Published in final edited form as:

Stat Med. 2011 April 15; 30(8): 866–876. doi:10.1002/sim.4157.

EXACT LOGISTIC MODELS FOR NESTED BINARY DATA

STEVEN TROXLER¹, TRENT LALONDE², and JEFFREY R. WILSON^{3,*}

¹Department of Statistics, University of California, 367 Evans Hall #3860, Berkeley, California 94720-386, U.S.A

²Department of Applied Statistics, University of Northern Colorado, 520 McKee Hall Box 124 Greeley, Colorado 80634, U.S.A

³NIH Alzheimer Research Center, Arizona State University, BAC 565, Tempe, Arizona, 85287-3806, U.S.A

SUMMARY

The use of logistic models for independent binary data has relied first on asymptotic theory and later on exact distributions for small samples. However, the use of logistic models for dependent analysis based on exact analysis is not as common. Moreover attention is usually given to one-stage clustering. In this paper we extend the exact techniques to address hypothesis testing (estimation is not addressed) for data with second-stage and probably higher levels of clustering. The methods are demonstrated through a somewhat generic example using C++ program.

Keywords

Correlated; dependent; clustering; hierarchical; multilevel; small sample

1. INTRODUCTION

The analysis of binary outcomes is common in many settings across several disciplines. For example, in the health sciences we often model the occurrence of a disease or the improvement of a patient. In marketing we might be interested in a consumer's decision to purchase a good. In educational research, graduating or not and the decision to continue to college may be useful outcomes to consider. In a toxicity study, survival or not may be the response, while in a quality-control study, the presence of defects might be the outcome of interest. It is customary to model these outcomes using a logistic model. The estimates of regression parameters in logistic models may be obtained using large sample theory, or with exact distribution theory. When the data are sparse, exact methods of estimation, based on sufficient statistics, are generally preferred. The large sample theory estimates have been shown to be unreliable when there are many covariates relative to the number of observations.

Clustered data occur frequently in statistical practice. In some areas of application, such as ophthalmology, clustered data are the rule rather than the exception. In such settings standard logistic regression models are invalid, because of the lack of independence among outcomes for individual sample points within a cluster. A one-level of nesting correlation structure is assumed as is typically found in ophthalmology, where person is the cluster and the individual eyes are the units within the cluster. Rosner [1] generalized to a two-level of nesting structure. Have, Kunselman, and Tran [2] compared mixed effects logistic

*Correspondence to: NIH Alzheimer's Center, BAC 565, Arizona State University, Tempe AZ 85287-3806, U.S.A .

regression models for binary response data with two nested levels of clustering. The comparison of these models occurs in the context of developmental toxicity data sets, for which multiple types of outcomes (first level) are measured on each rat pup (second level), nested within a litter (third level). Because the nested nature of such data is occasionally accommodated by ignoring one level of clustering, they considered three models: (i) a three-level model adjusting for clustering due to both pup and litter; (ii) a two-level model adjusting for just pup; and (iii) another two-level model adjusting for just litter.

The analysis of exact methods dates back to the Fisher's Exact Test for 2×2 tables and the multi hypergeometric version for larger dimensions. Corcoran, Ryan, Senchaudhuri, Mehta, Patel, and Monenbergs [3] presented exact methods to analyze correlated data. They proposed an exact trend test for clustered binary observations. The model was applied to a study of hospital infections, in which rates of patients in the same wing developing secondary infections are correlated. Similarly, in an educational study, we might expect peer pressures and the tendency of students to assist or be disruptive to each others' learning might cause correlation in their academic progress within a classroom. As such an analysis to determine which factors affect students' ability to advance may treat classrooms as clusters. Such inherent overdispersion cannot be ignored.

In this paper, we extend the work of Corcoran et al [3] as their model concentrated only on a first stage of clustering. In so doing we make use of two levels of clustering, a first stage and within this first stage we refer to as a second stage of clustering. Multiple levels of clustering are not uncommon. For example, in a regional study of secondary infections in hospitals, the infection rates of patients on the same floor may be strongly correlated and the rates for patients on different floors in the same hospital may also be correlated, though not as strong yet enough to contribute to the overall overdispersion. In such a setting, the hospital would be the first stage of clustering and the floor the second stage of clustering. In addition, consider a school district investigating graduation rates as they suspect that the academic progress of students in the same classroom are highly correlated, and the progress of students in different classrooms within the same school also exhibit some correlation, so that the first stage or primary cluster is the school and the second stage or secondary cluster is the classroom. Yet another example is to consider a district conducting a multi-year study of graduation rates among students. In higher grades where classes tend to be mixed, one might consider the school to be a first stage cluster, because the peer environment will affect student performance even across grade levels. One might treat the class year as a second stage cluster, as one may argue that students within the same class year will tend to influence one another more than students in different class years. Similarly, in a repeated-observation study in which we have already identified one level of clustering at any given time, we could treat those clusters as first stage clusters, and treat the observations on each individual as second stage clusters.

This paper extends the one-stage model to include a second stage of clustering. In so doing we allow correlation between clusters within the one-stage model in Section 3.2. It is also important to note that most of the work is with a single covariate. Further the covariate takes on single digit values. If this were not the case it seems very likely that the computation time will be excessive. As correctly noted by a reviewer in an earlier version the difficulty that is intrinsic to all exact methods for logistic regression namely over conditioning can occur leading to a single element in the reference set – in which case the approach fails.

