Dagens program:

Mandag den 28. februar

Generaliserede lineære modeller

- Motiverende eksempel, logistisk regression
- Eksponentielle dispersionsfamilier
- Generaliseret lineær model
- Estimation, fordeling af residualer mv
- Test for modelreduktion
- Eksempler på modeller

Datasæt musefostre og blommer

Toxitetsvurdering for kølevæske:

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

Andel døde musefostre som fkt af indgivet konc.:

Hvad er der galt med lineær model og mindste kvadraters metode (GLM) ?

- Andelen (responsen) skal ligge mellem nul og 1
- Variansen på responset er ikke konstant, men afhænger af middelværdien for responset, så observationerne bør ikke tillægges samme vægt

Det er i hvert fald et problem, når vi arbejder med "tælletal", dvs for binomialfordeling og Poissonfordeling

Eksponentielle familier af fordelinger:

Familie af fordelinger: En indekseret samling af fordelinger

Eksponentielle familier er sådan at **summen** af uafhængige obs fra fordelingen er **sufficient**, dvs at man kan tillade sig at reducere et observationssæt til summen af obs uden at miste information.

Exponentiel familie, fælles grundform:

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

Vi udvider til **eksponentielle dispersionsfamilier** med grundformen

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

Eksempel, binomialfordeling:

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

for $y = 0, 1/n, \ldots, 1$ og $0 \le p \le 1$.

Lad

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

så kan frekvensfunktionen udtrykkes som

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

med

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

og $\lambda = n$.

Altså to ækvivalente parametriseringer, ved p eller ved **kanonisk parameter** θ.

Kumulantfrembringer :

Funktionen $\kappa(\cdot)$ beskriver hele fordelingen.

Således fås middelværdi og varians for fordelingen svarende til en bestemt værdi af θ

$$\mathsf{E}\,[Y] = \kappa'(\theta) \tag{4.9}$$

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

Naturligt at kalde λ for præcisionsparameter. Jo større λ desto mindre varians (større præcision)

Middelværdiafbildning :

Funktionen

$$\tau(\theta) = \kappa'(\theta) \tag{4.11}$$

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

Afbildningen kaldes **middelværdiafbildningen**.

Den omvendte afbildning, $\theta = \tau^{-1}(\mu)$, der fører middelværdien over i den kanoniske parameter, kaldes den **kanoniske linkfunktion**

Variansfunktion :

(enheds) variansfunktionen udtrykker fordelingens varians som en funktion

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

af middelværdien.

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**, $V(\mu)$.

Enhedsdevians :

Vi indfører **enhedsdeviansen** ved

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

Tætheden for en observation kan udtrykkes ved enhedsdeviansen (dvs som funktion af μ) som

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

hvor $d(y;\mu)$ er enhedsdeviansen og λ er præcisionsparameteren

[Vi kunne også have indført enhedsdevians som logaritmen til den relative likelihood af μ (i forhold til maksimal likelihood)]

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

Enhedsdevianser svarende til forskellige 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	$(0,1)$	$\mu(1-\mu)$	$2\big\{y\ln(y/\mu)+$ $(1-y)\ln\big((1-y)/(1-\mu)\big)\big\}$	n [c]	$\ln\big(\mu/(1-\mu)\big)$
Neg Bin	$(0,1)$	$\mu(1+\mu)$	$2\big\{y\ln\big(y(1+\mu)/[(1+y)\mu]\big)$ $+\ln\big((1+\mu)/(1+y)\big)\big\}$	r [d]	$\ln(\mu)$
I Gauss	$(0,\infty)$	μ^3	$(y-\mu)^2/(y\times\mu^2)$		$1/\mu^2$
GHS [e]	$(-\infty,\infty)$	$1+\mu^2$	$2y\big\{\arctan(y)-\arctan(\mu)\big\}+$ $+\ln\big((1+\mu^2)/(1+y^2)\big)$	$1/\sigma^2$	$\arctan(\mu)$

