12
Analysis of Multiple Discrete Responses

In Chapter 11 we introduced some contemporary approaches to analyzing longitudinal data for which the responses are continuous measurements. In fact, most people imply continuous responses when they refer to longitudinal data. The analysis of discrete longitudinal data is a relatively new, though active, subject. Readers who are interested in methodological developments may find many unanswered questions in this chapter. The purpose of this chapter is to shed some light on this growing subject. In the statistical literature, the topic may be tagged with clustered or correlated discrete/binary outcomes. So far, most progress has been made toward the binary outcomes; hence, therein lies the focus of this chapter.

Sometimes, correlated discrete responses are generated from a single endpoint by repeatedly measuring it on individuals in a temporal or spatial domain. They are called longitudinal discrete responses. Examples 12.1 and 12.2 represent this class of data. Other times, as in Example 12.3 and in Section 12.3, the correlated responses consist of distinct endpoints. In recent years, we have witnessed more and more studies that involve both types of responses, such as Example 12.4.

Example 12.1 To investigate racial differences in the cause-specific prevalence of blindness, Sommer et al. (1991) used a randomly selected, stratified, multistage cluster sample of 2395 Blacks and 2913 Whites 40 years of age and older in East Baltimore. Those 5208 subjects underwent detailed ophthalmic examinations by a single team. In this study, the authors observed bivariate binary responses in a spatial domain for each subject, namely, the blindness of left and right eyes. The authors found that the leading causes

H. Zhang and B.H. Singer, *Recursive Partitioning and Applications*,
Springer Series in Statistics, DOI 10.1007/978-1-4419-6824-1_12,

of blindness were unoperated senile cataract, primary open-angle glaucoma, and age-related macular degeneration. They also concluded that the pattern of blindness in urban Baltimore appears to be different among Blacks and Whites. Whites are far more likely to have age-related macular degeneration, and Blacks to have primary open-angle glaucoma. Subsequently, Liang, Zeger, and Qaqish (1992) reanalyzed these data, comparing different statistical approaches.

Example 12.2 From 1974 to 1977, a team of investigators conducted a longitudinal study of the respiratory health effects of air pollutants among children and adults living in six cities in the United States. The study design was reported by Ferris et al. (1979) and Sommer et al. (1984). The selection of the cities was to cover a range of air quality based on their historic levels of outdoor pollution. In all but one small city, the initial examinations included all first- and second-grade school children. In the small city, children up to the fifth grade were included. The study subjects were reexamined annually for three years. At each visit, the investigators collected information regarding the number of persons living in the house, familial smoking habits, parental occupation and education background, the fuel used for cooking in the house, pulmonary function, respiratory illness history, and symptom history. In Ware et al. (1984), they selected 10,106 children 6 to 9 years of age at the enrollment and analyzed wheeze status (yes, no) of the children as a longitudinal binary outcome. Additional analyses have been conducted by Zeger, Liang, and Albert (1988) and Fitzmaurice and Laird (1993) among others.

Example 12.3 This is an example where the risk of two distinct, but presumably correlated, outcomes were studied, i.e., respiratory disease and diarrhea in children with preexisting mild vitamin A deficiency.

Sommer and colleagues (Sommer et al. 1983 and 1984) conducted a prospective longitudinal study of 4600 children aged up to 6 years at entry in rural villages of Indonesia between March 1977 and December 1978. Their research team examined these children every 3 months for 18 months. An average of 3135 children were free of respiratory disease and diarrhea at the examination. At each examination, they recorded interval medical history, weight, height, general health status, and eye condition. They found that the risk of respiratory disease and diarrhea were more closely associated with vitamin A status than with general nutritional status.

Example 12.4 Genes underlie numerous conditions and diseases. A vast number of genetic epidemiologic studies have been conducted to infer genetic bases of various syndromes. Multiple clustered responses naturally arise from such studies. For example, Scourfield et al. (1996) examined the gender difference in disorders of substance abuse, comorbidity anxiety, and sensation seeking, using the database from the Genetic Epidemiology Research Unit, Yale University School of Medicine, New Haven, Connecticut,

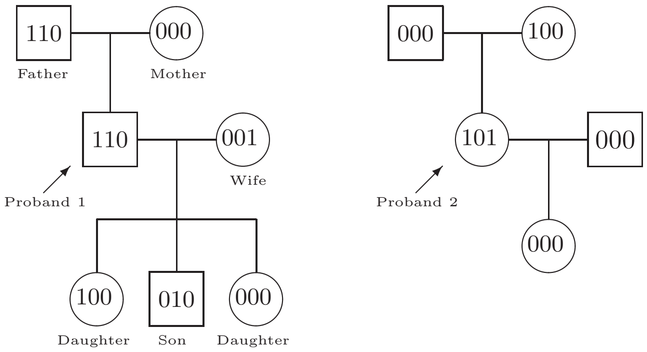

FIGURE 12.1. Two pedigrees of different family sizes. Each square or circle represents a family member. The left pedigree pinpoints the relationship of relatives to the proband. A sequence of three bits (0 or 1) is displayed within all squares and circles, marking the status of substance abuse, anxiety, and sensation seeking, respectively.

under the leadership of Professor Kathleen Merikangas. Two hundred sixty-two probands, through whom the other family members are ascertained, are included in the database. Information regarding a variety of psychiatric disorders and predictive covariates, e.g., gender, has been recorded for all probands and some of their relatives (parents, siblings, offspring, etc.). The pedigrees in Figure 12.1 illustrate typical family structures. We should note that the first proband has six relatives in the record, whereas the second one has four. In other words, the family size varies from pedigree to pedigree. It is also important to realize that multiple disorders, i.e., three, are evaluated for every member of a family.

12.1 Parametric Methods for Binary Responses

Suppose that $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iq_i})'$ is a vector of binary responses for subject i, $i = 1, \ldots, n$. In Example 12.1, $q_i = 2$ for all 2913 subjects, and (Y_1, Y_2) indicates the blindness of the left and right eyes. Likewise, we can easily define the response vector for Examples 12.2 and 12.3.

Parametric models have dominated the applications involving multiple binary responses. Log-linear and marginal models are in the spotlight in the literature. We give a brief discussion of these two models and strongly recommend reading related articles and books cited in this chapter.

12.1.1 Log-Linear Models

One of the most popular and conceptually simple models for multiple binary responses is the log-linear model that assumes the joint probability of $\mathbf{Y}_i$ to be of the form

$$\begin{aligned} I\!P\{\mathbf{Y}_i = \mathbf{y}_i\} &= \exp\left[\sum_{j=1}^{q_i} \theta_{ij} y_{ij} + \sum_{j_1<j_2} \theta_{ij_1j_2} y_{ij_1} y_{ij_2} + \cdots \right. \\ &\qquad \left. + \theta_{i1\cdots q_i} y_{i1} \cdots y_{iq_i} + A_i(\boldsymbol{\theta}_i)\right], \end{aligned} \tag{12.1}$$

where

$$\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iq_i}, \theta_{i12}, \ldots, \theta_{i,q_i-1,q_i}, \ldots, \theta_{1\cdots q_i})$$

is the $(2^{q_i-1} - 1)$-vector of canonical parameters and $\exp[A_i(\boldsymbol{\theta}_i)]$ is the normalizing constant.

Model (12.1) appears to involve too many parameters. In practice, however, it is usually greatly simplified. Two steps are critical to this simplification. First, most data are regular in the sense that the components of $\boldsymbol{\theta}_i$ correspond to fixed coordinates. In other words, $\boldsymbol{\theta}_i$ does not depend on i, and this subscript can be removed. In Examples 12.1–12.3, the vector of canonical parameters, $\boldsymbol{\theta}_i$, does not depend on i. For instance, Example 12.3 involves only $2^2 - 1 = 3$ parameters. Second, the canonical parameters with respect to the terms with the third- or higher-orders are generally hypothetically set to zero. The resulting models are referred to as the quadratic exponential model (see, e.g., Zhao and Prentice 1990; Fitzmaurice and Laird 1995). Estimating those "removed" parameters could otherwise raise a tremendous challenge to data analysis.

In family studies as illustrated by Example 12.4, the vector of canonical parameters, $\boldsymbol{\theta}_i$, may not have a fixed coordinate system. Although the number of interested disorders is three for every subject, the size of pedigree differs when the entire pedigree is regarded as a unit, or cluster. In such applications, it is vital to form a parametric system that reflects the nature of $\mathbf{Y}_i$. This practice depends, however, on individual applications.

