

Kapitel 4

Introduktion til generaliserede lineære modeller

4.0 Oversigt

Gennemgangen indledes med et motiverende eksempel (logistisk regression) i afsnit 4.2. Efter den indledende abstrakte gennemgang betragtes dette eksempel igen i afsnit 4.6.3 på side 128.

Lige som den generelle lineære model er intimt knyttet til normalfordelingsmodeller, er de generaliserede lineære modeller knyttet til nogle bestemte klasser af fordelinger, de såkaldte *eksponentielle dispersionsfamilier*.

Vi indleder derfor i afsnit 4.3.1 med at introducere disse familier og beskrive egenskaber ved netop disse familier. Ved et specielt valg af parametrisering, kan fordelingernes tætheder udtrykkes på en fælles grundform (den *kanoniske form*) (4.5), og med et parameterområde for den *kanoniske parameter*, der i mange tilfælde er hele den reelle akse.

Med henblik på anvendelserne er der imidlertid lige så praktisk at benytte en parametrisering ved middelværdien μ . Denne parametrisering indføres i (4.11). Nogle vigtige egenskaber for fordelingerne kan netop udtrykkes ved middelværdiparametriseringen:

- variansen er en kendt funktion (*variansfunktionen*) af middelværdien, μ ; (def. 4.3.4)

- logaritmen til sandsynlighedstætheden afhænger gennem en funktion, *deviansfunktionen* af observationen y og middelværdien μ ; (4.16)
- deviansfunktionen, der udtrykker afvigelsen mellem en observation y og modelværdien μ , er bestemt af fordelings variansfunktion; (4.14)

I afsnit 4.4 betragtes likelihoodfunktion, scorefunktion og informationsmatrix for et observationssæt fra en eksponentiel dispersionsparameterfamilie.

De generaliserede lineære modeller indføres nu i afsnit 4.5:

- I stedet for at betragte lineære modeller for middelværdien af observationerne, betragter vi lineære modeller for en funktion, *linkfunktionen* (4.32) af middelværdien
- i stedet for at minimere kvadratafgivelsessummer og opspalte kvadrat-afgivelsessummer, minimerer og opspalter vi *deviansen* ; (4.22)
- vi kan umiddelbart vurdere modeltilpasning (goodness of fit) ved at udføre et χ^2 -test for deviansen svarende til den maksimale model; (sætning 4.8.1)
- og vi kan tilsvarende udføre test for modelreduktion ved χ^2 -test for reduktion af deviansen (sætning 4.10.1) ; (4.87)

Fordelingsforhold for estimatorer, fittede værdier, residualer m.v. modsvarer situationen med normalfordelte observationer, blot gælder egenskaberne kun approximativt, da de bestemmes ved linearisering af den ikke-lineære model. Gennemgangen i afsnit 4.7 til 4.11 modsvarer derfor gennemgangen af den generelle lineære model - dog med den undtagelse at for generaliserede lineære modeller er det muligt at teste modeltilpasningen som vist i afsnit 4.8. Gennemgangen afsluttes i afsnit 4.11.1 med et detaljeret eksempel.

I afsnit 4.12 introduceres en udvidelse af modelklassen til at omfatte modeller med *overdispersion*, og i afsnit 4.13 skitseres håndteringen af generaliserede lineære modeller i dedikeret statistik-software (SAS og S-plus).

Den generelle lineære model for normalfordelte data omfatter både rene “regressionsmodeller” (med kontinuerte forklarende variable) og modeller med

“faktorstruktur” (med kategoriske forklarende variable) samt kombinationer af disse. Udvidelsen til forskellige responsfordelinger føjer yderligere en dimension til disse modeller. I afsnit 4.14 betragtes eksempler på “regressionsmodeller” for forskellige responsfordelinger, og i afsnit 4.15 betragtes eksempler på tofaktormodeller for forskellige responsfordelinger.

4.1 Baggrund og hovedideer

Den generelle lineære model er bekvem til analyse af lineære middelværdistrukturer, fordi

- man kan bestemme eksplicitte mindste kvadraters løsninger ved simple matrixoperationer,
- og fortolkningen ved begreber fra vektorrumsteorien giver indsigt i hvad der foregaar ved estimation og modelreduktion.

Den generelle lineære model og mindste kvadraters metode har været kendt, siden Gauss diskuterede dem i 1809 og 1811. I 1821 og 1823 viste Gauss, at disse mindste kvadraters estimatorer havde mindst varians blandt lineære estimatorer i normalfordelingsmodellen. Laplace (1816), (og også den danske statistiker Thiele (1889, 1897, 1903)) diskuterede ortogonaliseringen af modellen og Thiele beskrev en kanonisk (universel) form for den generelle lineære model. Formuleringen i abstrakt vektorrumsnotation udvikledes i 1960’erne. For en nogenlunde tidssvarende matematisk statistisk fremstilling af den generelle lineære model se fx Christensen [1].

Sideløbende med udviklingen af denne generelle indsigt i egenskaberne ved den generelle lineære model, opstod der i 1960’erne en række teoridannelser omkring andre, specielle problemstillinger og modeller. Blandt andet kan nævnes arbejder omkring modellering og analyse af *antalstabeller* (contingency-tables), der indpassede analyse af strukturer i antalstabeller i en generel modelklasse, *log-lineære modeller*. En hovedide i analysen af denne klasse er indlejringen af logaritmen til middelværdiparameteren i Poisson-fordelingen i et vektorrum og udnyttelse af vektorrumsstrukturen (ortogonalitet/uafhængighed) i formulering og analyse af modellerne. Se McCulloch og Nelder [6] for en beskrivelse af nogle af de udviklinger, der har dannet baggrund for den nuværende teori.

Samtidigt udvikledes også en større indsigt i de fundamentale egenskaber for de såkaldte *eksponentielle familier af fordelinger* jvf nedenfor afsnit 4.3.1. Med baggrund i denne indsigt formulerede Nelder og Wedderburn i 1972 en generel teori

for analyse af lineære strukturer for en funktion (linkfunktionen) af middelværdien af observationer, hvis fordeling tilhører en eksponential familie. Hovedreferencen til denne modelklasse er stadig McCullagh og Nelder [6]. Der er siden fremkommet en række andre lærebøger, der beskriver disse modeller, også ud fra et mere anvendelsesorienteret synspunkt . (fx. af Lindsey [10], af Fahrmeir and Tutz [8], Myers, Montgomery and Vining [12] og McCulloch and Searle [11])

Den generelle lineære model knytter sig stærkt til en modellering af observationerne som “middelværdi plus støj”, hvor “støjbidraget” antages at være normaltfordelt og med samme varians, og hvor middelværdi tænkes at have en lineær struktur. Mens disse antagelser kan tænkes opfyldt i nogen grad for kontinuerte målinger, stiller sagen sig anderledes når målingerne er resultatet af optællinger, eller er forhold mellem optællinger (andele). I sådanne situationer, som sædvanligvis modelleres ved Poisson- eller binomialfordelingsmodeller, er fordelingerne “skæve”, der er ikke samme varians, og middelværdistrukturen er ikke lineær. De generaliserede lineære modeller tilgodeser netop disse forhold.

4.2 Motiverende eksempel

Som introduktion til den abstrakte gennemgang betragter vi indledningsvist et eksempel.

Eksempel 4.2.1 *Fosterdødelighed hos mus*

Data fra Price et al. *The developmental toxicity of diethylene glycol dimethyl ether in mice* [35]

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

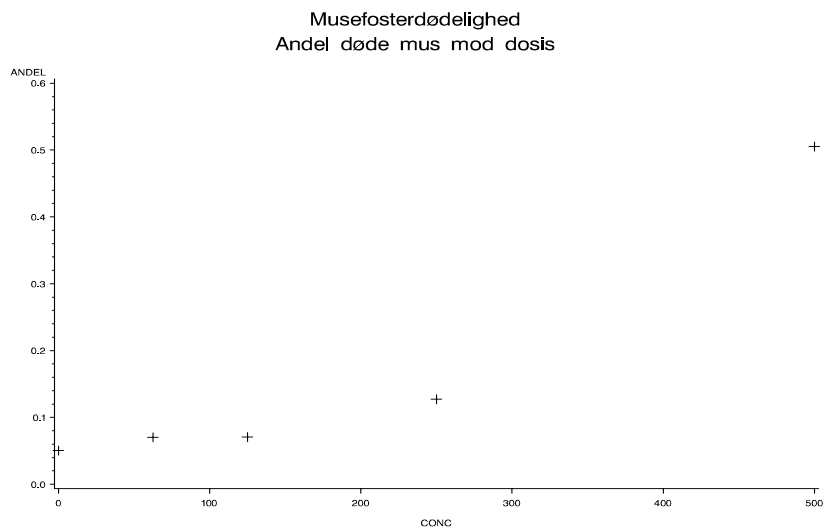
Antallet af døde fostre som funktion af den indgivne koncentration er vist i nedenstående tabel:

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

Den interessante størrelse er *andelen* af døde fostre; men en lineær model for andelen af døde fostre

$$p_i = \beta_1 + \beta_2 x_i$$

vil ikke afspejle det fænomen, at andelen ligger mellem nul og een, så allerede her er der et problem.



Endvidere er der også et problem med varianshomogenitet. Det vil være naturligt at modellere *antallet* af døde fostre, Z_i , ved en given koncentration, x_i ved en $B(n_i, p_i)$ -fordelt variabel, og dette indebærer for $Y_i = Z_i/n_i$, at

$$V[Y_i] = \frac{p_i(1 - p_i)}{n_i}$$

dvs variansen for den i 'te observation afhænger af antallet af eksponerede fostre, n_i , (som vi kender og måske nok kan korrigere for), og af dødssandsynligheden, p_i ved dosis x_i , og det er jo netop denne størrelse, p_i , som vi ønsker at bestemme.

Endelig kan man også bemærke, at den sædvanlige *middelværdi plus støjbidrag* model ikke er særligt hensigtsmæssig til at beskrive variationen af den observerede *andel*. Dels afhænger, som nævnt, variansen på støjbidraget af middelværdien, og dels kan det samlede resultat af middelværdi plus støjbidrag jo ikke overstige 1.

Vi kan altså ikke klare os med den generelle lineære model, men må udvide begrebsapparatet

I stedet for en lineær model for *middelværdien* (her p_i) af observationerne Y_i vil vi vælge en funktion, *linkfunktionen* af middelværdien; her vælges logit-transformationen

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

og formulere en *lineær model* for disse transformerede middelværdier

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

nemlig modellen med den *lineære prædiktør*

$$H_0 : \quad \eta_i = \beta_1 + \beta_2 x_i, \quad i = 1, 2, \dots, 5, \quad (4.1)$$

Transformationen tilbage til dødssandsynligheder, p_i er den *logistiske funktion*

$$H_0 : \quad p_i = \exp(\beta_1 + \beta_2 x_i) / \{1 + \exp(\beta_1 + \beta_2 x_i)\}, \quad i = 1, 2, \dots, k$$

og modellen kaldes undertiden en *logistisk regressionsmodel*.

Det bekvemme ved transformationen $\eta = \ln \left(p/(1-p) \right)$ er at den fører $p \in]0, 1[$ over i hele den reelle akse $\eta \in \mathbb{R}$, hvor det er naturligt at benytte lineære modeller.

▽

4.3 Naturlige eksponentielle familier, eksponentielle dispersionsfamilier

4.3.1 Naturlige eksponentielle familier

De fordelinger (Normal-, Binomial-, Negativ Binomial-, Poisson-, Gamma-fordeling), der sædvanligvis betragtes i indledende statistikkurser, 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 for middelværdiparameteren. 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*. (En familie af fordelinger er bare en parametriseret samling af fordelinger, fx samlingen af binomialfordelinger med antalsparameter n for alle sandsynlighedsparametre p).

Definition 4.3.1 *Naturlig eksponentiel familie*

De såkaldte *naturlige eksponentielle familier* er familier af fordelinger, hvis tætheder (eller sandsynlighedsfunktioner) med hensyn til et forelagt mål (oftest Lebesguemålet eller tællemålet) kan skrives på formen

$$f_X(x; \theta) = c(x) \exp\{\theta x - \kappa(\theta)\} \quad (4.2)$$

Formen (4.2) kaldes *den kanoniske form* (dvs grundformen). Når familien er fremstillet på denne form, kaldes parameteren θ for *den kanoniske parameter*, og funktionen $\kappa(\cdot)$ kaldes *kumulantfrembringeren*.

(For $X \sim N(\theta, 1)$ er tætheden for fordelingen af X netop på ovenstående form med $\kappa(\theta) = \theta^2/2$).

▽

Tabel 4.1 på side 115 indeholder en oversigt over nogle sædvanlige fordelinger, der kan udtrykkes som eksponentielle familier.

Eksempel 4.3.1 Bernoullifordeling som naturlig eksponentiel familie

Lad X følge en Bernoullifordeling med parameteren p . Frekvensfunktionen for fordelingen af X er da

$$g_X(x; p) = p^x(1-p)^{1-x} = (1-p) \left(\frac{p}{1-p} \right)^x \quad \text{for } x = 0, 1 \quad (4.3)$$

for $0 \leq p \leq 1$.

Indfører vi nu

$$\theta = \log \left(\frac{p}{1-p} \right) \quad \text{for } 0 < p < 1$$

ser vi, at frekvensfunktionen (4.3) kan skrives som

$$f_X(x; \theta) = \exp\{\theta x - \log(1 + \exp(\theta))\}$$

der netop er på formen (4.2) med

$$\kappa(\theta) = \log(1 + \exp(\theta))$$

Vi bemærker i øvrigt, at ved repræsentationen af familien af Bernoullifordelinger som en eksponentiel familie mister vi de to udartede fordelinger svarende til hhv $p = 0$ og $p = 1$. ∇

Eksempel 4.3.2 Exponentialfordeling som naturlig eksponentiel familie

Lad X følge en exponentialfordeling med skalaparameter β . Tæthedsfunktionen for X er da

$$g_X(x; \beta) = \frac{1}{\beta} \exp(-x/\beta) \quad \text{for } 0 < x \quad (4.4)$$

hvor $0 < \beta$.

Indfører vi nu

$$\theta = \frac{-1}{\beta} \quad \text{for } 0 < \beta$$

ser vi, at frekvensfunktionen (4.4) kan skrives som

$$f_X(x; \theta) = \exp\{\theta x - \log(-\theta)\}$$

der netop er på formen (4.2) med

$$\kappa(\theta) = \log(-\theta)$$

∇

Bemærkning 1 *Relation mellem observation og parameter*

Ved omskrivning af grundformen til

$$f_X(x; \theta) = c(x)b(\theta) \exp\{\theta x\}$$

ser man, at tætheden (og likelihoodfunktionen) faktoriseres i et produkt af en faktor, der alene afhænger af observationen, x , en faktor, der alene afhænger af parameteren, θ , og endelig faktoren, $\exp\{\theta x\}$, hvor både x og θ indgår. Dette indebærer, at ved et sæt af uafhængige, identisk fordelte observationer, $X_1, \dots, X_k$, vil $T = \sum X_i$ være *sufficient for* θ .

Ser vi specielt på bidraget til log-likelihood funktionen,

$$\ell(\theta; x) = \theta x - \kappa(\theta)$$

for to værdier, $x_1 < x_2$, finder man, at for ethvert θ_1 findes der en værdi, θ_2 med $\theta_1 < \theta_2$, sådan

$$\ell(\theta_1; x_1) = \ell(\theta_2; x_2)$$

En ændring, $x_1 \rightarrow x_2$ kan altså opvejes af en ændring, $\theta_1 \rightarrow \theta_2$ sådan at man opnår samme likelihood.

Man siger, at familien har *monoton likelihood -ratio*.

Ændringen, $\theta_1 \rightarrow \theta_2$ er bestemt af kumulantfrembringeren $\kappa(\cdot)$. Kumulantfrembringeren er således bestemmende for størrelsen af den ændring i θ , der modsvarer (i form af uændret likelihood) en ændring i x .

▽

4.3.2 Eksponentielle dispersionsfamilier

Af hensyn til anvendelserne i forbindelse med generaliserede lineære modeller indfører vi straks en ekstra parameter på den naturlige eksponentielle familie, en *precisionparameter*, som i denne fremstilling kaldes λ .

Definition 4.3.2 *Eksponentiel dispersionsfamilie*

En *eksponentiel dispersionsfamilie* af fordelinger for den variable Y har tæthedsfunktion (eller sandsynlighedsfunktion) på formen

$f_Y(y; \theta, \lambda) = c(y, \lambda) \exp[\lambda\{\theta y - \kappa(\theta)\}]$	(4.5)
--	-------

hvor parameterrummet, Ω for den *kanoniske parameter*, θ , er en delmængde af den reelle akse. For de fleste sædvanlige modeller, er parameterrummet for θ hele den reelle akse. Parameteren $\lambda > 0$ kaldes *præcisions parameteren*. (For $Y \sim N(\theta, \sigma^2)$ er tætheden for fordelingen af Y på formen (4.5) med $\kappa(\theta) = \theta^2/2$ og $\lambda = 1/\sigma^2$).

▽

Eksempel 4.3.3 Binomialfordelingen som eksponentiel dispersionsfamilie

Lad Z følge en Binomialfordeling med antalsparameter n og sandsynlighedsparameter p . Frekvensfunktionen for fordelingen af Z er da

$$g_Z(z; p) = \binom{n}{z} p^z (1-p)^{n-z} = \binom{n}{x} (1-p)^n \left(\frac{p}{1-p} \right)^z \quad \text{for } z = 0, \dots, n$$

for $0 \leq p \leq 1$.

og frekvensfunktionen for fordelingen af $Y = Z/n$ er derfor

$$g_Y(y; p) = \binom{n}{ny} p^{ny} (1-p)^{n(1-y)} = \binom{n}{x} (1-p)^n \left(\frac{p}{1-p} \right)^{ny} \quad (4.6)$$

for $y = 0, 1/n, \dots, 1$ og $0 \leq p \leq 1$.

Indfører vi som før

$$\theta = \log \left(\frac{p}{1-p} \right) \quad \text{for } 0 < p < 1$$

ser vi, at frekvensfunktionen (4.6) kan skrives som

$$f_Y(y; \theta) = \exp \{ n(\theta y - \log(1 + \exp(\theta))) \}$$

der netop er på formen (4.5) med

$$\kappa(\theta) = \log(1 + \exp(\theta))$$

og $\lambda = n$.

▽

Eksempel 4.3.4 Gammafordeling som eksponentiel dispersionsfamilie

Lad X følge en Gammafordeling $G(\alpha, \beta)$ med formparameter α og skalaparameter β . Tæthedsfunktionen for X er da

$$g_X(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta} \left(\frac{x}{\beta} \right)^{\alpha-1} \exp(-x/\beta) \quad \text{for } 0 < x$$

hvor $0 < \beta$.

Vi vil i stedet betragte $Y = X/\alpha \in G(\alpha, \beta/\alpha)$ som har tæthedsfunktionen

$$g_Y(y; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta/\alpha} \left(\frac{y}{\beta/\alpha} \right)^{\alpha-1} \exp(-y\alpha/\beta) \quad \text{for } 0 < y \quad (4.7)$$

Indfører vi som før

$$\theta = \frac{-1}{\beta} \quad \text{for } 0 < \beta$$

ser vi, at frekvensfunktionen (4.7) kan skrives som

$$f_Y(y; \alpha, \theta) = \frac{y^{\alpha-1} \alpha^\alpha}{\Gamma(\alpha)} \exp \left\{ \alpha \left(\theta y - \log(-\theta) \right) \right\}$$

der netop er på formen (4.2) med

$$\kappa(\theta) = \log(-\theta)$$

og $\lambda = \alpha$.

Fortolkningen af præcisionsparameteren λ som et antal “elementarobservationer” er i analogi med det foregående eksempel. For heltallige værdier af α kan fordelingen $X \in G(\alpha, \beta)$ nemlig opfattes som fordeling af summen af α uafhængige $\text{Ex}(\beta)$ fordelte variable. ∇

Afsnit 4.17 giver en uddybende beskrivelse af teorien for eksponentielle familier. For en udførligere introduktion til de eksponentielle dispersionsfamilier henvises til Jørgensen, [5].

Udvidelsen af klassen af naturlige eksponentielle familier med denne præcisionsparameter har til formål at adskille de middelværdirelaterede egenskaber, der er karakteriseret gennem *kumulantfrembringeren* $\kappa(\cdot)$ fra sådanne dispersionsegenskaber, som karakteriseres af stikprøvestørrelse, fælles varians, fælles overdispersion mv, der ikke er knyttet til middelværdien.

Vi vil se, at i nogle sammenhænge angiver denne præcisionsparameter et kendt antal observationer, i andre sammenhænge repræsenterer den en ukendt dispersion, som i normalfordelingen. I disse sidste situationer bruger man ofte symbolet $\sigma^2 = 1/\lambda$ for denne *dispersionsparameter*. Og i SAS bruges betegnelsen *scale* for en funktion af dispersionsparameteren $1/\lambda$; fx bruges betegnelsen *scale* for spredningen σ i normalfordelingen, så man bør være påpasselig

med fortolkningen af *scale* i SAS.

Bemærkning 1 *Dispersionsparameteren indgår sammen med y i normeringskonstanten*

Vi bemærker, at dispersionsparameteren, $\sigma^2 = 1/\lambda$, indgår i normeringskonstanten $c(y; \lambda)$, dvs sammenhængen mellem observationer og dispersionsparameter er ikke så simpel, som mellem observation og kanonisk parameter. Det er blandt andet dette forhold, der ført til indførelse af begrebet quasi-likelihood til håndtering af situationer med ukendt dispersionsparameter, se afsnit 4.12.1

▽

Bemærkning 2 *Andre grundformer af eksponentielle dispersionsfamilier*

Undertiden ser man exponentielle repræsenteret på formen

$$f(y; \theta, \phi) = \exp \left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right) \quad (4.8)$$

i lighed med McCulloch og Nelders oprindelige repræsentation. (Denne repræsentation benyttes blandt andet i SAS' referencemanual).

Den væsentligste forskel fra (4.5) er, at præcisionsparameteren λ er udtrykt som en funktion $1/a(\phi)$ af en parameter ϕ , der har relation til en såkaldt *scale parameter*.

▽

4.3.3 Middelværdiafbildning, kanonisk link, variansfunktion, og enhedsdevians

Egenskaberne for den eksponentielle dispersionsfamilie er i det væsentlige bestemt af kumulantfrembringeren $\kappa(\cdot)$ og præcisionsparameteren λ . Der gælder

$$\mathbb{E}[Y] = \kappa'(\theta) \quad (4.9)$$

$$\mathbb{V}[Y] = \frac{\kappa''(\theta)}{\lambda}, \quad (4.10)$$

og det følger derfor, at den monotone funktion

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

definerer en enentydig afbildning

$$\tau(\theta) = \kappa'(\theta) \quad (4.11)$$

$\mu = \tau(\theta)$ af parameterrummet Ω , for den kanoniske parameter, θ , ind på en delmængde, $\mathcal{M}$, af den reelle akse, *middelværdirummet*.

Groft taget er middelværdirummet det konvekse hylster for fordelings støtte. Middelværdirummet er der, hvor observationerne er, fx hvis observationerne er andele døde musefostre (med mulige værdier $0/n, 1/n, \dots, n/n$) er middelværdirummet delmængden $0 < \mu < 1$ af den reelle akse. Dette til forskel for den kanoniske parameter, som befinder sig i et mere abstrakt vektorrum.

Middelværdiafbildningen $\tau(\cdot)$ er monoton, idet vi har

$$\tau'(\theta) = \kappa''(\theta) = \lambda \mathbb{V}[Y] > 0$$

Definition 4.3.3 *Kanonisk link, linkfunktion, lineær prædikator*

Relationen mellem de to parametriseringer er bestemt ved $\mu = \tau(\theta)$ med $\tau(\cdot)$ givet ved (4.11). Den omvendte afbildning

$$\theta = \tau^{-1}(\mu) \quad (4.12)$$

kaldes *den kanoniske linkfunktion*.

