Published in final edited form as:

Stat Biosci. 2013 November 1; 5(2): 250–260. doi:10.1007/s12561-013-9094-9.

A Note on Penalized Regression Spline Estimation in the Secondary Analysis of Case-Control Data

Suzan Gazioglu,

Department of Mathematical Sciences, Montana Tech of the University of Montana, Butte MT 59701

Jiawei Wei,

Beijing Novartis Pharma Co. Ltd., Integrated Information Sciences (IIS), Building 3 Floor 3, No. 3728 Jinke Road, Pudong New District, Shanghai, 201203, China

Elizabeth M. Jennings, and

Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143

Raymond J. Carroll

Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143

Suzan Gazioglu: sgazioglu@mtech.edu; Jiawei Wei: jiawei.wei@novartis.com; Elizabeth M. Jennings: ejennings@stat.tamu.edu; Raymond J. Carroll: carroll@stat.tamu.edu

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 the case-control sampling, and to avoid the biased sampling that arises from the design, it is typical to use the control data only. In this paper, we develop penalized regression spline methodology that uses all the data, and improves precision of estimation compared to using only the controls. A simulation study and an empirical example are used to illustrate the methodology.

Some Key Words

Biased samples; B-splines; Homoscedastic regression; Nonparametric regression; Regression splines; Secondary data; Secondary phenotypes; Two-stage samples

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 of 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. Generally, ignoring the case-control

sampling scheme and doing a regression analysis among the covariates is inappropriate because of the biased nature of the sampling. That is, the regression of one covariate against another in the population differs from that in the case-control sample, see Section 3.1 and Figure 1 for illustration. 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: see Section 3.2 for discussion. However, Li et al. (2010) have shown that there are circumstances under which inclusion of cases can dramatically increase the efficiency of the secondary analysis.

Let disease status be denoted as D and let the covariates be (Y, X) , where Y and X are scalar. Let $f(\cdot)$ be an unknown smooth function. Our interest is in the nonparametric regression model

$$Y = f(X) + \varepsilon, \quad (1)$$

where ε has mean zero and is independent of X , but its distribution is otherwise not specified. We wish to estimate the unknown function $f(\cdot)$ efficiently when the data arise as part of a case-control sample.

We formalize model (1) for future reference as the following assumption.

Assumption 1 *Given X , Y follows the semiparametric model (1), where ε has mean zero and is independent of X , but the distributions of ε and X are otherwise not specified.*

In Section 2, we provide some well-known background to the penalized regression spline approach to fitting model (1) when the data are independent and identically distributed. In Section 3, we show how to modify the work of Wei, et al. (2013) on the secondary analysis of case-control data so that it applies to penalized spline regression. We use theory recently developed by Li and Ruppert (2008) to help guide our choice of the smoothing parameter. Sections 5 and 6 provide simulation evidence and an empirical example, respectively, showing that our new method can be much more efficient than the standard method that uses only the controls.

2 Penalized Regression Splines for Identically Distributed Data

2.1 Penalized Least Squares

Consider random sampling, rather than a case-control study. We will take up the case-control study subsequently in Section 3.

An increasingly popular way to estimate model (1) is through penalized regression splines, see for example Ruppert, et al. (2003) and Wood (2006). In this set up, one starts with a series of basis functions $B(x) = \{b_1(x), \dots, b_P(x)\}^T$, which might be B-splines, truncated polynomials, etc. Then one approximates $f(x) \approx \alpha + B^T(x)\beta$. The explicit use of an intercept is required when we try to use all the data in a case-control study, see Wei, et al. (2013) for an extensive discussion.

The use of the intercept α is standard in the truncated polynomial basis function approach. A truncated polynomial of order q with knots $(a_1, \dots, a_{P-q})$ has $b_j(x) = x^j$ for $j = 1, \dots, q$ and $b_j(x) = (x - a_j)_+^q$ for $j = q + 1, \dots, P$, where $c_+ = cI(c > 0)$. As pointed out by Currie and Durban (2002), B-splines with order q -penalty and $P + 1$ basis functions can also be placed into this framework.

