurdering af residualerne, men også med vurdering af andre hjælpestørrelser.

Blandt sådanne hjælpestørrelser er forklaringsgraden.

Betragt to modeller H_1 og H_m, hvor H_m angiver den mindst mulige model, der kan have interesse i den givne samenhæng. I en regressionsanalysesammenhæng kan H_m således angive modellen svarende til at alle regressionskoefficienter er nul, og kun en afskæring. I en variansanalysesammenhæng kan H_m være modellen fælles niveau og i antalstabeller kan H_m eksempelvis være en model med fuldstændig uafhængighed mellem inddelingskriterier.

$G^2(H_m)$ er et udtryk for den totale variation i data. Tilsvarende udtrykker størrelsen $G^2(H_m) - G^2(H_1)$ den variation, der er forklaret ved modellen H_1. Størrelsen

$$R^2 = \frac{G^2(H_m) - G^2(H_1)}{G^2(H_m)}$$

er således den andel af den totale variation, der kan forklares ved modellen H_1.

Størrelsen R^2 modsvarer udtrykket for forklaringsgraden, som ofte benyttes i lineær regressionsanalyse for normalt fordelte variable.

Jo større en model H_1, desto større vil forklaringsgraden være. Såfremt H_1 er den mættede model, vil $G^2(H_1) = 0$ og forklaringsgraden $R^2 = 1$.

2.14.1 Korrigeret forklaringsgrad R'^2

Såfremt man ønsker at sammenligne forklaringsgraden for to modeller, vil det derfor være naturligt at benytte den korrigerede forklaringsgrad

$$R'^2 = 1 - \frac{k - r}{k - r_0}(1 - R^2)$$

hvor r angiver dimensionen af modellen H_1 og r_0 angiver dimensionen af H_m.

Der gælder

$$R'^2 = 1 - \frac{G^2(H_1)/(k - r)}{G^2(H_m)/(k - r_0)}$$

dvs $1 - R'^2$ er forholdet mellem de to G^2-teststørrelser for tilpasning af modellerne H_1 og H_m, normeret med de tilhørende frihedsgrader. Vi erindrer om, at de to G^2-størrelser ikke er uafhængige.

En stor værdi af R'^2 indikerer, at modellen H_1 giver en god beskrivelse af data. Den største værdi af R'^2 fås for den model, H_1, der har den mindste værdi af $G^2(H_1)/(k-r)$. En brug af R'^2 som kriterium for valg af model vil således favorisere modeller med mange forklarende led.

2.14.2 Akaike's informationskriterium A_H

Akaike (1973) har foreslået et generelt kriterium for modelvalg baseret på et mål for den information, der er indeholdt i en given model. Akaike har foreslået at vælge den model, der maksimerer denne information. For generaliserede lineære modeller svarer dette kriterium til at vælge den model H, der maksimerer

$$A_H = G^2(H) - [k - 2r]$$

hvor $G^2(H)$ er likelihoodkvotient teststørrelsen for test af modellen H mod den mættede model, og hvor k er dimensionen af den mættede model, og r angiver dimensionen af modellen H.

2.15 Valg af model og modelkontrol

fil glm6.tex 1998-02-17

2.15.1 Generelt om modelvalg og kontrol

Valget af model til at beskrive et givet sæt data er oftest en iterativ proces bestående af følgende trin:

1. Valg af modelklasse
2. Valg af kovariable
3. Kontrol af modeltilpasning

Valg af modelklasse

McCullough og Nelder (1989) foreslår, at man tager udgangspunkt i nogle enkle principper ved modelformulering. Det første hovedprincip er, at alle modeller er forkerte; der er imidlertid nogle modeller, der er mere anvendelige end andre, og det drejer sig om at finde disse.

Et andet princip er, at man ikke skal blive så begejstret for en bestemt model, at man udelukker alternativer. Det er vigtigt at gøre sig klart, at data som regel vil pege med lige stor vægt på flere mulige modeller.

Et tredie princip anbefaler, at man kontrollerer tilpasningen af en model, for eksempel ved at analysere residualer og andre størrelser, der kan udledes af dataanalysen, for at vurdere afvigende observationer, og de enkelte observationers betydning for modeltilpasningen.

Først vælger man en klasse af modeller, for eksempel fordeling af responsvariabel, linkfunktion, og en passende samling forklarende variable, eventuelt transformationer af de forklarende variable.

Derefter tilpasser man en model med et passende sæt af kovariable ved at udfinde et sæt, som giver en rimelig fysisk og statistisk forklaring af data.

Og endelig udfører man en modelkontrol ved at vurdere om der er systematiske afvigelser fra modellen, eller om enkelte punkter kræver særlig opmærksomhed.

Det kan imidlertid ske, at selv om man har valgt sin modelklasse omhyggeligt, indikerer tilpasningen til data, at den resulterende model ikke er tlfredsstillende. Sådanne indikationer kan manifestere sig på forskellige måder. Det kan være at samlingen af data som helhed viser tegn på en systematisk afvigelse fra de fittede værdier, eller det kan være at nogle få datapunkter afviger fra det generelle mønster. I sådanne tilfælde begynder man forfra med at vurdere hvorvidt modelklassen bør ændres etc.

Valg af kovariable

Formålet med at tilpasse en model til de foreliggende data er i første omgang at erstatte datavektoren $\mathbf{y}$ med en vektor $\widehat{\boldsymbol{\mu}}$ af fittede værdier, som er beregnet ved brug af modellen. De fittede værdier er valgt sådan at et mål for afvigelse (f.eks. residualdeviansen) er minimal.

I én forstand drejer valget af kovariable sig blot om at finde et eller flere sæt af variable, der udspænder et passende underrum L.

Ved en første betragtning kunne man mene, at en god model er en model, der giver en meget god tilpasning til data, dvs en model, der bringer $\widehat{\boldsymbol{\mu}}$ meget tæt på $\mathbf{y}$. Hvis man inddrager tilstrækkelig mange variable i sin model, kan man imidlertid få så god en tilpasning, som man ønsker. Hvis man bruger lige så mange parametre, som der er observationer, kan man opnå en perfekt tilpasning, som vi så i afsnit 2.14. Gør man dette, har man imidlertid ikke opnået nogen reduktion i kompleksitet, dvs foreslået et enklere teoretisk mønster for de givne data. Enkelhed, eller parameterøkonomi, er også en ønskværdig egenskab ved en model. Man inddrager ikke unødvendige størrelser i modellen. En simpel model har ikke kun den fordel, at den gør det muligt at forstå og forklare de foreliggende data, den har også den fordel, at en simpel model, der stort set modsvarer virkeligheden, giver bedre forudsigelser, end en model, der indeholder unødige ekstra parametre.

Det er også vigtigt at tænke på modellens formål, eller anvendelsesområde, dvs under hvilke betingelser, modellen giver gode forudsigelser. Det kan undertiden være vanskeligt at formalisere modellens sigte, men som regel er det lettere at erkende sigtet. Man skal være forsigtig med at ekstrapolere i en model ud over det område, hvor den er tilpasset. Dette gælder både for de parametre, der er med i modellen, som for de omstændigheder, hvorunder de underliggende data er indsamlet.

Man bør holde sig for øje, at den resulterende model bør have en meningsfuld fysisk fortolkning. Som et minimum indebærer dette, at man kun skal medtage vekselvirkninger, hvis man også medtager de tilsvarende hovedvirkninger, at man kun skal medtage polynomier af højere orden, hvis man medtager de tilsvarende polynomier af lavere orden etc..

Hvis analysen skal bruges som et resume af en eller flere undersøgelser af det samme fænomen, kan det undertiden være en fordel at medtage hovedled, hvad enten de er signifikante, eller ej. Hvis man følger denne strategi vil det være lettere at sammenligne resultaterne af forskellige undersøgelser, og det vil afværge de problemer, der kan opstå, hvis man tilpasser forskellige modeller med forskellige variabelsæt i forskellige undersøgelser af samme fænonem. Faren er, at man kan risikere i én undersøgelse at fjerne et led med en bestemt variabel fordi den ikke er fundet at bidrage signifikant, mens den samme variabel kan have en tilsvarende koefficient i en anden undersøgelse, hvor den findes at have en signifikant indflydelse. Udelades denne variabel af modellen i det ene tilfælde, mens den bevares i det andet tilfælde, vil man stå med to forskellige modeller som tilsyneladende strider mod hinanden, mens de to undersøgelser reelt stemte overens.

Normalt skal man dog udvise den yderste sparsommelighed ved inddragelsen af kovariable (forklarende variable) i en model. En sådan sparsommelighed indebærer blandt andet at man normalt ikke skal medtage kovariable, som ikke øger forklaringen væsentligt. Det er dog lettere sagt end gjort. I nogle undersøgelser har man kun udvalgt nogle få - undertiden for få - kovariable til den statistiske analyse; i andre undersøgelser har man - for en sikkerheds skyld registreret alle størrelser som man mente kunne have relevans. Udvælgelsen af en brugbar gruppe af kovariable fra en sådan stor mængde er ikke nogen triviel opgave. Et af problemerne består i at opnå en passende balance mellem de to modsatrettede virkninger af at medtage et yderligere led i en model. Fordelen ved at medtage leddet er, at man opnår en bedre tilpasning mellem modellen og de foreliggende data. Ulempen er, at med mindre man på forhånd har en viden om at den variable har en ikke-forsvindende effekt på responset, vil medtagelsen af et yderligere led komplicere modellen og de konklusioner, man kan drage af modellen.

Atkinson (1981) anfører, at de fleste kriterier for inddragelse af yderligere led kan opfattes som specialtilfælde af problemet med at minimere størrelsen

$$Q = D + am\sigma^2 \,, \tag{2.15.1}$$

hvor D angiver deviansen, m er dimensionen af det lineære underrum, der udspænder modellen, σ^2 er dispersionsparameteren, og a er en størrelse, der udtrykker vægtningen mellem fordelen ved at reducere afvigelsen og ulempen ved at inddrage en yderligere variabel.

Brug af et 5 % signifikansvniveau ved et partielt test (afsnit 2.11.5) for inddragelsen af en variabel svarer til en a-værdi i (2.15.1) på ca. 4, mens Akaikes kriterium (afsnit 2.14.2) svarer til $a = 2$.

Man kan desuden overveje, om ikke størrelsen a i kriteriet bør afhænge af antallet af observationer; jo større observationsantal k, desto større værdi af a.

Modelkontrol

Modelkontrollen omfatter kontrol af såvel systematiske afvigelser fra modellen som af enkelte observationers afvigelse.

En systematisk afvigelse kan undertiden identificeres på et plot af residualerne imod de forklarende variable. Hvis tilpasningen er god, vil plottet af residualerne ligne "hvid støj". Hvis derimod plottet viser en koncentration af positive residualer i den ene ende, og negative residualer i den anden ende, kan dette være en indikation af at man burde have valgt en anden linkfunktion, eller at man mangler et led med kvadratet på den pågældende variable.

En isoleret modelafvigelse optræder, når nogle enkelte observationer giver anledning til residualer, der er større end resten. Et sådant billede kan f.eks. optræde i udkanten af variationsområdet for en forklarende variabel som en indikation af at model ikke er velegnet i dette område. Man skal heller ikke udelukke muligheden af at isolerede afvigelser fra modellen skyldes fejl i dataregistreringen.

2.15.2 Brug af residualer til kontrol af systematiske afvigelser fra modellen

Brug af standardiserede residualer

Hvis der er et rimeligt antal observationer, vil det være naturligt at forsøge at vurdere om der er indikation af systematiske afvigelser fra den fittede model.

Dette kan blandt andet gøres ved at optegne de standardiserede residualer mod den lineære prædiktor, $\widehat{\eta}$, eller mod de fittede værdier $\widehat{\mu}$.

Såfremt der ikke er systematiske afvigelser fra modellen, vil de standardiserede (og de studentiserede) residualer have middelværdi nul og konstant spredning. Typiske systematiske afvigelser vil være et krumt forløb af residualerne, eller en systematisk ændring af spredningen med stigende værdier af de fittede værdier.

Et krumt forløb kan skyldes et uheldigt valg af linkfunktion, et uheldigt valg af skala for en eller flere kovariater, eller at der burde være et kvadratisk led i en af kovariaterne.

Tilsvarende kan man tegne de standardiserede residualer op mod de forklarende variable. Også her skal man være opmærksom på de samme former for afvigelse som ovenfor.

Når man kontrollerer et residualplot for modelafvigelser, kan det undertiden forekomme at værdierne af den forklarende variable nogle steder ligger væsentligt tættere end andre steder sådan at der nogle steder er en større tæthed af punkter end andre. I sådanne tilfælde kan det være svært at vurdere eventuelle systematiske tendenser ved blot at se på punktsværmen. Det gælder jo, at jo flere punkter, der er, desto større er variationsbredden. Man kan da supplere visuelle vurdering af de plottede punkter ved at indtegne en linie eller udglattet kurve gennem punktsværmen. Selv om man skal være forsigtig med at overfortolke det mønster, der bliver indikeret af en sådan linie eller udglatttet kurve, kan den dog være en støtte ved vurderingen af residualplot.

Brug af partielle størrelser

Når der er flere end én kovariabel i modellen, kan det forekomme, at relationen mellem residualerne og en bestem kovariabel overskygges af effekten fra andre kovariable, sådan at den ikke er synlig på et plot af residualerne mod den kovariable.

For at afsløre sådanne situationer kan man optegne forskellige former for partielle residualer, dvs residualerne fra en model, hvor den pågældende variabel er udeladt.

I S-plus genereres de såkaldte partielle residualer. Det partielle residual

svarende til variabel j for den i'te observation er defineret som

$$r_i^j = x_{ij}\widehat{\beta}_j + (y_i - \widehat{\mu}_i)g'(\widehat{\mu}_i) \ .$$

Det partielle residual udtrykker således arbejdsresidualet (2.5.44) svarende til en model, hvor den j'te variabel er udeladt.

For hver forklarende variable er der således en vektor af partielle residualer, ét for hver observation.

I SAS-systemet konstrueres nogle partielle afvigelser, de såkaldte "partial leverage"-værdier. Værdierne af partial leverage for responsvariablen er residualer fra en model, hvor den pågældende variabel er udeladt. Værdierne af partial leverage for en forklarende variabel (kovariable) er residualerne fra en lineær model, hvor den udeladte kovariabel søges forklaret ved de resterende kovariable.

Betragt en model med den $k \times m$-dimensionale modelmatrix $\mathbf{X}$.

Lad $\mathbf{X}_{[j]}$ betegne den $k \times (m-1)$-dimensionale matrix, der fremkommer ved at man fjerner den j'te søjle, og lad $\mathbf{x}_j$ angive den j'te søjle.

Den j'te partielle leverage x-variabel er da residualet $\mathbf{r}_{x[j]}$ fra regressionen af $\mathbf{x}_j$ på $\mathbf{X}_{[j]}$, og den j'te partielle leverage y-variabel $\mathbf{r}_{y[j]}$ er vektoren af residualer fra den generaliserede lineære model med modelmatricen $\mathbf{X}_{[j]}$.

Idet vi erindrer, at residualerne er ortogonale på de fittede værdier ser vi, at plottet af partielle leverages kan opfattes som et plot af den afhængige variable, Y mod den uafhængige variable $\mathbf{x}_j$ efter de er gjort ortogonale på de øvrige led i modellen.

Man kan vise, at den i'te komponent i den j'te partielle leverage y-variabel kan udtrykkes som

$$r_{y[j]i} = r_{x[j]i}\widehat{\beta}_j + (y_i - \widehat{\mu}_i)g'(\widehat{\mu}_i) \qquad (2.15.2)$$

Punkterne vil således gruppere sig omkring en linie igennem $(0,0)$ med hældning $\widehat{\beta}_j$.

Plottet illustrerer forskellen i residualer, henholdsvis når $\mathbf{x}_j$ er med imodellen, og når den ikke er med.

For et givet punkt $(r_{x[j]i}, r_{y[j]i})$ vil den lodrette afstand til linien $r_{y[j]i} = 0$ angive residualet, såfremt den variable $\mathbf{x}_j$ ikke medtages, mens afstanden ned til linien igennem $(0,0)$ med hældning $\widehat{\beta}_j$ angiver residualet, såfremt $\mathbf{x}_j$ medtages i modellen.

2.15.3 Kontrol af enkeltobservationer, leverage

Som vi så i bemærkning 4 på side 212, er diagonalelementerne i hat-matricen udtryk for den "vægt" hvormed den enkelte observation bidrager til den fittede værdi for den pågældende datapunkt.

Dette følger direkte ved at betragte

$$\frac{\partial \widehat{\boldsymbol{\mu}}}{\partial \mathbf{y}} = \frac{\partial \widehat{\boldsymbol{\mu}}}{\partial \widehat{\boldsymbol{\eta}}} \frac{\partial \widehat{\boldsymbol{\eta}}}{\partial \mathbf{y}} .$$

Vi har

$$\frac{\partial \widehat{\boldsymbol{\mu}}}{\partial \widehat{\boldsymbol{\eta}}} = \operatorname{diag}\left\{ \frac{1}{g'(\widehat{\mu}_i)} \right\}$$

samt

$$\frac{\partial \widehat{\boldsymbol{\eta}}}{\partial \widehat{\boldsymbol{\beta}}} = \mathbf{X}$$

og endelig har vi fra lemma 2.5.2, at

$$\frac{\partial \widehat{\boldsymbol{\beta}}}{\partial \mathbf{y}} = [\mathbf{i}_\beta(\boldsymbol{\beta})]^{-1} [\mathbf{X}(\boldsymbol{\beta})]^T \, \mathbf{i}_\mu(\boldsymbol{\mu})$$

hvorfor vi har

$$\frac{\partial \widehat{\boldsymbol{\mu}}}{\partial \mathbf{y}} = \mathbf{H}(\widehat{\boldsymbol{\beta}})$$

hvor $\mathbf{H}(\boldsymbol{\beta})$ er givet ved (2.5.51).

Diagonalelementerne h_{ii} i hat-matricen $\mathbf{H}$ angiver således ændringen i den fittede værdi $\widehat{\mu}_i$ ved en ændring i y_i.

Eksempel 2.15.1 *Leverage i en enkel regressionsmodel*

For at illustrere begrebet leverage vil vi betragte en regressionsmodel svarende til normalfordelingen, dvs en model med en konstant varians.

Betragt modelmatricen svarende til et intercept-led og én uafhængig variabel

$$\mathbf{X} = \begin{pmatrix} 1 & (x_1 - \overline{x}.) \\ 1 & (x_2 - \overline{x}.) \\ \vdots & \vdots \\ 1 & (x_k - \overline{x}.) \end{pmatrix} ,$$

hvor vi har ortogonaliseret de to søjler ved at subtrahere gennemsnitsværdien af x_i fra samtlige x-værdier. Vi har da

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} k & 0 \\ 0 & \sum(x_i - \overline{x}_\cdot)^2 \end{pmatrix}$$

hvorfor

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 1/k & 0 \\ 0 & 1/\sum(x_i - \overline{x}_\cdot)^2 \end{pmatrix}$$

Det i'te diagonalled i hat-matricen er

$$\begin{pmatrix} 1 & (x_i - \overline{x}_\cdot) \end{pmatrix} (\mathbf{X}^T\mathbf{X})^{-1} \begin{pmatrix} 1 \\ x_i - \overline{x}_\cdot \end{pmatrix} = \frac{1}{k} + \frac{(x_i - \overline{x}_\cdot)^2}{\sum(x_j - \overline{x}_\cdot)^2} \,.$$

Det i'te diagonalled udtrykker således den euklidiske afstand mellem punktet $(1, x_i)$ svarende til det i'te datapunkt, og tyngdepunktet $(1/k, \overline{x}_\cdot)$ for alle datapunkter.

For en regressionsmodel med m forklarende variable finder man tilsvarende, at det i'te diagonalled i hat-matricen er

$$h_{ii} = \frac{1}{k} + \frac{(x_{i1} - \overline{x}_{\cdot 1})^2}{\sum(x_{j1} - \overline{x}_{\cdot 1})^2} + \cdots + \frac{(x_{im} - \overline{x}_{\cdot m})^2}{\sum(x_{jm} - \overline{x}_{\cdot m})^2} \,. \tag{2.15.3}$$

altså netop den euklidiske afstand mellem vektoren af forklarende variable for den i'te observation og tyngdepunktet $(1/k, \overline{x}_{\cdot 1}, \ldots, \overline{x}_{\cdot m})$ svarende til alle observationer.

Størrelsen (2.15.3) kaldes også Mahalanobis' afstand mellem datapunktet svarende til den i'te observation og punktsværmen bestående af samtlige observationer.

Størrelsen $h_{ii} - 1/k$ er således et invariant mål for den kvadratiske afstand mellem $\mathbf{x}_i$ og tyngdepunktet for alle n punkter i koefficientrummet. □

Diagonalelementerne i hat-matricen udtrykker således hvor meget den tilsvarende række i $\mathbf{X}$-matricen afviger fra samlingen af værdier af de forklarende variable.

I engelsksproget litteratur siger man, at diagonalelementerne udtrykker den leverage (vægtstangseffekt), der er knyttet til den pågældende observation.

Leverage er således en egenskab knyttet til modellen og "designet", dvs til værdierne af de forklarende variable; men som vi så i bemærkning 4 på side 212, er konsekvenserne af en stor leverage af betydning såvel for variansen på residualet, som for variansen på den fittede værdi.

Man skal have opmærksomheden rettet specielt mod observationer med stor leverage. Ikke fordi det nødvendigvis er en dårlig egenskab at en observation har stor leverage. Hvis et punkt med stor leverage i øvrigt er i overensstemmelse med resten af data, vil dette punkt blot tjene til at mindske usikkerheden på de estimerede koefficienter, se figur 2.12. Hvis en model er nogenlunde rimelig, og data er i overensstemmelse med modellen vil observationer med stor leverage være en fordel.

