LOGISTIC REGRESSION MODELS FOR ANALYSIS OF MULTISTAGE SURVEY DATA

Hongjian Yu, UCLA Center for Health Policy Research
William G. Cumberland, Department of Biostatistics, University of California, Los Angeles
Hongjian Yu, CHPR/CHS 10833 Le Conte Ave., Los Angeles, CA 90095

Key words: Logistic regression, Survey, Random effects, Generalized estimating equations.

1. Introduction

Logistic regression models of the form

$$\text{logit}[P(y_{ij} = 1)] = \mu + \beta X_{ij}$$

are extensively used in analyzing sample survey data to study the relationship between a binary response and a group of independent variables. Due to cost and efficiency considerations, stratified multistage samples are the norm. However, these samples, while efficient for estimation of the descriptive population quantities, pose challenges for model-based statistical inference. This sampling scheme often introduces multilevel correlation among the observations that can have implications for model parameter estimates. For multistage clustered samples, the dependence among observations often comes from several levels. Thus, drawing appropriate inferences from survey data may require complicated modeling techniques and very often, the computation required for this is very time consuming.

This paper is focused on model-based analysis for binary data with a structure similar to that of the National Health Interview Survey. It proposes a logistic regression model which fits the between-cluster variation with random effects. The model also takes into account the correlation among small groups of observations within each cluster. An algorithm is proposed that makes the computation feasible for the mixed logistic regression model on large survey data. Generalized estimating equations (Liang and Zeger, 1986) are used in the estimation procedure to accommodate the correlation among the observations within small groups, avoiding problems associated with the use of random effects to model correlations among large numbers of small groups. An adjustment is applied to eliminate the bias in estimation of fixed effects that exists in some procedures for random effects logistic regression noted by Rodriguez and Goldman, 1995.

2 National Health Interview Survey and Modeling Considerations

The current National Health Interview Survey (NHIS) is taken annually by the National Center for Health Statistics using a multistage sampling scheme. About 200 primary sampling units (PSU's) are selected from approximately 1900 geographically defined PSU's (each consisting of a single county or a group of contiguous counties) which collectively cover all 50 states in the United States and the District of Columbia. These PSU's are stratified using socioeconomic and demographic variables and are selected with probability proportional to population size within a stratum. Within each PSU, groups of households or neighborhoods are formed and subsampled. Once a neighborhood is selected, some or all the households in this neighborhood are selected for interview, depending on the size of the neighborhood. For the sampled households, information on all the members of the household is recorded. Each year a new sample consisting of about 50,000 households containing approximately 120,000 individuals is taken.

The sampling structure used by NHIS introduces multilevel correlation among the observations. A reasonable model must first consider the correlation within families. For example, in measuring the risk of disease, since household members are usually genetically related and live in close proximity, high correlation among the outcomes from the same household is possible. Other examples of correlated measures for household members include social and economic status, education, personal income and whether or not a person has health insurance coverage. The correlation is mainly caused by interdependence among the family members, economically, socially, or biologically.

Another correlation the model should consider is at the PSU level and is the result of an area effect, sometimes referred to as an "ecological" effect. The same effect applies to all the individuals in a PSU but varies from PSU to PSU. For example, if air pollution is a major risk factor for a given respiratory disease, higher rates for occurrence of the disease would be expected in highly polluted PSUs than in PSUs with lower pollution levels. The effect of pollution and its interaction with other effects on the risk of the disease is similar for people living in the same PSU but different from PSU to PSU. Other examples include the effect of the economic environment or government policy on individuals' social or economic status within different political boundaries.

3. Problems Associated with Estimation In Random Effects Logistic Regression.

One way to model the NHIS data is to use multilevel Random effects models with random effects specified at both family level and PSU level. However, estimation for random effects models is often complicated. Most of the estimation procedures are quite complex and calculations are time consuming, so that evaluation of their properties even through simulation is difficult. Properties associated with some of estimation procedures for logistic regression have become known only recently.