I teorien for generaliserede lineære modeller er en *link funktion* en vilkårlig monoton afbildning, $\eta = g(\mu)$ af middelværdirummet ind på (en delmængde)

af den reelle akse, der har til formål at danne *den lineære prædiktør*.

▽

Bemærkning 1 *Middelværdirum og kanonisk parameterum.
Prisen for at benytte den kanoniske parameter*

Middelværdirummet, $\mathcal{M}$, er billedmængden, $\tau(\Omega)$ for det kanoniske parameterområde, Ω , og er således indeholdt i det konvekse hylster for fordelings støtte.

Men betragt fx. binomialfordelingen, $Y = Z/n$, med $Z \sim \text{Bin}(n, p)$ og med kanonisk parameter, $\theta = \ln(p/(1-p)) \in]-\infty, \infty[$. Det konvekse hylster for fordelings støtte er det lukkede interval $[0, 1]$, men den kanoniske parameter θ kan ikke frembringe værdierne $p = 0$ og $p = 1$ (svarende til $\theta = \pm\infty$), så disse værdier af middelværdiparameteren, p , hører ikke med til middelværdirummet $\mathcal{M} =]0, 1[$.

Dette giver os problemer, når vi vil maksimum-likelihood estimere θ svarende fx til observationen $y = 0$, for “estimatet” er jo $\theta = -\infty$ som ikke eksisterer.

Havde vi bare holdt os til middelværdiparameteren p , og tilladt parameterummet, $p \in [0, 1]$ ville maksimum af likelihoodfunktionen ligge på randen af parameterområdet, men det ville såmænd ikke være så slemt.

Prisen for at bruge den kanoniske parameter kan altså være, at vi kan ramme ind i situationer, hvor maksimum-likelihood estimatet ikke eksisterer.

▽

Definition 4.3.4 *Enhedsvariansfunktion*

Ved at indføre middelværdiparameteren i (4.10) kan man udtrykke fordelings varians som en funktion, *enhedsvariansfunktionen* af middelværdien i fordelingen som

$$V(\mu) = \kappa''(\tau^{-1}(\mu)) \quad (4.13)$$

▽

Det kan vises (afsnit 4.17), at en naturlig eksponentiel familie er entydigt fastlagt af sin variansfunktion, og derfor er også en eksponentiel dispersionsfamilie entydigt bestemt af sin variansfunktion.

Der er således to ækvivalente parametriseringer af en eksponentiel dispersionsfamilie, nemlig

- ved *kanonisk parameter*, θ og *kumulantfrembringer*, $\kappa(\theta)$ med middelværdiafbildning mv.
- eller ved *middelværdiparameter*, μ og *variansfunktion*, $V(\mu)$.

BEMÆRK i øvrigt forskellen på *variansfunktionen*, $V(\cdot)$, som er en *reel funktion* af fordelings middelværdi, og *variansen*, $V[\cdot]$ som er knyttet til *fordelingen* af en stokastisk variabel.

Definition 4.3.5 *Devians (enhedsdevians)*

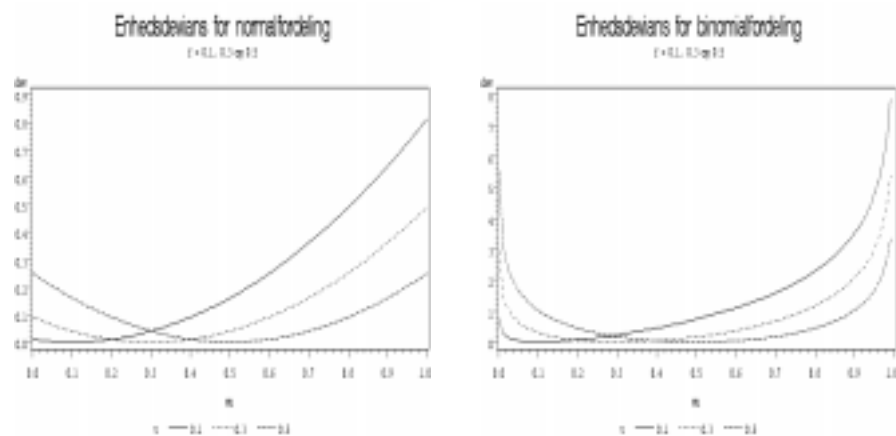
Parametriseringen ved middelværdiparameteren, μ , letter sammenligningen af observationer, y , med parameterværdier, μ , der er bestemt svarende til en bestemt model. Til brug for sammenligning af en observation, y , og en modelværdi, μ , indfører vi nu (forklaring kommer senere) *enhedsdeviansen*

$$d(y; \mu) = 2 \int_{\mu}^y \frac{y - u}{V(u)} du \quad , \quad (4.14)$$

hvor $V(\cdot)$ angiver variansfunktionen. Enhedsdeviansen udtrykker således afvigelsen mellem y og μ , vægtet med den reciproke varians for enhver værdi mellem y og μ . Og det er da egentlig i sig selv et meget fornuftigt princip at vægte disse punkter med præcisionen (den reciproke varians/usikkerhed) svarende til disse potentielle middelværdier.

For normalfordelingen, hvor variansen ikke afhænger af middelværdien, dvs $V(\mu) = 1$, er enhedsdeviansen blot kvadratafgigelsen $d(y; \mu) = (y - \mu)^2$. For en gammafordeling, er *variationskoefficienten* uafhængig af middelværdien, dvs $V(\mu) = \mu^2$, og enhedsdeviansen bliver da $d(y; \mu) = 2 \times \{y/\mu - \ln(y/\mu) - 1\}$. ∇

Figur 4.1 viser enhedsdeviansen svarende til hhv en normalfordelingsmodel og en binomialfordelingsmodel for observationerne $y = 0.1$, $y = 0.3$ og $y = 0.5$.



Figur 4.1: Enhedsdeviansen for normalfordeling (tv) og binomialfordeling (th)

Det ses, at mens enhedsdeviansen for normalfordelingsmodellen er symmetrisk, er enhedsdeviansen for binomialfordelingen “skæv”; jo nærmere observationen er ved 0 eller 1, desto mere udtalt er skævheden.

Bemærkning 1 *Taylorudvikling af enhedsdeviansen*

En Taylor-udvikling af $d(y; \mu)$ omkring $y = \mu$ giver approksimationen

$$d(y; \mu) \approx \frac{(y - \mu)^2}{V(\mu)}. \quad (4.15)$$

der tidligere er foreslået af Pearson som et mål for tilpasning.

▽

At enhedsdeviansen spiller en så væsentlig rolle, skyldes det faktum, at tæthedsfunktionen, (4.5), kan udtrykkes ved middelværdiparameteren, μ , igennem enhedsdeviansen som

$$g_Y(y; \mu, \lambda) = a(y, \lambda) \exp \left\{ -\frac{\lambda}{2} d(y; \mu) \right\}, \quad (4.16)$$

hvor $d(y; \mu)$ er enhedsdeviansen (4.14).

De to parametriseringer (kanonisk parameter, θ , og middelværdiparametrisering, μ) supplerer hinanden. Mens parametriseringen ved den kanoniske parameter, θ , har den fordel at parameterrummet er den reelle akse, og derfor er *velegnet til lineære operationer*, har parametriseringen ved middelværdiparameteren, μ , den fordel, at *observationer, y , direkte kan sammenlignes med den tilhørende middelværdiparameter, μ* , da de begge er udtrykt i samme måleenhed.

Det følger ved betragtning af (4.16) at enhedsdeviansen (4.14) måler 'nærheden' af y og μ udtrykt som forskel i log-likelihood (skaleret med en halv gange præcisionsparameteren λ).

Enhedsdeviansen kan tilsvarende opfattes som et udtryk for den *normerede*, eller *relative likelihood* for μ svarende til observationen y . Forholdet

$$R(\mu; y) = \frac{L(\mu; y)}{\max_{\mu} L(\mu; y)}$$

udtrykker jo netop den relative likelihood af μ (i forhold til den maksimale likelihood).

Tilsvarende er

$$d(y; \mu) = -2[\ell(\mu; y) - \max_{\mu} \ell(\mu; y)]$$

jo blot -2 gange logaritmen til denne relative likelihood, se også bemærkning 4 på side 206.

Tabel 4.1 angiver variansfunktion, $V(\mu)$, enhedsdevians, $d(y; \mu)$, og præcisionsparameter, λ , for de fordelinger, der hyppigst mødes i industrielle anvendelser. Bemærk, at vi har udvidet samlingen af sædvanlige fordelinger med den inverse Gauss-fordeling, da denne fordeling undertiden bruges til at beskrive højreskæve fordelinger, som man fx møder i forbindelse med analyse af pålidelighed og styrke, se fx Seshadri, [17], og for fuldstændighedens skyld tillige med den generaliserede hyperbolske secantfordeling, se fx Jørgensen [5].

Undertiden bruger man notationen $Y \sim \text{ED}(\mu, V(\mu)/\lambda)$ for at angive, at Y har tæthed (4.16) med deviansfunktionen $d(y; \mu)$ specificeret ved variansfunktionen

Tabel 4.1: Middelværdierum, enhedsvariansfunktion og enhedsdevians for en række sædvanlige fordelinger

Family	$\mathcal{M}$	$V(\mu)$	unit deviance $d(y, \mu)$	λ	θ
Normal	$(-\infty, \infty)$	1	$(y - \mu)^2$	$1/\sigma^2$	μ
Poisson	$(0, \infty)$	μ	$2\{y \ln(y/\mu) - (y - \mu)\}$	$-^a$	$\ln(\mu)$
Gamma	$(0, \infty)$	μ^2	$2 \times \{y/\mu - \ln(y/\mu) - 1\}$	α^b	$1/\mu$
Bin	$(0, 1)$	$\mu(1 - \mu)$	$2\{y \ln(y/\mu) + (1 - y) \ln((1 - y)/(1 - \mu))\}$	n^c	$\ln(\mu/(1 - \mu))$
Neg Bin	$(0, 1)$	$\mu(1 + \mu)$	$2\{y \ln(y(1 + \mu)/[(1 + y)\mu]) + \ln((1 + \mu)/(1 + y))\}$	r^d	$\ln(\mu)$
I Gauss	$(0, \infty)$	μ^3	$(y - \mu)^2/(y \times \mu^2)$		$1/\mu^2$
GHS ^e	$(-\infty, \infty)$	$1 + \mu^2$	$2y\{\arctan(y) - \arctan(\mu)\} + \ln((1 + \mu^2)/(1 + y^2))$	$1/\sigma^2$	$\arctan(\mu)$

^a The precision parameter λ can not be distinguished from the mean value.

^b Gamma distribution with shape parameter α and scale parameter μ/α .

^c $Y = Z/n$, where Z is the number of successes in n independent Bernoulli trials.

^d $Y = Z/r$, where Z is the number of successes until the r th failure in independent Bernoulli trials.

^e Generalized Hyperbolic Secant Distribution

$V(\cdot)$ gennem (4.14). Notationen er i analogi med notationen $N(\mu, \sigma^2)$, der angiver normalfordelinger. Udover at notationen er velegnet til at beskrive fordelingen af foldninger af eksponentielle dispersionsfamilier, tjener den også til at markere, at den sædvanlige fortolkning

$$\text{observation} = \text{' sand værdi ' + støj}$$

ikke er velegnet til at beskrive data for ikke-normalt fordelte målinger. Betragtes eksempelvis en Poisson-regression, er afvigelser fra linien ikke nødvendigvis heltal-lige, og derfor kan disse afvigelser jo ikke beskrives ved Poissonfordelte variable. Det afgørende er, at der er specificeret en model for middelværdien, μ , og at fordelingen af svarende til denne værdi af μ er karakteriseret ved fordelingen med variansfunktion $V(\mu)$.

Bemærkning 2 *Foldningsegenskaber*

Præcisionsparameteren, λ , spiller en vigtig rolle ved formulering af foldningsegenskaber for eksponentielle dispersionsfamilier: Lad $Y_1, \dots, Y_n$ være uafhængige stokastiske variable med $Y_i \sim \text{ED}(\mu, V(\mu)/\lambda_i)$, hvor $\lambda_1, \dots, \lambda_n$ er kendte konstanter. Lad $\lambda_* = \sum \lambda_i$, og betragt det præcisionsvægtede gennemsnit,

$$\bar{Y} = \frac{1}{\lambda_*} \sum_{i=1}^n \lambda_i Y_i. \quad (4.17)$$

Da gælder, at $Y \sim \text{ED}(\mu, V(\mu)/\lambda_*)$. Eksempelvis har man for $Y_1, \dots, Y_n$ uafhængige Bernoulli fordelte variable, $Y_i \sim \text{ED}(\mu, \mu(1-\mu)/1)$, at $\bar{Y} \sim \text{ED}(\mu, \mu(1-\mu)/n)$ svarende til en binomialfordeling af $Z = n \times \bar{Y}$. ∇

4.4 Model, likelihood og scorefunktion

4.4.1 Model for et observationssæt

Vi antager, at den statistiske model for observationssættet $Y_1, Y_2, \dots, Y_n$ er givet ved, at Y 'erne er indbyrdes uafhængige, tilhørende samme eksponentielle dispersionsparameterfamilie med kumulantfrembringer $\kappa(\cdot)$ og med præcisionsparameter for den i 'te observation givet som en kendt vægt, $\lambda_i = w_i$.

Antagelsen indebærer, at tætheden for Y_i er på formen (4.5), hvorfor den simultane tæthed er

$$f(\mathbf{y}; \boldsymbol{\theta}) = \exp \left[\sum_{i=1}^k w_i (\theta_i y_i - \kappa(\theta_i)) \right] \prod_{i=1}^k c(y_i, w_i) \quad (4.18)$$

Ved indførelse af middelværdiparameteren, $\mu = \tau(\theta)$ jvf. (4.11) og enhedsdeviansen (4.14) finder man ved benyttelse af (4.16), at den simultane tæthed kan udtrykkes ved middelværdiparametrene på den ækvivalente form

$$g(\mathbf{y}; \boldsymbol{\mu}) = \prod_{i=1}^k g_Y(y_i; \mu_i, w_i) = \exp \left[-\frac{1}{2} \sum_{i=1}^k w_i d(y_i; \mu_i) \right] \prod_{i=1}^k c(y_i, w_i) \quad (4.19)$$

4.4.2 Likelihoodfunktion, scorefunktion og informationsmatrix

Antag, at fordelingen af observationssættet $Y_1, Y_2, \dots, Y_n$ kan beskrives ved (4.18) eller den ækvivalente form (4.19).

Da er log-likelihoodfunktionen

$$\ell_{\theta}(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^k w_i (\theta_i y_i - \kappa(\theta_i)) \quad (4.20)$$

$$\ell_{\mu}(\boldsymbol{\mu}; \mathbf{y}) = -\frac{1}{2} \sum_{i=1}^k w_i d(y_i; \mu_i) = -\frac{1}{2} D(\mathbf{y}; \boldsymbol{\mu}) , \quad (4.21)$$

hvor

$$D(\mathbf{y}; \boldsymbol{\mu}) = \sum_{i=1}^k w_i d(y_i, \mu_i) . \quad (4.22)$$

Bemærk analogien med normalfordelingsmodellen (3.7) og størrelsen $D(\mathbf{y}; \boldsymbol{\mu})$ (3.6) i normalfordelingsmodellen.

Selv om $D(\mathbf{y}; \boldsymbol{\mu})$ givet ved (4.22) ikke er en egentlig norm, udtrykker den en form for “afstand” mellem observation og modelmiddelværdi i lighed med normalfordelingsmodellen.

Scorefunktionen mht den kanoniske parameter $\boldsymbol{\theta}$ fås direkte af (4.20) som

$$\frac{\partial}{\partial \theta_i} \ell_{\theta}(\boldsymbol{\theta}; \mathbf{y}) = w_i (y_i - \tau(\theta_i)) \quad (4.23)$$

eller på matrixform

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell_{\theta}(\boldsymbol{\theta}; \mathbf{y}) = \text{diag}(\mathbf{w})(\mathbf{y} - \boldsymbol{\tau}(\boldsymbol{\theta})) \quad (4.24)$$

hvor $\text{diag}(\mathbf{w})$ angiver en diagonalmatrix, hvis i 'te diagonalelement er w_i , og

$$\boldsymbol{\tau}(\boldsymbol{\theta}) = \begin{Bmatrix} \tau(\theta_1) \\ \tau(\theta_2) \\ \vdots \\ \tau(\theta_k) \end{Bmatrix}$$

angiver resultatet af den koordinatvise middelværdiafbildning.

Den observerede information (2.7) mht $\boldsymbol{\theta}$ er

$$\begin{aligned} \mathbf{j}(\boldsymbol{\theta}; \mathbf{y}) &= \text{diag}\{\mathbf{w}\} \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\tau}(\boldsymbol{\theta}) = \text{diag}\{w_i \kappa''(\theta_i)\} \\ &= \text{diag}\{w_i V(\tau(\theta_i))\} \end{aligned} \quad (4.25)$$

hvor $V(\tau(\theta))$ angiver variansfunktionens værdi svarende til værdien $\mu = \tau(\theta)$.

Vi bemærker, at da den observerede information ikke afhænger af observationen $\mathbf{y}$, er den forventede information lig med den observerede information (4.25).

Men i modsætning til normalfordelingssituationen er likelihoodfunktionens krumning ikke konstant, men afhænger af parameterværdien $\boldsymbol{\theta}$.

Ved benyttelse af (2.11) finder man scorefunktionen med hensyn til middelværdiparameteren

$$l'_{\mu}(\boldsymbol{\mu}; \mathbf{y}) = \text{diag} \left\{ \frac{w_i}{V(\mu_i)} \right\} (\mathbf{y} - \boldsymbol{\mu}) \quad (4.26)$$

Den observerede information er

$$\mathbf{j}(\boldsymbol{\mu}; \mathbf{y}) = \text{diag} \left\{ w_i \left[\frac{1}{V(\mu_i)} + (y_i - \mu_i) \frac{V'(\mu_i)}{V(\mu_i)^2} \right] \right\}, \quad (4.27)$$

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

$$\mathbf{i}(\boldsymbol{\mu}) = \text{diag} \left\{ \frac{w_i}{V(\mu_i)} \right\} , \quad (4.28)$$

hvilket naturligvis også kunne ses direkte ved brug af (2.12).

Vi bemærker, at den observerede information mht middelværdiparameteren $\boldsymbol{\mu}$ afhænger af observationen, $\mathbf{y}$.

4.5 Generaliserede lineære modeller

4.5.1 Generaliseret lineær model, dimension og model-matrix

Nu kan vi så formulere en *generaliseret lineær model* for observationssættet $y_1, \dots, y_k$.

Definition 4.5.1 *Generaliseret lineær model*

Antag, at fordelingen for $Y_1, Y_2, \dots, Y_k$ kan beskrives ved at $Y_1, Y_2, \dots, Y_k$ er uafhængige variable, hvis fordelinger kan beskrives ved en eksponentiel dispersionsmodel (4.16) med samme variansfunktion, $V(\mu)$.

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

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

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

$$H_0 : \boldsymbol{\eta} - \boldsymbol{\eta}_0 \in L , \quad (4.29)$$

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

Forskellen fra den generelle lineære model for normalfordelingen er

- dels, at fordelingen af observationerne nu beskrives ved en mere generel klasse af fordelinger, (eksponentielle dispersionsfamilier, hvis støtte ikke nødvendigvis er hele den reelle akse, som for normalfordelingen),
- og derfor måler man overensstemmelse mellem observation og parameter værdi på en måde, der afspejler egenskaberne ved den pågældende fordeling
- dels, at den lineære del ikke vedrører middelværdien, $\boldsymbol{\mu}$, men en funktion, $\boldsymbol{\eta}$ af middelværdien.

Når vi bare tilgodeser disse forskelle, kan vi kan genbruge en række begreber fra afsnittet om den generelle lineære model.

Nedenstående figur illustrerer den generaliserede lineære model i de to rum, den lineære prædiktør, $\boldsymbol{\eta}$, (til venstre) og middelværdirummet, $\boldsymbol{\mu}$, (til højre). Observationerne $\mathbf{y}$ identificeres naturligt som tilhørende middelværdirummet.

Definition 4.5.2 Dimension af generaliseret lineær model

I analogi med def. 3.2.2 kaldes dimensionen m af underrummet L for *modelens dimension*. ∇

Definition 4.5.3 Modelmatrix for generaliseret lineær model

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

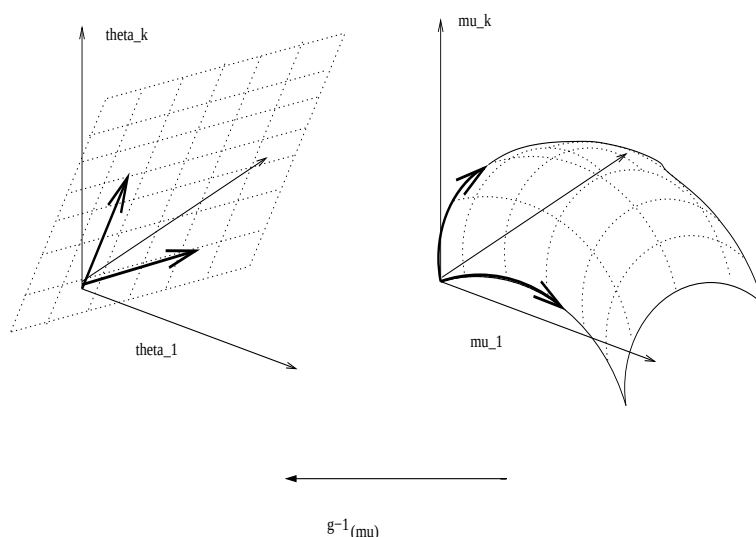
$$\boldsymbol{\eta} - \boldsymbol{\eta}_0 = \mathbf{X}\boldsymbol{\beta} \quad \text{med } \boldsymbol{\beta} \in \mathbb{R}^m, \quad (4.30)$$

hvor $\mathbf{X}$ har fuld rang, kaldes matricen $\mathbf{X}$ for hypotesens *modelmatrix* i lighed med definition 3.2.3. og vektoren

$$\mathbf{x}_i^* = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{im} \end{pmatrix}, \quad (4.31)$$

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

Figur 4.2: Lineær prædikator (tv) og middelværdirum (th) for en generaliseret lineær model



4.5.2 Komponenter i en generaliseret lineær model

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

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

En lineær komponent, den **lineære prædikator**, der beskriver den lineære afhængighed af de forklarende variable

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

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

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

Evt. en præcisionsangivelse λ_i

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 prædiktør η_i med middelværdiparameteren $\mu_i = \mathbb{E}[Y_i]$ ved relationen

$$\eta_i = g(\mu_i) \quad (4.32)$$

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

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

dvs

$$\mu_i = g^{-1}(\mathbf{x}_i' \boldsymbol{\beta}) = g^{-1} \left(\sum_j x_{ij} \beta_j \right) \quad (4.33)$$

Afsnit 4.16 angiver nogle sædvanlige linkfunktioner.

Ad præcisionsangivelsen λ_i :

Præcisionsangivelsen kan være formuleret som

- kendte *individuelle vægte*, $\lambda_i = w_i$. Man siger da, at man har en *vægtet* model,
- en (ukendt) *fælles dispersionsparameter*, σ^2 med $\lambda = 1/\sigma^2$. Man siger da, at man har en model med *overdispersion*.
- eller evt. som en *kombination af disse* $\lambda_i = w_i/\sigma^2$.

I den indledende gennemgang vil vi nøjes med at betragte en vægtet model. I afsnit 4.12 betragtes modeller med overdispersion.

