2008

Generalization of the Hybrid Logistic Model For More Than One Rare Risk Factor

Mamunur Rashid
DePauw University, mrashid@depauw.edu

Recommended Citation

Originally published: Rashid, Mamunur. Generalization of the Hybrid Logistic Model Under Case-Control Sampling Design. *Journal of Korean Data & Information Science Society*. 19(1): 361-373, 2008. http://www.kdiss.org/journal/list.html?pn=search&s_t=&s_a=mamunur+rashid&s_k=&s_v=&s_n=&x=0&y=0 This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Korean
Data & Information Science Society
2008, Vol. 19, No. 1, pp. 361 ~ 373

GENERALIZATION OF THE HYBRID LOGISTIC MODEL FOR MORE THAN ONE RARE RISK FACTOR

Mamunur Rashid[1]

Abstract

Logistic models are commonly used to analyze case-control data. For case-control studies, if there tends be rare disease in the control group with the risk factors, then the estimation procedure using logistic regression model for such factors becomes difficult. To overcomesuch situation, Chen et. al. (2003) proposed a hybrid logistic model in which they first estimate the problematic risk factor assuming that proportions having disease of such risk factor are treated equal for all permissible strata of the other risk factors, and then the residual of the risk factors are modeled by using the logistic regression model. The purpose of this paper is to extend theoretically the hybrid logistic model for case-control studies for more than one rare risk factor.

Keywords : Hybrid Logistic Model, Logistic Models, Logistic Regression, Risk Factors

1. Introduction

Logistic regression is a popular method for the analysis of binary data for case-control studies with widespread of applicability in different areas. In a case-control study design, diseased cases and non-diseased control study subjects areidentified and followed back to determine their exposure level especially with rare diseases (Prentice and Pyke, 1979). In such studies, if there tends to be rare disease in the control group with the risk factors, then the estimation of the parameters of those risk factors is difficult. The following table provides as an example of the rare risk factor for case-control study.

[1] Department of Mathematics and Statistics. Bowling Green State University. Bowling Green, OH 43403 USA
Email: mrashid@bgsu.edu

Table 1. Female adolescent suicides and controls by PAS
(Source: Chen et. al., 2003)

		Case	Control
Past attempt of suicide (PAS)	Yes	13	0
	No	8	40

In this example, PAS is problematic risk factors for female adolescent suicide because none of the young female had past attempt of suicide in the control group. In this situation, investigators (see, for example, Shaffer et al., 1996), in practice, do not include such risk factors and consider the other risk factors instead. Avoiding such important risk factors may overestimate the odds ratio for the remaining risk factors in the model (Brent et al., 1999). However, including all the risk factors into the model, the model does not converge with all the risk factors. Considering importance of such risk factors, as the existing logistic regression model cannot handle the troublesome risk factors, a hybrid logistic model (Chen et al.,2003) was proposed. In that procedure: first, troublesome risk factor was adjusted, and then the rest of the risk factors were modeled by using logistic regression. The specific form of the hybrid logistic model for case-control studies is expressed for one rare risk factor z and the other risk factors $x^T = (x_1, x_2, \dots, x_p)$ as

$$P(x, Z=z \mid Y=y, s=1) = \alpha^{zy}(1-\alpha)^{(1-z)y}\left(\frac{e^{\beta_0^*+\beta x}}{1+e^{\beta_0^*+\beta x}}\right)^{y}\left(\frac{1-z}{1+e^{\beta_0^*+\beta x}}\right)^{1-y}\frac{P(x \mid s=1)}{P(Y=y \mid s=1)} \quad (1)$$

where, s indicates whether a subject is sampled (1 = yes, 0 = no) and α is the proportion of the covariate $z = 1$ in the case group $P(z=1 \mid y=1, x, s=1) = \alpha$, α depends on x and $P(z=0 \mid y=1, x, s=1) = 1-\alpha$

The parameters in the model α and $\beta^T = (\beta_1, \beta_2, \dots, \beta_p)$ are estimated using the maximum likelihood estimation procedure. The purpose of this paper is to generalize the hybrid logistic model for case-control studies for more than one rare risk factor.

