Even though the number of $\{\beta_t\}$ does not increase as n does, the ordinary ML estimators $\{\hat{\beta}_t\}$ are not consistent. This happens in many models when the number of parameters has an order similar to that of the number of subjects. Asymptotic optimality properties of ML estimators, such as consistency, require the number of parameters to be fixed as n increases. For model (12.4), ML estimators of $\{\beta_t\}$ have bias of order $T/(T-1)$ (Andersen 1980, pp. 244–245). For the matched-pairs model (12.2), for instance, $\hat{\beta} \rightarrow 2\beta$ in probability (Problem 10.24).

For this reason, the preferable approach for the fixed effects model is *conditional ML*. One eliminates $\{u_i\}$ by conditioning on their sufficient statistics $\{S_i = \Sigma_t y_{it},\ i = 1, \dots, n\}$. In the item response context, these are the numbers of correct responses for each subject. Conditional on $\{S_i\}$, the distribution of $\{y_{it}\}$ is independent of $\{u_i\}$. Maximizing the resulting likelihood then yields consistent estimators of $\{\beta_t\}$. The analysis generalizes the one in Section 10.2.3 for the subject-specific logistic model (10.8) for matched pairs. See Andersen (1980) for details.

Compared with the random effects approach, the conditional ML approach has certain advantages. One does not need to assume a parametric distribution for $\{u_i\}$. It is difficult to check this assumption in the random effects approach. Conditional ML is also appropriate with retrospective sampling. In that case, bias can occur with a random effects approach because the clusters are not randomly sampled (Neuhaus and Jewell 1990b).

However, the conditional ML approach has severe disadvantages. It is restricted to the canonical link (the logit), for which reduced sufficient statistics exist for $\{u_i\}$. More important, as discussed in Section 10.2.7, it is restricted to inference about within-cluster fixed effects. The conditioning removes the source of variability needed for estimating between-cluster effects in models with explanatory variables such as those considered next. Also, this approach does not provide information about $\{u_i\}$, such as predictions of their values and estimates of their variability or of the probabilities they determine. Finally, in more general models with covariates, conditional ML can be less efficient than the random effects approach for estimating the fixed effects (see Note 12.2).

12.2 BINARY RESPONSES: LOGISTIC-NORMAL MODEL

The item response model (12.4) with random intercept is a special case of an important class of random effects models for binary data called *logistic-normal models*. With univariate random effect, the model form is

$$\text{logit}\left[P(Y_{it} = 1 \mid u_i)\right] = \mathbf{x}_{it}'\boldsymbol{\beta} + u_i \qquad (12.5)$$

where $\{u_i\}$ are independent $N(0, \sigma^2)$ variates. This is the special case of the GLMM (12.1) in which $g(\cdot)$ is the logit link and the random effects structure simplifies to a random intercept. The logistic-normal model has a long history, dating at least to Cox (1970, Prob. 20 in that text) for the matched-pairs model (12.3) and Pierce and Sands (1975).

More generally, the link function in model (12.5) can be an arbitrary inverse cdf. For such models, Y_{is} and Y_{it} are treated conditionally (given u_i) as independent but are marginally nonnegatively correlated. Let Φ denote the cdf that is the inverse link function. Then, for $s \neq t$,

$$\begin{aligned}\text{cov}(Y_{is}, Y_{it}) &= E\left[\text{cov}(Y_{is}, Y_{it} \mid u_i)\right] + \text{cov}\left[E(Y_{is} \mid u_i), E(Y_{it} \mid u_i)\right] \\ &= 0 + \text{cov}\left[\Phi(\mathbf{x}_{is}'\boldsymbol{\beta} + u_i), \Phi(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i)\right]. \qquad (12.6)\end{aligned}$$

The functions in the last covariance term are both monotone increasing in u_i, and hence are nonnegatively correlated. For common predictor value $\mathbf{x}$ at each t, the joint distribution for the model is exchangeable. This is often plausible for clustered data. In longitudinal studies, however, observations closer together in time may tend to be more highly correlated.

Usually, the main focus in using a GLMM is inference about the fixed effects. The random effects part of the model is a mechanism for representing how the positive correlation occurs between observations within a cluster. Parameters pertaining to the random effects may themselves be of interest, however. For instance, the estimate $\hat{\sigma}$ of the standard deviation of a random intercept may be a useful summary of the degree of heterogeneity of a population.