In Section 3.3, we allow the covariates to differ within clusters in a two-stage binomial model. We also discuss a generalization to a three- stage model. Gibbons and Hedeker [4] examined modeling random effects probit and logistic regression models for three-level data. The parameter estimation in that case is based on full-information maximum marginal

likelihood estimation. Their model is illustrated using data from 135 classrooms with 28 schools on the effects of two smoking cessation interventions. In Section 4 we consider numerical examples and the computations necessary for this analysis presented in a two-stage clustering. Some conclusions are given in Section 5.

2 CORRELATED LOGISTIC REGRESSION MODELS

2.1 Exact Logistic Regression

As a common means of modeling binary outcomes, the logistic model, a member of the class of generalized linear models, with link function and systematic component specified as

$$\ln \left[p_j / (1 - p_j) \right] = \alpha + x_j' \beta$$

where p_j is the probability of a favorable outcome on the j^{th} observation, α and β are fixed unknown parameters, and x_j is the vector of covariates for the j^{th} observation. Let $\mathbf{Y} = (Y_1, \dots, Y_n)$ denote a vector of n observations. If Y_j 's are independent, then the joint probability mass function,

$$\Pr(Y_1=y_1, \dots, Y_n=y_n) = \exp \left\{ \alpha + \sum_{j=1}^n y_j x_j' \beta + C \right\},$$

where C is a normalizing constant obtained by summing over all possible outcomes, Cox [5], Cox and Snell [6], and Collett [7]. To obtain estimates of α and β it is a common method to obtain their maximum likelihood estimates while obtaining Wald statistics and other large-sample inferential techniques. However, when the number of regressors is large relative to the number of observations these maximum likelihood methods are known to be unreliable, and in some cases if the data are too sparse then the unconditional likelihood function may not have a local maximum.

Cox [5] proposed a method for dealing with sparse data, in particular when interest is only on some components of β , and the other components are considered nuisance parameters. This approach relies in part on the fact that the logistic function, through its exponential form, can easily be used to determine sufficient statistics for the parameters of interest and ancillary statistics sufficient for the nuisance parameters. He suggested that one should explicitly compute the conditional distribution of the sufficient statistics for the parameters of interest over all possible outcomes that lead to the observed values of the ancillary statistics. This method at times can lead to a huge number of parameters, having exponential growth relative to the sample size, so this exact method may prove time consuming to implement. Computer speed was once a challenge, but researchers have developed efficient algorithms to compute the distribution in polynomial time, rather than exponential time, Trichter [8] and Hirji, Mehta, and Patel [9]. Derr [10] and also Mehta and Patel [11] have given a complete description of the theoretical aspects of exact logistic regression with a discussion of its implementation. Mehta and Patel [11] pointed out the major differences between the results based on the exact method as compared to those based on maximum likelihood estimation.

2.2 Correlated Data and the One-Stage Clustered Model

Mehta and Patel [11] addressed issues in exact analysis of data using the logistic model with independent observations, but the problem with correlated observations has received little attention. Several authors, in particular Connolly and Liang [12], have proposed methods of performing large-sample logistic estimation when the observations are correlated, but exact

methods for correlated data have only recently become the focus. Corcoran et al [3] proposed an exact method for clustered data that draws on conditioning arguments commonly used with exponential family models for samples of independent observations. We summarize their results here, and in subsequent sections describe how this approach can be extended to two- or three-stage study designs, and to data that include within-cluster covariates. We now describe their technique.

Let Y_{ij} denote the j^{th} observation ($j = 1, 2, \dots, n_i$) in the i^{th} cluster ($i = 1, 2, \dots, N$). Let $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})$ be the vector of responses for the i^{th} cluster and $Z_i = \sum_{j=1}^{n_i} Y_{ij}$. Suppose that the value of the covariate in the i^{th} cluster is x_i , the same for all observations within each cluster, and let $\mathbf{x} = (x_1, \dots, x_N)$. Then the probability mass function of $\mathbf{Y}_i$

$$Pr(\mathbf{Y}_i = \mathbf{y}_i | x_i) = \exp \{ (\alpha + \beta x_i) z_i - \delta_i z_i (n_i - z_i) + A_i(\alpha, \beta, \delta_i) \} \quad (1)$$

where A_i is a normalizing constant obtained by summing over all possible realizations of $\mathbf{Y}_i$, and the parameter δ_i is a dispersion parameter. The larger the value δ_i takes, the stronger is the correlation between observations within the i^{th} cluster. When $\delta_i = 0 \forall i$ then the resulting distribution is the logistic distribution for independent observations. Since the probability only depends on y_i only through its sum z_i , it follows that the probability mass function for Z_i

$$Pr(Z_i = z_i | x_i) = \binom{n_i}{z_i} \exp \{ (\alpha + \beta x_i) z_i - \delta_i z_i (n_i - z_i) + A_i(\alpha, \beta, \delta_i) \} \quad (2)$$

