Inhaltsverzeichnis

Kapitel 1

Das verallgemeinerte lineare Modell (GLM)

1.1 Daten:

Es ist y_i mit $i = 1, ..., N$ die abhängige Variable oder Zielvariable und $x_1, ..., x_p$ seien die erklärenden Variablen. In Matrixdarstellung ergibt sich also:

$$\underline{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}; \qquad \underline{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Np} \end{pmatrix} = \begin{pmatrix} \cdots & \underline{x}_1 & \cdots \\ \vdots & \ddots & \vdots \\ \cdots & \underline{x}_N & \cdots \end{pmatrix} \tag{1.1}$$

Zu Beachten ist: Die zweite Darstellung beinhaltet Zeilenvektoren und $\underline{x}_i$ ist die i-te Beobachtung.

1.1.1 Gruppierte / ungruppierte Daten:

Bis jetzt hatten $\underline{Y}$, $\underline{X}$ genau N Einheiten. Falls mehrere Zeilen der Designmatrix identisch sind (Varianzanalyse) kann man solche Zeilen zusammenfassen. In $\underline{X}$ stehen dann nur mehr verschiedene Zeilen, in der entsprechenden Zeile von $\underline{Y}$ steht dann die Summe der y-Werte mit dieser Merkmalskombination von x-Werten. Zusätzlich führt man einen Vektor mit den Anzahlen der Wiederholungen N_i (mit $\sum_{i=1}^{I} = N$) der entsprechenden Merkmalskombinationen ein.

Beispiel: gruppierte Daten:

$$\begin{pmatrix} N_1 \\ \vdots \\ N_I \end{pmatrix}; \qquad \begin{pmatrix} \sum_i y_{1i} \\ \vdots \\ \sum_i y_{Ii} \end{pmatrix}; \qquad \underline{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{I1} & \cdots & x_{Ip} \end{pmatrix} \tag{1.2}$$

Wir haben jetzt also I-Gruppen, vorher hatten wir N Einheiten.

1.2 Verteilungsannahmen:

Die Dichte von y_i ist aus einer einparametrischen Exponentialfamilie mit (natürlichem) Parameter Θ_i, einem Skalenparameter ϕ (von i unabhängig, vergleiche "Varianzhomogenität") und Gewichten w_i. Die Dichte von y_i ist also gegeben durch:

$$f(y_i|\Theta_i, \phi, w_i) = c(y_i, \phi, w_i)\ exp\ (\frac{\Theta_i y_i - b(\Theta_i)}{\phi}\ w_i) \tag{1.3}$$

Weiters gilt:

- ϕ ist unabhängig von i aber unbekannt (vgl. "Varianzhomogenität")
- Die Gewichte w_i sind:
  - $w_i = \frac{1}{N_i} \ldots$ bei gruppierten Daten
  - $w_i = 1 \ldots$ bei ungruppierten Daten
- Die Funktionen b und c bestimmen die jeweilige Exponentialfamilie.

1.3 Strukturannahme:

Es gilt folgendes: $\mu_i = E(y_i)$ ist eine Funktion des "linearen Prediktors" η. Daraus folgt:

$$\eta = \underline{X}\beta; \qquad \eta_i = \underline{x}_i\beta \tag{1.4}$$

wobei $\underline{x}_i$ die i-te Zeile von $\underline{X}$ ist.

1.4 Link-, Responsefunktion:

Es geht darum, einen Zusammenhang zwischen dem Erwartungswert für y, $E(y) = \mu$, und dem linearen Prediktor η herzustellen. Dies geschieht mit der Link- bzw. Responsefunktion. Es ist

- $\mu_i = h(\eta) \ldots h$ ist die "Responsefunktion"
- $\eta_i = g(\mu_i) \ldots g = h^{-1}$ ist die "Linkfunktion"

Weiters gelten folgende wichtige Zusammenhänge:

- $E(y_i) = \mu_i = b'(\Theta_i) = \frac{\vartheta b}{\vartheta \Theta}$
- $\Theta_i = \psi(\mu_i)$
- $Var(y_i) = \sigma^2; \qquad \sigma_i^2 = \frac{\phi \nu(\mu_i)}{w_i}; \qquad \nu(\mu_i) = b''(\Theta_i) = b''(\psi(\mu_i))$

Ein bestimmtes verallgemeinertes lineares Modell (GLM) ist durch den Typ der Exponentialfamilie, die Wahl der Link- bzw. Responsefunktion und die Definition und Auswahl der Regressoren vollständig bestimmt.

Beispiel: Normalverteilungsannahme

Sei $y_i \sim N(\mu_i, \sigma^2)$. Die Dichte ist dann gegeben als:

$$f(y_i|\mu_i, \sigma^2, 1) = \frac{1}{\sqrt{2\pi\sigma^2}}\ exp\ (-\frac{y_i^2 - 2\mu_i y_i + \mu_i^2}{2\sigma^2}) = \frac{1}{\sqrt{2\pi}\sigma}\ exp\ (\frac{-y_i^2}{2\sigma^2})\ exp\ (\frac{\mu_i y_i - \frac{\mu_i^2}{2}}{\sigma^2})$$

Nun muss geprüft werden, ob die oben genannten Zusammenhänge gelten:

- Voraussetzung: $\Theta_i = \mu_i; \qquad b(\Theta_i) = b(\mu_i) = \frac{\mu^2}{2}; \qquad w_i = 1$
- es gilt: $b'(\Theta_i) = b'(\mu_i) = \mu \Leftrightarrow E(y_i)$
- es gilt: $b''(\Theta_i) = \nu(\Theta_i) = 1$
- es gilt: $Var(y_i) = \frac{\phi \nu(\mu_i)}{w_i} = \frac{\sigma^2 1}{1} = \sigma^2 = Var(y_i)$

Beispiel: gruppierte binäre Zielvariable

Hier kann die Zielvariable y_i entweder den den Wert 0 oder den Wert 1 annehmen. Sei $y_i \sim B(N_i, p_i)$ mit $i = 1, ..., I$. Die Dichte ist dann gegeben als:

$$f(y_i|\Theta_i) = \begin{pmatrix} N_i \\ y_i \end{pmatrix} p_i^{y_i} \ (1-p_i)^{N_i - y_i} = \begin{pmatrix} N_i \\ y_i \end{pmatrix} exp \ (y_i \ ln \ (\frac{p_i}{1-p_i}) + N_i \ ln \ (1-p_i))$$

Es ergeben sich also folgende Parameterwerte: $\Theta_i = ln \ (\frac{p_i}{1-p_i}) = p_i(1 + exp \ (\Theta_i))$ und daraus folgt mit $exp \ (\Theta_i) = \dfrac{p_i}{1-p_i}$ und daraus schließlich:

$$p_i = \frac{exp \ (\Theta_i)}{1 + exp \ (\Theta_i)}; \qquad 1 - p_i = \frac{1}{1 + exp \ (\Theta_i)} \tag{1.5}$$

Die Funktion b ist gegeben durch: $-N_i \ ln \ (1-p_i) = N_i \ ln \ (\frac{1}{1-p_i}) = N_i \ ln \ (1 + exp \ (\Theta_i))$
Wieder müssen wir überprüfen, ob die Zusammenhänge gelten:

- es gilt: $b'(\Theta_i) = N_i \ \dfrac{exp \ (\Theta_i)}{1 + exp \ (\Theta_i)} = N_i \ p_i = E(y_i)$
- es gilt: $b''(\Theta_i) = N_i \ \dfrac{exp \ (\Theta_i) \ (1 + exp \ (\Theta_i)) - exp \ (\Theta_i)^2}{(1 + exp \ (\Theta_i))^2}$.
  Daraus ergibt sich: $N_i \ \dfrac{exp \ (\Theta_i)}{(1 + exp \ (\Theta_i))^2} = N_i \ p_i \ (1 - p_i) = Var(y_t)$
- als Parameterwerte ergeben sich außerdem: $\phi = 1$ und $w_i = 1$.

1.5 Wahl der Response-/Linkfunktion:

Die Wahl ist abhängig vom Typ der Zielvariablen und der speziellen Anwendung. Zu jeder Exponentialfamilie gibt es eine so genannte "kanonische" oder "natürliche Linkfunktion". Der natürliche/lineare Prediktor ist folgendermaßen definiert:

$$\Theta_i = \Theta(\mu_i) = \eta_i = x_i\beta \longrightarrow g(\mu_i) = \Theta(\mu_i) \tag{1.6}$$

Beispiel: Normalverteilungsannahme:

Aus der Annahme dass $y_i \sim N(\mu_i, \sigma^2)$ folgt für die Linkfunktion g: $\Theta_i = \mu_i = \Theta(\mu_i) \rightarrow g(\mu_i) = \mu_i$, also die "identische Funktion".

Beispiel: Binomialverteilungsannahme:

Sei $y_i \sim B^{sk}_{N_i,p_i}$ ("skalierte Binomialverteilung", dh: die Ausprägungen von y_i sind nicht $0, 1, 2, ..., N_i$ sondern $0, \frac{1}{N_i}, \frac{2}{N_i}, ...1$. Daraus folgt, dass die Werte von y_i alle zwischen 0 und 1 liegen.) Damit ergibt sich für den Erwartungswert und die Varianz von von y_i:

$$E(y_i) = p_i = \mu_i; \qquad Var(y_i) = \frac{1}{N_i} p_i(1 - p_i) \tag{1.7}$$

Damit ergibt sich für die Linkfunktion

$$\Theta = \Theta_i(p_i) = ln\ \frac{p_i}{1 - p_i} = \eta_i = x_i\beta \tag{1.8}$$

die sogenannte "Logit Funktion", die hier auch die "natürliche Linkfunktion" ist. Also $g(\mu_i) = ln\ (\frac{\mu_i}{1 - \mu_i})$.

1.6 Spezielle GLM:

1.6.1 Stetige Zielvariable:

Sei zb. $y_i \sim N(\mu_i, \sigma^2)$ mit der dazugehörigen natürlichen Linkfunktion. Daraus ergibt sich sofort das "klassische, lineare Modell" $\mu = \eta = X\beta$. Für manche Anwendungen kann ein nichtlinearer Ansatz $h(\eta) = \eta^2$ oder $h(\eta) = ln\ \eta$ sinnvoll sein (h ist hier auf jeden Fall eine nicht-lineare Funktion). Festzuhalten ist, dass auch solche nicht-linearen Ansätze mit GLMs behandelt werden können.

Beispiel:

Sei $y \sim \Gamma(\lambda, \mu)$ mit Dichte $f(y) = \frac{y^{\lambda-1}}{\Gamma(\lambda)\ \mu^\lambda}\ exp\ (-\frac{y}{\mu})$. Es ergibt sich für den Erwartungswert: $E(y) = \lambda\ \mu$. Dieser Ansatz ist etwa besonders geeignet für Regressionsanalysen mit einer nicht-negativen Zielvariable, wie sie etwa bei Lebensdaueranalysen oder bei Analysen von monetären Daten vorkommen. Ein wesentlicher Nachteil bei dieser Art der Modellierung ist jedoch folgender. Da der Erwartungswert von $\mu > 0$ ist, muss auch gelten dass $X\beta$ größer als Null ist. β ist daher beschränkt und eine einfache Residuenminimierung (wie bisher) führt nicht mehr zum Ziel. Statt dessen handelt es sich um eine "Optimierung mit Nebenbedingungen".

1.6.2 Binäre oder binomiale Zielvariable:

Für binäre Zielvariablen mit Ausprägungen 0 bzw. 1 gilt: $P(y = 1|x) = E(y|x) = p = \mu$. Gruppierte binäre Variable sind $B_{N,p}$ verteilt. Die Modelle für binäre und binomialverteilte Zufallsvariable werden durch ihre Response- beziehungsweise Linkfunktion spezifiziert. Die Linkfunktion stellt also den Zusammenhang zwischen $E(y_i)$ und dem linearen Prediktor $X\beta$ her.

- Lineares Wahrscheinlichkeitsmodell:
  Sei $g = h = id$ ("identische Funktion"). Das heißt: $p = g(\mu) = \eta = X\beta$ oder kurz $(p = X\beta)$. Da $p \in [0,1] \rightarrow X\beta \in [0,1]$. Dies führt zu einer "restringierten Optimierung", einer Optimierung mit Nebenbedingungen also, die sehr aufwändig sein kann. Andere Modelle vermeiden diesen Nachteil, indem sie $p = F(\eta)$ setzen, wobei F eine streng monoton wachsende Verteilungsfunktion ist. Es ist anzumerken, dass dadurch auch die "Invertierbarkeit" gegeben ist.

- Probit-Modell:
  Für die Funktion F wird die Verteilungsfunktion der $N(0,1)$ gewählt, also $F = \phi$. Daraus ergibt sich $p = \phi(\eta) = \phi(X\beta)$. Ein Nachteil ist allerdings, dass die numerische Auswertung der Verteilungsfunktion ϕ bei der Likelihood-Schätzung aufwendig sein kann.

- Logit-Modell:
  Für F wird die natürliche Linkfunktion / Logitfunktion $\eta = ln\ (\frac{p}{1-p})$ beziehungsweise die entsprechende Responsefunktion $p = h(\eta) = \frac{exp\ (\eta)}{1+exp\ (\eta)}$ verwendet. Die logistische Verteilungsfunktion geht für $\eta \rightarrow \pm\infty$ etwas langsamer gegen 0 bzw. 1 als die Verteilungsfunktion der $N(0,1)$. Probit- und Logitmodelle liefern aber ähnliche Schätzer, abgesehen von p-Werten, die nahe bei 0 oder 1 liegen.

- Komplementäres log-log Modell:
  Als Linkfunktion wählen wir die Verteilungsfunktion der Extremwertverteilung (Gumble-Verteilung) $g(p) = ln\ (-ln\ (1-p))$. Die Responsefunktion ergibt sich dann als $h(\eta) = 1 - exp\ (-exp\ (\eta))$. Festzuhalten ist, dass es sich im Gegensatz zu den Verteilungsfunktionen der Standardnormalverteilung beziehungsweise der Logitfunktion hier um eine asymetrische Verteilungsfunktion handelt, die für kleine p-Werte

ähnlich wie die logistische Verteilungsfunktion ist. Sie geht allerdings für große p-Werte schneller gegen 1.

1.6.3 Zähldaten als Zielvariable:

Wir betrachten Anzahlen von bestimmten Ereignissen in einem vorgegebenen Zeitraum beziehungsweise Häufigkeiten. Oft reicht die Normalverteilungsapproximation (vor allem wenn die Häufigkeiten groß sind) aus, bei seltenen Ereignissen ist aber die Poissonverteilung P_μ das passende Modell.

- Log-Lineares Poissonmodell:
  Es gilt $E(y) = \mu$. Daraus folgt die Wahl der natürlichen Linkfunktion $g(\mu) = ln\ \mu = \eta = X\beta$. Für die Prognosefunktion ergibt sich $\mu = exp\ (\eta)$. Es ergeben sich also log-lineare Modelle für Kontingenztafeln, falls alle Kovariablen kategorial sind.

- Lineares Poissonmodell:
  Wir setzen als Linkfunktion $g(\mu) = \mu = \eta = X\beta$. Dieses Modell verwendet man, wenn die Kovariablen additiv auf die Zielvariable y_i wirken. Der Nachteil ist wiederum dass $\mu > 0 \rightarrow X\beta > 0$ ist. Dies führt zu einer "restringierten Optimierung", einer Optimierung mit Nebenbedingungen also, die sehr aufwändig sein kann.

1.7 Schätzen und Testen:

In GLM verwendet man Likelihood-Schätzer. In der Vorlesung beschränken wir uns auf Maximum Likelihood Schätzer (ML-Schätzer). Folgende Annahmen werden getroffen:

- ϕ ... Skalierungsfaktor, der bekannt ist (für Bereichsschätzung)
- X ... Designmatrix, für die wir vollen Rang voraussetzen
- y_i ... stammt aus einer einparametrischen Exponentialfamilie mit Parameter Θ, Skalierungsparameter ϕ und Gewichten w_i.

Die Dichte ist also wie folgt definiert:

$$f(y_i|\Theta_i, \phi, w_i) = c(y_i, \phi, w_i)\ exp\ (\frac{\Theta_i y_i - b(\Theta_i)}{\phi}\ w_i) \tag{1.9}$$

Die Log-Likelihoodfunktion von y_i (einer Beobachtung) ergibt sich als

$$l_i(\Theta_i) = ln\ f(y_i|\Theta_i, \phi, w_i) = const + \frac{\Theta_i y_i - b(\Theta_i)}{\phi}\ w_i \tag{1.10}$$

wobei der konstante Faktor nicht nicht vom Parameter Θ_i abhängig ist. Nun kann man den Zusammenhang zwischen Θ_i und $E(y_i) = \mu_i$ einsetzen. Daraus folgt dann $l_i(\Theta_i)$.

Beispiel: Skalierte Binomialverteilung

Sei $y \sim B^{sk}_{N_i,p_i}$ mit $E(y_i) = p_i$. Daraus ergibt sich die Log-Likelihoodfunktion als:

$$l_i(\mu_i) = l_i(p_i) = N_i\ (y_i\ ln\ p_i + (1 - y_i)\ ln\ (1 - p_i)) + const \tag{1.11}$$

Setzt man nun die Erwartungswertfunktion (=Responsefunktion) $\mu_i = h(\eta) = h(x_i\beta)$ in die Log-Likelihoodfunktion ein, erhalten wir l_i als Funktion von β. Die Log-Likelihoodfunktion der i-ten Beobachtung ist daher gegeben als:

$$l_i(\beta) = l_i(h(\eta_i)) = l_i(h(x_i\beta)) \tag{1.12}$$

Sind die Beobachtungen unabhängig, dann ist die Log-Likelihoodfunktion der Stichprobe die Summe der Log-Likelihoodfunktionen der einzelnen Beobachtungen. Es gilt dann

$$l(\beta) = \sum_{i=1}^{n} l_i(h(\eta_i)) = l_i(h(x_i\beta)) \tag{1.13}$$

wobei β hier ein Vektor, bestehend aus p-Komponenten, ist.

1.7.1 Scorefunction:

Die so genannte "Scorefunktion" $s(\beta)$ ist folgendermaßen definiert

$$s(\beta) = \frac{\vartheta l}{\vartheta \beta} = \sum_{i=1}^{N} \frac{\vartheta l_i}{\vartheta \beta_i} = \sum_{i=1}^{n} s_i(\beta) \tag{1.14}$$

wobei $\frac{\vartheta l}{\vartheta \beta}$ ein Vektor, bestehend aus lauter einzelnen Ableitungen, ist.

Die "individuellen Scorefunktionen" können folgendermaßen dargestellt werden:

$$s_i(\beta) = \frac{x_i^T D_i(\beta)}{\sigma_i^2(\beta)} \ (y_i - \mu_i(\beta)) \tag{1.15}$$

Dabei ist:

- $\mu_i(\beta) = h(x_i\beta) = E(y_i) \ldots$ der Erwartungswert von y_i
- $\sigma_i^2(\beta) = Var(y_i) = \dfrac{\phi\nu(\mu_i)}{w_i} = \dfrac{\phi\nu(h(x_i\beta))}{w_i} \ldots$ die Varianz von y_i
- $D_i(\beta) = \dfrac{\vartheta h}{\vartheta \eta_i}(x_i\beta) \ldots$ 1. Ableitung von h nach η_i an der Stelle $x_i\beta$.

1.7.2 Fisher-Informationsmatrix:

Die erwartete Informationsmatrix (Fisher-Informationsmatrix) ist definiert als:

$$E(-\frac{\vartheta \ l}{\vartheta\beta \ \vartheta^T\beta}) = E(s(\beta)s^T(\beta)) = Kov(s(\beta)) = F(\beta) \tag{1.16}$$

Sind die Beobachtungen unabhängig, dann gilt:

$$F(\beta) = \sum_{i=1}^{n} \frac{x_i x_i^T D_i^2(\beta)}{\sigma_i^2(\beta)} \tag{1.17}$$

Verwendet man die "natürliche Linkfunktion" (Responsefunktion) vereinfachen sich die Formeln. Es gilt dann:

$$s(\beta) = \frac{1}{\phi} \sum_{i=1}^{N} w_i x_i^T \ (y_i - \mu_i(\beta)); \qquad w_i = \frac{D_i^2(\beta)}{\sigma_i^2(\beta)} \tag{1.18}$$

$$F(\beta) = \frac{1}{\phi} \sum_{i=1}^{N} w_i \nu(\mu_i(\beta)) \ x_i^T x_i; \qquad w_i = \frac{D_i^2(\beta)}{\sigma_i^2(\beta)} \tag{1.19}$$

Schreibt man diese Formen in Matrixschreibweise, so ergeben sich kompaktere Formeln:

$$\underline{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}; \qquad \underline{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Np} \end{pmatrix}; \qquad \mu(\beta) = \begin{pmatrix} \mu_1(\beta) \\ \vdots \\ \mu_N(\beta) \end{pmatrix} \tag{1.20}$$

$$\Sigma(\beta) = diag\ (\sigma_1^2(\beta), ..., \sigma_N^2(\beta))$$
$$W(\beta) = diag\ (w_1^2(\beta), ..., w_N^2(\beta))$$
$$D(\beta) = diag\ (D_1^2(\beta), ..., D_N^2(\beta)) \tag{1.21}$$

Daraus ergeben sich dann folgende Formeln:

$$s(\beta) = X^T D(\beta)\Sigma(\beta)^{-1}\ (Y - \mu(\beta)); \qquad F(\beta) = X^T W(\beta) X \tag{1.22}$$

Beispiel: "Normalverteilungsannahme"

Sei $y \sim N(\mu_i, \sigma^2)$. Wir verwenden die natürliche Linkfunktion $g(\mu) = \mu = \eta = X\beta$. Für die Log-Likelihoodfunktion der Einzelbeobachtung folgt:

$$l_i(\Theta_i) = -ln\ \sqrt{2\pi\sigma^2} - \frac{y_i^2 - 2y_i\mu_i + \mu_i^2}{2\sigma^2} \tag{1.23}$$

Nun setzen wir den Zusammenhang zwischen Θ_i und $E(y_i)$ ein. Das erübrigt sich allerdings hier, da zufällig Θ_i gleich dem Erwartungswert von y_i ist. Nun setzen wir die Responsefunktion $h(x_i\beta)$ ein wobei gilt $h(x_i\beta) = x_i\beta = \mu_i$. Für die Likelihoodfunktion der Einzelbeobachtung ergibt sich nun:

$$l_i(\beta) = -ln\ \sqrt{2\pi\sigma^2} - \frac{y_i^2 - 2y_i x_i\beta + (x_i\beta)^2}{2\sigma^2} \tag{1.24}$$

Für die gesamte Likelihoodfunktion, die es zu maximieren gibt, folgt:

$$l(\beta) = \sum_{i=1}^{N} l_i(\beta) = -N\ ln\ \sqrt{2\pi\sigma^2} - \sum_{i=1}^{N} \frac{y_i^2 - 2y_i x_i\beta + (x_i\beta)^2}{2\sigma^2} \tag{1.25}$$

β ist ein Vektor, daher muss die Funktion etwas anders angeschrieben werden, um später leichter nach β ableiten zu können. Denn in Matrixschreibweise ist $f(a) = a^T x \rightarrow \frac{\vartheta f}{\vartheta a} = x$. Setzt man nun $(x_i\beta)^2 = \beta^T x_i^T x_i \beta$ und $-2y_i x_i\beta = -2\beta^T x_i^T y_i$ und bildet man die 1. Ableitung nach β so ergibt sich:

$$\frac{\vartheta l}{\vartheta \beta} = \sum_{i=1}^{N} \frac{y_i x_i^T - 2x_i^T x_i\beta}{\sigma^2} \tag{1.26}$$

