.

Kategoriale Regression und Loglineare Modelle[1]

U. Mortensen

FB Psychologie und Sportwissenschaften, Institut III
Westfälischen-Wilhelms-Universität Münster

[1] Letzte Korrektur 09. 03. 2016

Inhaltsverzeichnis

Kapitel 1

Schätzen und testen

Wie in den meisten Wissenschaften sind auch in der Psychologie Beobachtungen im allgemeinen mit zufälligen Effekten behaftet. Diese Effekte werden durch eine Wahrscheinlichkeitsverteilung (wenn die zufälligen Größen diskret sind) oder durch eine Wahrscheinlichkeitsdichte (wenn sie stetig sind) charakterisiert. Ist x eine Beobachtung, so sei der Wert der Wahrscheinlichkeitsverteilung bzw. - dichte durch $f(x|\theta)$ gegeben; es ist im folgenden nicht nötig, jeweils zu spezifizieren, ob f eine Verteilung oder eine Dichte ist. θ ist ein *Parameter* von f. Oft wird f durch mehr als einen Parameter charakterisiert. Dann soll $\theta = (\theta_1, \cdots, \theta_k)'$ einen Vektor von Parametern bezeichnen. θ repräsentiert Eigenschaft des psychischen Prozesses, der den Beobachtungen x unterliegt, und/oder repräsentiert den Einfluß von Bedingungen, unter denen die Beobachtungen gemacht werden. Dementsprechend wird man versuchen, aus den Beobachtungen den im allgemeinen unbekannten Wert von θ zu schätzen. Es ergeben sich nun drei Themenbereiche:

1. die Methoden zur Schätzung von Parametern,
2. die Eigenschaften von Parameterschätzungen,
3. der Test von Hypothesen über solche Schätzungen.

Die Eigenschaften von Schätzungen können von der Methode der Schätzung abhängen; sie sollen deshalb zuerst angegeben werden, damit sie bei der Charakterisierung von Methoden benannt werden können. Der Test von Hypothesen wird anschließend allgemein beschrieben.

1.1 Eigenschaften von Schätzungen

Es sei eine Stichprobe $x_1, \cdots, x_n$ von n Beobachtungen gegeben. Der Einfachheit halber soll sie als Vektor $X = (x_1, \cdots, x_n)'$ angeschrieben werden. $\theta = (\theta_1, \cdots, \theta_k)'$ sei der unbekannte Vektor von Parametern der Verteilung bzw. Dichte $f(x|\theta)$. Es

sei nun

$$\hat{\theta} = T(X) \tag{1.1}$$

eine Schätzung von θ. Damit soll ausgedrückt werden, dass $\hat{\theta}$ anhand der $x_i, i = 1, \cdots, n$, also anhand der Komponenten von X, berechnet wird und sich eben als eine Funktion $T(X)$ aus X ergibt. T wird durch die Schätzmethode festgelegt.

X ist eine Stichprobe und damit mit zufälligen Effekten behaftet. Damit variiert auch $T(X)$ zufällig von einer Stichprobe zur nächsten. Besteht θ nur aus einem einzelnen Parameter, so ist $T_n(X)$ eine zufällige Veränderliche,ist θ ein Vektor mit mehr als einer Komponente, so ist $T_n(X)$ ein zufälliger Vektor.

Definition 1 *Es sei n der Stichprobenumfang, d.h. die Anzahl er Komponenten von X, und es sei $T_n = E(T_n(X))$ der Erwartungswert bzw. der Erwartungswertvektor für $T_n(X)$. Es gelte*

$$E(T_n) = \theta + b \tag{1.2}$$

Dann heißt b der Bias *der Schätzung; gilt insbesondere $b = 0$, so heißt die Schätzung* biasfrei *oder* erwartungstreu. *Gilt*

$$\lim_{n\to\infty} E(T_n) = \theta, \tag{1.3}$$

so heißt die Schätzung T_n asymptotisch erwartungstreu.

Der Bias einer Schätzung ist offenbar ein systematischer Fehler. Ist die Schätzung T_n asymptotisch erwartungstreu, so ist $E(T_n) = \theta + b_n$ und b_n strebt mit wachsendem n gegeb Null.

Beispiel 1 Es gelte $x = \mu + e$, $E(X) = \mu$, $Var(X) = \sigma^2$. In Band I wurde gezeigt, dass $T_n = \overline{x}$ für alle n eine erwartungstreue Schätzung für μ ist, $T_n = s^2 = \sum_i (x_i - \overline{x})^2/n$ aber keine erwartungstreue Schätzung für σ^2 ist; diese Schätzung ist allerdings asymptotisch erwartungstreu. Eine erwartungstreue Schätzung für σ^2 ist durch

$$\hat{s}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

gegeben. □

Eine Schätzung T_n kann zwar erwartungstreu für alle n sein, für eine gegebene Stichprobe aber mehr oder weniger vom wahren Wert θ abweichen. Deshalb ist die Varianz $Var(T_n)$ der zufälligen Veränderlichen T_n von Interesse. Nun ist es intuitiv naheliegend, zu vermuten, dass die Schätzungen für θ um so genauer werden, je größer der Stichprobenumfang n ist. Dies muß nicht so sein, - es ist ja möglich, dass bei der Erhebung der Stichprobe systematische Fehler begangen werden, oder dass die Methode der Schätzung von θ systematische Fehlschätzungen impliziert. Die in der folgenden Definition charakterisierte Eigenschaft einer Schätzung ist also nicht notwendig erfüllt:

Definition 2 *Es sei T_n eine Schätzung für θ bei einer Stichprobe vom Umfang n, und es gelte*

$$\lim_{n\to\infty} E(T_n) = \theta, \quad \lim_{n\to\infty} Var(T_n) = 0. \tag{1.4}$$

Dann heißt die Schätzung T_n konsistent[1].

Beispiel 2 Es gelte wieder $x = \mu + e$, $E(x) = \mu, Var(x) = \sigma^2$, und das arithmetische Mittel $\overline{x} = \sum_i x_i/n$ ist eine Schätzung T_n für μ. Dann ist $Var(T_n) = Var(\overline{x}) = \sigma^2/n$. Die Schätzung $\overline{x}$ ist offenbar konsistent, denm $\lim_{n\to\infty} Var(T_n) = 0$. □

Die Realisierungen x_i der Stichprobe seien stochastisch unabhängig voneinander; die Wahrscheinlichkeit, dass die Werte $x_1, \cdots, x_n$ beobachtet werden, ist dann durch das Produkt der Wahrscheinlichkeiten bzw. Dichten für die individuellen x_i gegeben. Da diese Wahrscheinlichkeiten von θ abhängen, kann man die Wahrscheinlichkeit einer speziellen Stichprobe als Funktion von θ auffassen. Dieser Sachverhalt führt zu der folgenden

Definition 3 *Es sei $X = (x_1, \cdots, x_n)'$ eine Stichprobe von Werten einer zufälligen Veränderlichen mit der Wahrscheinlichkeitsverteilung bzw. Dichte f mit dem Parameter bzw. dem Vektor von Parametern θ. Dann heißt*

$$L(X|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{1.5}$$

die Likelihood *bzw. die* Likelihoodfunktion *für X.*

$$\log L(X|\theta) = \sum_{i=1}^{n} \log f(x_i|\theta) \tag{1.6}$$

heißt Log-Likelihoodfunktion.

Nun ist *Likelihood* einfach ein anderer Ausdruck für *probability*, d.h. für Wahrscheinlichkeit, und in der Tat wird ja mit (1.5) eine Wahrscheinlichkeit bzw. Dichte bezeichnet. Der Ausdruck *Likelihood* wird aber im Folgenden nur für die in (1.5) charakterisierte Größe gebraucht.

Über die Likelihoodfunktion kann die Güte einer Schätzung charakterisiert werden; eine Schätzung soll ja nicht nur erwartungstreu sein, sondern auch eine möglichst kleine Varianz haben, denn ie Varianz der Schätzung ist einm Maß für ihre Genauigkeit. Ist T die Schätzung eines Parameters θ, so kann über L eine untere Grenze für die Varianz $Var(T)$ der Schätzung bestimmt werden. Es gilt der

[1] Man kann zwischen schwacher und starker Konsistenz unterscheiden; darauf soll hier nicht weiter eingegangen werden.

Satz 1 *Es sei X eine Stichprobe vom Umfang n und $T_n(\theta)$ sei eine Schätzung des Parameters θ der Verteilung von X. $b(\theta)$ sei der Bias der Schätzung T_n und es sei $db(\theta)/d\theta = b'(\theta)$. Dann gilt*

$$Var(T_n) \geq \frac{1 + b'(\theta)}{E[(\partial \log L(\theta)/\partial\theta)^2]} \tag{1.7}$$

Beweis: Es wird zunächst angemerkt, dass (i) L eine Wahrscheinlichkeitsdichte ist, denn nach (1.5) ist L ja ein Produkt der $f(x_i|\theta)$ und steht damit für die Dichte von unabhängigen Werten x_i; (ii) L eine zufällige Veränderliche ist, denn die x_i liegen als Realisationen der zufälligen Veränderlichen X vor und $f(x_i|\theta)$ ist, als Funktion einer zufälligen Veränderlichen, selbst eine zufällige Veränderliche.

Es wird nun, der Kürze wegen, X statt $x_1, \cdots x_n$ geschrieben; in der Tat ist X der Vektor $(x_1, \cdots, x_n)'$. Da L eine Dichte ist, folgt

$$\int L\, dX = 1, \tag{1.8}$$

Da L eine Funktion von θ ist, kann man $\int L\, dX$ ebenfalls als Funktion von θ auffassen. Die Ableitung $\partial L/\partial\theta$ drückt dann die Veränderung von L mit θ aus, und ist wiederum eine zufällige Veränderliche. Dementsprechend kann man $\int \partial L/\partial\theta dX$ betrachten. Das Integral ist aber nach (1.8) eine Konstante, und dementsprechend muß

$$\int \frac{\partial L}{\partial\theta}\, dX = 0 \tag{1.9}$$

gelten. Weiter gilt sicherlich

$$\frac{\partial L}{\partial\theta} = \frac{L}{L}\frac{\partial L}{\partial\theta} = \frac{1}{L}\frac{\partial L}{\partial\theta}L$$

Andererseits ist aber

$$\frac{1}{L}\frac{\partial L}{\partial\theta} = \frac{\partial \log L}{\partial\theta}$$

Zur Abkürzung werde

$$V = \frac{\partial \log L}{\partial\theta}$$

gesetzt. Man erhält

$$\int \frac{\partial L}{\partial\theta}\, dX = \int \frac{\partial \log L}{\partial\theta} L\, dX$$

Die rechte Seite ist aber gerade gleich dem Erwartungswert von V, denn V ist ja auch eine Funktion von X und L ist die Dichte von X. Also ist

$$E(V) = E\left(\frac{\partial \log L}{\partial\theta}\right) = \int \frac{\partial \log L}{\partial\theta} L\, dX = 0 \tag{1.10}$$

Weiter ist $Var(V) = E(V^2) - E^2(V)$; da nun $E(V) = 0$ ist, muß also $Var(V) = E(V^2)$ sein, d.h.

$$Var(V) = Var\left(\frac{\partial \log L}{\partial \theta}\right) = E\left(\left(\frac{\partial \log L}{\partial \theta}\right)^2\right) \tag{1.11}$$

Für die Kovarianz der Schätzung T_n und der Variablen V gilt allgemein $Kov(T_n, V) = E(TV) - E(T)E(V)$. Da $E(V) = 0$, folgt also $Kov(T_n, V) = E(T_n V)$, d.h.

$$Kov(T_n, V) = \int T_n V L \, dX = \int T_n \frac{\partial \log L}{\partial \theta} L \, dX$$

Andererseits ist $E(T_n) = \theta + b(\theta) = \int T_n L \, dX$ und

$$\begin{aligned} \frac{\partial E(T_n)}{\partial \theta} &= \int T_n \frac{\partial L}{\partial \theta} \, dX = \int T_n \frac{1}{L} \frac{\partial \log L}{\partial \theta} L \, dX \\ &= \int T_n V L \, dX = Kov(T_n, V) \end{aligned}$$

und wegen $\partial E(T_n)/\partial \theta = 1 + b'(\theta)$ folgt

$$Kov(T_n, V) = 1 + b'(\theta) \tag{1.12}$$

Für die Korrelation zwischen T_n und V gilt aber $\rho^2_{T_n V} = Kov^2(T_n, V)/(Var(T_n)Var(V)) \leq 1$ und damit $Kov^2(T_n, V)/Var(V) \leq Var(T_n)$, so dass

$$\frac{Kov^2(T_n, V)}{Var(V)} = \frac{1 + b'(\theta)}{Var(V)} \leq Var(T_n)$$

Setzt man hier den Ausdruck (1.11) für $Var(V)$ ein, so hat man die Aussage des Satzes. □

Die in (1.7) auftretende Größe $E[(\partial \log L(\theta)/\partial \theta)^2]$ hat einen speziellen Namen:

Definition 4 *Es sei X eine Stichprobe vom Umfang n und die Log-Likelihoodfunktion $\log L(X|\theta)$ sei mindestens zweimal bezüglich θ differenzierbar; da X zufällig ist, ist auch $\log L$ eine zufällige Veränderliche und damit auch $\partial^2 \log L/\partial \theta^2$. Der Erwartungswert (falls er existiert)*

$$I(\theta) := E\left[\left(\frac{\partial \log L}{\partial \theta}\right)^2\right] \tag{1.13}$$

heißt Information *in der Stichprobe.*

Anmerkungen:

1. Es gilt
$$E\left[\left(\frac{\partial \log L}{\partial \theta}\right)^2\right] = -E\left(\frac{\partial^2 \log L}{\partial \theta^2}\right) \tag{1.14}$$
Die Information wird dementsprechend auch durch den Ausdruck auf der rechten Seite von (1.14) definiert. Ist θ ein Vektor von Parametern, d.h. ist $\theta = (\theta_1, \cdots, \theta_p)'$, so ist die Information, der rechten Seite entsprechend, durch
$$I(\theta) = -E\left(\frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j}\right) \tag{1.15}$$
gegeben; die Matrix $I(\theta)$ heißt dann *Informationsmatrix*, oder *Hesse-Matrix*.
Um die Gleichung (1.14) einzusehen, differenziert man (1.10) ein zweites Mal nach θ, so erhält man
$$\int \left[\left(\frac{1}{L}\frac{\partial L}{\partial \theta}\right)\frac{\partial L}{\partial \theta} + L\frac{\partial}{\partial \theta}\left(\frac{1}{L}\frac{\partial L}{\partial \theta}\right)\right] dX = 0$$
Nun ist
$$\left(\frac{1}{L}\frac{\partial L}{\partial \theta}\right)\frac{\partial L}{\partial \theta} = \frac{L}{L^2}\left(\frac{\partial L}{\partial \theta}\right)^2$$
und
$$\frac{\partial}{\partial \theta}\left(\frac{1}{L}\frac{\partial L}{\partial \theta}\right) = \frac{\partial^2 \log L}{\partial \theta^2}$$
so dass
$$\int \left(\left(\frac{L}{L^2}\left(\frac{\partial L}{\partial \theta}\right)^2 + \frac{\partial^2 \log L}{\partial \theta^2}\right)\right) dX = 0$$
Dies ist nur möglich, wenn der Integrand verschwindet, d.h. es gilt (1.14).

2. Die Ungleichung (1.7) heißt auch *Cramér-Rao'sche* oder *Cramér-Rao-Frechetsche* Ungleichung. Sie kann dann in der Form
$$Var(T_n) \geq \frac{1 + b'(\theta)}{I(\theta)} \tag{1.16}$$
geschrieben werden.

Der Ausdruck *Information* für die Größe $I(\theta) = E[(\partial \log L/\partial \theta)^2]$ ergibt sich aus (1.16): die Varianz $Var(T_n)$ ist um so geringer, je größer $I(\theta)$ in der Stichprobe ist. Je genauer sich ein Parameter θ durch den aus den Messungen $x_1, \cdots, x_n$ bestimmten Größe T_n schätzen läßt, desto mehr "Information" enthalten dann die x_i über θ. insbesondere eignet sich $I(\theta)$ als Maß für diese Information, denn $L(X|\theta)$ ist ja eine Funktion der x_i, die von dem Parameter θ abhängt. Variiert L und damit auch $\log L$ für gegebene $X = (x_1, \cdots, x_n)'$ stark mit dem Wert von θ, so wird $\partial \log L/\partial \theta$ große Werte annehmen. Dies heißt ja, dass es eine ausgeprägte Abhängigkeit der

Wahrscheinlichkeit der Daten X von dem Parameter θ gibt, so dass eben auch umgekehrt die Messungen X für *bestimmte* Werte von θ sprechen; die Stichprobe X enthält dann eben viel Information über θ. Analog folgt für eine geringe Abhängigkeit der Daten X von θ, dass $\partial \log L/\partial\theta$ klein sein wird, denn eine geringe Abhängigkeit bedeutet ja, dass sich L bzw. $\log L$ nur wenig mit θ ändert. Dann enhält X nur wenig Information über θ und die Varianz der Schätzung T_n wird groß sein. Diese Deutung von $I(\theta)$ gilt nur dann uneingeschränkt, wenn $b'(\theta)$ in (1.16) verschwindet, insbesondere wenn $b(\theta) = b'(\theta) = 0$ gilt.

Die Information $I(\theta)$ ist aber unabhängig von der Schätzung T_n definiert. Damit können verschiedene Schätzungen, d.h. anhand verschiedener Methoden gewonnener Schätzungen T_n^1 und T_n^2 miteinander verglichen werden. Generell nennt man die Schätzung T_n^1 *wirksamer* als die Schätzung T_n^2, wenn $Var(T_n^1) \leq Var(T_n^2)$ gilt. Speziell gilt

Definition 5 *Es sei $\{T_n(\theta)\}$ eine Klasse von Schätzern für den Parameter θ. $T_n^o(\theta)$ heißt* wirksamster Schätzer *der Klasse, wenn T_n^o die kleinste Varianz der Schätzer in der Klasse hat. Insbesondere heißt T_n^o* effizient*, wenn T_n^o die kleinste Varianz in der Klasse der erwartungstreuen Schätzungen hat und für die das Gleichheitszeichen in der Ungleichung (1.7) gilt.*

Beispiel 3 Es sei $x \sim N(\mu, 1)$-verteilt, und $X = (x_1, \cdots, x_n)'$ sei eine Stichprobe dieser Variablen. Die Likelihoodfunktion ist dann durch

$$L(X|\mu, 1) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2}\right)$$

gegeben, und die Log-Likelihoodfunktion durch

$$\log L = -\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2} + \log\left(\frac{1}{2\pi}\right)^{n/2}$$

Dann ist

$$\begin{aligned} \frac{\partial \log L}{\partial \mu} &= \sum_{i=1}^{n}(x_i - \mu) \\ \frac{\partial^2 \log L}{\partial \mu^2} &= -n \end{aligned}$$

d.h. $I(\mu) = n$. Das arithmetische Mittel $\overline{x}$ ist eine Schätzung für μ, und $Var(\overline{x}) = 1/n$ und folglich $Var(T_n) = 1/I(\mu)$. Da $1/I(\mu)$ die untere Schranke für $Var(T_n) = Var(\overline{x})$ ist, kann es keine Schätzung für μ mit kleinerer Varianz als $Var(\overline{x})$ geben, mithin ist die Schätzung $\overline{x}$ für μ effektiv. □

Sicherlich ist es nützlich, wenn eine Schätzung T_n alle Information über den zu schätzenden Parameter θ, die in einer Stichprobe X enthalten ist, auch ausnützt.

Enthält also T_n alle Information über θ, so sollte die bedingte Wahrscheinlichkeit oder die bedingte Dichte der Daten X, gegeben T_n, gleich der bedingten Wahrscheinlichkeit bzw. Dichte von X, gegeben T_n sein; Kenntnis von θ sollte zur Vorhersage von X nicht mehr notwendig sein. Diese Eigenschaft läßt sich wie folgt charakterisieren:

Definition 6 *Es sei $X = (x_1, \cdots, x_n)'$ eine Stichprobe und θ sei ein zu schätzender Parameter. $T_n(\theta) = T(x_1, \cdots, x_n)$ sei eine Schätzung für θ. T_n heißt eine* hinreichende *oder* suffiziente *Schätzfunktion für θ, wenn die bedingte Verteilung $P(X|T_n = \hat{\theta})$ unabhängig von θ ist.*

Beispiel 4 Es werden n Bernoulli-Versuche gemacht; für den i-ten Versuch wird durch $x_i = 1$ ein "Erfolg" angezeigt, durch $x_i = 0$ ein "Mißerfolg". Die Wahrscheinlichkeit eines "Erfolges" sei $p(x_i = 1) = p, i = 1, \cdots, n$. Die Wahrscheinlichkeit von $S = x_1 + \cdots + x_n = k$ Erfolgen ist dann

$$P(S = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Es werde die bedingte Wahrscheinlichkeit einer bestimmten Folge $X = (x_1, x_2, \cdots, x_n)'$ unter der Bedingung, dass $S = k$ ist, betrachtet. Es ist

$$P(X|S = k) = \frac{p(X \cap S = k)}{p(S = k)}$$

Für $x_1 + x_2 + \cdots + x_n \neq k$ ist sicherlich $p(X \cap S = k) = 0$. Nun sei $x_1 + x_2 + \cdots + x_n = k$. Dann ist

$$p(X \cap S = k) = p(x_1 + \cdots x_n = k) = p^k (1-p)^{n-k}$$

Also ist

$$p(X|S = k) = \frac{p^k (1-p)^{n-k}}{p(S = k)} = \frac{1}{\binom{n}{k}}$$

Die bedingte Wahrscheinlichkeit $P(X|S = k)$ ist also unabhängig vom Parameter p. Andererseits ist $(x_1 + \cdots + x_n)/n = k/n$ eine Schätzung für p. Für diese Schätzung ist es unerheblich, in welcher Reihenfolge die "Erfolge" und "Mißerfolge" auftreten, die einzige Eigenschaft der Stichprobe $X = (x_1, \cdots, x_n)'$, die in die Schätzung eingeht, ist der Wert k der Summe der Werte. Bezüglich der Schätzung ist die Kenntnis des Wertes k bereits *hinreichend*, – also suffizient. □

Im Falle der Binomialverteilung ist $T(x_1, \cdots, x_n) = \sum_i x_i/n$ eine Schätzfunktion für den Parameter $\theta = p$. Die Frage ist, ob sich Schätzfunktionen generell bezüglich der Suffizienz der Schätzung charakterisieren lassen. Diese Frage wird im folgenden Satz beantwortet:

Satz 2 *Es sei T_n eine Schätzfunktion für θ. T_n ist suffizient genau dann, wenn sich die Likelihoodfunktion L in zwei Funktionen L_1 und L_2 faktorisieren läßt derart,*

dass L_1 nur von θ und von T_n abhängt und L_2 nur von θ und von den Stichprobenwerten $x_1, \cdots, x_n$ abhängt, d.h. wenn

$$L(X|\theta) = L_1(T(X), \theta)L_2(X) \tag{1.17}$$

gilt.

Anmerkung: L_1 soll also *nicht* explizit von X, sondern nur implizit, d.h. über den Wert von T_n, von X abhängen. Die Aussage des Satzes wird auch als *Faktorisierungskriterium* bezeichnet.

Beweis:

Beispiel 5 Die zufällige Veränderliche x sei $N(\mu, \sigma^2)$-verteilt, und $X = (x_1, \cdots, x_n)'$ sei die vorliegende Stichprobe; $\overline{x}$ sei die Schätzung für μ, und s^2 sei die Schätzung für σ^2. Die Likelihoodfunktion hat die Form

$$L(X|\mu, \sigma^2) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

Nun ist $\sum_i (x_i - \mu)^2 = \sum_i (x_i - \overline{x} + \overline{x} - \mu)^2$, oder

$$\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}(x_i-\overline{x}+(\overline{x}-\mu))^2 = \sum_{i=1}^{n}(x_i-\overline{x})^2+n(\overline{x}-\mu)^2-2(\overline{x}-\mu)\sum_{i=1}^{n}(x_i-\overline{x})$$

Aber $\sum(x_i-\overline{x}) = 0$ und $s^2 = \sum_i (x_i-\overline{x})^2/n$ so dass $\sum_i (x_i-\mu)^2 = n(s^2+(\overline{x}-\mu)^2)$. Dann ist

$$L(X|\mu, \sigma^2) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{n(s^2 + (\overline{x} - \mu)^2)}{2\sigma^2}\right)$$

Dies bedeutet, dass L von den Werten $x_1, \cdots, x_n$ nur über die Schätzungen $\overline{x}$ und s^2 abhängt. Das Faktorisierungskriterium ist damit erfüllt, denn man muß nur $L_2 = 1$ setzen; L_1 erfüllt, wie eben gezeigt, die Forderung des Kriteriums. $\overline{x}$ und s^2 sind also suffiziente Schätzungen für μ und σ^2. □

Definition 7 *Es sei $T_n(X) = T(x_1, \cdots, x_n)$ eine Schätzung für θ; θ kann ein Vektor $(\theta_1, \cdots, \theta_p)'$ sein. Die Schätzung T_n heißt* linear*, wenn T_n eine lineare Funktion der Beobachtungen X ist.*

Lineare Schätzungen spielen eine zentrale Rolle in der Regressionsrechnung und bei der Diskussion von Mittelwertsunterschieden; Beispiele für solche Schätzungen werden also in Zusammenhang mit diesbezüglichen Fragestellungen gegeben.

1.2 Schätzmethoden

1.2.1 Die Momentenmethode

Es seien p Parameter $\theta_1, \cdots, \theta_p$ zu schätzen. Aus der Stichprobe $X = (x_1, \cdots, x_n)'$ werden nun die ersten p Momente berechnet. Aus der angenommenen Verteilung der x_i werden nun p Gleichungen hergeleitet, in denen die Momente und die θ_k zueinander in Beziehung gesetzt werden. Diese Gleichungen werden dann nach den θ_k aufgelöst; die Lösungen sind die *Momentenschätzungen* $\hat{\theta}_k$ für die θ_k, $k = 1, \cdots, p$.

Die Stichprobenmomente sind konsistente Schätzungen der Momente der Verteilung, und deswegen sind die Momentenschätzungen für die θ_k konsistent. Sie sind aber nicht notwendig auch effizient. Sie werden deshalb gelegentlich als Startwerte für die Suche nach effizienten Schätzungen benützt.

Beispiel 6 X sei eine Stichprobe von $N(\mu, \sigma^2)$-verteilten Messungen. Dann sind $\overline{x}$ und s^2 Momentenschätzungen für μ und σ^2, denn $E(x) = \mu, Var(x) = \sigma^2$. □

Beispiel 7 Die zufällige Veränderliche X sei binomialverteilt mit den Parametern p und n, d.h. es werden n Bernoulli-Versuche durchgeführt, und es ist $x_i = 1$ mit der Wahrscheinlichkeit p und $x_i = 0$ mit der Wahrscheinlichkeit $1 - p$. Dann ist $X = x_1 + \cdots + x_n$ und $E(X) = \mu = np$, $Var(X) = \sigma^2 = np(1-p)$.

Nun ist $\overline{X} = \sum_i x_i/n = k$, wenn k die Anzahl der Versuche ist, bei denen $x_i = 1$ resultierte. Man kann nun $\hat{\mu} = \overline{X}$ gemäß der Momentenmethode als Schätzung für μ wählen. Nun ist $\overline{X}$ der durchschnittliche x_i-Wert, während sich $E(X) = np$ auf die Summe $X = x_1 + \cdots x_n$ bezieht. Dementsprechend hat man $n\hat{p} = n\overline{X}$ oder $\hat{p} = \overline{X} = k/n$. Für σ^2 verfährt man analog. □

1.2.2 Maximum-Likelihood-Schätzungen

Die Verteilung bzw. Dichte von x sei durch $f(x|\theta)$ gegeben, wobei θ ein unbekannter Vektor von Parametern sei. θ soll anhand einer Stichprobe $X = (x_1, \cdots, x_n)'$ geschätzt werden. Man kann nun von der Überlegung ausgehen, dass ja im allgemeinen diejenigen Werte X auftreten werden, die am wahrscheinlichsten sind. Deswegen stellt derjenige Vektor $\hat{\theta}$ eine mögliche Schätzung von θ dar, für den die Wahrscheinlichkeit der Beobachtungen X maximal wird. Diese Wahrscheinlichkeit ist aber durch die oben definierte Likelihood gegegen:

$$L(X|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{1.18}$$

sofern die x_i unabhängig voneinander sind.

Es werde also angenommen, dass die x_i stochastisch unabhängig sind. Dann kann $L(X|\theta)$ als Funktion von θ betrachtet werden:

Definition 8 *Diejenige Schätzung $\hat{\theta}$, für die die Likelihoodfunktion $L(X|\theta)$ maximal wird, heißt* Maximum-Likelihood-Schätzung *für θ.*

Nun ist der Logarithmus von L eine monoton wachsende Funktion von L. Dementsprechend kann man ebenso die log-Likelihoodfunktion $\log L$ bezüglich θ maximieren. Rechnerisch hat dieses Vorgehen oft Vorteile.

Das Maximum von L bzw. $\log L$ bezüglich θ findet man, indem man etwa $l(\theta) = \log L(\theta)$ nach θ differenziert und die Ableitung gleich Null setzt; diese Gleichung ist für die Maximum-Likelihood-Schätzung $\hat{\theta}$ erfüllt:

$$\left.\frac{\partial l(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0 \tag{1.19}$$

(1.19) heißt *Maximum-Likelihood-Gleichung* (*ML-Gleichung*).

Ist $\theta = (\theta_1, \cdots, \theta_p)'$ ein Vektor von Parametern, so erhält man einen Vektor

$$\vec{s}(\theta; X) := \frac{\partial l(\theta)}{\partial \theta} = \frac{1}{L}\frac{\partial L}{\partial \theta} \tag{1.20}$$

von ersten Ableitungen; dieser Vektor $s(\theta; X)$ heißt auch *Score-Funktion*. Die Gleichungen

$$\vec{s}(\theta; X) = 0$$

bilden ein System von ML-Gleichungen (vergl. (1.19)), und die Matrix $\partial^2 \log L / \partial\theta\partial\theta'$ heißt wieder *Informationsmatrix*.

Ein für inferenzstatistische Fragen sehr interessantes Resultat ist, dass ML-Schätzungen asymptotisch normalverteilt sind (Cramér (1946)); es gilt der

Satz 3 *Die zufälligen Variablen $x_1, \cdots, x_n$ seien unabhängig und identisch verteilt und θ sei der Vektor der Parameter der Verteilung der x_i, $i = 1, \cdots, n$. Unter Bedingungen B ist der Vektor $\hat{\theta}$der ML-Schätzungen asymptotisch konsistent und (multivariat) normalverteilt, d.h.*

$$(\hat{\theta} - \theta) \xrightarrow{D} N(0, I^{-1}(\theta)), \tag{1.21}$$

wobei $I^{-1}(\theta)$ die Inverse der Informationsmatrix ist, und $\xrightarrow{D}$ stochastische Konvergenz bedeutet.

Beweis: Der Beweis beruht auf einer Anwendung des Zentralen Grenzwertsatzes; Details findet man in Kendall und Stuart (1973), (Vol. II), Kapitel 18.

Die Bedingungen B beziehen sich auf die Differenzierbarkeit der Log-Likelihoodfunktion und auf Eigenschaften der Erwartungswerte der Ableitungen; sie werden im folgenden als gegeben vorausgesetzt. Die Bedingungen werden in Kendall und Stuart (1973) ausführlich diskutiert.

Für den Fall, dass nur ein Parameter θ geschätzt wird, gilt also für $n \to \infty$

$$\hat{\theta} = \theta + \xi, \quad E(\xi) = 0, \quad Var(\xi) = -1/E(\frac{\partial^2 \log L}{\partial \theta^2}) \tag{1.22}$$

Die Schätzung $\hat{\theta}$ ist also *für* $n \to \infty$ unverzerrt (d.h. sie hat keinen Bias); für kleinere Werte von n muß die Schätzung aber nicht unverzerrt sein.

Beispiel 8 Die zufällige Veränderliche X sei binomialverteilt, also $X \sim B(n,p)$, so dass

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$\theta = p$ ist der unbekannte Parameter. Die Likelihoodfunktion ist durch

$$L(X|p) = \binom{n}{k} p^k (1-p)^{n-k}$$

gegeben, und die Log-Likelihoodfunktion durch

$$\log L(X|p) = k \log p + (n-k) \log(1-p) + \log \binom{n}{k} \tag{1.23}$$

Dann ist

$$\frac{d \log L}{dp} = \frac{k}{p} - \frac{n-k}{1-p}\Big|_{p=\hat{p}} = 0$$

Also ist

$$\frac{n-k}{k} = \frac{1-\hat{p}}{\hat{p}}$$

woraus

$$\hat{p} = \frac{k}{n} \tag{1.24}$$

folgt.

Es werde noch der Fall $n \to \infty$ betrachtet, d.h. der Fall, dass n "groß" ist. Nach (1.22) ist dann die Schätzung $\hat{p}$ angenähert normalverteilt mit dem Erwartungswert p (Unverzerrtheit) und einer Varianz, die gemäß (1.22) bestimmt wird, indem man den Reziprokwert des Erwartungswertes von $\partial^2 \log L / \partial p^2$ bestimmt. Aus (1.23) folgt

$$\frac{\partial^2 \log L}{\partial p^2} = -\frac{k}{p} - \frac{n-k}{(1-p)^2}$$

Dann ist

$$-E\left(\frac{\partial^2 \log L}{\partial p^2}\right) = \frac{E(k)}{p^2} + \frac{n - E(k)}{(1-p)^2}$$

Aber $E(k) = np$, so dass

$$-E\left(\frac{\partial^2 \log L}{\partial p^2}\right) = \frac{np}{p^2} + \frac{n-np}{(1-p)^2} = \frac{n}{p} + \frac{n}{1-p} = n, \quad \frac{1}{p(1-p)}$$

Also folgt

$$Var(\hat{p}) = \frac{p(1-p)}{n}. \tag{1.25}$$

□

Beispiel 9 X sei die zufällige Veränderliche, die die Anzahl der Unfälle an einer Autobahnabfahrt während einer Woche repräsentiert. Da die Unfälle in guter Näherung unabhängig von einander geschehen und nur bekannt ist, dass $X \geq 0$, wird angenommen, dass die Verteilung von X durch die Poisson-Verteilung gegeben ist:

$$p(X = k) = e^{-\lambda}\frac{\lambda^k}{k!} \tag{1.26}$$

λ ist wieder der unbekannte Parameter, der geschätzt werden soll für verschiedene Abfahrten kann λ verschieden sein, und man will die Unfallträchtigkeit der verschiedenen Abfahrten untersuchen, oder man will die Wirksamkeit unfallverhütender Maßnahmen untersuchen).

