INSTITUT DE STATISTIQUE
BIOSTATISTIQUE ET
SCIENCES ACTUARIELLES
(I S B A)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION
PAPER

1033

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

WEI, J., CARROLL, R., MÜLLER, U., VAN KEILEGOM, I. and N. CHATTERJEE

Robust Estimation for Homoscedastic Regression in the Secondary Analysis of Case-Control Data

Jiawei Wei, Raymond J. Carroll, Ursula U. Müller
Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143
wjw@stat.tamu.edu, carroll@stat.tamu.edu, uschi@stat.tamu.edu

Ingrid Van Keilegom
Université catholique de Louvain
Voie du Roman Pays 20
B-1348 Louvain-la-Neuve, Belgium
ingrid.vankeilegom@uclouvain.be

Nilanjan Chatterjee
Division of Cancer Epidemiology and Genetics
National Cancer Institute, 6120 Executive Blvd, EPS 8038 Rockville MD 20852
chattern@mail.nih.gov

Abstract

Primary analysis of case-control studies focuses on the relationship between disease (D) and a set of covariates of interest (Y, X). A secondary application of the case-control study, often invoked in modern genetic epidemiologic association studies, is to investigate the interrelationship between the covariates themselves. The task is complicated due to case-control 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.

In this paper, we take up the issue of estimation of a regression function when Y given X follows a homoscedastic regression model, but otherwise the distribution of Y is unspecified. 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 empirical examples are used to illustrate the approach.

Some Key Words: Biased samples; Homoscedastic regression; Secondary data; Secondary phenotypes; Semiparametric inference; Two-stage samples;

Short title: Secondary Analysis of Case-Control Data

1 Introduction

Case-control designs are popularly used for studying risk-factors for rare diseases, such as cancers. Under this design, a fixed number of "cases" and "controls", i.e., subjects with and without the disease of interest, are sampled from an underlying base population. Data on various covariates on the subjects are then collected in a retrospective fashion so that they reflect history prior to the disease. The standard method for primary analysis of case-control data involves logistic regression modeling of the disease outcome as a function the covariates of interest. It is well known that such prospective logistic regression analysis for case-control data is efficient under a semiparametric framework that allows the "nuisance" distribution of the underlying covariates to be completely unspecified (Prentice and Pyke, 1979).

Epidemiologic researchers popularly use controls from case-control studies to examine the interrelationship between certain covariates themselves. Such secondary analysis of case-control studies have received increasing attention in genetic epidemiologic studies, where it is often of interest to investigate the effect of genetic susceptibility, such as SNP genotypes, not only on the primary disease outcome, but also on various secondary factors, such as smoking habits, that may themselves be associated with the disease of interest. For such secondary analysis, use of only controls is generally considered a robust approach since, when the disease is rare, the relationship between covariates in the controls should reflect that of the underlying population without any further model assumptions. It is, however, recognized that inclusion of cases in such analysis can increase efficiency provided appropriate adjustment can be made to account for non-random ascertainment in case-control sampling. Li et al. (2010), for example, noted that if two binary covariates have no interaction on the risk of the disease in a logistic scale, then the association between the factors in the cases remains the same as that for the underlying population and thus in such a setting inclusion of cases can dramatically increase the efficiency of the secondary analysis.

In this article, our goal is to develop an approach to secondary association analysis for a continuous covariate, say Y, in a case-control study setting so that both cases and controls can be used to increase efficiency and yet the resulting inference is robust to distributional assumptions about the covariates. 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 a vector of random covariates (Y, X), where Y is univariate and X is potentially multivariate, using a standard logistic regression formulation. Consider here the homoscedastic regression model

$$Y = \alpha_{\text{true}} + \mu(X, \beta_{\text{true}}) + \epsilon, \tag{1}$$

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 a predictor of disease status D, the sampling is biased and in the case-control sample model (1) will not hold.

This paper is organized as follows. In Section 2, 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 3 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 4 presents a series of simulation studies, while Section 5 presents analysis of two epidemiological data sets. Concluding remarks are in Section 6. Technical details are given in an appendix.

2 Efficient Parametric Estimation and Robustness

2.1 Framework

In this section we outline recent work on efficient estimation for case-control studies when the distribution of Y given X is specified up to a finite-dimensional parameter vector. 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 vector θ_1. For $d = 0, 1$, let $\pi_d = \text{pr}(D = d)$, and suppose there are n_1 cases with $D = 1$ and n_0 controls with $D = 0$.

We write $n = n_0 + n_1$ and introduce the parameter $\kappa = \theta_0 + \log(n_1/n_0) - \log(\pi_1/\pi_0)$. This reparameterization has the advantage that we can identify κ and θ_1, while we cannot identify θ_0, see, for example, Chatterjee and Carroll (2005).

Write the parametric model for Y given X as $f_\epsilon\{y - \alpha - \mu(x, \beta), \zeta\}$, where ζ is a finite-dimensional nuisance parameter. If in the population Y given X is normally distributed, then $\zeta = \mathrm{var}(\epsilon)$.

