SECONDARY ANALYSIS OF CASE-CONTROL STUDIES IN GENOMIC CONTEXTS

A Dissertation

by

JIAWEI WEI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2010

Major Subject: Statistics

SECONDARY ANALYSIS OF CASE-CONTROL STUDIES IN GENOMIC CONTEXTS

A Dissertation

by

JIAWEI WEI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Raymond J. Carroll
Committee Members,	Mohsen Pourahmadi
	Faming Liang
	Lan Zhou
	Nancy Turner
Head of Department,	Simon J. Sheather

August 2010

Major Subject: Statistics

ABSTRACT

Secondary Analysis of Case-Control Studies
in Genomic Contexts. (August 2010)
Jiawei Wei, B.S., Zhejiang University;
M.S., Texas A&M University
Chair of Advisory Committee: Dr. Raymond J. Carroll

This dissertation consists of five independent projects. In each project, a novel statistical method was developed to address a practical problem encountered in genomic contexts. For example, we considered testing for constant nonparametric effects in a general semiparametric regression model in genetic epidemiology; analyzed the relationship between covariates in the secondary analysis of case-control data; performed model selection in joint modeling of paired functional data; and assessed the prediction ability of genes in gene expression data generated by the CodeLink System from GE.

In the first project in Chapter II we considered the problem of testing for constant nonparametric effects in a general semiparametric regression model when there is the potential for interaction between the parametrically and nonparametrically modeled variables. We derived a generalized likelihood ratio test for this hypothesis, showed how to implement it, and gave evidence that it can improve statistical power when compared to standard partially linear models.

The second project in Chapter III addressed the issue of score testing for the independence of X and Y in the second analysis of case-control data. The semiparametric efficient approaches can be used to construct semiparametric score tests, but they suffer from a lack of robustness to the assumed model for Y given X. We showed how to adjust the semiparametric score test to make its level/Type I error correct

even if the assumed model for Y given X is incorrect, and thus the test is robust.

The third project in Chapter IV took up the issue of estimation of a regression function when Y given X follows a homoscedastic regression model. We showed how to estimate the regression parameters in a rare disease case even if the assumed model for Y given X is incorrect, and thus the estimates are model-robust.

In the fourth project in Chapter V we developed novel AIC and BIC-type methods for estimating the smoothing parameters in a joint model of paired, hierarchical sparse functional data, and showed in our numerical work that they are many times faster than 10-fold crossvalidation while at the same time giving results that are remarkably close to the crossvalidated estimates.

In the fifth project in Chapter VI we introduced a practical permutation test that uses cross-validated genetic predictors to determine if the list of genes in question has "good" prediction ability. It avoids overfitting by using cross-validation to derive the genetic predictor and determines if the count of genes that give "good" prediction could have been obtained by chance. This test was then used to explore gene expression of colonic tissue and exfoliated colonocytes in the fecal stream to discover similarities between the two.

ACKNOWLEDGMENTS

I would like to thank my committee chair, Dr. Carroll, and my committee members, Dr. Pourahmadi, Dr. Liang, Dr. Zhou and Dr. Turner, for their guidance and support throughout the course of this research.

This work would not be possible without the extensive help from my co-workers, Arnab Maity, Nilanjan Chatterjee, Ursula U. Müller, Ingrid Van Keilegom and Josue G. Martinez.

Thanks also go to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience.

Finally, thanks to my mother and father for their encouragement and to my husband for his patience and love.

TABLE OF CONTENTS

CHAPTER Page

I INTRODUCTION 1

II TESTING FOR CONSTANT NONPARAMETRIC EFFECTS IN GENERAL SEMIPARAMETRIC REGRESSION MODELS WITH INTERACTIONS 4

A. Introduction 4
B. Methodology 7
1. Basic Framework 7
2. Estimation of Model Components 7
3. Properties of Profile Estimates of Parameters and Functions 9
4. Generalized Likelihood Ratio Test 11
5. Test Statistic and Implementation 12
C. Simulation Study 13
D. Data Example 15
E. Discussion 16

III LOCALLY EFFICIENT SCORE TESTS FOR INDEPENDENCE IN THE SECONDARY ANALYSIS OF CASE-CONTROL DATA 17

A. Introduction 17
B. Efficient Parametric Methods and Robustness 19
1. Framework 19
2. Prior Results and Robustness 20
C. A Locally Efficient Robust Score Test 21
1. Preliminaries 21
2. The Theoretical Score Under the Null Hypothesis . . . 22
3. Practical Implementation and Asymptotic Theory . . 23
D. The Regression Case 25
E. Simulation Studies 25
1. When X is Binary 25
2. When X is Discrete 28
3. When X is Continuous and Scalar 29

CHAPTER Page

4. Model with Covariates 30
F. Applications 30
1. Prostate Cancer Example 31
2. Colorectal Adenoma 32
G. Discussion 33

IV LOCALLY EFFICIENT ESTIMATION FOR HOMOSCEDASTIC REGRESSION IN THE SECONDARY ANALYSIS OF CASE-CONTROL DATA 34

A. Introduction 34
B. Efficient Parametric Methods and Robustness 36
1. Framework 36
2. Prior Results and Robustness 36
C. Model-Robust Estimation 37
1. Preliminaries 37
2. The Methodology 38
3. Development of the Score when $f_X(\cdot)$ and α_{true} are Known 39
4. Implementation when $f_X(\cdot)$ is Unknown but α_{true} is Known 41
5. When the Intercept α_{true} is Unknown 42
6. Distribution Theory 43
D. Simulation 45
E. An Empirical Example 46
1. Prostate Cancer 46
2. Colorectal Adenoma 47
F. Discussion 48

V MODEL SELECTION IN JOINT MODELLING OF PAIRED FUNCTIONAL DATA 50

A. Introduction 50
B. Methodology 52
1. A Reduced Rank Model for Sparsely Observed Paired Curves 52
2. Selection of the Penalty Parameters 54
3. Selection of the Number of Important Principal Components 56
C. Application 56

CHAPTER Page

1. Selection of the Smoothing Parameters in Simulation Studies . . . 56
2. Selection of the Numbers of the Principle Components in Simulation Studies . . . 60
3. Model Selection in the Rreal Data Example . . . 60
D. Discussion . . . 62

VI PERMUTATION TEST FOR MICROARRAYS: COLONIC TUMORIGENESIS PREDICTION IN THE EARLY STAGES OF DEVELOPMENT . . . 64

A. Introduction . . . 64
1. Motivation for Genetic Influence on Colon Cancer . . . 64
B. Data . . . 66
C. Analysis . . . 69
1. Cross-validated Permutation Test (CPT) . . . 69
2. Cross-validated within Treatment Permutation Test (CWPT) . . . 72
3. Diagonal Linear Discriminant Analysis (DLDA) . . . 75
4. CWPT and DLDA on Breast Cancer Data . . . 77
D. Conclusion . . . 79

VII CONCLUSION . . . 80

REFERENCES . . . 83

APPENDIX A . . . 88

APPENDIX B . . . 97

APPENDIX C . . . 105

VITA . . . 113

LIST OF TABLES

TABLE		Page
1	Significance levels in the NAT2 example	16
2	Test levels of normal, chisquared and gamma distributions for three methods when X is binary	27
3	Powers of normal and gamma distributions when X is binary	28
4	Powers of normal and gamma distributions when X is discrete	28
5	Test levels of normal, chisquared and gamma distributions for three methods when X is discrete	29
6	Powers of normal and gamma distributions when X is continuous	29
7	Test levels of normal, chisquared and gamma distributions for three methods when X is continuous	30
8	P-values in the NAT2 example	33
9	Simulation study for the estimation of β_1	46
10	Results of the VDR data example	47
11	Results of the NAT2 data example	48
12	Average ratios of the computing time using AIC, BIC_s and BIC_o to that using 10-fold CV	57
13	Average smoothing parameters and corresponding degrees of freedom selected using 10-fold crossvalidation, AIC, BIC_s and BIC_o	58
14	Integrated mean squared errors	59
15	AIC, BIC_s and BIC_o scores on AIDS data	62

TABLE Page

16 Permutation test within treatment for each gene in the entire gene list of ACF and tumor stages. 74

17 DLDA in the entire gene list of ACF and tumor stages. 78

LIST OF FIGURES

FIGURE Page

1 Results on power and level in the simulation for testing whether $\theta(\cdot)$ is constant. 14

2 Fitted mean curves and principal component curves using 10-fold crossvalidation, AIC, BIC_s and BIC_o for AIDS data. 63

CHAPTER I

INTRODUCTION

This dissertation consists of five independent projects. In each project, a novel statistical method was developed to address a practical problem encountered in genomic contexts. In the first project, we considered the problem of testing for constant nonparametric effect in a general semiparametric regression model when there is the potential for interaction between the parametrically and nonparametrically modeled variables. The work was originally motivated by a unique testing problem in genetic epidemiology (Chatterjee, et al., 2006) that involved a typical generalized linear model but with an additional term reminiscent of the Tukey one-degree-of-freedom formulation. In this formulation, there are genetic variables, environmental variables, and demographic variables. The interest is in testing for main effects of the genetic variables, while gaining statistical power by allowing for a possible interaction between genes and the environment. Later work (Maity, et al., 2009) involved the possibility of modeling the environmental variable nonparametrically, but they focused on whether there was a parametric main effect for the genetic variables. In this study, we consider the complementary problem, where the interest is in testing for the main effect of the nonparametrically modeled environmental variable. We derive a generalized likelihood ratio test for this hypothesis, show how to implement it, and give evidence that it can improve statistical power when compared to standard partially linear models. An empirical example involving colorectal adenoma is used to illustrate the methodology.

The second project addressed the issure of score testing for independence in the

---

The journal model is *Journal of the American Statistical Association.*

secondary analysis of case-control data. Typical case-control studies focus on the relationship between disease D and covariates (Y, X). In the secondary analysis of case-control data, it is the relationship between Y and X that is of interest, but the analysis of this relationship is complicated by the case-control sampling framework, which is a type of biased sampling. Previous work has assumed a parametric distribution for Y given X and derived semiparametric efficient estimation and inference without any distributional assumptions about X: of course, the roles of X and Y can be interchanged. In this study, we take up the issue of score testing for the independence of X and Y. The semiparametric efficient approaches can be used to construct semiparametric score tests, but they suffer from a lack of robustness to the assumed model for Y given X. We take an entirely different and novel approach. We show how to adjust the semiparametric score test to make its level/Type I error asymptotically correct in the rare disease case even if the assumed model for Y given X is incorrect, and thus the test is model robust. Extensions to linear regression with additional covariates are discussed. Simulations and an empirical example are used to illustrate the approach.

The third project took up the issue of estimation of a regression function when Y given X follows a homoscedastic regression model in the secondary analysis of case-control data. The semiparametric efficient approaches can be used to construct semiparametric efficient estimates, but they suffer from a lack of robustness to the assumed model for Y given X. We take an entirely different and novel approach in the case that the disease is rare. We show how to estimate the regression parameters in the rare disease case even if the assumed model for Y given X is incorrect, and thus the estimates are model-robust. Simulations and an empirical example are used to illustrate the approach.

We developed novel AIC and BIC type methods for estimating smoothing param-

eters in a joint model of paired, sparse functional data in the fourth project. Utilizing penalized B-splines, a new approach proposed by Zhou, et al. (2008) jointly models a pair of sparsely observed functions through functional principal components. In their approach, a stepwise addition and deletion procedure is employed to decide upon the number of principal components (PCs) and crossvalidation is used to estimate penalty parameters. However the choice of the cutoff point in the stepwise addition and deletion procedure is subjective and the crossvalidation computation is very time consuming. In this project we propose to select the number of PCs and estimate the penalty parameters with a modified version of the Akaike information criterion (AIC) and two modified versions of the Bayesian information criterion (BIC). Our methods are computationally fast and straightforward to implement. We illustrate our methods with simulations and the empirical data example used by Zhou, et al. (2008).

In the fifth project, we introduced a practical permutation test that uses cross-validated genetic predictors to determine if the list of genes in question has "good" prediction ability. We call our the cross-validated permutation test. It avoids over-fitting by using cross-validation to derive the genetic predictor and determines if the count of genes that give "good" prediction could have been obtained by chance. This test is then used to explore gene expression of colonic tissue and exfoliated colonocytes in the fecal stream to discovery similarities between the two, done at each of the three stages of colonic tumorigenesis.

CHAPTER II

TESTING FOR CONSTANT NONPARAMETRIC EFFECTS IN GENERAL SEMIPARAMETRIC REGRESSION MODELS WITH INTERACTIONS

In this study, we consider the problem of testing for constant nonparametric effect in a general semiparametric regression model when there is the potential for interaction between the parametrically and nonparametrically modeled variables. The work was originally motivated by a unique testing problem in genetic epidemiology (Chatterjee, et al., 2006) that involved a typical generalized linear model but with an additional term reminiscent of the Tukey one-degree-of-freedom formulation. In this formulation, there are genetic variables, environmental variables, and demographic variables. The interest is in testing for main effects of the genetic variables, while gaining statistical power by allowing for a possible interaction between genes and the environment. Later work (Maity, et al., 2009) involved the possibility of modeling the environmental variable nonparametrically, but they focused on whether there was a parametric main effect for the genetic variables. In this study, we consider the complementary problem, where the interest is in testing for the main effect of the nonparametrically modeled environmental variable. We derive a generalized likelihood ratio test for this hypothesis, show how to implement it, and give evidence that it can improve statistical power when compared to standard partially linear models. An empirical example involving colorectal adenoma is used to illustrate the methodology.

A. Introduction

We consider the problem of testing for constant nonparametric effects in a general semiparametric regression model when there is the potential for interaction between the parametrically and nonparametrically modeled variables. The work was originally

motivated by a unique testing problem in genetic epidemiology. Chatterjee, et al. (2006) considered the following logistic regression type problem. Let Y be a binary response, X a set of covariates that might possibly interact with a scalar covariate Z, and let S be additional variables not thought to interact with Z. Let $H(\cdot)$ be the logistic distribution function. Then they propose the model

$$\text{pr}(Y = 1|X, S, Z) = H(\kappa_0 + X^{\text{T}}\beta_0 + S^{\text{T}}\eta_0 + Z\theta_0 + \gamma X^{\text{T}}\beta_0 Z\theta_0). \tag{2.1}$$

In their context, X represented a set of genetic variables such as single nucleotide polymorphisms (SNP), S were demographic variables and Z was an environmental effect. Their interest was in testing for a possible genetic main effect, $H_0 : \beta_0 = 0$ versus $H_A : \beta_0 \neq 0$. When $\gamma = 0$, this is nothing more than an ordinary logistic regression model, and thus the test is routine. However, Chatterjee, et al. argue that if there is a possible gene-environment interaction, then capturing it via the Tukey-like 1-degree of freedom term $\gamma X^{\text{T}}\beta_0 Z\theta_0$ has the potential to increase statistical power greatly. They document this increase in power both in simulations and in empirical work.

It is important to see that γ in (2.1) is not identifiable, because under the null hypothesis, it disappears from the model. Hence, it is not a parameter to be estimated per se, but is rather a tuning constant. Chatterjee, et al. fix γ along a range of values $\text{L} \leq \gamma \leq \text{R}$, compute the score test $\mathcal{T}(\gamma)$ for each γ, and then take the maximum value as the final test statistic. They then develop a simulation-based procedure for computing an overall p-value.

Maity, et al. (2009) generalized models such as (2.1) to allow the effect of the environmental variable Z to enter nonparametrically. Thus, for an unknown function

$\theta_0(\cdot)$, their generalization of (2.1) becomes

$$\text{pr}(Y = 1|X, S, Z) = H\left\{X^{\mathrm{T}}\beta_0 + S^{\mathrm{T}}\eta_0 + \theta_0(Z) + \gamma X^{\mathrm{T}}\beta_0\theta_0(Z)\right\}. \tag{2.2}$$

They developed a testing procedure in model (2.2) for testing $H_0 : \beta_0 = 0$ versus $H_A : \beta_0 \neq 0$, and demonstrated increased power both via simulations and via empirical work. As in Chatterjee, et al., Maity, et al. fix γ along a range of values $\mathrm{L} \leq \gamma \leq \mathrm{R}$, compute the profile likelihood score test $\mathcal{T}(\gamma)$ for each γ, and then take the maximum value as the final test statistic. They also develop a simulation-based procedure for computing an overall p-value.

Both Chatterjee, et al. (2006) and Maity, et al. (2009) were focused on testing for the main effect of a gene. However, testing for a main effect of the environmental variable is also of great interest. In this study, we take up the question of testing whether Z has any effect in model (2.2), i.e., $H_0 : \theta(z) = \text{constant}$. Of course, when we set $\gamma = 0$, the result is a standard partially linear logistic model. We will demonstrate that our testing procedure based on model (2.2) has the potential for great gains in power, with little loss of power if $\gamma = 0$ actually obtains. Similar to these papers, we will vary γ along a fixed range, form test statistics, maximize, and then use simulation to form a final p-value. We too will demonstrate the potential for an increase in power both in simulations and in empirical work.

An outline of this note is as follows. In Section B we will develop the statistical methodology for more general problems than logistic regression. Section C gives a simulation study, while Section D describes empirical work. The technical details justifying the method are described for the logistic case in the Appendix A.

B. Methodology

1. Basic Framework

Let (X, S) be vectors that do not have an entry 1.0 for an intercept.

Our methodology applies to general loglikelihood functions of the form

$$\mathcal{L}\left\{Y, X^{\mathrm{T}}\beta_0 + S^{\mathrm{T}}\eta_0 + \theta_0(Z) + \gamma X^{\mathrm{T}}\beta_0\theta_0(Z), \zeta_0\right\}, \tag{2.3}$$

where β_0 and η_0 are the main effects, $\theta_0(\bullet)$ is an unknown function, ζ_0 is a nuisance parameter and γ is the interaction effect that is not to be estimated directly since it is unidentified when either $\beta_0 = 0$ or $\theta_0(\cdot)$ is a constant. All technical details will be exhibited for the logistic model (2.2), although as we indicate below, the result holds much more generally. As stated previously, the null hypothesis is $H_0 : \theta_0(\cdot) =$ constant.

2. Estimation of Model Components

To test H_0, we use the concept of a generalized likelihood ratio test (Fan et al., 2001). To implement the testing procedure, we need to estimate the model components under the full and null models.

We use a kernel based profile method to estimate the parameters under the full model. Let $K(\cdot)$ be a symmetric density function and for any bandwidth h, let $K_h(t) = K(t/h)/h$. Then the local linear profile method works as follows: for any given $(\beta, \eta, \zeta) = (\beta^*, \eta^*, \zeta^*)$ and γ, we maximize the local loglikelihood

$$\sum_{i=1}^{n} K_h(Z_i - z_0)\mathcal{L}\left(Y_i, \alpha_0 + \alpha_1(Z_i - z_0) + X_i^{\mathrm{T}}\beta^*[1 + \gamma\{\alpha_0 + \alpha_1(Z_i - z_0), \zeta^*\}] + S_i^{\mathrm{T}}\eta^*\right)$$

with respect to α_0 and α_1, and set $\widehat{\theta}(z_0, \beta^*, \eta^*, \zeta^*, \gamma) = \widehat{\alpha}_0$. Then the profile estimates

of (β, η, ζ) is obtained by maximizing

$$\sum_{i=1}^{n} \mathcal{L}\left\{Y_i, X_i^{\mathrm{T}}\beta + S_i^{\mathrm{T}}\eta + \widehat{\theta}(Z, \beta, \eta, \zeta, \gamma) + \gamma X_i^{\mathrm{T}}\beta\widehat{\theta}(Z, \beta, \eta, \zeta, \gamma), \zeta\right\}. \tag{2.4}$$

Let the resulting estimator be $(\widehat{\beta}_F, \widehat{\eta}_F, \widehat{\zeta}_F)$, where it is understood that these estimates depend on the value of γ chosen.

Estimation under the null model is a purely parametric problem where one computes the MLE in the reduced model under H_0. We add an intercept κ so that we are maximizing

$$\sum_{i=1}^{n} \mathcal{L}\left(Y_i, \kappa + X_i^{\mathrm{T}}\beta + S_i^{\mathrm{T}}\eta, \zeta\right).$$

Let $(\widehat{\kappa}_R, \widehat{\beta}_R, \widehat{\eta}_R, \widehat{\zeta}_R)$ be the resulting null model estimates.

Remark 1 For specific models, the maximization of (2.4) is quite simple and can be implemented easily. For example, in logistic regression, the steps are as follows. There is no nuisance parameter ζ. For any given (β, η), define $\mathcal{U}_i = X_i^{\mathrm{T}}\beta + S_i^{\mathrm{T}}\eta$ and $\mathcal{V}_i = 1 + \gamma X_i^{\mathrm{T}}\beta$. Then $\widehat{\theta}(z_0, \beta, \eta, \gamma)$ is the estimated intercept ξ_0 in the linear logistic regression model

$$\mathrm{pr}(Y_i = 1) = H\left\{\mathcal{U}_i + \xi_0\mathcal{V}_i + \xi_1\mathcal{V}_i(Z_i - z_0)\right\}$$

with the weights $K_h(Z_i - z_0)$. This procedure is a weighted logistic regression with no intercept, an offset $\mathcal{U}_i$, and predictors $\mathcal{V}_i$ and $\mathcal{V}_i(Z_i - z_0)$, and is hence easily implemented. Computing $(\widehat{\beta}_F, \widehat{\eta}_F)$ is then done by performing maximum likelihood under the model $\mathrm{pr}(Y_i = 1) = H\{X_i^{\mathrm{T}}\beta + S_i^{\mathrm{T}}\eta + \widehat{\theta}(Z, \beta, \eta, \gamma) + \gamma X_i^{\mathrm{T}}\beta\widehat{\theta}(Z, \beta, \eta, \gamma)\}$ based on profile method. We used the function optim() in R with initial values estimated by backfitting.

3. Properties of Profile Estimates of Parameters and Functions

In order to be able to draw upon the work of Fan and Huang (2005) and Fan, et al. (2001), we require to know the properties of the parameter and function estimates under the null hypothesis of constant $\theta_0(\cdot)$.

The properties of profile estimates of parameters and function estimates have been well-studied in fairly general contexts, see for example Claeskens and Van Keilegom (2003), Claeskens and Carroll (2007), Van Keilegom and Carroll (2007) and Apanasovich, et al. (2009), among many others. Specifically, the parameter estimates are $n^{1/2}$-consistent and the function estimates have uniform linear expansions to order $o_p(n^{-1/2})$. Conditions, summarized in Apanasovich, et al. (2009) and translated to our context, are as follows. All assumptions are meant to apply to the null hypothesis, since our asymptotic results pertain only to the null hypothesis of constant $\theta_0(\cdot)$. This means that there are simplifications to the calculations of Apanasovich, et al. (2009), who also study misspecified models, a topic not of relevance in this study.

(C.1) The kernel function K is a symmetric, continuously differentiable density function on $[-1, 1]$ taking on the value zero at the boundaries.

(C.2) The bandwidth is $h \propto n^{-1/5}$.

(C.3) The random variables (X, S, Z) have compact support. The design density $f_Z(\cdot)$ of Z is strictly positive and twice continuously differentiable on its support.

(C.4) The parameter space, here denoted by $\mathcal{B}$, is compact. For any $(\beta^*, \eta^*, \zeta^*)$, let $\theta(z_0, \beta^*, \eta^*, \zeta^*, \gamma)$ be the maximizer in v of $E[\mathcal{L}\{Y, v + X^{\mathrm{T}}\beta^*(1 + \gamma v) + S^{\mathrm{T}}\eta^*, \zeta^*\}|Z = z_0]$, which is assumed to exist. The function $\theta(\cdot, \beta, \eta, \zeta, \gamma)$ has 3 continuous derivatives in its arguments. We also assume that the same calculations done by Claeskens and Carroll (2007) can be applied to our context.

(C.5) For each γ, and under the null hypothesis, $\theta(z, \beta, \eta, \zeta, \gamma)$ is constant in z, $(\beta_0, \eta_0, \zeta_0)$ is the unique maximizer of $E(\mathcal{L}[Y, \theta(Z, \beta, \eta, \zeta, \gamma) + X^{\mathrm{T}}\beta\{1+\gamma\theta(Z, \beta, \eta, \zeta, \gamma)\} + S^{\mathrm{T}}\eta, \zeta])$. In addition, the second total derivative of this function is uniformly negative definite in a neighborhood of $(\beta_0, \eta_0, \zeta_0)$.

(C.6) We can apply the results of Claeskens and Van Keilegom (2003) as needed. In particular, their assumptions imply that uniformly in z_0, for random variables R_i possessing sufficient moments, then if subscript (1) means first derivative, if

$$\begin{aligned} C_n &= n^{-1}\textstyle\sum_{i=1}^{n} K_h(Z_i - z_0)(Z_i - z_0)^j R_i \times \{\theta(z_0) + (Z_i - z_0)\theta^{(1)}(z_0)\}; \\ D_n &= n^{-1}\textstyle\sum_{i=1}^{n} K_h(Z_i - z_0)(Z_i - z_0)^j R_i \times \{\theta(Z_i)\}, \end{aligned}$$

then

$$\begin{aligned} \sup_{z_0}|C_n - E(C_n)| &= O_p[h^j\{\log(n)/(nh)\}^{1/2}]; \\ \sup_{z_0}|D_n - E(D_n)| &= O_p[h^j\{\log(n)/(nh)\}^{1/2}]. \end{aligned}$$

Under these assumptions, at the null hypothesis, their work can be easily extended to show that uniformly on compact sets of γ, $(\widehat{\beta}_F, \widehat{\eta}_F, \widehat{\zeta}_F)$ are $n^{1/2}$-consistent estimates of $(\beta_0, \eta_0, \zeta_0)$.

In addition, at the null hypothesis, we have the following result, also uniform in compact sets of γ. If the parameter space for (β, η, ζ) is $\mathcal{B}$, the true parameter value is $\mathcal{B}_0$, and if subscripts $\mathcal{L}_\theta(\cdot)$ and $\mathcal{L}_{\theta\theta}(\cdot)$ denote the first and second derivatives with respect to θ, respectively, define

$$\Omega(z_0, \mathcal{B}_0) = E(\mathcal{L}_{\theta\theta}[Y, \theta(z_0, \beta_0, \eta_0, \zeta_0) + X^{\mathrm{T}}\beta_0\{1 + \gamma\theta(z_0, \beta_0, \eta_0, \zeta_0)\} + S^{\mathrm{T}}\eta_0]|Z = z_0).$$

Then, with $\theta_0(z_0) \equiv \theta_0$ at the null hypothesis,

$$\begin{aligned} \widehat{\theta}(z_0, \beta_0, \eta_0, \zeta_0, \gamma) &= \theta_0 - n^{-1}\textstyle\sum_{i=1}^n K_h(Z_i - z_0)R_i + o_p(n^{-1/2}); \\ R_i &= -\frac{\mathcal{L}_{\theta\theta}\{Y_i, \theta_0 + X^{\mathrm{T}}\beta_0(1+\gamma\theta_0) + S^{\mathrm{T}}\eta_0, \zeta_0\}}{f_Z(z_0)\Omega(z_0, \mathcal{B}_0)} \end{aligned} \tag{2.5}$$

4. Generalized Likelihood Ratio Test

Given any fixed γ, the generalized likelihood ratio test statistic is given by

$$\begin{aligned} \Lambda_n(\gamma) = \textstyle\sum_{i=1}^n \Big[& \mathcal{L}\left\{ Y_i, X_i^{\mathrm{T}}\widehat{\beta}_F + S_i^{\mathrm{T}}\widehat{\eta}_F + \widehat{\theta}(Z, \widehat{\beta}_F, \widehat{\eta}_F, \widehat{\zeta}_F, \gamma) \right. \\ & \left. + \gamma X_i^{\mathrm{T}}\widehat{\beta}_F\widehat{\theta}(Z, \widehat{\beta}_F, \widehat{\eta}_F, \widehat{\zeta}_F, \gamma), \widehat{\zeta}_F \right\} \\ & - \mathcal{L}\left\{ Y_i, \widehat{\kappa}_R + X_i^{\mathrm{T}}\widehat{\beta}_R + S_i^{\mathrm{T}}\widehat{\eta}_R, \widehat{\zeta}_R \right\} \Big]. \end{aligned}$$

Under H_0, from Section 3 we have that the parameters are estimated $n^{1/2}$-consistently. As in Fan and Huang (2005), this means that the likelihood ratio statistic behaves asymptotically as if the parameters are known. As they note, it is easy to show that the likelihood ratio statistic is $\Lambda_n(\gamma) = \Lambda_n^*(\gamma) + O_p(1)$, where

$$\begin{aligned} \Lambda_n^*(\gamma) = \textstyle\sum_{i=1}^n \Big[& \mathcal{L}\Big\{ Y_i, X_i^{\mathrm{T}}\beta_0 + S_i^{\mathrm{T}}\eta_0 + \widehat{\theta}(Z, \beta_0, \eta_0, \zeta_0\gamma) + \gamma X_i^{\mathrm{T}}\beta_0\widehat{\theta}(Z, \beta_0, \eta_0, \zeta_0\gamma), \zeta_0 \Big\} \\ & - \mathcal{L}\Big\{ Y_i, \kappa_0 + X_i^{\mathrm{T}}\beta_0 + S_i^{\mathrm{T}}\eta_0, \zeta_0 \Big\} \Big]. \end{aligned}$$

This statistic is easily analyzed because of the expansion (2.5), and indeed that expansion allows us to use almost exactly the proof in Fan, et al. (2001), see also Fan and Jiang (2005). Useful special cases of this general framework are the partially linear Gaussian and partially linear logistic regression models. We will show the following result for the partially linear logistic regression model in Appendix A. However, it is clear from the proof that an analogous result can be easily established for general likelihood problems.