2. The hybrid logistic model: bivariate case

2.1 When the rare risk factors z_1 and z_2 are independent

Suppose $Y_1, Y_2, \ldots, Y_n$ 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. The set of risk factors (z^T, x^T) where $z^T = (z_1, z_2)$ is the rare risk factors and $x^T = (x_1, x_2, \ldots, x_p)$ is the other risk factors. These risk factors for subject i take the values $(z_{i1}, z_{i2}, x_{i1}, x_{i2}, \ldots, x_{ip})$. In this case, we consider the rare risk factor z_i, i = 1,2 takes two possible values with 1 (occurrence of the event) and 0 (not occurrence). The structure of the covariate z_i, i = 1,2 has the pattern with the outcome variable, Y for a sample of size n_1 cases ($y = 1$) and n_0 controls ($y = 0$) as shown in the following table.

Table 2. Cross-classification between the rare risk factor $z_i, i = 1,2$ versus y

		$y = 1$ (Case)	$y = 0$ (Control)
z_1	1	$n_{11}^{(1)}$	$n_{01}^{(1)} = 0$
	0	$n_{10}^{(1)}$	$n_{00}^{(1)}$
Total		$n_1^{(1)}$	$n_0^{(1)}$

		$y = 1$ (Case)	$y = 0$ (Control)
z_2	1	$n_{11}^{(2)}$	$n_{01}^{(2)} = 0$
	0	$n_{10}^{(2)}$	$n_{00}^{(2)}$
Total		$n_1^{(2)}$	$n_0^{(2)}$

According to Hosmer and Lameshow (2000), the full likelihood for a sample of n_1 cases $(y = 1)$ and n_0 controls $(y = 0)$ is,

$$\prod_{i=0}^{n_1} P(x_i, z_{i1}, z_{i2} \mid y_i = 1, s_i = 1) \prod_{i=1}^{n_0} P(x_i, z_{i1}, z_{i2} \mid y_i = 0, s_i = 1) \qquad (2)$$

For an individual term in the likelihood function shown in equation (1), the simplification is given using the Bayes theorem (Sheldon, 2001).

$$P(x,z_1,z_2 \mid y,s=1) = \frac{P(x,z_1,z_2 \mid s=1)\cdot P(y \mid x,z_1,z_2,s=1)}{P(y \mid s=1)} \tag{3}$$

The first term in the numerator of equation (3) yields,

$$P(x,z_1,z_2 \mid s=1) = P(z_1,z_2 \mid x,s=1)\cdot P(x \mid s=1) \tag{4}$$

The second term in the numerator of equation (3) yields,

$$P(y \mid x,z_1,z_2,s=1) = \frac{P(y \mid x,s=1)\cdot P(z_1 \mid z_2,y,x,s=1)\cdot P(z_2 \mid y,x,s=1)}{P(z_1,z_2 \mid x,s=1)} \tag{5}$$

Substituting (4) and (5) in (3), we get

$$P(x,z_1,z_2 \mid y,s=1) = P(y \mid x,s=1)\cdot P(z_1 \mid y,x,s=1)\cdot P(z_2 \mid y,x,s=1)\cdot\frac{P(x \mid s=1)}{P(y \mid s=1)} \tag{6}$$

as z_1 and z_2 are conditionally independent for given y.
Let α_1 be the proportion of the covariate $z_1=1$ and α_2 be the proportion of the covariate $z_2=1$ in the case group
$P(z_1=1 \mid y=1,x,s=1)=\alpha_1$ and $P(z_2=1 \mid y=1,x,s=1)=\alpha_2$, α_1 and α_2 depends on x
In the case when $z_1=0$ and $z_2=0$, we have
$P(z_1=0 \mid y=1,x,s=1)=1-\alpha_1$ and $P(z_2=0 \mid y=1,x,s=1)=1-\alpha_2$
The following model is obtained for the joint distribution of risk factors $P(x,z_1,z_2 \mid y,s=1)=P(x,Z_1=z_1,Z_2=z_2 \mid Y=y,s=1)$, $z_1=z_2=0,1$, $y=0,1$ in the case-control study,

$$P(x, Z_1=z_1, Z_2=z_2 \mid Y=y, s=1) =$$

$$\alpha_1^{z_1 y}(1-\alpha_1)^{(1-z_1)y}\cdot\alpha_2^{z_2 y}(1-\alpha_2)^{(1-z_2)y}\cdot\left(\frac{e^{\beta_0^*+\beta x}}{1+e^{\beta_0^*+\beta x}}\right)^{y}\left(\frac{(1-z_1)(1-z_2)}{1+e^{\beta_0^*+\beta x}}\right)^{1-y}\cdot\frac{P(x \mid s=1)}{P(Y=y \mid s=1)} \tag{7}$$

