Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Ruoxing Li Date

Comparisons of conditional logistic regression vs. a discriminant function approach in a case-control study where matching is performed

By

Ruoxing Li
MSPH
Biostatistics and Bioinformatics

Robert H. Lyles, Ph. D
Thesis Advisor

John Hanfelt, Ph. D
Reader

Comparisons of conditional logistic regression vs. a discriminant function approach in a case-control study where matching is performed

By

Ruoxing Li
MSPH
Biostatistics and Bioinformatics

B.S
University of Electronic Science and Technology of China
2018

Thesis Advisor: Robert H. Lyles, Ph. D

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics and Bioinformatics
2020

Abstract

Comparisons of conditional logistic regression vs. a discriminant function approach in a case-control study where matching is performed
By Ruoxing Li

The logistic regression model has been widely used for estimating adjusted odds ratio associated with a binary outcome in case-control study. When matching is involved, conditional logistic regression is more commonly used to estimate the odds ratio corresponding to a continuous predictor as an alternative to standard unconditional model to decrease the bias caused by sparse data. In this thesis the discriminant function approach is suggested to generate closed-form estimators, especially under conditions involving few or small matched sets. The application of this approach, which given a multiple regression model form with the continuous predictor of interest on the outcome, includes fixed intercept effects for each matched set. It is demonstrated that the estimator based on discriminant function approach outperform the usual maximum likelihood estimator from logistic regression based on our simulation works and examples. The advantages have seen in reducing bias and width of CI for odds ratio, as well as generating reliable estimator under separation situations where logistic regression fails. Potential improvements for this study are also talked in the end of the article.

KEY WORDS: Logistic regression; Bias; Discriminant function approach.

Comparisons of conditional logistic regression vs. a discriminant function approach in a case-control study where matching is performed

By

Ruoxing Li
MSPH
Biostatistics and Bioinformatics

B.S
University of Electronic Science and Technology of China
2018

Thesis Advisor: Robert H. Lyles, Ph. D

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics and Bioinformatics
2020

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Prof. Robert H. Lyles for his great support. He is such patient, knowledgeable and helpful. Whenever I am stuck by a problem, he always brings me light. Without his guidance, I would not be able to finish this work.

I also would like to thank Prof. John Hanfelt for being my reader and providing valuable comments. The motivating example he provided gives us fresh ideas about our approach.

I also thank to Dr. Donna Brogan for providing the example dataset which is a big help to enrich the thesis, and Dr. Dane van Domelen for past work that helped motivate this research.

Last but not least, I would like to thank my faculty advisor Traci Leong for her suggestions during the two years. I also sincerely appreciate all my professors and classmates for their teaching and accompany during my study.

Contents

Introduction

In epidemiology, a case-control study is usually designed for determining whether a particular disease is affected by a factor (e.g., an exposure variable), during which, matching is an intuitive approach for adjusting potential confounders by selecting controls with similar characteristics as cases. However, if many confounders are controlled simultaneously, this can result in a large number of strata of small size and/or numerous strata containing only cases or controls which are not informative for analysis. In addition, problems for extrapolation arise when a continuous risk factor is broken into different levels to form strata (Breslow and Day 1980).

In general, the logistic regression model provides a common tool for estimating adjusted odds ratios associated with a binary outcome. When matching is involved, conditional logistic regression is typically applied to decrease biases incurred by the standard unconditional model due to sparse data. However, researchers have found that conditional logistic regression can also suffer from bias, particularly when the model involves many covariates or very few matched sets (Greenland, Schwartzbaum and Finkle 2000). Under this situation, even a medium-size sample can result in infinite parameter estimates (Heinze and Schemper 2002).

Historically, the discriminant function approach to odds ratio estimation preceded the development of logistic regression (Cornfield 1962). However, criticism for the

discriminant function approach arose, as disadvantages were pointed out when the joint normality assumption of independent variables does not hold. This would generate bias for the estimator, and logistic regression became more popular for its robustness in this situation (Halperin 1971). But recently an adaptation of the approach has been considered, where the focus is on estimating the covariate-adjusted odds ratio corresponding to a single continuous exposure variable of interest (Lyles, Guo and Hill 2009). Under a more reasonable and testable univariate normality assumption, it was demonstrated that this approach can yield an unbiased log odds ratio estimator that is more precise than the estimator from standard logistic regression, and that this estimator avoids the risk of failure to converge due to the separation problems (Allison 2008, Heinze and Schemper 2002, Neuenschwander et al. 2000). The new look at the approach provided examples which confirmed the discriminant function method produced reliable estimates in no-covariate and covariate-adjusted cases and requires less strict assumptions relative to traditional discriminant function analysis in the latter case.

