A Fresh Look at the Discriminant Function Approach for Estimating Crude or Adjusted Odds Ratios

Robert Lyles, *Emory University*
Ying Guo, *Emory University*
Andrew Hill, *Emory University*

Journal Title: American Statistician

Volume: Volume 63, Number 4

Publisher: Taylor & Francis: SSH Journals | 2009

Type of Work: Article | Post-print: After Peer Review

Publisher DOI: 10.1198/tast.2009.08246

Permanent URL: http://pid.emory.edu/ark:/25593/fk55t

Final published version:
http://www.tandfonline.com/doi/abs/10.1198/tast.2009.08246#.UxoZAOddWgg

Published in final edited form as:
Am Stat. 2009 ; 63(4): . doi:10.1198/tast.2009.08246.

A Fresh Look at the Discriminant Function Approach for Estimating Crude or Adjusted Odds Ratios

Robert H. Lyles [Associate Professor], **Ying Guo [Assistant Professor]**, and **Andrew N. Hill [Adjunct Instructor]**
Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322

Robert H. Lyles: rlyles@sph.emory.edu

Abstract

Assuming a binary outcome, logistic regression is the most common approach to estimating a crude or adjusted odds ratio corresponding to a continuous predictor. We revisit a method termed the discriminant function approach, which leads to closed-form estimators and corresponding standard errors. In its most appealing application, we show that the approach suggests a multiple linear regression of the continuous predictor of interest on the outcome and other covariates, in place of the traditional logistic regression model. If standard diagnostics support the assumptions (including normality of errors) accompanying this linear regression model, the resulting estimator has demonstrable advantages over the usual maximum likelihood estimator via logistic regression. These include improvements in terms of bias and efficiency based on a minimum variance unbiased estimator of the log odds ratio, as well as the availability of an estimate when logistic regression fails to converge due to a separation of data points. Use of the discriminant function approach as described here for multivariable analysis requires less stringent assumptions than those for which it was historically criticized, and is worth considering when the adjusted odds ratio associated with a particular continuous predictor is of primary interest. Simulation and case studies illustrate these points.

Keywords

Bias; Efficiency; Logistic regression; Minimum variance unbiased estimator

1. INTRODUCTION

Odds ratios are staple measures of effect under many common study designs. Logistic regression is the most common analytic tool for estimating these measures based on univariate binary outcome data, particularly in the presence of one or more continuous predictors. Prior to the popularization of logistic regression, however, researchers had proposed other approaches. One of these, based on what is often referred to as the discriminant function approach (Anderson 1958; Cornfield 1962; Truett, Cornfield, and Kannel 1967; Halperin, Blackwelder, and Verter 1971; Efron 1975; Schlesselman 1982; Greenland 1987), is the subject of this article.

Cornfield (1962) and Truett, Cornfield, and Kannel (1967) recognized the connection between the logistic regression and discriminant function approaches, assuming normality of a predictor (X) or joint normality of a set of predictors ($\mathbf{X}$) conditional on a binary outcome (Y). Efron (1975) demonstrated a considerable efficiency advantage for the discriminant function approach over logistic regression for the population distinction problem under normality. However, many authors (Halperin, Blackwelder, and Verter 1971; Efron 1975; Schlesselman 1982; Greenland 1987) were quick to point out the disadvantage of requiring distributional assumptions for **X** given *Y* as logistic regression became more accessible, because the latter did not rely on such assumptions.

Whereas we agree that logistic regression enjoys more general application, we find motivation for a new look at the discriminant function approach as it relates to odds ratio estimation. In particular, historical criticism was primarily aimed at multivariate discriminant function analysis due to its seldom defensible assumption of joint normality of predictors (Cornfield 1962; Truett, Cornfield, and Kannel 1967). Although Halperin, Blackwelder, and Verter (1971) provided a detailed demonstration of the pitfalls of this assumption in the presence of categorical explanatory variables, there appears to have been little or no recognition of the fact that the discriminant function approach can be used to obtain valid adjusted odds ratio estimates under much less stringent conditions. For the common situation in which the adjusted odds ratio associated with a particular predictor is of main interest, we show that univariate normality of that predictor conditional on the outcome and other covariates (e.g., potential confounders) is all that is required. This approach is intuitive and illustrates an instructive connection between the multiple linear and logistic regression models. It also leads us to a uniformly minimum variance unbiased (UMVU) estimator for a crude or adjusted log odds ratio. This novel estimator is not only strictly unbiased, but can be notably more efficient than the traditional logistic regression estimator and leads to narrower confidence intervals for the odds ratio without sacrificing coverage. A further potential advantage is the availability of well-behaved estimators in certain situations where logistic regression fails to provide meaningful results due to a separation of data points.

We believe that the discriminant function approach, and particularly the multivariable version discussed below, has gone largely unrecognized by most epidemiologists and applied statisticians. It nevertheless offers an application tool of considerable potential value when justifiable. We demonstrate that seeking such justification may often be worthwhile in practice, as it only requires the application of standard linear regression diagnostic procedures.

2. METHODS AND EXAMPLES

Assume a binary outcome (*Y*), a continuous predictor of interest (*X*), and a study design (e.g., cross-sectional, prospective, retrospective) that supports estimating the odds ratio (OR) associated with a unit increase in *X*:

$$\mathrm{OR}=\frac{\Pr(Y=1|X=x+1)/\Pr(Y=0|X=x+1)}{\Pr(Y=1|X=x)/\Pr(Y=0|X=x)}. \quad (1)$$

Much of our motivation is founded on the equivalent representation of the OR that establishes its estimability by means of a case-control study. In the continuous predictor case, this is written in terms of conditional probability density functions (pdf's) as follows:

$$\mathrm{OR}=\frac{f_{X|Y=1}(x+1)/f_{X|Y=1}(x)}{f_{X|Y=0}(x+1)/f_{X|Y=0}(x)}. \quad (2)$$

Along similar lines, Cornfield (1962) applied Bayes' Rule to demonstrate the following connection:

$$\Pr(Y=1|X=x)=\frac{1}{1+((1-p_y)/p_y)(f_{X|Y=0}(x)/f_{X|Y=1}(x))}, \quad (3)$$

where p_y = Pr(Y = 1). In the presence of other covariates $\mathbf{C}'$ = (C_1, C_2, …, C_T), we rewrite (1) to define the adjusted OR associated with a unit increase in X, that is

$$\mathrm{OR}=\left(\Pr(Y=1|X=x+1,\mathbf{C}=\mathbf{c})/\Pr(Y=0|X=x+1,\mathbf{C}=\mathbf{c})\right)/\left(\Pr(Y=1|X=x,\mathbf{C}=\mathbf{c})/\Pr(Y=0|X=x,\mathbf{C}=\mathbf{c})\right). \quad (4)$$