Rodriguez and Goldman (1995) evaluated two software packages that are available for fitting multilevel models to binary data using a Monte Carlo study designed to represent the structure of a data set used in an analysis of health care utilization in Guatemala. In their study, two levels of random effects were included to model family and community clustering, in addition to several fixed effects. The results revealed substantial biases in the estimates of the fixed effects and/or variance components whenever

1. the random effects are sufficiently large to be interesting,
2. the number of observations within a given level of clustering (e.g. family) is small.

Rodriguez and Goldman showed that the multilevel estimates from the packages they evaluated are virtually the same as those obtained by using ordinary logistic regression models that ignore the hierarchical structure of the data. This means that in estimating random effects with the logistic regression model, the improvement through implementing a very complicated algorithm by these packages is minimal compared to simply applying ordinary logistic regression, which ignores all the complex model specifications. Neuhaus and Jewell (1990) showed that given the random effects model is correct, estimates from procedures that ignore these random effects, such as ordinary logistic regression, are biased toward zero, which is confirmed by Rodriguez and Goldman's study. In another recent evaluation of approximate methods of inference for generalized linear model with random effects, Breslow and Clayton (1993) also obtained similar results.

In the conclusion of their study, Rodriguez and Goldman stated that their simulation results leave open the question of whether random effects in the binary response model can ever be estimated at acceptable levels of bias and precision when the average size of clusters is modest. The sample size problem is unavoidable in a broad range of social, demographic and epidemiological studies when the lowest level of clustering is at the family level. Their findings highlight the need for alternative estimation procedures to handle multilevel models with binary response.

4. Fitting Multilevel Clustered Data: Hierarchical Logistic Regression Model.

The model proposed in this paper is a combination of the population average (PA) model and the random effects model. The population average model is used to take into account the correlation among observations from family members; the random effects model is employed to fit the between-PSU variation. This combination accommodates the multilevel correlations among the observations while avoiding the problems of estimation in random effects models when the sample sizes within each family are small.

In the hierarchical model studied by Wong and Mason (1985), covariates were divided into two groups: those at the micro level (individual level) and those at the macro level (PSUs level). The examples for the micro level covariates are age, sex, education, etc. The examples for macro level covariates are population size of PSU, pollution level of the PSU, etc. This model combines information at the macro (PSU) level with that at the micro level as follows: for y_{ijk}, the *k*th observation from the *j*th family in PSU *i* with micro level covariates X_{ijk}, the model is

$$\mathrm{logit}[p(y_{ijk} = 1)] = X_{ijk}\beta_i,$$

where β_i, which is $p \times 1$, is the vector of regression coefficients for PSU *i* and

$$\beta_i = Z_i\beta + \upsilon_i$$

where Z_i is a $p \times L$ macro level design matrix, and β is an overall fixed effect of order $L \times 1$. The error vectors υ_i are independent and identically distributed with mean zero and a $p \times p$ variance-covariance matrix Γ.

This random effects model can be extended to accommodate multilevel clustering of the data. However, in the NHIS, families define the lowest level of clustering. As discussed in section 3, due to small family sizes, specifying another set of random effects at the family level would not allow for satisfactory estimation. This problem can be avoided by using the PA model to fit the family correlations. The GEE estimation procedure for the PA model combines the samples of all families in a large cluster. So at the micro level, instead of introducing another set of random effects, the PA model is fitted to take the correlation into account. Accordingly, the final model for the vector of observations of family *j* in PSU *i* is

$$\mathrm{logit}[P(y_{ij} = 1)] = X_{ij}\beta_i$$

together with correlation matrices which specify the correlation structures among the observations within families. The distribution for β_i, the logistic regression coefficient for PSU *i*, is

$$\beta_i \sim MVN(Z_i\beta,\Gamma).$$

Since the full likelihood function for the GEE of logistic regression is not fully specified, the maximum likelihood estimation technique cannot be directly applied to this model. Korn and Whittemore (1979) introduced a two-step method that first uses a separate logistic regression of the binary response against micro level covariates in each PSU. These individual parameters are then combined to yield summary estimates for the overall effects of the covariates. This procedure provides a way to apply GEE to get estimates in each PSU taking into account the within family correlation and then to combine them to get second level estimates using currently available methods based on multivariate normal theory.