In this thesis, we consider the potential of the discriminant function approach as an alternative to conditional logistic regression, particularly in settings involving few or small matched sets. For situations in which matching is used to adjust for more than one potential confounder and there is a primary continuous exposure is of interest, we use simulations and a real example to compare performance of the discriminant

function approach with that of traditional conditional and unconditional logistic regression. We include a uniformly minimum variance unbiased estimator for the adjusted odds ratio (Lyles et al., 2009), which becomes available using the discriminant function method and potentially leads to narrower confidence intervals.

Methods

For a binary outcome Y, and a continuous explanatory variable X_1, the odds ratio corresponding to a unit increase in X_1 is defined as:

$$OR = \frac{\Pr(Y = 1|X_1 = x + 1)\,/\Pr(Y = 0|X_1 = x + 1)}{\Pr(Y = 1|X_1 = x)\,/\Pr(Y = 0|X_1 = x)} \tag{1}$$

Focusing on a case-control study, OR could also be written in terms of a conditional probability density function:

$$OR = \frac{\mathrm{f}_{\mathrm{X}_1|\mathrm{Y}=1}(x + 1)/\,\mathrm{f}_{\mathrm{X}_1|\mathrm{Y}=1}(x)}{\mathrm{f}_{\mathrm{X}_1|\mathrm{Y}=0}(x + 1)/\,\mathrm{f}_{\mathrm{X}_1|\mathrm{Y}=0}(x)} \tag{2}$$

When accounting for other potential risk factors or confounders, $\boldsymbol{C} = (X_2,\ X_3, \ldots,\ X_n)'$, formulae (1) and (2) could be written as：

$$OR = \frac{\Pr(Y = 1|X_1 = x + 1, \boldsymbol{C})/\Pr(Y = 0|X_1 = x + 1, \boldsymbol{C})}{\Pr(Y = 1|X_1 = x, \boldsymbol{C})/\Pr(Y = 0|X_1 = x, \boldsymbol{C})} \tag{3}$$

and

$$OR = \frac{\mathrm{f}_{\mathrm{X}_1|\mathrm{Y}=1,\mathbf{C}}(x + 1)/\,\mathrm{f}_{\mathrm{X}_1|\mathrm{Y}=1,\mathbf{C}}(x)}{\mathrm{f}_{\mathrm{X}_1|\mathrm{Y}=0,\mathbf{C}}(x + 1)/\,\mathrm{f}_{\mathrm{X}_1|\mathrm{Y}=0,\mathbf{C}}(x)} \tag{4}$$

Standard logistic regression

Sir Francis Galton developed regression analysis in the late 19th century (Kutner, Nachtsheim, Neter and Li, 2004). Now regression methods are widely used to describe the statistical relationships between an outcome variable and one or more explanatory variables (covariates). The linear regression model is probably the most familiar, where the response variable is assumed to be continuous. However, our outcome of interest outcome is often discrete or binary, like disease status, and of course problems would arise if we use the linear regression model. With a binary outcome (Y), the natural model is to allow covariates to impact the Bernoulli probability for each experimental unit. A clear problem with a multiple linear regression model for Y that the assumption of normally distributed error items cannot be tenable. For a given set of predictor (X) variables, each error item, $\epsilon_i = Y_i - (\beta_0 + \sum_{i=1}^{p} \beta_i X_i)$, can take on only two values (for Y_i equal to 1 or 0). A second problem is that error items would not have equal variances, i.e., $\sigma^2\{\epsilon_i\} = (\beta_0 + \sum_{i=1}^{p} \beta_i X_i)(1 - \beta_0 - \sum_{i=1}^{p} \beta_i X_i)$. According to this formula, its value would depend on X_i. Another problem is that when our outcome variable is a binary variable, the response function would be the probability that $Y_i = 1$ or 0, and should between 0 and 1. However, linear regression does not impose this constraint. Among others, these difficulties in applying linear regression led to the development of logistic regression, which is the most often used model in data analysis when describing the

relationship between a binary outcome and other explanatory variables.