4.5.3 Fuld model

Vi indfører lige nogle tekniske definitioner:

Definition 4.5.4 *Fuld model svarende til generaliseret lineær model*

Betragt en generaliseret lineær model for $Y_1, \dots, Y_k$. 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 .

▽

4.5.4 Lokal design matrix

Definition 4.5.5 *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 (4.30)

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

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

Matricen

$$\mathbf{X}(\boldsymbol{\beta}) = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} = \left[\frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}} \right]' \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}} = \text{diag} \left\{ \frac{1}{g'(\mu_i)} \right\} \mathbf{X} \quad (4.34)$$

kaldes den *lokale design matrix* svarende til parameterværdien $\boldsymbol{\beta}$.

I udtrykket (4.34) angiver $\text{diag}\{1/g'(\mu_i)\}$ en $k \times k$ dimensional diagonalmatrix, hvis i 'te diagonalelement er $1/g'(\mu_i)$, hvor μ_i bestemt ved (4.33), dvs. som middelværdien af den i 'te observation under hypotesen (4.30) svarende til parameterværdien $\boldsymbol{\beta}$, dvs

$$\mu_i = \mu_i(\boldsymbol{\beta}) = \mu(g^{-1}(\eta_i)) = \mu(g^{-1}(\mathbf{x}_i^* \boldsymbol{\beta})) , \quad (4.35)$$

hvor $\mathbf{x}_i^*$ er modelvektoren for den i 'te observation givet ved (4.31)
 For den kanoniske link er $g(\cdot) = \tau^{-1}(\cdot)$.

I tilfældet med den kanoniske link gælder derfor

$$g'(\mu) = \frac{1}{\tau'(\tau^{-1}(\mu))} = \frac{1}{V(\mu)} \quad (4.36)$$

▽

Bemærkning 1 *Fortolkning af lokal design matrix*

Den lokale design matrix $\mathbf{X}(\boldsymbol{\beta})$ udtrykker den lokale (for den aktuelle værdi af $\boldsymbol{\beta}$) reskalering af modelmatricen, $\mathbf{X}$, der er nødvendig for at tilgodese ikke-lineariteten i transformationen fra den lineære prædiktør $\boldsymbol{\eta}$ til $\boldsymbol{\mu}$. I figur 4.2 udspænder den lokale designmatrix tangentplanen til den krumme middelværdiflade i punktet $\boldsymbol{\mu}$. ▽

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 (4.36), hvorfor vi har

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

således at den lokale designmatrix bliver

$$\mathbf{X}(\boldsymbol{\beta}) = \text{diag} \{ V(\mu_i) \} \mathbf{X} \quad (4.37)$$

▽

Vi bemærker iøvrigt, at ikke-lineariteten (linkfunktionen) kun optræder *koordinatvis*, dvs vedrørende én koordinat ad gangen.

4.6 Maksimum likelihood estimation

4.6.1 Estimation

Sætning 4.6.1 *Estimation i generaliserede lineære modeller*

Betragt den generaliserede lineære model (4.29) for observationerne $Y_1, \dots, Y_k$ og antag at $Y_1, \dots, Y_k$ er indbyrdes uafhængige med fordelinger, der kan beskrives ved en eksponentiel dispersionsmodel med variansfunktion $V(\cdot)$, og eventuelt vægte w_i .

Såfremt L er parametriseret ved parametrene β svarende til modelmatricen $\mathbf{X}$ (4.30), bestemmes maksimum-likelihood estimatet for β ved at minimere totaldeviansen $D(\mathbf{y}; \mu)$ (4.22) med hensyn til β for $\mu = \mu(\beta)$. Estimatet $\hat{\beta}$ bestemmes som løsning til

$$[\mathbf{X}(\beta)]' \mathbf{i}_\mu(\mu)(\mathbf{y} - \mu) = \mathbf{0} , \quad (4.38)$$

hvor $\mathbf{X}(\beta)$ angiver den lokale designmatrix (4.34), og $\mu = \mu(\beta)$ angiver de fittede middelværdier svarende til parametersættet β ,

$$\mu_i(\beta) = g^{-1}(\mathbf{x}_i^{*'} \beta) , \quad (4.39)$$

og $\mathbf{i}_\mu(\mu)$ angiver den forventede information med hensyn til μ , (4.28).

Likelihoodligningen (4.38) må i almindelighed løses ved en iterativ procedure, se eksempel 4.6.2.

Hvis fordelingen af Y_i 'erne er en normalfordeling (med variansfunktion $V(\mu) = 1$) er løsningen den sædvanlige mindste kvadraters løsning.

Bevis:

Scorefunktionen med hensyn til middelværdiparameteren μ er jvf. (4.26)

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

Det følger da af (2.11), at scorefunktionen med hensyn til β er

$$l'_\beta(\beta; \mathbf{y}) = \left[\frac{\partial \mu}{\partial \beta} \right]' l'_\mu(\mu; \mathbf{y}) = [\mathbf{X}(\beta)]' \text{diag} \left\{ \frac{w_i}{V(\mu_i)} \right\} (\mathbf{y} - \mu(\beta)) . \quad (4.40)$$

Indsættes (4.28) i dette udtryk finder man

$$l'_\beta(\boldsymbol{\beta}; \mathbf{y}) = \frac{\partial}{\partial \beta} l_\beta(\boldsymbol{\beta}; \mathbf{y}) = [\mathbf{X}(\boldsymbol{\beta})]' \mathbf{i}_\mu(\boldsymbol{\mu}) (\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})) ,$$

der netop fører til (4.38). ∇

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 124 til $\text{diag}\{V(\mu_i)\}\mathbf{X}$, hvorfor likelihoodligningen (4.38) bliver

$$\mathbf{X}' \text{diag}\{V(\mu_i)\} \text{diag}\left\{\frac{w_i}{V(\mu_i)}\right\} (\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})) ,$$

dvs

$$\mathbf{X}' \text{diag}\{w_i\} \mathbf{y} = \mathbf{X}' \text{diag}\{w_i\} \boldsymbol{\mu} \quad (4.41)$$

Ligningen (4.41) kaldes *middelværdiligningen*, eller normalligningen i lighed med (3.15).

For en uvægtet model reduceres middelværdiligningen til den simple form

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\boldsymbol{\mu}(\boldsymbol{\beta}) \quad (4.42)$$

∇

Det følger af (4.38), at

Den fittede værdi, $\hat{\boldsymbol{\mu}}$ bestemmes altså ved projektion (mht informationsmatricen) ned på tangentplanen til modelfladen, og residualerne er ortogonale på denne tangentplan.

4.6.2 Fordeling af maksimum-likelihood estimat

Sætning 4.6.2 *Asymptotisk fordeling af maksimum-likelihood estimat*

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

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \overset{\text{as}}{\in} N_m(\mathbf{0}, \boldsymbol{\Sigma}) , \quad (4.43)$$

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 dispersionsmatricen Σ for $\widehat{\beta}$ er

$$D[\widehat{\beta}] = \Sigma = [\mathbf{X}'\mathbf{W}(\beta)\mathbf{X}]^{-1} \quad (4.44)$$

med

$$\mathbf{W}(\beta) = \text{diag} \left\{ \frac{w_i}{[g'(\mu_i)]^2 V(\mu_i)} \right\}, \quad (4.45)$$

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

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

▽

Bemærkning 1 *Dispersionsmatricen ved kanonisk link*

Såfremt specielt linkfunktionen er den kanoniske link, reduceres “vægtmatricen” $\mathbf{W}(\beta)$ (4.45) til

$$\mathbf{W}(\beta) = \text{diag} \{w_i V(\mu_i)\} \quad (4.46)$$

▽

Bemærkning 2 *Estimation af dispersionsmatricen* Man finder et estimat, $\widehat{\Sigma}$ for dispersionsmatricen ved at indsætte de estimerede værdier, $\widehat{\beta}$ i (4.44), hvorved man får

$$\widehat{D}[\widehat{\beta}] = \widehat{\Sigma} = [\mathbf{X}'\mathbf{W}(\widehat{\beta})\mathbf{X}]^{-1} \quad (4.47)$$

Diagonalelementet udskrives af SAS-Insight under Parameter Estimates. Matricen $\widehat{\Sigma}$ udskrives under Estimated Cov matrix.

▽

Bemærkning 3 *Fortolkning af “vægtmatricen” $\mathbf{W}(\beta)$*

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

Således ser vi af (4.45), at de vægte, der er tilknyttet observationerne i en vægtet model indgår i vægtmatricen.

Endvidere gælder, at jo større værdi af

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

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 $\mathbf{W}(\beta)$ jvf (4.46)

$$\mathbf{W}(\beta) = \text{diag} \{w_i V(\mu_i)\}$$

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. ∇

4.6.3 Eksempel på endimensional regression med kanonisk link

Eksempel 4.6.1 *Endimensional regression ved kanonisk link*

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

$$H_0 : \quad \theta_i = \alpha + \beta x_i, \quad i = 1, 2, \dots, k, \quad (4.48)$$

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

Hypotesen H_0 er på formen

$$\begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_k \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

dvs. netop af formen (4.30) 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

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

og tilsvarende

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

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

$$\hat{\mu}_i = \tau(\hat{\alpha} + \hat{\beta} x_i), \quad (4.49)$$

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

$$\sum_{i=1}^k y_i = \sum_{i=1}^k \tau(\alpha + \beta x_i) \quad (4.50)$$

$$\sum_{i=1}^k x_i y_i = \sum_{i=1}^k x_i \tau(\alpha + \beta x_i)$$

For en vægtet model med vægtene $w_1, \dots, w_k$ bestemmes $\hat{\alpha}$ og $\hat{\beta}$ ved

$$\sum_{i=1}^k w_i y_i = \sum_{i=1}^k w_i \tau(\alpha + \beta x_i) \quad (4.51)$$

$$\sum_{i=1}^k w_i x_i y_i = \sum_{i=1}^k w_i x_i \tau(\alpha + \beta x_i)$$

Eksempel 4.6.2 *Fosterdødelighed hos mus (fortsat)* *Estimation af parametre - Tastetryk for SAS-Insight*

Vi betragter atter data fra eksempel 4.2.1 på side 101.

```
* kursus/stat3/pricemus.sas ;
OPTIONS LINESIZE=75 PAGESIZE=58;
TITLE 'Musefosterd' '17'x 'delighed';
TITLE2 'Andel d' '17'x 'de mus mod dosis';
DATA PRICE;
INPUT CONC FOSTR DOD;
ANDEL = DOD/FOSTR;
ANDKOR = (DOD+0.5)/(FOSTR-DOD +0.5);
MYLGIT = LOG (ANDKOR);
CARDS;
  0    297  15
62.5  242  17
125   312  22
250   299  38
500   285 144
;
RUN;
```

I eksemplet opstillede vi regressionsmodellen for den kanoniske parameter:

$$H_0 : \theta_i = \alpha + \beta x_i, \quad i = 1, 2, \dots, k$$

Hypotesen er således et specialtilfælde af eksempel 4.6.1 på side 128, og vi finder at middelværdiligningen (4.51) til bestemmelse af α og β bliver

$$\begin{pmatrix} \sum_{i=1}^k n_i \exp(\alpha + \beta x_i) / \{1 + \exp(\alpha + \beta x_i)\} \\ \sum_{i=1}^k x_i n_i \exp(\alpha + \beta x_i) / \{1 + \exp(\alpha + \beta x_i)\} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^k z_i \\ \sum_{i=1}^k x_i z_i \end{pmatrix}, \quad (4.52)$$

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

$$\tilde{\beta} = \frac{\sum(\ell_i - \bar{\ell})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \quad (4.53)$$

og

$$\tilde{\alpha} = \bar{\ell} - \tilde{\beta}\bar{x} \quad (4.54)$$

Man finder $\tilde{\alpha} = -3.07497$ og $\tilde{\beta} = 0.005827$. 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 $\hat{\theta}_i$ (logit'er) og de tilsvarende fittede middelværdier $\hat{\mu}_i$ beregnes.

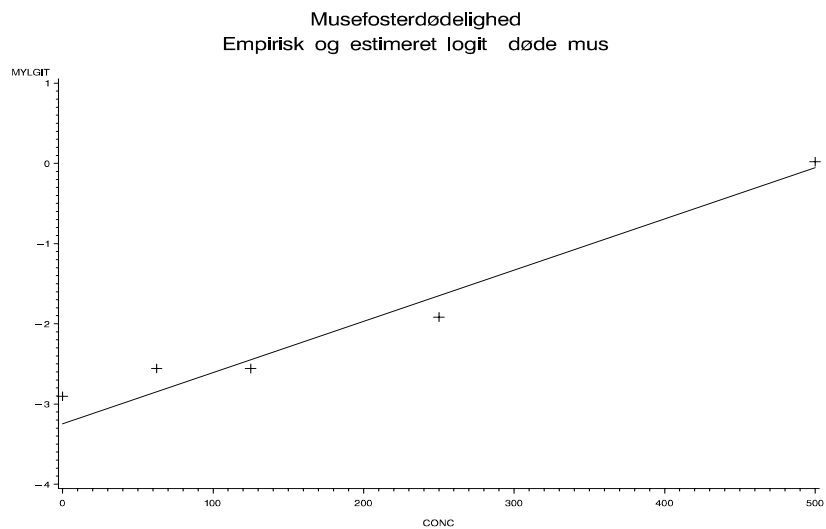
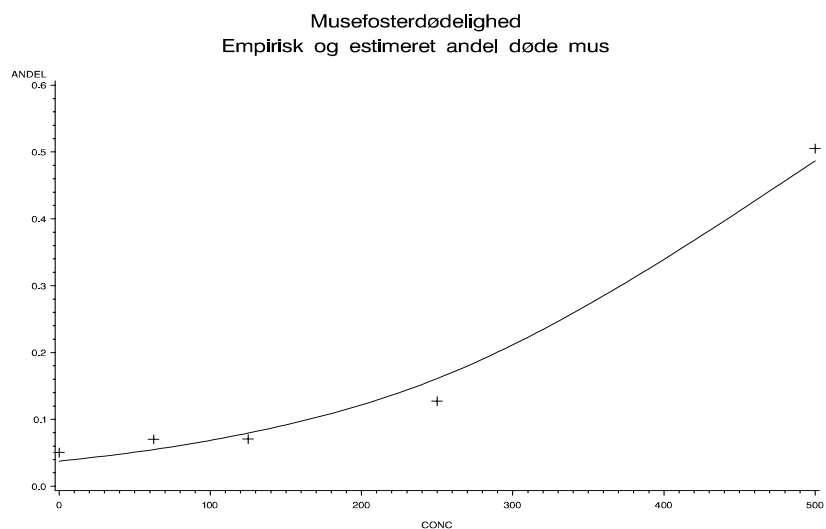
De fittede værdier af middelværdiparameteren p_i fås som

$$\hat{p}_i = \frac{\exp(-3.248 + 0.006389x_i)}{1 + \exp(-3.248 + 0.006389x_i)} \quad (4.55)$$

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
$\hat{\theta}_i$	-3.24800	-2.84869	-2.44938	-1.65075	-0.05350
$\hat{p}_i$	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 4.8 vil vi diskutere en metode til at vurdere tilpasningen.

Dispersionsmatricen for estimererne fås af (4.44), hvor vi benytter vægtmatricen $\mathbf{W}(\hat{\beta})$ bestemt ved (4.46) idet vi jo har brugt den kanoniske link, og hvor vi indsætter estimatet $\hat{\beta}$ i stedet for den ukendte sande værdi β . Man får

$$\hat{\mathbf{D}} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{pmatrix} 0.02486 & -0.00006 \\ -0.00006 & 0.000002 \end{pmatrix}$$

Ved brug af SAS INSIGHT eller PROC GENMOD finder man det likelihoodkvotientbaserede 95 % konfidensinterval for parametrene α og β som $\alpha \in [-3.5673, -2.9486]$ og $\beta \in [0.00555, 0.00726]$.

I SAS Insight vælges Fit menuen, antallet af døde fostre **dod** vælges som **Y**, koncentrationen **CONC** som **X**, og under **Method** vælges **Response dist** til **Binomial**, og antallet af fostre, **Fost**, placeres i **Binomial** boksen (for at angive antalsparameteren, n , i binomialfordelingen). Som link-funktion vælges **Canonical** eller **Logit** (det kommer ud på ét, da den kanoniske link for binomialfordelingen netop er logit-funktionen). For at få de fittede værdier af den lineære prædiktør ($\eta_i = \hat{\alpha} + \hat{\beta}x_i$) og af dødssandsynlighederne, p_i , vælges under **Output** $\rightarrow$ **Output Variables, Predicted** (for p_i) og **Linear Predictor** (for η_i). ∇

4.6.4 Fejlphobningsloven, iterativt genvægtet mindste kvadraters metode

I stedet for det simple mindste kvadraters estimat (4.53) og (4.54) kunne man være gået ét skridt videre i sin første approksimation ved at tage hensyn til at de enkelte datapunkter har forskellig varians.

$$V[\ln(Y_i/(1 - Y_i))] \approx (g'(\mu_i))^2 V[Y_i] = \frac{p_i(1 - p_i)/n_i}{[p_i(1 - p_i)]^2} = \frac{1}{n_i p_i(1 - p_i)}$$

Udtrykket følger af *fejlphobningsloven* (“Law of error propagation”), der udsiger at for $Z = g(Y)$ med $E[Y] = \mu$ gælder

$$E[Z] \approx g(\mu), \quad V[Z] \approx (g'(\mu))^2 V[Y]$$

For $g(y) = \ln(y/(1 - y))$, er

$$g'(y) = \frac{1}{y(1 - y)}$$

hvorfor

$$V[g(Y)] \approx (g'(\mu))^2 V[Y]$$

og så kan man jo udnytte (3.16) med

$$\Sigma = \text{diag} \left\{ \frac{1}{n_i p_i(1 - p_i)} \right\}$$

og godt nok kender vi ikke p_i , men vi kan indsætte estimatet, $\hat{p}_i = y_i$ og bestemme det vægtede mindste kvadraters estimat, hvor vægten

$$w_i^* = V(\hat{p}_i) \times n_i \quad (4.56)$$

angiver estimatet for *præcisionen* (den reciprokke varians).

Og når man har fået det første estimat, $(\hat{\alpha}, \hat{\beta})$, kan bruge dette estimat til beregning af fittede værdier, $\hat{p}_i$, i analogi med (4.55), og så kan vi genberegne vægten, w_i^* , jvf (4.56) svarende til dette forbedrede estimat $\hat{p}_i$. Og sådan fortsætter man.

Denne fremgangsmåde kaldes *iterativt genvægtet mindste kvadraters metode*, (IRWLS). Fremgangsmåden illustrerer netop den væsentlige egenskab ved de generaliserede lineære modeller, at de tilgodeser at variansen for de enkelte observationer afhænger af middelværdien.

Det gælder generelt for den kanoniske link, at mindste kvadraters vægten er

$$w_i^* = V(\hat{\mu}_i) \times w_i$$

hvor w_i angiver den eventuelle vægt jvf (4.19), idet nemlig

$$V[g(Y)] \approx \frac{V[Y]}{V(\mu)^2} = \frac{1}{V(\mu)w_i}$$

4.7 Fordeling af fittede værdier og residualer

4.7.1 Fittede værdier og residualer

Vi indfører først

Definition 4.7.1 *Fittede værdier for generaliseret lineær model*

Betragt den generaliserede lineære model, der blev behandlet i sætning 4.6.1.

Lad $\hat{\beta}$ angive maksimum-likelihood estimatet for parameteren β .

Ved de *fittede værdier* under hypotesen (4.30) vil vi forstå værdierne

$$\hat{\mu} = \mu(\mathbf{X}\hat{\beta}), \quad (4.57)$$

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

$$\hat{\mu}_i = g^{-1}(\hat{\eta}_i) \quad (4.58)$$

med den fittede værdi $\hat{\eta}_i$ af den lineære prædikator bestemt som

$$\hat{\eta}_i = \sum_{j=1}^m x_{ij} \hat{\beta}_j = (\mathbf{x}_i^*)' \hat{\boldsymbol{\beta}}, \quad (4.59)$$

hvor $\mathbf{x}_i^*$ angiver modelvektoren svarende til den i 'te observation givet ved (4.31).

Såfremt specielt linkfunktionen er den kanoniske link, bruges den kanoniske parameter θ som lineær prædikator, dvs (4.59) erstattes med

$$\hat{\theta}_i = \sum_{j=1}^m x_{ij} \hat{\beta}_j = (\mathbf{x}_i^*)' \hat{\boldsymbol{\beta}}, \quad (4.60)$$

▽

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 forbindelse med generaliserede lineære modeller ser man ofte forskellige former for residualer anvendt. Disse former adskiller sig ved den måde, hvorpå man beregner afvigelsen mellem observeret og fittet værdi. Det simpleste af disse er responsresidualet, der måler afvigelsen mellem observation og tilpasset (fitted) direkte i samme enheder, som observationerne y og middelværdierne μ . I afsnit 4.7.3 vil vi betragte andre former for residualer.

Definition 4.7.2 Responsresidual

Betragt den generaliserede lineære model (4.30) for observationerne $Y_1, \dots, Y_k$. Ved *responsresidualet* (eller blot residualet) svarende til modellen vil vi forstå værdierne

$$r_i^r = r_R(y_i; \hat{\mu}_i) = y_i - \hat{\mu}_i, \quad i = 1, 2, \dots, k, \quad (4.61)$$

hvor $\hat{\mu}_i$ angiver den fittede værdi (4.58) svarende til den i 'te observation.

▽

4.7.2 Fordeling af fittede værdier og residualer

Bemærkning 1 *Geometriske egenskaber for responsresidualer*

Udtrykkes maksimaliseringsligningen (4.38) ved vektoren af responsresidualer, $\mathbf{r} = (\mathbf{y} - \hat{\boldsymbol{\mu}})$, ser vi, at maksimum-likelihood estimatet skal opfylde

$$[\mathbf{X}(\hat{\boldsymbol{\beta}})]' \mathbf{i}_{\mu}(\hat{\boldsymbol{\mu}}) \mathbf{r} = \mathbf{0} ,$$

dvs ialt m ligninger af formen

$$\mathbf{u}_j' \mathbf{i}_{\mu}(\hat{\boldsymbol{\mu}}) \mathbf{r} = 0 ,$$

hvor vektorerne $\mathbf{u}_j$ er søjlerne i den lokale design matrix $\mathbf{X}(\hat{\boldsymbol{\beta}})$.

Det gælder således i analogi med normalfordelingsresultatet (bemærkning 2 på side 47), at *vektoren af responsresidualer er ortogonal på rummet udspændt af søjlerne i den lokale design matrix $\mathbf{X}(\hat{\boldsymbol{\beta}})$* , hvor ortogonalitet måles med hensyn til det indre produkt defineret ved $\mathbf{i}_{\mu}(\hat{\boldsymbol{\mu}})$ dvs

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}' \mathbf{i}_{\mu}(\hat{\boldsymbol{\mu}}) \mathbf{v} .$$

▽

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 responsresiduallet.

