Chapter 11

Review of Generalized Linear Models

11.1 INTRODUCTION

In Part II we considered methods for analyzing longitudinal data when the response variable is continuous. In many biomedical applications the longitudinal response is not continuous, for example, the presence or absence of respiratory illness, or counts of the number of epileptic seizures in a four-week interval. When the longitudinal response is discrete (e.g., binary, ordinal, or a count), the linear models discussed in Part II are no longer appropriate for relating changes in the mean response to covariates. Instead, we consider extensions of *generalized linear models* for analyzing discrete longitudinal data.

Generalized linear models provide a unified class of models for regression analysis of independent observations of a discrete or continuous response. A straightforward application of generalized linear models to longitudinal data is not appropriate due to the correlation (or lack of independence) among observations obtained from the same individual. Instead, we consider extensions of this broad class of models to handle longitudinal responses. There are many ways to extend generalized linear models to account for the correlation among longitudinal observations, we consider two general, but quite distinct, approaches in Chapters 12 through 16.

In Chapters 12 and 13 we present a unified methodology for analyzing longitudinal data when the response variable is discrete or continuous. It does not require distributional assumptions for the observations, only a regression model for the mean response. That is, we describe a general method for analyzing diverse types of longitudinal responses that avoids making assumptions about the distribution of the vector of responses; the method relies solely on assumptions about how the mean response is related to covariates. Recall that in previous chapters we noted that the multivariate normal assumption was not so crucial in longitudinal analysis of a continuous response, provided the number of subjects is relatively large in comparison to the number of repeated measures and any missing data are MCAR. In Chapter 13 we provide some rationale for why the distributional assumption for the vector of responses can be relaxed. In Chapters 14 and 15 we consider an alternative extension of generalized linear models that accounts for the correlation among longitudinal data via the introduction of random effects. These models extend in a natural way the conceptual approach represented by the linear mixed effects models discussed in Chapter 8. In Part III we focus primarily on longitudinal analysis of a discrete response, although the general methodology described in these chapters can be applied equally to continuous responses.

A characteristic feature of generalized linear models is that a suitable non-linear transformation of the mean response is related to a linear function of the covariates. This non-linearity raises some additional issues concerning the interpretation of the regression coefficients in models for longitudinal data. In Chapter 16 we emphasize that different approaches for accounting for the source of within-subject association in longitudinal data can lead to models having regression coefficients with quite distinct interpretations. As a result, for the same data, there will be differences between the estimated regression coefficients obtained from the two distinct classes of models described in Chapters 12 through 15. In general, the choice among different classes of models for discrete longitudinal data must be made on subject-matter grounds.

One of the underlying themes that will be emphasized in Part III is that different models for discrete longitudinal data have somewhat different targets of inference. Thus, to ensure that the regression model parameters bear directly on the question of scientific interest, somewhat greater care is needed in the choice of model for discrete longitudinal data.

11.2 SALIENT FEATURES OF GENERALIZED LINEAR MODELS

In this section we provide a non-technical summary of the most salient features of generalized linear models for a single, univariate response. In later chapters we discuss how generalized linear models can be extended to handle longitudinal responses. A good grasp of the material in this section is all that is required for an understanding of the methodology for longitudinal data that will be described in subsequent chapters. In Section 11.7 we present a detailed and somewhat more technical overview of generalized linear models. Many of our readers, in particular, those encountering this topic for the first time, may find the material in Section 11.7 challenging. While we encourage all of our readers to skim through Section 11.7, we note that it can be omitted without loss of continuity.

Generalized linear models provide a unified method for analyzing diverse types of univariate responses (e.g., continuous, binary, ordinal, and count data). Generalized linear models are actually a broad class or collection of regression models, and they include as special cases the standard linear regression and analysis of variance (ANOVA) models for a normally distributed continuous response, logistic regression models for a binary or dichotomous response, and log-linear or Poisson regression models for counts. Although generalized linear models encompass a much broader range of regression models, these three are among the most widely used regression models in biomedical research. In this section we focus primarily on generalized linear models for binary and count data since, with the exception of continuous responses, these two data types are by far the most commonly encountered in applications. In Section 11.4 we discuss generalized linear models for ordinal data; we devote a separate section to ordinal regression models because, when regarded as generalized linear models, the

models have certain non-standard features.

Notation

Throughout this chapter we assume that we have N *independent* observations of a response variable, Y. We let Y_i ($i = 1, \ldots, N$) denote the response variable for the i^{th} subject. Associated with each response, Y_i, is a $p \times 1$ vector of covariates,

$$X_i = \begin{pmatrix} X_{i1} \\ \vdots \\ X_{ip} \end{pmatrix}, \quad i = 1, \ldots, N;$$

where X_{ik} denotes the k^{th} covariate for the i^{th} subject. Typically, although not always, $X_{i1} = 1$ for all i, and then β_1 is the intercept term in the regression model. Generalized linear models extend the standard linear regression model in a number of important ways, while also retaining some of its distinctive features. In particular, a generalized linear model for Y_i has the following three-part specification:

1. a distributional assumption,

2. a systematic component, and

3. a link function.

We consider each of these three components in turn.

Distributional Assumption

Generalized linear models extend many of the basic concepts and ideas of standard linear regression analysis to settings where the response variable is discrete and can no longer be assumed to have a normal distribution. In particular, they extend the class of probability distributions for the response to include many of the distributions commonly used for modeling discrete responses. Generalized linear models assume that the response variable has a probability distribution belonging to the *exponential family* of distributions. The exponential family includes many distributions that the reader may already have encountered. For example, the normal, Bernoulli, binomial, and Poisson distributions all belong to the exponential family. The first component of a generalized linear model, the distributional assumption, specifies the *random component* of the model. That is, it specifies a probabilistic mechanism by which the responses are assumed to be generated.

Because the normal, Bernoulli, binomial, and Poisson distributions are members of the same family, they share some common statistical properties. In particular, the variance of the response can be expressed in terms of the product of a single scale or dispersion parameter, ϕ, and a *variance function*, denoted $v(\mu_i)$; the latter being a known function of the mean, μ_i. That is,

$$\text{Var}(Y_i) = \phi\, v(\mu_i),$$

where $\phi > 0$. The variance function, $v(\mu_i)$, describes how the variance is functionally related to

the mean of the response. The variance functions for the normal, Bernoulli, and Poisson distributions are summarized in Table 11.1. For many distributions for discrete data, ϕ is not a parameter that requires estimation but is a known constant (e.g., $\phi = 1$ for the Bernoulli and Poisson distributions); for other distributions ϕ is an unknown parameter (e.g., ϕ is the variance of the normal distribution).

Table 11.1 Canonical link and variance functions for the normal, Bernoulli, and Poisson distributions.

Distribution	Variance Function, $v(\mu)$	Canonical Link
Normal	$v(\mu) = 1$	Identity: $\mu = \eta$
Bernoulli	$v(\mu) = \mu(1 - \mu)$	Logit: $\log\left(\frac{\mu}{1-\mu}\right) = \eta$
Poisson	$v(\mu) = \mu$	Log: $\log(\mu) = \eta$

For the Bernoulli and Poisson distributions, the variance depends on the mean. This dependence of the variance on the mean is a characteristic feature of most distributions for discrete responses. On the other hand, for the normal distribution, the variance does not depend on the mean; that is, $\text{Var}(Y_i) = \phi$ (and the variance function, $v(\mu_i) = 1$). This provides some rationale for why the assumption of homogeneity of variance (or common variance) is generally adopted in the standard linear regression model for normally distributed responses. In some applications, however, the homogeneity of variance assumption is too restrictive and the variance may depend on covariates. In later sections we briefly mention how restrictive assumptions about the variance of Y_i can be relaxed.

Systematic Component

Generalized linear models not only share a common family of distributions, they also share a common regression formulation. An important aspect of the standard linear regression model that is retained in all generalized linear models is the linear regression component. This is the *systematic component* of a generalized linear model, and it specifies that the effects of the covariates, X_i, on the mean of Y_i can be expressed in terms of the following *linear predictor*, denoted by η_i,

$$\eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip},$$

where, typically, $X_{i1} = 1$ for all i, and then β_1 is the intercept. The linear predictor is simply a linear combination of the unknown regression coefficients, $\beta = (\beta_1, \ldots, \beta_p)'$ and the covariates, X_i.

The key word here is *linear*. The term "linear" in generalized linear models means that η_i must be linear in the regression parameters. This implies that the mean response (or any transformation of the mean response) can be expressed as a simple weighted sum of the regression parameters, β. For example,

$$\eta_i = \beta_1 + \beta_2 X_i,$$

$$\eta_i = \beta_1 + \beta_2 \log(X_i),$$

and

$$\eta_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2,$$

are all cases where η_i is linear in the regression coefficients, even if it is non-linear in X_i. However,

$$\eta_i = \beta_1 + e^{\beta_2 X_i},$$

and

$$\eta_i = \beta_1/(1 + \beta_2 e^{-\beta_3 X_i})$$

are examples where η_i is not linear in the regression parameters and the latter types of non-linearities are not included in the class of generalized linear models.

Thus the linearity strictly applies to the regression parameters, β, but not necessarily to the covariates. As a result the linearity restriction does not preclude relationships between the mean response and the covariates that are non-linear. This latter type of non-linearity is easily accommodated by taking appropriate transformations of the mean response (see below) and/or by transformation of the covariates (e.g., log(X) or X^2).

Link Function

The final way in which generalized linear models extend the standard linear regression model is by taking a suitable transformation of the mean response and relating the transformed mean response to the covariates. This is achieved by the introduction of a *link function*. The link function applies a transformation to the mean and then links the covariates, via the linear predictor, to the transformed mean of the distribution of the responses,

$$g(\mu_i) = \eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} = \sum_{k=1}^{p} \beta_k X_{ik} = X_i' \beta,$$

where the link function $g(\cdot)$ is some known function, for example, $\log(\mu_i)$. This implies that it is the transformed mean response that changes linearly with changes in the values of the covariates.

Thus, while in the standard linear regression model the mean response is related directly to a linear combination of the covariates, in generalized linear models, it is some appropriate transformation of the mean response, for example, $\log(\mu_i)$, that is related to a linear combination of the covariates. The linearity applies to a transformation of the mean response, or, put in a somewhat different way, the effects of covariates are assumed to be additive on a suitably transformed scale for the mean response.

The use of non-linear link functions, for instance, $\log(\mu_i)$, ensures that the model produces predictions of the mean response that are within the allowable range. For example, when analyzing a binary response, μ_i has interpretation in terms of the probability of "success" (with $0 < \mu_i < 1$). If the mean response, here the probability of success, is related directly to a linear combination of the covariates, the model can yield predicted probabilities outside of the range from 0 to 1. The use of certain non-linear link functions ensures that this cannot happen, while at the same time allowing an unbounded range of values for the regression parameters, β.

We can distinguish two main types of link functions, *canonical* link functions and *non-canonical* link functions. The former are unique and can be derived for any selected distribution; the latter are somewhat arbitrary and bear no direct relation to the selected distribution. For example, the logit link function is the canonical link function associated with the Bernoulli and binomial distributions; the probit link function is a non-canonical link function for these distributions that is often adopted for the analysis of binary data from toxicological experiments. Although, in principle, any suitable link function can be used to relate the mean response to the covariates, the choice of a canonical link function produces many of the most widely used regression models. The canonical link functions for the normal, Bernoulli, and Poisson distributions are summarized in Table 11.1.

In summary, in generalized linear models, the distribution of the response is assumed to belong to a single family of distributions known as the exponential family. The exponential family includes the normal, Bernoulli, binomial, and Poisson distributions. A transformation of the mean response is then linearly related to the covariates, via an appropriate link function. Because generalized linear models make distributional assumptions about the response variable, the regression parameters can be estimated using the method of maximum likelihood. The maximum likelihood estimates of the regression coefficients, β, are simply those values of β that are most probable (or most "likely") for the data that have actually been observed. The method of maximum likelihood provides a very general technique for estimation and for inference, that is, for estimating β, constructing confidence intervals, testing hypotheses, and assessing the adequacy of models. All of these ideas will be elaborated in Section 11.3, where we focus on two special cases of generalized linear models: logistic regression for a binary response and log-linear regression for counts. Although generalized linear models provide a very broad and flexible collection of regression models for analyzing diverse types of responses, they do have one very important restriction: they assume that observations on the response variable are independent of one another. In later chapters we will discuss how this restriction to independent observations can be relaxed to accommodate the correlated nature of the responses arising from longitudinal studies.

11.3 ILLUSTRATIVE EXAMPLES

To clarify the main ideas presented in the previous sections, we consider in greater detail two special cases of generalized linear models: logistic regression for a binary response and log-linear regression for counts. We consider each of these models in terms of their three-part specification as a generalized linear model. We also emphasize the interpretation of the regression coefficients, β, in these models. In Section 11.4 we discuss generalized linear models for ordinal data. The description of methods for extending generalized linear models for longitudinal responses presented in later chapters will assume a good working knowledge of these important regression models. As a result the reader is encouraged to master the material in this section (and Section 11.4) before proceeding to Chapters 12 through 16. This

section can be skimmed through for those with a strong background in logistic and log-linear regression models.

11.3.1 Logistic Regression for Binary Responses

Logistic regression is used widely to describe the relationship between a binary response variable (e.g., denoting "success" or "failure") and a set of covariates. In common with standard linear regression, the primary objective of logistic regression is to relate the mean of the response to a set of covariates. However, the response variable is binary rather than continuous and this has a number of consequences for modeling the mean. In this section we describe the main features of logistic regression and highlight its three-part specification as a generalized linear model. We also consider various aspects of interpretation of logistic regression coefficients. An example, using data of low-birth-weight infants, is used to illustrate the main ideas.

Let Y_i denote a binary response variable, whose two categories, for convenience, are often referred to as "success" or "failure." For example, Y_i might indicate the presence or absence of a disease. Denoting the two possible outcomes for Y_i by 1 (for "success") and 0 (for "failure"), the probability distribution of Y_i is Bernoulli, with $\Pr(Y_i = 1) = \mu_i$ (and correspondingly, $\Pr(Y_i = 0) = 1 - \mu_i$). The primary goal of logistic regression is to describe the effects of changes in a set of covariates, X_i, on the mean μ_i. For ease of exposition we first consider the simple case where there is only a single covariate, X_i. Generalizations to more than one covariate will be considered later.

