Harvard University

Harvard University Biostatistics Working Paper Series

Year 2013 *Paper* 155

A General Regression Framework for a Secondary Outcome in Case-control Studies

Eric J. Tchetgen Tchetgen[*]

[*]Harvard School of Public Health, etchetge@hsph.harvard.edu

Original Manuscript

A general regression framework for a secondary outcome in case-control studies

Eric J. Tchetgen Tchetgen

Departments of Biostatistics and Epidemiology, Harvard University

Corresponding author: Eric J. Tchetgen Tchetgen, Department of Epidemiology, Harvard School of Public Health 677 Huntington Avenue, Boston, MA 02115.

Abstract

Modern case-control studies typically involve the collection of data on a large number of outcomes, often at considerable logistical and monetary expense. These data are of potentially great value to subsequent researchers, who, although not necessarily concerned with the disease that defined the case series in the original study, may want to use the available information for a regression analysis involving a secondary outcome. Because cases and controls are selected with unequal probability, regression analysis involving a secondary outcome generally must acknowledge the sampling design. In this paper, the author presents a new framework for the analysis of secondary outcomes in case-control studies. The approach is based on a careful re-parametrization of the conditional model for the secondary outcome given the case-control outcome and regression covariates, in terms of (a) the population regression of interest of the secondary outcome given covariates, and (b) the population regression of the case-control outcome on covariates. The error distribution for the secondary outcome given covariates and case-control status is otherwise unrestricted. For a continuous outcome, the approach sometimes reduces to extending model (a) by including a residual of (b) as a covariate. However, the framework is general in the sense that models (a) and (b) can take any functional form, and the methodology allows for an identity, log or logit link function for model (a).

1 Introduction

Case-control studies typically collect information on a large number of outcomes, often at considerable cost. These data are of potentially great value for studying associations, involving a secondary outcome other than the disease outcome defining case-control status. For instance, secondary outcomes analyses are now routine in genetic epidemiology, with several recent papers on genetic variants influencing human quantitative traits such as height, body mass index and lipid levels, using data mostly from case-control studies of complex diseases (diabetes, cancer and hypertension) (Lettre et a, 2008, Loos et al, 2008, Sanna et al, 2008, Weedon et al, 2007). Other examples have emerged in environmental epidemiology, such as the recent study of Weuve et al (2009), which uses data taken, in part, from a case-control study nested within the Nurses' Health Study (NHS). In the NHS Lead Study, Boston-area NHS participants had extensive lead exposure assessment (bone and blood measures). Associations of lead measures with hypertension, bone mineral density/metabolism, and cognition were then assessed. However, the Lead Study selected women on the basis of their blood pressure status. Therefore, analyses that aim to evaluate risk factors of osteoporosis (a binary outcome) and cognitive function decline (a continuous outcome), may be affected by the case-control sampling design. In fact, Monsees et al (2009) and Lin and Zeng (2009) established that the non-random ascertainment from the study base, when ignored, can sometimes lead to inflated Type I error rate for tests of associations of a secondary outcome in re-purposed case-control samples. They further showed that commonly used analytic techniques, such as least-squares regression for quantitative traits, can sometimes give biased estimates, and that such bias can be present when covariates in the regression model in view, are associated with case-control outcome, which itself is independently associated with the secondary outcome.

A number of analytic strategies have been proposed to eliminate selection bias associated with oversampling of cases in analyses of secondary outcomes, see for instance Nagelkerke et al (1995), Lee et al (1997), Jiang et al (2006), Reilly et al (2005), Richardson et al (2007), Lin and Zeng (2009), Monses et al (2009), Li et al (2010), Wang and Shele (2011) and Wei et al (2013). Suggested strategies include:

(i) weighting the standard analysis by the inverse of sampling probabilities;

(ii) performing the analysis only in controls;

(iii) analyzing cases and controls separately, i.e., stratifying the analysis by case-control status;

(iv) including case-control status as a covariate in the regression model of the secondary outcome.

The first strategy (i) gives a viable simple solution as it recovers correct inferences about association measures, without the burden of additional modelling than would be required had data been sampled independently of case-control status. However, simply weighting by sampling rates will often be inefficient (Robins et al, 1994, Tchetgen Tchetgen, 2012). The second method is appropriate only when the disease status is rare in the population but does not use data on cases and therefore may be inefficient. Methods that adjust for the primary disease status by either (iii) or (iv) may yield flawed conclusions because the associations between a secondary outcome and an exposure of interest in the case and control groups can be quite different from the association in the underlying target population. More formal likelihood methods have also appeared in the literature. For instance:

(v) Jiang et al (2006) considered various likelihood methods for categorical secondary outcomes that can be more efficient than (i).

(vi) Recently, Lin and Zeng (2009) further generalized the likelihood framework for a continuous secondary outcome by assuming the latter follows a specific parametric distribution.

They also establish that the likelihood approach reduces to (iv) approximately under the following assumptions:

(LZ.1) a rare disease assumption about the disease outcome defining case-control status;

(LZ.2) no interaction between the secondary outcome and covariates in a regression model for the case-control outcome;

(LZ.3) the secondary outcome is normally distributed.

Thus, Lin and Zeng (2009) formally justify via a maximum likelihood argument, the conditional approach (iv) in settings where (LZ.1)-(LZ.3) hold. More recently,

(vi) Wei et al (2013) develop an estimating equations approach for a continuous secondary outcome that relaxes the distributional assumption made in (v) somewhat, and instead requires that the secondary outcome regression is "strongly homoscedastic" in the following sense. They assume that residuals from the secondary outcome regression are independent of covariates. In other words, they suppose that any association between the vector of covariates and the secondary outcome is completely captured by a location shift model. Their inferential framework relies crucially on this assumption, and may not be consistent if the assumption does not hold exactly.

In this paper, the author generalizes the conditional approach to allow for possible violation of any or all of assumptions (LZ.1)-(LZ.3), without assuming the location shift model of Wei et al (2013). The new approach is based on a careful nonparametric re-parametrization of the conditional model for the secondary outcome given the case-control outcome and regression covariates, in terms of (a) the population regression of interest for the secondary outcome given covariates, and (b) the population regression of the case-control outcome on covariates. As nonparametric models may not be feasible in settings with numerous covariates, parametric and semiparametric models are invariably used in practice for (a) or (b). The re-parametrization ensures models for (a) and (b) are variation independent, in the sense that a parametric or semiparametric model for (a) does not restrict the model used for (b) and vice-versa. The error distribution for the secondary outcome given covariates and case-control status is otherwise unrestricted. In the case of a continuous outcome, a simple version of the approach entails extending model (a) by including a residual of (b) into the regression model as a covariate which gives a conditional regression model given case-control status directly parametrized in terms of model (a). We show such a reparametrization appropriately accounts for selection bias without compromising inference about the population regression parameter. The framework is general in the sense that models (a) and (b) can take any functional form, and the methodology is developed to allow an identity, log or logit link function for model (a). For inference, a simple estimating equations framework is first developed, and a strategy for obtaining a semiparametric locally efficient estimator is subsequently described. Simulations and an empirical example are used to illustrate the approach.