Lemma : Fordeling af fittede værdier og residualer

Betragt observationssættet $Y_1, \dots, Y_k$ og antag, at $Y_1, \dots, Y_k$ er indbyrdes uafhængige, hvis fordelinger tilhører en eksponentiel dispersionsmodel med samme

variansfunktion $V(\mu)$, 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 (3.12), og lad $\hat{\beta}$ angive maksimaliseringsestimaten for parameteren β , og lad endvidere $\hat{\mu}$ angive de fittede værdier under denne hypotese.

$$\hat{\mu} = \mu(\mathbf{X}\hat{\beta}) ,$$

Lad $\tilde{\mathbf{y}}$ angive de *standardiserede værdier af y*

$$\tilde{y}_i = \frac{1}{\sqrt{V(\mu_i)/w_i}} y_i \quad (4.62)$$

og $\tilde{\mu}$ de *lokalt standardiserede fittede værdier*

$$\tilde{\mu}_i = \frac{1}{\sqrt{V(\mu_i)/w_i}} \hat{\mu}_i \quad (4.63)$$

og lad endelig $\tilde{\mathbf{r}}$ angive de tilsvarende *lokalt standardiserede responsresidualer*

$$\tilde{r}_i(\hat{\mu}) = \frac{1}{\sqrt{V(\mu_i)/w_i}} (y_i - \hat{\mu}_i) \quad (4.64)$$

Der gælder da:

$$\mathbf{D}[\tilde{\mathbf{r}}] \simeq (\mathbf{I}_n - \mathbf{H}(\beta)) \quad (4.65)$$

og altså specielt

$$\mathbf{V}[\tilde{r}_i] \simeq (1 - h_{ii}) \quad (4.66)$$

Endvidere gælder at varians-kovariansmatricen for de (standardiserede) observerede værdier $\tilde{\mathbf{y}}$ og de (lokalt standardiserede) prædikterede værdier $\tilde{\mu}$ er

$$\mathbf{D} \begin{bmatrix} \tilde{\mathbf{y}} \\ \tilde{\mu} \end{bmatrix} = \begin{pmatrix} \mathbf{I}_n & \mathbf{H} \\ \mathbf{H} & \mathbf{H} \end{pmatrix} \quad (4.67)$$

hvor $\mathbf{H}(\boldsymbol{\beta})$ er matricen:

$\mathbf{H}(\boldsymbol{\beta}) = \mathbf{W}(\boldsymbol{\beta})^{1/2} \mathbf{X} [\mathbf{X}' \mathbf{W}(\boldsymbol{\beta}) \mathbf{X}]^{-1} \mathbf{X}' \mathbf{W}(\boldsymbol{\beta})^{1/2} \quad (4.68)$
<p>med</p> $\mathbf{W}(\boldsymbol{\beta}) = \text{diag} \left\{ \frac{1}{(g'(\mu_i))^2 V(\mu_i)} \right\} \quad (4.69)$

Såfremt linkfunktionen $g(\cdot)$ specielt er den kanoniske link er

$$\mathbf{W}(\boldsymbol{\beta}) = \text{diag}(V(\mu_i)) \quad (4.70)$$

▽

Bemærk at vi har brugt betegnelsen *lokalt* standardiserede. Det fremgår jo af (4.66) at disse lokalt standardiserede residualer ikke har samme varians. De *standardiserede* residualer fremkommer af de lokalt standardiserede residualer ved yderligere at dividere med $(1 - h_{ii})$

4.7.3 Andre former for residualer

Ud over responsresidualerne vil vi her betragte deviansresidualer, Pearson-residualer og Wald-residualer. Disse residualer kan yderligere *standardiseres* ved division med $(1 - h_{ii})$ (jvf. (4.66)). Disse former adskiller sig ved den måde, hvorpå man beregner afvigelsen mellem observeret og fittet værdi.

Definition 4.7.3 *Deviansresidual*

Betragt den generaliserede lineære model (4.30) for observationerne $Y_1, \dots, Y_k$. Ved *deviansresidual*et svarende til modellen vil vi forstå værdierne

$$r_i^D = r_D(y_i; \hat{\mu}_i) = \text{sign}(y_i - \hat{\mu}_i) \sqrt{w_i d(y_i, \hat{\mu}_i)} \quad (4.71)$$

hvor funktionen $\text{sign}(x)$ er fortegnsfunktionen, $\text{sign}(x) = 1$ for $x > 0$ og $\text{sign}(x) = -1$ for $x < 0$, og hvor w_i angiver den eventuelle vægt, $d(y; \mu)$ angiver enhedsdeviansen og $\hat{\mu}_i$ 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 $\hat{\mu}_i$. Da fortegnet for

observationens afvigelse fra den fittede værdi kan have interesse ved residu-
alanalyser, har vi derfor introduceret dette fortegn i definitionen af devians-
residualet.

Betragtning af deviansresidualer er i god overensstemmelse med likelihood-
betragtninger, da de jo essentielt udtrykker forskelle i log-likelihood.

Vi skal senere (i afsnit 4.8) betragte kvadratsummen af deviansresidualerne
som et mål for modellens tilpasning.

Definition 4.7.4 *Pearson-residual*

Ved *Pearson-residualet* svarende til den generaliserede lineære model (4.30)
vil vi forstå værdierne

$$r_i^P = r_P(y_i; \hat{\mu}_i) = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)/w_i}} \quad (4.72)$$

Pearson residualet fremkommer af responsresidualet ved at skalere med $\sqrt{V[Y_i]}$.
Pearson-residualet måler således responsresidualet i enheder af den estime-
rede standardafvigelse for observationen. ∇

**Bemærkning 1 *Kvadratet på Pearson residualet approximerer de-
viansen***

Vi så i (4.15) (bemærkning 1 på side 113), at Taylorudviklingen af enheds-
deviansen $d(y; \mu)$ omkring $y = \mu$ er

$$d(y; \mu) \approx \frac{(y - \mu)^2}{V(\mu)}, \quad (4.73)$$

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

$$r_D(y; \hat{\mu})^2 \approx \frac{w(y - \hat{\mu})^2}{V(\hat{\mu})} = r_P(y; \hat{\mu})^2 \quad (4.74)$$

∇

Definition 4.7.5 Wald-residual

Betragt den generaliserede lineære model (4.30) for observationerne $Y_1, \dots, Y_k$ og lad middelværdiafbildningen (4.11) svarende til modellen være $\tau(\cdot)$ med den omvendte afbildning $\theta = \tau^{-1}(\mu)$. ∇

Definition 4.7.6 Wald-residual

Ved *Wald-residualet* svarende til den generaliserede lineære model (4.30) vil vi forstå værdierne

$$r_i^W = r_W(y_i; \hat{\mu}_i) = \{\tau^{-1}(y_i) - \tau^{-1}(\hat{\mu}_i)\} \sqrt{V(y_i)} \quad (4.75)$$

∇

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 akse svarende til den kanoniske parameter.

Idet

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

(jvf. fx note om fordelinger med anvendelse i statistik, side 35), har man, at afbildningen $\tau^{-1}(\cdot)$ også kan udtrykkes som

$$\tau^{-1}(\mu) = \int_{\mu_0}^{\mu} V^{-1}(u) du \quad (4.76)$$

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 1. ∇

Eksempel 4.7.1 Fosterdødelighed hos mus (fortsat)

Vi betragter atter data fra eksempel 4.2.1 på side 101.

I eksempel 4.6.2 på side 130 bestemte vi estimerne $\hat{\alpha} = -3.248$ og $\hat{\beta} = 0.006389$ for koefficienterne i den logistiske regressionsmodel $\theta_i = \alpha + \beta x_i$ og endvidere bestemte vi de fittede værdier svarende til denne model. Deviansen svarende til binomialfordelte observationer fås af tabel 4.1. Man har enhedsdeviansen

$$d(y; p) = 2 \times \{y \ln(y/p) + (1 - y) \ln((1 - y)/(1 - p))\}$$

og den vægtede devians bliver derfor

$$wd(y; p) = 2n \times \{y \ln(y/p) + (1 - y) \ln((1 - y)/(1 - p))\}$$

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
$\hat{\theta}_i$	-3.24800	-2.84869	-2.44938	-1.65075	-0.05350
$\hat{p}_i$	0.03740	0.05475	0.07948	0.16101	0.48663
$w_i d(y_i; \hat{p}_i)$	1.28117	1.03557	0.35573	2.70902	0.39599
$r_r(y_i; \hat{p}_i)$	0.01311	0.01550	-0.00897	-0.03392	0.01864
$r_D(y_i; \hat{p}_i)$	1.13189	1.01763	-0.59643	-1.64591	0.62928
$r_P(y_i; \hat{p}_i)$	1.19043	1.05984	-0.58585	-1.59571	0.62941
$r_W(y_i; \hat{p}_i)$	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. ∇

4.8 Test for modeltilpasning

Fra normalfordelingsmodellerne er vi vant til at *antage* at den “yderste model” er tilstrækkelig, og derefter at bruge residualerne fra denne model som estimat for variansen, σ^2 (evt lige efter et rimelighedscheck), og derefter bruge dette variansestimat til at vurdere betydningen af de øvrige led i modelformlen.

I almindelighed er det ikke sådan for generaliserede lineære modeller (med mindre der er en ukendt dispersionsparameter, σ^2). Når vi har specificeret fordelingsmodellen (eksempelvis *binomial*) ved vi jo, hvad variansen er (hvis ellers vi kendte middelværdien, men den estimerer vi jo). Så man kan bruge “residualerne” til at udføre et test for hvorvidt modellen kan antages at holde.

Vi indfører lige en definition:

Definition 4.8.1 *Residualdevians*

Betragt den generaliserede lineære model defineret i def. 4.5.1 på side 119 for observationerne $Y_1, \dots, Y_k$.

Ved *residualdeviansen* (eller blot *deviansen*) $D(\mathbf{y}; \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}))$ for hypotesen (4.30) vil vi forstå summen

$$D(\mathbf{y}; \boldsymbol{\mu}(\hat{\boldsymbol{\beta}})) = \sum_{i=1}^k w_i d(y_i; \hat{\mu}_i) \quad (4.77)$$

hvor $d(y_i; \hat{\mu}_i)$ angiver enhedsdeviansen svarende til observationen y_i og den fittede værdi $\hat{\mu}_i$ (4.14), og hvor w_i angiver de eventuelle vægte.

▽

Terminologien er lidt akavet, når vi lige har indført et *deviansresidual*, men den synes at have vundet hævd i statistikprogrammer, så derfor holder vi os denne terminologi.

Bemærkning 1 *Residualdeviansen er summen af de kvadrerede deviansresidualer*

Det følger af definitionen på deviansresidualet (def. 4.7.3 på side 138), at

$$r_D(y_i; \hat{\mu}_i)^2 = w_i d(y_i; \hat{\mu}_i)$$

▽

Sætning 4.8.1 Kvotienttest for modeltilpasning i generaliserede lineære modeller

Betragt den generaliserede lineære model (def. 4.5.1) for observationerne $Y_1, \dots, Y_k$.

Kvotientteststørrelsen for hypotesen

$$H_0 : \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} \quad (4.78)$$

imod den fulde model har teststørrelsen

$$G^2(H_0) = D(\mathbf{y}; \boldsymbol{\mu}(\hat{\boldsymbol{\beta}})) \quad (4.79)$$

hvor $D(\mathbf{y}; \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}))$ angiver residualdeviansen (4.77).

Under hypotesen (4.78) 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:

Kvotientteststørrelsen er

$$G^2(H_0) = 2[\sup_{\boldsymbol{\mu} \in M} l_{\boldsymbol{\mu}}(\boldsymbol{\mu}; \mathbf{y}) - \sup_{\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}} l_{\boldsymbol{\mu}}(\boldsymbol{\mu}; \mathbf{y})] \quad (4.80)$$

hvor $l_{\boldsymbol{\mu}}(\boldsymbol{\mu}; \mathbf{y})$ angiver loglikelihoodfunktionen (2.3).

Det følger da af definitionen på enhedsdeviansen (side 112), at den (bortset fra vægtfaktoren w_i) netop måler bidraget fra den enkelte observation til kvotientteststørrelsen (4.80).

▽

Bemærkning 1 *Pearson teststørrelse for modeltilpasning*

Det følger af ovenstående sætning 4.8.1 i forbindelse med (4.15), at størrelsen

$$X^2 = \sum_{i=1}^k w_i \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \quad (4.81)$$

approksimativt følger en $\chi^2(k - m)$ fordeling under modellen H_0 . Man kan således benytte X^2 til test for modeltilpasning i lighed med $D(\mathbf{y}; \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}))$. Størrelsen X^2 bestemt ved (4.81) kaldes *Pearson teststørrelsen* for modeltilpasning (goodness of fit).

▽

Sammenfatning

Vi kan vurdere tilpasningen ved den “yderste model” ved at sammenligne residualdeviansen, $D(\mathbf{y}; \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}))$ (eller Pearson teststørrelsen, X^2) med fraktilerne i en χ^2 -fordeling. Forkast for store værdier af residualdeviansen.

4.9 Test af enkelte parametre

Test

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 pågældende variable ikke har nogen betydning).

Sætning 4.9.1 *Test af hypoteser vedrørende enkelte værdier* β_j

Hypoteser $\beta_j = \beta_j^0$ vedrørende specifikke værdier for de enkelte parametre kan testes ved at betragte teststørrelsen

$$u_j = \frac{\hat{\beta}_j - \beta_j^0}{\sqrt{\hat{\sigma}_{jj}}} \quad (4.82)$$

hvor $\hat{\sigma}_{jj}$ angiver det j 'te diagonalelement i $\hat{\Sigma}$, og sammenligne med fraktiler i en $N(0, 1)$ 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 $\beta_j = 0$

$$u_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}_{jj}}} \quad (4.83)$$

Et ækvivalent test fås ved at betragte teststørrelsen

$$z_j = u_j^2 \quad (4.84)$$

og forkaste for $z_j > \chi_{1-\alpha}^2(1)$.

I SAS-Insight udskrives z_j i skemaet **Parameter estimates** under rubrikken **Chisq**.

Bevis:

Følger af sætning 4.6.1.

▽

Konfidensintervaller

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

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

hvor

$$l_0 = l(\hat{\beta}) - 0.5\chi_{1-\alpha}^2(1)$$

Intervalleret må bestemmes ved iteration. Intervalleret udskrives af SAS-Insight, hvis man under **Output** har valgt **C.I. (L.R.) for parameters**.

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

Det sædvanlige (Wald) konfidensinterval fås af (4.83) som

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

Disse udskrives af SAS-Insight, hvis man under **Output** har valgt **C.I. /Wald for parameters**.

4.10 Test af delhypotese

Ved en vurdering af hvilke koefficienter, der er nødvendige til beskrivelse af interessevariablen, kan man gå frem i analogi med fremgangsmåden for lineære normalfordelingsmodeller, beskrevet i afsnit 3.6.1.

Vi skal i det følgende introducere det begrebsapparat, der benyttes ved en sådan successiv testning af delhypoteser. Det vigtigste resultat er angivet i sætning 4.10.1, der angiver den fundamentale opspaltning af residualdeviansen $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ i additive bidrag, der måler afvigelsen mellem delmodeller i lighed med opspaltningen af kvadratafgivelsessummer i lineære normalfordelingsmodeller.

Betragt den fulde model

$$H_F : \boldsymbol{\mu} \in \Omega_F \subset \mathbb{R}^k$$

for $Y_1, \dots, Y_k$, og betragt modellen (den generaliserede lineære model)

$$H_0 : \boldsymbol{\eta} \in L \subset \mathbb{R}^m$$

svarende til en $k \times m$ -dimensional modelmatrix $\mathbf{X}_0$ af fuld rang m :

$$H_0 : \boldsymbol{\eta} = \mathbf{X}_0 \boldsymbol{\beta}, \quad \boldsymbol{\beta} \in B \subset \mathbb{R}^m, \quad (4.85)$$

Ved anvendelse af sætning 4.8.1 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

$$H_1 : \boldsymbol{\eta} \in L_1 \subset L$$

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:

$$H_1 : \quad \boldsymbol{\eta} = \mathbf{X}_1 \boldsymbol{\alpha} , \quad \boldsymbol{\alpha} \in A \subset \mathbb{R}^r , \quad (4.86)$$

hvor modelmatricen $\mathbf{X}_1$ svarende til H_1 tilfredsstiller

$$\mathbf{X}_1 = \mathbf{X}_0 \mathbf{G}$$

Lad $\boldsymbol{\mu}(\boldsymbol{\beta}(\hat{\boldsymbol{\alpha}}))$ angive de fittede værdier svarende til maksimum-likelihood estimatet af $\boldsymbol{\alpha}$ under hypotesen H_1

Man kan da benytte sætning 4.8.1 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, $G^2(H_1) = D(\mathbf{y}; \boldsymbol{\mu}(\boldsymbol{\beta}(\hat{\boldsymbol{\alpha}})))$, hvor residualdeviansen $D(\mathbf{y}; \boldsymbol{\mu}(\boldsymbol{\beta}(\hat{\boldsymbol{\alpha}})))$ 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 $G^2(H_0) = D(\mathbf{y}; \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}))$, og $G^2(H_1) = D(\mathbf{y}; \boldsymbol{\mu}(\boldsymbol{\beta}(\hat{\boldsymbol{\alpha}})))$ er ikke stokastisk uafhængige.

Da H_1 er en delhypotese af H_0 gælder åbenbart

$$\frac{\sup_{\mu \in \Omega_1} L_y(\mu)}{\sup_{\mu \in \Omega_F} L_y(\mu)} \leq \frac{\sup_{\mu \in \Omega_0} L_y(\mu)}{\sup_{\mu \in \Omega_F} L_y(\mu)} ,$$

hvor

$$\Omega_0 = \{\mu_1, \dots, \mu_k \in \mathbb{R}^k \mid g(\mu_1), \dots, g(\mu_k) \in L\}$$

og

$$\Omega_1 = \{\mu_1, \dots, \mu_k \in \mathbb{R}^k \mid g(\mu_1), \dots, g(\mu_k) \in L_1\}$$

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 fortolkning som anført i nedenstående sætning:

Sætning 4.10.1 Kvotienttest for delhypotese af generaliseret lineær model

Betragt den generaliserede lineære model

$$H_0 : \quad \boldsymbol{\eta} \in L \subset \mathbb{R}^m$$

svarende til matrixfremstillingen $\boldsymbol{\eta} = \mathbf{X}_0 \boldsymbol{\beta}$ (4.85), og betragt delhypotesen

$$H_1 : \quad \boldsymbol{\eta} \in L_1 \subset \mathbb{R}^r$$

med matrixfremstillingen $\boldsymbol{\eta} = \mathbf{X}_1 \boldsymbol{\alpha}$ (4.86).

Kvotienttestet for H_1 under antagelse af at modellen H_0 gælder, har da teststørrelsen

$$\begin{aligned} G^2(H_1|H_0) &= G^2(H_1) - G^2(H_0) \\ &= Dt(\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}); \boldsymbol{\mu}(\boldsymbol{\beta}(\hat{\boldsymbol{\alpha}}))) \end{aligned} \quad (4.87)$$

Under hypotesen H_1 vil teststørrelsen $D(\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}); \boldsymbol{\mu}(\boldsymbol{\beta}(\hat{\boldsymbol{\alpha}})))$ asymptotisk følge en $\chi^2(m - r)$ fordeling. Endvidere vil fordelingen af $D(\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}); \boldsymbol{\mu}(\boldsymbol{\beta}(\hat{\boldsymbol{\alpha}})))$ være asymptotisk uafhængig af $G^2(H_0) = D(\mathbf{y}; \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}))$.

Bevis:

Da H_1 er en delhypotese af H_0 , kan man faktorisere kvotientteststørrelsen $q_{F,1}$ for test af H_1 mod H_F i et produkt af kvotientteststørrelsen $q_{F,0}(\mathbf{y})$ for test af H_0 mod den fulde model, og kvotientteststørrelsen $q_{0,1}(\mathbf{y})$ for test af H_1 under H_0 :

$$q_{F,1}(\mathbf{y}) = \frac{\sup_{\mu \in \Omega_1} L_y(\mu)}{\sup_{\mu \in \Omega_F} L_y(\mu)} = \frac{\sup_{\mu \in \Omega_0} L_y(\mu)}{\sup_{\mu \in \Omega_F} L_y(\mu)} \frac{\sup_{\mu \in \Omega_1} L_y(\mu)}{\sup_{\mu \in \Omega_0} L_y(\mu)} = q_{F,0}(\mathbf{y}) q_{0,1}(\mathbf{y})$$

svarende til at logaritmen til likelihoodkvotienten spalter op i en sum

$$G^2(H_1) = G^2(H_0) + G^2(H_1|H_0) , \quad (4.88)$$

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

$$D(\mathbf{y}; \boldsymbol{\mu}(\boldsymbol{\beta}(\hat{\boldsymbol{\alpha}}))) = D(\mathbf{y}; \boldsymbol{\mu}(\hat{\boldsymbol{\beta}})) + D(\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}); \boldsymbol{\mu}(\boldsymbol{\beta}(\hat{\boldsymbol{\alpha}}))) . \quad (4.89)$$

Fordelingsforholdene følger af de generelle asymptotiske resultater for kvotienttestet. ∇

Bemærkning 1 Deviansanalyse

Spaltningen (4.89) kaldes en *deviansanalyse*, fordi den spalter deviansen svarende til H_1 i to relevante komponenter. Spaltningen er analog til opspaltningen af kvadratafgivelsessummer for normalt fordelte observationer. ∇

Bemærkning 2 Fortolkning af opspaltningen

Deviansopspaltningen (4.88) svarer til en opspaltning af residualdeviansen svarende til H_1 i et bidrag, $D(\mathbf{y}; \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}))$, der beskriver afvigelsen fra H_0 , og et bidrag, $D(\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}); \boldsymbol{\mu}(\boldsymbol{\beta}(\hat{\boldsymbol{\alpha}})))$, der beskriver afvigelsen af H_1 fra H_0 .

∇

Bemærkning 3 Bestemmelse af devianserne

Residualdeviansen $D(\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}); \boldsymbol{\mu}(\boldsymbol{\beta}(\hat{\boldsymbol{\alpha}})))$ kan udtrykkes ved enhedsdevianserne $d(\cdot; \cdot)$ (4.14) som

$$D(\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}); \boldsymbol{\mu}(\boldsymbol{\beta}(\hat{\boldsymbol{\alpha}}))) = \sum_{i=1}^k w_i d(\mu_i(\hat{\boldsymbol{\beta}}); \mu_i(\boldsymbol{\beta}(\hat{\boldsymbol{\alpha}}))) , \quad (4.90)$$

hvor vi har brugt betegnelsen $\mu_i(\hat{\boldsymbol{\beta}})$ for de fittede værdier under H_0 , og $\mu_i(\boldsymbol{\beta}(\hat{\boldsymbol{\alpha}}))$ for de fittede værdier under H_1 .

Når vi har bestemt de fittede værdier $\boldsymbol{\mu}(\hat{\boldsymbol{\beta}})$ og $\boldsymbol{\mu}(\boldsymbol{\beta}(\hat{\boldsymbol{\alpha}}))$ under de to modeller H_0 og H_1 , kan deviansbidragene $w_i d(\mu_i(\hat{\boldsymbol{\beta}}); \mu_i(\boldsymbol{\beta}(\hat{\boldsymbol{\alpha}})))$ altså bestemmes ved hjælp af tabel 4.1. ∇

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 interceptleddet, er en søjle med lutter ettaller.

Et sådant deviansanalyseskema ser ud som vist i tabel 4.2

∇

Testet for modelreduktion består således i at vurdere ændringen i residualdevians, jvf (4.89)

$$D(\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}); \boldsymbol{\mu}(\boldsymbol{\beta} = D(\mathbf{y}; \boldsymbol{\mu}(\boldsymbol{\beta}(\hat{\boldsymbol{\alpha}})))) - D(\mathbf{y}; \boldsymbol{\mu}(\hat{\boldsymbol{\beta}})) + \quad (4.91)$$

Hvis den reducerede model er tilfredsstillende, vil ændringen i residualdevians approksimativt følge en χ^2 -fordeling med et antal frihedsgrader, der svarer til ændringen i modellens dimension.