Model for et observationssæt:

observationssæt $Y_1, Y_2, \ldots, Y_n$. Uafhængige samme eksponentielle dispersionsparameterfamilie, kumulantfrembringer $\kappa(\cdot)$. Præcisionsparameter for fordeling af Y_i givet som $\lambda_i = w_i$ (kendt).

log-likelihood:

$$\ell_\theta(\boldsymbol{\theta};\mathbf{y}) \ = \ \sum_{i=1}^{k} w_i(\theta_i y_i - \kappa(\theta_i)) \tag{4.21}$$

$$\ell_\mu(\boldsymbol{\mu};\mathbf{y}) \ = \ -\frac{1}{2}\sum_{i=1}^{k} w_i d(y_i;\mu_i) = -\frac{1}{2}D(\mathbf{y};\boldsymbol{\mu}) \ , \tag{4.22}$$

hvor

$$D(\mathbf{y};\boldsymbol{\mu}) = \sum_{i=1}^{k} w_i d(y_i,\mu_i) \ . \tag{4.23}$$

Scorefunktion :

I analogi med normalfordelingsituationen ser vi først på scorefunktion og observeret information svarende til fuld model (hver obs sin egen parameter)

$$\ell'_\theta(\boldsymbol{\theta};\mathbf{y}) = \mathrm{diag}(\mathbf{w})(\mathbf{y}-\boldsymbol{\tau}(\boldsymbol{\theta})) \tag{4.25}$$

$$\ell'_\mu(\boldsymbol{\mu};\ \mathbf{y}) = \mathrm{diag}\left\{\frac{w_i}{V(\mu_i)}\right\}(\mathbf{y}-\boldsymbol{\mu}) \tag{4.27}$$

Er netop proportional med $(\mathbf{y}-\boldsymbol{\mu})$

Observeret information:

Observeret information

$$\mathbf{j}(\boldsymbol{\theta};\mathbf{y}) = \text{diag}\{w_i\, V(\tau(\theta_i))\} \tag{4.26}$$

Afhænger af parameterværdien $\boldsymbol{\theta}$

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

Afhænger også af $\mathbf{y}$

Generaliseret lineær model:

Ingredienser:

- observationssæt $Y_1, Y_2, \ldots, Y_k$
- eksponentiel dispersionsmodel med variansfunktion $V(\mu)$
- **linkfunktion** $\eta = g(\mu)$ der fører middelværdi over i **lineær prediktor** (koordinatvis)
- affin hypotese vedrørende $\eta_1, \eta_2, \ldots, \eta_k$

$$H_0 \;:\; \boldsymbol{\eta} - \boldsymbol{\eta}_0 \in L \;, \tag{4.30}$$

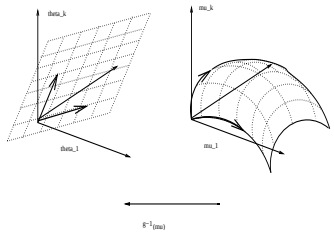
dvs $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$

Ideen er at vi laver en lineær model for en funktion (linkfunktionen) af middelværdien

Forskel fra den generelle lineære model (normalfordeling):

- fordelingen af observationerne 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 parameterværdi 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.
- I normalfordelingsmodeller skriver vi sædvanligvis obs = middelværdi plus støj, hvor støj $\sim \mathsf{N}(0, \sigma^2)$. Det kan vi ikke for de andre fordelingsmodeller. For dem siger vi $Y \sim \text{Ford}(\mu)$, og så formulerer vi modeller for μ (eller rettere for en funktion $\eta = g(\mu)$.