The corresponding analogues to (2) and (3) become

$$\mathrm{OR}=\frac{f_{X|Y=1,\mathbf{C}=\mathbf{c}}(x+1)/f_{X|Y=1,\mathbf{C}=\mathbf{c}}(x)}{f_{X|Y=0,\mathbf{C}=\mathbf{c}}(x+1)/f_{X|Y=0,\mathbf{C}=\mathbf{c}}(x)} \quad (5)$$

and

$$\Pr(Y=1|X=x,\mathbf{C}=\mathbf{c})=\left(1+\frac{(1-p_{y|\mathbf{c}})}{p_{y|\mathbf{c}}}\frac{f_{X|Y=0,\mathbf{C}=\mathbf{c}}(x)}{f_{X|Y=1,\mathbf{C}=\mathbf{c}}(x)}\right)^{-1}, \quad (6)$$

where $p_{y|\mathbf{c}}$ = Pr(Y = 1|**C** = **c**).

2.1 Conditional Normality of *X*: The No-Covariate Case

The common approach to estimating the OR in (1) is via maximum likelihood (ML) based on the logistic regression model

$$\mathrm{logit}[\Pr(Y=1|X=x)]=\alpha+\beta x. \quad (7)$$

For an alternative based on (2), however, assume that the observed data (X, Y) are consistent with the assumptions behind a standard two-sample t -test under homogeneity of variance. In other words, assume conditional independence of the X's such that

$$X|Y=y\sim \mathrm{N}(\mu_y,\sigma^2) \quad (y=0,1).$$

Inserting the corresponding normal pdf's into (2) and simplifying, we obtain

$$\mathrm{OR}=e^{(\mu_1-\mu_0)/\sigma^2}. \quad (8)$$

Moreover, Cornfield (1962) used (3) to verify that this conditional normality assumption implies the logistic model in (7). A closed-form competitor to the ML estimator of e^{β} based on (7) becomes

$$\widehat{\mathrm{OR}}_{\mathrm{samp}}=e^{(\overline{X}_1-\overline{X}_0)/S_p^2}, \quad (9)$$

where $\bar{X}_j$ is the sample mean in the $Y = j$ group (j = 0, 1), S_p^2 is the usual pooled sample variance, and the subscript "samp" indicates substitution of the usual sample-based estimates for unknown parameters in (8).

A delta method-based variance estimator for $\ln(\widehat{\mathrm{OR}})_{\mathrm{samp}}=(\overline{X}_1-\overline{X}_0)/S_p^2$ is readily obtained. However, we find improvements possible by taking advantage of the moment properties of the chi-squared distribution and the independence of the random variables ($\bar{X}_1 - \bar{X}_0$) and S_p^2. This leads to the following exact variance expression:

$$\mathrm{Var}[\ln(\widehat{\mathrm{OR}})_{\mathrm{samp}}]=\left(\frac{n-2}{n-4}\right)^2(\sigma^2)^{-2}\times\left[\left(\frac{n-4}{n-6}\right)\sigma^2(1/n_1+1/n_0)+2(\mu_1-\mu_0)^2/(n-6)\right],$$

where n_j is the number of subjects with $Y = j$ (j = 0, 1) and $n = n_0 + n_1$. An unbiased estimator of this exact variance becomes arguably ideal, and we derive it as follows:

$$\widehat{\mathrm{Var}}[\ln(\widehat{\mathrm{OR}})_{\mathrm{samp}}]=\left(\frac{n-2}{n-4}\right)(S_p^2)^{-2}\times[S_p^2(1/n_1+1/n_0)+2(\overline{X}_1-\overline{X}_0)^2/(n-2)]. \quad (10)$$

We relate in passing that the premultiplying factor of (n − 2)/(n − 4) in (10) distinguishes it slightly from a straight delta method-based estimator. This variance estimator also differs notably from one given by Greenland (1987), which omitted the second term inside the brackets and thus might tend to underestimate the variance when the OR is not near unity.

It is of interest to compare the finite-sample performance of the estimator in (9) and the accompanying confidence interval (CI) based on exponentiating the bounds of a standard normal theory-based CI for ln(OR) utilizing (10), versus that of the usual logistic regression-based point and interval estimators. An alternative to (9) is to exponentiate the uniformly minimum variance unbiased (UMVU) estimator for ln(OR), which we obtain by finding the following unbiased function of complete and sufficient statistics:

$$\ln(\widehat{\mathrm{OR}})_{\mathrm{umvu}}=\left(\frac{n-4}{n-2}\right)(\overline{X}_1-\overline{X}_0)/S_p^2, \quad (11)$$

with

$$\widehat{\mathrm{Var}}[\ln(\widehat{\mathrm{OR}})_{\mathrm{umvu}}]=\left(\frac{n-4}{n-2}\right)^2\widehat{\mathrm{Var}}[\ln(\widehat{\mathrm{OR}})_{\mathrm{samp}}]$$

and $\widehat{\mathrm{Var}}[\ln(\widehat{\mathrm{OR}})_{\mathrm{samp}}]$ as given in (10).

Note from (8) that under the assumption of conditional normality with homogeneous variance, OR = 1 if and only if $\mu_1 = \mu_0$. As a result, a uniformly most powerful unbiased size α test (Shao 2003) of H_0: OR = 1 under these conditions is based directly on the usual two-sample t -test, with the same power. In practice, the potentially improved estimators in (9) and (11) are justified when the assumptions of the two-sample t -test are met. Simulation studies (to follow) compare the estimators and address considerations of power and Type I error.

