4.2.4 Weitere ML Beispiele

(1) Gleichverteilung

$$f(x) = \frac{1}{b-a} \qquad a \leq x \leq b$$

$$L(\underline{x}; a, b) = \left(\frac{1}{b-a}\right)^n$$

u.d.B. $\quad b \geq \max x_i \quad$ und $\quad a \leq \min x_i$

Das Maximum kann man mit Lagrange Methoden finden:

$$\widehat{a} = \min x_i \quad \text{und} \quad \widehat{b} = \max x_i$$

(2) Negativ Binomial Verteilung

$$p(n) = \binom{n-1}{r-1} p^r q^{n-r}$$

Daten $n_1, n_2, \ldots, n_m$

(a) r bekannt

$$L = \prod \binom{n_i - 1}{r-1} p^r q^{n_i - r}$$

$$\log L = \sum \log \binom{n_i - 1}{r-1} + r \sum \log p + \sum (n_i - r) \log(1-p)$$

$$\frac{\partial \log L}{\partial p} = \frac{rm}{p} - \frac{\sum n_i - rm}{1-p}$$

$$\Rightarrow \widehat{p} = \frac{rm}{\sum n_i}$$

(b) r unbekannt
r muß ganzzählig sein und $1 \leq r \leq \min n_i$

Wir berechnen max L für $r = 1, \ldots, \min n_i$ und nehmen den größten.

4.2.5 Funktionen von Zufallsvariablen und ML Schätzer

$$X \sim f_X(x)$$

$$Y = g(X) \qquad 1-1$$

mit Dichte

$$h_Y(y) = f_X(g^{-1}(y))\frac{\partial g^{-1}}{\partial y}$$

Aus einer Stichprobe haben wir $(y_1, \ldots, y_n)$

$$L(\underline{y}, \underline{\theta}) = \prod h_Y(y_i, \underline{\theta})$$

$$= \prod f_X(g^{-1}(y_i))\frac{\partial g^{-1}}{\partial y_i}$$

$$\log L(\underline{y}, \underline{\theta}) = \sum \log f_X(g^{-1}(y_i)) + \sum \log \frac{\partial g^{-1}}{\partial y_i}$$

$$\frac{\partial log L(\underline{y}, \underline{\theta})}{\partial \theta} = 0 \text{ genau dann wenn}$$

$$\frac{\partial log L(\underline{x}, \underline{\theta})}{\partial \theta} = 0 \quad \text{mit} \quad \{x_i\} = \{(g^{-1}(y_i))\}$$

$\Rightarrow$ ML Schätzer für θ bleiben durch eine Transformation unverändert.

4.2.6 ML Intervallschätzer

Konfidenz Intervalle (KIs)

Bis jetzt haben wir nur Punktschätzer besprochen (d.h. es wird nur einen Schätzwert angegeben. Unsere Unsicherheit um den Wert wird nicht berichtet.) Jetzt führen wir Konfidenzintervalle ein — Intervalle, die mit einer bestimmten Wahrscheinlichkeit den wahren, unbekannten Wert enthalten.

Falls wir die Verteilung unseres Schätzers kennen, können wir vor der Ziehung Vorhersageintervalle berechnen:

$$P_{\widehat{\theta}}(a < \widehat{\theta} < b) = 1 - \alpha$$

z.B. $\quad \bar{X} \sim N(\mu, \sigma^2/n)$

$$P_{\bar{X}}(\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

Diese Wahrscheinlichkeit kann umgeschrieben werden:

$$P_{\bar{X}}(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

aber es handelt sich als Wahrscheinlichkeitsintervall immer noch um $\bar{X}$ und nicht um μ.

Wenn wir den Stichprobenwert $\bar{x}$ für $\bar{X}$ einsetzen, bekommen wir ein sogenanntes Konfidenzintervall, und wir sagen, dass

$$(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}})$$

ein 95% Konfidenzintervall für μ ist.

Wenn wir viele Stichproben ziehen und Konfidenzintervalle aus jeder berechnen, werden wir finden, dass ungefähr 95% der Intervalle den wahren Wert μ enthalten werden.

Falls σ unbekannt ist aber wir $X \sim N(\mu, \sigma^2)$ annehmen dürfen, dann können wir mit $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ arbeiten.