2.2 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. See also Monsees, et al. (2009). 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, with π_1 close to zero, which is typical for case-control sampling and the subject of this paper. For reasons of brevity we write $\Omega = (\kappa, \theta_1)$. It is easy to check that under the rare disease assumption the conditional density of D and Y given X is approximately

$$S_{\mathrm{par}}(d, y, x, \Omega, \alpha, \beta, \zeta) = f_\epsilon\{y - \alpha - \mu(x, \beta), \zeta\} \exp[d\{\kappa + m(y, x, \theta_1)\}]. \tag{2}$$

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

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

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 $L_{\mathrm{par}}(y, x, \alpha, \beta, \zeta) = \partial \log[f_\epsilon\{y - \alpha - \mu(x, \beta), \zeta\}]/\partial\beta$. Then the score for β is

$$\begin{aligned} \mathcal{K}_{\mathrm{par}}(D, Y, X, \Omega, \alpha, \beta, \zeta) &= \frac{\partial \log\{\mathcal{L}_{\mathrm{par}}(D, Y, X, \Omega, \alpha, \beta, \zeta)\}}{\partial\beta} \\ &= L_{\mathrm{par}}(Y, X, \alpha, \beta, \zeta) \\ &\quad - \frac{\int \sum_{d=0}^{1} L_{\mathrm{par}}(t, X, \alpha, \beta, \zeta) S_{\mathrm{par}}(d, t, X, \Omega, \alpha, \beta, \zeta) dt}{\int \sum_{d=0}^{1} S_{\mathrm{par}}(d, t, X, \Omega, \alpha, \beta, \zeta) dt}. \end{aligned} \tag{3}$$

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 (3) has mean zero under our rare disease assumption. Unfortunately, (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.

3 Model-Robust Estimation

3.1 Preliminaries

In this section we assume the same framework as in the previous section, with the exception that f_ϵ is now unknown. We pursue a sequential approach to derive an estimating equation for the parameters that determine the regression function.

- Estimate the true logistic regression parameters Ω_{true} by ordinary logistic regression of D on (Y, X). This can be done legitimately because it is known that ordinary logistic regression in a case-control study consistently estimates Ω_{true} (Chatterjee and Carroll, 2005). Call the result $\widehat{\Omega}$.

- Use a score function for β that would be an appropriate score function if the (Y, X) data arose from random sampling. Define $R(\beta) = Y - \mu(X, \beta)$. Then the simplest such score function is that from ordinary least squares, namely

$$L\{R(\beta), X, \alpha, \beta\} = \mu_\beta(X, \beta)\{R(\beta) - \alpha\}, \tag{4}$$

  where the subscript means differentiation with respect to β.

- The score (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 $\widehat{\alpha}(\beta, \Omega)$, which replaces α in the score (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.

Remark 1 The score function (4) is not the only one possible, of course, e.g., one could instead allow for robustness against outliers by replacing (4) by the estimating equation of an M-estimator with scale defined using the controls (Huber, 1981; Anderson, 2008).

3.2 The Estimation Algorithm

The development of our methodology is somewhat involved. Here we simply state our proposal, with its development given in the subsequent subsections 3.3-3.5. As before, define $R(\beta) = Y - \mu(X, \beta)$. Remember that estimation of α_{true} has to be done using an auxiliary equation, see (5) directly below. Define $\mathcal{K}\{R_i(\beta), x, \beta, \Omega\} = 1 + \exp[\kappa + m\{R_i(\beta) + \mu(x, \beta), x, \theta_1\}]$. For given (β, Ω), the estimator of α_{true} is justified in Section 3.5 and given by

$$\widehat{\alpha}(\beta, \Omega) = \frac{n^{-1}\sum_{i=1}^{n} R_i(\beta)\{n_0^{-1}\sum_{j=1}^{n}(1 - D_j)\mathcal{K}\{R_i(\beta), X_j, \beta, \Omega\}\}^{-1}}{n^{-1}\sum_{i=1}^{n}\{n_0^{-1}\sum_{j=1}^{n}(1 - D_j)\mathcal{K}\{R_i(\beta), X_j, \beta, \Omega\}\}^{-1}}. \tag{5}$$

Let $\mu_\beta(x, \beta) = \partial\mu(x, \beta)/\partial\beta$ and let $L\{R(\beta), X, \alpha, \beta\}$ be as in (4). Then define

$$\begin{aligned}\widehat{Q}_{n,\text{est}}(\beta, \Omega) \;=\; & n^{-1/2}\sum_{i=1}^{n}\Bigg[L\{R_i(\beta), X_i, \widehat{\alpha}(\beta, \Omega), \beta\} \\ & - \frac{n_0^{-1}\sum_{j=1}^{n}(1 - D_j)L\{R_i(\beta), X_j, \widehat{\alpha}(\beta, \Omega), \beta\}\mathcal{K}\{R_i(\beta), X_j, \beta, \Omega\}}{n_0^{-1}\sum_{j=1}^{n}(1 - D_j)\mathcal{K}\{R_i(\beta), X_j, \beta, \Omega\}}\Bigg].\end{aligned} \tag{6}$$

Our algorithm then is as follows.

- Estimate Ω by $\widehat{\Omega}$, the logistic regression estimates of D on (Y, X). As described previously, this is 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 (6), and at the end we describe the asymptotic distribution theory.

