Annu. Rev. Pub. Health 1993. 14:43–68

REGRESSION ANALYSIS FOR CORRELATED DATA

Kung-Yee Liang and Scott L. Zeger

Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205

KEY WORDS: estimating equations, odds ratio, marginal models, random effects models, transition models

INTRODUCTION

Regression analysis is among the most commonly used methods of statistical analysis in public health research. Its objective is to describe the relationship of a response with explanatory variables. One example of a regression problem is to identify factors associated with the racial difference in the risk of low birthweight (29). Regression includes the following as special cases: linear models for measured responses, logistic models for binary responses, and survival analyses for times to events. A basic assumption of regression analysis is that all observations are statistically independent, or at least uncorrelated with each other. In the low birthweight example, this assumption would mean that knowing one child's birthweight status provides no information as to whether another child in the study has a low birthweight. One may argue that the assumption of independence is unlikely to be true if children of the same mother are included in the sample. Due to their common household environment and genes, we would expect a child to have a greater chance of having a low birthweight if his/her sibling had. Data from this hypothetical example can usefully be thought of as being "clustered" into families. Birthweights from different families are likely independent; those from the same cluster are not. This dependence among observations from the same cluster must be accounted for in assessing the relationship between risk factors and health outcomes.

0163-7525/93/0510-0043$02.00

Clustered data are common in public health research. To illustrate the diversity of problems in which clustered data arise, we now briefly summarize seven research problems with dependent responses. The first example also illustrates the statistical methodology appropriate for clustered data.

1. Baltimore eye survey study. More than 5000 persons aged 40 years and older received a visual examination as part of a population-based prevalence study of ocular disorders (31). The objective is to identify demographic variables, such as age, race, education level, and access to medical care, which are associated with vision loss. Data are available on both eyes for all subjects. A single regression model expressing visual impairment in terms of demographic variables addresses the scientific objectives. But, the two eyes from the same person are unlikely to be independent, because many causes of impairment are binocular. This association must be considered.

2. Family study in liver cancer. Information regarding liver cancer on 371 siblings and children of 138 cases of primary hepatic carcinoma in Qi-dong county, Shanghai, was collected as a part of a survey conducted by Dr. F. M. Shen of the Shanghai Medical University (15a). The objective is to determine whether there is a familial aggregation of liver cancer as an important first step for studying the possible genetic explanation of the disease process. To reduce the influence of environmental factors, regression analysis must adjust for household and personal characteristics. The focus, however, is to estimate the degree of association between the liver cancer status for members of the same family.

3. Longitudinal study on numbers of sexual partners. The Multicenter AIDS Cohort Study (12) followed more than 4500 gay/bisexual men in Baltimore, Pittsburgh, Chicago, and Los Angeles since 1984 to study the natural history of AIDS. Investigators are interested in how sexual practice has changed since 1984 and which factors predict continued sexual behaviors that put men at high risk for HIV infection. The primary response variable used in this analysis is the number of sexual partners in the six months before each visit. There are up to 15 visits for each individual.

4. Sister chromatid exchange (SCE) study. A total of 14 hepatocellular carcinoma, 14 nasopharyngeal carcinoma, and 16 cervical cancer patients, and their age-sex-matched controls, were studied to compare the frequency of sister chromatid exchange (SCE) in their peripheral lymphocytes. Sister chromatid changes occur during cell replication when a chromosome duplicates its genetic material, thus forming a pair of chromosomes attached at the centromere (11). Elevated levels of SCEs would indicate that cells have been exposed to a mutagen potentially caused by chemical carcinogens. The hypothesis, which was tested in the study conducted by Dr. C. J. Chen, School of Medicine, National Taiwan University, is that cancer patients may have a higher frequency of SCE in lymphocytes than matched controls. The

outcome variable is the number of SCE per cell where 20 cells from each patient were cultured.

5. Family study on chronic obstructive pulmonary disease (COPD). Literature on obstructive pulmonary disease has suggested that there is a significantly increased risk for this common respiratory disease in relatives of patients with impaired pulmonary function (4, 14). Several studies have also shown that the observed familial aggregation in pulmonary function cannot be explained just by nongenetic risk factors, such as age and smoking (3). As part of a multidisciplinary study of COPD, 613 family members of 158 COPD cases seen at the Johns Hopkins Hospital were examined and given spirometry tests. The objective is to determine the percent of total variance attributable to unobserved genetic factors shared among siblings and their parents.

6. Growth study in Hmong refugee children. More than 1000 Hmong refugee children receiving health care at two Minnesota clinics between 1976 and 1985 were examined for their growth patterns (26). The objective is to study the pattern of growth and its association with the age at entry into the United States. Scientists believe that increases in stature are influenced by both genetic and environmental factors. When the offending environmental factors have been removed, the growth process often progresses at an accelerate rate, known as catch-up growth (36). This study allows investigators to address the questions as to whether there is indeed a period of accelerated growth following remediation, and if so, when this acceleration terminates. To study the growth, repeated observations of heights for each child were collected. The number of visits per child ranged from 1 to 15 and averaged 5. A regression model relating the rate of growth over time to the age at entering the United States will address the main objective. The correlation between repeated observations on height for each child is a nuisance, but cannot be ignored in regression analysis.

7. Indonesian children's health study. Approximately 3000 preschool children in Indonesia were medically examined in early 1980 as part of a survey of children's diseases. An objective of the study was to assess the role of vitamin A deficiency on the risk of respiratory infection and on body mass. A separate regression is needed for each of the two response variables, respiratory infection and weight for height. A secondary question is whether the response variables are correlated with each other, and if so, in what way?

1.1 Characteristics of Examples

Although different in their specific scientific objectives, the above examples have important characteristics in common that allow us to use a unified statistical method, which we present in Sections 3 and 4. Data in each of these examples are organized in clusters. For example, in the longitudinal studies (examples 3 and 6), a cluster comprises the repeated observations for an

individual. For family studies (examples 2 and 5), the clusters are formed by families. Table 1 summarizes the clusters for each example.

For a variety of reasons, the responses within a cluster are likely to be dependent, i.e. correlated with one another. In the Hmong study, a child who is smaller than expected at one visit is likely to remain below average at the next visit. This phenomenon is known as "tracking" and is commonly observed in longitudinal studies (5). In the COPD study, the response variable, forced expiratory volume (FEV), is likely to be correlated among siblings because of shared genes and/or shared environments. This notion of familial aggregation has been repeatedly observed in chronic diseases, such as breast cancer (24), and in psychiatric disorders, such as schizophrenia (21).

The second common feature is that the scientific questions for each of the studies above can be formulated as regression analyses. First, one seeks the regression of the mean response on the independent variables. For the eye survey example, we relate the risk of visual impairment to demographic variables, such as age and race; logistic regression is suitable. In the Hmong growth example, we study children's height as a function of age and date of entry to the US by using linear regression. In addition, the regression concept can be applied to the parameters that characterize the within-cluster dependence. For example, researchers in the Baltimore eye survey are interested in knowing whether the degree of correlation of visual impairment between eyes varies by age or race. A more detailed discussion on the choice of measures for within-cluster dependence is given in Section 2.