Da die Matrix X aus den einzelnen Zeilenvektoren $(\underline{x}_1, \ldots, \underline{x}_N)$ besteht ergibt sich in Matrixschreibweise:

$$\frac{1}{\sigma^2} \left(\sum_{i=1}^{N} y_i x_i^T - \sum_{i=1}^{N} x_i^T x_i \beta\right) = \frac{1}{\sigma^2} \left(X^T y - X^T X \beta\right) = s(\beta) \tag{1.27}$$

Ausserdem gilt natürlich $s_i(\beta) = \dfrac{x_i^T y_i - x_i^T x_i \beta}{\sigma^2}$. Die Bedingung für einen ML-Schätzer ist, dass die erste Ableitung (= Scorefunktion) Null gesetzt wird. Daraus folgt:

$$s(\beta) = 0 \longrightarrow \hat{\beta} = (X^T X)^{-1} X^T y \tag{1.28}$$

Nun überprüfung wir die Formeln. Es muss

$$s_i(\beta) = \frac{x_i^T D_i(\beta)}{\sigma_i^2(\beta)} \left(y_i - \mu_i(\beta)\right) \tag{1.29}$$

gelten, wobei

- $\mu_i(\beta) = h(x_i\beta) = h(\eta_i) = x_i\beta \ldots$ die identische Funktion
- $D_i(\beta) = \dfrac{\vartheta h}{\vartheta \eta_i} = 1$
- $\sigma_i^2(\beta) = \dfrac{\phi \nu(h(x_i\beta))}{w_i} = \dfrac{\sigma^2 1}{1} = \sigma^2$

ist. Daraus folgt für die individuelle Scorefunktion:

$$s_i(\beta) = \frac{x_i^T 1}{\sigma^2} \left(y_i - x_i\beta\right) = \frac{1}{\sigma^2}(x_i^T y_i - x_i x_i^T \beta) = s_i(\beta) \tag{1.30}$$

Für die Fisher-Informationsmatrix ergibt sich:

$$F(\beta) = Kov(s(\beta)) = \sum_{i=1}^{N} \frac{x_i^T x_i D_i(\beta)^2}{\sigma_i^2} = \frac{1}{\sigma^2} \sum_{i=1}^{N} x_i^T x_i = \frac{1}{\sigma^2} X^T X \tag{1.31}$$

Verwendet man die natürliche Linkfunktion, so vereinfachen sich die Formeln. Es gilt dann

$$s(\beta) = \frac{1}{\phi} \sum_{i=1}^{N} w_i x_i^T (y_i - \mu_i(\beta)) = \frac{1}{\sigma^2} \sum_{i=1}^{N} 1 x_i^T y_i - x_i^T x_i \beta = \frac{1}{\sigma^2} \left(X^T y - X^T X \beta\right) \tag{1.32}$$

sowie

$$F(\beta) = \frac{1}{\phi} \sum_{i=1}^{N} w_i \nu(\mu_i(\beta)) x_i^T x_i = \frac{1}{\sigma^2} (X^T X) \tag{1.33}$$

In Matrixschreibweise ergibt sich $s(\beta) = X^T D(\beta) \Sigma^{-1}(\beta)(Y - \mu(y))$ mit:

$$D(\beta) = I; \qquad \Sigma(\beta) = diag\ (\sigma_i^2); \qquad \mu(y) = X\beta \tag{1.34}$$

Daraus ergibt sich für die Scorefunktion $s(\beta) = \frac{1}{\sigma^2} X^T (Y - X\beta)$ und für die Fisher-Informationsmatrix $F(\beta) = X^T W(\beta) X$ mit $W(\beta) = diag\ (w_i(\beta) = \frac{1}{\sigma^2}\ I)$, woraus folgt: $F(\beta) = \frac{1}{\sigma^2} X^T X$

1.7.3 Numerische Berechnung des ML-Schätzers:

Gesucht ist eine Schätzer $\hat{\beta}$, das Maximum der Likelihoodfunktion, d.h die Nullstelle der Scorefunktion $s(\beta)$. Die Bestimmung des Maximums erfolgt iterativ, z.B mit dem Fisher-Scoring (=verallgemeinertes Newton-Verfahren). Ausgehend von einem Startwert β_0, der hinreichend nahe beim Maximum der Likelihoodfunktion sein soll, sind die Iterationen des Verfahrens folgendermaßen definiert

$$\hat{\beta}_{k+1} = \hat{\beta}_k + F^{-1}(\hat{\beta}_k) s(\hat{\beta}_k) \tag{1.35}$$

wobei $F^{-1}(\hat{\beta}_k)$ als die Jacobi Matrix von $s(\hat{\beta}_k)$ bezeichnet wird. Die Iteration wird durchgeführt bis ein Abbruchkriterium (z.B: die relative Änderung)

$$\frac{||\hat{\beta}_{k+1} - \hat{\beta}_k||}{||\hat{\beta}_k||} \leq \epsilon \tag{1.36}$$

erfüllt ist. ϵ entspricht etwa der Rechengenauigkeit des Computers oder auch einer vorgegebenen Zeitspanne.

Die nach heutigem Standard effizientesten Methoden zur Lösung von Maximierungsaufgaben sind die sogenannten Quasi-Newton Verfahren (darunter optimal die so genannte BFGS-Methode mit Liniensuche). Konvergenz tritt hier schon in $m \leq p$ Schritten ein, wobei p die Anzahl der Variablen im Optimierungsproblem bezeichnet.

Existenz und Eindeutigkeit von $\hat{\beta}_{ML}$:

Beim klassischen linearen Modell existiert ein eindeutiges $\hat{\beta}$ unter der Voraussetzung dass X vollen Rang besitzt. Beim GLM sind die Fragen der Existenz und Eindeutigkeit nicht allgemein beantwortbar. Falls jedoch X vollen Rang hat, ist die ML-Schätzer für die natürliche Linkfunktion eindeutig.

Asymptotische Eigenschaften von $\hat{\beta}_{ML}$:

Sei $\hat{\beta}_{ML}$ der ML-Schätzer gewonnen aus einer Stichprobe von Umfang N. Asymptotische Aussagen gelten für $N \longrightarrow \infty$, also müssen nicht die Messwiederholungen an jeder Messstelle gegen ∞ gehen.

Satz:

$\hat{\beta}_N$ existiert asymptotisch, ist schwach konsistent und asymptotisch normalverteilt. Es gilt also:

- $lim_{N\longrightarrow\infty} P(\hat{\beta}_N \; existiert) = 1 \ldots$ asymptotische Existenz
- $lim_{N\longrightarrow\infty} P(||\hat{\beta}_N - \beta|| > \epsilon) = 0 \ldots$ schwache Konvergenz
- $\hat{\beta}_N \sim N(\beta, F_N^{-1}(\beta)) \ldots$ asymptotische Normalität
- $F_n^{\frac{1}{2}}(\beta)\hat{\beta}_N - \beta \longrightarrow N(0, I) \ldots$ asymptotische Normalität

Wiederholung: starke/schwache Konsistenz

Starke Konvergenz bei stochastischen Regressoren ist gegeben, wenn gilt: $lim_{N\longrightarrow\infty} P(\hat{\beta}_N = \beta) = 1$. Sei nun $\hat{\beta}_N$ mit $N = 1, 2, ...,$ eine Folge von Zufallsvariablen mit Verteilungsfunktion $F_{\hat{\beta}_N}$ und $\beta \sim F(\beta)$. Falls die Folge $F_{\hat{\beta}_N}$ mit $(N = 1, 2, 3, ...)$ an jeder Stetigkeitsstelle von F_{β_N} gegen F_{β_N} konvergiert, dann gilt: $\hat{\beta} \longrightarrow \beta$ und $F_{\hat{\beta}_N} \longrightarrow F_{\beta_N}$, also die starke Konvergenz.

Dieser Satz gilt jedoch nur unter gewissen Regularitätsbedingungen, auf die hier nicht eigegangen wird. Einige wichtige Voraussetzungen sind aber, dass y_i, x_i iid verteilt sind und dass $y_i|x_i$ die Annahmen des GLM erfüllt.

1.8 Testen linearer Hypothesen:

Es gilt

$$H_0 : \ C\beta = \xi; \qquad H_1 : \ C\beta \neq \xi \tag{1.37}$$

wobei C eine (r, p)-Matrix ist, und $r \leq p$ und $Rg\ C = r$. Zum Testen können nun die Likelihood-Quotienten Statistik lq, die Score-Statistik u oder die Wald-Statistik w verwendet werden.

1.8.1 lq-Teststatistik:

Die Teststatistik ist gegeben als

$$lq = -2(l(\hat{\beta}) - l(\tilde{\beta})) \tag{1.38}$$

wobei $l(\hat{\beta})$ der Wert der Log-Likelihoodfunktion der Stichprobe an der Stelle des ML-Schätzers $\hat{\beta}_{ML}$ ist und $l(\tilde{\beta})$ der Wert der Log-Likelihoodfunktion der Stichprobe an der Stelle des restringierten Maximums (Nebenbedingung H_0) der Log-Likelihood-Funktion. Die lq-Teststatistik ist asymptotisch verteilt nach χ^2 mit r Freiheitsgraden.

Spezialfall:

Sei

$$H_0 : \ \beta_r = 0; \qquad H_1 : \ \beta_r \neq 0 \tag{1.39}$$

wobei β_r ein Teilvektor des Parametervektors β ist. Um die Teststatistik auszurechnen benötigen wir die Werte $\hat{\beta}_{ML}$, den ML-Schätzer des unrestringierten Modells und $\tilde{\beta}$, der ja der ML-Schätzer des entsprechenden Submodells unseres GLM ist (vgl. klassische Regressions- und Varianzanalyse).

1.8.2 Wald-Statistik:

$\hat{\beta}$ ist asymptotisch $\sim N(\beta, F_N^{-1}(\beta)) \rightarrow C\hat{\beta}$ asymptotisch $\sim N(C\beta, CF_N^{-1}(\beta)C^T)$, wobei $CF_N^{-1}(\beta)C^T$ eine asymptotische Kovarianzmatrix ist. Dabei gilt: C ist eine (r, p)-Matrix

mit $p =$ Anzahl der Parameter und $Rg\ C = r$ und $r \leq p$.

Stimmt nun $H_0 :\ C\beta = \xi$ dann ist $C\beta = \xi$ und als Teststatistik ergibt sich:

$$w = (C\hat{\beta} - \xi)\ (CF_N^{-1}(\beta)C^T)\ (C\hat{\beta} - \xi)^T \tag{1.40}$$

Auch die Wald-Teststatistik ist asymptotisch verteilt nach χ^2 mit r Freiheitsgraden.

1.8.3 Score-Statistik:

Die Score-Teststatistik ist gegeben durch

$$u = s(\tilde{\beta})^T\ F_N^{-1}(\beta)\ s(\tilde{\beta}) \tag{1.41}$$

wobei $\tilde{\beta}$ den ML-Schätzer unter der Berücksichtigung der Nullhypothese $H_0 :\ C\beta = \xi$ bezeichnet.

Satz: Asymptotische Äquivalenz der Teststatistiken:

Unter der Voraussetzung, dass $\hat{\beta}$ asymptotisch normalverteilt ist und dass die Nullhypothese $H_0 :\ C\beta = \xi$ stimmt, sind die 3 Teststatistiken lq, w und u asymptotisch äquivalent und approximativ verteilt nach χ^2 mit r Freiheitsgraden. Für kleine Stichprobenumfänge ($N < 50$) können sich die Werte der drei Teststatistiken allerdings deutlich voneinander unterscheiden.

1.9 Variablenselektion/Modellsuche:

Die Modellsuche verläuft ähnlich wie beim linearen Regressionsmodell. Man verwendet die "schrittweise Vorwärtsselektion" und die "Rückwärtselimination". Diese Verfahren sind numerisch effizient für den Score- und Waldtest. Zur Vorwärtsselektion wird der Scoretest verwendet, für die Rückwärtselimination der Waldtest zum Testen von einem Submodell gegen das aktuelle Modell.

1.10 Beurteilung der Anpassungsgüte des Modells:

Falls y stetig ist, wird wie beim linearen Regressionsmodell, das Gütemaß R^2 verwendet. Für diskretes (insbesonders kategoriales) y wird χ^2 oder die Devianz D verwendet. Voraussetzung für die Gültigkeit der gebräuchlichsten "Goodness-of-Fit"-Statistiken ist, dass die Daten gruppiert sind und $N_i,\ i = 1, 2, 3...$ sollte in allen Gruppen hinreichend gross sein.

- Pearson-Statistik:

  Die Pearson-Teststatistik ist definiert als

  $$\chi^2 = \sum_{i=1}^{I} \frac{y_i - \hat{\mu}_i}{\nu(\hat{\mu}_i) w_i} \tag{1.42}$$

  wobei $\hat{\mu}_t = h(x_i\hat{\beta})$ der geschätzte Erwartungswert von y_i ist und $\nu(\hat{\mu}_i) = \nu(h(x_i\hat{\beta}))$ die geschätzte Varianz von y_i ist.

- Devianz: Die Devianz ist definiert als

  $$D = -2\phi \sum_{i=1}^{I} (l_i(\hat{\mu}_i) - l_i(y_i)) \tag{1.43}$$

  mit $l_i(\hat{\mu}_i) = l_i(h(x_i\hat{\beta}))$ und $l_i(y_i)$ ist die Log-Likelihoodfunktion der i-ten Beobachtung mit der Beobachtung y_i anstelle von μ_i.

Die Devianz D und χ^2 sind approximativ verteilt nach χ^2_{I-p}, wobei I die Anzahl der Gruppen und p die Anzahl der zu schätzenden Parameter $(\beta_1, ..., \beta_p)$ ist. Große Werte für die Teststatistiken deuten auf ein schlecht angepasstes Modell (H_0 : die Modelle sind adequat) hin.

1.11 Mehrkategorielle Regressionsmodelle:

Wir haben bereits Modelle für univariate binäre Zielvariable kennen gelernt (Probit- / Logit- / komplementäres log-log Modell). Die Erweiterung auf mehrkategoriale Modelle ergibt im ersten Schritt ein spezielles multivariates GLM, im zweiten Schritt das allgemeine multivariate GLM.

Beispiel: Absatz von Tankstellen"

- $y \ldots$ ist die abhängige Variable "Absatz" mit den Kategorien "niedrig", "mittel" und "hoch"
- $x_1 \ldots$ ist die erklärende Variable "Ortsgröße" in den Kategorien "klein" und "gross"
- $x_2 \ldots$ ist die erklärende Variable "Angebotsform" mit den Kategorien "Selbstbedienung" und "Bedienung"

Es ist zu beachten, dass y jetzt mehrdimensional ist. Der Grund ist, da y auf folgende Art und Weise umparametrisiert werden kann.

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

wobei jetzt für y_1 der Wert 1 eingesetzt wird, wenn y_1 niedrig ist und Null, wenn y_1 nicht niedrig ist. Analog kann für y_2 der Wert 1 eingesetzt werden, wenn der Absatz mittel war und 0 sonst. Auch die dritte Ausprägungskategorie kann mit 0 oder 1 kodiert werden, je nachdem ob der Absatz hoch war oder nicht.

1.11.1 Allgemeines:

Allgemein ist zu sagen, dass die abhängige Variable Y kategorial mit insgesamt $R = q+1$ Kategorien ist.

1.11.2 Ziel:

Ziel der kategorialen Regression ist es, den Einfluss von $(x_1, ..., x_p)$ auf Y beziehungsweise den Vektor der Auftrittswahrscheinlichkeiten $p = (p_1, ..., p_q)^T$ zu erklären. Dabei gilt, dass $p_r = P(Y = r|x)$ ist, wobei $r = 1, ..., q$ und $p_R = 1 - \sum_{r=1}^{q} p_r$.

Die "letzte" Kategorie (R-te Ausprägung von Y) wird als "Referenzkategorie" bezeichnet. Für die i-te Beobachtung gilt $p_i = (p_{i1}, ..., p_{iq})^T$ mit $p_{ir} = P(y_i = r|x_i)$. Man muß aber aufpassen, denn Y ist nicht eindimensional, was eventuell durch die Zuweisung von Nummern für die einzelnen Kategorien suggeriert wird. Das wird sofort klar, wenn man Y in der Form einer multinomial-verteilten Zufallsvariable anschreibt. Es ist dann $y_i = (y_{i1}, ..., y_{iq})^T$ mit

$q = R - 1$ wobei y_{ir} den Wert 1 besitzt, wenn $y_{ir} = r$ und 0 sonst. Für die Referenzkategorie R gilt: $y_{iR} = 1 - \sum_{r=1}^{q} y_{ir}$.

Offensichtlich gilt, dass $y_i \sim B_{1,p_i}$. Es handelt sich also um eine Multinomialverteilung, wobei p_i den Vektor der Auftrittswahrscheinlichkeiten bezeichnet.

1.11.3 Gruppierte / ungruppierte Daten:

Seien die Daten gegeben als:

$$\begin{aligned} y_1^T, ..., y_N^T; \qquad & y_i^T = (y_{i1}, ..., y_{iq})^T \\ x = (x_1, ..., x_N); \qquad & x_i = (x_{i1}, ..., x_{ip}) \end{aligned} \tag{1.44}$$

Faßt man nun Beobachtungen mit gleichem Vektor x_i zusammen so ergeben sich insgesamt I Gruppen mit jeweils N_i Beobachtungen. Nun gibt es zwei Möglichkeiten für die Darstellung der unabhängigen Variable Y.

- Summe:
  Es gilt $y_i = \sum_{j=1}^{N_i} y_{ij}$ wobei y_{ij} jene y-Vektoren sind, die zum gleichen Beobachtungsvektor x_i gehören. Daraus ergibt sich, dass $y_i \sim B_{N_i,p_i}$ mit $p_i = P(y_i = r|x_i)$. Nachdem der Beobachtungsvektor x_i hier immer gleich ist, kann darauf kann das Additionstheorem der Multinomialverteilung angewandt werden.

- Mittelwert:
  Es gilt $\overline{y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}$ wobei y_{ij} jene y-Vektoren sind, die zum gleichen Beobachtungsvektor x_i gehören. Daraus ergibt sich nun dass $\overline{y}_i \sim B^{sk}_{N_i,p_i}$, der Mittelwert ist also nach der skalierten Multinomialverteilung verteilt. Jede Komponente der q-dimensionalen Zuvallsvariablen Y besitzt also statt den möglichen Ausprägungen von $(0, 1, 2, ..., N_i)$ die normierten / skalierten Ausprägungen $(0, \frac{1}{N_i}, \frac{2}{N_i}, ..., 1)$. Diese Variante ist für das Logit-Modell günstiger.

Damit würden sich für das Beispiel von vorher folgende endgültige Daten ergeben:

- Gruppe 1: N_1 Beobachtungen $\longrightarrow \overline{y}_1^T = (\overline{y}_{11}, ..., \overline{y}_{1q})$ und $x_1 = (x_{11}, ..., x_{1p})$
- Gruppe I: N_I Beobachtungen $\longrightarrow \overline{y}_I^T = (\overline{y}_{I1}, ..., \overline{y}_{Iq})$ und $x_I = (x_{I1}, ..., x_{Ip})$

1.11.4 Das mehrkategoriale Modell:

Die Responsekategorien für die abhängigen Variablen $(1, ..., R)$ müssen keine Ordnungsstruktur aufweisen. Das heißt, die Responsevariable Y besitzt nur Nominalskalenniveau. Unterschiede zum Logit-Modell für binäre Zufallsvariable ergeben sich nur dadurch, dass Y jetzt q-dimensional ist.

1.11.5 Erwartung:

Im ungruppierten und im gruppierten Fall mit Mittelwerten sind die Erwartungswerte gleich. Es gilt

$$E(y_i|x_i) = p_i = (p_{i1}, ..., p_{iq})^T \tag{1.45}$$

Bei gruppierten binären Zielvariablen mit logistischer Responsefunktion (hier ist das gleich der natürlichen Responsefunktion) ergab sich im eindimensionalen Fall folgender Zusammenhang:

$$p_i = \frac{exp\ (\Theta_i)}{1 + exp\ (\Theta_i)} = \frac{exp\ (x_i\beta)}{1 + exp\ (x_i\beta)} \tag{1.46}$$

Ein analoger Zusammenhang gilt auch das nominale nichtkategoriale Logit-Modell

$$p_{ir} = \frac{exp\ (x_i\beta_r)}{1 + \sum_{s=1}^{r} exp\ (x_i\beta_s)} \tag{1.47}$$

mit $r = 1, ..., q$. Für die Referenzkategorie R gilt:

$$p_{iR} = 1 - \sum_{r=1}^{q} p_{ir} = \frac{1}{1 + \sum_{s=1}^{r} exp\ (x_i\beta_s)} \tag{1.48}$$

Das binäre Logit-Modell ergibt sich als Spezialfall des allgemeinen Falles.

1.11.6 Alternative Darstellungsmöglichkeit:

Wir definieren $ln\ \frac{p_{ir}}{p_{iR}} = x_i\beta_r$ mit $r = 1, ..., q$. Das heißt, dass der lineare Prediktor $x_i\beta_r$ die logarithmierten Chancen (= "Wahrscheinlichkeiten" oder "Logits") zwischen den Responsekategorie r und der Referenzkategorie R bestimmt. Beim univariaten binären Logit-Modell wären das die Chancen zwischen Auftreten und Nicht-Auftreten des Ereignisses.

Kapitel 2

Das multivariate GLM:

Das multivariate GLM ist eine direkte Verallgemeinerung des univariaten linearen Modells. Benötigt wird wieder eine Verteilungsannahme sowie eine Strukturannahme.

2.1 Verteilungsannahme:

Die Dichte von y_i gehört zu einer q-parametrischen Exponentialfamilie mit natürlichem Parameter $\Theta_i = (\Theta_{i1}, ..., \Theta_{i_q})^T$, Skalenparameter ϕ und Gewichten w_i. Für die Dichte ergibt sich daher

$$f(y_i|\Theta_i, \phi, w_i) = c(y_i, \phi, w_i)\ exp\ (\frac{\Theta_i^T y_i - b(\Theta_i)}{\phi}\ w_i) \tag{2.1}$$

2.2 Strukturannahme:

$\mu_i = E(y_i|x_i)$ ist bestimmt durch den linearen Prediktor $\eta_i = X_i\beta$, und zwar über die Responsefunktion h beziehungsweise über die Linkfunktion g, die jetzt mehrdimensionale Funktionen sind. Wobei gilt:

- $\eta_i = (\eta_{i1}, ..., \eta_{iq})^T$
- $X_i = diag\ (\underline{x}_i)$, wobei $\underline{x}_i$ Zeilenvektoren sind
- $\beta = (\beta_1, ..., \beta_q)^T$ (insgesamt pq Zeilen, weil β_i ja ein p-Vektor ist)

Komponentenweises Anschreiben von $X_i\beta$ führt zu einem Gleichungssystem wie im eindimensionalen Fall.

$$\eta_i = \begin{pmatrix} \eta_{i1} = X_i\beta_1 \\ \vdots \\ \eta_{iq} = X_i\beta_q \end{pmatrix}$$

Also gilt $\mu_i = h(\eta_i) = h(X_i\beta)$ bzw. $\eta_i = X_i\beta = g(\mu_i)$ wobei h eine Funktion von $R^q \longrightarrow R^q$ ist und $g = h^{-1}$ und $h = (h_1, ..., h_q)^T$ gilt.

