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:

_____ _____

Huiqing Sun Date

Evaluating the performance of maximum likelihood estimation in a discriminant function framework to account for non-detectable exposure measurements in matched case-control studies

By

Huiqing Sun

Master of Science in Public Health

Biostatistics and Bioinformatics

_____

Robert H. Lyles, Ph. D

Thesis Advisor

_____

John Hanfelt, Ph. D

Reader

Evaluating the performance of maximum likelihood estimation in a discriminant function framework to account for non-detectable exposure measurements in matched case-control studies

By

Huiqing Sun

B.A.

University of Minnesota, Twin Cities

2021

Thesis Committee Chair: 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

2023

Abstract

Evaluating the performance of maximum likelihood estimation in a discriminant function framework to account for non-detectable exposure measurements in matched case-control studies

By

Huiqing Sun

For matched case-control studies, conditional logistic regression is the typical approach to be applied. With multivariate normality unnecessary, it is possible to investigate the discriminant function approach as an alternative to conditional logistic regression in matched case-control studies. Particularly when few or small matched sets were involved, the approach was found to give a more precise and unbiased estimator of the log odds ratio associated with a continuous predictor of primary interest. The most common method in environmental chemistry to deal with non-detects is simply substituting the detection limit or some fraction of it in place of the unknown exposure, which very likely give an estimator that far from the true value. This thesis specifically focuses on evaluating the performance of maximum likelihood estimation in a discriminant function framework to account for non-detectable exposure measurements in matched case-control studies. Compared with the expedient approach of plugging in the detection limit for non-detects and using regular or conditional logistic regression, the adjusted maximum likelihood estimation based on the discriminant function analysis shows less bias and the mean standard errors for ln(OR) are also noticeably reduced. Potential improvements could be sought to better adjust the MLE of the residual variance when using maximum likelihood accounting for nondetectable exposures in matched case-control studies.

KEY WORDS: Discriminant function approach; Maximum likelihood estimation; Logistic regression; Non-detects; Bias.

Evaluating the performance of maximum likelihood estimation in a discriminant function framework to account for non-detectable exposure measurements in matched case-control studies

By

Huiqing Sun

B.A.

University of Minnesota, Twin Cities

2021

Thesis Committee Chair: Robert H. Lyles, Ph. D

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

2023

Acknowledgements

First and foremost, I would like to thank my thesis supervisor Prof. Robert H. Lyles for his unwavering support and guidance throughout the entire process. His expertise, patience, and encouragement have been instrumental in shaping my research and writing skills.

I am grateful to Prof. John Hanfelt for being my reader and offering constructive feedback. I am also thankful to Dr. David Richardson for providing motivating study and the related data sets. Without their help and resources, I would not have been able to finish my research.

I would also like to extend my heartfelt thanks to my family and friends for their constant love and support. Their understanding and encouragement have been a constant source of motivation for me.

Lastly, I would like to express my appreciation to all the resources I have used in completing this thesis, including books, articles, databases, and online platforms. Their availability has allowed me to broaden my knowledge and understanding of the subject matter.

Table of Contents

Introduction

When the response variable is binary and one or more explanatory variables are continuous, use of the logistic regression model is the most common and traditional way to estimate the adjusted odds ratios associated with the binary outcome. For matched case-control studies, conditional logistic regression is the typical approach to be applied. However, a fresh look at a discriminant function framework for estimating crude or adjusted odds ratios has been suggested that offers potential benefits. This includes the availability of a uniformly minimum variance unbiased (UMVU) estimator for the adjusted log odds ratio in multivariable analysis involving a continuous exposure of primary interest (Lyles et al., 2009). The approach requires the assumption of normally distributed errors in a multiple linear regression model for the exposure, but these are less stringent assumptions than those of multivariate normality that caused interest to wane in the discriminant function approach as originally proposed decades ago (Halperin et al., 1971; Hosmer, Lemeshow and Sturdivant, 2003).

With multivariate normality unnecessary, it is possible to investigate the discriminant function approach as an alternative to conditional logistic regression in matched case-control studies. Particularly when few or small matched sets were involved, the approach was found to give a more precise and unbiased estimator of the log odds ratio associated with a continuous predictor of primary interest (Li 2020).

Non-detectable exposures (e.g., biomarker levels determined by laboratory assay) pose a very real challenge to many studies of environmental and infectious disease epidemiology. The most common method in environmental chemistry to deal with non-detects is simply substituting the detection limit or some fraction of it in place of the unknown exposure. However, this approach can produce inaccurate and irreproducible statistical summaries and inferences, resulting some estimates that are far from the true values and potentially obscuring patterns and trends in the data. Another common method is maximum likelihood estimation (MLE), which is a parametric, model-based method that can be used to estimate means and other

summary statistics with censored data (Lynn et al. 2001; Lyles et al., 2001). When data sets are small, for example, fewer than 30–50 detected values, in which one or two outliers throw off the estimation, or where there is insufficient evidence to know whether the assumed distribution fits the data well, maximum likelihood methods generally do not work well. For these cases, nonparametric methods that do not assume a specific distribution and shape of data might be preferred (Helsel, 2006).

In this thesis, we consider using maximum likelihood estimation to account for non-detectable exposures in a case-control study where matching is performed. First, we use simulations to compare the performance of the MLE in a discriminant function framework with that based on standard logistic regression for complete datasets. For situations when matching is involved in the case-control study, we perform simulations and examine a real data example to compare the performance of the MLE in discriminant framework with that of conditional logistic regression and that of using the detection limit to replace exposure values under the limit. We use a previously proposed multivariable discriminant function approach to estimate the adjusted log odds ratio (Lyles et al., 2009), based on a likelihood designed to account for non-detects and introducing adjustments to reduce bias in the resulting estimates of residual variance.

Motivating study

To better illustrate the discriminant function framework when accounting for non-detectable exposure measurements in matched case-control studies, we now introduce a real-world data example. The Colorado Plateau uranium miners' study conducted by Langholz, Thomas, Xiang, and Stram (1999) is a large occupational study of underground miners in which the exposure of primary interest, radon progeny, has been quantified for the individual miners. Other covariates include age, race, and smoking history. An incidence density matched case-control study was nested within the occupational cohort and the data set has already been prepared and shared by Dr. David Richardson. The radon exposure-lung cancer association is substantial in magnitude.