The studies also differ from another in important ways, such as their type of response variable. We have examples of continuous measurements, such as body weight and FEV, as well as discrete variables that can be either counts, such as number of sexual partners in six months, or dichotomous, such as the presence or the absence of visual impairment for an eye. The studies also

Table 1 Some features of seven examples introduction in Section 1

Example	Cluster	No. of clusters/ cluster size*	Response	Scientific focus Mean	Dependence
Eye survey	individual	5199/2	binary	primary	secondary
Liver cancer	family	138/3*	binary	nuisance	primary
SCE	individual	80/20	count	primary	nuisance
Sexual partners	individual	4500/8*	count	primary	nuisance
COPD	family	158/4*	continuous	nuisance	primary
Hmong growth	individual	1070/5*	continuous	primary	nuisance
Indonesian children	individual	275/2	continuous/ binary	primary	secondary

*Averaged cluster sizes if varied.

differ in the structure of the within-cluster dependence. For example, in the Baltimore eye survey, each cluster (individual) has two responses, visual impairment status for the left and right eyes. If the study included several members of the same family, a more complicated dependence structure would result, because family members would likely have correlated responses. Family studies can present different types of correlation. If first-degree relatives of cases are ascertained, different degrees of correlation would be expected between parents, parents and siblings, or siblings.

Finally, the examples differ from one another with respect to their focus. In some cases, e.g. examples 3 and 6, the regression of the response on explanatory variables is most important, and the within-cluster association is a nuisance. For the family studies (examples 2 and 5), regression for the dependence structure is of primary interest. Even so, regression adjustment for individual and shared household characteristics is crucial in order to separate out the environmental impacts from the genetic ones. Table 1 summarizes both the similarities and the differences among the examples.

1.2 Inadequacy of the Conventional Regression Approach

This section discusses the consequences on regression inferences of totally ignoring the dependence within clusters when it exists. The specific impact of ignoring dependence varies according to the type of response model, the degree of correlation, and other factors. Nevertheless, the patterns are common across a range of problems, which we can illustrate with a simple example.

Let Y_{ij} be the j^{th} observations from the i^{th} cluster, $j = 1, \ldots, n$, $i = 1, \ldots, K$. We assume

$$E(Y_{ij}) = \beta_0 + \beta_1 x_{ij}, \ \text{Var}(Y_{ij}) = \sigma^2,$$

$$\text{Cov}(Y_{ij}, Y_{ik}) = \sigma^2\rho, \ j < k = 1, \ldots, n.$$

This model assumes Y is a simple linear function of x and that the correlation between every pair of responses from a cluster is the same value, ρ. Let $\hat{\beta}_1$ be the ordinary least squares estimate of β_1 in which ρ is incorrectly assumed to be zero. Let $\tilde{\beta}_1$ be the best possible (weighted least squares) estimate obtained by properly accounting for the correlation. Although $\hat{\beta}_1$ remains unbiased, i.e. $E(\hat{\beta}_1) = \beta_1$, ordinary least squares faces two problems: The estimated variance for $\hat{\beta}_1$ is incorrect, and $\hat{\beta}_1$ may be more variable than $\tilde{\beta}_1$. Each consequence is considered in turn.

1.2.1 VARIANCE ESTIMATE Ignoring the correlation leads to the use of $V_1 = \sigma^2/\Sigma_i\Sigma_j(x_{ij} - \bar{x})^2 = \sigma^2/V_T$ as the variance of $\hat{\beta}_1$. The correct variance, V_2, of $\hat{\beta}_1$ has the form $V_2 = V_1[1 + \rho(n\phi - 1)]$ where

$$\phi = \sum_{i=1}^{K} n(\bar{x}_i - \bar{x})^2/V_T$$

is the fraction of the total variance among the xs that is caused by variation among cluster mean ($\bar{x}_i$)s, rather than variation in xs within clusters. Two important cases of ϕ deserve special attention. When $\phi = 0$, there is no between-cluster variation in x; that is, $\bar{x}_i$ is the same for all clusters. An example would be a longitudinal study in which every person is measured at the same set of times. In this case, β_1 is estimated by using only within-cluster changes in Y. On the other hand, $\phi = 1$ features the between-cluster comparison, because in this case $x_{i1} = \ldots = x_{in}$ for all i. This is typical for cluster-specific covariates, such as the race variable in the Baltimore eye survey example.

Figure 1 shows the plots of

$$f(\rho,n,\phi) = \log(V_1$$

against ρ for some selected ns and ϕs. The vertical axis is the logarithim of the ratio of incorrect versus correct variance of the ordinary least squares estimate; the horizontal axis is the actual correlation of responses from a cluster. A positive value of f indicates that the naive variance V_1 is too large and, hence, the confidence interval for β_1 based on $\hat{\beta}_1$ and V_1 is too wide. A negative value of f corresponds to confidence intervals for β_1 that are too narrow. The message from the plots is clear. For within cluster-comparisons, i.e. $\phi = 0$, the confidence interval based on V_1 is wider than it should be; the discrepancy between the incorrect and correct variances increases with ρ. On the other hand, the naive confidence interval is too narrow for between-cluster comparisons for which $\phi = 1$. In either case, invalid scientific conclusions may be drawn if V_1 is used as the variance estimate.

1.2.2 EFFICIENCY LOSS The second impact of ignoring the correlation and using $\hat{\beta}_1$ is a loss of efficiency, by which we mean that the uncertainty in $\hat{\beta}_1$ is greater than the uncertainty in the best unbiased estimate, $\tilde{\beta}_1$. The best estimate, $\tilde{\beta}_1$, has variance of the form

$$V_3 = V_1(1 - \rho)[1 + (n - 1)\rho]/[1 - \rho(1 - \phi)].$$

Figure 2 presents plots of

$$g(\rho,n,\phi) = V_3/V_2 = \left\{1 + \frac{n^2\rho^2\phi\ (1 - \phi)}{(1 - \rho)[1 + (n - 1)\rho]}\right\}^{-1}$$

against ρ for selected ns and ϕs. Interestingly, $\hat{\beta}_1$ is fully efficient in this example when either $\phi = 0$ or $\phi = 1$, irrespective of ρ and n. However,

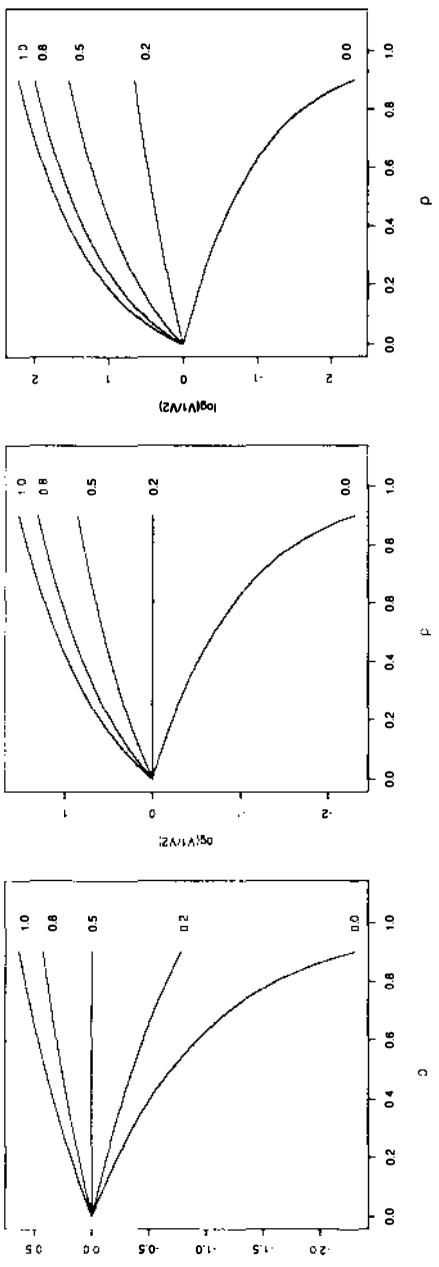

Figure 1 The plot of the logarithm of V_1/V_2 versus ρ for selected cluster sizes $n(2, 5, 10)$ and $\phi(0, 0.2, 0.5, 0.8, 1.0)$. Here, V_1 is the variance of the least squared estimate $\hat{\beta}_1$ when the within-cluster dependence is ignored and V_2 is the correct variance. We have assumed $E(Y_{ij}) = \beta_0 + \beta_1 x_{ij}$ and $\rho = \text{corr}(Y_{ij}, Y_{ik}), j < k = 1, \ldots n$; ϕ is the ratio of the between cluster variance to the total variance among the x_{ij}s.