Middelværdirum og rum for lineær prediktor:



Lineær prediktor til venstre, middelværdirum til højre

Estimation

Vi skal bare minimere **deviansen**

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

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

Scorefunktion mht $\boldsymbol{\beta}$

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

hvor

$$\mathbf{X}(\boldsymbol{\beta}) \stackrel{\text{DEF}}{=} \frac{\partial\mu}{\partial\beta} = \left[\frac{\mathsf{d}\mu}{\mathsf{d}\eta}\right]^T \frac{\partial\eta}{\partial\beta} = \mathsf{diag}\left\{\frac{1}{g'(\mu_i)}\right\}\mathbf{X} \tag{4.35}$$

er den **lokale design matrix**

For kanonisk link bestemmes løsningen fra **Middelværdiligningen**:

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\boldsymbol{\mu}(\boldsymbol{\beta}) \tag{4.43}$$

Eksistens og entydighed:

Højst én løsning pga konveksitet.

Man kan dog risikere at løsningen ligger på randen - og at randen ikke hører med til parameterområdet, nemlig feks en løsning $\theta = -\infty$ svarende til $p = 0$ i binomialfordelingssituation..

Musefostre, logistisk regression:

Middelværdiligning

$$\begin{pmatrix} \sum_{i=1}^k n_i \exp(\alpha+\beta x_i)/\{1+\exp(\alpha+\beta x_i)\} \\ \sum_{i=1}^k x_i n_i \exp(\alpha+\beta x_i)/\{1+\exp(\alpha+\beta x_i)\} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^k z_i \\ \sum_{i=1}^k x_i z_i \end{pmatrix}\ , \tag{4.53}$$

Sidespring om fejlophobningsloven :

law of error propagation

For $Z = g(Y)$ med E[Y] $= \mu$ gælder

$$\mathsf{E[Z]} \approx \mathsf{g}(\mu), \quad \mathsf{V[Z]} \approx \left(\mathsf{g}'(\mu)\right)^2 \mathsf{V[Y]}$$

Resultatet fremkommer ved Taylorudvikling (dvs linearisering) af $g(Y)$ omkring $Y = \mu$

$$g(Y) = g(\mu) + (Y-\mu)g'(\mu) + o(Y-\mu)$$

Man kan bestemme maksimum-likelihood løsningen ved iterativt genvægtet mindste kvadraters metode.

Hvad er fidusen ved en generaliseret lineær model ?

Eksempel: Estimation af hældning i regressionsmodeller

Data $(x_1, y_1); (x_2, y_2); \ldots; (x_n, y_n)$

Model:

$$\mathsf{E}[\mathsf{Y}] = \beta \mathsf{x}$$

Individuelle estimater for hældningen:

$$\widehat{\beta}_i = \frac{y_i}{x_i}$$

Hvad er fidusen ved en generaliseret lineær model ?

Normalfordeling: $Y \in \mathsf{N}(\beta x, \sigma^2)$. $\mathsf{V}[\mathsf{Y}] = \sigma^2$

$$\widehat{\beta} = \frac{\sum y_i x_i}{\sum x_i^2} = \frac{\sum x_i^2 \widehat{\beta}_i}{\sum x_i^2}$$

Poissonfordeling: $Y \in \mathsf{P}(\beta x)$. $\mathsf{V}[\mathsf{Y}] = \beta \mathsf{x}$

$$\widehat{\beta} = \frac{\sum y_i}{\sum x_i} = \frac{\sum x_i \widehat{\beta}_i}{\sum x_i}$$

Gammafordeling: $Y \in \mathsf{G}(\alpha, \beta x/\alpha)$. $\mathsf{V}[\mathsf{Y}] = (\beta \mathsf{x})^2$

$$\widehat{\beta} = \frac{\alpha \sum y_i/x_i}{\sum \alpha} = \frac{\sum \widehat{\beta}_i}{n}$$

Eksempel 4.14.6

Vejet gennemsnit af individuelle hældninger, hvor de individuelle hældninger vægtes med deres **præcision**

Fordeling af estimatorer

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \overset{\text{as}}{\in} \mathsf{N}_m(\mathbf{0}, \boldsymbol{\Sigma}) \,, \tag{4.44}$$

$$\mathbf{D}\left[\widehat{\boldsymbol{\beta}}\right] = \boldsymbol{\Sigma} = [\mathbf{X}^T \mathbf{W}(\boldsymbol{\beta}) \mathbf{X}]^{-1} \tag{4.45}$$

med

$$\mathbf{W}(\boldsymbol{\beta}) = \mathsf{diag}\left\{\frac{w_i}{[g'(\mu_i)]^2 V(\mu_i)}\right\} \,, \tag{4.46}$$

