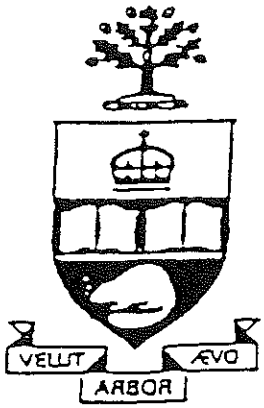

A Hybrid Logistic Model for Case-Control Studies

by

Tuhao Chen
Department of Mathematics and Statistics
Bowling Green State University

and

Fred M. Hoppe
Department of Mathematics and Statistics
McMaster University

and

Satish Iyengar
Department of Statistics
University of Pittsburgh

and

David Brent
Department of Psychiatry
University of Pittsburgh

Technical Report No. 0207 December 10, 2002

TECHNICAL REPORT SERIES

University of Toronto

Department of Statistics

A Hybrid Logistic Model for Case-Control Studies

Tuhao Chen* Fred M. Hoppe† Satish Iyengar‡ David Brent§

Abstract

For logistic regression in case-control studies, when risk factors associated with the outcome are exceedingly rare in the control group, the estimation of parameters in the model becomes difficult. In this paper we propose a two-stage hybrid method to achieve this. In the first stage we model the risk due to the rare factor, and in the second stage we model the residual risk due to the other factors using standard logistic model.

Keywords: Logistic regression, maximum likelihood estimation, case-control study, child and adolescent suicide

1 Introduction

Logistic regression is a standard method for the analysis of data from case-control studies when multiple risk factors are assessed. With a rare disease, for which the case-control approach is often the design of choice, it is also likely that the risk factors associated with the disease are exceedingly infrequent among corresponding matched controls. Such an

*Department of Mathematics and Statistics, Bowling Green State University

†Department of Mathematics and Statistics, McMaster University

‡Department of Statistics, University of Pittsburgh

§Department of Psychiatry, University of Pittsburgh

infrequency makes the estimation of risk for these factors difficult, even untractable. In practice, investigators normally steer clear of such troublesome risk factors by considering only the other risk factors in the model (see, for example, Shaffer et. al. [8]); however, such an approach spreads the contribution of the troublesome risk factor(s) among the remaining factors in the model and may consequently result in an overestimate of the odds ratios for the latter. In this paper we propose a hybrid procedure: first, we adjust for the troublesome risk factor, and then we model the residual using logistic regression. We illustrate with data from a large case-control study of adolescent suicide.

Let Y_1, ..., Y_n be a family of mutually independent {0,1} valued indicator random variables representing the cases ($Y = 1$) or controls ($Y = 0$) for n individuals in a case-control study. Recall that the controls are representative of the many individuals who failed to develop the disease [5]. We denote the explanatory or risk factors by (z, x^T) where z represents a rare risk factor and $x^T = (x_1, \cdots, x_p)$ the other risk factors. These take the values $(z_i, x_{i1}, \cdots, x_{ip})$ on subject i. For purposes of this paper z is univariate.

Let $p(y|z, x)$ denote the distribution of Y for an individual with risk vector (z, x), that is the conditional probability $P[Y = y|Z = z, x]$. In a prospective study (z, x) are regarded as fixed variables and the occurrence of disease is a random event. In such a study a cohort of disease-free individuals would be followed through a time period and $P[Y = 1|Z = z, x]$ would be estimated directly [2].

In contrast, in a case-control study separate samples of cases and controls are taken, meaning that the disease status is considered fixed and the risk factors are random conditional on disease status [2]. This changes the focus of the likelihood function because the quantity of primary interest is then not $p(y|z, x)$ but rather $p(z, x|y)$. For instance for $y = 1$ this represents the conditional probability $P[Z = z, x|Y = 1]$.

Logistic models were originally used in prospective studies. Traditionally, all risk

factors are treated equally, so there are no rare factors z and the linear logistic model is then expressed as

$$\pi(x) = P(Y = 1|x) = \frac{e^{\beta_0+\beta^T x}}{1 + e^{\beta_0+\beta^T x}} \tag{1}$$

for the probability of a case given that the subject has covariates x. Here β_0 and the $1 \times p$ row vector $\beta = (\beta_1, \cdots, \beta_p)$ are unknown parameters to be estimated by using the data.