where $\binom{n_i}{z_i}$ is the number of distinct vectors $\mathbf{y}_i$ that yield the same value of $z_i = \sum_{j=1}^{n_i} y_{ij}$. Assuming that $\delta_i = \delta, \forall i$ and those outcomes in different clusters are independent, the joint probability mass function for the vector of outcomes $\mathbf{Z} = (Z_1, \dots, Z_N)$,

$$\begin{aligned} Pr(\mathbf{Z} = \mathbf{z} | \mathbf{x}) &= \prod_{i=1}^N \binom{n_i}{z_i} \exp \{ (\alpha + \beta x_i) z_i - \delta_i z_i (n_i - z_i) + A_i(\alpha, \beta, \delta) \} \\ &= \left[\prod_{i=1}^N \binom{n_i}{z_i} \right] \exp \left\{ \alpha s_1 + \beta t - \delta s_2 + \sum_{i=1}^N A_i(\alpha, \beta, \delta) \right\} \end{aligned} \quad (3)$$

It follows that $s_1(\mathbf{z}) = \sum_{i=1}^N z_i$ is sufficient for α , $s_2(\mathbf{z}) = \sum_{i=1}^N z_i (n_i - z_i)$ is sufficient for δ , and $t(\mathbf{z}) = \sum_{i=1}^N x_i z_i$ is sufficient for β , since the probability mass function is a member of the exponential family. For ease of reference, henceforth we will use S_1 , S_2 , and t to refer to the observed values $S_1(\mathbf{Z})$, $S_2(\mathbf{Z})$, and $t(\mathbf{Z})$. As the main interest is to make inferences about the value of β , the other parameters are considered nuisance parameters. Thus, we base our inference on the conditional distribution of the sufficient statistic t given the observed values of the sufficient statistics s_1 , and s_2 . Define a reference set

$$\Gamma(s_1, s_2) = \{ \mathbf{z}^* : s_1(\mathbf{z}^*) = s_1, s_2(\mathbf{z}^*) = s_2 \}, \quad (4)$$

which contains all possible outcomes $\mathbf{z}$ that would have resulted in the observed values of s_1 and s_2 . The conditional probability for $\mathbf{Z}$, given the observed value of S_1 , and S_2 ,

$$Pr(\mathbf{Z} = \mathbf{z} | \mathbf{x}; s_1, s_2) = \frac{Pr(\mathbf{Z} = \mathbf{z} | \mathbf{x})}{\sum_{\mathbf{z}^* \in \Gamma(s_1, s_2)} Pr(\mathbf{Z} = \mathbf{z}^* | \mathbf{x})} \quad (5)$$

which reduces to

$$\Pr(\mathbf{Z}=\mathbf{z}|\mathbf{x};s_1, s_2) = \frac{\left[\prod_{i=1}^N \binom{n_i}{z_i} \right] \exp\{\beta t\}}{\sum_{\mathbf{z}^* \in \Gamma(s_1, s_2)} \left[\prod_{i=1}^N \binom{n_i}{z_i^*} \right] \exp\{\beta t(\mathbf{z}^*)\}} \quad (6)$$

Thus this conditional distribution can be used to test $H_0: \beta=0$ against $H_a: \beta > 0$. Under $\mathbf{H}_0$, the conditional distribution of $\mathbf{Z}$ reduces to

$$\Pr(\mathbf{Z}=\mathbf{z}|\mathbf{x};s_1, s_2) = \frac{\prod_{i=1}^N \binom{n_i}{z_i}}{\sum_{\mathbf{z}^* \in \Gamma(s_1, s_2)} \prod_{i=1}^N \binom{n_i}{z_i^*}} \quad (7)$$

Since the likelihood ratio for $\beta = 0$ versus $\beta > 0$ is decreasing in t , a one-sided p -value for this test,

$$Pr(t > t_{obs} | H_0, \mathbf{x}, s) = \sum_{\mathbf{z}^* \in \Gamma(s_1, s_2): t(\mathbf{z}^*) \geq t} \left[\frac{\prod_{i=1}^N \binom{n_i}{z_i}}{\sum_{\mathbf{z}^* \in \Gamma(s_1, s_2)} \prod_{i=1}^N \binom{n_i}{z_i^*}} \right] \quad (8)$$

We reject H_0 if the p -value above is less than our significance level α , or equivalently if $t_{obs} > t_a$ where $t_{1-\alpha}$ is the smallest value of t' satisfying $\Pr(t < t' | H_0, \mathbf{x}, s) = 1 - \alpha$.

3 TWO-STAGE CLUSTERED BINARY MODEL

3.1 A Two-Stage Model

We now extend the model of Section 2 in cases of two stages of clustering, where the design consists of first-stage clusters and second-stage clusters nested within the first-stage clusters. Let Y_{ijk} be the k^{th} observation from the j^{th} secondary cluster contained in the i^{th} primary cluster ($i = 1, 2, \dots, I; j = 1, 2, \dots, J_i; K = 1, 2, \dots, n_{ij}$). Let the value of the covariate for all observations in the j^{th} secondary cluster be x_{ij} . Define $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{iJ_i})$ and $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_I)$. Let $Z_{ij} = \sum_{k=1}^{n_{ij}} Y_{ijk}$. Then the probability mass function of $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, \dots, Z_{iJ_i})$,