To find the MLE of $\alpha_1, \alpha_2,$ and β, we substitute (7) in (2) for the n_1 cases and

n_0 controls, we have the likelihood is proportional to

$$L=\prod_{i=1}^{n}\alpha_1^{z_{i1}y_i}(1-\alpha_1)^{(1-z_{i1})y_i}\cdot\alpha_2^{z_{i2}y_i}(1-\alpha_2)^{(1-z_{i2})y_i}\cdot\left(\frac{e^{\beta_0^*+\beta x_i}}{1+e^{\beta_0^*+\beta x_i}}\right)^{y_i}\left(\frac{(1-z_{i1})(1-z_{i2})}{1+e^{\beta_0^*+\beta x_i}}\right)^{1-y_i} \tag{8}$$

Taking ln on both sides of the equation (8) and we get,

$$\ln L=\sum_{i=1}^{n}\Big[z_{i1}y_i\ln\alpha_1+(1-z_{i1})y_i\ln(1-\alpha_1)+z_{i2}y_i\ln\alpha_2+(1-z_{i2})y_i\ln(1-\alpha_2)$$

$$+y_i\ln\left(\frac{e^{\beta_0^*+\beta' x_i}}{1+e^{\beta_0^*+\beta' x_i}}\right)+(1-y_i)\ln\left(\frac{(1-z_{i1})(1-z_{i2})}{1+e^{\beta_0^*+\beta' x_i}}\right)\Big]$$

Setting $\dfrac{\partial(\ln L)}{\partial\alpha_k}=0$, $k=1,2$, we obtain

$$\sum_{i=1}^{n}\left[\frac{z_{i1}y_i}{\alpha_1}-\frac{(1-z_{i1})y_i}{1-\alpha_1}\right]=0$$

$$\sum_{i=1}^{n}\left[\frac{z_{i2}y_i}{\alpha_2}-\frac{(1-z_{i2})y_i}{1-\alpha_2}\right]=0$$

Since α_1 and α_2 depend on x, we let $n_{11}^{(ki)}, n_{10}^{(ki)}$, $k=1,2$ and $i=1,2,\dots,I$ represent the number of cases when $Z_k=1,0$ for k=1,2 for all permissible strata of the covariates respectively, and let α_{ki}, k=1,2 be the proportion of $Z_k=1$, k=1,2 in stratum numbered i. Therefore, we have

$$\frac{n_{11}^{(1i)}}{\alpha_{1i}}-\frac{n_{10}^{(1i)}}{1-\alpha_{1i}}=0 \quad\text{and}\quad \frac{n_{11}^{(2i)}}{\alpha_{2i}}-\frac{n_{10}^{(2i)}}{1-\alpha_{2i}}=0$$

This implies, $\hat{\alpha}_{1i}=\dfrac{n_{11}^{(1i)}}{n_{11}^{(1i)}+n_{10}^{(1i)}}=\dfrac{n_{11}^{(1i)}}{n_1^{(1i)}}$ and $\hat{\alpha}_{2i}=\dfrac{n_{11}^{(2i)}}{n_{11}^{(2i)}+n_{10}^{(2i)}}=\dfrac{n_{11}^{(2i)}}{n_1^{(2i)}}$

To obtain the variance of $\hat{\alpha}_k\equiv\hat{\alpha}_{ki}, k=1,2$, we take the second derivative and we get

$$\frac{\partial^2(\ln L)}{\partial\alpha_1^2}=\sum_{i=1}^{n}\left[-\frac{z_{i1}y_i}{\alpha_1^2}+\frac{(1-z_{i1})y_i}{(1-\alpha_1)^2}\right]=\frac{-n_{11}^{(1i)}}{\alpha_1^2}+\frac{-n_{10}^{(1i)}}{(1-\alpha_1)^2}$$

After simplifying the above, we have

$\hat{V}ar(\hat{\alpha}_1)=\dfrac{n_{11}^{(1i)}n_{01}^{(1i)}}{(n_1^{(1i)})^3}$ and likewise, $\hat{V}ar(\hat{\alpha}_2)=\dfrac{n_{11}^{(2i)}n_{01}^{(2i)}}{(n_1^{(2i)})^3}$

We consider the expression for $\hat{\alpha}_k, k=1,2$ can be simplified in the case where $\alpha_{ki}, k=1,2$ is the same across all permissible strata. Summarizing the above analysis we have the following theorem.