Theorem 1 Assume conditions (C.1)-(C.6). There is a constant r_K depending on the kernel function and a deterministic sequence $\mu_n(h) \propto h^{-1} \to \infty$ depending on the bandwidth h such that

$$r_K \{\Lambda_n^*(\gamma) - \mu_n(h)\} / \{2r_K \mu_n(h)\}^{1/2} \Rightarrow \text{Normal}(0,1). \tag{2.6}$$

A consequence of Theorem 1 is that because $\mu_n(h) \to \infty$,

$$r_K \{\Lambda_n(\gamma) - \mu_n(h)\} / \{2r_K \mu_n(h)\}^{1/2} \Rightarrow \text{Normal}(0,1). \tag{2.7}$$

Result (2.7) is the so-called Wilks-phenomenon, namely that the semiparametric likelihood ratio statistic has a common limiting distribution under the null hypothesis independent of the problem.

5. Test Statistic and Implementation

While (2.7) holds, it is not very useful in practice for decision making because it depends upon the bandwidth. This fact motivated Fan and Jiang (2005) to use a bootstrap-type test. Here we propose a parametric bootstrap-type test to overcome this problem (see below).

Since the true value of γ is unknown, we follow the idea of Davies (1987), Chatterjee, et al. (2006) and Maity, et al. (2009) and propose to use as the test statistic

$$\mathcal{T}_n^* = \max_{L \le \gamma \le R} \Lambda_n(\gamma),$$

where L and R are pre-specified lower and upper bounds for γ. A normalized version of $\mathcal{T}_n^*$ as in (2.7) converges to the maximum of a Gaussian process, see the Appendix A. However, this is not very useful in terms of setting a critical level due to the dependence upon the bandwidth. We propose instead a simulation based approach to compute p-values as follows.

- Let B be a large number, and for $b = 1, ..., B$, generate response data Y_{ib} from the null model fits.
- For each of the $b = 1, ..., B$ generated data sets, compute the test statistic $T^*_{n,b}$.
- The p-value is then computed as $B^{-1} \sum_{b=1}^{B} I(T^*_{n,b} > \mathcal{T}^*_n)$.

C. Simulation Study

We simulated data using the partially linear logistic model

$$\text{pr}(Y|X, Z) = H\{X^{\mathrm{T}}\beta_0 + \theta(z) + \gamma X^{\mathrm{T}}\beta_0\theta(z)\},$$

where $H(\cdot)$ denotes the logistic distribution function. The sample size was $n = 1,200$, X was standard bivariate normal, $\beta_0 = (1, -1)^{\mathrm{T}}$, and Z was uniform on $[-2, 2]$. We repeated the simulation $1,000$ times, for true values $\gamma_{\text{true}} = 0, 1, 2$. For each simulated data set, we fit the null model, namely logistic in X, then simulated from this null model $B = 1,000$ times to obtain a p-value. The values of γ used to construct our test statistic were 11 equally spaced values on the interval $[-2, 2]$. We used the Epanechnikov kernel to estimate the function $\theta(\cdot)$ and used bandwidth $h = \widehat{\sigma}_Z n^{-1/5}$, where $\widehat{\sigma}_Z$ is the standard deviation of Z. The results were not sensitive to varying h by factors of 3.0 in each direction.

In the null case, for nominal 5% tests, the actual significance level of our test was 3.9%, while the actual significance levels of the main effects test that set $\gamma = 0$ was 5.2%. For power calculations, the alternative values of the function were given as $\theta(z) = c\,\text{sine}(2z)$ for $c = 0.125, 0.250, 0.375$.

The results are given in Figure 1. It is evident that our method has near-nominal level, little power loss in the main effects only case ($\gamma_{\text{true}} = 0$), and considerable power gain when there is an interaction.

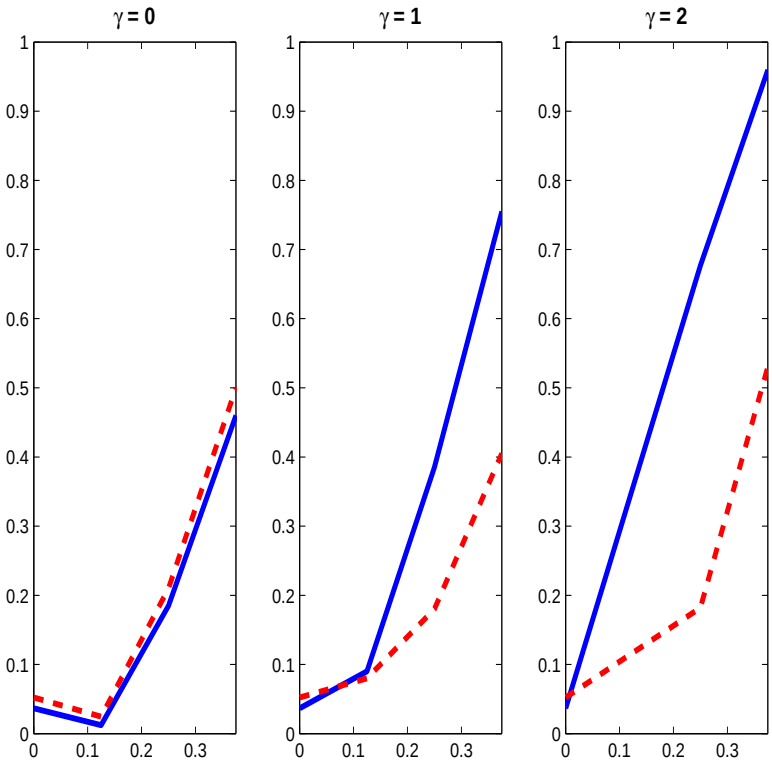

Figure 1. Results on power and level in the simulation for testing whether $\theta(\cdot)$ is constant.

Because of the Wilks phenomenon and the similarity between kernel regression and penalized spline regression, we also implemented the tests using penalized 2nd-order B-splines with equally spaced knots and 10 basis functions. Because our theory for kernel regression assumes that the same bandwidth is used for all values of γ, the penalty parameter was chosen with $\gamma = 0$ using GCV (Ruppert, et al., 2003). In fitting the non-null method, for any given γ we obtained estimates of β and $\theta(\cdot)$ by maximizing the loglikelihood function penalized by $-(\lambda/2)\zeta^{\mathrm{T}}\mathcal{K}\zeta$, where λ is the penalty parameter chosen as above, $B(z)$ are the basis functions, $\theta(z) = B^{\mathrm{T}}(z)\zeta$, and $\mathcal{K}$ is the penalty matrix. The results were almost identical to the kernel method.

D. Data Example

The data comes from a case-control study in Chatterjee et al. (2006). This study investigates the association between colorectal adenoma, a precursor of colorectal cancer, and NAT2, a candidate gene that is known to play an important role in detoxification of certain aromatic carcinogens in cigarette smoke. In our data set, we removed the nonsmokers, leaving 328 cases and 372 controls who were genotyped for six known functional polymorphisms related to NAT2 acetylation activity.

Maity et al. (2009) considered an application involving the three most common NAT2 diplotypes in comparison to the rest, which in our notation is X. The demographic variables S include gender and three indicator dummy variables for age level: between 60 and 65 years, between 65 and 70 years and more than 70 years. We explored three different environmental variables Z, namely CIG STOP, the number of years since stopping smoking, PhIP, 2-Amino-1-methyl-6-phenylimidazo[4,5-b]pyridine, which has been demonstrated to produce adenocarcinomas in mice, and Red Meat, daily grams of red meat intake.

The results are displayed in Table 1. We see that in all cases, the p-values using our method are smaller than that when γ is fixed to $= 0$. This is not a theorem of course, but it does show support with the results of the simulations, which indicate that if there is an interaction, our method will have greater statistical power.

Table 1. Significance levels in the NAT2 example

Environment	Number of Diplotypes	Our Method	Fixing $\gamma = 0$
CIG STOP	1	0.000	0.000
	2	0.000	0.000
	3	0.000	0.001
Red Meat	1	0.464	0.639
	2	0.381	0.595
	3	0.470	0.623
PhIP	1	0.984	0.935
	2	0.227	0.939
	3	0.162	0.938

E. Discussion

We have shown how to test for a constant environmental effect in the model (2.2). The methodology was described for kernel regression methods and justified in the important logistic regression case. Numerically, we have found that regression spline approaches are very close to being the same as kernel methods and much faster to compute, although their theory remains an open question in this context.

CHAPTER III

LOCALLY EFFICIENT SCORE TESTS FOR INDEPENDENCE IN THE SECONDARY ANALYSIS OF CASE-CONTROL DATA

Typical case-control studies focus on the relationship between disease D and covariates (Y, X). In the secondary analysis of case-control data, it is the relationship between Y and X that is of interest, but the analysis of this relationship is complicated by the case-control sampling framework, which is a type of biased sampling. Previous work has assumed a parametric distribution for Y given X and derived semiparametric efficient estimation and inference without any distributional assumptions about X: of course, the roles of X and Y can be interchanged.

In this study, we take up the issue of score testing for the independence of X and Y. The semiparametric efficient approaches can be used to construct semiparametric score tests, but they suffer from a lack of robustness to the assumed model for Y given X. We take an entirely different and novel approach. We show how to adjust the semiparametric score test to make its level/Type I error asymptotically correct in the rare disease case even if the assumed model for Y given X is incorrect, and thus the test is model robust. Extensions to linear regression with additional covariates are discussed. Simulations and an empirical example are used to illustrate the approach.

A. Introduction

Suppose that data are originally collected from a case-control study of a relatively rare disease. Let D be disease status, with $D = 1$ denoting a case and $D = 0$ denoting a control. Suppose also that D is to be modeled by covariates (Y, Z, X) using a standard logistic regression formulation.

There is growing awareness that such case-control data can also be exploited to

understand various facets of the relationship among (Y, Z, X). Although we deal with many different types of models, it is instructive to consider simpler cases in order to fix ideas. For example, suppose that one would like to model Y by (Z, X) using a homoscedastic additive regression model

$$Y = g_1(Z, \xi) + g_2(X, \beta) + \epsilon, \tag{3.1}$$

where $g_1(\cdot)$ and $g_2(\cdot)$ are known functions, and where ϵ has mean zero and variance σ^2 in the population, and is independent of (Z, X), but its distribution is otherwise not specified. Suppose we are further interested in knowing whether X is an independent predictor of Y given Z. We can formalize this by testing whether $\beta = 0$.

We cannot simply ignore the case-control sampling scheme and use the data *as is* to test the hypothesis that $\beta = 0$, because if (Y, X) are independent predictors of disease status D, the sampling is biased and in the case-control sample X is an independent predictor of Y. However, since the disease is rare, to a surprisingly good approximation we can test this hypothesis by simply using only the controls in the study. Strictly speaking this is asymptotically incorrect, but in practical data situations even with $5,000$ cases and $5,000$ controls, the level/Type I error of a regression test that uses the controls is very close to nominal.

The question we address here is whether in model (3.1) we can use both the cases and the controls to construct a test with greater power than using the controls only, without making strong distributional assumptions about the distribution of the experimental errors ϵ.

This chapter is organized as follows. In Section B, we start with the basic general problem that the covariates are simply (Y, X) and the interest is in knowing whether X is a predictor of Y. Here we describe recent work on case-control studies that allows an efficient score-test based solution if the distribution of Y given X is specified up to

parameters. While the solution is elegant, it suffers from the fact that the resulting test does not have the correct level if the hypothesized distribution for Y given X is misspecified, a fact we show both theoretically and in simulations (Section E).

Section C takes an entirely different and novel approach to the basic general problem, and describes a simple score-type test that is robust to misspecification of the distribution of Y given X. In Section D, we return to model (3.1) and describe a robust score-type test for the hypothesis that $\beta = 0$. Section E presents a series of simulation studies, while Section F presents a data analysis. Concluding remarks are in Section G.

B. Efficient Parametric Methods and Robustness

1. Framework

Before eventually providing a solution for model (3.1), here we consider the general problem that the covariates are (Y, X) and we wish to test whether Y and X are independent. We start with a logistic regression model underlying the case-control analysis, so that $\text{pr}(D = 1|Y, X) = H\{\theta_0 + m(Y, X, \theta_1)\}$, where $H(\cdot)$ is the logistic distribution function and $m(\cdot)$ is an arbitrary known function with unknown parameter θ_1. Let $\pi_d = \text{pr}(D = d)$, and suppose there are n_1 cases with $D = 1$ and n_0 controls with $D = 0$. Write $n = n_0 + n_1$ and define $\kappa = \theta_0 + \log(n_1/n_0) - \log(\pi_1/\pi_0)$. Parametric models start with a density/mass function for Y given X, written as $f_Y(y, x, \beta, \zeta)$, where $\beta = 0$ means that Y and X are independent, and ζ is a nuisance parameter.

2. Prior Results and Robustness

For this problem, Jiang, et al. (2006), Chen, et al. (2008) and Lin and Zheng (2009) derive the efficient profile likelihood, the latter importantly realizing that it can be used in our context. We use the notation of Chen, et al. (2008), and instead of proving formulae for the general case, we here provide formulae only for the rare disease case, the subject of this study. Define $\Omega = (\kappa, \theta_1)$ and

$$S_{\text{par}}(d, y, x, \Omega, \beta, \zeta) = f_Y(y, x, \beta, \zeta) \exp[d\{\kappa + m(y, x, \theta_1)\}]. \tag{3.2}$$

The previous authors show that the semiparametric efficient profile likelihood that makes no assumptions about the distribution of X when the distribution of Y given X is specified is, in the rare disease case, given by

$$\mathcal{L}_{\text{par}}(D, Y, X, \Omega, \beta, \zeta) = \frac{S_{\text{par}}(D, Y, X, \Omega, \beta, \zeta)}{\int \sum_{d=0}^{1} S_{\text{par}}(d, t, X, \Omega, \beta, \zeta) dt}.$$

Define $L(y, x, \zeta) = [\partial \log\{f_Y(y, x, \beta, \zeta)\}/\partial\beta]_{\beta=0}$. Then the score function for β evaluated at the null hypothesis $\beta = 0$ is

$$\begin{aligned} \mathcal{K}_{\text{par}}(Y, X, \Omega, \zeta) &= \frac{\partial \log\{\mathcal{L}_{\text{par}}(Y, X, \Omega, \beta, \zeta)\}}{\beta}\Big|_{\beta=0} \\ &= L(Y, X, \zeta) - \frac{\int \sum_{d=0}^{1} L(t, X, \zeta) S_{\text{par}}(d, t, X, \Omega, 0, \zeta) dt}{\int \sum_{d=0}^{1} S_{\text{par}}(d, t, X, \Omega, 0, \zeta) dt}. \end{aligned} \tag{3.3}$$

Because $\mathcal{L}_{\text{par}}(\cdot)$ is a legitimate semiparametric profile likelihood, when summed over the case-control data, the score statistic (3.3) has mean zero under our rare disease assumption. Implementation of course involves estimating Ω. This can be done either by maximizing the profile likelihood when $\beta = 0$ or much more easily from a logistic regression of D on (Y, X), because this yields a consistent estimate of Ω.

In the Appendix B, we show that if the distribution of Y given X is misspecified, even under the null hypothesis of independence between (Y, X), then the score

statistic does not in general have mean zero, and hence the score test is not model robust. This motivates our search for a robust score-type test, a topic we take up in the next section.

C. A Locally Efficient Robust Score Test

1. Preliminaries

First, in all our calculations and methods, we will estimate Ω consistently by a logistic regression of D on the covariates.

Now that we know that the semiparametric efficient score test is not robust to misspecification of the distribution of Y given X, we take up the topic of finding a robust test. The approach is entirely different from that described in Section 2.

We start with a conjectured model for Y given X, e.g., see just above (3.3). Let $L(Y, X, \zeta)$ be the null hypothesis score for this conjectured model. The idea is to center this null score at its expectation, where the expectation is computed *without any modeling assumptions about* Y. Remember that $\Omega = (\kappa, \theta_1)$, write the density function of X as $f_X(\cdot)$ and write the density function of Y under the null hypothesis generically as $f_Y(\cdot)$. For the moment we will assume that $f_X(\cdot)$ is known. Interest is in testing whether Y and X are independent in the population, with data from a case-control study. Define

$$S(d, y, x, \Omega) = \exp[d\{\kappa + m(y, x, \theta_1)\}]. \tag{3.4}$$

Of course, (κ, θ_1) can be estimated via ordinary logistic regression.

Under the hypothesis of independence, if Y is a continuous random variable, Spinka, et al. propose to pretend that it is discrete with support at the observed Y_i,

pretending that

$$\text{pr}(Y = Y_i) = p_{\text{est}}(Y_i) = \frac{n\pi_0}{n_0} n^{-1} \left\{ \int f_X(x) \textstyle\sum_{d=0}^{1} S(d, Y_i, x, \Omega) dx \right\}^{-1}. \tag{3.5}$$

Strictly speaking, this is only true when Y has a continuous density function, but in what follows we use (3.5) in such a way that our methods apply to the discrete case.

2. The Theoretical Score Under the Null Hypothesis

To derive the method, we consider the alternative formulation (Chen, et al., 2009) of case-control studies as random samples with missing data: of course, we use the only for intuition, and do all technical calculations in the actual case-control study. In this alternative formulation, we have random sampling and we observe (D, Y, X), which we write as $\delta = 1$, with $\text{pr}(\delta = 1 | D = d, Y, X) \propto n_d / n\pi_d$. Then, in this formulation

$$\begin{aligned} &\textstyle\sum_{d=0}^{1} \text{pr}(D = d, Y = y, X = x | \delta = 1) \\ &\quad = \frac{\{(n_d/(n\pi_d)\} \text{pr}(D = d | Y = y, X = x) \text{pr}(Y = y | X = x) f_X(x)}{\sum_{p=0}^{1} \{n_p/(n\pi_p)\} \int \text{pr}(D = p | Y = t, X = v) \text{pr}(Y = t | X = v) f_X(v)) dt dv} \\ &\quad = \frac{\sum_{d=0}^{1} S(d, y, x, \Omega) f_Y(y, x, \beta, \zeta) f_X(x)}{\sum_{p=0}^{1} \int S(p, t, v, \Omega) f_Y(t, v, \beta, \zeta) f_X(v) dt dv}. \end{aligned} \tag{3.6}$$

Then the score for β in this alternative formulation calculated under the null hypothesis is

$$L(Y, X, \zeta) - \frac{\sum_{d=0}^{1} \int L(t, x, \zeta) S(d, t, x, \Omega) f_Y(t) f_X(x) dt dx}{\sum_{d=0}^{1} \int S(d, t, x, \Omega) f_Y(t) f_X(x) dt dx}. \tag{3.7}$$

The problem of course is that we do not know the form of $f_Y(\cdot)$, so that the test statistic (3.7) cannot be implemented. The idea is to replace $f_Y(\cdot)$ in (3.7) by the

discrete distribution (3.5), leading to the test statistic

$$\begin{aligned}\mathcal{V}(\Omega) &= n^{-1}\textstyle\sum_{i=1}^{n} L(Y_i, X_i, \zeta) \\ &\quad - \dfrac{\sum_{i=1}^{n}\sum_{d=0}^{1}\int L(Y_i, x, \zeta) S(d, Y_i, x, \Omega) p_{\text{est}}(Y_i) f_X(x) dx}{\sum_{i=1}^{n}\sum_{d=0}^{1}\int S(d, Y_i, x, \Omega) p_{\text{est}}(Y_i) f_X(x) dx}. \end{aligned} \tag{3.8}$$

We now show how to simplify this test statistic, that it has mean zero under the null hypothesis *even if the model for Y is misspecified under the null hypothesis*, and that it does not have mean zero in general at alternatives.

Theorem 1 *Define*

$$U(Y, \Omega, \zeta) = \frac{\sum_{d=0}^{1}\int L(Y, x, \zeta) S(d, Y, x, \Omega) f_X(x) dx}{\sum_{d=0}^{1}\int S(d, Y, x, \Omega) f_X(x) dx}. \tag{3.9}$$

Then the test statistic $\mathcal{V}$ in (3.8) satisfies

$$\mathcal{V}(\Omega, \zeta) = n^{-1}\textstyle\sum_{i=1}^{n}\{L(Y_i, X_i, \zeta) - U(Y_i, \Omega, \zeta)\}. \tag{3.10}$$

In addition, if $f_X(\cdot)$ is specified correctly, the score test statistic (3.10) has mean zero in the case-control sampling scheme under the null hypothesis, but in general does not have mean zero at alternatives.

3. Practical Implementation and Asymptotic Theory

In order to implement the test statistic (3.10), we have to estimate $\Omega = (\kappa, \theta_1)$, ζ if applicable, and $f_X(\cdot)$. We do this as follows.

- It is well known that $\Omega = (\kappa, \theta_1)$ can be estimated consistently by ordinary logistic regression of D on (Y, X), and this is the estimate we use.

- The proof of Theorem 1 shows that as long as an estimate $\widehat{\zeta}$ is converging at the rate $O_p(n^{-1/2})$ to some value ζ_*, the robust score $\mathcal{V}(\widehat{\Omega}, \widehat{\zeta})$ will converge to zero under the null hypothesis. Based on the rare disease assumption, estimation of

ζ can be performed conveniently by using only the controls in the study, and doing the estimation under the null hypothesis using the conjectured model $f_Y(y, x, \beta = 0, \zeta)$.

- To estimate $f_X(\cdot)$, we use the rare disease assumption, namely that $f_X(\cdot) = f_{X,\text{cont}}(\cdot)$, then density of X among the controls. Then the integrals in (3.9) can be estimated unbiasedly as averages among the controls.

With these conventions, the test statistic becomes

$$\begin{aligned}\mathcal{V}(\widehat{\Omega}, \widehat{\zeta}) &= n^{-1}\textstyle\sum_{i=1}^n L(Y_i, X_i, \widehat{\zeta}) \\ &\quad -n^{-1}\textstyle\sum_{i=1}^n \dfrac{n_0^{-1}\sum_{j=1}^n (1-D_j)L(Y, X_j, \widehat{\zeta})S(d, Y, X_j, \widehat{\Omega})}{n_0^{-1}\sum_{j=1}^n (1-D_j)S(d, Y, X_j, \widehat{\Omega})}. \end{aligned} \tag{3.11}$$

We sketch a proof of the followings result in Appendix B, which uses U-statistic theory.

Theorem 2 *Assume that as $n \to \infty$, $n_0/n_1 \to c$, where $0 < c < \infty$. There is a function $\Lambda(Y, X, \Theta)$ defined in the Appendix B with the property that $E\{\Lambda(Y, X, \Theta)|D\} = 0$ such that under the null hypothesis,*

$$\begin{aligned} n^{1/2}\mathcal{V}(\widehat{\Omega}, \widehat{\zeta}) &= n^{-1/2}\textstyle\sum_{i=1}^n \Lambda(Y_i, X_i, \Theta) + o_p(1) \\ &\to \mathit{Normal}(0, \Sigma); \\ \Sigma &= \textstyle\sum_{d=0}^1 (n_d/n) cov\{\Lambda(Y, X, \Theta)|D = d\}. \end{aligned}$$

We also show in Appendix B how to estimate Σ by method of moment calculations, although we find it simpler to estimate it by using the bootstrap, resampling the cases and controls separately. With an estimate $\widehat{\Sigma}$, under the null hypothesis $n\mathcal{V}^{\mathrm{T}}(\widehat{\Omega}, \widehat{\zeta})\Sigma^{-1}\mathcal{V}(\widehat{\Omega}, \widehat{\zeta})$ is asymptotically χ^2_p, where $p = \dim(\beta)$, and the hypothesis of independence can be tested by referring to chisquared percentiles.

D. The Regression Case

We now return to model (3.1). We see that under the null hypothesis $\beta = 0$, X is independent of $Y - g_1(Z, \xi)$, where now the nuisance parameter in our general formulation is $\zeta = \xi$. Hence, all of our previous results apply if $L(Y, X, \zeta) = \{\partial g_2(X, \beta)/\partial\beta\}_{\beta=0}\{Y - g_1(Z, \xi)\}$, and, using the rare disease approximation, the nuisance parameter ξ is estimated from the regression of Y on Z among the controls.

E. Simulation Studies

We performed a series of simulation studies both at and away from an hypothesized normal model, with X binary, discrete and continuous. In general we will show that our proposed test statistic has near nominal level (Type I error) in all cases, while an implementation of the efficient test does not. We also show that while the test that uses only the controls has near nominal level, our method has much greater power.

1. When X is Binary

Our first simulation occurs when X is binary. Here we consider the case of a genotype, with minor allele frequency $p_A = 0.10$, and with genotypes generated under Hardy Weinberg Equilibrium. Assuming a dominant mode of inheritance, we have X as binary with $\text{pr}(X = 0) = (1 - p_A)^2 = 0.81$. The regression model for Y given X was taken as $Y = \beta_0 + \beta_1 X + \epsilon$, with $\beta_0 = 0$. We considered three distributions for ϵ. The hypothesized model was Normal$(0, \sigma^2)$ with $\sigma^2 = 1$. The misspecified models were (a) Chisquared(7) centered and standardized to have mean zero and variance one; and (b) centered and standardized Gamma(a, b), where $a = (0.4, 0.8, 1.4, 1.8)$ and $b = 1.9$. The logistic regression model has $m(Y, X, \theta_1) = \theta_{11} Y + \theta_{12} X$, with $\theta_{11} = 0.25$ and $\theta_{12} = 1$. The value of $\theta_0 = -3.50$ was chosen so that the rate of

disease in the population for the normal case was $\pi_1 = 0.045$. The case-control study had $n_1 = 500$ cases and $n_0 = 500$ controls. We generated $1,000$ simulated data sets. Values $\beta_1 = 0.00$ for the null hypothesis and $\beta_1 = (0.10, 0.15, 0.20, 0.30, 0.40, 0.50)$ were taken to investigate level and power. We made the rare disease assumption, so that $\Omega = \{\kappa, \theta_1 = (\theta_{11}, \theta_{12})\}$, and Ω was estimated by ordinary linear logistic regression of D on (Y, X).

Here $\zeta = (\beta_0, \sigma^2)$. We compared the efficient score test that assumes that $Y = \text{Normal}(\beta_0 + \beta_1 X, \sigma^2)$ but make no assumptions about the distribution of X with the robust method described, where for the former ζ was estimated efficiently at the null hypothesis using the controls. In both cases, the variance of the score statistic was estimated by 400 bootstrap samples. We use the rare disease assumption.

For the robust test, $\Omega = \{\kappa, \theta_1 = (\theta_{11}, \theta_{12})\}$, $\zeta = (\beta_0, \sigma^2)$, $f_Y(y, x, \beta, \zeta) = (2\pi\sigma^2)^{1/2}\phi\{(y - \beta_0 - x\beta_1)/\sigma\}$ and $L(y, x, \zeta) = [\partial \log\{f_Y(y, x, \beta, \zeta)\}/\partial\beta]_{\beta=0} = (y - \beta_0)x/\sigma^2$, where $\phi(\cdot)$ is the standard normal density function. We estimated β_0 as the mean of Y among the controls, and estimate σ^2 as the mean squared error of the linear regression of Y on X among the controls.

For the efficient test, the score function for β_1 evaluated at the null hypothesis $\beta_1 = 0$ is

$$\mathcal{K}_{\text{par}}(y, x, \Omega, \zeta) \;=\; L(y, x, \zeta) - \frac{\int \sum_{d=0}^{1} L(t, x, \zeta) S_{\text{par}}(d, t, x, \Omega, 0, \zeta) dt}{\int \sum_{d=0}^{1} S_{\text{par}}(d, t, x, \Omega, 0, \zeta) dt}.$$

Rather than implement the efficient Wald test, which requires repeated numerical integration, we instead estimate (Ω, ζ) as described above. Then, by simple algebra, the score for β_1 evaluated at $\beta_1 = 0$ is proportional to

$$\mathcal{T}_{\text{par}}(y, x, \Omega, \zeta) \;=\; xy - x\frac{\int \sum_{d=0}^{1} t S_{\text{par}}(d, t, x, \Omega, 0, \zeta) dt}{\int \sum_{d=0}^{1} S_{\text{par}}(d, t, x, \Omega, 0, \zeta) dt}. \tag{3.12}$$

In the Appendix B, we show how to compute (3.12) exactly without numerical integra-

tion. The score test statistic then becomes $\mathcal{V}_{\rm par}(\widehat{\Omega},\widehat{\zeta}) = n^{-1/2}\sum_{i=1}^{n}\mathcal{T}_{\rm par}(Y_i, X_i, \widehat{\Omega},\widehat{\zeta})$.

In Table 2, the first two columns show the test levels of the robust test and efficient score test with different distributions of ϵ. The robust test has near nominal level in all cases, as expected by the theory. Also as expected, the level of the efficient test is near nominal at the correctly specified normal distribution, but at the various Gamma distributions it has inflated Type I error.

There are of course alternative methods. One is to ignore the case-control sampling scheme entirely, and simply regress Y on X in the study data. As expected, this has inflated level in all cases, see Table 2. An alternative is to use the rare-disease assumption and regress Y on X among only the controls. Table 2 shows that this procedure has near-nominal level in all cases. However, unlike our method, it uses only 1/2 the data, and would be expected to suffer from lower power. In Table 3, we compare the power of our method to this regression of Y on X among the controls, both for the normal case and also one of the Gamma cases. In both, our method, while equally robust in terms of test level, has much greater power.