$$\Pr(\mathbf{Z}_i = \mathbf{z}_i | \mathbf{x}_i) = \left[\prod_{j=1}^{J_i} \binom{n_{ij}}{z_{ij}} \right] \exp \left\{ a s_{1i} + \beta t_i - \delta s_{2i} - \gamma \sum_{j=1}^{J_i} z_{ij} \left(\sum_{j=1}^{J_i} n_{ij} - \sum_{j=1}^{J_i} z_{ij} \right) + C_i(\alpha, \beta, \delta, \gamma) \right\}$$

where $s_{1i}(z_i) = \sum_{j=1}^{J_i} z_{ij}$, $s_{2i}(z_i) = \sum_{j=1}^{J_i} z_{ij} (n_{ij} - z_{ij})$, and $t_i(z_i) = \sum_{j=1}^{J_i} x_{ij} z_{ij}$, and C_i is a normalizing constant obtained by summing over all possible outcomes. The factor δ measures clustering within the second-stage clusters, while γ measures clustering within the first-stage clusters. The larger γ and δ are, the greater the dependency of the observations in the same first- and second-stage clusters, respectively. If we assume that the first-stage clusters are independent, then the probability mass function of the overall vector of responses $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_I)$,

$$Pr(\mathbf{Z}=\mathbf{z}|\mathbf{x}) = \left[\prod_{i=1}^I \prod_{j=1}^{J_i} \binom{n_{ij}}{z_{ij}} \right] \exp \left\{ \alpha s_1 + \beta t - \delta s_2 - \gamma s_3 + \sum_{i=1}^I C_i(\alpha, \beta, \delta, \gamma) \right\} \quad (10)$$

where

$$s_1(\mathbf{z}) = \sum_{i=1}^I s_{1i} = \sum_{i,j} z_{ij}$$

is sufficient for α ,

$$s_2(\mathbf{z}) = \sum_{i=1}^I s_{2i} = \sum_{i,j} z_{ij} (n_{ij} - z_{ij})$$

is sufficient for δ ,

$$s_3(\mathbf{z}) = \sum_{i=1}^I \left[\sum_{j=1}^{J_i} z_{ij} \left(\sum_{j=1}^{J_i} n_{ij} - \sum_{j=1}^{J_i} z_{ij} \right) \right]$$

is sufficient for γ , and

$$t(\mathbf{z}) = \sum_{i=1}^I t_i = \sum_{i,j} x_{ij} z_{ij}$$

is sufficient for β .

The probability mass function (10) bears some similarity to the probability mass function presented for the one-stage model in equation (3). The key difference is the introduction of an added dispersion parameter γ to account for clustering at the additional level. When $\gamma = 0$, the joint probability mass function in (10) reduces to (3), where only second-stage clusters are relevant. Likewise, if $\delta = 0$, then the joint probability mass function in (10) reduces to (3), where only the first-stage clusters are relevant.

To address statistical tests or inferences about β we focus on the conditional distribution of the sufficient statistic t , given the observed values of the sufficient statistics s_1, s_2 , and s_3 . Consider a reference set containing all the possible values of $\mathbf{z}$ that would have led to the same values of s_1, s_2 , and s_3 that we observed in the sample,

$$\Gamma(s_1, s_2, s_3 | \text{cluster sizes}) = \{ \mathbf{z} : s_j(\mathbf{z}) = s_j \text{ for } j=1, 2, 3 \}.$$

Then the conditional distribution of $\mathbf{Z}$ becomes

$$Pr(\mathbf{Z}=\mathbf{z}|\mathbf{x};s_1, s_2, s_3) = \frac{Pr(\mathbf{Z}=\mathbf{z}|\mathbf{x})}{\sum_{\mathbf{z}^* \in \Gamma(s_1, s_2, s_3)} Pr(\mathbf{Z}=\mathbf{z}^*|\mathbf{x})} \frac{\left[\prod_{i=1}^I \prod_{j=1}^{J_i} \binom{n_{ij}}{z_{ij}} \right] \exp\{\beta t\}}{\sum_{\mathbf{z}^* \in \Gamma(s_1, s_2)} \left[\prod_{i=1}^I \prod_{j=1}^{J_i} \binom{n_{ij}}{z_{ij}} \right] \exp\{\beta t(z^*)\}} \quad (11)$$

We make use of (11) to test $H_0: \beta = 0$ versus $H_a: \beta > 0$. Under H_0 , the terms involving t in the conditional distribution drop out, so the distribution of $\mathbf{Z}$ reduces to

$$Pr(\mathbf{Z}=\mathbf{z}|\mathbf{x};s_1, s_2, s_3) = \frac{\prod_{i=1}^K \prod_{i=1}^I \prod_{j=1}^{J_i} \binom{n_{ij}}{z_{ij}}}{\sum_{\mathbf{z}^* \in \Gamma(s_1, s_2)} \prod_{i=1}^I \prod_{j=1}^{J_i} \binom{n_{ij}}{z_{ij}}} \quad (12)$$

Thus the likelihood ratio for $\beta = 0$ versus $\beta > 0$ in (11) is decreasing in t , so a one-sided p -value for this test is

$$Pr(t > t_{obs} | H_0, \mathbf{x}, \mathbf{s}) = \sum_{\mathbf{z}^* \in \Gamma(s_1, s_2): t(\mathbf{z}^*) \geq t_{obs}} \left[\frac{\prod_{i=1}^I \prod_{j=1}^{J_i} \binom{n_{ij}}{z_{ij}}}{\sum_{\mathbf{z}^* \in \Gamma(s_1, s_2)} \prod_{i=1}^I \prod_{j=1}^{J_i} \binom{n_{ij}}{z_{ij}}} \right]$$

Thus we reject H_0 if the p -value is less than our significance level α , or equivalently, if $t_{obs} > t_\alpha$, where t_α is the smallest value of t for which $Pr(t > t_{obs} | H_0, \mathbf{x}, \mathbf{s}) = \alpha$.