Theorem 1. The estimates for Model (7) for the case-control data can be obtained by finding the MLE of α_k, k =1,2 in the above model,

$$\hat{\alpha}_k=\frac{n_{11}^{(k)}}{n_1^{(k)}}, \quad k=1,2$$

with the variance $\hat{V}ar(\hat{\alpha}_k)=\dfrac{n_{11}^{(k)}n_{01}^{(k)}}{(n_1^{(k)})^3}$, where $n_{10}^{(k)}, n_{11}^{(k)}$ are the number of cases in the $z_k=0$ and $z_k=1$, $k=1,2$ groups, respectively and $n_1^{(k)}=n_{10}^{(k)}+n_{11}^{(k)}$, $k=1,2$, the total number of cases.

For the other parameters involved in the model, Model (7) can be expressed in the following forms

$$P(x,z_1=1,z_2=1\mid y=1,s=1)=\alpha_1\cdot\alpha_2\cdot\frac{e^{\beta_0^*+\beta'x}}{1+e^{\beta_0^*+\beta'x}}\cdot\frac{P(x\mid s=1)}{P(Y=y\mid s=1)} \tag{9}$$

$$P(x,z_1=1,z_2=0\mid y=1,s=1)=\alpha_1\cdot(1-\alpha_2)\cdot\frac{e^{\beta_0^*+\beta'x}}{1+e^{\beta_0^*+\beta'x}}\cdot\frac{P(x\mid s=1)}{P(Y=y\mid s=1)} \tag{10}$$

$$P(x,z_1=0,z_2=1\mid y=1,s=1)=(1-\alpha_1)\cdot\alpha_2\cdot\frac{e^{\beta_0^*+\beta'x}}{1+e^{\beta_0^*+\beta'x}}\cdot\frac{P(x\mid s=1)}{P(Y=y\mid s=1)} \tag{11}$$

$$P(x,z_1=0,z_2=0\mid y=1,s=1)=(1-\alpha_1)\cdot(1-\alpha_2)\cdot\frac{e^{\beta_0^*+\beta'x}}{1+e^{\beta_0^*+\beta'x}}\cdot\frac{P(x\mid s=1)}{P(Y=y\mid s=1)} \tag{12}$$

$$P(x,z_1=1,z_2=1\mid y=0,s=1)=0 \tag{13}$$

$$P(x,z_1=1,z_2=0\mid y=0,s=1)=0 \tag{14}$$

$$P(x,z_1=0,z_2=1\mid y=0,s=1)=0 \tag{15}$$

$$P(x, z_1 = 0, z_2 = 0 \mid y = 0, s = 1) = \frac{1}{1 + e^{\beta_0^* + \beta' x}} \cdot \frac{P(x \mid s = 1)}{P(Y = y \mid s = 1)} \tag{16}$$

By combining equation (9)-(12), we have

Theorem 2. The estimates of $\beta_1^*, \dots, \beta_p^*$ for model (7) are the same as the estimates from

$$P(x \mid Y = 1, s = 1) = \frac{e^{\beta_0^* + \beta' x}}{1 + e^{\beta_0^* + \beta' x}} \quad \frac{P(x \mid s = 1)}{P(Y = 1 \mid s = 1)}$$

and $$P(x \mid Y = 0, s = 1) = \frac{1}{1 + e^{\beta_0^* + \beta' x}} \quad \frac{P(x \mid s = 1)}{P(Y = 1 \mid s = 1)}$$

2.2 Consider the case when the rare risk factors z_1 and z_2 are not independent

Suppose the rare risk factors z_1 and z_2 are not independent and the following table gives the cross-classification between these two variables

Table 3. Cross-classification between z_1 and z_2

		Z_2	
		1	0
Z_1	1	α_{11}	α_{10}
	0	α_{01}	α_{00}

Suppose that $P(z_1 = 1, z_2 = 1 \mid y = 1, x) = \alpha_{11}$

$P(z_1 = 1, z_2 = 0 \mid y = 1, x) = \alpha_{10}$

$P(z_1 = 0, z_2 = 1 \mid y = 1, x) = \alpha_{01}$

and $P(z_1 = 0, z_2 = 0 \mid y = 1, x) = \alpha_{00}$

where $\alpha_{11}, \alpha_{10}, \alpha_{01}$, and α_{00} depend on the covariate x.