2 Regression with an identity link function

2.1 Reparametrization of conditional regression function

Suppose one observes i.i.d case-control data consisting of case-control status D, a continuous secondary outcome Y, and covariates $\mathbf{X}$. Unless otherwise stated, assume that the sampling fractions for cases and controls are known, that is, similar to a number of previous papers (e.g. Jiang et al, 2006, Lin and Zeng, 2009 and Wei et al, 2013), we shall assume that disease prevalence is known to be $\overline{p} = \Pr(D = 1)$ in the target population, and $\overline{\pi} = \Pr(D = 1|S = 1)$ in the case-control sample, where S indicates inclusion into the case-control study. Formally, $\overline{\pi}$ may be taken as the limit of the proportion of cases in the case-control study as sample size grows to infinity. As seen later, the assumption that $\overline{p}$ is known is not needed when the disease is rare in the population within all levels of $\mathbf{X}$. The primary target of inference is the population mean model $\mu(\mathbf{X}) = E(Y|X)$. Likewise, let $\widetilde{\mu}(\mathbf{X}, D) = E(Y|\mathbf{X}, D) = E(Y|\mathbf{X}, D, S = 1)$ where the second equality holds since by design, membership into the case-control study is independent of $(Y, \mathbf{X})$ given D. Then, the following relation between $\mu(\mathbf{X})$ and $\widetilde{\mu}(\mathbf{X}, D)$ holds:

$$\begin{aligned}
\mu(\mathbf{X}) &= \widetilde{\mu}(\mathbf{X}, 1) \Pr(D = 1|\mathbf{X}) + \widetilde{\mu}(\mathbf{X}, 0) \Pr(D = 0|\mathbf{X}) \\
&\Leftrightarrow \begin{cases} \widetilde{\mu}(\mathbf{X}, 1) = \mu(\mathbf{X}) + (1 - \Pr(D = 1|\mathbf{X})) \{\widetilde{\mu}(\mathbf{X}, 1) - \widetilde{\mu}(\mathbf{X}, 0)\} \\ \widetilde{\mu}(\mathbf{X}, 0) = \mu(\mathbf{X}) + (0 - \Pr(D = 1|\mathbf{X})) \{\widetilde{\mu}(\mathbf{X}, 1) - \widetilde{\mu}(\mathbf{X}, 0)\} \end{cases} \\
&\Leftrightarrow \widetilde{\mu}(\mathbf{X}, D) = \mu(\mathbf{X}) + \{D - \Pr(D = 1|\mathbf{X})\} \{\widetilde{\mu}(\mathbf{X}, 1) - \widetilde{\mu}(\mathbf{X}, 0)\} \\
&= \mu(\mathbf{X}) + \{D - p(\mathbf{X})\} \gamma(\mathbf{X}) \qquad (1)
\end{aligned}$$

where $\gamma(\mathbf{X}) \equiv \{\widetilde{\mu}(\mathbf{X}, 1) - \widetilde{\mu}(\mathbf{X}, 0)\}$ describes the association between Y and D on the mean difference scale, within levels of $\mathbf{X}$, and $p(\mathbf{X}) \equiv \Pr(D = 1|\mathbf{X})$ is the population risk of D within levels of $\mathbf{X}$. Thus, one learns that the conditional mean function $\widetilde{\mu}(\mathbf{X}, D)$ can be directly parametrized in terms of the population regression function of interest $\mu(\mathbf{X})$, and the additional functions $\{p(\mathbf{X}), \gamma(\mathbf{X})\}$. These latter functions directly encode the selection bias due to an association between D and Y within levels of $\mathbf{X}$. Note that the proposed reparametrization is nonparametric and variation independent, and therefore does not a priori rule out any possible data generating mechanism. The reparametrization shows that the marginal and conditional regressions of Y on $\mathbf{X}$ coincide exactly when selection bias is absent on the additive scale, i.e. when $\gamma(\mathbf{X}) \equiv 0$, and further guarantees that even when $\gamma(\mathbf{x})$ is not zero for at least one level of $\mathbf{x}$, upon marginalization over D in the underlying population $\widetilde{\mu}(\mathbf{X}, D)$ reduces to $\mu(\mathbf{X})$ exactly. Furthermore, we also learn from the reparametrization that when:

(ETT.1) $\gamma(\mathbf{X}) = \gamma$ does not vary with $\mathbf{X}$, and

(ETT.2) the disease is rare in the population, such that $\widetilde{\mu}(\mathbf{X}, D = 0) \approx \mu(\mathbf{X})$ and $\widetilde{\mu}(\mathbf{X}, D = 1) = \mu(\mathbf{X}) + \{1 - p(\mathbf{X})\}\gamma \approx \mu(\mathbf{X}) + \gamma$,

one obtains $\widetilde{\mu}(\mathbf{X}, D) \approx \mu(\mathbf{X}) + D\gamma$, which implies that simply extending the population model of interest $\mu(\mathbf{X})$ by adding the main effect for D in order to adjust for case-control sampling is approximately correct. Although this approximate conditional regression is identical to that obtained by Lin and Zeng (2009), one should note that while their assumptions (LZ.1)-(LZ.3) imply assumptions (ETT.1) and (ETT.2), the converse is not generally true. Specifically, it is straightforward to verify that assumptions (LZ.2) and (LZ.3) imply the no-heterogeneity assumption (ETT.2). However, without the normality assumption, (LZ.2) and (ETT.2) are not necessarily equivalent. The appeal of (ETT.2) is that it does not require making any distributional assumption about the secondary outcome. One should finally note that (LZ.2) and (ETT.2) are empirically testable, and can be relaxed to account for possible effect heterogeneity. Specifically, as shown in the next section, (ETT.2) may be relaxed by modeling $\gamma(\mathbf{X})$, which lead to the following approximation $\widetilde{\mu}(\mathbf{X}, D) \approx \mu(\mathbf{X}) + D\gamma(\mathbf{X})$. Note that in this last approximation, possible interactions between D and $\mathbf{X}$ encoded in $\gamma(\mathbf{X})$ are not directly interpretable as part of the targeted marginal association between $\mathbf{X}$ and Y, only the first term of the expression encodes the marginal association of interest.

2.2 Inference via simple estimating equations

Next, let $\pi(\mathbf{X}) \equiv \Pr(D = 1|\mathbf{X}, S = 1)$ denote the risk function of D within levels of $\mathbf{X}$ in the case-control sample. $\pi(\mathbf{X})$ and $p(\mathbf{X})$ are well known to satisfy the following relation:

$$\text{logit}p(\mathbf{X}) = \text{logit}\pi(\mathbf{X}) + \log \frac{\overline{p}\,(1-\overline{\pi})}{\overline{\pi}\,(1-\overline{p})},$$

so that population and the case-control risks of D agree on the logit scale, up to a constant shift in the intercept. Next, suppose that the mean function $\mu(\mathbf{X})$ follows a parametric model $\mu(\mathbf{X};\beta_0)$, where $\mu(\cdot;\beta)$ is a known function with unknown parameter β_0 the main target of inference. A standard multiple linear regression might take $\mu(\mathbf{X};\beta) = (1, \mathbf{X}')\beta$, but more general functional forms could be specified involving interactions and nonlinear terms. Further suppose that $\pi(\mathbf{X})$ follows a logistic model

$$\text{logit}\pi(\mathbf{X};\psi_0,\eta_0) = \eta_0 + m\left(\mathbf{X};\psi_0\right), \tag{2}$$

where $m\left(\cdot;\psi\right)$ is a known function indexed by a parameter ψ satisfying $m\left(0;\psi\right) = 0$, with unknown intercept η_0 and slope ψ_0. Thus, $\text{logit}p(\mathbf{X};\eta_0,\psi_0) = m\left(\mathbf{X};\psi_0\right) + \eta_0 + \log \frac{\overline{p}(1-\overline{\pi})}{\overline{\pi}(1-\overline{p})}$. A standard logistic regression model might take the form $m\left(\mathbf{X};\psi\right) = \psi'\mathbf{X}$, but more general functional forms could be used. Finally, suppose $\gamma\left(\mathbf{X};\alpha_0\right)$ is used to model $\gamma\left(\mathbf{X}\right)$, with $\gamma\left(\cdot;\alpha\right)$ a known function, and unknown parameter α_0. A standard linear model might take $\gamma\left(\mathbf{X};\alpha\right) = \left(1, \mathbf{X}'\right)\alpha$, but again, more general functional forms could be considered. Together, these various models produce a corresponding model for $\widetilde{\mu}(\mathbf{X}, D)$:

$$\widetilde{\mu}(\mathbf{X}, D;\theta_0) = \mu(\mathbf{X};\beta_0) + \left\{D - p(\mathbf{X};\psi_0,\eta_0)\right\}\gamma\left(\mathbf{X};\alpha_0\right), \tag{3}$$

where $\theta_0 = \left(\beta_0', \eta_0, \psi_0', \alpha_0'\right)'$.