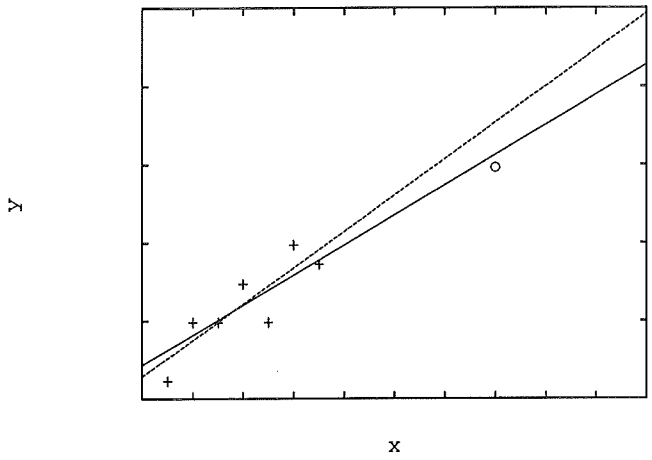
Observationer med stor leverage kan imidlertid være en indikation af at modellen ikke svarer til data. Det er velkendt, at man skal udvise stor varsomhed med at ekstrapolere en empirisk bestemt relation ud over det område i "designrummet", hvor den er tilpasset. En lineær model kan måske give en tilfredsstillende beskrivelse af data i et begrænset område af designrummet, men den behøver ikke nødvendigvis give en god beskrivelse uden for dette område.

Det er derfor ofte urealistisk at forvente, at man kan bestemme en lineær model, der giver en god beskrivelse i områder af designrummet, hvor der kun er få datapunkter. Datapunkter med stor leverage ligger netop i sådanne områder, og man kunne derfor formode, at sådanne observationer ikke beskrives særlig godt. Ofte ser man dog det modsatte fænomen. Modellen giver pæn tilpasning til datapunkter med stor leverage, mens tilpasningen til de øvrige datapunkter ikke er overvældende god, se f.eks. figur 2.13.

Belsley, Kuh og Welsch (1980) foreslår en afskæring på $2m/k$ for hat diagonalværdierne, hvor m er antallet af parametre i modellen, og k er antallet af observationer.

2.15.4 Kontrol af enkeltobservationers overensstemmelse, residual

Som vi har set, har de ustandardiserede residualer forskellig varians. Når man vil vurdere hvorvidt en enkelt observation afviger fra modellen, skal man derfor betragte de standardiserede eller de studentiserede residualer.

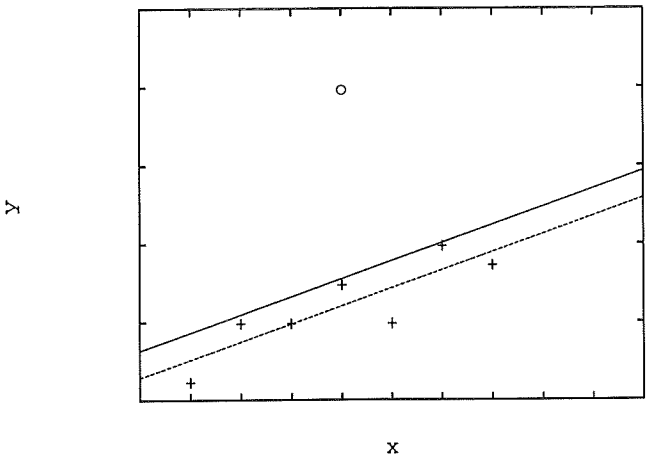


Figur 2.12. Illustration af et datapunkt ∘ med stor leverage

Den fuldt optrukne kurve angiver regressionslinien svarende til samtlige punkter. Den stiplede kurve viser regressionslinien estimeret under udeladelse af punktet ∘ med stor leverage.
(stor leverage, god konsistens, lille influens)

Det gælder for de skalerede og de studentiserede residualer, respons-, devians, og Pearson, der blev indført i afsnit 2.5.8, at de approximativt er normalt fordelt med middelværdi nul og spredning 1.

Et standardiseret residual, der er større end 2-3, er således tegn på en afvigende observation.



Figur 2.13. Illustration af et datapunkt ○ med stort residual
Den fuldt optrukne kurve angiver regressionslinien svarende til samtlige punkter. Den stiplede kurve viser regressionslinien estimeret under udeladelse af punktet ○ med et stort residual.
(lille leverage, ringe konsistens, lille influens)

2.15.5 Kontrol af enkeltobservationers indflydelse (influens)

Vi har allerede diskuteret "leverage" som et udtryk for datapunktets afstand fra de øvrige punkter i designrummet og herved også som et sammenfattende udtryk for den vægt, der tillægges det pågældende datapunkt i den samlede estimation.

Man kan imidlertid også have interesse i at vurdere nogle mere specifikke mål for de enkelte datapunkters indflydelse på de enkelte komponenter i estimationsproblemet.

Den enkleste måde, hvorpå man kan måle indflydelsen fra et datapunkt er ved at sammenligne de estimater man får, når det pågældende datapunkt

er medtaget i analysen og når det ikke er med.

I litteraturen er der foreslået en lang række mål til vurdering af indflydelsen fra et datapunkt. Nedenfor skal vi gennemgå en række af disse mål.

Belsley, Kuh og Welsch (1980) og Cook og Weisberg (1982) indeholder en mere udførlig beskrivelse.

I nedenstående gennemgang vil et indeks i rund parentes "(i)" indikere, at observation i er udeladt af estimationen.

De forskellige influensmål svarende til udeladelse af den i'te observation udregnes lettest ved at supplere $\mathbf{X}$-matricen med en ekstra søjle med nuller i alle positioner på nær den i'te række. Effekten af denne søjle er, at modellen udvides med et led, der blot tager sig af at estimere y_i. De øvrige søjler i $\mathbf{X}$ bruges da til at estimere en model for de øvrige observationer. Eksempelvis fås estimatet $\widehat{\sigma}^2_{(i)}$ som den valgte χ^2-teststørrelse for modeltilpasning (devians- eller Pearson-) svarende til denne udvidede model,

$$\widehat{\sigma}^2_{(i)} = \frac{G^2(H^*)}{k-m-1}$$

hvor H^* angiver den undersøgte model, udvidet med leddet svarende til den i'te observation.

Cook's D

Cook's D måler ændringerne i parameterestimater, hvis man udelader den i'te observation.

Cook's D svarende til den i'te observation er defineret som

$$D_i = \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})^T \mathbf{i}_\beta(\widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})}{\widehat{\sigma}^2 m} \tag{2.15.4}$$

hvor den forventede information $\mathbf{i}_\beta(\boldsymbol{\beta})$ er anført i lemma 2.5.1, formel (2.5.17).

$$\mathbf{i}_\beta(\boldsymbol{\beta}; \mathbf{y}) = \mathbf{X}^T\mathbf{W}\mathbf{X}$$

Til praktisk brug benyttes estimaterne for $\mathbf{W}$ bestemt af (2.5.31) eller (2.5.28).

(Se (2.5.30) eller (2.5.27)).

Ved at indsætte udtrykket (2.5.17) for $\mathbf{i}_\beta(\beta)$ ser man, at Cook's D kan skrives

$$D_i = \frac{(\widehat{\beta} - \widehat{\beta}_{(i)})^T \mathbf{X}^T \mathbf{W} \mathbf{X} (\widehat{\beta} - \widehat{\beta}_{(i)})}{\widehat{\sigma}^2 m} = \frac{\|\widehat{\eta} - \widehat{\eta}_{(i)}\|_W^2}{\widehat{\sigma}^2 m}$$

hvor $\widehat{\eta} = \mathbf{X}\widehat{\beta}$ og $\widehat{\eta}_{(i)} = \mathbf{X}\widehat{\beta}_{(i)}$. Cook's D måler således forskellen i den lineære prædiktor, når den i'te observation fjernes. Såfremt der indgår en estimeret dispersionsparameter, er Cook's D skaleret med dette estimat (i dette estimat indgår dog den i'te observation).

Cook (1977) foreslog at man sammenligner D_i med fraktilerne i en $\mathrm{F}(m, k-m)$-fordeling. En grov afskæring er ved $4/(k-m)$

Dffit og Dffits

Forskellen i fittet værdi, henholdsvis med og uden den i'te observation kaldes Dffit.

$$F_i = \frac{\widehat{\mu}_i - \widehat{\mu}_{(i)}}{\sqrt{\widehat{\sigma}^2 \, h_{ii}}}$$

Den tilsvarende skalerede værdi kaldes Dffits.

For den i'te observation har man

$$F_i^* = \frac{\widehat{\mu}_i - \widehat{\mu}_{(i)}}{\sqrt{\widehat{\sigma}^2_{(i)} h_{ii}}} \tag{2.15.5}$$

hvor h_{ii} angiver diagonalelementet i hat-matricen.

Sædvanligvis bruger man en afskæring på 2 for DFFITS, En antalsjusteret afskæring er $\pm 2\sqrt{m/(k-m)}$

Dfbetas

Den normaliserede forskel på estimatet af β_j henholdsvis med og uden den i'te observation kaldes Dfbetas.

For den i'te observation har man

$$B_{j,i} = \frac{\widehat{\beta}_j - \widehat{\beta}_{j(i)}}{\sqrt{\widehat{\sigma}^2_{(i)} [\mathbf{X}^T \mathbf{W} \mathbf{X}]^{-1}_{jj}}}, \tag{2.15.6}$$

hvor

$$\mathbf{X}^T\mathbf{W}\mathbf{X} = \mathrm{i}_\beta(\beta)$$

angiver informationsmatricen (med hensyn til β (2.5.17).

Til praktisk brug benyttes estimaterne for $\mathbf{W}$ bestemt af (2.5.31) eller (2.5.28).

Man får en hel matrix af $B_{j,i}$ værdier; for hver observation får man en værdi for hver af de uafhængige variable.

Værdier af $B_{j,i}$ større end 2 indikerer observationer, der er influentielle i estimationen af en β_j.

En anbefalet afskæringsværdi er $\pm 2/\sqrt{k}$ hvor k angiver antallet af observationer.

Covratio

Endelig kan man måle effekten af den i'te observation på den estimerede dispersionsmatrix (2.5.22) for parameterestimaterne. For den i'te observation definerer man

$$C_i = \frac{\widehat{\sigma}^2_{(i)}\ \det([\mathbf{X}_{(i)}{}^T\mathbf{W}_{(i)}\mathbf{X}_{(i)}]^{-1})}{\widehat{\sigma}^2\ \det([\mathbf{X}^T\mathbf{W}\mathbf{X}]^{-1})}\ , \tag{2.15.7}$$

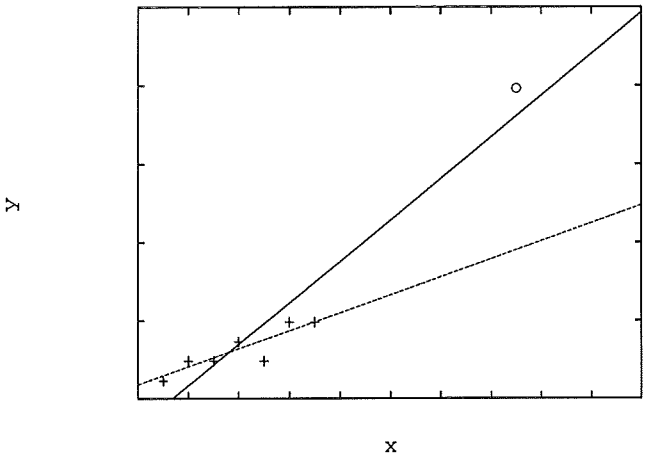
hvor $\det(\cdot)$ betegner determinanten, og $\mathbf{W}_{(i)}$ og $\mathbf{X}_{(i)}$ angiver $\mathbf{W}$ matricen og $\mathbf{X}$-matricen uden den i'te observation.

Størrelsen C_i kaldes almindeligvis for Covratio.

Værdier af C_i, der er nær 1 indikerer, at observationen ikke har stor effekt på estimationsnøjagtigheden

Værdier, der er mindre end 1, angiver, at estimationsnøjagtigheden forringes, når den pågældende observation medtages. Værdier, der er større end 1, angiver at estimationsnøjagtigheden forbedres, når observationen medtages.

Observationer med $|C_i - 1| \geq 3m/k$ har en væsentlig indflydelse på nøjagtigheden, og man bør derfor være sikker på at disse observationer ikke er fejlmålinger.



Figur 2.14. Illustration af et datapunkt ○ med stor influens

Den fuldt optrukne kurve angiver regressionslinien svarende til samtlige punkter. Den stiplede kurve viser regressionslinien estimeret under udeladelse af punktet ○ med stor influens.
(stor leverage, ringe konsistens, stor influens)

2.15.6 Vurdering af enkeltobservationer, sammenfatning

McCullough og Nelder (1989) foreslår at man sondrer mellem et datapunkts leverage, dets indflydelse (influens), målt ved et af de anførte influensmål, og dets overensstemmelse (konsistens), målt ved residualet.

Figurerne 2.12, 2.13 og 2.14 viser forskellige kombinationer af leverage, influens og konsistens.

Nedenstående tabel resumerer de forskellige mål for kontrol af enkeltobservationer:

Størrelse	Udtryk	Fortolkning	side	Afskæring
DFBETAS	$\frac{\widehat{\beta}_j - \widehat{\beta}_{j(i)}}{\sqrt{\widehat{\sigma}^2_{(i)}[\mathbf{X}^T\mathbf{W}\mathbf{X}]^{-1}_{jj}}}$	Ændringen i koefficienterne, skaleret med spredningen for de estimerede koefficienter	346	$\pm 2/\sqrt{k}$
DFFIT	$\frac{\widehat{\mu}_i - \widehat{\mu}_{(i)}}{\sqrt{\widehat{\sigma}^2\, h_{ii}}}$	Den standardiserede ændring i den fittede værdi, når den i'te observation udelades	346	
DFFITS	$\frac{\widehat{\mu}_i - \widehat{\mu}_{(i)}}{\sqrt{\widehat{\sigma}^2_{(i)} h_{ii}}}$	Som ovenfor med $\widehat{\sigma}^2_{(i)}$ som dispersionsparameter	346	$\pm 2\sqrt{m/(k-m)}$
Hat diagonal	h_{ii}	diagonal i projektionsmatrix	340	$2m/k$
Cooks D	se (2.15.4)	Gennemsnitlig ændring i parameterestimater	345	$\mathrm{F}(m, k-m)$

2.16 Referencer:

H. Akaike, H.: Information theory and an extension of the maximum likelihood principle. I *Proceedings of the 2nd International Symposium on Information.* B.N. Petrov and F.Czaki eds. Budapest: Akademiai Kiado 1973

Atkinson, A.C.: Likelihood ratios, posterior odds and information criteria. *Journ. Econometrics* **16**, 1981, pp. 15-20

Atkinson, A.C.: *Plots, Transformations and Regression.* Oxford, Clarendon Press (1985)

Barnard, G.A.: Statistical inference. *J. Roy. Statist. Soc.* **B 11**, pp 115 -149, 1949

Barnard, G.A.: The logic of statistical inference. *Brit. J. Phil. Sci.* **23**, pp 123 -132, 1972

Barndorff-Nielsen, O.E.: *Information and Exponential Families in Statistical Theory,* Wiley, New York ,1978.

Bartlett, M.S.: Properties of sufficiency and statistical tests. *Proc.Roy.Soc. A.* **160**, pp 268-82, 1937.

Belsley, D.A., Kuh, E. and Welsch, R.E. *Regression Diagnostics,* J.Wiley & Sons, New York 1980

Christensen, R.: *Log-Linear Models,* Springer, New York 1990

Clayton, M.K., Geisser, S. and Jennings, D. E.: A comparison of several model selection procedures. I *Bayesian Inference and Decision Techniques.* P. Goel and A. Zellner eds. Amsterdam, North Holland 1986.

Clemmesen, J., Hansen, G., Nielsen, A., Røjel, J.,Steensberg, J., Sørensen, S. & Toustrup, J.: Lungecancer og luftforurening i Fredericia, en epidemiologisk undersøgelse. *Ugeskr. Læg.* **136** (1974) pp. 2260-2268.

Cook, R.D, and Weisberg, S.: *Residuals and Influence in Regression.* Chapman and Hall, London 1982

D.R. Cox and E.J. Snell: A general definition of residuals. *Journ. Roy. Statist. Soc,* **B 30** (1968), pp. 248-275.

Dudewicz, E.J. and Mishra, S.N.: *Modern Mathematical Statistics,* Wiley, New York 1988.

Fisher, R.A.: *Philosophical Transactions*, Series A, Vol 222, 1922, pp. 309-368

Hald, A.: *Statistiske Metoder.* Akademisk forlag (1948).

Hansen, C.K. and Thyregod, P: Modelling and estimation of wafer Yields and Defect Densities from Microelectronics Test Structure Data. *Quality and Reliability Engineering International*, **12**, 1996, pp. 9-17

Hoblyn, T.N. and Palmer, R.C.: A complex experiment in the propagation of plum rootstocks from root cuttings, season 1931-32. *Journ. Pom. and Hort. Sci.*, **12**, 1934, pp 36-56

Jørsboe, O.G.: *Sandsynlighedsregning.* Matematisk Institut, DTH 1984

Jørgensen, B.: Exponential dispersion models and extensions: a review. *International Statistical Review* **60**, (1992), pp. 5-20.

Jørgensen, B.: *The theory of dispersion models.* Chapman & Hall, New York (1997).

Küchler, U.: Exponential Families of Markov Processes - Part I. General Results. *Math. Operationsforsch. Statist., Ser. Statistics* **13** (1982) pp 57-69.

Küchler, U.: Exponential Families of Markov Processes - Part II. Birth and death processes. *Math. Operationsforsch. Statist., Ser. Statistics* **13** (1982) pp 219-230.

Letac, G. (1992): *Lectures on Natural Exponential Families and their Variance Functions.* Monografias de Matematica **50**. Instituto de Matematica Pura e Aplicada. Rio de Janeiro.

McCullagh, P. and Nelder, J. A.: *Generalized Linear Models*, Second Edition, Chapman and Hall, London 1983

Müller-Funk, U. and Pukelsheim, F.: How Regular are conjugate exponential families ? *Statistics & Probability Letters* **7** (1989), pp 327-333

Morris, C.N.: Natural Exponential Families with Quadratic Variance Functions. *The Annals of Statistics* **10** (1982) pp. 65-80.

Price, C.J., Kimmel, C.A., George, J.D. and Marr, M.C: The developmental toxity of diethylene glycol dimethyl ether in mice. *Fund. Appl. Toxicol.* **8**, (1987), pp. 115-126.

Rockafeller, R.T.: *Convex Analysis*, Princeton University Press, (1970)

G.N. Wilkinson and C.E. Rogers: Symbolic description of factorial models for analysis of variance. *Appl.Statist.* **22**, (1978), pp 392-399.

D.A. Williams: Generalized linear model diagnostics using the deviance and single-case deletions. *Appl.Statist.* **36**, (1987), pp. 181-191

Afsnit 3

Modeller for binære responsvariable

fil bin.tex 1997-03-25

3.1 Binomialfordelingen som eksponentiel dispersionsparameterfamilie, kanonisk link

3.1.1 Odds, logit

Betragt n uafhængige gentagelser af et forsøg med to mulige responskategorier, A og A^c. Såfremt sandsynligheden $p = \mathsf{P}[A]$ er den samme i alle n forsøg, da vil fordelingen af antallet Y af gange, hændelsen A indtræffer i de n forsøg kunne beskrives ved en $\mathrm{B}(n,p)$ fordeling.

Vi indfører odds $\theta(p)$ for hændelsen A som

$$\theta(p) = \frac{p}{1-p}, \tag{3.1.1}$$

Afbildningen ved odds-funktionen er en monotont voksende afbildning, der afbilder $0 < p < 1$ ind på $0 < \theta(p) < \infty$.

Odds for hændelsen A angiver således, hvor mange gange mere sandsynligt det er, at hændelsen A indtræffer, end at den ikke indtræffer. Eller, udtrykt på en anden måde: Odds angiver forholdet mellem det forventede antal gange, hvor hændelsen indtræffer, og antallet af gange, hvor hændelsen ikke indtræffer i en serie af forsøg. Såfremt der eksempelvis er sandsynligheden $1/4$ for at en hændelse A indtræffer, da vil odds for hændelsen være $1:3$.

I bookmakersammenhænge, og i totalisatorspil har odds en lidt en anden betydning. Totalisatorodds, $Odds^*$ er nemlig det multiplum af indsatsen, der udbetales ved gevinst, samtidig med at indsatsen tilbagebetales.

Antag, at sandsynligheden for hændelsen A er p. Ved et "fair" spil vil den forventede gevinst være nul, dvs.

$$(Odds^* + 1)p - (1-p) = 0$$

svarende til at totalisatorodds er

$$Odds^* = \frac{1-p}{p} - 1$$

dvs. totalisatorodds svarer nærmere til sandsynligheden for ikke at vinde.

Odds-skalaen er i mange henseender bekvemmere, end p-skalaen. Da odds-skalaen udfylder hele den positive reelle akse, $\mathbb{R}_+$, deler den en række egenskaber med $\mathbb{R}_+$. Mængden af odds-værdier er således afsluttet overfor multiplikation med positive reelle tal. Specielt kan nye odds-værdier dannes ved at multiplicere odds-værdier. Der findes et neutralt element, $\theta = 1$ ved multiplikation (svarende til $p = 1/2$); enhver odds-værdi har en invers, (den reciprokke odds, svarende til odds for den komplementære hændelse A^c), og endelig kan vi dividere to odds-værdier med hinanden og resultatet bliver igen en odds-værdi.

Vi vil endelig indføre log-odds, ϑ, som

$$\vartheta(p) = \ln\left(\frac{p}{1-p}\right) \tag{3.1.2}$$

Logaritmen til odds-værdien for en hændelse benævnes ofte logit-værdien. Afbildningen ved logit-funktionen er en monotont voksende afbildning, der afbilder $0 < p < 1$ ind på $-\infty < \vartheta(p) < \infty$.

Da logit-skalaen udfylder hele den reelle akse, $\mathbb{R}$, deler den en række egenskaber med $\mathbb{R}$. Da logit-skalaen fremkommer ved en logaritmetransforma-

tion af odds-skalaen, overføres alle odds-skalaens egenskaber ved den isomorfi mellem $\mathbb{R}_+$ og $\mathbb{R}$, der herved introduceres. Målinger på logit-skalaen har en vektorrumsstruktur.

3.1.2 Sammenligning af hændelser

Ønsker vi at sammenligne to hændelser A og B med $\mathsf{P}[A] = p_1$ og $\mathsf{P}[B] = p_2$, har vi således en række forskellige sammenligningsmuligheder, svarende til valg af 'skala' for repræsentation af sandsynligheden.