The method of estimation of β remains the same for both case-control studies and cohort studies (see for example, [3], [6], [7]). In fact, as McCullogh and Nelder [6, page 111] point out "an important property of the logistic function is that it applies for both cohort (prospective) studies and case-control (retrospective) studies". Following [6], we can introduce a dummy variable S to denote whether an individual is sampled ($S = 1$) or not ($S = 0$). Denote the sampling proportions by

$$\tau_i = P(S = 1|x, Y = i) = P(S = 1|Y = i) \quad \text{for} \quad i = 0, 1.$$

Note that the selection process in the case and control groups is explicitly assumed independent of the covariates x as is customary. As in [2] we use a superscript * to represent a distribution conditional on being observed in the sample. Then for a case-control study the logistic model becomes

$$\pi^*(x) = P(Y = 1|x, S = 1) = \frac{P(Y = 1|x)P(S = 1|x, Y = 1)}{P(Y = 1|x)P(S = 1|x, Y = 1) + P(Y = 0|x)P(S = 1|x, Y = 0)}.$$

Substituting for τ_0 and τ_1, and recalling that the selection process is independent of covariates we arrive at

$$\pi^*(x) = P(Y = 1|x) = \frac{e^{\beta_0^*+\beta^T x}}{1 + e^{\beta_0^*+\beta^T x}} \tag{1b}$$

which is the same as model (1) except that the intercept is now $\beta_0^* = \beta_0 + ln(\frac{\tau_1}{\tau_0})$. Furthermore

$$p(x|Y = 1, S = 1) = \frac{\exp(\beta_0^* + \beta^T x)}{1 + \exp(\beta_0^* + \beta^T x)} \frac{p(x|S = 1)}{P(Y = 1|S = 1)}$$

where $p(x)$ is the marginal distribution of the covariates x in the population.

Models (1) and (1b) are widely used for binary data. However, we found them to be unsatisfactory for a study of risk factors for adolescent suicide [1]. The problematic risk factor in the study is the factor — past attempt of suicide (PAS): namely 62% of female suicides (the cases) had a past suicide attempt, but none had in the control group of young females (Table 1). Backward and forward stepwise logistic regression removed the risk factor PAS that is clinically regarded as important, while the regression model with all risk factors did not converge. If PAS is included in the model as a covariate then the confidence intervals for the odds ratios behave poorly because the corresponding variances become infinite. Moreover substance abuse is eliminated as a factor.

TABLE 1. Female Adolescent Suicides and Controls by PAS

	Case	Control	Total
PAS	13	0	13
Non-PAS	8	40	48
Total	21	40	61

Clearly the data set here has a problem of complete separation as described in [3] and neither exact inference [4] nor approximation methods could provide plausible estimates. Clinically PAS is known to be a strong predictor of child and adolescent suicide; therefore we had no theoretical reason to exclude it from the final model. On the other hand, it is unsuitable to develop a model involving only PAS because the purpose of the study was to identify a set of key risk factors that might aid in the prevention of adolescent suicide.

The same problem arose in the large case-control study of adolescent suicide by Shaffer et al. [8], in which study none of their female control subjects had made a previous suicide attempt. They reported this difficulty in their paper, noted that PAS is an important predictor, and then developed a logistic model based on risk factors other than PAS. This

way of handling the troublesome risk factor simplifies the data analysis but, as depicted in Table 5 of [1], it may spread out the risk due to PAS over the other risk factors thereby overestimating the risks associated with these other factors.

In this paper we propose a method that allows us to model factors such as PAS, which tend to be rare in the control group. In Section 2, we develop a statistical model for the case of one unusual risk factor. The newly developed method is applied to the adolescent suicide study and reported in Section 3.