Given n i.i.d samples on $(Y, \mathbf{X}, D)$, we propose to estimate (η_0, ψ_0') by standard maximum likelihood for the logistic regression model (2) using data $(\mathbf{X}, D)$,i.e. by maximizing $\mathbb{P}_n L(\psi_0, \eta_0)$ wrt (η_0, ψ_0') where $L(\psi_0, \eta_0) = D\text{logit}\pi(\mathbf{X}; \psi_0, \eta_0) + \log(1 - \pi(\mathbf{X}; \psi_0, \eta_0))$ and $\mathbb{P}_n(\cdot) = n^{-1}\sum_i (\cdot)_i$. Let $\varepsilon(\theta_0) = Y - \widetilde{\mu}(\mathbf{X}, D; \theta_0)$. Then, we propose to estimate (β_0', α_0') , with $\left(\widehat{\beta}', \widehat{\alpha}'\right)$ which solves $\mathbf{W}\left(\widehat{\theta}\right) = \mathbb{P}_{\mathbf{n}}\mathbf{U}\left(\widehat{\theta}\right) = 0$, where :

$$\mathbf{U}(\theta) = \frac{\partial \widetilde{\mu}(\mathbf{X}, D; \theta)}{\partial (\beta', \alpha')'} \varepsilon(\theta) \tag{4}$$

Note that for the following standard models, $m(\mathbf{X}; \psi) = \psi'\mathbf{X}, \gamma(\mathbf{X}; \alpha) = (1, \mathbf{X}')\alpha$ and $\mu(\mathbf{X}; \beta) = (1, \mathbf{X}')\beta$, one obtains

$$\mathbf{U}(\theta) = (1, \mathbf{X}', (1, \mathbf{X}')\{D - p(\mathbf{X}; \psi, \eta)\})' \varepsilon(\theta)$$

where

$$\varepsilon(\theta) = Y - (1, \mathbf{X}')\beta - \{D - p(\mathbf{X}; \psi, \eta)\}(1, \mathbf{X}')\alpha.$$

Further note that in general, for estimation the analyst could in principle specify any vector $\mathbf{h}(\mathbf{X}, D, \theta)$ of dimension $\dim((\beta_0', \alpha_0')')$ in place of $\partial\widetilde{\mu}(\mathbf{X}, D; \theta_0)/\partial\left(\beta', \alpha'\right)'$ in (4) , to obtain $\mathbf{U}(\theta, \mathbf{h}) = \mathbf{h}(\mathbf{X}, D, \theta)\varepsilon(\theta)$ provided the derivative of the resulting estimating equation, more precisely its expectation, is not singular, and the variance-covariance matrix of $\mathbf{U}(\theta, \mathbf{h})$ is finite. One can also verify using the proposition given in Section 5, that assuming $p(\mathbf{X})$ is known, the optimal choice of $\mathbf{h}$ is $\mathbf{h}_{opt}(\mathbf{X}, D, \theta) = \frac{\partial\widetilde{\mu}(\mathbf{X}, D; \theta)}{\partial\left(\beta', \alpha'\right)'} var(\varepsilon(\theta)|\mathbf{X}, D)^{-1}$, and therefore $\mathbf{U}(\theta, \mathbf{h}_{opt})$ would be optimal, in the sense of producing an estimator with minimal asymptotic variance among regular and asymptotically linear estimators (RAL), when $\varepsilon(\theta)$ is homoscedastic and $p(\mathbf{X})$ is known. A standard argument shows that under usual regularity assumptions, the resulting estimator $\widehat{\theta}$ is in large sample approximately:

$$\widehat{\theta} \dot{\sim} N\left(\theta_0, n^{-1}\Sigma(\theta_0)\right) \tag{5}$$

where $\Sigma(\theta)$ is the variance-covariance matrix of

$$\mathbb{E}\left[\partial\left(\mathbf{U}'(\theta),\mathbf{S}'(\psi,\eta)\right)/\partial\theta\right]^{-1}\times\left(\mathbf{U}'(\theta),\mathbf{S}'(\psi,\eta)\right)'$$

with $\mathbf{S}(\psi,\eta)=\partial L(,\eta)/\partial\left((\psi',\eta')'\right)$.

3 Regression with a Log link function

Here we give a generalization of the results presented in the previous section by considering regression analysis for a nonnegative outcome $Y \geq 0$ using a log-link function. In order to account for the retrospective sampling design, we again condition on case-control status in the regression model, while simultaneously obtaining inferences about a regression model that averages over disease status in the underlying population. To proceed, we now give a reparametrization of the mean function $E(Y|\mathbf{X},D)=\widetilde{\mu}(\mathbf{X},D)$ on the multiplicative scale, in terms of the population regression function of interest $\mu(\mathbf{X})=E(Y|\mathbf{X})$. One notes that:

$$\begin{aligned}E(Y|\mathbf{X},D) &= \frac{E(Y|\mathbf{X},D)}{E(Y|\mathbf{X})}\times E(Y|\mathbf{X}) \\ &= \frac{E(Y|\mathbf{X},D)}{E(Y|\mathbf{X},D=0)}\times\left\{\sum_{d^*=0}^{1}\frac{E(Y|\mathbf{X},D=d^*)}{E(Y|\mathbf{X},D=0)}\Pr(D=d^*|\mathbf{X})\right\}^{-1}\times E(Y|\mathbf{X}) \\ &= \exp\left[\log\mu(\mathbf{X})+\nu(\mathbf{X},D)-\overline{\nu}(\mathbf{X})\right]\end{aligned}$$

where $\nu(\mathbf{X},D)=\log E(Y|\mathbf{X},D)/E(Y|\mathbf{X},D=0)$ measures the multiplicative association between D and Y within levels of $\mathbf{X}$, and accounts for possible selection bias due to the retrospective sampling design. The term $\overline{\nu}(\mathbf{X})=\log\left\{\nu(\mathbf{X},D=1)\Pr(D=1|\mathbf{X})+\Pr(D=0|\mathbf{X})\right\}$ ensures that upon marginalization over D in the target population, the conditional mean function $E(Y|\mathbf{X},D)$

reduces exactly to $E(Y|\mathbf{X})$. As in the case of the identity link, we emphasize that the proposed reparametrization is completely nonparametric and variation independent, and therefore, except for the restriction that $E(Y|\mathbf{X},D) \geq 0$, does not a priori rule out any data generating mechanism.

A simplification occurs when D is rare in the population. Then, one observes that $E(Y|\mathbf{X}, D = 0) \approx E(Y|\mathbf{X})$ and therefore $\overline{\nu}(\mathbf{X}) \approx 1$, which gives $E(Y|\mathbf{X}, D = 1) \approx \exp\left[\log \mu(\mathbf{X}) + \nu(\mathbf{X}, 1)\right]$. Therefore $E(Y|\mathbf{X}, D) \approx \exp\left\{\log \mu(\mathbf{X}) + \nu(\mathbf{X}, 1)D\right\}$. Note here again, that only the first term on the exponential scale can be interpreted as an association measure between $\mathbf{X}$ and Y in the target population, any interaction between $\mathbf{X}$ and D encoded in the second term of the expression does not have such an interpretation. Under the assumption that the multiplicative association between D and Y is constant across levels of $\mathbf{X}$, simply adding the main effect for D to the population model of interest to obtain $E(Y|\mathbf{X}, D) \approx \exp\left\{\log \mu(\mathbf{X}) + \nu D\right\}$ is approximately correct.