2.2 Conditional Normality of *X*: The Multivariable (Covariate-Adjusted) Case

For estimating the covariate-adjusted OR in (4), the usual approach is to calculate the MLE of e^{β} based on the logistic regression model

$$\mathrm{logit}[\Pr(Y{=}1|X{=}x,\mathbf{C}{=}\mathbf{c})]{=}\alpha{+}\beta x{+}\gamma^{'}\mathbf{c}, \quad (12)$$

where $\gamma' = (\gamma_1, \gamma_2, \ldots, \gamma_T)$. It was quickly recognized (Cornfield 1962; Truett, Cornfield, and Kannel 1967; Halperin, Blackwelder, and Verter 1971) that joint multivariate normality of the predictors (X, $\mathbf{C}$) conditional on the outcome Y implies model (12). However, we suggest a much more palatable and defensible analogue to the result from the previous section that we refer to as a *multivariable* discriminant function approach (emphasizing control of covariates without requiring multivariate normality). This approach utilizes the following multiple linear regression model:

$$\mathrm{E}(X|Y{=}y,\mathbf{C}{=}\mathbf{c}){=}\alpha^{*}{+}\beta^{*}y{+}\gamma^{'*}\mathbf{c}. \quad (13)$$

Assuming independent and identically distributed (iid) normal errors with mean 0 and variance σ^2 under model (13), we obtain the following via (5):

$$\mathrm{OR}{=}e^{(\mu_{1\mathbf{c}}-\mu_{0\mathbf{c}})/\sigma^2}{=}e^{\beta^{*}/\sigma^2}, \quad (14)$$

where $\mu_{y\mathbf{c}} = \mathrm{E}(X|Y = y, \mathbf{C} = \mathbf{c})$ (y = 0, 1). These assumptions about the distribution of X conditional on (Y, $\mathbf{C}$) are far less restrictive than multivariate normality of (X, $\mathbf{C}$), and are readily assessed via standard linear regression diagnostics (Kutner et al. 2005). Equation (6) dictates that they imply the logistic model (12), provided logit[Pr(Y = 1|$\mathbf{C}$ = $\mathbf{c}$)] is linear in $\mathbf{c}$. Regardless of whether the latter is the case, (13) yields an analogue to the estimator from the previous section:

$$\widehat{\mathrm{OR}}_{\mathrm{samp}}{=}e^{\widehat{\beta}^{*}/\mathrm{MSE}}, \quad (15)$$

where $\widehat{\beta^{*}}$ and MSE represent the usual ordinary least squares estimator for β^{*} and the residual variance estimator based on model (13), respectively. An exact variance can again be derived, that is,

$$\mathrm{Var}[\ln(\widehat{\mathrm{OR}})_{\mathrm{samp}}]{=}\left(\frac{n{-}T{-}2}{n{-}T{-}4}\right)^{2}(\sigma^{2})^{-2}{\times}\left[\left(\frac{n{-}T{-}4}{n{-}T{-}6}\right)\mathrm{Var}(\widehat{\beta}^{*}){+}2\beta^{2}/(n{-}T{-}6)\right],$$

where Var($\widehat{\beta^{*}}$) is the true variance of $\widehat{\beta^{*}}$ under model (13) and $n - T - 2$ is the error degrees of freedom associated with that model. As before, we recommend an unbiased estimator for the exact variance, which we obtain as

$$\widehat{\mathrm{Var}}[\ln(\widehat{\mathrm{OR}})_{\mathrm{samp}}]{=}\left(\frac{n{-}T{-}2}{n{-}T{-}4}\right)\mathrm{MSE}^{-2}\times[\widehat{\mathrm{Var}}(\widehat{\beta}^{*}){+}2\widehat{\beta}^{*2}/(n{-}T{-}2)].$$

A UMVU estimator also remains feasible in the covariate-adjusted case. In particular, the analogue to (11) becomes

$$\ln(\widehat{\text{OR}})_{\text{umvu}}=\left(\frac{n-T-4}{n-T-2}\right)\hat{\beta}^{*}/\text{MSE}, \quad (16)$$

with

$$\widehat{\text{Var}}[\ln(\widehat{\text{OR}})_{\text{umvu}}]=\left(\frac{n-T-4}{n-T-2}\right)^{2}\widehat{\text{Var}}[\ln(\widehat{\text{OR}})_{\text{samp}}].$$

Equations (14)–(16) also hold under the no-covariate case of the previous section, as the elimination of covariates **C** in (13) reduces to the two-sample *t* -test scenario. Note from (14) that OR = 1 if and only if β^{*} = 0, so the usual two-sided partial *t* -test addressing $H_{0:}$ β^{*} = 0 produces a uniformly most powerful unbiased size α test for $H_{0:}$ OR = 1 assuming (13) with iid normal errors.

Example 1—Consider a study of low birth weight, used for logistic regression illustrations and freely available in the text by Hosmer and Lemeshow (2000). Original data on 189 births were obtained at Baystate Medical Center in Massachusetts, and were altered somewhat by the authors of the text. For the current example, we restrict attention to the 100 births for which the mother required no physician visits during the first trimester. The outcome *Y* characterizes infant birth weight (1 if ≥2500 g, 0 if <2500 g), and our predictor of primary interest is the natural log of the mother's weight at her last menstrual period (LOGLWT). Potential risk factors to be controlled for include: Mother's age, Race (1 if white, 0 otherwise), Smoking status during pregnancy (1 if yes, 0 if no), History of premature labor (1 if any, 0 if none), History of hypertension (1 if yes, 0 if no).

Table 1 provides estimated adjusted ORs corresponding to LOGLWT based on standard logistic regression and based on the multivariable discriminant function approach. In the latter case, both the sample estimator (15) and that based on exponentiating the UMVU estimator for ln(OR) (16) are tabulated. The OR estimates correspond to a one-unit increase in the natural log of maternal weight, which is approximately the largest increase supported by the observed data given that the smallest and largest maternal weights were 80 and 250 pounds. These estimates are large, with wide confidence intervals that overlap unity in each case but correspond to marginal significance of the adjusted OR. However, note the marked difference in the point estimate based on logistic regression (9.60) versus those based on the multivariable discriminant function approach (7.98 and 7.63). Standard errors are smaller and 95% CIs are substantially narrower for the latter two approaches, with a 34% reduction in width of the CI based on the UMVU estimator for ln(OR) relative to standard logistic regression. The fitted multiple linear regression model for this example is given in the Simulation Study section below.

Figure 1 provides evidence suggesting that the univariate normality assumptions required for the multivariable discriminant function-based analysis are reasonable for these data. In particular, histograms of the residuals from the multiple linear regression of LOGLWT on *Y* and the covariates listed above appear bell-shaped for both the high (*y* = 1) and low (*y* = 0) birth weight groups. The sample variances of these two sets of residuals are quite similar (0.040 versus 0.034), and Shapiro–Wilk tests fail to reject residual normality both overall and separately for the high and low groups (*p* = 0.27 and *p* = 0.82, respectively). Partial plots (not shown) reveal no other obvious problems with the typical multiple regression assumptions. Thus, it appears defensible to report the notably more precise results in Table 1 based on multivariable discriminant function analysis.

