Statistische Verfahren
FSU Jena - SS 2010
Notizen

Stilianos Louca

26. Juli 2010

Inhaltsverzeichnis

1 Vorwort

1.1 Was dies ist

Hierbei handelt es sich um persönliche Notizen des Stoffes, der im SS 2010 an der FSU Jena von Dr. J. Schumacher im Fach *Statistische Verfahren* gelehrt wurde.

1.2 Verbesserungen

Ich werde immer mal dieses Skript verbessern bzw. erweitern. Im Falle von Fehlern, ist mir Bescheid zu sagen das beste was du machen kannst, da so alle davon profitieren können. Wissen ist das einzige auf dieser Welt das vom Teilen mehr wird!
Ich bin zu erreichen unter *stilianos.louca@apfel.uni-jena.de*, ohne das *Obst*.

2 Grundbegriffe

2.0.1 Definition: Teststatistik

Seien E, T messbare Räume, $(\mathcal{P}_\theta)_{\theta\in\Theta}$ eine Familie von Wahrscheinlichkeitsmaßen auf E und $\mathcal{T} : E \to T$ eine messbare Abbildung. Dann heißt $\mathcal{T}$ *Teststatistik* bzw. *Stichprobenfunktion* zum stochastischen Modell $(E, \mathcal{A}, \mathcal{P}_\theta : \theta \in \Theta)$.

Sei nun Y eine E-wertige, gemäß $\mathcal{P}_\theta$ verteilte Zufallsgröße und H_o eine Hypothese über die Natur von θ. Dann hängt zwar die Verteilung von $\mathcal{T}(Y)$ vom konkreten (unbekannten) Parameterwert θ ab, doch lassen sich unter Annahme von H_o Aussagen über die Verteilung von $\mathcal{T}(Y)$ machen. Zu festgelegtem *Signifikanzniveau* $0 < \alpha \ll 1$ (typischerweise $\alpha \ll 1$) sei nun $T_\alpha \subseteq T$ eine feste messbare Teilmenge (*Annahmebereich*) so dass unter Annahme von H_0 gilt: $\mathcal{P}(\mathcal{T}(Y) \in T_\alpha) = (1-\alpha)$. Wird nun eine konkrete Stichprobe $y \in E$ zu Y gezogen die dazu führt dass $\mathcal{T}(y) \notin T_\alpha$, so ist die Hypothese H_0 abzulehnen. Sollte H_o doch stimmen, so ist die Wahrscheinlichkeit solch einer fälschlichen Ablehnung (*Fehler 1. Art*[1]) auf jeden Fall kleiner als α. Solch ein Test heißt *Hypothesentest* der Hypothese H_0 zum Signifikanzniveau α.

Ist die Teststatistik $\mathcal{T}(Y)$ unter der Nullhypothese t, χ^2 bzw. F verteilt, so heißt der Test jeweils *t-Test*, *χ^2-Test* bzw. *F-Test*.

Ist nun $\mathcal{T}_n : E \to T,\ n \in \mathbb{N}$ eine Folge von Teststatistiken so dass unter Annahme der Nullhypothese $\mathcal{P}(\mathcal{T}_n(Y) \in T_\alpha) \overset{n\to\infty}{\longrightarrow} (1-\alpha)$, so kann obiger Test für ein genügend großes $n \in \mathbb{N}$ durchgeführt werden. Das tatsächlich zugrundeliegende Signifikanzniveau des Bereichs T_α ist nunmehr nicht bekannt, doch konvergiert es mit wachsendem n gegen α. Solch ein Test heißt *asymptotischer Hypothesentest* der Hypothese H_0 und T_n *asymptotische Teststatistik*.

2.0.2 Definition: Konfidenzintervall

Sei $(\mathcal{P}_\theta)_{\theta\in\Theta}$ eine Familie von Wahrscheinlichkeitsmaßen auf dem messbaren Raum $(E, \mathcal{A})$, $\gamma : \Theta \to \Gamma$ eine *zu ermittelnde Kenngröße* des Parameters θ und $0 < \alpha < 1$. Eine Abbildung $K_\alpha : E \to \mathscr{P}(\Gamma)$ heißt *Konfidenzintervall* von γ zum *Signifikanzniveau* α, falls für jedes $\theta \in \Theta$ und gemäß $\mathcal{P}_\theta$ verteilte, E-wertige Zufallsgröße Y gilt

$$\mathcal{P}\left(\gamma(\theta) \in K_\alpha(Y)\right) = 1 - \alpha \quad . \tag{2.1}$$

Die Stichprobenfunktion K_α ordnet jeder Stichprobe (Realisierung) $y \in E$ eine Menge zu. Für konkreten Parameter θ, enthält diese mit einer Sicherheit von $(1-\alpha)$ den realen Wert $\gamma(\theta)$. **Beachte:** Der Parameter θ bzw. die Kenngröße $\gamma(\theta)$ sind keine Zufallsgrößen. Insbesondere ist bei konkreter Realisierung $y \in E$, $(1-\alpha)$ **nicht** die Wahrscheinlichkeit mit der $\gamma(\theta)$ in K_α liegt!

[1] Eine fälschliche Annahme der Nullhypothese heißt *Fehler 2. Art*.

Eine Folge von Abbildungen $K_{n,\alpha} : E \to \mathscr{P}(\Gamma),\ n \in \mathbb{N}$ heißt *asymptotisches Konfidenzintervall* von γ zum Signifikanzniveau α, falls für jedes $\theta \in \Theta$ und gemäß $\mathcal{P}_\theta$ verteilte, E-wertige Zufallsgröße Y gilt

$$\lim_{n\to\infty} \mathcal{P}\left(\gamma(\theta) \in K_{n,\alpha}(Y)\right) = 1 - \alpha \quad . \tag{2.2}$$

3 Verteilungsfamilien

3.1 Exponentialfamilien

3.1.1 Definition: Exponentialfamilie

Eine durch $\theta \in \Theta$ parametrisierte Familie E-wertige Zufallsgrößen Y_θ gehört zu einer *Exponentialfamilie*, falls ihre Dichten (bzgl. eines zugrundeliegenden Maßes auf E) die Form

$$f_\theta(y) = h(y) \cdot \exp\left[\boldsymbol{\eta}(\theta) \cdot \mathbf{T}(y) - C(\theta)\right] \tag{3.1}$$

annehmen, wobei $\boldsymbol{\eta} : \Theta \to \mathbb{R}^n$ beliebig und $\mathbf{T} : E \to \mathbb{R}^n$ messbar. Die Komponenten $\eta^i(\theta)$ heißen *natürliche Parameter*, die T_i *natürliche Statistiken* der Familie.

Die Exponentialfamilie befindet sich in *natürlicher Parametrisierung*, wenn sie durch $\boldsymbol{\eta}$ Parametrisiert die Darstellung

$$f_{\boldsymbol{\eta}}(y) = h(y) \cdot \exp\left[\boldsymbol{\eta} \cdot \mathbf{T}(y) - A(\boldsymbol{\eta})\right] \tag{3.2}$$

besitzt. Die Menge

$$\mathcal{H} := \left\{\boldsymbol{\eta} \in \mathbb{R}^n : h \cdot e^{\boldsymbol{\eta}\mathbf{T}} \in L_1(E)\right\} \tag{3.3}$$

heißt *natürlicher Parameterraum* der Familie.

3.1.2 Definition: Konjugierte Exponentialfamilie

Sei $\mathbf{Y}$ eine $\mathbb{R}^n$-wertige Zufallsgröße mit Dichte h (bzgl. eines zugrundeliegenden Maßes) und momenterzeugender Funktion

$$m(\boldsymbol{\theta}) := \mathbb{E}e^{\mathbf{Y}\boldsymbol{\theta}} \quad . \tag{3.4}$$

Dann heißt die Familie von Verteilungen auf $\mathbb{R}^n$, parametrisiert durch $\boldsymbol{\theta} \in \{\boldsymbol{\theta} \in \mathbb{R}^n : m(\boldsymbol{\theta}) < \infty\}$, mit Dichten

$$f_{\boldsymbol{\theta}}(\mathbf{y}) := h(\mathbf{y}) \cdot \exp\left[\boldsymbol{\theta} \cdot \mathbf{y} - A(\boldsymbol{\theta})\right] \quad , \quad e^{A(\boldsymbol{\theta})} := m(\boldsymbol{\theta}) \tag{3.5}$$

zu h konjugierte Exponentialfamilie. Sie besitzt die momenterzeugenden Funkionen

$$m_{\boldsymbol{\theta}}(\mathbf{k}) := \int e^{\mathbf{k}\mathbf{y}} f(\mathbf{y} \mid \boldsymbol{\theta})\ d\mathbf{y} = \frac{m(\mathbf{k}+\boldsymbol{\theta})}{m(\boldsymbol{\theta})} \quad . \tag{3.6}$$

Eine gemäß $f(\cdot \mid \boldsymbol{\theta})$ verteilte Zufallsgröße $\mathbf{Z}_{\boldsymbol{\theta}}$ besitzt die Momente

$$\mathbb{E}\mathbf{Z}_{\boldsymbol{\theta}} = \left.\frac{\partial A}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}} \quad , \quad \operatorname{Cov}\mathbf{Z}_{\boldsymbol{\theta}} = \left.\frac{\partial^2 A}{\partial \boldsymbol{\theta}^2}\right|_{\boldsymbol{\theta}} \tag{3.7}$$

3.2 Exponential-Dispersionsfamilien

3.2.1 Definition: Exponential-Dispersionsfamilie

Eine durch $\boldsymbol{\theta} \in H \subseteq \mathbb{R}^n$ und $\varphi \in \Phi \subseteq \mathbb{R} \setminus \{0\}$ parametrisierte Familie $\mathbb{R}^n$-wertiger Zufallsgrößen $\mathbf{Y}_{\boldsymbol{\theta},\varphi}$ gehört einer *Exponential-Dispersionsfamilie* (ED) an, falls sich ihre Dichten (bzgl. eines zugrundeliegenden Maßes) gemäß

$$f_{\boldsymbol{\theta},\varphi}(\mathbf{y}) = \exp\left[\frac{1}{\varphi}\left(\mathbf{y}\boldsymbol{\theta} - A(\boldsymbol{\theta})\right) - C(\mathbf{y},\varphi)\right] \tag{3.8}$$

darstellen lassen. Es sei angenommen dass $A : H \to \mathbb{R}$ 2 mal stetig differenzierbar und $\partial_{\boldsymbol{\theta}} A$ bijektiv ist.

Dabei heißt φ *Dispersionsparameter*, die θ^i *natürliche Parameter* der Familie. Deren Momenterzeugenden Funktionen $m_{\boldsymbol{\theta},\varphi}$ sind gegeben durch

$$m_{\boldsymbol{\theta},\varphi}(\mathbf{k}) = \exp\left[\frac{1}{\varphi}\big[A(\boldsymbol{\theta}+\varphi\mathbf{k}) - A(\boldsymbol{\theta})\big]\right] \tag{3.9}$$

Deren ersten beiden Momente sind gegeben durch

$$\mathbb{E}\mathbf{Y}_{\boldsymbol{\theta},\varphi} = \left.\frac{\partial A}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}} \quad , \quad \operatorname{Cov}\mathbf{Y}_{\boldsymbol{\theta},\varphi} = \varphi \cdot \left.\frac{\partial^2 A}{\partial \boldsymbol{\theta}^2}\right|_{\boldsymbol{\theta}} \tag{3.10}$$

Beispiele:

(i) Die **Normalverteilungsfamilie** mit Erwartungswert μ und Varianz σ^2 gehört zur Exponential-Dispersionsfamilie mit $\theta := \mu,\ \varphi := \sigma^2$ und Dichten::

$$\mathcal{N}_{\mu,\sigma^2}(y) = \exp\left\{\frac{1}{\varphi}\Big[y\theta - \underbrace{\frac{\theta^2}{2}}_{A(\theta)}\Big] - \underbrace{\Big[\frac{y^2}{2\varphi} + \frac{1}{2}\ln(2\pi\varphi)\Big]}_{C(y,\varphi)}\right\} \tag{3.11}$$

(ii) Die **Poissonverteilungsfamilie** mit Parameter λ gehört zur Exponential-Dispersionsfamilie mit $\theta := \ln\lambda,\ \varphi := 1$ und Dichten::

$$\text{Poisson}_\lambda(y) = \exp\left\{\frac{1}{\varphi}\Big[y\theta - \underbrace{e^\theta}_{A(\theta)}\Big] - \underbrace{\ln y!}_{C(y)}\right\} \tag{3.12}$$

(iii) Die **Γ-Verteilungsfamilie** mit Erwartungswert as und Varianz as^2 gehört zur Exponential-Dispersionsfamilie mit $\theta := -\frac{1}{as},\ \varphi := \frac{1}{a}$ und Dichten::

$$\Gamma_{a,s}(y) = \exp\left\{\frac{1}{\varphi}\Big[y\theta + \underbrace{\ln(-\theta)}_{-A(\theta)}\Big] - \underbrace{\big[\ln\Gamma(a) - a\ln a - (a-1)\ln y\big]}_{C(y,\varphi)}\right\} \tag{3.13}$$

(iv) Die **Binomialverteilungsfamilie** mit N Versuchen (fest) und Erfolgswahrscheinlichkeit p (variabel) gehört zur Exponential-Dispersionsfamilie mit $\theta := \ln\frac{p}{1-p},\ \varphi := 1$ und Dichten:

$$\mathcal{B}_{N,p}(y) = \exp\left\{\frac{1}{\varphi}\Big[y\theta - \underbrace{N\ln(1+e^\theta)}_{A(\theta)}\Big] + \underbrace{\ln\binom{N}{y}}_{-C(y)}\right\} \tag{3.14}$$

4 Maximum-Likelihood Schätzungen

4.1 Modelle ohne Kovariablen

4.1.1 Definition: Maximum-Likelihood Schätzer

Gegeben sei eine durch $\theta \in \Theta$ parametrisierte Familie von Verteilungen auf dem messbaren Raum E, mit Dichten f_θ bzgl. eines zugrundeliegenden Maßes. Zu gegebenen Realisierungen $y_1, .., y_n \in E$ heißt

$$\mathcal{L}(\theta \mid y_1, .., y_n) := \prod_{i=1}^{n} f_\theta(y_i) \tag{4.1}$$

Likelihood-Funktion der Familie. Sie stellt die Wahrscheinlichkeitsdichte, n unabhängig, gemäß f_θ verteilter, E-wertiger Zufallsgrößen, an der Stelle $(y_1, .., y_n) \in \bigotimes_{i=1}^n E$ dar.

Der (zu den konkreten $y_1, .., y_n$ gehörige) *Maximum-Likelihood Schätzer* (ML-Schätzer) $\widehat{\theta}$ ist nun definiert (falls eindeutig & existent) über

$$\mathcal{L}(\widehat{\theta} \mid y_1, .., y_n) \stackrel{!}{=} \sup_{\theta \in \Theta} \mathcal{L}(\theta \mid y_1, .., y_n) \quad . \tag{4.2}$$

Die Abbildung

$$l(\theta \mid y_1, .., y_n) := \ln \mathcal{L}(\theta \mid y_1, .., y_n) = \sum_{i=1}^n \ln f_\theta(y_i) \tag{4.3}$$

heißt *Loglikelihood Funktion*. Die Bestimmung des ML-Schätzers $\widehat{\theta}$ entspricht also der Maximierung der Loglikelihood-Funktion.

Bemerkungen:

(i) Als Funktion der konkreten $y_1, .., y_n \in E$, kann der Schätzer $\widehat{\theta}(y_1, .., y_n)$ auch als Zufallsgröße interpretiert werden (falls messbar), wenn die $y_1, .., y_n$ durch die gemäß f_θ verteilten $Y_1, ., Y_n$ ersetzt werden. Seine Verteilung hängt dann natürlich von der Verteilung der $Y_1, .., Y_n$ ab.

(ii) Sind $Y_1, .., Y_n \sim f_{\theta_r}$ unabhängige Zufallsgrößen, so stellt $\mathcal{L}(\theta \mid Y_1, .., Y_n)$ für festes θ eine Zufallsgröße dar und es gilt

$$\mathbb{E}\mathcal{L}(\theta_r \mid Y_1, .., Y_n) \geq \mathbb{E}\mathcal{L}(\theta \mid Y_1, .., Y_n) \quad \forall\ \theta \in \Theta \tag{4.4}$$

das heißt die Likelihood Funktion besitzt maximalen Erwartungswert für den *wahren* Wert θ_r.