Suppose that $\mu(\mathbf{X})$ follows a parametric model of the form $\exp\left\{t(\mathbf{X}; \beta_0)\right\}$ where $t(\cdot; \beta_0)$ is known up to the parameter of primary interest β_0. A familiar example of such a model is given by $t(\mathbf{X}; \beta_0) = \mathbf{X}'\beta_0$. Suppose also that the association function $\nu(\mathbf{X}, D)$ is modeled parametrically with $\nu(\mathbf{X}, D; \alpha_0)$ where α_0 is an unknown parameter, and $\nu(\mathbf{X}, D; \alpha)$ satisfies the restriction $\nu(\mathbf{X}, 0; \alpha) = \nu(\mathbf{X}, D; 0) = 0$. The resulting parametric model for $E(Y|\mathbf{X}, D)$ is given by:

$$\begin{aligned} \widetilde{\mu}(\mathbf{X}, D; \theta) &= \exp\left[t(\mathbf{X}; \beta) + \nu(\mathbf{X}, D; \alpha) - \overline{\nu}(\mathbf{X}; \psi, \eta, \alpha)\right] \qquad (6) \\ \theta_0 &= (\beta', \eta, \psi', \alpha')' \end{aligned}$$

Estimation and inference about θ_0 then proceeds as in the case of an identity link function, by solving the estimating equation $\mathbf{W}\left(\widehat{\theta}\right) = \sum_i \mathbf{U}_i\left(\widehat{\theta}\right) = 0$ given by (4) , upon substituting in equation (6) for the conditional mean model $\widetilde{\mu}(\mathbf{X}, D; \theta)$, and by letting $\left(\widehat{\psi}, \widehat{\eta}\right)$ be the mle defined in the previous section. The asymptotic distribution of $\widehat{\theta}$ is then given by (5) upon making the foregoing substitutions.

4 Regression with a logit link function

Next, suppose that the secondary outcome Y were binary. We give a novel reparametrization of $\mathbb{E}(Y|\mathbf{X},D) = \Pr(Y=1|\mathbf{X},D)$ on the logit scale, in terms of the population regression function of interest $\mathbb{E}(Y|\mathbf{X}) = \Pr(Y=1|\mathbf{X})$. To proceed, let $ODDS(\mathbf{X},D) = \Pr(Y=1|\mathbf{X},D)/\Pr(Y=0|\mathbf{X},D)$ denote the odds of $\{Y=1\}$ within levels of $(\mathbf{X},D)$. Likewise, let $ODDS(\mathbf{X}) = \Pr(Y=1|\mathbf{X})/\Pr(Y=0|\mathbf{X})$ denote the odds of $\{Y=1\}$ within levels of $\mathbf{X}$. Then, note that

$$\begin{aligned}
\frac{\widetilde{\mu}(\mathbf{X},D)}{1-\widetilde{\mu}(\mathbf{X},D)} &\equiv ODDS(\mathbf{X},D) \\
&= \frac{ODDS(\mathbf{X},D)}{ODDS(\mathbf{X})} \times ODDS(\mathbf{X}) \\
&= \frac{ODDS(\mathbf{X},D)}{ODDS(\mathbf{X},D=0)} \times \left\{ \sum_{d^*=0}^{1} \frac{ODDS(\mathbf{X},d^*)}{ODDS(\mathbf{X},D=0)} \Pr(D=d^*|\mathbf{X},Y=0) \right\}^{-1} \times ODDS(\mathbf{X}) \\
&= \exp\left\{ \log \frac{\mu(\mathbf{X})}{1-\mu(\mathbf{X})} + \nu(\mathbf{X},D) - \overline{\nu}(\mathbf{X}) \right\} \qquad (7)
\end{aligned}$$

where

$$\mu(\mathbf{X}) = \Pr(Y=1|\mathbf{X})$$

is the outcome risk function in the population,

$$\nu(\mathbf{X},D) = \log ODDS(\mathbf{X},D)/ODDS(\mathbf{X},D=0)$$

measures the log-odds ratio association between D and Y within levels of $\mathbf{X}$, and accounts for selection bias due to the sampling design. As formally shown later, the term

$$\overline{\nu}(\mathbf{X}) = \log\left\{\exp\left\{\nu(\mathbf{X}, D=1)\right\}\Pr(D=1|\mathbf{X}, Y=0) + \Pr(D=0|\mathbf{X}, Y=0)\right\}$$

ensures that upon marginalization over D in the target population, the conditional odds $ODDS(\mathbf{X}, D)$ reduces to the marginal odds of interest $ODDS(\mathbf{X})$, and therefore the corresponding mean function $\widetilde{\mu}(\mathbf{X}, D) = \Pr(Y=1|\mathbf{X}, D)$ marginalizes to $\mu(\mathbf{X}) = \Pr(Y=1|\mathbf{X})$ exactly. Interestingly, note that the population density of D used in the above re-parametrization conditions on $\{Y=0\}$ and hence differs from the density function of D involved in previous reparametrizations for the identity or log-link functions. This choice of parametrization is an immediate consequence of the following property of odds ratios. While $\mathbb{E}\left\{ODDS(\mathbf{X}, D)|\mathbf{X}\right\} \neq ODDS(\mathbf{X})$, it is however the case that $\mathbb{E}\left\{ODDS(\mathbf{X}, D)|\mathbf{X}, Y=0\right\} = ODDS(\mathbf{X})$, marginalization of the conditional odds with respect to disease status in the underlying population of individuals free of the secondary outcome gives the marginal odds function of primary interest. Equation (7) is equivalently written as a conditional logistic regression:

$$\widetilde{\mu}(\mathbf{X}, D) = \Pr\left(Y=1|D, \mathbf{X}\right) = \left[1+\exp\left\{-\log\frac{\mu(\mathbf{X})}{1-\mu(\mathbf{X})} - \nu(\mathbf{X}, D) + \overline{\nu}(\mathbf{X})\right\}\right]^{-1}$$

where

$$\Pr\left(Y=1|\mathbf{X}\right) = \left[1+\exp\left\{-\mu(\mathbf{X})\right\}\right]$$

Suppose that the log odds function $\log\left[\mu(\mathbf{X})/\left\{1-\mu(\mathbf{X})\right\}\right]$ follows a parametric model of the form $\mu^{\dagger}(\mathbf{X}; \beta_0)$ where $\mu^{\dagger}(\cdot; \beta_0)$ is known up to the parameter of primary interest β_0. A familiar example of such a model is given by $\mu^{\dagger}(\mathbf{X}; \beta_0) = \mathbf{X}'\beta_0$. Suppose also that the log-odds ratio function $\nu(\mathbf{X}, D)$ is modelled parametrically with $\nu(\mathbf{X}, D; \alpha_0)$ where α_0 is an unknown parameter, and $\nu(\mathbf{X}, D; \alpha)$ satisfies the restriction $\nu(\mathbf{X}, 0; \alpha) = \nu(\mathbf{X}, D; 0) = 0$. Let

$$\text{logit}\pi(\mathbf{X}; \psi_0, \eta_0) = \text{logit}\Pr(D = d^*|\mathbf{X}, Y = 0, S = 1; \psi_0, \eta_0) = \eta_0 + m\left(\mathbf{X}; \psi_0\right) \tag{8}$$

now denote a parametric model for $\Pr(D = d^*|\mathbf{X}, Y = 0, S = 1)$ in the case-control sample, with unknown parameter α_0. Let logit $\Pr(D = d^*|\mathbf{X}, Y = 0; \alpha_0) = \text{logit}\pi(\mathbf{X}; \alpha_0) + \log \frac{\overline{p}(1-\overline{\pi})}{\overline{\pi}(1-\overline{p})}$ denote the corresponding model in the population. The resulting parametric model for $\Pr\left(Y = 1|D, \mathbf{X}\right)$ is given by:

$$\Pr\left(Y = 1|D, \mathbf{X}; \theta_0\right) = \left[1 + \exp\left\{-\mu^{\dagger}(\mathbf{X}; \beta_0) - \nu(\mathbf{X}, D; \alpha_0) + \overline{\nu}(\mathbf{X}; \psi_0, \eta_0, \alpha_0)\right\}\right]^{-1} \tag{9}$$

$$\theta_0 = (\beta_0', \eta_0, \psi_0', \alpha_0')'$$

Estimation and inference about θ_0 can then proceed as in the identity or log link settings, by solving the estimating equation $\mathbf{W}\left(\widehat{\theta}\right) = \mathbb{P}_n\mathbf{U}\left(\widehat{\theta}\right) = 0$ given by (4), upon substituting in equation (9) for the conditional mean model $\widetilde{\mu}(\mathbf{X}, D; \theta)$, but with $\left(\widehat{\psi}, \widehat{\eta}\right)$ the mle obtained using the log-likelihood function $\mathbb{P}_n L\left(\psi_0, \eta_0\right)$ where $L\left(\psi_0, \eta_0\right) = (1 - Y)\left\{D_i \text{logit}\pi(\mathbf{X}; \psi_0, \eta_0) + \log(1 - \pi(\mathbf{X}; \psi_0, \eta_0))\right\}$. The asymptotic distribution of $\widehat{\theta}$ is then given by (5) once the above substitution is made. Finally, we briefly note that when D is rare, the logit link is well approximated by the log-link and $\Pr(D = 1|\mathbf{X}, Y = 0) \approx \Pr(D = 1|\mathbf{X})$ and therefore the approximate approach developed in the previous section can again be used for inference.