Next, let us take a look at the quadratic exponential model in which the canonical parameters have a fixed coordinate system:

$$I\!P\{\mathbf{Y} = \mathbf{y}\} = \exp\left[\sum_{j=1}^{q} \theta_j y_j + \sum_{j<k} \theta_{jk} y_j y_k + A(\boldsymbol{\theta})\right], \tag{12.2}$$

where

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q, \theta_{12} \cdots \theta_{q-1,q}).$$

Based on model (12.2), the canonical parameters have certain interpretations. Precisely, we have

$$\log\left[\frac{I\!P\{Y_j = 1 | Y_k = y_k, Y_l = 0, l \neq j, k\}}{I\!P\{Y_j = 0 | Y_k = y_k, Y_l = 0, l \neq j, k\}}\right] = \theta_j + \theta_{jk} y_k.$$

Thus, θ_j is the log odds for $Y_j = 1$ given that the remaining components of Y equal zero. In addition, θ_{jk} is referred to as an association parameter because it is the conditional log odds ratio describing the association between Y_j and Y_k provided that the other components of Y are zero. It is important to realize that the canonical parameters are the log odds or odds ratio under certain conditions, but we should be aware of the fact that these conditions may not always make sense.

Why is model (12.1) called a log-linear model? Let us consider a bivariate case. It follows from model (12.2) that the joint probability for the n bivariate vectors is

$$\exp[\theta_1(n_{21}+n_{22}) + \theta_2(n_{12}+n_{22}) + \theta_{12}n_{22} + nA(\boldsymbol{\theta})], \tag{12.3}$$

where $n_{11} = \sum_{i=1}^{n}(1-y_{i1})(1-y_{i2})$, $n_{12} = \sum_{i=1}^{n}(1-y_{i1})y_{i2}$, $n_{21} = \sum_{i=1}^{n} y_{i1}(1-y_{i2})$, and $n_{22} = \sum_{i=1}^{n} y_{i1}y_{i2}$ are the cell counts in the following 2 × 2 table:

		Y_2	
		0	1
Y_1	0	n_{11}	n_{12}
	1	n_{21}	n_{22}

It is easy to see that the expression in (12.3) equals

$$\frac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!} m_{11}^{n_{11}} m_{12}^{n_{12}} m_{21}^{n_{21}} m_{22}^{n_{22}},$$

where

$$\log(m_{jk}) = \mu + \lambda_j^{Y_1} + \lambda_k^{Y_2} + \lambda_{jk}^{Y_1Y_2}, \tag{12.4}$$

with

$$\begin{aligned} \mu &= (\theta_1+\theta_2)/2 + \theta_{12}/4 + A(\boldsymbol{\theta}), && (12.5)\\ \lambda_1^{Y_1} &= -\theta_1/2 - \theta_{12}/4 + A(\boldsymbol{\theta}), && (12.6)\\ \lambda_1^{Y_2} &= -\theta_2/2 - \theta_{12}/4 + A(\boldsymbol{\theta}), && (12.7)\\ \lambda_{11}^{Y_1Y_2} &= \theta_{12}/4, && (12.8) \end{aligned}$$

and $\lambda_2^{Y_1} = -\lambda_1^{Y_1}$, $\lambda_2^{Y_2} = -\lambda_1^{Y_2}$, and $\lambda_{12}^{Y_1Y_2} = \lambda_{21}^{Y_1Y_2} = -\lambda_{22}^{Y_1Y_2} = -\lambda_{11}^{Y_1Y_2}$. In other words, $(n_{11}, n_{12}, n_{21}, n_{22})$ follows a multinomial distribution with means specified by the log-linear effects in (12.4). This is usually how the log-linear models are introduced (e.g., Agresti 1990, Chapter 5). Further, Equations (12.5)–(12.8) provide another way to interpret the canonical parameters.

12.1.2 Marginal Models

As we mentioned earlier, the interpretation of canonical parameters in the log-linear model depends on certain conditions that are not always of clinical relevance. On the other hand, after the reformation of the log-linear model in (12.4), the canonical parameters have one-to-one relationships with the "marginal" parameters as delineated in (12.5)–(12.8). Here, the marginal parameters refer to the main and interactive effects in model (12.4). For many investigators, the question of utmost importance is related to the marginal parameters that are defined directly from the marginal distribution of the responses, unlike the canonical parameters, which involve all responses at once.

One possibility is to reparametrize the log-linear model in terms of marginal means, correlations, etc. In fact, the Bahadur representation is another typical method to represent the log-linear model, and it directly extends the multinomial distribution by including additional multiplicative factors to take into account the association among the components of $\mathbf{Y}$ (Bahadur 1961; Fitzmaurice et al., 1993; Diggle et al. 1991). In mathematical form, we have

$$\begin{aligned} I\!P\{\mathbf{Y}=\mathbf{y}\} &= \prod_{j=1}^{q} \mu_j^{y_j}(1-\mu_j)^{(1-y_j)} \\ &\times(1+\sum_{j_1<j_2}\rho_{j_1j_2}r_{j_1}r_{j_2}+\sum_{j_1<j_2<j_3}\rho_{j_1j_2j_3}r_{j_1}r_{j_2}r_{j_3}+\cdots+\rho_{1\cdots q}r_1\cdots r_q), \end{aligned}$$

where

$$\begin{aligned} \mu_j &= I\!E\{Y_j\}, \\ r_j &= (y_j-\mu_j)/\sqrt{\mu_j(1-\mu_j)}, \\ \rho_{j_1\cdots j_l} &= I\!E\{R_{j_1}\cdots R_{j_l}\}, \end{aligned}$$

$j=1,\ldots,q$.

The Bahadur representation is one step forward in terms of formulating the log-linear model as a function of the parameters such as means and correlations that we used to see in the analysis of continuous responses. This representation is, however, severely handicapped by the fact that the "hierarchal" correlations entangle the ones at lower orders and the means and that it is particularly problematic in the presence of covariates. To address the dilemma between the parameter interpretability and feasibility, Liang et al. (1992) proposed the use of marginal models parametrized by the means, the odds ratios, and the contrasts of odds ratios. Specifically, let

$$\begin{aligned} \gamma_{j_1j_2} &= OR(Y_{j_1},Y_{j_2})=\frac{I\!P\{Y_{j_1}=1,Y_{j_2}=1\}I\!P\{Y_{j_1}=0,Y_{j_2}=0\}}{I\!P\{Y_{j_1}=1,Y_{j_2}=0\}I\!P\{Y_{j_1}=0,Y_{j_2}=1\}}, \\ \zeta_{j_1j_2j_3} &= \log[OR(Y_{j_1},Y_{j_2}|Y_{j_3}=1)]-\log[OR(Y_{j_1},Y_{j_2}|Y_{j_3}=0)], \end{aligned}$$

and generally,

$$\zeta_{j_1\cdots j_l}=\sum_{y_{j_3},\ldots,y_{j_l}=0,1}(-1)^{b(\mathbf{y})}\log[OR(Y_{j_1},Y_{j_2}|y_{j_3},\ldots,y_{j_l})],$$

where $b(\mathbf{y}) = \sum_{k=3}^{l} y_{j_k} + l - 2$.

It is quite unfortunate that evaluating the full likelihood based on the new set of parameters, μ_j, $\gamma_{j_1 j_2}$, and $\zeta_{j_1 \cdots j_l}$, is generally complicated. To gain insight into where the complications arise, let us go through the details for the bivariate case. We need to specify the probability $I\!P\{Y_1 = y_1, Y_2 = y_2\} \overset{\text{def}}{=} p(y_1, y_2)$ for four possible combinations of (y_1, y_2). The following four equations can lead to the unique identification of the four probabilities:

$$\begin{aligned}
p(1,1) + p(1,0) &= \mu_1, \\
p(0,1) + p(1,1) &= \mu_2, \\
p(1,1) + p(1,0) + p(0,1) + p(0,0) &= 1, \\
p(1,1)p(0,0) &= \gamma_{12} p(0,1) p(1,0).
\end{aligned}$$

From the first three equations, we have $p(1,0) = \mu_1 - p(1,1)$, $p(0,1) = \mu_2 - p(1,1)$, and $p(0,0) = 1 - \mu_1 - \mu_2 + p(1,1)$. If we plug them into the last equation, we have a quadratic equation in $p(1,1)$,

$$(1 - \gamma_{12}) p^2(1,1) + [1 + (\gamma_{12} - 1)(\mu_1 + \mu_2)] p(1,1) - \gamma_{12} \mu_1 \mu_2 = 0,$$

and the solution for $p(1,1) \overset{\text{def}}{=} \mu_{11}$ is (Dale, 1986)

$$\begin{cases} \frac{1+(\gamma_{12}-1)(\mu_1+\mu_2)-\{[1+(\gamma_{12}-1)(\mu_1+\mu_2)]^2+4(1-\gamma_{12})\gamma_{12}\mu_1\mu_2\}^{-\frac{1}{2}}}{2(1-\gamma_{12})} & \text{if } \gamma_{12} \neq 1, \\ \mu_1 \mu_2 & \text{if } \gamma_{12} = 1. \end{cases}$$

Using this solution, it is easy to conclude that

$$p(y_1, y_2) = \mu_1^{y_1} (1 - \mu_1)^{1 - y_1} \mu_2 (1 - \mu_2)^{1 - y_2} + (-1)^{y_1 - y_2} (\mu_{11} - \mu_1 \mu_2).$$

When we have more than two responses, the problem could be intractable if we do not reduce the dimension of the parameters appropriately such as setting $\gamma_{j_1 j_2} = \gamma$.