Table 2. Test levels of normal, chisquared and gamma distributions for three methods when X is binary

	Score				
	Robust	Efficient	Regression	Regression(c)	$pr(D=1)$
Normal(0,1)	0.054	0.062	0.088	0.047	0.045
Chisquared(7)	0.050	0.051	0.069	0.048	0.048
G(0.4,0.9109)	0.049	0.100	0.093	0.045	0.039
G(0.8,0.9109)	0.047	0.125	0.104	0.035	0.039
G(1.4,0.9109)	0.065	0.097	0.082	0.053	0.039
G(1.8,0.9109)	0.056	0.105	0.077	0.034	0.039

Table 3. Powers of normal and gamma distributions when X is binary

	β_1	0.1	0.15	0.2	0.3	0.4	0.5
Robust	Normal(0,1)	0.183	0.381	0.698	0.951	0.999	1.000
	G(1.8,0.9)	0.150	0.316	0.537	0.926	0.998	1.000
Regression	Normal(0,1)	0.102	0.177	0.316	0.636	0.875	0.977
	G(1.8,0.9)	0.113	0.225	0.375	0.685	0.927	0.991

2. When X is Discrete

We next performed a similar simulation, with the only change being that X is discrete with support points $(-0.40, 0.15, 1.00, 1.30, 2.10)$, with respective probabilities $(0.04, 0.08, 0.16, 0.41, 0.31)$. We chose $\theta_0 = -4.68$ so that the probability of disease is about 0.04.

The gain in power over the method that regresses Y on X among the controls is demonstrated in Table 4. Table 5, shows that the robust test performs much better than the efficient score test in terms of test level when the distribution of Y given X is misspecified.

Table 4. Powers of normal and gamma distributions when X is discrete

	β_1	0.1	0.15	0.2	0.3	0.4	0.5
Robust	Normal(0,1)	0.416	0.747	0.872	0.997	1.000	1.000
	G(1.8,0.9)	0.443	0.768	0.884	0.994	1.000	1.000
Regression	Normal(0,1)	0.272	0.524	0.811	0.989	1.000	1.000
	G(1.8,0.9)	0.270	0.564	0.812	0.990	1.000	1.000

Table 5. Test levels of normal, chisquared and gamma distributions for three methods when X is discrete

	Score				
	Robust	Efficient	Regression	Regression(c)	$pr(D=1)$
Normal(0,1)	0.055	0.067	0.124	0.041	0.041
Chisquared(7)	0.053	0.089	0.157	0.061	0.042
G(0.4,0.9109)	0.031	0.398	0.156	0.053	0.042
G(0.8,0.9109)	0.041	0.257	0.143	0.053	0.042
G(1.4,0.9109)	0.036	0.173	0.140	0.061	0.042
G(1.8,0.9109)	0.041	0.133	0.153	0.058	0.042

3. When X is Continuous and Scalar

We formed a similar simulation as in the discrete case, with the only change being that X is generated as Uniform$(0,1)$. We chose $\theta_0 = -3.7$ so that the probability of disease is about 0.04. The gain in power over the method that regresses Y on X among the controls is demonstrated in Table 6. Table 7, shows that the robust test performs much better than the efficient score test in terms of test level when the distribution of Y given X is misspecified.

Table 6. Powers of normal and gamma distributions when X is continuous

	β_1	0.1	0.15	0.2	0.3	0.4	0.5
Robust	Normal(0,1)	0.131	0.235	0.367	0.700	0.925	0.994
	G(1.8,0.9)	0.104	0.190	0.315	0.650	0.881	0.979
Regression	Normal(0,1)	0.102	0.159	0.250	0.494	0.704	0.890
	G(1.8,0.9)	0.090	0.144	0.225	0.461	0.720	0.893

Table 7. Test levels of normal, chisquared and gamma distributions for three methods when X is continuous

	Score				
	Robust	Efficient	Regression	Regression(c)	$pr(D=1)$
Normal(0,1)	0.053	0.057	0.067	0.056	0.042
Chisquared(7)	0.041	0.082	0.069	0.049	0.042
G(0.4,0.9109)	0.052	0.378	0.083	0.048	0.042
G(0.8,0.9109)	0.042	0.398	0.074	0.056	0.042
G(1.4,0.9109)	0.046	0.385	0.079	0.059	0.042
G(1.8,0.9109)	0.047	0.396	0.067	0.041	0.042

4. Model with Covariates

In this section we add a covariate Z in our model since in practice we always have covariates. We generate Z as Uniform$(0,1)$. The regression model for Y given (X,Z) was then taken as $Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$, with $\beta_0 = 1$ and $\beta_2 = 0.5$. We use the controls to run the regression of Y on Z to estimate $\zeta = (\beta_1, \beta_2)$, then call $Y_* = Y - \hat{\xi}_0 - \hat{\xi}_1 Z$, the residuals, and run the test as if Y_* were our response. We got similar results as the without covariates case in terms of test levels and power.

F. Applications

We analyze two gene (X) environment (Y) interaction data sets, one in which Y and X are thought to be independent, and the other in which Y and X are thought to be related.

1. Prostate Cancer Example

Chen, et al. (2009) investigate a case-control study of prostate cancer. The sample includes 749 prostate cancer cases and 781 controls, also selected from the screening arm of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial at the National Cancer Institute, USA (Gohagan, Prorok, Hayes, and Kramer 2000; Moslehi et al. 2006). The main objective of the study is to examine the relationship between risk of prostate cancer and [25(OH)D], a serum level biomarker of vitamin D, that reflects both dietary and sunlight exposures. The anticancer effect of vitamin D is hypothesized due to the ability of prostate cells to convert [25(OH)D)] into 1,25-dihydroxy-vitamin D [1,25(OH)2D], the most active form of this vitamin, which regulates the gene transcription of many proteins involving cellular differentiation, proliferation, and apoptosis via the vitamin D receptor (VDR). Chen, et al. (2009) write "*Given the downstream role of the VDR gene in the vitamin-D pathway, it is very unlikely that these polymorphisms actually could influence the level of the [25(OH)D] itself. Thus, the gene-environment independence assumption in this application is likely to be valid*".

The notation of this chapter, D is the prostate cancer case-control status, Y is the level of 1,25-dihydroxy-vitamin D and X is one of three SNP, and Z is age level categorized into four groups and hence with three dummy variables. The quote above suggests that Y and X are independent in the population, and hence our test should find no evidence that any of the SNP are related to 1,25-dihydroxy-vitamin D level. Indeed, this is the case: all p-values are greater than 0.35, as expected.

2. Colorectal Adenoma

Chen, et al. also discuss a case-control study of colorectal adenoma, a precursor of colorectal cancer. The study sample includes 628 prevalent advanced adenoma cases and 635 gender-matched controls, also selected from the screening arm of the PLCO Study. One of the main objectives of this study is to assess whether the smoking-related risk of colorectal adenoma may be modified by certain haplotypes in NAT2, a gene known to be important in the metabolism of smoking related carcinogens. In addition, because NAT2 is involved in the smoking metabolism pathway, potentially it can influence an individuals addiction to smoking itself, causing the gene-environment independence assumption to be violated. In other words, here D is colorectal adenoma status, Y are various measures of smoking status, and X are various haplotypes. Chen, et al. use Z as age in years and gender. They claim that it is reasonable to suppose that in the population, Y and X are related.

The smoking variables used here are (a) years since stopping smoking censor at 45; (b) Number of packs smoked per day; and (c) pack years subtract 0.25 and censor at 100, i.e., packs per day times years smoked. As in Maity, et al. (2009), we let X be the indicator of the most common diplotype.

The results are given in Table 8. We see a statistically significantly protective of the most common diplotype for the years since stopping smoking. Crucially, all the p-values from the robust test are similar to those from regression among controls only, and are much less than those from the efficient score test. This is not a theorem of course, but it does show support with the results of the simulations, which indicate that the efficient score test can have the wrong level if the distribution of Y given X is misspecified. We also randomly sub-sampled 80% of the people 1,000 times for the smoking variable (a), and found that 61.2% of times our test rejected the null, while

the regression among controls rejected the null 34.9% of times. This indicated our test will have great gain in statistical power over the method that regresses Y on X among the controls.

Table 8. P-values in the NAT2 example

Environment	Robust	Non-Robust	Regression
CIG STOP	0.041	0.081	0.036
PACK YEARS	0.215	0.364	0.243
PACK DAY	0.198	0.382	0.227

G. Discussion

The study of the relationship among secondary variables in a case-control study is of great practical interest, because large case-control studies now exist and especially include predictors or phenotypes Y and demographic, environmental and genetic factors. As we have noted, the semiparametric efficient approaches can be used to construct semiparametric score tests, but they suffer from a lack of robustness to the assumed model for Y given X: it is possible to create skew distributions for the regression errors that result in bias when normality is assumed.

Our approach is entirely different. While we specify a target regression error distribution, we have shown that the test procedure is robust to violation of that target distribution, both theoretically and in a simulation study. In the rare disease case that would be the reason for a case-control study in the first place, an alternative is to simply use only the data for the controls. We have shown in simulations and in our two data examples that such throwing away of 50% of the data leads to a highly non-trivial loss of power compared to our method.

CHAPTER IV

LOCALLY EFFICIENT ESTIMATION FOR HOMOSCEDASTIC REGRESSION IN THE SECONDARY ANALYSIS OF CASE-CONTROL DATA

Typical case-control studies focus on the relationship between disease D and covariates (Y, X). In the secondary analysis of case-control data, it is the relationship between Y and X that is of interest, but the analysis of this relationship is complicated by the case-control sampling framework. Previous work has assumed a parametric distribution for Y given X and derived semiparametric efficient estimation and inference without any distributional assumptions about X.

In this project, we take up the issue of estimation of a regression function when Y given X follows a homoscedastic regression model. The semiparametric efficient approaches can be used to construct semiparametric efficient estimates, but they suffer from a lack of robustness to the assumed model for Y given X. We take an entirely different and novel approach in the case that the disease is rare. We show how to estimate the regression parameters in the rare disease case even if the assumed model for Y given X is incorrect, and thus the estimates are model-robust. Simulations and an empirical example are used to illustrate the approach.

A. Introduction

Suppose that data are originally collected from a case-control study of a relatively rare disease. Let D be disease status, with $D = 1$ denoting a case and $D = 0$ denoting a control. Suppose also that D is to be modeled by covariates (Y, X) using a standard logistic regression formulation.

There is growing awareness that such case-control data can also be exploited to understand various facets of the relationship among (Y, X). We consider here the

homoscedastic regression model

$$Y = \alpha_{\text{true}} + \mu(X, \beta_{\text{true}}) + \epsilon, \tag{4.1}$$

where α_{true} is an intercept, $\mu(\cdot)$ is a known function, and where ϵ has mean zero and is independent of X, but its distribution is otherwise not specified.

To estimate $(\alpha_{\text{true}}, \beta_{\text{true}})$, we cannot simply ignore the case-control sampling scheme and use the data *as is*, because if X is an independent predictor of disease status D, the sampling is biased and in the case-control sample model (4.1) will not hold. However, since the disease is rare, to a surprisingly good approximation we can indeed use only the controls and obtain an approximately consistent estimate.

The question we address here is whether in model (4.1) we can use both the cases and the controls to construct more efficient estimates of $(\alpha_{\text{true}}, \beta_{\text{true}})$ than using the controls only, at the same time without making distributional assumptions about the distribution of the experimental errors ϵ.

This chapter is organized as follows. In Section B, we describe recent work on case-control studies that allows efficient estimation if the distribution of Y given X is specified up to parameters. While the solution is elegant, it suffers from the fact that the resulting estimate is biased if the hypothesized distribution for Y given X is misspecified.

Section C takes an entirely different approach to the basic general problem, and describes a simple method that is robust to misspecification of the distribution of Y given X. Section D presents a series of simulation studies, while Section E presents a data analysis. Concluding remarks are in Section F. Technical details are given in an appendix.

B. Efficient Parametric Methods and Robustness

1. Framework

We start with a logistic regression model underlying the case-control analysis, so that $\text{pr}(D = 1|Y, X) = H\{\theta_0 + m(Y, X, \theta_1)\}$, where $H(\cdot)$ is the logistic distribution function and $m(\cdot)$ is an arbitrary known function with unknown parameter θ_1. Let $\pi_d = \text{pr}(D = d)$, and suppose there are n_1 cases with $D = 1$ and n_0 controls with $D = 0$, write $n = n_0 + n_1$ and define $\kappa = \theta_0 + \log(n_1/n_0) - \log(\pi_1/\pi_0)$. Write the parametric model for Y given X as $f_\epsilon\{y - \mu(x, \beta), \zeta\}$. If in the population Y given X is normally distributed, then $\zeta = \text{var}(\epsilon)$.

2. Prior Results and Robustness

For this problem, Jiang, et al. (2006), Chen, et al. (2008) and Lin and Zheng (2009) derive the efficient profile likelihood, the latter noting importantly that it can be used in our context. We use the notation of Chen, et al. (2008), and instead of proving formulae for the general case, we here provide formulae only for the rare disease case, the subject of this study. Define $\Omega = (\kappa, \theta_1)$ and

$$S_{\text{par}}(d, y, x, \Omega, \alpha, \beta, \zeta) = f_\epsilon\{y - \alpha - \mu(x, \beta), \zeta\} \exp[d\{\kappa + m(y, x, \theta_1)\}]. \qquad (4.2)$$

The previous authors show that the semiparametric efficient profile likelihood that makes no assumptions about the distribution of X when the distribution of Y given X is specified is

$$\mathcal{L}_{\text{par}}(D, Y, X, \Omega, \alpha, \beta, \zeta) = \frac{S_{\text{par}}(D, Y, X, \Omega, \alpha, \beta, \zeta)}{\int \sum_{d=0}^{1} S_{\text{par}}(d, t, X, \Omega, \alpha, \beta, \zeta) dt}.$$

Taking logarithms, summing over the observed data and then maximizing in the parameters yields semiparametric efficient inference. A difficulty arises however if the

density $f_\epsilon(\cdot)$ of ϵ is not specified properly.

To see what happens, we consider the score for β. Define $L_{\text{par}}(y, x, \alpha, \beta, \zeta) = \partial \log[f_\epsilon\{y - \alpha - \mu(x, \beta), \zeta\}]/\partial\beta$. Then the score for β is

$$\begin{aligned} \mathcal{K}_{\text{par}}(D, Y, X, \Omega, \alpha, \beta, \zeta) &= \frac{\partial \log\{\mathcal{L}_{\text{par}}(D, Y, X, \Omega, \alpha, \beta, \zeta)\}}{\partial\beta} \\ &= L_{\text{par}}(Y, X, \alpha, \beta, \zeta) \\ &\quad - \frac{\int \sum_{d=0}^{1} L_{\text{par}}(t, X, \alpha, \beta, \zeta) S_{\text{par}}(d, t, X, \Omega, \alpha, \beta, \zeta) dt}{\int \sum_{d=0}^{1} S_{\text{par}}(d, t, X, \Omega, \alpha, \beta, \zeta) dt}. \end{aligned} \tag{4.3}$$

Because $\mathcal{L}_{\text{par}}(\cdot)$ is a legitimate semiparametric profile likelihood, when summed over the case-control data and evaluated at the true parameters, the score (4.3) has mean zero under our rare disease assumption. Unfortunately, (4.3), when evaluated at the true parameter values, only has mean zero in general if the density $f_\epsilon(\cdot)$ of ϵ is specified properly, i.e., the approach is not model robust. This motivates our search for a robust estimation method, a topic we take up in the next section.

C. Model-Robust Estimation

1. Preliminaries

Our method involves a multi-step process.

- Estimate the logistic regression parameters Ω_{true} by ordinary logistic regression of D on (Y, X). This can be done legitimately because it is well known that ordinary logistic regression in a case-control study consistently estimates Ω_{true}. Call the result $\widehat{\Omega}$.

- Write a conjectured version of the model for Y given X. In our case, we conjecture a normal distribution with constant variance ζ. In general, we conjecture a distribution where ϵ depends on a nuisance parameter ζ.

- Compute the score function of the conjectured model for β and ζ. In the normal case that we pursue, define $R(\beta) = Y - \mu(X, \beta)$ in which case the score function for the conjectured model is proportional to ζ, which can be ignored, and hence is

$$L\{R(\beta), X, \alpha, \beta\} = \mu_\beta(X, \beta)\{R(\beta) - \alpha\}, \tag{4.4}$$

  where the subscript means differentiation with respect to β.

- The score (4.4) will not have mean zero in the case-control sampling scheme, so we adjust it so that it has mean zero in general, even if the conjectured model is not correct.

- For technical reasons described later, estimation of α_{true} has to be done via an auxiliary equation depending on the current values, which we generically call $\widehat{\alpha}(\beta, \widehat{\Omega})$, which replaces α in the score (4.4), see below for the definition.

- Solve the adjusted score equation to estimate β_{true} and hence α_{true}. Good starting values for β can be obtained by least squares regression among the controls.

2. The Methodology

The development of our methodology is somewhat involved. Here we simply state our proposal, with its development given in subsequent subsections. As before, define $R(\beta) = Y - \mu(X, \beta)$ and define $\mathcal{K}\{R_i(\beta), x, \beta, \Omega\} = 1 + \exp[\kappa + m\{R_i(\beta) + \mu(x, \beta), x, \Omega\}]$. Let

$$\widehat{\alpha}(\beta, \Omega) = \frac{n^{-1}\sum_{i=1}^n R_i(\beta)\{n_0^{-1}\sum_{j=1}^n (1 - D_j)\mathcal{K}\{R_i(\beta), X_j, \beta, \Omega\}\}^{-1}}{n^{-1}\sum_{i=1}^n \{n_0^{-1}\sum_{j=1}^n (1 - D_j)\mathcal{K}\{R_i(\beta), X_j, \beta, \Omega\}\}^{-1}}. \tag{4.5}$$

Let $\mu_\beta(x,\beta) = \partial\mu(x,\beta)/\partial\beta$ and let $L\{R(\beta), X, \alpha, \beta\}$ be as in (4.4). Then define

$$\widehat{Q}_{n,\text{est}}(\beta,\Omega) = n^{-1/2}\sum_{i=1}^n \Bigg[L\{R_i(\beta), X_i, \widehat{\alpha}(\beta,\Omega), \beta\} - \frac{n_0^{-1}\sum_{j=1}^n (1-D_j) L\{R_i(\beta), X_j, \widehat{\alpha}(\beta,\Omega), \beta\} \mathcal{K}\{R_i(\beta), X_j, \beta, \Omega\}}{n_0^{-1}\sum_{j=1}^n (1-D_j)\mathcal{K}\{R_i(\beta), X_j, \beta, \Omega\}} \Bigg]. \tag{4.6}$$

Our algorithm then is as follows.

- Estimate Ω by $\widehat{\Omega}$, the logistic regression estimates of D on (Y, X). These are known to produce consistent estimates of Ω_{true}.
- Solve $0 = \widehat{Q}_{n,\text{est}}(\beta, \widehat{\Omega})$ in β to obtain our estimate $\widehat{\beta}$.

In the next few subsections, we describe how we obtained (4.6), and at the end we describe the asymptotic distribution theory.

3. Development of the Score when $f_X(\cdot)$ and α_{true} are Known

We first describe how to proceed when the intercept α_{true} and the density $f_X(\cdot)$ of X in the population are known; of course they are not and we will show how to remove these restrictions in subsequent sections. In what follows, we use the notation E_{cc} as a short hand notation for expectation in the case-control sampling scheme. Thus, $E_{\text{cc}}\{G(D,Y,X)\} = n^{-1}\sum_{i=1}^n E\{G(D_i,Y_i,X_i)|D_i\} = \sum_{d=0}^1 (n_d/n) E\{G(D,Y,X)|D = d\}$.

To derive the method, we consider the alternative formulation (Chen, et al., 2009) of case-control studies as random samples with missing data. Of course, we use this only for intuition, and do all technical calculations in the actual case-control study. In this alternative formulation, we have random sampling and we observe (D, Y, X), thus setting the binary $\delta = 1$, with $\text{pr}(\delta = 1|D = d, Y, X) \propto n_d/n\pi_d$. Then, in this

formulation

$$\begin{aligned}&\mathrm{pr}(D=d,Y=y,X=x|\delta=1)\\&\quad=\frac{\{(n_d/(n\pi_d)\}\mathrm{pr}(D=d|Y=y,X=x\}\mathrm{pr}\{Y=y|X=x)f_X(x)}{\sum_{p=0}^1\{n_p/(n\pi_p)\}\int\mathrm{pr}(D=p|Y=t,X=v)\mathrm{pr}\{Y=t|X=v\}f_X(v)dtdv}.\end{aligned}$$

This means that the regression model in the alternative formulation is

$$\begin{aligned}&\mathrm{pr}(Y=y,X=x|\delta=1)\\&\quad=\frac{\sum_{d=0}^1(n_d/\pi_d)\mathrm{pr}(D=d|Y=y,X=x\}f_\epsilon\{y-\alpha_{\mathrm{true}}-\mu(x,\beta),\zeta\}f_X(x)}{\sum_{p=0}^1(n_p/\pi_p)\int\mathrm{pr}(D=p|Y=t,X=v)f_\epsilon\{t-\alpha_{\mathrm{true}}-\mu(v,\beta),\zeta\}f_X(v)dtdv}.\end{aligned}$$

We now make the rare disease approximation so that $\sum_{d=0}^1(n_d/\pi_d)\mathrm{pr}(D=d|Y=y,X=x\}=(n_0/\pi_0)+(n_1/\pi_1)\exp\{\theta_0+m(y,x,\theta_1)\}=(n_0/\pi_0)[1+\exp\{\kappa+m(y,x,\theta_1)\}]$. If the conjectured model is the normal distribution, then we can drop the term ζ and up to a constant of proportionality the score for β is

$$\begin{aligned}&\mu_\beta(X,\beta)\{Y-\alpha_{\mathrm{true}}-\mu(X,\beta\}\\&-\int\frac{\mu_\beta(v,\beta)\{t-\alpha_{\mathrm{true}}-\mu(v,\beta)\}f_\epsilon\{t-\alpha_{\mathrm{true}}-\mu(v,\beta)\}}{\int f_\epsilon\{t-\alpha_{\mathrm{true}}-\mu(v,\beta)\}[1+\exp\{\kappa+m(y,v,\theta_1)\}]f_X(v)dtdv}\\&\quad\times[1+\exp\{\kappa+m(y,v,\theta_1)\}]f_X(v)dtdv.\end{aligned}$$

We now make the change of variables $R(\beta)=Y-\mu(X,\beta)$, and recall that $\mathcal{K}(r,x,\beta,\Omega)=1+\exp[\kappa+m\{r+\mu(x,\beta),x,\Omega\}]$. Referring to (4.4), this means that the score for β in the alternative formulation of Chen, et al. (2009) is

$$L\{R(\beta_{\mathrm{true}}),X,\alpha_{\mathrm{true}},\beta\}-\frac{\int L\{t,x,\alpha_{\mathrm{true}},\beta\}\mathcal{K}(t,x,\beta,\Omega)f_\epsilon(t)f_X(x)dtdx}{\int\mathcal{K}(t,x,\beta,\Omega)f_\epsilon(t)f_X(x)dtdx}.\qquad(4.7)$$

The problem of course is that we do not know the form of $f_\epsilon(\cdot)$, so that the score (4.7) cannot be implemented. Spinka, et al. address this issue, although not directly in our context. Noting that $R(\beta_{\mathrm{true}})$ and X are independent, we propose to replace

$f_\epsilon(\cdot)$ by pretending that it is discrete with support at the observed $R_i(\beta)$, so that

$$\begin{aligned} \mathrm{pr}\{R(\beta) = R_i(\beta)\} &= p_{\mathrm{est}}\{R_i(\beta), \Omega\} \\ &= \frac{n\pi_0}{n_0} n^{-1} \left\{ \int f_X(x) \mathcal{K}\{R_i(\beta), x, \beta, \Omega\} dx \right\}^{-1}. \end{aligned} \tag{4.8}$$

When we make this substitution in (4.7) and sum over the data, the score becomes

$$\begin{aligned} &\textstyle\sum_{i=1}^n L\{R_i(\beta), X_i, \alpha_{\mathrm{true}}, \beta\} \\ &\quad - \frac{\sum_{i=1}^n \int L\{R_i(\beta), x, \alpha_{\mathrm{true}}, \beta\} \mathcal{K}\{R_i(\beta), x, \beta, \Omega\} p_{\mathrm{est}}\{R_i(\beta), \Omega\} f_X(x) dx}{n^{-1} \sum_{i=1}^n \int \mathcal{K}\{R_i(\beta), x, \beta, \Omega\} p_{\mathrm{est}}\{R_i(\beta), \Omega\} f_X(x) dx}. \end{aligned}$$

Because the denominator of this expression is π_0/n_0, by simple algebra is it readily seen that the normalized score function for estimating β can be defined as

$$\begin{aligned} 0 &= Q_n(\alpha_{\mathrm{true}}, \beta, \Omega) \\ &= n^{-1/2} \textstyle\sum_{i=1}^n \left[L\{R_i(\beta), X_i, \alpha_{\mathrm{true}}, \beta\} \right. \\ &\qquad \left. - \frac{\int L\{R_i(\beta), x, \alpha_{\mathrm{true}}, \beta\} \mathcal{K}\{R_i(\beta), x, \beta, \Omega\} f_X(x) dx}{\int \mathcal{K}\{R_i(\beta), x, \beta, \Omega\} f_X(x) dx} \right]. \end{aligned} \tag{4.9}$$

In Appendix C, we show that $E_{\mathrm{cc}}\{Q_n(\alpha_{\mathrm{true}}, \beta, \Omega)\}$ equals zero when evaluated at $(\alpha_{\mathrm{true}}, \beta_{\mathrm{true}}, \Omega_{\mathrm{true}})$, but not generally, and thus (4.9) is an unbiased estimating equation *in the case-control sampling scheme*, and not just the alternative formulation.

Remark 1 If the conjectured model $f_\epsilon\{y - \alpha_{\mathrm{true}} - \mu(x, \beta), \zeta\}$ is not the normal model, then ζ must also be estimated. This can be done by replacing $L\{R(\beta), X, \alpha_{\mathrm{true}}, \beta\}$ by $\partial \log[f_\epsilon\{Y - \alpha_{\mathrm{true}} - \mu(X, \beta), \zeta\}]/\partial(\beta, \zeta)$.

4. Implementation when $f_X(\cdot)$ is Unknown but α_{true} is Known

Of course, $f_X(\cdot)$ is not known. Because the disease is rare, we propose to approximate $f_X(\cdot)$ by $f_{X,\mathrm{cont}}(\cdot)$, the density of X among the controls. We then estimate the integrals in (4.9) unbiasedly by their averages among the controls, so that our estimating

equation is

$$\begin{aligned} 0 &= \widehat{Q}_n(\alpha_{\text{true}}, \beta, \Omega) \qquad (4.10) \\ &= n^{-1/2}\textstyle\sum_{i=1}^{n}\Big[L\{R_i(\beta), X_i, \alpha_{\text{true}}, \beta\} \\ &\qquad -\dfrac{n_0^{-1}\sum_{j=1}^{n}(1-D_j)L\{R_i(\beta), X_j, \alpha_{\text{true}}, \beta\}\mathcal{K}\{R_i(\beta), X_j, \beta, \Omega\}}{n_0^{-1}\sum_{j=1}^{n}(1-D_j)\mathcal{K}\{R_i(\beta), X_j, \beta, \Omega\}}\Big]. \end{aligned}$$

5. When the Intercept α_{true} is Unknown

In most cases, the mean function will include an intercept, although of course our methods are easily modified in case no intercept exists.

One might reasonably think that estimating the intercept is easy, e.g., simply supplement the score with the score for the intercept, so that $L\{R(\beta), X, \alpha, \beta\} = \{1, \mu_\beta^{\mathrm{T}}(X, \beta)\}^{\mathrm{T}}\{R(\beta) - \alpha\}$. The problem with this is that the first component of the estimating equation (4.10) would then be identically zero, and thus will not produce an estimate of the intercept. The reason for this is that the solution (4.8) was calculated nonparametrically under the assumption that $R(\beta_{\text{true}})$ and X are independent in the population. Since $Y - \alpha_{\text{true}} - X^{\mathrm{T}}\beta_{\text{true}}$ and $Y - X^{\mathrm{T}}\beta_{\text{true}}$ are both independent of X in the population, this means that (4.8) cannot lead to an estimate of the intercept. Hence, an alternative approach is required.

To overcome this problem, we estimate the intercept of $R(\beta)$ using the tilting suggested by Spinka, et al., i.e., if $f_X(\cdot)$ were known, then α could be estimated by

$$\widehat{\alpha}_1(\beta, \Omega) = \frac{n^{-1}\sum_{i=1}^{n} R_i(\beta) p_{\text{est}}\{R_i(\beta), \Omega\}}{n^{-1}\sum_{i=1}^{n} p_{\text{est}}\{R_i(\beta), \Omega\}},$$

a quantity that is free of the π_0 that shows up in (4.8). If we then invoke the rare disease approximation and replace the integral in the definition of $p_{\text{est}}(\cdot)$ by its therefore unbiased average $n_0^{-1}\sum_{j=1}^{n}(1-D_j)K\{R_i(\beta), X_j, \beta, \Omega\}$, we get exactly (4.5). Making this substitution in (4.10), we obtain (4.6). This completes the derivation of

our methodology.