2 The Model

As in section 1 let $\pi(x) = P(Y = 1|x)$ be the risk of suicides in the population, intrepreting suicide as disease, where now $x = (x_1, ..., x_p)^T$ is the vector composed of all risk factors except PAS, and let $Z = 1$ ($Z = 0$) indicate the presence (absence) of PAS. Instead of ignoring PAS, we consider a model in the following way: Since PAS increases the risk of suicide, we model the part of its contribution to π (between 0 and π) first, and then model the residual risk as a function of the remaining risk factors using logistic regression. Note that PAS is not present in the control group, and thus the associated risk in the control group can be explained by only the risk factors x.

Since the study is a case-control study, in the sequel, we state the setting for case-control studies, but keep in mind that the interpretation of odd ratios for risk factors x_1, ..., x_p remain the same as in cohort study. (See, for example, [3])

Let α be the proportion of PAS in the case group

$$P(Z = 1|Y = 1, x, S = 1) = \alpha$$

α depends on x. In the case where PAS = 0 , we have

$$P(Z = 0|Y = 1, x, S = 1) = 1 - \alpha.$$

We propose the following model for the joint distribution of risk factors $p(x, z|y, S = 1) = P(x, Z = z|Y = y, S = 1), z = 0, 1, y = 0, 1$ in the case-control study:

$$P(x, Z = z|Y = y, S = 1) =$$

$$\alpha^{zy}(1-\alpha)^{(1-z)y}\left(\frac{\exp(\beta_0^* + \beta^T x)}{1+\exp(\beta_0^* + \beta^T x)}\right)^y \left(\frac{1-z}{1+\exp(\beta_0^* + \beta^T x)}\right)^{1-y} \frac{p(x|S=1)}{P(Y=y|S=1)} \quad (2)$$

Model (2) is derived by expressing $P(x, Z = z|Y = y, S = 1)$ as

$$\frac{P(x, Z = z, Y = y|S = 1)}{P(Y = y|S = 1)}$$

and applying successive conditional probabilities.

We now find the MLE of α and β as follows. The full likelihood in this case-control study reads

$$\prod_{i=1}^{n_1} P(x_i, z_i|Y_i = 1, S_i = 1) \prod_{i=1}^{n_0} P(x_i, z_i|Y_i = 0, S_i = 1)$$

Substituting from (2) for the n_1 cases and the n_0 controls we see that the likelihood is proportional to

$$L = \prod_{i=1}^{n} \alpha^{z_i y_i}(1-\alpha)^{(1-z_i)y_i}\left[\frac{e^{\beta_0^* + \beta^T x_i}}{1+e^{\beta_0^* + \beta^T x_i}}\right]^{y_i}\left[\frac{1-z_i}{1+e^{\beta_0^* + \beta^T x_i}}\right]^{1-y_i} \quad (3)$$

Therefore, setting $\frac{d(ln(L))}{d\alpha} = 0$ yields

$$\sum_{i=1}^{n}\left[(z_i y_i)\frac{1}{\alpha} - (1-z_i)y_i\frac{1}{1-\alpha}\right] = 0$$

As α depends on x, let $n_{11}^{(i)}$, $n_{01}^{(i)}$, $i = 1, \cdots, I$ represent the number of cases when $Z = 1, 0$ for all permissible strata of covariates respectively, and let α_i be the corresponding proportion of PAS in stratum numbered i. (The first subscript is 1 for case, 0 for control, while the second subscript is 1 for PAS, 0 for Non-PAS – see Table 1.) We have

$$\frac{n_{11}^{(i)}}{\alpha_i} - \frac{n_{01}^{(i)}}{1-\alpha_i} = 0,$$

and

$$\hat{\alpha}_i = \frac{n_{11}^{(i)}}{n_1^{(i)}}.$$

To obtain the variance of $\hat{\alpha} \equiv \hat{\alpha}_i$, note that the second derivative of $ln(L)$ becomes

$$d^2 ln(L)/d\alpha^2 = \sum_{i=1}^{n} [(z_i y_i)/\alpha^2 + (1 - z_i) y_i (1 - \alpha)^{-2}],$$

