Stat. 7.5.2010

Kontingenstabeller

A frequency table.

	B_1	B_2	$\cdots$	B_c	Total
A_1	y_{11}	y_{12}	$\cdots$	y_{1c}	y_{1+}
A_2	y_{21}	y_{22}	$\cdots$	y_{2c}	y_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_r	y_{r1}	y_{r2}	$\cdots$	y_{rc}	y_{r+}
Total	y_{+1}	y_{+2}	$\cdots$	y_{+c}	$y_{++} = N$

To faktorer, A og B

A: med r niveauer

B: med c niveauer

Fremkomst af data

(1) Poissonfordelte observationer

$$Y_{hk} \sim \mathcal{P}(\mu_{hk}) \quad \text{uafh.}$$

simultant:

$$P(Y=y) = \prod_h \prod_k \frac{e^{-\mu_{hk}} \mu_{hk}^{y_{hk}}}{y_{hk}!}$$

(2) med $\sum_h \sum_k y_{hk} = N$ fast

$$P(Y=y \mid \sum_i Y_i = N)$$

$(Y_i = Y_{hk})$, $i = 1, \dots, N$; $h = 1, \dots, r$, $k = 1, \dots, c$

$$= \frac{P(Y=y)}{P(\sum_i Y_i = N)}$$

$$= \left[\prod_h \prod_k \frac{e^{-\mu_{hk}} \mu_{hk}^{y_{hk}}}{y_{hk}!} \right] \frac{N!}{e^{-\mu_{..}} \mu_{..}^{N}}$$

$$= N! \prod_h \prod_k \left[\frac{1}{y_{hk}!} \left(\frac{\mu_{hk}}{\mu_{\cdot\cdot}} \right)^{y_{hk}} \right]$$

$$= \binom{N}{y_{11}, \dots, y_{rc}} \prod_h \prod_k \pi_{hk}^{y_{hk}}, \quad \pi_{hk} = \frac{\mu_{hk}}{\mu_{\cdot\cdot}}$$

dvs.

$$(Y_{11}, \dots, Y_{rc}) \sim \underset{\substack{\uparrow \\ \text{multinomialfordelt}}}{m} (N; \pi_{11}, \dots, \pi_{rc})$$

(3) med $\sum_h y_{hk} = y_{\cdot k}$ fast for alle k

$$P(Y = y \mid Y_{\cdot k} = y_{\cdot k}, k = 1, \dots, c)$$

$$= \left[\prod_h \prod_k \frac{e^{-\mu_{hk}} \mu_{hk}^{y_{hk}}}{y_{hk}!} \right] \prod_k \frac{y_{\cdot k}!}{e^{-\mu_{\cdot k}} \mu_{\cdot k}^{y_{\cdot k}}},$$

$$= \prod_k \left[y_{\cdot k}! \prod_h \frac{1}{y_{hk}!} \left(\frac{\mu_{hk}}{\mu_{\cdot k}} \right)^{y_{hk}} \right]$$

$$= \prod_k \left[\binom{y_{\cdot k}}{y_{1k}, \dots, y_{rk}} \prod_h \pi_{h|k}^{y_{hk}} \right], \quad \pi_{h|k} = \frac{\mu_{hk}}{\mu_{\cdot k}}$$

produkt af c multinomialsands.,

dvs.

$$(Y_{1k}, \dots, Y_{rk}) \sim m(y_{\cdot k}; \pi_{1|k}, \dots, \pi_{r|k}),$$

$k = 1, \dots, c$

uafh.

Log-lineare modeller

En log-lineær model er en GLM med logaritmisk link, dvs. $\eta_i = \ln E[Y_i] = \ln \mu_i$

$(\eta_i = x_i^T \beta)$

$g(\mu) = \ln \mu$ er kanonisk link for Poissonfordelingen. Log-lineære modeller er derfor egnede til analyse af kontingenstabeller.

Betragt $Y_i \sim P(\mu_i)$

$$E[Y_i] = \mu_i = \mu_{..} \frac{\mu_i}{\mu_{..}} = \mu_{..} \pi_i$$

Test af uafhængighed mellem faktor A og faktor B i kontingenstabel.