3.3 Development of the Score when f_X and α_{true} are Known

3.3.1 The Alternative Formulation of Case-Control Studies

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, $E_{\text{cc}}\{G(D,Y,X)\} = n^{-1}\sum_{i=1}^{n} E\{G(D_i,Y_i,X_i)|D_i\} = \sum_{d=0}^{1}(n_d/n)E\{G(D,Y,X)|D=d\}$, where G is an arbitrary function.

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 $\text{pr}(\delta = 1|D=d,Y,X) = n_d/(n\pi_d)$. Then, in this formulation, with a slight abuse of notation,

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

Now use this and the assumed independence between X and ϵ to write the regression model as

$$\begin{aligned}&\text{pr}(Y=y,X=x|\delta=1) \qquad (7)\\ &\quad=\frac{\sum_{d=0}^{1}(n_d/\pi_d)\text{pr}(D=d|Y=y,X=x\}f_\epsilon\{y-\alpha_{\text{true}}-\mu(x,\beta),\zeta\}f_X(x)}{\sum_{p=0}^{1}(n_p/\pi_p)\int \text{pr}(D=p|Y=t,X=v)f_\epsilon\{t-\alpha_{\text{true}}-\mu(v,\beta),\zeta\}f_X(v)dtdv}.\end{aligned}$$

Recall that $\kappa = \theta_0 + \log(n_1/n_0) - \log(\pi_1/\pi_0)$. Under the rare disease approximation, which we assume in this paper, we have

$$\begin{aligned}\textstyle\sum_{d=0}^{1}(n_d/\pi_d)\text{pr}(D=d|Y=y,X=x\} &\doteq (n_0/\pi_0)+(n_1/\pi_1)\exp\{\theta_0+m(y,x,\theta_1)\}\\ &= (n_0/\pi_0)[1+\exp\{\kappa+m(y,x,\theta_1)\}].\end{aligned}$$

The next step is to differentiate (7) with respect to β. Let $L_\epsilon\{R(\beta),X,\alpha_{\text{true}},\beta,\zeta)$ be the derivative with respect to β of $\log[f_\epsilon\{y-\alpha_{\text{true}}-\mu(x,\beta),\zeta\}]$. Then, except for a constant of

proportionality, the score function for β becomes

$$L_\epsilon\{R(\beta), X, \alpha_{\text{true}}, \beta, \zeta) \quad (8)$$
$$-\frac{\int L_\epsilon\{R(\beta), v, \alpha_{\text{true}}, \beta, \zeta) f_\epsilon\{t - \alpha_{\text{true}} - \mu(v, \beta)\}[1 + \exp\{\kappa + m(y, v, \theta_1)\}] f_X(v) dt dv}{\int f_\epsilon\{t - \alpha_{\text{true}} - \mu(v, \beta)\}[1 + \exp\{\kappa + m(t, v, \theta_1)\}] f_X(v) dt dv}.$$

This expression of course is of no value because we do not know a parametric model for $f_\epsilon(\cdot)$. We propose a two-step process, see the next two subsections.

3.3.2 Replacing the Underlying Score

The first step in the process is to replace $L_\epsilon(\cdot)$ in (8) by the least squares score when (Y, X) come from a random sample, namely (4). More generally, $L_\epsilon(\cdot)$ is replaced with a score function that yields consistent estimation when (Y, X) come from a random sample. With this substitution, (8) becomes

$$\mu_\beta(X, \beta)\{Y - \alpha_{\text{true}} - \mu(X, \beta)\}$$
$$-\frac{\int \mu_\beta(v, \beta)\{t - \alpha_{\text{true}} - \mu(v, \beta)\} f_\epsilon\{t - \alpha_{\text{true}} - \mu(v, \beta)\}[1 + \exp\{\kappa + m(y, v, \theta_1)\}] f_X(v) dt dv}{\int f_\epsilon\{t - \alpha_{\text{true}} - \mu(v, \beta)\}[1 + \exp\{\kappa + m(t, v, \theta_1)\}] f_X(v) dt dv}.$$

We now make the change of variables $R(\beta) = Y - \mu(X, \beta)$, and recall that $\mathcal{K}(r, x, \beta, \Omega) = 1 + \exp[\kappa + m\{r + \mu(x, \beta), x, \theta_1\}]$. Referring to (4), this means that the score for β in the alternative formulation of Chen et al. (2009) is

$$L\{R(\beta), X, \alpha_{\text{true}}, \beta\} - \frac{\int L\{t, x, \alpha_{\text{true}}, \beta\} \mathcal{K}(t, x, \beta, \Omega) f_\epsilon(t - \alpha_{\text{true}}) f_X(x) dt dx}{\int \mathcal{K}(t, x, \beta, \Omega) f_\epsilon(t - \alpha_{\text{true}}) f_X(x) dt dx}. \quad (9)$$