Since the analytic goal is to investigate the relationship between μ_i and X_i, and since linear regression plays such a dominant role in applications, it may at first seem natural to assume a linear model relating the mean of Y_i to X_i,

$$E(Y_i|X_i) = \mu_i = \beta_1 + \beta_2 X_i.$$

However, this linear model for the probabilities has one obvious difficulty. Expressing μ_i as a linear function violates the restriction that probabilities must lie within the range from 0 to 1. For sufficiently large or small values of X_i, this regression model will yield predicted probabilities outside of the range from 0 to 1. A further difficulty with the linear model for μ_i is that we often expect a non-linear relationship between μ_i and X_i. For example, a 0.2 unit increase in μ_i might be considered more "extreme" when $\mu_i - 0.1$ than when $\mu_i = 0.5$. In terms of ratios, the change from $\mu_i = 0.1$ to $\mu_i = 0.3$ represents a three-fold or 200% increase, whereas the change from $\mu_i = 0.5$ to $\mu_i = 0.7$ represents only a 40% increase. In a sense, the units of measurement for a probability (or proportion) are often not considered to be constant over the range from 0 to 1. The linear probability model simply does not take this into consideration when relating μ_i to X_i. Note also that the usual assumption of homogeneity of variance (or constant variance) in linear regression would be violated since the variance of a

binary response explicitly depends on the mean, with

$$\text{Var}(Y_i) = \mu_i\,(1 - \mu_i)\,.$$

To circumvent these difficulties with the linear probability model, a non-linear transformation can be applied to μ_i and the transformed probabilities are related linearly to X_i. When the logit or logistic function, $\log\{\mu_i/(1 - \mu_i)\}$, is adopted, the resulting model

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \text{logit}(\mu_i) = \beta_1 + \beta_2 X_i,$$

is known as the *logistic regression* model. Recall that if μ_i is the probability of success, then $\mu_i(1 - \mu_i)$ is known as the *odds* of success. For example, if the probability of success is 0.8 then the odds of success is 4 (or 0.8/0.2) to 1. That is, the probability of success is 4 times as large as the probability of failure. Thus the logistic regression model assumes a linear relationship between the log odds of success and X_i. For the reader unfamiliar with logistic regression, it is useful to bear in mind that the transformation of μ_i in logistic regression has the following property: as the probability of success, μ_i, increases, so too does the odds of success and the log odds of success; similarly, as the probability of success decreases, so too does the odds of success and the log odds of success.

Next consider the interpretation of the logistic regression coefficients, β_1 and β_2. For the special case where the predictor variable X_i is dichotomous, taking values of 0 and 1, the logistic regression slope, β_2, has a simple and very attractive interpretation in terms of the log odds ratio (comparing the log odds of success when $X_i = 1$ to the log odds of success when $X_i = 0$). That is,

$$\text{logit}(\mu_i|X_i = 1) - \text{logit}(\mu_i|X_i = 0) = (\beta_1 + \beta_2) - \beta_1 = \beta_2.$$

Thus $\exp(\beta_2)$ has interpretation as the odds ratio of the response for the two possible values of the covariate.

In simple linear regression the interpretation of the slope of the regression is in terms of changes in the mean of Y_i for a single-unit change in X_i. Similarly, for arbitrary X_i, the logistic regression slope β_2 has interpretation as the change in the log odds (of success) for a unit change in X_i. Equivalently, a unit change in X_i increases or decreases the odds of success *multiplicatively* by a factor of $\exp(\beta_2)$. Also recall that the intercept in simple linear regression has interpretation as the mean value of the response variable when X_i is equal to zero. Similarly, the logistic regression intercept, β_1, has interpretation as the log odds (of success) when $X_i = 0$; alternatively,

$$\frac{\exp(\beta_1)}{1 + \exp(\beta_1)}$$

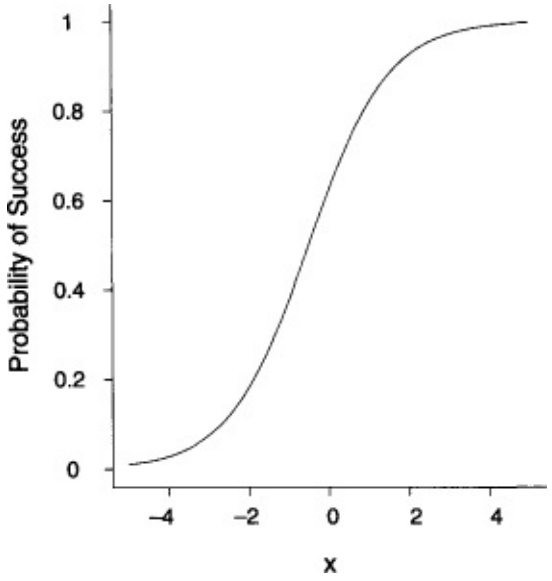
is the probability of success when $X_i = 0$.

The logistic regression model can also be expressed in terms of the probability of success, μ_i,

$$\mu_i = \frac{\exp(\beta_1 + \beta_2 X_i)}{1 + \exp(\beta_1 + \beta_2 X_i)}.$$

While the latter expression may appear to be somewhat more complicated, this is simply an equivalent way of expressing the logistic regression model. That is, logistic regression describes how the log odds, $\log(\frac{\mu_i}{1-\mu_i})$, has a linear relationship with X_i, which is equivalent to describing how μ_i has a sigmoidal or S-shaped relationship with increasing values of $\beta_{i2}X$. (See Figure 11.1 for a plot of μ versus X when $\beta_1 = 0.5$ and $\beta_2 = 0.9$.) Of note, the logistic transformation ensures that the predicted probabilities are restricted to the range from 0 and 1, while allowing an unbounded range for β_1 and β_2.

Fig. 11.1 Plot of logistic response function, with success probability, $\mu = \frac{e^{0.5+0.9X}}{1+e^{0.5+0.9X}}$.



When viewed as a generalized linear model, logistic regression is simply the special case where the distribution of Y_i is assumed to be Bernoulli (a member of the exponential family) and a logit link function, the canonical link function, has been adopted. Because the Bernoulli distribution is a one-parameter exponential family distribution, the variance of Y_i can be expressed explicitly in terms of the mean, via the following variance function:

$$\text{Var}(Y_i) = v(\mu_i) = \mu_i(1 - \mu_i)$$

and $\phi = 1$. For the Bernoulli distribution, the dispersion parameter is a fixed and known constant ($\phi = 1$).

When X_i is a discrete covariate with J distinct categories or levels (e.g., treatment groups), the binary responses for the N individuals can be grouped. Let m_j denote the number of individuals with the j^{th} covariate pattern, and let Y_j denote the number of successes among the m_j individuals, for $j = 1, \ldots, J$. We may provisionally assume that all individuals within a group respond independently with constant probability of success, μ_j, depending only on group. Then Y_j, the number of successes in the j^{th} group, has a binomial distribution with probability of success, μ_j. The binomial distribution belongs to the exponential family and the probability of success, μ_j, can be related to group using a logit link function. For the binomial distribution,

the mean or expected number of successes for the j^{th} covariate pattern is

$$E(Y_j) = m_j \mu_j.$$

There is a well-known relationship between the mean and variance of Y_j, with

$$\text{Var}(Y_j) = m_j \mu_j (1 - \mu_j).$$

However, in many biomedical applications, counts of the number of successes have variability that far exceeds that predicted by the binomial distribution; this phenomenon is often referred to as *overdispersion* (although underdispersion can also arise, it is far less common). Overdispersion is a common indicator of failure of the binomial assumptions: independent observations with constant probability of success. That is, overdispersion can be represented either by a positive correlation between the responses or by variation in the response probabilities. To allow for overdispersion or extra-binomial variation, a scale factor ϕ (with $\phi \neq 1$) is often included in the specification of the binomial variance,

$$\text{Var}(Y_j) = \phi m_j \mu_j (1 - \mu_j).$$

Failure to account for overdispersion has negligible impact of the estimated logistic regression coefficients. That is, the regression parameter estimates are consistent and there is usually little loss of efficiency. Neglecting overdispersion, however, results in the standard errors being underestimated and failure to make an adjustment to the nominal standard errors can result in misleading inferences concerning the regression coefficients (e.g., confidence intervals that are too narrow and *p*-values that are too small). Also model selection strategies based on likelihood ratio tests or on information criteria, such as the Akaike information criterion (AIC), will perform poorly. When overdispersion is ignored, a model with too many parameters is likely to be selected and thus can lead to overinterpretation of these parameters (e.g., unnecessary inclusion of interactions). Adjustments to the nominal standard errors to account for overdispersion can be made either by including a scale factor ϕ in the specification of the binomial variance,

$$\text{Var}(Y_j) = \phi m_j \mu_j (1 - \mu_j),$$

or by basing standard errors on the so-called "sandwich" estimator of Cov($\widehat{\beta}$); the latter will be discussed in greater detail in Chapter 13.

So far we have only considered the simple case where there is a single predictor X_i. Next, we consider the case where X_i is a $p \times 1$ vector of covariates. The logistic regression model becomes

$$\log\{\mu_i/(1 - \mu_i)\} = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip},$$

where $X_{i1} = 1$ for all $i = 1, \ldots, N$. The logistic regression coefficients in this model have the following interpretations. Each of the logistic regression slopes, β_k (for $k = 2, \ldots, p$), represents the change in the log odds (of success) for a unit change in X_{ik} given that all of the other predictor variables remain constant. This is completely analogous to the interpretation of the regression coefficients in multiple linear regression. Thus, by holding the remaining predictors at some fixed set of values and not allowing them to vary with any changes in X_{ik}, a single-unit increase in X_{ik} is predicted to increase or decrease the log odds of success by an amount β_k. Equivalently, a single-unit increase in X_{ik} increases or decreases the odds of

success *multiplicatively* by a factor of $\exp(\beta_k)$. The logistic regression intercept, β_1, now has interpretation as the log odds (of success) when all covariate values are set to zero. Alternatively,

$$\frac{\exp(\beta_1)}{1+\exp(\beta_1)}$$

is the probability of success when $X_{i2} = X_{i3} = \cdots = X_{ip} = 0$.

Finally, the logistic regression model for binary data can also be developed through the notion of an underlying latent variable distribution. Suppose that L_i is a latent (i.e., unobserved) continuous variable and that a positive response is observed only when L_i exceeds some threshold denoted by τ. The observed binary response can be thought of as a categorization of the unobservable latent variable, above and below the threshold τ. That is,

$$Y_i = 1 \text{ if } L_i > \tau,$$

$$Y_i = 0 \text{ if } L_i \leq \tau.$$

Suppose that the latent variable, L_i, has a standard logistic distribution. The standard logistic distribution (with mean zero and variance $\pi^2/3$) is a symmetric distribution and is very similar to the standard normal distribution, except that it has longer tails (and larger variance). Then, using calculus, it can be shown that the relationship between the observable binary response variable, Y_i, and the unobservable latent variable, L_i, is given by

$$\begin{aligned}\Pr(Y_i = 1) &= \Pr(L_i > \tau) \\ &= \int_\tau^\infty \frac{\exp(u)}{\{1+\exp(u)\}^2} du \\ &= \frac{\exp(-\tau)}{1+\exp(-\tau)}.\end{aligned}$$

Next suppose that the following linear model for L_i holds:

$$L_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + e_i = X_i'\beta + e_i,$$

where e_i (or $L_i - X_i'\beta$) is assumed to have a standard logistic distribution, with mean zero and variance $\pi^2/3$. Here we regard the threshold of L_i as fixed and the location or mean of the distribution of L_i as changing with X_i. Without loss of generality, we can assume the threshold for categorizing L_i is $\tau = 0$, since any non-zero values for the threshold would simply be absorbed into the intercept term in the linear predictor, $X'_i\beta$. Then the relationship between Y_i and L_i results in a logistic regression model for $\Pr(Y_i = 1)$. That is,

$$\begin{aligned}\Pr(Y_i = 1) &= \Pr(L_i > 0) \\ &= \Pr(L_i - X_i'\beta > -X_i'\beta) \\ &= \frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)}.\end{aligned}$$

Thus the linear model for L_i with standard logistic errors,

$$L_i = X_i'\beta + e_i,$$

implies the logistic regression model for Y_i,

$$\log\left\{\frac{\Pr(Y_i = 1)}{1 - \Pr(Y_i = 1)}\right\} = X_i'\beta.$$

Similarly, if a probit link function is adopted instead of a logit link function, the linear model for L_i with standard normal errors, instead of standard logistic errors, implies a probit regression model for Y_i.

Although the logistic regression model can be derived from the notion of a latent variable distribution, assuming the existence of a latent variable is not a necessary requirement for the use of logistic regression models (indeed, in practice, the existence of a latent variable is usually not verifiable from the data). In later chapters we use the notion of an underlying latent variable distribution to derive analogues of the between-subject and within-subject sources of variability in models for longitudinal binary responses.

Illustration

Next we consider an application of logistic regression to illustrate how the model can be used in practice. The data are from a study of low-birth-weight infants in a neonatal intensive care unit. In this example we are interested in the development of bronchopulmonary dysplasia (BPD), a chronic lung disease, in a sample of 223 infants weighing less than 1750 grams (Van Marter et al., 1990).

Let Y_i be a binary response, with $Y_i = 1$ if the i^{th} infant develops BPD by day 28 of life and $Y_i = 0$ otherwise (where BPD is defined by both oxygen requirement and compatible chest radiograph). To examine whether there is an association between the risk of BPD and birth weight (in grams $\times 10^{-2}$), we consider the following logistic regression model:

$$\log\{\mu_i/(1-\mu_i)\} = \beta_1 + \beta_2 \,\text{Weight}_i,$$

where $\mu_i = E(Y_i) = \Pr(Y_i = 1)$. For the 223 infants in the sample, the estimated logistic regression parameters (and standard errors), obtained using maximum likelihood, are displayed in Table 11.2.

Table 11.2 Estimated coefficients and standard errors for logistic regression of BPD on birth weight.