Man forkaster således modelreduktionen, for store ændringer i residualdeviansen

Bemærk analogien mellem (4.91) og (3.30)

Tabel 4.2: **Eksempel på deviansanalyseskema omkring intercepted**

Variationskilde	f	Devians	middeldevians	Goodness of fit fortolkning
Model H_0	$m - 1$	$D(\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}); \hat{\boldsymbol{\mu}}_M)$	$\frac{D(\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}); \hat{\boldsymbol{\mu}}_M)}{m - 1}$	$G^2(H_M H_0)$
Residual (Error)	$k - m$	$D(\mathbf{y}; \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}))$	$\frac{D(\mathbf{y}; \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}))}{k - m}$	$G^2(H_0)$
Korrigeret Total	$k - 1$	$D(\mathbf{y}; \hat{\boldsymbol{\mu}}_M)$		$G^2(H_M)$

Note: $\hat{\boldsymbol{\mu}}_M$ angiver estimatet for den fælles middelværdi.

Bemærkning 5 *Wald-test*

SAS-Insight giver også mulighed for en opspaltning ved de såkaldte Wald-tests. Ideen er her, at man betragter teststørrelsen

$$\mathbf{W} = (\hat{\boldsymbol{\theta}} - \hat{\hat{\boldsymbol{\theta}}})' \mathbf{i}(\hat{\boldsymbol{\theta}}) (\hat{\boldsymbol{\theta}} - \hat{\hat{\boldsymbol{\theta}}}) \quad (4.92)$$

hvor $\hat{\boldsymbol{\theta}}$ angiver estimatet under H_0 , $\hat{\hat{\boldsymbol{\theta}}}$ angiver estimatet under H_1 , og $\mathbf{i}(\hat{\boldsymbol{\theta}})$ angiver den observerede information under H_0 . Såfremt H_1 er sand, vil $\mathbf{W}$ approksimativt følge en $\chi^2(m - r)$ -fordeling.

I modsætning til likelihood-ratio testet, er Wald-testet ikke invariant under alternative parametriseringer af modellen.

▽

4.11 Successiv testning

Og hvis man mener at kunne acceptere modellen H_0 , kan man fortsætte med at spalte deviansen $D(\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}); \hat{\boldsymbol{\mu}}_M)$ i lighed med behandlingen af normalfordelingsmodellerne. Type I og Type III opspaltninger har samme fortolkning

som i normalfordelingsmodellerne. Kun bliver testene foretaget som test i χ^2 -fordelinger, da vi jo allerede kender variansforholdene under fordelingsmodellen.

4.11.1 Eksempel, tofaktor model med binomialt respons

Det følgende eksempel illustrerer formuleringen af en generaliseret lineær model svarende til en tofaktor model.

Eksempel 4.11.1 *Udplantning af blommestiklinger*

Data

Vi betragter data fra det forsøg, der blev omtalt i afsnit 3.7.5. Forsøget havde til formål at vurdere forskellige faktoreres 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.

Model

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 $B(n_{ij}, p_{ij})$ -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 præcisionsparameter (vægt) $w_{ij} = n_{ij}$.

Vi vælger at betragte den kanoniske link, og formulerer hypotesen svarende til forsvindende vekselvirkning i analogi med modellen for et tofaktorforsøg for normalfordelte målinger.

$$H_0 : \quad \eta_{ij} = \mu + \alpha_i + \gamma_j \quad (4.93)$$

Når vi vil formulere modellen ved en modelmatrix, skal vi sikre, at matrixen 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 modelmatrixen 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

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

Estimation

Middelværdiligningen bliver

$$\begin{aligned} \sum_{i=1}^2 y_{ij} &= \sum_{i=1}^2 \hat{p}_{ij}, \quad j = 1, 2, 3 \\ \sum_{j=1}^3 y_{ij} &= \sum_{j=1}^3 \hat{p}_{ij}, \quad i = 1, 2 \end{aligned}$$

med

$$\hat{p}_{ij} = \frac{\exp(\hat{\mu} + \hat{\alpha}_i + \hat{\gamma}_j)}{1 + \exp(\hat{\mu} + \hat{\alpha}_i + \hat{\gamma}_j)}$$

Ligningerne må løses ved iteration.

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 $\sqrt{\widehat{V}[\widehat{\beta}_j]} = \widehat{\sigma}_{jj}$ på den estimerede parameter værdi $\widehat{\beta}_j$, dvs kvadratroden af det tilsvarende diagonal-element i $\widehat{D}[\widehat{\beta}]$.

Test for modeltilpasning

Vi har mulighed for at foretage et test for forsvindende vekselvirkning i denne model, idet goodness of fit-teststørrelsen jo netop måler modeltilpasningen til denne model uden vekselvirkning.

Deviansanalyseskemaet er

Analysis of Deviance

Source	DF	Deviance	Deviance / DF	Scaled Dev	Pr >
Model	3	27.0298	9.0099	27.0298	.0001
Error	2	1.8854	0.9427	1.8854	.
C Total	5	28.9152	.	.	.

Residualdeviansen på 1.8854 svarer til 61% fraktilen i en χ^2_2 -fordeling. Der er således ikke noget, der taler imod at antage en binomialfordelingsmodel uden vekselvirkning for logit'erne.

Test af hypotese vedrørende enkelte koefficienter

Vi vil nu undersøge hypotesen om effekten af tykkelse “middel” adskiller sig fra referenceværdien, tykkelse “tynd”.

Parameteren γ_1 svarende til tykkelse “middel” blev estimeret til $\hat{\gamma}_1 = 1.6059$ 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 $P[N(0, 1) \geq 3.16] = 0.0079$ ses det, at hypotesen må afvises ved (tosidet) test på ethvert niveau større end 1.6 %.

Bestemmelse af konfidensinterval for enkelte koefficienter

Vurdering af 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 kan tilsvarende foretages ved brug af et konfidensinterval.

Parametrene til det sædvanlige Wald-konfidensinterval fås af udskriften. Parameteren γ_1 svarende til tykkelse “middel” blev estimeret til $\hat{\gamma}_1 = 1.6059$ med den estimerede spredning $\sqrt{\sigma_{33}} = 0.5082$. Wald-konfidensintervallet svarende til 95 % konfidenssandsynlighed fås da som

$$\hat{\gamma}_1 \pm 1.96 \sqrt{\sigma_{33}},$$

idet $u_{0.975} = 1.96$. Man får således intervallet

$$1.6059 \pm 1.96 \times 0.5082 = \begin{cases} 0.6098 \\ 2.6020 \end{cases}$$

Intervallerne omfatter ikke værdien nul.

Det likelihoodkvotientbaserede konfidensinterval må bestemmes ved iteration.

Logaritmen til likelihoodfunktionen i maksimumspunktet er

$$l(\hat{\beta}; \mathbf{y}) = -69.5127$$

og idet $\chi^2_{0.95}(1) = 3.841$, skal man altså bestemme de to værdier af γ_1 (henholdsvis mindre og større end $\hat{\gamma}_1 = 1.6059$) sådan at logaritmen til profillikelihooden $\tilde{l}(\gamma_1; \mathbf{y})$ netop er

$$\tilde{l}(\gamma_1; \mathbf{y}) = -69.5127 - 0.5 \times 3.841 = -71.4334$$

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			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	.	.

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 $\hat{\gamma}_1 = 1.6059$, men intervallet afviger dog ikke særlig meget fra det simplere Wald-konfidensinterval.

Successiv testning

95% C.I. (LR) for Parameters

Variable			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	.	.

Nedenstående inklusionsdiagram figur 4.3 illustrerer de to kæder, der kan konstrueres for et forsøg med to faktorer .

Det gælder imidlertid, 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 3.4 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.

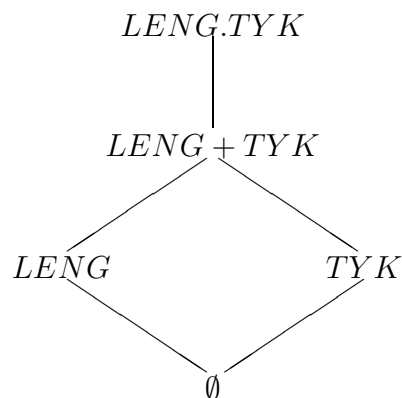
Type-I analyse i SAS og S-plus

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 4.3, hvor leddene i tabellen svarer til øgningen i modeldevians når man bevæger sig op igennem den venstre gren af inklusionsdiagrammet

Figur 4.3: Inklusionsdiagram svarende til tofaktormodeller for blommemforsøg

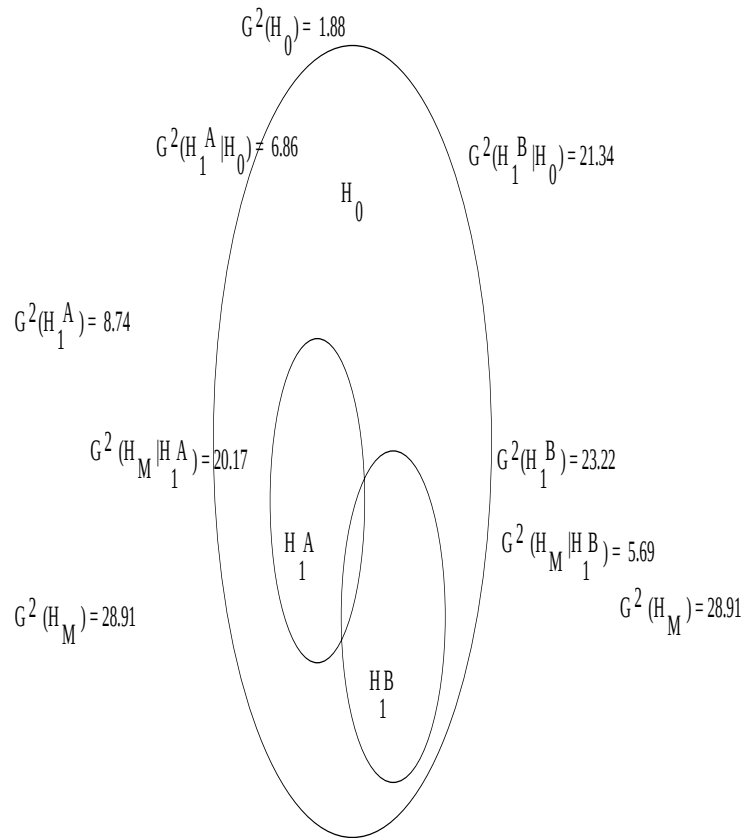


Tabel 4.3: Eksempel på udskrift af Type I-test, SAS procedure INSIGHT
Leddene tilføjes i rækkefølgen LENG og TYK.

Source	Type I (LR) Tests		
	DF	Chi-Sq	Pr > Chi-Sq
LENG	1.0000	5.6931	0.0170
TYK	2.0000	21.3367	0.0001

fig 3.4. Rækkefølgen svarer til den rækkefølge, hvori leddene er angivet i modelformlen. Ledet **TYK** i tabellen svarer derfor her til effekten af tykkelse, når der er korrigeret for den eventuelle effekt af længden.

Testene ved at gå forskellige veje i inklusionsdiagrammet er illustreret i nedenstående figur



▽

4.12 Modeller med overdispersion

Det kan forekomme, at man ikke opnår modeltilpasning ved den “yderste” model, og undertiden vil man da vælge en model med *overdispersion*, dvs en model af formen (4.16), hvor $\lambda_i = w_i/\sigma^2$ jvf. 122, hvor parameteren σ^2 angiver overdispersionen, se fx Lee og Nelder [7].

Tilstedeværelsen af en sådan (konstant) overdispersion ændrer ikke estimationen af parameteren β knyttet til middelværdien, idet overdispersionen blot indgår som en fælles faktor i (4.38).

Hvis man kender parameteren, σ^2 , på forhånd, kan man jo blot lade den indgå i w_i som en kendt vægt i modellen.

Sædvanligvis, når man vælger en model med overdispersion, vil parameteren σ^2 ikke være kendt på forhånd, men skal estimeres ud fra data, og så begynder problemerne at melde sig, for så indeholder likelihood-funktionen (4.19) også faktoren $\prod_{i=1}^k a(y_i, w_i/\sigma^2)$.

Hvis man har et eksplicit udtryk for denne normeringsfaktor, kan man principielt bestemme maksimum-likelihood estimatet for σ^2 . I praksis vælger man imidlertid ofte at bruge resultatet i sætning 4.8.1 som udgangspunkt. I tilfældet med overdispersion gælder nemlig at $D(\mathbf{y}; \mu(\hat{\beta}))$ approximativt følger en $\sigma^2\chi^2(k-m)$ -fordeling, og man kan da vælge estimatet

$$\hat{\sigma}_{dev}^2 = D(\mathbf{y}; \mu(\hat{\beta})) / (k - m) = \frac{1}{k - m} \sum_{i=1}^k w_i d(y_i; \hat{\mu}_i) \quad (4.94)$$

svarende til at man sætter residualdeviansen, $D(\mathbf{y}; \mu(\hat{\beta}))$ lig med sin middelværdi (Scale= Deviance).

Alternativt kunne man benytte Pearson-goodness of fit teststørrelsen (4.81), og vælge

$$\hat{\sigma}_{dev}^2 = \frac{1}{k - m} \sum_{i=1}^k w_i \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \quad (4.95)$$

(Scale= Pearson) jvf bemærkning 1 side 144.

Tabel 4.4: **Eksempel på deviansanalyseskema for ukendt dispersionsparameter**

Variationskilde	f	Devians	middeldevians	Test
Mellem H_1 og H_0	$m - r$	$D(\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}); \boldsymbol{\mu}(\hat{\boldsymbol{\gamma}}))$	$\frac{D(\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}); \boldsymbol{\mu}(\hat{\boldsymbol{\gamma}}))}{m - r}$	$\frac{D(\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}); \boldsymbol{\mu}(\hat{\boldsymbol{\gamma}}))/(m - r)}{D(\mathbf{y}; \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}))/(k - m)}$
Afvigelse fra H_0	$k - m$	$D(\mathbf{y}; \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}))$	$\frac{D(\mathbf{y}; \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}))}{k - m}$	
Total	$k - r$	$D(\mathbf{y}; \boldsymbol{\mu}(\hat{\boldsymbol{\gamma}}))$	$\frac{D(\mathbf{y}; \boldsymbol{\mu}(\hat{\boldsymbol{\gamma}}))}{k - r}$	

Og så har man et estimat for σ^2 , som man bruge til bestemmelse af en *skaleret devians* ved modelreduktion. Undertiden sammenligner man den således skalerede devians med fraktiler i en χ^2 -fordeling - selv om den jo har lidt mere variation pga divisionen med estimatet for σ^2 . Principielt burde de skalerede devianser (divideret med tællerens frihedsgrader) nok snarere sammenlignes med en F-fordeling, da de jo er en kvotient af to uafhængige, approximativt $\chi^2(f)/f$ -fordelte størrelser, og dette er også fremgangsmåden i PROC GENMOD. Se afsnit 4.13 for valg af SCALE option.

Bemærk analogien til normalfordelingssituationen, hvor vi også bruger residualvariationen svarende til den yderste model til estimation af dispersionsparameteren, σ^2 , dvs vi kan ikke teste goodness-of fit til den yderste model, men må tage denne model for givet.

4.12.1 Quasi-devians og quasi log-likelihood

Definition 4.12.1 *Quasi-devians og quasi log-likelihood*

Resultatet i sætning 4.6.1 at estimatet for $\boldsymbol{\beta}$ bestemmes ved at minimere totaldeviansen $D(\mathbf{y}; \boldsymbol{\mu})$ 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 (4.14) definere en *quasi-devians* ved

$$d(y; \mu) = 2 \int_{\mu}^y \frac{y - u}{V(u)} du \quad . \quad (4.96)$$

og derefter estimere μ ved at minimere deviansen som i sætning 4.6.1.

En sådan størrelse kaldes en quasi-devians (*quasi=lige som*), og den tilsvarende størrelse $D(\mathbf{y}; \mu)$ 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 . ∇

4.13 Værktøjer i SAS og S-plus

4.13.1 Generaliserede lineære modeller i SAS Insight

I SAS Insight analyseres generaliserede lineære modeller i lighed med de generelle lineære modeller ved brug af Fit menuen.

Knappen Method frembringer en menu, hvor man kan vælge

- **Response Dist.** Valget af af responsfordeling indebærer automatisk valg af *Variansfunktion* $V(\mu)$ og dermed af *enhedsdevians* $d(y; \mu)$ jvf formel (4.14) og oversigten i tabel. 4.1
- **Link Function:** Man kan her vælge linkfunktionen, dvs den funktion, der angiver, hvilken funktion af middelværdien, der skal være den *lineære prædiktør*.
Forhåndsvalget er *den kanoniske link* jvf. tabel 4.1
Ved linkfunktionen **Power** kan man vælge potensen i linien til højre.
- **Scale:** Her kan man så engang imellem vælge dispersions- eller præcisionsparameter (fast w eller σ^2) se nedenfor

Endvidere er der rubrikkerne

- **Binomial** som bruges, hvis man har valgt binomialfordeling som responsfordeling
Ved valg af binomialfordelingen kan responsvariablen Y enten have værdierne nul eller én svarende til en Bernoullivariabel. I dette tilfælde anføres ikke noget i rubrikken **Binomial**
Eller responsvariablen Y angiver det binomialfordelte *observerede antal*. I dette tilfælde placeres *antalsparameteren* n_i i rubrikken **Binomial**, og maskineriet regner så på andelene y_i/n_i og bruger (implicit) vægtene $w_i = n_i$. Fittede værdier er p 'er.
- **Offset** Her placeres den variable, der angiver et eventuelt *offset*, η_0 (se eksemplet i afsnit 4.14.5)

Scale parameteren

Her begynder det at blive lidt rodet. Scale parameteren er *relateret* til vores dispersionsparameter σ^2 , men på en lidt forskellig måde.

For normalfordelingen og den inverse gauss-fordeling er scale bare σ . For en gamma fordeling er scale lig med *formparameteren* α . For binomial og Poisson fordeling er scale bare 1 (med mindre man vil have overdispersion, så skal markere rubrikken **Quasi-likelihood** jvf nedenfor, afsnit 4.13.1).

Og for de fordelinger, der har tilknyttet en scale parameter, kan den så fastlægges på forhånd som en kendt værdi ved at angive denne værdi i feltet **Constant**

Eller man kan vælge at *estimere* denne scale ved brug af

- **MLE** (maximum-likelihood estimation), der tilgodeser normeringskonstanten $a(y_i, 1/\sigma^2)$ i (??)
- **Deviance** Middeldeviansen, $D(\mathbf{y}; \hat{\boldsymbol{\mu}})/(k - m)$, nemlig **Deviance** svarende til **Error** (goodness of fit) divideret med de tilhørende frihedsgrader
- **Pearson** Nemlig Pearson-teststørrelsen for goodness of fit divideret med de tilhørende frihedsgrader

Værdierne i rubrikken **Scaled deviance** i outputtet fremkommer ved at dividere med denne estimerede værdi, $\hat{\sigma}^2$.

Quasi-likelihood funktioner

For Poissonfordelings og binomialfordelingsmodeller med *overdispersion* kan man vælge at tilgodese denne ved at markere rubrikken **Quasi-Likelihood** og så bestemme hvilket overdispersionsestimat, man vil have, ved at vælge i rubrikken **Scale**.

Knappen **Exact Distribution** bruges kun i normalfordelingsmodeller med kanonisk link for at angive, at man vil have de eksakte fordelinger af teststørrelser, altså t-fordelinger mv.

4.13.2 Andre SAS-procedurer

Der er en række SAS-procedurer, der håndterer specifikke varianter af generaliserede lineære modeller.

PROC GENMOD procedureversionen af Fit- menuen til analyse af generaliserede lineære modeller. Indeholder yderligere modellering ved negativ binomialfordeling og multinomialfordeling, samt mulighed for selv at definere link-funktion. Valg af **SCALE** modsvarer i nogen grad valgene i SAS/INSIGHT. Der er dog en subtil sammenhæng mellem valgene **NOSCALE** og **SCALE =** i model sætningen. The **NOSCALE** option holds the scale parameter fixed. Otherwise, for the normal, inverse gaussian, and gamma distributions, the scale parameter is estimated by maximum likelihood. If you omit the **SCALE=** option, the scale parameter is fixed at the value 1. Using **NOSCALE** without specifying the value of the scale parameter in the **SCALE =**, the scale parameter is fixed at the value 1. (Specifying **SCALE = number** without specifying **NOSCALE** does not imply that the scale is fixed at that value, but only that *number* is used as an initial estimate of the scale parameter.) Det anbefales at bruge hjælp funktionen for at finde ud af, hvad den gør i det aktuelle valg af fordeling.

PROC CATMOD modeller for kategoriske data, log-lineære modeller, logistisk regression, repeated measurements analyser; diverse linkfunk-

tioner for kategorisk respons med flere end to responskategorier

PROC LOGISTIC logistisk regression (forskellige linkfunktioner) for binært, multinomialt (ordinalt og blot nominal) respons. For modeller med faktorvariable tillader den parametrisering ved forskellige kodninger (effekt, reference, polynomial og ortogonale polynomier). Giver mulighed for bestemmelse af *Receiver Operating Characteristic Curve* (ROC-kurve). Giver mulighed for eksakte tests

PROC PROBIT maximum-likelihood estimation i ikke-lineære regressionsmodeller for binært og multinomialt respons. Modellerne omfatter probit, logit, ordinal logistisk og ekstremværdi (gompit) responsmodeller.

PROC GLMMOD frembringer design matricen svarende til en generel lineær model. Er front-end for GLM proceduren. Kan bruges i forbindelse med andre regressionsprocedurer, eller med SAS/IML til at håndtere generelle lineære modeller, som ikke kan klares af PROC GLM. (PROC TRANSREG kan også lave design matricer for modeller med fuld rang, mindre end fuld rang, polynomier og splines)

4.13.3 Generaliserede lineære modeller i S-plus

S-plus funktionen `glm` producerer et `glm`-objekt, der indeholder et fit ved en generaliseret lineær model.

Modellen beskrives ved et `formula`-objekt og fordelingen angives ved et `family`-objekt, der specificerer variansfunktion, devians, link-muligheder mv.

Det resulterende `glm`-objekt indeholder estimer, fittede værdier, residualer, type-1 deviansopspaltning mv.

I den menu-orienterede version vælger man **Statistics** → **Regression** → **Generalized linear** og så har man de sædvanlige valgmuligheder.

4.14 Eksempler på “regressionsmodeller”

4.14.1 Logistisk link ved binomial respons, odds

Vi har allerede set eksempler på brug af den logistiske link ved binomialfordelte observationer, (variansfunktion $V(\mu) = \mu(1 - \mu)$). I det følgende vil

vi ofte bruge symbolet p for middelværdien μ , da det jo drejer sig om en *sandsynlighed*.

Den kanoniske link funktion er

$$\eta = \ln \left(\frac{p}{1-p} \right) , \quad (4.97)$$

altså netop logit-transformationen.

Den omvendte funktion er den logistiske funktion

$$p = \frac{\exp(\eta)}{1 + \exp(\eta)} \quad (4.98)$$

Middelværdiligningen (4.41) til bestemmelse af α og β bliver her

$$\begin{pmatrix} \sum_{i=1}^k n_i \exp(\alpha + \beta x_i) / \{1 + \exp(\alpha + \beta x_i)\} \\ \sum_{i=1}^k x_i n_i \exp(\alpha + \beta x_i) / \{1 + \exp(\alpha + \beta x_i)\} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^k z_i \\ \sum_{i=1}^k x_i z_i \end{pmatrix} , \quad (4.99)$$

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.

Den sædvanlige *logistiske regressionsmodel* er på formen

$$H_0 : \eta_i = \alpha + \beta x_i , i = 1, 2 \dots, k , \quad (4.100)$$

$$p(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

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,

$$\theta(x) = \frac{p(x)}{1-p(x)}$$

angiver netop forholdet mellem sandsynligheden for at den undersøgte hændelse indtræffer (fx at et foster dør) og sandsynligheden for at den ikke indtræffer.