Man kann wieder

$$L = e^{-\lambda}\frac{\lambda^k}{k!} \tag{1.27}$$

setzen; dies ist die Likelihood für die Beobachtung einer Woche. Dann ist

$$\log L = -\hat{\lambda} + k\log\hat{\lambda} - \log k!$$

und es ist

$$\frac{d\log L}{d\lambda} = -1 + \frac{k}{\lambda}$$

Dann ist

$$\left.\frac{d\log L}{d\lambda}\right|_{\lambda=\hat{\lambda}} = -1 + \frac{k}{\hat{\lambda}} = 0$$

woraus

$$\hat{\lambda} = k \tag{1.28}$$

folgt. Bei der Poisson-Verteilung ist also die Anzahl der Ereignisse bereits die beste (gemäß dem ML-Prinzip) Schätzung für den unbekannten Parameter λ.

Es soll noch die Varianz der Schätzung gemäß (1.22) bestimmt werden. Es ist

$$\frac{d^2\log L}{d\lambda^2} = -\frac{k}{\lambda^2}$$

und

$$E\left(-\frac{d^2\log L}{d\lambda^2}\right) = \frac{E(k)}{\lambda^2} = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}$$

so dass

$$Var(\hat{\lambda}) = \lambda. \tag{1.29}$$

Es werde nun angenommen, dass man die Abfahrt an mehreren Wochenenden, etwa n, beobachtet habe, und man habe die Unfallhäufigkeiten $k_1, \cdots, k_n$ gezählt. Weiter

werde angenommen, dass die k_i jedesmal gemäß der Verteilung (1.26) verteilt sind, der Parameter λ also für alle Zeitabschnitte gleich groß ist. Die Wahrscheinlichkeit, gerade diese Stichprobe von k_i-Werten zu bekommen, ist dann

$$L = \prod_{i=1}^{n} f(k_i) = \prod_{i=1}^{n} \left(e^{-\lambda} \frac{\lambda^{k_i}}{k_i!} \right) = e^{-n\lambda} \prod_{i=1}^{n} \frac{\lambda^{k_i}}{k_i!} \tag{1.30}$$

Dann ist

$$\log L = -n\lambda + \sum_{i=1}^{n} k_i \log \lambda - \sum_{i=1}^{n} \log k_i!$$

und

$$\frac{d \log L(\lambda)}{d\lambda} = -n + \frac{1}{\lambda} \sum_{i=1}^{n} k_i$$

Durch Nullsetzen dieser Gleichung erhält man daraus die Schätzung

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} k_i \tag{1.31}$$

also gerade das arithmetische Mittel der k_i. Für die Varianz dieser Schätzung ergibt sich

$$\frac{d^2 \log L}{d\lambda^2} = -\frac{1}{\lambda^2} \sum_{i=1}^{n} k_i$$

und

$$E\left(-\frac{d^2 \log L}{d\lambda^2} \right) = \frac{1}{\lambda^2} \sum_{i=1}^{n} E(k_i) = \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda},$$

so dass

$$Var(\hat{\lambda}) = \frac{\lambda}{n} \tag{1.32}$$

Es gibt eine weitere Möglichkeit, zu einer Schätzung von λ zu kommen, wenn die Anzahl der Wochenenden hinreichend groß ist. Man zählt zunächst aus, wie häufig Wochenenden mit genau i Unfällen auftraten, $i = 0, 1, 2, \cdots, m$, wobei m die größte Anzahl von Unfällen ist, die an einem Wochenende beobachtet wurde. Es sei k_i die Häufigkeit von Wochenenden mit genau i Unfällen. Die Bedeutung von k_i ist hier also eine andere als im vorangegangenen Fall: dort war k_i die Anzahl der Unfälle am i-ten Wochenende, hier ist k_i die Anzahl der Wochenende mit genau i Unfällen. Es wird angenommen, dass die Anzahl der Unfälle Poisson-verteilt ist mit einem Parameter λ, der für alle Wochenenden gleich ist. Die Wahrscheinlichkeit für i Unfälle an einem Wochenende ist

$$p(X = i) = e^{-\lambda} \frac{\lambda^i}{i!}, \quad i = 0, 1, 2, \cdots$$

Die Wahrscheinlichkeit, die Anzahlen $k_0, k_1, \cdots, k_n = m$ von Unfällen zu beobachten, ist durch die Likelihood

$$L = \prod_{i=0}^{m} \left(e^{-\lambda} \frac{\lambda^i}{i!} \right)^{k_i} = \prod_{i=0}^{m} \left(e^{-k_i \lambda} \frac{\lambda^i}{i!} \right) \tag{1.33}$$

gegeben. Dann ist

$$\log L = -\lambda \sum_i k_i + \log \lambda \sum_i i k_i$$

und

$$\frac{d \log L}{d\lambda} = -\sum_i k_i + \frac{1}{\lambda} \sum_i i k_i$$

Es ist

$$\left. \frac{d \log L}{d\lambda} \right|_{\lambda = \hat{\lambda}} = 0$$

für

$$\hat{\lambda} = \frac{\sum_i i k_i}{\sum_i k_i} = \frac{1}{N} \sum_i i k_i, \quad N = \sum_i k_i \tag{1.34}$$

Die Varianz der Schätzung ist

$$Var(\hat{\lambda}) = \frac{\lambda}{\sum_i i E(k_i)} = \frac{2\lambda}{n(n+1)} \tag{1.35}$$

□

Beispiel 10 (Multinomialverteilung) Gegeben sei eine Multinomialverteilung für k Kategorien mit den Parametern $\pi_1, \pi_2, \cdots, p_k$; hie sind die θ_i durch die π_i gegeben. Da natürlich

$$\sum_{i=1}^{k} \pi_i = 1$$

gilt, folgt, dass es hier nur $k-1$ freie Parameter gibt; für p_k etwa gilt

$$p_k = 1 - \sum_{i=1}^{k-1} \pi_i$$

Es seien die Häufigkeiten $n_1, \cdots, n_k$ beobachtet worden, mit

$$N = \sum_{i=1}^{k} n_i, \quad n_k = N - \sum_{i=1}^{k-1} n_i$$

Die Verteilung kann in der Form

$$p(n_1, n_2, \cdots, n_k | i_1, \cdots, i_k) = c(n_1, \cdots, n_k) \prod_{i=1}^{k-1} \pi_i^{n_i} \left(1 - \sum_{i=1}^{k-1} \pi_i \right)^{n_k}$$

geschrieben werden; hierbei ist $c(n_1, \cdots, n_k)$ ein Faktor, der nur von den n_i abhängt. Dann ist jedenfalls die Likelihoodfunktion durch

$$L(n_1, \cdots, n_k | \pi_1, \cdots, p_k) = c \prod_{i=1}^{k-1} \pi_i^{n_i} \left(1 - \sum_{i=1}^{k-1} \pi_i \right)^{n_k}, \tag{1.36}$$

gegeben, denn die rechte Seite ist ja gerade die Wahrscheinlichkeit dafür, die Häufigkeiten $n_1, n_2, \cdots, n_k$ zu beobachten. Die Loglikelihoodfunktion ist dann durch

$$\log L = \sum_{i=1}^{k-1} n_i \log \pi_i + n_k \log \left(1 - \sum_{i=1}^{k-1} \pi_i\right) + \log c \tag{1.37}$$

gegeben. Differenziert man nach $\pi_i, i = 1, \cdots, k-1$ und setzt die Ableitungen gleich Null, so erhält

$$\frac{\partial \log L}{\partial \pi_i} = \frac{n_i}{\pi_i} - \frac{n_k}{1 - \sum_{i=1}^{k-1} \pi_i}$$

Setzt man diese partiellen Ableitungen gleich Null, so erhält man die Schätzungen

$$\hat{\pi}_i = n_i \frac{p_k}{n_k} = n_i c, \quad c = \frac{p_k}{n_k} \tag{1.38}$$

Wegen $\sum_{i=1}^{k} \pi_i = 1$ folgt aber

$$\sum_{i=1}^{k} \pi_i = \sum_{i=1} k n_i c = cN = 1,$$

so dass schließlich

$$\hat{\pi}_i = \frac{n_i}{N}, \quad i = 1, \cdots, k \tag{1.39}$$

folgt; die ML-Schätzungen der π_i sind also gerade wieder durch die relativen Häufigkeiten der Zellbesetzungen gegeben. □

Beispiel 11 x sei $N(\mu, \sigma^2)$-verteilt, und es sei die Stichprobe $X = (x_1, \cdots, x_n)'$ gegeben. Gesucht sind die MLS für $\theta = (\theta_1, \theta_2)'$, wobei $\theta_1 = \mu$ und $\theta_2 = \sigma^2$ ist.

Die Log-Likelihoodfunktion ist durch

$$\log L(X|\theta) = -\sum_{i=1}^{n} \frac{(x_i - \theta_1)^2}{2\theta_2} - n \log \sqrt{2\pi\theta_2}$$

gegeben. Über die partiellen Ableitungen nach θ_1 bzw. θ_2 ergeben sich die folgenden Gleichungen:

$$\frac{\partial \log L}{\partial \theta_1} = 2\frac{\sum_{i=1}^{n}(x_i - \theta_1)}{2\theta_2} = \frac{n}{\theta_2}(\overline{x} - \theta_1) \tag{1.40}$$

$$\frac{\partial \log L}{\partial \theta_2} = \frac{\sum_{i=1}^{n}(x_i - \theta_1)^2}{2\theta_2^2} - \frac{n}{2\theta_2} \tag{1.41}$$

$$\frac{\partial^2 \log L}{\partial \theta_1 \partial \theta_2} = -\frac{n(\overline{x} - \theta_1)}{\theta_2^2} \tag{1.42}$$

Natürlich ist

$$\frac{\partial^2 \log L}{\partial \theta_2 \partial \theta_1} = \frac{\partial^2 \log L}{\partial \theta_1 \partial \theta_2}$$

Durch Nullsetzen der Gleichung (1.40) folgt

$$\hat{\theta}_1 = \overline{x}. \tag{1.43}$$

Das arithmetische Mittel ist also im Falle normalverteilter Variablen auch eine MLS für μ.

Setzt man (1.41) gleich Null, so erhält man

$$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = s^2 \tag{1.44}$$

Die Stichprobenvarianz ist also eine MLS für den Parameter σ^2. Man sieht, dass hier die ML-Schätzung keine Schätzung liefert, die *für alle* n unverzerrt (biasfrei) ist; eine unverzerrte Schätzung ist bekanntlich durch $\hat{s}^2 = \sum_i (x_i - \overline{x})^2/(n-1)$ gegeben. Aber die Schätzung ist konsistent, denn für $n \to \infty$ wird sie biasfrei.

Um die Varianz-Kovarianzmatrix der Schätzungen zu finden, wird zuerst die Informationsmatrix $I(\theta)$ bestimmt. Es ist

$$I(\theta) = -E\left(\frac{\partial^2 \log L}{\partial\theta_i \partial\theta_j}\right), \quad i = 1, 2$$

d.h.

$$I(\theta) = -E\begin{pmatrix} -n/\theta_2 & -n/(\overline{x}-\theta_1)/\theta_2^2 \\ -n(\overline{x}-\theta_1)/\theta_2^2 & -\sum_i (x_i-\theta_1)^2/\theta_2^3 + n/\theta_2^2 \end{pmatrix}$$

Bestimmt man die Erwartungswerte der Elemente dieser Matrix, so findet man $E(n(\overline{x}-\theta_1)/\theta_2^2) = 0$, so dass

$$I(\theta) = \begin{pmatrix} n/\theta_2 & 0 \\ 0 & n/\theta_2^2 - \theta_2/\theta_2^3 \end{pmatrix}$$

Die Varianz-Kovarianzmatrix der Schätzungen ist dann durch

$$I^{-1}(\theta) = \begin{pmatrix} \sigma^2/n & 0 \\ 0 & \sigma^4/(n-1) \end{pmatrix} \tag{1.45}$$

gegeben. Die Schätzungen für μ und σ^2 sind also bei der Normalverteilung unkorreliert. σ^2/n ist (wie bereits bekannt!) die Varianz der Schätzung $\overline{x}$ für θ_1, und $\sigma^4/(n-1)$ ist die Varianz der Schätzung für $\theta_2 = \sigma^2$. □

Die Matrix $I^{-1}(\theta)$ liefert hier also die Varianzen und Kovarianzen der Schätzungen der Komponenten von θ.

1.2.3 Die Methode der Kleinsten Quadrate (KQ)

Es werde die Variable Y gemessen, von der angenommen wird, dass sie als Funktion $\phi(X_1, \cdots, X_n)$ der ebenfalls gemessenen Variablen $X_1, \cdots, X_n$ darstellbar sei. Die

X_j, $1 \leq j \leq n$ sind nicht notwendig Realisationen von zufälligen Veränderlichen. Die Funktion(en) ϕ möge(n) weiter von unbekannten Parametern $\theta_1, \cdots, \theta_p$ abhängen, Die θ_k, $1 \leq k \leq p$ sollen so geschätzt werden, dass die Werte y_i der Variablen Y durch die "Vorhersagen" anhand der Funktion $\phi(x_{i1}, \cdots, x_{in}|\theta_1, \cdots, \theta_p)$ so gut wie nur irgend möglich sind; die x_{ij} sind dabei Werte der X_j, die zu den Werten $y_i, i = 1, \cdots, m$ von Y korrespondieren. Man macht den folgenden Ansatz:

$$y_i = \phi(x_{i1}, \cdots, x_{in}|\theta_1, \cdots, \theta_p) + e_{ij} \tag{1.46}$$

wobei e_{ij} der Fehler ist, der bei der Vorhersage von y_{ij} durch die ϕ begangen wird. Die unbekannten Parameter $\theta = (\theta_1, \cdots, \theta_p)'$ sollen so geschätzt werden, dass

$$Q(\theta) = \sum_{ij} e_{ij}^2 = \sum_{ij} (y_{ij} - \hat{\pi}_{ij})^2 \tag{1.47}$$

minimal wird. Man erreicht dies, indem man die Gleichungen

$$\left.\frac{\partial Q}{\partial \theta_r}\right|_{\theta_r = \hat{\theta}_r} = 0, \quad 1 \leq r \leq p \tag{1.48}$$

nach den $\hat{\theta}_r$ auflöst.

Ein außerordentlich wichtiger Fall ergibt sich, wenn die Funktionen ϕ linear sind. Die Gleichungen (1.46) nehmen dann die Form

$$y_j = \theta_0 + \theta_1 x_{j1} + \cdots + \theta_p x_{jp} + \epsilon_j \tag{1.49}$$

an. Die Beziehung zwischen den y_{ij} und den x_{ir} ist dann durch ein *lineares Modell* gegeben.

Das lineare Modell (1.49) kann in wesentlich kompakterer Form geschrieben werden, wenn von der Matrix- bzw. Vektorschreibweise Gebrauch gemacht wird. Es sei $y = (y_1, \cdots, y_m)'$, $X = (x_{jr})$ mit $j = 1, \cdots, m$ und $r = 1, \cdots, p$, $\theta = (\theta_1, \cdots, \theta_p)'$ und schließlich $\epsilon = (\epsilon_1, \cdots, \epsilon_p)'$. Dann läßt sich (1.49) in der Form

$$Y = X\,\theta + e \tag{1.50}$$

schreiben. Es ist $e = Y - X\,\theta$. Nun ist das Skalarprodukt

$$e'e = \sum_{j=1}^{n} \epsilon_j^2 = (Y - X\,\theta)'(Y - X\,\theta) = Q(\theta)$$

und das Prinzip der Kleinsten Quadrate kann in der Form

$$(Y - X\,\hat{\theta})'(Y - X\,\hat{\theta}) = \min_{\theta}(Y - X\,\theta)'(Y - X\,\theta) \tag{1.51}$$

geschrieben werden.

1.3 Likelihood-Quotienten-Tests

1.3.1 Allgemeine Definition

Die Wahrscheinlichkeiten $P(T(x) \in \Omega_1|H_0)$ und $P(T(x) \in \Omega_1|H_1)$ können über die Likelihoods der Messungen bzw. Daten $x = (x_1, \cdots, x_n)'$ ausgedrückt werden. Es sei $L(x|H_0)$ die Likelihood von x, wenn H_0 zutrifft, und $L(x|H_1)$ sei die Likelihood von x, wenn H_1 zutrifft; die tatsächliche Berechung der Likelihoods setzt natürlich die Kenntnis der entsprechenden Dichten voraus. Jedenfalls gilt allgemein

$$P(T(x) \in \Omega_1|H_0) = \alpha = \int_{\Omega_1} L(x|H_0)\, dx \tag{1.52}$$

und

$$P(T(x) \in \Omega_0|H_1) = \beta = \int_{\Omega_0} L(x|H_1)\, dx. \tag{1.53}$$

Für die Macht oder Güte gilt dann auch

$$1 - \beta = \int_{\Omega_1} L(x|H_1)\, dx \tag{1.54}$$

Der Integrand in 1.54 kann nun mit $L(x|H_0)$ erweitert werden; man erhält

$$1 - \beta = \int_{\Omega_1} \frac{L(x|H_1)}{L(x|H_0)} L(x|H_0)\, dx \tag{1.55}$$

Der hier auftretende Quotient

$$\Lambda(x) = \frac{L(x|H_1)}{L(x|H_0)} \tag{1.56}$$

heißt *Likelihood-Quotient*. Da x zufällig ist, ist $\Lambda(x)$ eine zufällige Veränderliche. $L(x|H_0)$ kann als Dichte von x unter der Bedingung, dass H_0 gilt, aufgefaßt werden. Dann bedeutet (1.55), dass die Güte $1 - \beta$ als Erwartungswert von $\Lambda(x)$ bezüglich Ω_1 interpretiert werden kann.

Es seien H_0 und H_1 einfache Hypothesen, und es soll anhand der Daten x eine Entscheidung bezüglich der Frage, welche dieser Hypothesen gilt, getroffen werden. Es ist intuitiv klar, dass man hierzu den Wert von $\Lambda(x)$ betrachten kann. Spricht x mehr für H_1, d.h. ist $L(x|H_1) > L(x|H_0)$ und damit $\Lambda(x) > 1$, so kann man sich für H_1 entscheiden, andernfalls für H_0. Die Frage ist, ob diese Entscheidung optimal ist; die Irrtumswahrscheinlichkeiten α und β können zu groß sein. Man kann nun umgekehrt vorgehen und den Wert von α vorgeben, wobei man z.B. Betrachtungen über die Kosten zugrundelegt, die entstehen, wenn man irrtümlich H_1 akzeptiert. Natürlich entstehen auch Kosten, wenn man irrtümlich H_0 akzeptiert. Dementsprechend wird man fordern, dass für einen gewählten Wert von α der Wert von $1 - \beta$ maximal sein soll, d.h. die Wahrscheinlichkeit, H_1 zu akzeptieren, wenn H_1 zutrifft, soll relativ zum Wert von α so groß wie nur irgend möglich sein. Es gilt dann der

Satz 4 *(Neyman-Pearson (1933)[2] Gegeben seien zwei einfache Hypothesen H_0 und H_1. Für vorgegebenes α existiert dann eine Zahl c_α derart, dass die Wahrscheinlichkeit $1-\beta$ des Ereignisses*

$$\Lambda(x) = \frac{L(x|H_1)}{L(x|H_0)} > c_\alpha \tag{1.57}$$

maximal ist.

Einen ausführlichen Beweis des Satzes findet man etwa in Lindgren (1962), p. 239; an dieser Stelle muß darauf nicht weiter eingegangen werden, zumal man es im allgemeinen mit zusammengesetzten Alternativhypothesen zu tun hat.

Beispiel 12 Die Stichprobe $x = (x_1, \cdots, x_n)'$ stamme aus einer normalverteilten Grundgesamtheit, die Dichte sei durch

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\mu)^2\right)$$

gegeben, d.h. die Daten seien $N(\mu, 1)$-verteilt. Betrachtet werden die Hypothesen $H_0 : \mu = \mu_0$ und $H_1 : \mu = \mu_1$ mit $\mu_1 \neq \mu_0$; H_1 ist also 2-seitig. Die Likelihood der Daten unter der Hypothese H_i, $i = 0, 1$ ist dann durch

$$L(x|H_i) = \prod_{j=1}^{n} f(x_i) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{j=1}^{n}(x_j - \mu_i)^2\right)$$

Es ist aber

$$\begin{aligned}
\sum_j (x_j - \mu)^2 &= \sum_j (x_j - \overline{x} + \overline{x} - \mu_i)^2 \\
&= \sum_j (x_j - \overline{x})^2 + \sum_j (\overline{x} - \mu_i)^2 - 2\sum_j (x_j - \overline{x})(\overline{x} - \mu_i) \\
&= n(s^2 + (\overline{x} - \mu_i)^2)
\end{aligned} \tag{1.58}$$

Dann ist

$$\frac{L(x|H_1)}{L(x|H_0)} = \exp\left(-\frac{n}{2}((\overline{x} - \mu_0)^2 + (\overline{x} - \mu_1)^2)\right) = \exp\left(\frac{n}{2}((\mu_0 - \mu_1)2\overline{x} + (\mu_1^2 - \mu_0^2))\right)$$

und H_0 wird verworfen, wenn

$$\frac{n}{2}((\mu_0 - \mu_1)2\overline{x} + (\mu_1^2 - \mu_0^2)) \geq c_\alpha \tag{1.59}$$

Man sieht, dass hier die $x_1, \cdots, x_n$ nur über den Mittelwert $\overline{x}$ eingehen; dies ist wieder ein Ausdruck der Tatsache, dass $\overline{x}$ eine suffiziente Statistik ist.

[2] Neyman, J., Pearson, E.S.: On the problem of most efficient tests of statistical hypotheses. Phil. Trans. A, **231**, 289

Der Wert von c_α ist noch nicht bestimmt. Um ihn zu finden, braucht man nur zu bedenken, dass es muß einen Wert $\overline{x}_\alpha$ geben muß, für den

$$\int_{\overline{x}_\alpha}^{\infty} L(x|H_0)\,dx = \alpha$$

ist. Nach Definition von $L(x|H_0)$ ist dann

$$\int_{\overline{x}_\alpha}^{\infty} = \left(\frac{n}{2\pi}\right)^{n/2} \exp\left(-\frac{n(\overline{x}-\mu_0)^2}{2}\right)^2)$$

Dem Wert von $\overline{x}_\alpha$ entspricht wieder ein z_α mit $P(Z > z_\alpha) = \alpha$, und

$$z_\alpha = \frac{\overline{x}_\alpha - \mu_0}{1/n}$$

und daraus folgt $\overline{x}_\alpha = z_\alpha/n + \mu_0$. Man entscheidet sich für H_1, wenn $\overline{x} > \overline{x}_\alpha$. Andererseits soll nach (1.59) für H_1 entschieden werden, wenn

$$\frac{c_\alpha - (\mu_1^2 - \mu_0^2)}{(\mu_0 - \mu_1)/n} \leq \overline{x}$$

gilt. Also muß

$$\overline{x}_\alpha = \frac{c_\alpha - (\mu_1^2 - \mu_0^2))}{(\mu_0 - \mu_1)/n}$$

gelten. Hieraus kann man sich den Wert von c_α ausrechnen. Das ist natürlich nicht nötig, denn die Entscheidung kann ja bereits anhand des Wertes von $\overline{x}_\alpha$ getroffen werden. □

Das Wesentliche an den vorangegangenen Betrachtungen ist, dass sie zeigen, dass die Entscheidungen über H_0 und H_1 Entscheidungen anhand eines Likelihood-Quotienten äquivalent sind. Dazu mußte der Quotient selbst eigentlich nicht berechnet werden. Bei vielen statistischen Entscheidungen geht man aber von Likelihood-Quotienten aus, um zu Prüfstatistiken zu gelangen.

Die Verteilung der x hänge von den Parametern $\theta = (\theta_r, \theta_s)$ ab, und es soll die Hypothese

$$H_0 : \theta_r = \theta_{r0} \tag{1.60}$$

getestet werden. Die θ_{r0} sind spezielle Werte der θ_r. Ist $s = 0$, so ist die Hypothese einfach, andernfalls ist sie zusammengesetzt. Die Alternativhypothese sei

$$H_1 : \theta_r \neq \theta_{r0} \tag{1.61}$$

Es können nun Maximum-Likelihood-Schätzungen $\hat{\theta}_r, \hat{\theta}_s$ für θ_r und θ_s gefunden werden, so dass $L(x|\hat{\theta}_r, \hat{\theta}_s)$ maximal wird. Weiter kann man die Maximum-Likelihood-Schätzungen $\hat{\hat{\theta}}_s$ für θ_s finden unter der Annahme bzw. Bedingung, dass H_0 gilt. Dieses sind *bedingte* ML-Schätzungen. Betrachtet werde nun der Likelihood-Quotient

$$\Lambda(x) = \frac{L(x|\theta_{r0}, \hat{\hat{\theta}}_s)}{L(x|\hat{\theta}_r, \hat{\theta}_s)} \tag{1.62}$$

Sicher ist $\Lambda(x) \geq 0$, da die Likelihoods ja Wahrscheinlichkeiten bzw. Wahrscheinlichkeitsdichten sind. Andererseits muß

$$L(x|\theta_{r0}, \hat{\hat{\theta}}_s) \leq L(x|\hat{\theta}_r, \hat{\theta}_s)$$

sein, denn die Likelihood $L(x|\hat{\theta}_r, \hat{\theta}_s)$ ist ja maximal, weil $\hat{\theta}_r, \hat{\theta}_s$ ML-Schätzungen sind, und weil die Wahl θ_{r0} supobtimal sein kann. Also muß

$$0 \leq \Lambda(x) \leq 1 \tag{1.63}$$

gelten. Weiter kann $\Lambda(x)$ als Maximum-Likelihood der Daten unter H_0 betrachtet werden als Anteil der maximal möglichen Likelihood. "Hinreichend" große Werte für $\Lambda(x)$ zeigen also an, dass H_0 korrekt, oder besser: akzeptierbar ist. Man kann also nach einem kritischen Wert c_α fragen, für den die Wahrscheinlichkeit, dass bei Geltung von H_0 das Ereignis $\Lambda(x) \leq c_\alpha$ beobachtet wird, gerade α ist. Dieser Wert läßt sich bestimmen, wenn man die Verteilung bzw. Dichte des Quotienten Λ kennt. Es sei g diese Dichte. Dann ist jedenfalls

$$\int_0^{c_\alpha} g(\Lambda)\, d\Lambda = \alpha \tag{1.64}$$

Beispiel 13 x sei $N(\mu, \sigma^2)$-verteilt und es soll die Hypothese $H_0 : \mu = \mu_0$ gestestet werden. Die Likelihood ist dann durch

$$L(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2}\sum_j \left(\frac{x_j - \mu}{\sigma}\right)^2\right)$$

und die ML-Schätzungen für die Parameter sind $\hat{\mu} = \overline{x}$ und $\hat{\sigma}^2 = s^2 = \sum_j (x_j - \overline{x})^2/n$; vergl. (1.43) und (1.44). Dementsprechend ist

$$L(x|\hat{\mu}, \hat{\sigma}^2) = (2\pi s^2)^{-n/2} \exp\left(-\frac{1}{2}n\right)$$

denn $\sum_j (x_j - \overline{x})^2 = ns^2$. Angenommen, es gelte H_0, so dass $\mu = \mu_0$. Dann ist

$$\begin{aligned}
\hat{\hat{\sigma}}^2 &= \frac{1}{n}\sum_j (x_j - \mu_0)^2 \\
&= \frac{1}{n}\sum_j (x_j - \overline{x} + \overline{x} - \mu_0)^2 \\
&= \frac{1}{n}\left(\sum_j (x_j - \overline{x})^2 + \sum_j (\overline{x} - \mu_0)^2 + 2\sum_j (x_j - \overline{x})(\overline{x} - \mu_0)\right) \\
&= s^2 + (\overline{x} - \mu_0)^2
\end{aligned}$$

und es folgt

$$L(x|\mu_0, \hat{\hat{\sigma}}^2) = (2\pi s^2 + (\overline{x} - \mu_0)^2)^{-n/2} \exp(-n/2)$$

Setzt man diese Ausdrücke für L in (1.62) ein, so erhält man

$$\Lambda(x) = \left(\frac{s^2}{s^2 + (\overline{x} - \mu_0)^2} \right)^{n/2} \tag{1.65}$$

Dividiert man den Quotienten innerhalb der Klammer auf der rechten Seite durch s^2, so erhält man den Ausdruck

$$\Lambda(x) = \left(\frac{1}{1 + (\overline{x} - \mu_0)^2 / s^2} \right)^{n/2}$$

Die Größe $(\overline{x} - \mu_0)^2/s^2$ ist aber gerade t^2, so dass

$$\Lambda(x)^{2/n} = \frac{1}{1 + t^2/(n-1)} \tag{1.66}$$

folgt. Nun ist aber $\Lambda^{2/n}$ eine monotone Funktion von t^2, und diese Größe ist $F(1, n)$-verteilt. Man kann also direkt die F-Verteilung benützen, um H_0 zu testen. Man sieht wieder, dass ein Test anhand der F-Verteilung einem Test anhand eines Likelihood-Quotienten äquivalent ist. □

Eine derart direkte Beziehung zwischen Λ und einer Größe, deren Verteilung bekannt ist, ist nicht immer gegeben. Für diesen Fall sind *asymptotische* Verteilungen für Λ bekannt.

1.3.2 Asymptotische Verteilungen für Λ; Wilks' G^2

Es sei θ der Parameter der Verteilung für x und $L(x|\theta)$ die Likelihood von x. Es sei θ eine Schätzung für θ. Es kann gezeigt werden (vergl. Kendall und Stuart (1973), Vol. II, p. 240), dass asymptotisch

$$L \propto \exp\left(\frac{1}{2} E \left(\frac{\partial^2 \log L}{\partial \theta^2} (\theta - \theta)^2 \right) \right), \tag{1.67}$$

d.h. die Likelihood-Schätzung θ und damit auch L ist asymptotisch normalverteilt.

Dies gilt auch, wenn θ ein Vektor von Parametern ist. Es gilt

$$L(x|\theta_{ro}, \hat{\hat{\theta}}_s) \propto \exp\left(-\frac{1}{2} (\theta_r - \theta_{r0})' I_r^{-1} (\theta_r - \theta_{r0}) \right), \tag{1.68}$$

d.h. die Parameterschätzungen sind asymptotisch normalverteilt mit einer Varianz-Kovarianz-Matrix I_r, und die Inverse von I_r ist durch

$$I_r^{-1} = -E \left(\frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j} \right) \tag{1.69}$$

Hieraus kann die asymptotische Verteilung des in (1.62) definierten Likelihood-Quotienten

$$\Lambda(x) = \frac{L(x|\theta_{r0}, \hat{\hat{\theta}}_s)}{L(x|\hat{\theta}_r, \hat{\theta}_s)}$$

hergeleitet werden. Es läßt sich zeigen (Kendall und Stewart (1973), p. 241), dass Λ multivariat normalverteilt ist, und dass dementsprechend

$$-2\log\Lambda(x) = (t_r - \theta_{0r})' I_r^{-1} (t_r - \theta_{r0}) \tag{1.70}$$

nicht-zentral χ^2-verteilt ist mit r Freiheitsgraden und dem Nichtzentralitätsparameter

$$\lambda = (\theta_r - \theta_{0r})' I_r^{-1} (\theta_r - \theta_{0r}).$$

I_r ist die Informatonsmatrix, (vergl. (1.15); Seite 9). Unter der Nullhypothese H_0 ist $\lambda = 0$ und $-2\log\Lambda(x)$ ist χ^2-verteilt mit r Freiheitsgraden. Dieses Resultat geht auf Wilks (1938)[3] zurück; die Größe

$$G^2 = -2\log\Lambda(x) = -2(\log L(x|\theta_{0r}, \hat{\hat{\theta}}_s) - \log L(x|\hat{\theta}_r, \hat{\hat{\theta}}_s)) \tag{1.71}$$

ist die in der Theorie der Kategorialen Regression und der loglinearen Modelle übliche Teststatistik. Die Bedeutung der Parameter ist wie in (1.62); G^2 wird benützt, um verschiedene Modelle gegeneinander zu testen.

1.3.3 Goodness-of-fit Statistiken

Um die Güte der Anpassung eines Modells zu testen, sind zwei Statistiken gebräuchlich, die Pearsonsche χ^2-Statistik sovie die *Deviance* D:

$$\chi^2 = \sum_{i=1}^{g} \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)} \tag{1.72}$$

$$D = -2\phi \sum_{i=1}^{g} (l_i(\hat{\mu}_i) - l_i(y_i)). \tag{1.73}$$

Dabei sind die $\hat{\mu}_i$ die geschätzten Mittelwerte und $v(\hat{\mu}_i)$ ist die entsprechende Varianzschätzung, und $l_i(y_i)$ sind die individuellen log-likelihoods; y_i ist die maximal mögliche Likelihood. g ist die Anzahl der Gruppierungen. ϕ ist ein Überdispersionsparameter (*overdispersion parameter*), der durch

$$\hat{\phi} = \frac{1}{g-p} \sum_{i=1}^{g} \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)/n_i} \tag{1.74}$$

geschätzt wird.

"Overdispersion" kann bei binomialverteilten Daten auftreten; die beobachtete Varianz der Daten ist dann größer als durch das nominale Modell vorhergesagt. Overdispersion tritt auf wenn

1. *unbeobachtete Heterogenität* nicht durch entsprechende Kovariaten im linearen Prädiktor in Rechnung gestellt wird, und/oder wenn eine

[3] Wilks, S.S.: The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Statist., 9, 60

2. *positive Korrelation* zwischen binären Antworten existiert. Die Varianz der Summe ist dann nicht mehr gleich der Summe der Varianzen, man hat

$$Var(\sum_i y_i) > \sum_i Var(y_i),$$

wobei $y_i = 0, 1$.

Kapitel 2

Kategoriale Regression

2.1 Vorbereitung: Einige deskriptive Statistiken

Noch einfügen: Marx - Smith: Principal Component Estimation for generalized linear regression

Gegeben sei eine Kontingenztabelle der allgemeinen Form Hierin ist n_{ij}, $i, j =$

Tabelle 2.1: 2×2-Tabelle: Häufigkeiten

	Abhäng. Variable (B)		
Unabh. Variable (A)	B_1 $(y = 1)$	B_2 $(y = 0)$	Σ
A_1 $(x = 1)$	n_{11}	n_{12}	n_{1+}
A_2 $(x = 0)$	n_{21}	n_{22}	n_{2+}
Σ	n_{+1}	n_{+2}	n

$1, 2$ die Häufigkeit, mit der bei der i-ten Ausprägung von x die j-te Ausprägung von y beobachtet wurde. dass x in nur zwei möglichen Ausprägungen vorkommen kann, ist ein Spezialfall, der natürlich verallgemeinert werden kann. Statt der Häufigkeiten n_{ij} kann man natürlich auch die Wahrscheinlichkeiten p_{ij} betrachten; sie entsprechen den relativen Häufigkeiten n_{ij}/n.

