In this chapter the concept of generalized linear models is extended to the case of a vector-valued response variable. Consider Example 2.1, where we were interested in the effect of risk factors and antibiotics on infection following birth by caesarian section. In this example the response was binary, distinguishing only between occurrence and nonoccurrence of infection, and thereby ignoring that the data originally provided information on the type of infection (type I or II) as well. It is possible, however, to use this information by introducing a response variable with three categories (no infection, infection type I, infection type II). Naturally, these categories cannot be treated as a unidimensional response. We have to introduce a (dummy) variable for each category, thus obtaining a multivariate response variable. Therefore, link and response functions for the influence term will be vector-valued functions in this chapter. The focus is on multicategorical response variables and multinomial models. Variables of this type are often called polychotomous, the possible values are called categories. Extension to other multivariate exponential family densities is possible but not considered in this text.

After an introductory section we first consider the case of a nominal response variable (Section 3.2). If the response categories are ordered, the use of this ordering yields more parsimoniously parameterized models. In Section 3.3 several models of this type are discussed. Section 3.4 outlines statistical inference for the multicategorical case. In many applications, more than one response variable is observed, for example, when several measurements are made for each individual or unit or when measurements are
observed repeatedly. Approaches for this type of multivariate responses with correlated components are outlined in Section 3.5.

### 3.1 Multicategorical Response Models

#### 3.1.1 Multinomial Distribution

For the categorical responses considered in this chapter, the basic distribution is the multinomial distribution. Let the response variable $Y$ have $k$ possible values, which for simplicity are labeled $1, \ldots, k$. Sometimes consideration of $Y \in \{1, \ldots, k\}$ hides the fact that we actually have a multivariate response variable. This becomes obvious by considering the response vector of the dummy variables $y' = (\tilde{y}_1, \ldots, \tilde{y}_q)$, $q = k - 1$, with components

$$\begin{align*}
\tilde{y}_r &= \begin{cases} 
1 & \text{if } Y = r, \quad r = 1, \ldots, q, \\
0 & \text{otherwise.}
\end{cases} 
\end{align*}
$$

Then we have

$$Y = r \iff y = (0, \ldots, 1, \ldots, 0).$$

The probabilities are simply connected by

$$P(Y = r) = P(y_r = 1).$$

Given $m$ independent repetitions $y_1, \ldots, y_m$ (or equivalently $Y_1, \ldots, Y_m$), it is useful to consider as a response variable the number of trials where we get outcome $r$. For the repetitions $(y_1, \ldots, y_m)$, we get the sum of vectors

$$y = \sum_{i=1}^{m} \tilde{y}_i.$$

Then the vector $y$ is multinomially distributed with distribution function

$$P(y = (m_1, \ldots, m_q)) = \frac{m!}{m_1! \cdots m_q!(m - m_1 - \ldots - m_q)!} \cdot \pi_1^{m_1} \cdots \pi_q^{m_q} (1 - \pi_1 - \ldots - \pi_q)^{m - m_1 - \ldots - m_q},$$

(3.1.2)

where $\pi_r = P(Y_i = r)$, $i = 1, \ldots, m$. The multinomial distribution of $y$ is abbreviated by

$$y \sim M(m, \pi), \quad \text{where } \pi' = (\pi_1, \ldots, \pi_q).$$