6. Distribution Theory

Let $(\Omega, \beta) = \Theta$, and let Θ_{true} denote its true value. Recall that $f_{X,\text{cont}}(\cdot)$ is the density function of X among the controls, and define

$$\begin{aligned}
\alpha(\beta, \Omega) &= \frac{E_{\text{cc}}(R(\beta)[\int f_{\text{cont}}(x)\mathcal{K}\{R(\beta), x, \beta, \Omega\}dx]^{-1})}{E_{\text{cc}}([\int f_{\text{cont}}(x)\mathcal{K}\{R(\beta), x, \beta, \Omega\}dx]^{-1})}; \\
\mathcal{T}\{R(\beta), X, \Theta, f_{X,\text{cont}}\} &= L\{R(\beta), X, \alpha(\beta, \Omega), \beta\} \\
&\quad - \frac{\int L\{R(\beta), x, \alpha(\beta, \Omega), \beta\}\mathcal{K}\{R(\beta), x, \Theta\}f_{X,\text{cont}}(x)dx}{\int \mathcal{K}\{R(\beta), x, \Theta\}f_{X,\text{cont}}(x)dx} \\
\mathcal{M}_\Omega &= E_{\text{cc}}\left[\frac{\partial \mathcal{T}\{R(\beta_{\text{true}}), X, \Theta_{\text{true}}, f_{X,\text{cont}}\}}{\partial \Omega^{\text{T}}}\right]; \\
\mathcal{M}_\beta &= E_{\text{cc}}\left[\frac{\partial \mathcal{T}\{R(\beta_{\text{true}}), X, \Theta, f_{X,\text{cont}}\}}{\partial \beta^{\text{T}}}\right].
\end{aligned}$$

Define $G_{\text{num}}(r, x, \Theta) = L\{r, x, \alpha(\beta, \Omega)\beta\}\mathcal{K}(r, x, \Theta)$ and $G_{\text{den}}(r, x, \Theta) = \mathcal{K}(r, x, \Theta)$ Define $\mathcal{A}_{\text{num}}(r, \Theta) = E\{G_{\text{num}}(r, X, \Theta)|D = 0\}$ and $\mathcal{A}_{\text{den}}(r, \Theta) = E\{G_{\text{den}}(r, X, \Theta)|D = 0\}$. Define

$$\mathcal{H}_n(\beta, \Theta) = n^{-1/2}\sum_{i=1}^n \left[\frac{n_0^{-1}\sum_{j=1}^n (1 - D_j)G_{\text{num}}\{R_i(\beta), X_j, \Theta\}}{n_0^{-1}\sum_{j=1}^n (1 - D_j)G_{\text{den}}\{R_i(\beta), X_j, \Theta\}} - \frac{\mathcal{A}_{\text{num}}\{R_i(\beta), \Theta\}}{\mathcal{A}_{\text{den}}\{R_i(\beta), \Theta\}}\right].$$

Define

$$\begin{aligned}
&W\{R_i(\beta), X_j, D_j, \Theta\} \\
&\quad = (1 - D_j)\frac{G_{\text{num}}\{R_i(\beta), X_j, \Theta\} - \mathcal{A}_{\text{num}}\{R_i(\beta), \Theta\}}{\mathcal{A}_{\text{den}}\{R_i(\beta), \Theta\}} \\
&\quad -(1 - D_j)\frac{\mathcal{A}_{\text{num}}\{R_i(\beta), \Theta\}[G_{\text{den}}\{R_i(\beta), X_j, \Theta\} - \mathcal{A}_{\text{den}}\{R_i(\beta), \Theta\}]}{\mathcal{A}^2_{\text{den}}\{R_i(\beta), \Theta\}}.
\end{aligned}$$

Also define

$$
\begin{aligned}
c_* &= \lim_{n\to\infty}(n/n_0); \\
\widetilde{Z}_i(\beta) &= \{R_i(\beta), X_i, D_i\}; \\
\widetilde{z} &= (r, x, d); \\
Q_1\{\widetilde{Z}_i(\beta), \widetilde{Z}_j(\beta), \Theta\} &= W\{R_i(\beta), X_j, D_j, \Theta\} + W\{R_j(\beta), X_i, D_i, \Theta\}; \\
Q_2(\widetilde{z}, \beta, \Theta) &= E[W\{R(\beta), x, d, \Theta\}|D = 1]; \\
h_1(d, \widetilde{z}, \beta, \Theta) &= E\{Q_1\{\widetilde{z}, \widetilde{Z}(\beta), \Theta\}|D = d\}; \\
h_2\{R_i(\beta), X_i, D_i, \Theta\} &= c_*(n_0/n)(1 - D_i)h_1\{D_i, \widetilde{Z}_i(\beta), \beta, \Theta\} \\
&\qquad + c_*(n_1/n)(1 - D_i)Q_2\{\widetilde{Z}_i(\beta), \beta, \Theta\}; \\
\Phi(y, x, d, \Omega) &= \{1, m_\Omega(y, x, \theta_1)\}^{\mathrm{T}}[D - H\{\kappa + m(y, x, \theta_1)\}]; \\
\mathcal{N}_\Omega &= -\left[E_{\mathrm{cc}}\left\{\partial\Phi(Y, X, D, \Omega)/\partial\Omega\right\}\right]^{-1}; \\
\mu_1(d) &= E[\mathcal{T}\{R(\beta_0), X, \Theta_0, f_X)|D = d]; \\
\mu_4(d) &= E\{\Phi(Y, X, D, \Omega_0)|D = d\}; \\
\Lambda(Y_i, X_i, D_i, \Theta_0) &= \mathcal{M}_\Omega(\Theta_0)\mathcal{N}_\Omega(\Omega_0)\{\Phi(Y_i, X_i, D_i, \Omega_0) - \mu_4(D_i)\} \\
&\qquad - h_2\{R_i(\beta_0), X_i, D_i, \Theta_0\} \\
&\qquad + \left[\mathcal{T}\{R_i(\beta_0), X_i, \Theta_0, f_X) - \mu_1(D_i)\right].
\end{aligned}
$$

The asymptotic distribution of our estimator in the rare event case is given in the following result, the proof of which is sketched in Appendix C.

Theorem 3 *Let $(\Omega, \beta) = \Theta$, and let Θ_{true} denote its true value. Assume that $n_1/n_0 \to c$, where $0 < c < \infty$. Also assume that $\mathcal{M}_\beta$ is invertible. Under the rare disease approximation that the distribution of X in the population is approximately the same as the distribution of X among the controls, $E\{\Lambda(Y, X, D, \Theta_{\mathrm{true}})|D\} = 0$*

and

$$n^{-1/2}(\widehat{\beta} - \beta_{\text{true}}) = n^{-1/2}\textstyle\sum_{i=1}^{n}\Lambda(Y,X_i,D,\Theta_{\text{true}}) + o_p(1) \quad (4.11)$$
$$\rightarrow \; \textit{Normal}\left[0, \Sigma = \textstyle\sum_{d=0}^{1}(n_d/n)\,cov\{\Lambda(Y,X,D,\Theta_{\text{true}})|D=d\}\right].$$

An estimate of Σ *can be obtained via the bootstrap or by substitution into the various terms composing* $\Lambda(\cdot)$ *and then using its sample covariance matrix.*

D. Simulation

We performed a small simulation study to assess the bias, coverage probability and efficiency of our method with respect to linear regression only among the controls.

In our simulation study we generated X as Uniform(0,1), the regression model for Y given X was taken as $Y = \beta_0 + \beta_1 X + \epsilon$, with $\beta_0 = \beta_1 = 0$. We considered three distributions for ϵ. The conjectured model was Normal$(0, \sigma^2)$ with $\sigma^2 = 1$. The misspecified models were (a) Chisquared(7) centered and standardized to have mean zero and variance one; and (b) centered and standardized Gamma(a, b), where $a = (0.4, 0.8, 1.4, 1.8)$ and $b = 1.9$. The logistic regression model has $m(Y, X, \theta_1) = \theta_{11}Y + \theta_{12}X$, with $\theta_{11} = 0.25$ and $\theta_{12} = 1$. The value of $\theta_0 = -3.70$ was chosen so that the rate of disease in the population for the normal case was $\pi_1 = 0.045$. The case-control study had $n_1 = 500$ cases and $n_0 = 500$ controls. We generated $1,000$ simulated data sets. We made the rare disease assumption, so that $\Omega = \{\kappa, \theta_1 = (\theta_{11}, \theta_{12})\}$, and Ω was estimated by ordinary linear logistic regression of D on (Y, X). For each simulated data set, the standard deviation of the $\widehat{\beta_1}$ was estimated by 500 bootstrap samples. We use the rare disease assumption.

The results are displayed in Table 9, and our easily summarized. First, our method is essentially unbiased. Second, it has actual coverage probabilities very close

to the nominal level. Third, our method is much more efficient than using only the controls, with mean squared error efficiencies ranging from 1.30 to 2.56.

Table 9. Simulation study for the estimation of β_1

	Mean Bias	S.D. S.D.	Mean of estimated S.D.	C.P. of 90% CI	C.P. of 95% CI	MSE Efficiency
Normal(0,1)	0.00	0.10	0.11	0.93	0.97	2.56
Chisquared(7)	-0.01	0.12	0.12	0.90	0.95	1.78
Gamma(0.4,0.91)	-0.02	0.13	0.14	0.91	0.96	1.30
Gamma(0.8,0.91)	-0.01	0.13	0.13	0.91	0.96	1.42
Gamma(1.4,0.91)	-0.01	0.12	0.12	0.89	0.95	1.51
Gamma(1.8,0.91)	-0.01	0.12	0.12	0.91	0.95	1.62

E. An Empirical Example

We used our methodology to investigate two examples described by Chen, et al. (2008). The basic purpose of the analysis is to show that in realistic settings, our methodology leads to much more precise inference that regression using only the controls.

1. Prostate Cancer

The first case-control study is one of prostate cancer. The sample includes 749 prostate cancer cases and 781 controls, selected from the screening arm of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial at the National Cancer Institute (Gohagan, et al., 2000; Moslehi et al., 2006). In the notation of this chapter, D is the prostate cancer case-control status and Y is the level of 1,25-

dihydroxy-vitamin D. Let Z be age level categorized into four groups and hence with three dummy variables. There are three single nucleotide polymorphisms (SNP) of interest, with possible values $0, 1, 2$: we call them SNP-1, SNP-2 and SNP-3. Then X combines Z with each SNP, with a linear regression model, and we are interested in the estimate of the SNP effect.

The results are given in Table 10. We see in this table that none of the coefficients for the SNP are statistically significantly different from zero, which is one of the expectations that Chen, et al. cite. Crucially, the 95% confidence intervals using our method are much shorter than using the control data only, and when translated into mean squared error efficiency, for the three SNP suggest gains in efficiency of 68%, 136% and 125%.

Table 10. Results of the VDR data example

	Robust			Regression			
X	Estimate	Lower CI	Upper CI	Estimate	Lower CI	Upper CI	MSE Efficiency
SNP-1	0.015	-0.165	0.195	-0.029	-0.262	0.204	1.68
SNP-2	0.023	-0.047	0.093	0.039	-0.069	0.146	2.36
SNP-3	0.015	-0.062	0.092	-0.045	-0.161	0.070	2.25

2. Colorectal Adenoma

Chen, et al. also discuss a case-control study of colorectal adenoma, a precursor of colorectal cancer. The study sample includes 628 prevalent advanced adenoma cases and 635 gender-matched controls, also selected from the screening arm of the PLCO Study. Here D is colorectal adenoma status and Y are various measures of smoking

status, namely (a) years since stopping smoking, which we censor at 45; (b) Number of packs smoked per day; and (c) pack years, i.e., packs per day times years smoked. For the latter, as in Chen, at al., we subtracted 0.25 and censored at 100. Here Z is age in years and gender, and X is Z combined with an indicator of the most common diplotype against the rest. We are interested in estimating the effect of the most common diplotype.

The results are given in Table 11. We see a statistically significantly protective of the most common diplotype for the years since stopping smoking. Again,crucially, the 95% confidence intervals using our method are much shorter than using the control data only, and when translated into mean squared error efficiency, for the three SNP suggest gains in efficiency of 95%, 143% and 148%.

Table 11. Results of the NAT2 data example

	Robust			Regression			
Y	Estimate	Lower CI	Upper CI	Estimate	Lower CI	Upper CI	MSE Efficiency
C_S	-3.501	-5.716	-1.318	-3.240	-6.307	-0.173	1.95
P_Y	-0.040	-0.199	0.120	0.210	-0.039	0.458	2.43
P_D	0.063	-0.058	0.184	0.135	-0.047	0.317	2.48

F. Discussion

The study of the relationship among secondary variables in a case-control study is of great practical interest, because large case-control studies now exist and especially include predictors or phenotypes Y and demographic, environmental and genetic factors. The homoscedastic regression model (4.1) is particularly important when

the predictors or phenotypes are continuous random variables, as they are in our two examples.

As we have noted, if one is willing to specify the distribution of the regression errors in the population up to a parameter, then it is possible to estimate the parameter β in model (4.1) in an efficient manner. However, we have shown that misspecification of that parameter model will lead to inconsistent estimation of β: it is possible to create skew distributions for the regression errors that result in bias when normality is assumed.

Our approach is entirely different. While we specify a target regression error distribution, we have shown that the estimation is robust to violation of that target distribution, both theoretically and in a simulation study. In the rare disease case that would be the reason for a case-control study in the first place, an alternative is to simply use only the data for the controls. We have shown in simulations and in our two data examples that such throwing away of 50% of the data leads to a highly non-trivial loss of efficiency compared to our method.

CHAPTER V

MODEL SELECTION IN JOINT MODELLING OF PAIRED FUNCTIONAL DATA

Utilizing penalized B-splines, a new approach proposed by Zhou, et al. (2008) jointly models a pair of sparsely observed functions through functional principal components. In their approach, a stepwise addition and deletion procedure is employed to decide upon the number of principal components (PCs) and crossvalidation is used to estimate penalty parameters. However the choice of the cutoff point in the stepwise addition and deletion procedure is subjective and the crossvalidation computation is very time consuming. In this project we propose to select the number of PCs and estimate the penalty parameters with a modified version of the Akaike information criterion (AIC) and two modified versions of the Bayesian information criterion (BIC). Our methods are computationally fast and straightforward to implement. We illustrate our methods with simulations and the empirical data example used by Zhou, et al. (2008).

A. Introduction

Recently, Zhou et al. (2008) proposed a modeling framework to study the relationship between two sparsely observed functional variables. In this framework, the data for each variable are viewed as smooth curves measured at discrete time-points plus random errors. While the curves for each variable are summarized using a few important principal components, the association of the two longitudinal variables is modelled through the association of the principal component scores. Penalized splines are used to model the mean curves and the principal component curves. The proposed model can be cast into a mixed effects model framework for model fitting, prediction and

inference, and the EM algorithm is used for computation.

There are two aspects of the methodology that need improvement. First, crossvalidation is used to estimate the penalty parameters, but this is computationally expensive. Second, stepwise addition and deletion based upon the changes of the estimated PC score variances to decide on the numbers of PCs. However, this stepwise method is ad hoc since setting up the thresholds for stoping addition and delection is a subjective choice.

The goal of this study is to develop a computationally efficient procedure for selecting the penalty parameters and an automatic procedure for selecting the number of principal components. We propose to apply widely used information criteria such as the Akaike's information criterion (AIC) (Akaike, 1973) and the Bayesian information criterion (BIC) (Schwarz, 1978) in this functional data analysis problem. The information criteria are applicable here because the modeling framework of Zhou et al. (2008) is likelihood based. Use of the information criteria helps us avoid multiple runs of the EM algorithm required by crossvalidation and thus can substantially speed up the penalty parameter selection process. It also makes it automatic to select the number of principal components.

While we focus on the modeling of paired curves in this project, we would like to point out there is a literature on modeling single curves using functional principal components. For example, James et al. (2000) and Rice and Wu (2001) used splines to model the functional PCs. Peng and Paul (2009) used crossvalidation for selecting the number of B-spline xobasis functions and developed a quadratic approximation for fast computation. Yao et al. (2005) proposed to estimate the functional principal components for sparse functional data through the eigen-decomposition of the covariance kernel estimated using two-dimensional smoothing. They also proposed to use AIC to select the number of principal components.

The rest of the chapter is organized as follows. In Section B, we first briefly review the modeling approach of Zhou et al. (2008), then define the degrees of freedom and the information criteria for the smoothing parameter and model selection. In Section C we apply our methods on simulated data sets and one real data example and compare our results with that using crossvalidation.

B. Methodology

1. A Reduced Rank Model for Sparsely Observed Paired Curves

Let $Y_i(t)$ and $Z_i(t)$ denote the two measurements at time t for the i^{th} individual. The joint model of Zhou et al. (2008) has the form

$$\begin{aligned} Y_i(t) &= \mu(t) + \sum_{j=1}^{k_\alpha} f_j(t)\alpha_{ij} + \epsilon_i(t) = \mu(t) + f(t)^{\mathrm{T}}\alpha_i + \epsilon_i(t), \\ Z_i(t) &= \nu(t) + \sum_{j=1}^{k_\beta} g_j(t)\beta_{ij} + \xi_i(t) = \nu(t) + g(t)^{\mathrm{T}}\beta_i + \xi_i(t). \end{aligned} \tag{5.1}$$

where $\mu(t)$ and $\nu(t)$ are the mean curves, $f = (f_1, f_2, .., f_{k_\alpha})^{\mathrm{T}}$ and $g = (g_1, g_2, .., g_{k_\beta})^{\mathrm{T}}$ are vectors of principal component curves, and $\epsilon_i(t)$ and $\xi_i(t)$ are experimental errors. The relationship between $Y_i(t)$ and $Z_i(t)$ is modeled through the correlation between the principal component scores α_i and β_i. The α_i's, β_i's, ϵ_i's and ξ_i's are assumed to have mean zero. The experimental errors $\epsilon_i(t)$ and $\xi_i(t)$ are assumed uncorrelated with constant variance σ_ϵ^2 and σ_ξ^2, respectively. It is also assumed that the α_i's, ϵ_i's and ξ_i's are mutually independent, as are the β_i's, ϵ_i's and ξ_i's. The principal components are subject to the orthogonality constraints $\int f_j f_l = \delta_{jl}$ and $\int g_j g_l = \delta_{jl}$, with δ_{kl} being the Kronecker delta.

For identifiability purpose, the principal component scores α_{ij}, $j = 1, \ldots, k_\alpha$, are assumed independent with strictly decreasing variances, and similarly, the principal

component scores β_{ij}, $j = 1, \cdots, k_\beta$, are also independent with strictly decreasing variances. Denote the diagonal covariance matrices of α_i and β_i by D_α and D_β, respectively. Denote $\mathrm{cov}(\alpha_i, \beta_i) = C$. It is assumed that α_i and β_i are jointly normally distribution so that

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} D_\alpha & C \\ C^{\mathrm{T}} & D_\beta \end{pmatrix} \right\}.$$

The observed data consist of $Y_i(t)$ and $Z_i(t)$ sampled at a finite number of observation times. For each individual i, let $t_{i1}, \ldots, t_{in_i}$ be the different time-points at which measures are available. Based on the observed data, we estimate the unknown functions using penalized splines where μ, ν, f and g are modeled as a member of the same space of spline functions with dimension q. Let $b(t) = \{b_1(t), \ldots, b_q(t)\}^{\mathrm{T}}$ be an orthonomal basis of the spline space where the basis functions satisfy $\int b_j(t) b_l(t)\, dt = \delta_{jl}$. Let θ_μ and θ_ν be q-dimensional vectors of spline coefficients such that

$$\mu(t) = b(t)^{\mathrm{T}} \theta_\mu, \qquad \nu(t) = b(t)^{\mathrm{T}} \theta_\nu. \tag{5.2}$$

Let Θ_f and Θ_g be respectively $q \times k_\alpha$ and $q \times k_\beta$ matrices of spline coefficients such that

$$f(t)^{\mathrm{T}} = b(t)^{\mathrm{T}} \Theta_f, \qquad g(t)^{\mathrm{T}} = b(t)^{\mathrm{T}} \Theta_g. \tag{5.3}$$

Write $Y_i = \{Y_i(t_{i1}), .., Y_i(t_{in_i})\}^{\mathrm{T}}$ and similarly for Z_i. Let $B_i = \{b(t_{i1}), .., b(t_{in_i})\}^{\mathrm{T}}$. The model for the observed data can be written as

$$\begin{aligned} Y_i &= B_i \theta_\mu + B_i \Theta_f \alpha_i + \epsilon_i, \\ Z_i &= B_i \theta_\nu + B_i \Theta_g \beta_i + \xi_i, \\ \beta_i &= \Lambda \alpha_i + \eta_i, \\ \epsilon_i \sim N(0, \sigma_\epsilon^2 I_{n_i}), &\quad \xi_i \sim N(0, \sigma_\xi^2 I_{n_i}), \quad \alpha_i \sim N(0, D_\alpha), \quad \beta_i \sim N(0, D_\beta). \end{aligned} \tag{5.4}$$

For identifiability of the model, we require that $\Theta_f^{\mathrm{T}} \Theta_f = I$ and $\Theta_g^{\mathrm{T}} \Theta_g = I$, and the

first nonzero element of each column of Θ_f and of Θ_g is positive.

Let $L(Y_i, Z_i)$ denote the contribution to the likelihood from subject i. The joint likelihood for the whole dataset is $\prod_{i=1}^{n} L(Y_i, Z_i)$. Define matrix $\mathcal{K} = \int b''(t)\{b''(t)\}^{\mathrm{T}} dt$. The method of penalized likelihood minimizes the criterion

$$\begin{aligned} -2\sum_{i=1}^{n} \log\{L(Y_i, Z_i; \theta_\mu, \theta_\nu, \Theta_f, \Theta_g, D_\alpha, D_\beta, C)\} \\ + \lambda_\mu \theta_\mu^{\mathrm{T}} \mathcal{K} \theta_\mu + \lambda_f \sum_{j=1}^{k_\alpha} \theta_{fj}^{\mathrm{T}} \mathcal{K} \theta_{fj} + \lambda_\nu \theta_\nu^{\mathrm{T}} \mathcal{K} \theta_\nu + \lambda_g \sum_{j=1}^{k_\beta} \theta_{gj}^{\mathrm{T}} \mathcal{K} \theta_{gj}, \end{aligned} \tag{5.5}$$

where θ_{fj} and θ_{gj} are, respectively, the j^{th} columns of Θ_f and Θ_g, and $\lambda_\mu, \lambda_\nu, \lambda_g, \lambda_g$ are four penalty parameters. The EM algorithm is employed to minimize the penalized likelihood criterion (5.5).

2. Selection of the Penalty Parameters

Because crossvalidation is computationally expensive in our context, as discussed in Section 1, we propose to use information criteria such as AIC and BIC to select the penalty parameters. Our proposal is motivated by existing results on asymptotic equivalence of the CV criterion and the information criteria. In particular, Stone (1977) and Shao (1997) showed that under some regularity conditions, the delete-one subject CV criterion is asymptotically equivalent to AIC, while Shao (1997) showed that the BIC and the delete-k subjects CV are asymptotically equivalent. Use of the information criteria can substantially speed up the optimization program for selecting the penalty parameters.

The AIC and BIC criteria are defined as

$$\mathrm{AIC} = 2\,K - 2\log(\widehat{L}_{max}), \qquad \mathrm{BIC} = \log(N) - 2\log(\widehat{L}_{max}),$$

where K is the number of parameters in the model, N is the sample size, and $\widehat{L}_{max}$

is the maximized value of the likelihood function for the estimated model. Adding penalties into the model fitting, it is not immediately clear how to count the number of parameters. We propose to follow Section 3.13 of Ruppert, et al. (2003) and use the effective degrees of freedom of the smoothers as the effective numbers of parameters. For a giving penalty parameter λ, the effective degrees of freedom is defined as

$$\mathrm{df}(\lambda) = \mathrm{trace}\left\{ \left(\sum_{i=1}^{n} B_i^{\mathrm{T}} B_i + \lambda \mathcal{K} \right)^{-1} \sum_{i=1}^{n} B_i^{\mathrm{T}} B_i \right\},$$

Taking into account the effects of four penalty parameters, the total number of effective degrees of freedom for all four smoothing operations is

$$\mathrm{df}(\lambda_\mu, \lambda_\nu, \lambda_f, \lambda_g) = \mathrm{df}(\lambda_\mu) + \mathrm{df}(\lambda_\nu) + k_\alpha \times \mathrm{df}(\lambda_f) + k_\beta \times \mathrm{df}(\lambda_g).$$

Then AIC in our context is defined as

$$\mathrm{AIC}(\lambda_\mu, \lambda_\nu, \lambda_f, \lambda_g) = 2\,\mathrm{df}(\lambda_\mu, \lambda_\nu, \lambda_f, \lambda_g) - 2\log(\widehat{L}_{max}).$$

We consider two versions of BIC where in the first version N is taken to be the number of subjects N_s and in the second version N is taken to be the total number of observations N_o. Specifically,

$$\begin{aligned} \mathrm{BIC}_s(\lambda_\mu, \lambda_\nu, \lambda_f, \lambda_g) &= \log(N_s)\,\mathrm{df}(\lambda_\mu, \lambda_\nu, \lambda_f, \lambda_g) - 2\log(\widehat{L}_{max}), \\ \mathrm{BIC}_o(\lambda_\mu, \lambda_\nu, \lambda_f, \lambda_g) &= \log(N_o)\,\mathrm{df}(\lambda_\mu, \lambda_\nu, \lambda_f, \lambda_g) - 2\log(\widehat{L}_{max}). \end{aligned}$$

We select the penalty parameters by minmizing the above AIC, BIC_s or BIC_o criteria.

Note that to evaluate an information criterion for a fixed set of penalty parameters, the EM procedure only needs to be run once, compared with d times when calculating the d-fold crossvalidation criterion. Therefore, roughly speaking, our information based procedure will take about $1/d$ of time compared with that using

d-fold crossvalidation. Using information criteria also has other advantages. In particular, since the full data set is used in the estimation, the estimates are more stable and in turn speeds up the EM procedure.

3. Selection of the Number of Important Principal Components

The information criteria defined above can also be used to select the number of important principal components. To be specific, we first identify the ranges for k_α and k_β, which usually start from one and the largest values depend on the size of the data. Next for a given pair of k_α and k_β, estimate the smoothing parameters following the procedure described in Section 2. Then record the AIC, BIC_s or BIC_o values at the estimated smoothing parameters, denoted as $AIC(k_\alpha, k_\beta)$, $BIC_s(k_\alpha, k_\beta)$ and $BIC_o(k_\alpha, k_\beta)$. The selected numbers of principle components are $(k_\alpha^{sel}, k_\beta^{sel}) = \text{argmin}\{(k_\alpha, k_\beta), AIC(k_\alpha, k_\beta)\}$, $(k_\alpha^{sel}, k_\beta^{sel}) = \text{argmin}\{(k_\alpha, k_\beta), BIC_s(k_\alpha, k_\beta)\}$, or $(k_\alpha^{sel}, k_\beta^{sel}) = \text{argmin}\{(k_\alpha, k_\beta), BIC_o(k_\alpha, k_\beta)\}$, respectively.

C. Application

1. Selection of the Smoothing Parameters in Simulation Studies

In this section we illustrate the proposed methods described in Section 2 and compare the performance of our methods with that of the crossvalidation using simulated data sets.

In each simulation run, we have $n = 50$ subjects. For simplicity of the presentation, we sample each subject on 11 equally spaced time points from 0 to 1. Let $\mu(t) = \sin(2\pi t)$, $\nu(t) = \sin(4\pi t)$. Let $f_1(t) = 1.4142\cos(2\pi t_{ij})$, $f_2(t) = 11.6508\{(t_{ij} - 1/2)^2 - \cos(2\pi t_{ij})/\pi^2\}$, $g_1(t) = 1.4142\cos(4\pi t_{ij})$ and $g_2(t) = 9.0613\{(x_j - 1/2)^2 - \cos(4\pi x_j)/(4\pi^2)\}$. Note that $f_1(t)$ and $f_2(t)$ are orthonormal, as are $g_1(t)$ and $g_2(t)$.

For subject i, $i = 1, \cdots, 50$, at time $t_{ij} = (j-1)/10$, $j = 1, \cdots, 10$, we simulate Y_{ij} and Z_{ij} from model (5.1) with $\sigma_\epsilon^2 = 1, \sigma_\xi^2 = 1$ under the following three scenarios:

- $k_\alpha = 1$, $k_\beta = 1$, $D_\alpha = 4$, $D_\beta = 2$ and $\text{cov}(\alpha_i, \beta_i) = \sqrt{2}$.
- $k_\alpha = 2$, $k_\beta = 1$, $\text{diag}(D_\alpha) = (4, 2)$, $D_\beta = 2$, $\text{cov}(\alpha_{i1}, \beta_i) = \sqrt{2}$ and $\text{cov}(\alpha_{i1}, \beta_i) = 0$.
- $k_\alpha = 2$, $k_\beta = 2$, $\text{diag}(D_\alpha) = (4, 2)$, $\text{diag}(D_\beta) = (2, 1)$, $\text{cov}(\alpha_{i1}, \beta_{i1}) = \sqrt{2}$ and $\text{cov}(\alpha_{i2}, \beta_{i2}) = 0.5$.

Under each scenario, 1,000 data sets were generated. As our focus here is on the smoothing parameter selection, in our simulation studies the data were all fit using the correct numbers of principal components. Ten-fold crossvalidation was used in all comparisons.

The simulated data were fit following the procedure described in Section B. We search over a four dimensional space for optimal smoothing parameters that minimize crossvalidation, AIC, BIC_s and BIC_o. Optimization was realized in R using L-BFGS-B by Byrd et. al. (1995). This optimizing method uses a limited-memory modification of the BFGS quasi-Newton method and requires preset lower and/or upper bounds, which were set at 0.1 and 10^5 respectively.

Table 12. Average ratios of the computing time using AIC, BIC_s and BIC_o to that using 10-fold CV

	CV/AIC	CV/BIC_s	CV/BIC_o
$k_\alpha = 1, k_\beta = 1$	12.0	8.7	6.9
$k_\alpha = 2, k_\beta = 1$	13.8	8.4	6.6
$k_\alpha = 2, k_\beta = 2$	11.7	10.8	8.7

Table 12 gives the ratios of the computing time between using 10-fold crossvalidation and the model selection criteria AIC, BIC_s and BIC_o. Under the three simulation scenario, AIC is between 11.7 and 13.8 times faster than the 10-fold crossvalidation; BIC_n is between 8.4 and 10.8 faster and BIC_o is between 6.6 and 8.7 times faster. While the improvement in computing time is consistent for AIC it is most pronounced in the scenario of $k_\alpha = 2$, $k_\beta = 2$.