Weitere Beispiele von KIs

(1) Gleichverteilung

$$X \sim G(0, b)$$

Wir beobachten $x = 7$, was ist eine 95% KI für b?

$$P_X(0 < X < 0.95b) = 0.95 \quad \Rightarrow \quad \left(\frac{x}{0.95}, \infty\right)$$

$$P_X(0.05b < X < b) = 0.95 \quad \Rightarrow \quad (x, 20x)$$

$$P_X(0.025b < X < 0.975b) = 0.95 \quad \Rightarrow \quad \left(\frac{x}{0.975}, 40x\right)$$

Meistens versuchen wir, das kurzeste Intervall zu finden. In diesem Fall haben wir die Intervalle

$$(7.37, \infty) \qquad (7, 140) \qquad (7.18, 280)$$

und wir würden das zweite nehmen.

(2) Proportionen — Binomial Verteilung

$$X \sim B(N, p)$$

Für N groß gilt

$$Y = \frac{X}{N} \sim N\left(p, \frac{p(1-p)}{N}\right)$$

und ein 95% Vorhersageintervall für Y wäre

$$P_Y\left(p - 1.96\sqrt{\frac{p(1-p)}{N}} < Y < p + 1.96\sqrt{\frac{p(1-p)}{N}}\right)$$

$$P_Y\left((Y-p)^2 < 1.96^2 \quad \frac{p(1-p)}{N}\right) = 0.95$$

Y und N sind bekannt. Für welche Werte von p gilt

$$(Y-p)^2 < 1.96^2 \quad \frac{p(1-p)}{N} \quad ?$$

Drei Möglichkeiten

(a) Als Approximation für die Varianz setzen wir $p = 0.5$ und

$$\frac{p(1-p)}{N} \leq \frac{1}{4N} \qquad \forall \quad p \in [0,1]$$

$$\Rightarrow P_Y(Y - 1.96\frac{1}{\sqrt{4N}} < p < Y + 1.96\frac{1}{\sqrt{4N}}) \geq 0.95$$

(b) “Plug-in Schätzer”

Wir benutzen $\frac{y(1-y)}{N}$ als Approximation für die Varianz

$$P_Y\left(Y - 1.96\sqrt{\frac{y(1-y)}{N}} < p < Y + 1.96\sqrt{\frac{y(1-y)}{N}}\right)$$

$$\approx 0.95$$

Dieser Schätzer ist besser als (a). Er ist am Rand (p nahe 0 oder 1) schlecht und die Coverage (Überdeckung) ist wegen der Normalapproximation unregelmäßig.

(c) Eine genauere Methode

Lösen wir die quadratische Gleichung

$$(y-p)^2 = 1.96^2 \quad \frac{p(1-p)}{N}$$

um Wurzeln p_1 und p_2 zu finden

$$\Rightarrow P_Y(p_1 < p < p_2) = 0.95$$

(Für eine eingehende Studie der vielen Alternativen, siehe Brown L, Cal, T.T., DasGupta, A. (2001) Interval Estimation for a Binomial Proportion. Statistical Science 16: 101-133)

Beispiele von Binomial KIs

(i) (Die Zeit 13. Februar, 2003 s30)
Paare sind auf Flughäfen und Bahnhöfen beobachtet worden (vermutlich in Deutschland, da ein Bochumer Wissenschaftler die Daten gesammelt hat). Aus 124 küssenden Paaren haben 80 den Kopf beim Küssen nach rechts und 44 den Kopf nach links gedreht.

$$\widehat{p} = 80/124 = 0.645$$

95% KIs

$$(a) \quad \text{Approx } p = 0.5 \quad (0.557, 0.733)$$

$$(b) \quad \text{"Plug-in" } p = \widehat{p} \quad (0.561, 0.729)$$

$$(c) \quad \text{Normal genau} \quad (0.558, 0.724)$$

(ii) Sonntagsfrage in der politischen Meinungsforschung z.B. 1000 Befragte CSU 55%, SPD 25%, FDP 3%