Um die Häufigkeiten bzw. Wahrscheinlichkeiten zu diskutieren, kann man die Differenzen der bedingten Wahrscheinlichkeiten in den beiden Zeilen miteinander vergleichen:

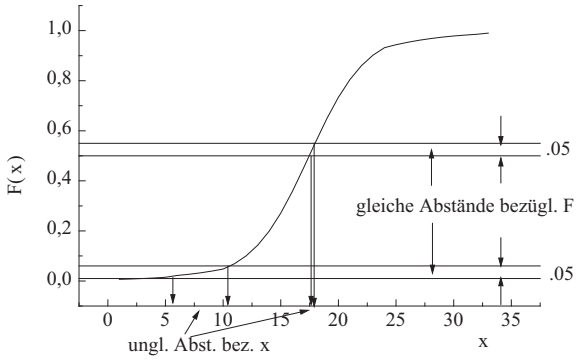
$$p(B_1|A_1) - p(B_1|A2) = (1 - p(B_2|A_1)) - (1 - p(B_2|A_1)) = p(B_2|A_2) - p(B_2|A_1) \tag{2.1}$$

Die Differenz dieser bedingten Wahrscheinlichkeiten liegt zwischen -1.00 und $+1.00$. Sind die Zeilen und Spaltenkategorien *stochastisch unabhängig*, so sind die Differenzen gleich Null, denn bei Unabhängigkeit gilt ja $p(B_2|A1) = p(B_2)$, $P(B_2|A_2) =$

$p(B_2)$, etc.

Ebenso kann man die Differenzen zwischen den Spalten bilden. Die Schwierigkeit bei der Betrachtung von Differenzen ist aber, dass die *Bedeutung* einer Differenz von den Werten der bedingten Wahrscheinlilchkeiten abhängt. Die Differenz zwischen .05 und .1 kann einen größeren Unterschied hinsichtlich der Größen, die den Wert der bedingten Wahrscheinlichkeiten bestimmen, bedeuten, als die Differenz zwischen .5 und .55. In Abbildung 2.1 wird dieser Sachverhalt verdeutlicht. Man be-

Abbildung 2.1: Wertedifferenzen für gleichgroße Wahrscheinlichkeitsdifferenzen



trachtet deshalb im allgemeinen Quotienten von bedingten Wahrscheinlichkeiten; in welcher Beziehung die Werte solcher Quotienten zur unabhängigen Variablen stehen, wird später verdeutlicht; hier sollen zunächst nur die wichtigsten Größen eingeführt werden.

Definition 9 *Der Quotient*

$$R = \frac{p(B_j|A_1)}{p(B_j|A_2)} \geq 0 \tag{2.2}$$

heißt relatives Risiko.

Im Falle der stochastischen Unabhängigkeit von unabhängiger und abhängiger Variable gilt $p(B_j|A_1) = p(B_j|A_2) = p(B_j)$; in diesem Fall ist das relative Risiko $R = 1$.

Definition 10 *Die Quotienten*

$$\Omega_1 = \frac{p(B_1|A_1)}{p(B_2|A_1)}, \quad \Omega_2 = \frac{p(B_1|A_2)}{p(B_2|A_2)} \tag{2.3}$$

heißen Odds *oder* Wettchancen.

Man kann z.B. wetten, dass B_1 bzw. B_2 eintritt, wenn entweder A_1 oder A_2 zutrifft. Die Chancen, zu gewinnen, sind dann durch Ω_1 bzw. Ω_2 gegeben.

Die Odds lassen sich leicht aus den Häufigkeiten n_{ij} ausrechnen. Denn es ist ja $P(A_i|B_j) = n_{ij}/n_{i+}$, und deshalb hat man sofort

$$\Omega_1 = \frac{n_{11}/n_{1+}}{n_{12}n_{1+}} = \frac{n_{11}}{n_{12}}, \quad \Omega_2 = \frac{n_{21}/n_{2+}}{n_{22}/n_{2+}} = \frac{n_{21}}{n_{22}} \tag{2.4}$$

Definition 11 *Das Verhältnis*

$$\Theta = \frac{\Omega_1}{\Omega_2} = \frac{p(B_1|A_1)p(B_2|A_2)}{p(B_2|A_1)p(B_1|A_2)} \geq 1 \tag{2.5}$$

heißt Kreuzproduktverhältnis*, oder auch* odds ratio.

Wegen (2.4) hat man sofort

$$\Theta = \frac{n_{11}n_{22}}{n_{12}n_{21}} \tag{2.6}$$

Θ ist ein Maß für die Stärke der Assoziation zwischen der unabhängigen und der abhängigen Variablen. Sind diese beiden Variablen stochastisch unabhängig, so ergibt sich $\Theta = 1$, wie man leicht überprüft. Die Assoziation zwischen den Variablen ist um so größer, je mehr Θ von 1 abweicht. Für $\Theta \to 0$ ist die Beziehung zwischen den Variablen *negativ*, d.h. $x = 1$ impliziert $y = 0$ und $x = 0$ bedeutet $y = 1$, und für $\Theta \to \infty$ ist sie *positiv*; für größer werdendes Θ gehen $\pi_{1|2} = p(B_1|A_2)$ und $\pi_{2|1} = p(B_2|A_1)$ gegen Null.

Beispiel 14 Aspirin lindert nicht nur den Kopfschmerz, sondern verringert auch das Risiko, einen Herzinfarkt zu erleiden. In einer längeren Studie hat man 11034 Ärzten ein Placebo und 11037 anderen Ärzten täglich eine Aspirintablette gegeben[1]. Dabei handelte es sich um einen Blindversuch: keiner der beteiligten Ärzte wußte, ob er ein Placebo oder eine Aspirintablette schluckte. Die Daten sind hier zu einer 2×2-Tabelle zusammengefaßt worden. Für das relative Risiko eines Herzinfarkts

Tabelle 2.2: Aspirin und die Wahrscheinlichkeit von Herzinfarkten

	Herzinfarkt		
Medikament	B_1 $(y = 1)$	B_2 $(y = 0)$	Σ
Aspirin $(A_1; x = 1)$	104	10933	11037
Placebo $(A_2; x = 0)$	189	10845	11034
Σ	293	21778	22071

erhält man

$$R_{HI} = \frac{p(B_1|A_1)}{p(B_1|A_2)} = \frac{104/11037}{189/11034} = \frac{104}{189} \cdot \frac{11034}{11037} = .5501$$

[1] Preliminary Report: Findings from the Aspirin Component of the Ongoing Physicians' Health Study (1988), N. Engl. J. Med. 318, 262-264

Der Anteil der Personen, die einen Herzinfarkt bekamen *und* Aspirin genommen haben ist nur etwas mehr als halb so groß wie der Anteil derjenigen Personen, die einen Herzinfarkt bekamen *und* kein Aspirin genommen haben. Man kann auch das relative "Risiko", *keinen* Herzinfarkt zu bekommen, berechnen:

$$R_{\neg HI} = \frac{p(B_2|A_1)}{p(B_2|A_2)} = \frac{10933}{10845} \cdot \frac{11034}{11037} = 1.00784$$

Man sieht, dass die beiden Arten von Risiken nicht komplementär sind. Der Wert von $R_{\neg HI}$ spiegelt die Tatsache wider, dass insgesamt die Wahrscheinlichkeit, einen Herzinfarkt zu bekommen, relativ klein ist ($p(HI) = 293/22071 = .01328$); der Effekt des Aspirins geht, betrachtet man die (Teil-)Population der Personen ohne Herzinfarkt, gewissermaßen verloren.

Die Odds, einen Herzinfarkt zu bekommen, wenn man Aspirin nimmt, sind

$$\Omega_1 = \frac{p(B_1|A_1)}{p(B_2|A_1)} = \frac{104/11037}{10933/11037} = \frac{104}{10933} = .00951$$

Die entsprechenden Odds für die Bedingung, ein Placebo verabreicht bekommen zu haben, sind

$$\Omega_2 = \frac{p(B_1|A_2)}{p(B_2|A_2)} = \frac{189/11034}{10845/11034} = \frac{189}{10845} = .01743$$

Offenbar sind die Odds für einen Herzinfarkt unter der Bedingung, Aspirin bekommen zu haben, geringer als die unter der Bedingung, nur ein Placebo verabreicht bekommen zu haben. Man enthält für das Kreuzproduktverhältnis

$$\Theta = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{104 \cdot 10845}{189 \cdot 10933} = \frac{.00951}{.01743} = .54584$$

Die Odds für einen Herzinfarkt unter der Bedingung, Aspirin eingenommen zu haben, sind fast auf die Hälfte reduziert relativ zu der Bedingung, nur ein Placebo verabreicht bekommen zu haben. □

2.2 Das Logit-Modell

2.2.1 Latente Variablen und zufällige Ereignisse

Im vorangegangenen Abschnitt ist ein Spezialfall betrachtet worden: es ist entweder $x = 0$ oder $x = 1$. Man kann aber auch den Fall betrachten, dass x mehr als nur einen Wert annehmen kann; sicherlich kann man die Aspirinmenge, die eine Person zu sich nimmt, dosieren und dann die Wahrscheinlichkeit eines Herzinfarkts in Abhängigkeit von der Aspirindosis betrachten. Die Frage ist jedenfalls, wie die Abhängigkeit der Wahrscheinlichkeit, etwa einen Herzinfarkt zu erleiden, von der Aspirindosis modelliert werden kann. Allgemein kann man nach einem Ausdruck für die (bedingte) Wahrscheinlichkeit $P(y = 1|x)$ fragen.

Eine interessante Möglichkeit der Modellierung besteht darin, eine *latente Variable* anzunehmen, in Abhängigkeit von deren Wert das Ereignis, das durch $y = 1$ charakterisiert wird, eintritt oder nicht. So repräsentiere ξ eine solche Variable, und es werde angenommen, dass $y = 1$ genau dann, wenn $\xi > S$, wobei S ein Schwellenwert ist. So könnte in bezug auf das Beispiel 14 die Variable ξ das Ausmaß der Verkalkung der Herzkranzgefäße abbilden. Ist die Verkalkung größer als ein bestimmter Wert S, so kommt es zu einem Infarkt. Dies ist sicherlich ein stark vereinfachtes Bild des Geschehens, aber es soll auch nur erläutert werden, wie man im Prinzip vorgehen könnte. Die Dosis x des Medikaments (hier: Aspirin) bestimmt die Verteilung von ξ; dass dies so ist, ist zunächst nur eine Hypothese, aber man kann diese Hypothese in dieser Form formulieren. Bedeutet $y = 1$ wieder, dass eine Person einen Infarkt erleidet, (oder dass allgemein das in Frage stehende Ereignis eintritt), und gibt x die Größe der Dosis an, so kann man dementsprechend

$$p(x) := P(y = 1|x) = P(\xi > S|x) \tag{2.7}$$

schreiben. Hier ist $p(x)$ nur eine abgekürzte Schreibweise für $P(\xi > S|x)$ bzw. für $P(y = 1|x)$. Die Schreibweise $P(\xi > S|x)$ deutet an, dass der Wert x in irgendeiner Weise auf die Parameter der Verteilung von ξ einwirken muß, so dass sich mindestens ein Parameter der ξ-verteilung als Funktion von x schreiben lassen muß.

Nun deutet (2.7) an, dass die "eigentliche" Variable, die das durch $y = 1$ kodierte Ereignis auslöst, eine zufällige Veränderliche ist; für einen gegebenen x-Wert können ganz verschiedene ξ-Werte auftreten, wenn auch mit einer von x abhängenden Wahrscheinlichkeit. Die Dosis x wird im allgemeinen nicht der einzige Faktor sein, der den Wert von ξ bestimmt; so hat die Verkalkung von Herzkranzgefäßen sicher einen multifaktoriellen Hintergrund, d.h. viele Faktoren wirken gleichzeitig. Analoge Betrachtungen gelten für psychische Zustände: ob es zu einem Panikanfall kommt oder nicht, hat sicher eine Reihe von Gründen, und die "kontrollierte", d.h. die betrachtete unabhängige Variable ist sicher nur eine von vielen, die auf das Panikgeschehen Einfluß hat.

2.2.2 Die logistische Verteilung

Wirken viele Einflußgrößen gleichzeitig auf eine Variable ein, so ist die Vermutung, dass die Variable Gauß-verteilt ist, vernünftig. Der Wert dieser Verteilung in Abhängigkeit von der oberen Grenze des Integrals ist aber bekanntlich schwierig zu berechnen. Einfacher im numerischen Umgang ist die *logistische Verteilung*:

$$P(\xi \leq S) = \frac{1}{1 + \exp\left(-\frac{(S-\mu)}{\sigma}\frac{\pi}{\sqrt{3}}\right)}, \tag{2.8}$$

oder

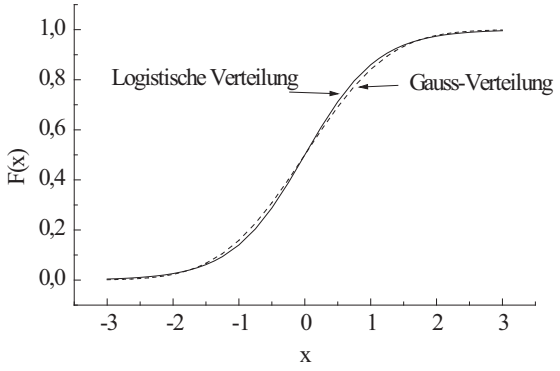
$$P(\xi \leq S) = \frac{1}{1 + \exp\left(-1.8138(S-\mu)/\sigma\right)}, \quad 1.8138 \approx \pi/\sqrt{3} \tag{2.9}$$

μ und σ sind die Parameter der Verteilung; es gilt

$$E(\xi) = \mu, \quad Var(\xi) = \sigma^2, \tag{2.10}$$

wobei hier $\pi = 3.14159\cdots$. Für $\mu = 0$ und $\sigma^2 = 1$ erhält man die standardisierte logistische Verteilung. Die Abbildung 2.2 zeigt die Verläufe von Gauß- und logisti-

Abbildung 2.2: Verläufe der standardisierten Gauß- und der logistischen Verteilung



scher Verteilung. Man sieht, dass die Unterschiede zwischen den Verteilungen für die meisten praktischen Zwecke vernachlässigen lassen.

Anmerkung zur Logistischen Verteilung: Man kann sich fragen, wie man zur Definition der Verteilung (2.8) kommt. Ausdrücke dieser Art ergeben sich, wenn man Wachstums- oder Ansteckungsprozesse betrachtet, in denen Sättigungen eine Rolle spielen. Der belgische Mathematiker Pierre Verhulst (1804 - 1849) entwickelte ein Modell für das Wachstum von Populationen, dass in der Differentialgleichung (2.11) zusamengefasst werden kann; (2.8) stellt eine Lösung dieser Differentialgleichung dar. 1838 wurde es von Verhulst angewendet, um den Bedarf an Wohnungen (*logis*) in Paris vorherzusagen, und seit dem spricht von logistischen Model. Das logistische Modell gilt auch als das Standardmodell der Epidemiologie, mit dem die Ausbreitung von Infektionen modelliert werden kann[2]. Hier wird nur eine kurze Skizze der Argumentation, die zur logistischen Verteilung führt, vorgestellt. Man betrachte die Ausbreitung von grippalen Infekten. Es sei $F(t)$ die Anzahl bereits Infizierter zum Zeitpunkt t, und $f(t) = dF(t)/dt$ sei die Veränderung von F zum Zeitpunkt t, - jeder Infizierte steckt ja Nichtinfizierte an. Macht man die vereinfachende Annahme, dass die Ansteckungsrate für alle Infizierte ungefähr gleich ist, so gilt zunächst $f(t) = dF(t)/dt = kF(t)$, dh die *Veränderung* des Anteils bereits Infizierter ist proportional zum Anteil bereits Infizierter. Diese Gleichung ist eine *Differentialgleichung*, weil das Differential f von F zu F in Beziehung gesetzt wird. Die Lösung einer Differentialgleichung besteht in der Angabe der Funktion F, die dieser Gleichung genügt, und man findet, dass $F(t) = k\exp(\lambda t)$ ist: die Anzahl Infizierter wächst demnach exponentiell. Aber je mehr Infizierte es gibt, desto weniger Nichtinfizierte gibt es, dh ein Infizierter kann immer weniger noch nicht Infizierte anstecken. Der Anteil der Nichtinfizierten ist $1 - F(t)$, und die Veränderungsrate soll nicht nur proportional zu

[2] vergl. das Skriptum *Logistische Funktionen und Verteilungen, Epidemien und die Wechselwirkung zwischen Hormonen* (epidemiologie.pdf))

$F(t)$ sein, sondern auch zu $1 - F(t)$. So kommt man zu der Gleichung

$$f(t) = \frac{dF(t)}{dt} = kF(t)(1 - F(t)). \tag{2.11}$$

Die Funktion F, die dieser Gleichung genügt, ist (2.8), dh

$$F(S) = P(\xi \leq S|x),$$

wobei t durch S ersetzt wurde. Bei der Bestimmung der Lösung der Differentialgleichung ergibt sich auch der Faktor $\pi/\sqrt{3}$. □

Um die logistische Verteilung für Zwecke der Regression nutzen zu können, muß gezeigt werden, in welcher Weise unabhängige Variablen, die möglicherweise auf die Wahrscheinlichkeit eines bestimmten Ereignisses einwirken, eingeführt werden können. Im *Allgemeinen Linearen Modell* (ALM) wird angenommen, dass Bedingungen, die auf eine beobachtete zufällige Variable wirken, insbesondere auf deren Erwartungswert μ wirken. Man nimmt also $y = \mu + \varepsilon$ an, wobei ϵ eine zufällige Veränderliche ist, die einen "Fehler" repräsentiert, $E(y) = \mu$, d.h. $E(\varepsilon) = 0$, postuliert wird, und μ hängt in irgendeiner Weise von der unabhängigen Variablen ab. die einfachste Annahme ist, dass diese Abhängigkeit linear ist, so dass $\mu = \alpha x + \beta$ postuliert wird. Diese Annahme wird auch hier gemacht:

Annahme: *Die Verteilung von ξ sei durch (2.8) gegeben. Die durch die Variable x repräsentierte unabhängige Variable wirkt linear auf $\mu = E(\xi)$ ein, d.h. es soll gelten*

$$\mu = \alpha x + \beta \tag{2.12}$$

Bemerkungen:

1. Die Annahme entspricht der in der Regressionsrechnung bzw. beim Vergleich von Mittelwerten (Varianzanalyse) gemachten Annahme, dass die unabhängigen Variablen auf den Erwartungswert μ der betrachteten zufälligen Veränderlichen wirken; μ ist gewissermaßen der "wahre" Wert, von dem die zufällige Veränderliche ξ nur zufällig abweicht.

2. Die Beziehung $\mu = \alpha x + \beta$ definiert eine Skalentransformation: wird die unabhängige Variable x auf einer bestimmten Skala gemessen, die durch die inhaltliche Bedeutung dieser Variablen gegeben ist, zB Aspirin in einer bestimmten Dosis. Der Wert von α wird dann einerseits von der gewählten Einheit, in der x gemessen wird, abhängen, andererseits von der Skala, auf der ξ definiert ist. ξ könnte die "Dünnflüssigkeit" des Blutes sein, oder das Ausmaß an Verkalkung der Herzkranzgefäße. In anderen Fragestellungen könnte ξ aber auch das Ausmaß an subjektiv empfundenem Stress sein, - die Maßeinheit ist dann nicht ganz klar.

Nach (2.9) hängt die Wahrscheinlichkeit $P(\xi \leq S)$ von der Größe $-(S - \mu)\pi/\sigma\sqrt{3}$ ab. Setzt man für μ nach (2.12) den Ausdruck $\alpha\mu + \beta$ ein, so erhält man

$$-\frac{(S - \mu)\pi}{\sigma\sqrt{3}} = -\frac{(S - \alpha x - \beta)\pi}{\sigma\sqrt{3}}.$$

S, α, β und σ sind unbekannte, d.h. *freie* Parameter, die aus den Daten geschätzt werden müssen, will man den Einfluß von x auf $P(\xi \leq S)$ bestimmen. Tatsächlich sind zunächst nur zwei Parameter zu schätzen, die sich wie folgt ergeben. Denn es ist

$$-\frac{(S-\alpha x-\beta)\pi}{\sigma\sqrt{3}} = -\frac{(S-\beta)\pi}{\sigma\sqrt{3}} + \frac{\alpha\pi x}{\sigma\sqrt{3}}.$$

Setzt man nun

$$A = \frac{\alpha\pi}{\sigma\sqrt{3}}, \quad B = -\frac{(S-\beta)\pi}{\sigma\sqrt{3}} \tag{2.13}$$

Also hat man

$$-\frac{(S-\alpha x-\beta)\pi}{\sigma\sqrt{3}} = Ax + B \tag{2.14}$$

und man erhält

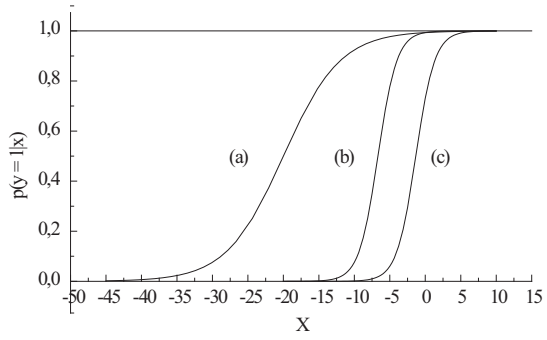
$$P(\xi \leq S|x) = \frac{1}{1+\exp(Ax+B)}. \tag{2.15}$$

Diese Gleichung impliziert[3]

$$p(x) = P(\xi > S|x) = \frac{1}{1+\exp(-(Ax+B))}. \tag{2.16}$$

Die Wahrscheinlichkeit $P(\xi \leq S|x)$ hängt also von x über die beiden freien Parame-

Abbildung 2.3: Logistische Funktionen für verschiedene A- und B-Werte; die Funktion (a) ist durch einen kleineren A-Wert (dh größeren σ-Wert relativ zu α) als die Funktionen (b) und (c) definiert; (c) unterscheidet sich von (b) nur hinsichtlich des Wertes von B.



ter A und B ab, deren Werte aus den Daten geschätzt werden müssen. Im Prinzip wird man zu diesem Zweck für x-Werte $x_1, \ldots, x_n$ die zugehörigen $P(\xi \leq S|x_i)$ abschätzen und dann eine geeignete Schätzmethode zum Auffinden von Schätzungen

[3] $P(\xi > S|x) = 1 - P(\xi \leq S|x)$, und

$$1 - \frac{1}{1+\exp(Ax+B)} = \frac{\exp(Ax+B)}{1+\exp(Ax+B)}.$$

Multiplikation von Zähler und Nenner mit $\exp(-(Ax+B))$ liefert (2.16).

$\hat{A}$ und $\hat{B}$ anwenden, um - analog zur gewöhnlichen Regression - die $P(\xi \leq S|x)$ "vorauszusagen". Zu diesem Zweck wird die folgende Beziehung hergeleitet:

Satz 5 *Es werde $p(x_i)$ für die x-Werte $x_1, \ldots, x_n$ bestimmt. Dann gilt*

$$\log \frac{p(x_i)}{1-p(x_i)} = Ax_i + B, \;\; i = 1, \ldots, n. \tag{2.17}$$

Beweis: Nach (2.7) ist

$$p(x) = P(y = 1|x) = P(\xi > S|x);$$

$p(x)$ ist nur eine abgekürzte Schreibweise für $P(y = 1|x)$ bzw. $P(\xi > S|x)$. Jedenfalls ist dann

$$1 - p(x) = P(\xi \leq S|x) = \frac{1}{1 + e^{Ax+B}}, \tag{2.18}$$

und

$$\begin{aligned} p(x) = P(\xi > S|x) &= 1 - \frac{1}{1 + e^{Ax+B}} = \frac{1 + e^{Ax+B} - 1}{1 + e^{Ax+B}} \\ &= \frac{e^{Ax+B}}{1 + e^{Ax+B}}, \end{aligned} \tag{2.19}$$

so dass

$$\frac{p(x)}{1-p(x)} = e^{Ax+B} \tag{2.20}$$

oder

$$g(p(x)) = \log \frac{p(x)}{1-p(x)} = Ax + B \tag{2.21}$$

folgt. □

Definition 12 *Die Größe* $\log p(x)/(1-p(x))$ *heißt* Logit*; die in (2.21) eingeführte Funktion g heißt* Link-Funktion*; sie verbindet $Ax + B$ mit der Wahrscheinlichkeit $p(x)$.*

Anmerkungen:

1. Ist $P(\xi \leq S|x_i)$ also durch die logistische Verteilung definiert, so ist die Link-Funktion g, dh $g(p(x_i)) = Ax_i + B$, durch die Logits $\log p(x_i)/(1 - p(x_i))$ gegeben. Für andere Verteilungen ist $Ax + b$ durch entsprechende andere Link-Funktionen definiert.

2. Nach (2.17) bzw. (2.21) sind die $Ax_i + B$ und die korrespondierenden Logits linear aufeinander bezogen; das heißt aber, dass die Beziehung zwischen den x_i und den $p(x_i)$ *nichtlinear* ist. Berechnet man also eine "normale" Regression zwischen x und $p(x)$ oder $1 - p(x)$, so kann es sich dabei bestenfalls um eine Approximation handeln.

Zur Interpretation der Parameter:

1. Die Größe $p(x)/(1-p(x))$ entspricht der in Definition 10, Seite 30, Gleichung (2.3) eingeführten Wettchance (Odds); der Quotient gibt ja für gegebenen Wert x das Verhältnis der Wahrscheinlichkeit, dass das in Frage stehende Ereignis eintritt, zur Wahrscheinlichkeit, dass es nicht eintritt, an. Nach (2.20) gilt

$$\frac{p(x)}{1-p(x)} = e^{Ax+B} = e^{Ax}e^{B}. \tag{2.22}$$

   Die Odds werden also als ein Produkt von Faktoren, $\exp(Ax)$ und $\exp(B)$ dargestellt. Die Interpretation orientiert sich demgemäß an den Werten dieser Faktoren, die wiederum von A und B abhängen: verändert man x um eine Einheit, so wirkt sich diese Veränderung über die Größe e^{Ax} auf den Quotienten $p(x)/(1-p(x))$ aus,

   Zur Illustration sei das in Frage stehende Ereignis der Herzinfarkt, der innerhalb eines bestimmten Zeitraums (1 Jahr, 5 Jahre, etc) eintritt oder auch nicht eintritt. Für $x = 0$ insbesondere erhält man

$$\frac{p(0)}{1-p(0)} = e^{B}.$$

   Die Bedeutung von B hängt jetzt von der betrachteten Variable x ab:

   (a) Es repräsentiere x das Ausmaß an Verkalkung der Herzkranzgefäße. Dann gibt $\exp(B)$ die Chance - also nicht die Wahrscheinlichkeit! - , einen Herzinfarkt zu erleiden an, wenn keine Verkalkung vorliegt. Angenommen, es sei $p(0) = .1$, so ist $\exp(B) = .1/.9 \approx .111$; was einem Wert $B = -2.197$ entspricht. Man interpretiert also nicht direkt den Wert von B, sondern den von $\exp(B)$.

   (b) Angenommen, x repräsentiere eine Dosis von Aspirin, etwa $x = 20$ Milligramm. Dann ist $p(0)$ die Wahrscheinlichkeit, einen Herzinfarkt zu erleiden, wenn man diese Dosis *nicht* täglich einnimmt, unabhängig vom Ausmaß der Verkalkung der Herzkranzgefäße. Es sei, für den gleichen Zeitraum wie eben, $p(0) = .3$; dann ist $p(0)/(1-p(0)) = .3/.7 \approx .43$, was dem Wert $B = -.847$ entspricht. Offenbar hängt der Wert von B von der betrachteten Variablen ab.

2. Es sei nun $x \neq 0$. Findet man $\hat{A} \approx 0$, so kann man $A = 0$ annehmen und folgern, dass die Variable x keinen Einfluß auf das Ereignis hat. Repräsentiert also x das genannte Ausmaß an Verkalkung, so hätte die Verkalkung der Herzkranzgefäße keinen Einfluß auf die Chance, einen Herzinfarkt zu erleiden. Repräsentiert x eine Dosis Aspirin, so bedeutet $A = 0$, dass die Einnahme von Aspirin keinen Einfluß auf das Herzinfarktrisiko hat.

   Es sei $A \neq 0$. Der Wert von A kann positiv oder negativ sein. Es sei $A > 0$. Wegen

$$p(x) = \frac{1}{1+e^{-(Ax+B)}}$$

(vergl. (2.16)) wird $p(x)$ mit steigendem x größer, denn $e^{-(Ax+B)}$ wird mit steigendem x kleiner. Repräsentiert x das Ausmaß an Verkalkung der Herzkranzgefäße, so wird man einen positiven Wert von A erwarten, denn die Verkalkung hebt den Wert von $p(x)$.

Den Fall $A < 0$ wird man erwarten, wenn ein steigender Wert von x den von $p(x)$ erniedrigt. So sei $A = -1$; dann ist $-(Ax + B) = -Ax - B = x - B$, und $\exp(x - B)$ wird größer mit steigendem x, so dass $p(x)$ kleiner wird. Repräsentiert x eine Dosis Aspirin, so ist dieser Fall zu erwarten, denn Aspirin senkt die Wahrscheinlichkeit, einen Herzinfarkt zu erleiden.

Es sei noch auf die folgenden Sachverhalte hingewiesen:

(a) Der Wert von A hängt u. U. von der Wahl der Maßeinheit für x ab. Repräsentiert x eine Dosis Aspirin, so kann man diese in Gramm, Milligramm, Kilogramm etc: die Wahl der Maßeinheit wird nach Maßgabe der Genauigkeit bestimmt, mit der man zwischen verschiedenen Quantitäten differenzieren will. Es sei A der Wert, wenn die Dosis in Gramm gemessen wird; für 1 Gramm hat man insbesondere $A+B$. Misst man in Milligramm, so hat man die Größe $A^*1000 + B^*$, die den Wert von $p(x)$ bestimmt. Für beide Maßeinheiten hat natürlich p den gleichen Wert. Eine analoge Aussage gilt für 1.5 Gramm oder 1500 Milligram. Also muß

$$\begin{aligned} A + B &= A^*1000 + B^* \\ A1.5 + B &= A^*1500 + B^* \end{aligned}$$

gelten. Man findet $B^* = B$, denn der Fall $x = 0$ hat unabhängig von der Maßeinheit für x stets die gleiche Bedeutung, und $A^* = (.5/500)A = .001A$. Wegen der Abhängigkeit von der gewählten Maßeinheit ist der numerische Wert von A zunächst schwer zu deuten.

(b) x kann auch eine Indikatorvariable sein. Dieser Fall liegt vor, wenn kategoriale Prädiktoren betrachtet werden. Zum Beispiel hat man $x = 1$, wenn Kategorie 1 gegeben ist (Proband ist weiblich), $x = 0$, wenn Kategorie 1 nicht vorliegt (Proband ist männlich). Dann ist $\mu_w = \alpha x + \beta = \alpha + \beta$, wenn der Proband weiblich ist, und $\mu_m = \beta$, wenn der Proband männlich ist. Dann ist die Differenz zwischen den Erwartungswerten gleich $\mu_w - \mu_m = \alpha$, und $A = \alpha\pi/\sigma\sqrt{3} = (\mu_w - \mu_m)\pi/\sigma\sqrt{3}$. A entspricht nun einer Effektgröße (den Faktor $\pi/\sqrt{3}$ kann man herausdividieren).

Die Frage ist nun, ob die Parameter S, α, β und σ ebenfalls geschätzt werden können. Können sie geschätzt werden, heißen sie *identifizierbar*, andernfalls heißen sich *nicht identifizierbar*. Man sieht nun anhand von (2.13), dass man nur zwei Gleichungen mit den bekannten - weil geschätzten - Größen A und B hat, aber vier Unbekannte finden muß, eben α, β, S und σ. Aus zwei Gleichungen lassen sich nicht vier Unbekannte bestimmen (man braucht vier Gleichungen dazu), also sind die Parameter S, α, β und σ *nicht identifizierbar*. Man kann allenfalls zu relativen Aussagen über α und σ kommen, etwa für den Fall, dass x eine Indikatorvariable ist.

Beispiel 15 Es werde die Interpretation der Parameter weiter diskutiert. Dazu betrachte man zwei Fälle: (i) x repräsentiere den täglichen Nikotinkonsum, der bekanntlich zu einer Erhöhung des Infarktrisikos führt, und (ii) x repräsentiere eine tägliche Aspirindosis, die zu einer Reduktion des Infarktrisikos führt.

Man betrachte die Infarkthäufigkeiten bei zwei Gruppen, etwa männlichen und weiblichen Probanden, deren jeweiliger Nikotinkonsum $x_1, \ldots, x_n$ Einheiten betrage; die Wahrscheinlichkeit eines Infarkts sei $p(x) = 1/(1 + \exp(-(Ax + B))$. Wenn x den Nikotinkonsum repräsentiert, wird $p(x)$ mit x größer werden; dies geschieht, wenn $\exp(-(Ax + B))$ kleiner wird, dh wenn $A > 0$. Die Personen in den beiden Gruppen seien vergleichbar hinsichtlich Alter, Lebensgewohnheiten etc. Es werde angenommen, dass $A_m < A_w$. Dann wird $\exp(-(Ax + B))$ für Männer langsamer kleiner als für Frauen, dh $p(x)$ wächst für Männer langsamer als für Frauen. Dies heißt nicht, dass das Risiko für Männer bei gleichem x geringer ist als das für Frauen, da der Wert von B mit berücksichtigt werden muß. Männer sind nur dann weniger gefährdet, wenn $A_m x + B_m < A_w x + B_w$. Die Parameter A_m und A_w bestimmen nur die Rate der Veränderung von $p(x)$.