Genauso wie im univariaten Fall ist ein multivariates GLM bestimmt durch:

- den Typ der Exponentialfamilie
- die Wahl der Link- / Responsefunktion
- Bestimmung und Auswahl der Regressoren bzw. gleichwertig der Designmatrix X

2.3 Beispiel: Das mehrkategoriale Logit-Modell

2.3.1 Verteilungsannahme:

Die Verteilungsannahme ($y_i \sim B^{sk}_{N_i,p_i}$) ist erfüllt mit:

$$\Theta_i = \begin{pmatrix} ln\ \frac{p_{i1}}{1-\sum_{j=1}^{q} p_{ij}} \\ \vdots \\ ln\ \frac{p_{iq}}{1-\sum_{j=1}^{q} p_{ij}} \end{pmatrix}; \qquad b(\Theta_i) = ln\ (1 - \sum_{j=1}^{q} p_{ij}); \qquad w_i = N_i; \qquad \phi_i = 1 \tag{2.2}$$

2.3.2 Strukturannahme:

Die Strukturannahme (mit natürlicher Linkfunktion) sieht folgendermaßen aus:

$$\eta_i = \begin{pmatrix} \eta_{i1} \\ \vdots \\ \eta_{iq} \end{pmatrix} = \begin{pmatrix} X_i\beta_1 \\ \vdots \\ X_i\beta_q \end{pmatrix}; \qquad p_i = h(\eta_i) = \begin{pmatrix} h_1(\eta_i) \\ \vdots \\ h_q(\eta_i) \end{pmatrix} = \begin{pmatrix} p_{i1} \\ \vdots \\ p_{iq} \end{pmatrix}; \tag{2.3}$$

Ausserdem gilt natürlich $p_{ij} = h_j(\eta_i)$ für $j = 1, ..., q$ und $i = 1, ..., I$.

2.3.3 Responsefunktion:

Die Responsefunktion h ist gegeben als:

$$h(\eta_i) = \begin{pmatrix} h_1(\eta_i) \\ \vdots \\ h_q(\eta_i) \end{pmatrix}; \qquad h_r(\eta_i) = \frac{exp\ (\eta_{ir})}{1 + \sum_{j=1}^{q} exp\ (\eta_{ij})} \tag{2.4}$$

In Matrixschreibweise gilt dann: $h(\eta_i) = h(X_i\beta) = p_i = (p_{i1}, ..., p_{iq})^T$

2.3.4 Linkfunktion:

Es gilt ganz allgemein: $g(p_i) = g(\mu_i) = X_i\beta$. In Vektorschreibweise erhält man:

$$g(p_i) = \begin{pmatrix} g_1(p_i) \\ \vdots \\ g_q(p_i) \end{pmatrix}; \qquad g_r(p_i) = ln\ \frac{p_{ir}}{1 - \sum_{j=1}^{q} p_{ij}} \tag{2.5}$$

2.4 Schätzen und Testen:

Wir behandeln hier nur Maximum-Likelihood (ML) Schätzer. Da die Dichte von y aus einer q-parametrischen Exponentialfamilie stammt, ist die Log-Likelihood Funktion der i-ten Beobachtung (bis auf eine additive Konstante) gegeben durch:

$$l_i(\beta) = \frac{(\Theta_i^T y_i - b(\Theta_i))w_i}{\phi} \tag{2.6}$$

Differenzieren nach β ergibt die individuellen Scorefunktionen $s_i(\beta) = X_i^T D_i(\beta)\Sigma_i^{-1}(\beta)\ (y_i - \mu_i(\beta))$ wobei $D_i(\beta) = \frac{\vartheta h(\eta_i)}{\vartheta \eta}$ gelten muss. Damit ergibt sich folgende Matrix,

$$\begin{pmatrix} \frac{\vartheta h_1}{\vartheta \eta_1} & \cdots & \frac{\vartheta h_1}{\vartheta \eta_q} \\ \vdots & \ddots & \vdots \\ \frac{\vartheta h_q}{\vartheta \eta_1} & \cdots & \frac{\vartheta h_q}{\vartheta \eta_q} \end{pmatrix} \tag{2.7}$$

die "Jakobimatrix" von h ausgewertet an der Stelle $\eta_i = X_i\beta$.

$\Sigma_i(\beta)$ ist die Kovarianzmatrix für $\mu_i(\beta) = h(X_i\beta)$. Für die skalierte Multinomialverteilung mit $\mu_i = p_i$ hat $\Sigma_i(\beta)$ folgende Form:

$$\Sigma_i(\beta) = \frac{1}{N_i} \begin{pmatrix} p_{i1}\ (1-p_{i1}) & \cdots & \cdots \\ \vdots & -p_{ij}\ p_{ik} & \vdots \\ \cdots & \cdots & p_{iq}\ (1-p_{iq}) \end{pmatrix} \tag{2.8}$$

wobei $-p_{ij}\ p_{ik}$ das Element der k-ten Spalte und j-ten Zeile ist. In Matrixdarstellung ergibt sich nun: $\Sigma_i(\beta) = \frac{1}{N_i}\ (diag\ (p_{ij})_{j=1,\ldots,q} - p_i p_i^T)$. Damit gilt für die Kovarianzmatrix $\Sigma_i^{-1}(\beta)$:

$$\Sigma_i(\beta) = Kov(\mu_i(\beta)) = Kov(h(X_i\beta)) \tag{2.9}$$

Für die Gewichte w_i gilt nun, $W_i(\beta) = D_i(\beta)\Sigma_i^{-1}(\beta)D_i^T(\beta)$. Die Fisher-Informationsmatrix und die Scorefunktion der i-ten Beobachtung sind daher gegeben als:

$$F_i(\beta) = X_i^T W_i(\beta) X_i \qquad s_i(\beta) = X_i^T W_i(\beta) D_i^{-T}(\beta)\ (y_i - \mu_i(\beta)) \tag{2.10}$$

Für die totale Scorefunktion und die totale Fisher Informationsmatrix fasst man die Matrizen X_i und Vektoren Y_i folgendermaßen zusammen:

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_q \end{pmatrix}; \qquad Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_q \end{pmatrix} \tag{2.11}$$

wobei X eine (Iq, pq)-Matrix und Y eine $(Iq, 1)$-Matrix ist. Es ergibt sich nun

- $\Sigma(\beta) = diag\ (\Sigma_i(\beta)_{i=1,\ldots,q})$
- $W(\beta) = diag\ (W_i(\beta)_{i=1,\ldots,q})$
- $D(\beta) = diag\ (D_i(\beta)_{i=1,\ldots,q})$

Es ist zu beachten, dass Σ, W und D sogenannte "Blockdiagonalmatritzen" sind. Die Diagonalelemente der Matrizen sind selbst wieder Matrizen der Dimension (q, q). Faßt man die Matrizen nun auf die obige Weise zusammen, ergibt sich für die totale Fisher-Informationsmatrix und die Scorefunktion:

$$F(\beta) = X^T W(\beta) X; \qquad s(\beta) = \sum_{i=1}^{I} s_i(\beta) = X^T W(\beta)\Sigma^{-1}(\beta)\ (Y - \mu(\beta)) \tag{2.12}$$

2.5 Numerische Berechnung des ML-Schätzer:

Die numerische Berechnung des ML-Schätzers, die asymptotischen Eigenschaften sowie das Testen von linearen Hypothesen sind exakt gleich dem eindimensionalen Fall.

2.6 Beurteilung der Anpassungsgüte:

Wie im eindimensionalen Fall für gruppierte Daten (und zwar nur für solche, das heisst keine stetigen Einflussgrößen mit $N_i = 1$) betrachten wir wieder die Pearson Statistik χ^2 und die Devianz D.

- Pearson-Statistik:

  Die Pearson-Statistik ist gegeben als:

$$\chi^2 = \phi \sum_{i=1}^{I} (y_i - \mu_i(\hat{\beta}))^T \; \Sigma_i^{-1}(\beta) \; (y_i - \mu_i(\hat{\beta})) \tag{2.13}$$

  Im Spezialfall des mehrkategoriellen Logit-Modells ergibt sich mit $\mu_i = p_i$ und $\phi = 1$ für die Pearson-Statistik

$$\chi^2 = \sum_{i=1}^{I} N_i \sum_{s=1}^{R} \frac{y_{is} - \hat{p}_{is}}{p_{is}}; \qquad \hat{p}_i = \begin{pmatrix} \hat{p}_{i1} \\ \vdots \\ \hat{p}_{iR} \end{pmatrix} \tag{2.14}$$

  wobei $\hat{p}_{ir} = p_{ir}(\hat{\beta})$ und y_{ir} die r-te Komponente des Vektors y_i (hier inklusive der R-ten (redundanten) Referenzkategorie!) ist.

- Devianz:

  Die Devianz D ist wie im eindimensionalen Fall gegeben als:

$$D = -2\phi \sum_{i=1}^{I} (l_i(\hat{\mu}) - l_i(y_i)) \tag{2.15}$$

Unter den gleichen Regularitätsvoraussetzungen wie im eindimensionalen Fall gilt für $\frac{N_i}{N} \to p_i \in (0,1)$ und $N \to \infty$, dass sowohl die Pearson Statistik χ^2 als auch die Devianz D aysmptotisch verteilt sind nach $\chi^2_{(I-p)q}$.

2.7 Residualanalyse:

Die generalisierte "Hat-Matrix" ist in der linearen Regression definiert als $H = X(X^TX)^{-1}X^T$ ("Hat Matrix" wegen: $\hat{y} = Hy = X\beta$). Es gilt, dass $h_{ii} = x_i(X^TX)^{-1}x_i^T$ der Einfluss der i-ten Beobachtung ist. Einflussreiche Punkte im Regressionsmodell haben große Werte h_{ii}. Gesucht ist nun ein Analogon für GLMs.

Lösung:

Indirekt kann man die Lösung über das Fisher Scoring herausfinden. Dort gilt ja bekanntlichermaßen $\hat{\beta}_{k+1} = \hat{\beta}_k + F^{-1}(\hat{\beta}_k)s(\hat{\beta}_k)$. Für die Lösung gilt daher: $\hat{\beta} = \hat{\beta} + F^{-1}(\hat{\beta}_k)s(\hat{\beta}_k)$. Einsetzen der Formeln für $F(\beta)$ und $s(\beta)$ liefert:

$$\hat{\beta} = \hat{\beta} + (X^TW(\hat{\beta})X)^{-1}\ X^TW(\hat{\beta})D^{-T}(\hat{\beta})\ (y - \mu(\hat{\beta})) \tag{2.16}$$

Mit $\tilde{y}(\hat{\beta}) = X\hat{\beta} + D^{-1}(\hat{\beta})\ (y - \mu(\hat{\beta}))$ gilt:

$$\begin{aligned} \hat{\beta} &= (X^TW(\hat{\beta})X)^{-1}\ X^TW(\hat{\beta})\ \tilde{y}(\hat{\beta}) \\ &= (X^TW^{\frac{T}{2}}(\hat{\beta})W^{\frac{1}{2}}(\hat{\beta})X)^{-1}\ X^TW^{\frac{T}{2}}(\hat{\beta})W^{\frac{1}{2}}(\hat{\beta})\ \tilde{y}(\hat{\beta}) \end{aligned} \tag{2.17}$$

Für $X_0 = W^{\frac{1}{2}}(\hat{\beta})X$ kann man das obige β auch als Schätzer für das lineare Regressionsproblem $Y_0 = X_0\beta + e$ sehen. Die dazugehörige Hat-Matrix hat die Form $H = X_0(X_0^TX_0)^{-1}X_0^T$. Einsetzen in die Hat-Matrix liefert $H = W^{\frac{1}{2}}(\hat{\beta})X\ (X^TW(\hat{\beta})X)^{-1}\ X^TW^{\frac{T}{2}}(\hat{\beta})$.

H ist eine (Iq, Iq)-Matrix, die sich in Blöcke H_{ij} der Größe (q, q) zerlegen lässt $(i, j = 1, ..., I)$. Den Einfluss der i-ten Beobachtung sieht man an H_{ij}. Beobachtungen mit großer Determinante $|H_{ij}|$ oder großer Spur $tr\ H_{ij}$ sind einflussreich. Die Spezialisierung auf den eindimensionalen Fall ist offensichtlich. Die Spur einer Matrix ist gleich der Determinante falls h_{ij} nur ein Skalar ist.

Mit

$$c_i = (\hat{\beta}_{(i)} - \hat{\beta})^T\ Cov(\hat{\beta})^{-1}\ (\hat{\beta}_{(i)} - \hat{\beta}) \tag{2.18}$$

wird die sogenannte "Cook-Distanz" definiert. Hier ist $\hat{\beta}_{(i)}$ der ML-Schätzer für β ohne

die i-te Beobachtung. Es gilt nun: Ist c_i groß, so ist die i-te Bobachtung relativ gesehen einflussreich. Verwendet man die asymptotische Verteilung von $\hat{\beta} \sim N(\beta, F^{-1}(\beta))$, erhält man

$$c_i = (\hat{\beta}_{(i)} - \hat{\beta})^T \; X_i^T W_i(\beta) X_i \; (\hat{\beta}_{(i)} - \hat{\beta}) \tag{2.19}$$

was schliesslich auf eine Chi-Quadrat verteilte Testgröße führt.

Der eindimensionale Spezialfall mit Normalverteilungsannahme (klassisches lineares Modell) führt zu: $c_i = (\hat{\beta}_{(i)} - \hat{\beta})^T \; X_i^T X_i \; (\hat{\beta}_{(i)} - \hat{\beta})$. Wieder weisen große Werte von c_i auf einflussreiche Beobachtungen hin.

Kapitel 3

Kontingenztafeln - Das log-lineare Modell

Bei einer kategoriale Regression herrscht eine asymetrische Fragestellung (Unterscheidung zwischen abhängigen und unabhängigen Variablen) vor. Hier liegt der Schwerpunkt bei symmetrischen Fragestellungen (keine Unterscheidung zwischen abhängigen und unabhängigen Variablen). Es soll nur die Frage beantwortet werden, ob ein Zusammenhang zwischen den Variablen besteht.

3.1 2-dimensionale Modelle:

Unterscheidung der Kontingenztabellen nach den Erhebungsschemata, die die Hypothesen und auch die Interpretationen einer Auswertung beeinflussen.

3.1.1 Produkt-multinomiales Erhebungsschema:

Ein Merkmal wird als Faktor im Sinne der Varianzanalyse betrachtet und die Untersuchungsobjekte werden zufällig den Faktorausprägungen (= Behandlungen) zugeordnet. Das abhängige Merkmal (= Reaktionsvariable) ist aber nicht wie in der Varianzanalyse metrisch sondern kategoriell. Folgende Tabelle soll den Zusammenhang verdeutlichen:

		Reaktionsvariable				
		1	2	...	J	$\sum$
	1	x_{11}				x_{1+}
Faktor	2		x_{22}			x_{2+}
	$\vdots$			x_{ij}		x_{i+}
	I				x_{IJ}	x_{I+}
	$\sum$	x_{+1}			x_{+J}	x_{++}

Bei dieser Kontingenztabelle ist festzustellen, dass sowohl die Randhäufigkeiten x_{i+} als auch der Stichprobenumfang x_{++} fest sind. Der Versuchsleiter bestimmt die Anzahl der Versuchseinheiten in den verschiedenen Faktorstufen x_{i+}, $i = 1, ..., I$. Damit ist auch der Stichprobenumfang x_{++} bestimmt. Die Rotationsvariable R besitzt für jeden Faktorwert eine Multinomialverteilung. Diese Verteilungen werden jetzt miteinander verglichen.

Sei $p_{ij} := P(\text{R fällt in Kategorie } j | \text{ Versuchsobjekt entstammt Faktorstufe } i)$. Die Nullhypothese ist, dass die Faktorausprägung keinen Einfluss auf das Reaktionsmerkmal hat. Diese Nullhypothese kann mathematisch gleich formuliert werden als:

$$H_0 : \ p_{1j} = p_{2j} = \ldots = p_{Ij} \qquad \forall j = 1, ..., J \tag{3.1}$$

Formulierung von H_0 mit zu erwartenden Häufigkeiten:

Sei x_{ij} die Anzahl der Versuchsobjekte mit Behandlung i, die in die Kategorie j fallen, dann ist die gemeinsame Verteilung aller IJ Häufigkeiten gegeben durch die sogenannte "Produkt-multinomiale Verteilung":

$$P(x_{11}, \ldots, x_{IJ}) = \prod_{i=1}^{I} \frac{x_{i+}!}{x_{i1}! x_{i2}! \cdots x_{iJ}!} \, p_{i1}^{x_{i1}} \cdots p_{iJ}^{x_{iJ}} \tag{3.2}$$

Die zu erwartende Häufigkeit der Zelle (i, j), m_{ij} ist gegeben durch $E(x_{ij}) = m_{ij} = x_{i+}p_{ij}$, wobei $m_{i+} = \sum_{j=1}^{J} x_{i+}p_{ij} = \sum_{j=1}^{J} m_{ij}$ und $m_{+j} = \sum_{i=1}^{I} x_{i+}p_{ij} = \sum_{i=1}^{I} m_{ij}$ die zu erwartenden Randhäufigkeiten sind. Man kann nun die Nullhypothese auch schreiben als

$$H_0 : \ m_{ij} = \frac{m_{i+}m_{+j}}{m_{++}} \tag{3.3}$$

wobei m_{++} den Stichprobenumfang bezeichnet.

3.1.2 Multinomiales Erhebungsschma:

Eine Stichprobe von festem Umfang (nur mehr x_{++} ist fest) wird der Population entnommen und sämtliche Merkmale sind Zufallsvariablen. Die Randsummen x_{i+} sind hier nicht fixiert sondern ebenfalls zufällig. x_{ij} besitzt also eine Multinomialverteilung mit IJ Ausprägungen. Sei Merkmal A die Klassifikation nach dem ersten Merkmal und B die Klassifikation nach dem zweiten Merkmal, dann ist die Wahrscheinlichkeit p_{ij} gegeben

durch: $p_{ij} := P(\text{R besitzt die Ausprägung } A_i \text{ und B besitzt die Ausprägung } B_i)$. Nun kann getestet werden, ob die Merkmale A und B unabhängig sind. Als Nullhypothese gilt daher:

$$H_0 : \ P(A_i \cap B_i) = P(A_i)P(B_i) \leftrightarrow p_{ij} = p_{i+}p_{+j} \tag{3.4}$$

Darstellung von H_0 mit zu erwartenden Häufigkeiten:

Die gemeinsame Verteilung der beobachteten Häufigkeiten ist eine Multinomialverteilung. Es gilt:

$$P(x_{11}, \ldots, x_{IJ}) = \frac{x_{++}!}{x_{11}! \cdots x_{IJ}!} \, p_{11}^{x_{11}} \cdots p_{IJ}^{x_{IJ}} \tag{3.5}$$

Mit zu erwartenden Häufigkeiten lässt sich die Nullhypothese wieder wie oben formulieren.

$$H_0 : \ m_{ij} = \frac{m_{i+}m_{+j}}{m_{++}} \tag{3.6}$$

3.1.3 Poisson-Erhebungsschema:

Hier ist auch der Stichprobenumfang N zufällig. Ein Versuch wird etwa nach einer fixen Zeitspanne und nicht nach einer festen Beobachtungszahl abgebrochen. Wir nehmen an, dass die beobachteten Häufigkeiten in den Zellen Realisierungen von unabhängigen Poissonprozessen sind. Daraus folgt, dass sich für jede Zellbesetzung x_{ij} eine Poissonverteilung mit Parameter μ_{ij} ergibt. Die gemeinsame Verteilung aller beobachteten Häufigkeiten ist dann:

$$P(x_{11}, \ldots, x_{IJ}) = \prod_{i,j} \frac{\mu_{ij}^{x_{ij}}}{x_{ij}!} \, exp\ (-\mu_{ij}) \tag{3.7}$$

Ein möglicher Test für den Zusammenhang zwischen den verschiedenen Zellhäufigkeiten ist gegeben durch:

$$H_0 : \ \mu_{ij} = \frac{\mu_{i+}\mu_{+j}}{\mu_{++}} \tag{3.8}$$

Man spricht hier von einer sogenannten "multiplikative Hypothese" oder etwas allgemeiner vom "multiplikativen Poissonmodell".

Formulierung von H_0 mit zu erwartenden Häufigkeiten:

Ist $m_{ij} = E(x_{ij})$, dann ist diese Nullhypothese wieder äquivalent zu den obrigen Nullhypothesen. Wir definieren $\tilde{x}_{ij} := x_{ij}|x_{++}$. Dann kann die Nullhypothese mit zu erwartenden Häufigkeiten wieder in folgender Form aufgeschrieben werden

$$H_0 : \tilde{m}_{ij} = \frac{\tilde{m}_{i+}\tilde{m}_{+j}}{\tilde{m}_{++}} \tag{3.9}$$

wobei gilt: $\tilde{m}_{ij} = E(\tilde{x}_{ij})$.

3.2 Das Unabhängigkeitsmodell:

Für alle obigen Erklärungsschemata ist die Nullhypothese von folgender Form:

$$H_0 : m_{ij} = \frac{m_{i+}m_{+j}}{m_{++}} \tag{3.10}$$

Logarithmiert man nun die Nullhypothese, so ergibt sich $ln\ m_{ij} = ln\ m_{i+} + ln\ m_{+j} - ln\ m_{++}$. Das heißt, dass bei Gültigkeit der Nullhypothese lässt sich die zu erwartende Häufigkeit einer Zelle darstellen als die Summe dreier Terme.

- ein nur von der Zeile abhängender Term ($ln\ m_{i+}$)
- ein nur von der Spalte abhängender Term ($ln\ m_{+j}$)
- ein nur von der Beobachtungszahl abhängender Term ($ln\ m_{++}$)

Die allgemeine Struktur der logarithmierten erwarteten Häufigkeiten ist also $ln\ m_{ij} = \mu + \mu_{A(i)} + \mu_{B(j)}$, wobei $\mu_{A(i)}$ nur von i, also vom ersten Merkmal und $\mu_{B(j)}$ nur von j, also dem zweiten Merkmal, abhängt. Für eine einfachere (eindeutige) Darstellung auch in höherdimensionalen Modellen wählt man für die Parameter folgende Restriktionen.

$$\sum_i \mu_{A(i)} = \sum_j \mu_{B(j)} = 0 \tag{3.11}$$

Die Struktur des Modells zusammen mit den Restriktionen ergibt also das log-lineare Unabhängigkeitsmodell.

Satz:

Die Hypothese $H_0: \ m_{ij} = \frac{m_{i+}m_{+j}}{m_{++}}$ ist

- für das Poisson-Erhebungsschema äquivalent zum log-linearen Unabhängigkeitsmodell
- für das multinomiale Erhebungsschema äquivalent zum loglinearen Unabhängigkeitsmodell mit zusätzlicher Nebenbedingung

$$x_{++} = \sum_{i,j} exp\ (\mu + \mu_{A(i)} + \mu_{B(j)}) = \sum_{i,j} m_{ij} \tag{3.12}$$

- für das Produkt-Multinomiale Erhebungsschema äquivalent zum loglinearen Unabhängigkeitsmodell mit zusätzlichen (insgesamt I für $i = 1, ..., I$) Nebenbedingungen

$$x_{i+} = \sum_{j} exp\ (\mu + \mu_{A(i)} + \mu_{B(j)}) = \sum_{j} m_{ij} \tag{3.13}$$

3.2.1 Parameter des log-linearen Unabhängigkeitsmodells:

Die Parameter des log-linearen Unabhängigkeitsmodells ergeben sich mit:

- $\mu = \frac{1}{IJ} \sum_i \sum_j ln\ m_{ij}$
- $\mu_{A(i)} = \frac{1}{J} \sum_j ln\ m_{ij} - \mu$
- $\mu_{B(j)} = \frac{1}{I} \sum_i ln\ m_{ij} - \mu$

Das log-lineare Unabhängigkeitsmodell entspricht der Unabhängigkeit von 2 Merkmalen. Mann kann das Modell einfach so erweitern, dass auch Abhängigkeiten (beliebige) zwischen den Merkmalen dargestellt werden können. Dies führt zum saturierten Modell.

