4. Modellwahlstrategien

Oft ist ein geeignetes Modell für die Daten nicht genau bekannt.

Modellwahlstrategien 4.1:

a) Anpassung eines kleinen / sparsamen Modells mit nur in jedem Falle notwendigen Strukturen sowie Analyse der Residuen (Modelldiagnose): Aufnahme zusätzlicher Strukturen, die ebenfalls notwendig erscheinen, zwecks schrittweiser Verbesserung des Modells.
   Hierbei sind z.B. Grafiken der Residuen nützlich.

b) Anpassung eines großen / allgemeinen Modells und Durchführung von Parametertests zur Überprüfung, ob das Modell verkleinert werden kann.
   Hierbei sind Bem. 3.15 und Beh. 3.16 nützlich.

c) Anpassung (vieler) verschiedener Modelle und Verwendung eines Modellwahlkriteriums zur Auswahl eines "besten" Modells.

Wir behandeln nachfolgend die Strategie c).

Definition 4.2:

Gegeben zwei Dichten f und g bezüglich des gleichen dominierenden Maßes μ nennen wir

$$I(f,g) = E_f \ln \frac{f}{g}$$

die gerichtete Kullback-Leibler-Information von g bezüglich f. Ist f die wahre Dichte der Beobachtungen und g die Dichte der Beobachtungen unter einem gegebenen statistischen Modell, so sollte g die KL-Information über neue Beobachtungen $Y^\star$ maximieren, nämlich

$$I(f,g) = E_f \ln(f(Y^\star)) - E_f \ln(g(Y^\star)) .$$

Enthält g unbekannte Parameter, $g = g_\theta$, so nutzen wir den ML-Schätzer $\hat{\theta}(Y)$ von θ gegeben Beobachtungen Y und streben die Maximierung an von

$$E_f I(f, g_{\hat{\theta}}) = E_f \ln(f(Y^\star)) - E_{f(Y)} E_{f(Y^\star)} \ln(g_{\hat{\theta}(Y)}(Y^\star)) .$$

Hierbei ist $E_f \ln(f(Y^\star))$ unabhängig vom betrachteten Modell, so dass wir den Term

$$AI = -2 E_{f(Y)} E_{f(Y^\star)} \ln(g_{\hat{\theta}(Y)}(Y^\star))$$

die Akaike Information nennen.

Das Akaike Informations-Kriterium ist sodann

$$AIC = -2 \ln(g_{\hat{\theta}(Y)}(Y^\star)) + 2d,$$

wobei d die Anzahl der freien Parameter in θ unter dem betrachteten Modell ist.

Beispiel 4.7:

$m = 10$ Herzinfarktpatienten wurde das Medikament Cadralazine verabreicht und die Konzentration des Wirkstoffs im Blut nach 2, 4, 6, 8, 10 und 24 Stunden gemessen ($n = 6$).
Die Konzentration C_t zum Zeitpunkt t in Abhängigkeit von der Dosis d wird beschrieben durch das Modell

$$C_t = (d/v)\exp(-kt),$$

wobei die unbekannten Parameter k und v die Eliminationsrate sowie ein Skalierungsfaktor sind, der mit dem Blutvolumen des Patienten identifiziert werden kann.
Nach Logarithmieren erhalten wir ein lineares Modell für

$$Y_{it} = \ln(C_{it}) - \ln(d_{it}) = -\ln(v_i) - k_i t_{ij} + \epsilon_{ij} \ ,$$

mit Y_{it} die Beobachtung für Individuum i zum Zeitpunkt t.

Modell		Population			Individuen		
Intercept	Steigung	AIC	d	Penalty	CAIC	d	Penalty
gemeinsam	gemeinsam	122.3	3	3.2	122.3	3	3.2
fest	gemeinsam	–	–	–	91.2	12	15.3
gemeinsam	fest	–	–	–	-7.8	12	15.3
fest	fest	–	–	–	-22.8	21	33.2
zufällig	gemeinsam	100.4	4	4.4	85.6	10.8	11.4
gemeinsam	zufällig	22.3	4	4.4	-12.0	12.0	12.6
zufällig	zufällig	12.6	6	6.8	-42.3	19.2	20.2