Variable	Estimate	SE	Z
Intercept	4.0343	0.6958	5.798
Birth Weight	−0.4229	0.0641	−6.599

The estimated logistic regression is

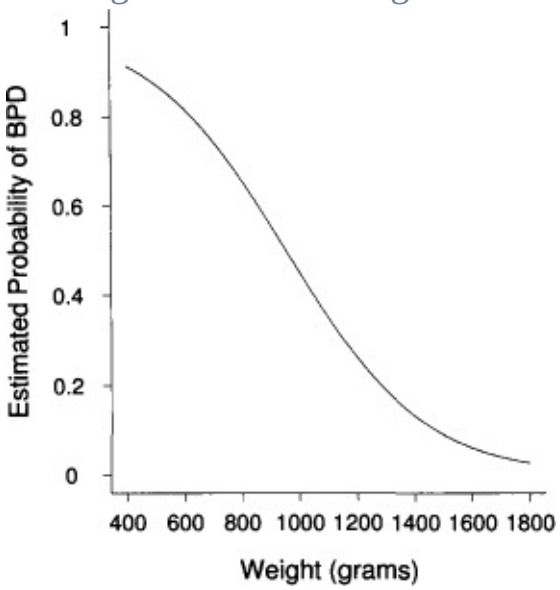
$$\log\{\hat{\mu}_i/(1-\hat{\mu}_i)\} = 4.0343 - 0.4229\,\text{Weight}_i.$$

When compared to its standard error, the ML estimate of β_2, the slope for birth weight, is significantly different from zero at the 0.05 level. The results from the logistic regression analysis indicate that the risk of BPD decreases with increasing birth weight. Specifically, the estimate of β_2 implies that for every 100 gram increase in birth weight, the log odds of BPD decreases by 0.42. For example, the odds of BPD for an infant weighing 1200 grams (approximately 2.5 pounds) is

$$\exp(4.0343 - 12 \times 0.4229) = \exp(-1.0405) = 0.353.$$

Thus the predicted probability of BPD is 0.353/ (1 + 0.353) ≈ 0.26. The estimated probability of BPD can be calculated at any birth weight and a plot of the estimated probability versus weight produces the characteristic sigmoidal or S-shaped curve displayed in Figure 11.2.

Fig. 11.2 Plot of estimated logistic response function of BPD on birth weight based on a sample of 223 infants with birth weight less than 1750 grams.



Next suppose that we include two additional covariates, gestational age (in weeks) and presence of toxemia (with 1 denoting the presence of toxemia and 0 its absence). That is, we consider the following logistic regression model:

$$\log\{\mu_i/(1-\mu_i)\} = \beta_1 + \beta_2\,\text{Weight}_i + \beta_3\,\text{Age}_i + \beta_4\,\text{Toxemia}_i.$$

For the 223 infants in the sample, the estimated logistic regression parameters (and standard errors) are displayed in Table 11.3.

Table 11.3 Estimated coefficients and standard errors for logistic regression of BPD on birth weight, gestational age and toxemia.

Variable	Estimate	SE	Z
Intercept	13.9361	2.9826	4.672
Birth Weight	−0.2644	0.0812	−3.254
Gestational Age	−0.3885	0.1149	−3.382
Toxemia	−1.3437	0.6075	−2.212

The estimated coefficient for birth weight has now decreased, when gestational age and toxemia are included in the analysis. Nonetheless, the estimate of β_2 remains significantly different from zero at the 0.05 level. The estimated coefficient for gestational age has interpretation in terms of the change in the log odds of BPD for a 1-week increase in gestational age, adjusting for birth weight and toxemia. Specifically, a 1-week increase in gestational age is associated with a 0.39 decrease in the log odds of developing BPD. Finally, the estimated coefficient for toxemia has interpretation in terms of the log odds ratio,

comparing mothers who were diagnosed with toxemia to mothers who were not, while adjusting for the effects of birth weight and gestational age. Specifically, the adjusted odds ratio is 0.26 (or $e^{-1.34}$) and indicates that infants of mothers diagnosed with toxemia have approximately a quarter the risk of developing BPD.

11.3.2 Log-Linear Regression for Counts

Log-linear regression, often referred to as Poisson regression, is used widely for the analysis of counts of the number of times some event occurs in either time or space. For example, the response variable might be the count of the number of epileptic seizures a particular patient experiences in a 4-week interval. Alternatively, the response might be a count of bacteria present in a fixed volume of bacterial suspension. In either case the response variable Y_i is a count, and the objective of log-linear regression is to relate the mean or expected count to a set of covariates.

If the occurrences of some event are counted within an interval of time (or sometimes volume or area), then the *rate* at which the event occurs is usually of more direct interest than the corresponding count. That is, the count or absolute number of events is often not satisfactory because any comparisons depend almost entirely on the "time at risk" (or, in other contexts, the sizes of the groups or areas of regions) that generated the observations. For example, it would not be very meaningful to compare counts of the number of seizures in a 4-week interval with counts of the number of seizures in a 12-month interval since it seems reasonable to suppose that the number of seizures is directly proportional to the period at risk. When the "time at risk" is not the same for all observations, a rate provides a meaningful basis for direct comparison. In either case the primary objective of log-linear regression is to relate the expected counts or rates to a set of covariates.

When the response is a count it is often reasonable to assume that Y_i has a Poisson distribution, although it is important to note that this is an assumption and it may not be valid. This is in contrast to the binary data case where the distribution of a binary response is always Bernoulli with mean μ_i. The Poisson distribution describes the probability that a specific number of events, say y_i, occurs,

$$\Pr(Y_i = y_i) = e^{-\mu_i}\mu_i^{y_i}/y_i!; \quad y_i = 0, 1, 2, \ldots$$

where $y! = y \times (y - 1) \times (y - 2) \times \ldots \times 2 \times 1$. The Poisson distribution is completely determined by a single parameter, $\mu_i = E(Y_i) > 0$, the mean number of events. A distinctive property of the Poisson distribution is that the mean and variance of Y_i are equal,

$$E(Y_i) = \mu_i = \mathrm{Var}(Y_i).$$

Note that μ_i is defined as the expected count or number of events. The expected rate is given by μ_i/T_i, where T_i is a relevant measure of the "time at risk" (e.g., T_i might be an interval of time, the person-years of observation, or the size of a group). In log-linear regression the goal is to describe the effects of a set of covariates, X_i, on the expected rate. Once again, for ease of

exposition, we will first consider the simple case where there is only a single covariate, X_i. Generalizations to more than one covariate will be considered later.

Because a rate of occurrence of some event cannot be negative, a standard linear regression model relating μ_i/T_i directly to X_i is somewhat unappealing. That is, for sufficiently large or small values of X_i, a standard linear regression model could yield predicted rates that are negative. Instead, we can relate a transformation of the rate directly to X_i. When a logarithmic transformation is adopted, the resulting model

$$\log(\mu_i/T_i) = \beta_1 + \beta_2 X_i$$

is known as the *log-linear regression* model. Recall that T_i is known and fully observed. As a result the log-linear regression model can also be expressed as

$$\log(\mu_i) = \log(T_i) + \beta_1 + \beta_2 X_i,$$

since $\log(\mu_i/T_i) = \log(\mu_i) - \log(T_i)$. Note that although $\log(T_i)$ appears on the right-hand side of the regression equation, it does not have a regression coefficient attached to it. That is, the regression parameter for $\log(T_i)$ is known to be equal to 1 and does not require estimation. We refer to $\log(T_i)$ as an *offset*. Thus the log-linear regression model assumes a linear relationship between the log rate of occurrence of some event and X_i.

When viewed as a generalized linear model, log-linear regression is simply the special case where the distribution of Y_i is assumed to be Poisson (a member of the exponential family) and a log link function, the canonical link function for the Poisson distribution, has been adopted. Because the Poisson distribution is a one-parameter exponential family distribution, the variance of Y_i can be expressed explicitly in terms of the mean, via the following variance function:

$$\text{Var}(Y_i) = v(\mu_i) = \mu_i.$$

For the Poisson distribution, the dispersion parameter is a fixed constant ($\phi = 1$). However, in many biomedical applications, count data have variability that far exceeds that predicted by the Poisson distribution. Overdispersion or extra-Poisson variation is a common indicator of failure of the Poisson assumption when dealing with count data. The implications of overdispersion for count data are the same as for grouped binary data. As discussed earlier, neglecting overdispersion results in the standard errors being underestimated and failure to make an adjustment to the nominal standard errors can result in misleading inferences concerning the regression coefficients (e.g., confidence intervals that are too narrow and *p*-values that are too small). Also model selection strategies based on likelihood ratio tests or on information criteria (e.g., AIC) will perform poorly. Adjustments to the nominal standard errors to account for overdispersion can be made either by including a scale factor ϕ in the specification of the Poisson variance,

$$\text{Var}(Y_i) = \phi\,\mu_i,$$

or by basing standard errors on the "sandwich" estimator of $\text{Cov}(\widehat{\beta})$; the "sandwich" estimator will be discussed in greater detail in Chapter 13. A third method of accounting for overdispersion is to incorporate an extra source of variability in the model for the counts; this

approach to handling overdispersion is discussed in detail in Section 11.5.

Next consider the interpretation of the log-linear regression coefficients, β_1 and β_2. For the special case where the predictor variable X_i is dichotomous, taking values of 0 and 1, the log-linear regression slope, β_2, has a simple and very attractive interpretation in terms of the log rate ratio (comparing the log expected rate when $X_i = 1$ to the log expected rate when $X_i = 0$). That is,

$$\log(\mu_i|X_i = 1) - \log(\mu_i|X_i = 0) = \{\log(T_i) + \beta_1 + \beta_2\} - \{\log(T_i) + \beta_1\} = \beta_2.$$

Thus $\exp(\beta_2)$ has interpretation as the rate ratio

$$\frac{(\mu_i|X_i = 1)}{(\mu_i|X_i = 0)}$$

for the two possible values of the covariate.

For arbitrary X_i the slope β_2 has interpretation as the change in the log expected rate for a single-unit change in X_i. Equivalently, a unit change in X_i increases or decreases (depending on the sign of β_2) the rate of occurrence of the event *multiplicatively* by a factor of $\exp(\beta_2)$. Thus, when exponentiated, the regression coefficients can be interpreted in terms of relative rates. This becomes more apparent if we express the log-linear regression model as

$$\mu_i = (\mu_i|X_i) = E(Y_i|X_i) = T_i \times e^{\beta_1} \times e^{\beta_2 X_i}.$$

From this expression it can be seen that a single-unit increase in X_i increases or decreases μ_i/T_i by a factor of e^{β_2}. That is,

$$(\mu_i|X_i + 1) = T_i \times e^{\beta_1} \times e^{\beta_2(X_i+1)} = T_i \times e^{\beta_1} \times e^{\beta_2 X_i} \times e^{\beta_2} = e^{\beta_2} \times (\mu_i|X_i).$$

On the other hand, the intercept, β_1, has interpretation as the log expected rate when $X_i = 0$; alternatively, $\exp(\beta_1)$ is the expected rate of occurrence of the event when $X_i = 0$.

So far we have only considered the simple case where there is a single predictor X_i. Next we consider the case where X_i is a $p \times 1$ vector of covariates. The log-linear regression model becomes

$$\log(\mu_i) = \log(T_i) + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip},$$

where $X_{i1} = 1$ for all $i = 1, \ldots, N$. The log-linear regression coefficients in this model have the following interpretations. Each of the log-linear regression slopes, β_k (for $k = 2, \ldots, p$), has interpretation as the change in the log expected rate for a unit change in X_{ik} given that all of the other covariates remain constant. Thus, holding the remaining covariates at some fixed set of values and not allowing them to vary with any changes in X_{ik}, we can predict a single-unit increase in X_{ik} to increase or decrease the log expected rate by an amount β_k. Equivalently, a single-unit increase in X_{ik} increases or decreases the expected rate *multiplicatively* by a factor of $\exp(\beta_k)$. The log-linear regression intercept, β_1, now has interpretation as the log expected rate of occurrence of the event when all covariate values are set to zero. Alternatively, $\exp(\beta_1)$ is the expected rate when $X_{i2} = X_{i3} = \cdots = X_{ip} = 0$.

Illustration

Next we consider an application of log-linear regression to illustrate how the model can be used in practice. The data for this illustration arise from a prospective study of potential risk factors for coronary heart disease (CHD) (Rosenman et al., 1975). The study observed 3154 men aged 40 to 50 for an average of 8 years and recorded the incidence of cases of CHD. The potential risk factors included smoking, blood pressure, and personality/behavior type. The data are summarized in Table 11.4.

Table 11.4 Data on incidence of CHD and associated risk factors.

Person-Years	Smoking[a]	Blood Pressure[b]	Behavior[c]	CHD
5268.2	0	0	0	20
2542.0	10	0	0	16
1140.7	20	0	0	13
614.6	30	0	0	3
4451.1	0	0	1	41
2243.5	10	0	1	24
1153.6	20	0	1	27
925.0	30	0	1	17
1366.8	0	1	0	8
497.0	10	1	0	9
238.1	20	1	0	3
146.3	30	1	0	7
1251.9	0	1	1	29
640.0	10	1	1	21
374.5	20	1	1	7
338.2	30	1	1	12

Source: From *Practical Biostatistical Methods*, 1st edition, by Steve Selvin. © 1995. Reprinted with permission of Brooks/Cole, a division of Thomson Learning: www.thomsonrights.com.

[a] 0: Non-smoker, 10: 1–10 cigarettes/day, 20: 11–20 cigarettes/day, 30: 30+ cigarettes/day.

[b] 0: < 140, 1: ≥ 140.

[c] 0: Type B Personality; 1: Type A Personality.

Let Y_i denote the count of the number of cases of CHD and T_i denote the person-years of follow-up. Person-years of follow-up is calculated as the total duration of observed follow-up, from entry into the study until either disease detection or end of follow-up, for the individuals in each risk group. To examine whether the rates of CHD are related to the smoking exposure variable we consider the following log-linear regression model:

$$\log(\mu_i/T_i) = \beta_1 + \beta_2 \text{Smoke}_i,$$

where $\mu_i = E(Y_i)$ and Smoke_i is a quantitative measure of smoking exposure (0: Non-smoker, 10: 1–10 cigarettes/day, 20: 11–20 cigarettes/day, 30: 30+ cigarettes/day). To adjust for differences in the total person-years of follow-up for each risk group, $\log(T_i)$ is included in the model for Y_i as an offset.

The ML estimate of the slope for smoking exposure, β_2, is 0.0318 (SE = 0.0056) and when compared to its standard error is significantly different from zero at the 0.05 level. This indicates that increases in the smoking exposure increase the log expected rate of CHD. That is, the expected rate of CHD for individuals who smoke one pack of cigarettes (or 20 cigarettes) per day is estimated to be approximately twice (or $e^{0.0318\times 20} = 1.88$ times) as high as the rate of CHD for non-smokers.