The following model is proposed for the joint distribution of risk factors where the variables z_1 and z_2 are not independent

$P(x, z_1, z_2 \mid y, s = 1) = P(x, Z_1 = z_1, Z_2 = z_2 \mid Y = y, s = 1),\ z_1 = z_2 = 0,1,\ y = 0,1$ in the case-control study,

$$P(x,z_1,z_2 \mid y)=\alpha_{11}^{z_1z_2y}\alpha_{10}^{z_1(1-z_2)y}\alpha_{01}^{(1-z_1)z_2y}\alpha_{00}^{(1-z_1)(1-z_2)y}\left(\frac{e^{\beta_0^*+\beta x}}{1+e^{\beta_0^*+\beta x}}\right)^y\left(\frac{(1-z_1)(1-z_2)}{1+e^{\beta_0^*+\beta x}}\right)^{1-y}\frac{P(x \mid s=1)}{P(Y=y \mid s=1)} \qquad (17)$$

Theorem 3. If $\alpha_{11}=\alpha_1\alpha_2$, then the model defined in equation (17) is similar to the model in equation (7).

Proof: Suppose $\alpha_{11}+\alpha_{10}=\alpha_1$, $\alpha_{11}+\alpha_{01}=\alpha_2$, and $\alpha_{00}=1-\alpha_1-\alpha_2+\alpha_{11}$

then the model (17) becomes

$$P(x,z_1,z_2 \mid y)=(\alpha_1\alpha_2)^{z_1z_2y}(\alpha_1-\alpha_{11})^{z_1(1-z_2)y}(\alpha_2-\alpha_{11})^{(1-z_1)z_2y}(1-\alpha_1-\alpha_2-\alpha_{11})^{(1-z_1)(1-z_2)y}$$
$$\left(\frac{e^{\beta_0^*+\beta' x}}{1+e^{\beta_0^*+\beta' x}}\right)^y\left(\frac{(1-z_1)(1-z_2)}{1+e^{\beta_0^*+\beta' x}}\right)^{1-y}\frac{P(x \mid s=1)}{P(Y=y \mid s=1)}$$

After simplification, we have

$$P(x,z_1,z_2 \mid y)=\alpha_1^{z_1z_2y}\alpha_2^{z_1z_2y}\alpha_1^{z_1y}\alpha_1^{-z_1z_2y}(1-\alpha_2)^{-z_1(1-z_2)y}\alpha_2^{z_2y}\alpha_2^{-z_1z_2y}(1-\alpha_1)^{-(1-z_1)z_2y}(1-\alpha_1)^{(1-z_1)y}(1-\alpha_1)^{-(1-z_1)z_2y}$$
$$(1-\alpha_2)^{(1-z_2)y}(1-\alpha_2)^{-z_1(1-z_2)y}\left(\frac{e^{\beta_0^*+\beta' x}}{1+e^{\beta_0^*+\beta' x}}\right)^y\left(\frac{(1-z_1)(1-z_2)}{1+e^{\beta_0^*+\beta' x}}\right)^{1-y}\frac{P(x \mid s=1)}{P(Y=y \mid s=1)}$$

and hence,

$$P(x,z_1,z_2 \mid y)=\alpha_1^{z_1y}(1-\alpha_1)^{(1-z_1)y}\alpha_2^{z_2y}(1-\alpha_2)^{(1-z_2)y}\left(\frac{e^{\beta_0^*+\beta' x}}{1+e^{\beta_0^*+\beta' x}}\right)^y\left(\frac{(1-z_1)(1-z_2)}{1+e^{\beta_0^*+\beta' x}}\right)^{1-y}\frac{P(x \mid s=1)}{P(Y=y \mid s=1)}$$

Which is the model shown in (7).
Hence the proof follows.

Now, to find the MLE of $\alpha_{11},\alpha_{10},\alpha_{01},\alpha_{00}$, and β, we substitute equation (17) in (2) for the n_1 cases and n_0 controls we have the likelihood is proportional to

$$L=\prod_{i=1}^{n}\alpha_{11}^{z_{i1}z_{i2}y_i}\alpha_{10}^{z_{i1}(1-z_{i2})y_i}\alpha_{01}^{(1-z_{i1})z_{i2}y_i}\alpha_{00}^{(1-z_{i1})(1-z_{i2})y_i}\cdot\left(\frac{e^{\beta_0^*+\beta' x_i}}{1+e^{\beta_0^*+\beta' x_i}}\right)^{y_i}\left(\frac{(1-z_{i1})(1-z_{i2})}{1+e^{\beta_0^*+\beta' x_i}}\right)^{1-y_i} \qquad (18)$$