3.3.3 Replacing the Unknown Error Density

The problem with (9) of course is that we do not know the form of $f_\epsilon(\cdot)$, so that the score (9) cannot be implemented. Similarly to Chatterjee and Carroll (2005) and Spinka et al. (2005), we therefore replace $f_\epsilon(\cdot)$ by a nonparametric maximum likelihood estimator. The idea is to take the observed $R_i(\beta) = Y_i - \mu(X_i, \beta)$ as the support, and to maximize the loglikelihood with respect to $\gamma_i = \text{pr}\{R(\beta) = R_i(\beta)\}$, $i = 1 \ldots, n$. The resulting estimator for $\text{pr}\{R(\beta) = R_i(\beta)\}$ is

$$p_{\text{est}}\{R_i(\beta), \Omega\} \quad = \quad \frac{\pi_0}{n_0} \left\{ \int f_X(x) \mathcal{K}\{R_i(\beta), x, \beta, \Omega\} dx \right\}^{-1}.$$

Its derivation is given in Appendix A.1. When we make this substitution in (9) and sum over the data, the score becomes

$$\sum_{i=1}^{n} L\{R_i(\beta), X_i, \alpha_{\text{true}}, \beta\} - \frac{\sum_{i=1}^{n} \int L\{R_i(\beta), x, \alpha_{\text{true}}, \beta\} \mathcal{K}\{R_i(\beta), x, \beta, \Omega\} p_{\text{est}}\{R_i(\beta), \Omega\} f_X(x) dx}{n^{-1} \sum_{i=1}^{n} \int \mathcal{K}\{R_i(\beta), x, \beta, \Omega\} p_{\text{est}}\{R_i(\beta), \Omega\} f_X(x) dx}.$$

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

$$\begin{aligned} 0 &= Q_n(\alpha_{\text{true}}, \beta, \Omega) \qquad (10) \\ &= n^{-1/2} \sum_{i=1}^{n} \left[L\{R_i(\beta), X_i, \alpha_{\text{true}}, \beta\} - \frac{\int L\{R_i(\beta), x, \alpha_{\text{true}}, \beta\} \mathcal{K}\{R_i(\beta), x, \beta, \Omega\} f_X(x) dx}{\int \mathcal{K}\{R_i(\beta), x, \beta, \Omega\} f_X(x) dx} \right]. \end{aligned}$$

In Appendix A.2, we show that under the rare disease assumption, $E_{\text{cc}}\{Q_n(\alpha_{\text{true}}, \beta, \Omega)\} = 0$ when evaluated at $(\alpha_{\text{true}}, \beta_{\text{true}}, \Omega_{\text{true}})$, but not generally, and thus (10) is indeed an (approximately) unbiased estimating equation in the *case-control sampling* scheme, and not just an estimating equation obtained with the alternative approach that uses the conditional density (7).

3.4 Implementation when f_X 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 $f_{X,\text{cont}}(\cdot)$, the density of X among the controls. We then estimate the integrals in (10) unbiasedly by their averages among the controls, so that our estimating equation is

$$\begin{aligned} 0 &= \widehat{Q}_n(\alpha_{\text{true}}, \beta, \Omega) \qquad (11) \\ &= n^{-1/2} \sum_{i=1}^{n} \left[L\{R_i(\beta), X_i, \alpha_{\text{true}}, \beta\} - \frac{n_0^{-1} \sum_{j=1}^{n} (1 - D_j) L\{R_i(\beta), X_j, \alpha_{\text{true}}, \beta\} \mathcal{K}\{R_i(\beta), X_j, \beta, \Omega\}}{n_0^{-1} \sum_{j=1}^{n} (1 - D_j) \mathcal{K}\{R_i(\beta), X_j, \beta, \Omega\}} \right]. \end{aligned}$$

3.5 Implementation 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 $L\{R(\beta), X, \alpha, \beta\} = \{1, \mu_\beta^\top(X, \beta)\}^\top\{R(\beta) - \alpha\}$. The problem with this is that the first component of the estimating equation (11) would then be identically zero, and thus will not produce an estimate of the intercept. The reason for this is that the solution (10) was calculated nonparametrically under the assumption that $R(\beta_{\text{true}})$ and X are independent in the population. Since $Y - \alpha_{\text{true}} - \mu(X, \beta_{\text{true}})$ and $Y - \mu(X, \beta_{\text{true}})$ are both independent of X in the population, this means that (10) 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 (10), i.e., if $f_X(\cdot)$ were known, then α_{true} could be estimated by

$$\widetilde{\alpha}(\beta, \Omega) = \frac{n^{-1}\sum_{i=1}^n R_i(\beta) p_{\text{est}}\{R_i(\beta), \Omega\}}{n^{-1}\sum_{i=1}^n p_{\text{est}}\{R_i(\beta), \Omega\}}, \tag{12}$$

a quantity that is free of the π_0 that shows up in (10). 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 $n_0^{-1}\sum_{j=1}^n (1 - D_j)\mathcal{K}\{R_i(\beta), X_j, \beta, \Omega\}$, we get exactly (5). Making this substitution in (11), we obtain (6). This completes the derivation of our methodology.