4.1.2 Definition: Score-Gleichung

Sei $f_{\boldsymbol{\theta}}$ eine nach $\boldsymbol{\theta} \in H \subseteq \mathbb{R}^n$ parametrisierte Familie von Verteilungsdichten auf dem messbaren Raum E mit Loglikelihood-Funktion l gemäß def. 4.1.1. Dann heißt

$$S(\boldsymbol{\theta} \mid y_1, .., y_n) := \frac{\partial l}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta} \mid y_1, .., y_n) \tag{4.5}$$

(falls existent) *Score Funktion* der Familie. Ihr Verlauf hängt von den konkreten $y_1, .., y_n$ ab. Im günstigen Fall ergibt sich der ML-Schätzer $\widehat{\theta}$ zu gegebenen $y_1, .., y_n$ als Lösung der *Score-Gleichung*

$$S(\widehat{\boldsymbol{\theta}} \mid y_1, .., y_n) = 0 \quad . \tag{4.6}$$

Ersetzt man die konkreten Realisierungen $y_1, .., y_n$ durch die unabhängigen, gemäß $f_{\boldsymbol{\theta}}$ verteilten, E-wertigen Zufallsgrößen $Y_1, .., Y_n$, so heißt (4.5), ausgewertet an der Stelle $\boldsymbol{\theta}$:

$$S(\boldsymbol{\theta} \mid Y_1, .., Y_n) \tag{4.7}$$

Score Statistik. Ihre Verteilung hängt nur von $\boldsymbol{\theta}$ ab.

Bemerkung: Es lässt sich zeigen, dass dann der Erwartungswert von $S(\boldsymbol{\theta} \mid Y_1, .., Y_n)$ (unter geeigneten Regularitätsvoraussetzungen an $\mathcal{L}$) gegeben ist durch

$$\mathbb{E}_{\boldsymbol{\theta}} S(\boldsymbol{\theta} \mid Y_1, .., Y_n) = 0 \tag{4.8}$$

4.1.3 Definition: Fisher Information

Sei $f_{\boldsymbol{\theta}}$ eine nach $\boldsymbol{\theta} \in H \subseteq \mathbb{R}^n$ parametrisierte Familie von Verteilungsdichten auf E mit Loglikelihood-Funktion l gemäß def. 4.1.1. Dann heißt für konkrete $y_1, .., y_n \in E$ die Matrix (falls existent)

$$\mathcal{I}_n(\boldsymbol{\theta} \mid y_1, .., y_n) := -\frac{\partial^2}{\partial \boldsymbol{\theta}^2} l(\boldsymbol{\theta} \mid y_1, .., y_n) \tag{4.9}$$

Fisher Information. Sie hängt sowohl von $\boldsymbol{\theta}$ als auch den konkreten $y_1, .., y_n$ ab. Wird sie an der Stelle $\widehat{\boldsymbol{\theta}}(y_1, .., y_n)$ ausgewertet, ist sie ein Maß für die *Güte* des ML-Schätzers $\widehat{\boldsymbol{\theta}}$ und heißt *beobachtete Fisher Information*.

Seien nun $Y_1, .., Y_n$ unabhängig, gemäß $f_{\boldsymbol{\theta}}$ verteilt, dann geht (4.9) über in

$$\mathcal{I}_n(\boldsymbol{\theta}) := -\frac{\partial^2}{\partial \boldsymbol{\theta}^2} l(\boldsymbol{\theta} \mid Y_1, .., Y_n) \quad . \tag{4.10}$$

Falls messbar, stellt sie (für konkretes $\boldsymbol{\theta}$) eine Zufallsgröße dar, deren Verteilung nur noch von $\boldsymbol{\theta}$ und n abhängt. Deren Erwartungswert

$$\mathcal{J}_n(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\theta}} \mathcal{I}_n(\boldsymbol{\theta}) \tag{4.11}$$

heißt *erwartete Fisher Information*, und beschreibt die erwartete *Güte*[2] der ML-Schätzung $\widehat{\boldsymbol{\theta}}$ als Funktion vom tatsächlichen, unbekannten $\boldsymbol{\theta}$.

Bemerkungen: Unter geeigneten Regularitätsvoraussetzungen an $\mathcal{L}$ gilt[1, 2]

(i) $\mathcal{J}_n(\boldsymbol{\theta}) = \mathrm{Cov}\left\{S(\boldsymbol{\theta} \mid Y_1, .., Y_n)\right\}$

(ii) $\mathcal{J}_n(\boldsymbol{\theta}) = n \cdot \mathcal{J}_1(\boldsymbol{\theta})$

4.1.4 Konsistenz des ML-Schätzers in ED-Familien

Sei $f_{\boldsymbol{\theta},\varphi}$ eine Exponential-Dispersionsfamilie auf $\mathbb{R}^m$ (vgl. def. 3.2.1) mit 2 mal stetig differenzierbarem A und erwarteter Fisher Information $\mathcal{J}_1$. Der natürliche Parameterraum $\Theta \subseteq \mathbb{R}^m$ der Familie sei eine offene Teilmenge des $\mathbb{R}^m$. Seien $\mathbf{Y}_1, \mathbf{Y}_2, ..$ unabhängig, gemäß $f_{\boldsymbol{\theta},\varphi}$ verteilte Zufallsgrößen, dazu der ML-Schätzer $\widehat{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_n(\mathbf{Y}_1, .., \mathbf{Y}_n)$ abhängig von den ersten n Realisierungen. Dann gilt:

$$\lim_{n\to\infty} \mathcal{P}\left(\widehat{\boldsymbol{\theta}}_n \text{ existiert als Lösung der Score-Gleichung}\right) = 1 \tag{4.12}$$

$$\sqrt{n} \cdot \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \overset{n\to\infty}{\underset{d}{\longrightarrow}} \mathcal{N}_{0,\mathcal{J}_1^{-1}(\boldsymbol{\theta})} \tag{4.13}$$

Beachte dass $\mathcal{J}_1(\boldsymbol{\theta}) = \mathrm{Cov}(S(\boldsymbol{\theta} \mid \mathbf{Y}_1)) = \mathrm{Cov}(\mathbf{Y}_1)$ positiv-definit und daher invertierbar ist.

4.2 Modelle mit Kovariablen

Man betrachte eine E-wertige Beobachtungsgröße Y (*Zielgröße*), deren genaue Verteilung von der (gegebenfalls mehrdimensionalen) unabhängigen Variablen $x \in \mathcal{X}$ (*Kovariable*, *Einflussgrößen*[3]) abhängt[4]. Deren Dichte (bzgl. eines zugrundeliegenden Maßes) gehöre jedoch stets zur, durch $\theta \in \Theta$ parametrisierten, Familie $(f_\theta)_{\theta\in\Theta}$, so dass letztendlich Y gemäß $f_{\theta=\zeta(x)}$ verteilt ist, sprich, mit θ als Funktion der Kovariablen x.

[2] Siehe zum Beispiel Abschnitt 4.2.9.

[3] Wobei sich hier die Bezeichnung *Kovariable* auf die zunächst abstrakte Variable x, die Bezeichnung *Einflussgröße* auf jede einzelne Komponente x^i von x beziehe.

[4] Z.B. Niederschlag Y in Abhängigkeit von Temperatur x. Es sei zunächst keinerlei Voraussetzung an die Indexmenge $\mathcal{X}$ gesetzt.

Ziel ist es nun, den fundamentalen Zusammenhang $\theta = \zeta(x)$ zu bestimmen[5]. Praktisch ist dies nur möglich, durch eine Einschränkung auf eine parametrisierte Familie $(\zeta_\beta)_{\beta \in B}$, wodurch sich das Problem auf die Schätzung des *unbekannten Parameters* (*Regressionsparameters*) $\beta \in B$ reduziert.

Insbesondere hängt der Erwartungswert von Y nun, gemäß einer durch β festgelegten Beziehung, von den kovariablen x ab. Dieser Zusammenhang

$$\mu_\beta(x) := \mathbb{E}Y = \mathbb{E}f_{\theta=\zeta_\beta(x)} \tag{4.14}$$

heißt *deterministischer Teil* des Modells, die betrachtete Familie (f_θ) *stochastischer Teil*. Falls $\mathcal{X} \subseteq \mathbb{R}^k$, so heißt k *Komplexität* des Modells.

Ziel der so genannten *Regressionsanalyse* ist es, β anhand gegebener Stichproben[6] zu schätzen. Letztere sollten dabei natürlich einen Umfang besitzen, der wesentlich größer als die Dimension von β bzw. x ist.

4.2.1 Definition: Likelihood Funktion und ML Schätzer

Betrachten das Modell mit stochastischem Teil $(f_\theta)_{\theta \in \Theta}$, mit Kovariablen $x \in \mathcal{X}$ und unbekanntem Parameter $\boldsymbol{\beta} \in B \subseteq \mathbb{R}^k$. Gegeben seien die konkreten Parameterwerte $x_1, .., x_n \in \mathcal{X}$, dazu die Realisierungen $y_1, .., y_n \in E \in \mathbb{R}^m$. Dann heißt

$$\mathcal{L}(\boldsymbol{\beta} \mid y_1, .., y_n) := \prod_{i=1}^{n} f_{\theta=\zeta_{\boldsymbol{\beta}}(x_i)}(y_i) \tag{4.15}$$

Likelihood Funktion und

$$l(\boldsymbol{\beta} \mid y_1, .., y_n) := \ln \mathcal{L}(\boldsymbol{\beta} \mid y_1, .., y_n) = \sum_{i=1}^{n} \ln f_{\theta=\zeta_{\boldsymbol{\beta}}(x_i)}(y_i) \tag{4.16}$$

Loglikelihood Funktion des Modells. Erstere stellt die Wahrscheinlichkeitsdichte eines Zufallsvektors $(Y_1, .., Y_n)$ aus unabhängigen Komponenten $Y_i \sim f_{\theta=\zeta_\beta(x_i)}$ an der Stelle $(y_1, .., y_n)$ dar. Die Abbildung

$$S(\boldsymbol{\beta} \mid y_1, .., y_n) := \frac{\partial l}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta} \mid y_1, .., y_n) \tag{4.17}$$

(falls existent) heißt *Score Funktion*. Der Maximierungswert $\widehat{\boldsymbol{\beta}}$ von $l(\cdot \mid y_1, .., y_n)$ (falls existent & eindeutig) heißt *Maximum-Likelihood Schätzer*. Im günstigen Fall ergibt dieser sich als Lösung der *Score Gleichung* $S(\widehat{\boldsymbol{\beta}} \mid y_1, .., y_n) \overset{!}{=} 0$. Mit Hilfe des Schätzers $\widehat{\boldsymbol{\beta}}$ lässt sich ein Schätzer $\widehat{y}_{n+i} := \mathbb{E}f_{\theta=\zeta_{\widehat{\boldsymbol{\beta}}}(x_i)} = \mu_{\widehat{\boldsymbol{\beta}}}(x_i)$ für *zukünftige Messungen* zum Kovariablenwert x_i angeben (*Punktprognose*). Die Quadratsumme

$$\mathrm{RSS} := \sum_{i=1}^{n} (y_i - \widehat{y}_{n+i})^2 \tag{4.18}$$

heißt *Residuenquadratsumme* und hängt von der konkreten Stichprobe $y_1, .., y_n$ ab.

Die Matrix

$$\mathcal{I}(\boldsymbol{\beta} \mid y_1, .., y_n) := -\frac{\partial^2 l}{\partial \boldsymbol{\beta}^2} l(\boldsymbol{\beta} \mid y_1, .., y_n) \tag{4.19}$$

heißt *Fisher Information*. Ausgewertet bei $\widehat{\boldsymbol{\beta}}(y_1, .., y_n)$ heißt sie *beobachtete Fisher Information*.

Seien nun $Y_i \sim f_{\theta=\zeta_{\boldsymbol{\beta}}(x_i)},\ i = 1, .., n$ unabhängig verteilt. Dann heißt der Erwartungswert

$$\mathcal{J}(\boldsymbol{\beta}) := \mathbb{E}_{\boldsymbol{\beta}} \mathcal{I}(\boldsymbol{\beta} \mid Y_1, .., Y_n) = \mathrm{Cov}_{\boldsymbol{\beta}} \{S(\boldsymbol{\beta} \mid Y_1, .., Y_n)\} \tag{4.20}$$

erwartete Fisher Information.

[5] Ideal wäre natürlich auch eine Erkenntnis über die *Ursache* dieses Zusammenhangs.
[6] Bestehend aus n Realisierungen $y_1, .., y_n \in E$ jeweils zu den Kovariablenwerten $x_1, .., x_n \in \mathcal{X}$.

4.2.2 Definition: Verallgemeinertes Lineares Modell (GLM)

Sei $(f_{\theta,\varphi})$, $\theta \in H \subseteq \mathbb{R}$, $\varphi \in \Phi \subseteq \mathbb{R} \setminus \{0\}$ eine Exponential-Dispersionsfamilie von Verteilungsdichten auf $E \subseteq \mathbb{R}$ (bzgl. eines zugrundeliegenden Maßes) und $h : \mathbb{R} \to \mathbb{R}$ ein $\mathcal{C}^2$ Diffeomorphismus[7]. Dann heißt das Model mit kovariablen $\mathbf{x} \in \mathcal{X} \subseteq \{1\} \times \mathbb{R}^{k-1}$, unbekannten Parametern $\boldsymbol{\beta} \in B \subseteq \mathbb{R}^k$ und $\varphi \in \Phi \subseteq \mathbb{R} \setminus \{0\}$, stochastischen Teil $(f_{\theta,\varphi})$ und deterministischen Teil

$$\mu_{\boldsymbol{\beta}}(\mathbf{x}) = h\big(\boldsymbol{\beta} \cdot \mathbf{x}\big) = h\left[\beta_1 + \sum_{i=2}^{k} \beta_i x^i\right] \quad , \quad \boldsymbol{\beta} \in B \subseteq \mathbb{R}^k,\ \mathbf{x} \in \mathcal{X} \tag{4.21}$$

Verallgemeinertes, Lineares Modell (GLM[8]) mit *Response Funktion* h bzw. *Link Funktion* h^{-1}. Letztere kann unabhängig von der Verteilungsfamilie, sprich dem stochastischen Teil, gewählt werden. Der Teil $\boldsymbol{\beta}\mathbf{x}$ heißt *linearer Prediktor* des Modells. Die Komponente β_1 heißt *intercept* (engl.) und $\beta_2, .., \beta_k$ *Regressionskoeffizienten*.

Beachte dass nach Voraussetzung (vgl. 3.2.1) θ eindeutig durch den Erwartungswert $\mu_{\boldsymbol{\beta}}(\mathbf{x})$ gegeben ist und

$$\theta = \theta(\mu_{\boldsymbol{\beta}}(\mathbf{x})) = (A')^{-1}\left[h(\boldsymbol{\beta} \cdot \mathbf{x})\right] \quad . \tag{4.22}$$

Im Falle $\theta = \boldsymbol{\beta} \cdot \mathbf{x}$ heißt die Link Funktion *kanonische Link Funktion* bzw. die Response Funktion *kanonische Response Funktion*. Erstere entspricht dann

$$h^{-1}(\mu) = \theta(\mu) = (A')^{-1}(\mu) \quad . \tag{4.23}$$

Der (unbekannte) Dispersionsparameter φ spielt oft nur eine unwesentliche Rolle bei der ML-Schätzung von $\boldsymbol{\beta}$ (siehe 4.2.4), und wird deshalb oft nicht mit betrachtet.

Zu konkreten Parameterwerten $\mathbf{x}_1, .., \mathbf{x}_n \in \mathcal{X}$ heiße die Matrix

$$\mathbb{X} := \begin{pmatrix} x_1^1 & \dots & x_1^k \\ \vdots & \ddots & \vdots \\ x_n^1 & \dots & x_n^k \end{pmatrix} \in \mathbb{R}^{n \times k} \tag{4.24}$$

Kovariablen Matrix (*Design Matrix*) und es sei stets angenommen[9] dass $\operatorname{rang}(\mathbb{X}) = k$. Zu konkreten Realisierungen heiße der Vektor $\mathbf{y}^T := (y_1, .., y_n)$ *Realisierungsvektor* und der Vektor mit Komponenten $u_i(\boldsymbol{\beta}) := (y_i - \mu_{\boldsymbol{\beta}}(\mathbf{x}_i))$ *Residuenvektor*. Ausgewertet am ML-Schätzwert $\widehat{\boldsymbol{\beta}}$ ergibt er den *geschätzten Residuenvektor* $\widehat{\mathbf{u}}$.

Mehr Informationen zu verallgemeinert LM sind in [8] zu finden.

Bemerkung zur Linearität: Ist der deterministische Teil anderseits in der *Multilinearen* Form

$$\mu_\beta(\mathbf{x}) = h\left(\beta(\mathbf{x})\right) = h\left[\sum_{j=1}^{N} \sum_{i_1,..,i_j} \beta_{i_1,..,i_j} \cdot x^{i_1}, .., x^{i_j}\right] \tag{4.25}$$

mit $\beta : \mathbb{R}^k \to \mathbb{R}$ als Linearkombination von Multilinearformen 1. bis N. Stufe (*Polynom* N. Grades in $\mathbf{x}$), so können die Produkte $x^{i_1,..,i_j} := x^{i_1} \cdot .. \cdot x^{i_j}$, insofern $\beta_{i_1,..,i_j} \neq 0$, für die Regressionsanalyse als weitere *unabhängige* Einflussgrößen betrachtet werden. Das Problem reduziert sich damit wieder auf den verallgemeinert, linearen Fall.