According to this previously published research based on this study, the relative risk of lung cancer associated with exposure to radon increases for around 8.5 years of exposure before decreasing and returning to background levels after roughly 34 years. The researchers strongly rejected the idea that the risk remains at its highest level with (p-value<0.001). Then they looked at how the effects varied across different subsets of the cohort based on factors like age, exposure level, exposure rate, and smoking. They found that only age had a significant impact on the results, with the decline in risk being much steeper among individuals over the age of 60 compared to younger individuals (Langholz et al., 1999).

The authors highlight the importance of accounting for the latency period between exposure and disease onset and describe the statistical methods used to analyze the data. We wish to use this real matched case-control dataset to illustrate the discriminant function approach. Specifically, we will use a randomly selected subset of the data, artificially produce some non-detectable exposures, and account for the non-detects by means of maximum likelihood.

Methods

When covariates are involved, a basic formula for estimating the odds ratio (OR) would be

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

where Y represents a binary outcome, X the exposure of interest, and C a vector of covariates. When thinking in terms of a case-control or a cross-sectional study, the OR can also be written as

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

where the specific form for the conditional densities is typically unknown but might be assumed for modeling purposes. Equation (2) in the case of a continuous exposure (X) forms the basis of the multivariable discriminant function approach utilized here (Lyles et al., 2009).

Standard Logistic Regression

The standard logistic model for the multivariable (covariate-adjusted) case is generally presented as

$$logit\{Pr(Y = 1|X = x, \boldsymbol{C} = \boldsymbol{c})\} = \beta_0 + \beta_1 X + \boldsymbol{\gamma}'\boldsymbol{c} \quad (3)$$

or

$$Pr(Y = 1|X, \boldsymbol{C}) = \frac{exp(\beta_0 + \beta_1 X + \boldsymbol{\gamma}'\boldsymbol{c})}{1 + exp(\beta_0 + \beta_1 X + \boldsymbol{\gamma}'\boldsymbol{c})} \quad (4)$$

where Y is the binary outcome with values either 0 or 1, $\boldsymbol{C}$ is a set of covariates and *X* is the continuous predictor of interest. The OR corresponding to a unit increase is e^{β_1}, which could be calculated based on the MLE of β (Hosmer, Lemeshow and Sturdivant, 2013).

Conditional Logistic Regression

Conditional logistic regression is a specialized type of logistic regression usually employed when case subjects with a particular condition or attribute are each matched with n control subjects without the condition. The form for the regression model is

$$logit\{Pr_k\} = \beta_{0k} + \beta_1 X + \boldsymbol{\gamma}'\boldsymbol{c} \quad (5)$$

where k represents strata (matched sets), and Pr_k is the probability that Y=1 in stratum *k* (Hosmer, Lemeshow and Sturdivant, 2013). To establish notation for the conditional likelihood for a 1-M matched case-control study, Hosmer, Lemeshow and Sturdivant (2013) write the conditional likelihood for the *kth* stratum as

$$l_k(\boldsymbol{\beta}) = \frac{\prod_{i=1}^{n_{1k}} e^{\beta_1 x_i + \boldsymbol{\gamma}'\boldsymbol{c_i}}}{\sum_{j=1}^{c_k} \prod_{i_j=1}^{n_{1k}} e^{\beta_1 x_{ji_j} + \boldsymbol{\gamma}'\boldsymbol{c}_{ji_j}}} \quad (6)$$

In equation (6), we assume the kth stratum contains n_{1k} cases and n_{0k} controls, c_k is the number of possible assignments of cases and controls to the total number of subjects, and j denotes any one of these

c_k assignments. Estimators of the vector $\boldsymbol{\beta}$ can be obtained by maximizing equation (6) with respect to those parameters.

Discriminant Function Approach

As noted previously, the discriminant function approach as studied by Lyles et al. (2009) was presented as an alternative to the standard logistic regression method for odds ratio estimation, which does not require multivariate normality. It leads us to a uniformly minimum variance unbiased (UMVU) estimator for a crude or adjusted log odds ratio. In the context of a matched case-control study, Li (2020) described that such a discriminant function approach can be based on the following multiple linear regression model:

$$\mu_{ij} = E\left(X_{ij} \middle| Y = y, \boldsymbol{C}\right) = \alpha^* + a_i + \beta_1^* y_{ij} + \gamma^{*'} \boldsymbol{C}_{ij} \tag{7}$$

where the a_is are fixed effects to index matched sets, and $j = 1, \ldots, M_i$ indexes individuals within the ith matched set. For the purposes of estimation, we make the further assumption of Gaussian errors in the linear model; that is, we assume that $X_{ij} = \mu_{ij} + \epsilon_{ij}$, where $\epsilon_{ij} \overset{(iid)}{\sim} \mathrm{N}(0, \sigma^2)$. The basic discriminant function-based estimator of the odds ratio associated with a unit increase in X_{ij} is:

$$\widehat{OR} = e^{\widehat{\beta_1^*}/MSE} \tag{8}$$

where the MSE is the standard mean squared error which estimates the residual variance in the multiple linear regression model in equation (7). The UMVU estimator of the log odds ratio under this assumed model can be written as follows (Lyles et al., 2009):

$$ln(\widehat{OR})_{umvu} = \left(\frac{n - T - 4}{n - T - 2}\right) \widehat{\beta_1^*}/MSE \tag{9}$$

where *T* represents the dimension of the covariate vector $\boldsymbol{C}$.

Handling Non-detectables

For a detectable X_{ij}, the basic form of the likelihood contribution is

$$f(x_{ij}) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2\sigma^2}(x_{ij} - \mu_{ij})^2} \tag{10}$$

For nondetectable exposures, the likelihood contribution is instead written as follows:

$$Pr(x_{ij} < LOD_{x_{ij}}) = Pr(\frac{x_{ij} - \mu_{ij}}{\sigma} < \frac{LOD_{x_{ij}} - \mu_{ij}}{\sigma}) = \mathbf{\Phi}(\frac{LOD_{x_{ij}} - \mu_{ij}}{\sigma}) \quad (11)$$

Sorting so that the first m of the n observations is detected, the likelihood function would be proportional to

$$L = \prod_{i=1}^{m} f(x_{ij}) \times \prod_{i=m+1}^{n} \mathbf{\Phi}(\frac{LOD_{x_{ij}} - \mu_{ij}}{\sigma}) \quad (12)\ ,$$

which we find to be conveniently maximized by using a general likelihood specification available in the NLMIXED procedure in SAS (SAS, n.d.).