3.6 Distribution Theory

Let $(\beta, \Omega) = \Theta$, 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

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

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

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

and

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

Also define

$$\begin{aligned} c_* &= \lim_{n\to\infty}(n/n_0); \\ \widetilde{Z}_i(\beta) &= \{R_i(\beta),X_i,D_i\}; \\ \widetilde{z} &= (r,x,d); \\ Q_1\{\widetilde{Z}_i(\beta),\widetilde{Z}_j(\beta),\Theta\} &= W\{R_i(\beta),X_j,D_j,\Theta\} + W\{R_j(\beta),X_i,D_i,\Theta\}; \\ Q_2(\widetilde{z},\beta,\Theta) &= E[W\{R(\beta),x,d,\Theta\}|D=1]; \\ h_1(d,\widetilde{z},\beta,\Theta) &= E[Q_1\{\widetilde{z},\widetilde{Z}(\beta),\Theta\}|D=d]; \\ h_2\{R_i(\beta),X_i,D_i,\Theta\} &= c_*(n_0/n)(1-D_i)h_1\{D_i,\widetilde{Z}_i(\beta),\beta,\Theta\} \\ &\qquad +c_*(n_1/n)(1-D_i)Q_2\{\widetilde{Z}_i(\beta),\beta,\Theta\}; \\ m_{\theta_1}(y,x,\theta_1) &= \partial m(y,x,\theta_1)/\partial\theta_1; \\ \Phi(y,x,d,\Omega) &= \{1,m_{\theta_1}(y,x,\theta_1)\}^{\rm T}[d - H\{\kappa + m(y,x,\theta_1)\}]; \\ \mathcal{N}_\Omega &= -\left[E_{\rm cc}\left\{\partial\Phi(Y,X,D,\Omega)/\partial\Omega\right\}\right]^{-1}\Big|_{\Omega=\Omega_{\rm true}}; \\ \mu_1(d) &= E[\mathcal{T}\{R(\beta_{\rm true}),X,\Theta_{\rm true},f_X\}|D=d]; \\ \mu_2(d) &= E\{\Phi(Y,X,D,\Omega_{\rm true})|D=d\}; \\ \Lambda(Y_i,X_i,D_i,\Theta_{\rm true}) &= \mathcal{M}_\Omega\mathcal{N}_\Omega\{\Phi(Y_i,X_i,D_i,\Omega_{\rm true}) - \mu_2(D_i)\} \\ &\qquad -h_2\{R_i(\beta_{\rm true}),X_i,D_i,\Theta_{\rm true}\} \\ &\qquad +\left[\mathcal{T}\{R_i(\beta_{\rm true}),X_i,\Theta_{\rm true},f_X\} - \mu_1(D_i)\right]. \end{aligned}$$

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 A.4. We assume that all terms in the above list of definitions exist. This means, in particular, that all differentiability conditions, all integrability conditions and all invertibility conditions are satisfied. In addition, we assume that $\widehat{Q}_{n,\text{est}}(\beta, \Omega)$ is twice continuously differentiable with respect to β and Ω.

Theorem 1 *Let $(\beta, \Omega) = \Theta$, 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, we can assume that the distribution of X in the population is the same as the distribution of X among the controls, $f_X = f_{X,\text{cont}}$. Then $E\{\Lambda(Y, X, D, \Theta_{\text{true}})|D\} = 0$ and*

$$n^{1/2}(\widehat{\beta} - \beta_{\text{true}}) \;\; = \;\; -n^{-1/2}\mathcal{M}_\beta^{-1}\textstyle\sum_{i=1}^{n}\Lambda(Y_i, X_i, D_i, \Theta_{\text{true}}) + o_p(1). \tag{13}$$

Therefore, $n^{1/2}(\widehat{\beta} - \beta_{\text{true}}) \to Normal(0, \Sigma)$,

$$\Sigma = \textstyle\sum_{d=0}^{1} c_*^{d-1}(1 - c_*^{-1})^d \mathcal{M}_\beta^{-1} cov\{\Lambda(Y, X, D, \Theta_{\text{true}})|D = d\}(\mathcal{M}_\beta^{-1})^{\text{T}}.$$

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

4 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 $\Omega = \{\kappa, \theta_1 = (\theta_{11}, \theta_{12})\}$, which of course 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 1, and are 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.

5 Empirical Examples

We used our methodology to investigate two examples described by Chen et al. (2008) and Maity, et al. (2009). The basic purpose of the analysis is to show that in realistic settings, our methodology leads to much more precise inference than regression using only the controls.

5.1 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 paper, D is the prostate cancer case-control status and Y is the level of 1,25-dihydroxy-vitamin D. Here X consists of three dummy variables for age groups, along with a genetic marker characterized as follows. 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 we combine the age dummy variables with each SNP separately, with a linear regression model, and we are interested in the estimate of the SNP effect.

The results are given in Table 2. 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. (2009) 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, the three SNP suggest gains in efficiency of 168%, 136% and 125%.

5.2 Colorectal Adenoma

Chen, et al. (2009) and Maity, et al. (2009) 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 is one of the following three measures of smoking status: (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. Of interest is the relationship of smoking with markers associated with the NAT2 gene, which is known to play an important role in detoxification of certain aromatic carcinogens in cigarette smoke, and hence possibly addiction to smoking. The cases and controls were genotyped for six known functional polymorphisms related to NAT2 acetylation activity. The genotype data were then used to construct diplotype information, i.e. the pair of haplotypes that the subjects carried along their pair of homologous chromosomes. Thus X combines age in years, gender, and the most common diplotype. Chen, et al. cite the expectation that the genetic marker is associated with smoking status.