Im Falle eines nicht-linearen Zusammenspiels der Einflussgrößen $x^1, .., x^k$ wie in (4.25) mit $\beta_{i_1,..,i_r,..,i_s..,i_j} \neq 0$ für mindestens ein $i_r \neq i_s$, spricht man von *Wechselwirkungen der Einflussgrößen*.

[7] Beachte dass dann insbesondere h streng monoton ist.

[8] **G**eneralized **L**inear **M**odel.

[9] Andernfalls spricht man von sogenannten *korrelierten* Kovariablen(werten). Wie sich z.B. im linearen Modell 4.2.7 zeigt, ist die *Unkorreliertheit* der Kovariablen(werte) eine wichtige Voraussetzung für die Eindeutigkeit von ML-Schätzungen. Mögliche Korrelationen sollten daher schon während der Stichprobenerfassung vermieden werden. Im Nachhinein korrigiert können sie z.B. durch geeignete Reduzierung der Kovariablendimension, sprich, Vernachlässigung einer oder mehrerer *Einflussgrößen*.

4.2.3 Kategorielle Kovariablen im GLM

Es ist in allgemeinen Modellen nicht selten der Fall, das manche Komponenten der Kovariable nur endlich viele Werte annehmen können, deren Interpretation als Zahlenwerte sich oft als kontraintuitiv erweist. Man spricht von sogenannten *kategoriellen Kovariablen* oder auch *Faktoren*. Eine Beschreibung der entsprechenden Abhängigkeiten in Form eines linearen Prediktors wie in 4.2.2, ist oft uneindeutig oder gar nicht umsetzbar.

Durch Einführung weiterer Kovariablen lässt sich dieses Problem auf den verallgemeinert linearen Fall zurückführen. Da man das Tupel der kategoriellen Komponenten einer Kovariablen zu einer einzigen kategoriellen Kovariable zusammenfassen kann, sei im folgenden nur der Fall eines einzigen Faktors betrachtet.

Betrachtet sei das Modell mit Exponential-Dispersionsfamilie $(f_{\theta,\varphi})$ auf $E \subseteq \mathbb{R}$, Kovariablen $\mathbf{x} \in \mathcal{X} \subseteq \{1\}\times\mathbb{R}^{k-1}$, $t \in \{1,..,T\}$, unbekanntem Parameter $\boldsymbol{\beta} \in B \subseteq \mathbb{R}^{T\times k}$ und deterministischen Teil

$$\mu_{\boldsymbol{\beta}}(\mathbf{x},t) = h\left[\sum_{\tau=1}^{T}\delta_{\tau,t}\cdot\sum_{i=1}^{k}\beta_{\tau,i}\cdot x^i\right] \tag{4.26}$$

wobei h sei wie in 4.2.2. Definiert man die transformierte Kovariable mit Komponenten

$$\widetilde{x}^{\tau,i} = \widetilde{x}^{\tau,i}(\mathbf{x},t) := \begin{cases} x^i & : \tau = 1 \\ \delta_{\tau,t}\cdot x^i & : \tau \in \{2,..,T\} \end{cases} \quad , \quad i \in \{1,..,k\} \tag{4.27}$$

und den neuen unbekannten Parameter

$$\widetilde{\beta}_{\tau,i} = \widetilde{\beta}_{\tau,i}(\boldsymbol{\beta}) := \begin{cases} \beta_{1,i} & : \tau = 1 \\ (\beta_{\tau,i} - \beta_{1,i}) & : \tau \in \{2,..,T\} \end{cases} \quad , \quad i \in \{1,..,k\} \tag{4.28}$$

so geht (4.26) über in

$$\mu_{\boldsymbol{\beta}}(\mathbf{x},t) = h\left[\sum_{i=1}^{k}\underbrace{\beta_{1,i}}_{\widetilde{\beta}_{1,i}}\cdot x^i + \sum_{\tau=2}^{T}\sum_{i=1}^{k}\underbrace{(\beta_{\tau,i}-\beta_{1,i})}_{\widetilde{\beta}_{\tau,i}}\cdot\underbrace{\delta_{\tau,t}\cdot x^i}_{\widetilde{x}^{\tau,i}}\right]$$

$$= h\left[\sum_{\tau=1}^{T}\sum_{i=1}^{k}\widetilde{\beta}_{\tau,i}\cdot\widetilde{x}^{\tau,i}\right] =: \mu_{\widetilde{\boldsymbol{\beta}}}\left(\widetilde{\mathbf{x}}\right) \quad . \tag{4.29}$$

Mit der Kovariablen $\widetilde{\mathbf{x}}$ und unbekanntem Parameter $\widetilde{\boldsymbol{\beta}}$ wird das Modell zu einem verallgemeinert linearen.

4.2.4 ML-Schätzungen im GLM

Gegeben sei das GLM mit stochastischem Teil $f_{\theta,\varphi}$, unbekannten Parametern $\boldsymbol{\beta} \in B \subseteq \mathbb{R}^k$ und $\varphi \in \Phi \subseteq \mathbb{R}\setminus\{0\}$, Response Funktion h und Kovariablen $\mathbf{x} \in \mathcal{X} \subseteq \{1\}\times\mathbb{R}^{k-1}$. Zu vorgegebenen Kovariablenwerten $\mathbf{x}_1,..,\mathbf{x}_n \in \mathcal{X}$ und Realisierungsvektor $\mathbf{y} \in \mathbb{R}^n$ ist die Loglikelihood Funktion gegeben durch

$$l(\boldsymbol{\beta},\varphi \mid y_1,..,y_n) = \sum_{i=1}^{n}\left\{\frac{1}{\varphi}\Big[y_i\cdot\theta\left(\mu_{\boldsymbol{\beta}}(\mathbf{x}_i)\right) - A\left(\theta(\mu_{\boldsymbol{\beta}}(\mathbf{x}_i))\right)\Big] - C(y_i,\varphi)\right\} \tag{4.30}$$

mit Ableitung

$$\frac{\partial l}{\partial\boldsymbol{\beta}} = \sum_{i=1}^{n}\frac{(y_i-\mu_{\boldsymbol{\beta}}(\mathbf{x}_i))}{\varphi}\cdot\frac{\partial\theta(\mu_{\boldsymbol{\beta}}(\mathbf{x}_i))}{\partial\boldsymbol{\beta}} = \sum_{i=1}^{n}\frac{(y_i-\mu_{\boldsymbol{\beta}}(\mathbf{x}_i))}{\sigma_i^2(\boldsymbol{\beta})}\cdot\frac{\partial\mu_{\boldsymbol{\beta}}(\mathbf{x}_i)}{\partial\boldsymbol{\beta}}$$

$$= \sum_{i=1}^{n}\frac{(y_i-\mu_{\boldsymbol{\beta}}(\mathbf{x}_i))}{\sigma_i^2(\boldsymbol{\beta})}\cdot h'\left(\boldsymbol{\beta}\cdot\mathbf{x}_i\right)\cdot\mathbf{x}_i =: S(\boldsymbol{\beta}\mid y_1,..,y_n) \tag{4.31}$$

wobei

$$\sigma_i^2(\boldsymbol{\beta}) := \varphi \cdot A''(\theta(\mu_{\boldsymbol{\beta}}(\mathbf{x}_i))) \tag{4.32}$$

die Varianz einer gemäß $f_{\theta(\mu_{\boldsymbol{\beta}}(\mathbf{x}_i)),\varphi}$ verteilten Zufallsgröße ist. Beachte dass der ML-Schätzer $\widehat{\boldsymbol{\beta}}$ als Nullstelle von $\frac{\partial l}{\partial \boldsymbol{\beta}}$, unabhängig vom Dispersionsparameter φ ist, weshalb auch oft $S(\boldsymbol{\beta} \mid y_1, .., y_n) := \partial_{\boldsymbol{\beta}} l$ als Score Funktion betrachtet wird.

Die entsprechende erwartete Fisher Information (vgl. def. 4.2.1) erhält man aus (4.31) gemäß

$$\mathcal{J}(\boldsymbol{\beta}) = \mathbb{X}^T \cdot \mathbb{H}(\boldsymbol{\beta}) \cdot \mathbb{X} \quad , \tag{4.33}$$

wobei $\mathbb{X}$ die Kovariablen Matrix ist und

$$\mathbb{H}(\boldsymbol{\beta}) := \begin{pmatrix} \frac{\left[h'(\boldsymbol{\beta}\mathbf{x}_1)\right]^2}{\sigma_i^2(\boldsymbol{\beta})} & \dots & 0 \\ & \ddots & \\ 0 & \dots & \frac{\left[h'(\boldsymbol{\beta}\mathbf{x}_n)\right]^2}{\sigma_i^2(\boldsymbol{\beta})} \end{pmatrix} \quad . \tag{4.34}$$

Aus der positiv-Definitheit[10] von $\mathbb{H}$ und $\text{rang}(\mathbb{X}) = k$, folgt die positiv-Definitheit von $\mathcal{J}(\boldsymbol{\beta})$.

Spezialfälle:

(a) Im Falle einer von $\mathbf{x}$ unabhängigen Varianz, sprich $\sigma_i(\boldsymbol{\beta}) = \sigma_j(\boldsymbol{\beta}) \ \forall \ i, j$, nimmt die Score Funktion (4.31) die Form

$$S(\boldsymbol{\beta} \mid y_1, .., y_n) = \sum_{i=1}^{n} (y_i - \mu_{\boldsymbol{\beta}}(\mathbf{x}_i)) \cdot \frac{\partial \mu_{\boldsymbol{\beta}}(\mathbf{x}_i)}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mu_{\boldsymbol{\beta}}(\mathbf{x}_i))^2 \tag{4.35}$$

an. Die Bestimmung des ML-Schätzers $\widehat{\boldsymbol{\beta}}$ entspricht also der Minimierung der Quadratsumme

$$\sum_{i=1}^{n} (y_i - \mu_{\boldsymbol{\beta}}(\mathbf{x}_i))^2 \quad . \tag{4.36}$$

(b) Sind anderseits die σ_i^2 zwar nicht identisch, dennoch bekannt, so entsprich eine ML-Schätzung der Minimierung der gewichteten Quadratsumme

$$\sum_{i=1}^{n} \frac{\left(y_i - \mu_{\boldsymbol{\beta}}(\mathbf{x}_i)\right)^2}{\sigma_i^2} \quad . \tag{4.37}$$

(c) Im Falle einer kanonischen Link Funktion vereinfacht sich (4.31) zu

$$S(\boldsymbol{\beta} \mid y_1, .., y_n) = \frac{1}{\varphi} \sum_{i=1}^{n} \left[y_i - \mu_{\boldsymbol{\beta}}(\mathbf{x}_i)\right] \cdot \mathbf{x}_i \quad . \tag{4.38}$$

und man erhält die Fisher Information

$$\mathcal{I}(\boldsymbol{\beta} \mid y_1, .., y_n) = \frac{1}{\varphi^2} \mathbb{X}^T \cdot \mathbb{S}(\boldsymbol{\beta}) \cdot \mathbb{X} \tag{4.39}$$

wobei $\mathbb{X}$ die Kovariablen Matrix ist und

$$\mathbb{S}(\boldsymbol{\beta}) := \begin{pmatrix} \sigma_1^2(\boldsymbol{\beta}) & \dots & 0 \\ & \ddots & \\ 0 & \dots & \sigma_n^2(\boldsymbol{\beta}) \end{pmatrix} \quad . \tag{4.40}$$

Aus der positiv-Definitheit von $\mathbb{S}$ und $\text{rang}(\mathbb{X}) = k$, folgt die positiv-Definitheit von $\mathcal{I} = -\partial_{\boldsymbol{\beta}}^2 l$. Die Loglikelihood Funktion $l(\boldsymbol{\beta} \mid y_1, .., y_n)$ ist daher streng konkav und der ML-Schätzer existiert als eindeutige Lösung der Score Gleichung.

[10] Beachte dass nach Voraussetzung $h' \neq 0$.

Bemerkung: Im allgemeinen entspricht die ML-Schätzung im GLM dem Auffinden der Nullstelle der Score Funktion (4.31). Dies ist meist nur mit numerischen Verfahren möglich, die ihrerseits allerdings oft geeignete *Startwerte* für $\widehat{\boldsymbol{\beta}}$ benötigen. Letztere können unter anderem durch folgende Ansätze gewonnen werden:

(i) Im Falle eines komplexen stochastischen Teils, können diese zuerst aus einem einfacheren Modell mit gleicher Link Funktion, sprich, gleichem deterministischen Teil, gewonnen werden.

(ii) Durch Unterdrückung der Abhängigkeit der Varianzen $\sigma_i(\boldsymbol{\beta})$ in (4.31) von $\boldsymbol{\beta}$ und $\mathbf{x}$, kann ein Startwert $\widehat{\boldsymbol{\beta}}^{(0)}$ durch Minimierung der Quadratsumme (4.36) geschätzt werden.

4.2.5 Beispiel: ML-Schätzung im kanonischen Poisson-Modell

Betrachtet sei das Poissonmodell mit Verteilungsdichten (bzgl. des Zählmaßes)

$$\text{Poisson}_\lambda(k) = \frac{\lambda^k}{k!} e^{-\lambda} = \exp\left[\theta k - A(\theta) - c(k)\right] \quad , \quad k \in \mathbb{N}_0 \tag{4.41}$$

Kovariable $\mathbf{x} \in \mathcal{X} \subseteq \{1\} \times \mathbb{R}^{k-1}$, Regressionsparameter $\boldsymbol{\beta} \in B \subseteq \mathbb{R}^k$ und deterministischen Teil $\mu_{\boldsymbol{\beta}}(\mathbf{x}) = e^{\boldsymbol{\beta}\mathbf{x}}$. Als Spezialfall einer Exponential-Dispersionsfamilie (3.8) mit

$$\theta := \ln\lambda \quad , \quad A(\theta) := e^\theta \quad , \quad \varphi := 1 \quad , \quad C(k) := \ln k! \tag{4.42}$$

macht die Poissonfamilie das Modell zu einem verallgemeinert linearen, mit kanonischer Link Funktion

$$h^{-1}(\mu) = \ln\mu \quad . \tag{4.43}$$

Aus der Score Funktion für Stichprobe $y_1, .., y_n$ zu den Kovariablenwerten $\mathbf{x}_1, .., \mathbf{x}_n \in \mathcal{X}$

$$S(\boldsymbol{\beta} \mid y_1, .., y_n) \overset{(4.38)}{=} \sum_{i=1}^n \left[y_i - e^{\boldsymbol{\beta}\mathbf{x}_i}\right] \cdot \mathbf{x}_i \tag{4.44}$$

ergibt sich der ML-Schätzer $\widehat{\boldsymbol{\beta}}$ als deren eindeutige Nullstelle[11] (siehe 4.2.4(c)).

Bemerkung: Ist der deterministische Teil zunächst in der nicht-kanonischen Form

$$\mu_{\boldsymbol{\beta}}(\mathbf{x}) = \beta_1 \cdot \prod_{i=2}^k \left[f_i(x^i)\right]^{\beta_i} \quad , \quad \mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^{k-1} \quad , \tag{4.45}$$

mit $f_i : U_i \subseteq \mathbb{R} \to \mathbb{R}$ invertierbar, so geht (4.45) nach der Kovariablentransformation[12]

$$\widetilde{x}^i := \begin{cases} 1 & : i = 1 \\ \ln f_i(x^i) & : 2 \le i \le k \end{cases} \tag{4.46}$$

und Parametertransformation

$$\widetilde{\beta}_i := \begin{cases} \ln\beta_1 & : i = 1 \\ \beta_i & : 2 \le i \le k \end{cases} \tag{4.47}$$

über in die Form

$$\mu_{\boldsymbol{\beta}}(\mathbf{x}) = \exp\left[\ln\beta_1 + \sum_{i=2}^k \beta_i \cdot \widetilde{x}^i\right] = e^{\widetilde{\boldsymbol{\beta}}\cdot\widetilde{\mathbf{x}}} =: \mu_{\widetilde{\boldsymbol{\beta}}}(\widetilde{\mathbf{x}}) \quad , \tag{4.48}$$

mit kanonischer Response Funktion $h(\widetilde{\boldsymbol{\beta}}\widetilde{\mathbf{x}}) = e^{\widetilde{\boldsymbol{\beta}}\widetilde{\mathbf{x}}}$. Im Falle einer vorliegenden Stichprobe kann somit zunächst $\widetilde{\boldsymbol{\beta}}$ anhand der transformierten Kovariablenwerte $\widetilde{\mathbf{x}}_1, .., \widetilde{\mathbf{x}}_n$ geschätzt werden. Ein Schätzer für $\boldsymbol{\beta}$ lässt sich dann durch Rücktransformation von (4.47) konstruieren[13].