In general, it is sufficient that the number of basis functions P be modestly large, e.g., 10–30 (Ruppert, 2002). If one simply substitutes $f(X)$ by $\alpha + B^T(X)\beta$ in (1), then generally the estimated curve is far too wiggly. Hence, the ordinary least squares likelihood function is penalized, so that one estimates (α, β) by minimizing

$$\sum_{i=1}^n \{Y_i - \alpha - B^T(X_i)\beta\}^2 + \lambda \beta^T \Delta \beta, \quad (2)$$

where Δ is a penalty matrix and λ is a penalty or smoothing parameter. As described in Ruppert, et al. (2003), the penalty matrix Δ for the truncated polynomial basis is diagonal with entries 0.0 corresponding to the polynomial terms in $B(x)$, and entries 1.0 corresponding to the other entries.

Differentiating with respect to β , we see that the score equations for estimating β are

$$0 = \sum_{i=1}^n B(X_i) \{Y_i - \alpha - B^T(X_i)\beta\} - \lambda \Delta \beta. \quad (3)$$

2.2 Smoothing Parameter Estimation for Identically Distributed Data

For independent and identically distributed data, the form (2) plays the role of a criterion function. In the literature, two methods for estimating λ are common (Ruppert, et al., 2003). The first is generalized crossvalidation (GCV), which is computationally fast and simple. The second is via a mixed-model approach, which is based on Y being normally distributed with mean $\alpha + B^T(X)\beta$ and variance σ^2 , and that β is also normally distributed with mean zero and a covariance matrix proportional to Δ^{-} , the generalized inverse of Δ .

As we make clear in Section 4.2 and Section 4.4, for case-control sampling, because of the semiparametric model inherent in Assumption 1, no simple criterion function is available, so there are no counterparts to GCV and mixed models.

3 Secondary Analysis and the Use of Only the Controls

3.1 The Case-Control Study and Biased Sampling

In a case-control study, there are n_0 observations of (Y, X) among the controls, for which $D = 0$, and there are n_1 observations of (Y, X) among the cases, for which $D = 1$. It is typical that in the population, the risk of disease is modeled by a logistic regression with $\text{pr}(D = 1 | Y, X) = H\{\theta_0 + m(Y, X, \theta_1)\}$, where $H(\cdot)$ is the logistic distribution function and $m(\cdot)$ is an arbitrary known function with unknown parameter vector θ_1 .

The first aspect of the case-control study is understanding what happens when logistic regression is used to estimate (θ_0, θ_1) but ignoring the case-control sampling scheme. For $d = 0, 1$, let $\pi_d = \text{pr}(D = d)$, the unknown probability of disease in the source population. As Prentice and Pyke (1979) noted, none of θ_0 , π_0 or π_1 cannot be identified in case-control studies. 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 $\kappa = \theta_0 + \log(n_1/n_0) - \log(\pi_1/\pi_0)$. Ordinary logistic regression consistently estimates κ as the intercept and θ_1 as the main risk parameter, see, for example, Chatterjee and Carroll (2005). Let $\Omega = (\kappa, \theta_1^T)^T$.

The second aspect of the case-control study is understanding what happens when one fits a regression model ignoring the case-control sampling scheme. For a numerical example, take the case that $n_0 = n_1 = 500$, $\theta_0 = -5.5$, $m(y, x, \theta_1) = y\theta_y + x\theta_x$ with $\theta_y = 0.5$ and $\theta_x = 1$, and,

in the population, $Y = \alpha + X\beta + \varepsilon$, with $\alpha = 0$, $\beta = 1$ and $\varepsilon = \text{Normal}(0, 1)$. In addition, in the population, $X = \text{Uniform}(0, 1)$. In this setup, the disease is rare, with $\text{pr}(D = 1) = 0.01$. Thus, while controls are 99% of the population, they are only 50% of the case-control study.

To understand the bias induced by ignoring the case-control sampling scheme, we generated 3,000 case control studies and computed the intercept and slope estimates using all the data versus using the controls only. Using the controls only, the mean estimated intercept and slope were 0.000 and 0.990, respectively, reflecting almost no bias. In addition, the coverage rate for the slope β for a nominal 95% confidence interval is 0.947. Using all the data, the mean estimated intercept and slope were 0.150 and 1.174, respectively, reflecting considerable bias, which leads to a coverage rate of only 0.666. Figure 1 shows the different regression functions compared to the true regression function.