Allgemein ist $A = \alpha\pi/\sigma\sqrt{3}$. Man hat demnach für $A_m < A_w$

$$\frac{\alpha_m \pi}{\sigma_m \sqrt{3}} < \frac{\alpha_w \pi}{\sigma_w \sqrt{3}},$$

also

$$\frac{\alpha_m}{\sigma_m} < \frac{\alpha_w}{\sigma_w}. \tag{2.23}$$

Nimmt man an, dass $\alpha_m = \alpha_w$, so bedeutet dies, dass die Kopplung an die infarktauslösende Variable ξ für beide Geschlechter die gleiche ist. Dann muß man aber wegen (2.23) $\sigma_m > \sigma_w$ folgern, dh dass ξ bei den Männern mehr streut als bei den Frauen. Repräsentiert x das Ausmaß des Nikotinkonsums, so muß die geringere (wegen $A_m < A_w$) Rate der Veränderung von $p(x)$ ein Resultat der größeren Streuung von ξ sein: der Effekt von x geht um so mehr in den Fluktuationen von ξ verloren, je größer die Varianz bzw. Streuung.

Man kann auch den Fall $\sigma_m = \sigma_w$ betrachten (dies kann übrigens eine überprüfbare Hypothese sein). Dann folgt $0 < \alpha_m < \alpha_w$, dh der Effekt des Nikotins auf die Verkalkung ist bei Männern *geringer* als bei den Frauen; Männer wären - statistisch gesehen - resistenter gegen eine Zunahme der Verkalkung als die Frauen. Warum das so sein soll, muß dann die weitere Forschung zeigen.

Es repräsentiere nun x eine Aspirindosis. Man wird $A_m < 0$ und $A_w < 0$ finden, da Aspirin das Risiko eines Infarkts reduziert: setzt man $A_w = -a_w$, $A_m = -a_m$ mit $a_w > 0$, $a_m > 0$, so hat man $\exp(-(A_m x + B)) = \exp(a_m x - B)$, und diese Größe wächst mit x, so dass $p(x) = 1/(1 + \exp(a_m x - B))$ kleiner wird mit x. Die Betrachtung für A_w ist analog. Wieder gelte $A_m < A_w$, also (2.23). Für Männer ist die Reduktion des Infarktrisikos geringer mit steigender Dosis als für Frauen. Gilt nun $\alpha_m = \alpha_w$, so ist die Kopplung des Aspirins an die infarkterzeugende Variable die gleiche für beide Geschlechter und es folgt wieder $\sigma_m > \sigma_w$. Die geringere Reduktion des Infarktrisikos bei den Männern ist ein Resultat der größeren Streuung

σ_m, der Effekt des Aspirins geht mehr in den Fluktuationen von ξ unter als bei den Frauen.

Gilt andererseits $\sigma_m = \sigma_w$, so folgt $\alpha_m < \alpha_w$. Die Reduktion des Infarktrisikos ist bei den Männern geringer als bei den Frauen, weil die Kopplung zwischen dem Aspirin und der infarktauslösenden Variable für die Männer schwächer ist.

Natürlich sind Annahmen der Art $\sigma_m = \sigma_w$ oder $\alpha_m = \alpha_w$ Spezialfälle. Aber für $A_m < A_w$ folgt aus (2.23)

$$\frac{\alpha_m}{\alpha_w} < \frac{\sigma_m}{\sigma_w}. \tag{2.24}$$

Hat man etwa Abschätzungen von σ_m und σ_w derart, dass $\sigma_m \neq \sigma_w$ angenommen werden muß, so liefert (2.24) eine Abschätzung von α_m/α_w. Gilt etwa $\sigma_m/\sigma_w \approx .75$, so folgt $\alpha_m/\alpha_w < .75$, dh $\alpha_m < .75\alpha_w$. Findet man andererseits $\sigma_m/\sigma_w \approx 1.75$, so folgt $\alpha_m < 1.75\alpha_w$, was auch den Fall $\alpha_m = \alpha_w$ nicht ausschließt. Man erhält zumindest Informationen über die Größenordnung des Unterschieds zwischen α_m und α_w.

Zusammenfassend kann man sagen, dass ein kleinerer Wert von $|A|$ bei gleicher Kopplung bedeutet, dass der Effekt von x durch die relativ zum Effekt von x großen Fluktuationen von ξ reduziert wird. Wegen (2.22) hängt die Veränderung der Logits mit x natürlich auch von B ab. □

2.2.3 Verallgemeinerung: nichtlineare Kopplungen

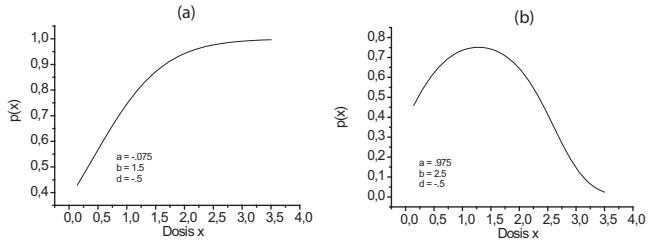
Der Standaranahme entsprechend geht die unabhängige Variable x über den linearen Ausdruck $Ax + B$ in den Ausdruck für $p(x)$ ein. Es ist aber denkbar, dass zB für kleinere x-Werte $p(x)$ mit größer werdendem x wächst (oder fällt), sich für größere Werte von x dieser Trend aber umkehrt. So kann es bei einer kleinen Dosis x eines Medikaments gegen Kopfschmerzen zu einer Linderung der Kopfschmerzen kommen, eine größere Dosis führt aber zu einer Intensivierung der Schmerzen. Dies kann geschehen, wenn die Gabe des Medikaments die Bildung von Stoffen auslöst, die den Effekt des Medikaments hemmen. Bei kleinen Dosierungen ist der hemmende Effekt vernachlässigbar, bei größeren Dosierungen kehren diese hemmenden Effekte die Wirkung um. Man kann solche Effekte nach Maßgabe des logistischen Modells[4] erklären, indem man zB

$$\mu(x) = x(B - Ax) + C \tag{2.25}$$

ansetzt. A, B und C sind wieder freie Parameter. Es ist $x(B-Ax)+C = Bx-Ax^2+C$; für Werte $x < 1$ ist $x^2 < x$ und man hat $\mu(x) \approx Bx + C$ und für $B > 0$ steigt $p(x)$ mit x. Der Term $-Ax^2$ drückt eine inhibierende Wirkung von x für größere x-Werte aus, und für geeignet gewählte Werte von A, B, und C kann $p(x)$ von einem bestimmten x-Wert an wieder kleiner werden. Abb. 2.4 zeigt zwei Verläufe von $p(x)$ als Funktion von x, wenn $\mu(x)$ wie in (2.25) definiert ist. In Abb. 2.4 (a)

[4] Verg. Skriptum *Logistische Funktionen und Verteilungen, Epidemien und die Wechselwirkung zwischen Hormonen*, epidemiologie.pdf

Abbildung 2.4: Nichtlineare (logistische) Kopplung von μ und x; Parameterwerte s. Text



ergibt sich ein monoton steigender Verlauf, während in Abb. 2.4 (b) ein nichtlinearer Verlauf gezeigt wird. In (a) sind die Parameterwerte $B = 1.5$, $A = -.075$ und $C = -.5$. In (b) sind die Werte $B = 2.5$, $A = .975$, und $C = -.5$. Wird $\mu(x) = Ax+B$ angenommen, ergibt sich für $p(x)$ stets ein monoton steigender oder fallender Verlauf. Wird $\mu(x)$ wie in (2.25) definiert, kann sich also sowohl ein monotoner wie auch ein nicht-monotoner Verlauf ergeben. Reflektieren die Parameter A, B und C physiologische Parameter, so ist durchaus denkbar, dass sie nicht stets den gleichen Wert haben, sondern vom Gesamtzustand des Organismus abhängen. Sie verändern sich nur relativ langsam, dh sie sind über relativ kurze Zeiträume relativ konstant. Sind Wechsel von Verläufen (a) zu (b) möglich, so ist klar, dass eine gegebene Dosis x nicht stets den gleichen Effekt haben kann; der Effekt hängt eben vom Gesamtzustand des Organismus ab. Eine Diskussion von Beziehungen der Form (2.25) und ihrer Anwendung auf physiologische Prozesse findet man in Murray (1989) [5].

2.2.4 Psychodiagnostik: das Rasch-Modell

Die logistische Verteilung hat Anwendung in der Psychodiagnostik gefunden. Dazu kann man annehmen, dass die Fähigkeit oder die Ausprägung eines Merkmals ξ bei jeder Person gewissen zufälligen Schwankungen unterliegt. Eine Aufgabe wird gelöst oder ein Persönlichkeits"item" wird positiv beantwortet, wenn $\xi > S_i$ ist, wobei S_i eine itemspezifische Schwelle ist; bei Intelligenztests kann man S_i mit der Schwierigkeit einer Aufgabe identifizieren. x sei die "wahre" Ausprägung der Fähigkeit oder des gesteten Merkmals einer Person; x ist ein Messwert auf einer Skala, und der Erwartungswert $E(\xi) = \mu$ von ξ für dieser Person wird eine Funktion von x sein, so dass man $\mu = \mu(x)$ schreiben kann. $p(x) = P(\xi > S|x)$ ist dann die Wahrscheinlichkeit, dass eine Person mit der Fähigkeit x eine bestimmte Aufgabe aus einem Intelligenztest löst, oder die Wahrscheinlichkeit, dass eine Person mit der Ausprägung x eines Persönlichkeitsmerkmals ein bestimmtes Item eines Fragebogens mit "ja" beantwortet. Die Anwendung der logistischen Funktion auf diagnostische Fragen soll kurz explizit gemacht werden.

[5] vergl. Murray, J.D.: *Mathematical Biology*, Berlin 1989

Es war

$$P(\xi \leq S|x) = \frac{1}{1 + \exp\left[-\left(\frac{S-\mu(x)}{\sigma}\right)\frac{\pi}{\sqrt{3}}\right]},$$

wobei μ bereits als Funktion von x aufgefasst worden ist. Sicherlich kann man die Größe $\pi/\sqrt{3}$ und $1/\sigma$ in die Parameter S und μ absorbieren[6], so dass

$$P(\xi \leq S|x) = \frac{1}{1 + \exp[-(S^* - \mu^*(x))]} = \frac{1}{1 + \exp[\mu^*(x) - S^*]}$$

geschrieben werden kann, wobei $S^* = S\pi/(\sigma\sqrt{3})$, $\mu^*(x) = \mu\pi/(\sigma\sqrt{3})$ gesetzt wurde. Die Differenz $\mu^*(x) - S^*$ wird demnach in $\sigma\sqrt{3}/\pi$-Einheiten gemessen. Man kann diesen Ausdruck so interpretieren, dass er die Wahrscheinlichkeit der Lösung einer Aufgabe oder der positiven Beantwortung eines Fragebogenitems angibt, wenn die Person mit der "wahren" Ausprägung $\mu^*(x)$ des gemessenen Merkmals bei der Beantwortung mit der tatsächlichen (= momentanen) Ausprägung des Merkmals unterhalb einer bestimmten Schwelle bleibt. Der Ausdruck "wahr" bezieht sich nur darauf, dass $\mu^*(x)$ eine für die Person konstante Grösse ist (bzw. auf die Annahme, dass diese Grösse konstant ist), die bis auf einen Skalenfaktor oder bis auf eine lineare Transformation eindeutig ist. $\mu^*(x)$ ist eine weitere numerische Repräsentation der Fähigkeit bzw Ausprägung u. Man hat also $\mu^* \mapsto x \mapsto u$. Da über die Abbildung $u \mapsto x$ nichts weiter ausgesagt wurde, kann man die Abbildung $\mu^* \mapsto x$ auch gleich in x umbenennen, so dass man $\mu^*(x) = x$ setzen kann; auf diese Weise hat man den Parameter μ der logistischen Verteilung mit der Skalierung der Ausprägung u in Zusammenhang gebracht.

Die Schwelle $S^* = S_i^*$ wird durch die "Schwierigkeit" des jeweiligen, dh des i-ten Items definiert; sie ist ebenfalls bis auf eine lineare Transformation eindeutig. Man kann deshalb S^* als einen Schwierigkeitsparameter, der ein bestimmtes Item charakterisiert, auffassen. Dementsprechend setzt man

$$\kappa_i = S_i^*; \text{ für das } i\text{-te Item.} \tag{2.26}$$

Es werde nun eine Indikatorvariable Y_i eingeführt:

$$Y_i = \begin{cases} 1, & \text{Item } i \text{ wird gelöst, oder mit "ja" beantwortet} \\ \\ 0, & \text{Item } i \text{ wird nicht gelöst, oder mit "nein" beantwortet} \end{cases} \tag{2.27}$$

Weiter sei u die Fähigkeit einer Person, bzw. die Ausprägung des getesteten Merkmals bei einer Person, und $x = x(u)$ sei die Repräsentation von u auf einer geeignet gewählten Skala. Gesucht ist die Wahrscheinlichkeit $P(Y_i = 1|u) = P(Y_i = 1|x)$, die Repräsentation x wird also als eindeutig vorausgesetzt. Also kann man

$$1 - P(Y_i = 1|u) = P(Y_i = 0|x) = P(\xi \leq S|x) = \frac{1}{1 + \exp(x - \kappa_i)}$$

[6] Man spricht von einer *Reparametrisierung*.

schreiben. Also hat man $P(Y_i = 1|x) = 1 - 1/(1 + \exp(x - \kappa_i))$, dh man erhält das

Rasch-Modell:

$$P(Y_i = 1|x) = \frac{\exp(x - \kappa_i)}{1 + \exp(x - \kappa_i)}. \tag{2.28}$$

Das Modell wurde zuerst von dem dänischen Statistiker G. Rasch vorgeschlagen[7]. Das Rasch-Modell ergibt sich zunächst als eine "natürliche", dh direkte Anwendung der logistischen Verteilung auf die Frage, wie die Wahrscheinlichkeiten von bestimmten Antworten zu modellieren seien. Es wird wie in der Klassischen Testtheorie (implizit) angenommen, dass die Ausprägung des gemessenen Merkmals zufällig um einen bestimmten Wert mit der Varianz σ^2 schwankt; ist die Ausprägung zum Zeitpunkt des Tests oberhalb einer durch die Aufgabe definierten Schwelle, so wird das Item positiv beantwortet, andernfalls nicht. Von dem gemessenen Merkmal wird angenommen, dass es logistisch statt normalverteilt ist. Diese Annahme scheint auf den ersten Blick keinen großen Unterschied zur üblichen Normalverteilungsannahme zu machen, da die logistische Verteilung bei geeigneter Parametrisierung kaum von der Normalverteilung zu unterscheiden ist.

Spezifische Objektivität: Tatsächlich ergeben sich markante Unterschiede zur "klassischen" Annahme der Normalverteilung. In der Klassischen Testtheorie (KT) wird die Schwierigkeit eines Items definiert als der Anteil der Population, die eine positive Antwort auf das Item geben; damit ist die Schwierigkeit des Items eine populationsabhängige Größe. Gleichzeitig wird der "Fähigkeitsparameter" in der KT relativ zur Schwierigkeit interpretiert, so dass sich eine gewisse begriffliche Zirkularität ergibt. Es zeigt sich aber, dass die Parameter x und κ_i in (2.28) in bestimmter Weise unabhängig voneinander betrachtet werden können. Dazu betrachte man den Logit-Wert

$$\log \frac{P(Y_i = 1|x)}{1 - P(Y_i = 1|x)} = \log \frac{P(Y_i = 1|x)}{P(Y_i = 0|x)} = x - \kappa_i. \tag{2.29}$$

Gegeben seien nun zwei Probanden mit den Fähigkeiten x_1 und x_2. Dann ist

$$\log \frac{P(Y_i = 1|x_1)}{P(Y_i = 0|x_1)} - \log \frac{P(Y_i = 1|x_2)}{P(Y_i = 0|x_2)} = (x_1 - \kappa_i) - (x_2 - \kappa_i) = x_1 - x_2. \tag{2.30}$$

Der Unterschied zwischen zwei Personen, ausgedrückt als Differenz der Fähigkeitsparameter, ist also unabhängig vom Schwierigkeitsparameter κ_i irgendeines Items. Auf analoge Weise zeigt man, dass

$$\log \frac{P(Y_i = 1|x)}{P(Y_i = 0|x)} - \log \frac{P(Y_j = 1|x)}{P(Y_j = 0|x)} = \kappa_j - \kappa_i, \tag{2.31}$$

dh die Differenz zweier Schwierigkeitsparameter ist unabhängig von x, der Fähigkeit eines Probanden, und damit für alle Probanden gleich. Diese in den Gleichungen

[7] Rasch, G.: Probabilistic models for some intelligence and attainment tests. Kopenhagen, Nissen & Lydecke, 1960

(2.30) und (2.31) ausgedrückte Unabhängigkeit der Unterschiede zwischen Personparametern $x_1, x_2, \ldots$ einerseits und Itemparametern $\kappa_1, \kappa_2, \ldots$ andererseits definiert den Begriff der *spezifischen Objektivität* des auf der logistischen Verteilung beruhenden Testmodells. Die Eigenschaft der spezifischen Objektivität ist charakteristisch für die logistische Verteilung; das Modell liefert also einen Ausweg aus der oben genannten Zirkularität von Schwierigkeit und Fähigkeitsabschätzung.

Es sei angemerkt, dass das Rasch-Modell oft in anderer Parametrisierung diskutiert wird. Es ist ja

$$\exp(x - \kappa_i) = \exp(x)\exp(\kappa_i) = \theta\delta_i, \tag{2.32}$$

mit $\theta = \exp(x)$ und $\delta_i = \exp(\kappa_i)$. Damit kann man

$$P(Y_i = 1|x) = \frac{\theta\delta_i}{1 + \theta\delta_i}. \tag{2.33}$$

schreiben; man spricht von *multiplikativer Parametrisierung*. Der Parameter δ_i heißt auch *Itemleichtigkeit*, denn $P(Y_i = 1|x)$ ist um so größer, je größer δ_i ist. Nun ist $P(Y_i = 0|) = 1 - \theta\delta_i(1+\theta\delta_i) = 1/(1+\theta\delta_i)$; das Rasch-Modell wird dementsprechend auch in der knappen Form

$$P(Y_i = y_i|x) = \frac{(\theta\delta_i)^{y_i}}{1 + \theta\delta_i}, \;\; y_i = 0, 1 \tag{2.34}$$

geschrieben; für $y_i = 1$ erhält man die Wahrscheinlichkeit einer positiven Antwort, für $y_i = 0$ resultiert die Wahrscheinlichkeit einer negativen Antwort.

Die spezifische Objektivität der Parameter für die Personen einerseits und die Items ist eine attraktive Eigenschaft des Rasch-Modells. Sie ist eine Implikation der Eigenschaft von ξ, logistisch verteilt zu sein. Obwohl die logistische Verteilung und die Normalverteilung *numerisch* sehr ähnlich sind, gilt die spezifische Objektivität für die Annahme der Normalverteilung, die in der KT gemacht wird, nicht. Die Frage ist nun, wie man entscheidet, dass entweder die logistische Verteilung oder die Normalverteilung die wahre Verteilung von ξ ist. Wegen der numerischen Ähnlichkeit läßt sich die Frage nicht durch Tests für die Güte der Anpassung entscheiden. Man benötigt also andere Kriterien für die Korrektheit der Annahme etwa der logistischen Verteilung und damit für die empirische Korrektheit der Gleichungen (2.30) und (2.31). Die Frage nach diesem Kriterium kann im Rahmen dieses Skriptums nicht weiter diskutiert werden.

2.2.5 Die Schätzung der Parameter

> **Hinweis:** Dieser Absschnitt ist nur für Personen gedacht, die die Details der Schätzmethode kennenlernen wollen; der Abschnitt ist nicht klausurrelevant! Es sei aber auf Beispiel 16, Seite 50, verwiesen, in dem ein wichtiger Spezialfall illustriert wird.

Obwohl x im Prinzip stetig variieren kann, wird man im allgemeinen die Wirkung von nur endlich vielen Werten $x_1, \cdots, x_k, \cdots, x_r$ untersuchen. Die abhängige Variable y nehme aber nur die zwei Werte 1 oder 0 an. Man hat dann eine Datentabelle von der Form der Tabelle 2.3.

Es werde nun angenommen, dass

$$p(x_k) = p(y = 1|x_k) = \frac{\exp(Ax_k + B)}{1 + \exp(Ax_k + B)}$$

gilt. Die Parameter A und B sind unbekannt und müssen aus den Daten, d.h. aus den n_{ks}, $s = 1, 2$ geschätzt werden. Dazu wird die Maximum-Likelihood-Methode angesetzt. Zur Vereinfachung werde

$$p_k = p(x_k), \quad n_k = n_{k1}, \quad N_k = n_{k+}$$

gesetzt. Sicherlich ist dann

$$N_k - n_k = n_{k2}$$

Die Likelihood-Funktion ist dann durch

Tabelle 2.3: Logistische Regression, r Prädiktorwerte

	Abhängige Variable		
Unabh. Variable	$y = 1$	$y = 0$	Σ
x_1	n_{11}	n_{12}	n_{1+}
x_2	n_{21}	n_{22}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	n_{k+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_r	n_{r1}	n_{r2}	n_{r+}
Σ	n_{+1}	n_{+2}	$n_{++} = n$

$$L = \prod_{k=1}^{r} p_k^{n_k} (1 - p_k)^{N_k - n_k} \tag{2.35}$$

gegeben, und daraus erhält man sofort die Log-Likelihood-Funktion

$$\log L = \sum_k (n_k \log p_k + (N_k - n_k)(1 - p_k)) \tag{2.36}$$

Um die Schätzungen für A und B zu gewinnen, müssen die partiellen Ableitungen von $\log L$ bezüglich A und B hergeleitet werden. Dazu wird der folgende Satz bewiesen:

Satz 6 *Es gelte*

$$P(y=1|x_k) = p_k = \frac{\exp(\phi_k(x_k))}{1+\exp(\phi_k(x_k))}, \quad \phi_k(x_k) = Ax_k + B \tag{2.37}$$

Weiter sei entweder $\theta = A$ oder $\theta = B$; ϕ_k kann dann jeweils als Funktion von θ aufgefaßt werden. Dann gilt

$$\frac{\partial \log L}{\partial \theta} = \sum_k (n_k - N_k p_k) \frac{\partial \phi_k}{\partial \theta}. \tag{2.38}$$

Beweis: Aus (2.36) erhält man sofort

$$\frac{\partial \log L}{\partial \theta} = \sum_k \left(n_k \frac{1}{p_k} \frac{\partial p_k}{\partial \theta} - (N_k - n_k) \frac{1}{1-p_k} \frac{\partial p_k}{\partial \theta} \right) = \sum_k \left(\frac{n_k}{p_k} - \frac{N_k - n_k}{1-p_k} \right) \frac{\partial p_k}{\partial \theta}$$

Nun ist

$$\frac{\partial p_k}{\partial \theta} = \frac{dp_k}{d\theta} \frac{\partial \phi_k}{\partial \phi_k},$$

und

$$\frac{dp_k}{d\phi_k} = \frac{e^{\phi_k}(1+e^{\phi_k}) - e^{2\phi_k}}{(1+\exp(\phi_k))^2} = \frac{e^{\phi_k}}{(1+\exp(\phi_k))^2} = \frac{p_k}{1+\exp(\phi_k)}$$

Deshalb erhält man

$$\begin{aligned} \frac{\partial \log L}{\partial \theta} &= \sum_k \frac{n_k - n_k p_k - N_k p_k + n_k p_k}{p_k(1-p_k)} \frac{\partial p_k}{\partial \theta} = \\ &= \sum_k \frac{n_k - N_k p_k}{p_k(1-p_k)} \frac{p_k}{(1+\exp(\phi_k))} \frac{\partial \phi_k}{\partial \theta} = \\ &= \sum_k \frac{n_k - N_k p_k}{(1-p_k)(1+\exp(\phi_k))} \frac{\partial \phi_k}{\partial \theta} \end{aligned}$$

Aber es ist

$$(1-p_k)(1+e^{\phi_k}) = \frac{1}{1+\exp(\phi_k)}(1+e^{\phi_k}) = 1,$$

und damit folgt (2.38). □

Die ML-Schätzungen für A und B erhält man aus (2.38), indem man die Gleichungen

$$\left. \frac{\partial \log L}{\partial \theta} \right|_{\theta=\hat{\theta}} = 0 \tag{2.39}$$

bildet.

Der allgemeine Fall

Nach Satz 6, Gleichung (2.38) gilt für die Likelihood-Funktion

$$\frac{\partial \log L}{\partial \theta} = \sum_k (n_k - N_k p_k) \frac{\partial \phi_k}{\partial \theta},$$

Da $\theta = A$ oder $\theta = b$ hat man zwei Gleichungen; gleich Null gesetzt erhält man

$$\sum_k n_k \frac{\partial \phi_k}{\partial \theta} = \sum_k N_k p_k \frac{\partial \phi_k}{\partial \theta} \tag{2.40}$$

Wegen

$$\frac{\partial \phi_k}{\partial \theta} = \begin{cases} x_k, & \theta = A \\ 1, & \theta = B \end{cases} \tag{2.41}$$

erhält man für (2.40) in ausführlicher Schreibweise

$$\sum_k n_k x_k = \sum_k N_k \hat{p}_k x_k \tag{2.42}$$

$$\sum_k n_k = \sum_k N_k \hat{p}_k \tag{2.43}$$

Hier ist $\hat{p}_k$ statt p_k geschrieben worden, weil die Gleichungen für die Schätzungen $\hat{A}$ und $\hat{B}$ gelten. Für $\hat{\pi}_k$ setzt man nun den in (2.37) gegebenen Ausdruck ein, der dann nach $\hat{A}$ und $\hat{B}$ aufgelöst werden muß. Da die Wahrscheinlichkeiten $\hat{\pi}_k$ in nichtlinearer Weise von $\hat{A}$ und $\hat{B}$ abhängen, kann man die Lösungen für $\hat{A}$ und $\hat{B}$ im allgemeinen Fall – wenn also x nicht mehr auf nur zwei Werte, 0 und 1, beschränkt ist – nicht mehr in geschlossener Form hinschreiben. Die Gleichungen müssen numerisch gelöst werden; hierzu kann z.B. die Newton-Raphson-Methode herangezogen werden. Computerprogramme zur logistischen Regression enthalten diesen oder einen ähnlichen Algorithmus.

Spezialfall: ein dichotomer Prädiktor

Einen einfachen Spezialfall hat man, wenn der Prädiktor x eine dichotome unabhängige Variable repräsentiert, wenn x also nur die Werte 0 oder 1 annehmen kann. Die Kontingenztabelle ist dann eine einfache 2×2-Tabelle: Nach (2.38) muß zunächst $\partial \phi_k / \partial \theta$ bestimmt werden. Für $\theta = A$ ist wegen $\phi_k = A x_k + B$

$$\frac{\partial \phi_k}{\partial \theta} = \frac{\partial \phi_k}{\partial A} = x_k$$

und da x_k nur die Werte 0 oder 1 annehmen kann ist

$$\frac{\partial \phi_k}{\partial A} = \begin{cases} 0, & x_k = 0 \\ 1, & x_k = 1 \end{cases}$$

Tabelle 2.4: Logistische Regression für einen dichotomen Prädiktor

	Abhängige Variable (y)	
Unabh. Variable (x)	$y = 1$	$y = 0$
$x = 1$	$e^{A+B}/(1 + e^{A+B})$	$1/(1 + e^{A+B})$
$x = 0$	$e^{B}/(1 + e^{B})$	$1/(1 + e^{B})$

Für $\theta = B$ findet man

$$\frac{\partial \phi_k}{\partial B} = 1$$

für beide x-Werte. Dementsprechend liefert (2.38) die Gleichungen

$$\left.\frac{\partial \log L}{\partial A}\right|_{A=\hat{A}} = n_1 - \frac{N_1 \exp(\hat{A}+\hat{B})}{1+\exp(\hat{A}+\hat{B})} = 0 \tag{2.44}$$

$$\left.\frac{\partial \log L}{\partial B}\right|_{B=\hat{B}} = n_1 - \frac{N_1 \exp(\hat{A}+\hat{B})}{1+\exp(\hat{A}+\hat{B})} + n_2 - \frac{N_2 \exp(\hat{B})}{1+\exp(\hat{B})} = 0 \tag{2.45}$$

Die Gleichung (2.45) vereinfacht sich aber wegen (2.44) sofort zu

$$\left.\frac{\partial \log L}{\partial B}\right|_{B=\hat{B}} = n_2 - \frac{N_2 \exp(\hat{B})}{1+\exp(\hat{B}))} = 0$$

woraus sofort

$$\frac{n_2}{N_2 - n_2} = \exp(\hat{B}) \tag{2.46}$$

folgt.

Aus (2.44) folgt

$$n_1(1 + \exp(\hat{A}+\hat{B})) = N_1 \exp(\hat{A}+\hat{B})$$

so dass

$$\frac{n_1}{N_1 - n_1} = \exp(\hat{A}+\hat{B}),$$

und wegen (2.46) hat man

$$\frac{n_1}{N_1 - n_1} = \frac{n_2}{N_2 - n_2}\exp(\hat{A})$$

so dass

$$\frac{n_1}{N_1 - n_1}\frac{N_2 - n_2}{n_2} = \exp(\hat{A}) \tag{2.47}$$

Nun war $n_k = n_{k1}$ und $N_k - n_k = n_{k2}$. Dementsprechend ist (vergl. (2.6))

$$\frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{\Omega_1}{\Omega_2} = \Theta = \exp(\hat{A}), \quad \text{bzw. } \log\Theta = \hat{A} \tag{2.48}$$

und (vergl. (2.4))

$$\frac{n_{21}}{n_{22}} = \Omega_2 = \exp(\hat{B}), \quad \text{bzw. } \log\Omega_2 = \hat{B} \tag{2.49}$$

Die ursprünglich für deskriptive Zwecke eingeführten Größen *Odds* und *Kreuzproduktverhältnis* liefern also gerade die ML-Schätzungen für die unbekannten Parameter A und B.

Beispiel 16 (Fortsetzung des Beispiels 14) In Beispiel 14 werde

$$x = \begin{cases} 0, & \text{Placebo} \\ 1, & \text{Aspirin} \end{cases} \tag{2.50}$$

definiert. Für die Odds hat man mit der Abkürzung HI für "Herzinfarkt"

$$\Omega_1 = \frac{P(\text{HI; ja}|\text{Aspirin})}{P(\text{HI; nein}|\text{Aspirin})} = e^{\hat{A}+\hat{B}}, \quad \Omega_2 = \frac{P(\text{HI; ja}|\text{Placebo})}{P(\text{HI; nein}|\text{Placebo})} = e^{\hat{B}},$$

und für das Kreuzproduktverhältnis erhält man

$$\Theta = \frac{\Omega_1}{\Omega_2} = \frac{\exp(\hat{A}+\hat{B})}{\exp(\hat{B})} = e^{\hat{A}};$$

hier reflektiert Θ den Zusammenhang zwischen der Medikamenteneinnahme (Placebo oder Aspirin) und der Erkrankung (Herzinfarkt oder nicht). Der Befund $\Theta = \exp(\hat{A})$ oder $\log\Theta = \hat{A}$ zeigt, dass tatsächlich θ und das Steigmaß A der Regressionsgeraden $Ax + B$ direkt aufeinander bezogen sind. Für $A = 0$ hängt die Erkrankung nicht von der Medikamention ab, und $\log\Theta = 0$ genau dann, wenn $\Theta = 1$, d.h. wenn $\Omega_1 = \Omega_2$.

Aus (2.48) erhält man die Schätzung

$$\hat{A} = \log\Theta = \log 1.833 = .60595$$

und aus (2.49)

$$\hat{B} = \log\frac{n_{11}}{n_{12}} = .01743$$

In Anmerkung 2b, Seite 39, wurde angemerkt, dass im Falle eines kategorialen Prädiktors der Parameter A wie eine Effektgrösse interpretiert werden kann, dh $A \propto (\mu_1 - \mu_2)/\sigma$. $\hat{A} = .60595$ bedeutet dann, dass die Differenz der Mittelwerte gerade $\approx .61$ σ-Einheiten beträgt.

Nach (1.21) enthält die Informationsmatrix

$$I(\theta) = -E\left(\frac{\partial^2 \log L}{\partial\theta_i\partial\theta_j}\right)$$

die Varianzen und Kovarianzen der Schätzungen für A und B. Da die Schätzungen darüber hinaus asymptotisch normalverteilt sind, können dann Konfidenzintervalle für die Schätzungen bestimmt werden. Diese Berechnungen werden dem Leser zur Übung überlassen.

□

2.2.6 Verallgemeinerung: Mehrere Prädiktorvariablen

In (2.20) sind die Logits

$$\log \frac{p(x)}{1-p(x)} = Ax + B$$

eingeführt worden. Dieser Ansatz kann sofort auf mehr als eine Prädiktorvariable verallgemeinert werden:

$$\log \frac{p(\vec{x})}{1-p(\vec{x})} = b_o + b_1 x_1 + \cdots + b_k x_k, \;\; \vec{x} = (x_1, \ldots, x_k)' \tag{2.51}$$

Hier können k "erklärende" Variable x_j zur Bestimmung von $p(\vec{x})$, herangezogen werden. Das Modell entspricht dem der multiplen Regression.

Es ist hier allerdings nicht notwendig, dass die x_j Intervallskalenniveau haben. Es ist möglich, dass die x_j kategoriale Variablen sind. Der Unterschied zu den später zu betrachtenden *loglinearen Modellen* besteht dann eigentlich nur noch darin, dass eine (oder mehrere) Variable als Antwortvariable ausgezeichnet werden, die Betrachtung in dieser Hinsicht also asymmetrisch ist.