Logistic regression derives its name from the logit transformation of a probability. Given a probability Pr, the logit transformation of Pr is

$$logit(Pr) = log\frac{Pr}{1-Pr} \tag{5}$$

With $Y = 1$ or 0, given a set of explanatory variables $X = (X_1, X_2, \dots, X_p)$, the standard logistic model is generally defined as

$$logit\{Pr(Y = 1|X)\} = \beta_0 + \sum_{i=1}^{p} \beta_i X_i \tag{6}$$

or,

$$Pr(Y = 1|X) = \frac{\exp\left(\beta_0 + \sum_{i=1}^{p} \beta_i X_i\right)}{1 + \exp\left(\beta_0 + \sum_{i=1}^{p} \beta_i X_i\right)} \tag{7}$$

The logistic regression method has seen wide use to make smooth estimates for risk ratios or odds ratios associated with model coefficients in epidemiology studies, such as cohort and case-control studies (Breslow and Day 1980). If X_1 is still our continuous predictor of interest, then the OR corresponding to a unit increase is e^{β_1}. The estimator could be calculated based on the MLE of β_1.

Conditional logistic regression for matched case-control studies

In a case-control study, stratification provides a common tool for control of

confounding, which is flexible to be introduced during the study design or at the analysis stage. Typically, we would seek to include the variables created for stratification to generalize the regression model. The new model is written as:

$$logit\{Pr_k\} = \beta_{0k} + \sum_{i=1}^{p} \beta_i X_i \quad (8)$$

where Pr_k is the probability that Y=1 in stratum k.

A "perfect" form of stratification is where each case is matched to several controls with similar values of confounding variables. A simple example is 1:1 pair matching, where only one case and one control comprise each stratum. If there are k pairs with p covariates, $p + k$ parameters need to be estimated using the sample of size $2k$. We can notice that the number of parameters would increase as the sample size increases. The optimality properties of the method of maximum likelihood do not work well under this situation and standard logistic regression could yield serious bias (Hosmer, Lemeshow and Sturdivant 2013). Breslow and Day (1980) demonstrate that in a matched study where each stratum has a matched case-control pair with a single binary predictor, the OR estimator given by standard logistic regression converges to the square of the correct value. In part, that is why it is important to consider conditional logistic regression for matched case-control data. Hosmer, Lemeshow and Sturdivant (2013) review this approach, which involves using conditional likelihood analysis to create a likelihood function generating maximum

likelihood estimators for our parameters associated with covariates, by eliminating the nuisance parameters indexing each stratum.

Consider a 1-M matched case-control study which means each case is matched to M controls. Or more generally, suppose there are k matched sets, and the ith stratum contains M_i controls. Denote by X_{i0} the column vector of exposures for the case in the stratum and by X_{ij} the vector of exposures for the jth control. Then we can write the conditional likelihood as:

$$\prod_{i=1}^{k} \frac{e^{\boldsymbol{\beta}' X_{i0}}}{\sum_{j=0}^{M_i} e^{\boldsymbol{\beta}' X_{ij}}} = \prod_{i=1}^{k} \frac{1}{1 + \sum_{j=1}^{M_i} e^{\boldsymbol{\beta}' (X_{ij} - X_{i0})}} \tag{9}$$

As for 1:1 pair matching, the above formula could be simplified to

$$\prod_{i}^{k} \frac{1}{1 + e^{\boldsymbol{\beta}' (X_{ik} - X_{i0})}} \tag{10}$$

Based on these formulas, we could get estimators of the vector $\boldsymbol{\beta}$. If any of the variables in X were used for matching, it would contribute no information to the likelihood and a corresponding parameter cannot be estimated. Moreover, a stratum also makes no contribution if it contains no case, or if the case and all controls have the same *X* values.

Discriminant function approach

A multivariable discriminant function approach, where multivariate normality is not

required, was suggested to be an alternative to the standard logistic regression method to give estimators of the odds ratio associated with an individual continuous exposure of interest (Lyles et al., 2009). Letting $\boldsymbol{C}$ denote controlled risk factors and with X_1 continuing to represent our continuous exposure, a discriminant function approach can be based on the following multiple linear regression model:

$$E(X_1|Y = y, \boldsymbol{C}) = \beta_0^* + \beta_1^* y + \gamma^{*\prime}\boldsymbol{C} \tag{11}$$

Assuming the independent and identically distributed error items $\varepsilon \sim N(0, \sigma^2)$, (11) yields the following expression for the odds ratio associating a unit increase in X_1 with Y:

$$OR = e^{\beta_1^*/\sigma^2} \tag{12}$$

and yields an estimator as the following:

$$\widehat{OR} = e^{\widehat{\beta_1^*}/MSE} \tag{13}$$

where $\widehat{\beta_1^*}$ is the ordinary least squares estimator of β_1^*, and MSE is the usual residual variance estimator based on (11).