3.2 Correlation between clusters in the One-Stage Model

Assume that, in the first-stage model described in Section 2.2, there is correlation not only between observations in the same cluster, but also correlation between any two observations in the sample. Have et. al [2] shared some possible examples. One way to model this correlation in a way that still permits exact analysis is to use the two-stage model described in Section 3.1 by treating the entire sample as a primary cluster, and each of the clusters as secondary clusters nested within the first stage cluster. In this case, the probability mass function becomes

$$Pr(\mathbf{Z}=\mathbf{z}|\mathbf{x}) = \left[\prod_{i=1}^N \binom{n_i}{z_i} \right] \exp\{\alpha s_1 + \beta t - \delta s_2 - \gamma s_3 + C(\alpha, \beta, \gamma)\} \quad (13)$$

where S_1, S_2, S_3 and t are defined as in (10). When there is only one first-stage cluster, the problem reduces by noting that s_1 and s_3 completely determine one another. Thus,

$$\begin{aligned} s_3 &= \sum_{i=1}^N \left(\sum_{i=1}^N n_i - \sum_{i=1}^N z_i \right) \\ &= s_1 \left(\sum_{i=1}^N n_i - s_1 \right). \end{aligned}$$

Since S_1 is sufficient for both γ and α , the reference set T used for conditional inference will contain the same values Z^* as the reference set used in the one-stage model with independent clusters. This phenomenon occurs because, although the probabilities in the

unconditional joint mass function of $\mathbf{Z}$ change when we add a term accounting for dependence between different clusters, the relative probabilities in the conditional distribution for any particular value of s_1 do not depend on γ . In fact, by conditioning on s_1 to account for the presence of α , we automatically account for possible dependence between clusters. This result holds true as long as the probability mass function includes a term in the

exponent of the form $F(\gamma)G\left(\sum_{i=1}^N z_i\right)$ regardless of the functional forms of F and G . Because the mass function is of exponential form, the sufficient statistics remain the same. The inference follows a similar argument as the original first-stage model.

3.3 Modeling Covariates within Clusters in the First-Stage Model

One drawback to the method proposed in [1] is that the covariates cannot differ within clusters. Modeling different covariates can be accomplished using the second-stage model and treating each individual observation as a second-stage cluster. In this case, the term in the probability mass function accounting for dependency in the second-stage cluster reduces to zero regardless of the outcome, so that the approach is equivalent to conditioning on the vectors $\mathbf{Y}_i$ instead of the z_i 's.

Consider a system with N clusters, the i^{th} cluster containing n_i observations $Y_{ij}, j = 1, \dots, n_i$.

Let $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})$, $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_N)$, and $Z_i = \sum_{j=1}^{n_i} Y_{ij}$. Let the value of the covariate for the ij^{th} observation be x_{ij} . Then based on equation (9), the joint probability,

$$Pr(\mathbf{Y}_i = \mathbf{y}_i) = \exp \left\{ \alpha z_i + \beta \sum_{j=1}^{n_i} y_{ij} x_{ij} - \delta z_i (n_i - z_i) + A_i(\alpha, \beta, \delta_i) \right\} \quad (14)$$

where we use δ , rather than γ , to denote the primary cluster dependency parameter. This allows notation to match with the original first-stage model. Assuming independence between clusters, the joint probability mass function of $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_N)$,

$$Pr(\mathbf{Y} = \mathbf{y}) = \exp \left\{ \alpha \sum_{i=1}^N z_i + \beta \sum_{i=1}^N \sum_{j=1}^{n_i} y_{ij} x_{ij} - \sum_{i=1}^N \delta z_i (n_i - z_i) + \sum_{i=1}^N A_i(\alpha, \beta, \delta_i) \right\} \quad (15)$$

which, expressed in terms of sufficient statistics,

$$Pr(\mathbf{Y} = \mathbf{y}) = \exp \left\{ \alpha s_1(\mathbf{y}) + \beta t(\mathbf{y}) - \delta s_2(\mathbf{y}) + \sum_{i=1}^N A_i(\alpha, \beta, \delta_i) \right\} \quad (16)$$

where

$$s_1(\mathbf{y}) = \sum_{i=1}^N z_i$$

is sufficient for α ,

$$s_2(\mathbf{y}) = \sum_{i=1}^N z_i (n_i - z_i)$$

is sufficient for δ , and

$$t(\mathbf{y}) = \sum_{i=1}^N \sum_{j=1}^{n_i} y_{ij} x_{ij}$$

is sufficient for β . Our reference set $\mathbf{T}(s_1, s_2)$,

$$\{\mathbf{y}^*: s_1(\mathbf{y}^*) = s_1(\mathbf{y}), s_2(\mathbf{y}^*) = s_2(\mathbf{y})\},$$