Because risk factors for CHD are likely to be correlated, we consider the impact of smoking on the rates of CHD after adjusting for the potential confounding effects of blood pressure and personality type. High blood pressure and Type A behavior pattern are known to be associated with high rates of CHD. Specifically, we consider the following log-linear regression model:

$$\log(\mu_i/T_i) = \beta_1 + \beta_2 \,\text{Smoke}_i + \beta_3 \,\text{BP}_i + \beta_4 \,\text{Type}_i,$$

where $\text{BP}_i = 1$ if blood pressure ≥ 140 and 0 otherwise; $\text{Type}_i = 1$ if Type A personality and $\text{Type}_i = 0$ if Type B personality.[1] The estimated log-linear regression parameters (and standard errors) are displayed in Table 11.5.

Table 11.5 Estimated coefficients and standard errors for log-linear regression of expected rate of CHD on smoking, blood pressure and personality type.

Variable	Estimate	SE	Z
Intercept	−5.4202	0.1308	−41.44
Smoke	0.0273	0.0056	4.88
Personality Type	0.7526	0.1362	5.53
Blood Pressure	0.7534	0.1292	5.83

The estimated coefficient for smoking, 0.027, has now decreased, when blood pressure and personality type have been controlled for in the analysis. Nonetheless, the estimate of β_2 remains significantly different from zero at the 0.05 level. The estimated coefficient for smoking has interpretation in terms of the change in the log expected rate of CHD, after adjusting for the effects of blood pressure and personality type. Specifically, the adjusted rate of CHD (controlling for blood pressure and behavior type) for individuals who smoke one pack of cigarettes (or 20 cigarettes) per day is estimated to be 1.7 (or $e^{0.027\times 20} = 1.704$) times higher than the rate of CHD for non-smokers.

There is also a very strong relationship between Type A behavior pattern and CHD incidence. The adjusted rate ratio (comparing Type A to Type B behavior pattern) is 2.12 (or $e^{0.7526}$), indicating that the rate of CHD among Type A individuals is approximately twice that among Type B individuals. Moreover this adjusted estimate of risk cannot be explained by the association of personality type with smoking and blood pressure since the latter two risk factors have been adjusted for in the analysis.

11.4 ORDINAL REGRESSION MODELS

In Section 11.3.1 we discussed regression models for a binary response, a categorical variable with only two levels. In this section we consider regression models for an ordinal response with three or more levels. The reason for devoting a separate section to ordinal regression models is that, when regarded as generalized linear models, they have certain non-standard features. As we will see, ordinal regression models can be regarded as *multivariate*, rather than *univariate*, generalized linear models. Although the main focus of this section is on methods for the analysis of an ordinal response, we conclude this section with a brief discussion of regression models for a nominal or unordered response having more than two levels.

Let us begin by defining what is meant by ordinal data. An ordinal response is a categorical variable whose categories can be naturally ordered, although the precise quantitative distance or spacing between categories is unknown. For example, "cancer stage" is a four-level ordinal variable that describes how far cancer has spread anatomically and can be used to categorize patients with similar prognosis. These four stages are denoted by the roman numerals I through IV, where stage I represents small localized cancers that are usually curable while stage IV represents inoperable or metastatic cancer. A second example of an ordinal variable is socioeconomic status (SES). Socioeconomic status is typically broken into three ordinal categories: high SES, middle SES, and low SES. When defined in this way, a family categorized as "high SES" is higher in the SES hierarchy than a family categorized as "middle SES" (or "low SES"); however, we cannot say "how much higher." Finally, ordinal variables are frequently used for subjective assessments of quality, importance, or relevance. For example, subjective scales often have response categories of "strongly agree," "agree," "disagree," and "strongly disagree."

11.4.1 Notation

Throughout the remainder of this section we assume that we have N independent observations of an ordinal response. We let Y_i ($i = 1, \ldots, N$) denote an ordinal response with K ordinal categories ($1, \ldots, K$) for the i^{th} subject. Note that the actual integer values, $1, \ldots, K$, are not particularly relevant except that larger values are assumed to correspond to "higher" outcomes and smaller values to "lower" outcomes. The distribution of Y_i is multinomial (a generalization of the binomial distribution), with K multinomial probabilities, $\Pr(Y_i = k)$ for $k = 1, \ldots, K$, for the distinct ordinal categories; note, there are only $K - 1$ non-redundant multinomial probabilities because the K probabilities are constrained to sum to 1. Associated with each response, Y_i, is a $p \times 1$ vector of covariates, $(X_{i1}, \ldots, X_{ip})'$. Next we consider a regression model for the ordinal response that is a direct extension of the familiar logistic regression model for a binary response. However, instead of directly applying the logit transformation to the mean of the ordinal response, the transformation is applied to the *cumulative* response probabilities. This leads to a regression model known as the proportional odds model. The proportional odds model is probably the most widely used model for the analysis of ordinal responses.

11.4.2 Proportional Odds Model

To develop the regression model, suppose that we dichotomize the ordinal outcome at 1 versus greater than 1, creating the binary response

$$U_{i1} = \begin{cases} 1 & \text{if } Y_i = 1, \\ 0 & \text{if } Y_i > 1. \end{cases}$$

Letting $F_{i1} = \Pr(U_{i1} = 1)$, it is natural to formulate a logistic regression model for the binary response U_{i1} by relating the logit transformation of F_{i1} to the covariates,

$$\text{logit}(F_{i1}) = \log\left(\frac{F_{i1}}{1 - F_{i1}}\right) = \alpha_1 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip};$$

although, in principle, any suitable link function (e.g., probit) could be used. Next, we can dichotomize the ordinal outcome at less than or equal to 2 versus greater than 2, creating a second binary response

$$U_{i2} = \begin{cases} 1 & \text{if } Y_i \leq 2, \\ 0 & \text{if } Y_i > 2, \end{cases}$$

with $F_{i2} = \Pr(U_{i2} = 1)$. Because U_{i2} is also binary, we can formulate a logistic regression model relating the logit of F_{i2} to the covariates,

$$\text{logit}(F_{i2}) = \log\left(\frac{F_{i2}}{1 - F_{i2}}\right) = \alpha_2 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}.$$

Note that we have allowed the intercepts for logit(F_{i1}) and logit(F_{i2}) to be different, but have assumed the β's for the covariates are the same; later we discuss the implications of this assumption. If we continue up the ordinal scale, dichotomizing the ordinal outcome above and below the remaining categories, we can generate a series of additional binary variables

$$U_{ik} = \begin{cases} 1 & \text{if } Y_i \leq k, \\ 0 & \text{if } Y_i > k, \end{cases}$$

and formulate a logistic regression model relating the logit of F_{ik} to the covariates (for $k = 1, \ldots, K - 1$),

$$\text{logit}(F_{ik}) = \log\left(\frac{F_{ik}}{1 - F_{ik}}\right) = \alpha_k + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}, \quad (11.1)$$

where $F_{ik} = \Pr(U_{ik} = 1 | X_{i1}, \ldots, X_{ip}) = \Pr(Y_i \leq k | X_{i1}, \ldots, X_{ip})$ is referred to as a "cumulative probability" of response and logit(F_{ik}) is referred to as a "cumulative log odds" (or "cumulative logit"). The model given by (11.1) is commonly called the *proportional odds model*, and it applies simultaneously to all $K - 1$ cumulative probabilities (or cumulative logits).

Thus the basic idea underlying the proportional odds model is a cumulative dichotomization of the ordinal variable going up (or down) the ordinal scale. A logistic regression model is assumed to hold simultaneously for each of these $K - 1$ dichotomous variables, in which the $K - 1$ intercepts (α_k's) are allowed to differ,[2] but the covariate effects (β's) are assumed to be the same. Therefore the proportional odds model can be thought of as a logistic regression model for the *cumulative probabilities* of response; specifically, it relates the cumulative log odds of

response, logit(F_{ik}), to the covariates.

Next we consider the interpretation of the regression parameters in the proportional odds model. For ease of exposition, suppose that we have a proportional odds model with two covariates, X_{i1} and X_{i2},

$$\text{logit}(F_{ik}) = \alpha_k + \beta_1 X_{i1} + \beta_2 X_{i2}.$$

We can interpret β_1 in a manner similar to how we interpret regression parameters in standard logistic regression, but recognizing that we are modeling the cumulative log odds (or cumulative logit). Thus β_1 has interpretation as the change in the cumulative log odds for each one unit increase in X_{i1}, while holding X_{i2} constant. That is, holding X_{i2} constant, β_1 can be thought of as the cumulative log odds ratio,

$$\log\left[\frac{F_{ik}(X_{i1} = c+1)/\{1 - F_{ik}(X_{i1} = c+1)\}}{F_{ik}(X_{i1} = c)/\{1 - F_{ik}(X_{i1} = c)\}}\right].$$

The sign of the coefficient for β_1 sometimes causes confusion about the direction of the relationship between the ordinal response and the covariate. Recall that (11.1) models the cumulative log odds of being in *lower-numbered* categories. Therefore larger values of β_1 X_{i1} are associated with an *increased* probability of being in the *lower-numbered* categories or, equivalently, a *decreased* probability of being in the *higher-numbered* categories. For example, when β_1 is positive this implies an inverse or negative relationship between X_{i1} and Y_i, with increases in X_{i1} associated with lower values of the ordinal scale. We caution the reader that some textbooks, statistical software, and alternative derivations of the model use the following convention for the proportional odds model,

$$\begin{aligned}\text{logit}(F_{ik}) &= \alpha_k - (\beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}) \\ &= \alpha_k - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots - \beta_p X_{ip},\end{aligned}$$

placing a negative sign in front of the β's so that larger values of $\beta_1 X_{i1}$ are associated with an *increased* probability of being in the *higher-numbered* categories.

The alert reader would have noticed that our interpretation of β_1 is not specific about which of the $K - 1$ cumulative log odds it refers to. The reason for this is that β_1 has the same interpretation for all $K - 1$ cumulative log odds. In the proportional odds model, the log odds ratio for a one unit increase in a covariate (while holding the other covariates constant) is the same for any of the cumulative probabilities,

$$\begin{aligned}\beta_1 &= \log\left[\frac{F_{i1}(X_{i1} = c+1)/\{1 - F_{i1}(X_{i1} = c+1)\}}{F_{i1}(X_{i1} = c)/\{1 - F_{i1}(X_{i1} = c)\}}\right] \\ &= \log\left[\frac{F_{i2}(X_{i1} = c+1)/\{1 - F_{i2}(X_{i1} = c+1)\}}{F_{i2}(X_{i1} = c)/\{1 - F_{i2}(X_{i1} = c)\}}\right] \\ &= \log\left[\frac{F_{i3}(X_{i1} = c+1)/\{1 - F_{i3}(X_{i1} = c+1)\}}{F_{i3}(X_{i1} = c)/\{1 - F_{i3}(X_{i1} = c)\}}\right].\end{aligned}$$

What this means is that if a unit increase in a covariate triples the odds of being in response level 1 (versus level 2 or higher), it also triples the odds of being in response level 2 or below (versus level 3 or higher), or in level 3 or below (versus level 4 or higher), and so on. This is the "proportionality assumption" that gives the model its name. This property of the

proportional odds model also implies that if you were to dichotomize an ordinal response (above and below a given level k) and use standard logistic regression as the method of analysis, the resulting odds ratios would be invariant to where you dichotomize the ordinal scale; only the intercept would depend on where you chose to dichotomize the scale.

One appealing property of the proportional odds model is that its regression parameters are invariant to collapsing of adjacent response categories. That is, we would not expect the results of an analysis to change much if we were to combine two adjacent categories. This feature of the model can be helpful when it is of interest to compare regression estimates from studies using different ordinal scales (e.g., one study using a five-level version of a quality-of-life scale, another using a three-level version of the same scale). Also this property of the model can be used to justify combining adjacent categories prior to analysis when data are sparse for certain response categories.

Recall from Section 11.3.1 that the logistic regression model for a binary response can be developed from the notion of an underlying latent variable distribution. The proportional odds model can also be developed from a linear regression model for a latent continuous variable. Suppose that L_i is a latent (i.e., unobserved) continuous variable, such that values of the ordinal response are observed only when L_i falls within one of K intervals determined by a set of "cut-points," α_k. In this latent variable formulation, the observed ordinal response can be thought of as a K – level categorization of the unobserved latent variable, with

$$Y_i = \begin{cases} 1 \text{ if } -\infty < L_i \le \alpha_1, \\ 2 \text{ if } \alpha_1 < L_i \le \alpha_2, \\ 3 \text{ if } \alpha_2 < L_i \le \alpha_3, \\ \vdots \\ K \text{ if } \alpha_{K-1} < L_i < \infty. \end{cases}$$

Next suppose that the following linear regression model holds for L_i.

$$L_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + e_i = X_i'\beta + e_i,$$

where e_i (or $L_i - X'_i\beta$) has a standard logistic distribution with mean zero and variance $\pi^2/3$. Here we regard the "cut-points" (α_k's) of L_i as fixed and the mean of the distribution of L_i as changing with X_i. It can be shown that this linear model for the latent variable (with logistic errors) implies a proportional odds model for the observed ordinal response with the same covariate effects (β). Therefore the latent variable formulation provides at least some motivation for the assumption of common covariate effects across the different cumulative logits in the proportional odds model.

As was mentioned earlier, the proportional odds model makes a strong assumption that the covariate effects (β's) are invariant to where you dichotomize the ordinal scale. This makes interpretation of the effects of covariates relatively straightforward when only a single parameter is required for each covariate. It is possible to relax the proportionality assumption and consider a "non-proportional" odds model, in which the covariate effects depend on the response level k,

(11.2) $$\text{logit}(F_{ik}) = \alpha_k + \beta_{k1} X_{i1} + \beta_{k2} X_{i2} + \cdots + \beta_{kp} X_{ip};$$

this model allows separate covariate effects for each cumulative logit. In the model given by (11.2) the log odds ratio now depends on k,

$$\beta_{k1} = \log\left[\frac{F_{ik}(X_{i1} = c+1)/\{1 - F_{ik}(X_{i1} = c+1)\}}{F_{ik}(X_{i1} = c)/\{1 - F_{ik}(X_{i1} = c)\}}\right].$$

Model (11.2) can be used to test the proportionality assumption based on a test of the null hypothesis,

$$H_0 : \beta_{1j} = \beta_{2j} = \cdots = \beta_{K-1,j} = \beta_j \quad \text{for all } p \text{ covariates } (j = 1, \ldots, p).$$

Under the null hypothesis that the proportional odds model holds, there are p distinct β's for the covariate effects. Under the alternative hypothesis there are $(K - 1) \times p$ distinct β's. So the test of the proportionality assumption has $df = (K - 1) \times p - p = (K - 2) \times p$. Furthermore the proportionality assumption can be relaxed for only a subset of the covariates; this leads to a *partial* proportional odds model where separate effects for each cumulative logit are fit for some but not all of the covariates. For example, the following *partial* proportional odds model,

$$\text{logit}(F_{ik}) = \alpha_k + \beta_{k1}X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_p X_{ip},$$

allows for separate (or "non-proportional") effects of X_{i1} but makes the proportionality assumption for the remaining covariates, X_{i2}, …, X_{ip}. One word of caution about model (11.2) and *partial* proportional odds models: By relaxing the proportionality assumption, the model no longer constrains the cumulative probabilities. As a result the fitting of model (11.2) can potentially lead to incoherent results where, for example, the estimate of $F_{i3} = \Pr(Y_i \leq 3|X_{i1} \ldots, X_{ip})$ is less than the estimate of $F_{i2} = \Pr(Y_i \leq 2|X_{i1}, \ldots, X_{ip})$ for some values of the covariates. This violates the proper order of cumulative probabilities and implies that $\Pr(Y_i = 3|X_{i1}, \ldots, X_{ip}) = (F_{i3} - F_{i2})$ must be negative!