An exact variance estimator for $\ln(\widehat{OR})$ based on the moment properties of the chi-squared distribution and the independence of the random variables $\widehat{\beta_1^*}$ and MSE can be derived as follows (Lyles et al., 2009):

$$var[\ln(\widehat{OR})] = \left(\frac{n-T-2}{n-T-4}\right)^2 (\sigma^2)^{-2}\left[\left(\frac{n-T-4}{n-T-6}\right) var(\widehat{\beta_1^*}) + \frac{2\beta_1^2}{n-T-6}\right] \quad (14)$$

where n is the number of subjects.

An unbiased estimator was suggested (Lyles et al., 2009) as:

$$\widehat{var}[\ln(\widehat{OR})] = \left(\frac{n-T-2}{n-T-4}\right)^2 MSE^{-2}\left[var(\widehat{\beta_1^*}) + \frac{2\widehat{\beta_1^*}^2}{n-T-2}\right] \quad (15)$$

where T is the dimension of the covariate vector $\boldsymbol{C}$. As the prior authors noted, it is also of interest to note that a uniformly minimum variance unbiased (UMVU) estimator of $\ln(OR)$ is available in this framework. This estimator of $\ln(OR)$ and its variance can be written as:

$$\ln(\widehat{OR})_{umvu} = \left(\frac{n-T-4}{n-T-2}\right)\widehat{\beta_1^*}/MSE \quad (16)$$

$$\widehat{var}\left[\ln(\widehat{OR})_{umvu}\right] = \left(\frac{n-T-4}{n-T-2}\right)^2 \widehat{var}[\ln(\widehat{OR})] \quad (17)$$

where $\widehat{var}[\ln(\widehat{OR})]$ in (17) is given as (15).

In a matched case-control study, specially we would involve fixed intercept effects for each dataset based on (11). Continuing on the case-control study with k matched sets, here we assume the ith stratum contains M_i subjects, model (11) could be modified as:

$$E(X_{ij}|Y = y, \boldsymbol{C}) = (\beta_0^* + a_i) + \beta_1^* y_{ij} + \gamma^{*\prime}\boldsymbol{C}_{ij} \quad (18)$$

Where a_is are the fixed effects for each matched set, $j = 1, \dots, M_i$.

With model (18) we could easily handle the large number of fixed effects. For computations with this model, we could use the "class" statement for matched set index 'i' and the "no intercept" option in SAS Proc GLM.

Examples

Example 1. Consider a study of risk factors for high blood pressure in Georgia. The dataset provided by Drs. Donna Brogan and John Hanfelt was generated from Georgia High Blood Pressure Survey (Brogan 1985). Conducted in Georgia, the complex sample survey interviewed over 6000 adults and aimed to estimate how many noninstitutionalized Georgia adults having blood pressure, hypertension, or related health conditions. Residents in prisons, convents, military barracks, college dormitories and nursing homes were not included in the survey. We analyzed data from 5902 subjects, including hypertensive outcomes and multiple demographic and risk factor variables. The individuals were matched according to small geographic residential regions, yielding a total of 465 strata. We defined hypertensive subjects as those who had mean readings of their last two DBP (diastolic blood pressure) measurements larger than 95 mmHg. Thus, the outcome Y characterizes subjects' blood pressure (1 if hypertensive, 0 otherwise). After some preliminary modeling, we selected two continuous risk factors (age and BMI) to be of primary interest, and

elected to control for 3 other baseline covariates. These included indicators of daily alcohol use (yes/no), black ethnicity (yes/no), and a three-category variable characterizing the habit of adding salt to food. The latter was indexed by two yes/no dummy variables ("adds significant salt" and "adds some salt").

Table 1 shows the estimators corresponding to age and BMI based on conditional logistic regression and the discriminant function approach. For the latter, UMVU estimators are also included in the table.

Table1 Analysis of Georgia blood pressure data. OR corresponding to one unit increase in age/BMI (controlling for race, alcohol use, salt consumption, and BMI or age)

		Conditional Logistic regression	Multivariable discriminant analysis	
			Discriminant function	UMVU
Age	$\widehat{OR}$	1.060	1.064	1.064
	$\ln(\widehat{OR})$(Std error)	0.058 (0.003)	0.062 (0.003)	0.062 (0.003)
	95% CI for OR	(1.054, 1.065)	(1.058, 1.070)	(1.058, 1.070)
BMI	$\widehat{OR}$	1.115	1.115	1.115
	$\ln(\widehat{OR})$(Std error)	0.109 (0.008)	0.109 (0.008)	0.109 (0.008)
	95% CI for OR	(1.098, 1.132)	(1.098, 1.133)	(1.098, 1.133)

During the analysis, we noticed that 59 strata contained no cases. These are called uninformative strata and make no contribution to the conditional logistic regression model fitting process. In order to investigate how the discriminant function approach handles such non-informative strata, we deleted those 59 strata (corresponding to a total of 530 individuals) and repeated the analysis. Table2 shows the estimators corresponding to age and BMI based on the discriminant function approach after

deleting uninformative strata.