Table 13. Average smoothing parameters and corresponding degrees of freedom selected using 10-fold crossvalidation, AIC, BIC_s and BIC_o

		λ_μ	λ_ν	λ_f	λ_g	df_{λ_μ}	df_{λ_ν}	df_{λ_f}	df_{λ_g}
$k_\alpha = 1$	CV	40.4	10.3	43.2	5.2	8.3	9.3	8.3	10.1
$k_\beta = 1$	AIC	64.7	8.0	86.0	6.5	7.2	9.3	6.6	9.6
	BIC_s	139.2	15.7	181.0	13.8	5.9	8.4	5.6	8.6
	BIC_o	197.3	23.4	276.5	22.0	5.5	7.9	5.3	8.0
$k_\alpha = 2$	CV	43.9	8.9	54.6	6.6	8.2	9.5	7.5	10.0
$k_\beta = 1$	AIC	76.4	6.6	65.3	7.2	7.0	9.7	6.8	9.6
	BIC_s	167.0	13.2	139.0	14.6	5.7	8.7	5.8	8.6
	BIC_o	243.7	20.0	209.9	22.6	5.3	8.1	5.4	8.0
$k_\alpha = 2$	CV	39.1	9.7	52.0	14.6	8.4	9.6	7.7	9.0
$k_\beta = 2$	AIC	74.8	6.9	65.9	14.3	7.1	9.8	6.9	8.9
	BIC_s	158.2	14.4	150.7	34.4	5.8	8.7	5.8	7.6
	BIC_o	224.3	22.8	244.9	56.0	5.4	8.1	5.3	6.9

Table 13 shows the smoothing parameters chosen by different criteria and the corresponding degrees of freedom defined in section 2. In general, BIC_o selected the largest smoothing parameters while 10-fold crossvalidation selected the smallest.

In Table 14 we present the integrated mean squared errors of the mean functions

and of the principal component functions, separably, as well as the average mean squared errors of the variances of the principal component scores and of the variances of measurement errors. With a few exceptions, our methods improve over that of the 10-fold crossvalidation. In most cases, BIC_s and BIC_o have smaller MSEs than AIC in the estimation of the mean functions μ and ν, as well as σ^2; AIC performs better in estimating the principle component functions and the variances of the principle component scores. All four methods have similar MSEs in estimating the covariance matrix R of the principle component scores.

Table 14. Integrated mean squared errors

		μ	ν	f	g	D_α	D_β	R	σ_ϵ^2	σ_ξ^2
$k_\alpha = 1$	CV	0.08	0.05	0.01	0.02	0.06	0.03	0.25	0.0251	0.0046
$k_\beta = 1$	AIC	0.08	0.05	0.01	0.02	0.04	0.02	0.25	0.0015	0.0018
	BIC_s	0.06	0.04	0.0064	0.02	0.04	0.04	0.25	0.0009	0.0012
	BIC_o	0.05	0.03	0.0070	0.03	0.05	0.06	0.26	0.0007	0.0007
$k_\alpha = 2$	CV	0.12	0.05	0.05	0.03	0.03	0.03	0.26	0.0035	0.0070
$k_\beta = 1$	AIC	0.11	0.05	0.05	0.03	0.03	0.03	0.26	0.0030	0.0019
	BIC_s	0.08	0.04	0.05	0.02	0.03	0.04	0.27	0.0016	0.0012
	BIC_o	0.07	0.04	0.05	0.02	0.04	0.06	0.26	0.0012	0.0007
$k_\alpha = 2$	CV	0.12	0.10	0.06	0.12	0.03	0.01	0.16	0.0034	0.0038
$k_\beta = 2$	AIC	0.11	0.09	0.05	0.12	0.03	0.01	0.16	0.0028	0.0042
	BIC_s	0.08	0.09	0.06	0.14	0.03	0.02	0.16	0.0014	0.0021
	BIC_o	0.08	0.09	0.06	0.19	0.03	0.03	0.16	0.0011	0.0012

To summarize, the three proposed model selection criteria show great improvement over using crossvalidation in model selection – all three methods dramatically

reduce the computing time while improving the estimation precision. Among the three criteria we proposed, AIC is fastest and BIC_o is slowest; in terms of estimating the population mean curves and the measurement error variances, our results show that BIC_o gives the smallest MSEs while AIC gives the largest MSEs; for the estimation of the principle component functions and scores variances, AIC generally gives the smallest MSEs and BIC_o gives the largest MSEs.

2. Selection of the Numbers of the Principle Components in Simulation Studies

We illustrate the selection of the numbers of the principle components with the proposed criteria using simulation setup 3 in Section 1, where there are two principle components for each variable. We generated 100 simulated data sets, applied the model selection procedure using 10-fold crossvalidation suggested by Zhou, et. al. (2008) as well as the procedure described in Section 3. In each simulation run, we set the selection ranges of the numbers of the principle components from 1 to 3. Therefore, there are 9 total possibilities for the pair (k_α, k_β). Among the 100 simulations, our proposed criteria picked the correct numbers of the PCs 100% of time while the crossvalidation method was correct in 63% of time and selected $(2, 1)$, $(2, 3)$, $(3, 1)$, $(3, 2)$ and $(3, 3)$ the remaining 37% of the time. Our simulation results suggest the use of one of the proposed criteria, gives not only faster, but also more accurate results.

3. Model Selection in the Rreal Data Example

Here we illustrate the proposed method using a data set from an AIDS clinical trial ACTG 315 (Lederman et al. 1998) conducted by the AIDS Clinical Trials Group; the same data was analyzed by Zhou, et al. (2008). In this clinical trial, forty-six HIV-1 infected patients were treated with potent antiviral drugs (ritonavir, 3TC and AZT). After initiation of the treatment on day 0, patients were followed for up to

10 visits. Scheduled visit times common to all patients are 7, 14, 21, 28, 35, 42, 56, 70, 84 and 168 days. Since the patients did not follow exactly the scheduled times and/or missed some visits, the actual visit times are irregularly spaced and different for different patients. The visit time varies from day 0 to day 196. In the notation of the joint model for paired functional data in Section 1, denote by Y the CD4+ cell counts divided by 100 and by Z the base-10 logarithm of plasma HIV RNA copies. To model the curves on the time interval $[0, 196]$, cubic B-splines with 10 interior knots were placed on scheduled visit times.

To fit this data, we need to estimate both the numbers of PCs and the smoothing parameters. Due to the limited sample size, we restrict our choices of the numbers of principle components (k_α, k_β) to $(1, 1)$, $(1, 2)$, $(2, 1)$, $(2, 2)$ and $(2, 3)$. We had difficulty in computation with more than two PCs for CD4+ cell counts or more than three PCs for the log viral load, largely because the variances of the PC scores were nearly zero. Zhou, et al (2008) used 10-fold crossvalidation and chose one PC for CD4+ cell counts and two PCs for the log viral load. Applying the procedure described in section 3, all three criteria chose two principle components for both variables. As shown in Table 15, all three criteria scores have same patterns: the scores drop from $(1, 1)$ to $(2, 1)$, $(1, 2)$ and reach the lowest level at $(2, 2)$. The biggest drop is from $(2, 1)$ to $(1, 2)$ and the scores increase a little from $(2, 2)$ to $(2, 3)$. Based on the criteria scores, one can either choose $(2, 1)$ for the parsimony, as did in Zhou, et al. (2008), or choose $(2, 2)$ which we believe is the best model for this data. Compare our methods with that of Zhou et al. (2008), in both models, the fitted values are similar; see Figure 2 for the mean curves and PC curves using $k_\alpha = 1$ and $k_\beta = 2$. On the other hand, our methods are much faster: when one principal component for CD4+ cell counts and two principal components for log viral load were used, AIC, BIC_s and BIC_o were about 19 times faster; when two principal components for both CD4+ cell counts and

log viral load were used, AIC, BIC_s and BIC_o were 10 - 12 times faster.

Table 15. AIC, BIC_s and BIC_o scores on AIDS data

(k_α, k_β)	AIC	BIC_s	BIC_o
$(1,1)$	-0.88	33.22	62.21
$(2,1)$	-11.82	29.13	62.26
$(1,2)$	-57.40	-28.79	5.02
$(2,2)$	-76.28	-45.34	-7.23
$(2,3)$	-75.11	-38.34	3.06

D. Discussion

We have shown how to select the smoothing parameters in joint model of paired sparse functional data using novel versions of AIC and BIC. The methodology was described for penalized spline mixed effects model and justified in the important paired sparse functional data case. Numerically, we have found that AIC and BIC are very close to being the same as 10-fold cross-validation in terms of model fits, while being much faster computationally.

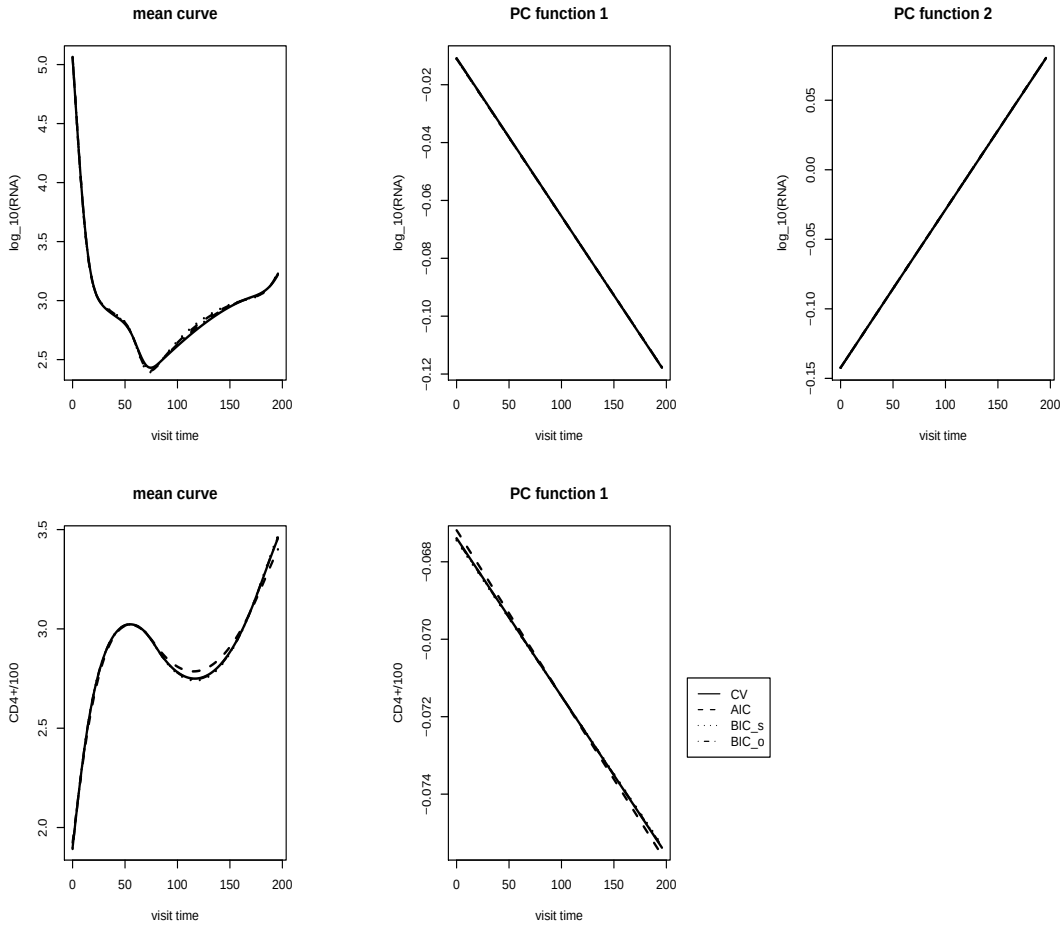

Figure 2. Fitted mean curves and principal component curves using 10-fold crossvalidation, AIC, BIC_s and BIC_o for AIDS data.

CHAPTER VI

PERMUTATION TEST FOR MICROARRAYS: COLONIC TUMORIGENESIS PREDICTION IN THE EARLY STAGES OF DEVELOPMENT

In this study, we introduce a practical permutation test that uses cross-validated genetic predictors to determine if the list of genes in question has "good" prediction ability. We call it the cross-validated permutation test. It avoids overfitting by using cross-validation to derive the genetic predictor and determines if the count of genes that give "good" prediction could have been obtained by chance. This test is then used to explore gene expression of colonic tissue and exfoliated colonocytes in the fecal stream to discover similarities between the two, done at each of the three stages of colonic tumorigenesis.

A. Introduction

1. Motivation for Genetic Influence on Colon Cancer

According to statistics compiled by the American Cancer Society in 2008 and 2009, colon cancer is the third deadliest form of cancer in the U.S. among both men and women. The prediction and eventual treatment of colon cancer at the early stages of development is crucial for the population as a whole.

Moreover, the study of the development of cancer is not only of interest for the benefit of public health as a whole, it is also an important topic for the successful development of long term space missions, like traveling to Mars. In such long missions humans would be exposed to prolonged doses of radiation which can potentially lead to all types of cancerous tumors in a very short period of time. The attenuation of these effects is therefore of great interest. To address these concerns our study

design not only looks at cancerous tumor development as a result of exposure to harmful chemical agents, it also considers the effects of radiation in tumorigenesis, and specifically as it occurs in the colon.

In this work we study the prediction ability of genes to foretell colonic tumorigenesis; however, to attack the problem of colon tumor development one must first understand how the development occurs. Below we provide a brief description of the stages of colonic tumor growth. The development of tumors in the colon happens in 3 main stages: initiation, aberrant crypt foci (ACF) and finally tumor stage. At the initiation stage the primary insult to the colon results in DNA damages. Constant exposure to reactive oxygen species results in the damage to DNA, leading to mutations. At the ACF or aberrant crypt foci stage, rogue colonic crypts begin to form clusters or high multiplicity ACFs (HMACF) that signal the potential for tumor development at these sites. Finally, the tumor stage is when actual tumors develop in the colonic walls. As with HMACFs, tumors can be measured and counted to serve as a phenotype that we can eventually predict.

It is well known that certain foods affect the development of cancer in the body via anti-inflammatory mediators. More specifically n-3 polyunsaturated fatty acids, found in fish oil, modulate the inflammatory process via genes such as interleukin-1 β, see Kim et al. (2009). Hong et al. (2005) show that fish oil is protective against oxidative DNA damage when compared against corn oil at the initiation stage. Moreover, Hong et al. (1997) also show that fish oil is protective at each of the three stages. Hence diet must be taken into account when studying the effects certain genes have on the development of colonic cancer. Our study incorporates diet by using a rat model where two different types of fiber (pectin and cellulose) and two different types of oil (corn oil and fish oil) are employed. Tumor development is induced by introducing both a chemical carcinogen, azoxymethane (AOM), as well

as an environmental carcinogen, radiation.

In a compilation of all relevant studies that look at the contribution of genes on gastric cancer in humans up to 2001, Gonzalez et al. (2002) found that interleukin-1 β and NAT 1 variants are associated with increased gastric cancer. Hence the expectation that certain genes may be able to predict colonic cancer is not without merit.

Most genetic studies of colon cancer use gene expression from mucosal tissue extracted directly from the colon. This procedure is invasive and is not easily applicable to human subjects. New ways to isolate mRNA from exfoliated colonocytes in the fecal stream enable us to study the development of colonic cancer without invasive procedures in both rats and humans, see Davidson et al. (1995) and Davidson et al. (2003).

Our goal is to compare the prediction ability of genes to foretell colonic tumorigenesis at the ACF and tumor stages in both mucosal and fecal colonocytes via microarray studies, albeit an indirect comparison but something that has never been done until now.

This chapter is organized as follows: Section B gives a short description of the data used to carry out the analysis, Section C introduces our permutation test and gives results that provide specific gene lists that are able to predict tumor and non tumor outcomes. Concluding remarks are in Section D.

B. Data

The data that will be used throughout this work are gene expression data generated by the CodeLink System from GE. This platform is a gold standard for rat genome studies and since our data come from rat models it is of course the method of choice.

The normalization scheme used to make gene expression comparable between independent arrays is performed by the GE system upon completion of the work, and for most cases this scheme performs well. In our study, we are dealing with gene expression derived from RNA found in mucosal tissue as well as in the fecal stream. The standard normalization scheme works well on arrays from mucosal tissue; however, because of the normal degradation of RNA in the colon gene expression from fecal matter is quite low and hence require a special normalization method. Liu et al. (2005) studied this problem and developed a two-stage normalization scheme specially targeted to handle the problem of partially degraded RNA and it is one of the methods we will use to normalize gene expression from the fecal stream.

In our aim to predict colonic tumorigenesis we will focus our efforts at both the aberrant crypt foci or (ACF) and tumor stages of development. We will do our prediction by using gene expression from RNA derived from both mucosal and fecal samples. Mucosally derived microarrays are normalized via the standard method while microarrays obtained from RNA in the fecal stream are normalized by both the standard and two-stage methods. We will split our analysis into two parts, one where all the genes in the microarray are considered as potential predictors and the other where only a select number of genes, which will refer to as the rat colonic biomaker list, is used. The genes in the rat colonic biomarker list are chosen because they are known to be involved in the development of colon cancer.

The reason we are concentrating our efforts at ACF and tumor stages is simply because the phenotypes are either inherently or can easily be dichotomized, and hence prediction via well known classification methods can be readily accomplished. At the tumor stage our outcomes are binary where a success indicates the existence of at least one tumor in the colon and a failure means that no tumors where found in the rat. At the ACF stage our binary outcomes where generated by dichotomizing the

count of high multiplicity aberrant crypt foci (HMACF). If the count of HMACF was greater than 4 we call that a success, which basically suggests a strong indication that a tumor will appear. If the count was less than or equal to 4 then we deemed that a failure. We label our successes at the ACF stage as tumor rats and our failures as healthy rats.

At the ACF stage, there are fecal arrays for 15 healthy rats and 18 tumor rats in the fecal arrays; while there are mucosal arrays, for 20 healthy rats and 13 tumor rats. At the tumor stage, there are 39 healthy rats and 10 tumor rats with fecal array data; while for the mucosal arrays, there are 61 healthy rats and 14 tumor rats.

Not only are we going to use gene expression to predict tumor outcomes, we will also use additional factors of interest that influence the incidence of tumor development. Recall that diet plays an important role in the development of colon cancer. Both dietary fiber and lipid sources were manipulated in this study. The fiber sources in the study were pectin and cellulose while the lipid sources were fish oil and corn oil. Additional interest was placed on the mechanism by which carcinogenesis was induced. In this study both a chemical carcinogen, azoxymethane or AOM, was employed as well as an environmental source: radiation. Overall the additional factors to consider are: Diet and Radiation. Diet includes the four possible treatment combinations: fish oil with pectin (fishpect), fish oil with cellulose (fishcell), corn oil with pectin (cornpect) and corn oil with cellulose (corncell). Radiation includes two treatments: radiated or not.

In the next section we will introduce our permutation test and show how genes are able to predict tumor outcomes and that their prediction ability surpasses that of other covariates like diet and radiation.

C. Analysis

In this section we are going to perform cross-validated permutation tests (CPT), cross-validated within treatment permutation tests (CWPT) and diagonal linear discriminant analysis (DLDA). We propose CPT to assess the prediction ability of the genetic predictors in our data, and illustrate this method using the data from Khan et al (2001). CWPT is proposed to assess the improvement that genetic predictors have over clinical predictors in our data, such as diet and radiation. We illustrate the use of the CWPT by applying it to the breast cancer data that is described in Hofling and Tibshirani (2008). In both the cross-validated permutation test and the cross-validated within treatment permutation test we classify the tumor and healthy subjects using linear discriminant analysis (LDA) using one gene at a time. We use DLDA to perform similar classification; however, we use DLDA on a larger set of genes to produce a classifier rule for tumor and healthy subjects. Missing data were imputed by nearest neighbor averaging (function pamr.knnimpute() in R).

1. Cross-validated Permutation Test (CPT)

The premise of our work revolves around the assumption that genes have the ability to predict tumor outcome in colonic tissue whether it be at the aberrant crypt foci stage or the final tumor stage. However it is difficult to discern how many genes play a vital role when discriminating between tumor or healthy outcomes. Recent work by Hua et al. (2009) shows that using more than 5 genes to build classifiers is not only computationally daunting, it does not improve error rates. This finding justifies a conservative approach and hence we will use at most one gene to build a classifier and then use it to predict tumor outcomes. We will do this at the ACF stage and then at the tumor stage.

Our classifier scheme is Fisher's LDA which assumes that each class in the data possesses a multivariate normal distribution with a mean μ_k and a covariance matrix Σ_k, where k is the class iterator. In our application our classes are binary and therefore k will be either 1 = tumor, or 2 = no tumor. A main assumption of the LDA is that the covariance is the same across class and therefore $\Sigma_k = \Sigma$ for all k, where Σ is some common covariance matrix. When applying the LDA one does not know the mean and variance parameters ahead of time so those must be estimated. We use leave-one-out cross validation to train, test and produce error rates from our classifier.

Using LDA, we produce a classifier that tries to predict tumor and non-tumor outcomes with one gene as the predictor. From this predictor we obtain classification errors in the form of false negative (FN) and false positive (FP) rates and do this for each of the 30,000+ genes found in the CodeLink microarray. We obtain these FN and FP values for each gene in arrays collected from the two samples: mucosal and fecal. For mucosal array normalization is standard; however, since normalization can be tricky for the arrays produced from fecal matter we produced classifiers for gene expressions that were normalized by two-stage method and also by the standard normalization generated by the GE system. All these steps were reproduced at both the ACF and tumor stages of colonic tumorigenesis.

The expectation is that after all this we will find genes that predict tumor outcomes. Low false positive and false negative rates are good indicators that a gene is able to do this. But how do we know that set of genes having low error rates, as defined by the procedure outlined above, did not occur by chance? We will test this hypothesis by doing a permutation test.

In this analysis we count the number of genes at both ACF and tumor stage whose false negative and false positive rates fall below an arbitrarily defined threshold b. In our work, we took $b = 0.2$. After obtaining this count of genes which we will denote

as N, we perform a permutation test to see if the count of genes, N, is significant. In this permutation test, we permutate the class labels 200 times. For each permutation, we then count the number of genes whose false negative and false positive rates fall below b=0.2 in each time. Let $B = 200$, for $b = 1, ..., B$, let N_b be the count of genes in the b^{th} permutation. Then our $p - value = \sum_{b=1}^{B}(N_b >= N)/B$. We refer to this as the cross-validated permutation test or (CPT). The cross-validation occurs when we are building the classifier using the LDA.

We applied our cross-validated permutation test to the complete set of genes found in the CodeLink arrays and on the shorter list of colonic biomarker genes. We found that among all the genes in the complete list, 5 genes are statistically significant with $p - value = 0.015$. These genes were found in the mucosal array at the ACF stage and the gene IDs for these 5 genes are: GE1155319, GE1170229, GE1266696, GE16725 and GE22209. We then performed CPT using one gene at a time plus diet and radiation covariates as predictors and found one gene to be statistically significant with $p - value = 0.04$. This gene was found in the two-stage normalized fecal array at the tumor stage; the gene ID for this gene is GE21877, and it is also in the rat colonic biomarker gene list.

We illustrate the ability of our method to detect genes of high predictive value by applying it to the data from gene expression microarray experiments on small, round blue-cell tumors (SRCTs), described in Khan et al (2001). There are four different SRBCT tumor types: neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL), and the Ewing family of tumors (EWS). We found 87 genes are statistically significant with $p-value = 0$. The counts of genes from 200 permutations have mean 0 and standard deviation 0.

Using linear discriminant analysis with one gene as a predictor we were able to find a short list of genes that can distinguish between individuals that have large

numbers of ACF and those that do not. Using our cross-validated permutation test we verified that the prediction ability of this set of genes did not occur at random, and hence further research should aim towards discovering the function and role these genes play in the development of colon cancer at the ACF stage. The inclusion of radiation and diet as predictors in the CPT further revealed that interlukin 1β, or GE21877, can distinguish between individuals that develop tumors and those that do not. Further exploration revealed that elevated levels of this gene's expression correlates with the incidence of tumor, indicating that a high expression of this gene could be a red flag if found in individuals at risk of colon cancer. However, only future research can determine the usefulness of this gene as a prediction.

An interesting question arises from this analysis. How do the additional covariates influence a genes ability to predict tumor outcomes? To answer this question we introduce a new permutation test which also uses cross validation but now permutes outcomes within a treatment class to asses the treatment's influence on the classifier.

2. Cross-validated within Treatment Permutation Test (CWPT)

In this section we propose a cross-validated within treatment permutation test (CWPT) to compare the prediction ability of genes to that of covariates like diet and radiation. In this scenario we are testing the significance of the count of genes whose joint false negative and false positive rates fall below that of the classifier formed by only employing treatments as predictors. Again we are using LDA as the mode of classification and the treatments used are covariates diet and radiation.

Define the false negative and false positive rates of a classifier that only includes covariates as the predictors as FN_c and FP_c. These rates are constant for all the genes. Let the false negative and false positive rates of classifiers that use both gene expression and covariates as predictors be defined as FN_{cgi} and FP_{cgi}, where i is the

gene iterator. Of course, FN_{cgi} and FP_{cgi} should not be greater than FN_c and FP_c for gene i. Define $T_i = I(FN_{cgi} < (1-a)FN_c) \times I(FP_{cgi} < (1-a)FP_c)$, where $a = 0.2, 0.3, 0.4, 0.5, 0.6, 0.7$. Here a indicates the improvement of the ith gene over the covariate predictors. The larger the value of a the more improvement the gene has over the covariates. If we define $\mathcal{S} = \sum_{i=1}^{N} T_i$, where N is the number of genes in the entire gene list, then a reasonable hypothesis is:

H_0: $\mathcal{S}$ has the same distribution as if the class labels were assigned at random within each treatment;

H_A: H_0 false.

We propose a p-value from a permutation test to obtain the significance of improvement of gene predictors over the diet and radiation covariates. For $b = 1, \ldots, B$, we resample the subjects with replacement within treatment, and define the false negative and false positive rates with covariates only as FN_c^b and FP_c^b; then we permutate the class labels within covariates, and define the false negative and false positive rates with both diet and radiation covariates and gene expression for gene i as FN_{cgi}^b and FP_{cgi}^b. Let $T_{bi} = I(FN_{cgi}^b < (1-a)FN_c^b) \times I(FP_{cgi}^b < (1-a)FP_c^b)$ and $S_b = \sum_{i=1}^{N} T_{bi}$, then our p-value is $p = B^{-1} \sum_{b=1}^{B} I(\mathcal{S}_b > \mathcal{S})$. We conclude that there are $\mathcal{S}$ genes that improve classification significantly if the p-value for that $\mathcal{S}$ is less than 0.10. We will refer to this as the cross-validated within treatment permutation test or CWPT. We were liberal in our significance threshold of 0.10 since 0.05 was too small to detect anything but one set of genes.

We performed the CWPT on the full list of genes from the two-stage normalized fecal arrays, GE normalized fecal arrays and mucosal arrays at ACF and tumor stages. Table 16 shows the number of genes that have false negative and false positive rates that fall below that of the classifier using covariate predictors only and p-values of the CWPT. In the two-stage normalized fecal array at ACF stage, there are 6980 genes

that improve classification significantly when $a = 0.2$, with p-value $p = 0.025$. In the GE normalized fecal array at the tumor stage, 415 and 50 genes improve classification significantly when $a = 0.4$ and 0.5, with p-value $p = 0.060$ and 0.080 respectively.

Table 16. Permutation test within treatment for each gene in the entire gene list of ACF and tumor stages.

		ACF		tumor	
	a	n	p	n	p
Fecal Arrays	0.2	6980(594)	0.025	1498(519)	0.484
Two-stage Normalized	0.3	393(176)	0.215	512(200)	0.469
	0.4	166(93)	0.175	483(195)	0.328
	0.5	15(12)	0.255	40(13)	0.359
	0.6	2(2)	0.215	1(0)	0.422
	0.7	1(0)	0.115	1(0)	0.125
Fecal Arrays	0.2	1652(1644)	0.260	2350(2296)	0.160
GE Normalized	0.3	385(385)	0.345	446(429)	0.245
	0.4	143(143)	0.255	415(398)	0.060
	0.5	19(19)	0.275	50(46)	0.080
	0.6	1(0)	0.370	0(0)	1
	0.7	0(0)	1	0(0)	1
Mucosal Arrays	0.2	689(687)	0.494	2(2)	0.760
	0.3	205(204)	0.551	0(0)	1
	0.4	44(44)	0.540	0(0)	1
	0.5	44(44)	0.381	0(0)	1
	0.6	2(2)	0.477	0(0)	1
	0.7	0(0)	1	0(0)	1

Our cross-validated within treatment permutation test shows that there are great number of genes that improve the classification of tumor phenotypes at both the tumor and ACF stages of colonic tumorigenesis. Expression of 6980 genes at the

ACF stage and that of 400+ genes at the tumor stage derived from fecal material shows the most promise in having the ability to predict tumor outcomes above that of treatment covariates: diet and radiation. This indicates that RNA harvested from fecal material has the potential to predict tumor outcomes non-invasively and before the actual development of the tumor.

The are a couple of drawbacks to this procedure. One is that now there are a very large number of genes that we have to sort through to find only a handful that are of true interest, and the other is that we used each gene as an independent predictor. This is a reasonable assumption; however, using more than one gene to do prediction might prove more powerful, if the selection process is appropriate. Inspired by these observations we introduce a permutation test that uses more than one gene to predict tumor outcomes.