12.2.1 Interpreting Heterogeneity in Logistic-Normal Models

When $\sigma = 0$, the logistic-normal model (12.5) simplifies to the ordinary logistic regression model treating all observations as independent. When $\sigma > 0$, how can we interpret the variability in effects this model implies?

Consider observation y_{it} at setting $\mathbf{x}_{it}$ of predictors and observation y_{hs} at setting $\mathbf{x}_{hs}$. Their log odds ratio is

$$\text{logit}\left[P(Y_{it} = 1 \mid u_i)\right] - \text{logit}\left[P(Y_{hs} = 1 \mid u_h)\right] = (\mathbf{x}_{it} - \mathbf{x}_{hs})'\boldsymbol{\beta} + (u_i - u_h).$$

We cannot observe $(u_i - u_h)$, which has a $N(0, 2\sigma^2)$distribution. However, $100(1-\alpha)\%$ of those log odds ratios fall within

$$(\mathbf{x}_{it} - \mathbf{x}_{hs})'\boldsymbol{\beta} \pm z_{\alpha/2}\sqrt{2}\,\sigma. \qquad (12.7)$$

When $\sigma = 0$, $(\mathbf{x}_{it} - \mathbf{x}_{hs})'\boldsymbol{\beta}$ is the usual form of log odds ratio for a model without random effects. When $\sigma > 0$, $(\mathbf{x}_{it} - \mathbf{x}_{hs})'\boldsymbol{\beta}$ is the log odds ratio for two observations in the same cluster ($h = i$) or with the same random effect value. Suppose that $\mathbf{x}_{it} = \mathbf{x}_{hs}$ for observations from different clusters. Then, using $z_{0.25} = 0.674$, the middle 50% of the log odds ratios fall within

$\pm 0.674\sqrt{2}\,\sigma = \pm 0.95\sigma$. Hence, the median odds ratio between the observation with higher random effect and the observation with lower random effect equals $\exp(0.95\sigma)$. With a single predictor and $x_{it} - x_{hs} = 1$, the median such odds ratio equals $\exp(\beta + 0.95\sigma)$. Larsen et al. (2000) presented related interpretations.

12.2.2 Connections between Conditional Models and Marginal Models

The fixed effects parameters $\boldsymbol{\beta}$ in GLMMs have conditional intepretations, given the random effect. Those fixed effects are of two types. First, consider an explanatory variable that varies in value among observations in a cluster. For instance, in a crossover study comparing T drugs, for each subject the drug taken varies from observation to observation in that subject's cluster of T observations. For such an explanatory variable, its coefficient in the model refers to the effect on the response of a within-cluster (e.g., subject-specific) 1-unit increase of that predictor. The random effect as well as other explanatory variables in the model are constant while that predictor increases by 1. The effect of that explanatory variable is a "within-cluster" or "within-subject" one.

Second, consider an explanatory variable with constant value among observations in a cluster. An example is gender when each subject forms a cluster. For such an explanatory variable, its coefficient refers to the effect on the response of a "between-cluster" 1-unit increase of that predictor. An example is a comparison of females and males using a dummy variable and its coefficient. However, this fixed effect in the GLMM applies only when the random effect (as well as other explanatory variables in the model) takes the same value in both groups: for instance, a male and a female with the same value for their random effects.

It is in this sense that random effects models are conditional models, as both within- and between-cluster effects apply conditional on the random effect value. By contrast, effects in marginal models are averaged over all clusters (i.e., population averaged), so those effects do not refer to a comparison at a fixed value of a random effect. In fact, a fundamental difference between the two model types is that when the link function is nonlinear, such as the logit, the population-averaged effects of marginal models often are smaller than the cluster-specific effects of GLMMs.

Specifically, the GLMM (12.1) refers to the conditional mean, $\mu_{it} = E(Y_{it} \mid \mathbf{u}_i)$. By inverting the link function,

$$E(Y_{it} \mid \mathbf{u}_i) = g^{-1}(\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_{it}'\mathbf{u}_i).$$

Marginally, averaging over the random effects, the mean is

$$E(Y_{it}) = E\big[E(Y_{it} \mid \mathbf{u}_i)\big] = \int g^{-1}(\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_{it}'\mathbf{u}_i) f(\mathbf{u}_i; \boldsymbol{\Sigma})\, d\mathbf{u}_i,$$

where $f(\mathbf{u}; \boldsymbol{\Sigma})$ is the $N(\mathbf{0}, \boldsymbol{\Sigma})$ density function for the random effects. For the identity link,

$$E(Y_{it}) = \int (\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_{it}'\mathbf{u}_i) f(\mathbf{u}_i; \boldsymbol{\Sigma})\, d\mathbf{u}_i = \mathbf{x}_{it}'\boldsymbol{\beta}\,.$$