2.3 The Heterogeneous Variance Case

Here we return to the setting discussed immediately following (7), but with heterogeneous variances. That is, assume conditional independence of the *X*'s such that

$$X|Y{=}1{\sim}\mathrm{N}(\mu_1,\sigma_1^2) \quad \text{and} \quad X|Y{=}0{\sim}\mathrm{N}(\mu_0,\sigma_0^2).$$

It then follows directly from (2) that

$$\mathrm{OR}{=}\frac{e^{-(x-\mu_1+0.5)/\sigma_1^2}}{e^{-(x-\mu_0+0.5)/\sigma_0^2}}{=}e^{\beta+\psi+2\psi x}, \quad (17)$$

where

$$\begin{gathered}\beta{=}\mu_1/\sigma_1^2{-}\mu_0/\sigma_0^2 \quad \text{and} \\ \psi{=}0.5(1/\sigma_0^2{-}1/\sigma_1^2).\end{gathered} \quad (18)$$

Thus, heterogeneity of variance implies that the OR associated with a unit increase in *X* depends on the value (*x*) of *X*. As recognized by multiple authors (Cornfield 1962; Truett, Cornfield, and Kannel 1967; Anderson 1975) and reviewed by Agresti (2002), an argument via Bayes' Rule like that leading to (3) implies a logistic regression model with a quadratic term:

$$\mathrm{logit}[\Pr(Y{=}1|X{=}x)]{=}\alpha{+}\beta x{+}\psi x^2. \quad (19)$$

Note that the OR corresponding to a unit increase in *X* based on (19) is identical to the right side of (17). Thus, evidence of heterogeneity in the variance of *X* across the values (0, 1) of *Y* is also evidence in favor of a second-order logistic model.

In keeping with the general theme, closed-form estimators for β and ψ in (18) make promising competitors to the MLEs based on model (19). In particular, it is natural to consider

$$\begin{gathered}\widehat{\beta}_{\text{samp}}{=}\overline{X}_1/S_1^2{-}\overline{X}_0/S_0^2 \quad \text{and} \\ \widehat{\psi}_{\text{samp}}{=}0.5(1/S_0^2{-}1/S_1^2),\end{gathered} \quad (20)$$

where S_j^2 is the sample variance of *X* among experimental units with *Y* = *j*. Corresponding delta method-based variance estimators are

$$\begin{gathered}\widehat{\mathrm{Var}}(\widehat{\beta}_{\text{samp}}){=}\sum_{j=0}^{1}\widehat{\mathrm{Var}}(\overline{X}_j/S_j^2) \quad \text{and} \\ \widehat{\mathrm{Var}}(\widehat{\psi}_{\text{samp}}){=}0.25\sum_{j=0}^{1}\widehat{\mathrm{Var}}(1/S_j^2),\end{gathered}$$

where $\widehat{\mathrm{Var}}(\overline{X}_j/S_j^2){=}(S_j^{-4})[S_j^2/n_j{+}2\overline{X}_j^2/(n_j{-}1)]$ and $\widehat{\mathrm{Var}}(1/S_j^2){=}2S_j^{-4}/(n_j{-}1)$ $(j{=}0.1)$. Alternatively, UMVU estimators are given by

$$\begin{aligned}\widehat{\beta}_{\text{umvu}}=&\left(\frac{n_1-3}{n_1-1}\right)\overline{X}_1/S_1^2-\left(\frac{n_0-3}{n_0-1}\right)\overline{X}_0/S_0^2,\\ \widehat{\psi}_{\text{umvu}}=&0.5\left\{\left[\frac{n_0-3}{(n_0-1)S_0^2}\right]-\left[\frac{n_1-3}{(n_1-1)S_1^2}\right]\right\},\end{aligned} \quad (21)$$

with the obvious minor adjustments to the variance estimators. These closed-form estimators are valid alternatives to those from logistic regression if the data on *X* are consistent with the assumptions of the two-sample *t* -test with heterogeneous variances across the levels of *Y*.

Example 2—For a real-life application of (20) and (21), we use data originally described by Hastie and Tibshirani (1990) and revisited by Agresti (2002, exercise 5.4, p. 199). The motivating study sought to associate the odds of kyphosis (marked flexion of the spine) subsequent to spinal surgery, with patient age in months at the time of the operation. These ages ranged from 12 to 157 months for 18 patients experiencing kyphosis, and from 1 to 206 months for 22 patients not experiencing kyphosis. Although there is some lack of symmetry in the age data for the no kyphosis group, Shapiro–Wilk tests fail to reject the null hypothesis for normality of age in either set of patients ($p = 0.35$ and $p = 0.09$ for the kyphosis and no kyphosis groups, respectively).

The mean (±SD) ages for those experiencing and not experiencing kyphosis were 93.1 (±43.1) months and 80.1 (±64.8) months, respectively. The apparent tendency toward greater variation in the no kyphosis group approaches statistical significance ($p = 0.09$) based on a standard *F* test for equality of variance, and suggests [see (18)] a potentially concave downward association between age and the logit of the response probability in model (19). Table 2 provides estimates of β and ψ based on standard logistic regression and based on the discriminant function estimators in (20) and (21). Note that standard errors are predictably smaller based upon the discriminant function approach, whereas the quadratic term is statistically significant by either method. Though not detailed here, a simulation study assuming normality of age in each group and mimicking the conditions of this example (including sample size) suggests roughly 25% gains in efficiency relative to logistic regression when using the UMVU estimators in (20) and (21). These simulations revealed noticeable positive and negative bias in the logistic regression-based estimates of β and ψ, respectively, as suggested by the results in Table 2.

2.4 Logistic Regression Convergence Failures

A well-documented occasional problem with logistic regression is a convergence failure due to nonexistence of the MLE for one or more model parameters. Whereas this is often due to the presence of one or more binary covariates that yield a complete or quasi-complete separation of data points (Allison 2004), it is also possible for such a separation to be produced by a continuous predictor of interest. In such an event, standard logistic regression fails to deliver a reliable estimate and standard error associated with that predictor.