Adjustment of the MLE of $\boldsymbol{\sigma^2}$

When maximum likelihood is applied to model (7), the issue of correcting the MLE for the residual variance (σ^2) should be considered. Specifically, for an MLR model with p predictors, the usual unbiased estimator of the residual variance is $\hat{\sigma}^2 = MSE = \frac{SSE}{n-p-1}$, where n is the sample size, MSE stands for mean squared error and SSE is sum of squares error. In contrast, the MLE under the typical "HEIL GAUSS" assumptions (Kutner et al.,2004) is $\hat{\sigma}^2{}_{MLE} = \frac{SSE}{n} = \frac{n-p-1}{n} MSE$. In most common MLE settings, n is large relative to p so the MLE is not biased much. However, in our matched sets scenario, p can potentially be very large so that the MLE for σ^2 is biased downward. Therefore, an adjustment should be performed for this bias after we obtain the MLEs numerically from the SAS NLMIXED procedure. Originally, the method $\hat{\sigma}^2{}_{adj} = \frac{N}{N-p-1} * \hat{\sigma}^2{}_{MLE}$ makes the obvious solution for datasets without non-detects. We also apply this same simple adjustment when accounting for non-detects using the discriminant function approach as an alternative to regular logistic regression (i.e., without matching).

When using the discriminant function approach as an alternative to conditional logistic regression with matching and in the presence of non-detectable exposures, we find that the standard adjustment is lacking as a bias correction to the MLE for σ^2, at least whenever there are any strata in which there are no detectable X values. After preliminary empirical investigations, we currently recommend the following adjusted estimator in the setting of k-to-1 matching:

$$\hat{\sigma}^2{}_{adj} = \frac{N-a}{(N-a)-p-1} * \hat{\sigma}^2{}_{MLE} \quad (13)$$

where $a = k * m$ and m is the number of strata containing all non-detects. While further work is needed to fully vet this adjustment and to explore the case of varying stratum-specific sample sizes, we note that it is equivalent to treating any stratum with all non-detects as if it provides only one effective observation. As such, we use that approach in our analysis of the motivating data (see Real Data Example).

Delta method

From the discriminant function approach, we use the estimator $ln(\widehat{OR}) = \frac{\widehat{\beta_1}}{\hat{\sigma}^2{}_{adj}}$, which is a nonlinear function of $\widehat{\beta_1}$ and $\hat{\sigma}^2{}_{adj}$:

$$ln(\widehat{OR}) = g\,(\widehat{\beta_1}, \hat{\sigma}^2{}_{adj}) = \widehat{\beta_1} * (\hat{\sigma}^2{}_{adj})^{-1} \quad (14)$$

The estimated derivatives of g with respect to β_1 and σ^2_{adj} are $\frac{\partial g}{\partial \beta_1} = (\hat{\sigma}^2{}_{adj})^{-1}$ and $\frac{\partial g}{\partial \hat{\sigma}^2{}_{adj}} = -\beta_1 * (\sigma^2_{adj})^{-2}$ respectively.

If the function of the random variables is differentiable with respect to each of its arguments and the random variables are approximately multivariate normal, the multivariate delta method allows us to estimate the distribution of the function of interest by taking the derivative of the function with respect to each of the random variables and using the covariance matrix of the random variables to calculate the variance of the function. Therefore, the variance of the function g is $var[g\,(\widehat{\beta_1}, \hat{\sigma}^2{}_{adj})] \doteq \widehat{D}\hat{\Sigma}\widehat{D}'$, where

$$\widehat{D} = [(\hat{\sigma}^2{}_{adj})^{-1}, -\widehat{\beta_1} * (\hat{\sigma}^2{}_{adj})^{-2}] \quad \text{and} \quad \hat{\Sigma} = \widehat{var}\begin{bmatrix}\widehat{\beta_1} \\ \hat{\sigma}^2{}_{adj}\end{bmatrix},$$

where we can obtain the latter matrix from SAS using the NLMIXED procedure. Hence,

$$var[g\,(\widehat{\beta_1}, \hat{\sigma}^2{}_{adj})] \doteq \widehat{d_1}^2 * \widehat{\sigma_1}^2 + \widehat{d_2}^2 * \widehat{\sigma_2}^2,$$

where $\widehat{d_1} = (\hat{\sigma}^2{}_{adj})^{-1}$, $\widehat{d_2} = -\widehat{\beta_1} * (\hat{\sigma}^2{}_{adj})^{-2}$, $\widehat{\sigma_1}^2 = \widehat{var}(\widehat{\beta_1})$, $\widehat{\sigma_2}^2 = \widehat{var}(\hat{\sigma}^2{}_{adj})$.

Simulation Studies and Results

We start with the setting where the typical approach would be standard logistic regression for complete data. The simulation studies were performed under the conditions of the one covariate case. Specifically, the binary outcome Y (1 if having the disease, 0 not having the disease) was generated randomly using binomial distribution with 0.2 prevalence. Age is the only covariate that was generated as normal with mean and variance equal 60 and 25 respectively, and the predictor of interest (X) is generated by discriminant function approach with true $\sigma^2 = 1$. This process was repeated for 2000 independent simulated datasets using different sample sizes (n=50, 100, 500 respectively) and different log odds ratios (true ln(OR)=0, 1, 2 respectively) to see if the log odds ratio estimates from the discriminant function approach indeed reflect less bias than those from standard logistic regression.