5. Estimation and Calculation

This section explains the estimation procedure in detail. Assuming normality for the random effects, the model defined in section 4 becomes

$$\text{logit}[p(y_{ijk}=1)] = X_{ijk}\beta_i$$

where $\beta_i = Z_i\beta + \upsilon_i$, $\upsilon_i \sim MVN(0,\,\Gamma)$ and Γ is a variance-covariance matrix.

Following the strategy proposed by Korn and Whittemore (1979), the estimation procedure for the fixed effect β splits into two steps:

Step 1. Conditioning on β_i, derive the estimated coefficient $\hat{\beta}_i$ and the variance $\hat{\eta}_i$ for logistic regression from the observations in each PSU by directly applying Liang and Zeger (1986).

Step 2. Assuming $\hat{\beta}_i \sim MVN(\beta_i,\eta_i)$, we have

$$\hat{\beta}_i \sim MVN(Z_i\beta,\eta_i+\Gamma) \qquad (1)$$

where η_i can be replaced by $\hat{\eta}_i$.

Racine-Poon (1985) gave a Bayesian approach which is among the best of its class (Gelfand et al. 1990) for the calculation in step 2 when Z_i equals the identity matrix. Following Racine-Poon, the prior distributions for β and Γ^{-1} are assumed, respectively, vague and Wishart with degrees of freedom ρ and matrix Q. Racine-Poon pointed out that to represent vague knowledge about Γ^{-1}, ρ should be chosen as small as possible (i.e., $\rho = p$ the number of covariates). Except for the case with very few groups, the choice of Q has little effect on the result. In this section, Racine-Poon's result is extended to the situation when Z_i is not the identity matrix.

To obtain the posterior distribution of β and Γ^{-1} given $\hat{\beta}_i$ for $i = 1, \ldots, I$, the joint density of $\hat{\beta}_i, \beta_i, \beta$, and Γ^{-1} has to be evaluated first. It is proportional to

$$\begin{aligned}&\left(\prod_{i=1}^{I}|\eta_i|^{-1/2}\right)\exp\left[-\tfrac{1}{2}\sum_{i=1}^{I}(\hat{\beta}_i-\beta_i)^T\eta_i^{-1}(\hat{\beta}_i-\beta_i)\right]\\ &\quad\times|\Gamma|^{-I/2}\exp\left[-\tfrac{1}{2}\sum_{i=1}^{I}(\beta_i-Z_i\beta)^T\Gamma^{-1}(\beta_i-Z_i\beta)\right] \qquad (2)\\ &\quad\times|\Gamma|^{-(\rho-p-1)/2}\exp\left[-\tfrac{1}{2}tr\Gamma^{-1}\cdot Q\right]\end{aligned}$$

This cannot be solved analytically. Following Racine-Poon's EM-type algorithm, first the conditional expectations of the $\hat{\beta}_i's$ and β are derived. Then they are plugged into (2) and the equation is maximized over Γ.

Following a theorem due to Lindley and Smith (1972) and assuming that knowledge of β is weak, the posterior of β given $\hat{\beta}_i$, $i = 1, \ldots, I$, and Γ is *L*-variate normal with mean β^* and covariance matrix *D*:

$$D = \left[\sum_{i=1}^{I} Z_i^T(\eta_i+\Gamma)^{-1}Z_i\right]^{-1}.$$

$$\beta^* = D\sum_{i=1}^{I}[Z_i^T(\eta_i+\Gamma)^{-1}\hat{\beta}_i]\cdot$$

Using a similar strategy, one can derive the posterior density of β_i given $\hat{\beta}_i$, η_i, β and Γ is normal with mean β_i^* and covariance matrix D_i:

$$D_i = (\eta_i^{-1}+\Gamma^{-1})^{-1}$$

$$\beta_i^* = D_i(\eta_i^{-1}\hat{\beta}_i+\Gamma^{-1}Z_i\beta)\,.$$

Applying the EM-type algorithm is then straightforward. At the *l*th iteration, let $\beta^{(l-1)}$, $\beta_i^{(l-1)}$, $\Gamma^{(l-1)}$ be the approximations of β, β_i, and Γ from the (*l*-1)th step, respectively.