*12.1.3 Parameter Estimation**

In the log-linear and marginal models we have not introduced covariates. As a matter of fact, the issue of most interest to us is modeling the distribution of $\mathbf{Y}$ in the presence of covariates as in the previous chapters. In principle, it is straightforward to incorporate a set of the covariates, $\mathbf{x}$, into the models. The canonical parameters $\boldsymbol{\theta}$ in the log-linear model (12.2) and the marginal parameters in the marginal models can be defined as a function of $\mathbf{x}$, which is called the link function in the context of generalized linear models (McCullagh and Nelder 1989, p. 27).

Depending on the specification of the link function, finding the maximum likelihood estimates of the parameters is not impossible; see, e.g., Section 12.2.3. Nevertheless, a more common practice is to make use of so-called generalized estimating equations (GEE), which simplify the estimation process while retaining some of the most important asymptotic properties of the estimates as elaborated below (Liang and Zeger 1986).

Now, let us turn back to model (12.2) and explain how to use the idea of generalized estimating equations. First, we reexpress the probability in vector form:

$$I\!\!P\{\mathbf{Y}=\mathbf{y}\}=\exp[\boldsymbol{\theta}'\mathbf{z}-A(\boldsymbol{\theta})], \tag{12.9}$$

where $\mathbf{z}=(\mathbf{y}',\mathbf{w}')'$ and $\mathbf{w}$ is a $\binom{q}{2}$-vector consisting of $(y_1y_2,\ldots,y_{q-1}y_q)'$.

For model (12.9), we assume that there exists a vectorial link function $\boldsymbol{\eta}$ that transforms $\mathbf{x}$ coupled with a condensed vector of parameters $\boldsymbol{\beta}$ to $\boldsymbol{\theta}$, e.g., $\boldsymbol{\theta}=\boldsymbol{\eta}(\mathbf{x}'\boldsymbol{\beta})$. Then, the GEE approach attempts to solve the unbiased estimating equations (Godambe 1960; Zhao and Prentice 1990)

$$U(\boldsymbol{\beta})=\sum_{i=1}^{n}JV_i^{-1}\begin{pmatrix}\mathbf{y}_i-\boldsymbol{\mu}\\ \mathbf{w}_i-\boldsymbol{\omega}\end{pmatrix}=0, \tag{12.10}$$

where $\boldsymbol{\omega}=I\!\!E\{\mathbf{w}\}$, $V_i=\mathrm{Cov}(\mathbf{z}_i)$, and $J=\partial\boldsymbol{\theta}/\partial\boldsymbol{\beta}'$.

Liang et al. (1992) called (12.10) GEE2, because it is a second-order extension of the estimating equations proposed by Liang and Zeger (1986). However, if we set the block off-diagonal matrices in J and V_i to zero in (12.10), then (12.10) becomes GEE1, which can be less efficient than GEE2 when the link function is misspecified. We should also note that the block off-diagonal elements of the covariance matrix V_i cannot be determined by $\boldsymbol{\mu}$ and $\boldsymbol{\omega}$. To avoid estimating additional parameters, so-called working matrices are usually used to replace the underlying matrices (Zhao and Prentice 1990).

The solution $\hat{\boldsymbol{\beta}}$ to (12.10) has, asymptotically as $n\to\infty$, a multivariate normal distribution with mean 0 and covariance matrix that can be consistently estimated by

$$\left(\sum_{i=1}^{n}JV_iJ'\right)^{-1}\left(\sum_{i=1}^{n}JV_i\begin{pmatrix}\mathbf{y}_i-\boldsymbol{\mu}\\ \mathbf{w}_i-\boldsymbol{\omega}\end{pmatrix}\begin{pmatrix}\mathbf{y}_i-\boldsymbol{\mu}\\ \mathbf{w}_i-\boldsymbol{\omega}\end{pmatrix}'V_iJ'\right)\left(\sum_{i=1}^{n}JV_iJ'\right)^{-1}$$

evaluated at $\hat{\boldsymbol{\beta}}$ (Liang et al., 1992). It also turns that $U(\boldsymbol{\beta})$ resembles the quasi-score function derived from the quasi-likelihood as introduced in (9.5) of McCullagh and Nelder (1989).

Likewise, if we are interested in the pairwise odds ratio and use the marginal models, then we assume a link function between parameters μ_j and γ_{jk}, and covariates $\mathbf{x}$. The rest of the derivation for GEE is identical to that above.

12.1.4 Frailty Models

In Example 12.4, we have encountered different numbers of binary responses among different measurement units, namely, families. Let the data for family i consist of binary responses Y_{ij} and covariates $\mathbf{x}_{ij}$, $j = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, I$. Here, I is the number of families, and n_i is the number of relatives in the ith family, $i = 1, 2, \ldots, I$.

In such family studies, the association of the health condition between relatives is of interest. One approach is to generalize the log-linear model introduced in Section 12.1.1 and to include higher-order interaction terms. Particularly, based on Connolly and Liang (1988), we may assume

$$\operatorname{logit} I\!\!P\{Y_{ij} = 1 | Y_{il}, l \neq j, \mathbf{x}_i\} = F_{n_i}(W_{ij}; \theta) + \mathbf{x}_{ij}\boldsymbol{\beta}, \tag{12.11}$$

where $W_{ij} = \sum_{l \neq j}^{n_i} Y_{il}$, F_{n_i} is an arbitrary function, and θ is a parameter. This leads to the joint probability for the outcome in the ith family

$$\log I\!\!P\{\mathbf{Y}_i = \mathbf{y}_i | \mathbf{x}_i\} = \alpha + \sum_{j=1}^{n_i} y_{ij}\mathbf{x}_{ij}\boldsymbol{\beta} + \sum_{l=0}^{W_i + y_{ij} - 1} F_{n_i}(l; \theta). \tag{12.12}$$

Related to model (12.12), Bonney (1986, 1987) introduced several classes of regressive logistic models, assuming simple Markovian structures of dependence among the traits of family members. In essence, these regressive logistic models are ordinary logistic regression models except that the "covariates" are derived from a set of common sense covariates and the outcomes of other family members. The regressive logistic models are practically appealing and have been widely used in segregation analysis.

Babiker and Cuzick (1994) noted two major problems with model (12.12) and its like. First, the parametrization depends on the family size n_i, and the coefficients obtained from different families are irreconcilable. Second, they pointed out that the conditional coefficients often are not easily converted to parameters of interest even when the family sizes are the same. For these concerns, they proposed the use of a simple frailty model. In most family studies, however, their simple one-frailty model cannot address questions of importance. To this end, it is useful to enhance the simple frailty model by considering the relationship among relatives.

Let us take the three-generation pedigree in Figure 12.1 as an example. We can introduce three types of unobserved frailties U_1^i, U_2^i, and U_3^i for the ith family that represent common, unmeasured environmental factors; genetic susceptibility of the family founders; and the transmission of relevant genetic materials from a parent to a child. Here, a family founder is an individual whose parents were not sampled in the pedigree. To avoid technical complications, suppose that these frailties are independent Bernoulli random variables; that is,

$$I\!\!P\{U_k^i = 1\} = \theta_k = 1 - I\!\!P\{U_k^i = 0\},$$

for $k = 1, 2, 3$. A critical assumption is that for the ith family and conditional on all possible U_k^i's, denoted by U^i, the health conditions of all family members are independent and

$$\operatorname{logit}(I\!P\{Y_j^i = 1|U^i\}) = \mathbf{x}_j^i\boldsymbol{\beta} + \mathbf{a}_j^i\boldsymbol{\gamma}, \tag{12.13}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of parameters, and

$$\mathbf{a}_j^i = (U_1^i, U_{2,2j-1}^i + U_{2,2j}^i, U_{2,2j-1}^i U_{2,2j}^i)'$$

harbors the frailties. The construction of $\mathbf{a}_j^i$ is based on assuming the existence of a major susceptibility locus with alleles A and a, as clarified below.

The frequency of allele A is θ_2, and $(U_{2,2j-1}^i, U_{2,2j}^i)$ indicate the presence of allele A in the two chromosomes of the jth member of the ith family. Based on the Mendelian transmission, $\theta_3 = 0.5$. The parameter interpretation in model (12.13) is most important. The β parameters measure the strength of association between the trait and the covariates conditional on the frailties, while the γ parameters indicate the familial and genetic contributions to the trait. Note that $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)'$. If $\gamma_2 = 0$ and $\gamma_3 \neq 0$, it suggests a recessive trait because a genetic effect is expressed only in the presence of two A alleles. On the other hand, if a completely dominant gene underlies the trait, genotypes Aa and AA give rise to the same effect, implying that $\gamma_2 = 2\gamma_2 + \gamma_3$, i.e., $\gamma_2 = -\gamma_3$.