Table 1 summarizes 2000 replications under the following conditions: n = 50/100/500, $\sigma^2 = 1$ (true OR =1, true ln(OR)=0). Although the average estimations for log odds ratios are all close to 0 if rounded up to two-decimal places, the empirical SDs and the mean estimated standard errors for the discriminant function-based ln(OR) estimators are smaller than logistic regression-based ln(OR) and the reduction in variance leads to narrower CIs via the discriminant function approach. Similar steps were used when increasing the value of true OR to 2.718, yields a lager true ln(OR) of 1. Results can be seen in Table 2, note the roughly 20% up-ward bias evident in $ln(\widehat{OR})_{log}$ when sample size n is 50, which is the estimate of the ln(OR) [β in formula (3)] based on regular logistic regression with covariates. This bias is reduced when sample size became large, for example, the bottom section for Table 2 suggests a less than 1% bias in log odds ratio estimations. Table 3 considers cases of larger true OR 7.389, and true ln(OR) is 2. In this scenario, the advantages over traditional logistic regression really begin to stand out.

Tables 1-3 agree with results from Lyles et al. (2009), confirming that the UMVU estimator from discriminant function approach yields average values closer to the true log odds ratio as well as better precision when compared with the MLE of the log OR from regular logistic regression.

Table1. Simulation results assessing alternative crude OR estimators based on 2000 replications with sample size n = 50/100/500 in each case; True OR=1 and True ln(OR)=0.

	$\widehat{ln(OR)}$			
	Mean (SD)	**Mean Estimated SE**	**95% CI Coverage %**	**Mean CI Width (Median)**
n=50, True ln(OR)=0				
Regular logistic	0.00 (0.44)	0.40	95.8%	1.98 (1.60)
Discriminant-UMVU	0.00 (0.39)	0.37	96.6%	1.76 (1.50)
n=100, True ln(OR)=0				
Regular logistic	0.00 (0.27)	0.26	96.2%	1.12 (1.04)
Discriminant-UMVU	0.00 (0.25)	0.26	96.5%	1.09 (1.01)
n=500, True ln(OR)=0				
Regular logistic	0.00 (0.12)	0.11	94.4%	0.45 (0.44)
Discriminant-UMVU	0.00 (0.12)	0.11	94.5%	0.45 (0.44)

Table2. Simulation results assessing alternative crude OR estimators based on 2000 replications with sample size n = 50/100/500 in each case; True OR=2.718 and True ln(OR)=1.

	$\widehat{ln(OR)}$			
	Mean (SD)	**Mean Estimated SE**	**95% CI Coverage %**	**Mean CI Width (Median)**
n=50, True ln(OR)=1				
Regular logistic	1.20 (1.00)	0.54	97.3%	1.2E29 (5.96)
Discriminant-UMVU	1.00 (0.44)	0.43	95.8%	6.57 (4.68)
n=100, True ln(OR)=1				
Regular logistic	1.07 (0.32)	0.32	96.3%	4.37 (3.62)

Discriminant-UMVU	1.00 (0.29)	0.29	95.6%	3.59 (3.21)
n=500, True ln(OR)=1				
Regular logistic	1.02 (0.14)	0.13	94.2%	1.50 (1.46)
Discriminant-UMVU	1.01 (0.13)	0.13	94.5%	1.42 (1.39)

Table3. Simulation results assessing alternative crude OR estimators based on 2000 replications with sample size n = 50/100/500 in each case; True OR=7.389 and True ln(OR)=2.

	$ln(\widehat{OR})$			
	Mean (SD)	**Mean Estimated SE**	**95% CI Coverage %**	**Mean CI Width (Median)**
n=50, True ln(OR)=2				
Regular logistic	3.26 (4.58)	1.80	98.2%	3.5E150 (41.20)
Discriminant-UMVU	1.99 (0.58)	0.56	94.5%	30.30 (17.37)
n=100, True ln(OR)=2				
Regular logistic	2.27 (0.90)	0.57	97.1%	4.3E18 (19.04)
Discriminant-UMVU	2.01 (0.38)	0.39	95.8%	14.41 (11.7)
n=500, True ln(OR)=2				
Regular logistic	2.04 (0.22)	0.21	95.0%	6.71 (6.17)
Discriminant-UMVU	2.00 (0.18)	0.17	94.6%	5.14 (4.94)

Next, we consider accounting for non-detects in standard logistic regression. To conduct the simulation, we empirically calculated LODs to yield 10%, 25% and 50% non-detectable exposures based on the assumed simulation conditions. We used these different LODs to study variations in estimation performance as the percentage of non-detects increases.

Simulation results from Tables 4-6 suggest that the performance of our adjusted maximum likelihood estimation (MLE_adj) in the discriminant function framework is more stable and yields average estimates

much closer to the true value than estimators from regular logistic regression, as well as those using the standard "plug in detection limit" (LOD) method.

Table4. Results of simulation to assess performance of the maximum likelihood estimation in a discriminant function framework when accounting for non-detects. This process was repeated for 2000 independent simulated datasets with sample size n = 50/100/500 in each case; True OR=1 and True ln(OR)=0; lod=6.99, 6.30, 5.69 in each case.

	$ln(\widehat{OR})$			
	Mean (SD)	**Mean Estimated SE**	**95% CI Coverage %**	**Mean CI Width (Median)**
n=50, True ln(OR)=0, lod=6.99				
Regular_plug in LOD	-0.51 (4.47)	1.76	97.4%	2.17E174 (3.28)
MLE_adj	-0.10 (0.90)	1.78	98.8%	7.18 (1.75)
n=50, True ln(OR)=0, lod=6.30				
Regular_plug in LOD	-0.01 (0.44)	0.40	96.0%	2.01 (1.59)
MLE_adj	0.00 (0.40)	0.38	96.7%	1.54 (1.49)
n=50, True ln(OR)=0, lod=5.69				
Regular_plug in LOD	-0.04 (0.48)	0.44	96.2%	2.17 (1.76)
MLE_adj	-0.02 (0.40)	0.39	97.3%	1.57 (0.28)
n=100, True ln(OR)=0, lod=6.99				
Regular_plug in LOD	-0.04 (0.51)	0.47	96.4%	2.15 (1.92)
MLE_adj	-0.02 (0.30)	0.30	97.0%	1.18 (1.16)
n=100, True ln(OR)=0, lod=6.30				
Regular_plug in LOD	0.00 (0.28)	0.26	94.5%	1.13 (1.04)
MLE_adj	0.00 (0.27)	0.26	95.1%	1.03 (1.02)
n=100, True ln(OR)=0, lod=5.69				
Regular_plug in LOD	-0.02 (0.30)	0.29	95.5%	1.21 (1.13)
MLE_adj	-0.01 (0.26)	0.26	96.4%	1.04 (1.02)