Taking ln on both sides of the equation (18) and we get,

$$\ln L = \sum_{i=1}^{n}\Big[z_{i1}z_{i2}y_i \ln\alpha_{11} + z_{i1}(1-z_{i2})y_i \ln\alpha_{10} + (1-z_{i1})z_{i2}y_i \ln\alpha_{01} + (1-z_{i1})(1-z_{i2})y_i \ln\alpha_{00}$$
$$+ y_i \ln\left(\frac{e^{\beta_0^* + \beta' x_i}}{1+e^{\beta_0^*+\beta' x_i}}\right) + (1-y_i)\ln\left(\frac{(1-z_{i1})(1-z_{i2})}{1+e^{\beta_0^*+\beta' x_i}}\right)\Big]$$

Now, we take the derivatives with respect to α_{11}, α_{10}, and α_{01} respectively and set equal to zero. This yields,

$$\sum_{i=1}^{n}\frac{z_{i1}z_{i2}y_i}{\alpha_{11}} = \sum_{i=1}^{n}\frac{(1-z_{i1})(1-z_{i2})y_i}{\alpha_{00}}$$

$$\sum_{i=1}^{n}\frac{z_{i1}(1-z_{i2})y_i}{\alpha_{10}} = \sum_{i=1}^{n}\frac{(1-z_{i1})(1-z_{i2})y_i}{\alpha_{00}}$$

$$\sum_{i=1}^{n}\frac{(1-z_{i1})z_{i2}y_i}{\alpha_{01}} = \sum_{i=1}^{n}\frac{(1-z_{i1})(1-z_{i2})y_i}{\alpha_{00}}$$

Thus, $\sum_{i=1}^{n}\frac{z_{i1}z_{i2}y_i}{\alpha_{11}} = \sum_{i=1}^{n}\frac{z_{i1}(1-z_{i2})y_i}{\alpha_{10}} = \sum_{i=1}^{n}\frac{(1-z_{i1})z_{i2}y_i}{\alpha_{01}} = \sum_{i=1}^{n}\frac{(1-z_{i1})(1-z_{i2})y_i}{\alpha_{00}} = \sum_{i=1}^{n} y_i$

as $\alpha_{11} + \alpha_{10} + \alpha_{01} + \alpha_{00} = 1$

Since α_{11}, α_{10}, α_{01}, and α_{00} depend on x, we let $n_{11}^{(i)}, n_{10}^{(i)}, n_{01}^{(i)}$, and $n_{00}^{(i)}$, $i = 1,2,\ldots,I$ represent the number of cases of the combination of the variables $z_1, z_2 = 1,0$ for all permissible strata of the covariates respectively, and let α_{11i}, α_{10i}, α_{01i}, and α_{00i} be the proportion of the combination of $z_1, z_2 = 1,0$ in stratum numbered i. Therefore, we have

$\frac{n_{11}^{(i)}}{\alpha_{11i}} = \sum_{i=1}^{n} y_i = n_1^{(i)}$, $\frac{n_{10}^{(i)}}{\alpha_{10i}} = n_1^{(i)}$, $\frac{n_{01}^{(i)}}{\alpha_{01i}} = n_1^{(i)}$, and $\frac{n_{00}^{(i)}}{\alpha_{00i}} = n_1^{(i)}$

This implies, $\hat{\alpha}_{11i} = \frac{n_{11}^{(i)}}{n_1^{(i)}}$, $\hat{\alpha}_{10i} = \frac{n_{10}^{(i)}}{n_1^{(i)}}$, $\hat{\alpha}_{01i} = \frac{n_{01}^{(i)}}{n_1^{(i)}}$, and $\hat{\alpha}_{00i} = \frac{n_{00}^{(i)}}{n_1^{(i)}}$

We consider the proportions are the same across all permissible strata. Therefore, summarizing the above we have the following statement.