Differens mellem hyppigheder (sandsynligheder)
En forskel mellem sandsynlighederne for hændelserne A og B kan udtrykkes som differensen (den absolutte forskel) $p_1 - p_2$ mellem sandsynlighederne. Der gælder

$$-1.0 \leq p_1 - p_2 \leq 1.0$$

Differensen mellem sandsynlighederne p_1 og p_2 for hændelserne A og B er lig med differensen mellem sandsynlighederne $1 - p_1$ og $1 - p_2$ for de komplementære hændelser A^c og B^c.

Såfremt sandsynlighederne p_1 og p_2 er udtrykt i procent, angiver man ofte differensen udtrykt i procentpoints, dvs man siger forskellen er "forskellen er x procentpoints" for at undgå den tvetydige formulering "forskellen er x %".

Relativ risiko
Da sandsynligheder er ikke-negative tal, kan man også vælge at sammenligne sandsynligheder ved at betragte forholdet p_1/p_2. En sådan betragtning kan specielt være relevant ved en sammenligning af små sandsynligheder, f.eks. knyttet til uønskede hændelser. En forskel mellem sandsynligheder p_1 og p_2 for havariprovokerende hændelser på $p_1 - p_2 = 0.009$ (dvs. 0.9 procentpoits) er nok af større interesse, når $p_1 = 0.010$ og $p_2 = 0.001$, end hvis forskellen er mellem sandsynlighederne $p_1 = 0.310$ og $p_2 = 0.301$.

Forholdet

$$p_1/p_2$$

betegnes ofte den relative risiko for A i forhold til B. Den relative risiko kan antage alle ikke-negative værdier. (Såfremt p_2 er givet, kan den relative risiko i forhold til p_2 dog ikke overstige $1/p_2$.

Den relative risiko for de komplementære hændelser A^c i forhold til B^c fås som $(1-p_1)/(1-p_2)$. Denne størrelse er forskellig fra den relative risiko p_1/p_2 for A i forhold til B.

Odds-ratioen

En sammenligning af hændelser, hvis sandsynligheder er udtrykt på odds-skalaen foretages naturligt ved at betragte odds-ratioen

$$\omega_{1,2} \stackrel{\text{DEF}}{=} \theta_1/\theta_2 = \frac{p_1}{(1-p_1)} \Big/ \frac{p_2}{(1-p_2)} = \frac{p_1(1-p_2)}{p_2(1-p_1)} \tag{3.1.3}$$

hvor $p_1 = \mathsf{P}\ [A]$ og $p_2 = \mathsf{P}\ [B]$

Relation mellem odds-ratio og relativ risiko
Det følger af definitionen på odds-ratioen, at

$$\text{odds-ratio} = \text{relativ risiko} \times \frac{1-p_2}{1-p_1}$$

Med mindre $p_1 = p_2$, vil den relative risiko være nærmere 1, end odds-ratioen. Såfremt de indgående risici, p_1 og p_2 begge er små, vil odds-ratioen og den relative risiko dog være tilnærmelsesvis ens.

Differens mellem logit-værdier

En sammenligning af to hændelser på logit-skalaen kan tilsvarende foretages ved at sammenligne differenser mellem hændelsernes logit-værdier. Differensen mellem logit-værdierne for hændelserne A og B er netop logaritmen til odds-ratioen for de to hændelser.

3.1.3 Generaliserede lineære modeller for binomialt fordelte variable

Vi minder om, at familien af $\mathrm{B}(n,p)$-fordelinger med $0 < p < 1$ udgør en fuld eksponentiel familie af orden 1. Den kanoniske parameter er netop logit'en $\vartheta = \vartheta(p)$, og den kanoniske stikprøvefunktion er $t(z) = z =$ antallet af gange, hændelsen A er indtruffet i de n forsøg. Det kanoniske parameterområde er åbent, hvilket giver formelle estimationsproblemer, når $z = 0$,

eller $z = n$, svarende til $p = 0$ eller $p = 1$, som jo ikke er omfattet af det kanoniske parameterområde.

Når vi modellerer binomialt fordelte variable, $Z \in \mathrm{B}(n,p)$,, ved generaliserede lineære modeller, vil vi i overensstemmelse med betragtningerne i afsnit 2.4.1 opstille generaliserede lineære modeller for middelværdien μ af $Y = Z/n$, dvs. for en $\mathrm{B}(n,p)/n$-fordelt variabel. I stedet for at bruge det generiske symbol μ for middelværdien, vil vi i binomialfordelingstilfældet ofte bruge symbolet p.

Det følger af eksempel 2.3.2, at den kanoniske link funktion er

$$\eta = \ln\left(\frac{p}{1-p}\right) , \tag{3.1.4}$$

altså netop logit-transformationen.

Den omvendte funktion er den logistiske funktion

$$p = \frac{\exp(\eta)}{1+\exp(\eta)} \tag{3.1.5}$$

Variansfunktionen er (jvf. tabel 2.2)

$$V(p) = p(1-p) ,$$

dispersionsparameteren $\delta = 1$, og vægten $w = n$.

I det følgende vil vi betragte et sæt, $Z_1, \ldots, Z_k$ af uafhængige binomialfordelte variable med $Z_i \in \mathrm{B}(n_i, p_i)$. For $Y_i = Z_i/n_i$ har vi således $Y_i \in \mathrm{B}(n_i, p_i)/n_i$

Deviansbidraget svarende til observationen $y = z/n$ og den fittede værdi $\widehat{p}$ fås af tabel 2.3 som

$$\begin{aligned} n\, d(y;\widehat{p}) &= 2n\{y \ln\left(\frac{y}{\widehat{p}}\right) + (1-y)\ \ln\left(\frac{1-y}{1-\widehat{p}}\right) \\ &= 2\{z \ln\left(\frac{z}{n\widehat{p}}\right) + (n-z)\ \ln\left(\frac{n-z}{n(1-\widehat{p})}\right) \end{aligned} \tag{3.1.6}$$

idet modellen er vægtet med vægten n.

Deviansresidualet (2.5.38) bliver her

$$r_D(y;\widehat{p}) = \text{sign}(y-\widehat{p})\sqrt{n\, d(y;\widehat{p})}\ . \tag{3.1.7}$$

hvor "sign" angiver signum-funktionen, med værdierne $+1$ eller -1 afhængigt af fortegnet for $y-\widehat{p}$.

For et sæt $y_1, y_2, \ldots, y_k$ af observationer fås devianstesteststørrelsen $D(\mathbf{y};\widehat{\mathbf{p}})$ for modeltilpasning til en model med de fittede værdier $\widehat{\mathbf{p}}$ jf. (2.6.6) som

$$D(\mathbf{y};\widehat{\mathbf{p}}) = \sum_{i=1}^{k} n_i\, d(y_i;\widehat{p}_i) = \sum_{i=1}^{k} r_D(y_i;\widehat{p}_i)^2 \tag{3.1.8}$$

Pearson-residualet (2.5.39) svarende til observationen $y = z/n$ og den fittede værdi $\widehat{p}$ er

$$\begin{aligned} r_P(y;\widehat{p}) &= \frac{y-\widehat{p}}{\sqrt{\widehat{p}(1-\widehat{p})/n}} \\ &= \frac{z-n\widehat{p}}{\sqrt{n\widehat{p}(1-\widehat{p})}}\ , \end{aligned} \tag{3.1.9}$$

og Pearson-teststørrelsen (2.6.7) for modeltilpasning svarende til observationssættet $y_1, y_2, \ldots, y_k$ og de fittede værdier $\widehat{p}_1, \ldots, \widehat{p}_k$ bliver

$$X^2 = \sum_{i=1}^{k} r_D(y_i;\widehat{p}_i)^2 = \sum_{i=1}^{k} \frac{(z_i - n_i\widehat{p}_i)^2}{n_i\widehat{p}_i(1-\widehat{p}_i)}\ , \tag{3.1.10}$$

hvor $y_i = z_i/n_i$

3.2 Regressionsmodeller

Vi vil i dette afsnit betragte modeller for sættet $y_1, y_2, \ldots, y_k$ af observationer, når der yderligere foreligger et sæt $x_1, x_2, \ldots, x_k$ af kontinuerte kovariable, de forklarende variable.

Vi vil betragte modeller med den parametriske repræsentation

$$H_0 : \eta_i = \alpha + \beta x_i \ , i = 1, 2 \ldots, k \ , \tag{3.2.1}$$

hvor

$$\eta = g(p)$$

angiver linkfunktionen.

I toksikologiske undersøgelser benyttes sådanne modeller blandt andet til at beskrive indvirkningen af en toksisk dosis på den hændelse, at et forsøgsobjekt dør. I en sådan sammenhæng kan man opfatte "responssandsynligheden"

$$p(x) = g^{-1}(\eta(x)) \tag{3.2.2}$$

som en fordeling af tolerancer i populationen.

Antag nemlig, at hvert individ har en tolerance, T, for den pågældende dosis, sådan at individet dør, $Z = 1$, hvis dosis x overstiger T, dvs $[Y = 1] \equiv [T \leq x]$. Antag, at tolerancerne, T, varierer henover populationen i overensstemmelse med en fordeling med den kumulerede fordelingsfunktion

$$F_T(t) = \mathsf{P}\ [T \leq t] \ .$$

For en given dosis, x, er sandsynligheden $p(x)$, for at et tilfældigt udvalgt individ dør ved et forsøg, hvor individet udsættes for dosen x, bestemt ved

$$p(x) = \mathsf{P}\ [Z = 1] = F_T(x) = \mathsf{P}\ [T \leq x] \tag{3.2.3}$$

Valget af den kanoniske linkfunktion (logit'en) svarer således til en logistisk fordeling $p(x)$ af tolerancer med middelværdi $\mathsf{E}\ [T] = -\alpha/\beta$ og varians $\mathsf{V}\ [T] = \pi/(|\beta|\sqrt{3})$.

For at kunne beskrive denne fordeling ved en fordeling med kendt form og for at kunne give parametrene α og β en fortolkning i relation til fordelingen af tolerancer er der en række link-funktioner, der er specielle for binomialfordelte observationer (jf. eksempel 2.3.3). En række af disse linkfunktioner er netop valgt sådan, at den inverse funktion (3.2.2) har en fortolkning som en kendt fordeling.

3.2.1 Logistisk regression

Når linkfunktionen $\eta = g(p)$ er logit-transformationen (3.1.4), kaldes modellen (3.2.1) for en logistiske regressionsmodel.

Da logit-transformationen er den kanoniske linkfunktion bliver estimation særlig simpel under denne model, idet man blot skal løse middelværdiligningen.

Vi har allerede tidligere (i eksempel 2.4.1, 2.5.4 og 2.6.1) diskuteret den logistiske regressionsmodel. Vi fandt middelværdiligningen (2.5.34). Ligningen må sædvanligvis løses ved iteration. Når den logistiske regressionsmodel bruges i dosis-respons sammenhænge til beskrivelse af toxitet fortolkes parametrene α og β i den kanoniske parametrisering

$$p(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

ofte ved betragtning af oddsfunktionen.

Odds,

$$\theta(x) \stackrel{\text{DEF}}{=} \frac{p(x)}{1 - p(x)}$$

angiver forholdet mellem sandsynligheden for at et foster dør og at det overlever to dage efter injektionen.

Idet den kanoniske parameter ϑ netop er logaritmen til odds

$$\vartheta(x) = \ln \theta(x)$$

ser vi, at den logistiske regressionsmodel kan udtrykkes som

$$\theta(x) = \exp(\alpha + \beta x) = \mathrm{e}^{\alpha}(\mathrm{e}^{\beta})^{x}$$

dvs at hver gang dosen forøges med én enhed, multipliceres odds med faktoren e^{β}

Den dosis, der dræber halvdelen af de eksponerede individer, betegnes LD_{50}. Dosen estimeres ved

$$\hat{x}_{0.50} = -\frac{\hat{\alpha}}{\hat{\beta}}$$

Vi vil her anføre endnu et eksempel på en logistisk regressionsanalyse med henblik på at illustrere brugen af SAS-proceduren GENMOD (til analyse af generaliserede lineære modeller).

Eksempel 3.2.1 *Afprøvning af isolationsanordning, logistisk respons*

Tabel 3.1 viser for en række spændingstrin antallet af overslag (gnister) ved afprøvning af en luftformig isolationsanordning. Ved hvert spændingstrin blev der 100 gange foretaget en spændingspåvirkning med en pulsspænding. For hver påvirkning blev det registreret, hvorvidt der var overslag eller ej, hvorefter isolatoren blev retableret, sådan at man kan antage at de enkelte udfald er uafhængige.

$n = 100$ spændingspåvirkninger per trin.

Spændingstrin i	Spænding x_i kV	Antal overslag z_i	Overslagshyppighed y_i	Spændingstrin i	Spænding x_i kV	Antal overslag z_i	Overslagshyppighed y_i
1	1065	2	0.02	7	1100	29	0.29
2	1071	3	0.03	8	1107	48	0.48
3	1075	5	0.05	9	1111	56	0.56
4	1083	11	0.11	10	1120	88	0.88
5	1089	10	0.10	11	1128	98	0.98
6	1094	21	0.21	12	1135	99	0.99

Tabel 3.1. Realiseret overslagsspænding ved afprøvning af luftformig isolering på fastlagte spændingstrin

Som en kontrol af uafhængigheden mellem de enkelte spændingspåvirkninger bør man vurdere, om forekomsten af overslag svarende til et givet spændingsniveau er tilfældigt fordelt henover sekvensen af påvirkninger. Tabel 3.2 viser de enkelte forsøgsresultater svarende til trin 7 (1100 [kV]).

Vi vil betragte modellen med den parametriske repræsentation (3.2.1), dvs.

$$H_0 : \eta_i = \alpha + \beta x_i \ , i = 1, 2 \ldots , k \ ,$$

hvor $\eta = g(p)$. Modellen svarer til modelformlen

$$H_0 : \mathsf{Oversl} = \mathsf{Spend}$$

med linkfunktionen logit.

Tabel 3.2. Realiseret overslagsspænding ved afprøvning af luftformig isolering ved en spænding på 1100 [kV]
$n = 100$ spændingspåvirkninger per trin.

Nr	Resultat: ⋆ = overslag, ○ = intet overslag										relativ hyppighed
	1	2	3	4	5	6	7	8	9	0	
1 – 10	○	○	⋆	○	○	○	⋆	⋆	○	○	0.30
11 – 20	○	⋆	○	○	⋆	○	○	○	○	○	0.20
21 – 30	⋆	○	○	⋆	○	○	○	⋆	⋆	○	0.40
31 – 40	○	○	⋆	○	○	○	○	○	○	⋆	0.20
41 – 50	○	○	○	○	⋆	○	⋆	⋆	○	○	0.30
51 – 60	○	○	⋆	○	○	⋆	○	○	⋆	○	0.30
61 – 70	○	○	○	⋆	○	○	○	⋆	○	○	0.20
71 – 80	⋆	⋆	⋆	○	○	⋆	○	○	⋆	○	0.50
81 – 90	○	○	⋆	○	○	○	○	○	○	○	0.10
91 – 100	⋆	○	○	○	⋆	○	⋆	○	○	⋆	0.40
Ialt											0.29

Den statistiske analyse kan for eksempel udføres med SAS proceduren GENMOD.

Nedenstående SAS-program indlæser data og kalder proceduren

```
DATA GNIST;
INPUT Nr Spend Oversl Antal hyp;
CARDS;
  1   1065    2   100   0.02
  2   1071    3   100   0.03
  3   1075    5   100   0.05
  4   1083   11   100   0.11
  5   1089   10   100   0.10
  6   1094   21   100   0.21
  7   1100   29   100   0.29
  8   1107   48   100   0.48
  9   1111   56   100   0.56
 10   1120   88   100   0.88
 11   1128   98   100   0.98
 12   1135   99   100   0.99
  ;
*
* Procedure GENMOD ;
* ;
PROC GENMOD;
model Oversl/Antal= SPEND /DIST=BINOMIAL OBSTATS TYPE1 ;
run;
```

Ordren `OBSTATS` bevirker, at proceduren udskriver de fittede værdier `Pred` $= \widehat{p}$), samt de fittede værdier `'Xbeta` $= \widehat{\eta}$ af den lineære prædiktor.

Endvidere udskrives den estimerede standardafvigelse `Std` for den lineære prædiktor og vægten `HessWgt` i Hessian-matricen svarende til den sidste iteration.

Endelig udskrives konfidensgrænser for de fittede værdier af p samt `'Resraw`, de rå residualer $y_i - \widehat{p}_i$; `Reschi`, Pearson-residualerne $r_P(y_i;\widehat{p}_i)$ (3.1.9), og endelig `Resdev`, deviansresidualerne $r_D(y_i;\widehat{\mu}_i)$ (3.1.7).

Udskriften fra programmet er vist nedenfor:

```
Analyse af Overslagshyppigheder
```

Ved forskellige spaendinger

The GENMOD Procedure

Model Information

Description	Value
Data Set	WORK.GNIST
Distribution	BINOMIAL
Link Function	LOGIT
Dependent Variable	OVERSL
Dependent Variable	ANTAL
Observations Used	12
Number Of Events	470
Number Of Trials	1200

The GENMOD Procedure

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	10	21.0178	2.1018
Scaled Deviance	10	21.0178	2.1018
Pearson Chi-Square	10	20.1442	2.0144
Scaled Pearson X2	10	20.1442	2.0144
Log Likelihood	.	-422.3334	.

The GENMOD Procedure

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	-127.7002	7.0615	327.0274	0.0000
SPEND	1	0.1155	0.0064	325.9821	0.0000
SCALE	0	1.0000	0.0000	.	.

NOTE: The scale parameter was held fixed.

The GENMOD Procedure

LR Statistics For Type 1 Analysis

Source	Deviance	DF	ChiSquare	Pr>Chi
INTERCEPT	783.1215	0	.	.
SPEND	21.0178	1	762.1038	0.0000

The GENMOD Procedure

Observation Statistics

OVERSL	ANTAL	Pred	Xbeta	Std	HessWgt
2	100	0.00889345	-4.7135072	0.26393812	0.88143547
3	100	0.01762552	-4.0206245	0.22805281	1.73148653
5	100	0.02768733	-3.5587026	0.2046321	2.69207386
11	100	0.06692837	-2.634859	0.1597937	6.24489652
10	100	0.1254309	-1.9419763	0.12928732	10.9697993
21	100	0.20349792	-1.364574	0.10777856	16.2086515
29	100	0.33811825	-0.6716912	0.09053976	22.3794299
48	100	0.5341149	0.13667196	0.08905974	24.8836173
56	100	0.64533452	0.59859379	0.09787625	22.8877877
88	100	0.83725143	1.6379179	0.13444121	13.6261475
98	100	0.9283597	2.56176155	0.17660391	6.65079644
99	100	0.9667577	3.37012475	0.21670169	3.2137249

The GENMOD Procedure

Observation Statistics

Lower	Upper	Resraw	Reschi	Resdev
0.0053207	0.01482947	1.11065519	1.18299794	1.01629344
0.01134465	0.02728778	1.23744756	0.94041001	0.85548461
0.01871066	0.04079164	2.23126734	1.3599036	1.22455991
0.04982881	0.08934417	4.30716278	1.72356895	1.58697586
0.10016661	0.15596297	-2.5430905	-0.7678254	-0.7922914

```
0.17138884  0.23988103  0.65020823  0.16150241  0.16087179
0.29961206  0.37889605  -4.8118249  -1.0171507  -1.0302561
0.49053065  0.57718445  -5.4114904  -1.0848261   -1.083212
 0.6003107  0.68792316  -8.5334522  -1.7837043  -1.7571526
0.79809301  0.87005191  4.27485734  1.15807072   1.2046593
0.90164382  0.94823553  5.16402965   2.0024049   2.3465774
 0.9500455  0.97800828  2.32422995  1.29650664  1.51704293
```

Vi finder Pearsons χ^2-teststørrelse for modeltilpasning er $X^2 = 20.14$ med 10 frihedsgrader, svarende til 97 % fraktilen i χ^2-fordelingen. Devianstesteststørrelsen er $G^2 = 21.02$.

Tilpasningen er ikke overbevisende; en analyse af residualerne antyder en vis systematik, og man kunne derfor forsøge med en asymmetrisk linkfunktion. I eksempel 3.2.2 vil vi belyse tilpasningen ved brug af andre linkfunktioner.

Parametrene i den logistiske model estimeres til $\alpha = -127.7$ og $\beta = 0.1155$. Begge parametre ses at være stærkt signifikante. □

3.2.2 Regression ved andre link-funktioner

Vi vil nu betragte regressionsmodellen (3.2.1) for andre link-funktioner, $\eta = g(p)$, end logitfunktionen.

Vi bemærker indledningsvist, at logitfunktionen er symmetrisk omkring $p = 0.5$, dvs. for logitfunktionen gælder $g(p) = -g(1-p)$. Dette indebærer specielt, at responsfunktionen $p(x)$ nærmer sig nul med samme hastighed, som $p(x)$ nærmer sig 1.

I praksis kan det forekomme, at de observerede responser vokser langsomt væk fra nul, hvorimod de nærmer sig 1 ad et forholdsvis stejlt forløb.

I toksikologiske sammenhænge bruger man ofte responsfunktionen

$$p(x) = 1 - \exp[-\exp(\alpha + \beta x)] \tag{3.2.4}$$

svarende til den komplementære log-log linkfunktion

$$\ln\{-\ln[1 - p(x)]\} = \alpha + \beta x \tag{3.2.5}$$

Linkfunktionen (3.2.5) benævnes i en række programmer blot log-log link funktionen. Vi vil her betegne den mere præcist ved den komplementære log-log link funktion for at undgå forveksling med linkfunktionen (3.2.7).