Finally, the regression parameters of the proportional odds model can be estimated by maximum likelihood (ML). This requires maximizing the multinomial likelihood for the ordinal response, with the response probabilities viewed as functions of the α's and β's. When regarded as a generalized linear model, the proportional odds model has certain non-standard features. In the proportional odds model we do not relate the mean of Y_i to the covariates via a logit link function (or any other suitable link function). Instead, we *jointly* relate the means of the $K - 1$ cumulative random variables, $(U_{i1}, \ldots, U_{i,K-1})$, to the covariates. Put another way, it is the cumulative probabilities, and not the mean of the ordinal response, that are simultaneously related to the covariates (via a logit link function). Fitting the proportional odds model requires computer algorithms that can handle the fact that it is a *multivariate*, rather than a *univariate*, generalized linear model for the $K - 1$ cumulative dichotomizations of the ordinal response. Procedures for fitting the model have been implemented in most of the commercially available statistical software packages.

11.4.3 Some Alternative Models for Ordinal Data

Although the proportional odds model is probably the most widely used model for the analysis

of ordinal responses, several alternative regression models are used. Here we briefly consider two models, both based on a logistic regression model for the ordinal response, known as the *adjacent-category* and *continuation-ratio* models. The basic idea underlying the *adjacent-category* logistic regression model is to compare each category of the response to the next largest level. So, with a *K*-level ordinal response we compare level 1 versus level 2, level 2 versus level 3, level 3 versus level 4, and so on. When these comparisons are made on the log odds scale, the following *adjacent-category* model is obtained,

$$\log\left\{\frac{\Pr(Y_i = k | Y_i = k \text{ or } k+1)}{\Pr(Y_i = k+1 | Y_i = k \text{ or } k+1)}\right\} = \alpha_k + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip},$$

where the left-hand side is the log odds of response at level *k*, given that the response is at either level *k* or level *k* + 1. This model assumes the effects of the covariates do not depend on the particular pair of adjacent categories being compared.

Instead of comparing each category of the response to the next largest level, the *continuation-ratio* model compares each category to all higher response levels. So with a *K*-level ordinal response we compare level 1 versus levels 2 through *K*, level 2 versus levels 3 through *K*, level 3 versus levels 4 through *K*, and so on. When these comparisons are made on the log odds scale, the following *continuation-ratio* model is obtained,

$$\begin{aligned}\text{logit}\{\Pr(Y_i = k | Y_i \geq k)\} &= \log\left\{\frac{\Pr(Y_i = k)}{\Pr(Y_i > k)}\right\} \\ &= \alpha_k + \beta_{k1} X_{i1} + \beta_{k2} X_{i2} + \cdots + \beta_{kp} X_{ip}.\end{aligned}$$

This model ordinarily assumes separate effects of the covariates on each of the $K - 1$ logits; when it seems plausible, it is possible to constrain covariate effects to be the same for each of the logits. The continuation-ratio model can be appealing when the categories of the ordinal response represent a natural sequence of stages in some progression (e.g., cancer stage). One unappealing feature of the model is that the results are not invariant to whether the categories have been ordered from low to high or from high to low.

11.4.4 Illustration

We illustrate the application of the proportional odds model to data from a clinical trial comparing auranofin therapy (3 mg of oral gold, twice daily) and placebo for the treatment of rheumatoid arthritis (Bombardier et al., 1986). In this six-month, randomized, double-blind trial, 303 patients with classic or definite rheumatoid arthritis were randomized to one of the two treatment groups and followed over time. The outcome variable of interest is a global impression scale (Arthritis Categorical Scale) at month 6. This is a self-assessment of a patient's current arthritis, measured on a five-level ordinal scale: (1) very good, (2) good, (3) fair, (4) poor, and (5) very poor. Data on this outcome variable are available for 293 of the patients who participated in this trial. The goal of the analysis is to determine whether treatment with auranofin therapy increases the odds of a more favorable response, after controlling for the baseline age of the patients. Consider the following proportional odds model:

$$\text{logit}(F_{ik}) = \log\left\{\frac{\Pr(Y_i \le k)}{\Pr(Y_i > k)}\right\} = \alpha_k + \beta_1 \text{Age}_i + \beta_2 \text{Trt}_i$$

where the covariates are baseline age in units of 10 years (Age) and treatment group (Trt = 1 if randomized to auranofin, Trt = 0 if randomized to placebo). Maximum likelihood estimates of the model parameters are presented in Table 11.6.

Table 11.6 ML estimates and standard errors from the proportional odds model for the arthritis clinical trial data.

Variable	Estimate	SE	Z
α_1	−1.2118	0.5316	−2.28
α_2	0.5251	0.5249	1.00
α_3	2.1494	0.5381	3.99
α_4	4.1364	0.6029	6.86
Age	−0.2048	0.0983	−2.08
Trt	0.6079	0.2142	2.84

The results in Table 11.6 indicate that there is a significant treatment effect ($p < 0.005$), with patients in the auranofin therapy group having an increased odds of a lower or more favorable response. Specifically, when adjusted for baseline age, patients in the auranofin therapy group have approximately twice (or $e^{0.6079} = 1.84$) the odds of a self-assessment of arthritis at response level k or lower (corresponding to a more favorable response) relative to patients in the placebo group. The estimated effect of age indicates that older patients in both treatment groups tend to report less favorable response. For example, a 10-year difference in baseline age decreases the odds of a more favorable response by a factor of 0.82 (or $e^{-0.205}$).

Finally, we can assess the assumption of proportionality by considering a more complex model that allows for separate effects of treatment and age on the four dichotomizations of the ordinal response. The resulting test of proportionality yields a chi-square statistic, $G^2 = 8.55$, with 6 degrees of freedom ($p > 0.20$). This test has 6 df because it allows for 6 additional regression parameters, 3 additional parameters for the treatment effect, and 3 additional parameters for the age effect. Because the more complex model does not fit significantly better ($p > 0.20$), the proportional odds assumption appears to hold for these data.

11.4.5 Regression Models for Nominal Responses

In this section we briefly discuss regression models when the response is nominal or unordered with more than two levels. Following the notation used in previous sections, we let Y_i denote a nominal response with K categories (1, …, K) for the i^{th} subject; the K unordered categories of the nominal variable are arbitrarily labeled with the integers 1, …, K. Associated with each response, Y_i, is a $p \times 1$ vector of covariates, $(X_{i1}, \ldots, X_{ip})'$. To develop a regression model relating Y_i to the covariates, we consider pairing each response category with a baseline or reference category. For convenience, we choose the last category as the reference category.

So with K categories we compare level 1 versus level K, level 2 versus level K, level 3 versus level K, and so on. When these comparisons are made on the log odds scale, the following *baseline-category* logistic model is obtained,

$$\log\left\{\frac{\Pr(Y_i = k)}{\Pr(Y_i = K)}\right\} = \alpha_k + \beta_{k1}X_{i1} + \beta_{k2}X_{i2} + \cdots + \beta_{kp}X_{ip},$$

for $k = 1, \ldots, K - 1$. The *baseline-category* logistic model is often referred to as the *multinomial* or *polytomous* logistic regression model.

This model is a very direct extension of standard logistic regression in the sense that it can be formulated as a series of $K - 1$ logistic regressions for dichotomizations comparing the outcome $Y_i = k$ versus $Y_i = K$ (for $k - 1, \ldots, K - 1$). For example, consider fitting a logistic regression model for the binary outcome $Y_i = 1$ versus $Y_i = K$ (ignoring all observations where $Y_i = 2, \ldots, K - 1$); then consider fitting a separate logistic regression model for the binary outcome $Y_i = 2$ versus $Y_i = K$ (ignoring all observations where $Y_i = 1, 3, 4, \ldots, K - 1$), and so on. Rather than fitting $K - 1$ separate logistic regressions to the data in this manner, the *baseline-category* logistic *model jointly* fits the $K - 1$ logistic regression models.

Interestingly the idea of fitting $K - 1$ separate logistic regressions to the data happens to very closely approximate the baseline-category multinomial logistic model. That is, the estimates obtained from fitting $K - 1$ separate logistic regressions will be similar to the corresponding ML estimates obtained by maximizing the multinomial likelihood for the nominal responses. In general, the latter method is preferred as it can yield more precise estimates of the regression parameters, although in practice, the differences may not be too discernible.

For the remainder of the book we do not focus on the analysis of nominal outcomes. The reasons for this decision are two-fold. First, in our experience, nominal responses are not commonly encountered in applications. With the exception of continuous responses, binary, ordinal, and count data are by far the most commonly encountered data types in longitudinal studies and are the main focus of subsequent chapters. Second, as noted earlier, the analysis of nominal data can always be considered a series of $K - 1$ separate, but otherwise standard, logistic regressions. As a result a comprehensive understanding of extensions of logistic regression models to handle longitudinal binary data provides the basis for a broader understanding of longitudinal methods for analyzing nominal data.

11.5 OVERDISPERSION

The Poisson distribution is generally considered to be the benchmark for count data. However, as noted in Section 11.3.2, in many applications count data exhibit far greater variability than is predicted by the Poisson distribution. This overdispersion has important implications for inference. In particular, neglecting overdispersion in regression models for count data results in standard errors being underestimated; failure to make an adjustment to the nominal standard errors can result in misleading inferences concerning the regression coefficients. In this section we describe a method of accounting for overdispersion by incorporating an extra source of

variability in the model for the counts. Although the focus of this section is on adjustments to regression models for Poisson count data, the same considerations apply to regression models for binomial counts of the number of successes (or to grouped binary data).

Recall that in the standard log-linear regression model for Poisson count data,

$$\log\{E(Y_i|X_i)\} = \log(\mu_i) = \log(T_i) + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip},$$

the variance of Y_i is expressed explicitly in terms of the mean,

$$\text{Var}(Y_i) = v(\mu_i) = \mu_i,$$

with the dispersion parameter held fixed at $\phi = 1$. Because overdispersion is more the rule than the exception with count data, it is advisable to include a scale factor ϕ in the specification of the variance,

$$\text{Var}(Y_i) = \phi\,\mu_i.$$

The scale factor ϕ can be estimated by the standard Pearson chi-squared statistic divided by its residual degrees of freedom. The estimated scale parameter affects the standard errors only; specifically, they are multiplied by a factor of $\sqrt{\hat{\phi}}$. This method of adjustment for overdispersion is very simple and somewhat ad hoc; however, in practice, it tends to work well. By including a scale factor ϕ, the variance is assumed to increase *linearly* with increases in the mean.

An alternative way to handle overdispersion is to consider it as an additional source of random variability. That is, overdispersion can be thought to arise due to unmeasured factors that vary among individuals. This suggests extending the log-linear model to incorporate an extra source of variability in the counts,

$$\log\{E(Y_i|X_i; e_i)\} = \log(T_i) + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + e_i,$$

where the e_i are random errors. Conditional on these random errors (and the covariates), it is assumed that the counts have a Poisson distribution. We consider two choices of distributions for the random errors, e_i. First, assume the errors have a normal distribution, with $e_i \sim N(0, \sigma^2_e)$. Then, it can be shown that the mean of Y_i, when averaged over the distribution of these errors (but conditional on the covariates), is given by,

$$\log\{E(Y_i|X_i)\} = \log(\mu_i) = \log(T_i) + \sigma^2_e/2 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}.$$

That is, the model for the mean is the same as the standard log-linear model except for the addition of the constant term $\sigma^2_e/2$. Thus, in a model where $X_{i1} = 1$ for all subjects, only the intercept changes (becoming $\beta_1 + \sigma^2_e/2$); all other regression parameters are unchanged. However, the inclusion of the normally distributed random errors implies that the variance of the counts (when averaged over the distribution of e_i) is

$$\text{Var}(Y_i) = \mu_i + (e^{\sigma^2_e} - 1)\mu_i^2,$$

which is larger than the mean, μ_i, when $\sigma^2_e > 0$. Therefore the inclusion of this additional source of variation in the log-linear model allows for overdispersion relative to Poisson variation. Of note, the model with normal errors does not have a closed-form likelihood; as a result ML estimation of the model parameters is not entirely straightforward and requires the use of so-called numerical integration techniques that can be computationally demanding.

Numerical integration techniques are discussed in Chapter 14.

A similar model that allows for overdispersion relative to Poisson variation can be obtained if, instead of a normal distribution, a gamma distribution is assumed for the errors. Specifically, if a gamma distribution is assumed for the exponentiated errors, $\exp(e_i)$, with mean of 1 and variance denoted by θ, then it can be shown that the mean of Y_i, when averaged over the distribution of these errors (but conditional on the covariates), is given by

$$\log\{E(Y_i|X_i)\} = \log(\mu_i) = \log(T_i) + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip},$$

and the corresponding variance of the counts is

$$\text{Var}(Y_i) = \mu_i + \theta\mu_i^2.$$

Therefore, assuming a gamma distribution for the (exponentiated) errors produces the same model for the mean as in a standard log-linear model, but allows for overdispersion relative to Poisson variability. The nature of the overdispersion is the same under the assumption that the errors have normal and gamma distributions, because θ corresponds to $e^{\sigma_e^2} - 1$ in the two expressions for $\text{Var}(Y_i)$ given above. In both cases the variance is assumed to increase as a *quadratic* function of the mean, allowing for overdispersion when either $\theta > 0$ or $\sigma^2_e > 0$. One appealing feature of assuming a gamma distribution for the (exponentiated) errors is that the model has a closed-form likelihood. Specifically, if a gamma distribution is assumed, then the distribution of the counts (when averaged over the distribution of e_i) is negative binomial and ML estimation of the model parameters is straightforward. That is, overdispersion can be accounted for by assuming the counts have a negative binomial rather than a Poisson distribution and fitting a log-linear model for the mean via a log link function rather than the canonical link function for the negative binomial (the canonical link for the negative binomial distribution is complicated and somewhat difficult to interpret).

