Minimal sufficiens

Stikprøvefunktionen $T(y)$ er minimal sufficient for θ, hvis $\forall\, y, z \in \mathcal{Y}$:

$$T(y) = T(z) \iff L(\theta; y) \propto L(\theta; z) \text{ for alle } \theta \in \Theta$$

(modo. og tilstr. bet.)

Likelihood ækvivalens bestemmer en klassedeling af $\mathcal{Y}$. Kald elementerne A_y.

En stikprøvefunktion, der antager en konstant værdi på ethvert A_y og forskellige værdier på forskellige A_y'er, er minimal sufficient.

T_1, T_2 minimal sufficiente $\Rightarrow$ $T_1 = g(T_2)$, g bijektiv

T_1 min. suff., T_2 suff. $\Rightarrow$ $T_1 = h(T_2)$

Bemærk

$\forall\, y, z \in \mathcal{Y}\ \forall \theta \in \Theta$:

$$T(y) = T(z) \iff \frac{L(\theta; y)}{L(\theta; z)} \text{ er uafh. af } \theta$$

eks. $y = (y_1, \dots, y_n)$, $Y_i \sim \text{Cauchy}(1, \theta)$ uafh.

$$L(\theta; y) = \prod_{i=1}^{n} \frac{1}{\pi} \frac{1}{1 + (y_i - \theta)^2}$$

$(y_{(1)}, y_{(2)}, \dots, y_{(n)})$ er suff. for θ (trivielt)

$$\frac{L(\theta; y)}{L(\theta; z)} = \frac{\prod_i (1 + (z_i - \theta)^2)}{\prod_i (1 + (y_i - \theta)^2)}$$

Begge polynomier har koefficienten 1 til θ^{2n}, kvot. derfor kun uafh. af θ, når pol. er ens, dvs. når $(z_1, \dots, z_n)$ er en permutation af $(y_1, \dots, y_n)$.

$(y_{(1)}, y_{(2)}, \dots, y_{(n)})$ er altså min. suff.

$\square$

Eksponentielle familier

$\mathcal{F} = \{ f(\cdot; \theta) \mid \theta \in \Theta \}$ udgør en eksponentiel familie, når

$$f(y;\theta) = q(y) \exp\left(\sum_i \psi_i(\theta)\, t_i(y) - \tau(\theta) \right) \quad *$$

uafh. af θ uafh. af y uafh. af θ uafh. af y

Tilhørende likelihood på eksponentiel form.

Bemærk $f(y;\theta) = 0 \iff q(y) = 0$, dvs. alle tætheder i en eksponentiel familie har samme støtte.

eks. $Y \sim b(n, \theta)$

$$f(y;\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}, \quad y = 0, 1, \dots, n$$

$$= \binom{n}{y} \exp\left(y \ln \frac{\theta}{1-\theta} - (-n \ln(1-\theta)) \right)$$

$\square$

eks. $y = (y_1, y_2, \dots, y_n)$, $Y_i \sim N(\mu, \sigma^2)$ uafh.

$$f(y;\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}}$$

$$= (2\pi)^{-\frac{n}{2}} \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_i y_i^2 - 2\mu \sum_i y_i + n\mu^2\right)\right)$$

$$= (2\pi)^{-\frac{n}{2}} \exp\left(\frac{\mu}{\sigma^2}\sum_i y_i - \frac{1}{2\sigma^2}\sum_i y_i^2 - \left(\frac{n\mu^2}{2\sigma^2} + \frac{n}{2}\ln\sigma^2\right)\right)$$

$\square$

eks. $y = (y_1, y_2, \dots, y_n)$, $Y_i \sim P(x_i\theta)$ uafh., x_i'erne givne

$$f(y;\theta) = \prod_{i=1}^{n} \frac{e^{-x_i\theta}(x_i\theta)^{y_i}}{y_i!} = \frac{e^{-\theta\sum x_i}\,\theta^{\sum y_i}\prod_i x_i^{y_i}}{\prod_i y_i!}$$

$$= \left(\prod_i \frac{x_i^{y_i}}{y_i!}\right)\exp\left(\ln\theta \sum_i y_i - \theta\sum_i x_i\right) \quad \square$$

Når $1, \psi_1(\theta), \psi_2(\theta), \dots, \psi_r(\theta)$ er lin. uafh., siges * at være på reduceret form, og r angiver den eksponentielle familiens orden

Sætning:

* er på reduceret form $\Rightarrow$

$T = (T_1(y), T_2(y), \dots, T_r(y))$ er min. suff. for θ

bevis

* viser, at T er sufficient (Neyman)

betragt $z, w \in \mathcal{Y}$, sæt $T_0(\cdot) = \ln q(\cdot)$

$$l(\theta;z) - l(\theta;w)$$
$$= T_0(z) - T_0(w) + \psi_1(\theta)(T_1(z) - T_1(w)) + \dots + \psi_r(\theta)(T_r(z) - T_r(w))$$