Commonly advocated solutions to the separation problem include the use of exact logistic regression (Cox 1970), or penalized maximum likelihood (Firth 1993; Heinze and Schemper 2002). These alternatives can be useful, although each poses its own set of problems. Exact logistic regression can be implemented via commercial software, but is computationally intensive and may be infeasible for larger datasets. In addition, the conditional MLE via the exact approach does not always exist, standard errors are not produced, and the method may not provide a valid or meaningful confidence interval when complete or quasi-complete separation occurs. The penalized likelihood approach circumvents some of these problems, but presents challenges for valid standard error estimation and may not be as accessible via common software. It is worth noting that the discriminant function approach (including the multivariable version as discussed in Section 2.2) is unhampered by the complete or quasi-complete separation problem. The following section presents an example to illustrate this point.

Example 3—The data in Table 3 were simulated under the conditions of the no-covariate case. Specifically, the 10 observations with $y = 1$ and the 10 observations with $y = 0$ were generated randomly from N (22, 4) and N (16, 4) distributions, respectively. From (8), this corresponds to a true OR of $e^{(22-16)/4} = 4.48$ for a unit increase in X.

Traditional logistic regression often fails under such conditions, as it does for the data in Table 3, because the X values essentially display no overlap across the values of y. In particular, the LOGISTIC procedure of SAS (SAS Institute, Inc. 2004) produces an error message indicating “Quasi-complete separation,” and the estimated OR at the last iteration balloons toward infinity. Results based on exact logistic regression as implemented via the LOGISTIC procedure are not much better, despite indicating a highly significant association between Y and X ($p < 0.001$). The estimated ln(OR) via exact logistic regression is 12.53, with reported 95% confidence limits (−15, 183) that are essentially useless and stand in contrast to the significance result. Table 4 provides the results based on the discriminant function analysis approach [(9) and (11)]. The estimated ORs and 95% CIs are quite reasonable in both of these cases, with the true value of 4.48 contained in each interval. This example thus illustrates a case in which taking advantage of an assumed conditional distribution of X given $Y = y$ via the discriminant function approach may be the only way to estimate the OR with reasonable precision.

3. SIMULATION STUDIES

This section details simulation studies to evaluate the odds ratio estimators derived based on the discriminant function approach, relative to those based on traditional logistic regression.

3.1 Without Adjustment for Other Covariates

The top section of Table 5 summarizes 2000 replications under the conditional normal model for $X|Y = y$, under the following conditions: $n = 50$, $n_1 = n_0 = 25$, $\mu_1 = 0.5$, $\mu_0 = 0$, and $\sigma^2 = 1$ (true OR = 1.649). Note the roughly 10% up-ward bias evident in $\ln(\widehat{\mathrm{OR}})_{\mathrm{log}}$, which is the estimate of the ln(OR) [β in model (7)] based on logistic regression. This bias is noticeably reduced for the discriminant function-based estimators $\ln(\widehat{\mathrm{OR}})_{\mathrm{samp}}$ and $\ln(\widehat{\mathrm{OR}})_{\mathrm{umvu}}$ [(9) and (11), respectively], with the UMVU estimator essentially free of any observed bias as expected. Along with these findings, there is a clear trend toward decreasing empirical standard deviations and mean estimated standard errors as we move from $\ln(\widehat{\mathrm{OR}})_{\mathrm{log}}$ to $\ln(\widehat{\mathrm{OR}})_{\mathrm{samp}}$ and finally to $\ln(\widehat{\mathrm{OR}})_{\mathrm{umvu}}$. These trends in terms of bias and efficiency are maintained when considering the corresponding OR estimators, although each displays some predictable upward bias due to the standard practice of exponentiating the log(OR) estimate. On average, the reduction in variance leads to narrower CIs via the discriminant function approach.

The second scenario in Table 5 (middle section) illustrates the same comparisons upon increasing the value of μ_1 to 1, to yield a larger true OR of 2.718. Note that the benefits of the discriminant function methods now become more pronounced, in terms of both bias and variability. In particular, the proposed estimator $\widehat{\mathrm{OR}}_{\mathrm{samp}}$ leads to a 12% decrease in average CI width relative to logistic regression, whereas the exponentiated UMVU estimator (11) yields a 21% decrease. Coverage rates of the approximate 95% CIs remain close to nominal for all three methods, and the empirical power associated with the test of $H_{0:}$ OR = 1 is similar in each case. As expected, the power associated with the standard two-sample t -test in connection with the discriminant function approach for addressing this null hypothesis is slightly higher.

The bottom section of Table 5 considers the case of $\mu_1 = 2$, $\mu_0 = 0$, and $\sigma^2 = 1$ (true OR = e^2 = 7.389). In this large true OR scenario, the advantages over traditional logistic regression really begin to stand out. Although the SAS LOGISTIC procedure declared a failure to converge in only 1 of the 2000 simulation runs, extremely large OR estimates and wide CIs were common. The results displayed for logistic regression are based on dropping the 17 largest OR estimates, each of which was >500. Even after this concession we see marked bias in the logistic regression-based log(OR) and (especially) OR estimator, with the observed mean in the latter case nearly doubling the true OR. The improvement via the discriminant function approaches (where no runs were dropped) is marked, and in this case we also see a clearer advantage of the UMVU estimator over its sample-based counterpart. Relative to logistic regression, the UMVU approach leads to a huge decrease in mean CI width and an impressive 35% reduction in median CI width.

Further simulations (not reported) were also conducted to assess OR estimators under the null case (true OR = 1). In these cases the traditional logistic regression and discriminant function-based estimators performed quite similarly, whereas the trend toward improved precision remained qualitatively the same as in the nonnull cases.

The simulation studies summarized in Table 5, and others like them, provide us with some useful observations. First, the sample-based and UMVU log(OR) estimators [see (9) and (11)] outperform traditional logistic regression when the assumptions of the discriminant function approach are met. The advantages in terms of both bias and efficiency clearly translate to potentially large mean squared error (MSE) benefits. For example, for the middle scenario in Table 5, empirical MSE estimates are 2.86 [$= (3.226 - 2.718)^2 + 1.614^2$], 1.91, and 1.50 for the logistic-, sample-, and UMVU-based OR estimators, respectively. We also find the nearly exact match between the empirical SDs and the mean estimated standard errors for the discriminant function-based ln(OR) estimators to be an advantage over logistic regression, and indicative of the benefits of the available unbiased estimator for the exact variance (10). Importantly, these studies indicate that the benefits of the discriminant function approach become more pronounced as the true OR increases (for a fixed sample size and residual variance, σ^2). Further empirical studies reveal that this is also generally true when reducing sample size (for a fixed true OR and σ^2), and when increasing σ^2 (for a fixed true OR and sample size). For a fixed sample size, the possibility of convergence problems in logistic regression due to separation issues also tends to grow as the true OR (with fixed σ^2) or σ^2 (with fixed true OR) increases.