[11] Im allgemeinen numerisch zu ermitteln.
[12] Beachte dass dann $\widetilde{\mathbf{x}} \in \widetilde{\mathcal{X}} \subseteq \{1\} \times \mathbb{R}^{k-1}$ für geeignetes $\widetilde{\mathcal{X}}$.
[13] Dieser muss allerdings nicht mehr Erwartungstreu sein.

4.2.6 Beispiel: ML-Schätzung im kanonischen Binomialmodell

Betrachtet sei das Modell mit der normierten Binomialfamilie[14] $(b_{N,p})_{0\leq p\leq 1}$ (N vorgegeben) als stochastischen Teil, Kovariablen $\mathbf{x} \in \mathcal{X} \subseteq \{1\} \times \mathbb{R}^{k-1}$, Regressionsparameter $\boldsymbol{\beta} \in B \subseteq \mathbb{R}^k$ und deterministischen Teil

$$\mu_{\boldsymbol{\beta}}(\mathbf{x}) = h(\boldsymbol{\beta}\mathbf{x}) := \frac{1}{1+e^{-\boldsymbol{\beta}\mathbf{x}}} \quad . \tag{4.49}$$

Als Spezialfall einer Exponential-Dispersionsfamilie (3.8) mit

$$\theta := \ln\frac{p}{1-p} \quad , \quad A(\theta) := \ln\left[1+e^{\theta}\right] \quad , \quad \varphi := \frac{1}{N} \quad , \quad C(y) := -\ln\binom{N}{Ny} \quad , \quad y \in \left\{\tfrac{0}{N},..,\tfrac{N}{N}\right\} \tag{4.50}$$

macht die Binomialfamilie das Modell zu einem verallgemeinert linearen, mit kanonischer Link Funktion

$$h^{-1}(\mu) = \ln\frac{\mu}{1-\mu} \quad . \tag{4.51}$$

Aus der Score Funktion für Stichprobe $y_1,..,y_n$ zu den Kovariablenwerten $\mathbf{x}_1,..,\mathbf{x}_n \in \mathcal{X}$

$$S(\boldsymbol{\beta} \mid y_1,..,y_n) \overset{(4.38)}{=} \sum_{i=1}^{n}\left[y_i - \frac{1}{1+e^{-\boldsymbol{\beta}\mathbf{x}_i}}\right]\cdot\mathbf{x}_i \tag{4.52}$$

ergibt sich der ML-Schätzer $\widehat{\boldsymbol{\beta}}$ als deren eindeutige Nullstelle[15] (siehe 4.2.4(c)). Die Fisher-Information erhält man gemäß (4.39) als

$$\mathcal{I}(\boldsymbol{\beta} \mid y_1,..,y_n) = \mathbb{X}^T \cdot \mathbb{S}(\boldsymbol{\beta}) \cdot \mathbb{X} = \mathcal{J}(\boldsymbol{\beta}) \quad , \tag{4.53}$$

mit

$$\mathbb{S}(\boldsymbol{\beta}) \overset{(4.40)}{=} \frac{1}{2}\begin{pmatrix} \left[1+\cosh(\boldsymbol{\beta}\mathbf{x}_1)\right]^{-1} & \dots & 0 \\ & \ddots & \\ 0 & \dots & \left[1+\cosh(\boldsymbol{\beta}\mathbf{x}_n)\right]^{-1} \end{pmatrix} \quad . \tag{4.54}$$

Bemerkung: Ist der deterministische Teil zunächst in der nicht-kanonischen Form

$$\mu_{\boldsymbol{\beta}}(\mathbf{x}) = \left[1+\beta_1\cdot\prod_{i=2}^{k}\left[f_i(x^i)\right]^{\beta_i}\right]^{-1} \quad , \quad \mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^{k-1} \quad , \tag{4.55}$$

mit $f_i : U_i \subseteq \mathbb{R} \to \mathbb{R}$ invertierbar, so geht (4.55) nach der Kovariablentransformation

$$\widetilde{x}^i := \begin{cases} 1 & : i = 1 \\ \ln f_i(x^i) & : 2 \leq i \leq k \end{cases} \tag{4.56}$$

und Parametertransformation

$$\widetilde{\beta}_i := \begin{cases} -\ln\beta_1 & : i = 1 \\ -\beta_i & : 2 \leq i \leq k \end{cases} \tag{4.57}$$

über in die Form

$$\mu_{\boldsymbol{\beta}}(\mathbf{x}) = \left[1+\exp\left[\ln\beta_1 + \sum_{i=2}^{k}\beta_i\cdot\widetilde{x}^i\right]\right]^{-1} = \frac{1}{1+e^{-\widetilde{\boldsymbol{\beta}}\cdot\widetilde{\mathbf{x}}}} =: \mu_{\widetilde{\boldsymbol{\beta}}}(\widetilde{\mathbf{x}}) \quad , \tag{4.58}$$

mit kanonischer Response Funktion $h(\widetilde{\boldsymbol{\beta}}\widetilde{\mathbf{x}}) = \left[1+e^{-\widetilde{\boldsymbol{\beta}}\widetilde{\mathbf{x}}}\right]^{-1}$. Im Falle einer vorliegenden Stichprobe kann somit zunächst $\widetilde{\boldsymbol{\beta}}$ anhand der transformierten Kovariablenwerte $\widetilde{\mathbf{x}}_1,..,\widetilde{\mathbf{x}}_n$ geschätzt werden. Ein Schätzer für $\boldsymbol{\beta}$ lässt sich dann durch Rücktransformation von (4.47) konstruieren[16].

[14] Entspricht der Verteilung einer Zufallsgröße Y/N, wobei Y Binomialverteilt mit N Versuchen und Erfolgswahrscheinlichkeit p ist.
[15] Im allgemeinen numerisch zu ermitteln.
[16] Dieser muss allerdings nicht mehr Erwartungstreu sein.

4.2.7 ML-Schätzung im klassischen, linearen Modell

Betrachten das klassische lineare Modell mit Kovariablen $\mathbf{x} \in \mathcal{X} \subseteq \{1\} \times \mathbb{R}^{k-1}$, unbekannten Parametern $\boldsymbol{\beta} \in B \subseteq \mathbb{R}^k, \sigma^2 \in (0,\infty)$, stochastischem Teil die Familie $\mathcal{N}_{\mu,\sigma^2}$ der Normalverteilungen und linearem deterministischen Teil

$$\mu_{\boldsymbol{\beta}}(\mathbf{x}) = \boldsymbol{\beta} \cdot \mathbf{x} \quad , \quad \boldsymbol{\beta} \in B \tag{4.59}$$

das heißt $Y = \boldsymbol{\beta} \cdot \mathbf{x} + \varepsilon$ mit $\varepsilon \sim \mathcal{N}_{0,\sigma^2}$. Zu vorgegebenen Kovariablenwerten $\mathbf{x}_1, .., \mathbf{x}_n \in \mathcal{X}$ und Realisierungen $y_1, .., y_n \in \mathbb{R}$ ist die Loglikelihood Funktion des Modells gegeben durch

$$l(\boldsymbol{\beta}, \sigma^2 \mid y_1, .., y_n) = -\frac{n}{2} \ln\left(2\pi\sigma^2\right) - \frac{(\mathbf{y} - \mathbb{X}\boldsymbol{\beta})^2}{2\sigma^2} \tag{4.60}$$

mit Ableitungen

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2} \cdot \mathbb{X}^T(\mathbf{y} - \mathbb{X}\boldsymbol{\beta})$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{(\mathbf{y} - \mathbb{X}\boldsymbol{\beta})^2}{2\sigma^4} \tag{4.61}$$

wobei $\mathbb{X}$ die Kovariablen Matrix und $\mathbf{y}$ der Realisierungsvektor sind. Die ML-Schätzer für $\boldsymbol{\beta}, \sigma^2$ ergeben sich als eindeutige Nullstellen der obigen Ableitungen (4.61) gemäß

$$\boxed{\widehat{\boldsymbol{\beta}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{y} \quad , \quad \widehat{\sigma}^2 = \frac{1}{n}\left(\mathbf{y} - \mathbb{X}\widehat{\boldsymbol{\beta}}\right)^2} \tag{4.62}$$

Der Vektor $\mathbb{X}\boldsymbol{\beta}$ entspricht genau dem Erwartungswert eines $\mathbb{R}^n$-Zufallsvektors, mit unabhängigen, gemäß $\mathcal{N}_{\mu_{\boldsymbol{\beta}}(\mathbf{x}_i),\sigma^2}$, $i = 1, .., n$ verteilten Komponenten. Insbesondere ist $\mathbb{X} \cdot \widehat{\boldsymbol{\beta}}(\mathbf{y})$ die Punktprognose zu den Kovariablenwerten $\mathbf{x}_1, .., \mathbf{x}_n$, basierend auf der vorliegenden Stichprobe $\mathbf{y}$. Der Schätzwert $\widehat{\sigma}^2$ entspricht daher genau dem Mittelwert der Quadrate der geschätzten Residuen $\widehat{u}_i := u_i(\widehat{\boldsymbol{\beta}})$, das heißt $\widehat{\sigma}^2 = \frac{1}{n}\widehat{\mathbf{u}}^2 = \frac{1}{n}\,\mathrm{RSS}(y_1, .., y_n)$.

Den Schätzwert $\widehat{\boldsymbol{\beta}}$ erhält man auch durch Minimierung der Norm $\|\mathbf{y} - \mathbb{X} \cdot \boldsymbol{\beta}\|$ (vgl. 4.2.4). Ist $P_{\mathbb{X}} : \mathbb{R}^n \to \mathrm{image}(\mathbb{X})$ der Orthogonalprojektor auf das Bild von $\mathbb{X}$ [17], so gilt

$$\mathbb{X}\widehat{\boldsymbol{\beta}} = P_{\mathbb{X}}\mathbf{y} \quad , \quad \widehat{\sigma}^2 = \frac{1}{n}\left[(\mathbb{1} - P_{\mathbb{X}})\mathbf{y}\right]^2 \tag{4.63}$$

Obiger Zusammenhang ist dargestellt in Abb. 4.1.

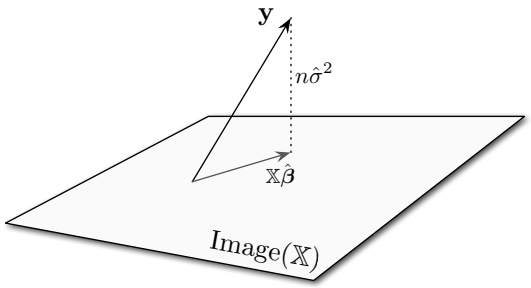

Abbildung 4.1: Zur Interpretation des ML-Schätzers als Kleinste-Quadrate-Schätzer im klassischen, linearen Modell.

Die ML-Schätzwerte (4.62) hängen natürlich stets von den konkreten Realisierungen $y_1, .., y_n$ ab (*Stichprobenfunktionen*). Ersetzt man diese durch unabhängige Zufallsgrößen $Y_i \sim \mathcal{N}_{\mu_{\boldsymbol{\beta}}(\mathbf{x}_i),\sigma^2}$, $i = 1, .., n$, so werden $\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2$ selbst zu Zufallsgrößen, die gemäß

$$\boxed{\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_{\boldsymbol{\beta},\ \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}} \quad , \quad n \cdot \frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-k}} \tag{4.64}$$

[17] Es ist $P_{\mathbb{X}} = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$.

verteilt sind. Zu erkennen ist, dass $\widehat{\boldsymbol{\beta}}$ zwar ein erwartungstreuer Schätzer, es jedoch $\widehat{\sigma}^2$ nicht ist, denn

$$\mathbb{E}\widehat{\sigma}^2 = \frac{(n-k)}{n} \cdot \sigma^2 \quad , \quad \mathbb{V}\widehat{\sigma}^2 = \frac{(n-k)}{n^2} \cdot 2\sigma^4 \tag{4.65}$$

Ein erwartungstreuer Varianzschätzer ist gegeben durch

$$\boxed{\widetilde{\sigma}^2 := \frac{n \cdot \widehat{\sigma}^2}{(n-k)}} \tag{4.66}$$

Der erwartungstreue Schätzer $\widetilde{\sigma}^2(\mathbb{X}^T\mathbb{X})^{-1}$ für die Kovarianz des ML-Schätzers $\widehat{\boldsymbol{\beta}}$ heißt *Standardfehler* und ist ein *Maß für die Ungenauigkeit* des letzteren.

Bemerke: Aus (4.63) und der Tatsache dass $\mathbf{Y}$ normalverteilt ist mit $\mathrm{Cov}(\mathbf{Y}) = \sigma^2 \cdot \mathbb{1}$, folgt die stochastische Unabhängigkeit der ML-Schätzer $\widehat{\boldsymbol{\beta}}$ und $\widehat{\sigma}^2$.

Spezialfall: Im Spezialfall $k = 1$, sprich $\boldsymbol{\beta} \in \mathbb{R}^1$, entspricht β genau dem festen (unbekannten) Erwartungswert $\mu = \beta$ der Zielgröße. Gemäß (4.62) nimmt dessen ML-Schätzer die vertraute Form

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y^i \tag{4.67}$$

an. Der erwartungstreue Schätzer (4.66) für die Varianz σ^2 der Zielgröße nimmt entsprechend die Form

$$\widetilde{\sigma}^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(y^i - \mu\right)^2 \tag{4.68}$$

an. Schließlich ist $\widetilde{\sigma}^2/n$ der oben eingeführte, erwartungstreue Schätzer für die Varianz von $\widehat{\mu}$ (Standardfehler).

4.2.8 Konfidenzintervalle und Prognosen im klassischen LM

Betrachtet sei das klassische, lineare Model mit deterministischen Teil $\mu_{\boldsymbol{\beta}}(\mathbf{x}) = \boldsymbol{\beta} \cdot \mathbf{x}$ und Verteilungsfamilie $\mathcal{N}_{\mu,\sigma^2}$. Seien $\mathbf{x}_1, .., \mathbf{x}_n \in \mathcal{X} \subseteq \{1\} \times \mathbb{R}^{k-1}$ konkrete Kovariablenwerte, dazu Zufallsvektor $\mathbf{Y}$ mit unabhängigen, gemäß $\mathcal{N}_{\mu_{\boldsymbol{\beta}}(\mathbf{x}_i),\sigma^2}$ verteilten Komponenten. Sind $\widehat{\boldsymbol{\beta}}(\mathbf{Y})$ und $\widetilde{\sigma}^2(\mathbf{Y})$ die entsprechenden Schätzer aus Abschnitt 4.2.7, so gilt für beliebiges $\mathbf{c} \in \mathbb{R}^k$:

$$\mathcal{T}_{\mathbf{c}}(\mathbf{Y}) := \frac{1}{\sqrt{\widetilde{\sigma}^2(\mathbf{Y})}} \frac{\mathbf{c}\widehat{\boldsymbol{\beta}}(\mathbf{Y}) - \mathbf{c}\boldsymbol{\beta}}{\sqrt{\mathbf{c}^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{c}}} \sim t_{n-k} \tag{4.69}$$

wobei t_{n-k} die Studentsche t-Verteilung sei[18]. Interpretationsgemäß stellt der Zähler in (4.69) die Abweichung einer geschätzten Linearkombination $\mathbf{c}\widehat{\boldsymbol{\beta}}$ vom wirklichen Wert $\mathbf{c}\boldsymbol{\beta}$ dar. Der Nenner in (4.69) ist lediglich die geschätzte Standardabweichung (*Standardfehler*) dieser Abweichung.

Betrachtet sei nun die Hypothese $H_o : \mathbf{c}\boldsymbol{\beta} = c_0$, für irgendwelche $\mathbf{c} \in \mathbb{R}^k,\ c_0 \in \mathbb{R}$. Dann gilt unter Annahme der Nullhypothese H_o:

$$\mathcal{T}_{\mathbf{c}}(\mathbf{Y}) = \frac{1}{\sqrt{\widetilde{\sigma}^2(\mathbf{Y})}} \frac{\mathbf{c}\widehat{\boldsymbol{\beta}}(\mathbf{Y}) - c_0}{\sqrt{\mathbf{c}^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{c}}} \sim t_{n-k} \quad . \tag{4.70}$$

Zu gegebenem *Signifikanzniveau* $0 < \alpha < 1$ sei $t_{n-k,\frac{\alpha}{2}} \in \mathbb{R}$ so dass $\mathcal{P}(\mathcal{T}_{\mathbf{c}} \leq t_{n-k,\frac{\alpha}{2}}) = \frac{\alpha}{2}$. Dann befindet sich $\mathcal{T}_{\mathbf{c}}(\mathbf{Y})$ mit einer Wahrscheinlichkeit $(1-\alpha)$ innerhalb des Intervalls

$$T_{\mathbf{c},\alpha} := \left[t_{n-k,\frac{\alpha}{2}},\ t_{n-k,1-\frac{\alpha}{2}}\right] \tag{4.71}$$

[18] Sie beschreibt die Verteilung einer Zufallsvariablen $\sqrt{\frac{n-k}{Y}} \cdot Z$ mit $Z \sim \mathcal{N}_{0,1}$ und $Y \sim \chi^2_{n-k}$.

bzw. mit einer Wahrscheinlichkeit α außerhalb dessen. Sollte nun ein konkreter Realisierungsvektor $\mathbf{y}$ (*Messungen*) tatsächlich dazu führen dass $\mathcal{T}_{\mathbf{c}}(\mathbf{y}) \notin T_{\mathbf{c},\alpha}$, so kann dies als Hinweis interpretiert werden, die Nullhypothese abzulehnen. Eine häufig zu testende Hypothese ist $H_0 : \beta_j = 0$. Sie ist in der Modellwahl von großer Bedeutung (siehe Abschnitt 5).