(a) $(0.519, 0.581)$ $(0.219, 0.281)$ $(-0.001, 0.061)$

(b) $(0.519, 0.581)$ $(0.223, 0.277)$ $(0.019, 0.041)$

(c) $(0.519, 0.581)$ $(0.224, 0.278)$ $(0.021, 0.043)$

4.2.7 Definition eines KIs

Sei X eine Statistik mit Verteilungsfunktion

$$F(x, \theta), \quad \theta \in \Theta$$

Eine Abbildung, C, die jedem möglichen Beobachtungsergebnis x eine Menge $C(x) \in \Theta$ zuordnet, heißt ein Konfidenzbereich für θ zum Niveau $100(1-\alpha)\%$ wenn

$$\inf_{\theta \in \Theta} P_\theta(x : C(x) \ni \theta) \geq 1 - \alpha$$

Wenn es ein Pivot (Drehpunkt) gibt, wie bei der Normalverteilung, sind KIs leicht zu berechnen, im allgemeinen nicht.

4.2.8 Assymptotische ML Konfidenz Intervalle

ML Schätzer sind assymptotisch normalverteilt.

Bedingungen

1. Der Parameterraum Ω hat endliche Dimension, ist geschlossen und kompakt und der wahre Wert θ liegt innerhalb Ω.

2. Die Wahrscheinlichkeitsverteilungen für zwei verschiedene Werte von θ sind verschieden.

3. $\frac{\partial}{\partial\theta}$, $\frac{\partial^2}{\partial\theta^2}$, $\frac{\partial^3}{\partial\theta^3}$, von $l(x;\theta)$ existieren fast sicher in der Nachbarschaft des wahren Wertes. Weiterhin in solch einer Nachbarschaft ist $\frac{1}{n}\left|\frac{\partial^3 l}{\partial\theta^3}\right| \leq$ eine Funktion von X deren Erwartungswert existiert.

4. $I(\theta) = E[(\frac{\partial l}{\partial\theta})^2] = -E[\frac{\partial l^2}{\partial\theta^2}]$ ist endlich und > 0 in der Nachbarschaft des wahren Wertes.

Sei θ der für uns interessante Parameter der (glatten) Dichte f. Die Loglikelihoodfunktion ist

$$l(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

für eine u.i.v. Stichprobe der Größe n.
Der ML Schätzer, $\widehat{\theta}$ muß ein stationärer Punkt von $l(\theta)$ sein:

$$l'(\widehat{\theta}) = 0$$

Aus einer Taylor Entwicklung von $l'(\widehat{\theta})$ um den unbekannten wahren Wert θ_0 haben wir

$$l'(\widehat{\theta}) \approx l'(\theta_0) + (\widehat{\theta} - \theta_0) l''(\theta_0)$$

$$\Rightarrow (\widehat{\theta} - \theta_0) \approx -\frac{l'(\theta_0)}{l''(\theta_0)} \qquad (1)$$

Betrachten wir

$$E[l'(\theta)] = \sum_{i=1}^{n} E\left[\frac{\partial}{\partial\theta} \log f(X_i|\theta)\right]$$

$$= \sum_{i=1}^{n} \int \left[\frac{\partial}{\partial\theta} \log f(x|\theta)\right] f(x|\theta) \mathrm{d}x$$

Es gilt

$$\int \left[\frac{\partial}{\partial\theta} \log f(x|\theta)\right] f(x|\theta)\mathrm{d}x = \int \frac{\partial}{\partial\theta} f(x|\theta)\mathrm{d}x \qquad (2)$$

und wenn wir sagen dürfen, dass

$$\int \frac{\partial}{\partial\theta} f(x|\theta)\mathrm{d}x = \frac{\partial}{\partial\theta} \int f(x|\theta)\mathrm{d}x$$

ist

$$\int [\frac{\partial}{\partial\theta} \log f(x|\theta)] f(x|\theta)\mathrm{d}x = 0$$

und

$$E[l'(\theta)] = 0$$

$$V[l'(\theta)] = \sum_{i=1}^{n} E[(\frac{\partial}{\partial\theta} \log f(X_i|\theta))^2]$$

$$= nI(\theta) \quad \text{per Definition}$$

$I(\theta)$ wird die Fisher-Information genannt. (Man findet auch den Ausdruck ”Scorefunktion” für $U_\theta = \frac{\partial}{\partial\theta} \log f(x|\theta)$).