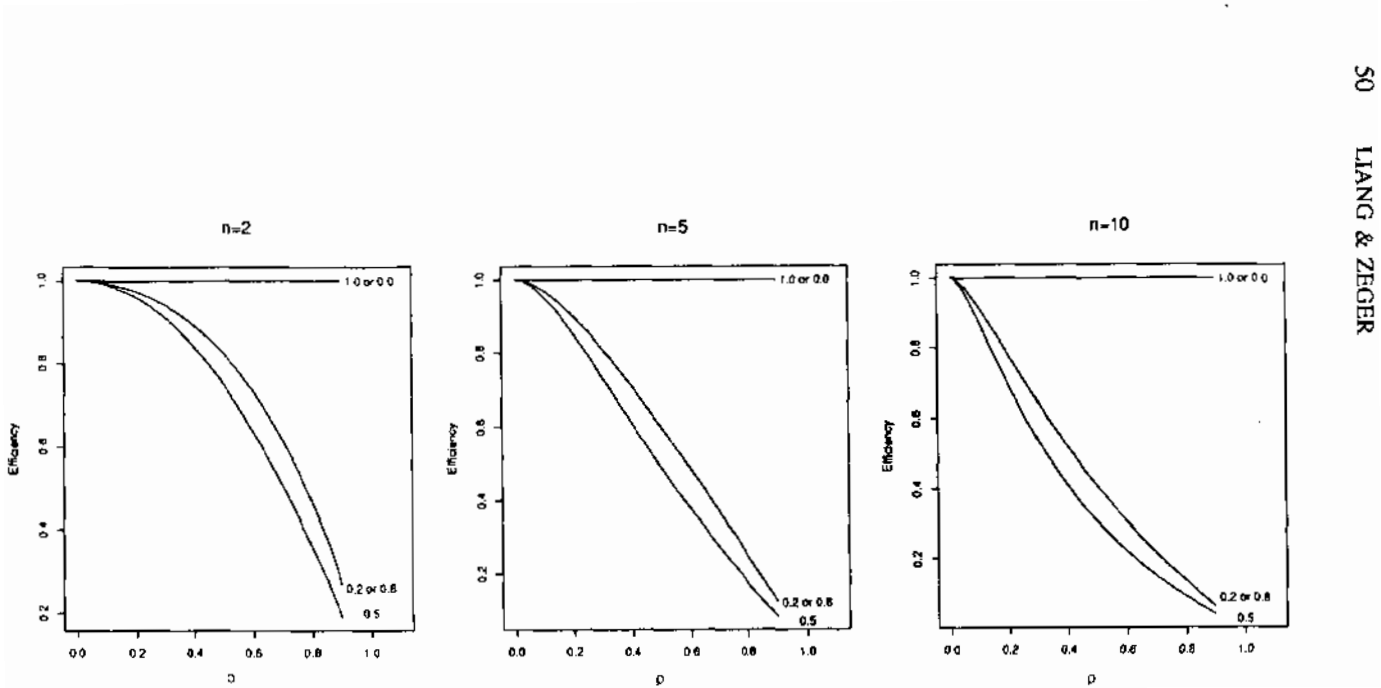

Figure 2 The plot of V_3/V_2 versus ρ for selected n (2, 5, 10) and ϕ (0, 0.2, 0.5, 0.8, 1.0). Here, V_2, ρ, ϕ, and the assumed model are the same as described in Figure 1; V_3 is the variance of the best unbiased estimate of β_1.

efficiency loss is evident when ϕ moves toward 0.5. This phenomenon is more apparent with increased ρ or n. The most important message in Figure 2 is that ignoring correlation can lead to a loss of power when both within-cluster and among-cluster information is being used to estimate β_1.

In Section 2, we provide statistical background for generalized linear models (GLMs) that have unified regression analysis for various types of independent data, and for measures of within-cluster dependence. Section 3 reviews three different approaches to modeling both the regression of Y on x and the within-cluster dependence. Section 4 describes briefly the statistical method called generalized estimating equations (GEE), which was designed to provide valid regression inferences with correlated data. We then illustrate GEE through an analysis of data from the Baltimore eye survey.

2 BACKGROUND

In describing the common features of the examples in Section 1.1, we identified two objectives in the analysis of clustered data: describing the response variable as a function of explanatory variables and measuring within-in-cluster dependence. In this section, we review briefly how commonly used regression models have been unified under the framework of GLMs when all the data are independent ("univariate case"). We also review two measures of within-cluster dependence: correlation for continuous data and odds ratio (OR) for dichotomous responses.

2.1 Generalized Linear Models

Generalized linear models (20, 22) are a unified class of regression methods for discrete and continuous response variables. Logistic regression for binary responses, linear regression for continuous responses, log-linear models for counts, and some survival analysis methods are special cases. There are two basic parts of a GLM, the systematic component and the random component. For the systematic component, one relates Y to x by assuming the averaged Y among individuals with a common value of x, $\mu = E(Y)$, satisfies

$$g(\mu) = x_1\beta_1 + \ldots + x_p\beta_p,$$

which we write in short-hand notation

$$g(\mu) = x'\beta \qquad 1.$$

where g is a prespecified function known as the "link function." The logistic regression model, frequently used for binary data, is a special case of Equation 1 with

$$\log\left(\frac{E(Y)}{1 - E(Y)}\right) = \log\frac{P(Y = 1)}{P(Y = 0)} = x'\beta.$$

For count data, the familiar log-linear model is a special case that assumes

$$\log E(Y) = x'\beta.$$

Table 2 provides a list of commonly used regression models in public health that are special cases of GLMs.

To account for the variability of the observed responses that is not explained by the systematic component caused by either measurement error or variation between individuals, GLMs assume that Y is generated from a distribution with likelihood function of the form

$$f(y) = \exp\{[y\theta + b(\theta)]/\phi + c(y, \phi)\}, \qquad 2.$$

known as the exponential family. The Poisson distribution is a special case of Equation 2 with

$$\begin{aligned} P(Y = y) &= \mu^y e^{-\mu}/y! \ , y = 0,1,2, \ldots \\ &= \exp\{y \log \mu - \mu - \log y!\}, \end{aligned}$$

so that $\theta = \log \mu$, $b(\theta) = -e^{\theta}$, $c(y) = -\log y!$, and $\phi = 1$. Other commonly used distributions are included in Table 2. The scale parameter ϕ in Equation 2 is called the "overdispersion" parameter. Many biomedical researchers have observed that the variation among count data is beyond that described by the Poisson distribution, which assumes that $\mathrm{Var}(Y) = \mu = E(Y)$. The introduction of ϕ in Equation 2 deals with overdispersion directly by allowing $\mathrm{Var}(Y) = \phi E(Y)$, $\phi > 1$ instead.

When both the systematic component, i.e. the link function, and the random component, i.e. Equation 2, are specified in a GLM, one can estimate the regression coefficients β by solving the estimating equation

$$U(\beta) = \sum_{i=1}^{K} \frac{\partial \mu_i(\beta)'}{\partial \beta} \mathrm{Var}(Y_i)^{-1}[Y_i - \mu_i(\beta)] = 0. \qquad 3.$$

As expected, $U(\beta)$ reduces to $\Sigma_{i=1}^{K} x_i(Y_i - x_i'\beta) = 0$ in the multiple linear regression case, i.e. when $Y_i \sim N(x_i'\beta, \phi^2)$. The form of the estimating equation is intuitively sensible. To estimate β, one equates the observed (Y_i) with the expected (μ_i) for each individual, and these $(O_i - E_i)$s are then combined across individuals with weights that are inversely proportional to the variability of the Y_is. We multiply in front by $\partial \mu_i/\partial \beta$ to change from units of Y to units of x. Because the same equation is solved for the entire class of GLMs, common software, theory, and model checking techniques can be used for logistic, log-linear, linear, and survival models. Finally, we note that the estimating equation above only uses the first two moments of the Ys, i.e. the mean and variance functions (34). This feature of $U(\beta)$ is especially important when one is uncertain about the full distribution for the data.