n=500, True ln(OR)=0, lod=6.99				
Regular_plug in LOD	-0.01 (0.20)	0.19	94.5%	0.77 (0.76)
MLE_adj	0.00 (0.13)	0.13	95.5%	0.49 (0.49)
n=500, True ln(OR)=0, lod=6.30				
Regular_plug in LOD	0.00 (0.11)	0.11	94.9%	0.45 (0.44)
MLE_adj	0.00 (0.11)	0.11	95.0%	0.44 (0.44)
n=500, True ln(OR)=0, lod=5.69				
Regular_plug in LOD	0.00 (0.13)	0.12	95.4%	0.49 (0.49)
MLE_adj	0.00 (0.12)	0.11	95.5%	0.45 (0.45)

Table5. Results of simulation to assess performance of the maximum likelihood estimation in a discriminant function framework when accounting for non-detects. This process was repeated for 2000 independent simulated datasets with sample size n = 50/100/500 in each case; True OR=2.718 and True ln(OR)=1; lod=7.19, 6.46, 5.79 in each case.

	$\widehat{ln(OR)}$			
	Mean (SD)	**Mean Estimated SE**	**95% CI Coverage %**	**Mean CI Width (Median)**
n=50, True ln(OR)=1, lod=7.19				
Regular_plug in LOD	1.46 (2.55)	0.95	94.3%	3.53E178 (12.29)
MLE_adj	1.06 (0.67)	0.55	96.0%	2.20 (2.07)
n=50, True ln(OR)=1, lod=6.46				
Regular_plug in LOD	1.29 (0.82)	0.56	97.4%	2.40E24 (7.43)
MLE_adj	1.07 (0.50)	0.47	96.7%	1.90 (1.83)
n=50, True ln(OR)=1, lod=5.79				
Regular_plug in LOD	1.20 (0.65)	0.52	96.7%	8.11E14 (6.24)
MLE_adj	1.05 (0.48)	0.45	95.5%	1.79 (1.74)
n=100, True ln(OR)=1, lod=7.19				
Regular_plug in LOD	1.39 (0.45)	0.42	87.7%	8.96 (6.86)
MLE_adj	1.04 (0.35)	0.36	96.7%	1.43 (1.39)

n=100, True ln(OR)=1, lod=6.46				
Regular_plug in LOD	1.16 (0.37)	0.34	94.4%	5.35 (4.29)
MLE_adj	1.01 (0.32)	0.32	95.9%	1.26 (1.23)
n=100, True ln(OR)=1, lod=5.79				
Regular_plug in LOD	1.09 (0.34)	0.32	95.8%	4.58 (3.73)
MLE_adj	1.02 (0.30)	0.30	94.9%	1.21 (1.19)
n=500, True ln(OR)=1, lod=7.19				
Regular_plug in LOD	1.31 (0.18)	0.18	58.5%	2.68 (2.59)
MLE_adj	1.01 (0.15)	0.15	96.4%	0.61 (0.61)
n=500, True ln(OR)=1, lod=6.46				
Regular_plug in LOD	1.11 (0.14)	0.14	89.6%	1.76 (1.70)
MLE_adj	1.00 (0.13)	0.14	96.2%	0.54 (0.54)
n=500, True ln(OR)=1, lod=5.79				
Regular_plug in LOD	1.04 (0.14)	0.14	94.5%	1.54 (1.49)
MLE_adj	1.00 (0.13)	0.13	95.6%	0.52 (0.52)

Table6. Results of simulation to assess performance of the maximum likelihood estimation in a discriminant function framework when accounting for non-detects. This process was repeated for 2000 independent simulated datasets with sample size n = 50/100/500 in each case; True OR=7.389 and True ln(OR)=2; lod=7.30, 6.50, 5.83 in each case.

	$\widehat{ln(OR)}$			
	Mean (SD)	**Mean Estimated SE**	**95% CI Coverage %**	**Mean CI Width (Median)**
n=50, True ln(OR)=2, lod=7.30				
Regular_plug in LOD	3.75 (5.61)	1.87	99.3%	8.11E112 (75.18)
MLE_adj	2.20 (0.78)	0.80	97.5%	3.21 (3.03)
n=50, True ln(OR)=2, lod=6.50				
Regular_plug in LOD	3.38 (4.50)	1.86	98.7%	1.02E108 (46.01)

MLE_adj	2.14 (0.67)	0.65	96.9%	2.63 (2.53)
n=50, True ln(OR)=2, lod=5.83				
Regular_plug in LOD	3.26 (4.48)	1.92	97.9%	1.88E165 (45.59)
MLE_adj	2.12 (0.63)	0.60	96.2%	2.41 (2.33)
n=100, True ln(OR)=2, lod=7.30				
Regular_plug in LOD	2.64 (0.75)	0.60	93.2%	4.13E14 (32.15)
MLE_adj	2.10 (0.50)	0.53	97.5%	2.12 (2.07)
n=100, True ln(OR)=2, lod=6.50				
Regular_plug in LOD	2.34 (0.80)	0.55	96.9%	6.58E21 (21.00)
MLE_adj	2.06 (0.43)	0.44	96.1%	1.75 (1.73)
n=100, True ln(OR)=2, lod=5.83				
Regular_plug in LOD	2.28 (0.82)	0.55	97.4%	7.11E18 (18.68)
MLE_adj	2.05 (0.41)	0.41	95.5%	1.62 (1.59)
n=500, True ln(OR)=2, lod=7.30				
Regular_plug in LOD	2.43 (0.24)	0.23	55.5%	11.34 (10.48)
MLE_adj	2.02 (0.21)	0.23	97.4%	0.90 (0.90)
n=500, True ln(OR)=2, lod=6.50				
Regular_plug in LOD	2.11 (0.21)	0.21	94.8%	7.16 (6.53)
MLE_adj	2.01 (0.19)	0.19	96.2%	0.75 (0.75)
n=500, True ln(OR)=2, lod=5.83				
Regular_plug in LOD	2.05 (0.22)	0.21	94.7%	6.83 (6.26)
MLE_adj	2.01 (0.18)	0.18	94.9%	0.70 (0.70)