Under $H_0: \beta = 0$, the conditional distribution of t ,

$$Pr(t > t_{obs} | H_0, \mathbf{x}, s) = \sum_{\mathbf{y}^* \in \Gamma(s_1, s_2): t(\mathbf{y}^*) \geq t_{obs}} \left[\frac{\prod_{i=1}^N \binom{n_i}{z_i}}{\sum_{\mathbf{y}^* \in \Gamma(s_1, s_2)} \prod_{i=1}^N \binom{n_i}{z_i}} \right] \quad (17)$$

Given the distribution of t , the inference follows an argument analogous to that in Section 3.1. The drawback to this method is that the set of possible values of $\mathbf{Y}$ is much larger than the set of possible values of $\mathbf{Z}$, so when the sample size is relatively large, the reference set might be too time-consuming to construct.

3.4 Non-Constant Dependency Between Clusters in the One-Stage Model

Similarly we can relax the requirement that the dispersion parameters δ_i must be the same for all clusters. Allow the covariates to be the same within clusters. If we were to condition on the z_i 's, then the reference set will contain at most two possible values of z_i : either the observed value or the difference between n_i and that value. The reference set collapses, giving a nearly degenerate distribution that cannot be used for inference. If covariates differ within clusters, however, we can condition on the y_{ij} 's, and the reference set will generally remain large enough for exact inference.

Based on (1), the probability mass function of the vector $\mathbf{Y}_i$ of responses in the i^{th} cluster,

$$Pr(\mathbf{Y}_i = \mathbf{y}_i) = \exp \left\{ \alpha z_i + \beta \sum_{j=1}^{n_i} y_{ij} x_{ij} - \delta_i z_i (n_i - z_i) + A_i(\alpha, \beta, \delta) \right\}. \quad (18)$$

Assuming independence between clusters and allowing the δ_i 's to be unequal, we have

$$Pr(\mathbf{Y} = \mathbf{y}) = \exp \left\{ \alpha \sum_{i=1}^N z_i + \beta \sum_{i=1}^N \sum_{j=1}^{n_i} y_{ij} x_{ij} - \sum_{i=1}^N \delta_i z_i (n_i - z_i) + \sum_{i=1}^N A_i(\alpha, \beta, \delta) \right\}, \quad (19)$$

which, written in terms of sufficient statistics, is

$$Pr(\mathbf{Y} = \mathbf{y}) = \exp \left\{ \alpha s_1(\mathbf{y}) + \beta t(\mathbf{y}) - \sum_{i=1}^N \delta_i s_{2i}(\mathbf{y}) + \sum_{i=1}^N A_i(\alpha, \beta, \delta) \right\}, \quad (20)$$

where

$$s_1 = \sum_{i=1}^N z_i$$

is sufficient for α ,

$$s_{2i} = z_i (n_i - z_i)$$

is sufficient for δ_i and

$$t = \sum_{i=1}^N \sum_{j=1}^{n_i} y_{ij} x_{ij}$$

is sufficient for β . Our reference set $\mathbf{T}(s_1, s_2)$ is then

$$\{\mathbf{y}^* : s_1(\mathbf{y}^*) = s_1(\mathbf{y}), s_{2i}(\mathbf{y}^*) = s_{2i}(\mathbf{y}), \quad i=1, \dots, I\}, \quad (21)$$

With a reference set identified, which is much more restrictive on the z_i 's but still somewhat sizable in the y_i 's, the inference proceeds as discussed in Section 3.1.

The methods and techniques used in the second-stage model presented in Section 3.1 can be extended to more than two stages. However the level of complexity of the algorithm for computing the reference set increases as more parameters are added, and as such the utility of additional levels of clustering may be limited. The amount and feasibility of the additional computational work is an open research question (in addition to other issues, like the consequences of assuming constant within-cluster dependency where the dependency actually differs). Nevertheless, we now outline and explore some thoughts of a model with three levels of clustering.

Let $Y_{ijk\ell}$ denote the ℓ^{th} response of the k^{th} third-stage cluster in the j^{th} second stage cluster of the i^{th} first stage cluster ($i = 1, \dots, I; j = 1, \dots, J_i; k = 1, \dots, K_{ij}; \ell = 1, \dots, n_{ijk}$). Let x_{ijk} denote the value of the covariate in the ijk^{th} third-stage cluster, and let $\mathbf{X} = [x_{ijk}]$ be an array containing the x_{ijk} 's. Define $Z_{ijk} = \sum_{\ell=1}^{n_{ijk}} Y_{ijk\ell}$ and let $\mathbf{Z} = [Z_{ijk}]$ be a three-dimensional array whose ijk^{th} entry is Z_{ijk} . Similar to the structure in (10), we take

$$Pr(\mathbf{Z}=\mathbf{z}|\mathbf{X}) = \prod_{i=1}^I \prod_{j=1}^{J_i} \prod_{k=1}^{K_{ij}} \binom{n_{ijk}}{z_{ijk}} \exp \{ \alpha s_1(z) + \beta t(z) - \delta s_2(z) - \gamma s_3(z) - \eta s_4(z) + C(\alpha, \beta, \gamma, \eta) \}$$