3.2 With Adjustment for Other Covariates

To assess the performance of the multivariable discriminant function approach based on multiple linear regression, we conducted an additional simulation study mimicking the conditions of Example 1. In particular, the fitted multiple linear regression model for the 100 observations in the birth weight dataset of Hosmer and Lemeshow (2000) was as follows:

$$\widehat{\mathrm{E}}(\mathrm{LOGLWT}) = 4.52 + 0.083Y + 0.11\mathrm{WHITE} - 0.07\mathrm{SMK} - 0.04\mathrm{HXPRELAB} + 0.26\mathrm{HXHYP} + 0.01\mathrm{AGE}, \quad (22)$$

where Y is the binary outcome (1 if birth weight ≥ 2500 g, 0 if <2500 g) and the meanings of the other variable names follow from the description under Example 1. The MSE associated with this model fit was 0.04.

For the simulation study, we generated binary covariates WHITE, SMK, HXPRELAB, and HXHYP as independent Bernoulli variates with probabilities equal to the corresponding observed proportions in the actual dataset. AGE was generated as normal with mean and variance equal to the observed sample mean and variance. The outcome Y was generated according to a logistic regression model with WHITE, SMK, HXPRELAB, HXHYP, and AGE as covariates, and with coefficients equal to those obtained when fitting such a model to the actual data. Finally, the predictor of interest (LOGLWT) was generated according to model (22), with $\sigma^2 = 0.04$. This process was repeated for 2000 independent simulated datasets, each of size $N = 200$.

Table 6 summarizes the results of the simulation study to evaluate adjusted OR estimators corresponding to the continuous predictor LOGLWT, where the true ln(OR) under the simulation conditions is 0.083/0.04 = 2.075, corresponding to an OR of 7.96. The patterns observed in Table 6 are consistent with Example 1, and with the simulation results for the no-covariate case. Specifically, we observe a positive bias in the logistic regression estimate of ln(OR). This bias is reduced when using the two discriminant function estimators, and eliminated in the case of the UMVU estimator. Mean standard errors associated with these estimators are also reduced, with the effect readily seen in terms of mean widths of CIs for the OR [e.g., the mean width based on logistic regression is 29% larger than that based on exponentiating the UMVU estimator for ln(OR)]. Whereas all three OR estimators display positive bias, this is clearly most pronounced for the logistic regression estimator. Empirical MSE estimates provided in Table 6 suggest approximate MSE efficiencies of only 63% and 57% for logistic regression when compared against the proposed sample- and UMVU-based discriminant function estimators. CI coverage rates are near nominal for all three approaches, and the power for the test of $H_{0:}$ OR = 1 is similar in each case and close to that achieved via the exact partial t -test for the parameter β in model (13).

4. DISCUSSION

The discriminant function approach is not new, and its implications for odds ratio estimation were largely recognized prior to the advent of logistic regression as the most popular tool for modeling binary outcome data. Due to historical criticisms and because that advent occurred many years ago, however, many current-day researchers may be unaware of the alternative approach to estimation that the discriminant function method provides. Most of the related prior work has focused on theoretical looks at the efficiency of the approach in the discrimination setting (Efron 1975), or efficiency comparisons for regression coefficient estimation in situations where multivariate normality of predictors is obviously violated (Halperin, Blackwelder, and Verter 1971). The current article represents a different perspective, focusing on an exposition and comparison (in terms of bias and efficiency) of OR estimators when the assumptions of normal discriminant function analysis are reasonable.

Most importantly in our view, this report is the first that we know of to highlight the ready possibility of relaxing the multivariate normality assumption when estimating the adjusted OR associated with a continuous predictor of interest via the discriminant function approach. Although the assumptions attendant with the linear regression of X on (Y, $\mathbf{C}$) in (13) with iid normal errors will not always be met in practice, there is considerable advantage to the fact that no assumptions about the distribution of the other covariates ($\mathbf{C}$) are needed. We feel this makes the multivariable approach introduced here much more defensible and practically useful than the multivariate discriminant function method, which was justifiably criticized (Halperin, Blackwelder, and Verter 1971; Hosmer and Lemeshow 2000).

Finally, we believe this to be the first report to exploit the discriminant function approach to develop UMVU estimators for crude and adjusted log(OR)s, offering further accuracy and precision benefits over logistic regression as well as over the more obvious sample-based estimators. These benefits are highlighted by our examples and simulation studies, and we find that their potential magnitude increases as true ORs become relatively large and/or as the operable residual variability (σ^2) increases.

The required assumptions for the proposed multivariable discriminant function approach are quite straightforward to assess via typical linear regression diagnostics, as partially exemplified in our discussion of Example 1. Standard techniques for this purpose [e.g., residual and partial plots to assess linearity and homoscedasticity with respect to each predictor in model (13), tests for approximate normality of residuals, influence diagnostics] are well-described in linear regression texts and readily implemented using standard software. Because linear regression diagnostics are in fact more fully understood and developed, a thorough investigation of the assumptions behind a discriminant function-based analysis is arguably as or more readily performed than one undertaken to support a logistic regression model. Nonetheless one must recognize that the assumption of normal errors is directly utilized when obtaining the discriminant function-based point estimator, thus placing a greater premium upon its assessment.

We believe the potential benefits often make the assessment of model (13) a worthwhile step when an adjusted OR associated with a particular continuous predictor is of primary interest, or at the very least whenever standard logistic regression encounters convergence problems in that setting. As a final cautionary note, however, the investigator should verify that sampling conditions align with the discriminant function approach before seriously considering its use. For example, the approach has natural appeal in retrospective or cross-sectional sampling settings, but its appropriateness may be in question for certain prospective settings in which the distribution of exposure (X) is under the control of the investigator.

Acknowledgments

This work was partially supported by National Institute of Environmental Health Sciences grant 2R01-ES012458-5 (R.H.L.) and by National Institutes of Health grant R01-MH079448-01 (Y.G.). The authors greatly appreciate valuable input from Drs. Larry Kupper and Dana Flanders, as well as the helpful comments of the editor, associate editor, and two referees.