3. Diagonal Linear Discriminant Analysis (DLDA)

As mentioned above, we now focus our methods so that multiple genes are used as the predictors in the classifier machinery. Up to this point we have used LDA to the classification, but because of the improved performance we will switch to the DLDA or diagonal linear discriminant analysis, see Dudoit et al. (2002). As the name suggests now the assumption on the covariance matrix of each class is equal but restricted to a diagonal. The switch from LDA to DLDA is not major since up to this point we have used only one gene as a predictor in the LDA, so in essence we have been doing DLDA since the covariance matrix is of dimension 1×1 and therefore diagonal.

As before we want to quantify the ability of genes to predict tumor outcomes above and beyond the prediction that additional covariates provide. Our goal therefore, is not much different than what we presented in the previous section; however, now we use DLDA and we use more than one gene as predictors.

We performed a preliminary selection of genes based on the ratio of their between-group to within-group sums of squares since lots of genes exhibit near-constant expression levels across samples. For gene j, the ratio is

$$BW(j) = \frac{\sum_i \sum_k I(y_i = k)(\bar{x}_{kj} - \bar{x}_{.j})^2}{\sum_i \sum_k I(y_i = k)(x_{ij} - \bar{x}_{kj})^2}$$

In the analysis, 15 and 25 genes with the largest BW ratio are selected at the ACF stage and tumor stage, respectively. We selected the top 15 and 25 at the ACF and tumor stages because they are approximately half the sample size found at each stage. We performed DLDA on these selected genes, using leave-one-out cross validation to build the classifier. We obtained the false negative and false positive rates using the short list of genes; we obtained FN and FP rates using both the short gene list as well as covariates and then performed our cross-validated within treatment permutation test to assess the improvement of classification that genes demonstrate over covariates.

Define the false negative rate and false positive rate of the classifier that only used the covariates as its predictors as FN_c and FP_c; and let the false negative rate and false positive rate with both treatments and gene expression for all selected genes be FN_{cg} and FP_{cg}. FN_{cg} and FP_{cg} should not be greater than FN_c and FP_c. A reasonable hypothesis is:

$$H_0 : FN_c = FN_{cg} \text{ and } FP_c = FP_{cg};$$
$$H_A : FN_{cg} < (1-a)FN_c \text{ and } FP_{cg} < (1-a)FP_c,$$

where $a = 0, 0.1, 0.2$. These hypotheses are reasonable because if the error rates of the classifier which uses genes as predictors equal that of the classifier that did not use genes as predictors, this means that the genes did not improve the classifier's ability to predict tumor outcomes. This is exactly what we want to find from these

data.

Define $T_1 = (1-a)FN_c - FN_{cg}$, $T_2 = (1-a)FP_c - FP_{cg}$, and let $G = (T_1^2 + T_2^2)I(T_1 > 0, T_2 > 0)$ be our test statistics. Then for $b = 1, \ldots, B$, we resample the animals with replacement within treatment, and define the false negative rate and false positive rate with treatments only as FN_c^b and FP_c^b; then we permutate the class labels within treatment, and define the false negative rate and false positive rate with both treatments and gene expression as FN_{cg}^b and FP_{cg}^b. Let $T_{b1} = (1-a)FN_c^b - FN_{cg}^b$, $T_{b2} = (1-a)FP_c^b - FP_{cg}^b$, and $G_b = (T_{1b}^2 + T_{2b}^2)I(T_{1b} > 0, T_{2b} > 0)$, then our p-value is $p = B^{-1}\sum_{b=1}^{B} I(G_b >= G)$.

Table 17 shows the p-values for the entire gene list of normalized fecal arrays at the ACF and tumor stages. At the tumor stage, the 25 selected genes in the GE normalized fecal arrays improve the classification significantly over covariates when $a = 0$ and 0.1. We did not find the genes in two-stage normalized fecal arrays and mucosal arrays improve the classification significantly over covariates.

4. CWPT and DLDA on Breast Cancer Data

To show that our methods are comparable to others we apply our cross-validated within treatment permutation tests to a published data set. We apply our CWPT using one gene as the predictor and using a pre-selected set of genes. We will refer to the one gene approach as the CWPT and the multigene approach simply as the DLDA because it was used to do the classification when multiple genes we included.

We apply the CWPT and DLDA to the breast cancer data that is described in Hofling and Tibshirani (2008) to illustrate our methods. This data consists of 4918 genes and 78 patients, in which 34 patients had poor prognosis and 44 patients had a good prognosis. There are 6 covariates in this data set:

Table 17. DLDA in the entire gene list of ACF and tumor stages.

		ACF	tumor
	a	p	p
Fecal Arrays	0	0.065	1.000
Two-stage Normalized	0.1	1.000	1.000
	0.2	1.000	1.000
Fecal Arrays	0	1.000	0.000
GE Normalized	0.1	1.000	0.000
	0.2	1.000	1.000
Mucosal Arrays	0	1.000	1.000
	0.1	1.000	1.000
	0.2	1.000	1.000

- Tumor grade (good: 1,2; poor: 3)
- Estrogen receptor (ER) status (good: <= 10; poor: > 10)
- Progesterone receptor (PR) status (good: <= 10; poor: > 10)
- Tumor size (mm) (good: <= 20; poor: > 20)
- Patient age (yrs) (good: <= 40; poor: > 40)
- Angioinvasion (good: 0; poor: 1)

Since our analysis considers the two-way interactions of the covariates, we selected the two covariates (ER and grade) that have the most significant main effects in ANOVA.

After applying the CWPT we found 0 genes that have false negative rate and false positive rate that fall below those when we use treatment covariates only. The

false negative rate and false positive rate of the classifier with treatment covariates as it only predictors are $FN_c = 0.500$, $FP_c = 0.147$.

We used the DLDA on these data, using 25 pre-selected genes, and found that this list of genes did not improve the classification significantly, which is consistent with the results from Hofling and Tibshirani (2008).

Our permutation tests are then shown to be comparable to pre-validation as described by Hofling and Tibshirani (2008).

D. Conclusion

In this study, we developed cross-validated permutation test (CPT), cross-validated within treatment permutation test (CWPT) and diagonal linear discriminant analysis (DLDA) to assess the prediction ability of the genetic predictors in the gene expression data that are generated by CodeLink System from GE. From CPT, we found 5 genes in the mucosal array of ACF stage were statistically significant in terms of predicting tumor outcome in colonic tissue; and 1 gene was significant in the two-stage normalized array of tumor stage. By CWPT, we found hundreds to thousands of genes improved classification significantly over covariates at both ACF and tumor stages. From DLDA, the 25 selected genes in the GE normalized fecal arrays improve the classification significantly over covariates at the tumor stage. The application on SRCT data and breast cancer data gives similar results to those in Khan et al. (2001) and Hofling and Tibshirani (2008).

CHAPTER VII

CONCLUSION

In the first project we have shown how to test for a constant environmental effect in general semiparametric regression models with interactions. The methodology was described for kernel regression methods and justified in the important logistic regression case. Numerically, we have found that regression spline approaches are very close to being the same as kernel methods and much faster to compute, although their theory remains an open question in this context.

The second project analyzed the relationship among secondary variables in a case-control study, which is of great practical interest, because large case-control studies now exist and especially include predictors or phenotypes Y and demographic, environmental and genetic factors. As we have noted, the semiparametric efficient approaches can be used to construct semiparametric score tests, but they suffer from a lack of robustness to the assumed model for Y given X: it is possible to create skew distributions for the regression errors that result in bias when normality is assumed. Our approach is entirely different. The score statistic has mean zero under the null hypothesis even the conjectured model is not correct, both theoretically and in a simulation study. An alternative is to simply use only the data for the controls. We have shown in simulations and in our two data examples that it suffer from lower power compared to our method since it used only half of the data.

The third project took up the issue of estimation of a regression function when Y given X follows a homoscedastic regression model in the secondary analysis of case-control data. The homoscedastic regression model is particularly important when the predictors or phenotypes are continuous random variables, as they are in our two examples. As we have noted, if one is willing to specify the distribution of the

regression errors in the population up to a parameter, then it is possible to estimate the parameter β in an efficient manner. However, we have shown that misspecification of that parameter model will lead to inconsistent estimation of β. Our approach is entirely different. While we specify a target regression error distribution, we have shown that the estimation is robust to violation of that target distribution, both theoretically and in a simulation study. In the rare disease case that would be the reason for a case-control study in the first place, an alternative is to simply use only the data for the controls. We have shown in simulations and in our two data examples that such throwing away of 50% of the data leads to a highly non-trivial loss of efficiency compared to our method.

In the fourth project, we have shown how to select the smoothing parameters in joint model of paired sparse functional data using novel versions of AIC and BIC. The methodology was described for penalized spline mixed effects model and justified in the important paired sparse functional data case. Numerically, we have found that AIC and BIC are very close to being the same as 10-fold cross-validation in terms of model fits, while being much faster computationally.

In the fifth project, we developed cross-validated permutation test (CPT), cross-validated within treatment permutation test (CWPT) and diagonal linear discriminant analysis (DLDA) to assess the prediction ability of the genetic predictors in the gene expression data that are generated by CodeLink System from GE. From CPT, we found 5 genes in the mucosal array of ACF stage were statistically significant in terms of predicting tumor outcome in colonic tissue; and 1 gene was significant in the two-stage normalized array of tumor stage. By CWPT, we found hundreds to thousands of genes improved classification significantly over covariates at both ACF and tumor stages. From DLDA, the 25 selected genes in the GE normalized fecal arrays improve the classification significantly over covariates at the tumor stage. The

application on khan data and breast cancer data gives similar results to those in Khan et al. (2001) and Hofling and Tibshirani (2008).

REFERENCES

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," *Second International Symposium on Information Theory*, Budapest: Akademiai Kiado, pp. 267-281.

Apanasovich, T. V., Carroll, R. J. and Maity, A. (2009), "SIMEX and variance estimation in semiparametric measurement error models." *Electronic Journal of Statistics*, 3, 318-348.

Billingsley, P. (1968), *Convergence of Probability Measures*, New York: Wiley.

Byrd, R. H., Lu, P., Nocedal, J., Zhu, C. (1995), "A limited memory algorithm for bound constrained optimization," *SIAM Journal on Scientific Computing*, 16, 1190-1208.

Chatterjee, N., Kalaylioglu, Z., Moslehi, R., Peters, U. and Wacholder, S. (2006), "Powerful multi-locus tests for genetic association in the presence of gene-gene and gene-environment interactions," *American Journal of Human Genetics*, 79, 1002-1016.

Chen, Y.-H., Carroll, R. J. and Chatterjee, N. (2008), "Retrospective analysis of haplotype-based case-control studies under a flexible model for gene-environment association," *Biostatistics*, 9, 81-99.

Chen, Y.-H., Chatterjee, N. and Carroll, R. J. (2009), "Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies," *Journal of the American Statistical Association*, 104, 220-233.

Claeskens, G. and Carroll, R. J. (2007), "An asymptotic theory for model selection inference in general semiparametric problems," *Biometrika*, 94, 249-265.

Claeskens, G. and Van Keilegom, I. (2003), "Bootstrap confidence bands for regression curves and their derivatives," *Annals of Statistics*, 31, 1852-1884.

Davidson, L. A., Jiang, Y. H., Lupton, J. R. and Chapkin, R. S. (1995), "Noninvasive detection of putative biomarkers for colon cancer using fecal messenger RNA," *Cancer Epidemiology, Biomarkers and Prevention*, 4, 643-647.

Davidson, L. A., Lupton, J. R., Miskovsky, E., Fields, H. P. and Chapkin, R. S. (2003), "Quantification of human intestinal gene expression profiles using exfoliated colonocytes: a pilot study," *Biomarkers*, 8, 51-61.

de Jong, P. (1987), "A Central Limit Theorem for generalized quadratic forms," *Probability Theory and Related Fields*, 75, 261-277.

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002), "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of American Statistical Association*, 97, 77-87.

Fan, J. and Huang, T. (2005), "Profile likelihood inferences on semiparametric varying-coefficient partially linear models," *Bernoulli*, 11, 1031-1057.

Fan, J. and Jiang, J. (2005), "Nonparametric inferences for additive models." *Journal of the American Statistical Association*, 100, 890-907.

Fan, J., Zhang, C. and Zhang, J. (2001), "Generalized likelihood ratio statistics and Wilks phenomenon," *Annals of Statistics*, 29, 153-193.

Gohagan, J. K., Prorok, P. C., Hayes, R. B., et al. (2000), "The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer screening trial of the National Cancer Institute: history, organization, and status," *Controlled Clinical Trials*, 21,(6 Suppl), 251S-272S.

Gonzalez, C. A., Sala, N. and Capella, G. (2002), "Genetic susceptibility and gastric cancer risk," *Int. J. Cancer*, 100, 249-260.

Hofling, H. and Tibshirani, R. (2008), "A study of Pre-validation," *Annals of Applied Statistics*, 2, 643-664.

Hong, M. Y., Bancroft, L. K., Turner, N. D., Davidson, L. A., Murphy, M. E., Carroll, R. J., Chapkin, R. S. and Lupton, J. R. (2005), "Fish oil decreases oxidative DNA damaga by enhancing apoptosis in rat colon," *Nutrition and Cancer*, 52(2), 166-175.

Hong, M. Y., Chang, W.-C. L., Chapkin, R. S. and Lupton, J. R. (1997), "Relationship among colonocyte proliferation, differentiation and apoptosis as a function of diet and carcinogen," *Nutrition and Cancer*, 28, 20-29.

Hua, J., Tembe, W. D. and Dougherty, E. R. (2009), "Performace of feature-selection methods in the classification of high-dimensional data," *Pattern Recognition*, 42, 409-424.

James G. M., Hastie, T. J. & Sugar, C. A. (2000), "Principal component models for sparse functional data," *Biometrika*, 87, 587-602.

Jiang, Y., Scott, A. J. and Wild, C. J. (2006), "Secondary analysis of case-control data," *Statistics in Medicine*, 25, 1323-1339.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescr, C. R., Peterson, C. and Meltzer, P. S. (2001), "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, 7, 673-679.

Kim, Y. S., Young, M. R., Bobe, G. Colburn, N. H. and Milner, J. A. (2009), "Bioactive food components, inflammatory targets, and cancer prevention," *Cancer Prev*

Res, 2(3), 200-208.

Lederman, M. M., Connick, E., Landay, A., Kuritzkes, D. R., Spritzler, J., Clair, M. S., Kotzin, B. L., Fox, L., Chiozzi, M. H., Leonard, J. M., Rousseau, F., Wade, M., D'Arc Roe, J., Martinez, A. and Kessler, H. (1998), "Immunologic responses associated with 12 weeks of combination antiretroviral therapy consisting of zidovudine, lamivudine, and ritonavir: results of AIDS clinical trials group protocol 315," *The Journal of Infectious Diseases*, 178, 70-99.

Lin, D. Y. and Zheng, D. (2009), "Proper analysis of secondary phenotype data in case-control association studies," *Genetic Epidemiology*, 33, 256-265.

Liu, L., Wang, N., Lupton J. R., Turner, N. D., Chapkin, R. S. and Davidson, L. A. (2005), "A two-stage normalization method for partially degraded mRNA microarray data," *Bioinformatics*, 21, 4000-4006.

Maity, A., Carroll, R. J., Mammen, E. and Chatterjee, N. (2009), "Testing in semiparametric models with interaction, with applications to gene-environment interactions," *Journal of the Royal Statistical Society, Series B*, 71, 75-96.

Moslehi, R., Chatterjee, N., Church, T. R., Chen, J., Yeager, M., Weissfield, J., Hein, D. W., and Hayes, R. B. (2006), "Cigarette smoking, N-acetyltransferase genes and the risk of advanced colorectal adenoma," *Pharmacogenomics*, 7, 819-829.

Peng, J. and Paul, D. (2009), "A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data," *Journal of Computational and Graphical Statistics* 18, 995-1015.

Rice, J. A. and Wu, C. (2001), "Nonparametric mixed effects models for unequally sampled noisy curves," *Biometrics* , 57, 253-259.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003), *Semiparametric Regression*,

Cambridge University Press.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6(2), 461-464.

Shao, J. (1997), "An asymptotic theory for linear model selection," *Statistica Sinica*, 7, 221-264.

Spinka, C., Carroll, R. J. and Chatterjee, N. (2005), "Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity," *Genetic Epidemiology*, 29, 108-127.

Stone, M. (1977). An asymptotic equivalence of choice of model by crossvalidation and Akaike's criterion, J. Roy. Statist. Soc. Ser. B 39, 44-47.

Van Keilegom, I. and Carroll, R. J. (2007), "Backfitting versus profiling in general criterion functions," *Statistica Sinica*, 17, 797-816.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005), "Functional data analysis for sparse longitudinal data," *Journal of the American Statistical Association* , 100, 577-590.

Zhou, L., Huang, J. and Carroll R. J. (2008), "Joint modeling of paired sparse functional data using principal components," *Biometrika*, 95, 601-619.

APPENDIX A

TESTING FOR CONSTANT NONPARAMETRIC EFFECTS IN GENERAL SEMIPARAMETRIC REGRESSION MODELS WITH INTERACTIONS

To illustrate the general result (2.7), we consider the partially linear logistic model (2.2). However, it is clear from the expansions described in Section 3, which are quite general, that the Wilks phenomenon result holds more generally.

Define $W(\gamma) = 1 + \gamma X^{\mathrm{T}}\beta_0$.

1. Testing Theory

For the null model, we have that

$$\mathrm{pr}(Y = 1|X, Z) = H(\kappa_0^* + X^{\mathrm{T}}\beta_0^* + S^{\mathrm{T}}\eta_0^*)$$

independent of γ. Define $\theta(\gamma)$ so that $\kappa_0^* + X^{\mathrm{T}}\beta_0^* + S^{\mathrm{T}}\eta_0^* = X^{\mathrm{T}}\beta_0 + S^{\mathrm{T}}\eta_0 + \theta(\gamma) + \gamma X^{\mathrm{T}}\beta_0\theta(\gamma)$. Then, under the null model,

$$\mathrm{pr}(Y = 1|X, Z) = H\{X^{\mathrm{T}}\beta_0 + S^{\mathrm{T}}\eta_0 + W(\gamma)\theta(\gamma)\}, \tag{A.1}$$

while under the full model,

$$\mathrm{pr}(Y = 1|X, Z) = H\{X^{\mathrm{T}}\beta_0 + S^{\mathrm{T}}\eta_0 + W(\gamma)\theta(Z, \beta_0, \eta_0, \gamma)\}, \tag{A.2}$$

and the null hypothesis is that

$$H_0 : \theta(z, \beta_0, \eta_0, \gamma) \equiv \theta(\gamma). \tag{A.3}$$

We have already described in Section 2 that we can treat (β_0^*, η_0^*) and (β_0, η_0) as if they were known. In this case, with the exception of the minor difference of

the offset $X^{\mathrm{T}}\beta_0 + S^{\mathrm{T}}\eta_0$ in (A.1) and (A.2), for any fixed γ the problem is exactly the same as a special case of that studied by Fan, Zhang and Zhang (2001) in their Theorem 10, which applies to generalized linear models. There are, however, huge simplifications because the null hypothesis (A.3) takes a particularly simple form.

Let $K_c(v) = K * K(v)$, the convolution function, and let $K_{ch}(v) = h^{-1}K_c(v/h)$. Define the volume of the support of Z to be $\mathcal{V} = E\{1/f_Z(Z)\}$, and also define $\mu_n(h) = h^{-1}\mathcal{V}\{K(0) - (1/2)\int K^2(t)dt\}$ and $\sigma_n^2(h) = 2\mu_n(h)/r_K$, where

$$r_K = \frac{K(0) - (1/2)\int K^2(t)dt}{\int\{K(t) - (1/2)K_c(t)\}^2 dt}.$$

Let $\Lambda_n(\gamma)$ be the likelihood ratio test statistic. Then by Fan et al. (2001), under the null hypothesis, independent of the value of γ, $\sigma_n^{-1}(h)\{\Lambda_n(\gamma) - \mu_n(h)\} \Rightarrow \mathrm{Normal}(0,1)$. This implies that the mean of $r_K\Lambda_n(\gamma)$ is $r_K\mu_n(h)$ and the variance is $2\mu_n(h)r_K$, as one would have with a chi-squared random variable with $r_K\mu_n(h)$ degrees of freedom, see their Theorem 5 on page 165 and Theorem 10 on page 174.

Define

$$\begin{aligned}
\Omega(z_0,\gamma) &= f_Z(z_0)E[W^2(\gamma)H^{(1)}\{X^{\mathrm{T}}\beta_0 + S^{\mathrm{T}}\eta_0 + W(\gamma)\theta(Z,\gamma)\}|Z = z_0]; \\
\epsilon &= Y - H\{X^{\mathrm{T}}\beta_0 + S^{\mathrm{T}}\eta_0 + W(\gamma)\theta(Z,\gamma)\}.
\end{aligned}$$

Using Fan et al (2001) (see their page 191), which applies to generalized linear models and allows heteroscedastic ϵ_i, we obtain

$$\begin{aligned}
\Lambda_n(\gamma) &= n^{-1}\textstyle\sum_{k=1}^{2}\sum_{i=1}^{n} K_h(Z_k - Z_i)\epsilon_k\epsilon_i W_i(\gamma)W_k(\gamma)/\Omega(Z_k,\gamma) \\
&\quad -(1/2)n^{-2}\textstyle\sum_{k=1}^{2}\sum_{i=1}^{n}\sum_{j=1}^{n} H^{(1)}\{X_k^{\mathrm{T}}\beta_0 + S_k^{\mathrm{T}}\eta_0 + W_k(\gamma)\theta(\gamma)\}\epsilon_i\epsilon_j W_i(\gamma) \\
&\qquad \times W_j(\gamma)W_k^2(\gamma)\{\Omega(Z_k,\gamma)\}^{-2}K_h(Z_i - Z_k)K_h(Z_j - Z_k) \times \{1 + o_p(1)\} \\
&= \{T_n(\gamma) - S_n(\gamma)\} \times \{1 + o_p(1)\}.
\end{aligned}$$

Also, modulo higher order terms, it follows that

$$\begin{aligned} T_n(\gamma) &= h^{-1}K(0)E\{1/f_Z(Z)\} + n^{-1}\sum_{k\neq i}^{n} K_h(Z_k - Z_i)\epsilon_k\epsilon_i W_k(\gamma)W_i(\gamma)/\Omega(Z_k,\gamma); \\ S_n(\gamma) &= (1/2)h^{-1}E\{1/f_Z(Z)\}\int K^2(t)dt \\ &\quad +(2n)^{-1}\sum_{k\neq i}^{n} K_{ch}(Z_i - Z_k)\epsilon_k\epsilon_i W_k(\gamma)W_i(\gamma)/\Omega(Z_k,\gamma). \end{aligned}$$

Make the definition

$$\mathcal{U}_n(\gamma) = n^{-1}\sum_{k\neq i}^{n}\{K_h(Z_k - Z_i) - (1/2)K_{ch}(Z_k - Z_i)\}\epsilon_k\epsilon_i W_k(\gamma)W_i(\gamma)/\Omega(Z_k,\gamma).$$

Then we have that

$$\Lambda_n(\gamma) = \{\mu_n(h) + \mathcal{U}_n(\gamma)\}\{1 + o_p(1)\}.$$

It is obvious that these results are uniform in compact sets for γ. Since the terms $\{1 + o_p(1)\}$ are actually $\{1 + o_p(h^{-1/2})\}$, see Fan et al (2001), page 183, if $\mathcal{D}$ is such a compact set, then uniformly in $\gamma \in \mathcal{D}$, $\sup_{\gamma\in\mathcal{D}}|h^{1/2}\{\Lambda_n(\gamma) - \mu_n(h)\} - h^{1/2}\mathcal{U}_n(\gamma)| = o_p(1)$. In section 2.b we show that $h^{1/2}\mathcal{U}_n(\gamma)$, as a process in γ, converges to a Gaussian process.

2. Weak Convergence of $\mathcal{U}_n$

To prove that the process $h^{1/2}\mathcal{U}_n(\gamma)$ converges to a Gaussian process in $\gamma \in \mathcal{D}$, we have to show that the finite dimensional distributions converge to normality, and that the process is tight.

Make the definitions

$$\begin{aligned} a_{ii}(\gamma) &= 0; \\ a_{ij}(\gamma) &= h^{1/2}\{K_h(Z_i - Z_j) - (1/2)K_{ch}(Z_i - Z_j)\}W_i(\gamma)W_j(\gamma)/\Omega(Z_i, \gamma); \\ c_{ij}(\gamma) &= n^{-1}\{a_{ij}(\gamma) + a_{ji}(\gamma)\}\epsilon_i\epsilon_j, \end{aligned}$$

the latter two are defined when $i \neq j$. Then we have that

$$h^{1/2}\mathcal{U}_n(\gamma) = \sum_{1\leq i<j\leq n} c_{ij}(\gamma),$$

where once again we note that (ϵ_i, ϵ_j) are independent of γ under the null hypothesis.

We use Proposition 3.2 in de Jong (1987) to show that $h^{1/2}\mathcal{U}_n(\gamma)$ converges to a Gaussian distribution for any fixed γ. Define

$$\begin{aligned} G_I &= \sum_{1\leq i<j\leq n} E\{c_{ij}(\gamma)^4\}; \\ G_{II} &= \sum_{1\leq i<j<k\leq n} \Big[E\{c_{ij}^2(\gamma)c_{ik}^2(\gamma)\} + E\{c_{ji}^2(\gamma)c_{jk}^2(\gamma)\} + E\{c_{ki}^2(\gamma)c_{kj}^2(\gamma)\}\Big]; \\ G_{IV} &= \sum_{1\leq i<j<k<l\leq n} \Big[E\{c_{ij}(\gamma)c_{ik}(\gamma)c_{lj}(\gamma)c_{lk}(\gamma)\} + E\{c_{ij}(\gamma)c_{il}(\gamma)c_{kj}(\gamma)c_{kl}(\gamma)\} \\ &\qquad + E\{c_{ik}(\gamma)c_{il}(\gamma)c_{jk}(\gamma)c_{jl}(\gamma)\}\Big]; \end{aligned}$$

To apply this proposition, we need to check the following conditions:

C1. $h^{1/2}\mathcal{U}_n(\gamma)$ is clean in the sense of de Jong (1987).

Using Definition 2.1 in de Jong (1987), we call $h^{1/2}\mathcal{U}_n(\gamma)$ is clean if the coditional expectations of c_{ij} vanish, and this is obviously true since $E(\epsilon|X, S, Z) = 0$.

C2. $\text{var}\{h^{1/2}\mathcal{U}_n(\gamma)\}$ converges to a finite quantity as $n \to \infty$.

C3. G_I is of smaller order than $\text{var}\{h^{1/2}\mathcal{U}_n(\gamma)\}$.

C4. G_{II} is of smaller order than $\text{var}\{h^{1/2}\mathcal{U}_n(\gamma)\}$.

C5. G_{IV} is of smaller order than $\text{var}\{h^{1/2}\mathcal{U}_n(\gamma)\}$.

In what follows, we check conditions C2 - C5 as condition C1 follows directly from the fact that $E(\epsilon|X,S,Z)=0$.

We use the following result:

Lemma 1 Let Z_i and Z_j are independent and identically distributed random variables with a strictly positive density and compact support and let $K(\cdot)$ be a symmetric kernel. Define $K_m(c)$ to be the $m-$ fold convolution of $K(c)$. Then

$$\begin{aligned}
&E[K_h^2(Z_i-Z_j)\Omega(Z_j,\gamma)/\{f_Z(Z_j)f_Z(Z_i)\Omega(Z_i,\gamma)\}] \\
&\quad = h^{-1}K_2(0)\ E\{1/f(Z_i)\}\{1+O(h)\} \\
&E[K_{2h}^2(Z_i-Z_j)\Omega(Z_j,\gamma)/\{f_Z(Z_j)f_Z(Z_i)\Omega(Z_i,\gamma)\}] \\
&\quad = h^{-1}K_4(0)\ E\{1/f(Z_i)\}\{1+O(h)\} \\
&E[K_{2h}(Z_i-Z_j)K_h(Z_i-Z_j)\Omega(Z_j,\gamma)/\{f_Z(Z_j)f_Z(Z_i)\Omega(Z_i,\gamma)\}] \\
&\quad = h^{-1}K_3(0)\ E\{1/f(Z_i)\}\{1+O(h)\}
\end{aligned}$$

To check condition C2, first observe that

$$\text{var}\{h^{1/2}\mathcal{U}_n(\gamma)\} \;=\; \sum_{i<j} E\{c_{ij}^2(\gamma)\}$$

Then we derive

$$\begin{aligned}
E\{c_{ij}^2(\gamma)\} &= (n^{-2}/4)E[\{a_{ij}(\gamma)+a_{ji}(\gamma)\}^2\epsilon_i^2\epsilon_j^2] \\
&= (n^{-2}/4)E[\{a_{ij}^2(\gamma)+a_{ji}^2(\gamma)+2a_{ij}(\gamma)a_{ji}(\gamma)\}\epsilon_i^2\epsilon_j^2] \\
&= A_1+A_2+A_3.
\end{aligned}$$

Now we see that, using Lemma 1,

$$\begin{aligned} A_1 &= hn^{-2}E\Big([\{K_h(Z_i - Z_j) - (1/2)K_{ch}(Z_i - Z_j)\}W_i(\gamma)W_j(\gamma)/\Omega(Z_i,\gamma)]^2\epsilon_i^2\epsilon_j^2\Big) \\ &= hn^{-2}E\Big([\{K_h(Z_i - Z_j) - (1/2)K_{ch}(Z_i - Z_j)\}]^2\Omega(Z_j,\gamma) \\ &\qquad \times \{f_Z(Z_j)f_Z(Z_i)\Omega(Z_i,\gamma)\}^{-1}\Big) \\ &= n^{-2}[K_2(0) - K_4(0) + (1/4)K_4(0)]E\{1/f_Z(Z)\}\{1 + O(h)\}. \end{aligned}$$

Similarly,

$$\begin{aligned} A_2 &= n^{-2}[K_2(0) - K_3(0) + (1/4)K_2(0)]E\{1/f_Z(Z)\}\{1 + O(h)\}; \\ A_3 &= n^{-2}[2K_2(0) - 2K_3(0) + (1/2)K_4(0)]E\{1/f_Z(Z)\}\{1 + O(h)\}. \end{aligned}$$

Hence we have

$$\begin{aligned} \lim_{n\to\infty} \mathrm{var}\{h^{1/2}\mathcal{U}_n(\gamma)\} &= \lim_{n\to\infty} \sum_{i<j} E\{c_{ij}^2(\gamma)\} \\ &= [4K_2(0) - 4K_3(0) + K_4(0)]E\{1/f_Z(Z)\}, \end{aligned}$$

and hence condition C2 is satisfied.