Aus (4.69) lässt sich zeigen, dass ganz allgemein gilt

$$(1-\alpha) \overset{(4.69)}{=} \mathcal{P}\left[\mathcal{T}_{\mathbf{c}}(\mathbf{Y}) \in T_\alpha\right] = \mathcal{P}\left[\mathbf{c}\boldsymbol{\beta} \in K_{\mathbf{c},\alpha}(\mathbf{Y})\right] \tag{4.72}$$

mit dem Konfidenzintervall

$$K_{\mathbf{c},\alpha}(\mathbf{y}) := \left[\mathbf{c}\widehat{\boldsymbol{\beta}}(\mathbf{y}) - t_{n-k,1-\frac{\alpha}{2}} \cdot \sqrt{\widetilde{\sigma}^2(\mathbf{y}) \cdot \mathbf{c}^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{c}},\ \mathbf{c}\widehat{\boldsymbol{\beta}}(\mathbf{y}) + t_{n-k,1-\frac{\alpha}{2}} \cdot \sqrt{\widetilde{\sigma}^2(\mathbf{y}) \cdot \mathbf{c}^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{c}}\right] \tag{4.73}$$

für die Linearkombination $\mathbf{c}\boldsymbol{\beta}$ zum Signifikanzniveau α.

Sei nun gegeben ein *neuer* Kovariablenwert $\mathbf{x}_{\text{neu}} \in \mathcal{X}$. Dann ist $K_{\mathbf{x}_{\text{neu}},\alpha}$ ein Konfidenzintervall für den (unbekannten) Erwartungswert $\mathbf{x}_{\text{neu}} \cdot \boldsymbol{\beta}$ einer gemäß $\mathcal{N}_{\boldsymbol{\beta}\mathbf{x}_{\text{neu}},\sigma^2}$ verteilten Zufallsgröße (*neue Messung*) Y_{neu} zum Signifikanzniveau α. Dabei ist bei Extrapolationen Vorsicht geboten, da die Linearität des Modells (deterministischer Teil) außerhalb des Beobachtungsbereich nicht mehr gegeben sein muss.

Als Schätzer für die zu $\mathbf{x}_{\text{neu}}$ gehörige Messung kann $\widehat{y}_{\text{neu}} := \mathbf{x}_{\text{neu}} \cdot \widehat{\boldsymbol{\beta}}$ angesetzt werden. Ähnlich wie $\widehat{\boldsymbol{\beta}}$ ist auch dieser zufällig verteilt, und es gilt

$$\frac{1}{\sqrt{\widetilde{\sigma}^2(\mathbf{Y})}} \frac{Y_{\text{neu}} - \widehat{y}_{\text{neu}}(\mathbf{Y})}{\sqrt{1 + \mathbf{x}_{\text{neu}}^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{x}_{\text{neu}}}} \sim t_{n-k} \tag{4.74}$$

insofern $\mathbf{Y}$ und Y_{neu} unabhängig sind. Aus (4.74) folgt dass

$$I_{\mathbf{x}_{\text{neu}},\alpha} := \left[\widehat{y}_{\text{neu}} - t_{n-k,1-\frac{\alpha}{2}} \cdot \sqrt{1 + \mathbf{x}_{\text{neu}}^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{x}_{\text{neu}}} \cdot \sqrt{\widetilde{\sigma}^2},\right.$$

$$\left.\widehat{y}_{\text{neu}} + t_{n-k,1-\frac{\alpha}{2}} \cdot \sqrt{1 + \mathbf{x}_{\text{neu}}^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{x}_{\text{neu}}} \cdot \sqrt{\widetilde{\sigma}^2}\right] \tag{4.75}$$

ein zufälliges Intervall ist, dass mit Wahrscheinlichkeit $(1-\alpha)$ den neuen Messwert Y_{neu} enthält.

4.2.9 Konsistenz des ML-Schätzers im GLM und asymptotische Tests

Betrachtet sei das verallgemeinerte lineare Modell mit stochastischem Teil $(f_{\theta,\varphi})_{\substack{\theta\in\Theta \\ \varphi\in\Phi}}$, Response Funktion $h : \mathbb{R} \to \mathbb{R}$, Kovariablen $\mathbf{x} \in \mathcal{X} \subseteq \{1\} \times \mathbb{R}^{k-1}$ und unbekannten Parametern $\boldsymbol{\beta} \in B \subseteq \mathbb{R}^k$, $\varphi \in \Phi$. Der zulässige Parameterraum B sei offen und konvex. Seien nun $\mathbf{x}_1, \mathbf{x}_2, .. \in \mathcal{X}$ konkrete Kovariablenwerte, dazu die unabhängigen $Y^i \sim f_{\theta(\mu_{\boldsymbol{\beta}}(\mathbf{x}_i)),\varphi}$. Seien $\mathcal{J}_n(\boldsymbol{\beta})$ die erwarteten Fisher Informationen und $\widehat{\boldsymbol{\beta}}_n$ die jeweiligen ML-Schätzer zu den ersten n Realisierungen. Dann gilt[19]:

1. $\mathcal{P}\left(\widehat{\boldsymbol{\beta}}_n \text{ existiert als Lösung der Score Gleichung}\right) \overset{n\to\infty}{\longrightarrow} 1.$
2. $\mathcal{P}\left(\left\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right\| \geq \varepsilon\right) \overset{n\to\infty}{\longrightarrow} 0 \quad \forall\, \varepsilon > 0$, das heißt $\widehat{\boldsymbol{\beta}}$ ist konsistenter Schätzer für $\boldsymbol{\beta}$.
3. $\mathcal{J}_n^{\frac{1}{2}}(\boldsymbol{\beta})^T \cdot (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \overset{\substack{n\to\infty \\ d}}{\longrightarrow} \mathcal{N}_{0,1_{k\times k}}.$

Der Schätzer $\widehat{\boldsymbol{\beta}}_n$ ist also für genügend großen Stichprobenumfang n, annähernd $\mathcal{N}_{\boldsymbol{\beta},\mathcal{J}_n^{-1}(\boldsymbol{\beta})}$ -verteilt und eine *größere* erwartete Fisher Information $\mathcal{J}_n$ entspricht einer *genaueren* Schätzung.

[19] Beachte dass nach 4.2.4 die erwartete Fisher Information $\mathcal{J}_n(\boldsymbol{\beta})$ positiv-definit ist, daher eine positiv-definite Wurzel $\mathcal{J}^{\frac{1}{2}}(\boldsymbol{\beta})$ besitzt, die eindeutig ist.

Unter Verwendung von Aussagen (2) & (3) lässt sich ein asymptotischer Test für die Hypothese $H_0 : \mathbf{c} \cdot \boldsymbol{\beta} = c_0$ angeben, wobei $\mathbf{c} \in \mathbb{R}^k \setminus \{0\},\ c_0 \in \mathbb{R}$. Die asymptotische Teststatistik

$$\mathcal{T}_n(\mathbf{Y}_n) := \frac{1}{\|\mathbf{c}\|} \mathcal{J}_n^{\frac{1}{2}}\left(\widehat{\boldsymbol{\beta}}_n(\mathbf{Y}_n)\right) \cdot \left[\mathbf{c}\widehat{\beta}_n(\mathbf{Y}_n) - c_0\right] \overset{\mathrm{n}\to\infty}{\underset{d}{\longrightarrow}} \mathcal{N}_{0,1}$$

$$n \in \mathbb{N},\ \mathbf{Y}_n := (Y^1, .., Y^n) \tag{4.76}$$

erfüllt unter Annahme der Nullhypothese H_0:

$$\lim_{n\to\infty} \mathcal{P}\big(\mathcal{T}_n(\mathbf{Y}_n) \in T_\alpha\big) = 1 - \alpha \quad , \tag{4.77}$$

mit $T_\alpha := \left(-N_{1-\frac{\alpha}{2}}, N_{1-\frac{\alpha}{2}}\right)$ und $N_{1-\frac{\alpha}{2}}$ als $\left(1 - \frac{\alpha}{2}\right)$ Quantil der Standardnormalverteilung. Ähnlicherweise lässt sich auch das asymptotisches Konfidenzintervall

$$K_{n,\alpha} := \left[\mathbf{c}\widehat{\boldsymbol{\beta}}_n - \|\mathbf{c}\|\, N_{1-\frac{\alpha}{2}} \cdot \mathcal{J}_n^{-\frac{1}{2}}(\widehat{\boldsymbol{\beta}}_n)\ ,\ \mathbf{c}\widehat{\boldsymbol{\beta}}_n + \|c\|\, N_{1-\frac{\alpha}{2}} \cdot \mathcal{J}_n^{-\frac{1}{2}}(\widehat{\boldsymbol{\beta}}_n)\right] \tag{4.78}$$

für die Linearkombination $\mathbf{c}\boldsymbol{\beta}$ mit Signifikanzniveau α angeben.

5 Modellwahl

Bisher wurde stets angenommen, dass die Verteilung der zu untersuchenden Zielgröße von einer Kovariablen $\mathbf{x} \in \mathcal{X} \subseteq \{1\} \times \mathbb{R}^{k-1}$ einer festen Dimension k (*Komplexität*) abhängt. Diese Abhängigkeit war durch einen unbekannten Parameter $\boldsymbol{\beta} \in B$ gegeben.

In der Praxis ist jedoch oft überhaupt nicht klar, welche Kovariablen $x^1, .., x^k$ denn letztendlich die Zielgröße beeinflussen. So stehen zunächst mehrere Modelle zur Auswahl, jedes beschrieben durch einen Kovariablenraum $\mathcal{X}_l \subseteq \{1\} \times \mathbb{R}^{l-1},\ l = 1, .., k$. Daher wird ein Verfahren zur Auswahl eines *passenden* Modells zur Notwendigkeit. Der stochastische Teil, sprich die verfügbare Verteilungsfamilie (f_θ) der Zielgröße, sei der Einfachheit halber für alle Modelle gleich.

Mehr Informationen zur Modellwahl finden sich in [7].

5.1 Modellvergleiche

5.1.1 Definition: Erwartete, Quadratische Prognosefehler (MSPSE)

Betrachten das Modell mit Verteilungsfamilie $(\mathcal{P}_\theta)_{\theta\in\Theta}$ auf $E \subseteq \mathbb{R}$, mit der Kovariablen $x \in \mathcal{X}$ und unbekanntem Parameter $\beta \in B$, sprich $\theta = \zeta_\beta(x)$. Zu konkreten Kovariablenwerten $x_1, .., x_n \in \mathcal{X}$ sei $\widehat{\beta}$ der ML-Schätzer des Modells für entsprechende Realisierungen $y_1, .., y_n \in E$.

Aus $\widehat{\beta}$ lassen sich Punktprognosen $\widehat{y}_{n+i} := \mathbb{E}\mathcal{P}_{\theta=\zeta_{\widehat{\beta}}(x_{n+i})}$ für *neue* Messungen zu Kovariablenwerten $x_{n+i} \in \mathcal{X}$, $i = 1, .., m$ konstruieren. Diese hängen als Stichprobenfunktionen natürlich von den ursprünglichen Realisierungen $y_1, .., y_n$ ab. Ersetzt man letztere durch unabhängige, gemäß $\mathcal{P}_{\theta=\zeta_\beta(x_i)}$ verteilte Zufallsgrößen $Y_i,\ i = 1, .., n$, so sind sowohl $\widehat{\beta}$ als auch die Punktschätzer $\widehat{y}_{n+i}$ zufällig verteilt. Sind nun $Y_{n+i} \sim \mathcal{P}_{\theta=\zeta_\beta(x_{n+i})},\ i = 1, .., m$ weitere unabhängige Zufallsgrößen, so heißt

$$\mathrm{MSPSE}(\beta; x_1, ..x_n; x_{n+1}, .., x_{n+m}) := \mathbb{E}\sum_{i=1}^{m}\left[Y_{n+i} - \widehat{y}_{n+i}(Y_1, .., Y_n)\right]^2 \tag{5.1}$$

erwartete, quadratischer Prognosefehler[20]. Der MSPSE lässt sich auch schreiben als

$$\mathrm{MSPSE} = \sum_{i=1}^{m}\left[\mathbb{V}Y_{n+i} + \mathbb{V}\widehat{y}_{n+i} + (\mathbb{E}\widehat{y}_{n+i} - \mathbb{E}Y_{n+i})\right] \tag{5.2}$$

und hängt nur vom unbekannten Parameter β und den konkreten $x_1, .., x_{n+m} \in \mathcal{X}$ ab. Er entspricht den erwarteten, quadratischen Abweichungen einer zweiten Messreihe von den, durch eine erste Stichprobe postulierten, Prognosen und ist somit ein Maß für die *Güte* des Modells.

[20] **M**ean **S**um of **P**rediction **S**quared **E**rror.

Bemerkung: Oft wird angenommen dass $n = m,\ x_i = x_{n+i}\ \forall\, i = 1, .., n$, das heißt alte und neue Stichprobe werden zu gleichen Kovariablenwerten gemessen. Ist $\mathrm{RSS}(y_1, .., y_n)$ die in (4.18) eingeführte Residuenquadratsumme, so lässt sich in dem Falle zeigen dass

$$\mathrm{MSPSE}(\beta; x_1, .., x_n; x_1, .., x_n) = \mathbb{E}\big\{\,\mathrm{RSS}(Y_1, .., Y_n)\big\} + 2\sum_{i=1}^{n} \mathbb{E}\big\{(Y_i - \mathbb{E}Y_i)(\widehat{y}_{n+i} - \mathbb{E}Y_i)\big\} \tag{5.3}$$

gilt.

5.1.2 MSPSE im klassischen, linearen Modell

Betrachtet sei das klassische LM, mit Verteilungsfamilie $\mathcal{N}_{\mu,\sigma^2}$, Kovariablen $\mathbf{x} \in \mathcal{X} \subseteq \{1\} \times \mathbb{R}^{k-1}$ und unbekannten Parametern $\boldsymbol{\beta} \in B \subseteq \mathbb{R}^k,\ \sigma^2 \in (0, \infty)$. Für konkrete Kovariablenwerte $\mathbf{x}_1, .., \mathbf{x}_n \in \mathcal{X}$ und $\mathbf{x}_{n+1}, .., \mathbf{x}_{n+m} \in \mathcal{X}$ seien jeweils $\mathbb{X} \in \mathbb{R}^{n\times k},\ \mathbb{X}_{\mathrm{neu}} \in \mathbb{R}^{m\times k}$ die Kovariablenmatrizen. Dann lässt sich mit Hilfe von Abschnitt 4.2.7 zeigen, dass der entsprechende MSPSE gegeben ist durch

$$\begin{aligned}
&\mathrm{MSPSE}(\boldsymbol{\beta}, \sigma^2; \mathbf{x}_1, .., \mathbf{x}_n; \mathbf{x}_{n+1}, .., \mathbf{x}_{n+m}) \\
&= \sigma^2 \cdot n + \sigma^2 \cdot \mathrm{trace}\left\{(\mathbb{X}_{\mathrm{neu}}^T \cdot \mathbb{X}_{\mathrm{neu}})(\mathbb{X}^T \cdot \mathbb{X})^{-1}\right\} + \sum_{i=1}^{m} \left(\mathbb{E}\widehat{y}_{n+i} - \mathbb{E}Y_{n+i}\right)^2 \quad .
\end{aligned} \tag{5.4}$$

Der erste Term $n\sigma^2$ ist ein *irreduzibler Prognosefehler*, der nicht von der Modellkomplexität k abhängt. Der zweite Term entspricht der Varianz der Prognosen, und wächst typischerweise mit steigender Modellkomplexität. Der letzte Term in (5.4), die sogenannte *Verzerrung*, sinkt mit steigender Modellkomplexität k.

Bemerkung: Im Falle dass $\mathbb{X} = \mathbb{X}_{\mathrm{neu}}$ vereinfacht dieser sich (5.4) zu

$$\mathrm{MSPSE}(\boldsymbol{\beta}, \sigma^2; \mathbf{x}_1, .., \mathbf{x}_n; \mathbf{x}_1, .., \mathbf{x}_n) = \sigma^2 \cdot n + \sigma^2 \cdot k + \sum_{i=1}^{n} \left(\mathbb{E}\widehat{y}_{n+i} - \mathbb{E}Y_{n+i}\right)^2 \quad . \tag{5.5}$$

Ist $\mathrm{RSS}(y_1, .., y_n)$ die entsprechende Residuenquadratsumme zur Stichprobe $y_1, .., y_n$ und Y_i unabhängige, gemäß $\mathcal{N}_{\boldsymbol{\beta}\mathbf{x}_i,\sigma^2}$ verteilte Zufallsgrößen, $i = 1, .., n$, so nimmt (5.3) die Form

$$\mathrm{MSPSE} = \mathbb{E}\big\{\,\mathrm{RSS}(Y_1, .., Y_n)\big\} + 2k \cdot \sigma^2 \tag{5.6}$$

an.