The marginal model has the same model form and effects $\boldsymbol{\beta}$. This is not true for other links. For instance, for the logistic-normal model (12.5),

$$E(Y_{it}) = E\left[\frac{\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i)}{1 + \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i)}\right].$$

This expectation does not have form $\exp(\mathbf{x}_{it}'\boldsymbol{\beta})/[1 + \exp(\mathbf{x}_{it}'\boldsymbol{\beta})]$ except when u_i has a degenerate distribution ($\sigma = 0$).

Approximate relationships exist between estimates from the two model types. In the logistic-normal case with effect $\boldsymbol{\beta}$ and small σ, Zeger et al. (1988) showed that

$$E(Y_{it}) \approx \exp(c\mathbf{x}_{it}'\boldsymbol{\beta})/[1 + \exp(c\mathbf{x}_{it}'\boldsymbol{\beta})], \qquad (12.8)$$

where $c = [1 + 0.6\sigma^2]^{-1/2}$. Since the effect in the marginal model multiplies that of the conditional model by about c, it is typically smaller in absolute value. The discrepancy increases as σ increases. For $\boldsymbol{\beta}$ near 0, Neuhaus et al. (1991) showed that the marginal model effect is approximately $\boldsymbol{\beta}(1 - \rho)$, where $\rho = \text{corr}(Y_{it}, Y_{is})$ at $\boldsymbol{\beta} = \mathbf{0}$. Again, the discrepancy increases as σ increases, since ρ increases with σ.

For Table 12.1 on ratings of the prime minister, the ML estimate for model (12.3) is $\hat{\beta} = -0.556$, with $\hat{\sigma} = 5.16$ for variability of $\{u_i\}$. Approximation (12.8) suggests that $\hat{\beta} = -0.556$ with $\hat{\sigma} = 5.16$ corresponds to a marginal estimate of about $[1 + 0.6(5.16)^2]^{-1/2}(-0.556) = -0.135$. The actual marginal estimate is the log odds ratio for the sample marginal distributions, equaling

$$\log[(880/720)/(944/656)] = -0.163.$$

In fact, the marginal effect is much smaller than the conditional effect, but this approximation connecting the two estimates works better for smaller $\hat{\sigma}$. At $\beta = 0$, the fit of the model is that of the symmetry model, for which $\hat{\mu}_{12} = \hat{\mu}_{21} = (n_{12} + n_{21})/2$. The correlation for that 2 × 2 table equals 0.699, from which the conditional estimate of -0.556 suggests a marginal estimate of $-0.556(1 - 0.699) = -0.167$, very close to the actual value of -0.163.

Figure 12.1 illustrates why the marginal effect is smaller than the conditional effect. For a single explanatory variable x, the figure shows subject-specific curves for $P(Y_{it} = 1|u_i)$ for several subjects when considerable heterogeneity exists. This corresponds to a relatively large σ for random effects.

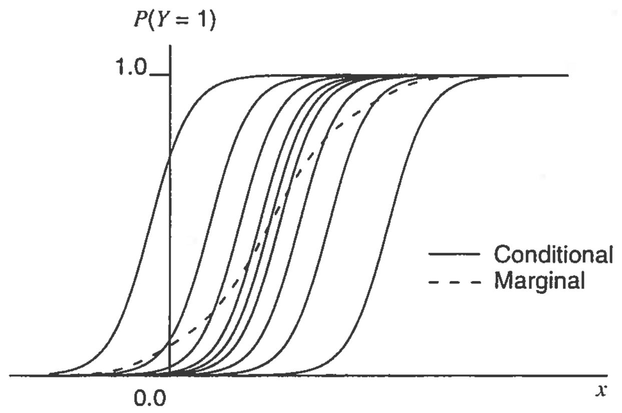

FIGURE 12.1 Logistic random-intercept model, showing the conditional (subject-specific) curves and the marginal (population-averaged) curve averaging over these.