5 Semiparametric locally efficient estimation

In this section, we present an alternative potentially more efficient strategy for estimating θ_0, based on semiparametric efficiency theory. To proceed, first note that as argued by Breslow et al (2000), the law of the observed data is formally given by the conditional density $f(Y,\mathbf{X}|D) = f(Y|\mathbf{X},D)f(\mathbf{X}|D)$ which is up to a proportionality constant equivalent to the density of an experiment in which D is itself randomly sampled from a Bernoulli density with known event probability equal to $\overline{\pi}$. Thus, we derive the efficient score for i.i.d data $(Y,\mathbf{X},D)$ sampled from the joint density

$$\begin{aligned}
& f(Y|\mathbf{X},D)f(\mathbf{X}|D)\overline{\pi}^{D}(1-\overline{\pi})^{1-D} \\
&= f(Y|\mathbf{X},D)\frac{f(D|\mathbf{X})f(\mathbf{X})}{f(D)}\overline{\pi}^{D}(1-\overline{\pi})^{1-D} \\
&\propto f(Y|\mathbf{X},D)f^{*}(D|\mathbf{X})f^{*}(\mathbf{X}) \qquad (10)
\end{aligned}$$

where $f(Y|\mathbf{X},D)$ is the population density of Y given $(\mathbf{X},D)$, $f(D)$ is the known marginal density of D in the target population; $f(D=1|\mathbf{X}) = p(\mathbf{X})$ is the population probability that $D=1$ given $\mathbf{X}$; logit$f^{*}(D=1|\mathbf{X})$ =logit$\pi(\mathbf{X})$ =logit$p(\mathbf{X}) - \log\frac{\overline{p}(1-\overline{\pi})}{\overline{\pi}(1-\overline{p})}$ is the probability that $D=1$ given $\mathbf{X}$ in the case-control sample; $f^{*}(\mathbf{X}) \propto f(\mathbf{X})\frac{f(D=0|\mathbf{X})}{f^{*}(D=0|\mathbf{X})}$ is the case-control density of $\mathbf{X}$. Define the semiparametric model $\mathcal{M}_1$, with sole restrictions given by the restricted mean model $\widetilde{\mu}(\mathbf{X},D;\theta)$ for Y given $(\mathbf{X},D)$, with identity link (equation (3)) or log link (equation (6)); and the parametric model (2) for D given $\mathbf{X}$. The model is otherwise nonparametric in the density of $\varepsilon(\theta) = Y - \widetilde{\mu}(\mathbf{X},D;\theta)$ given $(\mathbf{X},D)$, as well as in the population density $f(\mathbf{X})$ and thus in $f^{*}(\mathbf{X})$. To handle the logistic model, likewise define the semiparametric model $\mathcal{M}_2$ with sole restriction the parametric models (8) and (9), and the model is otherwise unrestricted in $f(\mathbf{X})$ and therefore in $f^{*}(\mathbf{X})$. Note that whereas $\mathcal{M}_1$ parametrizes $\Pr(D=d^{*}|\mathbf{X},S=1)$, $\mathcal{M}_2$ places a model for the density $\Pr(D = d^{*}|\mathbf{X}, Y = 0, S = 1)$. Nonetheless, as we show next, model (8) together with model (9) recover a parametric model for the conditional density $f(Y, D|\mathbf{X}, S = 1)$ using the following nonparametric characterization of a joint density (see for example Tchetgen Tchetgen et al, 2010 and Tchetgen Tchetgen and Rotnitzky, 2012):

$$\begin{aligned} f(Y, D|\mathbf{X}, S = 1) &= \frac{f(Y|D = 0, \mathbf{X}, S = 1)OR(Y, D|\mathbf{X}, S = 1)f(D|Y = 0, \mathbf{X}, S = 1)}{\sum_{d,y} f(Y|D = 0, \mathbf{X}, S = 1)OR(Y, D|\mathbf{X}, S = 1)f(D|Y = 0, \mathbf{X}, S = 1)} \\ &= \frac{f(Y|D = 0, \mathbf{X})OR(Y, D|\mathbf{X})f(D|Y = 0, \mathbf{X}, S = 1)}{\sum_{d,y} f(y|D = 0, \mathbf{X})OR(y, d|\mathbf{X})f(d|Y = 0, \mathbf{X}, S = 1)} \\ &= \frac{f(Y|D = 0, \mathbf{X})OR(Y, D|\mathbf{X})f(D|Y = 0, \mathbf{X}) \left\{\overline{p}\left(1 - \overline{\pi}\right)/\overline{\pi}\left(1 - \overline{p}\right)\right\}^{D}}{\sum_{d,y} f(y|D = 0, \mathbf{X})OR(y, d|\mathbf{X})f(d|Y = 0, \mathbf{X}, S = 1) \left\{\overline{p}\left(1 - \overline{\pi}\right)/\overline{\pi}\left(1 - \overline{p}\right)\right\}^{d}} \end{aligned} \tag{11}$$

where $OR(Y, D|\mathbf{X}, S = 1) = OR(Y, D|\mathbf{X}) =$

$$\begin{aligned} &\frac{f(Y|D, \mathbf{X})f(Y = 0|D = 0, \mathbf{X})}{f(Y|D = 0, \mathbf{X})f(Y = 0|D, \mathbf{X})} \\ &= \nu(\mathbf{X}, 1) \end{aligned}$$

is the odds ratio function relating D and Y within levels of $\mathbf{X}$, which yields under our choice of parametrization:

$$f(Y, D|\mathbf{X}, S = 1; \theta_0) = \frac{\exp\left\{Y\mu^{\dagger}(\mathbf{X}; \beta_0) + Y\nu(\mathbf{X}, D; \alpha_0) - Y\overline{\nu}(\mathbf{X}; \alpha_0, \psi_0, \eta_0) + D\eta_0 + Dm\left(\mathbf{X}; \psi_0\right)\right\}}{\sum_{d,y} \exp\left\{y\mu^{\dagger}(\mathbf{X}; \beta_0) + y\nu(\mathbf{X}, d; \alpha_0) - y\overline{\nu}(\mathbf{X}; \alpha_0, \psi_0, \eta_0) + d\eta_0 + dm\left(\mathbf{X}; \psi_0\right)\right\}} \tag{12}$$

This in turn implies a parametric model $f(D = 1|\mathbf{X}, S = 1; \theta_0) = \sum_{y} f(y, D = 1|\mathbf{X}, S = 1; \theta_0)$ for $\pi(\mathbf{X})$ in terms of θ_0. Note that in the target population, the analog to equation (11) is

$$f(Y, D|\mathbf{X}) = \frac{f(Y|D = 0, \mathbf{X})OR(Y, D|\mathbf{X})f(D|Y = 0, \mathbf{X})}{\sum_{d,y} f(y|D = 0, \mathbf{X})OR(y, d|\mathbf{X})f(d|Y = 0, \mathbf{X}, S = 1)},$$

which in turn can be used to verify that under the proposed parametrization (7) ,

$$
\begin{aligned}
\text{logit}\mathbb{E}\left\{\widetilde{\mu}(\mathbf{X},D)|\mathbf{X}\right\} &= \text{logit}\sum_{d} f(Y=1,D=d|\mathbf{X}) = \widetilde{\mu}(\mathbf{X},D) \\
&= \left[1+\exp\left\{-\mu(\mathbf{X})\right\}\right]^{-1} \\
&= f(Y=1|\mathbf{X})
\end{aligned}
$$

formally justifying the earlier claim that our choice of parametrization is made to ensure such marginalization whether nonparametric, semiparametric or parametric models are used.