For to værdier, x_1 og x_2, af den forklarende variable følger det af modellen (3.2.4), at

$$\frac{\ln[1-p(x_2)]}{\ln[1-p(x_1)]} = \exp[\beta(x_2-x_1)]$$

eller, at

$$1-p(x_2) = [(1-p(x_1)]^{\exp[\beta(x_2-x_1)]}$$

dvs såfremt den forklarende variable x forøges med én enhed, vil sandsynligheden for "negativt respons" blive opløftet til potensen $\exp(\beta)$

Responsfunktionen (3.2.4) kan også udtrykkes

$$p(x) = 1-\Lambda_1(-(\alpha+\beta x)) \ ,$$

hvor funktionen $\Lambda_1(\cdot)$ er den fra Statistik 1 kendte fordelingsfunktion for Max_1-fordelingen,

$$\Lambda_1(x) = \exp(-\mathrm{e}^{-x}) \ .$$

For $\beta > 0$ følger det derfor af bemærkningerne vedrørende ekstremværdifordelinger i Statistik 1, at fordelingen af tolerancer, T svarer til $T \in \text{Min}_1(-\alpha/\beta, 1/\beta)$, dvs $\mathsf{E}\,[T] = -(\alpha+\gamma)/\beta$, hvor γ er Eulers konstant, $\gamma \simeq 0.5772$, og $\mathsf{V}\,[T] = \pi^2/(6\beta^2)$.

Såfremt man i stedet har et forløb, hvor responset vokser stejlt væk fra nul, kan man benytte responsfunktionen

$$p(x) = \exp[-\exp(\alpha+\beta x)] \tag{3.2.6}$$

svarende til log-log linkfunktionen

$$\ln\{-\ln[p(x)]\} = \alpha+\beta x \tag{3.2.7}$$

For $\beta < 0$ vil (3.2.7) ligeledes være en voksende funktion af x.

Regressionsmodellen bestemt ved (3.2.6) kaldes undertiden en Gumbel-regression (idet extremværdifordelingen Max_1-fordelingen undertiden betegnes en Gumbelfordeling).

Det er klart, at såfremt $p(x)$ kan beskrives ved modellen (3.2.4), kan $1-p(x)$ beskrives ved (3.2.6) og omvendt. En komplementær log-log link for p er således blot en direkte log-log link for $1-p$.

Parametrene i modellen må estimeres ved at løse likelihoodligningen ved en iterativ metode.

Eksempel 3.2.2 ***Afprøvning af isolationsanordning, komplementær log-log respons***

Vi betragter atter data fra eksempel 3.2.1.

I eksempel 3.2.1 undersøgte vi tilpasningen ved en logistisk funktion svarende til at vi brugte logit linkfunktionen.

Da responset tilsyneladende steg langsommere i starten, og hurtigere mod slutningen, end det, der kunne tilpasses ved den logistiske funktion, vil vi forsøge at modellere data ved den komplementære log-log link (3.2.5) svarende til responsfunktionen (3.2.4)

$$p(x) = 1 - \exp[-\exp(\alpha + \beta x)]$$

I programsystemet "S-plus" vil kommandoen

glm(formula = cbind(oversl, ant - oversl) spend, family = binomial(cloglog))

resultere i udskriften:

```
Coefficients:
 (Intercept)      spend
   -91.10633 0.08190004

Degrees of Freedom: 12 Total; 10 Residual
Residual Deviance: 5.670954
```

dvs. estimaterne $\widehat{\alpha} = -91.10633$ og $\widehat{\beta} = 0.08190$, og en teststørrelse for modeltilpasning

$$G^2(H_0) = D(\mathbf{y}; \widehat{\mathbf{p}}) = 5.67\,,$$

Altså en klart bedre modeltilpasning, end ved den logistiske regression.

De fittede værdier, $\widehat{p}$, svarende til denne model er angivet nedenfor

OBS	SPEND	HYP	PRED
1	1065	0.02	0.02038262
2	1071	0.03	0.03310149
3	1075	0.05	0.04563605
4	1083	0.11	0.08601559
5	1089	0.10	0.1367225
6	1094	0.21	0.1986202
7	1100	0.29	0.303672
8	1107	0.48	0.4738202
9	1111	0.56	0.5897616
10	1120	0.88	0.844655
11	1128	0.98	0.9722783
12	1135	0.99	0.9982726

Devianstesteststørrelsen for modeltilpasning

$$G^2(H_0) = D(\mathbf{y}; \widehat{\mathbf{p}}) = 5.67\,,$$

skal sammenlignes med en $\chi^2(10)$ fordeling. Teststørrelsen svarer til 15 % fraktilen. Der er således ingen grund til afvisning af hypotesen. En vurdering af residualerne viser heller ingen tegn på systematik.

Usikkerhederne på parameterestimaterne udskrives også af ovenstående kommando. Man får

```
Coefficients:
                Value  Std. Error   t value
(Intercept) -91.10633347 4.601828609 -19.79785
      spend   0.08190004 0.004147065  19.74891

(Dispersion Parameter for Binomial family taken to be 1 )
```

Begge estimater er altså klart signifikant forskellige fra nul.

Testet for, hvorvidt koefficienten $\beta = 0$, kunne også udføres ved at estimere under hypotesen svarende til den minimale model $H_M : p_1 = \cdots = p_k$.

Under denne hypotese finder man testet for modeltilpasning $G^2(H_M) = D(\mathbf{y}; \widehat{\widehat{\mathbf{p}}}) = 783.1215$ svarende til 11 frihedsgrader.

Teststørrelsen for modelreduktion fås da af deviansanalyseskemaet

Variation	Devians	f
β	777.451	1
Residual	5.671	10
Total	783.122	11

Vi kan altså ikke reducere modellen yderligere.

For illustrationens skyld vil vi også illustrere tilpasningen til modellen (3.2.6) svarende til den direkte log-log link (for p).

Man finder estimaterne (ved itertiv løsning af likelihood ligningen):

```
Coefficients:
 (Intercept)       spend
    65.15193 -0.05936724

Degrees of Freedom: 12 Total; 10 Residual
Residual Deviance: 80.12935
```

dvs $\widehat{\alpha} = 65.15193$ og $\widehat{\beta} = -0.059367$.

Estimatet $\widehat{\beta}$ er negativt, svarende til at responsfunktionen er en voksende funktion af spændingen x.

Vi ser, at tilpasningen er ekstremt ringe. Teststørrelsen for modeltilpasning er

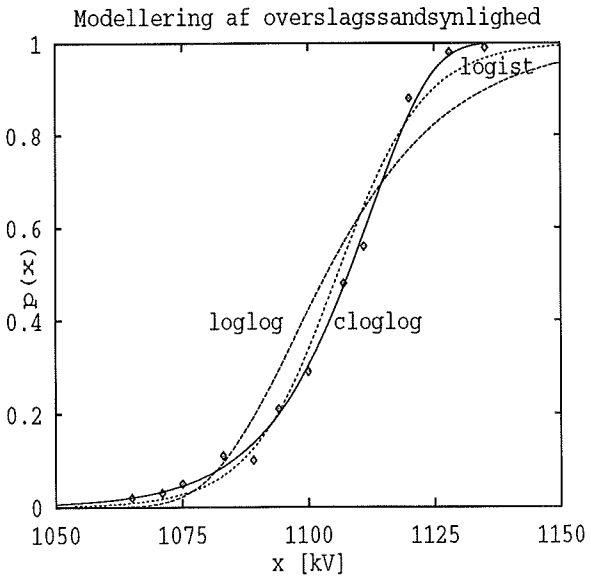
$$G^2(H_0) = D(\mathbf{y}; \widehat{\mathbf{p}}) = 80.12935$$

Til sammenligning ser vi, at $\chi^2(10)_{0.9995} = 31.4$. Hypotesen må klart afvises, selv ved test på et 0.05 % niveau.

Til sammenligning med de øvrige modeller har vi vist de fittede værdier under denne model i nedenstående tabel. Man ser klare systematiske tendenser i afvigelserne. Modellen vil gerne vokse hurtigere i starten, og langsommere til slut, end data indikerer.

OBS	SPEND	HYP	PRED
1	1065	0.02	0.00010
2	1071	0.03	0.00819
3	1075	0.05	0.02261
4	1083	0.11	0.09475
5	1089	0.10	0.19198
6	1094	0.21	0.29332
7	1100	0.29	0.42360
8	1107	0.48	0.56729
9	1111	0.56	0.63951
10	1120	0.88	0.76951
11	1128	0.98	0.84964
12	1135	0.99	0.89804

Nedenstående figur illustrerer tilpasningen ved de tre modeller.



Vi gør endelig opmærksom på betydningen af at have observationer i hele

det relevante variationsområde for den forklarende variable x, såfremt man skal have mulighed for at sondre mellem forskellige modeller. Hvis vi eksempelvis kun havde haft observationer svarende til spændingsniveauer $\leq$ 1090 kV ville vi formentlig ikke kunne have skelnet mellem de tre betragtede modeller, da de alle tre vil kunne tilpasses rimelig godt til forløbet i et begrænset p-interval. □

3.2.3 Regressionsmodeller med flere forklarende variable

Såfremt der foreligger flere forklarende variable, for eksempel x_{i1} og x_{i2} for hver y-værdi, kan vi benytte betragtningerne i afsnit 2.11.4 til at successivt at vurdere tilpasningen af en kæde af modeller.

Vi vil illustrere fremgangsmåden ved et eksempel.

Eksempel 3.2.3 ***Tidsmæssig udvikling i fødselsrisiko***

Nedenstående tabel viser for årene 1963 - 1976 antallet af fødsler, samt antallet af dødfødte børn blandt disse. (Kilde: Statistiske meddelelser 1978:1, Danmarks Statistik).

Ugifte mødre		
År	Ant. fødsl	heraf døde
x	n	z
1963	7420	103
1964	7877	93
1965	8223	111
1966	9085	102
1967	9114	92
1968	8395	85
1969	8145	92
1970	7898	87
1971	9370	85
1972	10980	103
1973	12419	93
1974	13515	104
1975	15792	129
1976	15795	123

Gifte mødre		
År	Ant. fødsl	heraf døde
x	n	z
1963	75941	845
1964	76385	813
1965	78815	831
1966	80122	773
1967	73011	623
1968	66784	551
1969	63764	520
1970	63508	517
1971	66609	535
1972	65102	474
1973	59999	430
1974	58253	337
1975	56762	354
1976	49903	308

Hvis man tegner andelen $y_i = z_i/n_i$ op mod årene ser man en klart faldende tendens med tiden, såvel for børn født af ugifte mødre, som børn født af gifte mødre.

Der ses en svag krumning i kurveforløbet, så det er naturligt at forsøge med en logistisk regressionsmodel for andelen af dødfødte børn. En optegning af de empiriske logit'er, $\ln(y_i/(1-y_i))$ mod årene indikerer også, at en lineær model for logit'erne kan være rimelig.

Man opstiller da modellen

$$H_0 : \mathsf{død} = \mathsf{stand} + \mathsf{år.stand}$$

med linkfunktionen logit. Modellen svarer til den parametriske model

$$H_0 : \eta_{i,j} = \alpha_j + \gamma_j x_i \ ,$$

hvor $j = 1, 2$ angiver civilstanden, $j = 1$ angiver ugift, og $j = 2$ angiver gift. Modellen angiver således en intercept-parameter α_j og en hældningsparameter, γ_j, (vekselvirkningsleddet mellem år og civilstand) for hver værdi af civilstanden.

Modellen svarer til at estimere to separate linier. Modellen har dimensionen $m = 4$ Man finder estimaterne (estimaterne må bestemmes ved iteration):

civilstand	$\widehat{\alpha}$	$\widehat{\beta}$
ugift	83.2876	-0.044622
gift	91.4479	-0.048872

og teststørrelsen for modeltilpasning (da dispersionsparameteren er 1, behøver man ikke skalere residualdeviansen),

$$G^2(H_0) = D(\mathbf{y}; \widehat{\mathbf{p}}) = 24.1119 ,$$

der skal sammenlignes med fraktilerne i en $\chi^2(24)$ fordeling. Da $\chi^2(24)_{0.60} = 25.1$ er der ingen grund til at afvise hypotesen om linearitet (dvs to linier).

Da de to skøn over hældningerne er af nogenlunde samme størrelse, forsøger vi om modellen kan reduceres til den simplere model, der udtrykker, at den årlige reduktion i dødsrisiko (odds) er den samme for de to grupper, svarende til $\gamma_1 = \gamma_2$.

Vi formulerer derfor hypotesen:

$$H_1 : \mathsf{d\o d} = \mathsf{stand} + \mathsf{\mathring{a}r}$$

svarende til den parametriske model

$$H_1 : \eta_{i,j} = \alpha_j + \beta x_i .$$

Modellen har dimensionen $r = 3$. Modellen er en delmodel af modellen svarende til H_0.

Vi skal reestimere alle tre parametre i denne model. Man finder estimaterne

civilstand	$\widehat{\alpha}$	$\widehat{\gamma}$
ugift	90.0256	-0.04814972
gift	90.2357	-0.04814972

og teststørrelsen for modeltilpasning,

$$G^2(H_1) = D(\mathbf{y}; \widehat{\widehat{\mathbf{p}}}) = 24.4748$$

Reduktionen i modeltilpasning,

$$G^2(H_1|H_0) = G^2(H_1) - G^2(H_0) = 0.363$$

skal sammenlignes med en $\chi^2(1)$ fordeling.

Da $\chi^2(1)_{0.50} = 0.455$ er der ingen grund til at afvise hypotesen om en fælles hældning.

De udførte tests svarer til deviansanalyseskemaet

Variation	Devians	f
Mellem hældninger	0.363	1
Omkring individuelle linier	24.112	4
Omkring model med fælles hældning	24.475	25

Den minimale model svarende til de betragtede variable er en model, hvor dødsfødselsrisikoen er konstant.

En hierarkisk organiseret følge af modeller er

$$H_M = 1 \subset \texttt{år} \subset \texttt{år + stand} \subset \texttt{stand +stand.år} \subset H_F$$

I dette tilfælde er der netop en forskel i dimension på 1 imellem hypoteserne, dvs. hver af devianstesttørrelserne har netop én frihedsgrad.

Deviansanalyseskemaet svarende til dette hierarki af modeller er

Variation	Devians	f
Mellem år	309.421	1
Mellem afskæringer	48.883	1
Mellem hældninger	0.363	1
Omkring individuelle linier	24.112	24
Total	382.779	27

Nedenstående tabel viser et eksempel på et typisk output fra et statistisk programsystem (her S-plus) svarende til denne analyse:

```
Terms added sequentially (first to last)
         Df Deviance Resid. Df Resid. Dev
NULL                      27     382.779
```

```
          aar  1  309.421        26     73.357
        stand  1   48.883        25     24.475
   stand:aar   1    0.363        24     24.112
```

□

3.3 Faktorielle opstillinger med binært respons

3.3.1 Opstillinger med to faktorer

Såfremt forsøgsfaktorerne er A og B, med index henholdsvis $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$, og antallene af forsøgsenheder under faktorkombination (i, j) er $n_{i,j}$ beskrives resultatet ofte ved en tre-dimensional antalstabel, hvor $x_{i,j,1}$ angiver antallet af enheder med respons i kategori 1, og $x_{i,j,2} = n_{i,j} - x_{i,j,1}$ angiver antallet af enheder med respons i den anden kategori.

Frekvensfunktionen bliver her

$$f(\underline{x}) = \prod_{i,j} \left(\begin{array}{c} n_{i,j} \\ x_{i,j} \end{array} \right) p_{i,j}^{x_{i,j}} (1-p)_{i,j}^{n_{i,j}-x_{i,j}} \tag{3.3.1}$$

Log-likelihoodfunktionen bliver tilsvarende

$$\begin{aligned} l(\underline{p}) &= \Sigma_{i,j} x_{i,j} \ln(p_{i,j}) + (n_{i,j} - x_{i,j}) \ln(1 - p_{i,j})) \\ &= \Sigma_{i,j} n_{i,j} \ln(1 - p_{i,j}) + \Sigma_{i,j} x_{i,j} \ln(p_{i,j}/(1 - p_{i,j})) \end{aligned} \tag{3.3.2}$$

Indfører vi den *kanoniske parameter* ved *logittransformationen*

$$\vartheta_{i,j} = \ln(p_{i,j}/(1 - p_{i,j})) \tag{3.3.3}$$

får vi log-likelihoodfunktionen

$$l(\underline{\vartheta}) = \Sigma_{i,j} x_{i,j} \vartheta_{i,j} - \Sigma_{i,j} n_{i,j} \ln(1 + \exp(\vartheta_{i,j})) \tag{3.3.4}$$

Eksempel 3.3.1 *Tofaktor forsøg med binomialt respons*

Vi betragter atter data fra eksempel 2.4.2.

Forsøget havde til formål at vurdere, hvorledes overlevelsesevnen afhænger af tykkelsen og længden af de udplantede stiklinger.

Ved forsøget betragtede man to længder, henholdsvis 6 [cm] og 12 [cm] og tre tykkelsesgrader, 3-6 [mm], 6-9 [mm] og 9-12 [mm], I oktober udplantede man 20 stiklinger af hver kombination af længde og tykkelse, og året efter vurderede man, hvor mange af disse, der stadig var i live.

Resultaterne i tabellen er angivet som 'antal overlevende/antal udplantede'

	længde	
	lang	kort
Tykkelse	12 cm	6 cm
tynd: 3-6 mm	6/20	4/20
mid: 6-9 mm	14/20	10 /20
tyk: 9-12 mm	18 /20	11 / 20

Stiklingerne blev afskåret som rodstumper af ældre træer i oktober 1931, og i oktober det følgende år blev det vurderet, hvorvidt stiklingerne havde overlevet.

Vi vælger som tidligere at modellere antallet $Z_{i,j}$ af overlevende stiklinger ved en $\mathrm{B}(n_{i,j}, p_{i,j})$-fordelt størrelse, hvor antallet af stiklinger, $n_{i,j} = 20$.

Vi betragtede en logistisk link:

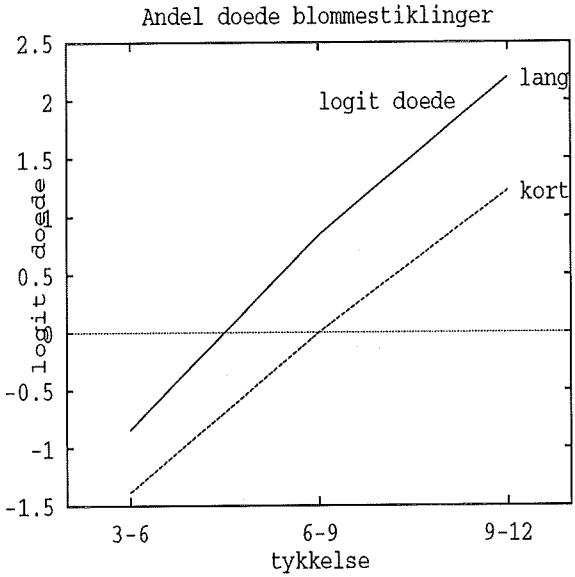
$$\eta(p) = \ln\left(\frac{p}{1-p}\right)$$

og vi formulerede modellen svarende til forsvindende vekselvirkning:

$$H_0: \quad \eta_{i,j} = \kappa + \alpha_i + \gamma_j$$

Hypotesen blev vurderet numerisk i eksempel 2.4.2. Man fandt teststørrelsen for modeltilpasning $G^2(H_0) = 1.88$ med 2 frihedsgrader.

Vi vil her supplere med at tegne et profilplot for logit'erne for at supplere dette test.



Under hypotesen

$$\eta_{i,j} = \kappa + \alpha_i + \beta_j$$

finder man de fittede værdier og de tilhørende deviansresidualer:

LEV	ANT	Pred	Resdev
6	20	0.34654756	-0.442544
14	20	0.72544766	-0.2526779
18	20	0.82800464	0.91233047
4	20	0.15345248	0.55647454
10	20	0.47455233	0.22775851
11	20	0.62199522	-0.657132

Estimaterne er bestemt i eksempel 2.4.2. Man fandt

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi

INTERCEPT		1	-0.6342	0.4048	2.4548	0.1172
LENG	KO	1	-1.0735	0.4211	6.4992	0.0108
LENG	LA	0	0.0000	0.0000	.	.
TYK	MID	1	1.6059	0.5082	9.9849	0.0016
TYK	TYK	1	2.2058	0.5335	17.0977	0.0000
TYK	TYN	0	0.0000	0.0000	.	.
SCALE		0	1.0000	0.0000	.	.

NOTE: The scale parameter was held fixed.

I eksempel 2.11.5 betragtede vi mulighederne for reduktion af modellen. Med udgangspunkt i de partielle test fandt vi, at der ikke var grundlag for at reducere modellen yderligere.

Vi vil derfor ikke diskutere denne reduktion yderligere, men i stedet betragte den resulterende additive model for logit'erne.

Nedenstående tabel viser de estimerede logit-parametre svarende til det valgte referenceniveau (Lang, Tynd). Vi ser, at logit'en for overlevelse mindskes hvis længden mindskes, og logit'en øges, når tykkelsen øges.

Estimerede kontraster svarende til logit model uden vekselvirkning

	Ref	længde	
	niveau	lang	kort
		12 cm	6 cm
	-0.6342	0.	-1.0735
Tykkelse			
tynd: 3-6 mm	0.		
mid: 6-9 mm	1.6059		
tyk: 9-12 mm	2.2058		

Nedenstående tabel viser de tilsvarende odds-kontraster (odds=exp(logit)).

Estimerede værdier af odds-kontraster svarende til logit model uden vekselvirkning

	Ref niveau 0.53	længde lang 12 cm 1.	kort 6 cm 0.34
Tykkelse			
tynd: 3-6 mm	1.0		
mid: 6-9 mm	4.98		
tyk: 9-12 mm	9.08		

For referencegruppen (Lang,Tynd) er odds for overlevelse $\omega = 0.53$. For korte stiklinger reduceres odds med faktoren 0.34. For middeltykke stiklinger øges odds med faktoren 5 (nemlig 4.98), og for tykke øges odds yderligere med faktoren 9.08 i forhold til de tynde, dvs. en yderligere øgning i forhold til de middeltykke med faktoren 1.82.

I stedet for log odds vil vi nu betragte de tilsvarende overlevelsessandsynligheder

$$\widehat{p}_{i,j} = \frac{\exp(\widehat{\kappa} + \widehat{\alpha}_i + \widehat{\beta}_j)}{1 + \exp(\widehat{\kappa} + \widehat{\alpha}_i + \widehat{\beta}_j)}$$

Ved indsættelse af estimaterne finder man

Fittede værdier af overlevelsessandsynligheder $\widehat{p}_{i,j}$ svarende til logit-model uden vekselvirkning

Tykkelse	længde lang 12 cm	kort 6 cm
tynd: 3-6 mm	0.35	0.15
mid: 6-9 mm	0.72	0.47
tyk: 9-12 mm	0.82	0.62

Vi lægger mærke til, at forholdet mellem overlevelsessandsynligheder for to tykkelser afhænger af længden. Således er

$$p_{mid,lang}/p_{tynd,lang} = \frac{0.72}{0.35} = 2.06$$

men

$$p_{mid,kort}/p_{tynd,kort} = \frac{0.47}{0.15} = 3.13$$

□

3.3.2 Vekselvirkning og valg af linkfunktion

Brug af den kanoniske link for en tofaktormodel uden vekselvirkning svarer til en additiv model for log odds.

