Statistics and Its Interface Volume 4 (2011) 475–487

Determination of proportionality in two-part models and analysis of Multi-Ethnic Study of Atherosclerosis (MESA)

Anna Liu, Richard Kronmal, Xiaohua Zhou and Shuangge Ma*

In MESA (Multi-Ethnic Study of Atherosclerosis), it is of interest to model the development and progression of CAC (coronary artery calcium). With about half of the CAC scores equal to zero and the rest continuously distributed, semiparametric two-part models are needed. Our main interest lies in determining the (partial) proportionality between the two covariate effects in two-part models. Such an investigation can provide important information on the mechanisms underlying CAC development. We propose a novel approach, which consists of penalized maximum likelihood estimation and a step-wise hypothesis testing procedure to determine proportionality. Simulation shows satisfactory performance of the proposed approach. Analysis of MESA suggests that proportionality holds for all covariates except LDL and HDL.

Keywords and phrases: Two-part models, Proportionality, Semiparametric estimation.

1. INTRODUCTION

The MESA (Multi-Ethnic Study of Atherosclerosis) is an ongoing study of the prevalence, risk factors, and progression of subclinical cardiovascular disease in a multi-ethnic cohort (Bild et al. 2002). It provides a valuable opportunity to study the development and progression of CAC (coronary artery calcium), which is an important risk factor for various coronary heart diseases. In MESA, the CAC is measured with the Agatston score, which is the amount of calcium at each lesion scaled by an attenuation factor and summed over all lesions. The histogram in Figure 1 shows that the CAC has a mixture distribution, with about half of the CAC scores equal to zero and the rest continuously distributed.

Data with characteristics similar to that of CAC has been referred to "zero-inflated data". Existing methods for analyzing such data include the marginal likelihood method, quasi-likelihood method (McCulloch and Searle 2001), penalized quasi-likelihood method (Yau and Lee 2001), nonparametric maximum likelihood method (Min and Agresti 2005), Bayesian method (Ghosh et al. 2006), penalized likelihood method (Ma 2009) and others. Among available models, two-part models have attracted extensive attention. Two-part models have a long history in economic, statistical, and biomedical literature. Unlike alternatives such as the promotion models (Thompson and Chhikara 2003), two-part models do not assume specific data generating mechanisms. On a special note, two-part models have been suggested as the default models for describing the CAC in MESA (http://mesa-nhlbi.org/).

In two-part models, there are two covariate effects. The focus of this study is on the determination of proportionality between them. Denote $X = (X_1, X_2, X_3)$ as the covariate. Motivated by Figure 1, we consider $Y = \log(1 + CAC)$ and the following two-part model. In the first part, assume

$$\phi^{-1}(Pr(Y > 0|X)) = h(X), \tag{1}$$

where ϕ is the link function, ϕ^{-1} is the inverse of ϕ, and $h(X)$ is the unknown covariate effect. In the second part of the model, assume

$$\text{for } Y > 0: \; Y|X = h^*(X) + \epsilon, \tag{2}$$

where $h^*(X)$ is the unknown covariate effect and ϵ is the random error.

With models (1) and (2), the two covariate effects are proportional if $h^*(X) = \tau h(X)$ with $\tau \neq 0$. In our study, a biologically meaningful result demands $\tau > 0$. In our data analysis, such a result is naturally obtained without any constraint. When the full proportionality does not hold, there can be multiple scenarios. Consider for example additive covariate effects with $h(X) = h_1(X_1) + h_2(X_2) + h_3(X_3)$. Partial proportionality holds if $h^*(X) = \tau(h_2(X_2) + h_3(X_3)) + (\tau h_1(X_1) + \tilde{h}(X_1))$ with $\tilde{h}(X_1) \neq 0$ and $\tau \neq 0$. That is, proportionality of covariate effects holds for X_2 and X_3 but not X_1. Other partial proportionality scenarios can be defined in a similar manner.

Determination of proportionality may provide a deeper understanding of CAC development. Models (1) and (2) describe the development of CAC in different ranges, with model (1) describing the development from zero to nonzero and model (2) describing the development above zero. If full proportionality holds, then the same mechanism – which

[*] Corresponding author.