For kanonisk link bliver den bare

$$\mathbf{W}(\boldsymbol{\beta}) = \mathsf{diag}\left\{w_i V(\mu_i)\right\} \tag{4.47}$$

Fittede værdier, residualer

Fittede værdier af lineær prædiktor, $\widehat{\eta}$ og af middelværdier $\widehat{\mu}$.

Residualer:

Deviansresidual

$$r_i^D = r_D(y_i; \widehat{\mu}_i) \overset{\text{DEF}}{=} \mathsf{sign}(y_i - \widehat{\mu}_i)\sqrt{w_i d(y_i, \widehat{\mu}_i)} \tag{4.72}$$

Måler i likelihoodværdier.

Pearson-residual

$$r_i^P = r_P(y_i; \widehat{\mu}_i) \overset{\text{DEF}}{=} \frac{y_i - \widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i)/w_i}} \tag{4.73}$$

Simpel, intuitiv.

Standardisering sikrer samme varians (ved division med $(1 - h_{ii})$)

Test for modeltilpasning, goodness of fit

Sætning 4.8.1 (side 143):

Teststørrelse $D(\mathbf{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))$.
Under $H_0 : \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ vil $D(\mathbf{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}))$ approximativt følge en $\chi^2(k-m)$-fordeling. Forkast for store værdier.

Evt kan man bruge Pearson teststørrelse.

Bemærk, at i modsætning til normalfordelingsmodellen kan vi her bruge residualdeviansen til at teste modeltilpasning.

Analysis of deviance tabel

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

Test for modelreduktion

Foregår ved opspaltning af deviansen:

$$D\big(\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}); \widehat{\boldsymbol{\mu}}_M\big) = D(\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}); \boldsymbol{\mu}(\widehat{\boldsymbol{\gamma}})) + D\big(\boldsymbol{\mu}(\widehat{\boldsymbol{\gamma}}); \widehat{\boldsymbol{\mu}}_M\big)$$

Under hypotesen $\boldsymbol{\eta} = \mathbf{Z}\boldsymbol{\gamma}$ vil $D(\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}); \boldsymbol{\mu}(\widehat{\boldsymbol{\gamma}}))$ approximativt følge en $\chi^2(m-r)$-fordeling.

Eksempel Blommeudplantning

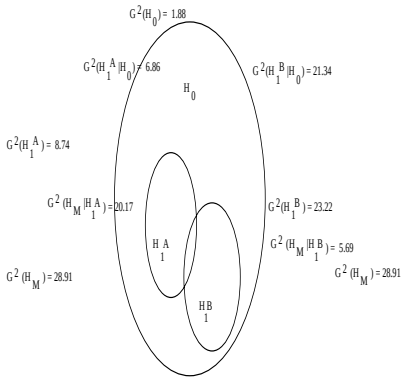
Eksempel 4.11.1 side 152.

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

Tofaktor forsøg med binomial respons

Type I opspaltninger, Blommer



Resume

- Vi har udvidet til andre responsfordelinger, end normalfordelingen
- Disse familier kan beskrives ved kun én parameter, fx middelværdien. Variansen afhænger af middelværdien
- Middelværdien kan ikke variere på hele den reelle akse, derfor betragter vi en funktion af middelværdien, **linkfunktionen**, og laver lineære modeller for denne
- **Kanonisk link** fører middelværdien over i kanonisk parameter (grundform)
- Vi kan måle afvigelsen mellem en observation og en fittet middelværdi ved **deviansen**, og deviansen tager netop hensyn til at variansen afhænger af middelværdien
- og så er resten næsten lige som normalfordelingen.
  Vi estimerer ved at minimere devians. Devianser er approximativt χ^2-fordelte, og vi kan **spalte** devianser.