Theorem 4. The estimates for Model (17) for the case-control data can be obtained by finding the MLE of $\alpha_{11}, \alpha_{10}, \alpha_{11}$, and α_{00} in the above model,

$$\hat{\alpha}_{11} = \frac{n_{11}}{n_1},\ \hat{\alpha}_{10} = \frac{n_{10}}{n_1},\ \hat{\alpha}_{01i} = \frac{n_{01}}{n_1},\ \text{and}\ \hat{\alpha}_{00} = \frac{n_{00}}{n_1}$$

and the estimated variance is obtained by

$$\hat{V}ar(\hat{\alpha}_{11}) = \frac{n_{11}(n_1 - n_{11})}{n_1^3},\quad \hat{V}ar(\hat{\alpha}_{10}) = \frac{n_{10}(n_1 - n_{10})}{n_1^3},\quad \hat{V}ar(\hat{\alpha}_{01}) = \frac{n_{01}(n_1 - n_{01})}{n_1^3},\ \text{and}$$

$\hat{V}ar(\hat{\alpha}_{00}) = \frac{n_{10}(n_1 - n_{10})}{n_1^3}$, where n_{ij}, $i,j = 1,0$ is the number of cases of the combination of variables $z_1, z_2 = 1,0$ and n_1 is the total number of cases.

For the other parameters involved in the model, Model (17) can be expressed in the following forms

$$P(x, z_1 = 1, z_2 = 1 \mid y = 1, s = 1) = \alpha_{11} \cdot \frac{e^{\beta_0^* + \beta' x}}{1 + e^{\beta_0^* + \beta' x}} \cdot \frac{P(x \mid s = 1)}{P(Y = y \mid s = 1)} \qquad (19)$$

$$P(x, z_1 = 1, z_2 = 0 \mid y = 1, s = 1) = \alpha_{10} \cdot \frac{e^{\beta_0^* + \beta' x}}{1 + e^{\beta_0^* + \beta' x}} \cdot \frac{P(x \mid s = 1)}{P(Y = y \mid s = 1)} \qquad (20)$$

$$P(x, z_1 = 0, z_2 = 1 \mid y = 1, s = 1) = \alpha_{01} \cdot \frac{e^{\beta_0^* + \beta' x}}{1 + e^{\beta_0^* + \beta' x}} \cdot \frac{P(x \mid s = 1)}{P(Y = y \mid s = 1)} \qquad (21)$$

$$P(x, z_1 = 0, z_2 = 0 \mid y = 1, s = 1) = \alpha_{00} \cdot \frac{e^{\beta_0^* + \beta' x}}{1 + e^{\beta_0^* + \beta' x}} \cdot \frac{P(x \mid s = 1)}{P(Y = y \mid s = 1)} \qquad (22)$$

$$P(x, z_1 = 1, z_2 = 1 \mid y = 0, s = 1) = 0 \qquad (23)$$

$$P(x, z_1 = 1, z_2 = 0 \mid y = 0, s = 1) = 0 \qquad (24)$$

$$P(x, z_1 = 0, z_2 = 1 \mid y = 0, s = 1) = 0 \qquad (25)$$

$$P(x, z_1 = 0, z_2 = 0 \mid y = 0, s = 1) = \frac{1}{1 + e^{\beta_0^* + \beta' x}} \cdot \frac{P(x \mid s = 1)}{P(Y = y \mid s = 1)} \qquad (26)$$

By combining equation (19)-(22), we have

Theorem 5. The estimates of $\beta_1^*, \ldots, \beta_p^*$ for model (17) are the same as the estimates from

$$P(x \mid Y=1, s=1) = \frac{e^{\beta_0^* + \beta' x}}{1+e^{\beta_0^* + \beta' x}} \; \frac{P(x \mid s=1)}{P(Y=1 \mid s=1)}$$

and $$P(x \mid Y=0, s=1) = \frac{1}{1+e^{\beta_0^* + \beta' x}} \; \frac{P(x \mid s=1)}{P(Y=1 \mid s=1)}$$

3. The hybrid logistic model: k-variate case

3.1 When the rare risk factors $z_1, z_2, \cdots, z_k$ are independent

In this case we consider k covariates, $z_1, z_2, z_3, \ldots, z_{k-1}$, and z_k have no event in the control group Consider (z^T, x^T) is a set of explanatory variables in the model, where $z^T = (z_1, z_2, z_3, \ldots, z_k)$ represents rare risk factors and $x^T = (x_1, x_2, \ldots, x_p)$ represents the other risk factors. In the case, when we assume $z_1, z_2, \cdots, z_{k-1}$, and z_k are independent and each variable consists two groups 1 and 0, then we propose the following model,