Hypotesen H_0: $\pi_{hk} = \pi_{h.} \pi_{.k}$ for alle h,k

Under H_0: $E[Y_i] = \mu_{..} \pi_{h.} \pi_{.k}$

$$\eta_{hk} = \ln \mu_{..} + \ln \pi_{h.} + \ln \pi_{.k} \qquad (*)$$

$$= \lambda + \lambda_h^A + \lambda_k^B \quad \text{med}$$

$$\left. \begin{array}{l} \sum_h \pi_{h.} = 1 \\ \sum_k \pi_{.k} = 1 \end{array} \right\} \text{ikke-lineære bånd}$$

antal parametre: $1 + (r-1) + (c-1) = r + c - 1$

Alternativ parametrisering:

$$\lambda = \ln \mu_{..} + \frac{1}{r} \sum_h \ln \pi_{h.} + \frac{1}{c} \sum_k \ln \pi_{.k}$$

$$\lambda_h^A = \ln \pi_{h.} - \frac{1}{r} \sum_h \ln \pi_{h.}$$

$$\lambda_k^B = \ln \pi_{.k} - \frac{1}{c} \sum_k \ln \pi_{.k}$$

med $\sum_h \lambda_h^A = 0$

$\sum_k \lambda_k^B = 0$

Referencemodel (den mættede model)

Eneste bånd : $\sum_h \sum_k \pi_{hk} = 1$

Den til (*) analoge log-lineære model

$$\eta_{hk} = \ln \mu_{..} + \ln \pi_{h.} + \ln \pi_{.k} + \ln \frac{\pi_{hk}}{\pi_{h.}\,\pi_{.k}}$$

$$= \lambda + \lambda_h^A + \lambda_k^B + \lambda_{hk}^{AB} \quad \text{med}$$

yderligere bånd (fx) $\sum_h \lambda_{hk}^{AB} = 0$

$\sum_k \lambda_{hk}^{AB} = 0$

antal parametre:

$$1 + (r-1) + (c-1) + (r-1)(c-1) = rc$$

Betragt også $Y_i \mid N \sim m(N; \pi_{11}, \dots, \pi_{rc})$

$$E[Y_i \mid N] = N\pi_i = N\,\pi_{hk}$$

$$= N\,\pi_{h.}\,\pi_{.k} \text{ under } H_0$$

Alle ovenstående regninger i forb. med Poissonmodellen kan opretholdes, når blot $\mu_{..}$ skiftes ud med N. Bemærk dog, at der er én parameter mindre i multinomialmodellen.

Likelihood i log-lineære modeller

l_p loglikelihood i Poissonmodel

l_m loglikelihood i tilsvarende multinomialmodel

$$l_p = \sum_i (y_i \ln \mu_i - \mu_i) + c, \quad \text{jf. AA s. 227}$$

$$l_m = \sum_i y_i \ln \pi_i + c, \quad \text{jf. AA s. 136}$$

$$= \sum_i y_i (\ln \mu_i - \ln \mu_{..}) + c, \quad \text{idet } \pi_i = \frac{\mu_i}{\mu_{..}}$$

$$= \sum_i (y_i \ln \mu_i - \mu_i - y_i \ln \mu_{..} + \mu_i) + c$$

$$= \sum_i y_i \ln \mu_i - \mu_i) - (N \ln \mu_{..} - \mu_{..}) + c$$

$$= l_p - l_p(\mu_{..}) + c$$

Bemærk, at $\left.\frac{\partial}{\partial \mu_{..}} (N \ln \mu_{..} - \mu_{..})\right|_{\mu_{..}=N} = \left.\frac{N}{\mu_{..}} - 1\right|_{\mu_{..}=N} = 0$

Når η_i indeholder et konstant led, så gælder $\hat{\mu}_{..} = \sum_i \hat{\mu}_i = \sum_i y_i = N$, jf. AA s. 236, dvs. $\hat{\mu}_i$ bestemt i Poissonmodellen har også gyldighed i multinomialmodellen.