At any fixed value of x, variability occurs in the conditional means, $E(Y_{it} | u_i) = P(Y_{it} = 1 | u_i)$. The average of these is the marginal mean, $E(Y_{it})$. These averages for various x values yield the superimposed curve. It has a shallower slope. In fact, it does not exactly follow the logistic formula. Similar remarks apply to other GLMMs. For the probit link with binary data, however, the conditional probit model with normal random effect does imply a marginal model of probit form (Problem 12.29). With univariate random intercept, the marginal effect equals the conditional effect multiplied by $[1 + \sigma^2]^{-1/2}$ (Zeger et al. 1988). In Section 13.5.1 we explore the conditional–marginal connection for loglinear GLMMs.

12.2.3 Comments about Conditional versus Marginal Models

Random effects models describe conditional (subject-specific) effects, whereas marginal models describe population-averaged effects. Some statisticians prefer one of these types, but most feel that both are useful, depending on the application.

The conditional modeling approach is preferable if one wants to specify a mechanism that could generate positive association among clustered observations, estimate cluster-specific effects, estimate their variability, or model the joint distribution. Latent variable constructions used to motivate model forms (e.g., the tolerance motivation for binary models of Section 6.6.1 and the related threshold motivation in Problem 6.28 and utility motivation in Problem 6.29) usually apply more naturally at the cluster level than at the marginal level. Given a conditional model, one can recover information about marginal distributions. That is, a conditional model implies a marginal model, but a marginal model does not itself imply a conditional model (although see Note 12.10 for an implicit connection).

In many surveys or epidemiological studies, a goal is to compare the relative frequency of occurrence of some outcome for different groups in a population. Then, quantities of primary interest include between-group odds ratios among marginal probabilities for the different groups. That is, effects of interest are between-cluster rather than within-cluster. When marginal effects are the main focus, it is usually simpler and may be preferable to model the margins directly. One can then parameterize the model so that regression parameters have a direct marginal interpretation. Developing a more detailed model of the joint distribution that generates those margins, as a random effects model does, provides greater opportunity for misspecification. For instance, with longitudinal data the assumption that observations are independent, given the random effect, need not be realistic. With the marginal model approach, we showed in Chapter 11 that ML is sometimes possible but that the GEE approach is computationally simpler and more versatile. A drawback of the GEE approach is that it does not explicitly model random effects and therefore does not allow these effects to be estimated. In addition, likelihood-based inferences are not possible because the joint distribution of the responses is not specified.

In Section 12.2.2 it was noted that conditional effects are usually larger than marginal effects, and increase as variance components increase. Usually, though, the significance of an effect (e.g., as measured by the ratio of estimate to standard error) is similar in the two model types. If one effect seems more important than another in a conditional model, the same is usually true with a marginal model. So the choice of the model is usually not crucial to inferential conclusions.

This statement requires a caveat, however, since sizes of effects in marginal models depend on the degree of heterogeneity in conditional models. In comparing effects for two groups or two variables that have quite different variance components, relative sizes of effects will differ for marginal and conditional models. From (12.8), with binary data the attenuation from the conditional to the marginal effect will tend to be greater for the group having the larger variance component. For instance, suppose that two groups, one young in age and the other elderly, both show the same conditional effect in a crossover study comparing two drugs. If the elderly group has more heterogeneity on the response, their marginal effect may be smaller than that for the younger group. The marginal effects differ even though the conditional effects are the same, because of the greater variance component for the elderly. In such cases, the conditional effect (appropriately modeled) may have more relevance.

Finally, with either marginal or conditional models, missing data are a common problem with multivariate responses. Unless data are missing at random, potential bias occurs in ML inference. GEE methods usually require the stronger condition that data are missing completely at random (Section 11.4.5). Thus, modeling missingness or conducting a sensitivity study to discern its potential effects can be an important component of an analysis.

Regardless of the choice of paradigm, it is a challenge for statisticians even to explain to practitioners why marginal and conditional effects differ with a nonlinear link function. Graphics such as Figure 12.1 can help. Neuhaus (1992) and Pendergast et al. (1996) surveyed ways of analyzing clustered binary data, including conditional and marginal models. Agresti and Natarajan (2001) surveyed conditional and marginal modeling of clustered ordinal data.

12.3 EXAMPLES OF RANDOM EFFECTS MODELS FOR BINARY DATA

In the next three sections we present a variety of examples of random effects models. In this section we consider binary responses.

12.3.1 Small-Area Estimation of Binomial Proportions

Small-area estimation refers to estimation of parameters for a large number of geographical areas when each has relatively few observations. For instance, one might want county-specific estimates of characteristics such as the unemployment rate or the proportion of families having health insurance coverage. With a national or statewide survey, some counties may have few observations. Then, sample proportions in the counties may poorly estimate the true countywide proportions. Random effects models that treat each county as a cluster can provide improved estimates. In assuming that the true proportions vary according to some distribution, the fitting process "borrows from the whole"—it uses data from all the counties to estimate the proportion in any given one.