Table 2 Some commonly used regression models in biomedical applications

Model	Response(Y)	Link function g	Distribution for Y	Scalar parameter (ϕ)	Variance function $\phi V(\mu)$
Linear multiple regression	continuous	Identity	Normal	ϕ	ϕ
		$\mu = x'\beta$	$\frac{1}{\sqrt{2\pi\phi}} e^{-(y-\mu)^2/2\phi}$		
Logistic regression	proportion	logit	Binomial		
	$0, 1/m, 2/m, \ldots, 1$	$\log \frac{\mu}{1-\mu} = x'\beta$	$\binom{m}{my} \mu^{my}(1-\mu)^{m-my}$	$1/m$	$\mu(1-\mu)/m$
Poisson regression (Log-linear)	count	log	Poisson		
	$0, 1, 2, \ldots$	$\log \mu = x'\beta$	$\mu^y e^{-\mu}/y!$	1	μ
Gamma regression	continuous	inverse	Gamma		
	(non-negative)	$\mu^{-1} = x'\beta$	$\frac{1}{\Gamma(1/\phi)(\mu\phi)^{1/\phi}} y^{1/\phi-1} e^{-y/(\mu\phi)}$	ϕ	$\phi\mu^2$

2.2 *Measures of Within-Cluster Dependence*

For continuous data, the most commonly used measure of dependence between two responses, Y_1 and Y_2, is the correlation coefficient

$$\rho = \text{Cov}(Y_1,Y_2)/[\text{Var}(Y_1)\text{Var}(Y_2)]^{1/2}$$

where $\text{Cov}(Y_1,Y_2)$ is the covariance. The correlation coefficient is dimensionless, taking values in the range $[-1, 1]$. The correlation ρ is close to 0 when there is little dependence. Strong dependence is indicated when ρ approaches either 1 or -1. A positive correlation indicates that Y_1 tends to be larger than expected if Y_2 is and vice versa. The correlation among observations for the same cluster can take a variety of forms; each can be summarized by the pattern of correlations, as we now illustrate.

In longitudinal studies, each cluster comprises repeated responses over time from an individual. For biological variables, such as blood pressures, cholesterol level, and body weight, the degree of correlation tends to be greater for observations that are closer in time than those that are far apart. One simple assumption is that the correlation between observations at times t_1 and t_2 has the form $\rho^{|t_1-t_2|}$, i.e. decays geometrically with $t_1 - t_2$. An alternate assumption is that correlation is the same for all pairs of observations for the same cluster. In a longitudinal study, this form arises if individuals tend to have their own levels (intercepts) for the response. In family studies, it is important to distinguish within-family correlation that is caused by the shared environment from that caused by shared genes. For families with data from both parents and offspring, one may assume three different correlation coefficients: between parents (ρ_{PP}), between siblings (ρ_{SS}), and between a parent and offspring (ρ_{PS}). Assuming all the relatives share the same environment, a genetic explanation for the trait may be warranted if ρ_{SS} is much greater than ρ_{PP}, as parents do not share genes in common as do siblings. For more details on correlation models, see Laird & Ware (15), Diggle (7), and Diggle et al (8).

For discrete, in particular dichotomous data, the correlation coefficient is a poor measure of association because it is constrained by the mean parameters μ_1, and μ_2. Specifically, for dichotomous variables, Y_1 and Y_2, the correlation is given by

$$\rho = \frac{\Pr(Y_1 = Y_1 = 1) - 2\mu_1\mu_2}{[\mu_1(1-\mu_1)\mu_2(1-\mu_2)]^{1/2}}.$$

However, the joint probability, $\Pr(Y_1 = Y_2 = 1)$ is constrained to satisfy

$$\max(0, \mu_1 + \mu_2 - 1) < \Pr(Y_1 = Y_2 = 1) < \min(\mu_1, \mu_2),$$

which narrows considerably the range ρ can take. For this reason, we prefer to use the OR

$$\mathrm{OR}(Y_1, Y_2) = \frac{\Pr(Y_1 = Y_2 = 1)\Pr(Y_1 = Y_2 = 0)}{\Pr(Y_1 = 1, Y_2 = 0)\Pr(Y_1 = 0, Y_2 = 1)},$$

which is not constrained by the means. For the Baltimore eye survey example, Y_1 and Y_2 would represent the visual impairment status for the left and right eyes, respectively, from each individual. An OR of 2 means that the odds of visual impairment for one eye increases twofold if the other eye was impaired.

Different measures of association between discrete variables have been suggested in the literature (9). The OR has been chosen as the primary measure of within-cluster dependence for discrete data, mainly because it is easy to interpret and is familiar to public health researchers.

3. STATISTICAL MODELS FOR CORRELATED DATA

To analyze clustered data, we must model both the regression of Y on x and the within-cluster dependence. If the responses are independent of each other, GLMs, reviewed in Section 2.1, can be used for diverse types of responses. For correlated data, GLMs are not sufficient, as we discussed in Section 1.3, and alternative approaches that address the dependence are needed. We now review three different modeling approaches: marginal, random effects, and observation driven. In marginal models, the regression of Y on x and the within-cluster dependence are modeled separately. The other approaches attempt to address both issues simultaneously through a single model. The public health investigator must select a model based upon one or a combination of these approaches whose parameters most nearly capture the scientific objectives, rather than on the basis of mathematical convenience.

Below, we let $Y_i = (Y_{i1}, \ldots, Y_{ij}, \ldots, Y_{ini})'$ be a $n_i \times 1$ vector of responses from the i^{th} cluster, $i = 1, \ldots, K$. Note that n_i, the cluster size, may vary. In addition to Y_{ij}, one observes a $p \times 1$ vector of explanatory variables, x_{ij}, thought to be related to Y_{ij}. We now consider each modeling approach in turn.

3.1 Marginal Models

In a marginal model, the regression of Y on x and the within-cluster dependence are modeled separately. For the former, we model the marginal expectation $E(Y_{ij})$ as a function of explanatory variables. The marginal expectation is the average response over the population of individuals with a common value of x, just as in the univariate case when $n_i = 1$ for each cluster. Specifically, we assume the following:

1. The marginal expectation or "population-average" of the response, $\mu_{ij} = E(Y_{ij})$ depends on the explanatory variables x_{ij} through $g(\mu_{ij}) = x_{ij}'\beta$,

where g is a known link function, such as the logit for dichotomous responses or log for counts just as in GLMs.

2. The marginal variance depends on the marginal mean by $\text{Var}(Y_{ij}) = V(\mu_{ij})\phi$ where V is a known variance function, such as $V(\mu_{ij}) = \mu_{ij}$ for count data, and ϕ is the overdispersion parameter to be estimated just as in GLMs.
3. The covariance between Y_{ij} and Y_{ik} is a function of the marginal means and perhaps of additional parameters α, i.e. $\text{cov}(Y_{ij}, Y_{ik}) = c(\mu_{ij}, \mu_{ik}; \alpha)$ where c is a known function.