Im Falle kategorialer erklärender Variablen entsprechen die b_j in (2.51) den Haupteffekten in einer Varianzanalyse. Dies führt zur *Effektkodierung* für die x_j: es gelte

$$x_j = \begin{cases} 1, & \text{es liegt Kateg. } j \text{ vor} \\ -1, & \text{es liegt Kateg. } k \neq j \text{ vor} \\ 0, & \text{sonst} \end{cases} \tag{2.52}$$

für $j = 1, \cdots, k-1$; es muß nur bis zur $(k-1)$-ten Kategorie betrachtet werden, dass die letzte bei Bestimmung dieser Kategorien festliegt. Dann gilt

$$b_k = -\sum_{j=1}^{k-1} b_j, \quad \text{oder} \quad \sum_{j=1}^{k} b_j = 0 \tag{2.53}$$

Dementsprechend gilt dann für eine gegebene Link-Funktion g

$$g(p(x)) = b_0 + b_j, \quad j = 1, \cdots, k-1 \tag{2.54}$$

und

$$g(p(x)) = b_0 - b_1 - \cdots - b_{k-1}, \quad j = k \tag{2.55}$$

Bei varianzanalytischen Designs ist es üblich, mehr als nur eine unabhängige Variable (Faktoren) zu betrachten. Es seien etwa die Faktoren A und B gegeben. Sie

können durch die Vektoren X^A und X^B repäsentiert werden, deren Komponenten durch

$$x_i^A, i = 1, \cdots, I-1; \quad x_j^B, j = 1, \cdots, J-1 \tag{2.56}$$

gegeben sind und die wie in (2.52) als Dummy-Variablen definiert sind. $x_i^A = x_j^B = 1$ heißt dann, dass die Bedingung $A_i \cap B_j$ (A_i und B_j) gegeben ist. Die beiden Vektoren können zu einem einzelnen Merkmalsvektor zusammengefaßt werden:

$$X = (x_1^A, \cdots, x_{I-1}^A, x_1^B, \cdots, x_{J-1}^B, x_1^A x_2^B, \cdots, x_1^A x_{J-1}^B, \cdots, x_{I-1}^A x_{J-1}^B)' \tag{2.57}$$

Die Komponenten $x_1^A x_1^B$ etc. repräsentieren die möglichen Wechselwirkungskomponenten. Dann ergibt sich das Modell

$$g(\pi(X)) = b_0 + X'b \tag{2.58}$$

wobei b ein Vektor ist, der die Haupt- und Wechselwirkungseffekte repräsentiert. Für g kann wieder die Logit-Transformation gewählt werden.

2.2.7 Parameterschätzungen

Es ist

$$p(x) = \frac{\exp(\vec{x}'b)}{1 + \exp(\vec{x}'b)} = \frac{1}{1 + \exp(-\vec{x}'b)}, \tag{2.59}$$

mit

$$\vec{x} = (1, x_1, \cdots, x_r)', \quad b = (b_1, b_2, \cdots, b_r)' \tag{2.60}$$

Für die abhängige Variable y wird wieder der binäre Fall angenommen, d.h. es soll

$$y = \begin{cases} 1 \\ 0 \end{cases}$$

gelten. Um die Parameter zu schätzen, kann der gleiche Ansatz wie in Satz 6 gemacht werden, nur, dass für $\phi_k(x)$ nun der multiple Regressionsansatz

$$\phi_k = b_0 + b_1 x_{k1} + \cdots + b_r x_{kr} \tag{2.61}$$

gemacht wird. Dann folgt die zu (2.38) analoge Aussage

$$\frac{\partial \log L}{\partial \theta} = \sum_k (n_k - N_k p_k) \frac{\partial \phi_k}{\partial \theta},$$

wobei nun allerdings

$$\frac{\partial \phi_k}{\partial \theta} = \frac{\partial \phi_k}{\partial b_j} = x_{kj}; \quad j = 1, \cdots, r \tag{2.62}$$

gilt, x_{kj} ist die k-te Realisierung der Prädiktorvariablen x_j.

Tabelle 2.5: Risiko einer Infektion beim Kaiserschnitt (KS); – Kaiserschnitt geplant, - nicht geplant

		KS geplant		KS nicht geplant	
		Infektion		Infektion	
Antibiotika	Risikofaktor (RF)	ja	nein	ja	nein
gegeben	vorhanden	1	17	11	87
	nicht vorhanden	0	2	0	0
nicht gegeben	vorhanden	28	30	23	3
	nicht vorhanden	8	32	0	9

Beispiel 17 Infektionsrisiko bei Kaiserschnittgeburten Es soll das Risiko einer Infektion bei einer Geburt mit Kaiserschnitt berechnet werden. Der Kaiserschnitt kann geplant oder nicht geplant durchgeführt werden, und die Patientin kann Risikofaktoren haben oder nicht. Man hatte die folgenden Daten zur Verfügung:

$$x_1 = \begin{cases} 1 & \text{nicht gepl} \\ 0 & \text{gepl} \end{cases}, \; x_2 = \begin{cases} 1 & \text{RF} \\ 0 & \text{kein RF} \end{cases}, \; x_3 = \begin{cases} 1 & \text{AB} \\ 0 & \text{kein AB} \end{cases} \tag{2.63}$$

Das Modell (2.64) ist das *Haupteffektmodell*, da keinerlei Interaktionen zwischen den drei unabhängigen Variablen angenommen werden.

$$\log \frac{p(\text{Infektion}|\vec{x})}{p(\text{keine Infektion}|\vec{x})} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3. \tag{2.64}$$

Daraus folgt

$$\frac{p(\text{Infektion}|\vec{x})}{p(\text{keine Infektion}|\vec{x})} = \exp(\beta_0)\exp(\beta_1 x_1)\exp(\beta_2 x_2)\exp(\beta_3 x_3) \tag{2.65}$$

Für einen speziellen Vektor $\vec{x}_i$ wird die abgekürzte Schreibweise

$$\rho_i = \frac{p(\text{Infektion}|\vec{x}_i)}{p(\text{keine Infektion}|\vec{x}_i)} = \frac{p(\text{I}|\vec{x}_i)}{p(\neg\text{I}|\vec{x}_i)} \tag{2.66}$$

eingeführt. Die Daten werden in der Tabelle 2.5 gezeigt, und die geschätzten Parameterwerte findet man in Tabelle 2.6. Ein nicht geplanter Kaiserschnitt erhöht nach (2.65) das Infektionsrisiko um den Faktor $\exp(\beta_1 x_1) = \exp(1.07) = 2.92$, ein vorhandener Risikofaktor erhöht das Infektionsrisiko um den Faktor $\exp(\beta_2 x_2) = \exp(2.03) = 7.6$, und ein Antibiotikum erniedrigt das Risiko um den Faktor $\exp(\beta_3 x_3) = \exp(-3.25) = .0388$, d.h. hat man

$$p(\text{Infektion})/p(\text{keine Infektion}) = 1$$

im Falle keiner Antibiotikagabe, so wird das Risko auf

$$p(\text{Infektion})/p(\text{keine Infektion}) = .0388$$

Tabelle 2.6: Parameterwerte

	Gewicht/Prädiktor			
	1	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Wert	-1.89	1.07	2.03	-3.25
$\sqrt{\text{Var}(\hat{\beta})}$	.41	.43	.46	.48
t	-4.61	2.49	4.41	-6.77

Tabelle 2.7: Mögliche Fälle; $\rho_i = p(I|\vec{x}_i)/p(\neg I|\vec{x}_i)$, $1 \leq i \leq 8$

Fall	x_1		x_2		x_3		ρ_i
1	1	npl	1	RF	1	AB	1.300
2	0	pl	1	RF	1	AB	.445
3	1	npl	0	kRF	1	AB	.171
4	1	npl	1	RF	0	kAB	33.509
5	1	npl	0	kRF	0	kAB	.445
6	0	pl	1	RF	0	kAB	11.476
7	0	pl	0	kRF	1	AB	.056
8	0	pl	0	kRF	0	kAB	1.510

bei Antibiotikagabe gesenkt. Es sind weitere Abschätzungen für verschiedene Vektoren $\vec{x}_i = (x_1, x_2, x_3)'$ möglich, vergl. Tabelle 2.7. Die x_1, x_2, x_3 nehmen Werte an, die durch den i-ten Fall angezeigt sind. Man kann jetzt noch Vergleiche von ρ_i mit ρ_j berechnen. So kann man die Wirkung von Antibiotika für den Fall bestimmen, dass der Kaiserschnitt nicht geplant war und Risikofaktoren vorliegen. Man findet

$$\frac{\rho_4}{\rho_1} = \frac{33.509}{1.300} = 25.776, \tag{2.67}$$

oder

$$\frac{\rho_1}{\rho_4} = \frac{1.300}{33.509} = .0388. \tag{2.68}$$

Bei einem nicht geplanten Kaiserschnitt im Fall von Risikofaktoren erhöht sich das Risiko einer Infektion fast um das 26-fache, wenn *keine* Antibiotika gegeben werden, bzw. reduziert sich auf das .04-fache, wenn Antibiotika gegeben werden.

Die bisherigen Betrachtungen illustrieren die Art und Weise, in der Parameter fürdas gegebene Modell interpretiert werden. Es ist natürlich möglich, zu testen, ob das Modell überhaupt mit den Daten verträglich ist; die Methodologie solcher Tests wird in Abschnitt 2.6 dargestellt. Hier soll nur angemerkt werden, dass die *Deviance* den Wert 10.997 hat, was bedeutet, dass mit Bezug auf $\alpha = .05$ das Modell *extrem schlecht passt*! Die in Tabelle 2.6 angegebenen t-Werte sind hochsignifkant, aber das besagt ja noch nicht, dass das Modell insgesamt mit den Daten kompatibel ist.

Man kann nun eine Erweiterung des Modells testen. Diese Erweiterung ist durch

$$\log \frac{p(\text{Infektion}|\vec{x})}{p(\text{keine Infektion}|\vec{x})} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 \tag{2.69}$$

gegeben. Es wird also noch eine Wechselwirkung zwischen der Planung des Kaiserschnitts und dem Risikofaktor angenommen. Es ergeben sich die folgenden Parameterschätzungen: Der Vergleich mit Tabelle 2.6 zeigt, dass nicht nur ein Parame-

Tabelle 2.8: Parameter für das Modell mit Wechselwirkung

	Gewicht/Prädiktor				
	1	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Wert	-1.39	-11.64	1.36	-3.83	13.49
$\sqrt{Var(\hat{\beta})}$	.40	136.20	.47	.60	136.20

terwert hinzugekommen ist, sondern sich die Werte für die ersten drei Parameter darüber hinaus verändert haben.

Man kann weiter Interaktionsterme x_1x_2 und x_1x_3 statt der Interaktion x_2x_3 betrachten. Die Deviancen für die drei Modelle sind in der Tabelle 2.9 zusammengefasst worden. Man entnimmt der Tabelle, dass die Differenz des Modells mit der

Tabelle 2.9: Vergleich verschiedener Interaktionsmodelle

	Deviance	Freiheitsgrade	Diff. zum Haupteffektmodell
Haupteffektmodell	10.997	3	·
Haupteffekt + x_1x_2	.955	2	10.04
Haupteffekt + x_1x_3	10.918	2	.078
Haupteffekt + x_2x_3	10.974	2	.023

Interaktion x_1x_2 (Planung × Risikofaktor) den größten Abstand zum Haupteffektmodell hat, während der Effekt der übrigen möglichen Wechselwirkungen vernachlässigbar erscheint. Es erscheint demnach sinnvoll, die Wechselwirkung Planung × Risikofaktor als wesentlich für die Beziehung zwischen den unabhängigen Variablen und der Infektionsgefahr mit einzubeziehen. Es sind keine individuellen p-Werte für die einzelnen Modelle angegeben worden, - sie würden für die gegebenen Daten die Vernachlässigung des Interaktionsterms anzeigen, was aber nicht sinnvoll wäre, da das Haupteffektmodell ja nicht passt. Die Differenzen der Deviancen in Tabelle 2.9 dagegen gelten als robuste Statistik. □

2.3 Das Probit-Modell

Man kann statt der logistischen Verteilung für ξ auch die Gauß-Verteilung annehmen. Dann soll also

$$P(\xi \leq S) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{S} f(\xi)\, d\xi \tag{2.70}$$

gelten, oder $\xi \sim N(\mu, \sigma^2)$. Um die Abhängigkeit von μ deutlich zu machen, geht man von ξ zu den standardisierten Werten $z = (\xi - \mu)/\sigma$ über. Es ist

$$P(\xi \leq S|x) = P\left(\frac{\xi - \mu}{\sigma} \leq \frac{S - \mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(S-\mu)/\sigma} f(z)\, dz = F_z(\alpha - \beta\mu) = p(x)$$

mit $\alpha = S/\sigma$, $\beta = 1/\sigma$. Man macht wieder die Annahme

$$\mu = ax + b \tag{2.71}$$

Dann ist

$$\alpha - \beta\mu = \alpha - \beta(ax + b) = Ax + B$$

mit $B = \alpha - b\beta$ und $A = -a\beta$. Die inverse Transformation

$$F_z^{-1}(p(x)) = Ax + B \tag{2.72}$$

liefert die Transformation von $p(x)$, für die eine in x lineare Beziehung angesetzt werden kann.

Definition 13 *Die Link-Funktion (2.72), wobei F_z die Verteilungsfunktion einer standardisierten Gauß-Verteilung ist, heißt* Probit-Transformation.

Hier ist $g(p(x)) = F_z^{-1}(p(x))$.

Die Logit- und die Probit-Transformationen sind Beispiele für *Link-Funktionen*, d.h. von Funktionen $g(\pi(X))$, die x bzw. $Ax + B$ mit der Wahrscheinlichkeit $p(x)$ des Eintretens des zur Diskussion stehenden Merkmals verbinden.

Beispiel 18 (Fortsetzung von Beispiel 17) Man kann die Daten der Tabelle 2.5 auch mit der Probit-Analyse untersuchen. Es ergeben sich die Parameterschätzungen

$$\hat{\beta}_0 = -1.09, \quad \hat{\beta}_1 = .61, \quad \hat{\beta}_2 = 1.20, \quad \hat{\beta}_3 = -1.90$$

Man überprüft leicht, dass die relativen Effekte nahezu die gleichen wie beim Logit-Modell sind. □

2.4 Modelle für Anzahlen

Bei der logistischen Regression liegt die Gesamtzahl der Beobachtungen fest und man bestimmt, für einen bestimmten Vektor $\vec{x} = (x_1, x_2, \ldots, x_k)'$ den Anteil der y-Werte, die gleich 1 sind, an der Gesamtzahl n der Beobachtungen. Gelegentlich ist aber n gar nicht bestimmt, d.h. $n_{y=1}$ liegt zwischen 0 und ∞, $0 \leq n_{y=1} < \infty$. Die einzelnen Beobachtungen seien unabhängig voneinander. die Verteilung der Häufigkeiten kann dann als Poisson-Verteilung angesehen werden:

$$P(n_{y=1} = k|\vec{x}) = e^{-\lambda}\frac{\lambda^k}{k!}, \;\; k = 0, 1, 2, \ldots \tag{2.73}$$

Es gilt

$$\mathbb{E}(n_{y=1}) = \lambda, \;\; Var(n_{y=1}) = \lambda. \tag{2.74}$$

λ kann als Funktion von $\vec{x}$ aufgefasst werden. Der einfachste Ansatz, diese Abhängigkeit auszudrücken, ist

$$\lambda = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k. \tag{2.75}$$

In diesem Fall werden keine Wechselwirkungen zwischen den Einflußfaktoren, die durch die x_i repräsentiert werden, angenommen. Eine Alternative ist, für den Fall $k = 2$,

$$\lambda = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2. \tag{2.76}$$

Hier wird angenommen, dass die beiden Einflußfaktoren miteinander wechselwirken, und dass diese Wechselwirkung durch das Produkt $x_1 x_2$ ausgedrückt werden kann.

Die Schätzung der $\beta_0, \beta_1, \ldots, \beta_k$ wird mittels der Maximum-Likelihood-Methode vorgenommen.

Die Ansätze (2.75) und (2.76) repräsentieren das *lineare Poisson-Modell*. Es hat den Nachteil, dass wegen der notwendigen Bedingung $\lambda > 0$ möglicherweise Restriktionen für die Parameter $\beta_0, \beta_1, \ldots, \beta_n$ betrachtet werden müssen. Deswegen wird häufig das *log-lineare Poisson-Modell*

$$\lambda = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) = \exp(\beta_0 \exp(\beta_1 x_1) \cdots \exp(\beta_n x_n) \tag{2.77}$$

(in diesem Fall ohne Wechselwirkungen) betrachtet, bei dem die Parameter keinerlei Restriktionen unterliegen.

Beispiel 19 In einer biomedizinischen Studie[8] soll die immunaktivierenden Fähigkeiten des Tumornekrosisfaktors (TNF) und des Interferons (IFN) in bezug auf die Zelldifferentiation untersucht werden. Dazu wurde die Anzahl der Zellen ausgezählt, die Differenzierungsmarker aufwiesen, nachdem sie der Wirkung von TNF und IFN ausgesetzt wurden. Es gab insgesamt 16 Dosiskombinationen TNF-IFN, und für jede Kombination wurden 200 Zellen beobachtet. Nimmt man an, dass

[8] Piegorsch, W.W., Weinberg, C.R., Margolin, B.H. (1988) Exploring simple independent action in multifactor tables of proportions. Biometrics, 44, 595-603

Tabelle 2.10: Zelldifferenzierung (Beispiel 19)

k	Dosis TNF [U/ml]	Dosis IFN [U/ml]
11	0	0
18	0	4
20	0	20
39	0	100
22	1	0
38	1	4
52	1	20
69	1	100
31	10	0
68	10	4
69	10	20
128	10	100
102	10	0
171	100	4
180	100	20
193	100	100

diese Anzahl groß ist im Vergleich zur Wahrscheinlichkeit des Auftretens einer Zelldifferentiation, also $200 \approx \infty$, so kann man für die Häufigkeiten der Marker eine Poisson-Verteilung annehmen. Es ergaben die Daten ind Tabelle 2.10.

Es wurde das log-lineare Modell mit Wechselwirkung TNF $\times$ IFN betrachtet, so dass

$$\lambda = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2) \tag{2.78}$$

gelten soll, wobei x_1 die TNF-Dosis, und x_2 die IFN-Dosis sein soll. Die Wechselwirkung $x_1 x_2$ soll einen synergistischen Effekt zwischen den beiden Substanzen reflektieren. Man erhält die Schätzungen

$$\hat{\beta}_0 = 3.436,\ \hat{\beta}_1 = .016,\ \hat{\beta}_2 = .009,\ \hat{\beta}_3 = -.001. \tag{2.79}$$

Der durch $x_1 x_2$ repräsentierte Effekt geht also praktisch nicht in die Vorhersage der Anzahl ein, dh es gibt keinen synergistischen Effekt, - vorausgesetzt, das angenommene Modell ist korrekt. □

2.5 Multikategoriale Daten

2.5.1 Kategorien ohne Rangordnung

Bisher wurde der Fall von zwei Kategorien betrachtet: die Dosis eines Medikaments hat gewirkt oder nicht gewirkt, ein Stimulus hat eine überschwellige Erregung erzeugt oder nicht, ein Studierender hat eine Prüfung bestanden oder nicht, etc.

Es soll jetzt der Fall betrachtet werden, dass eine Beobachtung in eine von mehr als zwei Kategorien fällt. So kann man im Krankenhaus mit einem Erreger vom Typ I, oder mit einem vom Typ II etc, oder mit keinem Erreger infizieren. Für eine spezielle Person, etwa die i-te, kann man dann einen Indikatorvektor $\vec{Y}_i$ einführen. Dessen Komponenten sind alle gleich Null, bis auf diejenige Komponente, die zu einer der Kategorien korrespondiert. Dies bedeutet, dass man die Kategorien so definiert, dass jeder Fall - hier also jede Person - in genau eine Kategorie eingeordnet werden kann. Summiert man alle diese Vektoren, so entsteht ein Vektor

$$\vec{y} = (n_1, n_2, \ldots, n_q)', \ q = k - 1; \tag{2.80}$$

Die n_j, $j = 1, 2, \ldots, q$ sind die Häufigkeiten, mit denen Fälle in die j-te Kategorie einsortiert wurden. Es werden nur $q = k - 1$ Komponenten betrachtet, da die Häufigkeit in der k-ten Kategorie festliegt, wenn insgesamt n Fälle betrachtet werden. Ziel der folgenden Betrachtungen ist, die Abhängigkeit dieser Zuordnungen von unabhängigen Variablen oder Prädiktoren $x_1, \ldots, x_p$ zu klären.

Es werde zunächst angenommen, dass die Zuordnungen unabhängig voneinander vorgenommen werden. Die Verteilung der Häufigkeiten $n_1, \ldots, n_q$ ist dann durch die Multinomialverteilung gegeben:

$$P(\vec{y} = (n_1, \ldots, n_q)') = A_{(n_1,\ldots,n_q)} \pi_1^{n_1} \cdots \pi_q^{n_q} (1 - \pi_1^{n_1} - \cdots - \pi_q^{n - n_1 - \cdots - n_q}) \tag{2.81}$$

mit

$$A_{(n_1,\ldots,n_q)} = \frac{n!}{n_1! \cdots n_q! (n - n_1 - \cdots - n_q)!} \tag{2.82}$$

Es sei $\vec{\pi} = (\pi_1, \ldots, \pi_q)'$ der Vektor der Wahrscheinlichkeiten π_r, $1 \leq r \leq q$,für die zufälligen Ereignisse $\omega \in C_r$. Die Schätzungen der π_r sind durch $\hat{\pi}_r = n_r/n$ gegeben, mit der Varianz-Kovarianz-Matrix

$$\mathrm{cov}(\bar{y}) = \frac{1}{n} \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & \cdots & -\pi_1\pi_q \\ -\pi_2\pi_1 & \pi_2(1-\pi_2) & \cdots & -\pi_2\pi_q \\ & & \vdots & \\ -\pi_q\pi_1 & -\pi_q\pi_2 & \cdots & \pi_q(1-\pi_q) \end{pmatrix} \tag{2.83}$$

Die Beobachtungen $Y_i = r$ sollen nun auf unabhängige Variablen (Prädiktoren) bezogen werden. Im einkategorialen Fall hat man die logistische Funktion

$$\pi_i = \frac{\exp(\vec{x}_i{}'\vec{\beta})}{1 + \exp(\vec{x}_i{}'\vec{\beta})} = \frac{1}{1 + \exp(-\vec{x}_i{}'\vec{\beta})}.$$

Dieser Ansatz läßt sich verallgemeinern:

$$P(Y_i = r) = \frac{\exp(\beta_{0r} + \vec{x}_i{}'\vec{\beta}_r)}{1 + \sum_{s=1}^{q} \exp(\beta_{0s} + \vec{x}_i{}'\vec{\beta}_s)} \tag{2.84}$$

Analog dazu hat man

$$P(Y_i = k) = \frac{\exp(\beta_{0k} + \vec{x}_i{}'\vec{\beta}_k)}{1 + \sum_{s=1}^{q} \exp(\beta_{0s} + \vec{x}_i{}'\vec{\beta}_s)} \tag{2.85}$$

Dann folgt

$$\frac{P(Y_i = r)}{P(Y_i = k)} = \frac{\exp(\beta_{0r} + \vec{x}_i{}'\vec{\beta}_r)}{\exp(\beta_{0k} + \vec{x}_i{}'\vec{\beta}_k)} = \exp(\beta_{0r} + \vec{x}_i{}'\vec{\beta}_r - \beta_{0k} - \vec{x}_i{}'\vec{\beta}_k),$$

und nach Logarithmierung erhält man

$$\log \frac{P(Y_i = r)}{P(Y_i = k)} = \beta_{0r} - \beta_{0k} + \vec{x}_i{}'(\vec{\beta}_r - \vec{\beta}_k) \tag{2.86}$$

Man kann diese logarithmierten Quotienten für verschiedene Werte von r bestimmen und dabei k konstant lassen; dieses Vorgehen ist von Interesse, wenn k eine Art von Nullkategorie repräsentiert. Im Beispiel zu den Infektionen bei Kaiserschnittgeburten wird man k für die Kategorie "keine Infektion" wählen und r bestimmte Arten von Infektionen, etwa vom Typ I und vom Typ II, repräsentieren lassen. $Y_1 = r_1$ heißt dann, dass es r_1 Fälle mit Infektionen vom Typ I gibt, und $Y_2 = r_2$ zeigt an, dass es r_2 Fälle mit Infektionen vom Typ II gibt. $Y_3 = n - r_1 - r_2$ ist die Anzahl der Fälle ohne eine Infektion. Dann kann man zu den folgenden Definitionen übergehen:

$$\beta_{r0} = \beta_{0r} - \beta_{0k} \tag{2.87}$$

$$\vec{\beta}_r = \vec{\beta}_r - \vec{\beta}_k; \tag{2.88}$$

man bemerke, dass (2.88) lediglich eine Umbenennung ist. (2.86) kann in der Form

$$\log \frac{P(Y_i = r)}{P(Y_i = k)} = \beta_{r0} + \vec{x}_i{}'\vec{\beta}_r \tag{2.89}$$

geschrieben werden. β_{r0} und $\vec{\beta}_r$ sind jetzt Parameter *relativ* zur Kategorie k.

Beispiel 20 (Fortsetzung des Beispiels 17) Es werden nun zwei mögliche Infektionen betrachtet, Infektion I und Infektion II. Die Tabelle 2.11 zeigt die Daten. Man

Tabelle 2.11: $x_1 = 1$ KS nicht geplant, $x_2 = 1$ RF, $x_3 = 1$ AB verabr.

Gruppe	n_i	y_{i1}	y_{i2}	x_1	x_2	x_3
7	$n_7 = 98$	4	7	1	1	1
4	$n_4 = 18$	0	1	0	1	1
8	$n_8 = 0$	0	0	1	0	1
6	$n_6 = 36$	10	13	1	1	0
5	$n_5 = 9$	0	0	1	0	0
2	$n_2 = 58$	11	17	0	1	0
3	$n_3 = 2$	0	0	0	0	1
1	$n_1 = 40$	4	4	0	0	0

kann nun die Risiken für die Infektionen I und II berechnen:

$$\log \frac{p(\text{Infektion I})}{p(\text{keine Infektion})} = \beta_{10} + \vec{x}_i{}'\vec{\beta}_1 \tag{2.90}$$

$$\log \frac{p(\text{Infektion II})}{p(\text{keine Infektion})} = \beta_{20} + \vec{x}_i{}'\vec{\beta}_2, \tag{2.91}$$

wobei $\vec{x}_i = (x_{i1}, \ldots, x_{ip})'$ und $\vec{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp})'$, $j = 1, 2$. Die Schätzungen der Parameter, dh der Komponenten der Vektoren $\vec{\beta}_1$ und $\vec{\beta}_2$, sind Man kann jetzt zB

Tabelle 2.12: Parameterschätzungen

$\hat{\beta}_{10}$	-2.621	$\hat{\beta}_{20}$	-2.560
$\hat{\beta}_{11}$	1.174	$\hat{\beta}_{21}$	.996
$\hat{\beta}_{12}$	-3.520	$\hat{\beta}_{22}$	-3.087
$\hat{\beta}_{13}$	1.829	$\hat{\beta}_{23}$	2.195

nach dem Effekt einer prophylaktischen Antibiotikagabe bei vorhandenen Risikofaktoren relativ zu nicht vorhandenen Risikofaktoren fragen. Für Infektionen vom Typ I haben prophylaktische Antibiotikagaben das "Gewicht" $\hat{\beta}_{12} = -3.520$, für Infektionen vom Typ II hat man $\hat{\beta}_{22} = -3.087$. Antibiotika reduzieren das Risiko einer Typ-I-Infektion mehr als das einer Typ-II-Infektion. Sind Risikofaktoren vorhanden, ist das relative Risiko für eine Typ-I-Infektion durch $\exp(-2.621+1.829) = .45$ mal das relative Risiko für die gleiche Art von Infektion wenn kein Risiko vorhanden ist. □

2.5.2 Kategorien mit Rangordnung: ordinale Regression

In diesem Abschnitt wird der Fall betrachtet, dass zwischen den Kategorien eine Rangordnung existiert. Die abhängige Variable Y bildet ein Merkmal ab, dass zwar kontinuierlich variiert, aber nicht auf einer Intervallskala, sondern nur auf einer Ordinalskala abgebildet wird.

Es sei ξ (sprich: ksi) die Variable, die mit Y erfasst wird, selbst aber nicht beobachtbar ist. $\vec{x}$ sei ein Vektor, dessen Komponenten Variablen repräsentieren, die einen Einfluß auf ξ haben bzw. haben könnten. Im "normalen" linearen Modell gelte die Beziehung

$$\xi = -\vec{\beta}'\vec{x} + \varepsilon, \tag{2.92}$$

wobei ϵ Fehler und den Effekt nichtkontrollierter Variablen repräsentiert. Das Minuszeichen vor $\vec{\beta}'\vec{x}$ dient nur der in (2.94) gebrauchten Schreibweise $F(\vec{x}{}'\vec{\beta} + \theta_r)$, es soll nicht bedeuten, dass $\vec{\beta}'\vec{x}$ nur negativ sein darf: wenn $\vec{\beta}'\vec{x}$ negativ ist, ist $-\vec{\beta}'\vec{x}$ eben positiv, und umgekehrt. Die zufällige Veränderliche ε habe die Wahrscheinlichkeitsverteilng F. Eine Beobachtung fällt in die r-te Kategorie, wenn ξ im

Intervall $[\theta_{r-1}, \theta_r)$ liegt:

$$Y = r \text{ genau dann, wenn } \theta_{r-1} \leq \xi < \theta_r,\ 1 \leq r \leq q \tag{2.93}$$

mit $-\infty = \theta_0 < \theta_1 < \cdots < \theta_R = \infty$. Die θ_j, $j = 1, \ldots, R$ heißen auch "Schwellen". Es gilt nun

$$\begin{aligned} P(Y \leq r) &= P(\xi \leq \theta_r) \\ &= P(-\vec{x}'\vec{\beta} + \varepsilon \leq \theta_r) \\ &= P(\varepsilon \leq \vec{x}'\vec{\beta} + \theta_r) \\ &= F(\vec{x}'\vec{\beta} + \theta_r) \end{aligned} \tag{2.94}$$

Dieses Modell heißt auch *kumulatives Schwellenmodell*[9], denn F ist ja eine Verteilungsfunktion, dh die kumulierte Wahrscheinlichkeitsverteilung von ξ. Es folgt

$$P(Y = 1|\vec{x}) = F(\theta_1 + \vec{\beta}'\vec{x}) \tag{2.95}$$

$$P(Y = r|\vec{x}) = F(\theta_r + +\vec{\beta}'\vec{x}) - F(\theta_{r-1} + \vec{\beta}'\vec{x}). \tag{2.96}$$

Für spezielle Wahlen für F entstehen spezielle Schwellenmodelle.

Das Modell der proportionalen kumulativen Chancen

Es seien $\vec{x}_i$ und $\vec{x}_j$ zwei verschiedene Realisierungen von $\vec{x}$, sie repräsentieren verschiedene Ausprägungen der unabhängigen Variablen. Gegeben seien die Quotienten

$$Q_i(r) = \frac{P(Y \leq r|\vec{x}_i)}{P(Y > r|\vec{x}_i)}, \quad Q_j(r) = \frac{P(Y \leq r|\vec{x}_j)}{P(Y > r|\vec{x}_j)}, \tag{2.97}$$

wobei die Ausdrücke $Q_i(r)$ und $Q_j(r)$ einfach zur Abkürzung eingeführt wurden. $Q_i(r)$ und $Q_j(r)$ heißen *kumulierte Chancen*.

Das logistische Schwellenmodell

Es sei F durch die logistische Funktion gegeben, also

$$P(Y \leq r|\vec{x}) = \frac{\exp(\theta_r + \vec{\beta}'\vec{x})}{1 + \exp(\theta_r + \vec{\beta}'\vec{x})}. \tag{2.98}$$

Man kann dann zu den Logits $\log p/(1-p)$ übergehen, wenn man $p = P(Y \leq r|\vec{x})$, $1 - p = 1 - P(Y \leq r|\vec{x}) = P(Y > r|\vec{x})$ setzt. Man erhält dann

$$\log \frac{P(Y \leq r|\vec{x})}{P(Y > r|\vec{x})} = \theta_r + \vec{\beta}'\vec{x}, \tag{2.99}$$

bzw.

$$\frac{P(Y \leq r|\vec{x})}{P(Y > r|\vec{x})} = \exp(\theta_r + \vec{\beta}'\vec{x}). \tag{2.100}$$

[9] McCullagh, 1980

Das logistische Modell der kumulativen Chancen

Für den Quotienten $Q_i(r)/Q_j(r)$ erhält man dann

$$\frac{Q_i(r)}{Q_j(r)} = \exp(\theta_r + \vec{\beta}'\vec{x}_i - \theta_r - \vec{\beta}'\vec{x}_j),$$

dh

$$\frac{Q_i(r)}{Q_j(r)} = \exp[\vec{\beta}'(\vec{x}_i - \vec{x}_j)]. \tag{2.101}$$

Obwohl, wie aus (2.97) ersichtlich, $Q_i(r)$ und $Q_j(r)$ von der Kategorie r abhängen, ist der Quotient $Q_i(r)/Q_j(r)$ von r unabhängig und hängt nur von der Differenz $\vec{x}_i - \vec{x}_j$ der Vektoren $\vec{x}_i$ und $\vec{x}_j$ ab. Es gilt

$$\log \frac{Q_i(r)}{Q_j(r)} = \vec{\beta}'(\vec{x}_i - \vec{x}_j), \tag{2.102}$$

der Logarithmus des Verhältnisses der kumulierten Chancen ist also, unabhängig von der Kategorie r, proportional zur Differenz $\vec{x}_i - \vec{x}_j$, so dass man vom *Modell der proportionalen kumulativen Chancen* spricht.

Anmerkung: (2.102) gilt nur dann, wenn die Verteilungsfunktion F durch die logistische Funktion gegeben ist. □

Das kumulierte Extremwertmodell

Eine alternative Annahme über die Verteilung F ist die *Extremwertverteilung*[10] Für $X = \max(X_1, \ldots, X_n)$, die X_j, $1 \leq j \leq n$ identisch, unabhängig und z. B. normalverteilt kann die Verteilung von X bei hinreichend großen Wert von n durch

$$F(x) = 1 - \exp(-e^{(ax+b)}) \tag{2.103}$$

approximíert werden; a und b sind freie Parameter. Für die (approximative) Verteilung $G(y)$ des Minimumns $Y = \min(X_1, \ldots, X_n)$ läßt sich zeigen, dass

$$G(y) = 1 - F(-x) = \exp(-e^{-(ay+b})) \tag{2.104}$$

gilt. Daraus folgt

$$P(Y \leq r|\vec{x}) = 1 - \exp(-\exp(\theta_r + \vec{\beta}'\vec{x})). \tag{2.105}$$

[10] Der Ausdruck Extremwertverteilung ergibt sich aus der Tatsache, dass diese Verteilung die Verteilung des Maximums X_+ bzw. des Minimums X_- von n unabhängigen, identisch verteilten zufälligen Veränderlichen $x_1, \ldots, x_n$ für hinreichend großen Wert von n ist, vorausgesetzt, die $x_1, \ldots, x_n$ haben einen bestimmten Verteilungstyp, zB die Normalverteilung, vergl Abschnitt 2.7. Außer der hier betrachteten Verteilung gibt es noch zwei weitere mögliche Verteilungen für $X = \max(X_1, \ldots, X_n)$, $n \to \infty$.