5.1.3 Definition: Saturiertes Modell & Devianz

Betrachten die Verteilungsfamilie $(f_\theta)_{\theta\in\Theta}$ auf $E \subseteq \mathbb{R}^m$. Zu jedem $k \in \mathbb{N}$ sei $\left(\zeta^{(k)}_{\beta^{(k)}}\right)_{\beta^{(k)}\in B^{(k)}}$ eine durch $\beta^{(k)} \in B^{(k)}$ parametrisierte Familie von Abbildungen $\zeta^{(k)}_{\beta^{(k)}} : \mathcal{X}^{(k)} \subseteq \mathbb{R}^k \to \Theta$. Dann entspricht jede Familie $\left(\zeta^{(k)}_{\beta^{(k)}}\right)_{\beta^{(k)}}$ zusammen mit $(f_\theta)_{\theta\in\Theta}$ einem stochastischen Modell $M^{(k)}$ mit Kovariablen[21] $\mathbf{x}^{(k)} \in \mathcal{X}^{(k)}$, Komplexität k und unbekanntem Parameter $\beta^{(k)} \in B^{(k)}$, das heißt $\theta = \zeta^{(k)}_{\beta^{(k)}}\big(\mathbf{x}^{(k)}\big)$.

Sei nun $y_1, .., y_n \in E$ eine vorgegebene Stichprobe zu den *vollen* Kovariablenwerten $\mathbf{x}_1, .., \mathbf{x}_n \in \mathbb{R}^\infty$, so dass $P_k\mathbf{x}_i \in \mathcal{X}^k\ \forall\ i = 1, .., n$ für jeden Projektor $P_k : \mathbb{R}^\infty \to \mathbb{R}^k$ auf die ersten k Koordinaten. Für $k \in \mathbb{N}$ sei $\widehat{\beta}^{(k)}$ der entsprechende ML-Schätzwert im Modell $M^{(k)}$, bzgl. der Kovariablenwerte $P_k\mathbf{x}_1, .., P_k\mathbf{x}_n \in \mathcal{X}^{(k)}$ und $\widehat{y}^{(k)}_{n+i}$ die daraus definierten Punktprognosen für die gleichen Kovariablenwerte (vgl. Abschnitt 4.2.1).

Dann heißt das einfachste Modell $M^{(k)}$ mit verschwindender Residuenquadratsumme, sprich $\widehat{y}^{(k)}_{n+i}(y_1, .., y_n) = y_i\ \forall\ i = 1, .., n$, *saturiertes Modell* zu dieser Stichprobe. Es entspricht gewissermaßen dem einfachsten Modell, das

[21] Es sei angenommen dass $P_{k,l}\mathcal{X}^{(k)} \subseteq \mathcal{X}^{(l)}$ für jeden Projektor $P_{k,l} : \mathbb{R}^k \to \mathbb{R}^l,\ l \leq k$ auf die ersten l von k Koordinaten.

den gesamten Datensatz exakt wiedergibt. Häufig ist es so, dass seine Dimension dem Stichprobenumfang entspricht.

Betrachtet sei nun der Spezialfall einer Verteilungsfamilie der Form $(f_{\theta,\varphi})_{\substack{\theta\in\Theta\\ \varphi\in\Phi}}$, mit so genanntem *Dispersionsparameter*[22] φ und $\mathbb{E}f_{\theta,\varphi}$ nur von θ abhängig. Entsprechend seien für das k-te Modell $M^{(k)}$ die unbekannten Parameter $\beta^{(k)} \in B^{(k)}$ und $\varphi \in \Phi$, so dass $\theta = \zeta^{(k)}_{\beta^{(k)}}\big(\mathbf{x}^{(k)}\big)$. Ist nun $M^{(k)}$ das saturierte Modell zur gegebenen Stichprobe und $l^{(k)}(\beta^{(k)}, \varphi \mid y_1, .., y_n)$ seine Loglikelihood Funktion, so heißt

$$l_{\text{sat}} := l^{(k)}\big(\widehat{\beta}^{(k)}, \varphi \mid y_1, .., y_n\big) \tag{5.7}$$

saturierter Loglikelihood Wert. Dieser hängt sowohl von der vorliegenden Stichprobe, als auch vom unbekannten Parameter φ ab! Zu jedem anderen Modell $M^{(j)}$ (typischerweise $j < k$) heißt

$$\mathfrak{D}\big(\beta^{(j)}, \varphi \big| y_1, .., y_n\big) := 2\left[l_{\text{sat}} - l^{(j)}\big(\beta^{(j)}, \varphi \mid y_1, .., y_n\big)\right] \tag{5.8}$$

Devianz (*Residuendevianz*) des Modells $M^{(j)}$. Ausgewertet bei $\widehat{\beta}^{(j)}, \varphi$ heißt sie *beobachtete Devianz* und hängt von der vorliegenden Stichprobe und dem Dispersionsparameter φ ab.

5.2 Modellwahl im klassischen LM

Eine Zielgröße sei beschrieben durch das *volle* lineare Modell mit Verteilungsfamilie $\mathcal{N}_{\mu,\sigma^2}$, Kovariablen $\mathbf{x} \in \mathcal{X} \subseteq \{1\} \times \mathbb{R}^{k-1}$ und unbekannten Parametern $\boldsymbol{\beta} \in B \subseteq \mathbb{R}^k$, $\sigma^2 \in (0, \infty)$. Zu bestimmen seien nun diejenigen Kovariablenkomponenten (*mögliche Einflussgrößen*) $M \subseteq \{1, .., k\}$, die *wirklich* einen Einfluss auf die Verteilung der Zielgröße haben.

Jede Wahl $M \subseteq \{1, .., k\}$ an mitbetrachteten Komponenten, entspricht für sich genommen einem anderen Modell der gleichen Zielgröße, das weniger oder mehr an die beobachteten Daten *angepasst* ist. In der Hinsicht ist die Wahl der wirklichen Einflussgrößen ein Spezialfall des allgemeineren Problems der Modellwahl. Mehr Informationen zur Modellwahl im linearen Fall finden sich in [6].

5.2.1 Vorwärts- & Rückwärtsselektion

Eine gegebenfalls vorhandene Hierarchie $x^1, x^2, .., x^k$ der Einflussgrößen spiegelt sich wieder in einer Hierarchie zwischen den infrage kommenden Modellen $M_l := \{1, .., l\}$, $l \leq k$. Solch eine Hierarchie, gibt Anlass zur Modellwahl durch die sogenannte *Vorwärtsselektion*, bei der beginnend mit dem einfachsten Modell M_1, die Zahl der Einflussgrößen (Dimension der Kovariablen) bis zu einer bestimmten Abbruchbedingung stückweise erhöht wird.

Eine typische Abbruchbedingung kann durch die den Hypothesentest aus Abschnitt 4.2.8 realisiert werden, mit Hilfe dessen die Hypothese $H_0 : \mathbf{e}_l\boldsymbol{\beta} = 0$, sprich, die *Unabhängigkeit* der Zielgröße von der l-ten Kovariablenkomponente, anhand des Modells M_l getestet wird. Die verwendete Stichprobe bzw. Kovariablenwerte sind für alle Modellkomplexitäten die gleiche:

for $l = 2$ to k **do**
  Test Hypothesis $H_0 : \mathbf{e}_l\boldsymbol{\beta} = 0$ for Modell M_l
  if Hypothesis is not rejected **then**
    return Model M_{l-1}
  end if
end for
return Model M_k

Ein ähnliches Verfahren ist das Rückwärtsselektionsverfahren, bei dem ausgehend vom vollen Modell M_k die Modellkomplexität gemäß der obigen Hierarchie stückweise verringert wird:

[22] Siehe 4.2.2.

for $l = k$ to 2 **do**
  Test Hypothesis $H_0 : \mathbf{e}_l\boldsymbol{\beta} = 0$ for Modell M_l
  if Hypothesis is rejected **then**
    return Model M_l
  end if
end for
return Model M_1

Beachte: Vorwärts- & Rückwärtsselektion führen **nicht** unbedingt auf das gleiche Modell.

Leider ist bei den potentiellen Einflussgrößen im Allgemeinen à priori keine Hierarchie vorgegeben, so dass zu jeder Komplexität $1 \leq l \leq k$ gleich mehrere Modellmöglichkeiten zur Verfügung stehen. Eine mögliche Lösung stellt folgende Variante der Vorwärtsselektion dar:

$M \leftarrow \{1\}$
for $l = 2$ to k **do**
  Choose new variable $j \notin M$, such that test statistic $\widehat{\beta}_j/\widehat{\sigma}$ maximizes for Model $M \cup \{j\}$
  Test Hypothesis $H_0 : \mathbf{e}_l\boldsymbol{\beta} = 0$ for Modell $M \cup \{j\}$
  if Hypothesis is not rejected **then**
    return Model M
  else
    $M \leftarrow M \cup \{j\}$
  end if
end for
return Model M

5.2.2 Modellwahl durch Minimierung des MSPSE

Der in Abschnitt 5.1.1 eingeführte bzw. in Abschnitt 5.1.2 behandelte Prognosefehler MSPSE, lässt sich als Auswahlkriterium zwischen den möglichen Modellen $M \subseteq \{1, .., k\}$ verwenden. Dabei werden natürlich stets die gleichen Kovariablenwerte $\mathbf{x}_1, .., \mathbf{x}_{n+m} \in \mathcal{X}$ angenommen. Da der wahre MSPSE_M zum Modell M unbekannt ist, muss er mit Hilfe der vorliegenden Stichproben geschätzt werden. Letztendlich wird das Modell $M \subseteq \{1, .., k\}$ ausgewählt, dass seinen Schätzwert MSPSE_M minimiert.

Folgende Herangehensweisen sind typisch:

1. Sei $y_1, .., y_n$ die vorhandene, den Kovariablenwerten $\mathbf{x}_1, .., \mathbf{x}_n \in \mathcal{X}$ entsprechende Stichprobe. Zu den Kovariablen $\mathbf{x}_{n+1}, .., \mathbf{x}_{n+m}$ wird eine neue Stichprobe $y_{n+1}, .., y_{n+m} \in E$ gezogen, und MSPSE_M hinsichtlich (5.1) durch

$$\widehat{\text{MSPSE}}_M := \sum_{i=1}^{m} (y_{n+i} - \widehat{y}_{n+i})^2 \tag{5.9}$$

   geschätzt, wobei $\widehat{y}_{n+i}$ die Punktprognosen basierend auf der alten Stichprobe für die neuen Kovariablenwerte $\mathbf{x}_{n+1}, .., \mathbf{x}_{n+m}$ sind.

2. Der vorliegende Datensatz $y_1, .., y_n$ wird im Voraus in zwei Hälften zerlegt. Die erste Hälfte wird als *Trainierdaten* zur Schätzung des Parameters $\boldsymbol{\beta}$ bzw. der Punktprognosen $\widehat{y}_{n+i}$ verwendet. Die zweite Hälfte wird als *Validierungsdaten* zur Schätzung von MSPSE_M wie im Fall (1) verwendet.

3. Der vorhandene Datensatz wird im Voraus in N ca. gleich-große Datensätze zerlegt. Zum r-ten Datensatz, werden die restlichen $(N-1)$ Datensätze als *Trainierdaten* zur Vorhersage des r-ten Datensatzes verwendet, um mit diesem dann gemäß Fall (1) den entsprechenden MSPSE_M zu Schätzen. Aus allen N Schätzwerten, wird der Mittelwert gebildet und als Vergleichsgröße zwischen den Modellen verwendet.

4. Sei RSS_M die der Stichprobe $y_1, .., y_n$ entsprechende Residuenquadratsumme im Modell $M \subseteq \{1, .., k\}$. Beziehung (5.6) liefert einen natürlichen Schätzer für $\mathrm{MSPSE}_M(\beta, \sigma^2; \mathbf{x}_1, .., \mathbf{x}_n; \mathbf{x}_1, .., \mathbf{x}_n)$:

$$\widehat{\mathrm{MSPSE}}_M := \mathrm{RSS}_M(y_1, .., y_n) + 2k \cdot \widehat{\sigma}^2(y_1, .., y_n) \tag{5.10}$$

der als Vergleich zwischen den Modellen M verwendet werden kann. Dabei sollte $\widehat{\sigma}^2$ die gleiche, möglichst unverzerrte Schätzung für σ^2 in allen Modellen sein. Sinnvollerweise wird deshalb $\widehat{\sigma}^2$ als ML-Schätzer im vollen Modell $\{1, .., k\}$ geschätzt.

Die Minimierung dieses Schätzwertes $\widehat{\mathrm{MSPSE}}_M$ unter den möglichen Modellen entspricht dabei der Minimierung des sogenannten *Mallowschen Komplexitätsparameters* (*Mallow's CP*):

$$\mathrm{CP}_M(y_1, .., y_n) := \frac{\mathrm{RSS}_M(y_1, .., y_n)}{\widehat{\sigma}^2(y_1, .., y_n)} - n + 2k \quad , \tag{5.11}$$

wobei auch hier $\widehat{\sigma}^2$ im vollen Modell geschätzt wird.

5.2.3 Das saturierte lineare Modell

Zu $k \in \mathbb{N}$ sei $M^{(k)}$ das klassisch, lineare Modell mit Verteilungsfamilie $\mathcal{N}_{\mu,\sigma^2}$, Kovariablen $\mathbf{x}^{(k)} \in \mathcal{X}^{(k)} \subseteq \{1\} \times \mathbb{R}^{k-1}$, unbekannten Parametern $\boldsymbol{\beta}^{(k)} \in B^{(k)} \subseteq \mathbb{R}^k$, $\sigma^2 \in (0, \infty)$ und deterministischen Teil $\mu_{\boldsymbol{\beta}^{(k)}}(\mathbf{x}) = \boldsymbol{\beta}^{(k)}\mathbf{x}^{(k)}$.

Sei $\mathbf{y} \in \mathbb{R}^n$ eine Stichprobe zu den *vollen* Kovariablenwerten $\mathbf{x}_1, .., \mathbf{x}_n \in \mathbb{R}^n$ so dass $P_k\mathbf{x}_i \in \mathcal{X}^{(k)}$ für jeden Projektor $P_k : \mathbb{R}^n \to \mathbb{R}^k$ in die ersten k Koordinaten. Seien $\widehat{y}^{(k)}_{n+i}$ die im Modell $M^{(k)}$ der Stichprobe $\mathbf{y}$ und Kovariablenwerten $P_k\mathbf{x}_1, .., P_k\mathbf{x}_n$ entsprechenden Punktprognosen. Nach Abschnitt 4.2.7 sind

$$\left(\widehat{y}^{(k)}_{n+1}, .., \widehat{y}^{(k)}_{2n}\right) = \mathbb{X}_{(k)}(\mathbb{X}^T_{(k)}\mathbb{X}_{(k)})^{-1}\mathbb{X}^T_{(k)}\mathbf{y} \quad , \tag{5.12}$$

mit $\mathbb{X}_{(k)} \in \mathbb{R}^{n\times k}$ als entsprechende Kovariablenmatrix. Das saturierte Modell[23] ist daher das Modell $M^{(k)}$ mit kleinstem $k \in \mathbb{N}$, so dass

$$\mathbf{y} \in \mathrm{image}\left\{\mathbb{X}_{(k)}\right\} \quad . \tag{5.13}$$

Zu erkennen ist, dass das saturierte Modell im schlimmsten Fall die Komplexität n besitzt. Für jedes andere Modell $M^{(j)}$ ist die beobachtete Devianz gegeben durch

$$\mathfrak{D}^{(j)} = \frac{\mathrm{RSS}_{(j)}}{\sigma^2} = \frac{1}{\sigma^2} \cdot \mathbf{y}^T \left[\mathbb{1} - \mathbb{X}_{(j)}(\mathbb{X}^T_{(j)}\mathbb{X}_{(j)})^{-1}\mathbb{X}^T_{(j)}\right] \mathbf{y} \quad . \tag{5.14}$$

mit $\mathrm{RSS}_{(j)}$ als Residuenquadratsumme. Ersetzt man die Werte $y_1, .., y_n$ durch die unabhängigen, gemäß dem Modell und den Kovariablenwerten $\mathbf{x}_1, .., \mathbf{x}_n$ verteilten $Y_1, .., Y_n$, so gilt nach (4.64):

$$\mathfrak{D}^{(j)} \sim \chi^2_{n-j} \quad . \tag{5.15}$$

Trotz möglicher Behauptungen in mancher Literatur, gilt (5.15) nicht für alle verallgemeinerten LM[8].