which results in

$$\hat{Var}(\hat{\alpha}) = \frac{n_{11}^{(i)} n_{01}^{(i)}}{(n_1^{(i)})^3}.$$

The expression for $\hat{\alpha}$ can be simplified in the case where α_i is the same across all permissible strata. For example, clinically in the case group, the fraction of PAS (due to some special genetic trend in the cases) is almost unchanged by the presence of other factors such as *mood disorder*, *substance abuse*, *conduct disorder*, *family history of disorder* 1^0 *in relative* and *gun in home*. When we compare the rates of PAS in the case group under different strata formed by risk factors using the data set in our study, the outcome shows a P-value of 0.27 in favor of equal proportions. This coincides with clinical experience. Thus in this special case, it is obvious from (3) that

$$\hat{\alpha} = \frac{n_{11}}{n_1} \quad \text{and} \quad \hat{Var}(\hat{\alpha}) = \frac{n_{11} n_{01}}{n_1^3}.$$

Summarizing the above analysis, we have the following statement:

Theorem 1.

The estimates for Model (2) for the case-control data can be obtained by finding the maximum likelihood estimate (MLE) of α in the above model,

$$\hat{\alpha} = \frac{n_{11}}{n_1}$$

with the variance $Var(\hat{\alpha}) = \frac{n_{11} n_{10}}{n_1^3}$, where n_{10}, n_{11} are the number of cases in the non-PAS and PAS groups respectively and $n_1 = n_{10} + n_{11}$, the total number of cases.

For the other parameters β involved in the model, since Model (2) can be expressed in the following forms:

$$P(x, Z = 1|Y = 1, S = 1) = \alpha \frac{\exp(\beta_0^* + \beta^T x)}{1 + \exp(\beta_0^* + \beta^T x)} \frac{p(x|S = 1)}{P(Y = 1|S = 1)}; \tag{4}$$

$$P(x, Z = 0|Y = 1, S = 1) = (1 - \alpha) \frac{\exp(\beta_0^* + \beta^T x)}{1 + \exp(\beta_0^* + \beta^T x)} \frac{p(x|S = 1)}{P(Y = 1|S = 1)}; \tag{5}$$

$$P(x, Z = 1|Y = 0, S = 1) = 0 \text{ (no PAS in the control group)}; \tag{6}$$

$$P(x, Z = 0|Y = 0, S = 1) = \frac{1}{1 + \exp(\beta_0^* + \beta^T x)} \frac{p(x|S = 1)}{P(Y = 0|S = 1)}. \tag{7}$$

By combining equations (4) and (5), we have

Theorem 2.

The estimates of $\beta_1^*, \cdots, \beta_n^*$ for Model (2) are the same as the estimates from

$$P(x|Y = 1, S = 1) = \frac{\exp(\beta_0^* + \beta^T x)}{1 + \exp(\beta_0^* + \beta^T x)} \frac{p(x|S = 1)}{P(Y = 1|S = 1)}$$

and

$$P(x|Y = 0, S = 1) = \frac{1}{1 + \exp(\beta_0^* + \beta^T x)} \frac{p(x|S = 1)}{P(Y = 0|S = 1)}.$$

One can then use a standard statistical package such as SAS to fit the model, obtain $\hat{\beta}$, and hence the odds ratios of the remaining risk factors.

Interpretation. The MLE of α is the estimate of the risk of PAS in the population of suicides. The conditional probability of suicide, given a past attempt, can then be evaluated by estimating the risk attributable to this factor from the overall risk and then modeling the residual risk over the rest of risk factors by a logistic regression model. The model proposed avoids the difficulty in dealing with zero cell caused by risk factor PAS, but takes into account the effect of PAS on suicide by making an appropriate transformation.

3 Applications

In this section, we discuss an example to illustrate our hybrid method. The study is from an investigation carried out at the Western Psychiatric Institute and Clinic of the University of Pittsburgh [1]. The data was gathered from 140 suicide victims and 131 community controls using a standard psychological autopsy protocol in order to examine the impact of various factors on adolescent suicide risk. Subjects were stratified according to age (older or younger than 16 years) and gender. The summary of the data for female adolescent suicide is given in Table 1.