So far we have described two distinct ways to account for overdispersion. The first approach is to include a scale factor in the specification of the variance (allowing for overdispersion when $\phi > 1$); the second approach is to incorporate an additional source of random variability in the model for the counts (allowing for overdispersion when either $\theta > 0$ or $\sigma^2_e > 0$). These two approaches make distinct assumptions about the relationship between the mean and variance. For the former, the variance is assumed to increase *linearly* with the mean, whereas for the latter, the variance is assumed to increase as a *quadratic* function of the mean. This implies that the two approaches weight observations differently when fitting the same regression model to the data at hand. In particular, because observations are weighted inversely proportional to their variance, the smallest and largest counts may be weighted somewhat differently by these two approaches. However, in practice, the regression parameter estimates tend to be relatively insensitive to these differences in weights.

Finally, we note that one potential source of overdispersion with count data is when there is a discernibly large number of zero counts. This excess of zeros will necessarily inflate the variability relative to that predicted by the Poisson model. In general, models that include an additional source of random variation (e.g., negative binomial model) allow for a larger probability of both high and low counts in the data. These models will predict, for example,

more zeros in the data than a standard Poisson model. However, in certain settings, the introduction of an additional source of random variation may only partially explain the excess number of zero counts. In such cases the large number of zero counts needs to be accounted for using regression techniques that explicitly model the production of zero counts. For example, so-called "zero-inflated Poisson" (ZIP) models have been developed to account for the extra zero counts. These models assume that there are two latent (unobserved) groups: one group can be considered the "always-zero group," the other the "sometimes-zero group." For example, in a study of the number of visits per year to a primary care physician, individuals can be thought of as belonging to one of two groups: those who would never visit a primary care physician (the "always-zero group") versus those who would visit a primary care physician whenever they are sufficiently ill (the "sometimes-zero group"). Observed counts of zero could arise from the former group and from a proportion of the latter group (with the proportion determined by the Poisson probability of a zero count). For example, zero counts of the number of visits per year to a primary care physician can arise from those who would never visit a primary care physician (regardless of their illness status) and from those who would visit but were not ill during the year. A second example arises in a study of the number of children women give birth to. In such a study there are two latent groups: one group of women who are fertile, another group who are infertile (or whose partners are infertile). That is, zero counts of the number of children can arise from fertile women who, although at risk of pregnancy and bearing children, do not have any children, and from women who are infertile (or whose partners are infertile). ZIP models combine a model for the probability of belonging to the two latent groups with a separate model for the Poisson counts; the latter model for the counts implicitly includes only zero counts from those belonging to the "sometimes-zero group." A more detailed discussion of ZIP models is beyond the scope of this chapter.

Illustration

In this section we consider the potential impact of overdispersion on the results of a log-linear regression analysis. We return to the data from the prospective study of potential risk factors for coronary heart disease (CHD) considered in Section 11.3.2. For illustrative purposes we focus on the simple model that examines the unadjusted effect of the smoking exposure variable on rates of CHD. In the standard Poisson log-linear regression model,

$$\log(\mu_i) = \log(T_i) + \beta_1 + \beta_2 \,\mathrm{Smoke}_i,$$

it is assumed that $\mathrm{Var}(Y_i) = \mu_i = E(Y_i)$. We consider three alternative approaches for allowing for overdispersion. First, we include a scale factor ϕ in the specification of the standard Poisson variance,

$$\mathrm{Var}(Y_i) = \phi\,\mu_i,$$

to be estimated from the data. Next, we extend the log-linear model to incorporate an extra source of variability in the counts,

$$\log\{E(Y_i|e_i)\} = \log(T_i) + \beta_1 + \beta_2 \,\mathrm{Smoke}_i + e_i,$$

where e_i are random errors. Two distributions for e_i are considered: normal and gamma. The

inclusion of normally distributed random errors, $e_i \sim N(0, \sigma^2_e)$, implies that the variance of the counts is

$$\text{Var}(Y_i) = \mu_i + (e^{\sigma^2_e} - 1)\mu_i^2,$$

whereas the assumption of a gamma distribution for the exponentiated errors, $\exp(e_i)$, with mean of 1 and variance θ, implies that

$$\text{Var}(Y_i) = \mu_i + \theta\mu_i^2.$$

For the model with gamma errors, ML estimation of the model parameters is straightforward. The model can be fit directly to the counts by assuming they have a negative binomial rather than a Poisson distribution and by assuming a log link function rather than the canonical link function for the negative binomial distribution. In contrast, ML estimation of the model with normal errors requires the use of numerical integration (numerical integration techniques will be discussed in Chapter 14). The results of fitting the four models to the data are summarized in Table 11.7.

Table 11.7 Estimated coefficients and standard errors for log-linear regression of expected rate of CHD on smoking from (a) standard Poisson model, (b) standard model with overdispersion factor ϕ, (c) Poisson model conditional on normal errors, and (d) Poisson model conditional on gamma errors.

Model	Variable	Estimate	SE	Z
(a) Poisson (fixed $\phi = 1$)	Intercept	−4.7993	0.0885	−54.22
	Smoke	0.0318	0.0056	5.65
(b) Poisson (unrestricted ϕ)	Intercept	−4.7993	0.2415	−19.87
	Smoke	0.0318	0.0153	2.07
	($\widehat{\phi} = 7.446$)			
(c) Poisson (normal errors)	Intercept	−4.7069	0.2558	−18.40
	Smoke	0.0282	0.0142	1.99
	($\widehat{\sigma}^2_e = 0.3133$)			
(d) Poisson (gamma errors)	Intercept	−4.5265	0.2536	−17.85
	Smoke	0.0263	0.0141	1.86
	($\widehat{\theta} = 0.2942$)			

The ML estimate of the slope for smoking exposure, β_2, from the standard Poisson regression model is 0.0318 (SE = 0.0056) and, when compared to its standard error (Z = 5.65, p < 0.0001), is significantly different from zero at the conventional 0.05 level. The results of the analysis that includes an overdispersion factor yield $\widehat{\phi}$ = 7.45. This implies the variability of the counts is approximately $7\frac{1}{2}$ times larger than that predicted by Poisson variation. The analysis adjusts the standard errors to account for this degree of overdispersion but does not alter the estimates of the regression coefficients. Specifically, the standard errors are simply multiplied by a factor of 2.7 ($\sqrt{7.45}$ = 2.73). When compared to its corrected standard error (SE = 0.0153). the estimate of the slope for smoking exposure remains significantly different from zero at the 0.05 level, although the p-value (Z = 2.07, $p \approx 0.039$) is substantially larger than it

was before adjustment for overdispersion. The results of the analysis that incorporates an extra source of variability, via the inclusion of normally distributed random errors, yield an estimate of the slope for smoking exposure of 0.0282 (SE = 0.0142). Although the estimated slope for smoking exposure is approximately 10% smaller than that obtained from the standard Poisson regression, the standard errors are 2 ½ times larger reflecting the correction made for overdispersion. Finally, the results from the analysis that includes random errors from a gamma distribution (or equivalently, assumes a negative binomial distribution for the counts) yield an estimated slope for smoking exposure of 0.0263 (SE = 0.0141). These results are very similar to those obtained assuming normal errors (albeit the test of slope, $Z = 1.86$, $p \approx 0.063$, is no longer significant at the conventional 0.05 level).

When taken together, the results in Table 11.7 indicate that the expected rate of CHD for individuals who smoke one pack of cigarettes (or 20 cigarettes) per day is estimated to be approximately twice (or $e^{0.03\times 20} = 1.82$ times) as high as the rate of CHD for non-smokers. The results also indicate that there is substantial overdispersion in these data. Failure to account for the overdispersion results in the standard errors being underestimated by a factor of approximately 2.5, leading to confidence intervals that are too narrow and *p*-values that are too small.

Finally, we note that the results presented in Table 11.5 for the joint analysis of the effects of smoking exposure, blood pressure, and personality type are far less affected by overdispersion. The degree of overdispersion seen earlier in Table 11.7 for the analysis of smoking exposure is dramatically attenuated when the effects of blood pressure and personality type are included in the regression model. Specifically, the estimated overdispersion factor, $\hat{\phi}$, decreases from 7.45 to 1.85. This indicates that these two covariates account for much of the excess variability that was evident in the earlier analysis reported in Table 11.7. That is, much of the overdispersion exhibited in the earlier analysis can be thought of as arising from inherent variation among individuals within each of the smoking exposure groups due to differences in blood pressure and personality type; this between-subject heterogeneity within each of the smoking exposure groups is the major cause of the overdispersion in the counts. When blood pressure and personality type are also included in the analysis, $\hat{\phi} = 1.85$. Thus, even when all three factors are controlled, the variability in the counts is almost twice as large as that predicted by Poisson variation. However, when the standard errors in Table 11.5 are corrected by multiplying by a factor of 1.36 (or $\sqrt{1.85} = 1.36$), the overall results are similar. For example, uncorrected for overdispersion, the estimated coefficient for smoking, 0.027 (with 95% confidence interval: 0.016, 0.038), is statistically significant ($Z = 4.50$, $p < 0.0001$); when corrected for overdispersion, it remains significantly different from zero ($Z = 3.58$, $p < 0.0005$) although the 95% confidence interval (0.012, 0.042) is slightly wider. In general, we recommend reporting results that are corrected for overdispersion, regardless of its magnitude, because they provide a more realistic estimate of the sampling variability.

11.6 COMPUTING: FITTING GENERALIZED LINEAR MODELS USING PROC GENMOD IN SAS

To fit generalized linear models we can use the PROC GENMOD procedure in SAS. The GENMOD procedure fits generalized linear models using maximum likelihood estimation. It includes many of the commonly used exponential family distributions for the response variable and a wide variety of link functions for relating the mean response to the covariates. PROC GENMOD can also be used to fit models to correlated responses using the generalized estimating equations approach. This latter aspect of the procedure will be described in Chapter 13.

For example, to fit a logistic regression model to data from two groups (e.g., treatment or exposure groups), we can use the illustrative SAS commands given in Table 11.8. Similarly, to fit a log-linear regression, with an offset, we can use the illustrative SAS commands given in Table 11.9. To fit a log-linear regression model to overdispersed counts, using a negative binomial rather than a Poisson distribution, we can use the illustrative SAS commands given in Table 11.10. To fit a proportional odds regression model to ordinal data, we can use the illustrative SAS commands given in Table 11.11. Finally, we note that the normal distribution and identity link function are the defaults in PROC GENMOD; this corresponds to the standard linear regression model with normal errors.

Table 11.8 Illustrative commands for logistic regression using PROC GENMOD in SAS.

```
PROC GENMOD DESCENDING;
  CLASS group;
  MODEL y=group / DIST=BINOMIAL LINK=LOGIT;
```

Table 11.9 Illustrative commands for log-linear regression, with an offset, using PROC GENMOD in SAS.

```
PROC GENMOD;
  CLASS group;
  MODEL y=group/DIST=POISSON LINK=LOG OFFSET=logtime;
```

Table 11.10 Illustrative commands for log-linear regression assuming negative binomial distribution, using PROC GENMOD in SAS.

```
PROC GENMOD;
  CLASS group;
  MODEL y=group / DIST=NEGBIN LINK=LOG OFFSET=logtime;
```

Table 11.11 Illustrative commands for proportional odds regression assuming multinomial distribution and cumulative logit link function, using PROC GENMOD in SAS.

```
PROC GENMOD;
```

```
CLASS group;
MODEL y=group / DIST=MULTINOMIAL LINK=CUMLOGIT;
```

Next we present a brief description of the most salient parts of the command syntax used in the four illustrations in Tables 11.8 through 11.11.

PROC GENMOD <options>;

This statement calls the procedure GENMOD in SAS. It can include an option for specifying the level of the response variable that is modeled. By default, the lower response level is modeled. For a binary response coded (0,1), it is the probability that $Y = 0$ that is modeled. Use of the DESCENDING option reverses the default ordering of the response levels, resulting in the highest response level being modeled (i.e., the probability that $Y = 1$ for binary data that are coded as 0 and 1).

CLASS variables;

The CLASS statement is used to identify all variables that are to be regarded as categorical or factors. By default, this statement will create indicator variables for each factor using a reference group coding, with the last level (where "last" here refers to the level with the largest alphanumeric value) regarded as the reference group. Different sort orders for the CLASS variables can be requested by the ORDER=<option> on the PROC GENMOD statement.

MODEL response = <effects> / <options>;

MODEL events/trials = <effects> / <options>;

The MODEL statement specifies the response variable and the covariate effects. The second form of the MODEL statement, with the events/`trials` syntax, allows the response to be in the form of a ratio of two variables (e.g., counts of the number of successes and the number of trials) and is used for binomial response data. The covariate effects determine the linear predictor and can include both discrete (defined on the CLASS statement) and quantitative (excluded from the CLASS statement) covariates. By default, PROC GENMOD includes a column of 1's for the intercept in the model.

Two important options need to be included on the MODEL statement. The DIST=*keyword* specifies a built-in response variable distribution, from the exponential family, that is assumed for the model. The LINK=*keyword* specifies the choice of built-in link function relating the mean response to the linear predictor. If the LINK=*keyword* is omitted, the default link function is the canonical link function for the distribution specified on DIST=*keyword*. If both the LINK=<option> and the DIST=<option> are omitted, the default is a normal distribution with an identity link function.

A final option that is often required when modeling count data is an offset. The OFFSET=*variable* specifies a variable to be used as an offset. Note that this variable cannot be a CLASS variable and it should not be included as one of the covariates listed on the MODEL statement.