Natürlich gilt $P(Y \leq r|\vec{x}) = 1 - P(Y > r|\vec{x})$, so dass

$$P(Y > r|\vec{x}) = \exp(-\exp(\theta_r + \vec{\beta}'\vec{x})), \tag{2.106}$$

und

$$\log(-\log P(Y > r|\vec{x})) = \theta_r + \vec{\beta}'\vec{x}. \tag{2.107}$$

Während in (2.99) $\theta_r + \vec{\beta}'\vec{x}$ auf den Logarithmus der "Wett"chance bezogen ist, bezieht sich dieser Ausdruck hier auf den Logarithmus von $-\log P(Y > r|\vec{x})$.

Das proportionale Hazardmodell

Man kann die Wahrscheinlichkeit betrachten, dass $Y \geq r$ beobachtet wird; für gegebenen Vektor $\vec{x}$ der unabhängigen Variablen ist sie durch $P(Y \geq r|\vec{x})$ gegeben. Man kann weiter nach der Wahrscheinlichkeit fragen, dass $Y = r$ gilt unter der Bedingung, dass schon bekannt ist, dass $Y \geq r$ ist. Diese Wahrscheinlichkeit ist durch die bedingte Wahrscheinlichkeit $P(Y = r|Y \geq r, \vec{x})$ gegeben. Bedingte Wahrscheinlichkeiten dieser Art werden insbesondere bei der Diskussion von Wartezeiten betrachtet: man möchte zB wissen, wann bei einer an multipler Sklerose erkrankten Person der nächste Krankheitsschub kommt. Die Zeit bis zum jeweils nächsten Schub ist nicht deterministisch festgelegt, sondern folgt einer bestimmten Wahrscheinlichkeitsverteilung, und die Evaluation von Therapien für diese Krankheit kann u. U. in der Überprüfung der möglichen Veränderung der Parameter dieser Verteilung bestehen. Die auch als *Hazard-Funktion* bekannte bedingte Wahrscheinlichkeit $P(Y = r|Y \geq r, \vec{x})$ spielt bei dieser Überprüfung eine zentrale Rolle. Diese Hazard-Funktion kann auch im Falle ordinal skalierter Kategorien bestimmt werden. Es gilt allgemein

$$P(Y = r|Y \geq r, \vec{x}) = \frac{P(Y = r \cap Y \geq r, \vec{x})}{P(Y \geq r, \vec{x})} \tag{2.108}$$

Nun fällt das Ereignis $\{Y = r \cap Y \geq r\}$ mit dem Ereignis $Y = r$ zusammen, so dass man

$$P(Y = r|Y \geq r, \vec{x}) = \frac{P(Y = r, \vec{x})}{P(Y \geq r, \vec{x})} \tag{2.109}$$

erhält. Jetzt ist sicherlich

$$P(Y = r, \vec{x}) = P(Y \leq r|\vec{x}) - P(Y \leq r - 1|\vec{x}) \tag{2.110}$$

Berücksichtigt man (2.105), so erhält man

$$\begin{aligned} P(Y = r|\vec{x}) &= 1 - \exp(-\exp(\theta_r + \vec{\beta}'\vec{x})) - 1 + \exp(-\exp(\theta_{r-1} + \vec{\beta}'\vec{x})) \\ &= \exp(-\exp(\theta_{r-1} + \vec{\beta}'\vec{x})) - \exp(-\exp(\theta_r + \vec{\beta}'\vec{x})). \end{aligned}$$

Nun ist $P(Y \geq r|\vec{x}) = P(Y > r - 1|\vec{x}) = \exp(-\exp(\theta_{r-1} + \vec{\beta}\,\vec{x}))$, so dass

$$\frac{P(Y = r|\vec{x})}{P(Y \geq r|\vec{x})} = \frac{\exp(-e^{\theta_{r-1}+\vec{\beta}'\vec{x}}) - \exp(-e^{\theta_r+\vec{\beta}'\vec{x}})}{\exp(-\exp(\theta_{r-1} + \vec{\beta}'\vec{x}))}$$

$$
\begin{aligned}
&= \left(\exp(-e^{\theta_{r-1}+\vec{\beta}'\vec{x}} - \exp(-e^{\theta_r+\vec{\beta}'\vec{x}})\right) \exp(\exp(\theta_{r-1}+\vec{\beta}'\vec{x})) \\
&= 1 - \exp((e^{\theta_{r-1}} - e^{\theta_r})e^{\vec{\beta}'\vec{x}})
\end{aligned}
$$

Man kann nun einen neuen Parameter $\tilde{\theta}_r = \log(e^{\theta_{r-1}} - e^{\theta_r})$ einführen; man spricht von einer *Reparametrisierung*. Dann ist jedenfalls

$$e^{\tilde{\theta}_r} = e^{\theta_{r-1}} - e^{\theta_r}$$

und man erhält

$$P(Y = r|Y \geq r, \vec{x}) = 1 - \exp(-\exp(\tilde{\theta}_r + \vec{\beta}'\vec{x})). \tag{2.111}$$

Man vergleiche diesen Ausdruck mit der Definition von F in (2.105). Die rechte Seite von (2.111) hat genau diese Form, nur dass $\tilde{\theta}_r$ statt θ_r auftaucht. Auf der linken Seite steht allerdings $P(Y = r|Y \geq r, \vec{x})$ statt $P(Y \leq r|\vec{x})$. Dies bedeutet, dass man $P(Y \leq r|\vec{x})$, wenn diese Wahrscheinlichkeit wie in (2.105) definiert und dem Parameter θ_r die Deutung von $\tilde{\theta}_r$ gegeben wird, wie in (2.111) als Hazard-Funktion betrachtet werden kann. Der Term $\exp(\tilde{\theta}_r + \vec{\beta}'\vec{x})$ in (2.111) kann in der Form

$$\exp(\tilde{\theta}_r + \vec{\beta}'\vec{x}) = \exp(\tilde{\theta}_r)\exp(\vec{\beta}'\vec{x})$$

geschrieben werden. Der Effekt der unabhängigen Variablen, $\exp(\vec{\beta}'\vec{x})$, ist also proportional zum "Effekt" der reparametrisierten Kategoriengrenze $\tilde{\theta}_r$, weswegen man von (2.111) als einem *proportionalen Hazard-Modell* oder vom *proportionalen Cox-Modell* spricht, nach Cox, ein analoges Modell für zeitliche Größen eingeführt hat.

Das sequentielle Modell

Gelegentlich kann eine abhängige Variable Y nur eine bestimmte Ausprägung haben, wenn sie vorher die geringeren Ausprägungen durchlaufen hat. Beispiele hierfür sind Lernvorgänge, aber auch Krankheitsverläufe. So kann eine Leber nur dann als "vergrößert" klassifiziert werden, wenn sie vorher "normal" war. Ein anderes Beispiel sind Mandeldrüsen bei Kindern, die mit *streptococcus pyogens* befallen sind:

Beispiel 21 Holms und Williams (1954)[11] berichteten die folgenden Daten: Es kann angenommen werden, dass der Prozess immer mit "Befallen, aber nicht vergrößert" beginnt, den Zustand "vergrößert" durchläuft und bei "stark vergrößert" endet. □

Man kann weiter annehmen, dass die Daten aufgrund eines nicht direkt beobachtbaren Prozesses erzeugt werden. Dieser kann durch eine"latente" zufällige Veränderliche U_r repräsentiert werden, wobei

$$U_r = -\vec{x}'\vec{\beta} + \varepsilon_r, \quad r = 1, \ldots, k-1. \tag{2.112}$$

[11] Holms, M.C., Williams, R.E.O. (1954) The distribution of carriers of *streptococcus pyogenes* among 2413 healthy children. J. Hyg. (Cambridge), 52, 165 - 179

Tabelle 2.13: Tonsillengröße und Streptokokkenbefall bei Kindern

	Befallen, aber nicht vergrößert	vergrößert	stark vergrößert
Träger	19	29	24
Nichtträger	497	560	269

ε_r ist eine zufällige Veränderliche, die unkontrollierte Effekte repräsentiert, mit der Verteilungsfunktion F. Die abhängige Variable Y wird durch den Wert von U_r bestimmt; es soll

$$Y = 1 \text{ genau dann, wenn } \quad U_1 \leq \theta_1. \tag{2.113}$$

θ_1 ist ein *Schwellenparameter*, der die obere Grenze der ersten Kategorie definiert. Der Prozess endet, wenn $U_1 \leq \theta_1$. Weiter hat man

$$Y = 2 \text{ genau dann, wenn } \quad \theta_1 < U_2 \leq \theta_2. \tag{2.114}$$

Allgemein gilt

$$Y = r \text{ genau dann, wenn } \quad \theta_{r-1} < U_r \leq \theta_r, \tag{2.115}$$

bzw.

$$Y = k \text{ genau dann, wenn } \quad U_r > \theta_{k-1}. \tag{2.116}$$

Es muß nun die Wahrscheinlichkeit für das Ereignis $\{Y = r\}$ gegeben $\vec{x}'\vec{\beta}$ bestimmt werden. Dies ist die Wahrscheinlichkeit, dass $U_r \leq \theta_r$ und $U_r > \theta_i$ für $i = 1, \ldots, r-1$. Demnach erhält man

$$P(Y = r|\vec{x}) = F(\theta_r + \vec{x}'\vec{\beta}) \prod_{i=1}^{r-1} (1 - F(\theta_i + \vec{x}'\vec{\beta})). \tag{2.117}$$

Das Modell nimmt eine spezifische Form an, wenn man eine bestimmte Wahl für die Verteilungsfunktion F vornimmt. Eine mögliche Wahl ist wieder die

Es werde die logistische Funktion $F(x) = 1/(1 - \exp(-x))$ angenommen. Dann erhält man insbesondere

$$P(Y = r|Y \geq r, \vec{x}) = \frac{\exp(\theta_r + \vec{x}'\vec{\beta})}{1 + \exp(\theta + \vec{x}'\vec{\beta})}. \tag{2.118}$$

Nun ist

$$P(Y > r|Y \geq r, \vec{x}) = 1 - F,$$

so dass

$$\frac{P(Y = r|\vec{x})}{P(Y > r|\vec{x}} = \frac{\exp(\theta_r + \vec{x}'\vec{\beta})}{1 + \exp(\theta + \vec{x}'\vec{\beta})} \left(\frac{1}{1 - \frac{\exp(\theta_r + \vec{x}'\vec{\beta})}{1 + \exp(\theta_r + \vec{x}'\vec{\beta})}} \right),$$

woraus

$$\frac{P(Y=r|\vec{x})}{P(Y>r|\vec{x})} = \exp(\theta_r + \vec{x}'\vec{\beta}), \tag{2.119}$$

und schließlich

$$\log\left(\frac{P(Y=r|\vec{x})}{P(Y>r|\vec{x})}\right) = \theta_r + \vec{x}'\vec{\beta} \tag{2.120}$$

folgt. Setzt man für F die Extremwertfunktion $G_{max}(x)$ ein und führt die entsprechenden Rechnungen durch, so erhält man

$$\log\left[-\log\left(\frac{P(Y=r|\vec{x})}{P(Y>r|\vec{x})}\right)\right] = \theta_r + \vec{x}'\vec{\beta}. \tag{2.121}$$

Der Vergleich von (2.120) und (2.121) zeigt, dass die Interpretation des Ausdrucks $\theta_r + \vec{x}'\vec{\beta}$ von der Wahl der Verteilungsfunktion F abhängt. Der Vergleich mit (2.111) wiederum zeigt, dass das sequentielle Modell zusammen mit der Wahl der Extremwertfunktion für F äquivalent mit dem kumulierten Extremwertmodell ist.

Beispiel 22 (Atemtests) Die folgenden, in Tabelle 2.14 angegebenen Daten sind einer Untersuchung von Forthofer und Lehnen (1981)[12] über die Lungenfunktion von Industriearbeitern in Texas entnommen. Die Tabelle 2.15 enthält die Parame-

Tabelle 2.14: Resultate der Atemtests bei Industriearbeiten in Houston

		Resultate		
Alter	Raucherstatus	normal	grenzwertig	abnormal
< 40	niemals geraucht	577	27	7
	ehem. Raucher	192	20	3
	gegenw. Raucher	682	46	11
40 - 59	niemals geraucht	164	4	0
	ehem. Raucher	145	15	7
	gegenw. Raucher	245	47	27

terwerte für drei Varianten des kumulativen Modells: (i) das kumulative logistische Modell, (ii) proportionale Hazardmodell, (iii) das Extremwertmodell. Die Werte für die zu schätzenden Parameter unterscheiden sich, weil ja unterschiedliche Modelle zugrundegelegt werden. Die Variablen sind effektkodiert, mit -1 für die letzte Kategorie. Positive Werte der Parameter indizieren eine Verschiebung auf der latenten Skala nach links und erzeugen damit höhere Werte für die Wahrscheinlichkeiten für die Kategorien 1 und 2 (normal und borderline). Wie zu erwarten ergeben sich bessere Werte für das Atmen wenn eine Person jung ist und wenig oder gar nicht raucht. Dem Deviance-Wert entsprechend passt das proportionale Hazardmodell am besten und das Extremwertmodell am schlechtesten. Nach Tabelle 2.15 hat

[12] Forthofer, R.N., Lehnen, R.G. (1981) *Public Program Analysis. A new Categorical Data Approach.* Belmont Calif., Lifetime Learning Publications.

Tabelle 2.15: Parameterschätzungen für verschiedene Modelle

	Kumul. logistisch		prop. Hazard		Extr'modell	
θ_1	2.370	.000	.872	.000	2.429	.000
θ_2	3.844	.000	1.377	.000	3.843	.000
Alter (1)	.114	.290	.068	.040	.095	.370
Rauchen (1)	.905	.000	.318	.000	.866	.190
Rauchen (2)	-.364	.010	-.110	.020	-.359	.140
Alter (1) × Rauchen (1)	-.577	.000	-.211	.000	-.529	.190
Alter (1) × Rauchen (2)	.015	.910	.004	.920	.021	.140
Deviance	8.146		3.127		9.514	

Tabelle 2.16: Proportionales Hazardmodell: Wechselwirkung zwischen Alter und Rauchen

	Raucherstatus		
	niemals Rauchen (1)	ehemals Rauchen (2)	gegenwärtig Rauchen (3)
Alter (1)	-.211	.004	.207
Alter (2)	.211	-.004	-.207

Rauchen (1) einen starken Einfluß (Haupteffekt:.318) auf das Atmen. Betrachtet man aber noch die Wechselkwirkung mit dem Alter (vergl. Tabelle 2.16), so sieht man, dass der Einfluß des Rauchens auf die Atmungsfunktion sich besonders stark im Alter auswirkt: Die WW Alter (2) × Rauchen (3) = -.207 muß zu den negativen Effekten des Alters (2) = -.114 und des Rauchens Rauchen (3) = -.541 addiert werden, um den Gesamteffekt zu erhalten. Das logistische Modell führt zu den gleichen Schlußfolgerungen, obwohl es weniger gut passt als das proportionale Hazardmodell. □

Beispiel 23 Berufliche Erwartungen An der Universität Regensburg wurden Studierende der Psychologie dar nach befragt, ob sie nach Beendigung des Studiums eine ihren Qualifikationen entsprechende Stelle zu finden erwarteten. Es gab drei Antwortkategorien: 1 = ich erwarte keine adäquate Stelle, 2 = ich bin nicht sicher, eine solche Stelle zu finden, 3 = ich erwarte eine solche Stelle unmittelbar nach Abschluß des Studiums. Die Tabelle 2.17 enthält die Daten. Der die Erwartung beeinflussende Faktor ist hier das Alter. Es sind zwei Modelle betrachtet worden: (i) das einfache kumulative logistische Modell, und (ii) die erweiterte Version dieses Modells:

$$P(Y \leq r|\text{Alter}) = F(\theta_r + \gamma \log \text{Alter}), \quad \text{(i)} \tag{2.122}$$

$$= F(\theta_r + \beta_r \log \text{Alter}), \quad \text{(ii)} \tag{2.123}$$

Tabelle 2.17: Gruppierte Daten zur beruflichen Erwartung; Kategorien = Antwortkategorien

Alter	Kategorien			Anzahl
(Jahre)	1	2	3	n_i
19	1	2	0	3
20	5	18	2	25
21	6	19	2	27
22	1	6	3	10
23	2	7	3	12
24	1	7	5	13
25	0	0	3	3
26	0	1	0	1
27	0	2	1	3
29	1	0	0	1
30	0	0	2	2
31	0	1	0	1
34	0	1	0	1

Die Parameterschätzungen und der Fit der Modelle wird in der Tabelle 2.18 wiedergegeben.

Tabelle 2.18: Kumulatives Modell für berufliche Erwartungen

	log Alter als globale Shift Variable		log Alter als Schwellenvariable	
Parameter	Param'wert	p-Wert	Param'wert	p-Wert
Schwelle 1	14.987	(.010)	9.467	(.304)
Schwelle 2	18.149	(.002)	20.385	(.002)
Steigung γ	-5.402	(.004)	-	-
Kat'spezif. Steig. β_1	-	-	-3.597	(.230)
Kat'spezif. Steig. β_2	-	-	-6.113	(.004)
Pearson's χ^2	42.696	(.007)	33.503	(.055)
Deviance	26.733	(.267)	26.063	(.248)

Das Modell (2.122) korrespondiert zu dem Modell, in dem log Modell als "Shift Variable" aufgeführt wird. Das Pearson–χ^2 ist auf dem .05-Niveau signifikant und passt auf jeden Fall schlechter als das Modell (2.123). Die Abweichung zwischen χ^2 und dem Deviance-Maß verweisen auf eine Verletzung der Annahmen bezüglich der Asymptotik für das χ^2. Die negativen Werte für $\hat{\beta}_r$ und die Steigung $\hat{\gamma}$ zeigen, dass ein größeres Alter eine geringere Wahrscheinlichkeit für niedrigere Kategorien

impliziert, d.h. es wird eher mit einer Stelle nach Abschluß des Studiums gerechnet: je näher die Studierenden an das Abschlußexamen komme, desto optimistischer werden sie. Die Effekte auf die kumulativen Odds

$$\frac{P(Y \leq r|x_1)/P(Y > r|x_1)}{P(Y \leq r|x_2)/P(Y > r|x_2)} = \exp(\beta_{0r} + \beta_r(x_1 - x_2)) \tag{2.124}$$

hängen aber von der Kategorie r ab. Man betrachte Altersgruppen $x_1 > x_2$. Die kumulativen Odds für $r = 1$ messen die Tendenz für eine ausgeprägte negative Erwartung (1: ich erwarte keine adäquate Anstellung). Für $r = 2$ messen die kumulativen Odds die Erwartung für 1 oder 2, also die Erwartung keiner Anstellung, zumindest Ungewißheit bezüglich der Anstellung, und zwar im Vergleich zu 3: Ich erwarte eine adäquate Anstellung. Es ist $\hat{\beta}_2 < \hat{\beta}_1$; dies bedeutet, dass der Effekt des Alters größer ist als die beiden anderen. Der p-Wert für β_1 ist .230. Dieser Wert zeigt, dass die negative Erwartung (Kategorie 1 versus Kategorien 2 und 3) möglicherweise durch das Alter beeinflußt wird. Der große χ^2-Wert legt aber nahe, dass das Modell den Daten nicht gerecht wird. Die Ursache liegt daran, dass beim kumulativen Modell angenommen wird, dass die Erwartung, eine Anstellung zu finden, mit dem Alter steigt. Es gibt aber eine(n) 29-Jährige(n), die entgegen dieser Annahme keine Anstellung erwartet. Dieser "Fall" repräsentiert einen Ausreisser, der den Fit des Modells reduziert. □

2.6 Der Test von Hypothesen

Es sei $\vec{b} = (b_1, \ldots, b_k)'$ der k-dimensionale Parametervektor, für den die Maximum-Likelihood-Schätzung $\hat{\vec{b}} = (\hat{b}_1, \ldots, \hat{b}_k)'$ vorliege. Es sollen Hypothesen über die Werte der Komponenten von $\vec{b}$ getestet werden. Die einfachsten Hypothesen sind von der Art $b_1 = b_3 = b_k = 0$, dh dass bestimmte Komponenten von $\vec{b}$ gleich Null sind. Man kann aber auch prüfen, ob bestimmte Komponenten gleich groß sind und einen von Nu ll verschiedenen Wert haben. Alle Hypothesen lassen sich in bezug auf einen Vektor $\vec{\xi}$ formulieren, der über eine geeignet definierte Matrix C gemäß

$$C\vec{b} = \vec{\xi} \tag{2.125}$$

definiert wird. Ein einfaches Beispiel erhält man, wenn man für C die k-dimensionale Einheitsmatrix wählt; dann hat man

$$C\vec{b} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ & & \vdots & \\ 0 & 0 & \ldots & 1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix} = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_k \end{pmatrix},$$

dh man testet, ob die b_j, $j = 1, \ldots, k$ gleich bestimmten Werte $\xi_1, \ldots, \xi_k$ sind. Im einfachsten Fall ist $\vec{\xi}$ der Nullvektor; man tested dann, ob die b_j signifikant von Null verschieden sind. Enthält C nur $s < k$ Zeilen, so wird dieser Test auf bestimmte, durch die Zeilen von C ausgewählte Komponenten beschränkt. Man kann auch Differenzen von Komponenten von $\vec{b}$ daraufhin überprüfen, ob sie bestimmte Werte annehmen. So sei zB $\vec{b} = (b_1, b_2, b_3)'$ und

$$C = \begin{pmatrix} 0 & -1 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

Dann ist

$$C\vec{b} = \begin{pmatrix} b_1 - b_2 \\ 0 \end{pmatrix} = \vec{\xi} = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}.$$

Für ξ_1 und ξ_2 können nun bestimmte Werte vorgegeben werden und man kann prüfen, ob die Schätzungen $\hat{b}_j$ diesen Werten entsprechen.

Die Hypothesen haben dann die allgemeine Form

$$H_0 : \; C\vec{b} = \vec{\xi} \text{ versus } H_1 : \; C\vec{b} \neq \vec{\xi}. \tag{2.126}$$

In Abschnitt 1.3 wurden die Likelihood-Quotienten eingeführt; in (1.71), Seite 27, wird insbesondere die log-Likelihood-Statistik G^2 definiert. Damit lassen sich nahezu beliebige Hypothesen über $\vec{b}$ testen.

Wald-Test Eine Approximation an G^2 ist der *Wald-Test*. Die Wald-Statistik ist durch

$$w = (C\hat{\vec{b}} - \vec{\xi})'(CI_r^{-1}(\hat{\vec{b}})C')^{-1}(C\hat{\vec{b}} - \vec{\xi}) \tag{2.127}$$

definiert.

Score-Statistik In (1.20), Seite 14, wurde die Score-Funktion definiert. Ein weiterer Test läßt sich über diese Funktion erreichen. Dazu definiert man die *Score-Statistik*

$$u = \vec{s}'(\tilde{\vec{b}}) I^{-1}(\tilde{\vec{b}}) \vec{s}(\tilde{\vec{b}}). \tag{2.128}$$

Ist $\hat{\vec{b}}$ die ML-Schätzung von $\vec{b}$, so ist $\vec{s}(\hat{\vec{b}}) = \vec{0}$ (s. Seite 14). Wird $\hat{\vec{b}}$ durch den unter H_0 spezifizierten Vektor $\tilde{\vec{b}}$ ersetzt, so weicht $\vec{s}(\tilde{\vec{b}})$ signifikant von Null ab, wenn H_0 nicht wahr ist. Die Distanz zwischen $\vec{s}(\tilde{\vec{b}})$ und $\vec{0}$ wird durch die Statistik u in (2.128) erklärt.

Gilt in (2.126) insbesondere $\vec{\xi} = \vec{0}$, so vereinfacht sich die Statistik zu

$$w = \hat{\vec{b}}' I^{-1} \hat{\vec{b}}. \tag{2.129}$$

Will man nur eine Komponente b_j testen, so erhält man den t-Wert

$$t = \frac{\hat{b}_j}{\sqrt{\mathrm{var}(\hat{b}_j)}}, \tag{2.130}$$

wobei $var(\hat{b}_j)$ durch das j-te Element der geschätzten Varianz-Kovarianz-Matrix für die Schätzungen der $b_1, \ldots, b_k$, I^{-1}, gegeben ist.

Unter H_0 sind die Teststatistiken asymptotisch äquivalent mit einer durch die χ^2-Verteilung gegebenen Grenzverteilung, dh

$$u, w \overset{a}{\sim} \chi^2(s) \tag{2.131}$$

mit s Freiheitsgraden. Dabei ist $s \leq k$ die Anzahl der getesteten Parameter.

2.7 Anhang: die Verteilung extremer Werte

Es soll der Begriff der Extremwertverteilung erklärt werden. Es werden die Werte der zufälligen Variablen $X_1, \ldots, X_n$ beobachtet. Einer dieser Werte wird der Größte sein; er wird mit $X = \max(X_1, \ldots, X_n)$ bezeichnet. Da die X_j zufällige Variablen sind, muß auch X eine zufällige Variable sein. Gesucht ist die Verteilung von X. Die Nützlichkeit solcher Verteilungen sieht man am Beispiel von Ebbe und Flut, einem Naturphänomen, das Küstenbewohnern gut bekannt ist. Ist X_j der Wasserstand bei der j-ten Flut, und hat man n-mal nacheinander den Flutwasserstand gemessen, so findet man, dass sich die Fluthöhen zufällig voneinander unterscheiden, und die Verteilung des Maximums X ist von Interesse, weil man dann die Höhe des Deiches berechnen kann, die sicherstellt, dass man in den nächsten 10 Jahren mit z.B. 95%-iger Sicherheit keine Überflutung erlebt. Analoge Betrachtungen kann man für das Minimum der X_j anstellen: jedes Schiff braucht mindestens eine Handbreit Wasser unter dem Kiel.

Der Einfachheit halber sei angenommen, dass die X_j stochastisch unabhängig und identisch verteilt sind; man schreibt kurz i.i.d. für diese Eigenschaft (i.i.d. steht für "independently and identically distributed"). Zieht man Stichproben vom Umfang n aus einer bestimmten Population, so hat man es mit i.i.d.-Werten zu tun. Maxima und Minima von Stichprobenwerten werden zusammenfassend Extremwerte genannt. Man findet ohne Mühe viele Beispiele von Extremwerten in der Psychologie: wie ist die maximale Reaktionszeit verteilt? (wichtig für den Abstand beim Autofahren), etc. Will man die Verteilung von X bestimmen und kann man von der Unabhängigkeit der X_j ausgehen, so betrachtet man die Wahrscheinlichkeit $P(X \leq x$, also die Wahrscheinlichkeit, dass X kleiner oder gleich x ist. Das ist dann der Fall, wenn $X_1 \leq x$ und $X_2 \leq x$ und schließlich $X_n \leq x$ ist, also

$$P(X \leq x) = P(X_1 \leq x \cap \ldots \cap X_n \leq x).$$

Wegen der Unabhängigkeit ist aber

$$P(X_1 \leq x \cap \ldots \cap X_n \leq x) = \prod_{j=1}^{n} P(X_j \leq x) = F^n(x).$$

Dabei ist $F(x) = P(X_j \leq x)$ für alle j. Man kann nun zeigen, dass für große Werte von n der Ausdruck

$$\prod_{j=1}^{n} P(X_j \leq x) = F^n(x)$$

gegen einen bestimmten Ausdruck $G_{max}(x)$ strebt, vorausgesetzt die Verteilung F gehört zu einer bestimmten Klasse von Verteilungen; für Verteilungen, die nicht zu dieser Klasse gehören, existiert eine solche Grenzverteilung nicht. Die Klasse von Verteilungen mit einer Grenzverteilung zerfällt in drei Teilklassen, die jeweils durch eine bestimmte Grenzverteilung charakterisiert sind. Existiert also für eine Verteilung F eine Grenzverteilung, so ist es eine Grenzverteilung von nur drei möglichen. Ist F die Gaussverteilung, so ist die zugehörige Grenzverteilung durch die *Extremwertverteilung*

$$G_{max}(x) = 1 - \exp(-e^x). \tag{2.132}$$

gegeben[13]. Dass $G_{max}(x)$ tatsächlich eine Verteilungsfunktion ist, sieht man, wenn man die Fälle $x \to -\infty$ und $x \to \infty$ betrachtet. Für $x \to \infty$ strebt e^x gegen unendlich und demnach $\exp(-\exp(-x))$ gegen 0, so dass $G(x)$ gegen 1 strebt. für $x \to -\infty$ strebt e^x gegen 0 und $\exp(-\exp(-x))$ gegen 1, also strebt $G(x)$ gegen 0; damit ist gezeigt, dass G für $-\infty < x < \infty$ tatsächlich $0 \leq G_{max}(x) \leq 1$ gilt. Auf die Details des Umgangs mit Extremwertfunktionen (Skalierung der Variablen für hinreichend großes, aber endliches n) kann hier verzichtet werden.

Für das Minimum (Wer hat die kleinste Reaktionszeit?) ergibt sich die Verteilung wie folgt. Ist x der Maximalwert der $X_1, \ldots, X_n$ und y der Minimalwert, so

[13] Die übrigen Grenzverteilungen sind die Pareto- und die Weibullverteilung.

ist $-y = \max(-X_1, \ldots, -X_n)$, d.h. $-y$ muß der Maximalwert der negativen Werte der X_j. Es läßt sich zeigen, dass

$$G_{min}(y) = 1 - G_{max}(-y) \tag{2.133}$$

gilt. Ist G_{max} wie in (2.132) definiert, erhält man

$$G_{min}(y) = \exp(-e^{-y}). \tag{2.134}$$

Für $y \to \infty$ folgt $e^{-y} \to 0$, so dass $G_{min}(y) \to 1$, und für $y \to -\infty$ folgt $e^{-y} \to \infty$, so dass $G_{min}(y) \to 0$; $G_{min}(y)$ ist also tatsächlich eine Verteilungsfunktion für $-\infty < y < \infty$.

Kapitel 3

Loglineare Modelle

Bei der kategorialen Regression ist die Fragestellung *asymmetrisch*: Eine bestimmte Variable gilt als abhängig, die übrigen als unabhängig. Bei loglinearen Modellen ist die Fragestellung dagegen *symmetrisch*: man ist an den Beziehungen zwischen Variablen interessiert. Es gibt drei mögliche Erhebungsschemata, die zuerst vorgestellt werden sollen:

3.1 Einführung, insbesondere 2-dimensionale Tabellen

Betrachtet werde zunächst ein einfaches Beispiel; die Tafel habe die Form der Tabelle 3.1. Die π_{ij} sind die Wahrscheinlichkeiten für das Auftreten der Kombination

Tabelle 3.1: $I \times J$-Kontingenztabelle

	Variable B				
Variable A	B_1	B_2	$\cdots$	B_J	Σ
A_1	π_{11}	π_{12}	$\cdots$	π_{1J}	π_{1+}
A_2	π_{21}	π_{22}	$\cdots$	π_{2J}	π_{1+}
$\vdots$			$\cdots$		$\vdots$
A_I	π_{I1}	π_{I2}	$\cdots$	π_{IJ}	π_{I+}
Σ	π_{+1}	π_{+2}	$\cdots$	π_{IJ}	π_{++}

(A_i, B_j). Es sei n die Gesamtzahl der Fälle in der Tafel. Mit $\hat{n}_{ij}$ sollen die *erwarteten* oder *wahren Häufigkeiten* bezeichnet werden; dementsprechend soll gelten

$$\pi_{ij} = \frac{\hat{n}_{ij}}{n}. \tag{3.1}$$

Für die *Randwahrscheinlichkeiten*, d.h. für die Wahrscheinlichkeit, überhaupt eine Beobachtung in einer bestimmten Kategorie zu machen, gelten dann die Beziehun-

gen

$$\pi_{i+} = \sum_j \pi_{ij} \tag{3.2}$$

$$\pi_{+j} = \sum_j \pi_{ij} \tag{3.3}$$

denn die Kategorien B_j, $j = 1, \ldots, J$ schließen einander aus; eine analoge Aussage gilt für die Kategorien A_i, $i = 1, \ldots, I$. Die Beziehungen (3.2) bzw. (3.3) machen natürlich nur Sinn, wenn bei einem kategorisierten Objekt oder bei einer Person nicht von vornherein feststeht, dass sie bereits zu einer Kategorie gehört; hierauf wird weiter unten noch eingegangen. Jedenfalls folgt noch

$$\sum_i \pi_{i+} = \sum_j \pi_{+j} = 1. \tag{3.4}$$

Betrachtet man die spezielle Hypothese H_0 : "A und B sind unabhängig voneinander", d.h. soll für die erwartete Häufigkeit in den Zellen

$$\hat{n}_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}, \tag{3.5}$$

so folgt sofort

$$\log \hat{n}_{ij} = \log n + \log \pi_{i+} + \log \pi_{+j}, \quad i = 1, \cdots, I; j = 1, \cdots, J. \tag{3.6}$$

Es sei insbesondere

$$\mu^A = \frac{1}{I} \sum_i \log \pi_{i+} \tag{3.7}$$

$$\mu^B = \frac{1}{J} \sum_j \log \pi_{+j} \tag{3.8}$$

$$\mu = \log n + \mu^A + \mu^B \tag{3.9}$$

Weiter sei

$$\mu_i^A = \log \pi_{i+} - \mu^A \tag{3.10}$$

$$\mu_j^B = \log \pi_{i+} - \mu^B \tag{3.11}$$

Durch Einsetzen überprüft man dann leicht, dass der Spezialfall (3.6) in

$$\log \hat{n}_{ij} = \mu + \mu_i^A + \mu_j^B \tag{3.12}$$

übergeht, wobei

$$\sum_i \mu_i^A = \sum_j \mu_j^B = 0 \tag{3.13}$$

gilt, wie man durch Einsetzen in (3.10) und (3.11) leicht nachweist.