Next, similar calculations as in Lemma 1 show that

$$\begin{aligned} E\{a_{ij}^4(\gamma)\epsilon_i^4\epsilon_j^4\} &= O(h^{-1}); \\ E\{a_{ij}^2(\gamma)a_{ji}^2(\gamma)\epsilon_i^4\epsilon_j^4\} &= O(h^{-1}), \end{aligned}$$

and it follows that $E\{c_{ij}^4(\gamma)\} = O(n^{-4}h^{-1})$. Hence we have that $G_I = O(n^{-2}h^{-1}) = o(1)$.

To check condition C4, we observe that,

$$\begin{aligned} E\{a_{ij}^2(\gamma)a_{ik}^2(\gamma)\epsilon_i^4\epsilon_j^2\epsilon_k^2\} = O(h^{-1}); \\ E\{a_{ij}(\gamma)a_{ji}(\gamma)a_{ik}(\gamma)a_{ki}(\gamma)\epsilon_i^4\epsilon_j^2\epsilon_k^2\} = O(h^{-1}), \end{aligned}$$

and similarly for other terms in the expansion of $c_{ij}^2(\gamma)c_{ik}^2(\gamma)$. It follows that $E\{c_{ij}^2(\gamma)c_{ik}^2(\gamma)\} = O(n^{-4}h^{-1})$. Hence we have that $G_{II} = O(n^{-1}h^{-1}) = o(1)$.

Finally, to check condition C5, we note that

$$E\{a_{ij}(\gamma)a_{jk}(\gamma)a_{k\ell}(\gamma)a_{\ell i}(\gamma)\epsilon_i^2\epsilon_j^2\epsilon_k^2\epsilon_\ell^2\} = O(h),$$

and similarly for other cross product terms, and thus implying $E\{c_{ij}(\gamma)c_{jk}(\gamma)c_{k\ell}(\gamma)c_{\ell i}(\gamma)\} = O(n^{-4}h)$. Hence we get $G_{IV} = O(h) = o(1)$.

We have thus shown that conditions C1-C5 are satisfied and hence the proof is complete.

We have to show that there exists $\zeta > 0$, $\eta > 1$, such that, for any $\gamma_1 < \gamma < \gamma_2$,

$$h^\zeta E\left\{|\mathcal{U}_n(\gamma) - \mathcal{U}_n(\gamma_1)|^\zeta |\mathcal{U}_n(\gamma) - \mathcal{U}_n(\gamma_2)|^\zeta\right\} \leq |\gamma_1 - \gamma_2|^\eta, \tag{A.4}$$

see Billingsley (1968, page 128). We show below that (A.4) holds for $\zeta = 1$.

Using the Cauchy-Schwarz inequality we observe

$$\begin{aligned} h^2E^2\left\{|\mathcal{U}_n(\gamma) - \mathcal{U}_n(\gamma_1)||\mathcal{U}_n(\gamma) - \mathcal{U}_n(\gamma_2)|\right\} &\leq E\left\{h^{1/2}|\mathcal{U}_n(\gamma) - \mathcal{U}_n(\gamma_1)|\right\}^2 \\ &\quad \times E\left\{h^{1/2}|\mathcal{U}_n(\gamma) - \mathcal{U}_n(\gamma_2)|\right\}^2. \end{aligned}$$

Recall that $h^{1/2}\mathcal{U}_n(\gamma) = \sum_{1\leq i<j\leq n} c_{ij}(\gamma)$. Let $c_{ij}^{(1)}(\gamma)$ denote the first derivative of $c_{ij}(\gamma)$ with respect to γ and similarly for $a_{ij}(\gamma)$. Now we see that

$$\begin{aligned} E\left\{h^{1/2}|\mathcal{U}_n(\gamma) - \mathcal{U}_n(\gamma_1)|\right\}^2 &= E\left[\sum_{1\leq i<j\leq n}\{c_{ij}(\gamma) - c_{ij}(\gamma_1)\}\right]^2 \\ &= E\left[\sum_{1\leq i<j\leq n}\{c_{ij}(\gamma) - c_{ij}(\gamma_1)\}^2\right] \\ &\leq (\gamma_1 - \gamma_2)^2 C, \end{aligned}$$

where $C = n^2\sup_{\gamma,\gamma_1\in[L,R]}E[\{c_{ij}(\gamma) - c_{ij}(\gamma_1)\}^2]$. Similar calculations can be done

for $E\left\{h^{1/2}|\mathcal{U}_n(\gamma)-\mathcal{U}_n(\gamma_2)|\right\}^2$, and hence $hE\left\{|\mathcal{U}_n(\gamma)-\mathcal{U}_n(\gamma_1)||\mathcal{U}_n(\gamma)-\mathcal{U}_n(\gamma_2)|\right\} \leq |\gamma_1-\gamma_2|^2C$. The only thing remaining is to show that C is finite, which follows immediately from condition C2.

Note that $K(s) = K(-s)$ and since Z has a positive density function on a compact support, we have $E\{1/f_Z(Z)\} = \int_z dz$. Then

$$\begin{aligned}
&E\{K_h^2(Z_i-Z_j)\Omega(Z_j,\gamma)/\{f_Z(Z_j)f_Z(Z_i)\Omega(Z_i,\gamma)\}\} \\
&\quad = h^{-2}\int K^2\{(z_j-z_i)/h\}\Omega(z_j,\gamma)/\{\Omega(z_i,\gamma)\}\, dz_i\, dz_j \\
&\quad = h^{-1}\int K^2(t)\Omega(z_i+th,\gamma)/\{\Omega(z_i,\gamma)\}\, dz_i\, dt \\
&\quad = h^{-1}\int K^2(t)\{\Omega(z_i,\gamma)+O(h)\}/\{\Omega(z_i,\gamma)\}\, dZ_i\, dt \\
&\quad = h^{-1}\int K^2(t)\, dz_i\, dt\{1+O(h)\} \\
&\quad = h^{-1}\int K^2(t)\, dt\; E\{1/f(Z_i)\}\{1+O(h)\} \\
&\quad = h^{-1}K_2(0)\; E\{1/f(Z_i)\}\{1+O(h)\}.
\end{aligned}$$

Also,

$$\begin{aligned}
&E\{K_{2h}^2(Z_i-Z_j)\Omega(Z_j,\gamma)/\{f_Z(Z_j)f_Z(Z_i)\Omega(Z_i,\gamma)\}\} \\
&\quad = h^{-2}\int K_2^2\{(z_j-z_i)/h\}\Omega(z_j,\gamma)/\{\Omega(z_i,\gamma)\}\, dz_i\, dz_j \\
&\quad = h^{-1}\int K_2^2(t)\Omega(z_i+th,\gamma)/\{\Omega(z_i,\gamma)\}\, dz_i\, dt \\
&\quad = h^{-1}\int K_2^2(t)\{\Omega(z_i,\gamma)+O(h)\}/\{\Omega(z_i,\gamma)\}\, dz_i\, dt \\
&\quad = h^{-1}\int K_2^2(t)\, dz_i\, dt\{1+O(h)\} \\
&\quad = h^{-1}\int K_2^2(t)\, dt\; E\{1/f(Z_i)\}\{1+O(h)\} \\
&\quad = h^{-1}K_4(0)\; E\{1/f(Z_i)\}\{1+O(h)\}.
\end{aligned}$$

Similarly, we derive

$$
\begin{aligned}
&E\{K_{2h}(Z_i - Z_j)K_h(Z_i - Z_j)\Omega(Z_j,\gamma)/\{f_Z(Z_j)f_Z(Z_i)\Omega(Z_i,\gamma)\}\} \\
&\quad = h^{-2}\int K_2\{(z_j - z_i)/h\}K\{(z_j - z_i)/h\}\Omega(z_j,\gamma)/\{\Omega(z_i,\gamma)\}\; dz_i\; dz_j \\
&\quad = h^{-1}\int K_2(t)K(t)\Omega(z_i + th,\gamma)/\{\Omega(z_i,\gamma)\}\; dz_i\; dt \\
&\quad = h^{-1}\int K_2(t)K(t)\{\Omega(z_i,\gamma) + O(h)\}/\{\Omega(z_i,\gamma)\}\; dz_i\; dt \\
&\quad = h^{-1}\int K_2(t)K(t)\; dz_i\; dt\{1 + O(h)\} \\
&\quad = h^{-1}\int K_2(t)K(t)\; dt\; E\{1/f(Z_i)\}\{1 + O(h)\} \\
&\quad = h^{-1}K_3(0)\; E\{1/f(Z_i)\}\{1 + O(h)\},
\end{aligned}
$$

completing the proof.

APPENDIX B

LOCALLY EFFICIENT SCORE TESTS FOR INDEPENDENCE IN THE SECONDARY ANALYSIS OF CASE-CONTROL DATA

1. Robustness of the Efficient Score Test

Here we show that if the distribution of Y given X is misspecified, even under the null hypothesis of independence between (Y, X), then the score statistic does not in general have mean zero, and hence the score test is not robust.

We will use a result of Spinka, et al. (2005), who show that for any function $R(D, Y, X)$, under the null hypothesis

$$E_{cc}\left\{n^{-1}\sum_{i=1}^{n} R(D_i, Y_i, X_i)\right\} = \frac{n\pi_0}{n_0}\int f_Y(t) f_X(x) \times \sum_{d=0}^{1} R(d, t, x) S(d, t, x, \Omega) dt dx, \tag{B.1}$$

where $E_{cc}(\cdot)$ means expectation in the case-control sampling scheme, i.e., $E_{cc}\{G(D, Y, X)\} = n^{-1}\sum_{i=1}^{n} E\{G(D_i, Y_i, X_i)|D_i\}$.

Theorem 4 *Let the true distribution of Y under the null hypothesis of independence be $f_{y,\text{true}}(y)$, while the distribution of X is given as $f_x(x)$. Then $E_{\text{cc}}\{\sum_{i=1}^{n} \mathcal{K}_{\text{par}}(Y_i, X_i, \Omega)\} = 0$ does not hold in general, and indeed the expectation is given as*

$$\begin{aligned}
&\frac{n\pi_0}{n_0}\int f_x(s)\left\{\int L(t, s) f_{y,\text{true}}(t)\textstyle\sum_{d=0}^{1}\mathcal{C}(d, t, s, \Omega) dt\right\} ds \\
&-\frac{n\pi_0}{n_0}\int f_x(s)\frac{\int f_{y,\text{true}}(t)\sum_{d=0}^{1}\mathcal{C}(d, t, s, \Omega) dt}{\int f_Y(t)\sum_{d=0}^{1}\mathcal{C}(d, t, s, \Omega) dt} \\
&\qquad\times\left\{\int L(u, s) f_Y(u)\textstyle\sum_{d=0}^{1}\mathcal{C}(d, u, s, \Omega) du\right\} ds.
\end{aligned}$$

The proof is a simple consequence of (B.1), plus some detailed algebra.

2. Proof of Theorem 1

Using (B.1), we see that

$$E_{cc}\left\{n^{-1}\textstyle\sum_{i=1}^{n} L(Y_i, X_i, \zeta)\right\} = \frac{n\pi_0}{n_0}\int f_Y(t) f_X(x) \times \textstyle\sum_{d=0}^{1} L(t, x, \zeta) S(d, t, x, \Omega) dt dx. \tag{B.2}$$

Now note that

$$\begin{aligned}
&\textstyle\sum_{i=1}^{n}\sum_{d=0}^{1}\int S(d, Y_i, x, \Omega) p_{\text{est}}(Y_i) f_X(x) dx \\
&\quad = \frac{n\pi_0}{n_0} n^{-1}\textstyle\sum_{i=1}^{n}\int f_X(x)\sum_{d=0}^{1} S(d, Y_i, x, \Omega) dx \\
&\qquad\qquad \times\{\int f_X(x)\textstyle\sum_{d=0}^{1} S(d, Y_i, x, \Omega) dx\}^{-1} \\
&\quad = \frac{n\pi_0}{n_0}.
\end{aligned}$$

Hence the right hand side of (3.8) is <u>exactly</u> equal to

$$\begin{aligned}
\Lambda &= \frac{n_0}{n\pi_0}\textstyle\sum_{i=1}^{n}\sum_{d=0}^{1}\int L(Y_i, x, \zeta) S(d, Y_i, x, \Omega) p_{\text{est}}(Y_i) f_X(x) dx \\
&= n^{-1}\textstyle\sum_{i=1}^{n}\sum_{d=0}^{1}\dfrac{\int L(Y_i, x, \zeta) S(d, Y_i, x, \Omega) f_X(x) dx}{\int f_X(x)\sum_{d=0}^{1} S(d, Y_i, x, \Omega) dx} \\
&= n^{-1}\textstyle\sum_{i=1}^{n}\dfrac{\sum_{d=0}^{1}\int L(Y_i, x, \zeta) f_X(x) S(d, Y_i, x, \Omega) dx}{\sum_{d=0}^{1}\int f_X(x) S(d, Y_i, x, \Omega) dx} \\
&= n^{-1}\textstyle\sum_{i=1}^{n} U(Y_i, \Omega, \zeta).
\end{aligned}$$

Now apply (B.1), so that

$$\begin{aligned} E_{cc}(\Lambda) &= \frac{n\pi_0}{n_0}\int f_Y(t)f_X(x)\textstyle\sum_{d=0}^{1}U(t,\Omega,\zeta)S(d,t,x,\Omega)dtdx \\ &= \frac{n\pi_0}{n_0}\int f_Y(t)\frac{\sum_{d=0}^{1}\int L(t,x,\zeta)f_X(x)S(d,t,x,\Omega)dx}{\sum_{d=0}^{1}\int f_X(x)S(d,t,x,\Omega)dx} \\ &\qquad \times \int f_X(x)\textstyle\sum_{d=0}^{1}S(d,t,x,\Omega)dxdt \\ &= \frac{n\pi_0}{n_0}\int f_Y(t)\frac{\int f_X(x)\sum_{d=0}^{1}L(t,x,\zeta)S(d,t,x,\Omega)dx}{\int f_X(x)\sum_{d=0}^{1}S(d,t,x,\Omega)dx} \\ &\qquad \times \int f_X(x)\textstyle\sum_{d=0}^{1}S(d,t,x,\Omega)dxdt \\ &= \frac{n\pi_0}{n_0}\int f_Y(t)f_X(x)\textstyle\sum_{d=0}^{1}L(t,x,\zeta)S(d,t,x,\Omega)dtdx. \qquad \text{(B.3)} \end{aligned}$$

Since (B.2) and (B.3) are identical, we have thus shown that the test statistic (3.8) numerically equals (3.10) and has mean zero under the hypothesis, as claimed.

Now we take up the question as to whether (3.10) has <u>non-zero</u> mean under the <u>alternative</u> hypothesis. Under the alternative, Y and X and dependent, and we write $\text{pr}(X = x|Y) = \mathcal{Q}(x, y, \kappa)$, where $\mathcal{Q}(x, y, \kappa = 0) = f_X(x)$. Define $S_{\text{alt}}(d, t, x, \Omega, \kappa)$ the same as $S(d, t, x, \Omega)$ except that $f_X(x)$ in the latter is replaced by $\mathcal{Q}(x, y, \kappa)$. As seen in Chen, et al. (2008), for any function $R(Y, X)$,

$$E_{\text{cc}}\{n^{-1}\textstyle\sum_{i=1}^{n}R(Y_i, X_i)\} = \frac{n\pi_0}{n_0}\int f_Y(t)f_X(x)\sum_{d=0}^{1}R(t,x)S_{\text{alt}}(d,t,x,\Omega,\kappa)dxdt.$$

Then we have that

$$\begin{aligned} E_{\text{cc}}\{\textstyle\sum_{i=1}^{n}U(Y_i,\Omega,\zeta)\} &= \frac{n\pi_0}{n_0}\int f_Y(t)f_X(x)\textstyle\sum_{d=0}^{1}U(t,\Omega,\zeta)S_{\text{alt}}(d,t,x,\Omega,\kappa)dxdt \\ &= \frac{n\pi_0}{n_0}\int f_Y(t)\frac{\sum_{d=0}^{1}\int L(t,v,\zeta)f_X(v)S(d,t,v,\Omega)dv}{\sum_{d=0}^{1}\int f_X(u)S(d,t,u,\Omega)du} \\ &\qquad \times \int f_X(x)\textstyle\sum_{d=0}^{1}S_{\text{alt}}(d,t,x,\Omega,\kappa)dxdt. \end{aligned}$$

Similarly,

$$E_{\rm cc}\{\sum_{i=1}^{n} L(Y_i, X_i, \zeta)\} = \frac{n\pi_0}{n_0} \int f_Y(t) f_X(x) \sum_{d=0}^{1} L(t, x, \zeta) S_{\rm alt}(d, t, x, \Omega, \kappa) dx dt.$$

Clearly, these expectations are generally different, and hence the test generally has non-zero mean under alternatives.

3. Computation of (3.12)

Let $\mathrm{erf}(x) = 2\Phi(\sqrt{2}x) - 1$, where $\Phi(\cdot)$ is the standard normal distribution function, and let $\phi(\cdot)$ be the standard normal density function. Then if $A = \int \sum_{d=0}^{1} t S_{\rm par}(d, t, x, \Omega, 0, \zeta) dt$ and $B = \int \sum_{d=0}^{1} S_{\rm par}(d, t, x, Omega, 0, \zeta) dt$, we have that

$$\begin{aligned} A &= \beta_0 + (\beta_0 + \theta_{11}\sigma^2) C; \\ B &= 1 + C. \end{aligned}$$

where $C = \exp(\kappa + \theta_{11}\beta_0 + \theta_{12}x + \theta_{11}^2\sigma^2/2)$. Both terms were computed using Mathematica, and checked numerically.

4. Asymptotic Distribution of $\mathcal{V}(\widehat{\Omega}, \widehat{\zeta})$ in (3.11)

As in Theorem 2, $n_0/n_1 \to c$, $0 < c < \infty$. Let $\Theta = (\Omega, \zeta)$ and write $T_n = n^{1/2}\mathcal{V}(\widehat{\Omega}, \widehat{\zeta})$. Define $G_{\rm num}(y, x, \Theta) = L(y, x, \zeta) \sum_{d=0}^{1} S(d, y, x, \Omega)$ and $G_{\rm den}(y, x, \Theta) = \sum_{d=0}^{1} S(d, y, x, \Omega)$. Then

$$T_n = n^{-1/2} \sum_{i=1}^{n} \left\{ L(Y_i, X_i, \widehat{\zeta}) - \frac{n_0^{-1} \sum_{j=1}^{n} (1 - D_j) G_{\rm num}(Y_i, X_j, \widehat{\Theta})}{n_0^{-1} \sum_{j=1}^{n} (1 - D_j) G_{\rm den}(Y_i, X_j, \widehat{\Theta})} \right\}.$$

Define $\mathcal{A}_{\rm num}(y, \Theta) = E\{G_{\rm num}(y, X, \Theta) | D = 0\}$ and $\mathcal{A}_{\rm den}(y, \Theta) = E\{G_{\rm den}(y, X, \Theta) | D = 0\}$. Define

$$\mathcal{M}(y, x, \Theta) = \frac{\partial}{\partial \Theta^{\rm T}} \left\{ L(y, x, \zeta) - \frac{\mathcal{A}_{\rm num}(y, \Theta)}{\mathcal{A}_{\rm den}(y, \Theta)} \right\}$$

Define $E_{\rm cc}\{\mathcal{M}(Y,X,\Theta)\} = n^{-1}\sum_{i=1}^{n} E\{\mathcal{M}(Y_i,X_i,\Theta)|D_i\}$. By simple calculations, using our estimators of Θ as described in Section 3, we have that for some Θ_0 and some estimating function $\Phi(Y,X,D,\Theta_0)$,

$$n^{1/2}(\widehat{\Theta}-\Theta_0) = n^{-1/2}\sum_{i=1}^{n}\Phi(Y_i,X_i,D_i) + o_p(1),$$

where $E_{\rm cc}\{\Phi(Y,X,D,\Theta_0)\} = \sum_{i=1}^{n} E\{\Phi(Y_i,X_i,D_i,\Theta_0)|D_i\} = 0$. Make the definition that $\mu_5(d) = E\{\Phi(Y,X,D,\Theta_0)|D=d\}$, and since $\sum_{i=1}^{n}\mu_r(D_i) = 0$, it follows that

$$\begin{aligned} T_n &= n^{-1/2}\sum_{i=1}^{n}\left\{L(Y_i,X_i,\zeta_0) - \frac{n_0^{-1}\sum_{j=1}^{n}(1-D_j)G_{\rm num}(Y_i,X_j,\Theta_0)}{n_0^{-1}\sum_{j=1}^{n}(1-D_j)G_{\rm den}(Y_i,X_j,\Theta_0)}\right\} \\ &\quad + E_{cc}\{\mathcal{M}(Y,X,\Theta_0)\}n^{-1/2}\sum_{i=1}^{n}\{\Phi(Y_i,X_i,D_i,\Theta_0) - \mu_5(D_i)\} + o_p(1). \end{aligned}$$

Define

$$\begin{aligned} Z_{\rm num}(Y,\Theta_0) &= n_0^{-1/2}\sum_{j=1}^{n}(1-D_j)\left\{G_{\rm num}(Y_i,X_j,\Theta_0) - \mathcal{A}_{\rm num}(Y,\Theta_0)\right\}; \\ Z_{\rm den}(Y,\Theta_0) &= n_0^{-1/2}\sum_{j=1}^{n}(1-D_j)\left\{G_{\rm den}(Y_i,X_j,\Theta_0) - \mathcal{A}_{\rm den}(Y,\Theta_0)\right\}. \end{aligned}$$

Since by assumption $n_0/n_1 \to c,\ \ 0<c<\infty$, we have that $Z_{\rm num}\{R(\beta_0),\Theta_0\} = O_p(1)$ and $Z_{\rm den}\{R(\beta_0),\Theta_0\} = O_p(1)$. Thus,

$$\begin{aligned} &\frac{n_0^{-1}\sum_{j=1}^{n}(1-D_j)G_{\rm num}(Y,X_j,\Theta_0)}{n_0^{-1}\sum_{j=1}^{n}(1-D_j)G_{\rm den}(Y,X_j,\Theta_0)} - \frac{\mathcal{A}_{\rm num}(Y,\Theta_0)}{\mathcal{A}_{\rm den}(Y,\Theta_0)} \\ &= \frac{\mathcal{A}_{\rm num}(Y,\Theta_0) + n_0^{-1/2}Z_{\rm num}(Y,\Theta_0)}{\mathcal{A}_{\rm den}(Y,\Theta_0) + n_0^{-1/2}Z_{\rm den}(Y,\Theta_0)} - \frac{\mathcal{A}_{\rm num}(Y,\Theta_0)}{\mathcal{A}_{\rm den}(Y,\Theta_0)} \\ &= \frac{n_0^{-1/2}Z_{\rm num}(Y,\Theta_0)}{\mathcal{A}_{\rm den}(Y,\Theta_0)} - \frac{\mathcal{A}_{\rm num}(Y,\Theta_0)}{\mathcal{A}_{\rm den}^2(Y,\Theta_0)}n_0^{-1/2}Z_{\rm den}(Y,\Theta_0) + o_p(n_0^{-1/2}). \end{aligned}$$

Under regulatory conditions it follows that

$$\begin{aligned} T_n &= n^{-1/2}\textstyle\sum_{i=1}^n \{L(Y_i, X_i, \Theta_0) - U(Y_i, \Theta_0)\} \\ &\quad + n^{-1/2}\textstyle\sum_{i=1}^n \left\{ U(Y_i, \Theta_0) - \dfrac{\mathcal{A}_{\rm num}(Y_i, \Theta_0)}{\mathcal{A}_{\rm den}(Y_i, \Theta_0)} \right\} \\ &\quad - n_0^{-1} n^{-1/2} \textstyle\sum_{i=1}^n \sum_{j=1}^n (1 - D_j) \dfrac{G_{\rm num}(Y_i, X_j, \Theta_0) - \mathcal{A}_{\rm num}(Y_i, \Theta_0)}{\mathcal{A}_{\rm den}(Y_i, \Theta_0)} \\ &\quad + n_0^{-1} n^{-1/2} \textstyle\sum_{i=1}^n \sum_{j=1}^n \dfrac{\mathcal{A}_{\rm num}(Y_i, \Theta_0)}{\mathcal{A}^2_{\rm den}(Y_i, \Theta_0)} (1 - D_j) \{G_{\rm den}(Y_i, X_j, \Theta_0) - \mathcal{A}_{\rm den}(Y_i, \Theta_0)\} \\ &\quad + E_{cc}\{\mathcal{M}(Y, X, \Theta_0)\} n^{-1/2} \textstyle\sum_{i=1}^n \Phi(Y_i, X_i, D_i) + o_p(1) \\ &= \mathcal{D}_1 + \mathcal{D}_2 + \mathcal{D}_3 + \mathcal{D}_4 + \mathcal{D}_5 + o_p(1). \end{aligned}$$

We now proceed to analyze these terms in turn. By the rare disease assumption, $\mathcal{D}_2 = 0$ since $U(Y_i, \Theta_0) \approx \mathcal{A}_{\rm num}(Y_i, \Theta_0)/\mathcal{A}_{\rm den}(Y_i, \Theta_0)$. We know from Theorem 1 that if $\mu_1(d) = E\{L(Y, X, \Theta_0) - U(Y, \Theta_0)|D = d\}$, then $\sum_{i=1}^n \mu_{1i}(\Theta_0) = 0$, this latter sum being the expectation in the case control sampling scheme. Hence

$$\begin{aligned} \mathcal{D}_1 + \mathcal{D}_2 + \mathcal{D}_5 &= n^{-1/2}\textstyle\sum_{i=1}^n \{L(Y_i, X_i, \zeta_0) - U(Y_i, \Theta_0) - \mu_1(D_i)\} \\ &\quad + n^{-1/2}\textstyle\sum_{i=1}^n E_{cc}\{\mathcal{M}(Y, X, \Theta_0)\}\{\Phi(Y_i, X_i, D_i) - \mu_5(D_i)\} \\ &= n^{-1/2}\textstyle\sum_{i=1}^n \mathcal{K}(Y_i, X_i, D_i, \Theta_0), \end{aligned}$$

say, where $E\{\mathcal{K}(Y, X, D, \Theta_0)|D\} = 0$.