Further simulations were also conducted to assess ln(OR) estimators under the matched case-control study setting. Specifically, we used 2 to 1 matching to illustrate the stable performance of the discriminant function method. We start with the complete data to perform similar simulations as the regular logistic

regression but including the stratum indicator a_i this time, which is generated randomly from $N(0,1)$ distribution. The binary outcome Y is generated randomly from binomial distribution but with prevalence related to a_i. The way that we set covariate Age and the predictor of interest (X) is as same as the first simulation study. Results were summarized under 2000 replications using different number of matched sets (k=25, 100 respectively) and different log odds ratio (true ln(OR)=0, 1, 2 respectively) to compare the log odds ratio estimators from both conditional logistic regression and discriminant function.

Tables 7-9 show that the UMVU estimator again provides estimates closer on average to the true log odds ratio compared with the estimator obtained by conditional logistic regression. The mean standard errors for ln(OR) are also noticeably reduced when we move from conditional logistic regression to the discriminant function analysis. Variance reduction leading to narrower CIs via discriminant function analysis can also be seen in Tables 7-9. These results agree in spirit with empirical studies presented by Li (2020).

Table7. Simulation results assessing alternative crude OR estimators under 2 to 1 matched study based on 2000 replications with number of matched sets k = 25/100 in each case; True OR=1 and True ln(OR)=0.

	$ln(\widehat{OR})$			
	Mean (SD)	**Mean Estimated SE**	**95% CI Coverage %**	**Mean CI Width (Median)**
k=25, True ln(OR)=0				
Conditional logistic	0.01 (0.27)	0.26	95.8%	1.13 (1.02)
Discriminant-UMVU	0.00 (0.25)	0.25	96.4%	1.07 (0.98)
k=100, True ln(OR)=0				
Conditional logistic	0.00 (0.12)	0.12	95.7%	0.50 (0.49)
Discriminant-UMVU	0.00 (0.12)	0.12	95.7%	0.49 (0.48)

Table8. Simulation results assessing alternative crude OR estimators under 2 to 1 matched study based on 2000 replications with number of matched sets k = 25/100 in each case; True OR=2.718 and True ln(OR)=1.

	$\widehat{ln(OR)}$			
	Mean (SD)	**Mean Estimated SE**	**95% CI Coverage %**	**Mean CI Width (Median)**
k=25, True ln(OR)=1				
Conditional logistic	1.14 (0.50)	0.39	97.2%	2517074.71 (4.22)
Discriminant-UMVU	1.00 (0.34)	0.33	94.5%	4.34 (3.43)
k=100, True ln(OR)=1				
Conditional logistic	1.03 (0.17)	0.17	95.4%	1.98 (1.84)
Discriminant-UMVU	1.00 (0.16)	0.16	95.0%	1.77 (1.70)

Table9. Simulation results assessing alternative crude OR estimators under 2 to 1 matched study based on 2000 replications with number of matched sets k = 25/100 in each case; True OR=7.389 and True ln(OR)=2.

	$\widehat{ln(OR)}$			
	Mean (SD)	**Mean Estimated SE**	**95% CI Coverage %**	**Mean CI Width (Median)**
k=25, True ln(OR)=2				
Conditional logistic	7.79 (36.29)	2177.98	90.4%	4.6E239 (40.16)
Discriminant-UMVU	2.01 (0.50)	0.49	94.0%	25.38 (14.97)
k=100, True ln(OR)=2				
Conditional logistic	2.04 (0.22)	0.21	95.0%	6.71 (6.17)
Discriminant-UMVU	2.00 (0.18)	0.17	94.6%	5.14 (4.94)

Last but not least, we conducted simulations accounting for non-detectables in X in the matched case-control setting. Simulations were conducted with 2000 replications for each case with 2 to 1 matching where the number of matched sets is set to 25, 100 and the true log odds ratio is set to 0, 1 and 2, respectively. In the same manner as when accounting for non-detects for standard logistic regression, we determined and used LODs consistent with 10%, 25% and 50% non-detectables.

Simulation results from Tables 10-12 suggest that the adjusted maximum likelihood estimation based on discriminant function analysis shows less bias compared to the expedient approach of plugging in the detection limit for non-detects and using conditional logistic regression. The mean standard errors for ln(OR) are also noticeably reduced when we move from the conditional logistic regression plug-in method to the discriminant function approach.

Table10. Results of simulation to assess performance of the maximum likelihood estimation in a discriminant function framework when accounting for non-detects in 2 to 1 matching. This process was repeated for 2000 independent simulated datasets with number of matched sets k = 25/100 in each case; True OR=0 and True ln(OR)=1; lod=7.14, 6.18, 5.34 in each case.

	$\widehat{ln(OR)}$		
	Mean (SD)	**Mean Estimated SE**	**95% CI Coverage %**
k=25, True ln(OR)=1, lod=7.14			
Regular_plug in LOD	1.48 (0.93)	0.57	97.6%
MLE_adj	1.17 (0.49)	0.42	95.5%
k=25, True ln(OR)=1, lod=6.18			
Regular_plug in LOD	1.24 (0.53)	0.43	97.2%
MLE_adj	1.08 (0.38)	0.18	94%
k=25, True ln(OR)=1, lod=5.34			
Regular_plug in LOD	1.18 (0.50)	0.40	97.5%
MLE_adj	1.06 (0.34)	0.17	93.3%
k=100, True ln(OR)=1, lod=7.14			
Regular_plug in LOD	1.30 (0.27)	0.24	82.5%
MLE_adj	1.09 (0.21)	0.20	94.2%
k=100, True ln(OR)=1, lod=6.18			
Regular_plug in LOD	1.12 (0.20)	0.19	93.2%
MLE_adj	1.04 (0.18)	0.16	93.2%
k=100, True ln(OR)=1, lod=5.34			
Regular_plug in LOD	1.06 (0.18)	0.18	95.9%
MLE_adj	1.02 (0.17)	0.15	93.4%

Table11. Results of simulation to assess performance of the maximum likelihood estimation in a discriminant function framework when accounting for non-detects in 2 to 1 matching. This process was repeated for 2000 independent simulated datasets with number of matched sets k = 25/100 in each case; True OR=1 and True ln(OR)=0; lod=7, 6.12, 5.32 in each case.