$$P(x, Z_1 = z_1, Z_2 = z_2, \ldots, Z_k = z_k \mid Y = y, s=1) = \prod_{j=1}^{k} \alpha_j^{z_{ji} y_i} (1-\alpha_j)^{(1-z_{ji}) y_i} \cdot \left(\frac{e^{\beta_0^* + \beta' x}}{1+e^{\beta_0^* + \beta' x}} \right)^{y}$$

$$\left(\frac{(1-z_1)(1-z_2)\ldots(1-z_k)}{1+e^{\beta_0^* + \beta' x}} \right)^{1-y} \cdot \frac{P(x \mid s=1)}{P(Y=y \mid s=1)} \qquad (27)$$

where α_j, j=1,2,⋯,k be the proportion of the covariate $z_j = 1$ in the case group $P(z_j = 1 \mid y=1, x, s=1) = \alpha_j$, j=1,2,⋯,k

The estimates for Model (27) for the case-control data can be obtained by finding the MLE of α_j, j=1,2,⋯,k which is similar to described in the Theorem 1. The estimates of other parameters can be obtained by applying Theorem 2.

3.2 When the rare risk factors $z_1, z_2, \cdots, z_k$ are not independent

Suppose the rare risk factors $z_1, z_2, \cdots, z_k$ are not independent and $x^T = (x_1, x_2, \ldots, x_p)$ represents the other risk factors. The following model are proposed for the joint distribution of risk factors

$P(x, z_1, z_2, ..., z_k \mid y, s=1) = P(x, Z_1 = z_1, ..., Z_k = z_k \mid Y = y, s=1),\ z_1 = ... = z_k = 0,1,\ y = 0$
in the case-control study,

$$P(x, z_1, z_2, ..., z_k \mid y) = \prod_{\substack{(i_1 i_2 .. i_k): \\ i_1, i_2, ... i_k = 0,1}} \alpha_{i_1 i_2 \Lambda i_k}^{\prod_{j=1}^{k} z_j^{i_j} (1-z_j)^{1-i_j}} \left(\frac{e^{\beta_0^* + \beta x}}{1 + e^{\beta_0^* + \beta x}} \right)^{y} \left(\frac{(1-z_1)...(1-z_k)}{1 + e^{\beta_0^* + \beta x}} \right)^{1-y} \frac{P(x \mid s=1)}{P(Y=y \mid s=1)} \quad (28)$$

The estimates for Model (28) for the case-control data can be obtained by finding the MLE of $\prod_{\substack{(i_1 i_2 ... i_k): \\ i_1, i_2, ..., i_k = 0,1}} \hat{\alpha}_{i_1 i_2 \Lambda i_k}$ which is similar to described in the Theorem 4. The estimates of other parameters can be obtained by applying Theorem 5.

4. Conclusion

This article theoretically extends the hybrid logistic model for case-control study for more than one rare risk factor. The methods provide here for modeling binary data when the risk factors associated with the outcome are exceedingly rare in thecontrol group, which is difficult to carry out with the conventional logistic regression method. In this approach, we estimate the rare risk factors assuming that proportions are equal for all permissible strata of the other risk factors because if the proportions are different, the estimates of the rare risk factors could be misleading. The estimates of remaining parameters of the model can be obtained using the standard statistical package such as SAS. In this paper, we consider that proportions of rare risk factors are the functions of the other risk factors as categorical. However, one could measure it with continuous scale risk factor that would be interesting.

Acknowledgement

The author is indebted to Dr. John T. Chen, Associate Professor and Director of Graduate Programs and Dr. Truc T. Nguyen, Professor of Emeritus, Department of Mathematics and Statistics, Bowling Green State University, USA for their suggestions. The author also thanks two anonymous referees for valuable comments.

References

1. Brent, D.A., Baugher, M, Bridge, J., Chen, T., and Chiappetta, L. (1999). Age- and Sex-related factors for adolescent suicide. *Journal of the American Academy of Child and Adolescent Psychiatry*, 38, 1497-1505.
2. Chen, T., Hoppe, F.M., Iyengar, S., and Brent, D. (2003). A Hybrid logistic Model for Case-Control Studies. *Methodology and Computing in Applied Probability*, 5, 419-426.
3. Hosmer, D.W. and Lameshow, S. (2000). *Applied Logistic Regression.* John Wiley and Sons: New York.
4. Prentice, R.L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* 66, 403-11
5. 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.* Vol. 53 pp. 339-348.
6. Sheldon, R. (2001). *A First Course in Probability (6th edition)*, Prentice Hall

[received date : Nov. 2007, accepted date : Feb. 2008]