where

$$s_1(z) = \sum_{i=1}^I \sum_{j=1}^{J_i} \sum_{k=1}^{K_{ij}} z_{ijk}$$

is sufficient for α ,

$$s_2(z) = \sum_{i=1}^I \sum_{j=1}^{J_i} \sum_{k=1}^{K_{ij}} z_{ijk} (n_{ijk} - z_{ijk})$$

is sufficient for δ ,

$$s_3(z) = \sum_{i=1}^I \sum_{j=1}^{J_i} \left(\sum_{k=1}^{K_{ij}} z_{ijk} \right) \left(\sum_{k=1}^{K_{ij}} (n_{ijk} - z_{ijk}) \right)$$

is sufficient for γ ,

$$s_4(z) = \sum_{i=1}^I \left(\sum_{j=1}^{J_i} \sum_{k=1}^{K_{ij}} z_{ijk} \right) \left(\sum_{j=1}^{J_i} \sum_{k=1}^{K_{ij}} (n_{ijk} - z_{ijk}) \right)$$

is sufficient η , and

$$t(z) = \sum_{i=1}^I \sum_{j=1}^{J_i} \sum_{k=1}^{K_{ij}} z_{ijk} x_{ijk}$$

is sufficient for β . In this three-stage model, the reference set,

$$\Gamma(s_1, s_2, s_3, s_4) = \{z^* : s_j(z^*) = s_j(z) \text{ for } j=1, 2, 3, 4\}$$

and the conditional distribution of t under $H_0 \nexists \beta = 0$ is

$$Pr(t > t_{obs} | H_0, \mathbf{x}, \mathbf{s}) = \sum_{\mathbf{z}^* \in \Gamma(s_1, s_2, s_3, s_4) : t(\mathbf{z}^*) \geq t_{obs}} \left[\frac{\prod_{i=1}^N \binom{n_i}{z_i}}{\sum_{\mathbf{z}^* \in \Gamma(s_1, s_2, s_3, s_4)} \prod_{i=1}^N \binom{n_i}{z_i^*}} \right].$$

Now that the reference set is defined, the inference proceeds similar to the method of Section 3.1.

4 NUMERICAL EXAMPLES AND COMPUTATIONS

Some examples with two-level clustering data were described by Have et. al [2]. Their examples included multiple family members followed longitudinally, multiple outcomes per subject followed longitudinally, and multiple developmental toxicity outcomes measured on each rat pup of a litter. As a demonstration of the application of the two-level clustered binary logistic model, somewhat artificial data based on some real (but without permission to use fully publicly) were used by the statistics research group at Arizona State University, W.P. Carey School of Business. In fact the data are void of its variables' labels to protect the sensitivity and the lack of permission to use with identifiers. These data have inherent and known nested correlation structure. The data with generic names are given in Table 1, where

group labels the first stage clusters, n_{ij} denotes the number of observations in each second stage cluster, z_{ij} is the number of positive responses in each second stage cluster, and the predictor is x .

We fitted a logistic regression model $\text{logit}(p) = \alpha + \beta x$, where $p = P[Y_{ij} = 1]$ is the probability of a single positive response. We first fit the model of [1], described in Section 2.2 using probability mass function (1) and (3), ignoring the first stage clustering of *group*. The p-value for the one-sided parameter test $H_0: \beta = 0$ vs $H_a: \beta > 0$ is $p_1 = 0.033$. The model described in Section 3.1 was also fitted using probability mass function (10), where the *group* level of correlation is accounted for in the model. This was fitted using an R program <http://www.public.asu.edu/~jeffreyw> that allows an exhaustive search through all possible outcomes to compute the conditional distribution. The p-value for the one-sided parameter test is now $p_2 = 0.095$. The test statistic for the model with only first stage clustering is clearly inflated by ignoring the additional level of clustering built into the data.

The data were analyzed using a C++ program <http://www.public.asu.edu/~jeffreyw> which can handle larger data sets than the R program. In addition to being faster and more efficient with memory, the C++ program uses feasibility checks, similar to those described in [9], to avoid computing the entire distribution. These feasibility checks allow the use of reasonably large data sets when the total number of successes remains small.

The data in Table 2 were analyzed similarly to the data in Table 1. The model described in Section 3.1 was fitted using probability mass function (10), where the group level of correlation is accounted for in the model. The C++ program computed the one-sided p-value of 0.0085 for the test of $H_0: \beta = 0$ vs $H_a: \beta > 0$. The R program was unable to complete this calculation.

For cases in which the data set may be very small, as in Table 1, an exhaustive program will suffice. But for a situation as in Table 2 (“count=8”), where the probability of a success is very low, the more efficient program is necessary. In fact, even with reasonably large sample sizes, the asymptotic theory will not apply well. In this case, for Table 2 the asymptotic generalized estimating equations approach fails to produce positive definite covariance structures for parameter estimates, but our exact method is able to perform a one-sided hypothesis test.