The frailty model (12.13) is closely related to many existing models for segregation analysis, all of which can be traced back to the classic Elston–Stewart (1971) model for the genetic analysis of pedigree data. The Elston–Stewart model was originally designed to identify the mode of inheritance of a particular trait of interest without considering the presence of covariates. The frailty model (12.13) is quite similar to the class D logistic regressive models of Bonney (1986, 1987). The major difference is the method for modeling familial correlations as a result of residual genetic effects and environment. The regressive models make use of the parental traits and assume the conditional independence among siblings on the parental traits. In contrast, the frailty model assumes the conditional independence among all family members on the frailty variable. Conceptually, frailty variables defined here are very similar to that of ousiotype introduced by Cannings et al. (1978) in pedigree analysis, where a unique ousiotype (essence) for each individual is assumed to represent unobservable genetic effects. Many other authors including Bonney (1986, 1987) adopted the ousiotype as the genotype. Frailty model (12.13) can be viewed as a further clarification of the ousiotype into a major genotype of focus and residual unobservable effects.

In terms of computation, when both U and Y are observable, the complete log-likelihood function is easy to derive, and the EM algorithm (Dempster, Laird, and Rubin 1977) can be applied to find the parameter estimates. A detailed development of the frailty model for segregation analysis will be presented elsewhere (Zhang and Merikangas 1999).

12.2 Classification Trees for Multiple Binary Responses

Many applications of parametric models have a notable common feature. That is, the models usually involve relatively few covariates, and there is little discussion of model selection. Although the theoretical models are not confined by the number of covariates, the reality of specifying parametric candidate models and then selecting the final model can be a serious challenge. To resolve this practical problem, Zhang (1998a) considered various automated approaches under the tree paradigm as a complement to the existing parametric methods. The discussions here are based on the work of Zhang (1998a).

12.2.1 Within-Node Homogeneity

Without exception, we need to define a new splitting function and cost-complexity in order to extend classification trees for the analysis of multiple discrete responses. First, we show how to generalize the entropy criterion (4.3) to the present situation making use of the log-linear model (12.9). We use the same idea as we derived (2.1). For the sake of simplicity, we assume that the joint distribution of $\mathbf{Y}$ depends on the linear terms and the sum of the second-order products of its components only. That is, we assume that the joint probability distribution of $\mathbf{Y}$ is

$$f(\mathbf{y};\Psi,\theta) = \exp(\Psi'\mathbf{y} + \theta w - A(\Psi,\theta)), \qquad (12.14)$$

where $w = \sum_{i<j} y_i y_j$. Now we define the generalized entropy criterion, or the homogeneity of node τ_L, as the maximum of the log-likelihood derived from this distribution, which equals

$$h(\tau_L) = \sum_{\{\text{subject } i\in\tau_L\}} (\hat{\Psi}'\mathbf{y}_i + \hat{\theta} w_i - A(\hat{\Psi},\hat{\theta})), \qquad (12.15)$$

where $\hat{\Psi}$ and $\hat{\theta}$ may be viewed as the maximum likelihood estimates of Ψ and θ, respectively. Obviously, the homogeneity of node t_R can be defined by analogy. The node impurity $i(\tau)$ can be chosen as $-h(\tau)$ if you will. Having defined the homogeneity (or impurity) measure, we plug it into (2.3) to form a splitting rule.

In addition to the homogeneity (12.15), there are other possibilities worth considering. If the responses were continuous, it would be natural to measure the node homogeneity through their covariance matrix. Therefore, it is reasonable to explore a homogeneity measure via a covariance matrix such as (11.38) for regression trees.

Within a node τ, we can measure its homogeneity (counter variation) in terms of the distribution of $\mathbf{Y}$ by

$$h_1(\tau) = -\log|V_\tau|, \qquad (12.16)$$

where $|V_\tau|$ is the determinant of the within-node sample covariance matrix of $\mathbf{Y}$. The use of the logarithm is to ensure the subadditivity

$$n_\tau h_1(\tau) \le n_{\tau_L} h_1(\tau_L) + n_{\tau_R} h_1(\tau_R),$$

where n_τ, n_{τ_L}, and n_{τ_R} are respectively the numbers of subjects in node τ and its left and right daughter nodes τ_L and τ_R.

When we have a single binary response, criterion (12.16) is essentially the Gini index in (4.4). This is because

$$|V_\tau| = \frac{n_\tau}{n_\tau - 1} p_\tau(1 - p_\tau),$$

where p_τ is the proportion of diseased subjects in node τ.

Further, as a direct extension from the criterion (11.38) used in the trees for continuous longitudinal data, another measure of within-node homogeneity that deserves our attention is

$$h_2(\tau) = -\frac{1}{n_\tau} \sum_{i \in \text{ node } \tau} (\mathbf{y}_i - \bar{\mathbf{y}}(\tau))' V^{-1} (\mathbf{y}_i - \bar{\mathbf{y}}(\tau)), \tag{12.17}$$

where V^{-1} is the covariance matrix of $\mathbf{Y}_i$ in the root node.

Finally, based on the discussion in the previous section, it would be more appropriate to replace the covariance matrix V_τ with a matrix constituted by the pairwise odds ratios when we deal with multiple binary responses. The consequence warrants further investigation.

12.2.2 Terminal Nodes

To construct a useful tree structure, a rigorous rule is warranted to determine the terminal nodes and hence the size of the tree. As in Section 4.2.2, we need to prepare a tree cost-complexity,

$$R_\alpha(\mathcal{T}) = R(\mathcal{T}) + \alpha|\tilde{\mathcal{T}}|,$$

as was first introduced in (4.7). Zhang (1998a) considered three definitions for the cost $R(\mathcal{T})$ with respect to h, h_1, and h_2. Using $h(\tau)$ he defined

$$R(\mathcal{T}) = -\sum_{\tau \in \tilde{\mathcal{T}}} \sum_{\{\text{subject } i \in \tau\}} \log f(\mathbf{y}_i; \hat{\Psi}, \hat{\theta}), \tag{12.18}$$

where f is introduced in (12.14), and $\hat{\Psi}$ and $\hat{\theta}$ are estimated from the learning sample. Note, however, that subject i may or may not be included in the learning sample.

Using $h_1(\tau)$ Zhang introduced

$$R_1(\mathcal{T}) = -\sum_{\tau \in \tilde{\mathcal{T}}} n_\tau \log |V_\tau|,$$

where V_τ is the covariance matrix of $\mathbf{Y}$ within node τ with the average obtained from the learning sample even though $\mathbf{Y}$ may not be included in the learning sample. It turned out that $h_1(\tau)$ and $R_1(\mathcal{T})$ are not as useful as the other choices. For the data in Section 12.3, $h_1(\tau)$ in (12.16) suffered an undesirable end-cut preference problem. This phenomenon was described at the end of Section 2.2 as a side effect of using the Gini index for a single binary outcome. Because $h_1(\tau)$ can be viewed as a generalization of the Gini index, it is not surprising that $h_1(\tau)$ manifested the problem. Thus, we remove $h_1(\tau)$ and $R_1(\mathcal{T})$ from further discussion.

Likewise, for $h_2(\tau)$ we have

$$R_2(\mathcal{T}) = -\sum_{\tau\in\tilde{\mathcal{T}}} \sum_{\{\text{subject } i\in\tau\}} (\mathbf{y}_i - \bar{\mathbf{y}}(\tau))' V^{-1} (\mathbf{y}_i - \bar{\mathbf{y}}(\tau)), \qquad (12.19)$$

where V and $\bar{\mathbf{y}}(\tau)$ are estimated from the learning sample only.

After $R_\alpha(\mathcal{T})$ is defined, the rest of the procedure is identical to that in Section 4.2.3. We should mention, however, that a theoretical derivation of the standard error for $R(\mathcal{T})$ seems formidable. As a start, Zhang (1998a) suggested repeating the cross-validation procedure ten times. This process results in an empirical estimate of the needed standard error. Although it was not explicitly stated, this in effect introduced the idea of bagging, except that it was for the purpose of determining the tree size.