Das Modell (3.12) enthält keinen Wechselwirkungsterm; dies ist eine Implikation der Unabhängigkeit der Faktoren A und B. Um den allgemeinen Fall, bei dem Abhängigkeiten zwischen irgendwelchen Kategorien A_i und B_j existieren dürfen, behandeln zu können, muß ein solcher Term μ_{ij}^{AB} eingeführt werden. Setzt man

$$\mu_{ij}^{AB} = \log \pi_{ij} - \mu_i^A - \mu_j^B, \tag{3.14}$$

so findet man

$$\sum_{i,j} \mu_{ij}^{AB} = \sum_{i,j} \log \pi_{ij} - \sum_{i,j} \mu_i^A - \sum_{i,j} \mu_j^B = 0 \tag{3.15}$$

Damit hat man mit

$$\log \hat{n}_{ij} = \mu + \mu_i^A + \mu_j^B + \mu_{ij}^{AB} \tag{3.16}$$

ein Modell für die Häufigkeiten $\hat{n}_{ij}$, das dem Strukturmodell einer 2-dimensionalen Varianzanalyse entspricht. In (3.16) kommt aber noch der Term μ_{ij}^{AB} hinzu, und damit kann die Häufigkeit $\hat{n}_{ij}$ vollständig erklärt werden, auch wenn die Hypothese der Unabhängigkeit (3.5) nicht gilt. In Definition 14 wird ein diesem Sachverhalt entsprechender Ausdruck für das Modell (3.15) gegeben werden. Anders als bei der Varianzanalyse ist man *nur* an dem Term μ_{ij}^{AB} interessiert, da er ja die möglichen Assoziationen zwischen den A_i und B_j repräsentiert; die "Haupteffekte" μ_i^A und μ_j^B sind im allgemeinen von untergeordneter Bedeutung: zum Teil werden ihre Werte sogar vom Experimentator kontrolliert, etwa dann, wenn z.B. für die Zeilenkategorien A_i bestimmte Stichproben, und damit die n_{i+}, festgelegt werden. Die Abhängigkeiten zwischen den Kategorien drücken sich in den μ_{ij} aus, und es sind diese Abhängigkeiten, die die man nicht vorgeben kann und die man untersuchen will.

Definition 14 *Gegeben sei eine 2-dimensionale Kontigenztafel, und es gelte*

$$\log \hat{n}_{ij} = \mu + \mu_i^A + \mu_j^B + \mu_{ij}^{AB} \tag{3.17}$$

wobei

$$\sum_i \mu_i^A = \sum_j \mu_j^B = \sum_i \mu_{ij}^{AB} = \sum_j \mu_{ij}^{AB} = 0$$

gelte. Hierbei bildet μ_{ij}^{AB} eine Interaktion zwischen A und B ab. Dann heißt (3.17) ein saturiertes Modell.

Zumindest die μ_{ij}^{AB} sind im saturierten Modell freie Parameter. Will man sie alle aus den Daten schätzen, so muß man so viele Parameter schätzen, wie es Zellen in der Tabelle gibt. Dies wäre eine etwas triviale Übung, weil man dann stets die μ_{ij}^{AB} so bestimmen kann, dass $\log \hat{n}_{ij} = \log n_{ij}$ und damit $\hat{n}_{ij} = n_{ij}$ gilt, d.h. das saturierte Modell mit freien Parametern μ_{ij}^{AB} paraphrasiert nur die gegebene Datenmatrix. "Interessante" Modelle ergeben sich erst, wenn man Restriktionen bezüglich dieser Parameter einführt; die allgemeine Unabhängigkeitshypothese $\mu_{ij}^{AB} = 0$ für alle i, j ist eine der möglichen Restriktionen.

Beziehung des Modells (3.17) zu Wahrscheinlichkeiten.

$$n = \sum_{r=1}^{I} \sum_{s=1}^{J} n_{rs}.$$

Nun ist aber im allgemeinen Fall $\log \hat{n}_{ij} = mu + \mu_i^A + \mu_j^B + \mu_{ij}^{AB}$, also folgt

$$\hat{n}_{ij} = e^{\mu + \mu_i^A + \mu_j^B + \mu_{ij}^{AB}}$$

und

$$n = \sum_{r=1}^{I} \sum_{s=1}^{J} e^{mu + \mu_r^A + \mu_s^B + \mu_{rs}^{AB}},$$

so dass

$$\pi_{ij} = \frac{\exp(\mu + \mu_i^A + \mu_j^B + \mu_{ij}^{AB})}{\sum_{r,s} \exp(\mu + \mu_r^A + \mu_r^B + \mu_{rs}^{AB})} \tag{3.18}$$

folgt.

Test für die Existenz von Abhängigkeiten: Die Existenz von Abhängigkeiten, d.h. von Assoziationen zwischen irgendwelchen A_i und B_j, läßt sich zunächst durch das Pearsonsche χ^2 testen:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}, \quad df = (I-1)(J-1),$$

wobei für die $\hat{n}_{ij}$ die Beziehung (3.5) angenommen wird. Ist der χ^2-Wert signifikant, so kann man davon ausgehen, dass Assoziationen zwischen Zeilen- und Spaltenkategorien existieren. Man weiß allerdings nicht, zwischen welchen Kategorien sie wirken. Man wird also Modelle testen wollen, bei denen für bestimmte i, j $\mu_{ij} \neq 0$ angenommen wird. Dazu müssen die entsprechenden μ_{ij} geschätzt werden. Die Schätzung geschieht i.a. nach der Maximum-Likelihood-Methode, hängt aber davon ab,in welcher Weise die Daten erhoben worden sind. Dazu werden im folgenden Abschnitt einige Erhebungsmethoden betrachtet.

3.1.1 Erhebungsweisen

Das produkt-multinomiale Schema

Dieses Schema ist dem einer Varianzanalyse analog: es gibt eine (oder mehrere) unabhängige Variable und verschiedene Ausprägungen einer abhängigen Variablen. Eine Gruppe von Vpn wird diesen Ausprägungen - Stufen - zugeteilt und es wird bestimmt, welche Ausprägung der abhängigen Variablen sich bei ihr findet. Im Unterschied zur VA wird aber nicht ein Meßwert erhoben, denn die Ausprägung der abhängigen Variablen ist ja nur kategorial. Schlußfolgerungen über die Wirkung der unabhängigen Variablen sollen über die Häufigkeiten, mit denen die Ausprägungen

Tabelle 3.2: Produkt-multinomiales Schema

Faktorstufen	Reaktion				Σ
1	n_{11}	n_{12}	$\cdots$	n_{1J}	n_{1+}
2	n_{21}	n_{22}	$\cdots$	n_{2J}	n_{2+}
$\vdots$					
I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	n_{I+}

der abhängigen Variablen auftreten, erreicht werden. Das Schema entspricht dem der Tabelle 3.2: Es gibt I Faktorstufen und J Ausprägungen bzw. Kategorien für die abhängige Variable. Die Randsummen n_{i+}, $i = 1, \cdots, I$ werden dabei vom Experimentator/Planer der Untersuchung festgelegt. Die Häufigkeiten n_{ij} in der i-ten Zeile sind jeweils multinomial verteilt:

$$p(n_{i1}, \cdots, n_{iJ}) = \prod_{i=1}^{I} \frac{n_{i+}!}{n_{i1}! n_{i2}! \cdots n_{iJ}!} \pi_{i1}^{n_{i1}} \cdots \pi_{iJ}^{n_{iJ}} \tag{3.19}$$

Unter der Hypothese H_0 :

$$\pi_{1j} = \pi_{2j} = \cdots = \pi_{Ij}, \quad \text{für alle } j \tag{3.20}$$

ist die *erwartete Häufigkeit*

$$\hat{n}_{ij} = n_{i+}\pi_{ij}$$

in der der Zelle (i, j) der Tafel 3.2 durch

$$\hat{n}_{ij} = \frac{n_{i+}n_{+j}}{n_{++}} \tag{3.21}$$

gegeben. (n_{i+} ist *vorgegeben*!).

Beispiel 24 Es ist bekannt, dass die Fokussierung der Aufmerksamkeit auf bestimmte Aspekte eines Reizmusters die Klassifikation des Musters beeinflußt. So kann die Fokussierung auf ein *irrelevantes* Merkmal die Wahrscheinlichkeit einer korrekten Klassifikation verändern, und zwar in Abhängigkeit von der SOA (Stimulus Onset Asynchrony); dies ist die Zeitdauer zwischen der Darbietung eines Reizes und eines Maskierungsreizes. Die unabhängige Variable sei der Wert der SOA, die "Reaktion" sei die Entscheidung für ein bestimmtes Muster M_i, $i = 1, \ldots, 4$, wobei das tatsächlich gezeigte Muster stets das gleiche ist: für jede SOA wird das Reizmuster also genau siebzigmal gezeigt; zu entscheiden ist, ob sich die Verteilungen der Antworten pro SOA unterscheiden. □

Das multinomiale Schema

Hier werden Klassen von Kategorien (unabhängige Variablen, Faktoren im Sinne der VA) definiert, dann wird eine Stichprobe mit festem Umfang ausgewählt, die

Tabelle 3.3: Klassifikation und SOA

	Reaktion (= Muster)				
SOA	M_1	M_2	M_3	M_4	Σ
30 ms	17	20	17	16	70
40 ms	13	16	20	21	70
50 ms	11	15	18	26	70
Σ	41	51	55	47	210

dann nach Maßgabe des Vorhandenseins einer spezifischen Kategorienkombination aufgeteilt wird.

Beispiel 25 Alle Insassen einer Reihe von psychiatrischen Landeskrankenhäusern werden (i) bezüglich ihres Körperbautyps und (ii) bezüglich ihrer psychischen Erkrankung klassifiziert. Es ergaben sich die folgenden Daten (Westphal (1931)) Mit

Tabelle 3.4: Körperbau und psychische Erkrankung

		Erkrankung			
Typ		man. dep.	Epilepsie	Schizophr.	Σ
pyknisch	n_{ij}	879	83	717	1679
erw.	$\hat{n}_{ij}$	282	312	1085	
athletisch	n_{ij}	91	435	884	1410
erw.	$\hat{n}_{ij}$	237	262	911	
leptosom	n_{ij}	261	378	2632	3271
erw.	$\hat{n}_{ij}$	549	608	2114	
dysplastisch	n_{ij}	15	444	550	1009
erw.	$\hat{n}_{ij}$	170	187	652	
atypisch	n_{ij}	115	165	450	730
erw.	$\hat{n}_{ij}$	123	136	471	
	Σ	1361	1505	5233	8099
		$\chi^2 = 2641.56$, $df = 8$, $p = .000$			

"erw." werden die unter der Annahme, dass Körperbau und Erkrankung unabhängig voneinander sind, erwarteten Häufigkeiten $\hat{n}_{ij}$ bezeichnet; die Differenzen zwischen

n_{ij} und $\hat{n}_{ij}$ legen die Existenz von Abhängigkeiten nahe; der gefundene χ^2-Wert ist hochsignifikant. In jedem Fall ist die Verteilung der Häufigkeiten in den Zeilen der Tabelle wieder multinomial. Im Unterschied zum produkt-multinomialen Schema liegen aber die Randhäufigkeiten *nicht* fest. □

Das Poisson-Schema

In Beispiel 25 ist die Anzahl der zu beobachtenden Personen *vor* der Untersuchung festgelegt worden, denn es sollten ja *alle* Patienten klassifiziert werden. Damit liegt die Gesamtzahl fest. Die Zeit, bis alle Käufer beobachtet wurden, ist damit nicht festgelegt.

Umgekehrt kann man die Zeitdauer festlegen und dafür die Anzahl der Personen offen lassen. Diese Anzahl wird dann eine zufällige Veränderliche. Man könnte etwa alle Neuzugänge in die Landeskrankenhäuser für eine bestimmte Zeitdauer - etwa ein Jahr - nach ihrem Körperbau und der Art ihrer Erkrankung klassifizieren. Da die einzelnen Personen unabhängig voneinander in ein Krankenhaus eingeliefert werden, kann man annehmen, dass die Häufigkeiten Poisson-verteilt sind, d.h.

$$p(n_{11}, \cdots, n_{IJ}) = \prod_{i,j} \frac{\nu_{ij}^{n_{ij}}}{n_{ij}!} e^{-n_{ij}} \tag{3.22}$$

Die erwartete Zellhäufigkeit ist dann

$$E(n_{ij}) = \nu_{ij} \tag{3.23}$$

Als Hypothese H_0 über den Zusammenhang zwischen den Zellhäufigkeiten wird im allgemeinen

$$\nu_{ij} = \frac{\nu_{i+}\nu_{+j}}{\nu_{++}} \tag{3.24}$$

angenommen. Diese Hypothese heißt *multiplikative Hypothese* oder *multiplikatives Poisson-Modell.*

Die Wahrscheinlichkeit π_{ij} hängt mit den Parametern ν_{ij} über die Beziehung

$$\pi_{ij} = \frac{\nu_{ij}}{\sum_{k,l} \nu_{kl}} \tag{3.25}$$

zusammen. Der Wert der Parameter ν_{ij} hängt natürlich von der gewählten Zeitdauer der Beobachtung ab. Es sei $\tilde{n}_{ij} = E(n_{ij})$ die erwartete Häufigkeit, d.h. es sei

$$\tilde{n}_{ij} = E(n_{ij}) = n_{++}\pi_{ij} = n_{++}\frac{\nu_{ij}}{\sum_{k,l} \nu_{kl}}$$

Unter der Nullhypothese (keine Abhängigkeiten zwischen Zeilen- und Spaltenkategorien) gilt dann

$$H_0 \Longleftrightarrow \tilde{n}_{ij} = \frac{\tilde{n}_{i+}\tilde{n}_{+j}}{\tilde{n}_{++}}, \tag{3.26}$$

Diese Beziehung entspricht der beim multinomialen Erhebungsschema.

Satz 7 *Gemäß der Unabhängigkeitshypothese gilt* $\hat{n}_{ij} = n_{i+}n_{+j}/n_{++}$. *Ist das Erhebungsschema*

1. *das* Poisson-Schema, *so ist diese Hypothese äquivalent dem Modell (3.12), ohne weitere Nebenbedingungen;*

2. *das* produkt-multinomiale Schema, *so ist die Hypothese äquivalent dem Modell (3.12) mit der Nebenbedingung*

$$n_{i+} = \sum_j e^{\mu+\mu_i^A+\mu_j^B}, \quad j = 1, \cdots, J. \tag{3.27}$$

3. *das* multinomiale Schema, *so ist die Hypothese äquivalent zu (3.12) mit der Nebenbedingung*

$$n = \sum_{i,j} e^{\mu+\mu_i^A+\mu_j^B}. \tag{3.28}$$

Beweis: Vergl. Fahrmeir und Hamerle (1984), p. 480. □

Spricht man also vom loglinearen Unabhängigkeitsmodell, so müssen die jeweiligen Restriktionen des betrachteten Schemas mitgedacht werden.

Für das Poisson-Schema müssen keine weiteren Restriktionen berücksichtigt werden. Für das produkt-multinomiale Schema hat man die Restriktion

$$n_{i+} = \sum_{i,j} e^{\mu+\mu_i^A+\mu_j^B+\mu_{ij}^{AB}}$$

und für das multinomiale Schema gilt

$$n_{++} = \sum_j e^{\mu+\mu_i^A+\mu_j^B+\mu_{ij}^{AB}}$$

3.1.2 Parameter, Logits und Kreuzproduktverhältnisse.

Der Einfachheit halber sei $J = 2$. Man kann dann die Logits

$$\log \frac{p(B_1|A_i)}{p(B_2|A_i)} = \log \frac{n_{i1}}{n_{i2}}$$

betrachten. Für die n_{ij} gelte das Modell (3.17). Es werde für den Augenblick angenommen, dass die Hypothese der Unabhängigkeit gilt, so dass die Interaktionsterme μ_{ij}^{AB} alle verschwinden. Eingesetzt ergibt sich

$$\log \frac{n_{i1}}{n_{i2}} = \mu + \mu_i^A + \mu_1^B - \mu - \mu_i^A - \mu_2^B = \mu_1^A - \mu_2^B \tag{3.29}$$

Die Hypothese der Unabhängigkeit impliziert also, dass die Logits für alle i identisch sind.

Die Betrachtung der Odd-Ratios erlaubt eine Interpretation der Parameter des loglinearen Modells. Es sei außerdem (der Einfachheit halber) $I = 2$. Das Kreuzproduktverhältnis für diese Tabelle ist

$$\Theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

Man findet

$$\begin{aligned} \log\Theta &= \log n_{11} + \log n_{22} - \log n_{12} - \log n_{21} \\ &= \mu + \mu_1^A + \mu_1^B + \mu_{11}^{AB} + \mu_2^A + \mu_2^B + \mu_{22}^{AB} \\ &\quad -(\mu + \mu_1^A + \mu_2^B + \mu_{12}^{AB}) - (\mu + \mu_2^A + \mu_1^B + \mu_{21}^{AB}) \\ &= \mu_{11}^{AB} + \mu_{22}^{AB} - \mu_{12}^{AB} - \mu_{21}^{AB} \end{aligned} \tag{3.30}$$

Es gelten die Nebenbedingungen $\sum_i \mu_{ij}^{AB} = \sum_j \mu_{ij}^{AB} = 0$, und $\mu_{11}^{AB} = \mu_{22}^{AB} = -\mu_{12}^{AB} = -\mu_{21}^{AB}$. Deshalb folgt

$$\log\Theta = 4\mu_{11}^{AB} \tag{3.31}$$

Andererseits ist Θ der Assoziationsparameter der 2×2-Tabelle. Gilt die Hypothese der Unabhängigkeit, so ist $\Theta = 1$ und $\log\Theta = 0$. Diese Bedingung ist genau dann erfüllt, wenn in (3.31) die Bedingung $\mu_{11}^{AB} = 0$ erfüllt ist.

Beispiel 26 Es soll die Hypothese, dass in den USA des Mordes angeklagte Schwarze häufiger zum Tode werden als des Mordes angeklagte Weiße. Man betrachte die folgende Tabelle[1]: Der Anteil zum Tode verurteilter Weißer ist $19/141 = .135$, der

Tabelle 3.5: Verhängung der Todesstrafe in den USA

	Todesstrafe		
Angeklagte	ja	nein	Σ
weiß	19	141	160
schwarz	17	149	166
Σ	36	290	326

Anteil zum Tode verurteilter Schwarzer ist $17/149 = .114$. Es scheint also eher so zu sein, dass Schwarze weniger häufig zum Tode verurteilt werden als Weiße. Betrachtet man noch den Odds-Ratio als Assoziationsmaß, so findet man

$$\Theta = \frac{19 \times 149}{141 \times 17} = 1.181$$

Es ist zwar $\Theta \neq 1$, so dass eine Abhängigkeit existieren könnte. Andererseits weicht der Wert nicht stark von 1 ab, so dass der Wert von Θ auch mit der Hypothese der Unabhängigkeit verträglich sein könnte.

[1] Rachelet, M.: Racial characteristics and imposition of the death penalty. Amer. Sociol. Review **46**, 918-927

Nach (3.31) ist nun $\log \Theta = 4\mu_{11}^{AB}$, und $\log 1.181 \approx .072$. In bezug auf (3.17) heißt dies, dass man $\mu_{11}^{AB} = 0$ vermuten kann, d.h. es liegt die Vermutung nahe, des es keinen Zusammenhang zwischen Verhängung der Todesstrafe und der Farbe der Angeklagten gibt. In bezug auf (3.33) entspricht dies $G^2 = .072$ bei $df = (I-1)(J-1) = 1$ Freiheitsgraden, und dieser Wert ist, wie auch der "klassische" χ^2 Wert $\chi^2 = .22$, nicht signifikant (Wahrscheinlichkeit eines solchen Wertes: $p = .6379$.) Es wird später gezeigt werden, dass der Sachverhalt nicht ganz so einfach ist wie er sich nach diesem Test darstellt. □

3.2 Tests für die Güte der Anpassung

Mit dem Pearsonschen χ^2-Test kann die Überzufälligkeit der gefundenen Assoziationen zwischen den Zeilen- und Spaltenkategorien geprüft werden; dies gilt auch für höherdimensionale Tabellen. Allgemein prüft man mit diesem Test, ob die Wahrscheinlichkeiten einer Multinomialverteilung bestimmten Hypothesen genügen. Für eine 2-dimensionale Tabelle hat man

$$X^2 = \sum_{i,j} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}, \quad df = (I-1)(J-1) \tag{3.32}$$

Für hinreichend große Stichproben gilt $X^2 \sim \chi^2$,d.h. die Verteilung von X^2 entspricht dann der einer χ^2-Verteilung.

Ein allgemeiner Ansatz, die Nullhypothese H_0 gegen eine Alternativhypothese H_1 zu testen, besteht darin, eien Likelihood-Ratio-Test zu konstruieren. Dazu wird die maximale Likelihood der Daten (i) unter H_0 und (ii) unter H_1 bestimmt und dann der Quotient Λ dieser maximalen Likelihoods berechnet. Wilks (1935, 1938) hat gezeigt, dass dann die Größe

$$G^2 \stackrel{def}{=} -2\log\Lambda \sim \chi^2, \quad H_0, \quad n \to \infty \tag{3.33}$$

erfüllt, d.h. $-2\log\Lambda$ ist für hinreichend großes n approximativ wie χ^2 unter der Nullhypothese verteilt. G^2 heißt auch *Likelihood-Ratio-χ^2-Statistik*. Für das multinomiale Schema gilt insbesondere

$$G^2 = -2\log\Lambda = 2\sum_{i,j} n_{ij}\log(n_{ij}/\hat{n}_{ij}), \quad \hat{n}_{ij} = \frac{n_{i+}n_{+j}}{n_{++}} \tag{3.34}$$

Die Parameter sind im allgemeinen Fall durch die π_{ij} gegeben, sie unterliegen der Nebenbedingung

$$\sum_{i,j} \pi_{ij} = 1$$

Deswegen können $IJ-1$ von ihnen frei gewählt werden, das IJ-te liegt dann fest. Damit ist die Anzahl der Freiheitsgrade gleich $IJ-1$. Unter H_0 gilt aber

$$\pi_{ij} = \pi_{i+}\pi_{+j}$$

Da wiederum $\sum_i \pi_{i+} = \sum_j \pi_{+j} = 1$ gilt, können $I-1$ Parameter π_{i+} und $J-1$ Parameter π_{+j} frei gewählt werden, also insgesamt $I-1+J-1$. Nun ist G^2 aber eine *Differenz* von approximativ χ^2-verteilten Größen; die Anzahl der Freiheitsgrade für G^2 ist dann durch

$$IJ - 1 - (I - 1 + J - 1) = IJ - I - (J - 1) = (I - 1)(J - 1).$$

gegeben.

X^2 konvergiert im allgemeinen schneller als G^2 gegen die χ^2-Verteilung; für $n/(IJ) < 5$ ist die Approximation der Verteilung für G^2 durch die χ^2-Verteilung eher schlecht.

Alle Betrachtungen übertragen sich auf den höherdimensionalen Fall.

3.3 Verallgemeinerung: 3-dimensionale Tafeln

Bekanntlich sind Korrelationen nur mit Vorsicht zu interpretieren: das bekannte Beispiel über den Zusammenhang zwischen Alkoholkonsum in den USA und der Häufigkeit, mit der dort der Priesterberuf gewählt wird, lehrt, dass erst die Betrachtung weiterer Variablen zu einer sinnvolleren Interpretation führt. Dieser Sachverhalt muß auch bei der Diskussion von Kontingenztabellen berücksichtigt werden. So kann ein Zusammenhang zwischen zwei Variablen durch die Wirkung einer oder mehrerer nicht berücksichtigter Variablen verdeckt werden, oder er existiert nur scheinbar. Zur Verdeutlichung werden jetzt 3-dimensionale Tabellen betrachtet.

Die drei Variablen (Klassen von Kategorien) seien A, B und C, mit jeweils I, J und K Kategorien A_i, B_j und C_k. n_{ijk} sei die Häufigkeit in der (i,j,k)-ten Zelle. Die Wahrscheinlichkeit, eine Beobachtung in der (i,j,k)-ten Zelle zu machen, sei π_{ijk}.

Aus der 3-dimensionalen Häufigkeitstabelle lassen sich auf verschiedene Weise 2-dimensionale Tabellen bilden:

1. *Partialtabellen:* Dies sind Tabellen, die durch einen "Schnitt" durch die 3-dimensionale Tabelle entstehen, der durch eine Stufe einer der drei Faktoren (Variablen, Klassen) entsteht. Man hält z.B. A_i fest und betrachtet *für diese Stufe* die Tabelle $B \times C$. In den Zellen dieser Tabelle stehen die Häufigkeiten $n_{i;jk}$ mit $i = \textit{konstant}$. Die Abhängigkeiten in einer Partialtabelle heißen "partielle Assoziationen".

2. *Marginaltabellen:* Tabellen dieser Art entstehen, wenn über einen Faktor (Variable, Klasse) *aggregiert*, d.h. summmiert wird. Summiert man über alle Stufen von A, so entsteht wieder eine $B \times C$-Tabelle, in deren Zellen die Häufigkeiten $n_{+jk} = \sum_i n_{ijk}$ stehen. Die Abhängigkeiten in einer Marginaltabelle heißen "marginale Assoziationen". Die Assoziationen in Marginaltabellen können sich sehr von denen in Partialtabellen unterscheiden; dieses Phänomen ist

als *Simpsons Paradoxon* bekannt. Bevor man eine Marginaltabelle betrachtet, muß die Frage der *Aggregierbarkeit* diskutiert werden; hierauf wird später noch ausführlich eingegangen.

Wie im 2-dimensionalen Fall werden die $\log n_{ijk}$ betrachtet. Diese Werte werden *parametrisiert*: es sei

$$\begin{aligned}
\mu &= \frac{1}{IJK}\sum_{i,j,k}\log n_{ijk} && (3.35)\\
\mu_i^A &= \frac{1}{JK}\sum_{j,k}\log n_{ijk} - \mu && (3.36)\\
\mu_j^B &= \frac{1}{IK}\sum_{i,k}\log n_{ijk} - \mu && (3.37)\\
\mu_k^C &= \frac{1}{IJ}\sum_{i,j}\log n_{ijk} - \mu && (3.38)\\
\mu_{ij}^{AB} &= \frac{1}{K}\sum_{k}\log n_{ijk} - \mu_I^A - \mu_j^B && (3.39)\\
\mu_{ik}^{AC} &= \frac{1}{J}\sum_{j}\log n_{ijk} - \mu_i^A - \mu_k^C - \mu && (3.40)\\
\mu_{jk}^{BC} &= \frac{1}{I}\sum_{i}\log n_{ijk} - \mu_j^B - \mu_k^C - \mu && (3.41)\\
\mu_{ijk}^{ABC} &= \log n_{ijk} - \mu_i^A - \mu_j^B - \mu_k^C - \mu_{ij}^{AB} - \mu_{ik}^{AC} - \mu_{jk}^{BC} && (3.42)
\end{aligned}$$

Dann lassen sich die $\log n_{ijk}$ stets in der Form

$$\log n_{ijk} = \mu + \mu_i^A + \mu_j^B + \mu_k^C + \mu_{ij}^{AB} + \mu_{ik}^{AC} + \mu_{jk}^{BC} + \mu_{ijk}^{ABC} \qquad (3.43)$$

darstellen. Man rechnet durch Einsetzen leicht nach, dass die folgenden Bedingungen erfüllt sind:

$$\begin{aligned}
\sum_i \mu_i^A &= \sum_j \mu_j^B = \sum_k \mu_k^C && (3.44)\\
\sum_i \mu_i^{AB} &= \sum_j \mu_{ij}^{AB} = 0\\
\sum_j \mu_{jk}^{BC} &= \sum_k \mu_{jk}^{BC}\\
\sum_i \mu_{ik}^{AC} &= \sum_k \mu_{ik}^{AC}\\
\sum_i \mu_{ijk}^{ABC} &= \sum_j \mu_{ijk}^{ABC} = \sum_k \mu_{ijk}^{ABC} = 0
\end{aligned}$$

Die Gleichung (3.43) zusammen mit den Nebenbedingungen (3.45) definiert das *saturierte 3-dimensionale loglineare Modell* dar. Es gelten die folgenden Bezeichnungsweisen:

- μ ist das *Gesamtmittel* der logarithmierten, zu erwartenden Häufigkeiten,
- μ_i^A, μ_j^B und μ_k^C heißen *Haupteffekte* der drei Variablen A, B und C,
- μ_{ij}^{AB}, μ_{ik}^{AC} und μ_{jk}^{BC} heißen *Wechselwirkungs-* oder *Interaktionsterme 1. Ordnung*,
- μ_{ijk}^{ABC} sind die *Wechselwirkungs-* oder *Interaktionsterme 2. Ordnung*, bzw. *Drei-Faktor-Interaktionen.*
- Mit μ^A, μ^B, $\cdots$, μ^{AB} etc werden die Haupt- und Interaktionsterme allgemein bezeichnet.

Das Modell (3.43) ist eigentlich "nur" eine Reparametrisierung der Daten und kann deshalb stets angepaßt werden. Interessanter sind deshalb Modelle, bei denen bestimmte Interaktionsterme weggelassen werden.

Weitere Nebenbedingungen ergeben sich durch die spezielle Erhebungsweise einer Untersuchung.

3.3.1 Typen von Unabhängigkeit

Das Modell ($AB/AC/BC$)

Eine erste Vereinfachung des Modells (3.43) ergibt sich, wenn $\mu^{ABC} = 0$; man erhält

$$\log n_{ijk} = \mu + \mu_i^A + \mu_j^B + \mu_k^C + \mu_{ij}^{AB} + \mu_{ik}^{AC} + \mu_{jk}^{BC} \tag{3.45}$$

In bezug auf das Beispiel 3.3.2 bedeutet $\mu^{ABC} = 0$, dass keine spezifischen Beziehungen zwischen A der Farbe des Täters, der des Opfers B und der Verhängung der Todesstrafe C bestehen; natürlich sind noch Interaktionen $\mu^{AB} \neq 0$, μ^{AC} und μ^{BC} möglich. Die Wechselwirkung $A \times B$ bedeutet, dass es eine Abhängigkeit zwischen der Farbe des Täters und der des Opfers gibt (Weiße bringen nur Schwarze um und umgekehrt, oder Weiße töten nur Weiße, Schwarze aber Weiße und Schwarze, etc.), Interaktionen der Form $A \times C$ und $B \times C$ signalisieren, dass die Todesstrafe in Abhängigkeit von der Hautfarbe des Täters und/oder des Opfers abhängt. $\mu^{ABC} = 0$ bedeutet, dass die Wirkung der Farbe des Opfers stets gleich ist, unabhängig von der Farbe des Täters, und dass die Wirkung der Farbe des Täters unabhängig von der Farbe des Opfers ist.

Es sei andererseits $\mu^{ABC} \neq 0$. Eine mögliche Form dieser Abhängigkeit besteht darin, dass etwa für $A \times C$ keine Abhängigkeit von B_2 (Farbe des Opfers ist schwarz) gibt; wenn ein Schwarzer wegen Mordes angeklagt wird, so kann das Todesurteil von seiner Hautfarbe begünstigt werden, aber die Tatsache, dass er einen Schwarzen getötet hat, ist ohne Belang. War sein Opfer aber weiß (B_2), so kann dies die Wahrscheinlichkeit eines Todesurteils erheblich erhöhen. Es gibt also einen Zusammenhang zwischen $A \times C$ und B_2.

Das Modell der bedingten Unabhängigkeit

Man kann die Beziehung zwischen A und B betrachten und dabei die "Werte" von C kontrollieren.

Definition 15 *Es sei C_k die k-te Kategorie des Faktors C, und es sei $T_{AB|C_k}$ die Kontingenztabelle für die Faktoren A und B für C_k; $T_{AB|C_k}$ ist die k-te Scheibe aus der 3-dimensionalen Tafel $A \times B \times C$. Weiter sei $\pi_{ij|k}$ die Wahrscheinlichkeit für eine Beobachtung A_i, B_j, gegeben die Kategorie C_k; es ist $\pi_{ij|k} = \pi_{ijk}/\pi_{++k}$. Gilt*

$$\pi_{ij|k} = \pi_{i+|k}\pi_{+j|k}, \quad \textit{für alle } i,j \tag{3.46}$$

so heißen A und B bedingt unabhängig, gegeben C_k. *Gilt*

$$\pi_{ij|k} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}}, \quad \textit{für alle } i,j,k \tag{3.47}$$

so heißen A und B unabhängig, gegeben C; *A und B sind dann unabhängig für alle Kategorien von C.*

Sind A und B bedingt unabhängig, gegeben C, so gilt das loglineare Modell

$$\log n_{ijk} = \mu + \mu_i^A + \mu_j^B + \mu_k^C + \mu_{ik}^{AC} + \mu_{jk}^{BC} \tag{3.48}$$

In diesem Modell soll also $\mu^{AB} = \mu^{ABC} = 0$ gelten. In bezug auf das Beispiel soll es also keine Interaktion zwischen der Hautfarbe des Täters und der des Opfers geben und darüber hinaus gibt es keine Abhängigkeit zwischen der Hautfarbe von Täter, Opfer und Verhängung der Todesstrafe. Die n_{ijk} lassen sich dann gemäß

$$n_{ijk} = \frac{n_{i+k}n_{+jk}}{n_{++k}} \tag{3.49}$$

voraussagen. Es sei n die Gesamtzahl der Beobachtungen; dann ist $n_{ijk} = n\pi_{ijk}$ und es folgt

$$\pi_{ijk} = \frac{\pi_{i+k}\pi_{+jk}}{i_{++k}} \tag{3.50}$$

Dividiert man beide Seiten noch einmal durch π_{++k}, so erhält man

$$\frac{\pi_{ijk}}{\pi_{++k}} = \frac{\pi_{i+k}}{\pi_{++k}}\frac{\pi_{+jk}}{\pi_{++k}} \tag{3.51}$$

Es ist $\pi_{ijk} = p(A_i \cap B_j \cap C_k)$, und es ist

$$p(A_i \cap B_j \cap C_k) = p(A_i \cap B_j|C_k)p(C_k)$$

oder

$$\frac{p(A_i \cap B_j \cap C_k)}{p(C_k)} = p(A_i \cap B_j|C_k)$$

Nach (3.51) muß aber auch

$$\frac{p(A_i \cap B_j \cap C_k)}{p(C_k)} = \frac{p(A_i \cap B_j)}{p(C_k)}\frac{p(B_j \cap C_k)}{p(C_k)}$$

gelten, und damit

$$p(A_i \cap B_j|C_k) = p(A_i|C_k)p(B_j|C_k).$$

Dies heißt aber, dass A und B *bedingt unabhängig, gegeben C_k* sind.

Unabhängigkeit von einer Variablen, z.B. AC/B

Man kann auch ein Modell betrachten, das entsteht, wenn $\mu^{ABC} = 0$ angenommen wird und darüber hinaus zwei Interaktionen 1. Ordnung vernachlässigt werden, d.h.

$$\mu^{ABC} = \mu^{AB} = 0$$

Definition 16 *Der Faktor B ist* gemeinsam unabhängig[2] *von A und C, wenn*

$$\pi_{ijk} = \pi_{i+k}\pi_{+j+} \tag{3.52}$$

gilt.

Hier kann man die Kombinationen von A und C als "Werte" einer neuen Variablen (Faktoren) AC ansehen, und die Faktoren B und AC sind unabhängig. Das entsprechende loglineare Modell ist

$$\log n_{ijk} = \mu + \mu_i^A + \mu_j^B + \mu_k^C + \mu_{ik}^{AC}. \tag{3.53}$$

Es folgt

$$n_{ijk} = \frac{n_{i+k}n_{+j+}}{n_{+++}} \tag{3.54}$$

und

$$p(A_i \cap B_j \cap C_k) = p(A_i \cap B_j)p(C_k) \tag{3.55}$$

Im Beispiel 3.3.2 bedeutet diese Form der Unabhängigkeit, dass das Ereignis, dass die Todesstrafe verhängt wird, unabhängig von der Täter-Opfer-Kombination (bezüglich der Farben) ist.