References

Agresti, A. Categorical Data Analysis. Hoboken, NJ: Wiley; 2002.

Allison, PD. Convergence Problems in Logistic Regression. In: Altman, M.; Gill, J.; McDonald, MP., editors. Numerical Issues in Statistical Computing for the Social Scientist. Hoboken, NJ: Wiley; 2004. p. 238-252.

Anderson JA. Quadratic Logistic Discrimination. Biometrika. 1975; 62:149–154.

Anderson, TW. An Introduction to Multivariate Statistical Analysis. New York: Wiley; 1958.

Cornfield J. Joint Dependence of Risk of Coronary Heart Disease on Serum Cholesterol and Systolic Blood Pressure: A Discriminant Function Analysis. Federal Proceedings. 1962; 21:58–61.

Cox, DR. Analysis of Binary Data. London: Chapman & Hall; 1970.

Efron B. The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis. Journal of the American Statistical Association. 1975; 70:892–898.

Firth D. Bias Reduction of Maximum Likelihood Estimates. Biometrika. 1993; 80:27–38.

Greenland S. Quantitative Methods in the Review of Epidemiologic Literature. Epidemiologic Reviews. 1987; 9:1–30. [PubMed: 3678409]

Halperin M, Blackwelder WC, Verter JI. Estimation of the Multivariate Logistic Risk Function: A Comparison of the Discriminant Function and Maximum Likelihood Approaches. Journal of Chronic Diseases. 1971; 24:125–158. [PubMed: 5094226]

Hastie, T.; Tibshirani, R. Generalized Additive Models. London: Chapman & Hall; 1990.

Heinze G, Schemper M. A Solution to the Problem of Separation in Logistic Regression. Statistics in Medicine. 2002; 21:2409–2419. [PubMed: 12210625]

Hosmer, DW.; Lemeshow, S. Applied Logistic Regression. 2. New York: Wiley; 2000.

Kutner, MH.; Nachtsheim, CJ.; Neter, J.; Li, W. Applied Linear Statistical Models. 5. Boston: McGraw-Hill; 2005.

SAS Institute, Inc. SAS/STAT 9.1 User's Guide. Vol. 4. Cary, NC: SAS Institute, Inc; 2004.

Schlesselman, JJ. Case-Control Studies. New York: Oxford University Press; 1982.

Shao, J. Mathematical Statistics. 2. New York: Springer-Verlag; 2003.

Truett J, Cornfield J, Kannel W. A Multivariate Analysis of the Risk of Coronary Heart Disease in Framingham. Journal of Chronic Diseases. 1967; 20:511–524. [PubMed: 6028270]

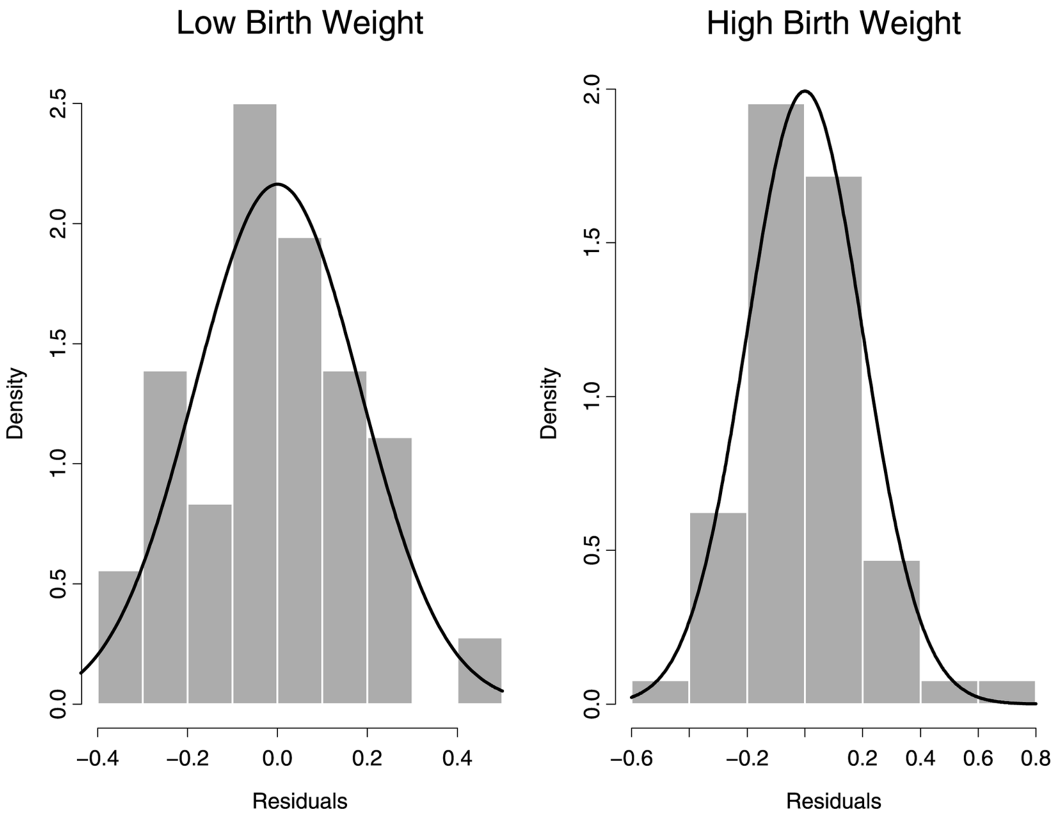

Figure 1.
Histograms of residuals. The two panels display residuals from the fit of model (13) to the data for Example 1, separately for the low (y = 0) and high (y = 1) birth weight groups. Normal curves based on the sample means and variances are overlaid, and Shapiro–Wilk tests are consistent with normality in each case.

Table 1

Analysis of birth weight data from Hosmer and Lemeshow (2000), restricting to mothers with no first trimester physician visits (*N* = 100); OR for one-unit increase in log(LWT), controlling for maternal age, race, smoking, history of premature labor, and history of hypertension.

		Multivariable discriminant function estimates	
	Logistic regression estimates	**Sample**[a]	**UMVU**[b]
$\ln(\widehat{\mathrm{OR}})$ (std. error)	2.26 (1.25)	2.08 (1.18)	2.03 (1.16)
$\widehat{\mathrm{OR}}$	9.60	7.98	7.63
95% CI for OR	(0.83, 111.79)	(0.78, 81.35)	(0.79, 74.01)

[a] See (15).