*12.2.3 Computational Issues**

Because each node may have many possible splits, the homogeneity (12.15) must be computed a large number of times. Therefore, it is important to reduce the computational burden as much as possible by designing efficient algorithms. Computing $\mathbf{y}$ and w is relatively simple, so the critical part is to find $\hat{\Psi}$ and $\hat{\theta}$. To simplify the notation, we attach w to $\mathbf{y}$ and θ to Ψ and let

$$\mathbf{z} = (\mathbf{y}', w)' \text{ and } \Phi = (\Psi', \theta)'.$$

According to Fitzmaurice and Laird (1993), $\hat{\Phi}$ can be found through the following updating formulas:

$$\Phi^{(J+1)} = \Phi^{(J)} + V^{-1}(\mathbf{y})(\bar{\mathbf{y}} - I\!\!E\{\mathbf{Y}\}), \qquad (12.20)$$

where $I\!\!E\{\mathbf{Y}\}$ and $V^{-1}(\mathbf{y})$ are the mean and covariance matrix of $\mathbf{Y}$ given model parameters at $\Phi^{(J)}$, respectively, and $\bar{\mathbf{y}}$ is the sample average of $\mathbf{Y}$ within a given node. Not surprisingly, the computation of $V(\mathbf{Y})$ requires more time. Moreover, it depends on the current value $\Phi^{(J)}$ and makes the updating formula more vulnerable to a poor initial value of Φ. Both numerical and theoretical evidence suggests that it is better to replace the theoretical value of $V(\mathbf{Y})$ with the sample covariance matrix V_0 of $\mathbf{Y}$ within a given node. In our application, the use of V_0 leads to satisfactory numerical results. From a theoretical point of view, as $\Phi^{(J)}$ converges to a stable point and if the sample size is sufficiently large, $\mathbb{E}\{\mathbf{Y}\}$ and $V(\mathbf{Y})$ should be close to $\bar{\mathbf{y}}$ and V_0, respectively. So, the following simplified updating formula takes over the one in (12.20):

$$\Phi^{(J+1)} = \Phi^{(J)} + V_0^{-1}(\bar{\mathbf{y}} - \mathbb{E}\{\mathbf{Y}\}). \tag{12.21}$$

*12.2.4 Parameter Interpretation**

We have noted earlier that the canonical parameters correspond to conditional odds or odds ratios and that the conditions in these odds may not be appropriate. We illustrate here how to transform the canonical parameters to the marginal parameters that have natural interpretations.

Let $\gamma = \mathbb{E}(w)$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_q)' = \mathbb{E}(\mathbf{Y})$. Now we introduce an "overall" measure of pairwise correlations:

$$\rho = \frac{\gamma - \sum_{i<j} \mu_i \mu_j}{\sqrt{\sum_{i<j} \mu_i(1-\mu_i)\mu_j(1-\mu_j)}}. \tag{12.22}$$

Next, we show how to derive the estimates of marginal distribution parameters, $\boldsymbol{\mu}$ and ρ, and their standard errors by making use of those of $\Phi = (\Psi', \theta)'$. The estimates for $\hat{\boldsymbol{\mu}}$ and $\hat{\rho}$ can be directly computed by substituting $\hat{\Phi}$ into the distribution function. What follows explains how to find the standard errors.

It is easy to see that

$$\frac{\partial \boldsymbol{\mu}}{\partial \Phi'} = \mathrm{Cov}(\mathbf{Y}, \mathbf{Z}'), \text{ and } \frac{\partial \gamma}{\partial \Phi'} = \mathrm{Cov}(w, \mathbf{Z}').$$

By the chain rule, we have

$$\begin{aligned} \frac{\partial \rho}{\partial \Phi'} &= \frac{\partial \rho}{\partial \gamma}\frac{\partial \gamma}{\partial \Phi'} + \frac{\partial \rho}{\partial \boldsymbol{\mu}'}\frac{\partial \boldsymbol{\mu}}{\partial \Phi'} \\ &= \mathrm{Cov}(w, \mathbf{Z}')\frac{\partial \rho}{\partial \gamma} + \frac{\partial \rho}{\partial \boldsymbol{\mu}'}\mathrm{Cov}(\mathbf{Y}, \mathbf{Z}'). \end{aligned}$$

Therefore,

$$\begin{pmatrix} \frac{\partial \boldsymbol{\mu}}{\partial \Phi'} \\ \frac{\partial \rho}{\partial \Phi'} \end{pmatrix} = \begin{pmatrix} I & 0 \\ \frac{\partial \rho}{\partial \boldsymbol{\mu}'} & \frac{\partial \rho}{\partial \gamma} \end{pmatrix} \mathrm{Cov}(\mathbf{Z}) \stackrel{\text{def}}{=} JV.$$

Since V is the information matrix with respect to Φ, the information matrix for $\boldsymbol{\mu}$ and ρ is

$$\mathcal{I}(\boldsymbol{\mu}, \rho) = (VJ')^{-1}V(JV)^{-1} = (J^{-1})'V^{-1}J^{-1}.$$

Considering potential model misspecification as discussed by Fitzmaurice and Laird (1993) and Zhao and Prentice (1990), we should adopt a robust estimate for the covariance matrix of $\hat{\boldsymbol{\mu}}$ and $\hat{\rho}$ from Royall (1986) as follows:

$$\begin{aligned}\hat{V}(\hat{\boldsymbol{\mu}}, \hat{\rho}) &= [n_\tau \mathcal{I}(\hat{\boldsymbol{\mu}}, \hat{\rho})]^{-1} \sum \left[(\hat{V}\hat{J}')^{-1} \begin{pmatrix} \mathbf{y}_i - \hat{\boldsymbol{\mu}} \\ w_i - \hat{\gamma} \end{pmatrix} \right. \\ &\qquad \left. \times \begin{pmatrix} \mathbf{y}_i - \hat{\boldsymbol{\mu}} \\ w_i - \hat{\gamma} \end{pmatrix}' (\hat{J}\hat{V})^{-1} \right] [n_\tau \mathcal{I}(\hat{\boldsymbol{\mu}}, \hat{\rho})]^{-1} \\ &= \frac{1}{n_\tau^2} \hat{J} \sum \begin{pmatrix} \mathbf{y}_i - \hat{\boldsymbol{\mu}} \\ w_i - \hat{\gamma} \end{pmatrix} \begin{pmatrix} \mathbf{y}_i - \hat{\boldsymbol{\mu}} \\ w_i - \hat{\gamma} \end{pmatrix}' \hat{J}',\end{aligned}$$

where n_τ is the number of subjects in node τ and the summation is over all subjects in node τ. From the formula above it is numerically straightforward to compute the standard errors for $\hat{\boldsymbol{\mu}}$ and $\hat{\rho}$.

12.3 Application: Analysis of BROCS Data

12.3.1 Background

Building-related occupant complaint syndrome (BROCS) is a nonspecific set of related symptoms of discomfort reported by occupants of buildings. It occurs throughout the world in office buildings, hospitals, etc. The most common symptoms of BROCS include irritation of the eyes, nose, and throat; headache; and nausea. The cause of BROCS is generally not known. To enhance the understanding of BROCS, Zhang (1998a) analyzed a subset of the data from a 1989 survey of 6800 employees of the Library of Congress and the headquarters of the Environmental Protection Agency in the United States. The discussion here is similar to the analysis of Zhang (1998a). In his analysis, Zhang built trees using the entire sample. But in order to validate the trees, we divide the sample equally into two sets: one to build the tree and one to validate it. Again, we also considered 22 predictors as the risk factors of BROCS (represented by 22 questions in Table 12.1) and 6 binary responses (each of which includes a number of specific health discomforts as given in Table 12.2). The purpose is to predict the risk of BROCS by identifying contributing factors.

12.3.2 Tree Construction

Since some of the predictors have missing information, the missings together strategy described in Section 4.8.1 is adopted in the tree construction. To ensure that there is a reasonable number of subjects in every node, taking into account both the study sample size and the number of responses, Zhang (1998a) suggested not partitioning any node that has fewer than 60 subjects. In addition, the entire sample is equally divided into a learning

TABLE 12.1. Explanatory Variables in the Study of BROCS

Predictor	Questions
x_1	What is the type of your working space? (enclosed office with door, cubicles, stacks, etc.)
x_2	How is your working space shared? (single occupant, shared, etc.)
x_3	Do you have a metal desk? (yes or no)
x_4	Do you have new equipment at your work area? (yes or no)
x_5	Are you allergic to pollen? (yes or no)
x_6	Are you allergic to dust? (yes or no)
x_7	Are you allergic to molds? (yes or no)
x_8	How old are you? (16 to 70 years old)
x_9	Gender (male or female)
x_{10}	Is there too much air movement at your work area? (never, rarely, sometimes, often, always)
x_{11}	Is there too little air movement at your work area? (never, rarely, sometimes, often, always)
x_{12}	Is your work area too dry? (never, rarely, sometimes, often, always)
x_{13}	Is the air too stuffy at your work area? (never, rarely, sometimes, often, always)
x_{14}	Is your work area too noisy? (never, rarely, sometimes, often, always)
x_{15}	Is your work area too dusty? (never, rarely, sometimes, often, always)
x_{16}	Do you experience glare at your workstation? (no, sometimes, often, always)
x_{17}	How comfortable is your chair? (reasonably, somewhat, very uncomfortable, no one specific chair)
x_{18}	Is your chair easily adjustable? (yes, no, not adjustable)
x_{19}	Do you have influence over arranging the furniture? (very little, little, moderate, much, very much)
x_{20}	Do you have children at home? (yes or no)
x_{21}	Do you have major childcare duties? (yes or no)
x_{22}	What type of job do you have? (managerial, professional, technical, etc.)