In the Baltimore eye survey study, a marginal model can be used to assess the dependence of visual impairment for either eye on demographic variables, such as sex, age, and race. For simplicity, let x_i indicate the sex (1-male; 0-female) of a subject. A simple marginal model has the form

$$\text{logit}\mu_{ij} = \log[\mu_{ij}/(1 - \mu_{ij})] = \beta_0 + \beta_1 x_i,\ j = 1,2,$$

$$\text{Var}(Y_{ij}) = \mu_{ij}(1 - \mu_{ij}),$$

$$\log \text{OR}\ (Y_{i1}, Y_{i2}) = \alpha_0.$$

The parameter, $\exp(\beta_1)$, is the odds of visual impairment for any eye among males relative to the odds among females. In other words, we have assumed in this model that the difference between men and women in the prevalence of visual impairment is the same for right and left eyes. Note that $\exp(\beta_1)$ is a ratio of population frequencies, so we refer to it as a population-averaged parameter. The degree of correlation between two eyes, measured by the OR, was assumed to be the same for each individual. This simplified assumption may be checked and modified, if necessary, by imposing a regression model on $\text{OR}(Y_{i1}, Y_{i2})$, such as

$$\log \text{OR}(Y_{i1}, Y_{i2}) = \alpha_0 + \alpha_1 x_i.$$

A negative α_1 would suggest that the degree of correlation between two eyes is stronger among females.

Because the marginal model coefficients, β, describe the effects of covariates on the marginal expectation of the response variables, β has the same interpretation regardless of the cluster size, n_i, which may vary among clusters. In particular, the interpretation of β is the same as from a GLM for independent data, which would be appropriate if $n_i = 1$ for all clusters. The interpretation of β is not altered by the magnitude of within-cluster dependence as described by the parameter α.

3.2 *Random Effects Models*

The essence of the random effects model is the assumption that parameters vary from cluster to cluster, thus reflecting natural heterogeneity caused by

unmeasured factors. Suppose Hmong children grow roughly as a linear function of age over the span of our study, so that each child's growth can be summarized by a baseline height (intercept) and growth rate (slope). Children obviously enter the study at different heights and have different growth rates. This heterogeneity in intercepts and slopes is beyond what can be explained by such predictor variables as age and sex. A random effects model is a reasonable description of the data if the collection of intercepts and slopes across children can be thought of as a sample from a distribution. The correlation of repeated observations of heights from the same child arises because of the heterogeneity among children in their true growth curves, which cannot be observed.

To be more specific, suppose a subset of the Hmong data can be described by the following simple model for the height, Y_{ij}, of child i at age j:

1. $Y_{ij} = (\beta_1^* + b_{i1}) + (\beta_2^* \text{age}_{ij} + b_{i2}\text{age}_{ij}) + \beta_3^* \text{sex}_i + \beta_4^* (\text{age}_{ij} \times \text{sex}_i) + \epsilon_{ij}$ where $\text{sex}_i = 1$, if boy and 0 if girl, age_{ij} = the age of the i^{th} child at the j^{th} visit, and $\epsilon_{ij} \sim N(0, \sigma^2)$, $j = 1, \ldots, n_i$.
2. $\epsilon_{i1}, \ldots, \epsilon_{in_i}$ are independent of one another, given $b_i = (b_{i1}, b_{i2})$.
3. b_i follows a bivariate normal distribution with mean = (0,0) and covariance matrix

$$\begin{pmatrix} \delta_{11} & \delta_{12} \\ \delta_{21} & \delta_{22} \end{pmatrix}.$$

In this model, β_1^* and $\beta_1^* + \beta_3^*$ represent the average baseline heights for girls and boys, respectively, whereas β_2^* and $\beta_2^* + \beta_4^*$ describe the average growth rates for girls and boys. A significantly positive β_4^* indicates that, on average, boys grow faster than girls. A negative b_{i2} indicates that the growth rate of the i^{th} child is lower than the average. The random effects variances , δ_{11} and δ_{22}, measure the variability of the initial heights and growth rates among children that cannot be explained by the gender difference. The correlation $\delta_{12}/(\delta_{11}\delta_{22})^{1/2}$ measures the association between the child-specific deviations in initial height and growth rate.

To accommodate the variety of responses seen in public health research, a random effects GLM can be described as follows:

1. Given random effects, b_i, which are specific to the i^{th} cluster, the conditional distribution of Y_{ij} follows a GLM with $g[E(Y_{ij}|b_i)] = x_{ij}'\beta^* + z_{ij}'b_i$, where z_{ij}, a $q \times 1$ vector of covariates, is a subset of x_{ij}.
2. Given b_i, $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$ are statistically independent.
3. The b_is are independent observations from a distribution, $F(\cdot;\delta)$, indexed by some possibly unknown parameters, δ. The term "random effects" was coined for b_i, because we treat the b_i as a random sample from F. The key assumption is that a cluster's b_i is not related to the x_{ij}s.

The random effects model is especially useful when the objective is to make inference about individuals, such as the children in the Hmong study, rather than just about the population average, which can be done equally well with the marginal approach. A more detailed discussion on the difference and the connection between β^* in the random effects model and β in the marginal model is given in Zeger et al (41), Neuhaus et al (23), and Zeger & Liang (40).

3.3 *Observation-Driven Models*

Cox (6) coined the phrase "observation-driven" model for the situation in which the correlation within a cluster arises because one response is explicitly caused by others. An example is infectious disease incidence for children from the same family. If Y_{ij} indicates whether a child j in family i has an incident case, it is likely that Y_{ij} will be directly caused by the responses for the other children, Y_{ik}. $k \neq j$, because infections are often passed among siblings.

Observation-driven models are commonly used in longitudinal and times series studies where they are termed "Markov" or "transition" models (33). Here, the conditional distribution of the response Y_{ij} at time t_{ij} given the entire past, $Y_{ij-1}, Y_{ij-2}, \ldots, Y_{i1}$, is modeled as a function of the explanatory variables, x_{ij}, and explicitly as a function of the past responses themselves. That is, the past outcomes, or functions thereof, are treated as additional explanatory variables. To simplify the analysis, we commonly assume that only the most recent past observations affect the current response. For example, the probability of respiratory infection for a given child at visit, j, depends only on whether the child had infection at visit, $j - 1$, as well as on explanatory variables, x_{ij}.

The autoregressive model for equally spaced, continuous observations is an example of an observation-driven model. In the simplest first order case, we assume

$$Y_{ij} = x_{ij}'\beta^{**} + \epsilon_{ij}, \qquad 4.$$

where

$$\epsilon_{ij} = \alpha\epsilon_{ij-1} + a_{ij} \qquad 5.$$

and a_{ij} are independent, mean-zero innovations. The residual from the linear regression at time, t_{ij}, depends explicitly on the residual the previous time. Thus, the past directly influences the present. The GLM extension of the transition model is straightforward. We model the conditional distribution of Y_{ij}, given the past as an explicit function of the preceding responses. To illustrate, consider the logistic regression model to study the association between vitamin A deficiency and respiratory infection, using data on chil-

dren from the Indonesian Children's Health Study. Here, the response is whether a child had infection, and the major explanatory variable is whether the child was vitamin A deficient. But, children infected at one visit are more likely to be infected at the next. An observation-driven model could be written as

$$\text{logit } E(Y_{ij}|y_{ij-1}, y_{ij-2}, \ldots, y_{i1}) = \text{logit } E(Y_{ij}|y_{ij-1}) = x_{ij}'\beta^{**} + \alpha y_{ij-1}. \quad 6.$$

To specify the general transition model, let $P_{ij} = \{y_{ij-1}, \ldots, y_{i1}\}$ be the past responses at time, t_{ij}, and let $\mu_{ij}^c = E(Y_{ij}|P_{ij})$ and $v_{ij}^c = \text{var}(Y_{ij}|P_{ij})$ be the conditional mean and variance of Y_{ij} given the past responses and the explanatory variables. Analogous to the GLM for independent data, we assume

$$g(\mu_{ij}^c) = x_{ij}'\beta^{**} + \Sigma_{v=1}^{q} f_v(P_{ij}; \alpha)$$

$$v_{ij}^c = V(\mu_{ij}^c).$$

The past outcomes, after transformation by the known functions, f_v, are treated as additional explanatory variables. If the model for the conditional mean is correctly specified, we can treat the repeated transitions for a person as independent events and use standard statistical methods. See Zeger & Qaqish (42), Korn & Whittemore (13), and Ware et al (33) for additional discussion.