Uddybende argumentation

$$\eta_i = \ln \mu_i = \alpha + z_i^T \beta \quad \text{(konstantleddet er skilt ud)}$$

$$l_p = \sum_i (y_i \ln \mu_i - \mu_i)$$

$$= \sum_i y_i (\alpha + z_i^T \beta) - \mu_{..}$$

$$= N\alpha + \sum_i y_i (z_i^T \beta) - \mu_{..}$$

$$= \sum_i y_i (z_i^T \beta) - N \ln \sum_i \exp z_i^T \beta$$
$$+ N\alpha + N \ln \sum_i \exp z_i^T \beta - \mu_{..}$$

$$= \sum_i y_i \ln (\exp z_i^T \beta) - \sum_i y_i \ln \sum_j \exp z_j^T \beta$$
$$+ N(\alpha + \ln \sum_i \exp z_i^T \beta) - \mu_{..}$$

$$= \sum_i y_i \ln \frac{\exp z_i^T \beta}{\sum_j \exp z_j^T \beta} + N \ln \mu_{..} - \mu_{..}$$

$$\alpha + \ln \sum_i \exp z_i^T \beta$$
$$= \ln \exp(\alpha + \sum_i z_i \beta)$$
$$= \ln \mu_{..}$$

$$\frac{\exp z_i^T \beta}{\sum_j \exp z_j^T \beta} = \frac{\exp(\alpha + z_i^T \beta)}{\exp(\alpha + \sum_j z_j^T \beta)}$$
$$= \frac{\mu_i}{\mu_{..}} = \pi_i$$

$$= \sum_i y_i \ln \pi_i + N \ln \mu_{\cdot\cdot} - \mu_{\cdot\cdot}$$

dvs.

$$l_p(\mu_{\cdot\cdot}, \beta) = l_m(\beta) + \underbrace{l_p(\mu_{\cdot\cdot})}_{\text{uafh. af } \beta}$$

$l_p(\mu_{\cdot\cdot}, \beta)$ og $l_m(\beta)$ maksimeres for samme $\hat{\beta}$.

Hvis rækkesummerne er faste, eller hvis søjlesummerne er faste, kan tilsvarende ækvivalens mellem Poissonfordelingens likelihood og de indgående multinomialfordelingers likelihood eftervises.

Se tilføjelse side 10

Odds

En hændelse med sands. p har odds $\frac{p}{1-p}$

Bemærk, at ln odds = logit

Betragt skema

A 2 × 2 probability table

Exposure to risk	Occurrence of event $\bar{E}$	E	Total
$\bar{X}$	π_{00}	π_{01}	π_{0+}
X	π_{10}	π_{11}	π_{1+}
Total	π_{+0}	π_{+1}	1

Prospektivt studie

X og $\tilde{X}$ valgt på forhånd

ford. på E og $\tilde{E}$ observeres

$$\operatorname{logit} P(E\mid\tilde{X}) - \operatorname{logit} P(E\mid X)$$

$$= \ln \frac{\frac{\pi_{01}}{\pi_{0\cdot}}}{\frac{\pi_{00}}{\pi_{0\cdot}}} - \ln \frac{\frac{\pi_{11}}{\pi_{1\cdot}}}{\frac{\pi_{10}}{\pi_{1\cdot}}} = \ln \frac{\pi_{01}\,\pi_{10}}{\pi_{00}\,\pi_{11}}$$

$$= -\ln \omega$$

Retrospektivt studie

Der foreligger obs. af E og $\tilde{E}$

De klassificeres efter X og $\tilde{X}$

$$\operatorname{Logit} P(X\mid\tilde{E}) - \operatorname{logit} P(X\mid E)$$

$$= \ln \frac{\frac{\pi_{10}}{\pi_{\cdot 0}}}{\frac{\pi_{00}}{\pi_{\cdot 0}}} - \ln \frac{\frac{\pi_{11}}{\pi_{\cdot 1}}}{\frac{\pi_{01}}{\pi_{\cdot 1}}} = \ln \frac{\pi_{01}\,\pi_{10}}{\pi_{00}\,\pi_{11}}$$

$$= -\ln \omega$$

'Krydsproduktforholdet' er altså ens

$\omega = 1$ svarer til uafhængighed

$\omega > 1$ — " — positivt samspil af faktorer

$\omega < 1$ — " — negativt samspil af faktorer

$$-\ln\omega = \ln\left(\frac{\pi_{01}}{\pi_{00}+\pi_{01}} \cdot \frac{\pi_{00}+\pi_{01}}{\pi_{00}} \; \frac{\pi_{10}}{\pi_{10}+\pi_{11}} \cdot \frac{\pi_{10}+\pi_{11}}{\pi_{11}}\right)$$

$$= \lambda^{XE}_{01} - \lambda^{XE}_{00} + \lambda^{XE}_{10} - \lambda^{XE}_{11}$$