The following theorem gives the efficient score for θ_0 in models $\mathcal{M}_j$, $j=1,2$.

Proposition 1 *The efficient score of* θ_0 *in model* $\mathcal{M}_1$ *is given by*

$$
\mathbf{R}\left(\theta_0\right) = \begin{pmatrix} \mathbf{R}_{(\beta,\alpha)}\left(\theta_0\right) \\ \mathbf{R}_{(\eta,\psi)}\left(\theta_0\right) \end{pmatrix}
$$

where

$$
\mathbf{R}_{(\beta,\alpha)} = \frac{\partial\widetilde{\mu}(\mathbf{X},D;\theta)}{\partial\left(\beta',\alpha'\right)'}\left\{var\left(\varepsilon(\theta)|\mathbf{X},D\right)\right\}^{-1}\varepsilon(\theta),
$$

and

$$
\mathbf{R}_{(\eta,\psi)}\left(\theta\right) = \mathbf{S}\left(\psi,\eta\right) + \frac{\partial\widetilde{\mu}(\mathbf{X},D;\theta)}{\partial\left(\psi',\eta\right)'}\left\{var\left(\varepsilon(\theta)|\mathbf{X},D\right)\right\}^{-1}\varepsilon(\theta).
$$

The efficient score in model $\mathcal{M}_2$ *is given by the score equation of* θ *corresponding to the log-likelihood* $\mathbb{P}_n \log f(Y,D|\mathbf{X},S=1;\theta)$ *defined in equation* (12) .

Next, suppose that $\widehat{\sigma}^2\left(\mathbf{X},D,\widehat{\theta}\right) = \widehat{var}\left(\varepsilon(\widehat{\theta})|\mathbf{X},D\right)$ is a consistent estimate of the conditional variance $\sigma^2\left(\mathbf{X},D,\theta_0\right) = var\left(\varepsilon(\theta_0)|\mathbf{X},D\right)$, then, upon defining $\widehat{\mathbf{R}}\left(\theta\right)$ as $\mathbf{R}\left(\theta\right)$ by replacing $\sigma^2\left(\mathbf{X},D\right)$ with $\widehat{\sigma}^2\left(\mathbf{X},D,\widehat{\theta}\right)$, the estimator $\widehat{\theta}_{eff}$ that solves $\mathbb{P}_n\widehat{\mathbf{R}}\left(\widehat{\theta}_{eff}\right) = 0$ is regular and asymptotically linear, with large sample variance the semiparametric efficiency bound in $\mathcal{M}_1$ which is given by $\mathbb{E}\left\{\mathbf{R}\left(\theta_0\right)\mathbf{R}^T\left(\theta_0\right)\right\}^{-1}$. In practice, $\widehat{\sigma}^2\left(\mathbf{X}, D, \widehat{\theta}\right)$ may be based on a parametric/semiparametric model, and therefore, may be inconsistent if modeling error were present. Then, $\widehat{\theta}_{eff}$ would still be RAL, although not asymptotically efficient. For this reason, $\widehat{\theta}_{eff}$ is known as a semiparametric locally efficient estimator that is consistent and asymptotically normal regardless of whether $\widehat{\sigma}^2\left(\mathbf{X}, D, \widehat{\theta}\right)$ is consistent or not, and that is asymptotically efficient at the submodel where $\widehat{\sigma}^2\left(\mathbf{X}, D, \widehat{\theta}\right)$ is consistent. The result also states that when Y is binary, the semiparametric efficiency bound is achieved by the maximum likelihood estimator that solves $\mathbb{P}_n\mathbf{R}_{bin}\left(\theta\right) = \mathbb{P}_n\partial\log f(Y, D|\mathbf{X}, S = 1; \theta)/\partial\theta = 0$, with variance obtained by an empirical version of $\mathbb{E}\left\{\mathbf{R}_{bin}\left(\theta_0\right)\mathbf{R}_{bin}^T\left(\theta_0\right)\right\}^{-1}$. This results follows from standard maximum likelihood theory.

Interestingly, upon close inspection of the efficient score $\mathbf{R}_{(\eta,\psi)}\left(\theta\right)$ one notes that information about (ψ, η) the parameter indexing the density of D given $\mathbf{X}$, naturally comes from the score of the corresponding factor of the likelihood function, i.e. $\mathbf{S}\left(\psi, \eta\right)$; however, additional information is obtained from the factor corresponding to the conditional density of Y given $\left(D, \mathbf{X}\right)$. Although unusual, this is not entirely surprising given that this density was carefully reparametrized to depend on $\left(\psi, \eta\right)$. This further reveals that the simple estimating equations approach that gave $\widehat{\theta}$ in previous sections, do not generally exploit this additional information since $\left(\widehat{\psi}, \widehat{\eta}\right)$ solve the score equation $\mathbb{P}_n\left\{\mathbf{S}\left(\psi, \eta\right)\right\} = 0$ instead of the efficient score equation $\mathbb{P}_n\left\{\mathbf{R}_{(\eta,\psi)}\left(\theta\right)\right\} = 0$, and is therefore generally inefficient, except perhaps when the disease is rare.

6 A simulation study

We performed a simulation study to compare in the context of simple linear regression, the performance of the locally efficient estimator to that of two common strategies used in practice. The first approach involves inverse-probability weighting (ipw) by the selection probability given case-control status, while the second approach involves including case-control status as a covariate in the regression for the secondary outcome. We also compared these methods to ordinary linear regression based on the entire data set, which one expects to be significantly biased. We generated X from a mixture of normals with density $N(0, 4)$ with probability 0.88 and density $N(2, 4)$ otherwise. The logistic model is $\text{logit}\Pr(D = 1|X) = -2.5 + \psi_0 X$, where $\psi_0 = 0.5$. The model for Y given X is the linear regression model, $Y = 50 + \beta_1 X + \epsilon$, where $\epsilon|X$ is a mean zero residual error, that is generated such that model (3) holds with $\gamma(X; \alpha_0) = 3 + 2X$, and $\varepsilon(\theta_0)|D, X \sim N(0, 4)$. The simulation study explores both null ($\beta_1 = 0$) and non-null ($\beta_1 = 4$) conditions. The rate of disease is approximately 0.12 in the target population and therefore, the rare disease approximation does not hold. The case-control study has 500 cases and 500 controls, we generated 1000 simulated data sets.

For the simulation study, the locally efficient approach is implemented by maximizing the log-likelihood $\log\{f(\varepsilon(\theta)|X, D)f^*(D|X; \eta, \psi)\}$ which corresponds exactly to solving the efficient score of Proposition 1, under homoscedastic normal error, i.e. assuming $\varepsilon(\theta)|X, D \sim N(0, \sigma^2)$. This specific choice of likelihood model facilitates the implementation of the locally efficient approach using standard off-the shelf software, we used Proc NLMIXED in SAS to implement the approach.

Insert Table 1

The simulation results given in Table 1 confirm that ipw and the locally efficient approach both have small bias and produce 95% confidence intervals with appropriate coverage under either the null or the alternative hypothesis. In contrast, as expected, ordinary linear regression using the entire sample and ignoring the sampling design is noticeably biased with disastrous coverage (= 0%)

in all scenarios. Simply adding a main effect for disease status corrects some of the bias but still produces 95% confidence intervals with poor coverage. In terms of efficiency, as expected, locally efficient estimation clearly outperforms ipw in both scenarios with relative efficiency sometimes greater than 200%.