E-step: Conditioning on $\Gamma^{(l-1)}$ the posterior expectation for β is given by

$$\beta^{(l)} = [D^{(l)}]\sum_{i=1}^{I}[Z_i^T(\eta_i+\Gamma^{(l-1)})^{-1}]\hat{\beta}_i \qquad (3)$$

where

$$D^{(l)} = \left[\sum_{i=1}^{I} Z_i^T(\eta_i+\Gamma^{(l-1)})^{-1}Z_i\right]^{-1},$$

By conditioning on $\beta = \beta^{(l)}$ and $\Gamma = \Gamma^{(l-1)}$ the Bayes estimates for $\beta_i's$, $i = 1,\cdots,I$ are given by

$$\beta_i^{(l)} = (\eta_i^{-1} + \Gamma^{(l-1)^{-1}})^{-1}(\eta_i^{-1}\hat{\beta}_i + \Gamma^{(l-1)^{-1}} Z_i \beta^{(l)}).$$

***M*-step:** Conditioning on $\beta_i = \beta_i^{(l)}$, the log transformation of (2) is

$$c - \tfrac{1}{2}(I + \rho - p - 1)\log|\Gamma|$$
$$- \tfrac{1}{2}\left\{ tr\Gamma^{-1}\left[Q + \sum_{i=1}^{I} (\beta_i^{(l)} - Z_i\beta^{(l)})(\beta_i^{(l)} - Z_i\beta^{(l)})^T \right]\right\}.$$

The conditional posterior model can be obtained as

$$\Gamma^{(l)} = \frac{[(Q + \sum (\beta_i^{(l)} - Z_i\beta^{(l)})(\beta_i^{(l)} - Z_i\beta^{(l)})^T]}{(I + \rho - p - 1)}.$$

The ***E*** and ***M*** steps are repeated until Γ^{-1} converges. Reasonable starting values of β and Γ are

$$\beta^{(0)} = (A^T A)^{-1} A^T y$$

$$\Gamma^{(0)} = \frac{[Q + \sum (\hat{\beta}_i - Z_i\beta^{(0)})(\hat{\beta}_i - Z_i\beta^{(0)})^T]}{(I + \rho - p - 1)}$$

where A = $\{ Z_1^T, \cdots, Z_I^T \}^T$.

The key assumption is the normality of the estimates $\hat{\beta}_i$'s from each cluster. Hence, the number of observations within each cluster should be sufficiently large to ensure (asymptotically) the normality assumption. In this study, the clusters are PSUs which usually have sample sizes much larger than most longitudinal studies. Liang and Zeger (1986) showed that the GEE estimators $\hat{\beta}_i$ are asymptotically normal so the normality assumptions needed for the algorithm are appropriate for our application.

Intuitively, this two-step algorithm derives the overall estimate by combining the estimates from each individual PSU. The contributions of these individual estimates towards the overall estimate are weighted by their precision matrices (inverse of the variance). If Z_i' s are identity matrices, the overall estimate $\hat{\beta}$ is the weighted average. As Stiratelli, Laird and Ware (1984) stated, this two-step algorithm has much to recommend it. It not only greatly simplifies the estimation procedure but can also accommodate the situation when directly applying the maximum likelihood estimation is not possible.

Preliminary simulation showed that without any adjustment, the two-step algorithm would also give biased estimates for the fixed effects. In the two-step method, the bias mainly comes from the fact that the variance of estimates of logistic regression parameters is directly related to the estimates themselves. To remedy this problem, in equation (3), η_i is replanced by

$$C_i = \frac{\sum n_l^2 \hat{\eta}_l}{\sum n_l} \Big/ n_i ,$$

where n_i is the sample size in PUS *i*, and l = 1, …, *I*.

This adjustment connects the weights of $\hat{\beta}_i$ only to the sample size in a cluster and dissociate them from the estimates themselves.