This table is reproduced from Table 1 of Zhang (1998a).

TABLE 12.2. Six Clusters of BROCS

Response	Cluster	Included Symptoms
y_1	CNS	difficulty remembering/concentrating, dizziness, lightheadness, depression, tension, nervousness
y_2	Upper Airway	runny/stuffy nose, sneezing, cough, sore throat
y_3	Pain	aching muscles/joints, pain in back/shoulders/ neck, pain in hands/wrists
y_4	Flu-like	nausea, chills, fever
y_5	Eyes	dry, itching, or tearing eyes, sore/strained eyes, blurry vision, burning eyes
y_6	Lower Airway	wheezing in chest, shortness of breath, chest tightness

This table is reproduced from Table 2 of Zhang (1998a).

and a validation sample in order to assess the performance of various approaches. The learning sample is used to construct trees and the validation sample to compare the predictive power of the constructed trees.

When $h(\tau)$ in (12.15) is used as a measure of node homogeneity, we obtained an initial tree with 65 nodes. Applying $R(\mathcal{T})$ defined in (12.18) as the tree cost, we derived a sequence of 33 nested optimal subtrees from the initial tree. Figure 12.2(a) plots the log cost of these subtrees against their complexity. In contrast, the use of $h_2(\tau)$ in (12.17) results in a starting tree of 199 nodes. Then, we obtained a sequence of 69 nested optimal subtrees using $R_2(\mathcal{T})$ in (12.19) as the tree cost. See Figure 12.2(b).

The subtree cost estimate and its standard error were derived from ten repetitions of 5-fold cross-validation. Each time, we have a 5-fold cross-validation estimate of the cost for every subtree. Repeating ten times gives ten such estimates. The average and the square root of the sample variance of these ten estimates are used as the tree cost estimate and its standard error, respectively. Based on Figure 12.2, we selected a 6-terminal-node final subtree from the initial tree using $h(\tau)$ shown in Figure 12.3 and a 7-terminal-node final subtree from the other initial tree depicted in Figure 12.4.

12.3.3 Description of Numerical Results

Table 12.3 suggests that terminal node 7 in Figure 12.3 is most troublesome. Subjects in this terminal node complained about more problems in nearly all clusters than everyone else. This is because the air quality in their working area was poor, namely, often too stuffy and dusty. For the same reasons, subjects in terminal nodes 5 and 6 also reported relatively more symptoms. In contrast, subjects in terminal node 10 experienced the least

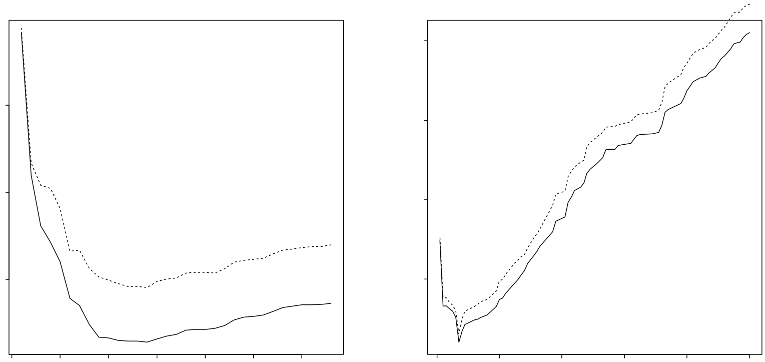

FIGURE 12.2. Cost-complexity for two sequences of nested subtrees. Panels (a) and (b) come from trees using $h(\tau)$ and $h_2(\tau)$, respectively. The solid line is the log cross-validation (CV) estimates of cost, and the dotted line is the log of one standard error above the estimated cost estimated by cross-validation

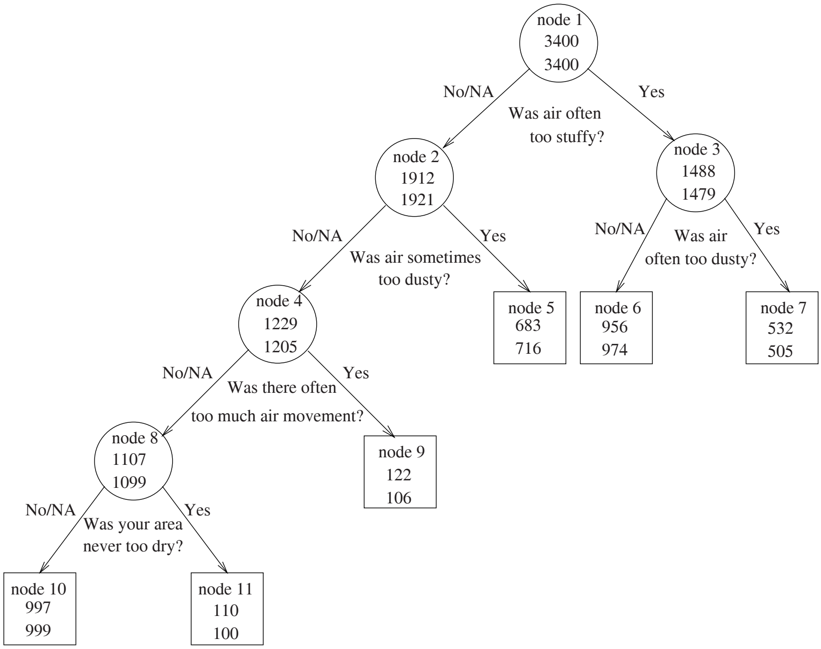

FIGURE 12.3. Tree structure for the risk factors of BROCS based on $h(\tau)$. Inside each node (a circle or a box) are the node number and the numbers of subjects in the learning (middle) and validation (bottom) samples. The splitting question is given under the node

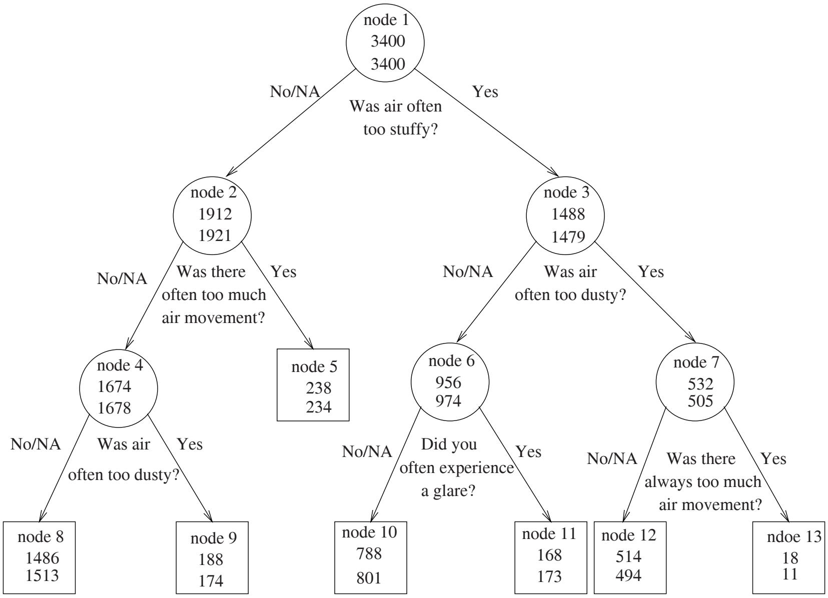

FIGURE 12.4. Tree structure for the risk factors of BROCS based on $h_2(\tau)$. Inside each node (a circle or a box) are the node number and the numbers of subjects in the learning (middle) and validation (bottom) samples. The splitting question is given under the node

TABLE 12.3. Estimates of Symptom Prevalence Rates in the Terminal Nodes of the Tree in Figure 12.3

Terminal	Cluster of symptoms					
node #	CNS	U.A.	Pain	Flu-like	Eyes	L.A.
5	0.14[†]	0.29	0.29	0.15	0.03	0.10
	0.14[¶]	0.24	0.24	0.14	0.02	0.08
6	0.21	0.30	0.35	0.16	0.05	0.07
	0.20	0.31	0.35	0.19	0.05	0.07
7	0.29	0.49	0.51	0.29	0.08	0.12
	0.27	0.49	0.47	0.25	0.06	0.11
9	0.10	0.19	0.16	0.15	0.02	0.27
	0.08	0.20	0.17	0.13	0.01	0.18
10	0.07	0.09	0.10	0.06	0.01	0.03
	0.07	0.11	0.12	0.06	0.01	0.02
11	0.21	0.26	0.24	0.17	0.05	0.09
	0.08	0.14	0.26	0.08	0.04	0.04

[†] Based on the learning sample.