	$ln(\widehat{OR})$		
	Mean (SD)	**Mean Estimated SE**	**95% CI Coverage %**
k=25, True ln(OR)=0, lod=7			
Regular_plug in LOD	-0.01 (0.50)	0.45	96.4%
MLE_adj	0.01 (0.33)	0.25	98.1%
k=25, True ln(OR)=0, lod=6.12			
Regular_plug in LOD	0.00 (0.35)	0.34	95.8%
MLE_adj	0.00 (0.28)	0.20	96.2%
k=25, True ln(OR)=0, lod=5.32			
Regular_plug in LOD	0.01 (0.30)	0.29	95.7%
MLE_adj	0.01 (0.27)	0.17	94.8%
k=100, True ln(OR)=0, lod=7			
Regular_plug in LOD	0.00 (0.20)	0.20	95.5%
MLE_adj	0.00 (0.15)	0.15	97.0%
k=100, True ln(OR)=0, lod=6.12			
Regular_plug in LOD	0.00 (0.15)	0.15	95.6%
MLE_adj	0.00 (0.13)	0.13	95.9%
k=100, True ln(OR)=0, lod=5.32			
Regular_plug in LOD	0.01 (0.14)	0.13	94%
MLE_adj	0.00 (0.13)	0.12	94%

Table12. Results of simulation to assess performance of the maximum likelihood estimation in a discriminant function framework when accounting for non-detects in 2 to 1 matching. This process was repeated for 2000 independent simulated datasets with number of matched sets k = 25/100 in each case; True OR=7.39 and True ln(OR)=2; lod=7.14, 6.17, 5.33 in each case.

	$ln(\widehat{OR})$		
	Mean (SD)	**Mean Estimated SE**	**95% CI Coverage %**
k=25, True ln(OR)=2, lod=7.14			
Regular_plug in LOD	4.95 (21.08)	553.62	94.9%
MLE_adj	2.47 (0.84)	0.19	91.3%
k=25, True ln(OR)=2, lod=6.17			

Regular_plug in LOD	6.80 (31.19)	1881.87	92.7%
MLE_adj	2.22 (0.60)	0.18	92.5%
k=25, True ln(OR)=2, lod=5.33			
Regular_plug in LOD	6.39 (30.82)	1598.23	92.3%
MLE_adj	2.10 (0.52)	0.44	91.3%
k=100, True ln(OR)=2, lod=7.14			
Regular_plug in LOD	2.66 (0.61)	0.47	83.0%
MLE_adj	2.33 (0.35)	0.29	82.0%
k=100, True ln(OR)=2, lod=6.17			
Regular_plug in LOD	2.27 (0.46)	0.39	96.3%
MLE_adj	2.10 (0.27)	0.23	91.3%
k=100, True ln(OR)=2, lod=5.33			
Regular_plug in LOD	2.16 (0.43)	0.38	97.3%
MLE_adj	2.04 (0.25)	0.22	91.2%

Real data example

For a real-life application of (9) and (12), we use data originally described in the Colorado Plateau uranium miners' study (Langholz et al., 1999), which had 263 matched sets with one case each and an average of 40 controls. Usually, the potential bias and precision gains of the discriminant approach are better highlighted when the study is not extremely large. At the same time, the sample residuals from the linear model look relatively bell shaped but are a bit left-skewed if we use the full data, but this problem is less pronounced based on the randomly subsetted data. Therefore, we randomly selected 100 of those sets, and within each one we randomly sampled controls with 1/8 selection probability, leading to varying stratum sizes but an average of close to 8 controls per matched set. For the exposure variable, we used the natural log of the total radon exposure, which is referred to as "totrdn". At the same time, we control for total cumulative smoking ("totsmk"), and for the number of years exposed ("numyrsexposed", which we calculated as the difference between two variables, "rendage" minus "rdnstar"). With the case-control indicator denoted as "ccind_k", the applicable conditional logistic regression model is presented as the following:

$$logit(ccind\_k) = \beta_{0k} + \beta_1 * log\_totrdn + \beta_3 * numyrsexposed + \beta_4 * totsmk$$

We logged the total radon exposure variable because the alternative approach we want to use assumes normality of the errors in a multiple linear regression model that flips the problem around, with the exposure as the outcome and the set indicators, case status, and the control variables as the predictors. The normality approximation when looking at the model residuals appears much better if we first log the exposure variable. The estimated coefficient for exposure in the standard conditional logistic regression model is 0.336, which represents the estimated ln(OR) for a one natural log increase in the total radon exposure as estimated using conditional logistic regression.

Based on the discriminant function approach applied to the complete data (100 matched sets), our continuous predictor log_totrdn would be estimated based on the following fitted model:

$$\hat{E}(log\_totrdn) = a_i + 0.00014 * totsmk + 0.18 * numyrsexposed + 0.311 * ccind + \epsilon$$

where a_i denotes the stratum indicators and i=1, 2, …, 99. From SAS output, the MSE for log_totrdn is 1.03, and the point estimate of the ln(OR) for a one log increase in total radon exposure based on the subsetted datasets is calculated as 0.311/1.03=0.302 based on (8). We note the similarity between this estimate and the estimate (0.336) that was obtained via conditional logistic regression.

Then LODs at 10%, 25% and 50% of the overall exposure distribution were set. These are 4.13, 5.26, and 6.11, respectively, allowing us to demonstrate results accounting for the non-detects by means of maximum likelihood in this example. There are 2 matched sets with all non-detects (yielding values of $\hat{\sigma}^2$ and adjusted $\hat{\sigma}^2$ equal to 0.52 and 0.64, respectively), when LOD is 6.11. There is 1 matched set with all non-detects, yielding $\hat{\sigma}^2$ = 0.57 and adjusted $\hat{\sigma}^2$ = 0.70 if the LOD is 5.26. With LOD=4.13, there were no matched sets with all non-detects, and estimaes of $\hat{\sigma}^2$ and adjusted $\hat{\sigma}^2$ were 0.69 and 0.85, respectively.