The results are given in Table 3. We see a statistically significant protective effect 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, the three SNP suggest gains in efficiency of 95%, 143% and 148%.

6 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 predic-

tors or phenotypes Y and demographic, environmental and genetic factors. The homoscedastic regression model (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 (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 score function for the regression parameters, we have shown that the estimation is robust and makes no assumptions about the distribution of the regression errors ϵ, 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.

Acknowledgments

This paper represents part of the first author's Ph.D. dissertation at Texas A&M University. Wei and Carroll's research were supported by a grant from the National Cancer Institute (R37-CA057030). Carroll was also supported by Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST). Chatterjee's research was supported by a gene-environment initiative grant from the National Heart Lung and Blood Institute (RO1-HL091172-01) and by the Intramural Research Program of the National Cancer Institute. Müller was supported by a National Science Foundation grant (DMS-0907014). Van Keilegom gratefully acknowledges financial support from IAP research network P6/03 of the Belgian Government (Belgian Science Policy), and from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650.

References

Anderson, R. (2008). *Modern Methods for Robust Regression*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-152.

Carroll, R. J., Wang, C. Y. and Wang, S. (1995). Prospective analysis of logistic case-control studies. *Journal of the American Statistical Association*, 90, 157-169.

Chatterjee, N. and Carroll, R. J. (2005). Semiparametric maximum likelihood estimation in case-control studies of gene-environment interactions. *Biometrika*, 92, 399-418.

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.

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.

Huber, P. J. (1981). *Robust Statistics*. John Wiley and Sons.

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

Li, H., Gail, M. H., Berndt, S. and Chatterjee, N. (2010). Using cases to strengthen inference on the association between single nucleotide polymorphisms and a secondary phenotype in genome-wide association studies. *Genetic Epidemiology*, 34, 427-433. Published Online: Jun 25 2010 3:42PM DOI: 10.1002/gepi.20495.

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

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.

Monsees, G., Tamimi, R. and Kraft, P. (2009). Genomewide association scans for secondary traits using case-control studies. *Genetic Epidemiology*, 33, 717728.

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.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66, 403-411.

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.

Appendix

A.1 Derivation of the Error Density Estimator (10)

The key idea of the approach is to introduce discrete probabilities $\gamma_i = \mathrm{pr}\{R(\beta) = R_i(\beta)\}$, $i = 1, \ldots, n$, which yields

$$\mathrm{pr}(D = d) = \textstyle\sum_{i=1}^n \mathrm{pr}\{D = d | R(\beta) = R_i(\beta)\}\gamma_i,$$

and to work with the maximum likelihood estimates, i.e. with those γ_i that maximize the retrospective loglikelihood

$$\begin{aligned}
& \textstyle\sum_{i=1}^n \log \mathrm{pr}\{R(\beta) = R_i(\beta) | D = D_i\} \\
= \; & \textstyle\sum_{i=1}^n \log \dfrac{\mathrm{pr}\{R(\beta) = R_i(\beta)\}\mathrm{pr}\{D = D_i | R(\beta) = R_i(\beta)\}}{\mathrm{pr}(D = D_i)} \\
= \; & \textstyle\sum_{i=1}^n \log \sum_{k=1}^n \gamma_k 1\{R_i(\beta) = R_k(\beta)\} + \sum_{i=1}^n \log \dfrac{\mathrm{pr}\{D = D_i | R(\beta) = R_i(\beta)\}}{\sum_{k=1}^n \mathrm{pr}\{D = D_i | R(\beta) = R_k(\beta)\}\gamma_k}.
\end{aligned}$$

Taking the derivative with respect to γ_k, $k = 1, \ldots, n$, gives

$$\begin{aligned}
& \frac{\sum_{i=1}^n 1\{R_i(\beta) = R_k(\beta)\}}{\gamma_k} - \sum_{i=1}^n \frac{\mathrm{pr}\{D = D_i | R(\beta) = R_k(\beta)\}}{\sum_{k=1}^n \mathrm{pr}\{D = D_i | R(\beta) = R_k(\beta)\}\gamma_k} \\
& \quad = \gamma_k^{-1} - \sum_{i=1}^n \frac{\mathrm{pr}\{D = D_i | R(\beta) = R_k(\beta)\}}{\mathrm{pr}(D = D_i)} \\
& \quad = \gamma_k^{-1} - \sum_{d=0}^1 \mathrm{pr}\{D = d | R(\beta) = R_k(\beta)\}\frac{n_d}{\pi_d}.
\end{aligned}$$

Now set this equal to zero to obtain

$$\begin{aligned}
\gamma_k &= \Big[\textstyle\sum_{d=0}^1 \mathrm{pr}\{D = d | R(\beta) = R_k(\beta)\}\dfrac{n_d}{\pi_d}\Big]^{-1} \\
&= \Big[\int \textstyle\sum_{d=0}^1 \mathrm{pr}\{D = d | R(\beta) = R_k(\beta), X = x\} f_X(x)\, dx \dfrac{n_d}{\pi_d}\Big]^{-1} \\
&\doteq \Big[\int (n_0/\pi_0)\big(1 + \exp[\kappa + m\{R_i(\beta) + \mu(x, \beta), x, \theta_1\}]\big) f_X(x)\, dx\Big]^{-1}.
\end{aligned}$$

In the last step we used the rare disease assumption, as already carried out in Section 3.3. By definition of $\mathcal{K}$ this is the desired formula (10).