When the cluster is formed by a factor, such as family, rather than time, the observations within a cluster do not have a natural, complete ordering. Hence, although it is still possible to specify an observation-driven model in terms of conditional distributions of Y_{ij} given Y_{ik}, $k \neq j$, greater care is required. It is easy to choose apparently sensible conditional models that cannot exist. See Besag (1) for details in the context of spatial data.

With binary responses, Liang & Zeger (17) show that the joint distribution of $Y_{i1}, \ldots, Y_{in_i}$ can be fully specified by n_i logistic models for Y_{ij} given Y_{ik}, $k \neq j$. Rosner (27) has developed beta-binomial models by using this approach. However, it has an inherent limitation. The interpretation for β^{**} is not the same for clusters of different size. For example, in a family study, suppose we regress the response for one member on the responses for others, as well as on explanatory variables. The interpretation of the coefficients for the explanatory variables is different when the number of family members is different. This is a serious problem that limits the use of models that are specified in terms of conditional distributions.

This problem is less acute in longitudinal studies. In the linear regression case, it is possible to specify the model so that the interpretation of the regression coefficients does not change as the number of previous responses used to predict the current value changes. With nonlinear links, such as the

log and logit, this is not generally possible. The meaning and values of regression coefficients, β^{**}, change when the model for the dependence on the prior outcomes is changed. This is in contrast to the marginal model, in which the regression of Y on x can be separated from the within-cluster association. In most observation-driven models, the two objectives are intertwined into one modeling equation.

4. STATISTICAL INFERENCE

This section reviews statistical inference for parameters specified by models discussed in Section 3. The focus is centered upon the application of the estimating procedure, GEE (16, 19, 25, 39), to the marginal models for which the GEE method was originated. Extension of the GEE method to the other two models is also briefly reviewed.

4.1 Marginal Model

When the regression analysis for the mean is the primary interest, the β coefficients specified in Section 3.1 can be estimated by solving the estimating equation

$$U_1(\beta,\alpha) = \sum_{i=1}^{K} \left(\frac{\partial \mu_i}{\partial \beta}\right)' [\mathrm{Cov}(Y_i;\ \beta,\ \alpha)]^{-1}(Y_i - \mu_i(\beta)) = 0 \qquad 7.$$

where $\mu_i(\beta) = E(Y_i)$, the marginal expectations for Y_i. Note that U_1 in Equation 7 has exactly the same form as $U(\beta)$ in Equation 3, except that Y_i is now a $n_i \times 1$ vector, which comprises the n_i observations from the i^{th} cluster, and the covariance matrix, $\mathrm{cov}(Y_i)$, for Y_i depends not only on β but on α, which characterizes the within-cluster dependence. This additional complication can be alleviated by iterating until convergence between solving $U_1[\beta,\hat{\alpha}(\beta)] = 0$ and updating $\hat{\alpha}(\beta)$, an estimate of α (16). The GEE approach is simply to choose parameter values $\hat{\beta}$ so that the expected $\mu_i(\beta)$ is as close to the observed Y_i as possible, weighting each cluster of data inversely to its variance matrix $\mathrm{var}(Y_i;\ \beta,\ \alpha)$, which is a function of the within-cluster dependence.

Generalized estimating equations have some theoretical and practical advantages. First, no joint distribution assumption for $Y_i = (Y_{i1}, \ldots, Y_{in_i})$ is required to use the method. This is especially important for discrete responses for which there are no simple and sensible classes of joint distributions. Second, $\hat{\beta}$, the solution of $U_1[\beta,\ \hat{\alpha}(\beta)] = 0$, has high efficiency compared with the maximum likelihood estimate of β in many cases studied. Third, White (35), Gourieroux et al (10), and Liang & Zeger (16) proposed use of a a robust variance, $V_{\hat{\beta}}$, of $\hat{\beta}$, which, in conjunction with $\hat{\beta}$, often provides valid

inferences for β, even when the covariance structure $c(\mu_{ij}, \mu_{ik}; \beta, \alpha)$ in Section 3.2 is misspecified. Specifically, suppose the investigators mistakenly assume that the observations from the same cluster are independent of each other. The 95% confidence interval for each regression coefficient β_j, $j = 1, \ldots, p$, based upon

$$\hat{\beta}_j \pm 1.96(V_{\hat{\beta}_j})^{1/2}$$

remains valid. Thus, investigators are protected against misspecification of the within-cluster dependence structure. This is especially appealing when the data set comprises a large number of small clusters, as is the case for nearly all the examples considered in Section 1.

When the within-cluster dependence is the primary interest, as is true for most family studies, this first procedure, which we call "GEE1," has an important limitation (18). In GEE1, we estimate β and α, acting as if they are independent of each other. Consequently, very little information from β is used when estimating α. This can lead to a significant loss of α information. As a remedy, Prentice (25) and Liang et al (18) discuss estimating $\delta = (\beta, \alpha)$ jointly by solving

$$U_2(\beta,\alpha) = \sum_{i=1}^{K} \left(\frac{\partial \mu_i^*}{\partial \delta}\right)' [\mathrm{Cov}(Z_i; \delta)]^{-1}(Z_i - \mu_i^*(\delta)] = 0, \qquad 8.$$

where

$$Z_i = (Y_{i1}, \ldots, Y_{in_i}, Y_{i1}^2, \ldots, Y_{in_i}^2, Y_{i1}Y_{i2}, Y_{i1}Y_{i3}, \ldots, Y_{in_i} \ldots, Y_{in_i-1}Y_{in_i})' \qquad 9.$$

and $\mu_i^* = E(Z_i; \delta)$, which is completely specified by the modeling assumptions made in Section 3.1. We call this expanded procedure, which uses both the Y_{ij}s and $Y_{ij}Y_{ik}$s, the "GEE2."

GEE2 appears to have high efficiency for both β and α (18). On the other hand, the robustness property for β of GEE1 is no longer true. Hence, correct inferences about β require correct specification of the within-cluster dependence structure given by $c(\mu_{ij}, \mu_{ik}; \beta, \alpha)$. The same authors suggest using a sensitivity analysis when making inference on β. That is, one may repeat the procedure with different models for the within-cluster dependence structure to examine the sensitivity of $\hat{\beta}$ to choice of dependence structure.

4.2 Random Effects Models

The GEE approach can also be used to estimate β^* in some random effects models. For illustration, let us assume that investigators collect count data and use a random intercept model to account for heterogeneity among clusters. That is, they assume the following:

1. Given a scalar random effect b_i, the counts $Y_{i1}, \ldots, Y_{in_i}$ are independent, and each has a Poisson distribution whose mean follows

$$\log E(Y_{ij}|b_i) = x_{ij}'\beta^* + b_i.$$

2. b_i follows a normal distribution with mean zero and variance δ.