PROC GENMOD provides many options for handling the dispersion parameter, ϕ, in the exponential family distribution. Recall that for many discrete response distributions (e.g.,

Bernoulli, binomial, and Poisson), the dispersion parameter is a fixed constant ($\phi = 1$) and not a parameter to be estimated. As discussed earlier, in many applications the data display more variability than is predicted by the variance–mean relationship for the assumed distribution of the response. Neglecting overdispersion (e.g., greater variability than that predicted by the binomial or Poisson distributions) results in standard errors being underestimated. To allow for overdispersion, PROC GENMOD provides options for estimating ϕ and making suitable adjustments to standard errors and test statistics. Strictly speaking, in these cases where ϕ is estimated, rather than assumed to be fixed, we no longer have a legitimate distribution for the response variable and the function that is maximized is referred to as a quasi-likelihood function rather than a likelihood function. Alternatively, for overdispersed counts, a negative binomial rather than a Poisson distribution can be assumed (see Table 11.10). Finally, an adjustment to the nominal standard errors to account for overdispersion can be made by basing standard errors on the "sandwich" estimator of Cov($\hat{\beta}$); the "sandwich" estimator will be described in Chapter 13.

11.7 OVERVIEW OF GENERALIZED LINEAR MODELS*

In this section[†] we present a somewhat more technical and detailed overview of generalized linear models that supplements the material presented in Section 11.2. Generalized linear models are a broad class of regression models suitable for analyzing diverse types of univariate responses (e.g., continuous, binary, counts). As was mentioned in Section 11.2, a generalized linear model for Y_i has a three-part specification:

1. a distributional assumption,

2. a systematic component, and

3. a link function,

and we consider each of these three components in turn.

Distributional Assumption

Generalized linear models are an extended family of probability models for a univariate response variable, Y_i. The family of probability distributions, known as the exponential family, includes the normal distribution for a continuous response, the Bernoulli (or binomial) distribution for a binary response, and the Poisson distribution for counts. The exponential family also includes many other distributions, for example, the gamma, beta, and negative binomial distributions.

Any distribution that belongs to the exponential family can be expressed in the same general form. Before we describe that general form we want to emphasize that our motivation for doing so is three-fold. First, we want to demonstrate that probability distributions for seemingly quite

different data types (e.g., continuous, binary, and count data) have much in common as members of the exponential family of distributions. Second, we want to emphasize the importance of the canonical "location" parameter in exponential family distributions; the canonical location parameter is closely related to, but generally not equal to, the mean of the distribution. Third, we want to emphasize that the variance of many exponential family distributions depends on the mean, via a "variance function." We caution the reader that the material in the remainder of this section is somewhat technical in nature, but we strongly encourage the reader to stay the course.

All distributions that belong to the exponential family can be expressed as follows:

$$f(y_i;\theta_i,\phi) = \exp\left[\{y_i\theta_i - a(\theta_i)\}/\phi + b(y_i,\phi)\right], \tag{11.3}$$

for some specific functions $a(\cdot)$ and $b(\cdot)$. The specific functions $a(\cdot)$ and $b(\cdot)$ associated with an exponential family distribution distinguish one member of the family from another. For example, the normal, Bernoulli, and Poisson distributions can all be expressed in the same form, albeit with different functions $a(\cdot)$ and $b(\cdot)$. This expression for the exponential family has two parameters, θ_i and ϕ. The first parameter, θ_i, is a location parameter (and is sometimes referred to as the "canonical" location parameter); the second parameter, ϕ, is a scale or dispersion parameter. As these terms imply, θ_i is related to the mean of the distribution (but θ_i is not necessarily the mean), while ϕ is related to the variance. For many distributions for discrete data, ϕ is not a parameter that requires estimation but is a known constant; for other distributions, ϕ is an unknown parameter. When ϕ is known, Y_i is said to have a one-parameter exponential family distribution, while when ϕ is unknown, it has a two-parameter exponential family distribution.

While many elegant statistical properties can be derived for distributions that belong to the exponential family, the main concept we want to emphasize in this section is that the exponential family provides some unification for distributions that are commonly assumed for seemingly diverse types of responses variables (e.g., probability distributions for continuous and binary responses).

To fix ideas, we will demonstrate how three of the most commonly encountered distributions in biomedical applications, the normal, Bernoulli, and Poisson distributions, can be expressed in the exponential family form given in (11.3). Recall that the probability density function for the normal distribution (see Section 3.2) is usually written as

$$f\left(y_i;\mu_i,\sigma^2\right) = \left(2\pi\sigma^2\right)^{-1/2}\exp\left\{-\left(y_i-\mu_i\right)^2/2\sigma^2\right\}.$$

However, it is possible to re-arrange the terms in this expression for the normal density to obtain

$$\begin{aligned} f\left(y_i;\mu_i,\sigma^2\right) &= \exp\left\{-1/2\log\left(2\pi\sigma^2\right)\right\}\exp\left\{-\left(y_i-\mu_i\right)^2/2\sigma^2\right\} \\ &= \exp\left\{-\left(y_i^2-2y_i\mu_i+\mu_i^2\right)/2\sigma^2 - 1/2\log\left(2\pi\sigma^2\right)\right\} \\ &= \exp\left[\left\{y_i\mu_i-\mu_i^2/2\right\}/\sigma^2 - 1/2\left\{y_i^2/\sigma^2+\log\left(2\pi\sigma^2\right)\right\}\right]. \end{aligned}$$

When expressed in this form, the normal distribution is seen to be an exponential family

distribution with canonical location parameter, $\theta_i = \mu_i$, and scale parameter, $\phi = \sigma^2$ (with $v(\mu_i) = 1$). Also

$$a(\theta_i) = \mu_i^2/2 = \theta_i^2/2,$$

and

$$\begin{aligned} b(y_i, \phi) &= -1/2\left\{y_i^2/\sigma^2 + \log\left(2\pi\sigma^2\right)\right\} \\ &= -1/2\left\{y_i^2/\phi + \log\left(2\pi\phi\right)\right\}. \end{aligned}$$

Thus, for the normal distribution the location parameter, θ_i, happens to be the mean of the response and the scale parameter happens to be the variance.

Two important exponential family distributions for discrete response data are the Bernoulli and the Poisson distributions. The Bernoulli distribution is ordinarily expressed as

$$f(y_i; \mu_i) = \mu_i^{y_i}(1-\mu_i)^{1-y_i},$$

where $\mu_i = E(Y_i) = \Pr(Y_i = 1)$. At first glance it is not obvious that the Bernoulli distribution also belongs to the exponential family. However, the Bernoulli distribution can also be re-expressed as

$$\begin{aligned} f(y_i; \mu_i) &= \mu_i^{y_i}(1-\mu_i)^{(1-y_i)} \\ &= \exp\{y_i \log(\mu_i) + (1-y_i)\log(1-\mu_i)\} \\ &= \exp[y_i \log\{\mu_i/(1-\mu_i)\} + \log(1-\mu_i)]. \end{aligned}$$

When expressed in this form, the Bernoulli distribution is seen to be a one-parameter exponential family distribution with location parameter,

$$\theta_i = \log\{\mu_i/(1-\mu_i)\} = \text{logit}(\mu_i),$$

and $\phi = 1$ is simply a fixed and known constant. Finally, the Poisson distribution is ordinarily expressed as

$$f(y_i; \mu_i) = e^{-\mu_i}\mu_i^{y_i}/y_i!$$

but it too can be re-expressed as

$$f(y_i; \mu_i) = e^{-\mu_i}\mu_i^{y_i}/y_i! = \exp\{y_i \log \mu_i - \mu_i - \log(y_i!)\}.$$

When written in this form, the Poisson distribution is also a one-parameter exponential family distribution with location parameter,

$$\theta_i = \log(\mu_i),$$

and $\phi = 1$, a fixed and known constant.

The exponential family unifies many probability distributions for diverse types of response variables. Moreover it is possible to derive some elegant statistical properties for distributions belonging to this family. The two properties that we focus on here are the mean and variance of exponential family distributions. It can be shown (although it requires the use of calculus) that the mean of Y_i can be expressed as

$$E(Y_i) = \mu_i = \frac{\partial a(\theta_i)}{\partial \theta},$$

where $\frac{\partial a(\theta_i)}{\partial \theta}$ denotes differentiation of the function $a(\theta_i)$ with respect to θ. For readers unfamiliar with calculus, $\frac{\partial a(\theta_i)}{\partial \theta}$ can simply be thought of as another known function of θ_i. Thus μ_i, the mean of Y_i, is simply a known function of θ_i, and vice versa. The second property that we

are interested in is the variance of exponential family distributions. The variance of Y_i can be expressed as

$$\text{Var}(Y_i) = \phi \frac{\partial^2 a(\theta_i)}{\partial \theta^2},$$

where $\frac{\partial^2 a(\theta_i)}{\partial \theta^2}$ (known in calculus as the second derivative of $a(\theta_i)$ with respect to θ) is simply another known function of θ_i. Thus the variance of Y_i for distributions belonging to the exponential family can be expressed as the product of ϕ, the dispersion parameter, and some known function of θ_i. The latter function is referred to as the "variance function." However, recall that θ_i can be expressed as some known function of the mean, μ_i (since earlier we showed that μ_i is a known function of θ_i). Because θ_i and μ_i are functionally related to each other, the variance of Y_i can be expressed as the product of ϕ and some known function of μ_i. When expressed in terms of the mean, the variance function is denoted by $v(\mu_i)$ and

$$\text{Var}(Y_i) = \phi v(\mu_i). \quad (11.4)$$

Thus, for distributions belonging to the exponential family, the variance of Y_i can be expressed in terms of a scale or dispersion parameter ϕ and some known function of the mean, $v(\mu_i)$. For the normal distribution, the variance of Y_i is

$$\text{Var}(Y_i) = \sigma^2 = \phi,$$

and $v(\mu_i) = 1$. For the Bernoulli distribution,

$$\text{Var}(Y_i) = v(\mu_i) = \mu_i(1 - \mu_i),$$

and $\phi = 1$; while for the Poisson distribution,

$$\text{Var}(Y_i) = v(\mu_i) = \mu_i,$$

and $\phi = 1$. For one-parameter exponential family distributions (e.g., Bernoulli and Poisson), the variance of Y_i is simply a known function of the mean, μ_i; that is, the variance is completely determined by the mean response.

In summary, generalized linear models assume that the response, Y_i, has a probability distribution that belongs to the "exponential family." This extended family of distributions includes, among others, the normal, Bernoulli, and Poisson distributions. Some exponential family distributions (e.g., Bernoulli and Poisson) have only a single "location" (or canonical) parameter, and this parameter is related to (but it is not necessarily) the mean of the distribution. For one-parameter exponential family distributions, the variance of Y_i is a known function of the mean, referred to as the variance function. For two-parameter exponential family distributions (e.g., the normal distribution), there is an additional "scale" parameter, often referred to as a dispersion parameter. In two-parameter exponential family distributions the variance can be expressed as the product of the scale parameter and a variance function, where the latter is a known function of the mean.

Systematic Component

The systematic component of the generalized linear model specifies that the effects of the

covariates, X_i, on the mean of the distribution of Y_i can be expressed via the following "linear predictor":

$$\eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}.$$

The linear predictor is simply a linear combination of the unknown vector of regression coefficients, $\beta = (\beta_1, \ldots, \beta_p)'$, and the vector of covariates, X_i,

$$\eta_i = \sum_{k=1}^{p} \beta_k X_{ik}. \quad (11.5)$$

The term "linear," as used in this context, means that η_i must be linear in the regression parameters.

We remind the reader that the restriction that η_i be linear in the regression parameters does not preclude relationships between the mean response and covariates that are curvilinear or non-linear. This latter type of non-linearity can be accommodated by taking appropriate transformations of the covariates (e.g., log(X)) and/or by including a polynomial in X). The inclusion of transformed covariates does not violate in any way the requirement that η_i be linear in the regression parameters.

Link Function

Finally, the formulation of a generalized linear model is completed by specifying the connection between the random and systematic components of the model through a "link function." The link function describes the relation between μ_i, the mean of Y_i, and the linear predictor, η_i, given by (11.5). Specifically, the link function is some known function $g(\cdot)$ such that

$$g(\mu_i) = \eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}. \quad (11.6)$$

In the case of the standard linear regression model, the random and systematic components are directly related, with

$$E(Y_i) = \mu_i = \eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}.$$

When viewed as a generalized linear model, the standard linear regression model adopts an identity link function, $g(\mu_i) = \mu_i$.

The primary motivation for considering link functions other than the identity is to ensure that the linear predictor produces predictions of the mean response that are within the allowable range. For example, when analyzing a binary response, μ_i has interpretation in terms of the probability of "success." As a result we must have $0 < \mu_i < 1$ and the identity link is not appealing since, for sufficiently large or small values of the covariates, it can yield predicted probabilities outside of the range from 0 to 1. It is preferable to use a link function that takes a non-linear transformation of μ_i, mapping the range of μ_i from [0,1] onto the unrestricted range $(-\infty, \infty)$.

In principle, any function $g(\cdot)$ can be chosen to link the mean of Y_i to the linear predictor. However, every distribution that belongs to the exponential family has a special link function

called the *canonical* link function. The canonical link function is defined as that function $g(\cdot)$ such that

$$g(\mu_i) = \theta_i,$$

where θ_i is the canonical location parameter (recall that μ_i is a known function of θ_i, and vice versa). Although there is no a priori reason why the covariate effects should necessarily be additive (or linear) on the particular scale defined by the canonical link function, generalized linear models with canonical link functions produce the most widely used regression models in biomedical applications.

For example, the canonical link function for the normal distribution is the identity link function,

$$g(\mu_i) = \mu_i,$$

and this gives the standard linear regression model,

$$\mu_i = \eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}.$$

For counts from a Poisson distribution, where we must have $\mu_i > 0$, the canonical link function is the log link function,

$$g(\mu_i) = \log(\mu_i),$$

and this gives the log-linear regression model,

$$\log(\mu_i) = \eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}.$$

For the Bernoulli distribution, where $0 < \mu_i < 1$, the canonical link function is the logistic or logit link function,

$$g(\mu_i) = \log\{\mu_i/(1-\mu_i)\}$$

and this gives the logistic regression model

$$\log\{\mu_i/(1-\mu_i)\} = \eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}.$$

We can, however, choose other link functions when they seem appropriate to the application at hand. For example, when Y_i is Bernoulli, we would generally prefer a link function that transforms the interval [0, 1] on to the entire real line, $(-\infty, \infty)$. The complementary log-log link function,

$$g(\mu_i) = \log\{-\log(1-\mu_i)\},$$

and the probit link function,

$$g(\mu_i) = \Phi^{-1}(\mu_i),$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function, both have this property. In some applications it may be of interest to consider a link function that does not transform the range of μ_i on to the entire real line $(-\infty, \infty)$. For example, in modeling the prevalence of a disease, and the impact of risk factors, it may be preferable on scientific grounds to use a log link function since the resulting regression coefficients, β, have interpretation in terms of the log relative risk of disease. The relative risk of disease is simply the ratio of the probability of disease when a risk factor is present to the probability of disease when a risk factor is absent. The relative risk is an index of association that is favored by many empirical researchers. Although, in principle, link functions that do not transform the range of μ_i on to $(-\infty, \infty)$ can be adopted, in practice, this can result in problems with predictions that are out of range (e.g.,

predicted probabilities less than zero or greater than 1) and problems with convergence of the model-fitting algorithm.