Vi så i eksempel 3.3.1, at den resulterende model for p'erne ikke gav anledning til simple relationer direkte mellem p'erne.

De differentielle effekter (2.12.2)

$$\Delta^A_{i,i';j} = \eta_{i,j} - \eta_{i',j}$$

bliver her differenser mellem logit'er, dvs

$$\Delta^A_{i,i';j} = \ln\left(\frac{p_{i,j}}{(1-p_{i,j})}\right) - \ln\left(\frac{p_{i',j}}{(1-p_{i',j})}\right) \tag{3.3.5}$$

eller

$$\Delta^A_{i,i';j} = \ln\left(\frac{p_{i,j}(1-p_{i',j})}{(1-p_{i,j})p_{i',j}}\right) = \ln(\omega^A_{i,i';j}) \ , \tag{3.3.6}$$

hvor vi har indført den differentielle effekt på odds-skalaen

$$\omega^A_{i,i';j} \stackrel{\text{DEF}}{=} \frac{p_{i,j}/(1-p_{i,j})}{p_{i'j}/(1-p_{i'j})} \tag{3.3.7}$$

Den betragtede additive model for logit'erne er ensbetydende med at de tilsvarende differentielle effekter $\Delta^A_{i,i';j}$ eller $\omega^A_{i,i';j}$ ikke afhænger af j, eller analogt, at de differentielle effekter $\Delta^B_{i;j,j'}$ eller $\omega^B_{i;j,j'}$ ikke afhænger af i.

Vi bemærker, at modellen baseret på logit'erne er symmetrisk i p og $1-p$. Hvis der gælder en additiv model for logit(p), da gælder der også en additiv model for logit$(1-p) = -$logit(p).

Logaritmisk Link

Undertiden kan det være mere relevant at betragte en anden linkfunktion. Antag f.eks. at man har et system bestående af to komponenter A og B, som er serieforbundne. Systemet fejler, hvis blot en af komponenterne fejler. Lad p_A angive sandsynligheden for at A-komponenten er fejlfri, og p_B

sandsynligheden for at B-komponenten er fejlfri. Hvis de to komponenter fejler uafhængigt af hinanden gælder

$$\mathrm{P}\ [\text{System fejlfri}] = p_A p_B$$

Antag nu, at A foreligger i r varianter, A_i, $i = 1, 2, \ldots, r$ og B foreligger i s varianter, B_j, $j = 1, 2, \ldots, s$, hver med sin fejlsandsynlighed. Under antagelse af at der er uafhængighed mellem fejl i A og i B, da er sandsynligheden for at et system bestående af en A_i og en B_j komponent er fejlfrit

$$\mathrm{P}\ [\text{System fejlfri}] = p_{A_i} p_{B_j}$$

Betragt nu et eksperiment, hvori $n_{i,j}$ apparater af typen A_i, B_j undersøges for fejl. Antallet $Z_{i,j}$ af fejlfrie apparater vil da kunne beskrives ved en $\mathrm{B}(n_{i,j}, p_{i,j})$-fordelt variabel. Hypotesen om uafhængighed mellem A og B svarer til den additive model for $\ln(p_{i,j})$

$$\eta_{i,j} = \ln(p_{i,j}) = \kappa + \alpha_i + \beta_j \tag{3.3.8}$$

dvs en generaliseret lineær model for $Y_{i,j} = Z_{i,j}/n_{i,j}$ med en logaritmisk linkfunktion.

Hypotesen om uafhængighed svarer altså til hypotesen om forsvindende vekselvirkning i denne model.

De differentielle effekter svarende til (3.3.8) er

$$\Delta^A_{i,i';j} = \ln\left(\frac{p_{i,j}}{p_{i',j}}\right) = \ln(\omega^A_{i,i';j}) \ , \tag{3.3.9}$$

hvor ratioen

$$\omega^A_{i,i';j} \overset{\mathrm{DEF}}{=} \frac{p_{i,j}}{p_{i'j}} \tag{3.3.10}$$

angiver forholdet mellem sandsynlighederne $p_{i,j}$ og $p_{i'j}$.

Den anden ordens differentielle effekt er

$$\Delta^{AB}_{i,i';j,j'} \overset{\mathrm{DEF}}{=} \Delta^B_{i;j,j'} - \Delta^B_{i';j,j';j} = \ln\left(\frac{p_{i,j}p_{i',j'}}{p_{i,j'}p_{i',j}}\right) = \ln(\omega^{AB}_{i,i';j,j'}) \ , \tag{3.3.11}$$

hvor den tilsvarende differentielle effekt for sandsynlighederne er

$$\omega^{AB}_{i,i';j,j'} \overset{\mathrm{DEF}}{=} \frac{p_{i,j}}{p_{i,j'}} \Big/ \frac{p_{i'j}}{p_{i',j'}} \tag{3.3.12}$$

Eksempel 3.3.2 *Tofaktor forsøg med binomialt respons, logaritmisk link*

Nedenstående data hidrører fra en amerikansk undersøgelse foretaget af Lombard & Doering (1947) af sammenhænge mellem folks viden om cancer og forskellige baggrundsvariable.

Ved undersøgelsen blev svarene fra 1729 personer klassificeret efter personernes eksponering overfor forskellige medier samt deres viden om cancer.

Nedenstående tabel viser svarene for 759 personer, som kun havde været udsat for avislæsning og/eller "solid reading" sammen med svarene fra 477 personer, som slet ikke havde været eksponeret overfor nogle medier.

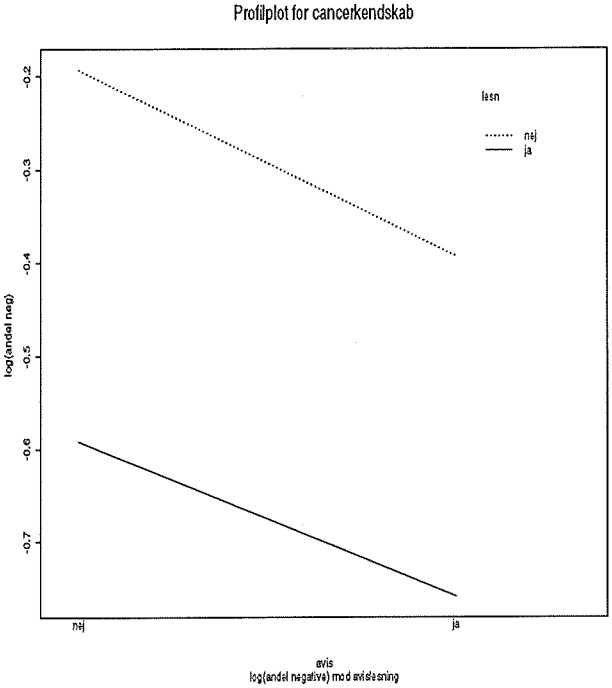
Antal personer med ringe viden om cancer/antal udspurgte

	læsning	
Avislæsn.	Nej	Ja
Nej	393/477	83/150
Ja	156/231	177/378

Det kan her være naturligt at forsøge at formulere en multiplikativ model for sandsynligheden for at have ringe viden om cancer i afhængighed af informationskilderne. (Rationalet bag denne model er - i lighed med betragtningen af det serielle system - at ringe viden kræver såvel ringe udbytte af sædvanlig læsning som ringe udbytte af avislæsning. Har man fået udbytte af én af disse kilder, gør det jo ikke så meget, at man ikke får noget ud af den anden.

Nedenstående figur viser et profilplot for logaritmen til andelen af personer med ringe viden om cancer.

Det ses, at de to linier stort set er parallelle. Der er altså ingen tegn på vekselvirkning svarende til denne link funktion.



Man formulerer hypotesen

$$\ln(p_{i,j}) = \kappa + \alpha_i + \beta_j \tag{3.3.13}$$

Man finder estimaterne :

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	-0.1950	0.0209	87.4668	0.0000
AVIS	J	1	-0.1915	0.0436	19.3178	0.0000
AVIS	N	0	0.0000	0.0000	.	.
LAES	J	1	-0.3813	0.0515	54.7124	0.0000
LAES	N	0	0.0000	0.0000	.	.
SCALE		0	1.0000	0.0000	.	.

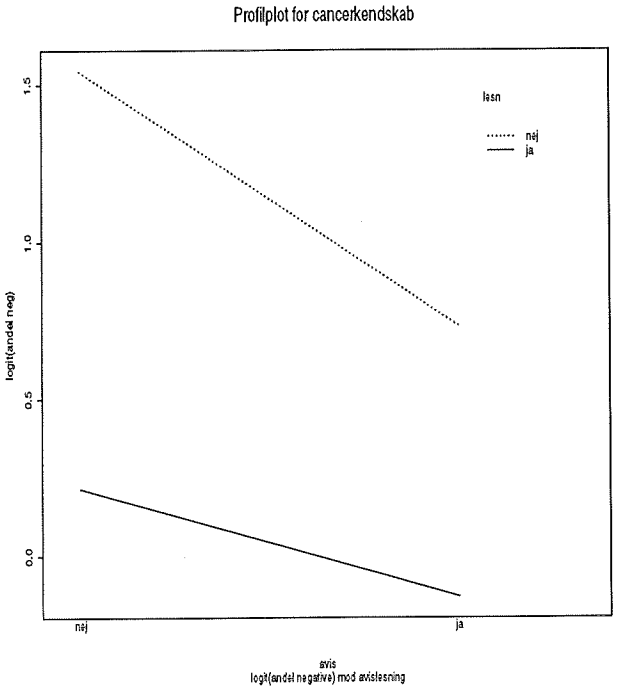
NOTE: The scale parameter was held fixed.

og en devianstesteststørrelse for modeltilpasning til hypotesen (3.3.13), $D(\mathbf{y}; \widehat{\mathbf{p}}) = 0.0941$ Der er altså ingen grund til at afvise hypotesen.

Såfremt man havde valgt at betragte en additiv model for logit'erne

$$\ln(p_{i,j}/(1-p_{i,j})) = \kappa + \alpha_i + \beta_j \tag{3.3.14}$$

havde man fundet nedenstående profilplot:



Estimaterne under denne hypotese er

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	1.4518	0.1053	190.1955	0.0000
AVIS	J	1	-0.5880	0.1347	19.0583	0.0000
AVIS	N	0	0.0000	0.0000	.	.

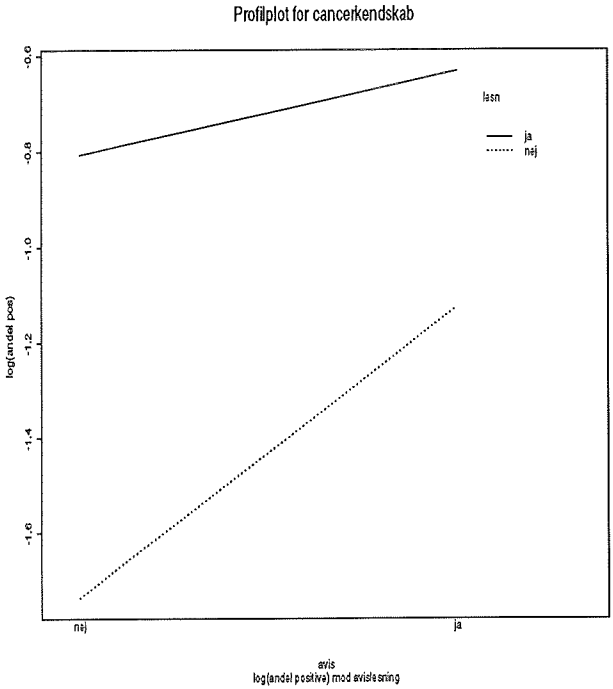
```
LAES        J     1      -1.0602       0.1336      62.9474   0.0000
LAES        N     0       0.0000       0.0000            .        .
SCALE             0       1.0000       0.0000            .        .

NOTE:  The scale parameter was held fixed.
```

og devianstesttørrelsen for modeltilpasning til denne hypotese er $D(\mathbf{y};\widehat{\mathbf{p}}) = 3.0741$, altså en væsentlig ringere tilpasning, hvilket også fremgår af profilplottet.

Endelig - for illustrationens skyld - vil vi betragte en multiplikativ model i sandsynligheden for at have godt kendskab til cancer.

$$\ln(1-p_{i,j}) = \kappa + \alpha_i + \beta_j \tag{3.3.15}$$

Profilplottet svarende til denne hypotese er vist nedenfor:



Også under denne model ses en klar tendens til vekselvirkning.

Man finder estimaterne

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	-1.6113	0.0786	420.1369	0.0000
AVIS	J	1	0.3420	0.0900	14.4505	0.0001
AVIS	N	0	0.0000	0.0000	.	.
LAES	J	1	0.6687	0.0902	54.9054	0.0000
LAES	N	0	0.0000	0.0000	.	.
SCALE		0	1.0000	0.0000	.	.

NOTE: The scale parameter was held fixed.

Teststørrelsen for modeltilpasning til denne model er $D(\mathbf{y};\widehat{\mathbf{p}}) = 6.2690$, altså en klar forkastelse.

Dvs at den simpleste model (en model uden vekselvirkning) fås ved at modellere $\ln(p)$. Modellen har den yderligere fordel, at den kan fortolkes ved binær additionsregel for kendskab.

Den multiplikative model for p (den additive model for $\ln(p)$) svarer til at krydsprodukt ratioen er 1, nemlig til proportionalitet mellem andelene uden cancerkendskab:

$$p_{1,1}p_{2,2} = p_{1,2}p_{2,1}$$

□

3.3.3 Yule's krydsprodukt ratio og betingede odds

Definition 3.3.1 ***Yule's krydsprodukt ratio for hændelser***

Betragt to hændelser A og B samt de komplementere hændelser A^c og B^c.

Størrelsen

$$\text{cpr}(A,B) = \frac{\mathsf{P}\ [A \cap B]}{\mathsf{P}\ [A^c \cap B]} \frac{\mathsf{P}\ [A^c \cap^c B]}{\mathsf{P}\ [A \cap B^c]} \tag{3.3.16}$$

kaldes Yule's krydsprodukt ratio for hændelserne A og B.

Hændelserne er illustreret i nedenstående skema

	B	B^c
A	$A \cap B$	$A \cap B^c$
A^c	$A^c \cap B$	$A^c \cap B^c$

□

Bemærkning 1 ***Krydsproduktratioen for hændelser er forholdet mellem de betingede odds***

Indfører vi de betingede odds for A givet B som

$$\theta(A|B) = \frac{\mathsf{P}\ [A|B]}{\mathsf{P}\ [A^c|B]} \tag{3.3.17}$$

ser man, at Yule's krydsproduktratio kan udtrykkes som

$$\mathrm{cpr}(A,B) = \frac{\mathsf{P}\ [A|B]}{\mathsf{P}\ [A^c|B]} \frac{\mathsf{P}\ [A^c|B^c]}{\mathsf{P}\ [A|B^c]} = \frac{\theta(A|B)}{\theta(A|B^c)}$$

Yules krydsproduktratio for hændelser kan således udtrykkes som forholdet mellem de betingede odds.

Vi bemærker i øvrigt, at krydsproduktratioen er symmetrisk i hændelserne A og B. Der gælder således

$$\mathrm{cpr}(A,B) = \frac{\theta(B|A)}{\theta(B|A^c)}$$

□

3.3.4 Rasch model for itemanalyse, latente parametre

Den såkaldte Rasch-model for item analyse har sin oprindelse i beskrivelse af svarene for en gruppe personer $i = 1, 2, \ldots, r$, der udsættes for et batteri af prøver, $j = 1, \ldots, s$, hvor svaret på den enkelte prøve kan klassificeres som rigtigt eller forkert.

Sættet af svar beskrives ved et tosidet skema af uafhængige Bernoulli-variable

$$Y_{i,j}, \quad i = 1, 2, \ldots, r;\ j = 1, 2, \ldots, s$$

med

$$\mathsf{P}\ [Y_{i,j} = 1] = p_{i,j}, \quad 0 < p_{i,j} < 1$$

$r \times s$-matricen $\mathbf{Y}$ af observationer følger en eksponential dispersionsparameterfamilie af orden $r \times s$ med kanoniske parametre $\vartheta_{i,j} = \log(p_{i,j}/(1-p_{i,j}))$, og med kanonisk parameterområde $D = \mathbb{R}^{r \times s}$.

Rasch-modellen (Rasch (1960)) er den generaliserede lineære model bestemt ved

$$H_0 : \quad \eta_{i,j} = \alpha_i + \beta_j$$

hvor η er den lineære prædiktor svarende til den kanoniske link, (logitfunktionen)

$$\eta = g(p) = \ln\left(p_{i,j}/(1 - p_{i,j})\right)$$

Estimaterne under denne hypotese bestemmes ved løsning af middelværiligningen

$$y_{i+} = \exp(\alpha_i) \sum_j \exp(\beta_j)/[1 + \exp(\alpha_i + \beta_j)]$$

$$y_{+j} = \exp(\beta_j) \sum_i \exp(\alpha_i)/[1 + \exp(\alpha_i + \beta_j)] \; ;,$$

hvor $y_{i+} = \sum_j y_{i,j}$ angiver rækkesummen svarende til den i'te række, og tilsvarende $y_{+j} = \sum_i y_{i,j}$ angiver søjlesummen svarende til den j'te søjle

Vi bemærker, at sættet af række-og søjlesummer er sufficiente for parametrene $\alpha_1, \ldots, \alpha_r; \beta_1, \ldots, \beta_s)$.

Endvidere bemærker vi, at den betingede fordeling af observationerne for givet rækkesum y_{i+} ikke afhænger af parameteren α_i.

Sædvanligvis vil batteriet af spørgsmål, s være væsentligt mindre, end antallet af testpersoner. Modelantagelsen kan da kontrolleres ved at gruppere personerne efter deres score (antal rigtige svar), og for hver gruppe i og hvert sprørgsmål j bestemme logit'en $l_{i,j}$ til andelen af korrekte svar på spørgsmål j. For hvert j skal punkterne $(l_{i,\cdot}, l_{i,j})$, $i = 1, \ldots, r$ ligge på en ret linie, og linierne svarende til forskellige værdier af j, skal være parallelle.

Når modellen anvendes til beskrivelse af svarene overfor et batteri af tests, kan modellens parametre tillægges en umiddelbar fortolkning, α_i som den i'te persons dygtighed, og β_j som den j'te opgaves "lethed". Sædvanligvis benyttes en parametrisering ved $-\beta_j$, idet $-\beta_j$ da kan fortolkes som prøvens sværhedsgrad.

Rasch model ved tilfældigt udvalgte personer

Hvis man udvælger en gruppe af personer tilfældigt, og udsætter dem for testbatteriet, vil Rasch-modellen indebære at parameteren α_i fortolkes som en stokastisk variabel. I en sådan sammenhæng kalder man parameteren α_i for en latent variabel.

Generelt siger man, at en parameter θ er en latent variabel, hvis man har et flerdimensionalt respons, $y_{i1}, y_{i2}, \ldots, y_{is}$, der er indbyrdes afhængige i den simultane fordeling af observationerne, men hvor der findes en parameter, θ, sådan at $y_{i1}, y_{i2}, \ldots, y_{is}$ er uafhængige i den betingede fordeling af $y_{i1}, y_{i2}, \ldots, y_{is}$ for givet θ.

Eksempel 3.3.3 ***Holdninger til vold***

Som led i en undersøgelse af forskellige befolkningsgruppers holdning overfor vold optog man en række videosekvenser med et stigende voldeligt indhold.

Sekvenserne blev vist for en række tilfældigt udvalgte personer, der blev bedt om at karakterisere hver sekvens som henholdsvis voldelig eller ikke voldelig.

Nedenstående tabel viser fordelingen af responser, V=voldelig, E=ej voldelig for 100 skoleelever til 3 af disse sekvenser, A, B og C, hvor A er tilsigtet at have det mindst voldelige indhold, B et større, og C et endnu større voldeligt indhold.

Tabel 3.3. Fordelingen af bedømmelsen af voldssekvenser for 100 skoleelever

Respons	Antal pers.
EEE	16
EEV	10
EVE	6
EVV	16
VEE	2
VEV	5
VVE	2
VVV	43

Tabellen angiver den marginale fordeling af det multivariate respons.

Man kan opfatte responset fra en person som en vektor af tre variable, (X_i^A, X_i^B, X_i^C), hvor hvert X_i^ν antager en af værdierne E eller V. Koder vi E som nul og V som 1, er responset fra hver person en tredimensional størrelse, der beskriver gentagne målinger fra denne person.

Såfremt man formulerer en parametrisk tæthed for sandsynlighedsfordelingen af den latente parameter θ, kan man estimere parametrene i fordelingen af θ ud fra den observerede marginale fordeling i tabel 3.3. □

3.4 Tovejs antalstabeller svarende til binært respons

3.4.1 Indledning

En tovejs antalstabel fremkommer ved klassifikation af data efter to kriterier A og B. Hvis A har r niveauer, og B har s niveauer siger man, at man har en $(r \times s)$-tabel.

Vi vil i dette afsnit betragte såkaldte 2×2-tabeller, dvs tabeller indeholdende antal, organiseret med to rækker og to søjler, hvor rækkerne repræsenterer en klassifikation efter ét kriterium (A) og søjlerne et andet kriterium (B), hver med to niveauer.

Sådanne tabeller har den generelle form som vist i tabel 3.4, hvor vi har brugt symbolet Z til at angive det antal, der står på den pågældende plads i skemaet. Vi har indiceret række- og søjlesummer (marginalsummerne) med et index for den pågældende række/søjle og et "+" på pladsen svarende til det index, der er summeret over. Selv om der således i denne generelle opskrivning er brugt symbolet Z for række- og søjlesummer, indikerer dette ikke nødvendigvis, at størrelsen er en stokastisk variabel. Vi vil også betragte situationer, hvor f.eks. rækkesummerne er faste. Modelleringen af en given tabel afhænger imidlertid af hvilke marginalsummer, der antages faste, dvs hvordan data er udvalgt.