Idet den lineære prædiktor η netop er logaritmen til odds

$$\eta(x) = \ln \theta(x)$$

ser vi, at den logistiske regressionsmodel kan udtrykkes som

$$\theta(x) = \exp(\alpha + \beta x) = e^\alpha (e^\beta)^x$$

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

$$\hat{x}_{0.50} = -\frac{\hat{\alpha}}{\hat{\beta}}$$

SAS-systemet indeholder en procedure, PROC LOGISTIC, der er specielt indrettet mod logistisk regression.

4.14.2 Andre linkfunktioner for binært respons

Eksempel 4.14.1 “Dosis-respons” funktion for isolationsanordning

Datasættet **Gnist** indeholder samhörørende værdier af spænding (i kV), antal spændingspåvirkninger og antal overslag fra et forsøg med 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.

Man kan nu prøve at modellere overslagssandsynlighederne ved forskellige sigmoide funktioner:

logit

$$\begin{aligned} g(p) &= \ln \left(\frac{p}{1-p} \right) = \alpha + \beta x \\ p(x) &= \frac{\exp(\eta)}{1 + \exp(\eta)} = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \end{aligned} \quad (4.101)$$

probit

$$\begin{aligned}g(p) &= \Phi^{-1}(p) = \alpha + \beta x \\p(x) &= 1 - \exp \Phi(\eta) = \Phi(\alpha + \beta x)\end{aligned}\tag{4.102}$$

svarende til den kumulerede fordelingsfunktion for normalfordelingen.

komplementær log-log

$$\begin{aligned}g(p) &= \ln(-\ln(1-p)) = \alpha + \beta x \\p(x) &= 1 - \exp[-\exp(\alpha + \beta x)]\end{aligned}\tag{4.103}$$

Pas på: i en række programmer kaldes denne linkfunktion blot log-log link funktionen. Responsfunktionen kaldes undertiden *Gombit*-funktion.

Responsfunktionen (4.103) 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,

$$\Lambda_1(x) = \exp(-e^{-x}) .$$

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 $E[T] = -(\alpha + \gamma)/\beta$, hvor γ er Eulers konstant, $\gamma \simeq 0.5772$, og $V[T] = \pi^2/(6\beta^2)$.

Responsfunktionen er *asymmetrisk*; den vokser langsomt væk fra nul, hvori-mod den nærmer sig 1 ad et forholdsvis stejlt forløb.

For to værdier, x_1 og x_2 , af den forklarende variable følger det af modellen (4.103), 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 $\exp(\beta)$

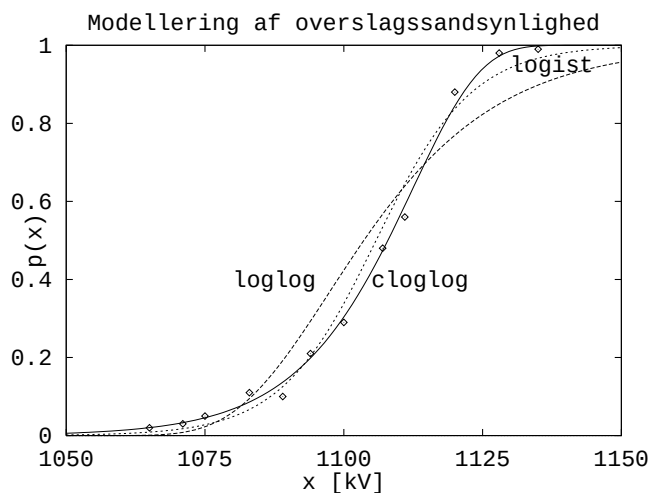
log-log link Såfremt man i stedet har et forløb, hvor responset vokser stejlt væk fra nul, kan man benytte responsfunktionen

$$\begin{aligned} g(p) &= \ln(-\ln(1-p)) = \alpha + \beta x \\ p(x) &= 1 - \exp[-\exp(\alpha + \beta x)] \end{aligned} \quad (4.104)$$

Denne model kaldes undertiden en *Gumbel*-regression, idet ekstremværdifordelingen Max_1 -fordelingen undertiden betegnes en Gumbel-fordeling.

Det er klart, at såfremt $p(x)$ kan beskrives ved modellen (4.103), kan $1-p(x)$ beskrives ved (4.104) og omvendt. En komplementær log-log link for p er således blot en direkte log-log link for $1-p$ funktion.

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 ≤ 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.

▽

Eksempel 4.14.2 *Bestemmelse af bakterietæthed fra fortyndings-eksperiment*

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,

$$p_i = P[P(\lambda x_i) = 0] = \exp(-\lambda x_i) . \quad (4.105)$$

Lad nu Z_i betegne antallet af glas fra den i 'te prøve, der ikke viste vækst. Der gælder da, at $Z_i \in B(n_i, p_i)$, hvor p_i er givet ved (4.105).

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

$$\ln p_i = -\lambda x_i , \quad (4.106)$$

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 $B(n_i, p_i)/n_i$. Udtrykket (4.106) viser, at vi søger en lineær prædiktør 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 fejlintensiteterne se f.eks. [19]

Eksempel 4.14.3 *Fortyndingseksperiment igen*

Datasættet `Yield` indeholder resultater fra afprøvningen af testchips i en produktion af mikroelektroniske kredsløb.

Disse chips var konstrueret med en specifik leder, i samme længde for alle chips, men i forskellig bredde, således at arealet af lederen var forskelligt.

I 10 chips på hver wafer var arealet 1 [arealenhed], i 10 andre var det 4 [arealenheder] og i de 10 sidste var det 16 [arealenheder].

Ved afprøvningen kunne man ikke registrere antallet af fejl på den enkelte chip, man registrerede kun om chip'en var tilfredsstillende, eller om den havde fejl.

Datasættet indeholder arealet (`Size`), antallet af defekte chips, (`Def`). Man ønsker nu at estimere fejlintensiteten, dvs man bruger en binomialfordelingsmodel for antallet af ikke fejlende kredsløb (10 minus `def`) og linkfunktionen $\ln(\cdot)$. jvf eksempel 4.14.2

```
* kursus/stat3/yieldind.sas ;
TITLE 'Yield data fra wafer produktion';
DATA YIELD;
INPUT SIZE WAF DEF ;
ANT = 10;
OK = ANT-DEF;
Cards;
1 1 0
1 2 2
1 3 0
1 4 0
1 5 0
4 1 0
4 2 0
4 3 0
4 4 0
4 5 0
4 6 0
```

```

16 1 2
16 2 1
16 3 3
16 4 1
16 5 0
16 6 2
;
RUN;

```

▽

4.14.3 Indskud om sensitivitet, specificitet af klassifikation, ROC-kurve

Et binært respons optræder ofte i forbindelse med *klassifikation* og udvikling af diagnostiske tests. Ofte bruges en logistisk regression til udvikling af sådanne tests. SAS proceduren PROC LOGISTIC indeholder diverse værktøjer til vurdering af sådanne test. Vi vil derfor kort introducere nogle begreber, der ofte benyttes i sådanne sammenhænge.

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 (tærskelværdier), eller under forskellige omstændigheder, se fx Dorfman og Alf [20] og [21]. En nylig reference er Zhang et al [22].

SAS-proceduren PROC LOGISTICS giver mulighed for konstruktion af datasæt til brug for beregning af ROC-kurver. I MODEL sætningen vælger man option OUTROC = dataset (se også optionen ROCEPS=

4.14.4 Ordnet kategorisk respons

Undertiden har man et *kategorisk respons* med flere svarkategorier, og så kan man selvfølgelig godt lave en logistisk regression for hver af disse kategorier, men antallet af responser i de forskellige kategorier er ikke uafhængige (man skal jo svare i en af kategorierne). Den naturlige responsfordeling er en *Multinomial fordeling* (Polynomialfordeling), dvs man har et sæt af sandsynligheder, $p_1(x), \dots, p_r(x)$ svarende til de r responskategorier.

Bemærkning 1 *Likert skalaen* Ved tilfredshedsundersøgelser benyttes ofte en variant af en såkaldt *Likert skala* benævnt efter den amerikanske statistiker, Rensis Likert (1903-1981), der introducerede denne metode i sin Ph.D. afhandling ved Columbia University's Department of Psychology, *A Technique for the Measurement of Attitudes* fra 1932. Likerts oprindelige ide var at lade respondenterne angive sin grad af enighed/uenighed med en forelagt holdning på en skala med et neutralt midtpunkt. Likert viste i sin afhandling, at denne metode var lige så effektiv, som den mere komplicerede Thurstone metode, der bygger på en psykofysisk metode med intervaller, der fremtræder lige store. ▽

Skalaen gav anledning til et *ordnet kategorisk respons*. Der er en række forskellige måder at vælge en logit-transformation af et sådant ordnet respons, se Agresti [23]. Vi vil her betragte en parametrisering ved de såkaldte *kumulative logit'er*, der fremkommer ved en succesiv aggregering af kategorierne. Vi indfører de kumulerede sandsynligheder, Π_j ved

$$\Pi_j = p_1 + p_2 + \dots + p_j, \quad j = 1, 2, \dots, r - 1 \quad (4.107)$$

De *kumulative odds* indføres da som

$$\theta_j^K = \frac{\Pi_j}{1 - \Pi_j}, \quad j = 1, 2, \dots, r - 1 \quad (4.108)$$

og de tilsvarende *kumulative logit'er* er bestemt ved

$$\theta_j^K = \ln(\theta_j^K) = \ln\left(\frac{\Pi_j}{1 - \Pi_j}\right), \quad j = 1, 2, \dots, r - 1 \quad (4.109)$$

Vi bemærker, at alle r kategorier indgår i definitionen af de enkelte kumulative logit'er.

Eksempel 4.14.4 Tilfredshedsangivelser fra buspassagerer, kumulative odds

Som led i de løbende undersøgelser af kundetilfredsheden foretager et busselskab stikprøveundersøgelser blandt passagerne, hvor passagererne bliver bedt om at svare på en række spørgsmål vedrørende servicekvaliteten. Samtidig med registreringen af kundernes tilfredshed registrerer man også forskellige objektive størrelser, herunder bussens eventuelle forsinkelse.

Et af spørgsmålene vedrører kundens tilfredshed med overholdelsen af køreplanen. Spørgsmålet har svarkategorierne "Meget utilfreds", "Utilfreds", "Både og", "Tilfreds", "Meget tilfreds".

Nedenstående tabel viser fordelingen af svarene på 12 233 spørgeskemaer indsamlet blandt kvindelige passagerer. Svarene er opdelt efter bussens forsinkelse i minutter.

Forsinkelse min.	Meget utilfr.	Utilfr.	Både og	Tilfr.	Meget tilfr.	Ialt
0	234 2.3 %	559 5.4 %	1 157 11.2 %	5 826 56.4 %	2 553 24.7 %	10 329
2	41 3.6 %	100 8.9 %	145 12.9 %	602 53.5 %	237 21.1 %	1 125
5	42 7.9 %	76 14.3 %	89 16.7 %	254 47.7 %	72 13.5 %	533
7	35 14.3 %	48 19.7 %	39 16.0 %	95 38.9 %	27 11.1 %	244

Vi ser indledningsvist på responserne svarende til busser uden forsinkelse. Nedenstående tabel viser de beregnede (observerede) kumulative odds og logit'er for de forskellige svarkategorier.

	Meget utilfr.	Utilfr.	Både og	Tilfr.	Meget tilfr.	Ialt
Antal	234	559	1 157	5 826	2 553	10 329
Kumu leret antal	234	793	1 950	7 776	10 329	
Odds	0.0232	0.0832	0.2327	3.0458	-	
Logit	-3.7645	-2.4870	-1.4579	1.1138	-	

▽

Såfremt der optræder forklarende variable, $\mathbf{x}^*$, kan man betragte en lineær (affin) model for de kumulative logit'er,

$$\theta_j^K(\mathbf{x}) = \alpha_j + \mathbf{x}^{*'}\boldsymbol{\beta}, \quad (4.110)$$

hvor $\mathbf{x}^*$ angiver en (søjle)-vektor af kendte koefficienter, fælles for samtlige kategorier.

Sættet af parametre, $(\alpha_1, \dots, \alpha_{r-1})$ kaldes *afskæringspunkter* (eng. cutpoints). Det gælder, at α_i er ikke-aftagende som funktion af i .

For modellen bestemt ved (4.110) gælder, at odds-ratioen svarende til to forskellige værdier, $\mathbf{x}_1^*$ og $\mathbf{x}_2^*$ af de forklarende variable er den samme for alle kategorier, j . Der gælder nemlig

$$\theta_j^K(\mathbf{x}_1^*) - \theta_j^K(\mathbf{x}_2^*) = \ln \left(\frac{\Pi_j(\mathbf{x}_1^*)/(1 - \Pi_j(\mathbf{x}_1^*))}{\Pi_j(\mathbf{x}_2^*)/(1 - \Pi_j(\mathbf{x}_2^*))} \right) = (\mathbf{x}_1^* - \mathbf{x}_2^*)'\boldsymbol{\beta} \quad (4.111)$$

Odds ratioen i udtrykket (4.111) kaldes en *kumulativ odds ratio*.

Logaritmen til den kumulative odds ratio er proportional med forskellen mellem værdierne af de forklarende variable. Modellen (4.110) kaldes derfor en *proportional odds model*.

Bemærkning 1 Fortolkning af modeller for kumulative logit'er

Betragt den simple model for de kumulative logit'er

$$\theta_j^K = \alpha_j, j = 1, \dots, r - 1 \quad (4.112)$$

Modellen svarer til, at de kumulative sandsynligheder, Π_j er parametriseret som

$$\Pi_j = \frac{\exp(\alpha_j)}{1 + \exp(\alpha_j)}, j = 1, \dots, r - 1 \quad (4.113)$$

Lad $R \in L(0, 1)$ være en logistisk fordelt stokastisk variabel. Der gælder da
 Definer nu den variable J ved

$$J = j \quad \text{hvis } \alpha_{j-1} < R \leq \alpha_j, \quad (4.114)$$

hvor $-\infty = \alpha_0 < \alpha_1 < \dots < \alpha_r = \infty$. Da gælder netop, at $P[R \leq \alpha_j] = \Pi_j$, $j = 1, \dots, r-1$, hvorfor vi har, at $P[J = j] = P[A_j] = p_j$ for $j = 1, \dots, r-1$. Den variable J angiver således indeks for den hændelse, der indtræffer i et enkelt forsøg, og man kan opfatte R som en finere angivelse af responset. R betegnes undertiden den *latente variable*.

▽

Hvis responsfordelingen i stedet havde været givet ved (4.110) ville man kunne anstille analoge betragtninger for $R \in L(-\mathbf{x}^{*\prime}\boldsymbol{\beta}, 1)$, dvs en positionsforskydning $\mathbf{x}^{*\prime}\boldsymbol{\beta}$ enheder.

Specielt, hvis der kun er en enkelt, kontinuert, forklarende variabel, x , kan vi betragte *responskurven* svarende til det j 'te kumulative respons,

$$F_j(x) = P[A_1 \cup A_2 \dots \cup A_j] = P[R \leq \alpha_j] \quad (4.115)$$

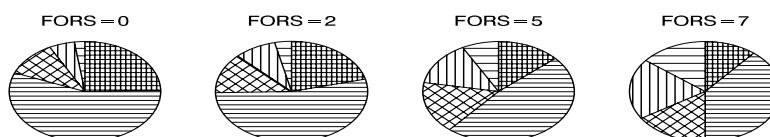
Responskurven for de kumulative responser svarer til de logistiske regressionsmodeller, der blev diskuteret ovenfor. For modellen (4.110) med proportionale odds med en enkelt, kontinuert, forklarende variabel gælder altså

$$F_\nu(x) = F_j(x + (\alpha_\nu - \alpha_j)/\beta) \quad (4.116)$$

De enkelte responskurver fremkommer altså af hinanden ved translation $(\alpha_\nu - \alpha_j)/\beta$ enheder i x -aksens retning.

Betragt nu fordelingen af svarene ved de forskellige forsinkelseskategorier. Nedenstående figur illustrerer disse fordelinger grafisk.

Passagertilfredshed ved forskellige forsinkelser



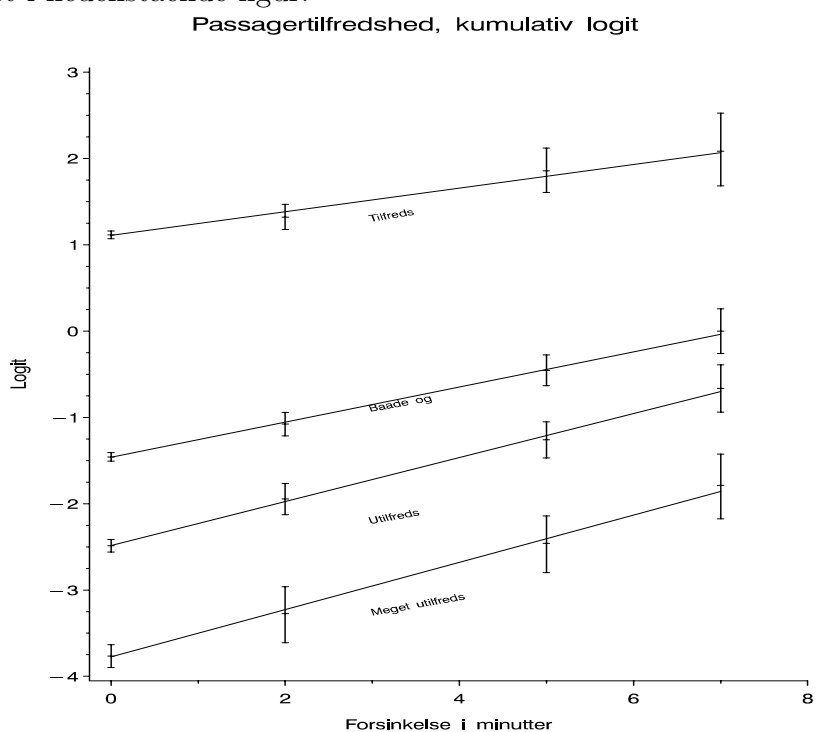
SVAR m.utilfr utilfr baade og m.tilfr

Svarkategorierne er *ordnede*, gående fra “Meget utilfreds” til “Meget tilfreds”. Ordningen er afspejlet i figuren ved at andelen af besvarelser af kategorien “Meget utilfreds” er markeret umiddelbart til venstre for klokken 12, derefter fortsættes mod uret indtil kategorien “Meget tilfreds”, der er afbildet umiddelbart til højre for klokken 12. Det ses, at andelen af utilfredse kunder stiger, jo større forsinkelsen er, og tilsvarende falder andelen af kunder, der erklærer sig tilfredse eller meget tilfredse.

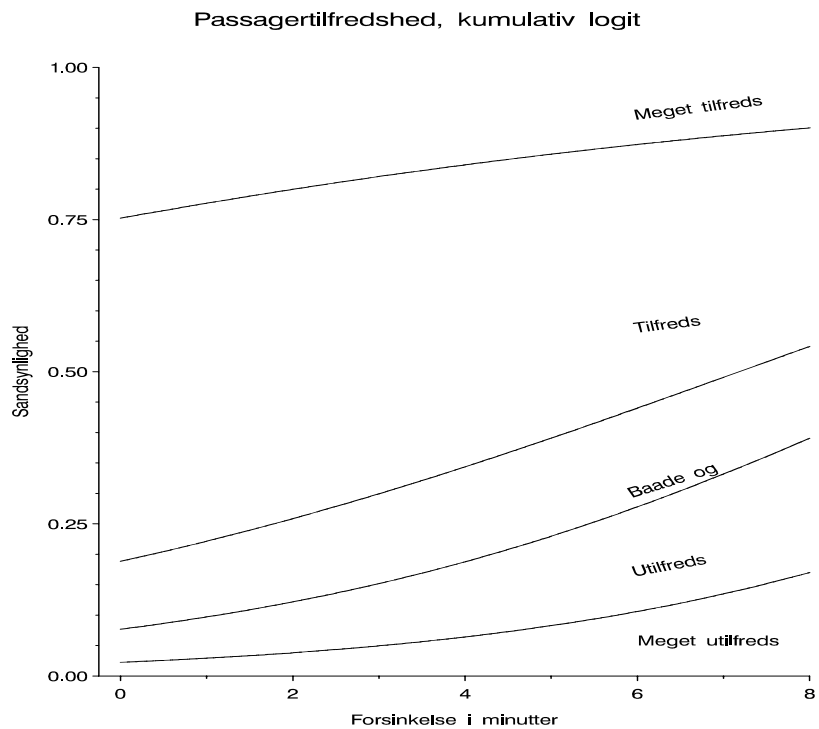
Man kan nu tilpasse en logistisk model for de kumulative logit’er svarende til et binomialt respons for hver kategori for sig. Vi bemærker, at residualdevianserne svarende til de forskellige kategorier ikke er uafhængige. Vi tillader os alligevel som en grov modelkontrol at beregne en total ved addition.

Svar-kategori	α_j	β_j	$D(\mathbf{y}, \hat{\Pi})$
Meget Utilfreds	-3.7733	0.2736	0.3744
Utilfreds	-2.4849	0.2550	0.4157
Både og	-1.4603	0.2033	0.2157
Tilfreds	1.1101	0.1367	1.0090
Ialt			1.7991

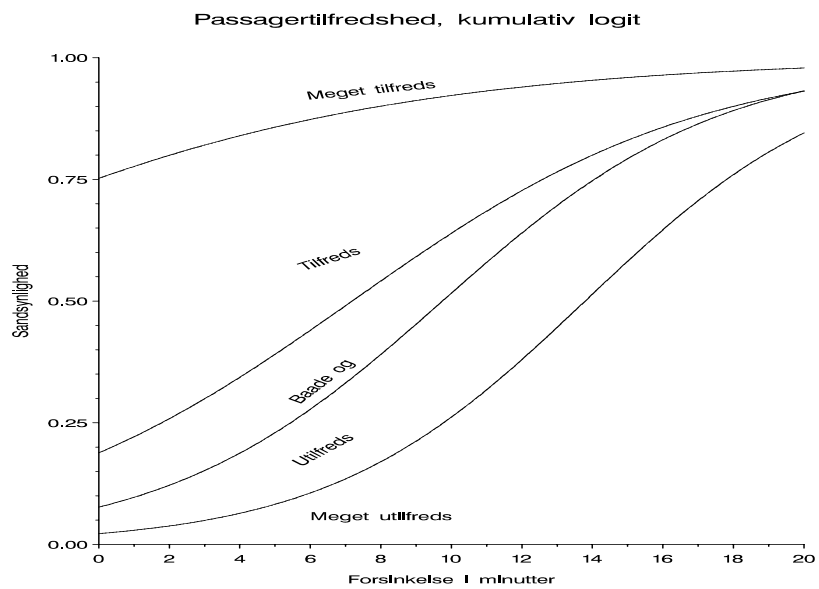
Residualdevianserne giver ikke anledning til afvisning af modellen.
De estimerede relationer mellem de kumulative logit'er og forsinkelsen er illustreret i nedenstående figur.



Forløbet af de estimerede sandsynligheder for svar i de respektive kategorier er illustreret i nedenstående figur.



Nedenstående figur viser det extrapolerede forløb af sandsynlighederne for de forskellige svarkategorier ved forsinkelser op til 20 minutter.



Vi bemærker, at vi tillod forskellige hældninger for de forskellige kumulative logit'er. Det kan imidlertid ofte være ønskeligt (og rimeligt) at benytte en model med samme hældning. Dette kan foretages i proceduren GENMOD (Generaliserede lineære modeller)

SAS-procedurer

Procedurerne LOGISTIC og GENMOD kan håndtere multinomialfordelinger data med linkfunktionerne, probit, logit, komplementær log-log for forskellige valg af odds.