3.2.2 Das saturierte Modell:

Das saturierte Modell ist gegeben als

$$ln\ m_{ij} = \mu + \mu_{A(i)} + \mu_{B(j)} + \mu_{AB(i,j)} \tag{3.14}$$

mit Nebenbedingungen

$$\sum_i \mu_{A(i)} = \sum_j \mu_{B(j)} = 0 \qquad \sum_i \mu_{AB(i,j)} = \sum_j \mu_{AB(i,j)} = 0; \quad \forall i,j \tag{3.15}$$

Weiters werden noch zusätzliche Restriktionen eingeführt:

- $x_{++} = \sum_{i,j} exp\,(\mu + \mu_{A(i)} + \mu_{B(j)}) = \sum_{i,j} m_{ij}$ für das multinomiale Erhebungsschema
- $x_{i+} = \sum_j exp\,(\mu + \mu_{A(i)} + \mu_{B(j)}) = \sum_j m_{ij}; \quad \forall i = 1, ..., I$ für das produkt-multinomiale Erhebungsschema

3.2.3 Bezeichnung der Parameter:

Die Bezeichnung der Parameter ist ähnlich wie in der Varianzanalyse, obgleich die Interpretation der Parameter nicht immer übereinstimmt. Es ist also:

- $\mu_{A(i)} + \mu_{B(j)} \ldots$ Haupteffekte
- $\mu_{AB(i,j)} \ldots$ Interaktionseffekt

Die Anzahl der frei variierenden Parameter des saturierten Modells ist (unter Berücksichtigung der Nebenbedingungen):

- $1 + (I-1) + (J-1) + (I-1)(J-1) = IJ$ im Poisson-Erhebungsschema
- $IJ - 1$ im multinomialen Erhebungsschema
- $I(J-1)$ im produkt-multinomialen Erhebungsschema:

Durch das saturierte Modell lässt sich also jede (beliebige) Menge $\{m_{ij}\}$ von zu erwartenden Häufigkeiten beschreiben. Das log-lineare Unabhängigkeitsmodell ist ein Spezialfall mit $\mu_{AB(i,j)} = 0$ ($\mu_{AB(i,j)} = ln\ m_{ij} - \mu - \mu_{A(i)} - \mu_{B(j)}$). Die Anzahl der freien Parameter im log-linearen Unabhängigkeitsmodell ist also:

- $(I-1) + (J-1) = I + J - 1$ im Poisson-Erhebungsschema
- $I + J - 2$ im multinomialen Erhebungsschema
- $J - 1$ im produkt-multinomialen Erhebungsschema:

3.2.4 Zusammenhang mit der Varianzanalyse:

Die Modellgleichung für eine zweifaktorielle Varianzanalyse mit Effektparametrisierung ist gegeben durch:

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} \tag{3.16}$$

Es zeigt sich eine starke Ähnlichkeit zum saturierten log-linearen Modell, allerdings mit einem wesentlichen Unterschied. In der Varianzanalyse wird der Mittelwert (=Erwartungswert) der abhängigen Variable als Summe der Effekte von Faktoren modelliert. Im log-linearen Modell wird die logarithmierte zu erwartende Zellhäufigkeit als Summe von Effekten parametrisiert.

Betrachtet man die logarithmierten Zellhäufigkeiten als Zielvariable (abhängige Variable) und interessiert man sich für den Effekt der zugehörigen Merkmalskombination auf die Häufigkeiten, dann lassen sich die Parameter insbesondere im Poisson-Erhebungsschema analog zur Varianzanalyse interpretieren. Die Interpretation der log-linearen Modelle zielt aber weiters auf den Zusammenhang der Merkmale ab, und daraus folgt der wesentliche Unterschied zur Varianzanalyse.

3.2.5 Unterschiede im log-linearen Modell:

Im log-linearen Modell gehören sowohl zu unabhängigen als auch zur abhängigen Variable Haupteffekte μ_A, μ_B. Verschwinden im log-linearen Modell die Interaktionseffekte μ_{AB}, dann bedeutet das die Unabhängigkeit der Merkmale A, B. Hingegen äußert sich ein fehlender Einfluss des Faktors (entspricht Unabhängigkeit zwischen erklärenden und zu erklärendende Variable) in der Varianzanalyse durch Verschwinden der Haupteffekte.

Ein Effekt ist in der Varianzanalyse also immer auf eine Zielvariable ausgerichtet. Im log-linearen Modell hingegen besitzen die Haupteffekte nur einen Zusammenhang mit den Häufigkeiten und erst die Interaktionseffekte erlauben eine Beurteilung des Zusammenhangs zwischen verschiedenen Merkmalen.

3.3 Das Kreuzprodukt-Verhältnis α

Das Kreuzprodukt-Verhältnis ("odds-ratio-Assoziation") ist folgendermaßen definiert:

$$\alpha_{i_1,i_2,j_1,j_2} = \frac{m_{i_1,j_1} m_{i_2,j_2}}{m_{i_1,j_2} m_{i_2,j_1}} = \frac{P(B_{j_1}|A_{i_1})/P(B_{j_2}|A_{i_1})}{P(B_{j_1}|A_{i_2})/P(B_{j_2}|A_{i_2})} \tag{3.17}$$

Das Kreuzprodukt-Verhältnis stellt den Zusammenhang des Merkmals A in den Ausprägungen A_{i_1}, A_{i_2} mit Merkmal B in den Ausprägungen B_{j_1}, B_{j_2} dar. Die Kreuzprodukte α_{i_1,i_2,j_1,j_2} enthalten die gesamte Information über den Zusammenhang von Merkmal A und B.

- $\alpha_{i_1,i_2,j_1,j_2} = 1$ gilt genau dann, wenn A, B unabhängig sind $(\forall i_1 \neq i_2, j_1 \neq j_2)$
- Kennt man sämtliche α_{i_1,i_2,j_1,j_2}, dann lassen sich bei Kenntnis der Randverteilungen p_{i+} und p_{+j} sämtliche Auftrittswahrscheinlichkeiten $P(A_i \cap B_j) = p_{ij}$ beziehungsweise $P(A_i|B_j)$ oder $P(B_j|A_i)$ bestimmen.

Ausserdem ist α_{i_1,i_2,j_1,j_2} eine direkte Funktion des Interaktionsparameter μ_{AB} des saturierten log-linearen Modells (zweite Differenzen!). Es gilt nämlich:

$$\alpha_{i_1,i_2,j_1,j_2} = exp\left((\mu_{AB(i_1,j_1)} - \mu_{AB(i_1,j_2)}) - (\mu_{AB(i_2,j_1)} - \mu_{AB(i_2,j_2)})\right) \tag{3.18}$$

Im Unabhängigkeitsmodell sind sämtliche Interaktionen 0, damit sämtliche zweite Differenzen ebenfalls 0, woraus folgt, dass $exp\,(0) = 1 = \alpha_{i_1,i_2,j_1,j_2}$ ist.

3.4 3-Dimensionale Modelle:

Mit der Erhöhung der Variablenzahl geht eine wesentliche Erweiterung der möglichen Zusammenhänge der beteiligten Variablen einher (zb. Alle Variablen sind voneinander unabhängig; 2 Variablen hängen voneinander ab, die 3. Variable ist aber unabhängig,...). Es wird also untersucht, ob ein Zusammenhang zwischen den Variablen besteht und wenn ja, von welcher Form der Zusammenhang ist.

3.4.1 Beispiel: Modelle mit 3 Variablen

Wir betachten die 3 Variablen A, B, C wobei A die Ausprägungen $i = 1, ..., I$, B die Ausprägungen $j = 1, ..., J$ und C die Ausprägungen $j = 1, ..., k$ besitzt. Weiters gilt $m_{ijk} =$

$E(x_{ijk})$ und $p_{ijk} = \frac{m_{ijk}}{x_{+++}}$. Die insgesamt 9 verschiedenen Modelltypen unterscheiden sich in der Art des Zusammenhanges zwischen den Variablen.

1. Das saturierte Modell (ABC)

   Das saturierte Modell ist gegeben als durch die Haupteffekte, die 2-er Interaktionen und eine Dreier-Interaktion. In Formeln ergibt sich:

   $$ln\ m_{ijk} = \mu + \mu_{A(i)} + \mu_{B(j)} + \mu_{C(k)} + \mu_{AB(i,j)} + \mu_{AC(i,k)} + \mu_{BC(j,k)} + \mu_{ABC(i,j,k)}$$

   Die Nebenbedingungen (Normierung) sind:

   - $\sum_i \mu_{A(i)} = \sum_j \mu_{B(j)} = \sum_k \mu_{C(k)} = 0$
   - $\sum_i \mu_{AB(i,j)} = \sum_j \mu_{AB(i,j)} = 0$
     $\sum_i \mu_{AC(i,k)} = \sum_c \mu_{AC(i,k)} = 0$
     $\sum_j \mu_{BC(j,k)} = \sum_c \mu_{BC(j,k)} = 0$
   - $\sum_i \mu_{ABC(i,j,k)} = \sum_j \mu_{ABC(i,j,k)} = \sum_k \mu_{ABC(i,j,k)} = 0$

   Jede beliebige Menge $\{m_{ijk}\}$ ist also durch das saturierte Modell darstellbar.

2. Modell ohne 3-er Interaktionen: (AB|AC|BC)

   Es ist die erste Vereinfachung der Zusammenhangstruktur. Es gilt: Der Zusammenhang zwischen zwei Merkmalen ist unbeeinflusst von der dritten Variablen. Bedingungen des Modells an die m_{ijk} bzw. an die bedingten Kreuzproduktverhältnisse (und zwar für alle möglichen) sind:

   $$\alpha_{i_1,i_2,j_1,j_2}(c_{k_1}) = \frac{m_{i_1,j_1,k_1} m_{i_2,j_2,k_1}}{m_{i_2,j_1,k_1} m_{i_1,j_2,k_1}} = \frac{m_{i_1,j_1,k_2} m_{i_2,j_2,k_2}}{m_{i_2,j_1,k_2} m_{i_1,j_2,k_2}} = \alpha_{i_1,i_2,j_1,j_2}(c_{k_2}) \qquad (3.19)$$

3. Modelle der bedingten Unabhängigkeit: (z.B: AC|BC)

   Hier gibt es keine 3-er Interaktionen und nur zwei 2-er Interaktionen. Die Bedingungen des Modells an m_{ijk} sind:

   $$\begin{aligned} m_{ijk} = \frac{m_{i+k} m_{+jk}}{m_{++k}} \Leftrightarrow p_{ijk} = \frac{p_{i+k} p_{+jk}}{p_{++k}} \Leftrightarrow [\frac{p_{ijk}}{p_{++k}} = \frac{p_{i+k}}{p_{++k}} \frac{p_{+jk}}{p_{++k}}] \Leftrightarrow \\ \frac{P(A_i \cap B_j \cap C_k)}{P(C_k)} = \frac{P(A_i \cap C_k)}{P(C_k)} \frac{P(B_j \cap C_k)}{P(C_k)} \Leftrightarrow \\ P(A_i \cap B_j | C_k) = P(A_i | C_k) P(B_j | C_k) \qquad (3.20) \end{aligned}$$

   Aus diesen Bedingungen folgt auch der Name "bedingte Unabhängigkeitsmodelle".

4. Modelle der Unabhängigkeit einer Variable: (z.B: AB|C)
   Nur Haupteffekte und eine Zweierinteraktion kommen im Modell vor (hier AB). Die dritte Variable (C) ist unabhängig von den anderen Variablen. Die Bedingungen des Modells an m_{ijk} bzw. p_{ijk} sind:

$$m_{ijk} = \frac{m_{i+k}m_{+jk}}{m_{+++}} \Leftrightarrow p_{ijk} = p_{ij+}p_{++k} \Leftrightarrow P(A_i \cap B_j \cap C_k) = P(A_i \cap B_j)P(C_k)$$

   Das heißt, dass A, B zusammen unabhängig von C sind.

5. Modelle der totalen Unabhängigkeit der Variablen: (A|B|C)
   Bei dieser Modellspezifikation gibt es keine Interaktionen sondern nur Haupteffekte. Die Bedingungen sind folgende:

$$P(A_i \cap B_j \cap C_k) = P(A_i)P(B_j)P(C_k)$$

   Weitere Modelltypen sind nicht so interessant, der Vollständigkeit halber aber erwähnt.

6. saturiertes Modell für 2 Variablen: (z.B: AC)

$$ln\ m_{ijk} = \mu + \mu_{A(i)} + \mu_{C(k)} + \mu_{AC(i,k)} \tag{3.21}$$

7. Modelle mit nur 2 Haupteffekten: (z.B: A|C)

$$ln\ m_{ijk} = \mu + \mu_{A(i)} + \mu_{C(k)} \tag{3.22}$$

8. Modelle mit nur einem Haupteffekt: (z.B: B)

$$ln\ m_{ijk} = \mu + \mu_{B(j)} \tag{3.23}$$

9. Nullmodell:

$$ln\ m_{ijk} = \mu \Leftrightarrow P(A_i \cap B_j \cap C_k) = \frac{1}{IJK} \tag{3.24}$$

Die Modelle 3 bis 9 heißen "multiplikative Modelle", weil sich deren gemeinsame Dichten durch Randdichten faktorisieren lassen.

3.4.2 Erhebungsschemata im 3-dimensionalen Modell:

Wir behandeln nun Modelle mit den Merkmalen A,B und C. Das multinomiale sowie das Poisson-Erhebungsschema sind völlig analog zu zweidimensionalen Modellen definiert. Beim Produkt-Multinomialen gewichteten Schema muss man aber unterscheiden, ob für ein oder zwei Merkmale feste Stichprobenumfänge vorgegeben sind. Man spricht von produkt-multinomialen Erhebungsschemata in einem bzw. in zwei Merkmalen. Wird z.B. nur nach dem Merkmal A geschichtet, d.h. die Randsummen x_{i++} sind fix, dann gilt:

$$P(x_{111}, ..., x_{IJK}) = \prod_i \frac{x_{i++}!}{\prod_{j,k} x_{i,j,k}!} \prod_{j,k} {p_{i,j,k}}^{x_{i,j,k}} \tag{3.25}$$

Eine Kontingenztafel, die nach dem produkt-multinomialen Schema erhoben wird, engt die Menge der möglichen Modelle (1,...9 → vgl. oben) ein und erfordert zusätzliche Nebenbedingungen.

Beispiel:

Es sind nun - wie oben - die Randsummen x_{i++} vorgegeben. Das log-lineare Modell muss daher die folgenden Nebenbedingungen erfüllen:

$$x_{i++} = \sum_{j,k} exp[\mu + \mu_{A(i)} + \mu_{B(j)} + \mu_{C(k)} + \mu_{AB(i,j)} + ...]; \quad i = 1, ...I \tag{3.26}$$

Das funktioniert allerdings nur dann, wenn der Haupteffekt $\mu_{A(i)}$ im Modell enthalten ist.

Beispiel:

Nun seien die Randsummen für 2 Merkmale vorgegeben. Es gilt:

$$x_{ij+}; \qquad i = 1, ...I; \quad j = 1, ...J \tag{3.27}$$

Die Nebenbedingungen sind also gegeben durch:

$$x_{ij+} = \sum_k exp[\mu + \mu_{A(i)} + \mu_{B(j)} + ...] \qquad i = 1, ...I \quad j = 1, ...J \tag{3.28}$$

Das heisst, alle zulässigen Modelle müssen die Interaktion $\mu_{AB(i,j)}$ enthalten. Auch müssen natürlich alle Haupteffekte im Modell vorhanden sein.

Allgemeine Beschränkung:

Die allgemeine Beschränkung für die Zulässigkeit von Modellen ist folgendermaßen definiert: Wird eine Randsumme x_M für eine Merkmalskombination M festgelegt, dann muß das Modell den Paramter μ_M enthalten. Entspricht M z.B. der Kombination BC, d.h der Festlegung von x_{+jk}, dann sind nur Modelle zulässig, die μ_{BC} enthalten.

3.4.3 Schätzen und Teststatistiken:

Die Statistiken sind für alle drei Erhebungsschemata dieselben, wenn man sich auf zulässige Modelle beschränkt.

3.4.4 Modellhierarchie:

Die Modellhierarchie ist völlig identisch wie bei der Varianzanalyse.

3.4.5 Höherdimensionale Modelle:

Höherdimensionale Modelle sind analog zu zwei- beziehungsweise drei-dimensionalen Modellen definiert.

Kapitel 4

Log-Lineare Modelle als Spezialfall von GLM:

4.1 Einführendes Beispiel:

Wir haben eine $(2,2)$ Kontingenztafel mit Poisson-Erhebungsschema. Es handelt sich also um ein multiplikatives Poisson-Modell, das wir auch als "Unabhängigkeitsmodell" bezeichnet haben. Das Modell ist gegeben als:

$$ln(m_{ij}) = \mu + \mu_{A(i)} + \mu_{B(j)} \qquad i,j = 1,2 \tag{4.1}$$

In Matrixschreibweise ergibt sich nun folgende (überparametrisierte) Schreibweise:

$$\begin{pmatrix} ln\ m_{11} \\ ln\ m_{12} \\ ln\ m_{21} \\ ln\ m_{22} \end{pmatrix} = \mu \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \mu_{A(1)} \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} + \mu_{A(2)} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} + \mu_{B(1)} \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} + \mu_{B(2)} \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} \tag{4.2}$$

Wichtig ist aber, daß auch die folgenden Nebenbedingungen erfüllt sind.

$$\mu_{A(1)} + \mu_{A(2)} = 0; \qquad \mu_{B(1)} + \mu_{B(2)} = 0 \tag{4.3}$$

Mit Hilfe dieser Nebenbedingungen lässt sich das Modell viel kürzer formulieren. Es gilt dann:

$$\begin{pmatrix} ln\ m_{11} \\ ln\ m_{12} \\ ln\ m_{21} \\ ln\ m_{22} \end{pmatrix} = \mu \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \mu_{A(1)} \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix} + \mu_{B(1)} \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix} \tag{4.4}$$

Das heisst also, falls das Modell stimmt (falls also wirklich Unabhängigkeit zwischen den Merkmalen vorherrscht), dann muß der Vektor der logarithmierten erwarteten Häufigkeiten aus dem Unterraum des R^k sein, der von den Vektoren

$$\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}; \qquad \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}; \qquad \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix} \tag{4.5}$$

aufgespannt wird. Faßt man nun diese Vektoren zu einer $(4, 3)$-Matrix X zusammen, dann kann mit

$$ln\ \underline{m} = \begin{pmatrix} ln\ m_{11} \\ ln\ m_{12} \\ ln\ m_{21} \\ ln\ m_{22} \end{pmatrix}; \qquad \underline{\mu} = \begin{pmatrix} \mu \\ \mu_{A(1)} \\ \mu_{B(1)} \end{pmatrix}; \qquad X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{pmatrix} \tag{4.6}$$

das Modell folgendermaßen angeschrieben werden:

$$ln\ \underline{m} = X\underline{\mu} \tag{4.7}$$

Das heisst, man erhält ein verallgemeinertes lineares Modell.

Der Zusammenhang der log-linearen Tabelle zu den verallgemeinerten linearen Modellen wird noch deutlicher, wenn man die beobachteten Zellhäufigkeiten einer mehrdimensionalen Kontingenztabelle mit y_i (nur ein Index i, der sämtliche Zellen der Tafel durchläuft; $i = 1, ...N$ mit $N \ldots$ Anzahl der Zellen) bezeichnet. Weiters gilt nun,

$$E(y_i) = m_i; \qquad i = 1, ..., N \tag{4.8}$$

also

$$y_i = m_i + \epsilon_i; \qquad i = 1, ..., N \tag{4.9}$$

wobei ϵ wieder als "Residuum" bezeichnet wird.

Für die systematische Komponente des Modelles wählen wir den Ansatz $m_i = exp(x_i\underline{\mu})$, wobei x_i die i-te Zeile der analog zusammengestellten Designmatrix X ist. Man erhält schließlich:

$$y_i = exp(x_i\underline{\mu}) + \epsilon_i \qquad i = 1, ..., N \tag{4.10}$$

also ein spezielles GLM mit Linkfunktion $g(m_i) = ln\ (m_i)$. Es stehen also sämtliche theoretische Resultate von GLMs, die wir bereits hergeleitet haben, auch für log-lineare Modelle zur Verfügung.

4.2 Parameterschätzung und Modellanpassung:

Wir machen in diesem Fall ML-Schätzung. Allerdings existieren oft einfachere Methoden zur direkten Schätzung der Modellparameter.