I det følgende afsnit vil vi diskutere en række forskellige problemstillinger, der alle giver anledning til en 2×2 antalstabel.

Tabel 3.4. Generel 2×2 tabel af antal

Kriterium A	Kriterium B		Ialt
	1	2	
1	$Z_{1,1}$	$Z_{1,2}$	$Z_{1,+}$
2	$Z_{2,1}$	$Z_{2,2}$	$Z_{2,+}$
Ialt	$Z_{+,1}$	$Z_{+,2}$	$Z_{+,+}$

3.4.2 Konfidensintervaller ved sammenligning af to hyppigheder

Vi indleder med nogle resultater vedrørende fordelinger, der optræder ved sammenligning af to hyppigheder.

Vi vil betragte to binomialfordelte størrelser X og Y, hvor

$$X \in \mathrm{B}(n, p_1) \quad \text{og} \quad Y \in \mathrm{B}(m, p_2)\,, \quad X \text{ og } Y \text{ er uafhængige} \tag{3.4.1}$$

Maksimaliseringsestimaterne for p_1 og p_2 er de relative hyppigheder

$$\widehat{p}_1 = \frac{X}{n}\,, \quad \widehat{p}_2 = \frac{Y}{m} \tag{3.4.2}$$

Sætning 3.4.1 ***Approximativ fordeling for den estimerede forskel i responssandsynlighed***

Lad fordelingen af X og Y være som i (3.4.1), og betragt differensen

$$\Delta = p_1 - p_2$$

mellem responssandsynlighederne.

Maksimaliseringsestimatet for Δ er

$$\widehat{\Delta} = \widehat{p}_1 - \widehat{p}_2 = \frac{X}{n} - \frac{Y}{m}$$

Der gælder

$$\mathsf{E}\,[\widehat{\Delta}] = p_1 - p_2$$

og

$$\mathsf{V}\,[\widehat{\Delta}] = \frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}$$

Variansen estimeres ved

$$\widehat{\mathsf{V}}\,[\widehat{\Delta}] = \frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}\,,$$

hvor $\widehat{p}_1$ og $\widehat{p}_2$ er givet ved (3.4.2).

Et approximativt $100(1-\alpha)$ % konfidensinterval for Δ fås som

$$\widehat{\Delta} \pm u_{1-\alpha/2} \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}} \tag{3.4.3}$$

Bevis:
Følger af den sædvanlige normalfordelingsapproximation til binomialfordelingen □

Sætning 3.4.2 ***Approximativ fordeling for den estimerede relative risiko***

Lad fordelingen af X og Y være som i (3.4.1), og betragt den relative risiko

$$\rho = p_1/p_2$$

for hændelsen optalt ved X i forhold til hændelsen optalt ved Y.

Maksimaliseringsestimatet for ρ er

$$\widehat{\rho} = \widehat{p}_1/\widehat{p}_2 = \frac{X}{n}/\frac{Y}{m}$$

Estimatet er ikke centralt.

Den asymptotiske varians for $\ln(\widehat{\rho})$ er

$$\mathsf{V}\left[\ln(\widehat{\rho})\right] \simeq \frac{1-p_1}{np_1} + \frac{1-p_2}{mp_2}$$

Estimatet

$$\ln(\tilde{\rho}) = \ln\left(\frac{X+1/2}{n+1/2}\right) - \ln\left(\frac{Y+1/2}{m+1/2}\right) \tag{3.4.4}$$

for logaritmen til den relative risiko er mindre skævt, end $\ln(\widehat{\rho})$.

Et approximativt $100(1-\alpha)$ % konfidensinterval for $\ln(\rho)$ fås som

$$\ln(\tilde{\rho}) \pm u_{1-\alpha/2} \sqrt{\frac{1}{x+1/2} - \frac{1}{n+1/2} + \frac{1}{y+1/2} - \frac{1}{m+1/2}} \quad (3.4.5)$$

Grænserne for konfidensintervallet for den relative risiko ρ fås ved at tage exponentialfunktionen af grænserne for intervallet for $\ln(\rho)$.

Bevis:

Følger ved Taylorudvikling af $\ln(\widehat{\rho})$ □

Sætning 3.4.3 ***Approximativ fordeling for den estimerede odds-ratio***

Lad fordelingen af X og Y være som i (3.4.1), og betragt odds ratioen

$$\psi = \frac{p_1(1-p_2)}{(1-p_1)p_2}$$

Maksimaliseringsestimatoren for ψ er

$$\widehat{\psi} = \frac{\widehat{p}_1(1-\widehat{p}_2)}{(1-\widehat{p}_1)\widehat{p}_2}$$

for $y > 0$ og $x < n$. Hvis $y = 0$ eller $x = n$, eksisterer maksimaliseringsestimatoren ikke.

Fordelingen af $\widehat{\psi}$ og $\ln(\widehat{\psi})$ har derfor ikke nogen middelværdi og varians.

Fordelingen af

$$\widetilde{\psi} = \frac{(X+1/2)(m-Y+1/2)}{(n-X+1/2)(Y+1/2)} \quad (3.4.6)$$

og af $\ln(\widetilde{\psi})$ har middelværdi og varians.

Fordelingen af $\ln(\widetilde{\psi})$ er approximativt en normal fordeling med middelværdi $\ln(\psi)$ og varians

$$\mathsf{V}\left[\ln(\widetilde{\psi})\right] \simeq \frac{1}{np_1} + \frac{1}{n(1-p_1)} + \frac{1}{np_2} + \frac{1}{n(1-p_2)}$$

Et approximativt $100(1-\alpha)$ % konfidensinterval for $\ln(\psi)$ fås som

$$\ln(\tilde{\psi}) \pm u_{1-\alpha/2}\sqrt{\frac{1}{x+1/2}+\frac{1}{n-x+1/2}+\frac{1}{y+1/2}+\frac{1}{m-y+1/2}} \tag{3.4.7}$$

Grænserne for konfidensintervallet for odds-ratioen ψ fås ved at tage exponentialfunktionen af grænserne for intervallet for $\ln(\psi)$.

Bevis:

Se Haldane (1955), Gart (1962), Gart og Zweiful (1967), Gart og Thomas (1972) □

Lemma 3.4.1 ***Eksakt fordeling af odds-ratio ved sammenligning af to binomialfordelinger***

Lad fordelingen af X og Y være som i (3.4.1).

Da afhænger den betingede fordeling af X givet $X+Y=t$ kun af odds-ratioen

$$\psi = \frac{p_1}{(1-p_1)} \Big/ \frac{p_2}{(1-p_2)}$$

Fordelingen har frekvensfunktionen

$$f_X(x;\psi) = \frac{\binom{t}{x}\binom{n-t}{t-x}\psi^x}{\sum_{u=x_-}^{x_+}\binom{t}{u}\binom{n-t}{t-u}\psi^u} \quad \text{for } x_- \le x \le x_+ , \tag{3.4.8}$$

hvor

$$x_- = \max\{0, t-m\} , \quad x_+ = \min\{t, n\} ,$$

Bevis:

Der gælder

$$\begin{aligned} \mathsf{P}\ [X = x\,|X + Y = t] &= \frac{\mathsf{P}\ [X = x \cap X + Y = t]}{\mathsf{P}\ [X + Y = t]} \\ &= \frac{\mathsf{P}\ [X = x \cap Y = t - x]}{\mathsf{P}\ [X + Y = t]} \\ &= \frac{\mathsf{P}\ [X = x]\,\mathsf{P}\ [Y = t - x]}{\mathsf{P}\ [X + Y = t]} \end{aligned} \tag{3.4.9}$$

Da X og Y er uafhængige har vi, at

$$\mathsf{P}\ [X + Y = t] = \sum_{u=x_-}^{x_+} \mathsf{P}\ [X = u]\mathsf{P}\ [Y = t - u]$$

$$= \sum_{u=x_-}^{x_+} \binom{n}{u} p_1^u (1-p_1)^{n-u} \binom{m}{t-u} p_2^{t-u} (1-p_2)^{m-t+u}$$

$$= (1-p_1)^n (1-p_2)^m \left(\frac{p_2}{1-p_2}\right)^t$$

$$\sum_{u=x_-}^{x_+} \binom{n}{u} \binom{m}{t-u} \left(\frac{p_1}{1-p_1}\right)^u \left(\frac{p_2}{1-p_2}\right)^{-u}$$

$$= (1-p_1)^n (1-p_2)^m \left(\frac{p_2}{1-p_2}\right)^t \sum_{u=x_-}^{x_+} \binom{n}{u} \binom{m}{t-u} \psi^u$$

hvor faktoren foran summationstegnet optræder både i tælleren og nævneren af (3.4.9).

Fordelingen med frekvensfunktion (3.4.8) er den betingede fordeling på skrålinierne $x + y = t$, som illustreret i figur 3.1

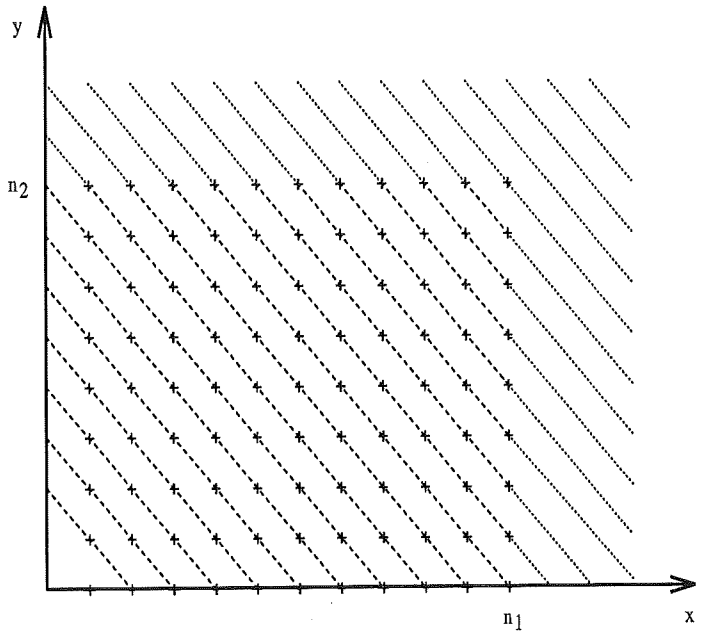
For $p_1 = p_2$ er $\psi = 1$, og nævneren i (3.4.8) bliver

$$\sum_{u=x_-}^{x_+} \binom{n}{u} \binom{m}{t-u} = \binom{n+m}{t} .$$

I dette tilfælde bliver den betingede fordeling (3.4.8) af $X|X + Y = t$ blot en $\mathrm{H}(t, n, m + n)$-fordeling som vist i Stat 1, afsnit 3.2.1

□

Figur 3.1. Illustration af gitter for den betingede fordeling givet en sum af to uafhængige binomialfordelte størrelser



Definition 3.4.1 *Betinget maksimaliseringsestimator for odds-ratioen*

Lad fordelingen af X og Y være som i (3.4.1).

Ved den betingede maksimaliseringsestimator for odds-ratioen

$$\psi = \frac{p_1}{(1-p_1)} \Big/ \frac{p_2}{(1-p_2)}$$

vil vi forstå den værdi af ψ, der maksimerer den betingede frekvensfunktion (3.4.8). □

Bemærkning 1 *Den betingede maksimaliseringsestimator er entydigt bestemt*

Hvis man differentierer log-likelihood'en svarende til (3.4.8) for en given observation x og t finder man, at estimatet $\widehat{\psi}$ tilfredsstiller middelværdiligningen

$$x = \mathsf{E}\left[X|X+Y=t\right]$$

Ligningen har en entydig løsning. Ligningen må løses ved iteration. (Se f.eks. Cornfield (1956), Cox and Snell (1989) og Plackett (1981, p.45). □

Sætning 3.4.4 ***Eksakt konfidensinterval for odds-ratioen***

Lad fordelingen af X og Y være som i (3.4.1).

Et eksakt $100(1-\alpha)$ % konfidensinterval for odds ratioen ψ svarende til observationen $X = x$ og $X + Y = t$ fås ved at bestemme ψ_L og ψ_U sådan at

$$\sum_{u=x}^{x+} f_X(x;\psi_L) = \alpha/2$$

og

$$\sum_{u=x_-}^{x} f_X(x;\psi_U) = \alpha/2$$

Da fordelingen af X er diskret, er konfidensgraden af dette interval mindst $1-\alpha$.

Bevis:

Følger af monotoniceten af likelihoodkvotienten i ψ. □

Baptista (1977) har angivet en algoritme til bestemmelse af konfidensinterval og de tilsvarende tests.

Mange af de programsystemer, der er dedikeret til analyse af diskrete data, kan beregne dette konfidensinterval.

Bemærkning 1 ***Betinget test for hypoteser vedrørende odds ratioen***

Et niveau α-test for test af hypotesen $H_0 : \psi = \psi_0$ imod det ensidede alternativ $H_1 : \psi > \psi_0$ har det kritiske område: forkast for $x \geq x_u$, hvor x_u er bestemt så

$$\sum_{u=x_u}^{x+} f_X(x;\psi_0) \leq \alpha$$

Tilsvarende har niveau α-testet for hypotesen $H_0 : \psi = \psi_0$ imod det ensidede alternativ $H_1 : \psi < \psi_0$ det kritiske område: forkast for $x \leq x_l$, hvor x_l er bestemt så

$$\sum_{u=x_-}^{x_l} f_X(x;\psi_0) \leq \alpha$$

For $H_0 : \psi_0 = 1$ reducerer testene til det fra Statistik 1, afsnit 3.2.1 kendte test for sammenligning af to binomialfordelinger ved beregning af sandsynligheder i en $\mathrm{H}(t, n, n+m)$-fordeling.

Testet svarende til $\psi_0 = 1$ kaldes Fisher's eksakte test. □

3.4.3 Prospektive og retrospektive undersøgelser

Vi vil nu specielt betragte sådanne antalstabeller, hvor den ene klassifikation repræsenterer en forklarende variabel, en faktor eller en stimulus, og den anden klassifikation repræsenterer en responsvariabel. Som hidtil vil vi indskrænke os til at betragte stimuli med to niveauer, og binære responser, dvs. 2×2-tabeller.

Vi forestillller os således data organiseret i en tabel som vist i tabel 3.5

Tabel 3.5. Klassifikation af observationsenheder efter stimulus of respons

Stimulus	Respons B		
A	1	2	Ialt
1	$Z_{1,1}$	$Z_{1,2}$	$Z_{1,+}$
2	$Z_{2,1}$	$Z_{2,2}$	$Z_{2,+}$
Ialt	$Z_{+,1}$	$Z_{+,2}$	$Z_{+,+}$

Tabellen har samme struktur som den generelle tabel 3.4, blot har vi markeret at kriterium A angiver en stimulus. Som tidligere har vi benyttet symbolet Z for alle celler og alle marginaler i tabellen for ikke på forhånd at binde os til at opfatte nogle af marginalerne som fastlagt.

De ialt $Z_{+,+}$ enheder i tabellen kan være udvalgt på forskellig måde. Man taler således om:

Prospektive undersøgelser: Hvis enhederne udvælges før responset har manifesteret sig, siger man, at man har en prospektiv undersøgelse.

Dette er f.eks. situationen i et sædvanligt kontrolleret forsøg, hvor enhederne udvælges tilfældigt, og allokeres til en af de to stimuli, hvorefter man registrerer responset. Eksemplet 3.3.1 med blommestiklingerne er et eksempel på et sådant kontrolleret forsøg.

Specielt i epidemiologiske sammenhænge møder man undersøgelser, hvor allokeringen af enheder til de mulige stimuli ikke er kontrolleret (styret på forhånd af forsøgsdesignet). Sådanne undersøgelser, hvor man udvælger et antal enheder og derefter registrerer, hvilke stimuli, de udsættes for, og hvilket respons, der resulterer, kaldes kohortestudier. I epidemiologiske sammenhænge kan en sådan forsøgsplan optræde, fx hvis man udvælger en gruppe kvinder, registrerer deres valg af antikonceptionsmiddel og forekomsten af blodpropper i gruppen, og relaterer forekomsten af blodpropper til brugen af antikonceptionsmiddel.

I industrielle sammenhænge kan man udvælge en samling produktenheder som følges i deres livscyklus eller i en fastlagt periode; dvs. man registrerer hvilke belastninger de udsættes for og forekomsten af fejl, og relaterer forekomsten af fejl til belastningen.

Retrospektive undersøgelser: Hvis enhederne udvælges efter at responset har manifesteret sig, siger man, at man har en retrospektiv undersøgelse.

Stikprøveudvælgelsen kan foretages ved at man udvælger et antal enheder tilfældigt, hvorefter man klassificerer dem efter henholdsvis stimulus-og responsklassifikationen. I et sådant tilfælde kalder man undersøgelsen en tværsnitsundersøgelse, eller evt en udsnitsundersøgelse. Eksempel 3.3.2 med undersøgelsen af kilder til viden om cancer er et eksempel på en sådan tværsnitsundersøgelse.

Udvælgelsen kan også foretages ved at man udvælger et antal enheder blandt enheder som har det pågældende respons, hvorefter de klassificeres efter, hvorvidt de har været udsat for den betragtede stimulus, eller ej. Tilsvarende udvælges et antal kontrolenheder blandt enheder som ikke har det pågældende respons, hvorefter også disse enheder klassificeres efter, hvorvidt de har været udsat for den betragtede stimulus, eller ej. En sådan udvælgelsesstrategi kaldes et case-control studie.

3.4.4 Modeller for prospektive studier

Kontrollerede forsøg

I et kontrolleret forsøg er antallet af enheder, der udsættes for en given stimulus, fastlagt på forhånd.

I opstillingen i tabel 3.5 svarer det til, at rækkemarginalerne $Z_{i+} = n_i$ er fastlagt. Stikprøvefordelingen af Z_{i1} vil sædvanligvis være

$$Z_{i1} \in \mathrm{B}(n_i, p_{1|i}) ,$$

hvor sandsynlighedsparameteren $p_{1|i}$ angiver sandsynligheden for respons 1 ved stimulus i.

De relevante sandsynligheder er angivet i nedenstående tabel:

Stimulus	Respons B	
A	1	2
1	$p_{1\|1}$	$p_{2\|1} = 1 - p_{1\|1}$
2	$p_{1\|2}$	$p_{2\|1} = 1 - p_{1\|1}$

Eksempel 3.4.1 ***Forekomst af hjerteanfald klassifceret efter aspirinindtagelse***

Nedenstående tabel angiver hyppighederne af hjerteinfarkt hos to grupper amerikanske læger, hvoraf den ene gruppe indtog en aspirintablet hver dag i undersøgelsesperioden. Den anden gruppe indtog en placebotablet, dvs. en tilsvarende tablet, som blot ikke indeholdt noget aktivt stof.

Studiet er et kontrolleret forsøg. Stikprøvestørrelserne i de to grupper, Placebo og Aspirin, er fastsat til $n_1 = 11\,034$ og $n_2 = 11\,037$. Stikprøvestørrelsen var fastlagt af antallet af læger, der var villige til at indgå i forsøget, og der benyttedes en randomiseringsmekanisme til at allokere de deltagende læger til Aspiringruppen/Placebogruppen,

Studiet var tilrettelagt som et blindforsøg: de deltagende læger vidste ikke, hvorvidt det præparat, de indtog, indeholdt aktivt stof, eller ej.

Responsvariablen er "Forekomst af hjerteanfald" med værdierne "Anfald" og "Ingen anfald".

Studiet er et prospektivt studie. Vi kan umiddelbart af data estimere sandsynligheden for en bestemt værdi af responsvariablen.

Tabel 3.6. Forekomst af hjerteanfald klassificeret efter aspirinindtagelse, prospektivt studie

	Forekomst af hjerteanfald		
Aspirin-indtagelse	Anfald	Ingen anfald	Ialt
Placebo	189	10 845	11 034
Aspirin	104	10 933	11 037

Kilde: Findings from the Aspirin Component of the Ongoing Physicians' Health study, *N.Engl.J. Med.* **38**: 262 - 264, (1988).

En passende model kunne være $Z_{i1} \in \mathrm{B}(n_i, p_{1|i})$, $i = 1, 2$, hvor Z_{i1} angiver antallet af personer med hjerteanfald i den i-te gruppe. De observerede andele $\widehat{p}_{1|1} = 189/11034 = 0.0171$ og $\widehat{p}_{1|2} = 104/11037 = 0.0094$ er direkte estimater for populationsandelene, dvs. for risikoen for hjerteanfald for henholdsvis ikke-Aspirinspisende og Aspirinspisende læger.

Den absolutte forskel i risikoen er $\widehat{p}_{1|1} - \widehat{p}_{1|2} = 0.0077$, dvs. 0.77 procentpoints. Et approximativt 95 % konfidensinterval for forskellen fås af sætning 3.4.1 til

$$\begin{aligned} 0.0077 &\pm 1.96 \times \sqrt{0.00000152 + 0.00000084} \\ &= 0.0077 \pm 1.96 \times 0.0015 = 0.0077 \pm 0.0030 \end{aligned}$$

Den relative overrisiko for ikke-aspirinspisende personer estimeres til $\widehat{\rho} = \widehat{p}_{1|1}/\widehat{p}_{1|2} = 0.0171/0.0094 = 1.82$. Et approximativt 95 % konfidensinterval for overrisikoen fås af sætning 3.4.2. Man får $\ln(\widehat{\rho}) = 0.5955$ og intervallet for $\ln(\rho)$ bliver

$$\begin{aligned} 0.5955 &\pm 1.96 \times \sqrt{0.005277 - 0.000091 + 0.009569 - 0.000091} \\ &= 0.5955 \pm 1.96 \times 0.121 = 0.5955 \pm 0.237\,, \end{aligned}$$

dvs konfidensintervallet for ρ bliver $(1.43; 2.30)$

Endelig finder man, at stikprøveforholdet mellem odds for ikke-aspirinspisende og aspirinspisende læger er $\widehat{\psi} = (189/10845)/(104/10933) = 1.83$. Et approximativt 95 % konfidensinterval for ψ fås af sætning 3.4.3. Man får det approximative interval for $\ln(\psi)$

$$\begin{aligned} 2.9425 &\pm 1.96\sqrt{0.005277 + 0.000092 + 0.009569 + 0.000091} \\ &= 2.9425 \pm 1.96 \times \sqrt{0.01503} = 2.94225 \pm 0.2403 \end{aligned}$$

dvs intervallet for ψ bliver (14.9; 24.1) □

Kohortestudier

I et kohortestudium er kun det totale antal enheder $Z_{+,+} = n$ fastlagt på forhånd. Stikprøvefordelingen af sættet $(Z_{1,1}, Z_{1,2}, Z_{2,1}, Z_{2,2})$ vil sædvanligvis være en multinomialfordeling

$$(Z_{1,1}, Z_{1,2}, Z_{2,1}, Z_{2,2}) \in \text{Mult}(n; \pi_{1,1}, \pi_{1,2}, \pi_{2,1}, \pi_{2,2})$$

Vi vil ikke her komme nærmere ind på analysen af multinomialfordelte størrelser.