Table2 Analysis of Georgia blood pressure data. OR corresponding to one unit increase in age/BMI. (controlling for race, alcohol use, salt consumption, and BMI or age)

		Multivariable discriminant analysis	
		Discriminant Function	UMVU
Age	$\widehat{OR}$	1.063	1.063
	$\ln(\widehat{OR})$(Std error)	0.061(0.003)	0.061(0.003)
	95% CI for OR	(1.057, 1.069)	(1.057, 1.067)
BMI	$\widehat{OR}$	1.118	1.118
	$\ln(\widehat{OR})$(Std error)	0.111 (0.008)	0.111 (0.008)
	95% CI for OR	(1.100, 1.136)	(1.100, 1.136)

In general, we note that the estimators from the different approaches in Table1 are almost identical. Compared to Table2, the estimators based on the discriminant function approach changed slightly, although the difference was essentially negligible. Nevertheless, this suggests that (unlike conditional logistic regression), the discriminant function approach does make some use of traditionally non-informative strata.

Example 2. A problem in standard logistic regression modeling is the possibility of a failure of convergence, commonly caused by complete or quasi-complete separation (Allison 2004). Penalized maximum likelihood has been demonstrated as a useful approach to deal with this problem (Firth 1995; Heinze and Schemper 2002), and exact logistic regression offers another possibility (Mehta and Patel 1995). However, common software may not always allow access to the penalized maximum likelihood approach, and exact logistic regression might not work when the dataset becomes

very large. It is encouraging to see that the discriminant function approach works well as an alternative to standard logistic regression when separation problems occur (Lyles et al., 2009). Our goal in this section is to demonstrate that the separation problem can also plague conditional logistic regression, and that the discriminant function continues to offer a feasible remedy.

Table3. Simulated data illustrating separation problems

Pair #	Y	X	Y	X
1	0	5.0012	1	15.6853
2	0	4.2045	1	14.9077
3	0	6.6476	1	9.6886
4	0	3.3183	1	11.1500
5	0	5.8709	1	13.6066
6	0	4.9812	1	13.8029
7	0	1.4460	1	12.4873
8	0	8.9625	1	11.9442
9	0	6.1540	1	11.0878
10	0	7.7568	1	14.2354
11	0	4.8829	1	12.4879
12	0	0.1809	1	13.2911
13	0	5.7517	1	12.0722
14	0	2.1629	1	13.8244
15	0	3.0822	1	11.3953

The data in table 3 were simulated under the condition of no covariates, where separation is occurring due to a lack of overlap in the X distribution for cases and controls. The exposure X was generated from $N(4,4)$ and $N(13,4)$ distributions, respectively, when Y equals to 0 and 1. The true OR could be calculated from $e^{(13-4)/4} = 9.488$. Conditional logistic regression failed under this situation, and gave estimates approaching infinity. Results based on standard logistic regression were not

reasonable, with a 95% CI for the OR of (1.008, 1.199). The top section of table 4 shows the estimates based on the discriminant function approach. Note that the true OR was contained in the 95% CIs which is a noticeable improvement.

Table4. Results based on dataset with separation problems. True OR=9.488, and ln(OR)=2.25.

	Logistic regression	Conditional Logistic regression	Multivariable discriminant analysis	
			Discriminant Function	UMVU
$\widehat{OR}$	1.100	Infinity	5.820	3.696
95% CI for OR	(1.008, 1.199)	-	(1.359, 24.931)	(1.300, 15.747)
$\ln(\widehat{OR})$(Std error)	0.095(0.044)	-	1.761(0.742)	1.509(0.636)
$\widehat{OR}$	1.117	Infinity	10.212	9.510
$\ln(\widehat{OR})$(Std error)	0.111(0.004)	Infinity	2.282(0.276)	2.252(0.272)
Mean CI width (median)	0.066(0.065)	Infinity (18.403)	11.495(10.185)	10.968(9.740)
95% CI coverage for OR (%)	0	8.0	93.7	93.4

The bottom section of Table 4 summarizes a simulation under the same conditions in which we increased the dataset size to 300, with 1000 replications. Note that estimates based on the discriminant function approach (particularly based on the UMVU approach) now show very little bias, while standard and conditional logistic regression still failed.