Beispiel:

Wir betrachten ein Modell mit zwei Merkmalen und ein Poisson-Erhebungsschema. Die Likelihoodfunktion ist gegeben als:

$$L(m_{11}, ..., m_{IJ}; x_{11}, ..., x_{IJ}) = \prod_{i,j} \frac{{m_{ij}}^{x_{ij}}}{x_{ij}!} exp(-m_{ij}) \tag{4.11}$$

Nun stellen wir die Likelihoodfunktion als Exponentialfamilie dar. Es gilt nun also:

$$\begin{aligned} L() &= \prod_{i,j} exp\ (-m_{ij})\ exp\ (x_{ij} ln\ m_{ij} - ln\ x_{ij}!) \\ &= exp\ (\sum_{i,j} x_{ij} ln\ m_{ij})\ exp\ (-\sum_{i,j} ln\ x_{ij}!)\ \prod_{i,j} exp\ (-m_{ij}) \end{aligned} \tag{4.12}$$

Setzt man nun $ln\ m_{ij} = \mu + \mu_{A(i)} + \mu_{B(j)} + \mu_{AB(i,j)}$ ein, so ergibt sich:

$$\begin{aligned} L() &= exp\ (\sum_{i,j} x_{ij}\ (\mu + \mu_{A(i)} + \mu_{B(j)} + \mu_{AB(i,j)})) \\ & exp\ (-\sum_{i,j} ln\ x_{ij}!)\ \prod_{i,j} -(\mu + \mu_{A(i)} + \mu_{B(j)} + \mu_{AB(i,j)}) \end{aligned} \tag{4.13}$$

Die Likelihoodfunktion kann also vereinfacht folgendermaßen formuliert werden:

$$c\ (x_{ij}) a_0(\mu + \mu_{A(i)} + \mu_{B(j)} + \mu_{AB(i,j)})\ exp\ (x_{++}\mu + \sum_i x_{i+}\mu_{A(i)} + \sum_j x_{+j}\mu_{B(j)} + \sum_{i,j} x_{ij}\mu_{AB(i,j)}) \tag{4.14}$$

Die Likelihoodfunktion ist also von der Form einer Exponentialfamilie mit den Parametern μ, $\mu_{A(i)}$, $\mu_{B(j)}$, $\mu_{AB(i,j)}$. Gleiches gilt für multinomial beziehungsweise produktmultinomiale Erhebungsschemata, einzig die Funktionen $c(x_{11}, ..., x_{IJ})$ und $a_0(\mu + \mu_{A(i)} + \mu_{B(j)} + \mu_{AB(i,j)})$ unterscheiden sich. Die Verteilung von $(x_{11}, ..., x_{IJ})$ folgt aber für jedes Erhebungsschema einer Exponentialfamilie. Die Darstellung der Exponentialfamilie ist hier nicht minimal, da die Bedingungen $\sum_i \mu_{A(i)} = \sum_j \mu_{B(j)} = \sum_j \mu_{AB(i,j)} = 0$ nicht berücksichtigt sind. Die Likelihoodfunktion ist daher gegeben als:

$$L(\mu, \mu_{A(1)}, ..., \mu_{A(I-1)}, \mu_{B(1)}, ..., \mu_{B(J-1)}, \mu_{AB(1,1)}, ...\mu_{AB(I-1,J-1)}; x_{11}, ..., x_{IJ}) = c(x_{11}, ..., x_{IJ})\ {a_0}^*\ (\mu, ..., \mu_{AB(I-1,J-1)})\ exp\ (x_{++}\mu + \sum_{i=1}^{I-1} \mu_{A(i)}(x_{i+} - x_{I+}) + \sum_{j=1}^{J-1} \mu_{B(j)}(x_{+j} - x_{+J}) + \sum_{i,j=1}^{I-1;J-1} \mu_{AB(i,j)}(x_{ij} - x_{I+} - x_{+J} + x_{IJ})) \tag{4.15}$$

$(x_{11}, ..., x_{IJ})$ gehören zu einer einfachen Exponentialfamilie, folglich ergeben sich die ML-Gleichungen durch:

- $\hat{m}_{++} = x_{++}$
- $\hat{m}_{i+} - \hat{m}_{I+} = x_{i+} - x_{I+} \qquad i = 1, ..., I-1$
- $\hat{m}_{+j} - \hat{m}_{+J} = x_{+j} - x_{+J} \qquad j = 1, ..., J-1$
- $\hat{m}_{ij} - \hat{m}_{i+} - \hat{m}_{+j} + \hat{m}_{++} = x_{ij} - x_{i+} - x_{+j} + x_{++}$

Äquivalent dazu ist die Formulierung $\hat{m}_{ij} = x_{ij}$ aus der sich schließlich ergibt:

- $\hat{\mu} = \frac{1}{I\ J} \sum_{i,j}^{I,J} ln\ (x_{ij})$
- $\hat{\mu}_{A(i)} = \frac{1}{J} \sum_j ln\ (x_{ij}) - \hat{\mu}$
- $\hat{\mu}_{B(j)} = \frac{1}{I} \sum_i ln\ (x_{ij}) - \hat{\mu}$
- $\hat{\mu}_{AB(i,j)} = ln\ (x_{ij}) - (\hat{\mu} - \hat{\mu}_{A(i)} - \hat{\mu}_{B(j)})$

Beim Unabhängigkeitsmodell ist $\hat{\mu}_{AB(i,j)} = 0$, das heißt, man erhält die Gleichungen:

$$\hat{m}_{++} = x_{++} \qquad \hat{m}_{i+} = x_{i+} \qquad \hat{m}_{+j} = x_{+j} \tag{4.16}$$

Auflösen der Gleichungen nach $\hat{m}_{ij}$ mit Berücksichtigung der Bedingungen des Unabhängigkeitsmodells $m_{ij} = \frac{m_{i+}\ m_{+j}}{m_{++}}$ liefert die folgenden Gleichungen:

- $\hat{m}_{ij} = \frac{x_{i+}\ x_{+j}}{x_{++}}$
- $\hat{\mu} = \frac{1}{I}\sum_i ln\ (x_{i+}) + \frac{1}{J}\sum_j ln\ (x_{+j}) - ln\ (x_{++})$
- $\hat{\mu}_{A(i)} = ln\ (x_{i+}) - \frac{1}{I}\sum_i ln\ (x_{i+})$
- $\hat{\mu}_{B(j)} = ln\ (x_{+j}) - \frac{1}{J}\sum_j ln\ (x_{+j})$

Im 2-dimensionalen Unabhängigkeitsmodell lässt sich $\hat{m}_{ij}$ darstellen als Produkt beziehungsweise Quotient von den Randsummen x_{i+}, x_{+j}, x_{++}. Das ist eine wesentliche Eigenschaft und Grund für die Berechnung der ML-Schätzer. Schon im 3-dimensionalen Modell ohne 3er-Interaktionen $(AB|AC|BC)$ ist das aber nicht mehr möglich (kein multiplikatives Modell!), hier sind iterative Verfahren zur Bestimmung der ML-Schätzer nötig. Es gibt also zwei Klassen von log-linearen Modellen, solche für die direkte Schätzmethoden existieren (multiplikative Modell) und solche, die nur indirekte (iterative) Schätzverfahren zulassen.

4.3 Problem der leeren Zellen:

Normalerweise wird im log-linearen Modell davon ausgegangen, dass alle Merkmalskombinationen positive Auftrittswahrscheinlichkeiten besitzen. Dann sind nämlich in genügend großen Stichproben keine leeren Zellen zu erwarten. In der Praxis treten insbesondere bei höherdimensionalen Kontingenztafeln leere Zellen auf. Für eine leere Zelle muß die geschätzte zu erwartende Häufigkeit nicht Null sein. Probleme gibt es, wenn mehrere Zellen leer sind und insbesondere dann, wenn dadurch Randsummen Null werden. Dann erhält man für bestimmte Modelle auch Schäzungen $\hat{m}_i$, die Null sind. In diesem Fall ist aber eine Schätzung der μ-Paramter nicht mehr möglich ($ln(0) = ?$).

4.3.1 Konzepte zum Umgehen dieser Schwierigkeiten:

- zu jeder Zelle $\frac{1}{2}$ addieren

- leere Zellen mit $\frac{1}{R}$ besetzen, wobei R die Anzahl der Zellen ist.

4.4 Anpassungs-Tests:

Hat man $\hat{m}_i$ geschätzt ($i \ldots$ Multiindex; für m-Merkmale gilt: $i = i_1, ...i_m$), dann kann man die Modellanpassung für alle Modelle M und für alle Erhebungsschemata einheitlich testen:

- Person-Statistik: $\chi^2(M) = \sum_i \frac{(\hat{m}_i - x_i)^2}{\hat{m}_i}$
- Likelihood-Quotienten Statistik: $lq(M) = 2 \sum_i x_i ln\ (\frac{x_i}{\hat{m}_i})$

Der Likelihoodquotient $lq\ (M)$ ist eine Transformation des Likelihoodquotienten $\lambda(M)$, der Maximum-Likelihood des Modells M dividiert durch die Maximum-Likelihood des saturierten Modells. Es gilt also:

$$lq(M) = -2ln\ (\lambda(M)) \tag{4.17}$$

$\chi^2(M)$ und $lq(M)$ sind asymptotisch mit denselben Freiheitsgraden nach χ^2-verteilt, d.h. bei großen Stichproben haben die beiden Statistiken annähernd den gleichen Wert. Für die Freiheitsgrade gilt nun also: df = Anzahl der Zeilen - Anzahl der geschätzten Parameter. Achtung: Es interessiert hier aber nur die Anzahl der tatsächlich geschätzten Parameter. Parameter, die sich aus den Restriktionen des untersuchten Modells bzw. des gewählten Erhebungsschemas ergeben, werden nicht geschätzt. Es interessiert also die minimale Anzahl von Schätzungen, mit denen alle Parameter des Modells bestimmt sind.

Beispiel:

Wir betrachten ein 3-dimensionales Modell in einem Poissonerhebungsschema. Das gewählte Modell ist: $AB|BC$. Die 3 Merkmale haben folgende Ausprägungen:

$$A : i = 1, .., I; \qquad B : j = 1, .., J; \qquad C : k = 1, .., K \tag{4.18}$$

Die Modellgleichung ist also:

$$ln\ m_{ijk} = \mu + \mu_{A(i)} + \mu_{B(j)} + \mu_{C(k)} + \mu_{AB(ij)} + \mu_{BC(jk)} \tag{4.19}$$

Die Anzahl der zu schätzenden Parameter (=Freiheitsgrade für dieses Modell) ergibt sich nun als:

$$1 + (I-1) + (J-1) + (K-1) + (I-1)(J-1) + (J-1)(K-1) = J(I-1)(K-1)$$

Beispiel:

Wir betrachten ein produkt-multinomiales Schema, wo nach den Merkmalen B und C geschichtet wird. Die Randsummen für B und C sind also vorgegeben. Das Modell, das wir betrachten, ist das gleiche, wie im obigen Beispiel. In diesem Beispiel hat man $(J+K-1)$-Freiheitsgrade weniger als das Poissonschema, da eben im produkt-multinomialen Erhebungsschema gewisse Randsummen fixiert werden.

$\chi^2(M)$ und $lq(M)$ sind asymptotisch nach χ^2 verteilt, d.h. für gültige Aussagen sind gewisse Mindestgrößen des Stichprobenumfanges nötig. Ein Stichprobenumfang wird für die Anwendung eines Anpassungstest als hinreichend groß angesehen, falls die kleinste zu erwartende Zellbesetzung den Wert 5 (bei manchen Autoren 3) nicht unterschreitet.

4.4.1 leere Zellen:

Falls leere Zellen so angeordnet sind, dass Randsummen, die bei multiplikativen Modellen zum Faktorisieren verwendet werden, Null sind, dann muß das auch bei den Freiheitsgraden des Modells berücksichtigt werden. Dies ist klar, weil dann sämtliche zu erwartenden Zellhäufigkeiten der betroffenden Zellen Null sein müssen. Es ergibt sich daher eine Einschränkung der zu schätzenden Parameter. Mathematisch wird diese Einschränkung definiert als

$$df = (Z - Z_0) - (P - P_0) \tag{4.20}$$

mit:

- Z... Anzahl der Zellen
- Z_0... Anzahl der Zellen, wo gilt: $\hat{m}_i = 0$
- P... Anzahl der freien Parameter unter Berücksichtigung der Restriktionen

- P_0... Anzahl der Parameter, die nicht geschätzt werden können, weil eine Randsumme den Wert Null annimmt.

4.5 Submodelle gegeneinander testen:

Mit $\chi^2(M)$ oder $lq(M)$ wird das Modell M gegen das saturierte Modell getestet. Will man zwei verschiedene log-lineare Modelle gegeneinander testen (LQ-Test), dann geht das im Allgemeinen nur, wenn M_1 ein Submodell von M_2 ist, d.h. die Parametermenge von M_1 ist eine echte Teilmenge der Parametermenge von M_2. Es gilt also: $M_1 \subset M_2$.

Seien $\hat{m}_i$ die ML-Schätzer von M_1 und $\hat{\hat{m}}_i$ die ML-Schätzer von M_2, dann gilt:

$$lq(M_1) = 2 \sum_i x_i ln\ (\frac{x_i}{\hat{m}_i}) \qquad lq(M_2) = 2 \sum_i x_i ln\ (\frac{x_i}{\hat{\hat{m}}_i}) \tag{4.21}$$

Aus $M_1 \subset M_2$ folgt, dass $lq(M_1) \geq lq(M_2)$ und für die Differenzenbildung folgt:

$$lq(M_1) - lq(M_2) = 2 \sum_i x_i\ ln\ (\frac{x_i}{\hat{m}_i}) - 2 \sum_i x_i\ ln\ (\frac{x_i}{\hat{\hat{m}}_i}) \tag{4.22}$$

Dies lässt sich aber vereinfachen (Logarithmusregeln!) und es ergibt sich schließlich:

$$lq(M_1) - lq(M_2) = 2 \sum_i x_i\ (ln\ (\frac{x_i}{\hat{m}_i}) - ln\ (\frac{x_i}{\hat{\hat{m}}_i}) = 2 \sum_i x_i\ ln\ (\frac{\hat{m}_i}{\hat{\hat{m}}_i}) \tag{4.23}$$

$lq(M_1|M_2)$ ist also die bedingte Teststatistik für M_1 unter der Bedingung "M_2 gilt". Die Nullhypothese lautet nun: $H_0 : M_1$ gilt. Unter H_0 besitzt $lq(M_1|M_2)$ eine asymptotische χ^2-Verteilung mit Freiheitsgraden df = Anzahl der Parameter von M_2 - Anzahl der Parameter von M_1. Man kann also $lq(M_1|M_2)$ als bedingtes Maß für die Güte der Anpassung verwenden. Dies ergibt sich auch dadurch, dass $lq(M_1|M_2)$ auf folgende Weise dargestellt werden kann.

$$lq(M_1) = lq(M_2) + lq(M_1|M_2) \tag{4.24}$$

Ist $lq(M_1|M_2)$ klein, dann tragen die zusätzlichen Parameter von M_2 wenig zur Verbesserung der Anpassung bei, d.h. $lq(M_1|M_2)$ testet die Signifikanz der Parameter, die in M_2 aber nicht in M_1 vorkommen.

Dueses Prinzip der Aufteilung der Güte der Modellanpassung auf Gruppen von Parametern kann auf ganze Hierarchien von log-linearen Modellen der Form $M_1 \subset M_2 \subset ... \subset M_m$ verallgemeinert werden. Es gilt dann:

$$lq(M_1) = lq(M_m) + lq(M_{m-1}|M_m) + ... + lq(M_1|M_2) \tag{4.25}$$

Diese Zerlegung der lq-Teststatistik kann auch zur Modellwahl in hierarchisch angeordneten Modellen verwendet werden.

4.6 Modellsuche:

Bis jetzt haben wir die Gültigkeit eines konkreten Modells beziehungsweise H_0 : bestimmte (Interaktions-) Parameter sind Null getestet. Die Methode war ein Anpassungstest zum Niveau α, wobei α die Wahrscheinlichkeit ist, einen Fehler 1. Art zu machen.

Bei der Modellsuche ist nun aber die Situation anders. Man will feststellen, welche unter allen im Rahmen des log-linearen Modells beschriebenen Zusammenhängsstrukturen bestehen. Bei einem solchen explorativen Vorgeben muß man im Allgemeinen mehrere Tests (= zusammengesetzter Test) durchführen, wobei sich dann Probleme ergeben, das Signifikanzniveau des Gesamttests angeben zu können. Das Problem ergibt sich vor allem deshalb, weil die Einzeltests voneinander nicht unabhängig sind. Ein exaktes Testen der auf exporativem, datengesteuertem Weg gefundenen Modellen kann dann nur auf Grundlage einer (neuen) Stichprobe erfolgen. Wir unterscheiden nun 2 verschiedene Szenarien.

- Die Modellhierarchie ist vorgegeben:
  $M_1 \subset M_2 \subset ... \subset M_m$ ist also ausformuliert und die Modelle sind bekannt. Ausgegangen wird von der Zerlegung der Anpassungsstatistik:

  $$lq(M_1) = lq(M_m) + lq(M_{m-1}|M_m) + ... + lq(M_1|M_2) \tag{4.26}$$

  Man betrachtet zu einem vorher festgesetzten Niveau α sukkzessive die Signifikanz der Anpassungsstatistik $lq(M_k)$ und der "Abweichungsstatistik" $lq(M_k|M_{k+1})$. Begonnen wird beim allgemeinsten Modell M_m. Ist $lq(M_m)$ signifikant (großer Wert der

Teststatistik → das Modell passt nicht zu den Daten), dann ist auch das allgemeinste Modell der Hierarchie ungeeignet, die Daten zu erklären. Ist $lq(M_m)$ jedoch nicht signifikant (kleiner Wert der Teststatistik → das Modell passt gut zu den Daten), dann betrachtet man $lq(M_{m-1}|M_m)$ und $lq(M_{m-1})$. Ist eine der beiden χ^2-verteilten Statistiken signifikant, wählt man das Modell M_m (und hört zu testen auf), ansonsten betrachtet man $lq(M_{m-2}|M_{m-1})$ und $lq(M_{m-1})$.

Warum bleibt man im Fall der Signifikanz einer der beiden Teststatistiken beim allgemeineren Modell? Ist $lq(M_{m-1}|M_m)$ signifikant, dann bedeutet das, dass die weggelassenen Interaktionen (Parameter, von M_m, die nicht in M_{m-1} enthalten sind) relevant sind (= von Null verschieden!). Ist $lq(M_{m-1})$ signifikant, dann deutet das auf mangelnde Anpassung des speziellen Modells (M_{m-1}) hin. Dieses Vorgehen ist heuristisch und erlaubt keine Bestimmung einer Wahrscheinlichkeit für den Fehler 1. Art.

Hilfsmittel: ist die Berechnung der Wahrscheinlichkeit für den experimentweisen Fehler. Die Modellwahl ist ein multiples Testproblem in einer vorgegebenen Modellhierarchie mit Nullhypothesen $H_k : M_k$ stimmt und den Alternativhypothesen $H_{k+1} : M_{k+1}$ stimmt. Die Wahrscheinlichkeit, dass mindestens eine der Nullhypothesen fälschlicherweise verworfen wird, ist definiert als die Wahrscheinlichkeit für den experimentweisen Fehler α_{EW}.

Sei α_k das Niveau des Tests H_k gegen H_{k+1} (die dazugehörige Teststatistik ist: $lq(M_k|M_{k+1})$). Sei weiters $k^* := max\ \{k \in N : H_k\ wird\ verworfen\}$. Das heisst, dass alle Modelle M_k mit $k > k^*$ sind gültige Modelle, alle Modelle M_k mit $k \leq k^*$ sind ungültige Modelle. Es gilt nun. $lq(M_1)$ und $lq(M_2)$ sind verteilt nach χ^2 mit df_1 beziehungsweise df_2 Freiheitsgraden. Die bedingte Teststatistik für M_1 gegeben M_2 ist ebenfalls verteilt nach χ^2 mit $df = df_1 - df_2$ Freiheitsgraden. $lq(M_2)$ und $lq(M_1|M_2)$ sind asymptotisch unabhängig, was aus dem Additionstheorem der χ^2-Verteilung folgt.

Allgemein gilt nun, dass die Teststatistiken $lq(M_k|M_{k+1})$ asymptotisch unabhängig

sind und $\alpha_{EW} = 1 - \prod_{k=1}^{m-1}(1 + \alpha_k)$ gilt.

- Die Modellhierarchie ist nicht vorgegeben:

  Hier geht es um simulatane Tests der Ordnung k. Die Modellwahl passiert in 2 Stufen. Zuerst sucht man ein Basismodell, das dann weiter analysiert wird.

  Definition: "Modell der Ordnung k: M(k)"

  Das hierarchische Modell, das sämtliche k-Faktoren/Interaktionen enthält aber keine Interaktion höherer Ordnung nennt wird Modell der Ordnung k genannt. Für m Merkmale (m-dimensionale Tabelle) erhält man die Modellhierarchie M(0) $\subset$ M(1) $\subset$ ... $\subset$ M(m).

  - M(0) ... keine Haupteffekte, keine Interaktionen
  - M(1) ... nur Haupteffekte
  - M(m) ... Haupteffekte und 2er Interaktionen

  Mit der bedingten Teststatistik $lq\ (M(k-1)|M(k))$ prüft man die Signifikanz aller Interaktionen an denen k-Faktoren beteiligt sind gleichzeitig. Daher auch der Name "simultaner Test".
  $k^* = max\ \{k \in N :\ lq\ (M(k-1)|M(k))\ ist\ signifikant\}$ deutet auf die Ordnung des Modells hin, üblicherweise wählt man M(k^*) als Basismodell. Will man das Gesamtniveau α kontrollieren (vgl. oben), dann ist für die einzelnen Teststatistiken (lokales Signifinanzniveau) α_i folgendermaßen zu wählen:

$$\alpha_{EW} = 1 - \prod_{i=0}^{m-1}(1 - \alpha_i) = 1 - (1 - \alpha_i)^m \Leftrightarrow \alpha_i = 1 - (1 - \alpha_{EW})^{\frac{1}{m}} \tag{4.27}$$

Zu einem Basismodell der Ordnung k kann man auch mit Hilfe der Anpassungsstatistik $lq\ (M(k))$ kommen.

4.6.1 Stufenweises Aufbauverfahren:

Man beginnt mit dem Modell $M(0)$ und geht schrittweise zu Modellen höherer Ordnung über, bis ein Modell den Daten zum vorher festgelegten Signifikanzniveau entspricht.

4.6.2 Stufenweises Abbauverfahren:

Das stufenweise Abbauverfahren funktioniert analog zum stufenweisen Aufbauverfahren, allerdings in die umgekehrte Richtung.

4.6.3 Feinsuchverfahren:

Ausgehend von einem Basismodell wird die Modellsuche abgeschlossen. Ein solches Verfahren ist die sogenannte Testprozedur von Goodman. Genauso wie beim Testen von Submodellen gegeneinander benutzt dieses Verfahren die χ^2-Zerlegung der lq (M)-Statistik. Als Basismodell kann auch das Nullmodell $M(0)$ bei Vorwärtsselektion und das saturierte Modell $M(m)$ bei Rückwärtselimination gewählt werden. Diese Vorgehensweise ist allerdings nicht empfehlenswert, da es effizienter ist, ein Basismodell $M(k)$ zu wählen, von dem man für die Modellsuche ausgeht.

4.6.4 Vorwärtsselektion:

Sei M^1 ein nicht-saturiertes Modell (hierarchisch) und seien ${M^2}_t; \quad t = 1, .., T$ die Modelle, die sich von M^1 durch die Hinzunahme eines Haupteffektes oder einer Interaktion ergeben (Supermodelle von M^1). Mit der Teststatistik $lq(M^1|{M^2}_t)$ kann man die zusätzlichen Effekte auf ihre Signifikanz prüfen. Darauf baut die Prozedur von Goodman, deren einzelnen Schritte nun beschrieben werden, auf.

1. Wähle ein Signifikanzniveau α
2. Wähle ein Basismodell M^1
3. Berechne lq $({M^2}_t) \quad \forall t = 1, ..., T$ und wähle das Modell ${M^2}_{t*}$, das den Daten am besten angepasst ist (minimales lq $({M^2}_t)$
4. Berechne lq $(M^1|{M^2}_{t*})$ und teste den zusätzlichen Parameter zum Niveau α.
5. Ist lq $(M^1|{M^2}_{t*})$ signifikant oder das Modell ${M^2}_{t*}$ den Daten schlecht angepasst, setzte $M^1 = {M^2}_{t*}$ und gehe zu Schritt 3, ansonsten wähle M^1 als bestes Modell und der Algorithmus endet.

In Schritt 3 wird also derjenige Effekt hinzugenommen, der den signifikantesten Beitrag leistet, also derjenige Effekt für den die lq-Teststatistik am kleinsten ist.

4.6.5 Rückwärtselimination:

Die Rückwärtselimination funktioniert analog, nur in die andere Richtung. Jetzt sind die Modelle $M^2{}_t: \quad t = 1, .., T$ aber Submodelle von M^1, die durch Entfernung eines einzigen Effekts (Interaktion) entstehen. Es wird derjenige Effekt eliminiert, der den kleinsten nicht-signifikanten Beitrag leistet. Natürlich ist auch eine Modifikation wie bei der Varianzanalyse denkbar: Also eine Vorwärtsselektion und nach jedem Auswahlschritt (ein zusätzlicher Effekt wird aufgenommen) folgt ein Ausschlußschritt (ein überflüssiger Effekt wird eliminiert).

4.7 Wh: ML-Schätzung:

Für die ML-Schätzung von m_{ij} im log-linearen Modell ergeben sich die entsprechenden ML-Gleichungen dadurch, dass die gemeinsame Dichte die Form einer Exponentialfamilie hat.