5.3 Modellwahl in allgemeineren Modellen

5.3.1 Der Likelihood-Quotienten-Test

Sei $(\mathcal{P}_{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\Theta}$ eine Familie von Wahrscheinlichkeitsmaßen parametrisiert durch $\boldsymbol{\theta} \in \Theta \in \mathbb{R}^k$ und $\Theta_0 \subseteq \Theta$. Dabei seien Θ_0, Θ offen. Zu $n \in \mathbb{N}$ definiere die so genannte *Likelihood Quotientenstatistik*

$$\mathcal{T}_n(y_1, .., y_n) := 2 \cdot \ln \frac{\sup_{\boldsymbol{\theta}\in\Theta} \mathcal{L}(\boldsymbol{\theta} \mid y_1, .., y_n)}{\sup_{\boldsymbol{\theta}\in\Theta_0} \mathcal{L}(\boldsymbol{\theta} \mid y_1, .., y_n)} \quad . \tag{5.16}$$

[23] Siehe Abschnitt 5.1.3.

Sei $D \subseteq \mathbb{R}^l,\ l \leq k$ offen und $\gamma : D \to \Theta_0$ ein $\mathcal{C}^2$-Diffeomorphismus. Sind nun $Y_1, Y_2, ..$ unabhängig, gemäß $\mathcal{P}_{\boldsymbol{\theta}}$ verteilt, so geht unter Annahme der Nullhypothese $H_0 : \boldsymbol{\theta} \in \Theta_0$:

$$\mathcal{T}_n(Y_1, .., Y_n) \xrightarrow[\mathrm{d}]{n\to\infty} \chi^2_{k-l} \quad . \tag{5.17}$$

[3] Die asymptotische Teststatistik $\mathcal{T}_n$ kann also für einen asymptotischen Test der Nullhypothese $H_0 : \boldsymbol{\theta} \in \Theta_0$ verwendet werden. Ist $\chi^2_{k-l,1-\alpha}$ das $(1-\alpha)$ Quantil der χ^2_{k-l} Verteilung, so gilt unter Annahme der Nullhypothese:

$$\lim_{n\to\infty} \mathcal{P}\left(\mathcal{T}_n(Y_1, .., Y_n) \geq \chi^2_{k-l,1-\alpha}\right) = \alpha \tag{5.18}$$

Die Teststatistik $\mathcal{T}_n$ stellt gewissermaßen einen Vergleich zwischen den erzielbaren Maximalwerten der Loglikelihood Funktion mit und ohne die Einschränkung $\boldsymbol{\theta} \in \Theta_0$ dar. Beziehung (5.18) besagt interpretationsgemäß, dass eine zu große Differenz der beiden Werte, sprich ein Überschreiten der Schranke $\chi^2_{k-l,1-\alpha}$, einen starken Hinweis auf die Unkorrektheit der Hypothese darstellt.

5.3.2 Der Likelihood-Quotiententest im LM

Betrachtet sei das klassische lineare Model $\mathcal{N}_{\mu,\sigma^2}$ mit unbekannten Parametern $\boldsymbol{\beta} \in B \subseteq \mathbb{R}^k,\ \sigma^2 \in (0, \infty)$ und Kovariablen $\mathbf{x} \in \mathcal{X} \subseteq \{1\} \times \mathbb{R}^{k-1}$. Die Varianz sei zunächst als konstant angenommen. Zu Testen sei die Nullhypothese $H_0 : \beta_{l+1}, .., \beta_k = 0$.

Zu Stichprobenumfang n und konkreten Kovariablenwerten $\mathbf{x}_1, .., \mathbf{x}_n \in \mathcal{X}$ seien $\widehat{\boldsymbol{\beta}}^0_n$ & RSS^0_n bzw. $\widehat{\boldsymbol{\beta}}_n$ & RSS_n jeweils die ML-Schätzer & Residuenquadratsummen für $\boldsymbol{\beta}$ mit bzw. ohne die Bedingung H_0. Seien $Y^i \sim \mathcal{N}_{\boldsymbol{\beta}\mathbf{x}_i,\sigma^2}$ unabhängig, $i = 1, .., n$. Dann ist die Statistik

$$\widetilde{\mathcal{T}}_n(\mathbf{Y}) := 2 \cdot \ln \frac{\mathcal{L}(\widehat{\boldsymbol{\beta}}_n(\mathbf{Y}), \sigma^2 \mid \mathbf{Y})}{\mathcal{L}(\widehat{\boldsymbol{\beta}}^0_n(\mathbf{Y}), \sigma^2 \mid \mathbf{Y})} = \frac{1}{\sigma^2}\left(\mathrm{RSS}^0_n(\mathbf{Y}) - \mathrm{RSS}_n(\mathbf{Y})\right) \tag{5.19}$$

verteilt gemäß χ^2_{k-l}. Ersetzt man den (unbekannten) σ^2 mit dem ML-Schätzer $\widehat{\sigma}^2_n$ des komplexeren Modells (sprich ohne H_0), erhält man aus (5.19) die Likelihood Quotientenstatistik

$$\mathcal{T}_n(\mathbf{Y}) = \frac{1}{\widehat{\sigma}^2_n(\mathbf{Y})}\left(\mathrm{RSS}^0_n(\mathbf{Y}) - \mathrm{RSS}_n(\mathbf{Y})\right) \quad . \tag{5.20}$$

Diese erfüllt

$$\frac{\mathcal{T}_n(\mathbf{Y})}{k-l} \sim F_{k-l,n-k} \tag{5.21}$$

wobei $F_{k-l,n-k}$ die Fisher-Verteilung[24] ist. Mit (5.21) lässt sich ein F-Test zur Nullhypothese konstruieren, der z.B. Modellwahl Anwendung findet.

5.3.3 Definition: Kullback-Leibler-Divergenz

Es seien $\mathcal{P}, \mathcal{Q}$ Wahrscheinlichkeitsmaße auf dem messbaren Raum $(E, \mathcal{A})$ so dass $\mathcal{P} \ll \mathcal{Q}$. Dann heißt

$$D_{\mathrm{KL}}(\mathcal{P} \mid \mathcal{Q}) := \int_E \ln \frac{d\mathcal{P}}{d\mathcal{Q}}\, d\mathcal{P} \tag{5.22}$$

(falls existent) *Kullback-Leibler-Divergenz* (oder *Kullback-Leibler-Abstand*[25]) *von $\mathcal{P}$ nach $\mathcal{Q}$*. Ähnlicherweise, sind p, q zwei Wahrscheinlichkeitsdichten auf dem messbaren Raum $(E, \mathcal{A})$ bzgl. eines zugrundeliegenden Maßes

[24] Die Fisher-Verteilung $F_{m,n}$ ist definiert als Verteilung einer reellen Zufallsgröße X/Y, wobei $mX \sim \chi^2_m$ und $nY \sim \chi^2_n$ unabhängig sind.

[25] Beachte dass D_{KL} nicht symmetrisch in seinen Argumenten ist!

μ, so dass $\{q = 0\} \subseteq \{p = 0\}$, dann heißt

$$D_{\mathrm{KL}}(p \mid q) := \int_E [\ln p - \ln q] \cdot p \; d\mu \quad . \tag{5.23}$$

(falls existent) *Kullback-Leibler-Divergenz von p nach q*. Die KL-Divergenz ist stets nicht-negativ, und es gilt $D_{\mathrm{KL}}(\mathcal{P} \mid \mathcal{Q}) = 0$ genau dann wenn $\mathcal{P} = \mathcal{Q}$.

5.3.4 Das Akaike Informationskriterium (AIC)

Betrachten Klasse an Wahrscheinlichkeitsdichten $(f_\theta)_{\theta\in\Theta}$ auf $\mathbb{R}$ (*Modell*). Sei g ebenfalls eine Wahrscheinlichkeitsdichte (*Realität*) auf $\mathbb{R}$ so dass $\{f = 0\} \subseteq \{g = 0\}$. Dann heißt

$$D_{\mathrm{KL}}\left(g \mid f_\theta : \theta \in \Theta\right) := \inf_{\theta\in\Theta} D_{\mathrm{KL}}\left(g \mid f_\theta\right) \tag{5.24}$$

Kullback-Leibler-Divergenz von g nach $(f_\theta)_{\theta\in\Theta}$. Sei nun $\Theta \subseteq \mathbb{R}^k$ offen, $\mathcal{L}_n$ die Likelihood Funktion und $\widehat{\theta}_n$ der eindeutige ML-Schätzer des Modells $(f_\theta)_{\theta\in\Theta}$ zu Stichproben mit Umfang $n \in \mathbb{N}$ und $g_n(\mathbf{y}) := \prod_{i=1}^n g(y^i),\ \mathbf{y} \in \mathbb{R}^n$. Es sei angenommen das ein eindeutiges $\theta_{\mathrm{KL}} \in \Theta$ existiert, so dass

$$D_{\mathrm{KL}}\left(g \mid f_{\theta_{\mathrm{KL}}}\right) = D_{\mathrm{KL}}\left(g \mid f_\theta : \theta \in \Theta\right) \quad . \tag{5.25}$$

Sind $Y_1, Y_2, ..$ unabhängige, gemäß g verteilte Zufallsgrößen, so lässt sich unter geeigneten Regularitätsvoraussetzungen zeigen, dass

$$\widehat{\theta}_n(Y_1, .., Y_n) \xrightarrow[\mathrm{p}]{n\to\infty} \theta_{\mathrm{KL}} \tag{5.26}$$

und

$$D_{\mathrm{KL}}\big[g_n \mid \mathcal{L}_n\big(\widehat{\theta}_n\big)\big] := \int_{\mathbb{R}^n} g_n(\mathbf{y}) \ln \frac{g_n(\mathbf{y})}{\mathcal{L}_n\big[\widehat{\theta}_n(\mathbf{y}) \mid \mathbf{y}\big]} \; d\mathbf{y} \overset{n\to\infty}{\longrightarrow} D_{\mathrm{KL}}\left(g \mid f_{\theta_{\mathrm{KL}}}\right) + \frac{k}{2} \tag{5.27}$$

ML-Schätzungen *minimieren* sozusagen den Kubblack-Leilber-Abstand des Modells zur *Realität*.

Ist $y_1, .., y_n$ eine Stichprobe unabhängiger, gemäß g verteilter Zufallsgrößen, so ist ein geeigneter Schätzer zur (unbekannten) Divergenz $D_{\mathrm{KL}}\big[g_n \mid \mathcal{L}_n\big(\widehat{\theta}_n\big)\big]$ gegeben durch

$$\widehat{D}_{\mathrm{KL}}\big[g_n \mid \mathcal{L}_n\big(\widehat{\theta}_n\big)\big] := -l(\widehat{\theta}_n(y_1, .., y_n) \mid y_1, .., y_n) + k + \int_{\mathbb{R}^n} g_n(\mathbf{y}) \ln g_n(\mathbf{y}) \; d\mathbf{y} \tag{5.28}$$

so dass für die vorgegebene Stichprobe $y_1, .., y_n$ (und genügend große n), das *beste* Modell dasjenige ist, das das so genannte *Akaike Informationskriterium* (AIC)

$$\widehat{\mathrm{AIC}}(y_1, .., y_n) := -2 \cdot l(\widehat{\theta}_n(y_1, .., y_n) \mid y_1, .., y_n) + 2k \tag{5.29}$$

minimiert. Häufig, aber nicht immer, liefern das AIC und Mallows CP (5.11) das gleiche *beste* Modell. Dazu kommt, dass die Wahl des letzteren abhängig von der Stichprobe und damit letztendlich *zufällig* ist.

5.4 Überdispersion

5.4.1 Definition: Überdispersion

Das Vorhandensein einer Variabilität der Messdaten weit über dem vom zugrunde liegenden Modell vorhergesagten Wert, heißt *Überdispersion*. Dies ist häufiger bei einfacheren Modellen der Fall, in denen die Verteilungsfamilie zu wenig freie Parameter enthält um die Varianz der empirischen Daten vollständig wiederzugeben (z.B. einfache Poissonfamilie). Die Überdispersion betrifft daher hauptsächlich den stochastischen Teil des Modells.

Alternativ zur Erhöhung der Modellkomplexität, kann dieses Problem durch eine Angabe allein einer Erwartungswert-Varianz Beziehung der Verteilung behandelt werden, z.B. in der Form

$$\mathrm{Cov}(\mathbf{Y}) = \varphi \cdot V(\mathbb{E}\mathbf{Y}) \tag{5.30}$$

mit der *Varianzfunktion* $V : \mathbb{R}^m \to \mathbb{R}^{m\times m}$ als hinreichend glatte Funktion und φ als sogenannten *Überdispersionsparameter*. Beachte die Ähnlichkeit zur Exponential-Dispersionsfamilie 3.2.1, wo genau $V(\mathbb{E}\mathbf{Y}) = \partial^2_{\boldsymbol{\theta}} A\big|_{\boldsymbol{\theta}(\mathbb{E}\mathbf{Y})}$. Im allgemeinen lässt sich allerdings durch (5.30) nicht auf die ursprüngliche Exponential-Dispersionsfamilie (bzw. allgemeine Verteilungsfamilie) schließen.

In den meisten Fällen ist $\mathbb{E}\mathbf{Y} = \boldsymbol{\mu}_\beta(x)$, gemäß dem deterministischen Modellteil abhängig von einer Kovariablen $x \in \mathcal{X}$ und einem (unbekannten) Regressionsparameter $\beta \in B$. Sowohl β als auch φ müssen dann anhand der Stichproben geschätzt werden.

5.4.2 Definition: Quasi-Likelihood-Funktion

Betrachtet sei ein Modell mit gegebenem deterministischen Teil $\mathbb{E}Y = \mu_{\boldsymbol{\beta}}(x)$, Kovariablen $x \in \mathcal{X}$, Regressionsparameter $\boldsymbol{\beta} \in B \subseteq \mathbb{R}^k$, ohne festgelegte Verteilungsfamilie. Unter Verwendung einer Varianz-Erwartungswert Beziehung wie in (5.30), lässt sich[26] ohne Voraussetzung einer entsprechenden Verteilungsfamilie, für gegebene Stichprobe $y_1, .., y_n$ zu konkreten Kovariablenwerten $x_1, .., x_n \in \mathcal{X}$ die so genannte *Quasi-Likelihood-Funktion*

$$l_q(\mu_1, ..., \mu_n \mid y_1, .., y_n) := \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - z}{\varphi \cdot V(\mu_i)}\, dz \tag{5.31}$$

in Analogie zur bekannten **Log**likelihood Funktion definieren[4, 5]. Fasst man die $\mu_i := \mu_{\boldsymbol{\beta}}(\mathbf{x}_i)$ als Funktionen der (festen) Kovariablen x_i und des Regressionsparameter $\boldsymbol{\beta} \in B \subseteq \mathbb{R}^k$ auf, so wird

$$l_q(\boldsymbol{\beta} \mid y_1, .., y_n) := K\big(\mu_{\boldsymbol{\beta}}(x_1), .., \mu_{\boldsymbol{\beta}}(x_n) \mid y_1, .., y_n\big) \tag{5.32}$$

zu einer Funktion des Parameters $\boldsymbol{\beta}$ mit Ableitung

$$S_q(\boldsymbol{\beta} \mid y_1, .., y_n) := \frac{\partial l_q}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{y_i - \mu_{\boldsymbol{\beta}}(x_i)}{\varphi \cdot V\left(\mu_{\boldsymbol{\beta}}(x_i)\right)} \cdot \frac{\partial \mu_{\boldsymbol{\beta}}(x_i)}{\partial \boldsymbol{\beta}} \tag{5.33}$$

in voller Analogie zur Score Funktion (4.31) im Falle einer Exponential-Dispersionsfamilie. Die Funktion (5.33) wird daher *Quasi-Score Funktion* genannt. Die Lösung der *Quasi-Score Gleichung* ergibt einen Schätzwert $\widehat{\boldsymbol{\beta}}_q$ für den unbekannten Parameter $\boldsymbol{\beta}$, der unabhängig vom Dispersionsparameter φ ist.

Beachte: In diesem Modell wird nicht angenommen dass die Quasi-Likelihood Funktion aus einer Verteilungsfamilie resultiert, insbesondere auch nicht, dass die Zielgröße Exponential-Dispersiv verteilt ist. Die Quasi-Likelihood Schätzmethode kann eingesetzt werden, wo zwar eine genaue *Kenntnis* der Verteilungsfamilie der Zielgröße fehlt, jedoch ein Zusammenhang zwischen Varianz und Erwartungswert vermutet wird. Unabhängig davon, wird natürlich stets ein deterministischer Teil benötigt.