Jetzt betrachten wir den Nenner von (1)

$$l''(\theta) = \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f(X_i|\theta)$$

Wir zeigen, dass

$$E[l''(\theta)] = -nI(\theta)$$

$$E[l''(\theta)] = E\left[\frac{\partial}{\partial \theta} l'(\theta)\right] \neq \frac{\partial}{\partial \theta} E[l'(\theta)]$$

$$E[l'(\theta)] = \int \left[\frac{\partial}{\partial \theta} \log f(x|\theta)\right] f(x|\theta) \mathrm{d}x = 0$$

$\Rightarrow$ (nach Ableitung und Wiederbenutzung von (2))

$$\int \left[\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2}\right] f(x|\theta)\mathrm{d}x + \int \left[\frac{\partial \log f(x|\theta)}{\partial \theta}\right]^2 f(x|\theta)\mathrm{d}x$$

$$= 0$$

$$\Rightarrow E[l''(\theta)] = -V[l'(\theta)]$$

$$= -nI(\theta)$$

Für n groß, nach dem GGZ

$$\frac{1}{n}l''(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f(x_i,\theta_0) \to -I(\theta_0)$$

und daher aus (1)

$$\sqrt{n}(\hat{\theta}-\theta_0) \approx \frac{n^{-1/2}l'(\theta_0)}{I(\theta_0)}$$

und

$$E[\sqrt{n}(\hat{\theta}-\theta_0)] = 0$$

$$V[\sqrt{n}(\hat{\theta}-\theta_0)] \approx \frac{1}{I(\theta_0)}$$

$$V[(\hat{\theta}-\theta_0)] \approx \frac{1}{nI(\theta_0)}$$

$$l'(\theta_0) = \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(x_i|\theta)|_{\theta=\theta_0}$$

ist eine Summe von u.i.v. Zufallsvariablen. Nach dem ZGS

$$\sqrt{nI(\theta_0)}(\hat{\theta}-\theta_0) \to N(0,1)$$

Beispiel (1)

X hat die Dichte

$$f(x) = (\theta + 1)x^{\theta} \qquad 0 \leq x \leq 1$$

eine Beta Verteilung mit $a = \theta + 1$ und $b = 1$

$$E[X] = \frac{\theta + 1}{\theta + 2}$$

$$l(x; \theta) = n \log(\theta + 1) + \theta \sum \log x_i$$

$$\widehat{\theta} = -1 - \frac{n}{\sum \log x_i}$$

$$\frac{\partial}{\partial \theta} \log f = \frac{1}{\theta + 1} + \log x$$

$$\Rightarrow I(\theta) = E[(\frac{1}{\theta + 1} + \log X)^2]$$

$$\frac{\partial^2}{\partial \theta^2} \log f = -\frac{1}{(\theta + 1)^2}$$

$$\Rightarrow I(\theta) = \frac{1}{(\theta + 1)^2}$$

$$\Rightarrow V[\widehat{\theta}] = \frac{(\theta+1)^2}{n}$$

Assymptotisch

$$\widehat{\theta} = (-1 - \frac{n}{\sum \log x_i}) \sim N(\theta, \frac{(\theta+1)^2}{n})$$

Im allgemeinen gilt

$$P(-1.96 < \sqrt{nI(\theta)}(\widehat{\theta} - \theta) < 1.96) = 0.95$$

$$P(|\widehat{\theta} - \theta| < \frac{1.96}{\sqrt{n}}(\theta+1)) = 0.95$$

$$P_{\widehat{\theta}}\left(\frac{\widehat{\theta} - \frac{1.96}{\sqrt{n}}}{1 + \frac{1.96}{\sqrt{n}}} < \theta < \frac{\widehat{\theta} + \frac{1.96}{\sqrt{n}}}{1 - \frac{1.96}{\sqrt{n}}}\right) = 0.95$$

Ein 95% KI für θ wenn wir $\widehat{\theta}$ mit dem beobachteten Wert ersetzen.