Satz:

Sind $X = (x_1, ..., x_N)$ unabhängig mit Dichtefunktion

$$f_n(x_n|\theta) = c_n(x_n)\ exp\{\theta^T\ t_n(x_n) - b_n(\theta)\} \qquad n = 1, ..., N \tag{4.28}$$

aus den Exponentialfamilien mit $t_n(x_n) = (t_{n1}(x_n), ..., t_{nm}(x_n))^T$, dann ist die gemeinsame Dichte von $X = (x_1, ..., x_N)^T$ gegeben durch

$$f_n(x_n|\theta) = C(X)\ exp\{\theta^T\ t^{(N)}(X) - B(\theta)\} \tag{4.29}$$

wobei

$$C(X) = \prod_{n=1}^{N} c_n(x_n); \qquad B(\theta) = \sum_{n=1}^{N} b_n(\theta); \qquad t^{(N)}(X) = \sum_{n=1}^{N} t_n(x_n) \tag{4.30}$$

ist, also wieder aus einer Exponentialfamilie in θ.

4.7.1 ML-Schätzung bei Exponentialfamilien:

$X = (x_1, ..., x_N)^T$ sind unabhängig verteilt mit Dichte wie oben. Dann ist die Log-Likelihood Funktion gegeben durch:

$$l(\theta, X) = ln\ C(X) + \theta^T\ t^{(N)}(X) - B(\theta) = \sum_{n=1}^{N} ln\ c_n(x_n) + \sum_{j=1}^{m} \theta_j\ t_j^{(N)} - \sum_{n=1}^{N} b_n(\theta) \quad (4.31)$$

Wir differenzieren nun und es ergibt sich:

$$\frac{\vartheta l}{\vartheta \theta} = t^{(N)}(X) - \frac{\vartheta B(\theta)}{\vartheta \theta} \quad (4.32)$$

Nullsetzen führt dann zu folgenden ML-Gleichungen:

$$t^{(N)}(X) = \frac{\vartheta B(\hat{\theta})}{\vartheta \hat{\theta}} \quad (4.33)$$

beziehungsweise äquivalent dazu:

$$t_j^{(N)}(X) = \frac{\vartheta B(\hat{\theta})}{\vartheta \theta_j} = \sum_{n=1}^{N} \frac{\vartheta b(\hat{\theta})}{\vartheta \theta_j} \qquad j = 1, ..., m \quad (4.34)$$

$t^{(N)}$ gehört also zu einer einfachen Exponentialfamilie mit $t^{(N)}(X)$ statt X und $B(\theta)$ statt $b(\theta)$. Daher gilt:

$$\frac{\vartheta B(\hat{\theta})}{\vartheta \hat{\theta}} = E(t^{(N)}(\theta)) \quad (4.35)$$

Für die ML-Gleichungen gilt daher:

$$t^{(N)} = \frac{\vartheta B(\hat{\theta})}{\vartheta \hat{\theta}} = E(t^{(N)}(\hat{\theta})) = \widehat{E(t^{(N)})} \quad (4.36)$$

4.7.2 Anwendung auf das log-lineare Modell:

Wir haben bereits gezeigt, egal welches Erhebungsschema vorliegt, die Likelihoodfunktion ist von der Form einer einfachen Exponentialfamilie mit den Parametern μ, , $\mu_{A(i)}$, $\mu_{B(j)}$

und $\mu_{AB(i,j)}$.

$$L(\mu, ..., \mu_{AB(I-1,J-1)}, x_{11}, ..., x_{IJ}) =$$
$$c(x_{11}, ..., x_{IJ})\ exp\ (x_{++}\ \mu + \sum_{i=1}^{I-1}(x_{i+} - x_{I+})\ \mu_{A(i)} + \sum_{j=1}^{J-1}(x_{+j} - x_{+J})\ \mu_{B(j)} +$$
$$+ \sum_{i=1}^{I-1}\sum_{j=1}^{J-1}(x_{ij} - x_{I+} - x_{+J} + x_{IJ})\ \mu_{AB(ij)})\ {a_0}^*(\mu, ..., \mu_{AB(ij)}) \qquad (4.37)$$

Die Funktionen $t^{(N)}$ sind also:

- $x_{++} = \widehat{E(x_{++})} = \hat{m}_{++}$
- $x_{i+} - x_{I+} = E(\widehat{x_{i+} - x_{I+}}) = \hat{m}_{i+} - \hat{m}_{I+} \qquad i = 1, ..(I-1)$
- $x_{+j} - x_{+J} = E(\widehat{x_{+j} - x_{+J}}) = \hat{m}_{+j} - \hat{m}_{+J} \qquad j = 1, ..(J-1)$
- $x_{ij} - x_{I+} - x_{+J} + x_{IJ} = E(x_{ij} - \widehat{x_{I+} - x_{+J}} + x_{IJ}) = \hat{m}_{ij} - \hat{m}_{I+} - \hat{m}_{+J} + \hat{m}_{IJ}$

mit $i = 1, ..., (I-1)$ und $j = 1, ..(J-1)$.

Kapitel 5

Diskriminanzanalyse:

Die Problemstellung ist folgende: Die Grundgesamtheit besteht aus mehreren Teilgesamtheiten (Gruppen), jedes Objekt gehört genau einer dieser Gruppen an. Ziel ist es nun, ein Objekt aus der Grundgesamtheit, dessen Klassenzugehörigkeit unbekannt ist, aufgrund seiner beobachteten Merkmalsausprägungen der richtigen Klasse zuzuordnen.

5.1 Der allgemeine entscheidungsorientierte Ansatz:

Gegeben sei die Grundgesamtheit $\Omega = \bigcup_{k=1}^{g} \Omega_k$ mit paarweise disjunkten Teilgesamtheiten $\Omega_1, .., \Omega_g \quad g \geq 2$. Jedem Objekt ω wird einerseits die Ausprägung von $x = (x_1, ..., x_p)$ und andererseits ein Index k der Teilgesamtheit Ω_k, aus der es stammt, zugeordnet. Die Aufgabe ist nun: Der unbekannte Index k soll aufgrund der Daten x eindeutig bestimmt werden. Gesucht ist also eine Entscheidungsfunktion

$$e : \Omega_x \mapsto \{1, ..., g\} \qquad x \mapsto \hat{k} = e(x) \tag{5.1}$$

Es gilt nun:

- $\hat{k}$ ist ein Schätzer für den wahren aber unbekannten Klassenindex k.
- $e(x)$ soll so sein, dass "möglichst wenige" Fehlentscheidungen passieren.

Seien nun sowohl x als auch k Zufallsvariablen. Sei weiters $P(k) = P(\omega \in \Omega_k)$ die "a-priori-Wahrscheinlichkeit" für $\omega \in \Omega_k$ (zb: die Anzahl der Elemente in Ω_k dividiert durch die Gesamtanzahl aller Elemente der Grundgesamtheit Ω). $f(x|k)$ sei die Dichte der Verteilung

von x in Ω_k. Die Dichte der Verteilung von x auf Ω ist dann:

$$f(x) = \sum_{k=1}^{g} f(x|k)\ p(k) \tag{5.2}$$

Für das Klassifikationsproblem interessiert vor allem die "a-posteriori-Wahrscheinlichkeit" $f(k|x)$ oder äquivalent dazu $p(k|x)$, die Verteilung des Klassenindex k bei gegebenen Daten x. Nach dem Theorem von Bayes gilt:

$$f(k|x) = p(k|x) = \frac{f(x|k)\ p(k)}{f(x)} \tag{5.3}$$

In der Praxis muß davon ausgegangen werden, dass sowohl $p(k)$ (die a-priori-Wahrscheinlichkeit) als auch $f(x|k)$ (die Likelihood) unbekannt sind und erst aus einer sogenannten "Lernstichprobe" geschätzt werden müssen (in der Lernstichprobe sind sowohl der Klassenindex k als auch die Daten x bekannt). Wir nehmen im Folgenden nun an, dass die Verteilungen $p(k)$ und $f(x|k)$ beziehungsweise deren Parameter bereits geschätzt wurden und daher bekannt sind.

5.1.1 Fehlerklassifikationswahrscheinlichkeiten:

Wir führen Bezeichnungen für Fehlerklassifikationswahrscheinlichkeiten ein. Es gilt:

- Verwechselungswahrscheinlichkeiten / individuelle Fehlerraten:
  $\epsilon_{k,\hat{k}}(e) := P(e(x) = \hat{k}|k)$ mit $k \neq \hat{k}$
- Bedingte Fehlerraten:
  $\epsilon(e|x) := P(e(x) \neq k|x)$
- Gesamtfehlerrate:
  $\epsilon(e) := P(e(x) \neq k)$ (von den Daten x unabhängig!!)
- Trefferrate:
  $1 - \epsilon(e)$

Es gilt:

$$\epsilon(e) = \sum_{k=1}^{g} \sum_{\hat{k} \neq k=1}^{g} \epsilon_{k,\hat{k}}\ p(k) = \int_{\Omega_x} \epsilon(e|x)\ f(x)\ dx \tag{5.4}$$

5.2 Bayes- und ML-Entscheidungsregel:

5.2.1 Bayes-Entscheidungsregel:

Wähle $\hat{k}$ so, dass $p(\hat{k}|x) = max\ \{p(k|x)\}$ für $k = 1, .., g$ beziehungsweise folgt aus dem Bayes-Theorem dass $f(x|\hat{k})\ p(\hat{k}) = max\ \{f(x|k)\ p(k)\}$ für $k = 1, .., g$ ist. $f(x)$ ist hier vernachlässigbar, da es von k unabhängig ist. Bei gegebenem x entscheidet man sich also für jene Klasse, die die größte a-posteriori-Wahrscheinlichkeit besitzt. Das Maximum muss nicht eindeutig sein, das beeinflußt aber die Optimalitätseigenschaften der Entscheidungsregel nicht.

5.2.2 Maximum-Likelihood-Entscheidungsregel:

Der Spezialfall, dass gleiche a-priori-Wahrscheinlichkeiten ($p(1) = p(2) = ... = p(g)$) betrachtet werden, führt zur Maximum-Likelihood-Entscheidungsregel, die folgendermaßen definiert ist:

Wähle $\hat{k} = e(x)$ so, dass $f(x|\hat{k}) = max\ \{f(x|k)\}$ für $k = 1, .., g$ ist. Bei gegebenem x wählt man also $\hat{k}$ so, dass die Likelihoodfunktion $L(\hat{k}|x) = f(x|\hat{k})$ maximal wird. Die ML-Entscheidungsregel ist insbesondere dann zur Klassifikation geeignet, wenn man die Klassenzugehörigkeit nicht als Zufallsvariable sondern als unbekannten Parameter auffasst (keine a-priori-Wahrscheinlichkeiten bekannt!).

Satz: Optimalität der Bayes-Entscheidungsregel

Die Bayes-Entscheidungsregel besitzt (unter allen Entscheidungsregeln) für alle x die kleinste bedingte Fehlerrate und damit auch die kleinste Gesamtfehlerrate.

Beweis:

Die bedingte Fehlerrate ist gegeben durch:

$$\epsilon(e|x) = P(e(x) \neq k|x) = 1 - P(e(x) = k|x) = 1 - p(k|x) \tag{5.5}$$

Für die Bayes-Entscheidungsregel gilt: $p(\hat{k}|x) \geq p(k|x)$ und damit ergibt sich

$$\epsilon(e|x) = P(e(x) \neq k|x) = 1 - P(e(x) = k|x) = 1 - p(k|x) \geq 1 - p(\hat{k}|x) \geq \epsilon(e_B|x) \quad (5.6)$$

also $\epsilon(e|x) \geq \epsilon(e_B|x)$. Der Beweis dass die Bayes-Entscheidungsregel ebenfalls die kleinste Gesamtfehlerrate besitzt ist nun einfach. Da $f(x) \geq 0$ für alle x ist, gilt:

$$\epsilon(e) = \int_{\Omega_x} \epsilon(e|x)\ f(x)\ dx \geq \int_{\Omega_x} \epsilon(e_B|x)\ f(x)\ dx = \epsilon(e_b) \quad (5.7)$$

Folgerung:

Die ML-Entscheidungsregel ist optimal, wenn die a-priori-Wahrscheinlichkeiten gleich groß sind, d.h. wenn keine Vorinformation über k vorliegt.

5.3 Kostenfunktionen:

Bei der Minimierung der Fehlerraten wurden alle Fehlklassifikationen gleich bewertet, egal aus welcher Klasse das Objekt stammt und welcher Klasse das Objekt zugeordnet wird. Unterschiedliche Bewertungen von Fehlklassifikationen kann man mit Kostenfunktionen erledigen. Das ist wichtig in der Praxis, da etwa die Einstufung eines Kranken als gesund schwerwiegender ist als die Einstufung eines Gesunden als krank.

Definition:

Seien $C(k, \hat{k}) :=$ jene Kosten, die entstehen, wenn die Entscheidungsregel auf $\hat{k}$ fällt, aber k stimmt. Das impliziert, dass $C(k, k) = 0$ ist. Sei nun $e(x) = \hat{k}$ dann ergeben sich daraus die bedingten erwarteten Kosten als

$$C(\hat{k}|x) = \sum_{k=1}^{g} C(k, \hat{k})\ p(k|x) \quad (5.8)$$

beziehungsweise äquivalent dazu als:

$$\overline{C}(\hat{k}|x) = \sum_{k=1}^{g} C(k, \hat{k})\ f(x|k)\ p(k) \quad (5.9)$$

Das heisst also, die Kosten werden mit ihren a-posteriori-Wahrscheinlichkeiten gewichtet.

5.4 Kostenoptimale Entscheidungsregel:

Wähle $\hat{k}$ so, dass die bedingten Kosten C bzw. $\overline{C}$ minimal werden.

Folgerung:

Sind für alle x die bedingten Kosten minimal, dann sind auch die Gesamtkosten minimal (gleiches Argument wie im Beweis des obigen Satzes!). Die *Gesamtkosten* ergeben sich als

$$C(\hat{k}) = \int_{\Omega_x} C(\hat{k}|x)\ f(x)\ dx \tag{5.10}$$

beziehungsweise als:

$$\overline{C}(\hat{k}) = \int_{\Omega_x} \overline{C}(\hat{k}|x)\ f(x)\ dx \tag{5.11}$$

5.4.1 Spezielle Kostenfunktionen:

- einfache symmetrische Kostenfunktion
  $c_e(k,\hat{k}) = 0$ für $k = \hat{k}$; $\quad c_e(k,\hat{k}) = c \geq 0$ für $k \neq \hat{k}$
  Fehlklassifikationen werden mit gleichen Kosten bewertet. Die kostenoptimale Entscheidungsregel für diese Kostenfunktion ist die Bayes-Entscheidungsregel.

- umgekehrt-proportionale Kostenfunktion
  $c_p(k,\hat{k}) = 0$ für $k = \hat{k}$; $\quad c_e(k,\hat{k}) = \frac{c}{p(k)}$ für $k \neq \hat{k}$
  Hier werden die Kosten für die Fehlklassifikation von Objekten aus Klassen mit geringer a-priori-Wahrscheinlichkeit stark vergrößert (z.b: selten auftretende, ernste Krankheiten). Die optimale Entscheidungsregel für diese Kostenfunktion ist die ML-Entscheidungsregel.

5.5 Die Diskriminanzfunktion:

Die obigen Entscheidungsregeln basieren auf folgendem Prinzip. Berechne für jedes gegebene x (Daten) die g Diskriminanzfunktionen

$$d_1(x), ..., d_g(x): \quad \Omega_x \mapsto R \tag{5.12}$$

Die Einheit ω wird dann $\Omega_{\hat{k}}$ zugeordnet mit:

$$d_{\hat{k}}(x) = max\{d_k(x)\} \quad k = 1, .., g \tag{5.13}$$

Für jede Entscheidungsregel gibt es eine eigene Diskriminanzfunktion:

- Bayes-Entscheidungsregel: $d_k(x) = p(k|x)$ oder äquivalent $f(x|k)\ p(k)$
- ML-Entscheidungsregel: $d_k(x) = f(x|k)$
- Kostenoptimale Regel: $d_k(x) = -C(k|x)$ oder äquivalent $-\overline{C}(k|x)$

Für manche Anwendungen ist es günstiger mit $f(d_1(x), ..., f(d_g(x))$ anstelle der Diskriminanzfunktionen $d_1(x, ..., d_g(x)$ zu rechnen, wobei f eine streng monoton (monoton wachsende) Funktion ist.

Beispiel:

Sei $(x|k) \sim N(\mu, \sigma^2)$, dann ist die Bayes-Diskriminanzfunktion gegeben als:

$$d_k(x) = p(k)\ f(x|k) \tag{5.14}$$

Äquivalent dazu, allerdings leichter auszurechnen ist folgende Darstellung

$$ln\ d_k(x) = ln\ p(k) + ln\ f(x|k) \tag{5.15}$$

5.6 Klassengebiete:

Durch jede Entscheidungsregel wird Ω_x in disjunkte Klassengebiete $D_1, ..D_g$ zerlegt (es gilt dann: $\Omega_x = \bigcup_{i=1}^{g} D_i$), die folgendermaßen definiert sind:

Vorbemerkung:

Sei $\circ A$ das "Innere" von D_k, ϑD_k der "Rand" von D_k und $\overline{D_k}$ der "Abschluss" von D_k, wobei gilt: $\overline{D_k} = \overline{D_k} \bigcup \vartheta D_k$.

$$\circ D_k := \{x \in \Omega_x | d_k(x) > d_i(x); \forall i \neq k\} = D_k \setminus\ \vartheta D_k \tag{5.16}$$

In D_k ist die Diskriminanzfunktion der k-ten Klasse maximal. ϑD_k sind die Trennflächen zwischen den Klassengebieten, für die gilt: $\exists i = k \quad d_i(x) = d_k(x)$. Auch diese Trennflächen (Hyperflächen) müssen eindeutig einem Klassengebiet zugeordnet werden, damit die Zerlegung von Ω_x wohldefiniert ist (etwa die Trennflächen der Klasse mit dem niedrigsten/höchsten Index zuordnen).

Wozu macht man das?:

Oft wird anstelle der Entscheidungsregel e die durch die Entscheidungsregel induzierte Zerlegung von Ω_x angegeben. Es gilt:

$$e(x) = \hat{k} \Longleftrightarrow x \in D_{\hat{k}} \tag{5.17}$$

Verwendet man die Zerlegung in Klassengebiete, dann kommt man zu anderen, einfacheren Darstellungen der Fehlerraten.

Beispiel: Gesamtfehlerrate; g=2 (2 Klassengebiete)

Es gilt also:

$$\epsilon(e) = \epsilon(D_k, f) = P(e(x) \neq k) = P(\omega \; wird \; fehlklassifiziert) \tag{5.18}$$

Da $\Omega_x = (D_1, D_2)$ ist, gilt:

$$\begin{gathered} P(x \in D_2, k = 1) + P(x \in D_1, k = 2) = \\ P(x \in D_2 | k = 1) \; p(1) + P(x \in D_1 | k = 2) \; p(2) \\ \int_{D_2} f(x|1) \; p(1) \; dx + \int_{D_1} f(x|2) \; p(2) \; dx \end{gathered} \tag{5.19}$$

Beispiel: individuelle Fehlerraten

Es gilt:

$$\epsilon_{k,\hat{k}} = P(x \in D_{\hat{k}} | k) = \int_{D_{\hat{k}}} f(x|k) \; dx \tag{5.20}$$

5.7 Geschätzte Entscheidungsregeln und Fehlerraten:

$p(k)$, $f(x|k) \rightarrow p(k|x)$ wurden bisher als bekannt vorausgesetzt. In der Praxis müssen diese Verteilungen bzw. deren Parameter erst geschätzt werden. Meist trifft man dabei gewisse Verteilungsannahmen und schätzt dann die entsprechenden Parameter. Man setzt also die Verteilungen, zb: $f(x|\Theta_k)$, $p(k|x,\Theta)$ oder auch die Diskriminanzfunktionen $d_k(x|\Theta_k)$ in parametrischer Form an.

5.8 Schätzmethoden für die Parameter:

ML-Schätzer, LS-Schätzer und andere Schätzverfahren sind möglich und zulässig, es gibt hier keinerlei Einschränkungen. Die geschätzten Entscheidungsregeln sind von der Schätzmethode und der Stichprobenerhebungsart abhängig.

- ML-Schätzung + uneingeschränkte Zufallsauswahl (Gesamtstichprobe)
  $(x_i, k_i) \quad i = 1, ..N$ sind unabhängige Beobachtungen der Zufallsvariablen (x, k). Für die Likelihoodfunktion ergibt sich, wenn man sie über die a-priori-Wahrscheinlichkeit $p(k)$ berechnet:

$$L_G = \prod_{i=1}^{N} f(x_i, k_i) = \prod_{i=1}^{N} f(x_i|k_i) \prod_{i=1}^{N} p(k_i) \tag{5.21}$$

  Anmerkung: Man muss aber beachten, dass $f(x_i|k_i)$ eigentlich auch eine Funktion von Θ ist, also gilt: $f(x_i|\Theta_{ki})$
  L_G ist also eine Funktion von $(\Theta_1, ...\Theta_g, p(k))$. Maximiert man nun L_G so führt das zu den Schätzern $(\hat{\Theta}_1, ...\hat{\Theta}_g, p(k) = \frac{N_k}{N})$ wobei N_k die Anzahl der Beobachtungen in der Klasse k bezeichnet.

  Die Diskriminanzfunktion der Bayes-Entscheidungsregel, die gegeben ist als

$$d_k(x) = f(x|k)\ p(k) \tag{5.22}$$

  wird ersetzt durch die geschätzte Diskriminanzfunktion (ist die a-priori-Verteilung $p(k)$ bekannt, so verwendet man natürlich die bekannte Verteilung und nicht die

geschätzte Verteilung $p(\hat{k})$.

$$d_k(x|\hat{\Theta}_k) = f(x|\hat{\Theta}_k)\ p(\hat{k}) \tag{5.23}$$

Analog ergibt sich die geschätzte ML-Diskriminanzfunktion als:

$$d_k(x|\hat{\Theta}_k) = f(x|\hat{\Theta}_k) \tag{5.24}$$

Die Likelihoodfunktion kann aber genauso über die a-posteriori-Wahrscheinlichkeiten berechnet werden. Es ist dann:

$$L_G = \prod_{i=1}^{N} f(x_i, k_i) = \prod_{i=1}^{N} p(k_i|x_i, \Theta_i) \prod_{i=1}^{N} f(x_i) \tag{5.25}$$

Die gechätzte Bayes-Diskriminanzfunktion ergibt sich dann als:

$$d_k(x|\hat{\Theta}_k) = p(k|x, \hat{\Theta}_k) \tag{5.26}$$

Achtung: Auch die Mischverteilung $f(x)$ kann Informationen über den Parameter Θ_k enthalten! (nicht nur über den ersten Faktor von L_G maximieren!).

- ML-Schätzung einer nach Klassenzugehörigkeit geschichtete Stichprobe:
  Aus jeder der g Klassen werden N_k (fest) unabhängige Beobachtungen x gezogen ($f(x|k)$). Dies ist notwendig, falls einzelne Klassen sehr klein sind. Die Schätzer für diese Klassen wären auch bei großen Stichproben zu ungenau. Mit der Likelihoodfunktion

$$L_K = \prod_{i=1}^{N} f(x_i|\Theta_{ki}) \tag{5.27}$$

  erhält man Schätzer $(\Theta_1, .., \Theta_g)$. Die a-priori-Verteilung $p(k)$ kann nicht geschätzt werden, (N_k ist fest!!), dh. es gibt nur geschätzte ML-Diskriminanzfunktionen. Schwierigkeiten gibt es ebenfalls bei der ML-Schätzung der a-posteriori Wahrscheinlichkeiten $p(k|x, \Theta_k)$.