Der Schätzwert $\widehat{\boldsymbol{\beta}}_q(y_1, .., y_n)$ hängt zwar bei gegebener Stichprobe nicht vom Dispersionsparameter φ ab, doch sehr wohl die Verteilung des Schätzers $\widehat{\boldsymbol{\beta}}_q(Y_1, .., Y_n)$. Ist z.B. $\widehat{\boldsymbol{\beta}}_q(\mathbf{y}) = \mathbb{B}\mathbf{y}$ linear abhängig von der Stichprobe $\mathbf{y} \in \mathbb{R}^n$ wobei $\mathbb{B} \in \mathbb{R}^{k\times n}$, so gilt

$$\mathbb{E}\widehat{\boldsymbol{\beta}}_q = \mathbb{B}\mathbb{E}\mathbf{Y} \quad , \quad \mathrm{Cov}\left(\widehat{\boldsymbol{\beta}}_q\right) = \mathbb{B}^T \underbrace{\mathrm{Cov}(\mathbf{Y})}_{\propto\varphi} \mathbb{B} \quad , \tag{5.34}$$

wie schon im linearen Modell 4.2.7 der Fall war.

[26] Wobei wir uns im folgenden auf reelle Zufallsgrößen beschränken.

6 Kontingenztafeln

6.0.3 Definition: Kontingenztafel

Eine Auflistung gemessener (relativer) Häufigkeiten von Kombinationen bestimmter *qualitativer Merkmale* (*Faktoren*, *Kenngrößen*) der Untersuchungsobjekte heißt *Kontingenztafel*. Die einfachste Form einer Kontingenztafel wäre eine $N \times M$ Tabelle von Häufigkeiten, entsprechend der $N \times M$ möglichen Kontingenzwerte zweier Kenngrößen, die jeweils N und M Werte annehmen können.

Tabelle 6.1 stellt ein Beispiel einer 2-dimensional Kontingenztafel dar.

	Weiblich	Männlich	Hermaphrodit	Zeilensumme
Blau	5	3	0	8
Braun	34	40	1	75
Schwarz	30	27	1	58
Grün	6	2	0	8
Spaltensumme	75	72	2	149

Abbildung 6.1: Fiktives Beispiel einer Kontingenztafel bzgl. des festgestellten Geschlechts und Augenfarbe bei 149 untersuchten Personen.

Sowohl die gemessenen Häufigkeiten als auch deren Gesamtzahl sind im allgemeinen Zufällig! Deren Verteilung bzw. mögliche Korrelation der Kenngrößen gilt es dann aus vorgegebenen Kontingenztafeln zu bestimmen.

Die Anzahl der betrachteten Merkmale entspricht der *Dimension* der Kontingenztafel. Im Falle einer Dimension größer als 2 werden Kontingenztafeln schnell unübersichtlich und mögliche Zusammenhänge sind oft nur schwer zu erkennen.

6.1 Stochastische Modellierung bei fester Gesamtzahl

6.1.1 Multinomialmodell

Betrachten die Verteilung von $k \in \mathbb{N}$ Kenngrößen, die jeweils die Werte $1, .., r_j$ annehmen können, $j = 1, .., k$. Deren Verteilung sei modelliert durch die, zunächst nicht unbedingt unabhängigen, Zufallsgrößen $X_1, .., X_k$.

Seien nun

$$Y_{i_1,..,i_k} \quad , \quad i_j \in \{1, .., r_j\}\,, \; j \in \{1, .., k\} \tag{6.1}$$

die (zufälligen) Häufigkeiten mit der die Tupel $(i_1, .., i_k)$ als Werte der Kontingenzwerte $(X_1, .., X_k)$ nach N unabhängigen Versuchen eintreten. Dann ist deren Gesamtheit (Kontingenztafel) multinomialverteilt mit Wahrscheinlichkeiten

$$p_{i_1,..,i_k} := \mathcal{P}\left(X_1 = i_1, .., X_k = i_k\right) \quad , \tag{6.2}$$

das heißt

$$\mathcal{P}\big(\,(Y_{1,..,1}, .., Y_{r_1,..,r_k}) = (y_{1,..,1}, .., y_{r_1,..,r_k})\,\big) = N! \prod_{i_1=1}^{r_1} \cdots \prod_{i_k=1}^{r_k} \frac{(p_{i_1..i_k})^{y_{i_1..i_k}}}{y_{i_1..i_k}!} \quad . \tag{6.3}$$

Unbekannte Parameter des Modells sind hier genau die Wahrscheinlichkeiten $p_{i_1,..,i_k}$. Deren Kenntnis würde unter anderem auch mögliche Abhängigkeiten zwischen den Kenngrößen aufdecken. Viele Ergebnisse aus Abschnitt 4 können hier sofort angewandt werden, wobei man beachten sollte dass die Zielgröße hier die gesamte Kontingenztafel ist, deren Verteilung durch die Einzelwahrscheinlichkeiten $p_{i_1..i_k}$ gegeben ist. Dabei gilt per

Konstruktion die Einschränkung

$$\sum_{i_1=1}^{r_1} \cdots \sum_{i_k=1}^{r_k} p_{i_1..i_k} = 1 \quad . \tag{6.4}$$

6.1.2 ML-Schätzung der Einzelwahrscheinlichkeiten im Multinomialmodell

Zu gegebener Realisierung $(y_{i_1..i_k})$ der Kontingenztafel aus 6.1 lautet nach (**??**) die Loglikelihood Funktion

$$l\big((p_{i_1..i_k}) \mid (y_{i_1..i_k})\big) = \ln N! + \sum_{i_1=1}^{r_1} \cdots \sum_{i_k=1}^{r_k} \Big[y_{i_1..i_k} \ln p_{i_1..i_k} - \ln\left[y_{i_1..i_k}!\right] \Big] \quad . \tag{6.5}$$

Die ML-Schätzwerte für $p_{i_1..i_k}$ ergeben sich durch Maximierung von (6.5) unter der Nebenbedingung (**??**), mit Hilfe der Methode der Lagrange Multiplikatoren als

$$\boxed{\widehat{p}_{i_1..i_k} = \frac{Y_{i_1..i_k}}{N} \quad , \quad i_j \in \{1,..,r_j\} \quad ,} \tag{6.6}$$

was genau der intuitiven Erwartung entspricht. Insbesondere stimmen die Prognosen $\widehat{y}_{i_1..i_k} = \widehat{p}_{i_1..i_k} \cdot N$ für eine wiederholte Messung der Kontingenztafel mit der ursprünglichen Überein, sprich, das Modell ist saturiert mit saturiertem Loglikelihood Wert

$$l_{\text{sat}}\left((y_{i_1..i_k})\right) = \ln N! + \sum_{i_1=1}^{r_1} \cdots \sum_{i_k=1}^{r_k} \left[y_{i_1..i_k} \ln \frac{y_{i_1..i_k}}{N} - \ln\left[y_{i_1..i_k}!\right] \right] \quad . \tag{6.7}$$

Unter der Nullhypothese H_0 der Unabhängigkeit der Kenngrößen $X_1, .., X_k$ muss gelten

$$H_0 \ : \ p_{i_1..i_k} = \prod_{j=1}^{k} \rho_{i_j}^j \tag{6.8}$$

wobei

$$\rho_{i_j}^j := \mathcal{P}\left(X_j = i_j\right) \quad , \quad i_j \in \{1,..,r_j\} \quad . \tag{6.9}$$

Die Loglikelihood Funktion (6.5) nimmt unter H_0 die Form

$$l_{H_0}\big((\rho_{i_j}^j) \mid (y_{i_1..i_k})\big) = \ln N! + \sum_{i_1=1}^{r_1} \cdots \sum_{i_k=1}^{r_k} \Big[y_{i_1..i_k} \ln \rho_{i_1}^1 \cdot .. \cdot \rho_{i_k}^k - \ln\left[y_{i_1..i_k}!\right] \Big] \quad . \tag{6.10}$$

an. Die ML-Schätzer für $\rho_{i_j}^j$ bzw. $p_{i_1..i_k}$ ergeben sich durch deren Maximierung unter den Einschränkungen

$$\sum_{i_j=1}^{r_j} \rho_{i_j}^j = 1 \quad , \quad j = 1,..,k \tag{6.11}$$

gemäß

$$\widehat{\rho}_{i_j}^j = \frac{\Sigma_j(i_j)}{N} \tag{6.12}$$

bzw.

$$\boxed{\widehat{p}_{i_1..i_k}\big|_{H_0} = \frac{1}{N^k} \prod_{j=1}^{k} \Sigma_j(i_j)} \tag{6.13}$$

wobei

$$\boxed{\Sigma_j(i_j) := \sum_{s_1=1}^{r_1} \cdots \sum_{s_{j-1}=1}^{r_{j-1}} \sum_{s_{j+1}=1}^{r_{j+1}} \cdots \sum_{s_k=1}^{r_k} y_{s_1..s_{j-1}i_j s_{j+1}..s_k}} \tag{6.14}$$

Die beobachtete Devianz ergibt dieses sich für dieses Modell als

$$\mathfrak{D}\big|_{H_0} = 2\left[l_{\text{sat}} - l_{H_0}\left(\left(\widehat{p}_{i_1..i_j}\big|_{H_0}\right)\right)\right] = 2\sum_{i_1=1}^{r_1} \cdots \sum_{i_k=1}^{r_k} y_{i_1..i_k} \cdot \ln\left[\frac{y_{i_1..i_k}}{N\widehat{p}_{i_1..i_k}\big|_{H_0}}\right] \tag{6.15}$$

wobei unter Annahme der Nullhypothese $D\big|_{H_0} \overset{N\to\infty}{\underset{\text{d}}{\longrightarrow}} \chi^2_{(r_1-1)\cdot\ldots\cdot(r_k-1)}$.

6.2 Stochastische Modellierung bei zufälliger Gesamtzahl

6.2.1 Poissonmodell

In Abschnitt 6.1 wurde davon ausgegangen, dass die Kontingenztafel durch die Beobachtung von N unabhängigen Kopien des Versuchsobjektes, sprich, N unabhängigen Versuchen, konstruiert wird. Dies ist z.B. bei einer Befragungsreihe von N Versuchspersonen oder einer Untersuchung einer festen Exemplarzahl bei Produktionslinien der Fall. Es könnte jedoch sehr wohl eine Situation vorliegen, in der die erfasste Personenzahl an sich auch zufällig ist.

Betrachten dazu den typischen Fall, dass N Poissonverteilt ist mit Erwartungswert $\overline{N}$. Dann sind die einzelnen Häufigkeiten $Y_{i_1..i_k}$ auch poissonverteilt mit Erwartungswerten $\mu_{i_1..i_n} := \overline{N} \cdot p_{i_1..i_k}$, das heißt

$$\mathcal{P}\left(Y_{i_1..i_k} = y_{i_1..i_k}\right) = \sum_{n=y_{i_1..i_k}}^{\infty} \mathcal{P}\left(Y_{i_1..i_k} = y_{i_1..i_k} \mid N = n\right) = \frac{e^{-\overline{N}p_{i_1..i_k}}}{y_{i_1..i_k}!} \cdot \left(\overline{N} \cdot p_{i_1..i_k}\right)^{y_{i_1..i_k}} \quad . \tag{6.16}$$

Die Schätzung der unbekannten Modellparameter $\overline{N}, p_{i_1..i_k}$ unter der Nebenbedingung (6.4) ist dabei äquivalent zur Schätzung der $\mu_{i_1..i_k}$.

6.2.2 ML-Schätzung der Einzelerwartungswerte im Poissonmodell

Zu gegebener Realisierung der Kontingenztafel $(y_{i_1..i_k})$ ist im Modell aus 6.2.1 die Loglikelihood Funktion gegeben durch

$$l\left((\mu_{i_1..i_k}) \mid (y_{i_1..i_k})\right) = \sum_{i_1=1}^{r_k} \cdots \sum_{i_k=1}^{r_k} \left[y_{i_1..i_k} \cdot \ln \mu_{i_1..i_k} - \ln\left[y_{i_1..i_k}!\right] - \mu_{i_1..i_k}\right] \quad . \tag{6.17}$$

Die ML-Schätzwerte für $\mu_{i_1..i_k}$ ergeben sich durch Maximierung von (6.17) als

$$\boxed{\widehat{\mu}_{i_1..i_k} = y_{i_1..i_k}} \tag{6.18}$$

ganz analog zu Abschnitt 6.1.2. Das Modell entspricht daher dem saturierten mit saturiertem Loglikelihood Wert

$$l_{\text{sat}} = \sum_{i_1=1}^{r_k} \cdots \sum_{i_k=1}^{r_k} \left[y_{i_1..i_k} \cdot \ln y_{i_1..i_k} - \ln\left[y_{i_1..i_k}!\right] - y_{i_1..i_k}\right] \quad . \tag{6.19}$$

Unter der Nullhypothese H_0 der Unabhängigkeit der einzelnen Kenngrößen muss analog zu Abschnitt 6.1.2 gelten

$$H_0 \;:\; \mu_{i_1..i_k} = \prod_{j=1}^{k} \nu_{i_j}^j \tag{6.20}$$

wobei

$$\nu_{i_j}^j := \sqrt[k]{N} \cdot \mathcal{P}\left(X_j = i_j\right) \quad , \quad i_j \in \{1, .., r_j\} \quad . \tag{6.21}$$

Die Loglikelihood Funktion (6.17) nimmt in diesem Modell die Form

$$l_{H_0}\left((\nu_{i_j}^j) \mid (y_{i_1..i_k})\right) = \sum_{i_1=1}^{r_1} \cdots \sum_{i_k=1}^{r_k} \left[y_{i_1..i_k} \cdot \sum_{j=1}^{k} \ln \nu_{i_j}^j - \ln\left[y_{i_1..i_k}!\right] - \prod_{j=1}^{k} \nu_{i_j}^j \right] \tag{6.22}$$

an. Durch Maximierung von (6.22) erhält man die ML-Schätzer für die $\nu_{i_1..i_k}$ bzw. für $\overline{N}, p_{i_1..i_j}$ gemäß

$$\widehat{\nu}_{i_j}^j = \left[\prod_{l \neq j} \sum_{s_l=1}^{r_l} \widehat{\nu}_{s_l}^l \right]^{-1} \cdot \Sigma_j(i_j) = \frac{\Sigma_j(i_j)}{N^{\frac{(k-1)}{k}}} \tag{6.23}$$

bzw.

$$\boxed{\widehat{\mu}_{i_1..i_k}\big|_{H_0} := \prod_{j=1}^{k} \widehat{\nu}_{i_j}^j = N^{1-k} \cdot \prod_{j=1}^{k} \Sigma_j(i_j) \quad , \quad \widehat{\overline{N}}\big|_{H_0} := \sum_{i_1=1}^{r_1} \cdots \sum_{i_k=1}^{r_k} \widehat{\mu}_{i_1..i_k} = N} \tag{6.24}$$

wobei

$$N := \sum_{i_1}^{r_1} \cdots \sum_{i_k=1}^{r_k} y_{i_1..i_k} \tag{6.25}$$

die Größe der Stichprobe und $\Sigma_j(i_j)$ wie in (6.14) definiert ist. Schließlich erhält man für $p_{i_1..i_k}$ die Schätzwerte

$$\boxed{\widehat{p}_{i_1..i_k}\big|_{H_0} := \widehat{\overline{N}}^{-1} \cdot \widehat{\mu}_{i_1..i_k}\big|_{H_0} = \frac{1}{N^k} \prod_{j=1}^{k} \Sigma_j(i_j)} \tag{6.26}$$

in voller Übereinstimmung mit Abschnitt 6.1.2. Die beobachtete Devianz dieses Modells ergibt sich als

$$\mathfrak{D} := 2\Big[l_{\text{sat}} - l_{H_0}\left((\widehat{\mu}_{i_1..i_k})\right)\Big] = \sum_{i_1=1}^{r_1} \cdots \sum_{i_k=1}^{r_k} y_{i_1..i_k} \cdot \ln\left[\frac{y_{i_1..i_k}}{\widehat{\mu}_{i_1..i_k}\big|_{H_0}}\right] \tag{6.27}$$

und ist gleich der beobachteten Devianz (6.15) im Multinomialmodell mit vornherein fester Beobachtungszahl N.

Literatur

[1] J. I. Myung, D. J. Navarro, *Information Matrix*
Ohio State University
http://faculty.psy.ohio-state.edu/myung/personal/info_matrix.pdf (12.06.2010)

[2] J. Rougier, *Lecture Notes on Statistics 2: Fisher Information*
October 2008
http://www.maths.bris.ac.uk/∼mazjcr/stats2/HOfisher.pdf (12.06.2010)

[3] P. J. Bickel, *Mathematical Statistics*
Holden Day, 1991

[4] R. W. M. Wedderburn, *Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method.*
Biometrika, Vol. 61, No. 3, pp. 439-447, 1974

[5] P. McCullagh, *Quasi-Likelihood Functions*
Ann. Stat., Vol. 11, Issue 1, pp. 59-67, 1983

[6] W. Stahel, *Lineare Regression - Vorlesungsunterlagen*
http://stat.ethz.ch/ stahel/courses/regression/reg1-script.pdf (21.06.2010)

[7] D. R. Cox, E. J. Snell, *Applied Statistics - Principles and Examples*
CRC Press, 1981

[8] P. McCullagh, J. Nelder, *Generalized Linear Models*
Chapman & Hall, 1983

Index