3.2 Rare Disease and the Use of Controls Only

Case-control studies almost exclusively focus on rare diseases. This has no precise definition, with some stating that it means $< 1\%$ and others allowing it to mean $< 5\%$. In the latter case, with over 95% of the population being controls ($D = 0$), the analysis of the controls generally provides an excellent approximation to the regression of Y on X , see Section 3.1 for a good illustration, and also see Section 5. Because case-control studies are conducted for rare outcomes, the rare disease approximation is natural in most applications. It is also widely used, a very non-exhaustive list of which includes Piegorsch, et al. (1994), Epstein and Satten (2003), Lin and Zeng (2006), Modan, et al. (2001), Zhao, et al. (2003), Kwee, et al. (2007), Yang, et al. (1997), Lin and Zeng (2009) and Hu, et al. (2010). Along with its success as an approximation, it also results in considerable computational simplification and avoids thorny issues of identifiability associated with not knowing the disease rate in the population (Chatterjee and Carroll, 2005; Kwee, et al., 2007; Lobach, et al., 2008).

At the request of a referee, we summarize this discussion in the following assumption.

Assumption 2 Define $\pi_1 = \text{pr}(D = 1)$, the unknown probability of disease in the source population. We assume that $\pi_1 < 0.05$ or $\pi_1 < 0.01$.

We thus let $(\hat{\lambda}_{\text{cont}}, \hat{\alpha}_{\text{cont}}, \hat{\beta}_{\text{cont}})$ be the estimates of (λ, α, β) obtained from any version of a penalized spline regression, but applied to the controls only.

The major issue with using the controls only is that only a part of the data are being used, i.e., the cases ($D = 1$) are being ignored. Many case-control studies, such as the one in our empirical example, have as many cases as there are controls, so that as much as 50% of the data are being ignored. Our goal is to use all the data to obtain smaller variability in the estimates of the regression function.

4 Secondary Analysis Using Cases and Controls

4.1 The Case-Control Likelihood Function for Known $f_\varepsilon(\cdot)$

Recall that the logistic disease model is $\text{pr}(D = 1 | Y, X) = H\{\theta_0 + m(Y, X, \theta_1)\}$ and that $\kappa = \theta_0 + \log(n_1/n_0) - \log(\pi_1/\pi_0)$ and $\Omega = (\kappa, \theta_1)$. Let $f_\varepsilon(\cdot)$ be the unknown density function in the population for ε , and let $f_X(\cdot)$ be the unknown density function for X in the population. Define

$$\mathcal{L}(d, y, x, \Omega, \theta_0) = \exp[d\{\kappa + m(y, x, \theta_1)\}][1 + \exp\{\theta_0 + m(y, x, \theta_1)\}]^{-1}.$$

Because of its retrospective nature, the likelihood function for the case-control data is $\text{pr}(Y = y, X = x \mid D = d)$. If $f_\varepsilon(\cdot)$ were known, then the profile semiparametric likelihood function (Chen, et al., 2008, 2009; Lin and Zeng, 2009) for model is known to be

$$\mathcal{L}(D, Y, X, f, \Omega, \theta_0) = \frac{f_\varepsilon\{Y - f(X)\} \mathcal{S}(D, Y, X, \Omega, \theta_0)}{\sum_{d=0}^1 \int f_\varepsilon\{v - f(X)\} \mathcal{S}(d, v, X, \Omega, \theta_0) dv}. \quad (4)$$

If we had a parametric form for $f_\varepsilon(\cdot, \zeta)$, we could then estimate the spline parameters (α, β) by maximizing a penalized form of the loglikelihood

$$\sum_{i=1}^n \log[\mathcal{L}\{D_i, Y_i, X_i, \alpha + B^T(X_i)\beta, \Omega, \theta_0, \zeta\}].$$