- ML-Schätzung nach Merkmalsausprägungen von x geschichteter Stichprobe:
  Für N vorgegebene Werte $x_1, ..., x_N$ wird die Klassenzugehörigkeit unabhängig als Realisierung eines Zufallsexperiments $k|x_1, ..., k|x_N$ betrachtet (geplante Versuche,

zb. in der Medizin). Die Parametrisierung passiert naheliegenderweise über die a-posteriori Verteilung. Es gilt:

$$L_X = \prod_{i=1}^{N} p(k|x_i) = \prod_{i=1}^{N} p(k_i|x_i\Theta_{ki}); \qquad p(k|x) = \frac{f(x|k)p(k)}{f(x)} \tag{5.28}$$

Enthält die Mischverteilung $f(x)$ keine Information über die Parameter Θ, dann erhält man dieselben Schätzer wie bei der uneingeschränkten Zufallsauswahl, denn es gilt:

$$L_G = L_X = \prod_{i=1}^{N} f(x_i|\Theta_i) = \prod_{i=1}^{N} f(x_i) \tag{5.29}$$

5.9 Fehlerraten:

Jede theoretische Entscheidungsregel impliziert eine Zerlegung $D = (D_1, ..., D_g)$. Beim Übergang zu geschätzten Entscheidungsregeln muß man dann anstelle der theoretischen (aber unbekannten) Fehlerrate $\epsilon(D, f)$ geschätzte Fehlerraten verwenden.

5.9.1 Theoretische Fehlerrate:

Sei $g = 2$ (eine Verallgemeinerung auf $g > 2$ ist leicht möglich!). Die theoretische Fehlerrate ist definiert als

$$\epsilon(D, f) = p(1) \int_{D_2} f(x|1)dx + p(2) \int_{D_1} f(x|2)dx = \epsilon_{12}(D, f) + \epsilon_{21}(D, f) \tag{5.30}$$

wobei $\epsilon_{12}(D, f)$ sowie $\epsilon_{21}(D, f)$ als "individuelle Fehlerraten" bezeichnet werden. Die theoretische Fehlerrate ist optimal (minimal) bei Verwendung der Bayes-Entscheidungsregel. Jetzt haben wir allerdings geschätzte Entscheidungsregeln.

5.9.2 Tatsächliche Fehlerrate:

Wir betrachten die Zerlegung $\hat{D} = (\hat{D}_1, ..., \hat{D}_g)$ und wir bestimmen die tatsächliche Fehlerrate $\epsilon(\hat{D}, f)$. Es gilt:

$$\epsilon(\hat{D}, f) = p(1) \int_{\hat{D}_2} f(x|1)dx + p(2) \int_{\hat{D}_1} f(x|2)dx = \epsilon_{12}(\hat{D}, f) + \epsilon_{21}(\hat{D}, f) \tag{5.31}$$

Diese Fehlerrate $\epsilon(\hat{D}, f)$ ist eine Zufallsvariable, d.h. in der Praxis sind wir an ihrem Erwartungswert interessiert, also der zu erwartenden tatsächlichen Fehlerrate $E(\epsilon(\hat{D}, f))$. Der Erwartungswert wird über die Zufallsvariablen (k_i, x_i) aus der Lernstichprobe bestimmt.

Das alles ist allerdings noch immer theoretisch. Für die Praxis brauchen wir einen Schätzer für die tatsächliche Fehlerrate. Es bietet sich wieder ein plug-in Schätzer an. Wir ersetzen wirkliche, aber unbekannte Verteilungen durch geschätzte. Es gilt dann:

$$E(\epsilon(\hat{D}, \hat{f})) = \hat{p}(1) \int_{\hat{D}_2} \hat{f}(x|1)dx + \hat{p}(2) \int_{\hat{D}_1} \hat{f}(x|2)dx \tag{5.32}$$

Ist nämlich $\hat{f}(x|k)$ ein erwartungstreuer Schätzer für $f(x|k)$, dann gilt für die Bayes-Regel:

$$E(\epsilon(\hat{D}, \hat{f})) \leq \epsilon(D, f) \leq E(\epsilon(\hat{D}, f)) \tag{5.33}$$

5.9.3 Konvergenz:

Wir betrachten nun die Konvergenz der geschätzten Diskriminanzfunktion und der tatsächlichen Fehlerraten.

Satz:

Gilt $\hat{f}(x|k) \longrightarrow f(x|k)$ für $N \to \infty$ für alle k mit positiver a-priori Wahrscheinlichkeit $p(k)$, dann konvergiert auch die geschätzte Diskriminanzfunktion gegen die optimale Bayes-Diskriminanzfunktion. Also gilt:

$$\hat{p}(k)\hat{f}(x|k) \longrightarrow p(k)f(x|k); \qquad k = 1, ..., g \tag{5.34}$$

Gilt zusätzlich

$$\int_{\Omega} \sum_{k=1}^{g} \hat{p}(k)\hat{f}(x|k)dx \longrightarrow 1 \tag{5.35}$$

(das ist bei parametrischer Schätzung von $\hat{f}(x|k) = f(x|\hat{\Theta}_k)$ immer erfüllt, da $f(x|\hat{\Theta}_k)$ eine Dichte ist), dann konvergiert auch die tatsächliche Fehlerrate $\epsilon(\hat{D}, f)$ gegen die theo-

retische Fehlerrate $\epsilon(D, f)$. Also gilt für $N \to \infty$:

$$\epsilon(\hat{D}, f) \longrightarrow \epsilon(D, f) \tag{5.36}$$

Bemerkung:

Bei der parametrischen Schätzung von $f(x|\hat{\Theta}_k)$ konvergiert $\hat{f}(x|k)$, wenn die Schätzer $\hat{\Theta}_k$ konsistent sind und $f(x|\Theta_k)$ stetig in Θ_k ist. Dann konvergiert (nach obigem Satz) die geschätzte Diskriminanzfunktion.

5.10 Klassische Diskriminanzanalyse:

Wir betrachten nun den Spezialfall, wobei die Merkmale normalverteilt sind. Wir sprechen hier von der sogenannten "klassischen Diskriminanzanalyse". Die Voraussetzungen sind, dass $(x|k) \sim N(\mu_k, \Sigma_k)$, wobei x p-dimensional ist. Bei Verwendung der Bayes-Regel $p(k)f(x|k) \to max$ ergibt sich folgende Diskriminanzfunktion, die wir in logarithmierter Form anschreiben

$$d_k(x) = ln\ p(k) + ln\ f(x|k) = ln\ p(k) + \frac{p}{2}\ ln\ 2\pi - \frac{1}{2}\ ln\ |\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)$$

wobei der Term $\frac{p}{2}ln2\pi$ für die Maximierung unerheblich ist und daher weggelassen werden kann. Die Diskriminanzfunktion für die ML-Regel erhält man, indem man einfach die a-priori Dichte $ln\ p(k)$ weglässt.

Erster Spezialfall:

Sei $\Sigma_k = \sigma^2 I$, das heisst die Merkmale sind unabhängig verteilt mit gleicher Varianz σ^2. Es gilt nun:

$$|\Sigma_k| = \sigma^{2p} \qquad \Sigma_k^{-1} = \frac{1}{\sigma^2}I \tag{5.37}$$

Daher ist

$$d_k(x) = ln\ p(k) + p\ ln\ \sigma - \frac{1}{2\sigma^2}(x-\mu_k)^T(x-\mu_k) \tag{5.38}$$

und da der Term $p\ ln\sigma$ nichts mit der Berechnung des Maximums zu tun hat und daher weggelassen werden kann und damit die letzte quadratische Form vereinfacht werden kann,

ergibt sich schliesslich:

$$d_k(x) = ln\ p(k) - \frac{||x - \mu_k||^2}{2\sigma^2} \tag{5.39}$$

Sind die a-priori Wahrscheinlichkeiten gleich groß bzw. kennt man sie nicht, so kannt man zur ML-Diskriminanzfunktion $d_k(x) = -||x - \mu_k||^2$, das heisst, ω wird zur Klasse k zugeordnet, deren Mittelpunkt μ_k den kleinsten (euklidischen) Abstand zu x hat (Maximum-Distanz-Klassifikation). Die Bayes-Diskriminanzfunktion ist in Wirklichkeit linear zu x. Denn es ist

$$d_k(x) = ln\ p(k) - \frac{1}{2\sigma^2}(||x||^2 - 2{\mu_k}^T x + ||\mu_k||^2) = ln\ p(k) + \frac{1}{\sigma^2}{\mu_k}^T x - \frac{||\mu_k||^2}{2\sigma^2} = a_k^T x + a_{k0}$$

weil ja der Term $\frac{1}{2\sigma^2}(||x||^2)$ ebenfalls nichts mit der Maximierung zu tun hat und weggelassen werden kann. Die Annahme $\Sigma_k = \sigma^2 I$ ist sehr einschränkend, etwas allgemeiner ist der nächste Spezialfall.

Zweiter Spezialfall:

Wir nehmen nun an, dass $\Sigma_k = \Sigma$ ist, wir also klassenweise identische Kovarianzmatrizen vorliegen haben. Die Diskriminanzfunktion ergibt sich daher als:

$$d_k(x) = ln\ p(k) - \frac{1}{2}\ ln\ |\Sigma| - \frac{1}{2}\ (x - \mu_k)^T \Sigma^{-1}(x - \mu_k) \tag{5.40}$$

Der Term $\frac{1}{2} ln\ |\Sigma|$ ist für die Suche des Maximums wieder unerheblich und kann weggelassen werden. Weuters wurd der Term $(x - \mu_k)^T \Sigma^{-1}(x - \mu_k)$ auch "Makalanobis-Distanz" genannt. Wiederum ist die Diskriminanzfunktion in Wirklichkeit linear in x. Es gilt nämlich:

$$d_k(x) = \mu_k^T \Sigma^{-1} x - \frac{1}{2} ln\ p(k) - \frac{1}{2}\ \mu_k^T \Sigma^{-1} \mu_k = a_k^T x + a_{ko} \tag{5.41}$$

Nur bei klassenweise verschiedenen Kovarianzen ist die Diskriminanzfunktion quadratisch in x. Dann gilt:

$$d_k(x) = x^T A_k x + a_k^T x + a_{k0} \tag{5.42}$$

mit

$$A_k = -\frac{1}{2}\ \Sigma_k^{-1} \qquad a_k = \Sigma_k^{-1} \mu_k \qquad a_{k0} = ln\ p(k) - \frac{1}{2}\ |\Sigma_k| - \frac{1}{2}\ \mu_k^T \Sigma_k^{-1} \mu_k \tag{5.43}$$

In den gängigen Programmpaketen wird aber standardmäßig eine "lineare Diskriminanzanalyse" durchgeführt, es liegt hier also die Annahme gleicher Kovarianzmatrizen von x in den k Gruppen zugrunde. Also $(x|k) \sim N(\mu_k, \Sigma)$.

Wie sehen die geschätzten Diskriminanzfunktionen aus?

Man setzt einfach die unverzerrten Schätzer $\overline{x_k}$ für μ_k, für $k = 1, ..., g$ ein. Damit ergibt sich (für Σ)

$$S = \frac{1}{N-g} \sum_{k=1}^{g} \sum_{i=1}^{N_g} (x_{ki} - \overline{x_k})(x_{ki} - \overline{x_k})^T \tag{5.44}$$

woraus die geschätzte Diskriminanzfunktion folgt, die gegeben ist durch:

$$\hat{d_k}(x) = ln\ (N_k) - ln\ N + \overline{x_k}^T S_x^{-1} - \frac{1}{2} \overline{x_k}^T S^{-1} \overline{x_k} \tag{5.45}$$

wobei der Term $ln\ N$ wieder weggelassen kann.

Dritter Spezialfall:

Behandeln wir nun den Spezialfall, wo wir nur $g = 2$ Klassen haben, ergibt sich folgendes Ergebnis. Das Objekt ω wird der Klasse 1 zugeordnet, falls $(x - \frac{1}{2}(\overline{x_1} - \overline{x_2}))^T a > ln\ \frac{p(2)}{p(1)}$ ist, wobei $a = \int^{-1} (\overline{x_1} - \overline{x_2})$ ist. Ist keine a-priori-Verteilung bekannt oder will man anstelle der Bayes-Regel die ML-Regel verwenden, muss man einfach $ln\ p(k)$ Null setzen.

5.11 Verteilungsfreier Ansatz von Fischer:

Gegeben sei der Vektor $(x_1, ...x_p)^T$. Die Idee ist nun, das p-dimensionale Problem auf ein eindimensionales Problem überzuführen. Dies funktioniert mit einer Linearkombination des Merkmalsvektor x: Es gilt dann:

$$y = a^T x \qquad a^T = (a_1, ..., a_p) \tag{5.46}$$

Das heisst, x_{ki} mit $i = 1, ..., N_k$ wird transformiert zu $y_{ki} = a^T x_{ki}$. Für $||a|| = 1$ ist $a^T x$ die Projektion der Daten auf eine Gerade mit Richtung a. Allgemein soll a so gewählt werden, dass $Q(a) = \frac{\overline{y_1} - \overline{y_2}^2}{S_1^2 + S_2^2}$ maximal wird, wobei $\overline{y_k} = a^T \overline{x_k}$ und $S_k^2 = \sum_{i=1}^{N_k} (y_{ki} - \overline{y_k})^2$ ist. Vergleiche dazu die Regressions- und Varianzanalyse.

Alternative Darstellung von $S_1^2 + S_2^2$:

$S_1^2 + S_2^2$ lassen sich auch anders darstellen. Es gilt nämlich

$$S_1^2 + S_2^2 = \sum_{i=1}^{N_1} a^T(x_{1i} - \overline{x_1})(x_{1i} - \overline{x_1})^T a + \sum_{i=1}^{N_2} a^T(x_{2i} - \overline{x_2})(x_{2i} - \overline{x_2})^T a = a^T W a \quad (5.47)$$

woraus folgt:

$$Q(a) = \frac{(a(\overline{x_1} - \overline{x_2}))^2}{a^T W a} \quad (5.48)$$

$Q(a)$ sollte nun durch Variation von a maximiert werden. Dies geschieht durch Ableiten der Funktion nach a und Null setzen, also:

$$\frac{\vartheta Q(a)}{\vartheta a} = \frac{2(a(\overline{x_1} - \overline{x_2}))(\overline{x_1} - \overline{x_2})a^T W W a - 2Wa(a^T(\overline{x_1} - \overline{x_2}))^2}{(a^T W a)^2} = 0 \quad (5.49)$$

Äquivalent dazu ist folgende Darstellung:

$$\Leftrightarrow (\overline{x_1} - \overline{x_2})a^T W a = W a a^T(\overline{x_1} - \overline{x_2})$$

Da $a^T W a$ ein Skalar ist dann durch diese Zahl durchdividiert werden und die Matrix W auf die andere Seite der Gleichung gebracht werden. Daraus folgt:

$$W^{-1}(\overline{x_1} - \overline{x_2}) = a \, \frac{a^T(\overline{x_1} - \overline{x_2})}{a^T W a} \quad (5.50)$$

Die rechte Seite besteht aus dem Vektor a und einem Skalar, der die Richtung von a aber nicht beeinflusst.

Die lineare Fischer-Diskriminanzfunktion $y = a^T x$ ist für $g = 2$ bis auf einen konstanten Term C identisch mit der Diskriminanzfunktion bei Normalverteilungsannahme mit klassenweise identischen Kovarianzmatrizen und unter Voraussetzung der ML-Regel.

Begründung:

Die Fisher-Regel ist ja folgendermaßen definiert. Sei x eine Beobachtung mit unbekanntem Klassenindex. Berechne $y = a^T x$. Das Objekt x gehört dann zur Gruppe 1, falls y näher

bei $\overline{y_1}$ als bei $\overline{y_2}$ liegt und umgekehrt. Äquivalent dazu ist folgende Darstellung:

$$|y - \overline{y_1}| < |y - \overline{y_2}| \Leftrightarrow y > \frac{1}{2}(\overline{y_1} + \overline{y_2}) \Leftrightarrow y - \frac{1}{2}(\overline{y_1} + \overline{y_2}) > 0 \Leftrightarrow a^T(x - \frac{1}{2}(\overline{x_1} + \overline{x_2})) \quad (5.51)$$

Folgerung:

Die lineare Diskriminanzanalyse ist relativ robust. Die Ergebnisse sind auch bei Verletzung der Voraussetzung $\Sigma_k = \Sigma$ brauchbar.

5.11.1 Allgemeiner Fall:

Wir betrachten nun den allgemeinen Fall mit g Klassen. Das Trennkriterium ist nun definiert als:

$$Q(a) = \frac{\sum_{k=1}^{g} N_k(\overline{y_k} - \overline{y})^2}{\sum_{k=1}^{g} S_k^2} \quad (5.52)$$

$Q(a)$ soll nun wie im zweidimensionalen Fall über a maximiert werden. Eine andere Darstellung für $Q(a)$ ergibt sich mit

$$Q(a) = \frac{a^T B a}{a^T W a} \quad (5.53)$$

wobei gilt:

$$W = \sum_{k=1}^{g} W_k; \qquad W_k = \sum_{i=1}^{n_k} (x_{ki} - \overline{x_k})(x_{ki} - \overline{x_k})^T \quad (5.54)$$

$$B = \sum_{k=1}^{g} N(\overline{x_k} - \overline{x})(\overline{x_k} - \overline{x})^T \quad (5.55)$$

Seiten nun $\lambda_1 > \lambda_2 > ... > \lambda_q > 0$ $(q \leq min\ \{g-1, p\})$ die positiven Eigenwerte von $W^{-1}B$ und $a_1, ..., a_q$ die dazugehörigen Eigenvektoren. Dann gilt:

$$y_k = a_k x \quad (5.56)$$

geben in der Reihenfolge $(k = 1, ..., q)$ die vorgegebene Zerlegung am besten wieder. Die y_k können alle (oder auch nur zum Teil) zur Dimensionsreduktion von x verwendet werden. Dann es gilt:

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_r \end{pmatrix} = \begin{pmatrix} a_1^T x \\ \vdots \\ a_r^T x \end{pmatrix} \qquad r \leq q \tag{5.57}$$

Das heisst für die Fischer-Regel: Wähle $\hat{k}$ so, dass

$$\sum_{l=1}^{r} (a_l^T (x - \overline{x_{\hat{k}}}))^2 \leq \sum_{l=1}^{r} (a_l^T (x - \overline{x_k}))^2 \tag{5.58}$$

für $k = 1, ...g$ gilt. Auch diese Regel ist für gleiche a-priori Wahrscheinlichkeiten ($p(1) = p(2) = ... = p(g)$) äquivalent zur Bayes-Regel (ML-Regel) bei gruppenweise identischen Σ (lineare Diskriminanzanalyse).

Kapitel 6

Clusteranalyse

Die Clusteranalyse ist ein Verfahren zur Gruppenbildung. Objekte derselben Gruppen sollen einander "möglichst ähnlich" sein. Objekte in verschiedenen Gruppen sollen "möglichst verschieden" sein.

6.1 Daten:

Wir betrachten eine Menge von Objekten $I = \{1, ..., N\}$, wobei an jedem Objekt p Merkmale gemessen werden. Die Daten werden anschließend zur Datenmatrix X zusammengefasst. Der Merkmalsvektor und die Datenmatrix sind gegeben als:

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}; \qquad X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Np} \end{pmatrix} \tag{6.1}$$

Es gibt Klassifikationsverfahren, bei denen die Merkmalswerte direkt zur Klassenbildung verwendet werden. Zunächst behandeln wir aber Verfahren, bei denen aus den Merkmalswerten zunächst "Ähnlichkeiten" oder "Distanzen" für die $\begin{pmatrix} \mathrm{N} \\ \\ 2 \end{pmatrix}$ Objektpaare berechnet werden und dann diese Daten zur Klassenbildung verwendet werden.

6.2 Ähnlichkeitsmaße:

Definition:

Sei $s : IxI \longmapsto R$ eine Abbildung zweier Objekte aus I auf eine Zahl aus R mit folgenden Eigenschaften:

- $s_{nm} = s_{mn}\ \forall m, n \in I$ (Symmetrie)
- $s_{nm} \leq s_{nn}\ \forall m, n \in I$

Manchmal wird zusätzlich verlangt, dass $s_{nm} \in [0, 1]$ gelten muss. Objekte sind umso ähnlicher, je größer s_{nm} ist.

6.3 Distanzmaß:

Definition:

Sei $d : IxI \longmapsto R$ eine Abbildung zweier Objekte aus I auf eine Zahl aus R mit folgenden Eigenschaften:

- $d_{nm} = d_{mn}\ \forall m, n \in I$ (Symmetrie)
- $d_{nn} = 0\ (d_{nm} \Leftrightarrow m = n)\ \forall m, n \in I$

6.4 Metrik:

Definition:

Eine Metrik ist ein Distanzmaß mit einer zusätzlichen Eigenschaft. Es muss nämlich die Dreiecksungleichung gelten:

$$\forall l, m, n \in I :\ d_{l,n} \leq d_{lm} + d_{mn} \tag{6.2}$$

In der Praxis werden oft Metriken gewählt, weil sie unserer räumlichen Vorstellung entsprechen. Ähnlich definiert man Ähnlichkeitsmaße und Distanzmaße für Teilmengen von I (= Gruppen, Klassen).

Definition:

Sei $\{\} \neq C_i \subseteq I$ wobei $(C_i \in P_*(I))$ (also die Potenzmenge von I ohne die leere Menge). $S(C_i, C_j)$ beziehungsweise $D(C_i, C_j)$ sind Ähnlichkeits- bzw. Distanzmaße für Klassen, wenn sie folgende Eigenschaften erfüllen:

- $S, D : P_*(I) \longmapsto R$
- $S(C_i, C_j) = S(C_j, C_i) \leq S(C_i, C_i)$
- $D(C_i, C_j) = D(C_j, C_i) \geq 0; \quad D(C_i, C_i) = 0$

Meist gilt auch folgende Monotoniebedingung:

- $C_i \subseteq C_j \Longrightarrow S(C_i, C_k) \geq S(C_j, C_k) \quad \forall i, j, k$
- $C_i \subseteq C_j \Longrightarrow D(C_i, C_k) \leq D(C_j, C_k) \quad \forall i, j, k$

Es werden aber auch Maße S, D verwendet, die diese Eigenschaft nicht besitzen (z.B: Zentroid-Methode).

6.5 Beispiele für Ähnlichkeits- und Distanzmaße:

1. Nominal-skalierte binäre Merkmale (Merkmalsausprägungen sind 0/1)

   Man kann eine Kontingenztabelle von C_n mit C_m für p binäre Merkmale erstellen. Es ist:

C_n/C_m	1	0	$\sum$
1	a	c	a+c
0	b	e	b+e
$\sum$	a+b	c+e	p

   Wie häufig stimmen die Ausprägungen von p Merkmalen überein? Genau $(a + e)$ Mal. Wie oft stimmen die Ausprägungen nicht überein? Insgesamt $(b + c)$ Mal.

   M-Koeffizient: ("matching-coefficient")

   Der M-Koeffizient ist definiert als: $S_{nm} = \frac{a+e}{p}$. Will man Übereinstimmung und Nicht-Übereinstimmung verschieden gewichten, dann wählt man ein Gewicht $u \in]0, 1[$ (Gewicht für die Übereinstimmung). Als Ähnlichkeitsmaß erhält man dann:

$$S_{nm} = \frac{u\ (a + e)}{u\ (a + e) + (1 - u)\ (b + c)} \tag{6.3}$$

Es gilt für jedes Gewicht u, dass $S_{nm} \in [0,1]$ ist. Jede Wahl von u erzeugt auch die selbe Ähnlichkeitsrangordnung.