Simulation Studies and Results

In our hypertension example with such a large sample, conditional logistic regression performed essentially equivalently to the discriminant function approach. We suspected that if we reduced the sample size, the differences between the estimators would get larger. Thus, we conducted simulations loosely mimicking the real dataset

and assuming smaller overall samples of size 200 or 500. All binary covariates (representing black ethnicity, significant ("salt_add") and moderate ("saltsome") use of salt, and daily alcohol use) were generated as independent Bernoulli variates with probabilities equal to the corresponding observed proportions in the actual dataset. The binary outcome Y ("hypertension") was generated as Bernoulli with 20% prevalence. The small geographic region stratification variable ("geo") was generated as normal with mean and variance equal to 0 and 0.25, respectively. Age was generated according to a fitted multivariable linear model in the Georgia blood pressure dataset, with random errors distributed as normal with mean 0 and variance 300:

$$\hat{E}(age) = 48.9 - 9.2salt\_add - 5.5saltsome - 0.6alc - 3.2black$$

BMI was generated according to the following discriminant function model, also based on estimates obtained using the real data:

$$BMI = 20 + 2.49Y + .03age - .84alt\_add$$
$$-.31saltsome - .56alc + 1.14black + geo + \epsilon$$

where Y is the outcome (1 if hypertensive, 0 otherwise), and $\epsilon \sim N(0, 22.79)$. This process was repeated for 1000 independent simulated datasets. Note that the latter model dictates a true ln(OR) of $2.49/22.79 = 0.109$.

Table5 Simulation mimicking Georgia blood pressure data (N=500/200, 1000 iterations). OR

corresponding to one unit increase in BMI. Controlling some baseline covariates, such as race, et al. True OR=1.115, ln(OR)=0.109.

		Logistic regression	Conditional Logistic regression	Multivariable discriminant analysis	
				Discriminant function	UMVU
Size 200	$\ln(\widehat{OR})$(Std error)	0.226(0.096)	0.136 (0.085)	0.110 (0.058)	0.107 (0.056)
	$\widehat{OR}$	1.270	1.150	1.118	1.114
	95% CI coverage for OR (%)	77.0	96.0	88.4	88.0
	Mean CI width (median)	0.525(0.415)	0.352(0.302)	0.210 (0.208)	0.218 (0.217)
Size 500	$\ln(\widehat{OR})$(Std error)	0.184(0.050)	0.118(0.040)	0.110(0.034)	0.109(0.034)
	$\widehat{OR}$	1.205	1.126	1.117	1.115
	95% CI coverage for OR (%)	66.2	94.2	94.3	94.0
	Mean CI width (median)	0.238(0.230)	0.175(0.171)	0.155(0.154)	0.153(0.153)

Table5 summarizes 1000 replications for each case when sample size is set to 200 and 500. The estimators of $\ln(OR)$ based on discriminant analysis (both UMVU and discriminant function estimators) show less bias compared to estimators based on conditional logistic regression (true $\ln(OR)=0.109$). The mean standard errors for $\ln(OR)$ are also noticeably reduced when we move from conditional logistic regression to discriminant analysis. Bias reduction is maintained when we consider OR estimators, along with variance reduction leading to narrower CIs via discriminant analysis.

In order to assess the performance of the discriminant function approach more comprehensively, another simulation was conducted to mimic the common practice of case-control study matching based on gender and age. The outcome Y (disease status) and gender were generated as Bernoulli variables with probabilities 0.25 and 0.5 respectively. We assumed the age of patients was normally distributed with the

mean 60 and the standard deviation 10. Our continuous predictor X was generated based on the following model:

$$X = 4 + Y + 0.05age - 0.5gender + \epsilon$$

where $\epsilon \sim N(0, 0.5)$. Note that this model dictates a true ln(OR) of $1/0.5 = 2$.

Sizes of dataset and matched sets varied with our settings (shown in the following table). Every time the dataset generated, objects were matched based on their gender and ages using K-means clustering method (Mitchell Lyles et al. 2014). Although desired number of matched sets could be automatically generated through this way, case number in each set differs: some may have more than one case while some having none.

Table6 Results of simulation to assess performance of the discriminant function. This process was repeated for 4000 independent simulated datasets with matched sets of varying size obtained via k-means clustering (see text). When 'Correlation' is designated, this means 'Age' was generated as linearly associated with 'Y'. True OR=7.389, ln(OR)=2.