To use the GEE method, one needs to compute the marginal expectations of the Z_is given in Equation 9. These computations are tedious yet straightforward. For example,

$$E(Y_{ij}) = E[E(Y_{ij}|b_i)] = E(e^{x'_{ij}\beta^* + b_i}) = e^{\delta/2 + x'_{ij}\beta^*}$$

and for $j \neq k$

$$E(Y_{ij}Y_{ik}) = E[E(Y_{ij}Y_{ik}|b_i)] = E[E(Y_{ij}|b_i)E(Y_{ik}|b_i)]$$
$$= E(e^{x'_{ij}\beta^* + x'_{ik}\beta^* + 2b_i}) = e^{2\delta + x'_{ij}\beta^* + x'_{ik}\beta^*}.$$

Thus, with minor modifications. statistical software that is suitable for fitting marginal models can be used to make inference about the fixed effects, β^*, and the variance of the random effects, δ (41). However, when one is interested in the estimation of the random effects, b_i, a different strategy is needed. Interested readers are referred to Laird & Ware (15), Stiratteli et al (30), Zeger & Karim (38), Breslow & Clayton (2), Schall (28), and Waclawiw & Liang (32).

4.3 *Observation-Driven Models*

As stated above, we treat other responses like explanatory variables in observation-driven models. Hence, we can fit this class using GEE by simply adding the necessary outcomes or functions thereof to the list of predictors for each observation. To illustrate, suppose we assume the logistic model for respiratory infection given in Equation 6. Then, we must add the infection status at the previous visit to vitamin A status in the list of explanatory variables and, in this case, use standard software for fitting a logistic model. If the first order Markov assumption is correct, the inferences will be correct. If the respiratory response at one time, given the entire past, depends on more than the last outcome as assumed, the inferences will be incorrect in much the same way as discussed in Section 1.2.1.

With clusters that are not ordered by time, the same strategy is used, but the repeated conditional events, Y_{ij} given Y_{ik}, $k \neq j, j = 1, \ldots, n_i$, will not in general be independent, as was possible for time-ordered data. Hence, the models must be fit by using GEE or some other approach that accounts for the dependence.

5. BALTIMORE EYE SURVEY DATA

As stated earlier, the main objective of the Baltimore eye survey (31) was to identify demographic variables, such as age, race, and education level, that may be associated with the prevalence of visual impairment (VI). Liang et al (18) have provided descriptive statistics on prevalence of VI for both eyes for each race × age combination. In short, the prevalence increases with age and is higher among blacks; the discrepancy between blacks and whites increases with age. In addition, risks of VI were apparently similar between left and right eyes for each race × age combination. Table 3 gives the estimates and the standard errors of regression coefficients when separate logistic regression models have been fit for the right and left eyes. For example, coefficients for the variable "race" reveal that at age 60, the prevalence of VI for left eyes is 43% ($= e^{0.356} - 1$) higher among blacks than whites; similar results held for the right eye: 35% ($= e^{0.314} - 1$) higher among blacks. That the race-associated difference in prevalence is greater for the left eye than for the right appears to have occurred by chance, as the test statistic

$$Z^2 = (0.356 - 0.314)^2/[(0.132)^2 + (0.127)^2 + 0.0056^* 2] = 0.039$$

is not significant even at the 0.10 level. Note that the number 0.0056 was produced from the GEE procedure to account for the fact that the data from two correlated eyes for each person were used to compute Z^2. The estimated correlation coefficient 0.334 = 0.0056/(0.132*0.127) between the two estimates of the race effect is indirect evidence of strong within-person dependence between the two eyes.

For the rest of the analyses, we assumed that the logistic regression coefficients for the left and right eyes are the same. We fit a sequence of models that differ only in the manner of modeling the between-eye dependence. In model 1, we incorrectly ignored the dependence to illustrate the consequences. In models 2, 3, and 4, different assumptions about the associa-

Table 3 Regression estimates and standard errors (in parentheses) fitting separate logistic regression models to the right and left eyes from the Baltimore eye survey

	Logistic regression model						
Eye	Intercept	Race	Age-60	(Age-60)^2	Race(Age-60)	Race*(Age-60)^2	Education
Left	−2.870	0.356	0.050	0.0007	−0.003	−0.0009	−0.067
	(0.098)	(0.132)	(0.009)	(0.0004)	(0.011)	(0.0006)	(0.020)
Right	−2.781	0.314	0.048	0.002	0.004	−0.0012	−0.052
	(0.093)	(0.127)	(0.008)	(0.0004)	(0.011)	(0.0006)	(0.020)

tion between eyes were made, as summarized in Table 4. The key features of the results, presented fully in Table 4, can be summarized as follows:

1. Results from model 1, in which the dependence between eyes has been ignored, show the naive variance estimates of $\hat{\beta}$ are, in general, too small. For estimating the race effect, for example, the naive variance estimate is 28% [$= 1 - (0.089/0.105)^2$] smaller than the correct one. In this example, ignoring dependence does not lead to qualitatively different conclusions. However, ignoring dependence often leads to serious scientific mistakes, as can been seen, for example, in an analysis of a 2 × 2 cross-over trial (40).
2. The frequency of VI increases with age, more rapidly in later life. Blacks have roughly 35% more VI than whites at age 60. Persons with higher education have lower rates of VL.
3. A comparison of the logistic regression coefficients from models 2, 3, and 4 suggests that the regression inferences using GEE are not sensitive to how one models the dependence between eyes. This provides the investigators more confidence in the validity of the results.
4. The last column of Table 4 gives the ratio of the variance estimates from model 2 to the variance estimates where separate regression coefficients were being fitted for each eye. If the data from the two eyes of each person were indeed independent of each other, one would expect the ratio to be close to two because of doubling of the sample size. The ratios presented here (known as the design effect in the context of sample survey) are less than two, which indicates a strong degree of correlation between pairs of eyes.
5. One sees a strong within-person dependence for whites and blacks. For whites, the risk of VI for one eye is inflated by a factor of 9.8 ($e^{2.286}$) should the other eye also be affected. Among blacks, the corresponding OR is estimated as 17.17 ($e^{2.286+0.0557}$). This observation is apparently consistent with the long-standing clinical finding in this country that blacks have higher incidence of glaucoma and of diabetic mellitis, both of which tend to be bilateral diseases.

6. CONCLUDING REMARKS

Clustered data are increasingly common in public health research for many reasons. The search for earlier risk factors, such as biomolecular markers of the disease process, has increased the need for longitudinal studies. The advances in our understanding of the genetic roots of disease have made family studies more attractive. The increased appreciation for the social and behavioral contributions to disease has made multivariate measures neces-

Table 4 Regression estimates and standard errors (in parenthesis) for the visual impairment data for the Baltimore eye survey

Variable	Model 1	2	3	4	Variance ratio†
Intercept	−2.821 (0.076) (0.067)	−2.824 (0.076)	−2.824 (0.076)	−2.824 (0.076)	1.66
Race (1-B;0-W)	0.332 (0.105) (0.089)*	0.334 (0.105)	0.334 (0.015)	0.334 (0.105)	1.58
Age-60	0.049 (0.007) (0.006)*	0.049 (0.007)	0.049 (0.007)	0.049 (0.007)	1.65
(Age-60)^2	0.0018 (0.0004) (0.0003)*	0.0018 (0.0003)	0.0018 (0.0003)	0.0018 (0.0003)	1.60
Race*(Age-60)	0.001 (0.0009) (0.0008)*	0.0007 (0.0005)	0.0007 (0.009)	0.0007 (0.009)	1.60
Race*(Age-60)^2	−0.001 (0.0005) (0.0004)*	−0.001 (0.0005)	−0.001 (0.0005)	−0.001 (0.0005)	1.43
Education	−0.059 (0.017) (0.013)*	−0.060 (0.017)	−0.060 (0.017)	−0.060 (0.017)	1.39
log-odds ratio					
Intercept	—	2.555 (0.126)	2.286 (0.176)	2.390 (0.205)	
Race	— —	— —	0.557 (0.252)	0.500 (0.256)	
Age	—	—	—	−0.010 (0.011)	

*Naive standard error.
†Ratio of the variance estimates from Table 3 (left eye) to the variance estimates from Model 2.

sary. At the same time, increased computing power has made regression analysis more accessible to public health investigators. Hence, routine analyses attempt to characterize the nature of the dependence of a response on explanatory variables, rather than only asking whether such a relationship exists.