Beispiel (2)
$\{X_i\}$ u.i.v. von einer Rayleigh Dichte

$$f(x) = \frac{x}{\theta^2} e^{-\frac{x^2}{2\theta^2}} \qquad x \geq 0$$

(Die Rayleigh Verteilung folgt aus Y, W u.i.v. $\sim N(0, \sigma^2)$. Nach der Transformation in Polarkoordinaten (r, ϕ) hat r eine Rayleigh Verteilung.)

$$l(x;\theta) = -2n \log \theta + \sum \log x - \frac{1}{2\theta^2} \sum x^2$$

$$\frac{\partial l}{\partial \theta} = 0 \quad \Rightarrow \quad -\frac{2n}{\theta} + \frac{1}{\theta^3} \sum x^2 \quad \Rightarrow \quad \widehat{\theta} = \sqrt{\frac{\sum x^2}{2n}}$$

$$\frac{\partial}{\partial \theta} \log f = -\frac{2}{\theta} + \frac{x^2}{\theta^3}$$

$$\frac{\partial^2}{\partial \theta^2} \log f = \frac{2}{\theta^2} - 3\frac{x^2}{\theta^4}$$

$$E[X^2] = 2\theta^2 \quad \Rightarrow \quad I(\theta) = \frac{4}{\theta^2}$$

und assymptotisch

$$\widehat{\theta} = \sqrt{\frac{\sum x^2}{2n}} \sim N\left(\theta, \frac{\theta^2}{4n}\right)$$

Beste Schätzer

Um Resultate über Optimalität zu erzielen, beschränken wir uns auf erwartungstreue Schätzer für reguläre Modelle. Ein einparametriges Modell heißt regulär wenn

1. Θ ist ein offenes Intervall in $\mathbb{R}$

2. Die Likelihoodfunktion ist auf X x Θ strikt positiv und nach θ differenzierbar. Somit existiert die Scorefunktion
$$U_\theta(x) = \frac{\partial}{\partial\theta} \log f(x|\theta)$$

3. Für jedes $\theta \in \Theta$ existiert $V_\theta[U_\theta] = I(\theta)$ und ist nicht 0. Es gilt die Vertauschungsrelation
$$\int \frac{\partial}{\partial\theta} f(x|\theta) \mathrm{d}x = \frac{\partial}{\partial\theta} \int f(x|\theta) \mathrm{d}x$$

Die Cramér-Rao Ungleichung

Gegeben sei ein reguläres statistisches Modell $P_X(x;\theta)$, $\theta \in \Theta$, ein Intervall. Dann gilt für jeden erwartungstreuen Schätzer, $\tilde{\theta}$ von θ, dass

$$Var_\theta[\tilde{\theta}] \geq \frac{1}{I(\theta)}$$

Beweis

$$E[\tilde{\theta}] = \theta$$

und für die Scorefunktion $U_\theta = \frac{\partial}{\partial\theta} \log f(x|\theta)$ gilt

$$E[U_\theta] = 0$$

Deshalb haben wir

$$Cov[\tilde{\theta}, U_\theta] = E[\tilde{\theta}U_\theta]$$

$$E[\tilde{\theta}U_\theta] = \int_X \tilde{\theta}\frac{\partial}{\partial\theta} \log f(x|\theta) f(x|\theta) \mathrm{d}x$$

$$= \int_X \tilde{\theta}\frac{\partial}{\partial\theta} f(x|\theta) \mathrm{d}x$$

$$= \frac{\partial}{\partial\theta}\int_X \tilde{\theta} f(x|\theta)\mathrm{d}x = \frac{\partial}{\partial\theta}E[\tilde{\theta}] = 1$$

Betrachten wir die Varianz von der Funktion $(\tilde{\theta} - U_\theta/I(\theta))$

$$0 \leq V[\tilde{\theta} - U_\theta/I(\theta)]$$

$$= V[\tilde{\theta}] - 2Cov[\tilde{\theta}, U_\theta]/I(\theta) + V[U_\theta]/I(\theta)^2$$

$$= V[\tilde{\theta}] - 1/I(\theta)$$

Q.E.D.

Cramér-Rao liefert uns eine untere Grenze für die Varianz von erwartungstreuen Schätzern für reguläre Modelle (z.B. es gilt nicht für das Beispiel mit der Gleichverteilung). Es sagt uns nicht, wie wir einen optimalen Schätzer finden. Deshalb sprechen wir von der Effizienz eines Schätzers:

Gegeben einen erwartungstreuen Schätzer, $\tilde{\theta}$, eines Parameters für ein reguläres Modell, dann ist die Effizienz von $\tilde{\theta}$

$$eff(\tilde{\theta}) = \frac{1}{I(\theta)Var[\tilde{\theta}]}$$