Situation	Approach	$\ln(\widehat{\mathrm{OR}})$ (Mean estimated SE)	$\widehat{\mathrm{OR}}$	Mean CI width (Median)	95% CI coverage (%)
#Clusters: 50 Dataset size: 300	Logistic regression	2.241(0.914)	12.443	21.68(16.76)	67.18
	Conditional Logistic regression	2.037(0.325)	8.110	11.63(9.79)	95.35
	Discriminant function	2.013(0.275)	7.783	9.12(8.40)	95.33
	UMVU	1.986(0.272)	7.191	8.89(8.20)	95.18
#Clusters: 50 Dataset size: 200	Logistic regression	2.763(0.935)	26.269	232.12(34.70)	70.07
	Conditional Logistic regression	2.100(0.480)	9.402	24.42(13.68)	95.65

Situation	Approach	$\ln(\widehat{OR})$ (Mean estimated SE)	$\widehat{OR}$	Mean CI width (Median)	95% CI coverage (%)
	Discriminant function	2.032(0.365)	8.186	13.06(11.20)	95.00
	UMVU	2.005(0.360)	7.418	12.48(10.73)	94.40
#Clusters: 50 Dataset size: 100	Logistic regression	3.983(1.879)	1107.32	$>10^5$(335.17)	88.95
	Conditional Logistic regression	2.578(2.495)	3.159	$>10^5$(42.49)	95.70
	Discriminant function	2.087(0.665)	10.538	62.83(24.19)	95.80
	UMVU	2.059(0.656)	8.496	49.16(20.95)	94.90
#Clusters: 30 Dataset size: 300	Logistic regression	1.416(1.209)	7.130	9.82(8.89)	56.95
	Conditional Logistic regression	2.025(0.303)	7.956	10.15(8.84)	95.28
	Discriminant function	2.006(0.266)	7.709	8.56(7.91)	95.40
	UMVU	1.983(0.262)	7.186	8.35(7.73)	95.25
#Clusters: 30 Dataset size: 200	Logistic regression	1.759(1.282)	10.774	24.19(15.38)	62.38
	Conditional Logistic regression	2.053(0.413)	8.579	15.65(11.77)	96.02
	Discriminant function	2.014(0.349)	7.988	11.64(10.15)	95.58
	UMVU	1.990(0.344)	7.333	11.19(9,78)	95.48
#Clusters: 30 Dataset size: 100	Logistic regression	2.924(1.742)	1028.43	$>10^5$(68.00)	77.90
	Conditional Logistic regression	2.258(0.998)	12850.53	$>10^5$(24.36)	96.18
	Discriminant function	2.061(0.557)	9.370	27.63(17.88)	95.70
	UMVU	2.036(0.550)	8.056	24.53(16.22)	94.98
#Clusters: 30 Dataset size: 300 Correlation	Logistic regression	2.238(0.579)	10.714	17.61(14.04)	85.65
	Conditional Logistic regression	2.042(0.330)	8.167	11.87(10.04)	95.70
	Discriminant function	2.016(0.275)	7.806	9.30(8.64)	95.43
	UMVU	1.993(0.271)	7.244	9.08(8.44)	95.18

Situation	Approach	$\ln(\widehat{OR})$ (Mean estimated SE)	$\widehat{OR}$	Mean CI width (Median)	95% CI coverage (%)
#Clusters: 30 Dataset size: 200 Correlation	Logistic regression	2.450(0.731)	16.164	50.53(22.94)	84.90
	Conditional Logistic regression	2.075(0.454)	9.055	20.20(13.27)	95.48
	Discriminant function	2.023(0.361)	8.099	12.74(11.04)	94.98
	UMVU	1.999(0.357)	7.384	12.24(10.64)	94.78
#Clusters: 30 Dataset size: 100 Correlation	Logistic regression	3.316(1.624)	36116 56.58	$>10^5$(103.93)	88.83
	Conditional Logistic regression	2.370(1.729)	$>10^5$	$>10^5$(29.29)	96.58
	Discriminant function	2.060(0.571)	9.464	29.77(19.02)	96.10
	UMVU	2.036(0.564)	7.985	26.41(17.27)	9

Table 6 summarizes the simulation results and provides us useful observations. Starting from the first situation (50 clusters and the sample size=300), ln(OR) estimators based on the discriminant function approach and UMVU estimators noticeably outperform estimators based on logistic regression, reducing bias and standard errors. Similar benefits can be observed with respect to the OR estimates, and variance reduction leads to narrower CIs via discriminant analysis. When more information is contained in each cluster (decreasing the number of clusters), the ln(OR) estimators based on the discriminant function approach and UMVU estimators appear to have smaller standard errors, along with 95% CIs for the OR becoming narrower. When sample size and number of clusters go down, logistic regression failed to give reasonable results sometimes (e.g., 30 clusters and the sample size=100), and the discriminant function approach shows marked benefits

under this situation. As we expected, the generation of crude correlation between Y and age had some negative influences on the ln(OR) estimates in terms of precision. Nevertheless, the advantages of the discriminant function over standard and conditional logistic regression were still maintained.