$l(\theta;z) - l(\theta;w) = 0 \iff T_j(z) = T_j(w)$, $j = 0, 1, \dots, r$

$L(\theta;z) \propto L(\theta;w) \iff T_j(z) = T_j(w)$, $j = 1, \dots, r$

T er altså min. suff. for θ

eks. logistisk regression

$y = (y_1, y_2, \dots, y_n)$, $Y_i \sim b(1, \pi_i)$ uafh.

$$\pi_i = \frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)}, \quad i = 1, \dots, n$$

(omv. fkt. $\alpha + \beta x_i = \ln \frac{\pi_i}{1-\pi_i} = \text{logit } \pi_i$)

(x_i, π_i) ligger på $\pi(x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$,

grafer kaldes logistiske kurver

likelihood

$$L(\alpha,\beta;y) = \prod_{i=1}^{n} \left(\frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)}\right)^{y_i} \left(\frac{1}{1+\exp(\alpha+\beta x_i)}\right)^{1-y_i}$$

$$= \prod_{i=1}^{n} (\exp(\alpha+\beta x_i))^{y_i} (1+\exp(\alpha+\beta x_i))^{-1}$$

$$= \exp \sum_{i=1}^{n} ((\alpha+\beta x_i) y_i - \ln(1+\exp(\alpha+\beta x_i)))$$

$$= \exp\left(\alpha \sum_i y_i + \beta \sum_i x_i y_i - \sum_i \ln(1+\exp(\alpha+\beta x_i))\right)$$

$(\sum_i y_i, \sum_i x_i y_i)$ er min. suff. for (α, β)

$r = 2$

$\square$

Gentagne observationer (uafh.)

$y = (y_1, y_2, \dots, y_n)$ reduceret form

$$f(y_j;\theta) = q(y_j) \exp\left(\sum_{i=1}^{r} \psi_i(\theta) t_i(y_j) - \tau(\theta)\right)$$

$$\Rightarrow f(y;\theta) = \left(\prod_{j=1}^{n} q(y_j)\right) \exp\left(\sum_{i=1}^{r} \psi_i(\theta) \sum_{j=1}^{n} t_i(y_j) - n\tau(\theta)\right)$$

dvs. $\left(\sum_{j=1}^{n} t_1(y_j), \sum_{j=1}^{n} t_2(y_j), \dots, \sum_{j=1}^{n} t_r(y_j)\right)$ er

min. suff. for θ, dim. fortsat r uafh. af n

Tilstrækkelige betingelser (ikke nødvendige) for eksponentiel familie (u/ bevis)

- uafh. og identisk ford. variable
- støtte afh. ikke af θ
- eksistens af ikke-triviel suff. stikpr.-fkt. med dim. mindre end y's dim. og og dim. uafh. af n

Flere egenskaber ved eksponentielle familier

- $\tau(\theta)$ vilk. ofte diff., når $\varphi_j(\theta)$, $j = 1, \dots, r$ er diff.
- $$\frac{d^s}{d\theta^s} \int_{\mathcal{Y}} g(y) f(y;\theta)\, d\nu(y) = \int_{\mathcal{Y}} \frac{d^s}{d\theta^s} g(y) f(y;\theta)\, d\nu(y),$$ når integralerne eksisterer

(u/ bevis)

Regulære eksponentielle familier

$\mathcal{F} = \{ f(\cdot\,;\theta) \mid \theta \in \Theta \}$ med f som * er regulær, når

- $\Theta = \left\{ \theta \,\middle|\, \int_{\mathcal{Y}} g(y) \exp\left(\sum_i \varphi_i(\theta) t_i(y)\right) d\nu(y) < \infty \right\}$
- Θ åben delmængde af $\mathbb{R}^k$
- dim. af Θ er lig med dim. af den min. suff. stikpr.-fkt.
- afb. $\theta \mapsto \varphi(\theta) = (\varphi_1(\theta), \varphi_2(\theta), \dots, \varphi_r(\theta))$ er bijektiv
- φ_j'erne vilk. ofte diff. mht. θ's komponenter

eks. $y = (y_1, y_2, \dots, y_n)$

$$Y_1 \sim N\left(0, \frac{\sigma^2}{1-\rho^2}\right), \quad |\rho| < 1$$

$$Y_t = \rho Y_{t-1} + E_t, \quad t = 2, \dots, n, \quad E_t \sim N(0, \sigma^2) \text{ uafh.}$$

bemærk, at Y_t'erne er afh.

$$Y \sim N_n\left(0, \frac{\sigma^2}{1-\rho^2}\Omega\right), \quad \Omega = \begin{bmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & & \rho^{n-2} \\ \vdots & \vdots & & \vdots \\ \rho^{n-1} & \rho^{n-2} & \cdots & 1 \end{bmatrix}$$

(autoregressiv proces af 1. orden)