Let π_i denote the true proportion in area i, $i = 1, \dots, n$. These areas may be all the ones of interest, or only a sample. Let $\{y_i\}$ denote independent $\text{bin}(T_i, \pi_i)$ variates; that is, $y_i = \sum_{t=1}^{T_i} y_{it}$, where $\{y_{it}, t = 1, \dots, T_i\}$ are independent with $P(Y_{it} = 1) = \pi_i$ and $P(Y_{it} = 0) = 1 - \pi_i$. The sample proportions $\{p_i = y_i/T_i\}$ are ML estimates of $\{\pi_i\}$ for the fixed-effects model

$$\text{logit}(\pi_i) = \alpha + \beta_i, \qquad i = 1, \dots, n.$$

This model is saturated, having n nonredundant parameters (with a constraint such as $\sum_i \beta_i = 0$) for the n binomial observations.

For small $\{T_i\}$, $\{p_i\}$ have large standard errors. Thus, $\{p_i\}$ may display much more variability than $\{\pi_i\}$, especially when $\{\pi_i\}$ are similar. Then, it is helpful to shrink $\{p_i\}$ toward their overall mean. One can accomplish this with the random effects model

$$\text{logit}\left[P(Y_{it} = 1 \mid u_i)\right] = \alpha + u_i, \tag{12.9}$$

where $\{u_i\}$ are independent $N(0, \sigma^2)$ variates. This model is a logit analog of one-way random effects ANOVA. When $\sigma = 0$, all π_i are identical.

For this model,

$$\hat{\pi}_i = \exp(\hat{\alpha} + \hat{u}_i)/\left[1 + \exp(\hat{\alpha} + \hat{u}_i)\right].$$

This estimate differs from the sample proportion p_i. If $\hat{\sigma} = 0$, then all $\hat{u}_i = 0$. Then, the random effects estimate of each π_i is $(\sum_{i=1}^{n} \sum_{t=1}^{T_i} y_{it})/(\sum_i T_i)$, the overall sample proportion after pooling all n samples. When truly all π_i are equal, this is a much better estimator of that common value than the sample proportion from a single sample.

Generally, the random effects model estimators shrink the separate sample proportions toward the overall sample proportion. The amount of shrinkage decreases as $\hat{\sigma}$ increases. The shrinkage also decreases as the $\{T_i\}$ grow; as each sample has more data, we put more trust in the separate sample proportions. The predicted random effect $\hat{u}_i$ is the estimated mean of the distribution of u_i, given the data (see Section 12.6.7). This prediction depends on all the data, not just data from area i. A benefit is potential reduction in the mean-squared error of the estimates around the true values.

We illustrate model (12.9) with a simulated sample of size 2000 to mimic a poll taken before the 1996 U.S. presidential election. For T_i observations in state i ($i = 1, \dots, 51$, where $i = 51$ is DC = District of Columbia), y_i is $\text{bin}(T_i, \pi_i)$, where π_i is the actual proportion of votes in state i for Bill Clinton in the 1996 election, conditional on voting for Clinton or the Republican candidate, Bob Dole. Here, T_i is proportional to the state's population size, subject to $\sum_i T_i = 2000$. Table 12.2 shows $\{T_i\}$, $\{\pi_i\}$, and $\{p_i = y_i/T_i\}$.

For the ML fit of model (12.9), $\hat{\alpha} = 0.163$ and $\hat{\sigma} = 0.29$. The predicted random effect values (obtained using PROC NLMIXED in SAS) yield the proportion estimates $\{\hat{\pi}_i\}$, also shown in Table 12.2. Since $\{T_i\}$ are mostly small and since $\hat{\sigma}$ is relatively small, considerable shrinkage of these estimates occurs from the sample proportions toward the overall proportion supporting Clinton, which was 0.548. The $\{\hat{\pi}_i\}$ vary only between 0.468 (for TX = Texas) and 0.696 (for NY = New York), whereas the sample proportions vary between 0.111 (for Idaho) and 1.0 (for DC). Sample proportions based on fewer observations, such as DC, tended to shrink more. Although the estimates incorporating random effects are relatively homogeneous, they tend to be closer than the sample proportions to the true values.