Discussion

The discriminant function approach is a historical one and its application to odds ratio estimation preceded the development of logistic regression. As historical criticism for the discriminant function approach exists, it has perhaps been somewhat underestimated as an alternative to logistic regression. Recently, a fresh look at the approach highlighted the ready possibility of relaxing the multivariate normality assumption when estimating the covariate-adjusted odds ratio corresponding to a single continuous exposure variable of interest (Lyles, Guo and Hill 2009). They also derived UMVU estimators for adjusted odds ratios from the approach. Based on this newfound variation on the historical discriminant function approach, the current thesis specifically focuses on estimating the adjusted OR associated with a continuous predictor in situations (e.g., a matched case-control study) in which conditional logistic regression would be the typical approach.

The discriminant function approach is not a replacement to logistic regression, but an alternative way to give a more precise estimator of odds ratio (particularly when

logistic regression is unstable or fails due to separation problems). As we could see from Example 1, when the sample size is large enough the conditional logistic regression model has little drawback and the discriminant function approach would not show much benefit. But in many real studies, the collected data are comprised of fewer observations, or complete records, for each patient. Problems can pop up, like our example 2 or the first simulation, leading to the nonexistence of the MLE estimate based on conditional logistic regression. In these instances, the discriminant function approach would clearly show performance benefits.

Overall, the examples and simulation work in this thesis have shown the discriminant function to be a useful tool to estimate covariate-adjusted odds ratio relating to a continuous predictor in matched case-control studies. Its required assumptions are straightforward to assess (though normality of errors is required), which makes it potentially of practical use. Example 2 shows that discriminant function uses traditionally non-informative strata, which is not the case with conditional logistic regression. In later simulation work which mimicked our example data, we only focused on changes in the sample size. Further work could be done to drop non-informative strata during the simulation process to assess the estimators based on the discriminant function approach. Another potential improvement could be done to better assess the discriminant function approach in comparison to using penalized maximum likelihood to deal with separation problems in the conditional logistic

regression setting (Heize and Puhr 2010). Although we obtained a SAS macro from the authors (G. Heinze, personal communication), we were unable to implement it at the time of this writing.

Reference

[1] Lyles, R. H., Guo, Y., & Hill, A. N. (2009). A fresh look at the discriminant function approach for estimating crude or adjusted odds ratios. *American Statistician*, *63*(4), 320–327.

[2] Halperin, M., Blackwelder, W. C., & Verter, J. I. (1971). Estimation of the multivariate logistic risk function: A comparison of the discriminant function and maximum likelihood approaches. *Journal of Chronic Diseases*, *24*(2–3), 125–158.

[3] Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, *21*(16), 2409–2419.

[4] Heinze, G., & Puhr, R. (2010). Bias‐reduced and separation‐proof conditional logistic regression with small or sparse data sets. Statistics in medicine, 29(7‐8), 770-777.

[5] Neuenschwander, B. E., Zwahlen, M., & Greenland, S. (2000). Problems due to small samples and sparse data in conditional logistic regression analysis. *American Journal of Epidemiology*, *152*(7), 688–689.

[6] Firth, D. (1995). Bias reduction of maximum likelihood estimates. *Biometrika*, *82*(3),

[7] Allison, P. (2004). Convergence Problems in Logistic Regression.

[8] Breslow NE, & Day NE. Statistical methods in cancer research.Vol 1. The analysis of case-control studies. (IARC ScientificPublication no. 32). Lyon, France: International Agency forResearch on Cancer, 1980.

[9] David W. Hosmer, Jr., Stanley Lemeshow, & Rodney X. Sturdivant. (2013). Applied Logistic regression. Third edition.

[10] Emily M. Mitchell, Robert H. Lyles, Amita K. Manatunga, Neil J. Perkins, & Enrique F. Schistermana. (2014). A highly efficient design strategy for regression with outcome pooling. *Statistics in Medicine,*

[11] Mehta, C. R., &Patel, N. R. (1995). Exact logistic regression: theory and examples. Statistics in medicine, 14(19), 2143-2160.

[12] Cornfield, J. (1962). Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis. In Fed Proc (Vol. 21, No. 4, pp. 58-61).

[13] Brogan, D. (1985) 1983 Georgia high blood pressure survey. Technical Report. Department of Biostatistics, Emory University, Atlanta