4.2 No Case-Control Likelihood Function for Unknown $f_\varepsilon(\cdot)$

In our Assumption 1, we are not assuming a distributional form for $f_\varepsilon(\cdot)$. This presents a major problem, because not only is (4) not calculable, but θ_0 , the intercept in the underlying logistic model, is not even identifiable (Prentice and Pyke, 1979; Chatterjee and Carroll, 2005). If we make the rare disease approximation, formalized as Assumption 2, then $\text{pr}(D = 0 \mid Y, X) = [1 + \exp\{\theta_0 + m(Y, X, \theta_1)\}] \approx 1$, and then $\mathcal{S}(d, y, x, \Omega, \theta_0) \approx \exp[d\{\kappa + m(y, x, \theta_1)\}]$, which does not depend on θ_0 and then, at least approximately, θ_0 disappears from the (approximate) retrospective likelihood function. This is a major simplification. Based upon it, this allowed Wei, et al. (2013) to devise an estimate of the regression parameters in the homoscedastic linear model under Assumption 1, based on using all the data. We summarize this next in our context.

4.3 Estimating Equations

Here we form an estimating equation that is the analogue of (3). We approximate the function $f(x) \approx \alpha + B^T(x)\beta$, which is a linear model, then apply the estimating equations in Wei, et al. (2013) to this case, for (α, β) and then add a penalty as in (3). The direct analogue of the method of Wei, et al. is the following, for fixed λ .

Remark 1 For technical reasons explained in Wei, et al., the intercept α needs a special treatment. The reason for this is that, in the linear model $f(x) = \alpha + B^T(X)\beta$, there is an equivalent semiparametric formulation of Assumption 1 where there is no α , and we simply assume that ε is independent of X but its distribution is otherwise not specified.

- Run an ordinary logistic regression of D on (Y, X) to get a consistent estimate $\hat{\Omega}$. The consistency is well-known, e.g., Prentice and Pyke (1979) and Chatterjee and Carroll (2005).
- For any β , define $R(\beta) = Y - B^T(X)\beta$. Further define $\mathcal{S}\{R_i(\beta), b, x, \beta, \Omega\} = 1 + \exp[\kappa + m\{R_i(\beta) + b^T\beta, x, \theta_1\}]$. Finally, define $L\{R(\beta), X, \alpha, \beta\} = B(X)\{R(\beta) - \alpha\}$.
- For any given (β, Ω) , estimate α as

$$\hat{\alpha}(\beta, \Omega) = \frac{n^{-1} \sum_{i=1}^n R_i(\beta) \{n_0^{-1} \sum_{j=1}^n (1 - D_j) \mathcal{S}\{R_i(\beta), B(X_j), X_j, \beta, \Omega\}\}^{-1}}{n^{-1} \sum_{i=1}^n \{n_0^{-1} \sum_{j=1}^n (1 - D_j) \mathcal{S}\{R_i(\beta), B(X_j), X_j, \beta, \Omega\}\}^{-1}}. \quad (5)$$

- Define

$$\hat{Q}_{n,\text{est}}(\beta, \Omega) = \sum_{i=1}^n \left[L\{R_i(\beta), X_i, \hat{\alpha}(\beta, \Omega), \beta\} - \frac{n_0^{-1} \sum_{j=1}^n (1 - D_j) L\{R_i(\beta), X_j, \hat{\alpha}(\beta, \Omega), \beta\} \mathcal{S}\{R_i(\beta), B(X_j), X_j, \beta, \Omega\}}{n_0^{-1} \sum_{j=1}^n (1 - D_j) \mathcal{S}\{R_i(\beta), B(X_j), X_j, \beta, \Omega\}} \right]. \quad (6)$$

- Solve for β in the equation

$$0 = \hat{Q}_{n,\text{est}}(\beta, \hat{\Omega}) - \lambda \Delta \beta. \quad (7)$$

4.4 Estimation of λ

The most difficult issue is the estimation of the smoothing parameter λ . This is because, as emphasized in Section 4.2, there is no likelihood function available, and alternatives are needed. It is not clear how this can be done. For example, ordinary crossvalidation makes no sense, because as shown in Section 3.1, the case-control sample is a biased sample from the population with cases vastly over-represented. We cannot do a weighted crossvalidation, because the weights depend on the unknown π_d in the population. We have experimented with crossvalidation applied to the controls only, as well as AIC applied to the controls only, and the results are much worse than our simple alternative presented below.