Similarly, note that

$$\begin{aligned} 0 &= \mu_3(d, y) = E\left\{ (1 - D) \frac{G_{\rm num}(y, X, \Theta_0) - \mathcal{A}_{\rm num}(y, \Theta_0)}{\mathcal{A}_{\rm den}(y, \Theta_0)} | D = d \right\} \\ 0 &= \mu_4(d, y) = E\left[(1 - D) \frac{\mathcal{A}_{\rm num}(y, \Theta_0)\{G_{\rm den}(y, X, \Theta_0) - \mathcal{A}_{\rm den}(y, \Theta_0)\}}{\mathcal{A}^2_{\rm den}(y, \Theta_0)} | D = d \right]. \end{aligned}$$

Let $c_* = n/n_0$. Hence,

$$\mathcal{D}_3 + \mathcal{D}_4 = c_* n^{-3/2} \textstyle\sum_{i=1}^n \sum_{j=1}^n W(Y_i, X_j, D_j, \Theta_0),$$

where

$$\begin{aligned} W(Y_i, X_j, D_j, \Theta_0) &= -(1-D_j)\frac{G_{\text{num}}(Y_i, X_j, \Theta_0) - \mathcal{A}_{\text{num}}(Y_i, \Theta_0)}{\mathcal{A}_{\text{den}}(Y_i, \Theta_0)} \\ &\quad +(1-D_j)\frac{\mathcal{A}_{\text{num}}(Y_i, \Theta_0)\{G_{\text{den}}(Y_i, X_j, \Theta_0) - \mathcal{A}_{\text{den}}(Y_i, \Theta_0)\}}{\mathcal{A}^2_{\text{den}}(Y_i, \Theta_0)}. \end{aligned}$$

Notice that $W(y, x, d = 1, \Theta_0) = 0$ and $E\{W(y, X, d = 0, \Theta_0)|D = 0\} = 0$. Without loss of generality, we can make the first n_0 observations be controls, and the last $n - n_0$ observations be the cases. Define $\widetilde{Z}_i = (Y_i, X_i, D_i)$, $Q_1(\widetilde{Z}_i, \widetilde{Z}_j, \Theta_0) = W(Y_i, X_j, D_j, \Theta_0) + W(Y_j, X_i, D_i, \Theta_0)$, $Q_2(\widetilde{z}, \Theta_0) = E\{W(Y, x, d, \Theta_0)|D = 1\}$ and $h_1(\widetilde{z}, \Theta_0) = E\{Q_1(\widetilde{z}, \widetilde{Z}, \Theta_0)|D = 0\}$. Then

$$\begin{aligned} \mathcal{D}_3 + \mathcal{D}_4 &= c_* n^{-3/2} \textstyle\sum_{i=1}^{n} \sum_{j=1}^{n_0} W(Y_i, X_j, D_j, \Theta_0) \\ &= c_* n^{-3/2} \textstyle\sum_{i=1}^{n_0} \sum_{j=1}^{n_0} W(Y_i, X_j, D_j, \Theta_0) \\ &\qquad + c_* n^{-3/2} \textstyle\sum_{i_0+1}^{n} \sum_{j=1}^{n_0} W(Y_i, X_j, D_j, \Theta_0) \\ &= c_* n^{-3/2} \textstyle\sum_{i=1}^{n_0} \sum_{j<i}^{n_0} Q_1(\widetilde{Z}_i, \widetilde{Z}_j, \Theta_0) \\ &\qquad + c_* n_1 n^{-3/2} \textstyle\sum_{j=1}^{n_0} n_1^{-1} \sum_{i_0+1}^{n} W(Y_i, X_j, D_j, \Theta_0) + o_p(1). \end{aligned}$$

An easy calculation shows that

$$\text{var}\left[n^{-3/2} \textstyle\sum_{j=1}^{n_0} \sum_{i_0+1}^{n} W(Y_i, X_j, D_j, \Theta_0) - n_1 n^{-3/2} \sum_{i=1}^{n_0} Q_2(\widetilde{Z}_i, \Theta_0)\right] \to 0.$$

Hence we have shown that

$$\begin{aligned} \mathcal{D}_3 + \mathcal{D}_4 &= c_* (n_0/n)^{3/2} n_0^{-3/2} \textstyle\sum_{i=1}^{n_0} \sum_{j<i}^{n_0} Q_1(\widetilde{Z}_i, \widetilde{Z}_j, \Theta_0) \\ &\qquad + c_* n_1 n^{-3/2} \textstyle\sum_{i=1}^{n_0} Q_2(\widetilde{Z}_i, \Theta_0) + o_p(1). \end{aligned}$$

Except for the factor $c_*(n_0/n)^{3/2}$, the first term above is a classical symmetric U-statistic of order two applied to independent and identically distributed observations, since by convention the first n_0 observations are the controls. It then follows from

standard U-statistic results that

$$\begin{aligned}
\mathcal{D}_3 + \mathcal{D}_4 &= c_*(n_0/n)^{3/2} n_0^{-1/2} \textstyle\sum_{i=1}^{n_0} h_1(0, \widetilde{Z}_i, \Theta_0) + c_* n_1 n^{-3/2} \sum_{i=1}^{n_0} Q_2(\widetilde{Z}_i, \Theta_0) + o_p(1) \\
&= c_*(n_0/n) n^{-1/2} \textstyle\sum_{i=1}^{n} (1 - D_i) h_1(D_i, \widetilde{Z}_i, \Theta_0) \\
&\qquad + c_*(n_1/n) n^{-1/2} \textstyle\sum_{i=1}^{n} (1 - D_i) Q_2(\widetilde{Z}_i, \Theta_0) + o_p(1) \\
&= n^{-1/2} \textstyle\sum_{i=1}^{n} h_2\{R_i(\beta_0), X_i, D_i, \Theta_0\} + o_p(1),
\end{aligned}$$

say, where of course $E[h_2\{R(\beta_0), X, D, \Theta_0\}|D] = 0$. Define $\Lambda(Y, X, D, \Theta_0) = \mathcal{K}(Y, X, D, \Theta_0) + h_2(Y, X, D, \Theta_0)$. Because of the way we have set things up, $E_{\mathrm{cc}}\{\Lambda(Y, X, D, \Theta_0)\} = 0$ and the normalized test statistic satisfies

$$\begin{aligned}
T_n &= n^{-1/2} \textstyle\sum_{i=1}^{n} \Lambda(Y_i, X_i, D_i, \Theta_0) + o_p(1) \to \mathrm{Normal}(0, \Sigma); \\
\Sigma &= = \mathrm{cov}_{\mathrm{cc}}(Y, X, D, \Theta_0) = \textstyle\sum_{d=0}^{1} (n_d/n) \mathrm{cov}\{\Lambda(Y, X, D, \Theta_0)|D = d\}.
\end{aligned}$$

In principle, all the terms in $\Lambda(\cdot, \Theta_0)$ can be estimated, and a method of moments covariance matrix can constructed separately for both the cases and the controls in order to estimate Σ.

APPENDIX C

LOCALLY EFFICIENT ESTIMATION FOR HOMOSCEDASTIC REGRESSION IN THE SECONDARY ANALYSIS OF CASE-CONTROL DATA

1. Unbiasedness of the Estimation Function (4.9)

Since this is a case-control sampling scheme, all expectations are conditional on $(D_1, ..., D_n)$. Let E_{cc} denote the expectation under the case-control sampling scheme and G an arbitrary function. Then, with $(\beta_{\text{true}}, \Omega_{\text{true}})$ the true parameter, β an arbitrary value, $\tau(x, \beta, \beta_{\text{true}}) = \mu(x, \beta_{\text{true}}) - \mu(x, \beta)$, and with $R(\beta) = Y - \mu(X, \beta)$,

$$E_{\text{cc}}\left[G\{R(\beta), X\}\right] = \textstyle\sum_{d=0}^{1}(n_d/n)E[G\{Y - \mu(X, \beta), X\}|D = d].$$

In order to derive the conditional density given the disease state we use the fact that we assume a logistic model, $P(D = 1|Y, X) = H\{\theta_0 + m(Y, X, \theta_1)\}$, with $H(x)$ is the logistic distribution function, for which $H\{\theta_0 + m(Y, X, \theta_1)\} = [1 - H\{\theta_0 + m(Y, X, \theta_1)\}] \exp\{\theta_0 + m(Y, X, \theta_1)\}$. Now write $f_{YX}(\cdot)$ as the joint density/mass function of (Y, X) in the population. Then, with θ_0 and θ_1 denoting the true parameters,

$$\begin{aligned}
\pi_d &= P(D = d) \\
&= \int H\{\theta_0 + m(y, x, \theta_1)\}^d [1 - H\{\theta_0 + m(y, x, \theta_1)\}]^{1-d} f_{YX}(y, x)\, dy\, dx \\
&= \int [1 - H\{\theta_0 + m(y, x, \theta_1)\}] \exp[d\{\theta_0 + m(y, x, \theta_1)\}] f_{YX}(y, x)\, dy\, dx.
\end{aligned}$$

It then follows that the density/mass function of (Y, X) given D

$$f_{YX|D=d}(y, x) = \frac{\exp[d\{\theta_0 + m(y, x, \theta_1)\}] f_{YX}(y, x)}{[1 + \exp\{\theta_0 + m(y, x, \theta_1)\}]\pi_d}.$$

If we make the rare disease assumption, this becomes $\exp[d\{\theta_0+m(y,x,\theta_1)\}]f_{YX}(y,x)$ $\times\pi_d^{-1}$. Recall that $\kappa=\theta_0+\log(n_1/n_0)-\log(\pi/\pi_0)$. The above expectation can now be computed as

$$\begin{aligned}
& E_{\text{cc}}\left[G\{R(\beta),X\}\right] \\
& = \textstyle\sum_{d=0}^{1}\frac{n_d}{n\pi_d}\int G\{y-\mu(x,\beta),x\}\exp[d\{\theta_0+m(y,x,\theta_1)\}]f_{YX}(y,x)\,dy\,dx \\
= & \ \frac{n_0}{n\pi_0}\int \textstyle\sum_{d=0}^{1} G\{y-\mu(x,\beta),x\}\frac{n_d/n_0}{\pi_d/\pi_0}\exp[d\{\theta_0+m(y,x,\theta_1)\}]f_{YX}(y,x)\,dy\,dx \\
= & \ \frac{n_0}{n\pi_0}\int G(r,x)[1+\exp[\kappa+m\{r+\mu(x,\beta),x,\theta_1\}]f_{YX}\{r+\mu(x,\beta),x\}\,dr\,dx.
\end{aligned}$$

We now note that the joint density/mass function of (Y,X) in the population is $f_{YX}(y,x)=f_\epsilon\{y-\alpha_{\text{true}}-\mu(x,\beta_{\text{true}})\}f_X(x)$. Hence, $f_{YX}\{r+\mu(x,\beta),x\}=f_\epsilon\{r-\alpha_{\text{true}}-\tau(x,\beta,\beta_{\text{true}})\}f_X(x)$. Thus,

$$\begin{aligned}
& E_{\text{cc}}\left[G\{R(\beta),X\}\right] \\
= & \ \frac{n_0}{n\pi_0}\int G(r,x)[1+\exp[\kappa+m\{r+\mu(x,\beta_{\text{true}})-\tau(x,\beta,\beta_{\text{true}}),x,\theta_1\}] \\
& \qquad\qquad \times f_\epsilon\{r-\alpha_{\text{true}}-\tau(x,\beta,\beta_{\text{true}})\}f_X(x)\,dr\,dx \\
= & \ \frac{n_0}{n\pi_0}\int G\{r+\tau(x,\beta,\beta_{\text{true}}),x\}[1+\exp[\kappa+m\{r+\mu(x,\beta_{\text{true}}),x,\theta_1\}] \\
& \qquad\qquad \times f_\epsilon(r-\alpha_{\text{true}})f_X(x)\,dr\,dx.
\end{aligned}$$

Now, since $\mathcal{K}(r,x,\beta_{\text{true}},\Omega_{\text{true}})=1+\exp[\kappa+m\{r+\mu(x,\beta_{\text{true}}),x,\theta_1\}]$, we have that

$$\begin{aligned}
& E_{\text{cc}}\left[G\{R(\beta),X\}\right] \qquad\qquad (C.1) \\
& = \frac{n_0}{n\pi_0}\int f_\epsilon(r-\alpha_{\text{true}})f_X(x)\mathcal{K}(r,x,\beta_{\text{true}},\Omega_{\text{true}})G\{r+\tau(x,\beta,\beta_{\text{true}}),x\}dr\,dx.
\end{aligned}$$

It follows from (C.1) that

$$\begin{aligned}
&\frac{n\pi_0}{n_0} E_{\mathrm{cc}}\{Q_n(\alpha_{\mathrm{true}}, \beta, \Omega_{\mathrm{true}})\} \\
&= n^{1/2} \int f_\epsilon(r - \alpha_{\mathrm{true}}) f_X(x) \mathcal{K}(r, x, \beta_{\mathrm{true}}, \Omega_{\mathrm{true}}) \\
&\times \Bigg[L\{r + \tau(x, \beta, \beta_{\mathrm{true}}), x, \alpha(\beta, \Omega_{\mathrm{true}}), \beta\} \\
&- \int \frac{L\{r + \tau(x, \beta, \beta_{\mathrm{true}}), v, \alpha(\beta, \Omega_{\mathrm{true}}), \beta\} \mathcal{K}\{r + \tau(x, \beta, \beta_{\mathrm{true}}), v, \beta, \Omega_{\mathrm{true}}\}}{\int \mathcal{K}\{r + \tau(x, \beta, \beta_{\mathrm{true}}), s, \beta, \Omega_{\mathrm{true}}\} f_X(s) ds} \\
&\quad \times f_X(v) dv dx dr \Bigg].
\end{aligned}$$

If $\beta = \beta_{\mathrm{true}}$, since $\tau(x, \beta_{\mathrm{true}}) = 0$, it follows directly that $E_{\mathrm{cc}}\{Q_n(\alpha_{\mathrm{true}}, \beta_{\mathrm{true}}, \Omega_{\mathrm{true}})\}$ $= 0$, and hence that $Q_n(\beta, \Omega)$ is an unbiased estimating equation. If $\beta \neq \beta_{\mathrm{true}}$, then in general we will have $0 \neq E_{\mathrm{cc}}\{Q_n(\alpha_{\mathrm{true}}, \beta, \Omega_{\mathrm{true}})\}$.

2. A Technical Lemma

The following Lemma is used in our analysis, including for the intercept. Refer to the definitions before the statement of Theorem 3.

Lemma 1 *Under regulatory conditions, as* $(n_0, n_1) \to \infty$ *such that* $n_0/n_1 \to c$, *with* $0 < c < \infty$,

$$\mathcal{H}_n(\beta, \Theta) = n^{-1/2} \textstyle\sum_{i=1}^n h_2\{R_i(\beta), X_i, D_i, \Theta\} + o_p(1), \tag{C.2}$$

where $E[h_2\{R(\beta), X, D, \Theta\}|D] = 0$.

Sketch of Proof: Define

$$\begin{aligned}
Z_{\mathrm{num}}\{R(\beta), \Theta\} &= n_0^{-1/2} \textstyle\sum_{j=1}^n (1 - D_j) \left[G_{\mathrm{num}}\{R(\beta), X_j, \Theta\} - \mathcal{A}_{\mathrm{num}}\{R(\beta), \Theta\} \right]; \\
Z_{\mathrm{den}}\{R(\beta), \Theta\} &= n_0^{-1/2} \textstyle\sum_{j=1}^n (1 - D_j) \left[G_{\mathrm{den}}\{R(\beta), X_j, \Theta\} - \mathcal{A}_{\mathrm{den}}\{R(\beta), \Theta\} \right].
\end{aligned}$$

Since by assumption $n_0/n_1 \to c, \;\; 0 < c < \infty$, we have that $Z_{\mathrm{num}}\{R(\beta), \Theta\} = O_p(1)$

and $Z_{\rm den}\{R(\beta),\Theta\} = O_p(1)$. Thus, by a Taylor series expansion

$$\begin{aligned}
&\frac{n_0^{-1}\sum_{j=1}^n (1-D_j)G_{\rm num}\{R(\beta),X_j\}}{n_0^{-1}\sum_{j=1}^n (1-D_j)G_{\rm den}\{R(\beta),X_j\}} - \frac{\mathcal{A}_{\rm num}\{R(\beta),\Theta\}}{\mathcal{A}_{\rm den}\{R(\beta),\Theta\}} \\
&= \frac{\mathcal{A}_{\rm num}\{R(\beta),\Theta\} + n_0^{-1/2}Z_{\rm num}\{R(\beta),\Theta\}}{\mathcal{A}_{\rm den}\{R(\beta),\Theta\} + n_0^{-1/2}Z_{\rm den}\{R(\beta),\Theta\}} - \frac{\mathcal{A}_{\rm num}\{R(\beta),\Theta\}}{\mathcal{A}_{\rm den}\{R(\beta),\Theta\}} \\
&= \frac{n_0^{-1/2}Z_{\rm num}\{R(\beta),\Theta\}}{\mathcal{A}_{\rm den}\{R(\beta),\Theta\}} - \frac{\mathcal{A}_{\rm num}\{R(\beta),\Theta\}}{\mathcal{A}^2_{\rm den}\{R(\beta),\Theta\}} n_0^{-1/2} Z_{\rm den}\{R(\beta),\Theta\} + o_p(n_0^{-1/2}).
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathcal{H}_n(\beta,\Theta) &= n_0^{-1}n^{-1/2}\sum_{i=1}^n\sum_{j=1}^n (1-D_j)\frac{G_{\rm num}\{R_i(\beta),X_j,\Theta\} - \mathcal{A}_{\rm num}\{R_i(\beta),\Theta\}}{\mathcal{A}_{\rm den}\{R_i(\beta),\Theta\}} \\
&\quad -n_0^{-1}n^{-1/2}\sum_{i=1}^n\sum_{j=1}^n \frac{\mathcal{A}_{\rm num}\{R_i(\beta),\Theta\}}{\mathcal{A}^2_{\rm den}\{R_i(\beta),\Theta\}} \\
&\qquad \times(1-D_j)\left[G_{\rm den}\{R_i(\beta),X_j,\Theta\} - \mathcal{A}_{\rm den}\{R_i(\beta),\Theta\}\right] + o_p(1) \\
&= \mathcal{D}_1 + \mathcal{D}_2 + o_p(1).
\end{aligned}$$

By definition, $E(\mathcal{D}_1|D_1,...,D_n) = E(\mathcal{D}_2|D_1,...,D_n) = 0$. Let $c_* = n/n_0$. By the definition of $G_{\rm num}$, etc.,

$$\mathcal{D}_1 + \mathcal{D}_2 = c_* n^{-3/2}\sum_{i=1}^n\sum_{j=1}^n W\{R_i(\beta),X_j,D_j,\Theta\},$$

Notice that $W(r,x,d=1,\Theta) = 0$. Without loss of generality, we can make the first n_0 observations be the controls, and the last $n-n_0$ observations be the cases. Then

$$\begin{aligned}
\mathcal{D}_1 + \mathcal{D}_2 &= c_* n^{-3/2}\sum_{i=1}^n\sum_{j=1}^{n_0} W\{R_i(\beta),X_j,D_j,\Theta\} \\
&= c_* n^{-3/2}\sum_{i=1}^{n_0}\sum_{j=1}^{n_0} W\{R_i(\beta),X_j,D_j,\Theta\} \\
&\qquad + c_* n^{-3/2}\sum_{i_0+1}^{n}\sum_{j=1}^{n_0} W\{R_i(\beta),X_j,D_j,\Theta\} \\
&= c_* n^{-3/2}\sum_{i=1}^{n_0}\sum_{j<i}^{n_0} Q_1\{\widetilde{Z}_i(\beta),\widetilde{Z}_j(\beta),\Theta) \\
&\qquad + c_* n_1 n^{-3/2}\sum_{j=1}^{n_0} n_1^{-1}\sum_{i_0+1}^{n} W\{R_i(\beta),X_j,D_j,\Theta\} + o_p(1).
\end{aligned}$$

An easy calculation shows that

$$\text{var}\left[n^{-3/2}\sum_{j=1}^{n_0}\sum_{i_0+1}^{n}W\{R_i(\beta),X_j,D_j,\Theta\}-n_1n^{-3/2}\sum_{i=1}^{n_0}Q_2\{\widetilde{Z}_i(\beta),\beta,\Theta\}\right]\to 0.$$

Hence we have shown that

$$\begin{aligned}\mathcal{D}_1+\mathcal{D}_2 &= c_*(n_0/n)^{3/2}n_0^{-3/2}\sum_{i=1}^{n_0}\sum_{j<i}^{n_0}Q_1\{\widetilde{Z}_i(\beta),\widetilde{Z}_j(\beta),\beta,\Theta\}\\ &\quad +c_*n_1n^{-3/2}\sum_{i=1}^{n_0}Q_2\{\widetilde{Z}_i(\beta),\beta,\Theta\}+o_p(1).\end{aligned}$$

Except for the factor $c_*(n_0/n)^{3/2}$, the first term above is a classical symmetric U-statistic of order two applied to independent and identically distributed observations, since by convention the first n_0 observations are the controls. It then follows from standard U-statistic results that

$$\begin{aligned}\mathcal{D}_1+\mathcal{D}_2 &= c_*(n_0/n)^{3/2}n_0^{-1/2}\sum_{i=1}^{n_0}h_1\{0,\widetilde{Z}_i(\beta),\beta,\Theta\}\\ &\quad +c_*n_1n^{-3/2}\sum_{i=1}^{n_0}Q_2\{\widetilde{Z}_i(\beta),\beta,\Theta\}+o_p(1)\\ &= c_*(n_0/n)n^{-1/2}\sum_{i=1}^{n}(1-D_i)h_1\{D_i,\widetilde{Z}_i(\beta),\beta,\Theta\}\\ &\quad +c_*(n_1/n)n^{-1/2}\sum_{i=1}^{n}(1-D_i)Q_2\{\widetilde{Z}_i(\beta),\beta,\Theta\}+o_p(1)\\ &= n^{-1/2}\sum_{i=1}^{n}h_2\{R_i(\beta),X_i,D_i,\Theta\}+o_p(1).\end{aligned}$$

This completes the sketch of Lemma 1.

3. Sketch of the Asymptotic Theory for $\widehat{\beta}$

Under the rare disease approximation, the estimate is consistent for β_{true}, and $\alpha(\beta_{\text{true}},\Omega_{\text{true}})=\alpha_{\text{true}}$. Define $\mathcal{M}_\Omega$, $\mathcal{T}\{R(\beta),X,\Theta,f_{X,\text{cont}}\}$, and $\mathcal{M}_\beta$ as in Section 6.

Define

$$\begin{aligned}
\mathcal{J}\{R(\beta), X, \beta, \Omega) &= \mu_\beta(X,\beta) - \frac{\int \mu_\beta(x,\beta)\mathcal{K}\{R(\beta),x,\beta,\Omega\}f_{X,\text{cont}}(x)dx}{\int \mathcal{K}\{R(\beta),x,\beta,\Omega\}f_{X,\text{cont}}(x)dx};\\
c_{1n}(\beta,\Omega) &= n^{-1}\textstyle\sum_{i=1}^n \mathcal{J}\{R_i(\beta), X_i, \beta, \Omega);\\
c_1(\beta,\Omega) &= E_{\text{cc}}[\mathcal{J}\{R(\beta),X,\beta,\Omega)\}.
\end{aligned}$$

We are solving $0 = \widehat{Q}_{n,\text{est}}(\widehat{\beta},\widehat{\Omega})$. By a Taylor series expansion,

$$\begin{aligned}
0 &= \widehat{Q}_{n,\text{est}}(\beta_{\text{true}},\Omega_{\text{true}}) + \frac{\partial}{\partial\beta^{\text{T}}}n^{-1/2}\widehat{Q}_{n,\text{est}}(\beta_{\text{true}},\Omega_{\text{true}})n^{1/2}(\widehat{\beta}-\beta_{\text{true}})\\
&\quad + \frac{\partial}{\partial\Omega^{\text{T}}}n^{-1/2}\widehat{Q}_{n,\text{est}}(\beta_{\text{true}},\Omega_{\text{true}})n^{1/2}(\widehat{\Omega}-\Omega_{\text{true}}) + o_p(1).
\end{aligned}$$

However, it is clear that for any (β,Ω), $n^{-1/2}\widehat{Q}_{n,\text{est}}(\beta,\Omega) = E_{cc}[\mathcal{T}\{R(\beta),X,\Theta,f_{X,\text{cont}}\}]$ $+o_p(1)$. Hence it follows that

$$0 = \widehat{Q}_{n,\text{est}}(\beta_{\text{true}},\Omega_{\text{true}}) + \mathcal{M}_\beta n^{1/2}(\widehat{\beta}-\beta_{\text{true}}) + \mathcal{M}_\Omega n^{1/2}(\widehat{\Omega}-\Omega_{\text{true}}) + o_p(1).$$

Because of its form,

$$\begin{aligned}
\widehat{Q}_{n,\text{est}}(\beta_{\text{true}},\Omega_{\text{true}}) &= \widehat{Q}_n(\alpha_{\text{true}},\beta_{\text{true}},\Omega_{\text{true}})\\
&\quad + c_1(\beta_{\text{true}},\Omega_{\text{true}})n^{1/2}\{\widehat{\alpha}(\beta_{\text{true}},\Omega_{\text{true}}) - \alpha(\beta_{\text{true}},\Omega_{\text{true}})\} + o_p(1).
\end{aligned}$$

However, under the rare disease approximation, when we replace $f_{\text{cont}}(\cdot)$ by $f_X(\cdot)$ in the definition of $\mathcal{J}(\cdot)$, by the same argument as in Section 1, $c_1(\beta_{\text{true}},\Omega_{\text{true}}) = 0$. In addition, using the same tools as in Lemma 1, $n^{1/2}\{\widehat{\alpha}(\beta_{\text{true}},\Omega_{\text{true}}) - \alpha(\beta_{\text{true}},\Omega_{\text{true}})\} = O_p(1)$. We have thus shown that

$$n^{1/2}(\widehat{\beta}-\beta_{\text{true}}) = -\mathcal{M}_\beta\left\{\widehat{Q}_n(\alpha_{\text{true}},\beta_{\text{true}},\Omega_{\text{true}}) + \mathcal{M}_\Omega n^{1/2}(\widehat{\Omega}-\Omega_{\text{true}})\right\} + o_p(1). \quad \text{(C.3)}$$

Remember that $\mathcal{K}(r,x,\Theta) = 1+\exp[\kappa + m\{r+\mu(x,\beta),x,\Omega\}]$. Define $\Phi(y,x,d,\Omega) =$

$\{1, m_\Omega(y,x,\theta_1)\}^{\mathrm{T}}[D - H\{\kappa + m(y,x,\theta_1)\}]$ and

$$\mathcal{N}_\Omega = -\left[E_{\mathrm{cc}}\left\{\partial\Phi(Y,X,D,\Omega)/\partial\Omega\right\}\right]^{-1}.$$

Because $\Omega = (\kappa, \theta_1)$ is estimated by ordinary logistic regression, it follows from standard theory that

$$n^{-1/2}(\widehat{\Omega} - \Omega_{\mathrm{true}}) = n^{-1/2}\textstyle\sum_{i=1}^n \mathcal{N}_\Omega \Phi(Y_i, X_i, D_i, \Omega_{\mathrm{true}}).$$

We thus have from (C.3) that

$$\begin{aligned} n^{1/2}(\widehat{\beta} - \beta_{\mathrm{true}}) &= -\mathcal{M}_\beta\Big\{\widehat{Q}_n(\alpha_{\mathrm{true}}, \beta_{\mathrm{true}}, \Omega_{\mathrm{true}}) \\ &\qquad + \mathcal{M}_\Omega n^{-1/2}\textstyle\sum_{i=1}^n \mathcal{N}_\Omega \Phi(Y_i, X_i, D_i, \Omega_{\mathrm{true}})\Big\} + o_p(1). \quad \text{(C.4)} \end{aligned}$$

We are now in a position to apply Lemma 1 to $\widehat{Q}_n(\alpha_{\mathrm{true}}, \beta_{\mathrm{true}}, \Omega_{\mathrm{true}})$. In order to apply Lemma 1, we define $G_{\mathrm{num}}(r,x,\Theta) = L\{r, x, \alpha(\beta,\Omega), \beta\}\mathcal{K}(r,x,\Theta)$ and $G_{\mathrm{den}}(r,x,\Theta) = \mathcal{K}(r,x,\Theta)$. Invoking Lemma 1, it follows that

$$\begin{aligned} \widehat{Q}_n(\beta_{\mathrm{true}}, \Theta_{\mathrm{true}}) &= n^{-1/2}\textstyle\sum_{i=1}^n \mathcal{T}\{R_i(\beta_{\mathrm{true}}), X_i, \Theta_{\mathrm{true}}, f_{X,\mathrm{cont}}) \\ &\qquad - n^{-1/2}\textstyle\sum_{i=1}^n h_2\{R_i(\beta_{\mathrm{true}}), X_i, D_i, \Theta_{\mathrm{true}}\} + o_p(1). \end{aligned}$$

We now make the rare disease assumption, and thus set $f_X(x) = f_{X,\mathrm{cont}}(x)$, and we have

$$\begin{aligned} \widehat{Q}_n(\beta_{\mathrm{true}}, \Theta_{\mathrm{true}}) &= n^{-1/2}\textstyle\sum_{i=1}^n \mathcal{T}\{R_i(\beta_{\mathrm{true}}), X_i, \Theta_{\mathrm{true}}, f_X) \\ &\qquad - n^{-1/2}\textstyle\sum_{i=1}^n h_2\{R_i(\beta_{\mathrm{true}}), X_i, D_i, \Theta_{\mathrm{true}}\} + o_p(1). \end{aligned}$$

We have shown in Section 1 that the first term has mean zero. That is, if

$$\mu_1(d) = E[\mathcal{T}\{R(\beta_{\mathrm{true}}), X, \Theta_{\mathrm{true}}, f_X)|D = d],$$

then $\sum_{i=1}^{n}\mu_1(D_i) = 0$. Hence we have shown that

$$\begin{aligned}\widehat{Q}_n(\beta_{\text{true}}, \Theta_{\text{true}}) &= n^{-1/2}\textstyle\sum_{i=1}^{n}\left[\mathcal{T}\{R_i(\beta_{\text{true}}), X_i, \Theta_{\text{true}}, f_X) - \mu_1(D_i)\right] \\ &\quad -n^{-1/2}\textstyle\sum_{i=1}^{n} h_2\{R_i(\beta_{\text{true}}), X_i, D_i, \Theta_{\text{true}}\} + o_p(1).\end{aligned}$$

Remember that $E[h_2\{R(\beta_{\text{true}}), X, D, \Theta_{\text{true}}\}|D] = 0$. Now let $\mu_4(d) = E\{\Phi(Y, X, D, \Omega_{\text{true}})|D = d\}$, and because of the unbiasedness of the estimating equation for logistic regression, $\sum_{i=1}^{n}\mu_4(D_i) = 0$. Summarizing, we have shown that

$$\begin{aligned}n^{-1/2}(\widehat{\beta} - \beta_{\text{true}}) &= -\mathcal{M}_\beta^{-1}(\Theta_{\text{true}})n^{-1/2}\textstyle\sum_{i=1}^{n}\Lambda(Y_i, X_i, D_i, \Theta_{\text{true}}) + o_p(1); \\ \Lambda(Y_i, X_i, D_i, \Theta_{\text{true}}) &= \mathcal{M}_\Omega(\Theta_{\text{true}})\mathcal{N}_\Omega(\Omega_{\text{true}})\{\Phi(Y_i, X_i, D_i, \Omega_{\text{true}}) - \mu_4(D_i)\} \\ &\quad -h_2\{R_i(\beta_{\text{true}}), X_i, D_i, \Theta_{\text{true}}\} \\ &\quad +\left[\mathcal{T}\{R_i(\beta_{\text{true}}), X_i, \Theta_{\text{true}}, f_X) - \mu_1(D_i)\right]; \\ 0 &= E[\Lambda(Y, X, D, \Theta_{\text{true}})|D],\end{aligned}$$

as claimed.

VITA

Name:	Jiawei Wei
Address:	Department of Statistics
	c/o Dr. Raymond J. Carroll
	Texas A&M University
	College Station, TX 77843-3122
Email Address:	wjw@stat.tamu.edu
Education:	B.S., Statistics, Zhejiang University, China, 2005
	M.S., Texas A&M University, USA, 2007
	Ph.D., Texas A&M University, USA, 2010