6. Simulation Study

6.1. Generating Correlated Binary Data.

It is fairly straightforward to generate correlated binary variables using a random effects logistic model. It is more difficult to generate such variables following a population average model. In the latter case, the difficulty arises because the marginal distribution of these variables has to be kept in logistic form. In this section a method is proposed to generate correlated binary responses for given individual covariates, keeping the marginal in logistic form.

Let $i = 1, \cdots I$ be index for PSUs, $j = 1, \cdots J_i$ be index for families, and $k = 1, \cdots, K_{ij}$ be index for observations from family *j*. For this very simple situation where only family correlation considered, an outline of the procedure is as follows:

(1) Calculate the corresponding probability based on logistic equation. For given covariates x_{jk} and β,

$$p_{jk} = \frac{\exp(x_{jk}\beta)}{1 + \exp(x_{jk}\beta)} \qquad (4)$$

(2) Generate correlated normal variables. For given *j*,

$$U_{jk} = \mu_j + \varepsilon_{jk},$$

where μ_j and $\mu_{j'}$ are independent for $j \neq j'$ and

$$\mu_j \sim N(0, \sigma_\mu^2),$$

where ε_{jk}'s are independent of each other and of the μ_i's and

$$\varepsilon_{jk} \sim N(0,\ 1)$$

for $k = 1, \cdots, K_{ij}$. So

$$\mathrm{var}(U_{jk}) = \sigma_\mu^2 + 1,$$
$$\mathrm{cov}(U_{jk}, U_{jk'}) = \sigma_\mu^2$$

and

$$\mathrm{cov}(U_{jk}, U_{j'k'}) = 0 \;\text{ for }\; j \neq j'$$

Derive correlated normally distributed variable with mean zero and variance one

$$V_{jk} = U_{jk} \Big/ \sqrt{\sigma_\mu^2 + 1} \,.$$

The correlation between V_{jk} and $V_{jk'}$ is

$$\gamma = \sqrt{\frac{\sigma_\mu^2}{1+\sigma_\mu^2}} \,.$$

(3) Generate correlated binary data. Set

$$Y_{jk} = I\left[\Phi(V_{jk}) < p_{jk}\right]$$

where Φ is the CDF of the standard normal distribution and I is an indicator function. The resulting random variables Y_{jk} and $Y_{jk'}$ are binary and correlated and Y_{jk} is independent of $Y_{j'k'}$ if $j \neq j'$. The probability that $Y_{jk}=1$ is determined by equation (4) through the p_{jk}'s. So the marginal distribution of the binary random variables follows an ordinary logistic regression. The dependence of the correlation between Y_{jk} and $Y_{jk'}$ on their covariates means that γ is not the direct measure of this correlation. But this correlation is still positively related to γ . So γ can be defined as the ***index*** of correlation, or ***IC*** in short, between Y_{jk} and $Y_{jk'}$. This dependence of the correlation between responses on their covariates is small as long as the difference of the between the expected value of the binary responses is not too large. To derive variables with between-PSU variation, β_i is first generated using multivariate normal distribution and then plugged in equation (4).

6.2. Simulation Results

Since the magnitude of the between-PSU variance plays an important role in the consistency of the estimates of fixed effects (as reported by Rodriguez and Goldman, 1995), two between-PSU variances are chosen: Γ_1, "small", the standard deviation for the slope being about 40% of the magnitude of the parameter, and Γ_2, "large", the standard deviation of the slope being about 100% of the magnitude of the parameter.

Simulation results showed that when the between PSU variance is small these is no bias in the estimates given by the two-step algorithm. In the estimates by ordinary logistic regression which ignores the hierarchical structure of the data, these is tendency of bias toward zero, although the magnitude of the bias is small. However when the between PSU variance is large, while the estimates from the two-step algorithm still remain unbiased, there is a big bias in the estimates from the ordinary logistic regression applied to the hierarchical data and the bias is toward zero. This illustrates that the adjustment proposed does alleviate the bias problem which plaques some of the estimating procedures for random effect logistic models.