Alternativ specifikation af variable:

$$Y_1 \sim N\left(0, \frac{\sigma^2}{1-\rho^2}\right), \quad |\rho| < 1$$

$$Y_t \mid (Y_1, Y_2, \dots, Y_{t-1}) \sim N(\rho y_{t-1}, \sigma^2), \quad t = 2, \dots, n$$

$$f(y; \rho, \sigma^2) = \frac{\sqrt{1-\rho^2}}{\sqrt{2\pi}\,\sigma} e^{-\frac{(1-\rho^2)y_1^2}{2\sigma^2}} \cdot \prod_{t=2}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(y_t - \rho y_{t-1})^2}{2\sigma^2}}$$

log likelihood:

$$2\ell(\rho, \sigma^2; y) = c + \ln(1-\rho^2) - n \ln \sigma^2 - \frac{1}{\sigma^2} y_1^2 + \frac{\rho^2}{\sigma^2} y_1^2 - \frac{1}{\sigma^2} \sum_{t=2}^{n} \left(y_t^2 - 2\rho y_t y_{t-1} + \rho^2 y_{t-1}^2\right)$$

$$= c + \ln(1-\rho^2) - n \ln \sigma^2 - \frac{1}{\sigma^2}(-\rho^2 y_1^2) - \frac{1}{\sigma^2}\left(\sum_{t=1}^{n} y_t^2 - 2\rho \sum_{t=2}^{n} y_t y_{t-1} + \rho^2 \sum_{t=1}^{n-1} y_t^2\right)$$

$$= c + \ln(1-\rho^2) - n \ln \sigma^2 - \frac{1}{\sigma^2}\left(d_{00} - 2\rho d_{01} + \rho^2 d_{11}\right),$$

hvor $d_{rs} = \sum_{t=s+1}^{n-r} y_t y_{t+r-s}$

(d_{00}, d_{01}, d_{11}) er min. suff. for (ρ, σ^2) □

Ikke-regulær eksponentiel familie, da $\dim(d_{00}, d_{01}, d_{11}) = 3$ og $\dim(\rho, \sigma^2) = 2$

Egenskaber for regulære eksponentielle familier af orden 1

$$f(y;\theta) = q(y)\exp(\psi(\theta)t(y) - \tau(\theta))$$

bemærk $\frac{\partial}{\partial\theta}\int_y f(y;\theta)\,d\nu(y) = 0$

$$\Rightarrow \int_y \frac{\partial}{\partial\theta} f(y;\theta)\,d\nu(y) = 0$$

$$\int_y f(y;\theta)(\psi'(\theta)t(y) - \tau'(\theta))\,d\nu(y) = 0$$

$$\psi'(\theta)\int_y t(y) f(y;\theta)\,d\nu(y) - \tau'(\theta)\int_y f(y;\theta)\,d\nu(y) = 0$$

$$\psi'(\theta)\,E[t(Y)] - \tau'(\theta) = 0$$

$$E[t(Y)] = \frac{\tau'(\theta)}{\psi'(\theta)}$$

ved to gange diff.

$$\int_y f(y;\theta)\Big(\psi''(\theta)t(y) - \tau''(\theta) + (\psi'(\theta)t(y) - \tau'(\theta))^2\Big)\,d\nu(y) = 0$$

$$\psi''(\theta)\,E[t(Y)] - \tau''(\theta) + (\psi'(\theta))^2 E[(t(Y) - E[t(Y)])^2] = 0$$

$$(\psi'(\theta))^2\,\mathrm{Var}[t(Y)] = \tau''(\theta) - \psi''(\theta)\frac{\tau'(\theta)}{\psi'(\theta)}$$

$$\mathrm{Var}[t(Y)] = \frac{\tau''(\theta)\psi'(\theta) - \psi''(\theta)\tau'(\theta)}{(\psi'(\theta))^3}$$

ved yderligere diff. kan højere ordens momenter tilsvarende bestemmes

Kanonisk parameter

Ved reparametrisering, så $\psi = \psi(\theta)$ vælges som ny parameter, bliver udtrykkene for $E[t(Y)]$ og $Var[t(Y)]$ (og for højere ordens momenter) en hel del simplere.

ψ kaldes den kanoniske parameter

$$h(y;\psi) = q(y) \exp(\psi t(y) - \tau_*(\psi))$$

eks. $Y \sim b(n, \theta)$

$$f(y;\theta) = \binom{n}{y} \exp\left(y \ln \frac{\theta}{1-\theta} - (-n \ln(1-\theta))\right)$$

sæt $\psi = \operatorname{logit} \theta = \ln \frac{\theta}{1-\theta}$

$$\Leftrightarrow \quad 1-\theta = \frac{1}{1+e^{\psi}}$$

$$h(y;\psi) = \binom{n}{y} \exp(y\psi - n \ln(1+e^{\psi}))$$

$\square$