A.2 Unbiasedness of the Estimation Function (10)

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

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

In order to derive the conditional density given the disease state we use the fact that we assume a logistic model, $\text{pr}(D = 1|Y, X) = H\{\theta_0 + m(Y, X, \theta_1)\}$, 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 function of (Y, X) in the population. Then, with θ_0 and θ_1 denoting the true parameters,

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

It then follows that the density of (Y, X) given D is

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

Under the rare disease assumption this is approximately $\exp[d\{\theta_0 + m(y, x, \theta_1)\}] f_{YX}(y, x)\pi_d^{-1}$. Recall that $\kappa = \theta_0 + \log(n_1/n_0) - \log(\pi_1/\pi_0)$. The above expectation can now be computed as

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

The joint density 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,

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

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

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

It follows from (A.1) that

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

If $\beta = \beta_{\text{true}}$, since $\tau(x, \beta_{\text{true}}, \beta_{\text{true}}) = 0$, it follows directly that the last term is zero, and therefore $0 \doteq E_{\text{cc}}\{Q_n(\alpha_{\text{true}}, \beta_{\text{true}}, \Omega_{\text{true}})\}$. Hence $Q_n(\alpha_{\text{true}}, \beta, \Omega_{\text{true}}) = 0$ is an approximately unbiased estimating equation. If $\beta \neq \beta_{\text{true}}$, then in general we will have $0 \neq E_{\text{cc}}\{Q_n(\alpha_{\text{true}}, \beta, \Omega_{\text{true}})\}$.

A.3 A Technical Lemma

The following Lemma is used in our analysis, including for the intercept. Refer to the definitions before the statement of Theorem 1.

Lemma 1 *Suppose the conditions of Theorem 1 are satisfied. Let $n_0 \to \infty$ and $n_1 \to \infty$ such that $n_0/n_1 \to c$, with $0 < c < \infty$. Then*

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

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

Sketch of Proof: Define

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

Since by assumption $n_0/n_1 \to c, \quad 0 < c < \infty$, we have that $Z_{\text{num}}\{R(\beta),\Theta\} = O_p(1)$ and $Z_{\text{den}}\{R(\beta),\Theta\} = O_p(1)$. Thus, by a Taylor series expansion,

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

Thus,

$$\begin{aligned} \mathcal{H}_n(\beta,\Theta) &= c_*n^{-3/2}\{1+o(1)\}\Big(\textstyle\sum_{i=1}^{n}\sum_{j=1}^{n}(1-D_j)\dfrac{G_{\text{num}}\{R_i(\beta),X_j,\Theta\}-\mathcal{A}_{\text{num}}\{R_i(\beta),\Theta\}}{\mathcal{A}_{\text{den}}\{R_i(\beta),\Theta\}} \\ &\qquad -\textstyle\sum_{i=1}^{n}\sum_{j=1}^{n}\dfrac{\mathcal{A}_{\text{num}}\{R_i(\beta),\Theta\}}{\mathcal{A}^2_{\text{den}}\{R_i(\beta),\Theta\}} \\ &\qquad \times(1-D_j)\left[G_{\text{den}}\{R_i(\beta),X_j,\Theta\}-\mathcal{A}_{\text{den}}\{R_i(\beta),\Theta\}\right]\Big)+o_p(1) \\ &= (\mathcal{D}_1+\mathcal{D}_2)\{1+o_p(1)\}+o_p(1). \end{aligned}$$

By definition, $E(\mathcal{D}_1|D_1,...,D_n) = E(\mathcal{D}_2|D_1,...,D_n) = 0$. Recall that $c_* = \lim_{n,n_0\to\infty} n/n_0$. By the definition of $W\{R_i(\beta),X_j,D_j,\Theta\}$,

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

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,

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

An easy calculation shows that

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

Hence we have shown that

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

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

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

This completes the sketch of proof.

A.4 Sketch of the Asymptotic Theory for $\widehat{\beta}$

Under the rare disease approximation, because of the unbiasedness of the estimating function (11) and the fact that (12) is consistent and asymptotically normally distributed for α_{true} when evaluated at $(\beta_{\text{true}},\Omega_{\text{true}})$, 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 3.6. Set

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

We use the fact that $0=\widehat{Q}_{n,\text{est}}(\beta,\widehat{\Omega})\big|_{\beta=\widehat{\beta}}$. 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, since $\widehat{\alpha}(\beta_{\text{true}},\Omega_{\text{true}})$ is a consistent estimator for α_{true}, it is clear that $n^{-1/2}\frac{\partial}{\partial\beta^{\text{T}}}\widehat{Q}_{n,\text{est}}(\beta,\Omega_{\text{true}})\big|_{\beta=\beta_{\text{true}}}=\mathcal{M}_\beta+o_p(1)$ and $n^{-1/2}\frac{\partial}{\partial\Omega^{\text{T}}}\widehat{Q}_{n,\text{est}}(\beta_{\text{true}},\Omega)\big|_{\Omega=\Omega_{\text{true}}}=\mathcal{M}_\Omega+$