In summary, the link function connects the random and systematic components of the generalized linear model. It relates the mean of Y_i to the linear predictor and determines the scale on which the additive effects of covariates have an impact on the mean response. Each distribution has a special link function called the canonical link function and adoption of the canonical link function gives rise to many of the widely used regression models in biomedical applications (e.g., linear regression for a normally distributed continuous response, logistic regression for a Bernoulli response, and log-linear regression for Poisson counts). In principle, other link functions can be selected and these link functions bear no relationship to the assumed distribution for the response. Instead, a non-canonical link function may be chosen because additivity of the covariate effects is more appropriate on that scale or because it yields regression coefficients, β, that have somewhat more useful interpretations.

Estimation

Next we very briefly discuss estimation of the regression coefficients in a generalized linear model. This section is somewhat more technical and can be omitted on a first reading of this chapter. Recall that in the standard linear regression setting we estimate the linear regression coefficients using the method of least squares. The least squares criterion chooses values for the regression coefficients that minimize the sum of squared deviations of the observed Y_i from their predicted values, denoted by $\widehat{Y}_i = \widehat{\mu}_i$, under the assumed regression model,

$$\mu_i = \eta_i = \sum_{k=1}^{P} \beta_k X_{ik}.$$

The least squares method yields estimates of the regression coefficients that are also the *maximum likelihood* (ML) estimates when Y_i is assumed to have a normal distribution with constant variance. We use the more general method of maximum likelihood estimation for estimating the parameters of a generalized linear model.

Recall from Chapter 4 (Section 4.2) that the method of maximum likelihood estimation chooses values of the regression coefficients that are most likely (or most probable) to have generated the observed data. This is achieved by maximizing the *likelihood function* for the data. Construction of the likelihood function requires an assumption about the probability distribution of Y_i. In generalized linear models the response is assumed to have a distribution belonging to the exponential family of distributions. Assuming independent observations of the response, Y_i, and the covariates X_{i1}, X_{i2}, …, X_{ip} available on N individuals, the joint probability of (Y_1, …, Y_N) is the product of the N probability (density) functions. Thus the likelihood function can be expressed as the product

$$L = \prod_{i=1}^{N} \exp\left[\{y_i\theta_i - a(\theta_i)\}/\phi + b(y_i, \phi)\right].$$

It is this function, or equivalently, the logarithm of this likelihood function, that must be

maximized. Note that the likelihood is a function of the unknown regression coefficients, β, since θ_i is a known function of the mean, μ_i, and

$$\mu_i = g^{-1}\left(\sum_{k=1}^{p} \beta_k X_{ik}\right),$$

where $g^{-1}(\cdot)$ denotes the inverse link function; for instance, if $g(\cdot) = \log(\cdot)$, then $g^{-1}(\cdot) = \exp(\cdot)$. The maximum likelihood estimates of β are obtained by substituting the expression above for μ_i into the likelihood function and finding those values of the regression coefficients that produce the largest value for the likelihood function. Ordinarily the likelihood function has only a single maximum.

Instead of maximizing the likelihood, it is usually more convenient to maximize the log-likelihood. We maximize the log-likelihood with respect to β by taking the derivative of the log-likelihood with respect to β, and then finding the values of β that make those derivatives equal to 0. Given

$$l = \log L = \sum_{i=1}^{N} \left[\{y_i\theta_i - a(\theta_i)\}/\phi + b(y_i, \phi)\right],$$

the derivative of the log-likelihood with respect to β can be shown (with the aid of calculus) to be the vector,

$$\partial l/\partial \beta = \sum_{i=1}^{N} (\partial\theta_i/\partial\boldsymbol{\beta})(y_i - \mu_i)/\phi.$$

When a canonical link function, $g(\mu_i) = \theta_i = \eta_i$, has been assumed,

$$\partial l/\partial\beta = \sum_{i=1}^{N} X_i (y_i - \mu_i)/\phi.$$

Solving this set of equations,

$$\sum_{i=1}^{N} X_i (y_i - \mu_i) = 0,$$

yields the ML estimates of β. In general, this requires an iterative procedure that has been implemented in many statistical software packages (e.g., PROC GENMOD in SAS, the `glm` function in R and S-Plus, and the `glm` command in Stata). What is quite remarkable about ML estimation for generalized linear models (with canonical link functions) is that it requires the solution to the exact same set of equations, regardless of the type of response variable.

Finally, estimates of the standard errors of the estimated regression coefficients can readily be obtained using the method of maximum likelihood estimation; in addition, likelihood ratio tests can be constructed by comparing nested models. Interestingly the solution to this set of equations, $\widehat{\beta}$, is consistent for β (i.e., with very high probability, $\widehat{\beta}$ is close to the population regression parameters β for sufficiently large N) even if the variance of Y_i is misspecified; the only requirement is that the model for the mean response (the link function and linear predictor) has been correctly specified. However, when the variance of Y_i is misspecified, standard errors for components of $\widehat{\beta}$ should be based on the "sandwich" estimator of Cov($\widehat{\beta}$); the "sandwich" estimator is discussed in Chapter 13.

11.8 FURTHER READING

A general overview of logistic regression, Poisson regression, and generalized linear models can be found in Chapter 14 of Neter et al. (1996). The textbooks by Dobson (1990) and Gill (2000) provide excellent introductions to generalized linear models. Hosmer and Lemeshow (2000) provide an accessible and comprehensive description of logistic regression models for binary data. Agresti (2010) provides a comprehensive description of regression models for ordinal data.

Bibliographic Notes

Generalized linear models were introduced in a seminal paper by Nelder and Wedderburn (1972). McCullagh and Nelder (1989) is the definitive textbook on this topic, providing a comprehensive description of the theory and application of generalized linear models. Firth (1991) presents a concise but remarkably lucid review of generalized linear models; also see Chapter 2 of Fahrmeir and Tutz (2001) and Chapter 5 of McCulloch and Searle (2001). For a comprehensive survey on recent developments in statistical methods for the analysis of ordinal data, see the review article by Liu and Agresti (2005). Cameron and Trivedi (1998) present an overview of regression models for count data, including negative binomial regression models for overdispersed count data; also see Hilbe (2007) for a comprehensive discussion of negative binomial regression models.

Problems

11.1 In an experimental study of patients with bladder cancer conducted by the Veterans Administration Cooperative Urological Research Group (Byar and Blackard, 1978; Wei et al., 1989), patients underwent surgery to remove tumors. Following surgery, patients were randomized to either placebo or treatment with thiotepa. Subsequently patients were examined at 18, 24, 30, and 36 months. For this problem set we focus only on the data for month 18. The response variable is binary, indicating whether or not there is a new tumor (Y = 1, if new tumor; Y = 0, if no new tumor) at the 18-month visit. The objective of the analysis is to determine the effect of treatment on tumor recurrence by month 18.

The raw data are stored in an external file: `tumor.dat`

Each row of the data set contains the following three variables:

ID Treatment Y

Note: The response variable Y is coded 1 = new tumor, 0 = no new tumor. The categorical variable Treatment is coded 1 = thiotepa, 0 = placebo.

11.1.1 Assuming a Bernoulli distribution for the recurrence of tumor at month 18, fit the following logistic regression model relating the mean or probability of recurrence (μ_i) to Treatment:

$$\text{logit}(\mu_i) = \beta_1 + \beta_2 \, \text{Treatment}_i.$$

11.1.2 What are the interpretations of β_1 and β_2?

11.1.3 From the results obtained in Problem 11.1.1, what can you conclude about the effect of treatment on tumor recurrence at month 18?

11.1.4 What is the *estimated* probability of recurrence of a new tumor among those who received placebo?

11.1.5 What is the *estimated* probability of recurrence of a new tumor among those who received thiotepa?

11.1.6 Construct a 95% confidence interval for the log odds ratio, comparing thiotepa to placebo.

11.1.7 Construct a 95% confidence interval for the odds ratio, comparing thiotepa to placebo.

11.2 In a clinical trial of patients suffering from epileptic seizures (Thall and Vail, 1990), patients were randomized to receive either a placebo or the drug progabide, in addition to standard therapy. A baseline count of the number of epileptic seizures in an 8-week period prior to randomization was obtained. In addition, counts of the number of epileptic seizures in each of four successive 2-week (post-baseline) treatment periods were obtained. For this problem set, we focus only on the data from the last 2-week treatment period. The goal of the analysis is to make a comparison between the two treatment groups in terms of the counts of the number of seizures in the final 2-week period of the study. The question we want to address is whether treatment with progabide is effective in reducing epileptic seizures.

The raw data are stored in an external file: `seizure.dat`

Each row of the data set contains the following four variables:

ID Treatment Age Y

Note: The response variable *Y* is a count of the number of epileptic seizures in a 2-week interval. The categorical variable Treatment is coded 1 = progabide, 0 = placebo. The variable Age is the age of each patient (in years) at baseline.

11.2.1 Assuming a Poisson distribution for the counts, fit the following model relating the mean number of seizures (μ_i) to Treatment:

$$\ln(\mu_i) = \beta_1 + \beta_2 \,\text{Treatment}_i.$$

11.2.2 What are the interpretations of β_1 and β_2?

11.2.3 From the results obtained in Problem 11.2.1, what can you conclude about the effect of progabide in reducing the number of epileptic seizures.

11.2.4 Construct a 95% confidence interval for the rate ratio, comparing progabide to placebo.

11.2.5 Redo the analysis in Problem 11.2.1, adjusting for the effect of baseline age of the patient:

$$\ln(\mu_i) = \beta_1 + \beta_2 \,\text{Treatment}_i + \beta_3 \,\text{Age}_i.$$

11.2.6 Based on the results of the analysis for Problem 11.2.5, construct a 95% confidence interval for the age-adjusted rate ratio, comparing progabide to placebo.

11.2.7 Redo the analysis in Problem 11.2.5, allowing for potential overdispersion (i.e., variability greater than that predicted by the Poisson distribution).

11.2.8 Construct a 95% confidence interval for the age-adjusted rate ratio, comparing progabide to placebo, after taking account of any potential overdispersion.

11.3 In a study of mental health conducted on a random sample of 40 adult residents of Alachua County, Florida, mental impairment was measured on a four-level ordinal scale with four categories (well, mild symptom formation, moderate symptom formation, impaired); these data are from Chapter 3 (Table 3.3) of Agresti (2010). The goal of the study was to relate mental impairment to several covariates, including an index of life events. The life events (LE) index is a composite measure of the number and severity of important life events that occurred within the past three years (e.g., birth of a child, new job, divorce, death of a family member). The main objective of the analyses is to assess whether the odds of a more favorable mental impairment response is related to the index of life events. Because socioeconomic status (SES) is considered to be a potential confounding variable, it is also of interest to assess the relationship between mental impairment and life events adjusted for SES.

The raw data are stored in an external file: `impairment.dat`

Each row of the data set contains the following four variables:

ID LE SES Y

Note: The ordinal response variable *Y*, denoting subjects' reported mental impairment, has four categories coded 1=well, 2=mild symptom formation, 3=moderate symptom formation, and 4=impaired. The categorical variable SES is coded 1 = high SES, 0 = low SES. The variable LE is a quantitative measures of the number and severity of important life events.

11.3.1 Assuming a multinomial distribution for the ordinal response, fit the following proportional odds model relating mental impairment to life events (LE):

$$\log\left\{\frac{\Pr(Y_i \le k)}{\Pr(Y_i > k)}\right\} = \alpha_k + \beta_1 \mathrm{LE}_i.$$

11.3.2 What is the interpretation of the estimate of β_1?

11.3.3 Construct a test of the null hypothesis of no effect of life events on the cumulative log odds of response. What conclusions do you draw about the effect of life events on mental impairment?

11.3.4 Based on the results from Problem 11.3.1, estimate the odds ratio of a more favorable response for subjects with no life events (LE=0) relative to subjects with 6 life events (LE=6).

11.3.5 The proportional odds model in Problem 11.3.1 makes the assumption of a common effect of life events (β_1) across the different cumulative logits. Provide a formal or informal assessment of the "proportionality assumption." What do you conclude?

11.3.6 Redo the analysis in Problem 11.3.1, adjusting for the effect of SES:

$$\log\left\{\frac{\Pr(Y_i \le k)}{\Pr(Y_i > k)}\right\} = \alpha_k + \beta_1 \mathrm{LE}_i + \beta_2 \mathrm{SES}_i.$$

11.3.7 Based on the results from Problem 11.3.6, what are the interpretations of the

estimates of β_1 and β_2?

11.3.8 Construct a test of the null hypothesis of no effect of life events on the cumulative log odds of response, after adjusting for SES. What conclusions do you draw about the adjusted effect of life events on mental impairment?

11.3.9 Combine the two adjacent categories, mild and moderate symptom formation, to form a three category ordinal response. With the three category ordinal response, redo the analysis in Problem 11.3.6:

$$\log\left\{\frac{\Pr(Y_i \leq k)}{\Pr(Y_i > k)}\right\} = \alpha_k + \beta_1 \mathrm{LE}_i + \beta_2 \mathrm{SES}_i.$$

11.3.10 Compare and contrast the estimate of β_1 obtained from Problem 11.3.9 with the corresponding estimate of β_1 obtained from Problem 11.3.6. Does β_1 have the same interpretation in the model from Problem 11.3.9 as it does in the model from Problem 11.3.6?

[1] Type A personalities are characterized by impatience, competitiveness, aggressiveness, a sense of time urgency, and tenseness; Type B personalities are the opposite of Type A and exhibit traits such as being easy going, more relaxed about time, not competitive, and not easily angered or agitated.

[2] Note that because this representation of the proportional odds model includes separate parameters (α_k) for the $K - 1$ intercepts, we no longer assume that $X_{i1} = 1$ for all i; this is a slight departure from the notation used in earlier sections and chapters of the book.

[†] This section provides a more technical presentation of generalized linear models and can be omitted without loss of continuity.