We also implemented the inefficient estimating equations of Section 2.2, together with standard logistic maximum likelihood estimation of ψ_0. Although both approaches show little bias (results not shown), as projected by Proposition 1, the locally efficient estimator outperforms this alternative strategy in terms of efficiency and demonstrates remarkable efficiency gain not only for the parameter of primary interest β_0 $(ARE\,(\beta_0) \ = 115\%, ARE\,(\beta_1) = 180\%)$, where $ARE\,(\beta) = var(\widehat{\beta})/var(\widehat{\beta}_{eff})$ but also for the logistic regression parameter ψ_0 $(ARE\,(\psi_0) \ = 300\%)$. This result confirms that as projected by Proposition 1, the locally efficient approach can, when the disease is not rare, recover information about ψ_0 that standard logistic regression cannot exploit.

7 An empirical application

This section illustrates the locally efficient approach in an analysis of data from a population-based case-control study of ovarian cancer (Modan et al, 2001). Two controls per case were selected from a central population registry in Israel, matching on age within two years, area of birth and place and length of residence. Blood samples were collected on both cases and controls and were tested for the presence of mutation in two major breast and ovarian cancer susceptibility genes BRCA1 and BRCA2. Additional data were collected on reproductive and gynecologic history, such as parity, number of years of oral contraceptive use and gynecologic surgery. The main objective of the study was to examine the interplay of the BRCA1/2 genes and known reproductive/gynecologic risk factors for ovarian cancer. In reanalyses of these data, a number of authors have exploited a gene-environment independence assumption to obtain more efficient estimates of interactions between BRCA1/2, and parity and oral contraceptive use respectively (Chatterjee and Carroll, 2005, Tchetgen Tchetgen and Robins, 2010, Tchetgen Tchetgen, 2011). Specifically, they assumed that in the target population BRCA1/2 is jointly independent of parity and oral contraceptive within levels of covariates. As a secondary analysis, we evaluate this hypothesis empirically and estimate the mean association in the target population, between BRCA1/2 status and years of oral contraceptive use (Y_1) and parity (Y_2) respectively, adjusting for covariates. Thus, let $\mathbf{X}$ =(BRCA1/2, age (categorical defined by decades), ethnic background (Ashkenazi or non-Ashkenazi), the presence of personal history of breast cancer, a history of gynecologic surgery, and family history of breast or ovarian cancer (no cancer vs one breast cancer in the family vs one ovarian cancer or two or more breast cancer cases in the family)). The analysis uses data on 832 cases and 747 controls who did not have bilateral oophorectomy and who were interviewed for risk factor information and successfully tested for BRCA1/2 mutations. To illustrate the method with both identity and log link functions, Y_1 is coded as number of years of oral contraceptive use and a linear regression of Y_1 on $\mathbf{X}$ is evaluated, while Y_2 is a count of live births, and a log-linear model is assumed for the regression of Y_2 on $\mathbf{X}$. As suggested by Chatterjee and Carroll (2005), we set the population rate of ovarian cancer to $\overline{p} = 8.7 \times 10^{-4}$ which implies the rare disease approximation is appropriate, and thus an estimate of the risk of ovarian cancer as a function of $\mathbf{X}$ is not strictly needed. Nonetheless, we performed both analyses, with and without the rare disease approximation, and obtained identical results.

For each outcome, we compare inferences based on standard OLS ignoring case-control status, IPW and the locally efficient approach with and without possible effect heterogeneity by BRCA1/2 in the case-control adjustment, i.e. $\gamma(\mathbf{X}; \alpha_0) = \alpha_0$ vs $\gamma(\mathbf{X}; \alpha_0) = \alpha_0 + \alpha_1 \times$BRCA1/2.

Insert Table 2 here.

Table 2. summarizes the results for BRCA1/2 associations with Y_1 and Y_2. In both sets of analyses, standard OLS gives the largest point estimates for the effect of BRCA1/2 on the average years of oral contraceptive use and parity, respectively. For both outcomes, IPW and the locally efficient approach incorporating a $D \times BRCA1/2$ interaction correct the OLS estimate, nonetheless the three methods agree in their conclusion and none rejects the null hypothesis of no gene-environment association at the α=0.05 level. Interestingly, not including the interaction in the locally efficient approach has different effects in the two analyses. For Y_1, not including the interaction leads to a wider Wald 95% confidence interval that rejects the null hypothesis of no $BRCA1/2$ association, which suggests the need to account for the interaction. In contrast, removing the interaction in the Y_2 regression leads to a shorter confidence interval without altering the overall conclusion, suggesting that perhaps the interaction is not necessary.

8 Conclusion

In this paper, we have described a general yet simple framework for performing regression analysis for a secondary outcome in the context of case-control sampling. The current results focused on the three most common link functions used in practice, the identity link typically used for a continuous outcome, the log link typically used with counts, and the logit link typically used for binary data. A simple set of estimating equations is described for inference, and a potentially more efficient approach is also given. A particular appeal of the approach is that it is readily implemented with off-the-shelf statistical software. The framework also gives a formal justification for including the case-control status as a covariate in the regression model in view to account for study design when the case-control disease is rare, without requiring the distributional assumptions that have previously appeared in the literature. It is also straightforward to extend our basic argument to justify this type of conditional approach for other link functions, such as the complementary log-log link, or the probit link, under rare disease. When the disease is not rare, the approach requires that sampling fractions are known for cases and non-case controls, which may be a challenge in certain settings, but is usually feasible if the case-control sample is nested within a well-defined cohort study. It is also straightforward to extend our framework to the context of matched case-control studies, the simplest strategy would be to include matching factors into the regression model.

Finally, an interesting and important direction for future work is to further develop the framework to handle settings where the secondary outcome is a vector of correlated variables, arising either from a longitudinal process, or due to spatial or other potential sources of clustering.

APPENDIX

PROOF OF PROPOSITION 1: Let L_2^0 denote the Hilbert space of mean zero functions of $O = (Y, D, \mathbf{X})$, with inner product given by the expectation wrt F_O the case-control distribution of O with density equivalently written $f(\varepsilon(\theta_0)|\mathbf{X}, D)f^*(D|\mathbf{X}; \psi_0, \eta_0)\ f^*(\mathbf{X})$. The model is semiparametric in the sense that the conditional density of the residual $\varepsilon(\theta_0)$ given $(\mathbf{X}, D)$ and the case-control density of $\mathbf{X}$ are left unrestricted. Throughout, assume that the population disease prevalence is known. The nuisance tangent space Λ_{nuis} for the model is given by the closed linear span of all regular parametric scores for the conditional density of $\varepsilon(\theta_0)$ given $(\mathbf{X}, D)$ and of $f^*(\mathbf{X})$. Then, one can verify that

$$\Lambda_{nuis} = \left\{ \begin{array}{c} a_1(O) + a_2(\mathbf{X}) : \text{such that} \\ \mathbb{E}\left\{a_1(O)|\mathbf{X}, D\right\} = \mathbb{E}\left\{\varepsilon(\theta_0)a_1(O)|\mathbf{X}, D\right\} = \mathbb{E}\left\{a_2(\mathbf{X})\right\} = 0 \end{array} \right\} \cap L_2^0$$

It follows that the set of all influence functions is contained in the ortho-complement of Λ_{nuis} :

$$\Lambda_{nuis}^{\perp} = \left\{h_1(\mathbf{X}, D)\,\varepsilon(\theta_0) + h_2(\mathbf{X})\left\{D - \Pr(D = 1|\mathbf{X}, S = 1; \psi_0, \eta_0)\right\} :\ h_1, h_2\right\} \cap L_2^0$$