3.4.5 Retrospektive studier

Tværsnitsundersøgelser

I en tværsnitsundersøgelse (eng. *Cross-sectional study*) er kun det totale antal enheder $Z_{+,+} = n$ fastlagt på forhånd. Stikprøvefordelingen af sættet $(Z_{1,1}, Z_{1,2}, Z_{2,1}, Z_{2,2})$ vil sædvanligvis være en multinomialfordeling

$$(Z_{1,1}, Z_{1,2}, Z_{2,1}, Z_{2,2}) \in \text{Mult}(n; \pi_{1,1}, \pi_{1,2}, \pi_{2,1}, \pi_{2,2})$$

I afsnit 3.4.6 og 4 vil vi belyse forskellige eksempler på tværsnitsundersøgelser.

Case-control studier

I en case-control undersøgelse er antallet af enheder med en given respons fastlagt på forhånd.

Sædvanligvis vælger man et antal, n_1 "cases" med det respons, (respons nr. 1), der er af interesse. Derefter vælger man et antal, n_2 "kontrolenheder", som såvidt muligt vælges fra den samme population som gruppen af cases, men som blot adskiller sig fra gruppen af cases ved at have respons nr. 2 (hvilket oftest vil sige "ingen respons", dvs ikke ramt af sygdommen, ikke fejlbehæftet etc.).

I opstillingen i tabel 3.5 svarer det til, at søjlemarginalerne $Z_{+j} = n_j$ er fastlagt. Stikprøvefordelingen af $Z_{i,j}$ vil sædvanligvis være

$$Z_{1j} \in \text{B}(n_j, p^*_{1|j}) \,,$$

I opstillingen i tabel 3.5 svarer det til, at søjlemarginalerne $Z_{+j} = n_j$ er fastlagt. Stikprøvefordelingen af $Z_{i,j}$ vil sædvanligvis være

$$Z_{1j} \in \mathrm{B}(n_j, p^*_{1|j}) \ ,$$

hvor sandsynlighedsparameteren

$$p^*_{1|j} = \mathsf{P}\ [\text{stimulus } 1|\text{respons } j]$$

angiver sandsynligheden for at et element med respons j har været udsat for stimulus 1.

Interessen knytter sig imidlertid ikke til denne størrelse, men snarere til de betingede sandsynligheder

$$p_{1|i} = \mathsf{P}\ [\text{respons } 1|\text{stimulus } i\ ]$$

Antag nu, at stimulus nr. 1 optræder med hyppigheden $\pi_{1,+}$, og stimulus nr. 2 med hyppigheden $\pi_{2,+} = 1 - \pi_{1,+}$ i populationen.

Såfremt sandsynligheden for et givet respons kun afhænger af den anvendte stimulus, kan populationsandelene af de fire kombinationer af stimulus og respons udtrykkes ved hyppighederne $\pi_{1,+}$ og $\pi_{2,+}$ samt responssandsynlighederne $p_1|i$:

$$\begin{aligned} \pi_{1,1} &= \pi_{1,+}\, p_{1|1} \\ \pi_{1,2} &= \pi_{1,+}\, (1 - p_{1|1}) \end{aligned} \qquad (3.4.10)$$

$$\begin{aligned} \pi_{2,1} &= \pi_{2,+}\, p_{1|2} \\ \pi_{2,2} &= \pi_{2,+}\, (1 - p_{1|2}) \end{aligned}$$

Nedenstående tabel illustrerer de simultane og de marginale sandsynligheder svarende til populationen:

Yules krydsprodukt ratio (3.3.16) for hændelserne A og B i det tosidede skema i tabel 3.7 er

$$\mathrm{cpr}(A, B) = \frac{\pi_{1,1}\pi_{2,2}}{\pi_{1,2}\pi_{2,1}} \qquad (3.4.11)$$

Indfører vi responssandsynlighederne $p_{1|i}$ i (3.4.11) får vi

$$\mathrm{cpr}(A, B) = \frac{p_{1|1}}{(1 - p_{1|1})} \Big/ \frac{p_{1|2}}{(1 - p_{1|2})} \ , \qquad (3.4.12)$$

Tabel 3.7. Simultan fordeling i populationen af kombinationer af stimuli og responser

Stimulus A	Respons B 1	Respons B 2	Ialt
1	$\pi_{1,1}$	$\pi_{1,2}$	$\pi_{1,+} = \pi_{1,1} + \pi_{1,2}$
2	$\pi_{2,1}$	$\pi_{2,2}$	$\pi_{2,+} = \pi_{2,1} + \pi_{2,2}$
Ialt	$\pi_{+,1}$	$\pi_{+,2}$	1

$\pi_{+,1} = \pi_{1,1} + \pi_{2,1}$; $\pi_{+,2} = \pi_{1,2} + \pi_{2,2}$

altså netop forholdet mellem de betingede odds for responserne givet stimulus.

Sandsynlighederne $p^*_{i|j}$ svarende til den anvendte stikprøveudtagning er

$$\begin{aligned} p^*_{1|1} &= \frac{\pi_{1,1}}{\pi_{+,1}} \\ p^*_{1|2} &= \frac{\pi_{1,2}}{\pi_{+,2}} \end{aligned} \tag{3.4.13}$$

Forholdet mellem de betingede odds svarende til disse sandsynligheder er

$$\omega' = \frac{p^*_{1|1}}{1-p^*_{1|1}} \Big/ \frac{p^*_{1|2}}{1-p^*_{1|2}} = \frac{\pi_{1,1}\pi_{2,2}}{\pi_{1,2}\pi_{2,1}}\,, \tag{3.4.14}$$

altså netop $\mathrm{cpr}(A,B)$ (3.4.11).

Vi har altså i overenstemmelse med bemærkningen i afsnit 3.3.3:

> Ratioen mellem de betingede odds svarende til responssandsynlighederne $p_{1|i}$ er den samme som ratioen mellem de betingede odds svarende til stimulussandsynlighederne $p^*_{1|j}$ i case- og i kontrolgrupperne.

Selv om data ikke giver mulighed for at estimere selve responssandsynlighederne, har vi altså mulighed for at estimere odds ratioen svarende til responssandsynlighederne.

Eksempel 3.4.2 ***Retrospektivt studie, case-control studie***

Nedenstående tabel angiver to grupper amerikanske kvinder, klassifikationen efter brugen af orale antikonceptionsmidler, nemlig dels hos 58 kvindelige patienter, indlagt på grund af hjerteanfald, og dels en kontrolgruppe udvalgt som 166 andre kvindelige patienter på samme hospital.

Tabel 3.8. Forekomst af hjerteanfald klassificeret efter brug af orale antikonceptionsmidler, retrospektivt studie

	Forekomst af hjerteanfald	
Brug af orale Antikonceptions-midler	Anfald	Ingen anfald
P-piller	23	34
Ingen piller	35	132
Ialt	58	166

Kilde: J.I.Mann, M.P.Vessey, M.Thorogood, and R.Doll: Myocardial infarction in young women with special reference to oral contraceptive practice. *British J. Med.* **2**: 241- 245, (1975).

Undersøgelsen er retrospektiv. De to værdier af responsvariablen afgrænser forskellige populationskategorier i modsætning til en prospektiv undersøgelse, hvor kategorierne af responsvariablen er de samme for alle betragtede kombinationer af faktorvariable.

Den egentlige interesse knytter sig til forekomsten af hjerteanfald som responsvariabel og brugen af P-piller som forklarende variabel. Vi ville gerne kunne sige noget om andelen $p_{1|1}$ af hjerteanfald blandt kvinder, som bruger P-piller, sammenlignet med andelen $p_{1|2}$ af hjerteanfald blandt kvinder, som ikke bruger P-piller, eller mere præcist formuleret, om den betingede sandsynlighed $p_{1|1}$ for hjerteanfald, givet en tilfældigt udtaget kvinde er P-pillebruger, versus den betingede sandsynlighed $p_{1|2}$ for hjerteanfald, givet hun ikke har brugt P-piller.

Som anført bygger stikprøveplanen imidlertid på responsvariablen. Der er udtaget n_1 personer med hjerteanfald, og n_2 uden hjerteanfald, hvorefter de er klassificeret efter deres brug af P-piller. Disse andele afspejler de

betingede fordelinger af brugen af P-piller blandt kvinder med hjerteanfald og kvinder uden hjerteanfald.

En naturlig model er $Z_{1,1} \in \mathrm{B}(n_1, p^*_{1|1})$ og $Z_{1,2} \in \mathrm{B}(n_2, p^*_{1|2})$, hvor $p^*_{1|1} = \pi_{1,1}/\pi_{+,1}$ og $p^*_{1|2} = \pi_{2,1}/\pi_{+,2}$.

Vi kan imidlertid bestemme et estimat for oddsratioen

$$\omega' = \frac{p^*_{1|1}}{1-p^*_{1|1}} \Big/ \frac{p^*_{1|2}}{1-p^*_{1|2}}$$

som jvf det foregående også er oddsratioen svarende til responssandsynlighederne.

Vi finder jvf sætning 3.4.3 estimatet (3.4.6)

$$\widetilde{\psi} = \frac{23.5 \times 132.5}{35.5 \times 34.5} = 2.54$$

med $\ln(\widetilde{\psi}) = 0.9331$ med det approximative 95 % konfidensinterval for $\ln(\psi)$

$$\begin{aligned} 0.9331 &\pm 1.96 \times \sqrt{0.0426 + 0.0282 + 0.0290 + 0.0075} \\ &= 0.9331 \pm 1.96 \times \sqrt{0.1073} = 0.9331 \pm 0.6419 \end{aligned}$$

dvs intervallet for ψ bliver $(1.34; 4.83)$.

Intervallet omfatter ikke $\psi = 1$, hvorfor man kan slutte, at det tilsvarende test ville afvise hypotesen $H_0 : \psi = 1$ ved et tosidet test på niveau $\alpha = 0.05$.

Havde vi i stedet benyttet en algoritme til bestemmelse af det eksakte konfidensintervallet i overensstemmelse med sætning 3.4.4, havde vi fundet intervallet $(1.46; 4.38)$. Approximationen var altså rimeligt god. □

Eksempel 3.4.3 *Case-kontrol analyser i industrielle sammenhænge*

Ved produktionen af isoleringsruder har den anvendte lim en vis betydning for udviklingen af utætheder.

Ved produktionen af en bestemt type isoleringsrude har man skiftet mellem to typer lim, type A og type B. Risikoen for udvikling af utætheder er ikke nødvendigvis den samme for de to typer.

Limtype	Funktion	
	Fejlfunktion	Tilfredstillende funktion
Type A	$p_{1\|1}$	$p_{2\|1} = 1 - p_{1\|1}$
Type B	$p_{1\|2}$	$p_{2\|2} = 1 - p_{1\|2}$

Ved et kontrolleret forsøg vil man være i stand til direkte at estimere $p_{1|1}$ og $p_{1|2}$ og derved at estimere den relative risiko ved at bruge type A i forhold til type B, eller odds-ratioen for type A i forhold til B.

I praksis er fejlene lang tid om at udvikle sig, og selv om man eventuelt kan benytte forskellige former for accelleret prøvning, er det af værdi også at vurdere forskellen på de to typer ud fra markdata.

Man kunne indsamle en stikprøve af ibrugtagne ruder, og klassificere dem efter anvendt limtype og forekomsten af fejl. Da fejlandelene er relativt små, kræver det imidlertid en forholdsvis stor stikprøve, hvis man skal have et passende antal fejlende ruder i stikprøven. Ydermere er der det problem, at fejlende ruder udskiftes, når de fejler, hvorfor de ikke vil blive repræsenteret i stikprøven i det rette forhold.

Man kan derfor vælge en case-control undersøgelse, hvor der fra reparatørerne udtages et antal fejlende ruder (cases), og fra det samme kundegrundlag udtages et antal ikke-fejlende ruder.

Med udgangspunkt i andelene af type A og type B ruder i det to stikprøver kan man da bestemme odds ratioen for type A i forhold til type B. □

Eksempel 3.4.4 ***Sensitivitet og specificitet af binær klassifikationsrutine***

Betrag en diagnostisk prøvningsmetode, der sigter imod at konstatere tilstedeværelse/ikke-tilstedeværelse af en given egenskab.

Prøvningsmetodens operationskarakteristik er givet ved matricen

sand tilstand	Respons	
	Positivt	Negativt
Til stede	$p_{1\|1}$	$p_{2\|1} = 1 - p_{1\|1}$
Ej til stede	$p_{1\|2}$	$p_{2\|2} = 1 - p_{1\|2}$

hvor et positivt prøvningsresultat angiver et resultat, der signalerer tilstedeværelse af egenskaben.

De to sandsynligheder svarende til de korrekte klassifikationer betegnes ofte metodens sensitivitet ($p_{1|1}$), og prøvens specificitet ($p_{2|2}$).

Antag, at man har udført et prospektivt forsøg med en bestemt prøvningsmetode, hvor n_1 enheder med den givne egenskab og n_2 enheder uden egenskaben blev prøvet ved metoden.

Resultaterne kan sammenfattes i skemaet

sand tilstand	**Respons**		**Ialt**
	Positivt	Negativt	
Til stede	$x_{1\|1}$	$n_1 - x_{1\|1}$	n_1
Ej til stede	$n_2 - x_{2\|2}$	$x_{2\|2}$	n_2

Sensitiviteten estimeres da ved $x_{1|1}/n_1$, og specificiteten ved $x_{2|2}/n_2$.

I nogle sammenhænge møder man betegnelsen Receiver operating characteristic (ROC) for en afbildning af samhørende værdier af størrelserne $p_{1|1}$ og $1 - p_{2|2}$ for forskellige definitioner af et positivt prøvningsresultat, eller under forskellige omstændigheder, se Tosteson og Begg (1988).

I den specifikke anvendelse af klassifikationsproceduren knytter interessen sig imidlertid til den sande tilstand. Man vil således gerne kunne angive sandsynligheden for at egenskaben er til stede, givet at klassifikationen har indikeret tilstedeværelse. Denne sandsynlighed kan imidlertid ikke umiddelbart beregnes, da den afhænger af den ubetingede sandsynlighed for at egenskaben er tilstede i prøvningsmaterialet.

Lad π angive sandsynligheden for tilstedeværelse af den betragtede egenskab. I analogi med tabel 3.5 finder vi da matricen af simultane sandsynligheder

sand tilstand	**Respons**		Marg. ssh.
	Til stede	Ej til stede	
Til stede	$p_{1\|1} \times \pi$	$p_{2\|1} \times \pi$	π
Ej til stede	$p_{1\|2} \times (1-\pi)$	$p_{2\|2} \times (1-\pi)$	$1-\pi$
Marg. ssh.	$p_{+,1}$	$p_{+,2}$	1.0

hvor

$$p_{+,1} = p_{1|1} \times \pi + p_{1|2} \times (1 - \pi)$$

angiver den marginale sandsynlighed for at klassifikationen indikerer tilstedeværelse af egenskaben. og tilsvarende $p_{+,2} = 1 - p_{+,1}$ angiver sandsynligheden for at klassifikationen indikerer ikke-tilstedeværelse.

Den ønskede sandsynlighed for at egenskaben er til stede, givet at klassifikationen har indikeret tilstedeværelse er da

$$\frac{p_{1|1} \times \pi}{p_{1|1} \times \pi + p_{1|2} \times (1 - \pi)} \tag{3.4.15}$$

Odds for at egenskaben er til stede, imod at den ikke er til stede, er

$$\theta = \frac{p_{1|1}}{p_{1|2}} \times \frac{\pi}{1-\pi} = \frac{p_{1|1}}{1 - p_{2|2}} \times \frac{\pi}{1-\pi}$$

dvs odds er produktet af apriori-odds for tilstedeværelse med forholdet mellem sensitiviteten $p_{1|1}$, og sandsynligheden $1 - p_{2|2}$ for falsk positiv Yules krydsprodukt-ratio (3.3.16) for hændelserne "sand tilstand til stede" og "Respons til stede" i det tosidede skema bliver

$$\frac{(p_{1|1} \times \pi)[p_{2|2} \times (1-\pi)]}{[p_{1|2} \times (1-\pi)](p_{2|1} \times \pi)} = \frac{p_{1|1}}{1 - p_{1|1}} \Big/ \frac{p_{1|2}}{1 - p_{1|2}} \tag{3.4.16}$$

I overensstemmelse med bemærkningen i afsnit 3.3.3 ser vi, at krydsprodukt-ratioen for hændelserne ikke afhænger af den marginale sandsynlighed π for tilstedeværelse, men kun af forholdet mellem odds svarende til sensitiviteten og odds svarende til specificiteten. □

3.4.6 Modeller for gentagne målinger

Vi betragter atter den generelle 2×2-tabel, tabel 3.4. Hvis hver observation i tabellen repræsenterer et observationspar, f.eks. en før (kriterium A) og efter (kriterium B) måling på det samme individ, eller en respons ved stimulus A, og en respons for det samme individ ved stimulus B, siger vi, at vi har en situation med gentagne målinger (eng. *repeated measurements*).

I en sådan situation vil man betragte det samlede antal observationer $n = Z_{+,+}$ som fast. Enhver observation placeres i en af de fire celler, og hvis observationerne er indbyrdes uafhængige vil man modellere antallet af observationer i hver af de fire celler $(1,1); (1,2); (2,1); (2,2)$ ved en multinomialfordeling med antalsparameter $n = Z_{+,+}$ og sandsynlighedsparameter π, hvor π består af de fire sandsynligheder $(\pi_{1,1}, \pi_{1,2}, \pi_{2,1}, \pi_{2,2})$.

Indiceringen af de enkelte sandsynligheder modsvarer indiceringen af observationerne. Tilsvarende vil vi indicere sandsynlighederne for de marginale observationer ved et "+" svarende til det index, der er summeret over.

Tabel 3.9. Sandsynligheder svarende til en multinomial fordeling for en 2×2 antalstabel.

Kriterium A	Kriterium B		Ialt
	1	2	
1	$\pi_{1,1}$	$\pi_{1,2}$	$\pi_{1,+}$
2	$\pi_{2,1}$	$\pi_{2,2}$	$\pi_{2,+}$
Ialt	$\pi_{+,1}$	$\pi_{+,2}$	1

Da vi har brugt den sædvanlige repræsentation af multinomialfordelingen er fordelingen overparametriseret. Vi har for eksempel

$$\pi_{2,2} = 1 - (\pi_{1,1} + \pi_{1,2} + \pi_{2,1}) , \tag{3.4.17}$$

og tilsvarende

$$\pi_{+,2} = 1 - \pi_{+,1} , \qquad \pi_{2,+} = 1 - \pi_{1,+} \tag{3.4.18}$$

Eksempel 3.4.5 ***Effekt af lokalbedøvelsesmiddel***

Som led i effektvurderingen af et nyt lokalbedøvelsesmiddel B foretoges blandt andet en sammenligning af effekten af dette middel med et anerkendt middel A.

Man udvalgte 100 forsøgspersoner, som på to på hinanden følgende dage fik injiceret bedøvelsesvæsken ved en bestemt nerve i kæben. Efter 5 minutter målte man om bedøvelsen virkede eller ej. (For hver person foretoges en lodtrækning med ssh. 1/2 om man skulle begynde med middel A eller middel B).

Resultaterne er angivet i nedenstående tabel:

Middel A	Middel B		Ialt
	Effekt	Ej effekt	
Effekt	84	1	85
Ej effekt	9	6	15
Ialt	93	7	100

□

Man er sædvanligvis interesseret i at sammenligne de marginale sandsynligheder $\pi_{1,+} = \pi_{1,1} + \pi_{1,2}$ og $\pi_{+,1} = \pi_{1,1} + \pi_{2,1}$ Vi opstiller derfor hypotesen om marginal symmetri

$$H_0 : \quad \pi_{1,+} = \pi_{+,1} \tag{3.4.19}$$

med alternativet $H_1 : \pi_{1,+} \neq \pi_{+,1}$.

(Hvis $\pi_{1,+} = \pi_{+,1}$ da er også $\pi_{2,+} = \pi_{+,2}$).

Vi bemærker iøvrigt, at forskellen

$$\pi_{1,+} - \pi_{+,1} = \pi_{1,2} - \pi_{2,1} \tag{3.4.20}$$

da sandsynligheden $\pi_{1,1}$ jo indgår i begge de marginale sandsynligheder. Dvs for en 2×2-tabel medfører marginal symmetri, at der er symmetri omkring hoveddiagonalen (og omvendt).

Sætning 3.4.5 ***Test for marginal symmetri i 2×2-tabel for gentagne målinger.***

Lad $(Z_{1,1}, Z_{1,2}, Z_{2,1}, Z_{2,2}) \in \text{Mult}(n; \pi_{1,1}, \pi_{1,2}, \pi_{2,1}, \pi_{2,2})$.

Det betingede test for hypotesen (3.4.19) om marginal symmetri udføres ved at sammenligne antallet $Z_{1,2}$ med fraktilerne i en $\text{B}(Z_{1,2} + Z_{2,2}, 0.5)$-fordeling. Hypotesen forkastes for små og for store værdier af $Z_{1,2}$.

Et approximativt test fås ved at betragte teststørrelsen

$$Q = \frac{(Z_{1,2} - Z_{2,1})^2}{Z_{1,2} + Z_{2,1}} \tag{3.4.21}$$

Under hypotesen vil Q approximativt følge en $\chi^2(1)$-fordeling. Hypotesen forkastes for store værdier af Q.

Bevis:

Betragt først det betingede test.

Idet hypotesen om marginal symmetri er ensbetydende med hypotesen

$$H_0^* : \quad \pi_{1,2} = \pi_{2,1}$$

om symmetri omkring hoveddiagonalen.