Die vollständige Unabhängigkeit, A/B/C

Definition 17 *Die Faktoren A, B und C heißen* wechselseitig unabhängig, *wenn*

$$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k} \tag{3.56}$$

gilt.

Dann folgt

$$\log \pi_{ijk} = \log \pi_{i++} + \log \pi_{+j+} + \log \pi_{++k}$$

bzw.

$$\log n_{ijk} = \mu + \mu_i^A + \mu_j^B + \mu_k^C. \tag{3.57}$$

Die wechselseitige Unabhängigkeit bedeutet dann, dass jeder Faktor gemeinsam unabhängig von allen anderen ist.

[2] jointly independent

Hierarchische Modelle

Sind die Faktoren wechselseitig unabhängig, so ist B gemeinsam unabhängig von A und C und A und B sind bedingt unabhängig.

Alle hier betrachteten Modelle entstanden sukzessive aus dem saturierten Modell; es wurde zuerst die Interaktion 2. Ordnung (μ^{ABC}) fortgelassen, aber unter Beibehaltung aller Interaktionen 1. Ordnung. Dann wurden Interaktionen der 1. Ordnung weggelassen, aber unter Beibehaltung aller Haupteffekte. Schließlich kann man auch noch Haupteffekte vernachlässigen. Jedenfalls werden die untergeordneten Effekte beibehalten, bis die jeweiligen übergeordneten Effekte alle verschwunden sind. Modelle dieser Art heißen *hierarchische Modelle*.

Tabelle 3.6: Typen von Unabhängigkeit

Typ	π_{ijk}	Assoziation
wechselseitig	$\pi_{i++}\pi_{i+k}\pi_{++k}$	keine
B gemeinsam von A und C	$\pi_{i+k}\pi_{+j+}$	μ_{ik}^{AC}
A, B bedingt unabh. von C	$\pi_{i+k}\pi_{+jk}/\pi_{++k}$	$\mu_{ik}^{AC} + \mu_{jk}^{BC}$

3.3.2 Gesamtzahl möglicher Modelle

Bei einem 3-faktoriellen Design sind eine Anzahl von Modellen möglich; für ein solches Design hat man insgesamt die in Tabelle 3.7 angegebenen Kombinationen. Es mag in speziellen Situationen sinnvoll sein, alle Modelle zu diskutieren, sehr häufig wird man aber nur an einigen der überhaupt möglichen Modelle interessiert sein. In Beispiel wird eine 3-dimensionale Tabelle diskutiert werden.

3.3.3 Interpretation der Parameter

Die Parameter eines loglinearen Modells werden über Odds-Ratios interpretiert.

Drei-Faktor-Interaktion

Um die Bedeutung der Drei-Faktor-Interaktion μ^{ABC} zu finden, betrachtet man das bedingte Kreuzproduktverhältnis:

Tabelle 3.7: Mögliche Modelle bei 3-dimensionalem Design

Modell	Bedeutung
A, B, C	Vollst. Unabhängigkeit
AB, C	Assoz. $A \times B$ unabh. v. C
AC, B	Assoz. $A \times C$ unabh. v. B
BC, A	Assoz. $B \times C$ unabh. v. A
AB, AC	Assoz. $A \times B$, $A \times C$
AB, BC	Assoz. $A \times B$, $B \times C$
AC, BC	Assoz. $A \times C$, $B \times C$
ABC	Assoz. $A \times B \times C$

Definition 18 *Es bezeichnen i_1 und i_2 irgendwelche Stufen von A, und j_1, j_2 mögen irgendzwei Stufen von B bezeichnen. Für die feste Stufe C_k von C heißt*

$$\Theta_{i_1 i_2; j_1 j_2|k} = \frac{p(A_{i_1}|B_{j_1} \cap C_k)/p(A_{i_2}|B_{j_1} \cap C_k)}{p(A_{i1}|B_{j_2} \cap C_k)/p(A_{i_2}|B_{j_2} \cap C_k)} = \frac{n_{i_1 j_1 k} n_{i_2 j_2 k}}{n_{i_1 j_2 k} n_{i_2 j_1 k}} \tag{3.58}$$

das bedingte Kreuzproduktverhältnis, gegeben C_k.

Ist das bedingte Kreuzproduktverhältnis identisch für alle C_k, so hat C keinen Einfluß auf die Beziehung zwischen (d.h. auf die Assoziation von) A und B, die Verhängung der Todesstrafe wäre dann unabhängig von der speziellen Täter-Opfer-Kombination. Ist das Verhältnis nicht unabhängig von C, so hängt das Urteil eben *von der Kombination* ab.

Logarithmiert man $\Theta_{i_1 i_2; j_1 j_2|k}$ und setzt man für $\log n_{i_1 j_1 k}$ die entsprechenden Ausdrücke des saturierten Modells ein, so ergibt sich (längliche Rechnung)

$$\begin{aligned} \log \Theta_{i_1 i_2; j_1 j_2|k} \quad = \quad & (\mu^{ABC}_{i_1 j_1 k_1} - \mu^{ABC}_{i_2 j_1 k_1}) - (\mu^{ABC}_{i_1 j_2 k_1} - \mu^{ABC}_{i_2 j_2 k_1}) \\ & -(\mu^{ABC}_{i_1 j_2 k_2} - \mu^{ABC}_{i_2 j_1 k_2}) - (\mu^{ABC}_{i_1 j_2 k_2} - \mu^{ABC}_{i_2 j_2 k_2}) \end{aligned} \tag{3.59}$$

$\Theta_{i_1 i_2; j_1 j_2|k}$ bildet die Wirkung von C auf den Zusammenhang zwischen A und B ab, und (3.59) zeigt, dass diese Einwirkung auf die Drei-Faktor-Interaktionen zurückzuführen ist. Für $\mu^{ABC} = 0$ ist $\Theta_{i_1 i_2; j_1 j_2|k} = 1$ für alle C_k.

Zwei-Faktor-Interaktionen

Für die Interpretation der Zwei-Faktor-Interaktionen betrachtet man die bedingten Kreuzprodukte

$$\Theta_{i_1 i_2; j_1 j_2|k} = \frac{n_{i_1 j_1 k} n_{i_2 j_2 k}}{n_{i_1 j_2 k} n_{i_2 j_1 k}} \tag{3.60}$$

und die Logarithmierung ergibt

$$\begin{aligned}\log \frac{\Theta_{i_1 i_2;j_1 j_2|k_1}}{\Theta_{i_1 i_2;j_1 j_2|k_2}} &= (\mu^{AB}_{i_1 j_1} - \mu^{AB}_{i_2 j_1}) - (\mu^{AB}_{i_1 j_2} - \mu^{AB}_{i_2 j_2}) \\ &\quad + \mu^{ABC}_{i_1 j j_1 k} + \mu^{ABC}_{i_2 j_2 k} - \mu^{ABC}_{i_1 j_2 k} - \mu^{ABC}_{i_2 j_1 k}\end{aligned} \tag{3.61}$$

Demnach wird der Zusammenhang zwischen A und B durch die Interaktion μ^{AB} sowie durch die Drei-Faktoren-Interaktion μ^{ABC} bestimmt. Für $\mu^{AB} = 0$ ist der Zusammenhang (die Interaktion) zwischen A und B nicht mehr von C abhängig und $\Theta_{i_1 i_2;j_1 j_2|k} = \Theta_{i_1 i_2;j_1 j_2}$, d.h. das Kreuzproduktverhältnis ist identisch mit dem entsprechenden Verhältnis der $A \times B$-Tafel. Diese entsteht durch *Aggregation* (d.h. durch Summation) über C.

Ein-Faktor-Effekte

Auch Haupteffekte können diskutiert werden. Dazu werden die bedingten Odds (Wettchancen) 1. Ordnung betrachtet:

$$\Theta_{i_1 i_2}(B_j \cap C_k) = \frac{p(A_{i1}|B_j \cap C_k)}{p(A_{i2}|B_j \cap C_k)} \tag{3.62}$$

und

$$\begin{aligned}\log \Theta_{i_1 i_2}(B_j \cap C_k) &= \mu^{A}_{i_1} - \mu^{A}_{i_2} + \mu^{AB}_{i_1 j} - \mu^{AB}_{i_2 j} \\ &\quad + \mu^{AC}_{i_1 k} - \mu^{AC}_{i_2 k} + \mu^{ABC}_{i_1 j k} - \mu^{ABC}_{i_2 j k}\end{aligned} \tag{3.63}$$

Hier wird die Wahrscheinlichkeit des Auftretens von A_{i_1} relativ zu der von A_{i_2} unter der Bedingung, dass die Kombination $(B_j \cap C_k)$ vorliegt, betrachtet. Verschwinden alle Interaktionsterme μ^{AB}, μ^{AC} und μ^{ABC}, so hängt dieses Verhältnis nur von der Differenz $\mu^{A}_{i_1} - \mu^{A}_{i_2}$ ab.

Beispiel 27 Zur Illustration werden das oben gegebene Beispiel zur Verhängung der Todesstrafe in den USA vervollständigt; es handelt sich um eine $2\times2\times2$-Tabelle. Radelet (1981) veröfffentlichte die folgenden Daten: Es kann angenommen werden,

Tabelle 3.8: Verhängung der Todestrafe in den USA

		Todesstrafe		
Angeklagte	Opfer	ja	nein	Anteil (ja)
weiß	weiß	19	132	.126
	schwarz	0	9	.000
schwarz	weiß	11	52	.175
	schwarz	6	97	.058

dass es sich um ein multinomiales Erhebungsschema handelt. Die Parameter des Modells müssen dann der Bedingung

$$n_{+++} = \exp(\mu + \mu_i^A + \mu_j^B + \mu_k^C + \mu_{ij}^{AB} + \mu_{ik}^{AC} + \mu_{jk}^{BC} + \mu_{ijk}^{ABC}) \tag{3.64}$$

genügen. Dieses Modell "erklärt" die Daten in *jedem* Fall; - es ist also "trivial". Die Frage ist, ob nicht ein einfacheres Modell die Daten ebenfalls erklärt.

Die erste Frage ist, welche der möglichen Modelle inhaltlich interessant sind. Die Verhängung der Todesstrafe kann als Variable, die von der Farbe des Opfers einerseits und von der Farbe des Täters andererseits abhängt, aufgefaßt werden. Es stehe A für die Todesstrafe, B für das Opfer und C für den Täter[3] Die Modelle werden der Reihe nach diskutiert:

1. *Das Modell (A, B, C)*: Dies ist das "Nullmodell", d.h. es wird keinerlei Abhängigkeit zwischen den Faktoren Hautfarbe des Täters, Hautfarbe des Opfers und Verhängung der Todesstrafe angenommen. Wie man der Tabelle 3.9 entnehmen kann, ist der G^2-Wert hochsignifikant, d.h. das Modell ist nicht mit den Daten verträglich.

Tabelle 3.9: Modelle für Abhängigkeiten zwischen Strafe und Hautfarbe

Modell	G^2	*df*	p
A, B, C	137.93	4	.000
A, BC	8.13	3	.043
AB, C	131.68	3	.000
AC, B	137.71	3	.000
AB, AC	131.46	2	.000
AB, BC	1.88	2	.390
AB, AC, BC	.701	1	.403
ABC	.000	0	1.000
A Bestrafung, B Opfer,C Täter			

2. *Das Modell (A, BC)*: Es wird angenommen, dass die Todesstrafe unabhängig von der Hautfarbe weder des Täters noch des Opfers verhängt wird; die Abhängigkeiten in der Tabelle können, dieser Hypothese entsprechend,

[3]Oder die Täterin, auf die Männlich-weiblich-Unterscheidung wird im Folgenden der Einfachheit wegen verzichtet.

durch Assoziationen zwischen Tätern und Opfern erklärt werden: Weiße töten überzufällig häufig Weiße, und/oder Schwarze überzufällig häufig Schwarze, und/oder Schwarze überzufällig häufig Weiße und/oder umgekehrt. Gemäß Tabelle 3.9 weicht auch dieses Modell signifikant von den Daten ab ($p = .043 < .05$).

3. *Das Modell (AB, c)*: Hier wird eine Beziehung zwischen der Farbe des Opfers und der Art der Bestrafung angenommen, die unabhängig von der Farbe des Täters ist. Der G^2-Wert is hochsignifikant, d.h. das Modell ist nicht mit den Daten kompatibel.

4. *Das Modell (AC, B)*: Hier wird eine Abhängigkeit zwischen der Hautfarbe des Täters und der Verhängung der Todesstrafe angenommen; die Farbe des Opfers spielt keine Rolle. Der G^2-Wert ist offenbar hochsignifikant, d.h. das Modell ist nicht mit den Daten kompatibel.

5. *Das Modell (AB, AC)*: Hier wird angenommen, dass es (i) eine Abhängigkeit zwischen der Verhängung der Todesstrafe und Farbe des Täters einerseits und (ii) zwischen Verhängung der Todesstrafe und Farbe des Opfers andererseits gibt. So kann die Tatsache, dass ein Täter schwarz ist, die Wahrscheinlichkeit des Todesurteils erhöhen, und unabhängig davon (d.h. unabhängig davon davon, dass der Täter weiß oder schwarz ist) von der Farbe des Opfers; die Richter können es als besonders verwerflich empfinden, dass ein Weißer umgebracht wurde, einen Schwarzen zu töten kann als läßliche Sünde gelten. Nach Tabelle 3.9 kann das Modell nicht akzeptiert werden.

6. *Das Modell (AB, BC)*: Hier wird eine Abhängigkeit zwischen Verhängung der Todesstrafe einerseits und der Farbe des Opfers (z.B. wird das Töten eines Weißen als besonders verwerflich betrachtet) andererseits postuliert, und darüber hinaus wird angenommen, dass es noch eine Assoziation zwischen Opfern und Tätern gibt. Nach Tabelle 3.9 ist dieses Modell mit den Daten verträglich, - es ist aber die Frage, ob es auch das *beste* Modell ist.

7. *Das Modell (AB, AC, BC)*: Hier werden paarweise Abhängigkeiten zwischen der Verhängung der Todesstrafe und (i) der Farbe des Opfers, (ii) der Farbe des Täters und schließlich (iii) zwischen Opfer und Täter angenommen. Dieses Modell ist sicher mit den Daten verträglich, aber wieder stellt sich die Frage, ob es das beste Modell ist, - schhließlich werden mehr Parameter geschätzt als bei dem Modell (AB, BC),was automatisch eine bessere Anpassung impliziert.

8. *Das Modell (ABC)*: Hier wird postuliert, dass die jeweilige Kombination von Opfer, Täter und Verhängung der Todesstrafe ganz spezifisch ist.

Betrachtet man alle Modelle, so kommt man zu dem Schluß, dass das Modell (AB, BC) die beste Beschreibung der Daten liefert: es kommt mit einem Parameter weniger als das Modell (AB, AC, BC) aus und erzeugt einen G^2-Wert, dessen

Wahrscheinlichkeit unter der Nullhypothese nur unwesentlich von dem des komplexeren Modells abweicht. Die Bestrafung hängt also im wesentlichen (i) von der Farbe des Opfers ab - es ist schlechter für den Täter, wenn er einen Weißen getötet hat,unabhängig davon, ob er selbst weiß oder schwarz ist - und (ii) von einer Assoziation zwischen Täter und Opfer - Weiße töten eher Weiße, und Schwarze eher Schwarze.

□

3.3.4 Aggregierbarkeit

Allgemeine Definition

Gegeben sei eine $I \times J \times K$-Tabelle, d.h. es werden drei Faktoren A, B und C betrachtet. Es mag der Übersichtlichkeit dienen, über die Stufen/Kategorien eines Faktors zu summieren (zu *aggregieren*), so dass eine 2-dimensionale Kontingenztabelle entsteht. Man spricht auch von einer *Marginaltabelle*. Betrachtet man nur die k-te Scheibe des 3-dimensionalen Datenkubus, so erhält man eine 2-dimensionale *Partialtabelle*. Die Partialtabellen enthalten gewissermaßen bedingte Häufigkeiten: sie reflektieren den Zusammenhang zwischen A und B für eine gegebene Kategorie von C. Partialtabellen können stark von den Marginaltabellen abweichen.

Es zeigt sich nun, dass die Schlußfolgerungen, die man anhand der aggregierten Daten zieht, falsch, zumindest aber irreführend sein können. Dieser Sachverhalt weist natürlich auch darauf hin, dass es irreführend sein kann, von vornherein nur eine 2-dimensionale Tabelle zu betrachten: andere Faktoren oder Variablen können dann konfundierend das Bild über die Zusammenhänge zwischen den betrachteten Faktoren stören.

In Beispiel 26 wurde die Tabelle 3.5 vorgestellt. Dieser Tabelle zufolge haben Weiße, entgegen einer häufig vorgebrachten Vermutung, eine *höhere* Chance, zum Tode verurteilt zu werden, als Schwarze. Die Tabelle ist eine Marginaltabelle von Tabelle 3.8; hier wurde über die Hautfarbe des Opfers summiert. Man kann ebenso über die Farbe der Täter aggregieren. Man erhält dann die Tabelle 3.10. Es ergibt

Tabelle 3.10: Verhängung der Todesstrafe in den USA; Aggreg. über Täter

	Todesstrafe		
Opfer	ja	nein	Σ
weiß	30	184	214
schwarz	6	106	112
Σ	36	290	326

sich der Odds-Ratio $\Theta = (30 \times 106)/(6 \times 184) = 2.88$. Addiert man den Wert .5 zu den Häufigkeiten (die Gesamttabelle enthält den Eintrag 0), so ergibt sich ein

Odds-Ratio von 2.71. Hier scheint eine starke Abhängigkeit auf: ob man zum Tode verurteilt wird oder nicht, hängt stark von der Farbe des- oder derjenigen ab, die oder den man umbringt. Das Kreuzproduktverhältnis ist hier das Verhältnis der Häufigkeiten "Opfer weiß und Todesstrafe" × "Opfer schwarz und keine Todesstrafe" und "Opfer schwarz und Todesstrafe" × "Opfer weiß und keine Todesstrafe". Der Wert von Θ legt also nahe, dass die Todesstrafe insbesondere dann verhängt wird, wenn eine weiße Person Opfer wird; die Farbe des Täters ist dann von untergeordneter Bedeutung. Ein Schwarzer wird also nicht deshalb zum Tode verurteilt, weil er schwarz ist, sondern weil er einen Weißen umgebracht hat, - und dies gilt (fast) ebenso für einen Weißen! Der Rassismus drückt sich gewissermaßen in der Haltung aus, dass es weniger schlecht ist, einen Schwarzen zu töten als einen Weißen.

Man kann auch über die Variable "Todesstrafe" aggregieren: Der Odds-Ratio

Tabelle 3.11: Todesstrafe in den USA; aggregiert über Strafe

	Opfer		
Täter	weiß	schwarz	Σ
weiß	151	9	160
schwarz	63	103	166
Σ	214	112	326

beträgt hier $\Theta = (151 \times 103)/(63 \times 9) = 27.43$. Dieser Wert weist auf einen extrem ausgeprägten Zusammenhang zwischen der Hautfarbe des Täters und des Opfers hin: Weiße töten Weiße und Schwarze töten Schwarze. Betrachtet man die Odds, dass ein Weißer einen Weißen tötet relativ zu dem Ereignis, dass er einen Schwarzen tötet, so erhält man

$$\Omega_w = \frac{\pi_{w|w}}{\pi_{s|w}} = \frac{\pi_{ww}/\pi_{w+}}{\pi_{sw}/\pi_{w+}} = \frac{151}{9} = 16.78$$

Die Chance, dass ein Weißer einen Weißen tötet, ist fast 17-mal größer als die, dass er einen Schwarzen tötet. Analog erhält man für den schwarzen Täter

$$\Omega_s = \frac{63}{103} = .61$$

Die Chance, dass ein Schwarzer einen Weißen tötet, ist nur etwas mehr als die Hälfte so groß wie die, einen Schwarzen zu töten. Da die Tötung eines Weißen mit größerer Wahrscheinlichkeit die Todesstrafe nach sich zieht, ist man als Weißer also gefährdeter, – einfach deshalb, weil man eben eher einen Weißen umbringt als einen Schwarzen.

Bis hierher sind nur Marginaltabellen betrachtet worden. Es sollen noch die *Partialtabellen* angegeben werden: bei diesen Tabellen wird die jeweils dritte Variable nicht *ignoriert* (indem über sie summiert wird), sondern sie wird *kontrolliert*,

indem jeweils ein Wert konstant gehalten wird. In der folgenden Tabelle sind Marginalassoziationen und ihre zugehörigen Partialassoziationen angegeben. A stehe wieder für die Strafe, B für das Opfer und C für den Täter. Die Spalte A-C bezieht sich auf die Marginaltabelle 3.5; bei dieser Tabelle wurde über die Farbe des Opfers (B) summiert, und der Odds-Ratio beträgt $\Theta = (19.5 \times 149.5)/(17.5 \times 141.5) = 1.177 \approx 1.18$: Der Wert .5 wurde zu den Werten der Tabelle 3.5 addiert, weil in der Tabelle 3.8 die Häufigkeit 0 vorkommt. Man kann nun die Partialassoziationen für den Zusammenhang zwischen Farbe des Täters und der Todesstrafe in Abhängigkeit von der Farbe des Opfers berechnen. Ist das Opfer weiß, so erhält man aus Tabelle 3.8 den Wert $\Theta = (19.5 \times 52.5)/(11.5 \times 132.5) = .672$, d.h. die Odds für Weiße, zum Tode verurteilt zu werden, ist .67-mal so hoch für Weiße als für Schwarze, wenn das Opfer weiß ist. Ist das Opfer aber schwarz, so ergibt sich der Wert $\Theta = (.5 \times 97.5)/(6.5 \times 9.5) = .789 \approx .79$. Die Odds, zum Tode verurteilt zu werden, sind .79-mal so hoch für Weiße als für Schwarze, wenn das Opfer schwarz ist. Man betrachte nun die Spalte $A - B$; hier wird der Zusammenhang zwischen der Farbe des Opfers (*V*ictim) und der Todesstrafe charakterisiert. Der Odds-Ratio beträgt 2.71, wie bei 3.10 angegeben. Man kann nun den Odds-Ratio für weiße und schwarze

Tabelle 3.12: Partialassoziationen: Strafe

		Variablen		
Assoziation		A-C	A-B	B-C
Marginal		1.18	2.71	25.99
Partial	weiß	.67	2.80	22.04
	schwarz	.79	3.29	25.90

Täter getrennt bestimmen; dies sind die korrespondierenden Partialassoziationen. Für weiße Täter erhält man den Wert $\Theta = (19.5 \times 9.5/(.5 \times 132.5) = 2.796 \approx 2.8$, und für schwarze Täter $\Theta = (11.5 \times 97.5)/(6.5 \times 52.5) = 3.285 \approx 3.29$. Schließlich kann man noch die Marginal- und Partialassoziationen für die Kombination $B - C$ berechnen: sie geben die Beziehungen zwischen Täter und Opfer an, vergl. Tabelle 3.11. Die Marginalassoziation beträgt (wenn .5 addiert wird), 25.99, und die Partialassoziationen für weiße bzw. schwarze Täter betragen 22.04 bzw. 25.90. Die Assoziation zwischen der Farbe des Opfers und der des Täters ist sehr stark, und weiter zeigen die Odds-Ratios an, dass die Todesstrafe leichter verhängt wird, wenn die Farbe des Opfers weiß ist. Daraus ergibt sich, dass die Marginalassoziation Täter-Todesstrafe eine größere Tendenz, Weiße zum Tode zu verurteilen als Schwarze, anzeigt, als die entsprechenden Partialassoziationen.

Simpsons Paradox und die Aggregierbarkeit von Tabellen

Zeigen die Marginal- und Partialassoziationen verschiedene Richtungen des Zusammenhanges an, spricht man von *Simpson's Paradox*. Dieses Paradox läßt sich

allgemein charakterisieren:

Definition 19 *Es seien U, V und W zufällige Ereignisse, und U^c, V^c und W^c seien die entsprechenden Komplementärereignisse. Es gelte die folgende Kombination von Aussagen:*

$$\{P(U|V) > P(U|V^c)\} \cap \begin{cases} \{P(U|V \cap W) < P(U|V^c \cap W)\} \\ \\ \{P(U|V \cap W^c) < P(U|V^c \cap W^c)\} \end{cases} \tag{3.65}$$

gilt. Dann erfüllen die Ereignisse U, V und W Simpsons Paradox.

Beispiel 28 Es sei U das Ereignis, dass die Todesstrafe verhängt wird, V das Ereignis, dass der Täter weiß ist, und W sei das Ereignis, dass das Opfer weiß ist. Die Häufigkeiten in Tabelle 3.8 genügen diesen Bedingungen für Simpsons Paradox.

Um die Bezeichnungen etwas suggestiver zu machen, werde $U = Td$, $V = A_w$, $V^c = A_s$, $W = O_w$ und $W^c = O_s$ geschrieben. Aus Tabelle 3.8 findet man

$$\begin{aligned} p(Td|A_w) &= \frac{19}{160} = .119, \quad p(Td|A_s) = \frac{17}{166} = .102 \\ p(Td|A_w \cap O_w) &= \frac{19}{151} = .126, \quad p(Td|A_s \cap O_w) = \frac{11}{63} = .174 \\ p(Td|A_w \cap O_s) &= \frac{0}{103} = 0, \quad p(Td|A_s \cap O_s) = \frac{6}{103} = .058 \end{aligned}$$

Offenbar gilt die der Aussage (3.65) entsprechende Aussage

$$\{P(Td|A_w) > P(Td|A_s)\} \cap \begin{cases} \{p(Td|A_w \cap O_w) < p(Td|A_s \cap O_w)\} \\ \\ \{p(Td|A_w \cap O_s) < p(Td|A_s \cap O_s)\} \end{cases} \tag{3.66}$$

Links findet man die "allgemeine" Aussage, dass die Todesstrafe bei weißen Angeklagten (A_w) häufiger verhängt wird als bei schwarzen (A_s). Rechts findet man dagegen die Aussage, dass weiße Angeklagte *weniger* häufig zum Tode verurteilt werden. Natürlich wird die Auflösung dieses Paradoxes gleich mitgeliefert, denn als Bedingung geht rechts nicht nur die Farbe des Täters, sondern auch die des Opfers ein, während bei der Aussage auf der linken Seite die Farbe des Opfers nicht explizit mit eingeht, weil über die Farbe des Opfers aggregiert wurde.

Tabelle 3.5 ist eine Marginaltabelle der Tabelle 3.8; es ist über die Farbe des Opfers aggregiert worden. Diese Aggregation suggeriert, dass Weiße häufiger als Schwarze zum Tode verurteilt werden, entgegen dem "Vorurteil", Farbiger würde häufiger zum Tode verurteilt als Weiße. Aber (3.66) zeigt eben, dass die "Evidenz" der Tabelle 3.5 nur scheinbar ist, denn auf der rechten Seite von (3.66) wird deutlich, dass Farbige eben *doch* häufiger zum Tode verurteilt werden als Weiße. Die Bildung der Marginaltabelle durch Aggregation über die Farbe des Opfers ist nicht zulässig, weil sie die Interaktionen zwischen den Faktoren A: Todesstrafe (ja-nein), B Farbe des Täters (weiß-schwarz) und C Farbe des Opfers (weiß-schwarz) nicht berücksichtigt. □

Diese Betrachtungen haben Implikationen für die Aggregierbarkeit von 3-dimensionalen Kontingenztabellen. Aggregiert man über eine Variable bzw. über einen Faktor, etwa C, so kann sich zwischen den beiden betrachteten Faktoren A und B ein Zusammenhang zeigen, der nicht *an sich* existiert, *sondern nur durch den Faktor C bedingt wird.* Die Situation ist ähnlich wie bei Korrelationen: die Korrelation zwischen Alkoholkonsum und der Entscheidung, Priester zu werden, bricht zusammen, wenn die dritte Variable "ökonomische Situation" herauspartialisiert. Es gilt dementsprechend der folgende

Satz 8 *Die Variable C ist bezüglich der Interaktion von A und B aggregierbar, wenn C bedingt unabhängig von A oder B ist. C ist in bezug auf den Haupteffekt von A oder B aggregierbar, wenn die Interaktion zwischen C und A oder zwischen C und B verschwindet.*

Beweis: Bishop, Fienberg und Holland (1975), p. 39 □

Hamerle und Tutz (1984) weisen darauf hin, dass die Bedingungen in Satz 8 nicht notwendig, aber hinreichend sind.

3.4 Die Beziehung zwischen logistischer Regression und log-linearen Modellen

In Tabelle 3.8 wurden Daten präsentiert, die Informationen über die Beziehungen zwischen der Hautfarbe (i) des Opfers, (ii) des Täters und (iii) der Verhängung der Todesstrafe enthalten. Die Daten können durch ein log-lineares Modell beschrieben werden. Andererseits kann die Variable "Todesstrafe" mit den Werten "ja" (verhängt) oder "nein" (nicht verhängt) als abhängige Variable, und die Farben von Opfer und Täter als unabhängige Variablen betrachtet werden, so dass man auch eine logistische Regression rechnen könnte. Es zeigt sich nun, dass eine Teilmenge der überhaupt möglichen log-linearen Modelle der logistischen Regression äquivalent sind.

Die Farbe des Täters werde mit A bezeichnet; $A = A_w$, wenn der Täter weiß ist, $A = A_s$, wenn er schwarz ist. B stehe für die Farbe des Opfers, $B = B_w$, wenn das Opfer weiß, und $B = B_s$, wenn das Opfer schwarz ist. C sei der Faktor "Todesstrafe": $C = C_j$, wenn sie verhängt wird, $C = C_n$, wenn sie nicht verhängt wird. Es sei p die Wahrscheinlichkeit (relative Häufigkeit), dass die Todesstrafe verhängt wird, und $1 - p$ dementsprechend die Wahrscheinlichkeit, dass sie nicht verhängt wird. Gemäß dem Ansatz der kategorialen Regression wird

$$\log \frac{p_{ij1}}{1 - p_{ij1}} = \log \frac{n_{ij1}}{n_{ij2}} = \beta_0 + \beta_i^A + \beta_j^B \tag{3.67}$$

betrachtet. Im log-linearen Ansatz kann das Modell (AB, AC, BC) diskutiert werden; hier treten also nicht nur die Beziehungen von A und B zu C auf, sondern es kommt noch ein Interaktionsterm AB hinzu.

Es läßt sich nun zeigen, dass dieses Modell das Regressionsmodell (3.67) impliziert. Das log-lineare Modell lautet

$$\begin{aligned}\log \frac{n_{ij1}}{n_{ij2}} &= \log n_{ij1} - \log n_{ij2} \\ &= (\mu + \mu_i^A + \mu_j^B + \mu_1^C + \mu_{ij}^{AB} + \mu_{i1}^{AC} + \mu_{j1}^{bc}) \\ &\quad -(\mu + \mu_i^A + \mu_j^B + \mu_2^C + \mu_{ij}^{AB} + \mu_{i2}^{AC} + \mu_{j2}^{bc}) \\ &= (\mu_1^C - \mu_2^C) + (\mu_{i1}^{AC} - \mu_{i2}^{AC}) + (\mu_{j1}^{BC} - \mu_{j2}^{BC})\end{aligned}$$

Nun muß die Bedingung, dass sich Effekte zu Null summieren, berücksichtigt werden; es gilt

$$\sum_k \mu_k^C = \sum_k \mu_{ik}^{AC} = \sum_k \mu_{jk}^{BC} = 0.$$

Daraus folgt

$$\mu_1^C = -\mu_2^C \tag{3.68}$$
$$\mu_{j1}^{AC} = -\mu_{i2}^{AC} \tag{3.69}$$
$$\mu_{j1}^{BC} = -\mu_{j2}^{BC}. \tag{3.70}$$

Daraus ergibt sich

$$\log \frac{n_{ij1}}{n_{ij2}} = 2\mu_1^C + 2\mu_{i1}^{AC} + 2\mu_{j1}^{BC} \tag{3.71}$$

und diese Gleichung entspricht (3.67), denn $2\mu_{i2}^{AC}$ ist der i-te Effekt von A auf das Logit von C, d.h. $\mu_{i1}^{AC} = \beta_i^A$, und $2\mu_{j1}^{BC}$ ist der j-te Effekt von B auf das Logit von C, d.h. $2\mu_{j1}^{BC} = \beta_j^B$. Zum Schluß erhält man noch $2\mu_1^C = \alpha$.

Die logistische oder kategoriale Regression (3.67) enthält keinen der Interaktionsterme μ_{ij}^{AB}, der im allgemeinen log-linearen Ansatz enthalten ist. Der Grund dafür ist, dass sich diese Terme im Logit $\log n_{ij1}/n_{ij2}$ herauskürzen.

Es ist aber nicht so, dass sich grundsätzlich alle Interaktionsterme zwischen den unabhängigen Variablen herauskürzen. Man kann z.B. eine 4-fach Klassifikation vorliegen haben mit den Faktoren A, B, C und D, wobei D eine binärre abhängige Variable ist. Betrachtet man nun das Logit

$$\log \frac{n_{ijk1}}{n_{ijk2}} = \alpha + \beta_k^A + \beta_i^B + \beta_j^C \tag{3.72}$$

und gleichzeitig die möglichen log-linearen Modelle, so findet man, dass das Modell

$$(ABC, AD, BD, CD)$$

diesem Ansatz entspricht; hier ist also die 3-fach Interaktion ABC enthalten. Man kann weiter den Regressionsansatz (3.72) um einen Term β_{ki}^{AB} erweitern, und dann entspricht das log-lineare Modell (ABC, ABD, CD) diesem Regressionsansatz.

Literaturverzeichnis

[1] Agresti, A.: Categorical Data Analysis. New York 1990

[2] Fahrmeir, L., Hamerle, A., Tutz, G. (Hrsg) Multivariate statistische Verfahren, Walter De Gruyter, Berlin 1996

[3] Fahrmeir, L., Tutz, G.: Multivariate statistical modelliung based on generalized linear models. Springer-Verlag, New York 1994

[4] Johnson, R.A., Wichern, D.W.: Applied multivariate statistical analysis. Prentice Hall, Upper Saddle River 2002

[5] Marx, B.D., Smith, E.P. (1990) Principal component analysis for generalized linear regression. *Biometrika*, 77 (1), 12 – 3

Index