Table13 summarizes the results of evaluating adjusted OR estimators corresponding to the continuous predictor log_totrdn accounting for non-detects using the subsetted data from the Colorado Plateau

uranium miners' study, where the ln(OR) estimate for the complete data without non-detects is 0.302 under the discriminant function analysis (corresponding to an estimated OR of 1.35). Compared with the traditional plug-in detection limit method under conditional logistic regression, the adjusted MLE appears to reflect less bias with reduced associated standard errors. At the same time, the proposed method provides accompanies the point estimates with narrower 95% confidence intervals.

Table 13. Colorado Plateau uranium miners' study results assessing alternative crude OR estimators based on random subset data with detection limit lod=6.11 /5.26/4.13 in each case; subsetted data using discriminant function: OR=1.35 and ln(OR)=0.302 ; subsetted data using conditional logistic regression: OR=1.40 and ln(OR)=0.336

	$\widehat{ln(OR)}$		
	Mean	**Mean Estimated SE**	**95% CI**
Conditional Logistic_complete data	0.34	0.13	(0.09, 0.58)
Discriminant Fuction_complete data	0.30	0.12	(0.07, 0.53)
Lod=6.11			
Plug in LOD	0.59	0.20	(0.19, 0.98)
MLE_adj	0.48	0.17	(0.14, 0.81)
Lod=5.26			
Plug in LOD	0.48	0.16	(0.17,0.79)
MLE_adj	0.46	0.15	(0.17, 0.74)
Lod=4.13			
Plug in LOD	0.40	0.14	(0.13, 0.67)
MLE_adj	0.36	0.13	(0.10, 0.61)

Discussion

A fresh look at the discriminant function approach for multivariable analysis with fewer strict requirements was proposed by Lyles, Guo and Hill (2009) and the UMVU estimator they derived yields a more precise estimate of odds ratio for a continuous exposure of interest (relative to logistic regression) when the assumed model holds. Similarly, Li (2020) showed that the discriminant function approach performs better when estimating a covariate-adjusted odds ratio relating to a continuous predictor in matched case-control studies compared with conditional logistic regression, especially when logistic regression is unstable or fails due to separation problems. Based on these related prior works, this thesis specifically focuses on evaluating the performance of maximum likelihood estimation in a discriminant function framework to account for non-detectable exposure measurements in matched case-control studies.

We first used simulation studies to confirm that the UMVU estimator provides estimates closer on average to the true log odds ratio compared with the estimators obtained by either regular logistic regression in an unmatched study or by conditional logistic regression in a 2 to 1 matching case; the latter is consistent with the results presented by Li (2020).

To account for non-detectables in X, we outlined the necessary likelihood contributions and discussed the need for adjustments to the MLE for the residual variance that is utilized in the discriminant function approach. We conducted simulations for both regular logistic regression and the matched case-control setting (2 to 1 specifically). For each case, we used different sample sizes or different numbers of matched sets, different true log odds ratios and different LODs that were consistent with 10%, 25% and 50% non-detectables. Simulation results all suggest that compared with the expedient approach of plugging in the detection limit for non-detects and using regular or conditional logistic regression, the adjusted maximum likelihood estimation based on the discriminant function analysis shows less bias and the mean standard errors for ln(OR) are also noticeably reduced. We conducted an analysis of a real data

example involving matched sets, and the discriminant function approach continued to show potential performance benefits compared to the expedient approach when accounting for non-detects.

For future work, potential improvements could be sought to better adjust the MLE of the residual variance when maximum likelihood is applied to the discriminant function model in matched studies involving nondetectable exposures. The original obvious adjustment applies naturally for datasets without non-detects or when accounting for non-detects using the discriminant function approach without matching involved. However, the MLE for σ^2 cannot be corrected for bias through this simple adjustment, particularly in cases where there are strata with all non-detectable X values. Although we currently propose a new adjustment (13) in the setting of k-to-1 matching, we believe this method can still be improved and we also would like to further explore the case of varying stratum-specific sample sizes.

References

[1] Lyles, R. H., Guo, Y., & Hill, A. F. (2009). A Fresh Look at the Discriminant Function Approach for Estimating Crude or Adjusted Odds Ratios. *The American Statistician*, *63*(4), 320–327. https://doi.org/10.1198/tast.2009.08246

[2] Li, R. (2020) *Comparisons of conditional logistic regression vs. a discriminant function approach in a case-control study where matching is performed | ID: 6h440t63f | Tesis y Disertaciones Electrónicas de Emory*. (n.d.). https://etd.library.emory.edu/concern/etds/6h440t63f?locale=es

[3] Helsel, D. R. (2006). Fabricating data: How substituting values for nondetects can ruin results, and what can be done about it. *Chemosphere*, *65*(11), 2434–2439. https://doi.org/10.1016/j.chemosphere.2006.04.051

[4] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2005). Applied Logistic Regression. *Wiley Series in Probability and Statistics*. https://doi.org/10.1002/9781118548387

[5] *Kutner, et al. (2004) Applied Linear Statistical Models. 4th Edition, McGraw-Hill, New York. - References - Scientific Research Publishing*. (n.d.). https://www.scirp.org/(S(351jmbntvnsjt1aadkozje))/reference/referencespapers.aspx?referenceid=2657083

[6] Langholz, B., Thomas, D.C., Xiang, A., & Stram, D.O. (1999). Latency analysis in epidemiologic studies of occupational exposures: application to the Colorado Plateau uranium miners cohort. *American journal of industrial medicine*, *35 3*, 246-56 .

[7] Halperin, M., Blackwelder, W. C., & Verter, J. (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. https://doi.org/10.1016/0021-9681(71)90106-8

[8] *SAS/STAT 14.1 User's Guide The NLMIXED Procedure*. (n.d.). Retrieved March 21, 2023, from https://support.sas.com/documentation/onlinedoc/stat/141/nlmixed.pdf