Instead, we use the recent theory of Li and Ruppert (2008) to suggest a quick approach that requires no further calculations, and which works well in our simulations. Their asymptotic analysis allows the number of basis functions P to increase with the sample size. Let λ be the smoothing parameter, define $M = n_0/P$ and define $\lambda_* = \lambda/M$. By the result at the end of their Theorem 2 on their page 423, they show that, when using the controls only, the optimal rate of the smoothing parameter is $\lambda_* \sim (1/4)P^4 h_n^4$, where the equivalent bandwidth is $h_n = c n_0^{-1/9}$ and c is a constant. Thus, with simple algebra, $\lambda = \lambda_* M = (1/4)P^3 c^4 n_0^{5/9}$. Thus, if the number of controls increases from n_0 to $n_0 + m$, then the smoothing parameter would increase by the multiplicative factor $\{(n_0 + m)/n_0\}^{5/9}$. Although we are not increasing the number of controls, our approach is to use as the smoothing parameter $\lambda = \{(n_0 + n_1)/n_0\}^{5/9} \hat{\lambda}_{\text{cont}}$, where $\hat{\lambda}_{\text{cont}}$ is the smoothing parameter obtained by GCV when (3) is solved among the controls. By the way, we have experiments with just using $\hat{\lambda}_{\text{cont}}$, but this also leads to less efficient results than our simple device.

4.5 Remarks on Asymptotic Theory

Asymptotic theory when the number of basis functions is fixed follows along the lines of Yu and Ruppert (2002) and Wei, et al. (2013). There are three essential assumptions. Recall that $n = n_0 + n_1$. Then we assume that (a) $n \rightarrow \infty$ and $n_0/n \rightarrow c_*$ with $0 < c_* < \infty$; (b) $f(x) = B^T(x)\beta$, the spline approximation is exact, and (c) $\lambda = o(n^{1/2})$. Then it is easy to show that $n^{1/2}(\hat{\beta} - \beta) \rightarrow \text{Normal}(0, \Sigma)$, where Σ is exactly as given in Theorem 1 of Wei, et al. (2013), modified for the rare disease approximation as in their Section 4.1, and with their $\mu(x, \beta) = B^T(x)\beta$. In addition, bootstrap inference is asymptotically correct, as argued in Section 3.7 of Wei, et al. (2013).

5 Simulation

Our main intention in this simulation is to demonstrate that our new method, which uses all the data and not just the controls, is more efficient than using the controls data only, without incurring additional bias.

In this simulation, in all cases, we generated 1,000 data sets. We first generated a very large random sample from a population in which $X = \text{Uniform}(0, 1)$, and for different functions $f(x)$, $Y = f(X) + \varepsilon$, where $\varepsilon = \text{Normal}(0, 1)$. We then generated D by using the logistic model $\text{pr}(D = 1|X, Y) = H(\theta_0 + \theta_1 Y + \theta_2 X)$, where H is the logistic function and $\theta_0 = -3.70$, $\theta_1 = 0.25$ and $\theta_2 = 1.00$, so that the rate of disease in the population = 0.045. Having generated this random sample from the population, we then created a case-control study with $n_0 = 100$ controls and $n_1 = 100$ cases.

We used five functions for $f(x)$: $f(x) = \sin(x)$, $f(x) = \sin(\pi x)$, $f(x) = \sin(2\pi x)$, $f(x) = \sin(4\pi x)$ and $f(x) = x^2$. In approximating these functions, we used a truncated polynomial basis spline of degree = 2 along with 10 knots. The initial estimate of λ was formed by GCV among the controls. When using all the data, we used the method described in Sections 4.3 and 4.4.

For the case that $f(x) = \sin(\pi x)$, we present detailed results in the top left and top right portion of Figure 2. In the top left portion of that figure, we plot pointwise means of the methods and the true function: the lack of any noticeable bias for our new method is evident. In the top right portion, we plot the ratio of the pointwise variances for using controls only divided by using all the data: the increase in variance efficiency when using all the data is also evident.