[¶] Based on the validation sample.

discomfort because they had the best air quality. Overall, Figure 12.3 and Table 12.3 show the importance of air quality around the working area.

Based on a different criterion, $h_2(\tau)$, Figure 12.4 demonstrates again the importance of air quality. It uses nearly the same splits as Figure 12.3 except that "experiencing a glare" also emerged as a splitting factor. By comparing terminal nodes 10 and 11 in Figure 12.4, it appears that "experiencing a glare" resulted in more discomfort for all clusters of symptoms.

12.3.4 Alternative Approaches

We mention two alternative approaches that make direct use of the tree methods for a single outcome as described in earlier chapters. First, we could grow separate trees for individual clusters of symptoms and then attempt to summarize the information. Depending on the number of clusters, this approach could be very laborious and not necessarily as productive, as explained by Zhang (1998a). The second approach is to create a surrogate response variable. This surrogate response can be taken as the sum of the positive responses in the six clusters or a more sophisticated linear combination derived from a descriptive principal components analysis (Kleinbaum et al. 1988, p. 604). It is regarded as descriptive because the responses are binary, which do not satisfy the conditions of principal components analysis. Then, we can treat the surrogate response as a numerical variable and grow a regression tree for it. After such a regression tree is grown, we can regard it as a classification tree for the original binary outcomes. We refer to Zhang (1998a) for details.

TABLE 12.4. Estimates of Symptom Prevalence Rates in the Terminal Nodes of the Tree in Figure 12.4

Terminal	Cluster of symptoms					
node #	CNS	U.A.	Pain	Flu-like	Eyes	L.A.
5	0.15[†]	0.27	0.27	0.21	0.04	0.24
	0.12[¶]	0.25	0.26	0.18	0.02	0.21
8	0.09	0.13	0.14	0.08	0.01	0.04
	0.08	0.14	0.15	0.07	0.01	0.03
9	0.16	0.41	0.34	0.20	0.04	0.10
	0.18	0.29	0.28	0.17	0.01	0.05
10	0.19	0.29	0.30	0.13	0.04	0.06
	0.18	0.30	0.32	0.16	0.04	0.06
11	0.31	0.36	0.57	0.28	0.10	0.08
	0.28	0.37	0.51	0.30	0.08	0.08
12	0.28	0.48	0.51	0.28	0.08	0.10
	0.27	0.49	0.47	0.25	0.06	0.11
13	0.56	0.61	0.44	0.56	0.22	0.61
	0.18	0.45	0.36	0.18	0.18	0.27

[†] Based on the learning sample.

[¶] Based on the validation sample.

12.3.5 Predictive Performance

To compare the predictive performance of the trees constructed in Figures 12.3 and 12.4, we produce ROC curves (see Section 3.2 for the description of ROC curves) for individual clusters. Figure 12.5 displays two sets of ROC curves: one from the prediction rule based on Figure 12.3 and the other on Figure 12.4. In addition, the areas under the ROC curves are listed. Each panel of Figure 12.5 corresponds to a cluster. The performance of the two trees is very close, as indicated by both the ROC curves and the areas under the curves, although Figure 12.4 is decisively better than Figure 12.3 for the clusters of “flu-like” and “lower airway.”

12.4 Ordinal and Longitudinal Responses

The homogeneity $h(\tau)$ can be further extended to analyze longitudinal binary responses and polytomous responses. For longitudinal data, the time trend can be incorporated into the parameters introduced in (12.14), hence allowing $h(\tau)$ to be a function of time.

For ordinal responses, we describe the method proposed by Zhang and Ye (2008). Let z_{ij} be the jth ordinal response in the ith subject, taking a value of $1, \ldots, K$. Note here that K is the same for all response variables, although in principle we can create extra levels with zero frequency to accommodate

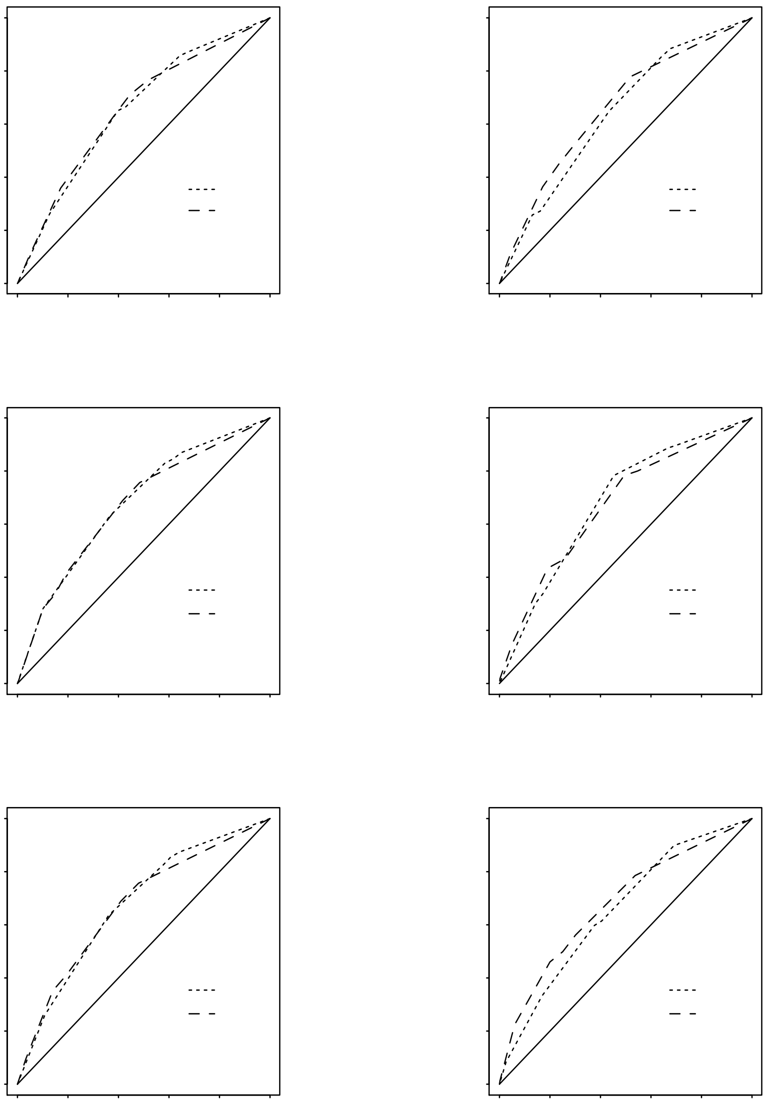

FIGURE 12.5. Comparison of ROC curves for the classifications tree in Figures 12.3 and 12.4 among individual clusters. The true positive probability (TPP) is plotted against the false positive probability (FPP). The solid line indicates the performance of a random prediction. The dotted and dashed ROC curves respectively come from Figures 12.3 and 12.4, and the areas under them are also reported

different K's. We define $K-1$ indicator variables $y_{ijk} = I(z_{ij} > k)$, for $k = 1, \ldots, K-1$. Recall $I(\cdot)$ is the indicator function. Let

$$\begin{aligned} y_{ij} &= (y_{ij,1}, \cdots, y_{ij,K-1})', \\ \mathbf{y}_i &= (y'_{i1}, \cdots, y'_{in})', \end{aligned} \qquad (12.23)$$

Then, the observed responses from the ith unit can be rewritten as

$$\mathbf{y}_i = (y_{i1,1}, \cdots, y_{i1,K-1}, \cdots, y_{in,1}, \cdots, y_{in,K-1})'.$$

Now, the components of the $\mathbf{y}_i$ are binary, and hence we can use the same procedure as described in Section 12.2.1.

12.5 Analysis of the BROCS Data via Log-Linear Models

Building a log-linear model with standard software such as SAS and SPLUS is usually a prohibitive task when we include a large number of factors into the model and consider their higher-order interactions. In the present application, given the six response variables it is not realistic to scrutinize all 22 covariates in the same model. In fact, it was still computationally too ambitious when we entered only four variables (in their original scale) that appeared in Figures 12.3 and 12.4. As a compromise, we dichotomized the four variables based on the splits and created four dummy variables: $z_1 = I(x_{10} > 3)$, $z_2 = I(x_{12} > 3)$, $z_3 = I(x_{13} > 3)$, and $z_4 = I(x_{15} > 3)$. In log-linear models, we assume that the sample counts for the 2^{10} cross-classification cells of y's and z's are independent Poisson random variables with expected values to be modeled.