This paper has reviewed approaches to regression analysis of correlated data organized in clusters. We have focused on extensions of GLMs, so that the types of outcomes common in public health—continuous measures, binary indicators of disease counts, or times to events---can be treated in a unified fashion. We believe that marginal, random effects, and observation-driven models, or combinations thereof, provide researchers with many of the tools necessary to infer answers to their scientific questions.

Several issues remain to be addressed. In longitudinal studies, subjects sometimes enter and drop out of cohorts because of factors related to their response variables. Many current methods of analysis give biased inferences in these situations (8, 37). In observation-driven models, the regression inferences depend on the choice of the within-cluster dependence model. Better methods for choosing the best model from a set of candidates are needed. Random effects models are very attractive, but difficult to estimate except in the linear and log-linear case (41). Finally, many of the methods described above have relatively recently been put into practive. We need integrated software that will make it easy for the public health scientist to try different models in an effort to appreciate the sometimes subtle distinctions between them.

ACKNOWLEDGMENTS

This work was partially supported by grants #GM39622 and AI125529 from the National Institutes of Health.

Literature Cited

1. Besag, J. 1974. Spatial interaction and the statistical analyses of lattice systems. *J. R. Stat. Soc. B* 36:192–236
2. Breslow, N. E., Clayton, D. G. 1992. Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.* In press
3. Cohen, B. H. 1980. Chronic obstructive pulmonary disease: a challenge in genetic epidemiology. *Am. J. Epidemiol.* 112:274–88
4. Cohen, B. H., Ball, W. C. Jr., Brashears, S., Diamond, E. L., Kreiss, P., et al 1977. Risk factors in chronic obstructive pulmonary disease (COPD). *Am. J. Epidemiol.* 105:223–32
5. Cook, N. R., Ware, J. H. 1983. Design and analysis methods for longitudinal research. *Annu. Rev. Public Health* 4:1–23
6. Cox, D. R. 1981. Statistical analysis of time series, some recent developments. *Scand. J. Stat.* 8:93–115
7. Diggle, P. J. 1988. An approach to the analysis of repeated measurements. *Biometrics* 44:959–72
8. Diggle, P. J., Liang, K.-Y., Zeger, S. L. 1993. *Analysis of Longitudinal Data.* Oxford: Oxford Univ. Press
9. Goodman, L. A., Kruskal, W. H. 1979. *Measures of Association for Cross Classifications.* New York: Springer-Verlag

10. Gourieroux, C., Monfort, A., Trognon, A. 1984. Pseudo maximum likelihood methods: theory. *Econometrica* 52:681–700
11. Hulka, B. S., Wilcosky, T. C., Griffith, J. D. 1990. *Biological Markers in Epidemiology*. New York: Oxford Univ. Press
12. Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, F., et al. 1987. The Multicenter Cohort Study: rationale organization and selected characteristics of participants. *Am. J. Epidemiol.* 126: 310–18
13. Korn, E. E., Whittemore, A. S. 1979. Methods for analyzing panel studies of acute health effects of air pollution. *Biometrics* 35:795–802
14. Kueppers, F., Miller, R. D., Gordon, H., Hopper, N. G., Offord, K. 1977. Familial prevalence of chronic obstructive pulmonary disease in a matched prior study. *Am. J. Med.* 63:366–72
15. Laird, N. M., Ware, J. H. 1982. Random effects models for longitudinal studies. *Biometrics* 38:963–74
15a. Liang, K.-Y., Beaty, T. H. 1991. Measuring familial aggregation by using odds-ratio regression models. *Genetic Epidemiol.* 8:361–70
16. Liang, K.-Y., Zeger, S. L. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73:13–22
17. Liang, K.-Y., Zeger, S. L. 1989. A class of logistic regression models for multivariate binary time series. *J. Am. Stat. Assoc.* 84:447–51
18. Liang, K.-Y., Zeger, S. L., Qaqish, B. 1991. Multivariate regression analyses for categorical data (with discussion). *J. R. Stat. Soc. B* 54:3–40
19. Lipsitz, S. R. 1989. *Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association,* tech. rep. Boston: Dep. of Biostat. Harvard School of Public Health
20. McCullagh, P., Nelder, J. A. 1989 *Generalized Linear Models*. London: Chapman & Hall. 2nd ed.
21. Murray, R. M., Reveley, A. M., McGuffin, P. 1986. Genetic vulnerability to schizophrenia. *Psychiatr. Clin. North Am.* 9:3–16
22. Nelder, J. A., Wedderburn, R. W. M. 1972. Generalized linear models. *J. R. Stat. Soc. A* 135:370–84
23. Neuhaus, J. M., Kalbfleisch, J. D., Hauck, W. W. 1991. A comparison of cluster-specific and population averaged approaches for analyzing correlated binary data. *Int. Stat. Rev.* 59:25–36
24. Ottman, R., Pike, M. C., King, M. C., Casagrande, J. T., Henderson, B. E. 1986. Familial breast cancer in a population-based series. *Am. J. Epidemiol.* 123:15–21
25. Prentice, R. L. 1988. Correlated binary regression with covariate specific to each binary observation. *Biometrics* 44:1022–48
26. Rorabaugh, M. L. 1990. *Catch-up growth in young Hmong refugee children.* Doctoral thesis. Johns Hopkins School of Hygiene and Public Health, Baltimore
27. Rosner, B. 1984. Multivariate methods in ophthalmology with application to paired data situations. *Biometrics* 40: 961–71
28. Schall, R. 1991. Estimation in generalized linear models with random effects. *Biometrika* 78:719–27
29. Starfield, B., Shapiro, S., Weiss, J., Liang, K.-Y., Ra, K., et al 1991. Race, family income and low birthweight. *Am. J. Epidemiol.* 134:1167–74
30. Stiratteli, R., Laird, N., Ware, J. H. 1984. Random-effects models for serial observations with binary response. *Biometrics* 40:961–71
31. Tielsch, J. M., Sommer, A., Witt, K., Katz, J., Royall, R. M. 1990. Blindness and visual impairment in an American urban population: Baltimore eye survey. *Arch. Ophthalmol.* 108:286–90
32. Waclawiw, M. A., Liang, K.-Y. 1993. Prediction of random effects in the generalized linear model. *J. Am. Statist. Assoc.* In press
33. Ware, J. H., Lipsitz, S., Speizer, F. E. 1988. Issues in the analysis of repeated categorical outcomes. *Stat. Med.* 7:95–107
34. Wedderburn, R. W. M. 1974. Quasi-likelihood function, generalized linear models and the Gaussian method. *Biometrika* 61:439–47
35. White, H. 1982. Maximum likelihood estimation of misspecified models. *Econometrics* 50:1–25
36. Williams, J. P. G. 1981. Catch-up growth. *J. Embryol. Exp. Morphol.* 65: 89–101
37. Wu, M. C., Carroll, R. J. 1988. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* 44:175–88
38. Zeger, S. L., Karim, M. R. 1991. Generalized linear models with random effects; a Gibbs sampling approach. *J. Am. Stat. Assoc.* 86:79–86
39. Zeger, S. L., Liang, K.-Y. 1986. Longi-

tudinal data analysis for discrete and continuous outcomes. *Biometrics* 42:121–30
40. Zeger, S. L., Liang, K.-Y. 1992. An overview of methods for the analysis of longitudinal data. *Stat. Med.* In press
41. Zeger, S. L., Liang, K.-Y. & Albert, P. 1988. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 44:1049–60
42. Zeger, S. L., Qaqish, B. 1988. Markov regression models for time series: a quasilikelihood approach. *Biometrics* 44:1019–31