Verwendet man also ein Klassifikationsverfahren, das zur Klassenbildung nur die Ähnlichkeitsrangordnung verwendet (zb: single-linkage, complete-linkage), dann ist die Wahl des Gewichtes u für obiges S_{nm} gleichgültig. Eine weitere wichtige Eigenschaft des obigen Koeffizienten S_{nm} ist, dass er invariant bezüglich bijektiven Abbildungen eines oder mehrerer Merkmale von x ist.

S-Koeffizient: ("similarity-coefficient")

Negative Übereinstimmungen (beide Objekte haben in einem Merkmal die Ausprägung 0 $\Leftrightarrow$ "Fehlen" der untersuchten Eigenschaft) werden nicht gezählt. Es ergibt sich dann wie oben (ungewichtet / gewichtet):

$$S_{nm} = \frac{a}{a+b+c} \qquad S_{nm} = \frac{ua}{ua + (1-u)(b+c)} \tag{6.4}$$

Wieder erzeugt jede Wahl von u dieselbe Ähnlichkeitsrangordnung, die aber nicht mit der des M-Koeffizienten übereinstimmen muss. Der S-Koeffizient ist nicht invariant bezüglich bijektiven Transformationen der Merkmale x.

Korrelationskoeffizienten:

Man muss sich hier die Objekte und Merkmale vertauscht denken und dann die "Pearson-Korrelation" zwischen binären Merkmalen rechnen. Es ergibt sich:

$$S_{nm} = \frac{ae - bc}{\sqrt{(a+c)(b+e)(a+b)(c+e)}} \tag{6.5}$$

Dieser Koeffizient ist ebenfalls invariant gegenüber bijektiven Abbildungen von x. Probleme ergeben sich allerdings, falls der Nenner Null wird. Man könnte aber wie bereits in der VL besprochen eine Vorbesetzung der Zellen mit kleinen Werten vornehmen.

2. Nominal skalierte mehrstufige Merkmale (mehr als 2 Ausprägungen)
   Der verallgemeinerte M-Koeffizient ist gegeben als $S_{nm} = \frac{u_{nm}}{p}$ wobei u_{nm} die Anzahl

der übereinstimmenden Komponenten von x_n sowie x_m ist.

Will man die Anzahl der Merkmalsausprägungen berücksichtigen (eine Übereinstimmung in einem Merkmal mit "vielen" Ausprägungen zählt mehr als in einem Merkmal mit wenigen Ausprägungen), dann wählt man

$$S_{nm} = \frac{1}{m^*} \sum_{i=1}^{p} m_i \sigma(x_{ni}, x_{mi}) \tag{6.6}$$

wobei m_i die Anzahl der Ausprägungen von x_i bezeichnet und $m^* = \sum_{i=1}^{p} m_i$ die Summe aller Ausprägungen ist. $\sigma(x_{ni}, x_{mi})$ zählt einfach die Übereinstimmung zwischen zwei Merkmalen und ist 1, wenn $x_{ni} = x_{mi}$ und 0, wenn $x_{ni} \neq x_{mi}$ ist. Wichtig ist wieder, dass die Koeffizienten invariant bezüglich bijektiver Transformationen von x sind.

3. Ordinale skalierte Merkmale

   Für die m_i Ausprägungen von x_i gilt die Ordnungsrelation $x_{1i} < x_{2i} < \ldots < x_{mi}$ für $i = 1, ..., p$. Objekte werden als umso ähnlicher betrachtet, je näher die Werte x_{ni} und x_{mi} hinsichtlich der obigen Rangordnung sind. Das geschieht, indem man für jedes ordinale Merkmal soviele binäre Hilfsvariablen ein, wie das Merkmal Ausprägungen hat einführt. Hat ein Merkmalswert x_{ni} in der Rangordnung die Position j, dann weist man den ersten j Hilfsvariablen den Wert 1 zu, den restlichen den Wert 0. Schlußendlich verwendet man Ähnlichkeitsmaße für binäre Variablen.

4. Quantitative Merkmale:

   Sowohl Meßskale als auch der Koordinatenursprung können frei gewählt werden. Für Ähnlichkeits- bzw. Distanzmaße ist es daher von Interesse, ob sie skaleninvariant (Maß hängt nicht von der Meßskala ab) und translationsinvariant (Maß hängt nicht von der Wahl des Nullpunkts ab) sind.

   Definition: "Skaleninvarianz"

   Sei $C = diag\ (c_1, ..., c_p)$ mit $c_i > 0$, dann berechnet man $\tilde{x}_n = Cx_n$ (n-tes Merkmal). Das Maß d ist genau dann skaleninvariant, falls $d(x_n, x_m) = d(\tilde{x}_n, \tilde{x}_m)$.

Definition: "Translationsinvarianz"

Sei $b \in R^p$, dann berechnet man $\tilde{x}_n = x_n + b$ (n-tes Merkmal). Das Maß d ist genau dann translationsinvariant, falls $d(x_n, x_m) = d(\tilde{x}_n, \tilde{x}_m)$.

6.6 LQ-Distanz

Definition:

Eine LQ-Distanz ist genau dann, gegeben, wenn gilt:

$$d_q(n,m) := (\sum_{i=1}^{p} |x_{ni} - x_{mi}|^q)^{\frac{1}{q}}; \qquad q \geq 1 \tag{6.7}$$

Vor Berechnung der LQ-Distanz müssen die Merkmale auf gemeinsame Maßeinheiten gebracht werden (Normierung). Meist wird folgende Normierung verwendet:

$$\tilde{x}_{ni} = \frac{x_{ni} - \overline{x}_i}{s_{i,q}} \qquad n = 1, ..., N \qquad i = 1, ..., p \qquad s_{i,q} = (\frac{1}{N} \sum_{n=1}^{N} |x_{ni} - \overline{x}_i|^q)^{\frac{1}{q}} \tag{6.8}$$

6.7 Häufig verwendete Distanzmaße:

- $d_1(n,m) = \sum_{i=1}^{p} |x_{ni} - x_{mi}| \ldots$ "City Block Metrik"
- $d_2(n,m) = \sqrt{\sum_{i=1}^{p} (x_{ni} - x_{mi})^2} = ||x_n - x_m||_2 \ldots$ "euklidische Norm"

Die euklidische Norm entspricht genau unseren geometrischen Distanzvorstellungen. Sie ist auch invariant bezüglich orthogonaler Transformationen von x, dh: ist C eine orthogonale Matrix, dann ist $d_2(x_n, x_m) = d(Cx_n, Cx_m)$. Auch ändert sich die euklidische Distanz bei Drehung und Spiegelung des Koordinatensystems (äquivalent dazu wäre die Drehung bzw. Spiegelung der Meßwerte selbst) also nicht.

6.8 Mahalanobis-Distanz:

Sei $d_M(n,m) = \sqrt{(x_n - x_m)^T K^{-1} (x_n - x_m)}$ wobei $K = \frac{1}{N} \sum_{n=1}^{N} (x_n - \overline{x})(x_m - \overline{x})^T$ die empirische Kovarianzmatrix ist. Die Mahalanobis-Distanz d_M ist invariant bezüglich beliebiger nicht-singulärer linearer Transformationen. Das heisst, ist C eine beliebige, reguläre

(p,p)-Matrix und $b \in R^p$ beliebig, und ist $\tilde{x}_n = Cx_n + b$, dann gilt:

$$d_M(x_n, x_m) = d_M(\tilde{x}_n, \tilde{x}_m) \tag{6.9}$$

Eine weitere wichtige Eigenschaft ist folgende. Sei $\tilde{x}_n = K^{-\frac{1}{2}} x_n \Rightarrow \tilde{x}_n$ sind unkorreliert und es gilt:

$$d_M(x_n, x_m) = \sqrt{(x_n - x_m)^T K^{\frac{-T}{2}} K^{\frac{-1}{2}} (x_n - x_m)} \tag{6.10}$$

$$= \sqrt{(\tilde{x}_n - \tilde{x}_m)^T (\tilde{x}_n - \tilde{x}_m)} = d_2(\tilde{x}_n, \tilde{x}_m) = ||\tilde{x}_n - \tilde{x}_m||_2 \tag{6.11}$$

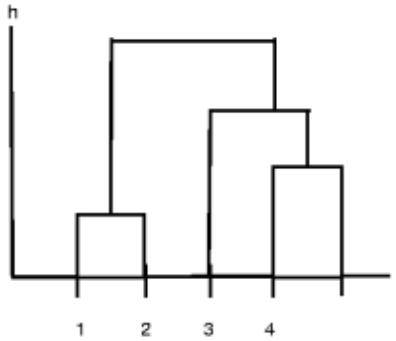
Die Berechnung der Mahalanobis-Distanz ist äquivalent zu folgender Darstellung:

- Merkmale in unkorrelierte Merkmale transformieren und
- dann die euklidische Distanz zwischen den unkorrelierten Merkmalen berechnen.

6.9 Hierarchische Klassifikationsverfahren:

Eine Folge von Partitionen (Aufteilungen) in Gruppen der Objektmenge $I = \{1, ..., N\}$ wird erzeugt. Dabei wird die Klassenzahl schrittweise verkleinert ("agglomerative Verfahren") oder erhöht ("diversive Verfahren"). Eine Erhöhung der Klassenzahl ist mit einer Erhöhung der Homogenität in den Klassen verbunden. Bei agglomerativen Verfahren werden die Klassen sukzessive vereinigt, bei diversiven Verfahren sukzessive aufgeteilt.

6.9.1 Darstellung im Dendogramm:



h mißt die Homogenität der Klassen, je kleiner der Wert von h bei einer Vereinigung von Klassen ist, umso ähnlicher sind die Objekte in den Klassen:

6.9.2 Vorteil hierarchischer Klassen:

Die (angestrebte) Klassenzahl muß nicht angegeben werden und auch die Vorgabe einer Homogenitätsschranke ist nicht notwendig. Außerdem ist die Berechnung sehr einfach. Andere Klassifikationsverfahren oder andere Distanzmaße erzeugen naheliegenderweise andere Dendogramme.

Definition: "Hierarchien":

Sei $I = \{\ 1, ..., N\} = \{\ I_1, ..., I_n\}$ und $P_*(I)$ sei die Potenzmenge von I ausgenommen der leeren Menge $\{\}$. Ein System von Mengen $H \subset P_*(I)$ heißt Hierarchie, falls für zwei verschiedene Mengen B und $C \in H$ genau eine der folgenden 3 Möglichkeiten gilt:

$$B \bigcap C = \{\} \qquad B \in C \qquad C \in B \tag{6.12}$$

Definition: "totale Hierarchie"

Eine Hierarchie H, die sowohl die Objektmenge I als auch alle Objekte $I_1, ..., I_n$ als Klassen enthält, bezeichnet man als "totale Hierarchie".

Definition: "Index einer Hierarchie H"

Eine für alle Klassen $C \in H$ definierte, nicht-negative Funktion h mit folgenden Eigenschaften

- $h(C) \leq h(B) \Leftarrow C \subseteq B$
- $h(C) = 0 \Rightarrow C$ (enthält nur äquivalente Objekte $x_n = x_m$)

heisst "Index" einer Hierarchie H. h mißt die Homegenität in den Klassen, je kleiner h ist, umso homogener sind die Klassen. Es gibt aber Klassifikationsverfahren (zb: Zentroid-Methode) bei denen das Homogenitätsmaß h kein Index ist. Es kann dann sein, dass $h(C) > h(B)$ obwohl C eine echte Teilmenge von B ist.

6.10 Agglomerative Verfahren:

Voraussetzung ist, dass für beliebige, nicht-leere Teilmengen von Objekten ein Distanzmaß D beziehungsweise ein Ähnlichkeitsmaß S vorgegeben sein muß. Der algorithmische Ablauf des Verfahrens ist wie folgt:

1. Sei $\nu = 0$ und $C^0 = \{\{I_1\}, \{I_2\}, ..., \{I_n\}\}$ sind die feinsten Partitionen von I.

2. Die Partition C^ν mit $\nu \geq 1$ erhält man aus $C^{\nu-1}$ indem man diejenigen Klassen von $C^{\nu-1}$ vereinigt, für die D minimal beziehungsweise S maximal wird (unter allen möglichen Paaren von Klassen aus $C^{\nu-1}$).

3. Wiederhole Schritt 2 solange bis $C^\nu = \{I\}$

Dadurch entsteht eine totale Hierarchie. Die Hierarchie kann indiziert werden, indem der in Schritt ν durch Vereinigung der Klassen C_j und C_k entstandenen Klasse folgender Index zugewiesen wird:

$$h_\nu = D_\nu = min_{k \neq j}\ D(C_j, C_k) \tag{6.13}$$

$$h_\nu = S_\nu = max_{k \neq j}\ S(C_j, C_k) \tag{6.14}$$

Ausserdem setzt man $h_0 = D_0 = 0$ beziehungsweise $S_0 = h_0 = max_{n \in I}\{S_{NN}\}$.

Ist im obigen Algorithmus das Paar von Klassen mit minimalem Abstand bzw. maximaler Ähnlichkeit nicht eindeutig, dann vereinigt man die Klassen auf eine der beiden folgenden Möglichkeiten:

- Man vereinigt so, dass möglichst viele Paare fusioniert werden können (es werden aber immer nur 2 Klassen miteinander fusioniert)

- Es werden alle Klassen C_k und C_j vereinigt, für die es eine Kette von Klassen mit minimaler Distanz bzw. maximaler Ähnlichkeit gibt. Es gilt also:

$$D(C_k, C_{k1}) = D(C_{k1}, C_{k2}) = \ldots = D(C_t, C_j) = min_{C_1, C_2 \in C^{\nu-1}}\ \{D(C_1, C_2)\}$$
$$S(C_k, C_{k1}) = S(C_{k1}, C_{k2}) = \ldots = S(C_t, C_j) = max_{C_1, C_2 \in C^{\nu-1}}\ \{D(C_1, C_2)\}$$

  Man wählt die Elemente am Ende der Kette aus, um die Kette nicht zu "zerreißen".

6.11 Rekursion zur Berechnung der Klassendistanzen für neue Klassen:

C_v und C_w werden zur neuen Klasse C fusioniert. Es gilt dann folgende Rekursionsbeziehung zur Berechnung der Distanz der neuen Klasse C zu einer anderen Klasse C_k:

$$D(C, C_k) = \alpha_v D(C_v, C_k) + \alpha_w D(C_w, C_k) + \beta D(C_v, C_w) + \gamma |D(C_v, C_k) - D(C_w, C_k)|$$

Gilt $\alpha_v + \alpha_w + \beta = 1$, dann kann das Distanzmaß D durch ein Ähnlichkeitsmaß S ersetzt werden.

6.12 Spezielle agglomerative Verfahren:

6.12.1 Single-Linkage-Verfahren:

Das Single-Linkage-Verfahren wird auch als "minimal-distance-method" oder "nearest-neighbour-method" bezeichnet. Die Rekursionsparameter zur Berechnung der Distanzen der neuen Klassen sind gegeben als:

$$\alpha_v = \alpha_w = \frac{1}{2}; \qquad \beta = 0; \qquad \gamma = -\frac{1}{2}$$

Die Distanz zwischen zwei Klassen C_j und C_k ist gleich der kleinsten Distanz zwischen einem Objekt aus C_j und einem Objekt aus C_k. Es gilt also:

$$D(C_j, C_k) = min_{n \in C_j, m \in C_k} \{d_{nm}\} \tag{6.15}$$

Im ν-ten Iterationsschritt werden die Klassen C_v und $C_w \in C^{\nu-1}$ fusioniert, für die gilt:

$$D(C_v, C_w) = D_\nu = h_\nu = min_{k \neq j} \{min_{n \in C_j, m \in C_k} \{d_{nm}\}\} \tag{6.16}$$

Mit dieser Methode werden auch Klassen miteinander vereinigt, wenn nur ein einziges Objekt der einen Klasse nahe bei einem einzigen Objekt der anderen Klasse liegt. Die anderen Objekte können aber durchaus sehr weit auseinander liegen. Ein weiterer Nachteil ist, dass Klassen, die duch sogenannte "Brücken" miteinander verbunden sind, nicht erkannt werden und auch eine sogenannte "Kettenbildung" auftreten kann.

6.12.2 Complete-Linkage-Verfahren:

Das Complete-Linkage-Verfahren wird auch als "maximum-distance-method" oder "furthest-neighbour-method" bezeichnet. Die Rekursionsparameter zur Berechnung der Distanzen der neuen Klassen sind gegeben als:

$$\alpha_v = \alpha_w = \frac{1}{2}; \qquad \beta = 0; \qquad \gamma = \frac{1}{2}$$

Die Distanz zwischen zwei Klassen C_j und C_k ist gleich der maximalen Distanz zwischen einem Objekt aus C_j und einem Objekt aus C_k. Es gilt also:

$$D(C_j, C_k) = max_{n \in C_j, m \in C_k} \{d_{nm}\} \tag{6.17}$$

Im ν-ten Iterationsschritt werden die Klassen C_v und $C_w \in C^{\nu-1}$ fusioniert, für die gilt:

$$D(C_v, C_w) = D_\nu = h_\nu = min_{k \neq j} \{max_{n \in C_j, m \in C_k} \{d_{nm}\}\} \tag{6.18}$$

Alle Objekte, die nach dem ν-ten Schritt in einer Klasse sind, haben zueinander höchstens den Abstand D_ν. Wie beim Single-Linkage-Verfahren benötigt man hier keinen Index, es reicht die Kenntnis der Distanzrangordnung.

6.12.3 Average-Linking-Method:

Die Rekursionsparameter zur Berechnung der Distanzen der neuen Klassen sind gegeben als:

$$\alpha_v = \frac{n_v}{n_v + n_w}; \qquad \alpha_w = \frac{n_w}{n_v + n_w}; \qquad \beta = \gamma = 0$$

Die Distanz zwischen zwei Klassen C_j und C_k ist gleich dem Durchschnitt aller Distanzen zwischen Objekten aus C_j und Objekten aus C_k. Es gilt also:

$$D(C_j, C_k) = \frac{1}{n_j n_k} \sum_{n \in C_j} \sum_{m \in C_k} d_{nm} \tag{6.19}$$

Im ν-ten Iterationsschritt werden die Klassen C_v und $C_w \in C^{\nu-1}$ fusioniert, für die gilt:

$$D(C_v, C_w) = D_\nu = h_\nu = min_{k \neq j} \{\frac{1}{n_j n_k} \sum_{n \in C_j} \sum_{m \in C_k} d_{nm}\} \tag{6.20}$$

Durch den Effekt der Durchschnittsbildung können in einer Klasse durchaus auch weit voneinander entfernte Objekte liegen, wenn dies duch viele, sehr nahe beieinanderliegende Objekte derselben Klasse kompensiert wird.

6.12.4 Verfahren von Ward:

Man benötigt für das Verfahren quantitative Merkmale. Die Rekursionsparameter sind gegeben als

$$\alpha_v = \frac{n_v + n_k}{n_v + n_w + n_k}; \qquad \alpha_w = \frac{n_w + n_k}{n_v + n_w + n_k}; \qquad \beta = -\frac{n_k}{n_v + n_w + n_k}; \qquad \gamma = 0$$

wobei gilt:

- $n_v \ldots$ Anzahl der Objekte in der Klasse C_v.
- $n_w \ldots$ Anzahl der Objekte in der Klasse C_w.
- $n_k \ldots$ Anzahl der Objekte in der Klasse C_k.

Die Homogenität einer Partition $C^{\nu-1}$ wird gemessen als Summe der Streuungen innerhalb der Klassen. Es ist

$$H(C^{\nu-1}) = \sum_{k=1}^{g} \sum_{n \in C_k} ||x_n - \overline{x}_k||^2 \tag{6.21}$$

wobei insgesamt g Klassen in der Partition $C^{\nu-1}$ vorhanden sind und $\overline{x}_k$ das Mittel der x-Werte der k-ten Klasse C_k bezeichnet.

Entsteht C^ν aus $C^{\nu-1}$ durch Vereinigung von C_v und C_w zu C, dann ist die Homogenität gegeben durch:

$$H(C^\nu) = \sum_{k \neq v,w} \sum_{n \in C_k} ||x_n - \overline{x}_k||^2 + \sum_{n \in C} ||x_n - \overline{x}_c||^2 \tag{6.22}$$

Nun soll der Homogenitätsverlust (höhere Werte für H bedeutet, dass die Homogenität geringer wird) im ν-ten Schritt minimal sein. Es gilt dann:

$$h_\nu = D(C_v, C_w) = min_{k \neq j}\{\frac{n_k n_j}{n_k + n_j}||\overline{x}_k - \overline{x}_j||^2\} = min_{k \neq j}\{D(C_k, C_j)\} \tag{6.23}$$

6.12.5 Zentroid-Methode:

Man benötigt für das Verfahren quantitative Merkmale. Die Rekursionsparameter sind gegeben als

$$\alpha_v = \frac{n_v}{n_v + n_w}; \qquad \alpha_w = \frac{n_w}{n_v + n_w}; \qquad \beta = -\frac{n_v + n_w}{(n_v + n_w)^2}; \qquad \gamma = 0$$

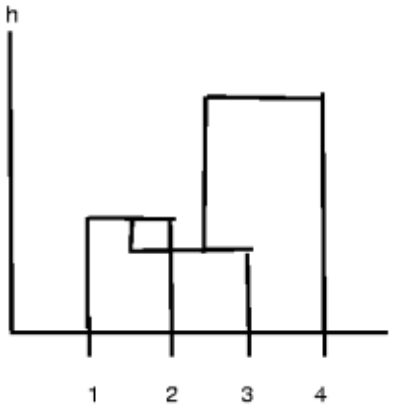
Jede Klasse wird repräsentiert duch den Klassenschwerpunkt, der gegeben ist als:

$$\overline{x}_k = \frac{1}{n_k} \sum_{n \in C_k} x_n \tag{6.24}$$

Die Distanz zwischen 2 Klassen ist gleich der quadratischen Euklid'schen Distanz der Klassenschwerpunkte. Im ν-ten Schritt werden die Klassen C_v, $C_w \in C^{\nu-1}$ fusioniert, für die gilt:

$$h_\nu = D_\nu = D(C_v, C_w) = min_{k \neq j}\{||\overline{x}_j - \overline{x}_k||^2\} \tag{6.25}$$

Hier ist die Indexbedingung $C \subseteq B \Rightarrow h(C) \leq h(B)$ für $C, B \in H$ (Hierarchie) nicht erfüllt. Es kann vorkommen, dass die im ν-ten Schritt gebildete Klasse $C_v \cup C_w$ homogener (gemessen im Homogenitätsmaß h_ν) ist, als die Klassen C_v, C_w selbst. In einem Dendogramm würde sich damit folgendes Bild ergeben:



Auch hier ist ähnlich wie beim Average-Linkage-Verfahren ein Effekt der Durchschnittsbildung zu beobachten. Wählt man auch beim Average-Linkage-Verfahren quadrierte Euklid'sche Distanzen, so ergibt sich

$$d_{AL;k,j} = \frac{1}{n_k n_j} \sum_{n \in C; j, m \in C_k} ||x_n - x_m||^2 = ||\overline{x}_k - \overline{x}_j||^2 + s_k^2 + s_j^2 \tag{6.26}$$

wobei $\overline{x}_k, \overline{x}_j$ sowie k, j und s_k^2, s_j^2 die Mittelwerte bzw. Varianzen der Klassen k, j sind.

Beim Average-Linking-Verfahren können also die Abstände der Schwerpunkte zweier zu fusionierender Klassen auch größer sein als bei der Zentroid-Methode, wenn die die Klassen sehr homogen sind.