$o_p(1)$. 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 $f_{X,\text{cont}}(\cdot)$ by $f_X(\cdot)$ in the definition of $\mathcal{J}(\cdot)$, by the same argument as in Appendix A.2, $c_1(\beta_{\text{true}}, \Omega_{\text{true}}) = 0$. 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

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

Remember that $\mathcal{K}(r, x, \Theta) = 1 + \exp[\kappa + m\{r + \mu(x, \beta), x, \Omega\}]$. Because $\Omega = (\kappa, \theta_1)$ is estimated by ordinary logistic regression, it follows from standard theory that

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

We thus have from (A.3) that

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

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

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

We now make the rare disease assumption which allows us to replace $f_{X,\text{cont}}$ in the above formula by f_X. We have shown in Appendix A.2 that the first term has mean zero, i.e., $\sum_{i=1}^n \mu_1(D_i) = 0$. Hence we have

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

Remember that $E[h_2\{R(\beta_{\text{true}}), X, D, \Theta_{\text{true}}\}|D] = 0$. Because of the unbiasedness of the estimating equation for logistic regression, $\sum_{i=1}^{n} \mu_2(D_i) = 0$. Summarizing, we have shown that

$$\begin{aligned}
n^{1/2}(\widehat{\beta} - \beta_{\text{true}}) &= -\mathcal{M}_\beta^{-1} n^{-1/2} \textstyle\sum_{i=1}^{n} \Lambda(Y_i, X_i, D_i, \Theta_{\text{true}}) + o_p(1); \\
\Lambda(Y_i, X_i, D_i, \Theta_{\text{true}}) &= \mathcal{M}_\Omega \mathcal{N}_\Omega \{\Phi(Y_i, X_i, D_i, \Omega_{\text{true}}) - \mu_2(D_i)\} \\
&\qquad -h_2\{R_i(\beta_{\text{true}}), X_i, D_i, \Theta_{\text{true}}\} \\
&\qquad + [\mathcal{T}\{R_i(\beta_{\text{true}}), X_i, \Theta_{\text{true}}, f_X) - \mu_1(D_i)]\,; \\
0 &= E[\Lambda(Y, X, D, \Theta_{\text{true}})|D],
\end{aligned}$$

as claimed.

	Mean	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(0.4,1.91)	-0.02	0.13	0.14	0.91	0.96	1.30
Gamma(0.8,1.91)	-0.01	0.13	0.13	0.91	0.96	1.42
Gamma(1.4,1.91)	-0.01	0.12	0.12	0.89	0.95	1.51
Gamma(1.8,1.91)	-0.01	0.12	0.12	0.91	0.95	1.62

Table 1: Simulation study for the estimation of β_1. Displayed is the mean of estimates across the simulation, their standard deviation, the mean of the estimated standard deviation from the bootstrap and the coverage probability of nominal 90% and 95% confidence intervals. Six distributions for the mean of Y given X were used, namely Normal(0,1), Chisquared(7), Gamma(0.4,1.91), Gamma(0.8,1.91), Gamma(1.4,1.91) and Gamma(1.8,1.91). Also displayed is the mean squared error efficiency ("MSE Efficiency") of our method compared to linear regression among the controls only.

	Our Method			Controls Only			
X	Estimate	Lower limit	Upper limit	Estimate	Lower limit	Upper limit	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

Table 2: Results of the VDR data example in Section 5.1. Three analyses are displayed, one each when X is SNP-1, SNP-2 and SNP-3, respectively. Displayed are the parameter estimates of the slope for X, ("Estimate"), and lower ("Lower") and upper ("Upper") 95% bootstrap confidence intervals. Our method is contrasted with using linear regression among the controls only. Also displayed are the mean squared error efficiencies of our method ("MSE Efficiency") calculated by the square of the ratio of the lengths of the confidence intervals.

	Our Method			Controls Only			
Y	Estimate	Lower limit	Upper limit	Estimate	Lower limit	Upper limit	MSE Efficiency
CIG STOP	-3.501	-5.716	-1.318	-3.240	-6.307	-0.173	1.95
PACK YEARS	-0.040	-0.199	0.120	0.210	-0.039	0.458	2.43
PACK DAY	0.063	-0.058	0.184	0.135	-0.047	0.317	2.48

Table 3: Results of the NAT2 data example in Section 5.2. Three analyses are displayed, one each when Y is years since stopping smoking ('CIG STOP'), number of packs smoked per day ('PACK DAY') and pack years ('PACK YEARS'). Displayed are the parameter estimates of the slope for X, ('Estimate') and lower ('Lower') and upper ('Upper') limits of the 95% bootstrap confidence intervals. Our method is contrasted with linear regression among the controls only. Also displayed are the mean squared error efficiencies of our method ('MSE Efficiency') calculated by the square of the ratio of the lengths of the confidence intervals.