For the other four functions, a similar lack of bias was found, as was a similar gain in efficiency, see Table 1.

In summary, this simulation indicates that the new methodology improves significantly upon spline regression using the controls only.

6 Empirical Example

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 Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial at the National Cancer Institute (Gohagan et al., 2000; Moslehi et al., 2006). Here D is colorectal adenoma status. In our case, we are interested in the relationship between Y , the heterocyclic amine PhIP (2-AMINO-1-METHYL-6-PHENYLIMIDAZO[4,5-b]PYRIDINE) and X , red meat consumption. Heterocyclic amines are chemicals that are formed when meat, poultry, or fish is cooked at high temperatures, such as frying, broiling, and barbecuing, and they are known to be carcinogens. Both Y and X were transformed by $\log(1 + x)$ and then standardized to the unit interval.

We applied our method to these data. The bottom left portion of Figure 2 displays the function fits, which are similar as expected from the simulation of Section 5. We bootstrapped the data according to procedure of Wang, et al. (1997), with 500 bootstrap data sets. The bottom right portion of Figure 2 shows the ratio of the pointwise bootstrap variances, those for using the controls only divided by those using all the data. The increase in variance efficiency for our new method is as expected from the simulation.

As seen in Figure 2, the fitted function seems also adequately modeled by a quadratic function. As in all spline fitting, starting with a properly specified parametric model instead of using a spline will result in more efficient estimation. However, our primary focus here is on spline modeling, not fully parametric modeling, which has been covered by Wei, et al. (2013).

Acknowledgments

Jennings, Wei and Carroll's research were supported by a grant from the National Cancer Institute (R37-CA057030). This publication is based in part on work supported by Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST).

References

- Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation in case-control studies of gene-environment interactions. *Biometrika*. 2005; 92:399–418.
- Chen Y-H, Carroll RJ, Chatterjee N. Retrospective analysis of haplotype-based case-control studies under a flexible model for gene-environment association. *Biostatistics*. 2008; 9:81–99. [PubMed: 17490987]
- Chen Y-H, Chatterjee N, Carroll RJ. Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. *Journal of the American Statistical Association*. 2009; 104:220–233. [PubMed: 19430598]
- Currie ID, Durban M. Flexible smoothing with P-splines: a unified approach. *Statistical Modeling*. 2002; 4:333–349.
- Epstein M, Satten GA. Inference on haplotype effects in case-control studies using unphased genotype data. *American Journal of Human Genetics*. 2003; 73:1316–1329. [PubMed: 14631556]
- Gohagan JK, Prorok PC, Hayes RB, et al. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: history, organization, and status. *Controlled Clinical Trials*. 2000; 21(6 Suppl):251S–272S. [PubMed: 11189683]
- Hu YJ, Lin DY, Zeng D. A general framework for studying genetic effects and gene-environment interactions with missing data. *Biostatistics*. 2010; 11:583–598. [PubMed: 20348396]
- Kwee LC, Epstein MP, Manatunga AK, Duncan R, Allen AS, Satten GA. Simple methods for assessing haplotype-environment interactions in case-only and case-control studies. *Genetic Epidemiology*. 2007; 31:75–90. [PubMed: 17123302]
- Li H, Gail MH, Berndt S, Chatterjee N. Using cases to strengthen inference on the association between single nucleotide polymorphisms and a secondary phenotype in genome-wide association studies. *Genetic Epidemiology*. 2010; 34:427–433. [PubMed: 20583284]
- Li Y, Ruppert D. On The asymptotics of penalized splines. *Biometrika*. 2008; 95:415–436.
- Lin DY, Zeng D. Likelihood-based inference on haplotype effects in genetic association studies (with discussion). *Journal of the American Statistical Association*. 2006; 101:89–118.
- Lin DY, Zeng D. Proper analysis of secondary phenotype data in case-control association studies. *Genetic Epidemiology*. 2009; 33:256–265. [PubMed: 19051285]
- Lobach I, Carroll RJ, Spinka C, Gail MH, Chatterjee N. Haplotype-based regression analysis of case-control studies with unphased genotypes and measurement errors in environmental exposures. *Biometrics*. 2008; 64:673–684. [PubMed: 18047538]
- Maity A, Carroll RJ, Mammen E, Chatterjee N. Testing in semiparametric models with interaction, with applications to gene-environment interactions. *Journal of the Royal Statistical Society, Series B*. 2009; 71:75–96.
- Modan MD, Hartge P, Hirsh-Yechezkel G, Chetrit A, Lubin F, Beller U, Ben-Baruch G, Fishman A, Menczer J, Struwing JP, Tucker MA, Wacholder S. for the National Israel Ovarian Cancer Study Group. Parity, oral contraceptives and the risk of ovarian cancer among carriers and noncarriers of a BRCA1 or BRCA2 mutation. *The New England Journal of Medicine*. 2001; 345:235–240. [PubMed: 11474660]
- Moslehi R, Chatterjee N, Church TR, Chen J, Yeager M, Weissfield J, Hein DW, Hayes RB. Cigarette smoking, N-acetyltransferase genes and the risk of advanced colorectal adenoma. *Pharmacogenomics*. 2006; 7:819–829. [PubMed: 16981843]
- Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population based case-control studies. *Statistics in Medicine*. 1994; 13:153–162. [PubMed: 8122051]

- Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1979; 66:403–411.
- Ruppert D. Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*. 2002; 11:735–757.
- Ruppert, D.; Wand, MP.; Carroll, RJ. *Semiparametric Regression*. Cambridge University Press; 2003.
- Wang CY, Wang S, Carroll RJ. Estimation in choice-based sampling with measurement error and bootstrap analysis. *Journal of Econometrics*. 1997; 77:65–86.
- Wei J, Carroll RJ, Müller U, Van Keilegom I, Chatterjee N. Locally efficient estimation for homoscedastic regression in the secondary analysis of case-control data. *Journal of the Royal Statistical Society, Series B*. 2013; 75:185–206.
- Wood, SN. *Generalized Additive Models: An Introduction with R*. CRC Press; 2006.
- Yang Q, Khoury MJ, Flanders WD. Sample size requirements in case-only designs to detect gene-environment interaction. *American Journal of Epidemiology*. 1997; 146:713–720. [PubMed: 9366618]
- Yu Y, Ruppert D. Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*. 2002; 97:1042–1054.
- Zhao LP, Li SS, Khalid N. A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. *American Journal of Human Genetics*. 2003; 72:1231–1250. [PubMed: 12704570]

Figure 1.

Illustration of the bias induced by the case-control sampling scheme, see Section 3.1. Black dash-dotted line is the true regression function, red solid line is the regression function when using only the controls, and the blue dashed line is the regression function when using all the data and ignoring the case-control sampling scheme.

Figure 2.

Top left: Results of the simulation when $f(x) = \sin(\pi x)$. Plotted are the true function (solid line), the pointwise mean using all the data (dashed line) and the pointwise mean using the controls only (dotted line). This figure illustrates the lack of significant bias in the function fits. Top right: Results of the simulation when $f(x) = \sin(\pi x)$. Plotted are the ratio of the pointwise variances for using controls only divided by using all the data. This figure illustrates the gain in efficiency when using our method. Bottom left: Data analysis in Section 4. Plotted are the function estimates. The solid line uses all the data, while the dashed line uses the controls only. Bottom right: Data analysis in Section 6. Plotted is the pointwise ratio of bootstrap variances, controls only divided by using all the data. The gain in efficiency is as expected from the simulations.

Table 1

Results of the simulation study. Here, “MSE” indicates the average mean squared error, averaged across the simulations and a grid of 101 equally spaced values on the unit interval. “MSE Efficiency” is the ratio of the mean squared error for the analysis that uses only the controls divided by the mean squared error for the analysis that uses all the data, i.e., the ratio of the second column to the third.

Condition	MSE, Controls only	MSE, New all data	MSE efficiency
$f(x) = \sin(x)$	0.044	0.024	1.791
$f(x) = \sin(\pi x)$	0.044	0.024	1.802
$f(x) = \sin(2\pi x)$	0.060	0.033	1.825
$f(x) = \sin(4\pi x)$	0.084	0.045	1.863
$f(x) = x^2$	0.043	0.024	1.795