SAS-proceduren PROC GENMOD giver mulighed for tilpasning ved kumulative logit'er med

```
DATA bustilf;
INPUT fors RESP $ antal total ;
CARDS;
0 AMU 234 10239
0 AU 559 10239
0 BO 1157 10329
0 CT 5826 10329
0 MT 2553 10329
2 AMU 41 1125
2 AU 100 1125
2 BO 145 1125
2 CT 602 1125
2 MT 237 1125
5 AMU 42 533
5 AU 76 533
5 BO 89 533
5 CT 254 533
5 MT 72 533
7 AMU 35 244
7 AU 48 244
7 BO 39 244
7 CT 95 244
7 MT 27 244
;
PROC GENMOD ;
```

```

CLASS RESP;
MODEL RESP = fors /DIST=MULT LINK=CLOGIT;
WEIGHT antal;
RUN;

```

Proceduren estimerer forskellige intercepter, men samme hældning.

4.14.5 Poisson regression, brug af OFFSET

Vi vil her betragte en “regressionsmodel” for Poissonfordelte data.

Eksempel 4.14.5 *Antal trafikulykker som funktion af trafikmængde*

Datasættet `modsvjan` indeholder de registrerede uheldstal for januar kvartal for en række år svarende til uheldskategorien “modsving” for ulykker i dagslys. Endvidere er angivet trafikindekset, der er et mål for trafikmængden.

Man kunne nu forestille sig en model, $Y_i \sim P(\mu_i)$, hvor variationen i antallet af uheld beskrives ved en Poissonfordeling, dvs variansfunktionen $V(\mu) = \mu$.

Tilsvarende kunne man forestille sig, at uheldsantallet er proportionalt med trafikmængden,

$$\mu_i = \rho x_i$$

Den kanoniske link for Poissonfordelingen er

$$g(\mu) = \ln(\mu)$$

dvs udtrykt ved den kanoniske link er modellen

$$\eta_i = \beta + \ln(x_i)$$

med $\beta = \ln(\rho)$, eller

$$\eta_i - \ln(x_i) = \beta$$

dvs en model med *offset* lig med $\ln(x_i)$ og blot et intercept led. Proportionalitetsfaktoren $\rho = \exp(\beta)$

Bemærk, at en model af formen `UHELD = LINDEX` med Distribution Poisson

og **Link Logarithm** ville have givet anledning til at der blev estimeret en ko-efficient til $\ln(x_i)$ i den lineære prædikator (svarende til at ulykkerne voksede som en estimeret *potens* af trafikmængden - og altså ikke nødvendigvis proportionalt med x_i som ønsket).

Goodness-of fit teststørrelsen udtrykker godheden af tilpasningen til denne linie.

Estimationen er ganske simpel:

$$\hat{\rho} = \sum y_i / \sum x_i$$

altså det totale antal ulykker divideret med den totale trafikmængde.

Fortolket ved de individuelle ulykkesintensiteter $\hat{\rho}_i = y_i/x_i$ finder man

$$\hat{\rho} = \sum x_i \hat{\rho}_i / \sum x_i$$

altså et *vejet gennemsnit med trafikmængderne som vægte*.

Nemlig fordi (under modellen har man)

$$V[\hat{\rho}_i] = V[Y_i]/x_i^2 = (\rho x_i)/x_i^2 = \rho/x_i$$

og så vægter vi bare de individuelle hældningsestimater med deres præcision.

▽

4.14.6 Empiriske varianser for normalfordelte observationer

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

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

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

Vi ved fra statistik 1, at

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

med $f = n-1$.

Men da denne fordeling er den samme som en $G(f/2, \sigma^2/(f/2))$ -fordeling (jvf. afsnit 1.4.2), har vi altså idet familien af gammafordelinger er en eksponentiel dispersionsmodel at familien af fordelinger for S^2 er en *vægtet* eksponentiel dispersionsmodel med middelværdiparameter

$$E[S^2] = \sigma^2 ,$$

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

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

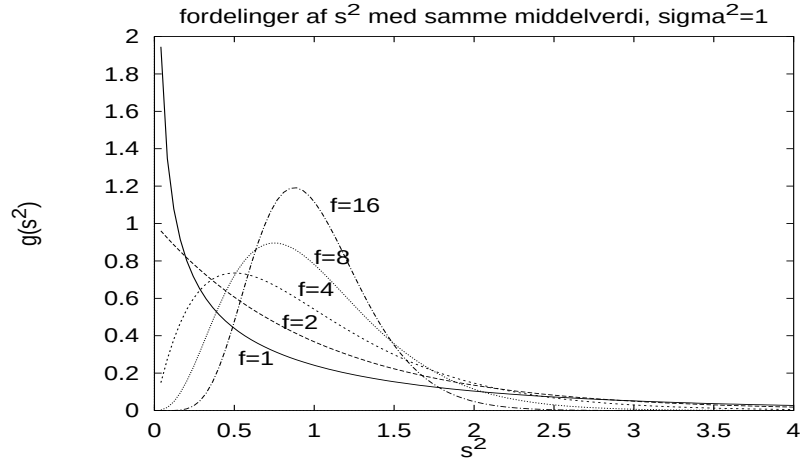
Man kunne godt vælge at fastsætte dispersionsparameteren **Scale** til $2/f$, hvis alle observationer har samme antal frihedsgrader, herved opnår man at nogle af hjælpestørrelserne bliver skalerede, men sædvanligvis vælger man at benytte en vægtet model med dispersionsparameter 1, og vægten $f/2$.

Der er en enentydig sammenhæng mellem variationskoefficienten i fordelingen og formparameteren, $f/2$, nemlig

$$\frac{2}{f} = \frac{V[S^2]}{(E[S^2])^2} = V[S^2/\sigma^2] \quad (4.117)$$

Fordelingen af den empiriske varians for normalfordelte 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 $\sigma^2 = 1$.



Eksempel 4.14.6 *Gamma regression*

Lad observationerne $Y_1, Y_2, \dots, Y_k$ være indbyrdes uafhængige gammafordelte variable med $Y_i \in G(\alpha_i, \mu_i/\alpha_i)$.

Modellen

$$H_0 : \mu_i = \gamma x_i, \quad i = 1, 2, \dots, k, \quad (4.118)$$

hvor $\mu_i = E[Y_i]$ og $x_1, x_2, \dots, x_k$ er kendte kovariater, kaldes *Gamma regressionsmodellen*.

Idet den kanoniske linkfunktion for gammafordelingen er den reciproke $\theta = 1/\mu$, (jvf. tabellen over kanoniske links) ser vi, at modellen (4.118) svarer til

$$H_0 : \theta_i = \frac{1}{\gamma x_i},$$

dvs

$$H_0 : \theta_i = \frac{1}{\gamma} t_i$$

med $t_i = 1/x_i$. Modellen er således af formen svarende til en kanonisk regressionsmodel, hvor interceptleddet α antages at være nul. Middelværdiligningen reduceres her til den enkelte ligning

$$\sum_{i=1}^k \alpha_i y_i \frac{1}{x_i} = \sum_{i=1}^k \alpha_i \gamma x_i \frac{1}{x_i} \quad (4.119)$$

med løsningen

$$\hat{\gamma} = \sum_{i=1}^k \alpha_i \frac{y_i}{x_i} / \sum_{i=1}^k \alpha_i \quad (4.120)$$

og de fittede værdier $\hat{\mu}_i$ under modellen (4.118) bliver

$$\hat{\mu}_i = \hat{\gamma} x_i \quad (4.121)$$

Såfremt formparameteren er den samme for alle observationer, $\alpha_i = \alpha$, bliver estimatet blot

$$\hat{\gamma} = \frac{1}{k} \sum_{i=1}^k \frac{y_i}{x_i} \quad (4.122)$$

Vi bemærker, at gammafordelingens variansstruktur, $V(\mu) = \mu^2$, svarende til en konstant variationskoefficient $\sqrt{V[Y_i]}/E[Y_i]$, har den konsekvens, at estimatet for den fælles hældning netop bliver det *simple aritmetiske gennemsnit af de individuelle hældninger*. ∇

Eksempel 4.14.7 *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 at de empiriske varianser, $s_i^2 \in G(f_i/2, \sigma_i^2/(f_i/2))$ med $E[s_i^2] = \sigma_i^2$.

Man ønsker at vurdere modellen

$$H_0 : \sigma_i^2 = \gamma/x'_i, \quad i = 1, 2, 3,$$

dvs netop regressionsmodellen (4.118) med $x_i = 1/x'_i$.

Vi får derfor estimatet (4.119)

$$\hat{\gamma} = \sum_{i=1}^3 (f_i/2) s_i^2 x'_i / \sum_{i=1}^3 (f_i/2) = 5.57$$

Nedenstående tabel viser de fittede værdier og deviansresidualerne, bestemt ved hjælp af enhedsdeviansen for Gammafordelingen (jvf tabellen over enhedsdevianser).

$$d(y; \mu) = 2 \times \{y/\mu - \ln(y/\mu) - 1\}$$

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
$d(s_i^2; \hat{\sigma}_i^2)$	0.11785	0.00900	0.00130
$(f_i/2)d(s_i^2; \hat{\sigma}_i^2)$	0.11785	0.01800	0.00391
$r_D(s_i^2; \hat{\sigma}_i^2)$	-0.34329	0.13416	0.06254

Man finder goodness of fit teststørrelsen $D^*(\mathbf{s}^2; \hat{\boldsymbol{\sigma}}^2) = 0.1398$, 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.

Hvis frihedsgraderne er de samme for alle observationer, kan man i stedet i SAS vælge en fast `Scale` som $f_i/2$. ∇

4.14.7 Bartlett's test

Det fra statistik 1 kendte test for varianshomogenitet, Bartlett's test, Sætning 11 i stat 1 noten, afsnit 5.3 er netop test for tilpasning til en model med bare et interceptled i en generaliseret lineær model for gammafordelte empiriske varianser.

Testet kaldes også Bartlett's test efter den engelske statistiker M.S.Bartlett. Bartlett (1937) angav ikke blot teststørrelsen, men udledte desuden en korrektionsfaktor,

$$c = 1 + \frac{1}{3(k-1)} \left(\sum_{i=1}^k \frac{1}{f_i} - \frac{1}{f} \right), \quad (4.123)$$

til teststørrelsen $D(\mathbf{s}^2; \hat{\sigma}^2)$, sådan at fordelingen af

$$\frac{1}{c} D(\mathbf{s}^2; \hat{\sigma}^2)$$

svarer bedre til $\chi^2(k-1)$ -fordelingen. Princippet for bestemmelsen af denne korrektionsfaktor kan generaliseres til den generelle kvotientteststørrelse. En sådan korrektion kaldes "Bartlett-korrektion".

4.15 Tofaktormodeller, vekselvirkning

Betragt en model med to forklarende faktorvariable, og antag at den additive model

$$\eta_{p,q} = \alpha_p + \beta_q \quad (4.124)$$

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*

$$\Delta_{p,p';q}^A = \eta_{p,q} - \eta_{p',q} \quad (4.125)$$

mellem niveau p og p' af faktoren A , ser vi, at under den additive model (4.124) gælder:

$$\Delta_{p,p';q}^A = \alpha_p + \beta_q - (\alpha_{p'} + \beta_q) = \alpha_p - \alpha_{p'} ,$$

der ikke afhænger af q .

Sætning 4.15.1 *Konstant differentiel effekt indebærer additiv model uden vekselvirkning*

Betragt en model med to faktorer. Hvis den differentielle effekt $\Delta_{p,p';q}^A$ ikke afhænger af niveauet q , da gælder den additive model (4.124)

Bevis:

Hvis den differentielle effekt $\Delta_{p,p';q}^A$ ikke afhænger af q , da kan den udtrykkes som

$$\Delta_{p,p';q}^A = \eta_{p,q} - \eta_{p',q} = \delta(p, p') \quad (4.126)$$

hvor $\delta(p, p')$ ikke afhænger af q .

Sætter vi $p' = 1$ har vi af (4.126), at

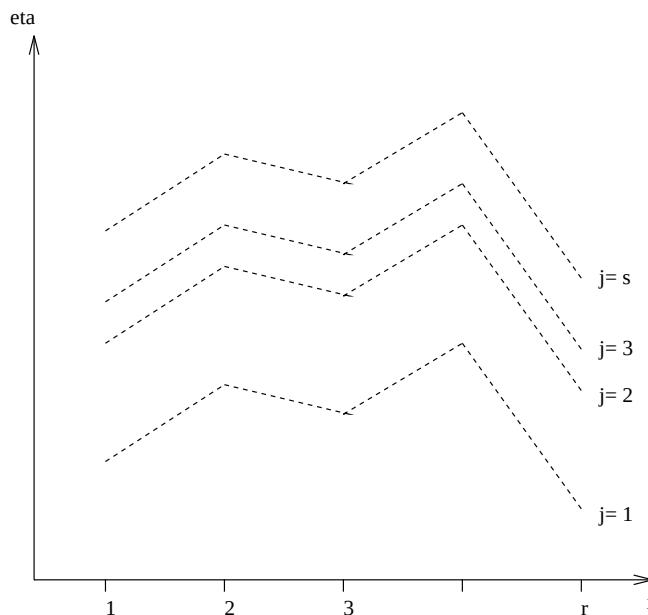
$$\eta_{p,q} = \eta_{1,q} + \delta(p, 1)$$

altså netop på formen svarende til en model uden vekselvirkning. ∇

4.15.1 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 $g(y_{1,q}), g(y_{1,q}), \dots, g(y_{r,q})$. Under den additive model vil profilerne af faktor A svarende til forskellige niveauer af faktor B være parallelle.



Figur 4.4: Eksempel på profilplot

$$\eta_{i,j} = \alpha_i + \beta_j$$

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 til at specificere en bestemt celle i tabellen.

Hvis der var vekselvirkning, vil et udsagn om en forskel på eksempelvis to rækker afhænge af den specifikke vægtning af søjlerne.

?? giver et eksempel på en sådan opspaltning af modeldeviansen.

4.15.2 Tofaktormodel for Poissonfordelte data

Eksempel 4.15.1 *Kvartalsvise uheldstal for motorkøretøjer*

Nedenstående tabel viser de kvartalsvise uheldstal for uheld med personskaade med motorkøretøjer for uheldskategorien “møde” i dagslys for ikke-spirituspåvirkede førere for årene 1987-89 i Danmark

Indeks	Antal uheld	år	kvartal
i	y_i	z_{i1}	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 personskaade i årene 1987-1989

Kategori "møde"					
År	Kvartal				Ialt
	1	2	3	4	
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 4.5.

Liniestykkerne er med god tilnærmelse parallelle.

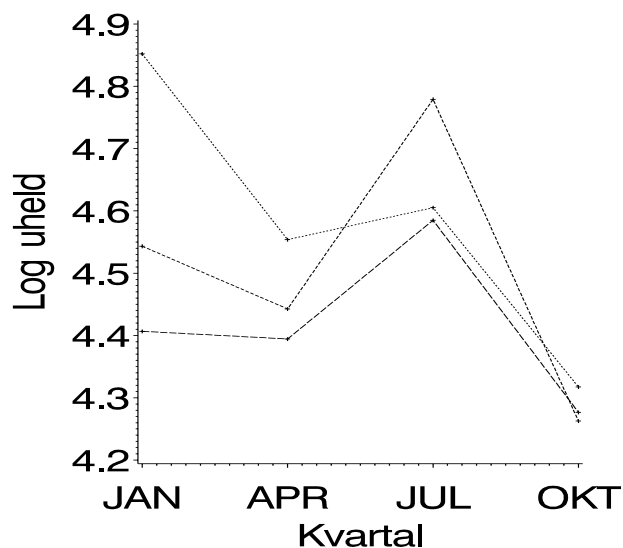
Vi har imidlertid mulighed for at teste denne antagelse om forsvindende vekselvirkning.

Den *fulde model*, der tilknytter hver observation sin egen middelværdi er af formen $\text{UHELD} = \text{AAR KVT AAR} \cdot \text{KVT}$. Vi vælger at tilpasse en model uden vekselvirkning udtrykt som $\text{UHELD} = \text{AAR KVT}$, og vælger Poissonfordeling og kanonisk link.

Deviansanalyseskemaet er

Analysis of Deviance					
Source	DF	Deviance	Deviance / DF	Scaled Dev	Pr >
Model	5	28.3293	5.6659	28.3293	<.0001
Error	6	8.9533	1.4922	8.9533	.
C Total	11	37.2826	.	.	.

Profilplot for log personskadeuheld



Figur 4.5: Profilplot for logaritmen til kvartalsvise uheldstal i årene 1987-1989

Uheldstal for samme år er forbundet med rette linier

Residualdeviansen på 8.9533 svarer til 82% fraktilen i en χ^2_6 -fordeling. Der er således ikke noget, der taler imod at antage en Poissonfordelingsmodel uden vekselvirkning.

Man kan derfor forsøge successivt at fjerne led i modellen.

Type III opspaltningen af residualdeviansen er

Type III (LR) Tests			
Source	DF	ChiSq	Pr > ChiSq
AAR	2	5.8096	0.0548
KVT	3	22.5197	<.0001

der viser, at når der korrigeres for effekten af kvartalerne, er der ikke en sig-

nifikant indflydelse af årene. Når der korrigeres for effekten af årene, er der en stærkt signifikant indflydelse af kvartalerne. Man kunne derfor overveje at reducere modellen ved at fjerne AAR fra modellen.

Vi bemærker i øvrigt, at i dette Poissonfordelingstilfælde er opspaltningen i bidragene fra AAR og KVT ortogonal. Type I opspaltningen er nemlig

Type I (LR) Tests

Source	DF	ChiSq	Pr > ChiSq
AAR	2	5.8096	0.0548
KVT	3	22.5197	<.0001

Den additive model for η_{pq} modsvarer af en multiplikativ model for $\mu_{pq} = E[Y_{pq}]$, dvs en model af formen

$$\mu_{pq} = \alpha_p^* \beta_q^* \quad (4.127)$$

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 $\Delta_{p,p';q}^A$ for η_{pq} (se afsnit 4.15) 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.

▽

4.15.3 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 med blommestiklingerne, at den resulterende model for p 'erne ikke gav anledning til simple relationer direkte mellem p 'erne.

De differentielle effekter (4.125)

$$\Delta_{i,i';j}^A = \eta_{i,j} - \eta_{i',j}$$

bliver her differenser mellem logit'er, dvs

$$\Delta_{i,i';j}^A = \ln \left(\frac{p_{i,j}}{(1 - p_{i,j})} \right) - \ln \left(\frac{p_{i',j}}{(1 - p_{i',j})} \right) \quad (4.128)$$

eller

$$\Delta_{i,i';j}^A = \ln \left(\frac{p_{i,j}(1 - p_{i',j})}{(1 - p_{i,j})p_{i',j}} \right) = \ln(\omega_{i,i';j}^A), \quad (4.129)$$

hvor vi har indført den differentielle effekt på odds-skalaen

$$\omega_{i,i';j}^A = \frac{p_{i,j}/(1 - p_{i,j})}{p_{i',j}/(1 - p_{i',j})} \quad (4.130)$$

Den betragtede additive model for logit'erne er ensbetydende med at de tilsvarende differentielle effekter $\Delta_{i,i';j}^A$ eller $\omega_{i,i';j}^A$ ikke afhænger af j , eller analogt, at de differentielle effekter $\Delta_{i,j,j'}^B$ eller $\omega_{i,j,j'}^B$ 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 $\text{logit}(p)$, da gælder der også en additiv model for $\text{logit}(1 - p) = -\text{logit}(p)$.

Logaritmisk Link ved binomialt respons

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

$$P[\text{System fejlfri}] = p_A p_B$$

Antag nu, at A foreligger i r varianter, A_i , $i = 1, 2, \dots, r$ og B foreligger i s varianter, B_j , $j = 1, 2, \dots, s$, 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

$$P[\text{System fejlfri}] = p_{A_i} p_{B_j}$$

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 $B(n_{i,j}, p_{i,j})$ -fordelt variabel. Hypotesen om uafhængighed mellem A og B svarer til den additive model for $\ln(p_{i,j})$

$$\eta_{i,j} = \ln(p_{i,j}) = \kappa + \alpha_i + \beta_j \quad (4.131)$$

dvs en generaliseret lineær model for $Y_{i,j} = Z_{i,j}/n_{i,j}$ med en *logaritmisk link-funktion*.

Hypotesen om uafhængighed svarer altså til hypotesen om forsvindende vekselvirkning i denne model.

De differentielle effekter svarende til (4.131) er

$$\Delta_{i,i';j}^A = \ln\left(\frac{p_{i,j}}{p_{i',j}}\right) = \ln(\omega_{i,i';j}^A), \quad (4.132)$$

hvor ratioen

$$\omega_{i,i';j}^A = \frac{p_{i,j}}{p_{i',j}} \quad (4.133)$$

angiver forholdet mellem sandsynlighederne $p_{i,j}$ og $p_{i',j}$.

Den anden ordens differentielle effekt er

$$\Delta_{i,i';j,j'}^{AB} = \Delta_{i,j,j'}^B - \Delta_{i',j,j'}^B = \ln\left(\frac{p_{i,j}p_{i',j'}}{p_{i,j'}p_{i',j}}\right) = \ln(\omega_{i,i';j,j'}^{AB}), \quad (4.134)$$

hvor den tilsvarende differentielle effekt for sandsynlighederne er

$$\omega_{i,i';j,j'}^{AB} = \frac{p_{i,j}}{p_{i,j'}} / \frac{p_{i',j}}{p_{i',j'}} \quad (4.135)$$

Forholdet (4.135) kaldes undertiden *Yules krydsprodukt ratio*.

Eksempel 4.15.2 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.

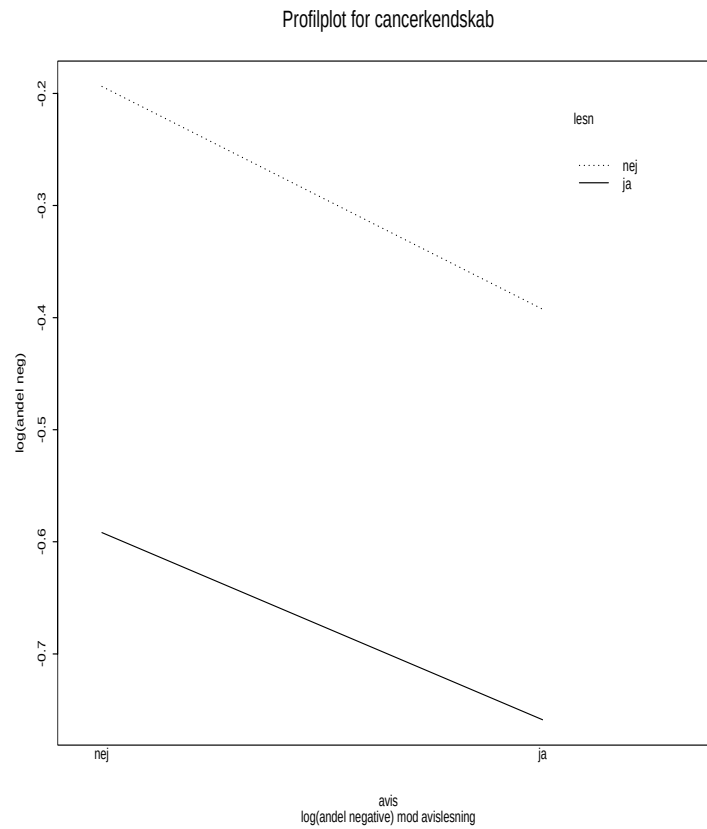
Antal personer med ringe viden om cancer/antal udspurgte

Avislæsn.	læsning	
	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

$$\ln(p_{i,j}) = \kappa + \alpha_i + \beta_j \quad (4.136)$$

Man finder estimatorne :

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	.	.