[b] See (16).

Table 2

Analysis of kyphosis data from Hastie and Tibshirani (1990) and revisited by Agresti (2002).

		Discriminant function estimates	
	Logistic regression estimates	**Sample**[a]	**UMVU**[b]
$\hat{\beta}$ (std. error)	0.060 (0.027)	0.031 (0.019)	0.027 (0.017)
$\hat{\psi}$ (std. error)[c]	−0.330 (0.156)	−0.150 (0.099)	−0.129 (0.088)

[a] See (20).

[b] See (21).

[c] Estimates and standard errors multiplied by 10^3.

Table 3

Simulated data illustrating separation of data points with $y = 1$ generated from N (22, 4); data with $y = 0$ data generated from N (16, 4).

Obs. #	*y*	*x*	Obs. #	*y*	*x*
1	0	11.07	11	1	18.74
2	0	12.15	12	1	19.87
3	0	13.54	13	1	20.42
4	0	14.93	14	1	20.57
5	0	15.37	15	1	21.16
6	0	17.21	16	1	21.66
7	0	18.33	17	1	21.71
8	0	18.44	18	1	21.92
9	0	18.54	19	1	22.98
10	0	18.76	20	1	23.85

Table 4

Results for data illustrating separation of data points based on simulated data shown in Table 3 (true OR = 4.48); note that traditional logistic regression fails to provide reliable point or interval estimates for the OR.

		Discriminant function estimates	
	Exact logistic regression estimates[c]	**Sample**[a]	**UMVU**[b]
$\ln(\widehat{\mathrm{OR}})$ (std. error)	12.53	1.05 (0.43)	0.94 (0.38)
$\widehat{\mathrm{OR}}$	$e^{12.53}$	2.86	2.55
95% CI for OR	(e^{-15}, e^{183})	(1.24, 6.61)	(1.21, 5.36)

[a] See (9).

[b] See (11).

[c] Standard error not provided by software; results unreliable.

Table 5

Simulation results assessing alternative crude OR estimators based on 2000 replications with n = 50 and $n_1 = n_0$ = 25 in each case; each OR estimator obtained by exponentiating the corresponding ln(OR) estimator.

	$\ln(\widehat{\mathrm{OR}})$		$\widehat{\mathrm{OR}}$				
	Mean (SD)	**Mean std. error**	**Mean (SD)**	**Mean CI width (median)**	**95% CI coverage**	**Rejection rate for H_0: OR = 1**	**Rejection rate: 2-sample *t*-test**
$\mu_1 = 0.5$, $\mu_0 = 0$, $\sigma^2 = 1 \Rightarrow \ln(\mathrm{OR}) = 0.5$, OR = 1.649							
Logistic	0.547 (0.333)[a]	0.319	1.832 (0.679)	2.579 (2.170)	96.2%	39.1%	
Discri-Sample	0.526 (0.319)[b]	0.319	1.783 (0.618)	2.490 (2.140)	96.5%	35.2%	42.0%
Discri-UMVU	0.504 (0.306)[c]	0.305	1.737 (0.574)	2.306 (1.997)	95.8%	35.2%	
$\mu_1 = 1$, $\mu_0 = 0$, $\sigma^2 = 1 \Rightarrow \ln(\mathrm{OR}) =1$, OR = 2.718							
Logistic	1.086 (0.391)[a]	0.376	3.226 (1.614)	5.948 (4.367)	95.7%	92.2%	
Discri-Sample	1.041 (0.366)[b]	0.368	3.045 (1.343)	5.248 (4.220)	95.7%	90.6%	93.8%
Discri-UMVU	0.997 (0.351)[c]	0.353	2.898 (1.211)	4.725 (3.852)	94.7%	90.6%	
$\mu_1 = 2$, $\mu_0 = 0$, $\sigma^2 = 1 \Rightarrow \ln(\mathrm{OR}) =2$, OR = 7.389							
Logistic[d]	2.266 (0.725)[a]	0.661	14.28 (25.81)	265.29 (23.89)	96.9%	100%	
Discri-Sample	2.090 (0.529)[b]	0.527	9.511 (7.311)	28.92 (17.94)	95.7%	100%	100%
Discri-UMVU	2.003 (0.506)[c]	0.505	8.593 (6.173)	24.29 (15.59)	94.2%	100%	

[a] The usual ML estimator for β based on the logistic regression model in (5).

[b] Natural log of OR estimator in (9).

[c] UMVU estimator for ln(OR), (11).

[d] After dropping 17 logistic regression OR estimates that exceeded 500.

Table 6

Simulation results assessing alternative adjusted OR estimators based on 2000 replications with n = 200 in each case and data generated to mimic birth weight example as described in text; true ln(OR) = 2.075 ⇒ OR = 7.96; each OR estimator obtained by exponentiating the corresponding ln(OR) estimator.

	Mean (SD) [MSE][a]	Mean std. error	Mean CI width (median)	95% CI coverage	Power for H_0: OR = 1
$\ln(\widehat{\mathrm{OR}})_{\mathrm{log}}$[b]	2.17 (0.89)	0.86	–		
$\widehat{\mathrm{OR}}_{\mathrm{log}}$	13.33 (15.93) [282.60]	–	76.53 (43.07)	94.8%	71.1%
$\ln(\widehat{\mathrm{OR}})_{\mathrm{samp}}$[c]	2.09 (0.85)	0.84	–		
$\widehat{\mathrm{OR}}_{\mathrm{samp}}$	11.75 (12.78) [177.69]	–	63.58 (37.94)	95.7%	69.2%
$\ln(\widehat{\mathrm{OR}})_{\mathrm{umvu}}$[d]	2.07 (0.84)	0.83	–		
$\widehat{\mathrm{OR}}_{\mathrm{umvu}}$	11.41 (12.23) [161.48]	–	60.51 (36.48)	95.7%	69.2%
Partial *t*-test	–	–	–	–	70.3%

[a] Empirical mean squared error (MSE) estimate = (mean OR estimate − 7.96)^2 + SD^2.

[b] The usual ML estimator for β based on the logistic regression model in (12).

[c] Natural log of OR estimator in (15).

[d] UMVU estimator for ln(OR), (16).