As a comparison of the efficiency for the estimation, the empirical variances of the slope estimates from the two-step algorithm with GEE and that with ordinary logistic regression are compared. The algorithm with GEE is more efficient when the within-PSU sample size is small, but this gain of the efficiency is less when the within-PSU sample size increases. The main reason for this is that the variance of the estimates for the fixed effect β is a function of two parts: variance due to within-PSU variance and variance due to between-PSU variance Γ. GEE only affects the first one, the within-PSU variance, and has no effect on the second one. When the within-PSU variance dominates the between-PSU variance, using GEE will affect the overall variance of the estimates for β. As the within-PSU sample size increases, the within-PSU variance decreases, so the dominance of the within-PSU variance over between-PSU variance diminishes as a result of this. Therefore the gain of efficiency through GEE also vanishes.

Simulation on two step algorithm with the PSU level covariates Z_i's other than identity was conducted, Again the estimates showed no signs of any bias.

7. Conclusion and Discussion

Model-based logistic regression analysis of survey data has been difficult because of two things: taking into account the hierarchical or nested structure of the data, and the nonlinear link function. Statistical techniques required to analyze data with this kind of structure have been developed relatively recently. They are often very complicated and beyond the easy grasp of researchers outside the statistical field. This makes the descriptive approach, which is usually based on very simple model specifications more appealing to many people. However, since observations from the same cluster tend to be more alike than those from different clusters, the basic assumptions, such as independence between observations, for many simple model specifications are likely to be violated. As a consequence, these " shortcut" analyses may result in biased estimates of the parameters. This problem is especially serious when the link function between the outcome variable and the explanatory variables is nonlinear.

The model which combines the random effects model and population average model and the two-step algorithm developed in this paper provide a new approach to logistic regression analysis of multistage stage survey data. This approach provides a way to model the multilevel correlation among the observations and avoids problem of small sample size at low levels of clustering, which prevents satisfactory estimation by applying the random effects model at this low level. Applying the GEE technique not only takes into account the correlation among the observations in small groups but also improves the efficiency of the estimation. The adjustment developed in the two-step algorithm eliminates the bias in estimating of the fixed effects that plagues in some currently available estimating procedures in analysis of hierarchical logistic regression. Also by splitting the estimation into two relatively simple steps, the two-step algorithm greatly increases the computation efficiency.

Finally, model-based analysis eliminates some need for some sample design information such as the inclusion probability of the sample, which is necessary for a descriptive analysis. Thus it greatly simplifies the analysis in this respect. Sample design information may not always be available or may not be accurate. Even when the information is available, fully understanding it and using it appropriately in the analysis might not be feasible for many researchers without an extensive statistical background.

References

Gelfand, A. E., Hills, S. E., Racine-Poor, A. and Smith, A. F. M. (1990), "Illustration of Baysian Inference in Normal Data Models Using Gibbs Sampling," *Journal of the American Statistical Association,* **85**, 972-985.

Korn, E. L. and Whittemore, A. S. (1979), "Methods for Analyzing Panel Studies of Acute Health Effects of Air Pollution," *Biometrics* **35**, 795-802.

Liang, K. and Zeger, S. (1986), "Longitudinal Data Analysis Generalized Linear Models," *Biometrka.* **73** 13-22.

Neuhaus, J. and Jewell, N. (1990), "Some Comments on Rosner's Multiple Logistic Model for Clustered Data," *Biometrics* **46**, 523-534.

Rodriguez, G. and Goldman, N. (1995), "An Assessment of Estimation Procedures for Multilevel Models with Binary Responses," *Journal of the Royal Statistical Society,* **158**, PART 1, 73-89.

Racine-Poon, A. (1986), "A Bayesian Approach to Nonlinear Random Effects Models," *Biometrics,* **41**, 1015-1023.

Stiratelli, R., Laird, N. and Ware, J. (1984), "Random-Effects Model for Serial Observations with Binary Response," *Biometrics,* **40**, 961-971

Wong, G. and Mason, W. M. (1985), "The Hierarchical Logistic Regression Model for Multilevel Analysis," *Journal of the American Statistical Association,* **80**, 513-524,