Next, let $\mathbf{S}_{\theta_0}(O; \theta_0) = \partial \log f(\varepsilon(\theta_0)|\mathbf{X}, D)/\partial\theta_0 + \partial \log f^*(D|\mathbf{X}; \psi_0, \eta_0)/\partial\theta_0$ denote the score wrt

$\theta_0 = (\beta_0', \alpha_0', \eta_0, \psi_0')'$, Then:

$$\begin{aligned} S_{\theta_0}(O;\theta_0) &= \mathbf{S}^1_{\theta_0}(O;\theta_0) + \mathbf{S}^2_{\theta_0}(O;\theta_0) \\ &= -\frac{\partial f(\varepsilon(\theta_0)|\mathbf{X},D;\theta_0)}{\partial \varepsilon(\theta_0)} \times \frac{1}{f(\varepsilon(\theta_0)|\mathbf{X},D;\theta_0)} \times \frac{\partial \widetilde{\mu}(\mathbf{X},D;\theta)}{\partial \theta}|_{\theta_0} \\ &+ \begin{pmatrix} 0 \\ 1 \\ \frac{\partial m(\mathbf{X};\psi_0)}{\partial \psi_0} \end{pmatrix} \{D - \Pr(D=1|\mathbf{X}, S=1; \psi_0, \eta_0)\} \end{aligned}$$

therefore, the efficient score of θ_0 is given by the orthogonal projection of $\mathbf{S}_{\theta_0}(O;\theta_0)$ onto $\Lambda^{\perp}_{nuis}$. Upon noting that $\mathbb{E}\left(\frac{\partial f(\varepsilon(\theta_0)|\mathbf{X},D;\theta_0)}{\partial\varepsilon(\theta_0)} \times \frac{1}{f(\varepsilon(\theta_0)|\mathbf{X},D;\theta_0)} \times \varepsilon(\theta_0)|\mathbf{X}\right) = -1$, it is straightforward to verify that this projection is given by $\mathbf{R}_{(\eta,\psi)}(\theta_0)$, with $\mathbf{S}^2_{\theta_0}(O;\theta_0) = \mathbf{S}(\psi_0, \eta_0)$.

□

References

[1] Breslow NE, Robins JM, Wellner JA (2000). On the semiparametric efficiency of logistic regression under case-control sampling. Bernoulli. 6(3):447-455.

[2] Jiang, Y., Scott, A. J. and Wild, C. J. (2006) Secondary analysis of case-control data. Statist. Med., 25, 1323–1339.

[3] Lee AJ, McMurchy L, Scott AJ. Re-using data from case–control studies. Statistics in Medicine 1997; 16:1377–1389.

[4] Lettre G, Jackson A, Gieger C, Schumacher FR, Berndt S, Hirschhorn J. 2008 Identification of ten loci associated with height and previously unknown biological pathways in human growth. Nat Genet 40(5):584–591.

[5] 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. Genet. Epidem., 34, 427–433.

[6] Lin, D. Y. and Zeng, D. (2009) Proper analysis of secondary phenotype data in case-control association studies. Genet. Epidem., 33, 256–265.

[7] Loos R, Lindgren CM, Li S, Wheeler E, Zhao J. 2008 Association studies involving over 90,000 samples demonstrate that common variants near MC4R influence fat mass, weight and risk of obesity. Nat Genet. Nat Genet, 40(6): 768–775.

[8] Modan, M. D., Hartge, P. et al. (2001). Parity, oral contraceptives and the risk of ovarian cancer among carriers and noncarriers of a BRCA1 or BRCA2 mutation. *New England Journal of Medicine.* 345, 235–40.

[9] Monsees, G., Tamimi, R. and Kraft, P. (2009) Genomewide association scans for secondary traits using case-control samples. Genet. Epidem., 33, 717–728.

[10] Nagelkerke NJD, Moses S, Plummer FA, Brunham RC, Fish D. Logistic regression in case–control studies: the effect of using independent as dependent variables. Statistics in Medicine 1995; 14:769–755.

[11] Reilly M, Torrang A, Klint A. Reuse of case–control data for analysis of new outcome variables. Statistics in Medicine 2005; 24:4009–4019.

[12] Richardson DB, Rzehak P, Klenk J, Weiland SK. Analysis of case–control data for additional outcomes. Epidemiology 2007; 18:441–445.

[13] Robins JM, Rotnitzky A, Zhao LP. (1994). Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association, 89:846-866. Reproduced courtesy of the American Statistical Association.

[14] Sanna S, Jackson AU, Nagaraja R, Willer CJ, Chen WM, Bonnycastle LL, Shen H, Timpson N, Lettre G, Usala G and others. 2008. Common variants in the GDF5-UQCC region are associated with variation in human height. Nat Genet 40(2):198-203.

[15] Tchetgen Tchetgen E, Robins J and Rotnitzky A. On Doubly robust estimation of a semi-parametric odds ratio model. Biometrika. 2010, vol. 97(1), pages 171-180.

[16] Tchetgen Tchetgen EJ, Rotnitzky A: Double-robust estimation of an exposure-outcome odds ratio adjusting for confounding in cohort and case-control studies. Stat Med; 2011 Feb 20;30(4):335-47.

[17] Tchetgen Tchetgen E and Robins J. The semi-parametric case-only estimator. Biometrics. Biometrics. 2010 Dec;66(4):1138-44. doi: 10.1111/j.1541-0420.2010.01401.

[18] Tchetgen Tchetgen E . Robust Discovery of Genetic Associations incorporating Gene-Environment Interaction and Independence. (2011) Epidemiology. Volume 22 ;2; 262-272.

[19] Tchetgen Tchetgen EJ. Leveraging auxiliary information to enhance power in the analysis of nested case-control GWAS. (2012) Technical Report. Harvard University.

[20] Wang J, Shete S. Estimation of odds ratios of genetic variants for the secondary phenotypes associated with primary diseases. Genetic Epidemiology 2011; 35:190–200.

[21] Weedon MN, Lettre G, Freathy RM, Lindgren CM, Voight BF, Perry JR, Elliott KS, Hackett R, Guiducci C, Shields B and others. 2007. A common variant of HMGA2 is associated with adult and childhood height in the general population. Nat Genet 39(10):1245-50.

[22] Wei, J., Carroll, R. J., Muller, U., Van Keilegom, I. and Chatterjee, N. (2013). Locally efficient estimation for homoscedastic regression in the secondary analysis of case-control data. Journal of the Royal Statistical Society, Series B, 75, 186-206.

[23] Weuve J, Korrick S, Weisskopf M, Ryan L, Schwartz J, Nie H, Grodstein F, Hu H, Cumulative exposure to lead in relation to cognitive function in older women, Environ Health Perspect 2009;117:574-80.

Table 1. Simulation results

	absolute bias	variance	Coverage
$\beta_1 = 0$			
Standard OLS	0.734	2.2×10^{-3}	0.000
Conditional OLS	0.227	2.8×10^{-3}	4×10^{-3}
IPW	1.95×10^{-4}	3.3×10^{-3}	0.970
Locally Efficient	1.11×10^{-3}	1.8×10^{-3}	0.960
$\beta_1 = 4$			
Standard OLS	0.730	2.3×10^{-3}	0.000
Conditional OLS	0.231	2.7×10^{-3}	2×10^{-3}
IPW	4.0×10^{-3}	3.4×10^{-3}	0.957
Locally Efficient	4.2×10^{-4}	2.0×10^{-3}	0.956

Table 2. Parameter estimates (standard errors) of mean effect of BRCA1/2 on oral contraceptive use and Parity.

	Y_1	Y_2
	BRCA1/2 (se)	BRCA1/2 (se)
Standard OLS	0.212 (0.144)	-0.053 (0.047)
IPW	0.327 (0.570)	-5$\times10^{-4}$ (0.142)
Locally Efficient without interaction	0.332 (0.152)	-0.020 (0.033)
Locally Efficient with interaction	0.287 (0.109)	0.094 (0.175)

Analyses further adjust for age, ethnic background, personal history of breast cancer, history of gynecologic surgery, and family history of breast or ovarian cancer.