Bånd mellem parametrene

$$\lambda^{XE}_{00} + \lambda^{XE}_{10} = 0 \qquad \text{Heraf} \quad \lambda^{XE}_{10} = -\lambda^{XE}_{00}$$

$$\lambda^{XE}_{01} + \lambda^{XE}_{11} = 0 \qquad \lambda^{XE}_{11} = -\lambda^{XE}_{01}$$

$$\lambda^{XE}_{00} + \lambda^{XE}_{01} = 0 \qquad \lambda^{XE}_{01} = -\lambda^{XE}_{00}$$

$$\lambda^{XE}_{10} + \lambda^{XE}_{11} = 0 \qquad \lambda^{XE}_{11} = -(-\lambda^{XE}_{00})$$

$$\ln\omega = \lambda^{XE}_{00} + \lambda^{XE}_{00} + \lambda^{XE}_{00} + \lambda^{XE}_{00} = 4\lambda^{XE}_{00}$$

Kvasi likelihood

Antag $\quad E[Y_i] = \mu_i$

$\quad Var[Y_i] = \varphi V(\mu_i)$

Sæt $\quad u = u(Y;\mu,\varphi) = \dfrac{Y-\mu}{\varphi V(\mu)}$ (indeks i udeladt)

Heraf $\quad E[u] = 0$

$$Var[u] = \frac{1}{(\varphi V(\mu))^2} Var\, Y = \frac{1}{\varphi V(\mu)}$$

$$\frac{\partial u}{\partial \mu} = \frac{\varphi V(\mu)(-1) - (Y-\mu)\varphi V'(\mu)}{(\varphi V(\mu))^2}$$

$$\Rightarrow -E\left[\frac{\partial u}{\partial \mu}\right] = \frac{1}{\varphi V(\mu)}$$

dvs. u opfører sig analogt til en scoringsfunktion

Kvasi likelihood defineres derfor som

$$Q(\mu;y) = \int_y^\mu u(t,y)\,dt$$

$$= \int_y^\mu \frac{y-t}{\varphi V(t)}\,dt$$

Når flere uafhængige observationer, så

$$Q(\mu;y) = \sum_i Q(\mu_i;y_i)$$

Når Q er en egentlig loglikelihoodfkt., dvs. når der findes en funktion l, hvor

$$\frac{\partial l}{\partial \mu} = \frac{y-\mu}{\varphi V(\mu)}, \quad E[Y]=\mu, \quad Var[Y] = \varphi V(\mu),$$

så foreligger der en exponentiel familie (u/ bevis).

I en regressionsmodel er μ_i en fkt. af β. Kvasi likelihoodligningerne

$$\frac{\partial Q}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j} = 0, \quad j=1,\dots,r$$

ses at være

$$\sum_i \frac{y_i-\mu_i}{V(\mu_i)}\frac{\partial \mu_i}{\partial \beta_j} = 0, \quad j=1,\dots,r$$

og afhænger ikke af dispersionsparameteren φ.

Kvasi devians

$$D(y;\hat\mu) = -2\varphi\, Q(\hat\mu;y)$$

$$= 2\int_{\hat\mu}^{y} \frac{y-t}{V(t)}\,dt \geq 0$$

Når uafhængige observationer

$$D(y;\hat{\mu}) = \sum_i D(y_i;\hat{\mu}_i)$$

eks. Overdispersion i Poissonmodellen vedr stofdata, jf. AA s. 225-226

$D_i \sim P(\beta_i L_i)$

GLM : $E[D_i] = \beta_i L_i$

$Var[D_i] = \varphi E[D_i] = \varphi \beta_i L_i$

faktoren φ modellerer for overdispersion

Smlgn. fig. 6.4 s. 247 i AA og fig. 6.6 s. 262 i AA

Estimation af φ (jf. AA s. 240):

$$\hat{\varphi} = \frac{1}{n-p} \sum_i \frac{w_i (y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

$$= \frac{1}{n-1} \sum_i \frac{(D_i - \hat{\beta} L_i)^2}{\hat{\beta} L_i}$$

Fortsat fra side 6:

Idet maksimum likelihood estimaterne er ens i hhv. den rene Poissonmodel, i multinomialmodellen og i tilfældet med uafhængige multinomialmodeller, kan vi, hvadenten vi ønsker at teste for uafhængighed mellem to faktorer, eller vi ønsker at teste for homogenitet mellem en faktors niveauer, i alle tilfælde benytte Pearsons X^2 som teststørrelse, jf. AA s. 137 og s. 140 samt note 7 s. 10 og s. 11.