Under denne hypotese er den betingede fordeling af $Z_{1,2}$ givet summen $Z_{1,2} + Z_{2,2} = t$ en $\mathrm{B}(t,p)$-fordeling med

$$p = \frac{\pi_{1,2}}{\pi_{1,2} + \pi_{2,1}} = \frac{1}{2}$$

(se Mosteller 1952).

Hypotesen testes derfor ved at sammenligne $Z_{1,2}$ med fraktilerne i en $\mathrm{B}(t, 0.5)$-fordeling.

Det approximative test fås ved at betragte teststørrelsen

$$D = \widehat{\pi}_{1,+} - \widehat{\pi}_{+,1} = \frac{Z_{1,2} - Z_{2,1}}{n}$$

Under H_0 er $\mathsf{E}[D] = 0$.

Idet $\mathsf{COV}[Z_{1,2}, Z_{2,1}] = -n\pi_{1,2}\pi_{2,1}$ finder man

$$\mathsf{V}[D] = \frac{\pi_{1,2}(1-\pi_{1,2}) + \pi_{2,1}(1-\pi_{2,1}) + 2\pi_{1,2}\pi_{2,1}}{n} \tag{3.4.22}$$

Under hypotesen om marginal symmetri er der symmetri om hoveddiagonalen, dvs $\pi_{1,2} = \pi_{2,1}$, og man får derfor

$$\mathsf{V}[D] = \frac{2[\pi_{1,2}(1-\pi_{1,2}) + \pi_{1,2}\pi_{2,1}]}{n} = 2\,\frac{\pi_{1,2}}{n} \tag{3.4.23}$$

Under H_0 estimeres $\pi_{1,2}$ ved

$$\widehat{\pi}_{1,2} = \frac{Z_{1,2} + Z_{2,1}}{2n}$$

hvorfor vi har estimatet for $\mathsf{V}[D]$:

$$\widehat{\mathsf{V}}[D] = \frac{Z_{1,2} + Z_{2,1}}{n^2}$$

Under H_0 følger teststørrelsen

$$T = \frac{D - 0}{\sqrt{\widehat{\mathsf{V}}[D]}} = \frac{Z_{1,2} - Z_{2,1}}{\sqrt{Z_{1,2} + Z_{2,1}}}$$

approximativt en N(0,1)-fordeling, og derfor vil

$$Q = T^2 = \frac{(Z_{1,2} - Z_{2,1})^2}{Z_{1,2} + Z_{2,1}}$$

approximativt følge en $\chi^2(1)$-fordeling.

□

Det approximative test kaldes McNemar's test. McNemar (1947).

Bemærkning 1 *Sammenligning med en "uparret situation"*

Vi bemærker at begge de anførte test alene betragter de observationssæt (blandt de n), hvor de to kriterier giver forskelligt resultat. (I eksemplet ovenfor svarer dette til de tilfælde, hvor de to bedøvelsesmidler giver forskelligt resultat). Når man er interesseret i forskellen på de to, er de tilfælde, hvor responserne er ens, jo uden interesse.

Denne angrebsvinkel svarer til brugen af det parrede t-test i normalfordelingssituationen.

Havde man i stedet haft to uafhængige stikprøver, hver på n observationer, og for den ene bestemt andelen af respons 1 ved kriterium A (bedøvelsesmiddel A), og for den anden andelen af respons 1 ved kriterium B (bedøvelsesmiddel B), ville det approximative test også her bestå i en vurdering

$$D = \widehat{\pi}_{1,+} - \widehat{\pi}_{+,1} ,$$

men i dette tilfælde ville man have

$$\mathsf{V}[D] = \frac{\pi_{1,2}(1 - \pi_{1,2}) + \pi_{2,1}(1 - \pi_{2,1})}{n} . \tag{3.4.24}$$

For at sammenligne denne størrelse med variansen i den parrede situation omskriver vi (3.4.22) til

$$\mathsf{V}[D] = \frac{\pi_{1,+}(1 - \pi_{1,+}) + \pi_{+,1}(1 - \pi_{+,1}) - 2(\pi_{1,1}\pi_{2,2} - \pi_{1,2}\pi_{2,1})}{n} \tag{3.4.25}$$

Vi ser, at såfremt

$$\frac{\pi_{1,1}\pi_{2,2}}{\pi_{1,2}\pi_{2,1}} > 1 , \tag{3.4.26}$$

da vil variansen (3.4.25) svarende til den parrede situation være mindre end variansen (3.4.24) svarende til den uparrede situation.

Størrelsen (3.4.26) er netop Yule's krydsprodukt ratio (afsnit 3.3.3, side 389). Hvis der er positiv afhængighed, (Yule's krydsprodukt ratio større end 1), da vil variansen svarende til den parrede situation være mindre end variansen svarende til den uparrede situation. □

Eksempel 3.4.6 *Effekt af lokalbedøvelsesmiddel (fortsat)*

Vi betragter atter situationen i eksempel 3.4.5.

Vi vil vurdere den hypotese, at de to midler har samme effekt, dvs $\pi_{1,+} = \pi_{+,1}$.

Idet $z_{1,2} = 1$ og $z_{2,1} = 9$ finder man, at teststørrelsen for det betingede test for denne hypotese er $z_{1,2} = 1$, som skal sammenlignes med fraktilerne i en B(10, ; 0.5)-fordeling.

Idet P [B(10; 0.5) $\leq$ 1] = 0.0107 ser vi, at hypotesen må afvises ved et (tosidet) test på 5 %-niveauet.

Der er altså forskel på effektiviteten af de to midler. Det fremgår af observationerne, at middel B er mere effektivt, end middel A.

Det approximative test har teststørrelsen

$$Q = \frac{(1-9)^2}{1+9} = 6.4$$

der skal sammenlignes med fraktilerne i en $\chi^2(1)$-fordeling. Idet $\chi^2_{0.95}(1) = 3.84$ leder også det approximative test til afvisning af hypotesen. □

Bemærkning 1 *Modellering ved individuelle responssandsynligheder*

Cox (1958) og Cox og Snell (1989) har foreslået en model for ovenstående situation, der direkte modellerer individeffekten.

Antag, at de n individer har hver sit sæt af responssandsynligheder og at responssandsynlighederne for det ν'te individ er:

$$\pi_{1,+,\nu} = \frac{\exp(\alpha_\nu)}{1+\exp(\alpha_\nu)}$$

for at respons A er 1, og

$$\pi_{+,1,\nu} = \frac{\exp(\alpha_\nu + \beta)}{1 + \exp(\alpha_\nu + \beta)}$$

for at respons B er 1, hvor α_ν er karakteristisk for det ν'te individ, og β er fælles for alle n individer.

Modellen udtrykker, at odds-ratioen (3.1.3)

$$\frac{\pi_{+,1,\nu}}{1-\pi_{+,1,\nu}} \Big/ \frac{\pi_{1,+,\nu}}{1-\pi_{1,+,\nu}} = \exp(\beta) \tag{3.4.27}$$

ikke afhænger af individparameteren $\alpha_n u$.

Odds-ratioen $\exp(\beta)$ (3.4.27) udtrykker forholdet mellem odds for respons 1 ved A og odds for respons 1 ved B. For $\beta = 0$ er der marginal symmetri.

Antag, at responsen ved de to kriterier er stokastisk uafhængige og lad $(X_{1,+,\nu}, X_{+,1,\nu})$ angive responserne for det ν'te individ, hvor

$$X_{1,+,\nu} = \begin{cases} 1 & \text{hvis respons } A \text{ er } 1 \\ 0 & \text{hvis respons } A \text{ er } 2 \end{cases}$$

og tilsvarende

$$X_{+,1,\nu} = \begin{cases} 1 & \text{hvis respons } B \text{ er } 1 \\ 0 & \text{hvis respons } B \text{ er } 2 \end{cases}$$

Den simultane sandsynlighed for observationssættet $(X_{1,+,\nu}, X_{+,1,\nu})$, $\nu = 1, 2, \ldots, n$ er da på grund af uafhængigheden produktet af to Bernoullifordelte størrelser:

$$\begin{aligned} f(\mathbf{x}; \alpha_1, \ldots, \alpha_n, \beta) &= \prod_{\nu=1}^{n} \left(\frac{\exp(\alpha_\nu)}{1+\exp(\alpha_\nu)}\right)^{x_{1,+,\nu}} \left(\frac{1}{1+\exp(\alpha_\nu)}\right)^{1-x_{1,+,\nu}} \\ &\times \left(\frac{\exp(\alpha_\nu+\beta)}{1+\exp(\alpha_\nu+\beta)}\right)^{x_{+,1,\nu}} \left(\frac{1}{1+\exp(\alpha_\nu+\beta)}\right)^{1-x_{+,1,\nu}} \end{aligned}$$

Idet nævneren $\left[\left(1+\exp(\alpha_\nu)\right)\left(1+\exp(\alpha_\nu+\beta)\right)\right]^n$ ikke afhænger af observationerne, finder vi at den observationsafhængige del af sandsynligheden er proportional med

$$\exp\left[\sum_{\nu=1}^{n} \alpha_\nu (x_{1,+,\nu} + x_{+,1,\nu}) + \beta \left(\sum_{\nu=1}^{n} x_{+,1,\nu}\right)\right]$$

Vi indfører nu symbolet

$$x^*_{+,+,\nu} = x_{1,+,\nu} + x_{+,1,\nu}$$

for summen af de to Bernoullivariable. Summen $x^*_{+,+,\nu}$ kan antage værdierne 0, 1 og 2.

Vi betragter nu den betingede fordeling af $(x_{1,+,\nu}, x_{+,1,\nu})$, givet summen $x^*_{+,+,\nu}$. Når $x^*_{+,+,\nu} = 0$ er $x_{1,+,\nu}$ og $x_{+,1,\nu}$ begge 0 med sandsynligheden 1, og når $x^*_{+,+,\nu} = 2$ er $x_{1,+,\nu}$ og $x_{+,1,\nu}$ begge 1 med sandsynligheden 1. Det er kun hvis $x^*_{+,+,\nu} = 1$, at den betingede fordeling af $(x_{1,+,\nu}, x_{+,1,\nu})$ overhovedet afhænger af parametrene.

Man finder

$$\mathsf{P}\left[X_{1,+,\nu} = x_{1,+,\nu}\,, X_{+,1,\nu} = x_{+,1,\nu} \mid X^*_{+,+,\nu} = 1\right]$$

$$= \begin{cases} \exp(\beta)/[1+\exp(\beta)] & \text{for } (x_{1,+,\nu}, x_{+,1,\nu}) = (0,1) \\ 1/[1+\exp(\beta)] & \text{for } (x_{1,+,\nu}, x_{+,1,\nu}) = (1,0) \end{cases}$$

Vi bemærker, at den betingede fordeling af $(x_{1,+,\nu}, x_{+,1,\nu})$, givet summen $x^*_{+,+,\nu}$ ikke afhænger af individparameteren α_ν, men kun af interesseparameteren β. Ydermere er observationer, hvor $x^*_{+,+,\nu}$ er 0 eller 2 uden information om parameteren β, hvorfor det kun er de observationspar, for hvilke responserne ved de to kriterier A og B er forskellige (dvs $x^*_{+,+,\nu} = 1$), der er relevante for testet af β.

Ialt er der $t = z_{1,2} + z_{2,1}$ observationer med $x^*_{+,+,\nu} = 1$.

Den betingede fordeling af observationssættet, givet $x^*_{+,+,\nu}$, $\nu = 1, \ldots, n$ har tætheden

$$\prod_{\nu: x^*_{+,+,\nu}=1} \left(\frac{1}{1+\exp(\beta)}\right)^{x_{1,+,\nu}} \left(\frac{\exp(\beta)}{1+\exp(\beta)}\right)^{x_{+,1,\nu}} = \left[\exp(\beta)\right]^{z_{2,1}} \left[1+\exp(\beta)\right]^{-t}. \tag{3.4.28}$$

Testet for hypotesen (3.4.19) om marginal symmetri kan således udføres ved at teste om $\beta = 0$ i ovenstående model. Idet $\beta = 0$ indebærer, at $\exp(\beta)/[1+\exp(\beta)] = 0.5$, er det altså det samme test som det eksakte test i sætning 3.4.5.

Vi bemærker i øvrigt, at maksimum likelihood estimatet for $\exp(\beta)$ i den betingede fordeling af observationerne netop er

$$\exp(\widehat{\beta}) = \frac{z_{2,1}}{z_{1,2}}$$

J.Gart (1969) har udvidet ovenstående model til at inddrage information om den rækkefølge, hvori de to behandlinger udføres. I denne udvidelse af modellen kan man formelt teste om rækkefølgen har en betydning. □

3.5 Modeller for parvise sammenligninger

Betragt en situation, hvor k objekter, $\nu = 1, 2, \ldots, k$ sammenlignes ved parvise sammenligninger, og hvor hver sammenligning resulterer i en angivelse af hvilket af de to objekter, der foretrækkes.

Der er ialt $\binom{n}{2} = n(n-1)/2$ mulige par, idet vi ser bort fra rækkefølgen, og opfatter de enkelte par som uordnede.

Vi antager, at en sammenligning af objekterne i og j har de to svarmuligheder $i \succ j$ svarende til "i foretrækkes for j", og $i \prec j$, svarende til "j foretrækkes for i". (Vi tillader altså ikke "uafgjort").

Metoden bruges ofte ved smagstest og vurdering af andre imponderalia, hvor det kan være vanskeligt for de enkelte bedømmere at rangordne samtlige n objekter, men hvor bedømmeren er i stand til at angive sin præference, når han blot skal sammenligne to objekter.

Lad

$$p_{i,j} = \mathsf{P}\ [i \succ j]\ , i = 1, \ldots, k\ ;\ , j = 1, \ldots, k\ ;\ i \neq j \tag{3.5.1}$$

angive sandsynligheden for at objekt i foretrækkes for objekt j.

Antagelsen om "tvunget valg" indebærer, at

$$p_{j,i} = 1 - p_{i,j}$$

3.5.1 Bradley-Terry modellen

I Bradley-Terry modellen for parvise sammenligninger (Bradley, R. A. and Terry, M. E. (1952)) antages, at der findes et sæt af ikke-negative parametre, $\pi_1, \pi_2, \ldots, \pi_k$, sådan at

$$p_{i,j} = \frac{\pi_i}{\pi_i + \pi_j} \tag{3.5.2}$$

Indfører vi nu

$$\phi_i = \ln(\pi_i) \tag{3.5.3}$$

ser vi, at logit-værdierne svarende til modellen (3.5.2) kan udtrykkes som

$$\eta_{i,j} = \text{logit}(p_{i,j}) = \ln\left(\frac{p_{i,j}}{1-p_{i,j}}\right) = \phi_i - \phi_j \tag{3.5.4}$$

Antag nu, at der er udført $n_{i,j}$, $i = 1, 2, \ldots, k$, $j = i+1, \ldots, k$ sammenligninger mellem objekt i og objekt j, og lad $z_{i,j}$ angive antallet blandt disse, hvor i blev foretrukket for j.

Vi har da, at $Z_{i,j} \in \text{B}(n_{i,j}, p_{i,j})$, hvorfor loglikelihoodfunktionen svarende til dette eksperiment er

$$\begin{aligned} l(\mathbf{p};\mathbf{z}) &= \sum_{i=1}^{k}\sum_{j=i+1}^{k}[z_{i,j}\text{logit}(p_{i,j}) + n_{i,j}\ln(1-p_{i,j})] \\ &= \sum_{i=1}^{k}\sum_{j=i+1}^{k} z_{i,j}(\phi_i - \phi_j) + \sum_{i=1}^{k}\sum_{j=i+1}^{k} n_{i,j}\ln(1-p_{i,j}) \end{aligned}$$

Den sidste dobbeltsum afhænger ikke af observationerne. Vi betragter nu den første sum:

$$\sum_{i=1}^{k}\sum_{j=i+1}^{k} z_{i,j}(\phi_i - \phi_j) = \sum_{i=1}^{k}\phi_i \sum_{j=i+1}^{k} z_{i,j} - \sum_{i=1}^{k}\sum_{j=i+1}^{k} z_{i,j}\phi_j$$

Men idet

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=i+1}^{k} z_{i,j}\phi_j &= \sum_{j=1}^{k}\sum_{i=1}^{j} z_{i,j}\phi_j \\ &= \sum_{j=1}^{k}\phi_j\sum_{i=1}^{j} z_{i,j} \\ &= \sum_{i=1}^{k}\phi_i\sum_{j=1}^{i} z_{j,i}\end{aligned}$$

har vi at

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=i+1}^{k} z_{i,j}(\phi_i-\phi_j) &= \sum_{i=1}^{k}\phi_i\left[\sum_{j=i+1}^{k} z_{i,j}-\sum_{j=1}^{i} z_{j,i}\right] \\ &= \sum_{i=1}^{k}\phi_i S_i\end{aligned}$$

hvor

$$S_i = \sum_{j=i+1}^{k} z_{i,j} - \sum_{j=1}^{i} z_{j,i}$$

angiver i'te objekts score, nemlig antallet af sammenligninger, hvor i indgik og blev foretrukket minus antallet af sammenligninger, hvor i indgik og ikke blev foretrukket. Vi ser altså, at sættet af scores S_i er sufficient for parametrene ϕ_i.

For balancerede situationer (dvs alle $n_{i,j}$ lige store) gælder, at den rangorden, der defineres ved score, overføres i ordningen af $\{\phi_i\}$.

Eksempel 3.5.1 *Superligaen i fodbold*

Nedenstående tabel viser resultaterne for udvalgte hold i superligaen i fodbold sommeren 1994:

	udehold			
	Brøndby	Fremad A	FC Køb.	Ikast
Brøndby		3-0	0-4	1-0
Fremad A	1-6		3-2	3-2
FC Københ.	2-1	2-3		3-2
Ikast	0-2	4-0	2-0	

Vi reducerer tabellen til at angive antallet af vundne og tabte kampe:

	taber			
	Brøndby	Fremad A	FC Køb.	Ikast
Brøndby		2	0	2
Fremad A	0		2	1
FC Køenh.	2	0		1
Ikast	0	1	1	

Nedenstående tabel giver antal kampe $n_{i,j}$, og antal vundne kampe $z_{i,j}$ for hver af de 6 parvise møder:

```
Brond FrA   2  2
Brond FcK   2  0
Brond Ika   2  2
FrA   FcK   2  2
FrA   Ika   2  1
FcK   Ika   2  1
```

Vi kan altså formulere en generaliseret lineær model med en logistisk linkfunktion til beskrivelse af de 6 observationer $z_{i,j}$.

Idet observationerne betragtes i den ovenstående rækkefølge har vi modelmatricen

```
 FrA FcK Ika

-1   0   0
 0  -1   0
 0   0  -1
 1  -1   0
 1   0  -1
 0   1  -1
```

hvor vi har sat $\phi_{Brond} = 0$, og hvor der ikke er noget interceptled.

I S-plus kan parametrene estimeres ved kald af proceduren `glim`. Kaldet er vist nedenfor:

```
fodb <- glim(modmatr,sejr,antal, error="binomial",
   link="logit",intercept=F)
```

hvor `modmatr` angiver modelmatricen og `sejr` og `antal` angiver vektorerne med henholdsvis antallet af sejre og antallet af kampe for de seks sammenligninger

Man får koefficienterne ϕ_i:

```
     FrA          FcK          Ika

-0.528031    -0.528031    -1.056077
```

svarende til π-værdierne

```
Bron          FrA          FcK          Ika
1.000     0.5897651    0.5897651    0.3478176
```

Den herved etablerede ordning er naturligvis i overensstemmelse med de opnåede scores (i betydningen antal vundne minus antallet tabte kampe) for de fire hold. Man har scores:

```
Bron       FrA          FcK          Ika
2           0            0           -2
```

Da Fremad Amager og FC København har opnået samme score, tillægges de også samme π-værdi. □

3.6 Referencer

Agresti, A. (1990): *Categorical Data Analysis*, New York, Wiley

Baptista, J and Pike, M. C.: (1977) Algorithm AS 115: Exact two-sided confidence limits for the odds ratio in a 2 × 2 table. *Appl.Statist.* **26** pp 214-220

Bradley, R. A. and Terry, M. E. (1952): Rank analysis of incomplete block designs I: The method of paired comparisons. *Biometrika* **39**, pp 324-345

Cornfield, J. (1956): A statistical problem arising from retrospective studies. *Proc. Third Berkeley Symposium on Math.Statist. and Probab.*, J. Neyman, ed. **4**, pp. 135-148

Cox, D.R. (1958): Two further applications of a model for binary regression, *Biometrika*, **45**, 562-565.

Cox, D. R. and Snell, E. J. (1989): *The Analysis of Binary Data*, London: Chapman and Hall

Dyke, A.V. and Patterson H.D. (1952): Analysis of factorial models when the data are proportions. *Biometrics* **8**, pp 1-12 .

Gart, J. J. (1962): Approximate confidence limits for the relative risk, *Journ. Roy Statist. Soc.* **B 24** pp. 454-463

Gart, J. J. and Zweiful, J. R. (1967): On the bias of various estimators of the logit and its variance with respect to quantal bioassay. *Biometrika* **54** pp. 181-187

Gart, J. J. (1969): An exact test for comparing matched proportions in crossover designs. *Biometrika*, **56**, pp 75-80.

Gart, J. J. and Thomas, D. G. (1972): Numerical results on approximate confidence limits for the odds ratio. *Journ. Roy Statist. Soc.* **B 34** pp. 441-447

Haldane, J. B. S. (1955) The estimation and significance of the logarithm of a ratio of frequencies. *Ann. Human Genet.* **20**, pp. 309-311

Lombard, H. L. and Doering, C. R. (1947) Treatment of the four-fold table correlations as it relates to public heath problems. *Biometrics* **3**, pp 123-128.

McNemar, Q. (1947): Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, **12**, pp 153-157.

Mosteller, F. (1952): Some statistical problems in measuring the subjective response to drugs. *Biometrics*, **8**, pp 220-226.

Plackett, R. L. (1981): *The Analysis of Categorical Data*, 2nd edition, London, Griffin.

Rasch, G: (1960) *Probabilistic Models for Some Intelligence and Attainment Tests*, Danmarks Pædagogiske Institut.

Tosteson, A. N. A. og Begg, C. B. (1988): A general regression methodology for ROC curve estimation. *Medical Decision Making* **8:** pp 204-215.

Indeks