The program run with data with 6 positive responses (Table 2, “count=6”) took about 4 seconds to produce a p-value of 0.0404827. The program run with data with 7 positive responses (Table 2, “count=7”) took about 22 seconds to produce a p-value of 0.0194951. The original data set with 8 positive responses (Table 2, “count=8”) took about one minute and 48 seconds (108 seconds) to produce a p-value of 0.0085. The program run with data with 9 positive responses (Table 2, “count=9”) took about 9 minutes and 1 second (541 seconds) to produce a p-value of 0.00523589. However, when we tried 10 or more responses the program took at least 4 hours before we stopped it. The program was run on an Inspiron 531 machine with a 2.80 GHz AMD Athlon 64 X2 Dual Core Processor (5600+) with approximately 2.0 GB (1982 MB) RAM.

5 CONCLUSIONS

The use of large sample theory for fitting logistic models is well documented both for correlated and for uncorrelated observations. However the use of exact analysis is not as well known. This paper presented exact models for correlated data through different levels of clustering. We concentrated on a one-sided test, $H_0: \beta = 0$ vs $H_a: \beta > 0$. To conduct the test $H_0: \beta = 0$ vs $H_a: \beta < 0$, we could either perform the given test using $Y'_{ij} = 1 - Y_{ij}$, or we

could equivalently use a p-value obtained by summing over $t < t_{obs}$ instead of $t > t_{obs}$. To conduct a two-sided test of $H_0: \beta = 0$ vs $H_a: \beta \neq 0$ given that the conditional distribution of t is not symmetric, we should find $t_{\alpha/2}$ and $t_{1-\alpha/2}$ where, for $0 < p < 1$, t_p is the smallest value of t satisfying $Pr(t < t_p | H_0, x, s) = p$. We reject the null hypothesis if the observed value of t , t_{obs} is either greater than $t_{1-\alpha/2}$ or less than

The use of C++ to analyze these correlated data with different levels of clustering will certainly enhance the use of logistic models with sparse data.

Acknowledgments

The authors wish to thank two referees and the editor for helping to add clarity to the issues addressed in this paper. This research was supported in part by a National Institute of Health Alzheimer's Grant in the Arizona Alzheimer Center at Arizona State University.

Contract/grant sponsor: National Institute of Health Alzheimer's Grant

REFERENCES

1. Rosner B. Multivariate methods for clustered binary data with more than one level of nesting. *Journal of American Statistical Association*. 1989; 84(406):373–380.
2. Have TR, Kunselman AR, Tran L. A comparison of mixed effects logistic regression models for binary response data with two nested levels of clustering. *Statist. Med*. 1999; 18:947–960.
3. Corcoran CL, Ryan PS, Mehta C, Patel N, Monenbergs G. An exact trend test for correlated binary data. *Biometrika*. 2001; 57:941–948.
4. Gibbons RD, Hedeker DH. Random Effects Probit and Logistic Regression Models for Three-Level Data. *Biometrics*. 1997; 53:1527–1537. [PubMed: 9423267]
5. Cox, DR. *Analysis of Binary Data*. Chapman and Hall; New York: 1970.
6. Cox, DR.; Snell, EG. *Analysis of Binary Data*. 2nd edition. Chapman and Hall; New York: 1989.
7. Collett, D. *Modelling Binary Data*. Chapman and Hall; New York: 1991.
8. Tritchler D. An algorithm for exact logistic regression. *Journal of the American Statistical Association*. 1984; 79(300):709–711.
9. Hirji KF, Mehta CR, Patel NR. Computing distributions for exact logistic regression. *Journal of the American Statistical Association*. 1987; 82(400):1110–1117.
10. Derr, RE. Performing exact logistic regression with the SAS system. *Proceedings of the Twenty-fifth Annual SAS Users Group International Conference*; 2000.
11. Mehta CR, Patel NR. Exact logistic regression: theory and examples. *Statistics in Medicine*. 1995; 14:2143–2160. [PubMed: 8552893]
12. Connolly MA, Liang K. Conditional logistic regression models for correlated binary data. *Biometrika*. 1988; 75(3):501–506.

Table 1

Group	Number n_{ij}	Count z_{ij}	x
1	4	2	2
1	2	0	4
1	6	2	3
2	5	0	0
2	5	0	1
3	3	1	7
3	7	6	6
3	4	2	7
3	2	1	5

Table 2

Group	Number n_{ij}	x	“Count=6” z_{ij}	“Count=7” z_{ij}	“Count=8” z_{ij}	“Count=9” z_{ij}
1	21	0	0	0	0	0
1	20	0	0	0	0	0
1	20	0	0	0	0	0
1	19	0	0	0	0	0
1	17	1	0	0	0	0
1	15	1	0	0	0	0
1	19	2	1	1	1	1
2	21	0	0	0	0	0
2	18	1	0	0	0	0
2	20	2	1	1	1	1
2	15	2	1	1	1	1
3	17	0	0	0	0	0
3	19	1	0	0	0	0
3	18	1	0	0	0	0
3	19	2	1	1	1	1
3	17	2	0	0	0	0
4	20	0	0	0	0	0
4	20	1	1	1	1	1
4	20	2	0	1	1	1
5	14	0	0	0	0	0
5	12	0	0	0	0	0
5	14	1	1	1	1	1
5	13	1	0	0	0	0
5	15	2	0	0	0	1
5	14	2	0	0	1	1