We started with a model that allows for third-order interactions between two of the six response variables and one of the four dummy variables. The first `PROC CATMOD` statement of the SAS program in Table 12.5 carried out the estimation for the initial model. Insignificant terms (p-value ≥ 0.05) were removed from the model sequentially, which led to the final log-linear model with the expected cell counts specified by

$$\begin{aligned}\exp\Bigg[&\mu+\sum_{k=1}^{6}\lambda_{i_k}^{y_k}+\sum_{k=1}^{4}\lambda_{j_k}^{z_k}\\&+\left(\sum_{k=4,6}\lambda_{j_1 i_k}^{z_1 y_k}+\sum_{k\neq 3,5}\lambda_{j_2 i_k}^{z_2 y_k}+\sum_{k\neq 2,5}\lambda_{j_3 i_k}^{z_3 y_k}+\sum_{k=1}^{4}\lambda_{j_4 i_k}^{z_4 y_k}\right)\\&+\left(\sum_{l=1,2}\sum_{k=3}^{6}\lambda_{i_l i_k}^{y_l y_k}+\sum_{k=4}^{6}\lambda_{i_3 i_k}^{y_3 y_k}+\sum_{k=5}^{6}\lambda_{i_4 i_k}^{y_4 y_k}+\lambda_{i_5 i_6}^{y_5 y_6}\right)\\&+\left(\lambda_{j_2 i_2 i_4}^{z_2 y_2 y_4}+\sum_{k=4,6}\lambda_{j_3 i_3 i_k}^{z_3 y_3 y_k}+\sum_{k=1,3}\lambda_{j_4 i_k i_4}^{z_4 y_k y_4}\right)\Bigg].\end{aligned}\tag{12.24}$$

The second PROC CATMOD statement of the SAS program in Table 12.5 performed the computation for model (12.24). The results were organized in Table 12.6 in five categories based on the grouping of the terms in model (12.24).

Interpreting Table 12.6 rigorously and thoroughly would be difficult and may be even impossible because of the mutual relationship among the responses and covariates. Our attempt here is merely to extract the major message in a descriptive manner. Table 12.6 confirms the correlation between the 6 response variables. Conditional on everything else, the first response variable (CNS) appears to be uncorrelated with the second (upper airway) response because the final model does not contain the interaction: $y_1 * y_2$. Five of the 14 significant correlations between the 6 responses may be mediated by the three dummy variables z_2 to z_4. The dummy variable z_1 (air movement) has significant effects only on the mean frequency of the fourth (flu-like) and sixth (lower airway) clusters of symptoms. The air dryness (z_2) may not be significantly associated with the pain (y_3) and lower airway (y_6) symptoms. Although we have seen the importance of air stuffiness (z_3) in the tree-based analysis, the log-linear model does not suggest that it significantly affects the upper airway (y_2) and eye (y_5) problems. Finally, the dusty air (z_4) did not express significant association with the eye (y_5) and lower airway (y_6) symptoms although we expect that the dusty air would cause more eye discomfort. One might think that relatively few reports in the eye cluster perhaps limited our power; however, the model reveals its significant association with air dryness. One good explanation comes from the tree in Figure 12.4, where we see that the combination of dusty air with movement resulted in many more eye problems. Due to practical limitations, it was not possible to consider the interactions between the covariates in the initial model. As a matter of fact, the interaction,

TABLE 12.5. SAS Program for the Analysis of BROCS Data

```
data one;
infile 'BROCS.DAT';
input x1-x22 y1-y6;
run;
data two; set one;
where x10 ne . and x12 ne . and x13 ne . and x15 ne .;
z1 = (x10 > 3); z2 = (x12 > 3);
z3 = (x13 > 3); z4 = (x15 > 3);
proc sort; by z1 z2 z3 z4 y1 y2 y3 y4 y6;
proc freq noprint;
     tables z1*z2*z3*z4*y1*y2*y3*y4*y5*y6
            /list out=counts;
run;
proc catmod data=counts; weight count;
model z1*z2*z3*z4*y1*y2*y3*y4*y5*y6 = _response_
      /ml noprofile noresponse noiter;
loglin y1|y2|z1 y1|y2|z2 y1|y2|z3 y1|y2|z4
       y1|y3|z1 y1|y3|z2 y1|y3|z3 y1|y3|z4
       y1|y4|z1 y1|y4|z2 y1|y4|z3 y1|y4|z4
       y1|y5|z1 y1|y5|z2 y1|y5|z3 y1|y5|z4
       y1|y6|z1 y1|y6|z2 y1|y6|z3 y1|y6|z4
       y2|y3|z1 y2|y3|z2 y2|y3|z3 y2|y3|z4
       y2|y4|z1 y2|y4|z2 y2|y4|z3 y2|y4|z4
       y2|y5|z1 y2|y5|z2 y2|y5|z3 y2|y5|z4
       y2|y6|z1 y2|y6|z2 y2|y6|z3 y2|y6|z4
       y3|y4|z1 y3|y4|z2 y3|y4|z3 y3|y4|z4
       y3|y5|z1 y3|y5|z2 y3|y5|z3 y3|y5|z4
       y3|y6|z1 y3|y6|z2 y3|y6|z3 y3|y6|z4
       y4|y5|z1 y4|y5|z2 y4|y5|z3 y4|y5|z4
       y4|y6|z1 y4|y6|z2 y4|y6|z3 y4|y6|z4
       y5|y6|z1 y5|y6|z2 y5|y6|z3 y5|y6|z4;
run;
proc catmod data=counts; weight count;
model z1*z2*z3*z4*y1*y2*y3*y4*y5*y6 = _response_
      /ml noprofile noresponse noiter;

loglin y1*z2 y1*z3 y2*z4 y1*y3 y2*y3 y4*z1 y5*z2
       y2*y5 y3*y5 y1|y4|z4 y1|y5 y1|y6 y6|z1
       y2|y4|z2 y2|y6 y3|y4|z3 y3|y4|z4 y3|y6|z3
       y4|y5 y4|y6 y5|y6;
run;
```

TABLE 12.6. SAS Program for the Analysis of BROCS Data

```
Effect Estimate Error Prob.   Effect Estimate Error Prob.
---------------------------   ---------------------------
Y1       0.218 0.049 0.0000   Y2      -0.212 0.055 0.0001
Y3      -0.230 0.058 0.0001   Y4       0.137 0.055 0.0125
Y5       1.116 0.054 0.0000   Y6       0.427 0.053 0.0000
---------------------------   ---------------------------
Z1       0.604 0.030 0.0000   Z2       0.019 0.038 0.6161
Z3      -0.103 0.029 0.0004   Z4       0.322 0.022 0.0000
---------------------------   ---------------------------
Z1*Y4    0.112 0.028 0.0001   Z1*Y6    0.421 0.028 0.0000
Z2*Y1    0.117 0.019 0.0000   Z2*Y2    0.157 0.019 0.0000
Z2*Y4    0.080 0.020 0.0000   Z2*Y5    0.137 0.038 0.0002
Z3*Y1    0.146 0.020 0.0000   Z3*Y3    0.106 0.028 0.0001
Z3*Y4    0.100 0.021 0.0000   Z3*Y6   -0.090 0.026 0.0007
Z4*Y1    0.068 0.022 0.0016   Z4*Y2    0.214 0.018 0.0000
Z4*Y3    0.085 0.022 0.0001   Z4*Y4    0.090 0.022 0.0001
---------------------------   ---------------------------
Y1*Y3    0.210 0.020 0.0000   Y1*Y4    0.277 0.023 0.0000
Y1*Y5    0.202 0.043 0.0000   Y1*Y6    0.166 0.031 0.0000
Y2*Y3    0.345 0.017 0.0000   Y2*Y4    0.137 0.022 0.0000
Y2*Y5    0.290 0.047 0.0000   Y2*Y6    0.157 0.030 0.0000
Y3*Y4    0.189 0.023 0.0000   Y3*Y5    0.197 0.050 0.0001
Y3*Y6    0.088 0.032 0.0053   Y4*Y5    0.127 0.047 0.0061
Y4*Y6    0.250 0.032 0.0000   Y5*Y6    0.257 0.052 0.0000
---------------------------   ---------------------------
Z2*Y2*Y4 0.079 0.019 0.0000
Z3*Y3*Y4 0.063 0.020 0.0015   Z3*Y3*Y6 0.111 0.026 0.0000
Z4*Y1*Y4 0.067 0.021 0.0018   Z4*Y3*Y4 0.093 0.020 0.0000
---------------------------------------------------------
```

$z_2 * z_4 * y_5$, would be extremely significant if we knew that it should be included.

In retrospect, log-linear models provide us with the opportunity to explore the association among many categorical variables. Due to the model's complexity, we are usually confined to simplistic choices of log-linear models and have to give up the chance of exploring some important relationships. The tree-based analysis offers a fruitful complement to the use of log-linear models, particularly in dimension reduction and model specification.