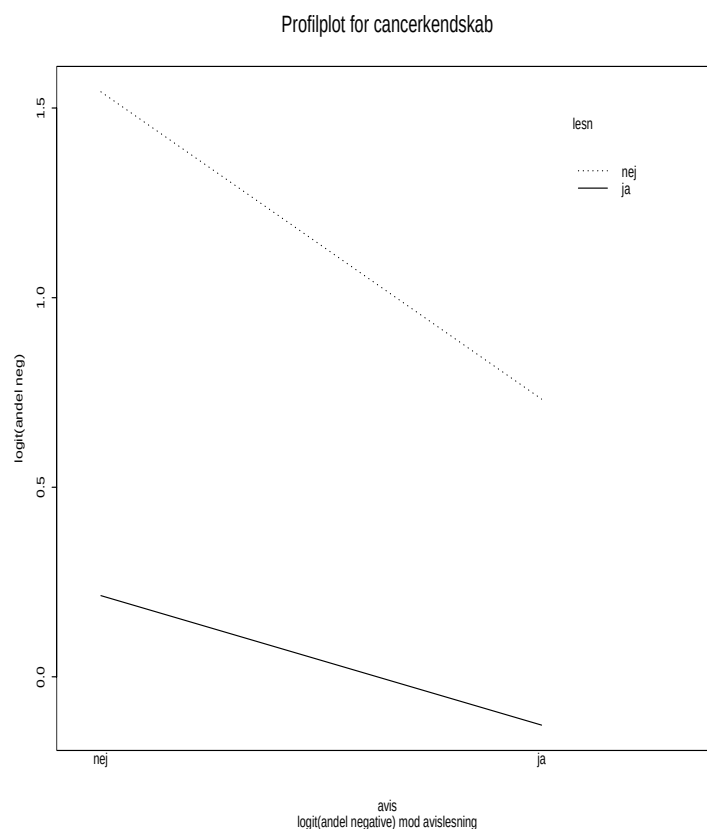
NOTE: The scale parameter was held fixed.

og en deviansteststørrelse for modeltilpasning til hypotesen (4.136), $D(\mathbf{y}; \hat{\mathbf{p}}) = 0.0941$. Der er altså ingen grund til at afvise hypotesen.

Såfremt man havde valgt at betragte en additiv model for logit'erne

$$\ln(p_{i,j}/(1 - p_{i,j})) = \kappa + \alpha_i + \beta_j \quad (4.137)$$

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	.	.

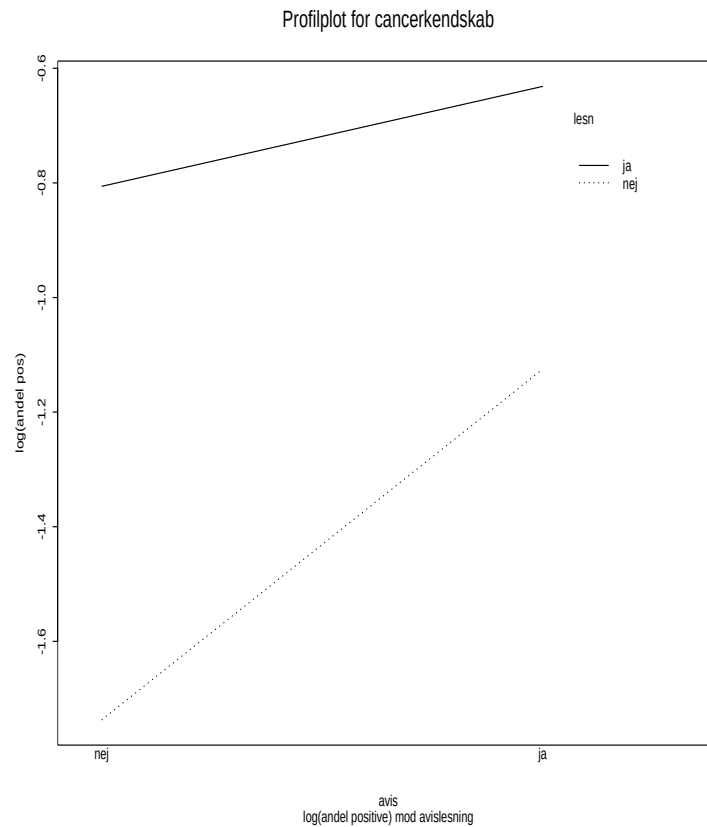
NOTE: The scale parameter was held fixed.

og deviansteststørrelsen for modeltilpasning til denne hypotese er $D(\mathbf{y}; \hat{\mathbf{p}}) = 3.0741$, 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.

$$\ln(1 - p_{i,j}) = \kappa + \alpha_i + \beta_j \quad (4.138)$$

Profilplottet svarende til denne hypotese er vist nedenfor:



Også under denne model ses en klar tendens til vekselvirkning.

Man finder estimerne

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 $D(\mathbf{y}; \hat{\mathbf{p}}) = 6.2690$, 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:

$$p_{1,1}p_{2,2} = p_{1,2}p_{2,1}$$

▽

4.16 Linkfunktioner

Menuen **Fit** i SAS-insight giver mulighed for valg mellem en række forskellige linkfunktioner. SAS-proceduren **GENMOD** (Generaliserede lineære modeller giver endvidere mulighed for at man selv programmerer sin link-funktion). Nedenfor gives en oversigt over de sædvanlige linkfunktioner.

Eksempel 4.16.1 *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

$$\eta = \mu^k = \begin{cases} \mu^k & \text{for } k \neq 0 \\ \log(\mu) & \text{for } k = 0 \end{cases} \quad (4.139)$$

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

Eksempel 4.16.2 *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. ∇

4.17 Lidt mere om eksponentielle familier

De eksponentielle familier spiller en væsentlig rolle i forbindelse med teorier for statistisk inferens. Dette skyldes blandt andet, at for disse familier er det muligt at vise eksistens (og ofte entydighed) af “optimale” inferentielle størrelser.

I dette afsnit gives en uddybende introduktion til de eksponentielle som supplement til introduktionen i afsnit 4.3. Afsnittet tjener hovedsageligt referencemål for den mere matematisk orienterede læser.

Bemærkning 1 *Flerdimensional eksponentiel familie*

Definition 4.3.1 på side 104 vedrører blot en endimensional eksponentiel familie.

Eksponentielle familier spiller en væsentlig rolle i matematisk statistisk teori, da de giver mulighed for at opnå principiel indsigt i problemstillinger knyttet til parametrisk statistisk inferens.

Lad $\nu\{\cdot\}$ være et σ -endeligt mål¹ på $\mathbb{R}^k$, som ikke er koncentreret på et affint underrum af $\mathbb{R}^k$.

Betragt

$$\psi(\theta) = \int \exp\{\theta'x\} \nu\{dx\}, \theta \in \mathbb{R}^k. \quad (4.140)$$

og sæt

$$D = \{\theta \in \mathbb{R}^k : \psi(\theta) < \infty\} \quad (4.141)$$

Mængden D er konveks. Lad $\text{int}(D)$ angive det indre af D .

Lad endelig

$$\kappa(\theta) = \ln(\psi(\theta)) \quad \text{for } \theta \in D. \quad (4.142)$$

Såfremt $\text{int}(D) \neq \emptyset$, kaldes familien af fordelinger med tætheder (mht. $\nu\{\cdot\}$)

$$f(x; \theta) = \exp\{\theta'x - \kappa(\theta)\} \quad \text{for } \theta \in D \quad (4.143)$$

for den naturlige eksponentielle familie frembragt af målet $\nu\{\cdot\}$ på $\mathbb{R}^k$. Såfremt specielt målet $\nu\{\cdot\}$ er en sandsynlighedsfordeling, $F(\cdot)$, siger man, at familien er frembragt af fordelingen $F(\cdot)$. En naturlig eksponentiel familie kan frembringes af enhver af fordelingerne i familien.

Den stokastiske variable X kaldes den *kanoniske stikprøvefunktion*, parameteren θ kaldes den *kanoniske parameter*, og mængden D kaldes det *kanoniske parameterområde* for familien. Det positive heltal, k , kaldes for familiens *orden*.

¹Sædvanligvis vil $\nu\{\cdot\}$ være på formen $\nu\{dx\} = c(x)dx$, hvor dx er enten Lebesguemålet (kontinuert tæthed) eller tælleområdet (diskret tæthed)

Det gælder, at den kanoniske stikprøvefunktion er *sufficient* for den kanoniske parameter.

▽

Bemærkning 2 *Den eksponentielle familie er bestemt af det frembringende mål $\nu(\cdot)$*

Det ses, at såvel $\psi(\cdot)$ (og dermed også $\kappa(\cdot)$) og parameterområdet D er bestemt af målet $\nu(\cdot)$.

Såfremt man har blot ét sandsynlighedsmål ν på $\mathbb{R}$, kan man altså benytte dette mål til at frembringe en eksponentiel familie ved (4.143). ▽

Teorien for eksponentielle familier knytter sig stærkt til teorien for vektorrum. I ovenstående definition antog vi, at der var valgt basis i vektorrummet således at det kunne identificere med $\mathbb{R}^k$. Dette er imidlertid ikke nødvendigt. Man kan definere en naturlig eksponentiel familie koordinatfrit på et abstrakt, endeligdimensionalt vektorrum. Vi vil her skitsere denne definition.

Lad E være et vilkårligt endeligdimensionalt vektorrum og lad E^* angive det duale rum (mængden af linearformer på E). Lad desuden $\langle \theta, x \rangle$ angive den kanoniske bilinearform på $E^* \times E$.

Lad endelig $\nu\{\cdot\}$ være et mål på E og betragt

$$\psi(\theta) = \int \exp\{\langle \theta, x \rangle\} \nu\{dx\}, \theta \in E^*.$$

Lad

$$D = \{\theta \in E^* : \psi(\theta) < \infty\}$$

og sæt

$$\kappa(\theta) = \ln(\psi(\theta)) \quad \text{for } \theta \in D.$$

Lad $\mathcal{P}$ angive samlingen af sandsynlighedsmål, der har tætheden

$$f(x; \theta) = \exp\left\{\langle \theta, x \rangle - \kappa(\theta)\right\} \quad \text{for } \theta \in D \quad (4.144)$$

med hensyn til målet ν .

Familien $\mathcal{P}$ med elementer $P \in \mathcal{P}$ kaldes den *naturlige eksponentielle familie* frembragt af målet $\nu\{\cdot\}$ og funktionen $\kappa(\cdot)$ kaldes *kumulantfrembringeren* for familien. Det væsentlige indhold i denne koordinatfri fremstilling er, at den tydeliggør at de kanoniske parametre er elementer i det duale vektorrum til vektorrummet, E , for observationerne.

Middelværdien svarende til fordelingen $P \in \mathcal{P}$ er den linearform på E^* , der fører $\ell \in E^*$ over i $\int \ell(x)P\{dx\} = \mathbb{E}[\ell(X)]$, dvs middelværdien er et element i E^{**} (mængden af linearformer på E^*), der er isomorf med E .

Afbildningen

$$\tau(\theta) = \frac{\partial}{\partial \theta} \kappa(\theta) = \int x \exp \left\{ \langle \theta, x \rangle - \kappa(\theta) \right\} \nu\{dx\}$$

er en reel analytisk, bijektiv afbildning, der afbilder $\text{int}(D)$ ind på middelværdimængden, $\mathcal{M}$.

Kovariansen, Σ svarende til fordelingen $P \in \mathcal{P}$ er den symmetriske, positiv semi-definite bilinearform på $E^* \times E^*$, der fører (ℓ_1, ℓ_2) over i

$$\mathbb{E}[(\ell_1(X) - \mathbb{E}[\ell_1(X)])(\ell_2(X) - \mathbb{E}[\ell_2(X)])]$$

Nu er imidlertid mængden af bilinearformer på $E^* \times E^*$ isomorf med mængden af lineære afbildninger $E^* \rightarrow E$. Man kan således opfatte en kovarians Σ som en lineær afbildning $E^* \rightarrow E$. (Afbildningen er symmetrisk, dvs den er lig med sin transponerede).

Bemærkning 3 *Dualitet mellem variation og usikkerhed*

Denne koordinatfri fremstilling antyder noget om spillet mellem *variation*, som den er udtrykt gennem fordelingen (frekvensfunktionen) for Y , og *usikkerheden* om parameter værdien, θ , idet jo observation og parameter ligger i hver sit duale vektorrum, bundet sammen gennem det indre produkt.

Fordelingsforholdene for Y for en bestemt parameter værdi, θ , er fuldstændigt fastlagt af kumulantfrembringeren $\kappa(\cdot)$. ∇

Bemærkning 4 *Alternativt udtryk for enhedsdeviansen*

De fleste fremstillinger af generaliserede lineære modeller tager udgangspunkt i den kanoniske parametrisering (4.2), Enhedsdeviansen fremkommer da ved at betragte log-likelihoodfunktionen som funktion af $\mu = \tau(\theta)$.

Logaritmen til likelihoodfunktionen svarende til observationen x er jvf (4.2)

$$l(\theta; x) = \theta x - \kappa(\theta) \quad (4.145)$$

Betragter vi i stedet likelihoodfunktionen som funktion af middelværdiparameteren, μ , får vi

$$l_\mu(\mu; x) = l(\tau^{-1}(\mu); x) = \tau^{-1}(\mu)x - \kappa(\tau^{-1}(\mu))$$

Sætter man differentialkvotienten af (4.145) med hensyn til θ lig med nul, finder man (idet $\kappa'(\theta) = \tau(\theta)$) netop løsningen $\hat{\theta} = \tau^{-1}(x)$ svarende til at observationen sættes lig med sin middelværdi (idet vi har antaget, at $x \in \mathcal{M}$).

Deviansen, $d(x; \mu)$ defineres sædvanligvis som

$$\begin{aligned} d(x; \mu) &= 2\{l_\mu(x; x) - l_\mu(\mu; x)\} \\ &= 2[x\{\tau^{-1}(x) - \tau^{-1}(\mu)\} - \{\kappa(\tau^{-1}(x)) - \kappa(\tau^{-1}(\mu))\}] \end{aligned} \quad (4.146)$$

altså forskellen mellem maksimum af loglikelihoodfunktionen og værdien af loglikelihoodfunktionen svarende til μ .

Man kan vise (sætning 4.17.2), at denne definition er ækvivalent med definitionen (4.14). ∇

En naturlig eksponentiel familie er fastlagt af sin variansfunktion

Sætning 4.17.1 *En naturlig eksponentiel familie er fastlagt ved sin variansfunktion*

Bevis:

Vi skitserer beviset for en endimensional familie. Beviset i den generelle situation er anført af Letac (1992).

Idet

$$\tau(\theta) = \frac{\partial}{\partial \theta} \kappa(\theta)$$

og $\theta = \tau^{-1}(\mu)$ følger det af relationen

$$\frac{\partial}{\partial \theta} \tau(\theta) = \frac{\partial}{\partial \theta'} \left(\frac{\partial}{\partial \theta} \kappa(\theta) \right) = \frac{\partial^2}{\partial \theta' \partial \theta} \kappa(\theta) \quad (4.147)$$

at

$$\frac{\partial}{\partial \mu} \tau^{-1}(\mu) = \{V(\mu)\}^{-1}, \quad (4.148)$$

og endvidere har vi af (4.11), at

$$\frac{\partial}{\partial \theta} \kappa(\theta) = \tau(\theta). \quad (4.149)$$

For en given variansfunktion, $V(\cdot)$, kan man således bestemme κ ved at løse differentialligningen (4.148), invertere τ^{-1} , og derefter løse differentialligningen (4.149). Når $\kappa(\cdot)$ er fastlagt, er familien (4.143) fastlagt. ∇

Vi bemærker endvidere af (4.148), at variansfunktionen $V(\cdot)$ kan udtrykkes som

$$V(\mu) = \left\{ \frac{\partial}{\partial \mu} \tau^{-1}(\mu) \right\}^{-1} \quad (4.150)$$

Eksempel 4.17.1 *Bestemmelse af eksponentiel familie med $V(\mu) = \mu$*

Antag, at vi ønsker at bestemme en naturlig eksponentiel familie af fordelinger sådan at $V(\mu) = \mu$, dvs sådan at

$$V[Y] = E[Y]$$

for enhver parameterværdi.

Det følger af (4.148), at den inverse afbildning til middelværdiafbildningen skal opfylde

$$\frac{d\theta}{d\mu} = \frac{1}{\mu},$$

dvs

$$\tau^{-1}(\mu) = \theta(\mu) = \int^{\mu} \frac{1}{t} dt = \ln(\mu) - \ln(c)$$

Middelværdiafbildningen er altså af formen

$$\tau(\theta) = c_1 \exp(\theta)$$

dvs at (4.149) bliver

$$\frac{d\kappa}{d\theta} = c_1 \exp(\theta)$$

hvorfor

$$\kappa(\theta) = \int^{c_1 \exp(\theta)} c_1 \exp(t) dt = c_1 \exp(\theta) - c_1 \exp(c_2) = C \exp(\theta)$$

Man kan overbevise sig om, at dette netop er kumulantfrembringeren for Poisson-fordelingen. Ellers kan man kigge i bogen om fordelinger med anvendelse i statistik.

▽

Bemærkning 1 *Lokale egenskaber for enhedsdeviansen*

Man finder de første to afledede af enhedsdeviansen

$$\frac{\partial}{\partial \mu} d(y; \mu) = -2 \frac{y - \mu}{V(\mu)} \quad (4.151)$$

$$\frac{\partial^2}{\partial \mu^2} d(y; \mu) = 2 \left\{ \frac{1}{V(\mu)} + (y - \mu) \frac{V'(\mu)}{V(\mu)^2} \right\} \quad (4.152)$$

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

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

▽

Sætning 4.17.2 *Relation mellem variansfunktion og enhedsdevians*

Betragt en naturlig eksponentiel familie med kumulantfrembringer $\kappa(\cdot)$, dvs med tætheden (4.2). Lad enhedsdeviansen $d(\cdot; \cdot)$ være givet ved (4.14) og variansfunktionen $V(\cdot)$ ved (4.13).

Der gælder da, at variansfunktionen kan udtrykkes ved enhedsdeviansen som

$$V(\mu) = 2 \Big/ \left[\frac{\partial^2}{\partial \mu^2} d(y; \mu) \right]_{y=\mu} \quad (4.153)$$

og omvendt kan enhedsdeviansen udtrykkes som

$$d(y; \mu) = 2 \int_{\mu}^y \frac{y-u}{V(u)} du \quad . \quad (4.154)$$

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

▽

Bemærkning 1 *Momenter for enhedsdeviansen*

Ved benyttelse af bemærkning 1 på side 208 finder man, at såfremt fordelingen af Y tilhører en reproduktiv eksponentiel dispersionsmodel med enhedsdeviansen $d(y; \mu)$, variansfunktion $V(\mu)$ og præcisionsparameter λ gælder

$$\mathbb{E} \left[\frac{\partial}{\partial \mu} d(Y; \mu) \right] = 0 \quad (4.155)$$

$$\mathbb{V} \left[\frac{\partial}{\partial \mu} d(Y; \mu) \right] = \frac{4\lambda}{V(\mu)} \quad (4.156)$$

$$-\mathbb{E} \left[\frac{\partial^2}{\partial \mu^2} d(Y; \mu) \right] = \frac{2\lambda}{V(\mu)} \quad (4.157)$$

NB CHECK λ

▽

Normalfordelinger som eksponentielle familier

Sætning 4.17.3 *Familien af $N(\mu, \sigma^2)$ -fordelinger er en eksponentiel familie af orden 2*

Familien af $N(\mu, \sigma^2)$ -fordelinger for $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$ udgør en eksponentiel familie af orden 2 med kanonisk parameter

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix}, \quad (4.158)$$

kanonisk stikprøvefunktion

$$\mathbf{t} = \begin{pmatrix} t_1(x) \\ t_2(x) \end{pmatrix} = \begin{pmatrix} x \\ x^2 \end{pmatrix}$$

og kumulantfrembringer

$$\kappa(\boldsymbol{\theta}) = \ln \left(\sqrt{\pi} \right) - \frac{1}{2} \ln(-\theta_2) - \frac{\theta_1^2}{4\theta_2}.$$

Det kanoniske parameterområde er $D = \mathbb{R} \times]-\infty, 0[$, som er åbent. Familien er regulær.

Bevis:

Omskrivningen

$$f(x; \mu, \sigma^2) = \frac{\exp \left(-\mu^2/(2\sigma^2) \right)}{\sqrt{2\pi\sigma^2}} \exp \left(\frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} x^2 \right) \quad (4.159)$$

viser, at tætheden kan udtrykkes på formen (4.143). ∇

Sætning 4.17.4 *For fastholdt σ^2 er familien af $N(\mu, \sigma^2)$ -fordelinger en naturlig eksponentiel familie*

Bevis:

Følger ved at betragte tætheden (4.159). ∇

Normalfordelingen som eksponentiel dispersionsmodel

Lad $Y \in N(\mu, \sigma^2)$. Vi så ovenfor, at familien af $N(\mu, \sigma^2)$ -fordelinger for $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$ udgør en eksponentiel dispersionsfamilie med kanonisk

parameter $\theta = \mu$ og med enhedskumulantfrembringer $\kappa(\theta) = \theta^2/2$. Parameterområdet er $D = \mathbb{R}$ og $\Delta = \mathbb{R}_+$.

Da normalfordelingen netop i forvejen er parametriseret ved sin middelværdi, μ , og da den kanoniske parameter netop er μ , bibringer middelværdiafbildningen

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

os ikke noget nyt. Vi bemærker dog, at den kanoniske link,

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

er den identiske afbildning.

Enhedsvariansfunktionen er

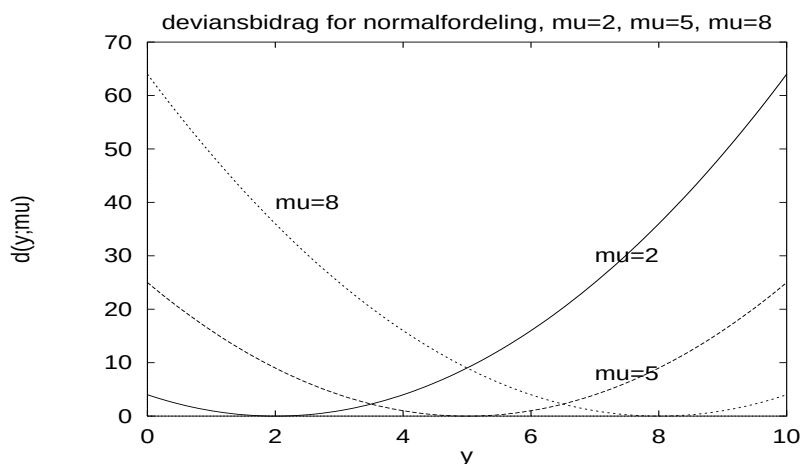
$$V_n(\mu) = \tau'(\tau^{-1}(\mu)) = 1 .$$

Variansen afhænger således ikke af middelværdien, men er alene udtrykt gennem dispersionsparameteren σ^2 .

Enhedsdeviansen for $N(\mu, \sigma^2)$ fordelingen er

$$d(y; \mu) = (y - \mu)^2 \quad (4.160)$$

Nedenstående figur viser enhedsdeviansen for normalfordelingen svarende til $\mu = 2$, $\mu = 5$ og $\mu = 8$. Det ses at grafen af enhedsdeviansen er symmetrisk omkring $y = \mu$ og at grafen svarende til forskellige værdier af μ fremkommer ved en parallelforskydning.



Den sædvanlige repræsentation af tætheden for en $N(\mu, \sigma^2)$ -fordeling er netop udtrykt ved enhedsdeviansen $d(y; \mu)$ på formen (4.16) med

$$a(y; \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}$$