The aim of the data analysis was to identify major predictors for female adolescent suicide from a group of potential risk factors: PAS, age, handgun in home, substance abuse, family history of abuse, mood disorder, loss of boyfriend or girlfriend, and any psychological disorder. Backward and forward stepwise logistic regression removed the risk factor PAS that is clinically regarded as important, while the regression model with all risk factors did not converge. Grouped logistic regression presented two problems: first, the variances of all estimates were infinite; and second, the risk factor substance abuse was not included in the final model, once again contradicting clinical experience. Also the exact procedures did not converge for this data either. However, our hybrid method provided the following results. First we estimated the proportion of risk contributed by PAS using $\hat{\alpha} \equiv \frac{n_{11}}{n_1} = 13/21 = 0.619$ (with standard deviation 0.106). We then adjusted the overall risk in each stratum by $\hat{\alpha}$ (this is the essence of (4) and (5)), and then we estimated the contribution (probabilities) of the remaining risk factors by logistic regression using Theorem 2. The major risk factors for female suicide identified by this method were mood disorder (OR = odds ratio = 34.3), substance abuse (OR = 36.6) and handgun in home (OR = 15.0) along with PAS. The main risk factors identified by the hybrid model were conisistent with univariate analysis, the literature, and clinical

experience; this analysis provided numerical measures of these risk factors. A description of the clinical background and the data may be found in [1].

4 Discussion

The hybrid logistic regression method presented here provides an alternative approach for modeling binary data when there are rare risk factors. The method proposed is easy to carry out in practice, especially when there is only one rare risk factor. This method can also be viewed as an alternative to exact procedures, which are often used in modeling binary data when the MLE does not exist in a conventional logistic regression.

Our model was tailored to the demands of clinical experience in assuming that PAS in the control group is 0. It would be a generalization of our results to relax this assumption and allow the conditional probability of the rare risk factor to be non-zero in the control group. This would entail modifying equations (6) and (7) and it should not be difficult to handle this extension, but we do not have a data set or situation at hand to which to apply such results. For the important problem of identifying key risk factors for preventing adolescent suicide our assumption of 0 is reasonable and consistent with the data. Of course, it may well be that some controls preferred not to admit any possible past attempt at suicide.

In our application, clinical experience dictates that the fraction of PAS in the cases is almost unchanged by the presence of other factors. This allowed us to pool the estimates for α_i improving the accuracy of the estimated standard deviation. If the alphas vary, the quality of the estimates may be misleading and a simulation study to determine small sample properties would be interesting. Also, modelling alpha as a function of x would be useful for risk factors measured on a continuous scale. Here the risk factors are categorical.

Acknowlegement

We thank the referee for suggesting some generalizations for future work which are discussed in the last section.

References

[1] Brent, D.A., Baugher M., Bridge, J., et al (1999) Age- and sex- related risk factors for adolescent suicide. *Journal of the American Academy of Child and Adolescent Psychiatry* **38**, 12, 1497-1505.

[2] Farewell, V. T. (1979) Some results on the estimation of logistic models based on retrospective data. *Biometrika* **66**, 1, 27-32.

[3] Hosmer D.W. and Lemeshow S. (1989) *Applied Logistic Regression.* John Wiley and Sons, New York.

[4] *LogXact (Version 2.0)* (1996) Cytel Software Corporation, Cambridge, MA.

[5] Mantel, N. (1973) Synthetic retrospective studies and related topics. *Biometrics* **29**, 479-486.

[6] McCullogh, P. and Nelder, J.A. (1989) *Generalized Linear Models, 2nd ed.* Chapman and Hall, London.

[7] Ryan, T.P. (1997) *Modern Regression Methods.* John Wiley and Sons, New York.

[8] Shaffer D, Gould M.S., Fisher P., Trautman P., Moreau D., Kleinman M. and Flory M. (1996) Psychiatric Diagnosis in Child and Adolescent Suicide. *Arch. Gen. Psych.* **53**, 339–348.