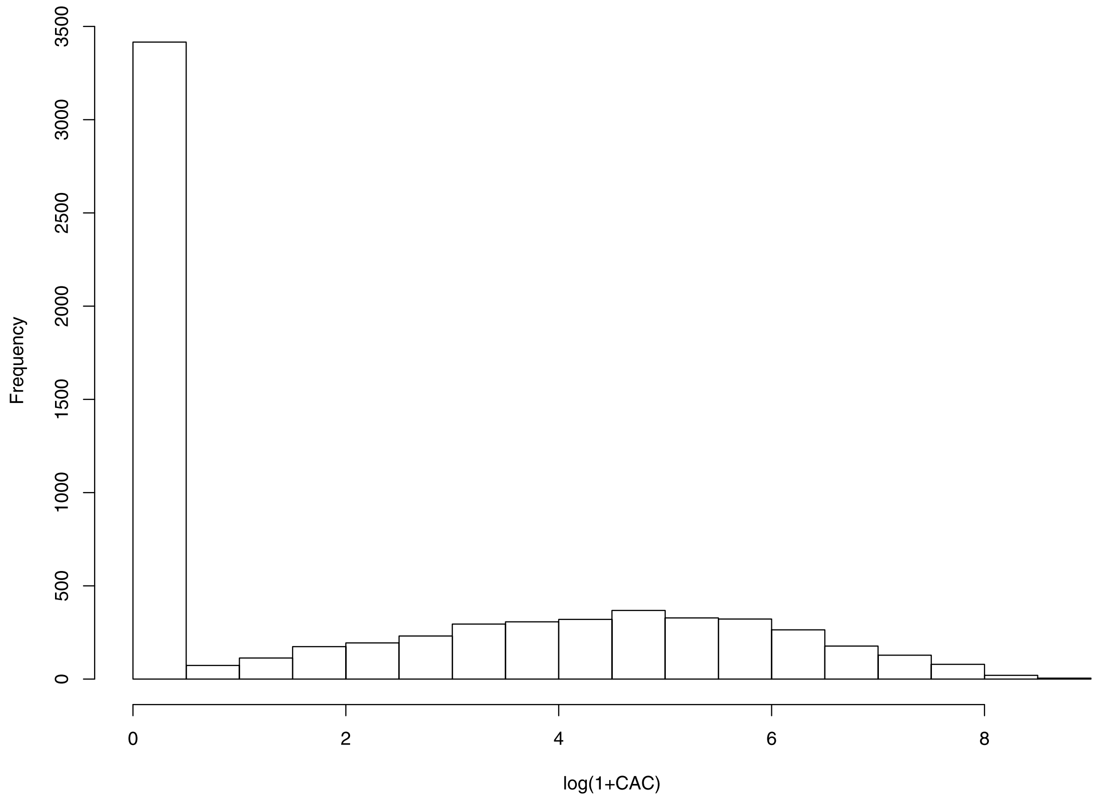

Figure 1. Analysis of MESA: histogram of $\log(1+CAC)$.

corresponds to $h(X)$ – determines development in both ranges. In contrast, under the partial proportionality described above, there may be two mechanisms. The first corresponds to $h_2(X_2) + h_3(X_3)$, which remains the same in both ranges of CAC values. In contrast, the mechanism corresponding to X_1 differs between the two ranges. We note that models (1) and (2) have different link functions and are on different scales. However, when investigating covariate effects, we are more interested in contributions of covariates relative to each other. Thus, it is meaningful to compare $h(X)$ against $h^*(X)$.

Published proportionality studies include the zero-inflated Poisson regression model in Lambert (1992) and Albert et al. (1997), logit-(log) gamma two-part model in Moulton et al. (2002), and logit-linear two-part model in Han and Kronmal (2006). These studies show that determining proportionality structure can provide insights into the biological mechanisms underlying (for example) disease development. In addition, compared with models without proportionality constraints, models with fully or partially proportional covariate effects have fewer unknown parameters and can be more accurately estimated.

The aforementioned proportionality studies have assumed parametric covariate effects. For the CAC in MESA, McClelland et al. (2006) and our analysis suggest that semiparametric models may be needed. With semiparametric two-part models, we conjecture that determination of proportionality with respect to parametric covariate effects can be achieved using the hypothesis testing approach in Han and Kronmal (2006), although such a possibility has not been investigated. On the other hand, determination of proportionality with respect to nonparametric covariate effects has not been studied.

In this article, we investigate determination of proportionality of covariate effects with semiparametric two-part models. This study advances from published literature along the following aspects. First, it advances from existing proportionality studies by adopting flexible semiparametric models. Second, the hypothesis testing approach (for determining proportionality) advances from published studies by investigating semiparametric models and adopting a stepwise approach that can accommodate multiple covariate effects. Third, this study advances from published two-part model studies by investigating different models and more importantly developing an effective approach for determining proportionality. Last, this study provides comprehensive analysis of CAC, which may help advance our understanding of the development of coronary heart diseases.

The rest of the article is organized as follows. We introduce the data and model settings in Section 2. We describe the proposed method in Section 3. We consider a penalized maximum likelihood approach for estimation and a hypothesis testing approach for determination of proportionality. We conduct simulation in Section 4 and analyze the MESA data in Section 5. The article concludes with discussion in Section 6.

2. DATA AND MODEL

Let $Y = \log(1+CAC)$. Without loss of generality, denote $X = (X_1, X_2, X_3)'$ and $Z = (Z_1, Z_2, Z_3)'$ as covariates. In the first part of the two-part model, assume that

$$\begin{aligned}(3)\quad \phi^{-1}(Pr(Y>0|X,Z)) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \\ &\quad + f_1(Z_1) + f_2(Z_2) + f_3(Z_3) \\ &= \beta'\tilde{X} + f(Z),\end{aligned}$$

where ϕ is the link function and ϕ^{-1} is its inverse. $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)'$, $\tilde{X} = (1, X')'$, and $f(Z) = f_1(Z_1) + f_2(Z_2) + f_3(Z_3)$. In the second part of the model, assume that for Y>0:

$$\begin{aligned}(4)\quad Y|X,Z &= \tau(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + f_1(Z_1) \\ &\quad + f_2(Z_2) + f_3(Z_3)) + \alpha_0 + \alpha_2 X_2 + \alpha_3 X_3 \\ &\quad + g_1(Z_1) + g_2(Z_2) + g_3(Z_3) + \epsilon \\ &= \tau(\beta'\tilde{X} + f(Z)) + \alpha'\tilde{\tilde{X}} + g(Z) + \epsilon,\end{aligned}$$

where $\alpha = (\alpha_0, \alpha_2, \alpha_3)'$, $\tilde{\tilde{X}} = (1, X_2, X_3)'$, $g(Z) = g_1(Z_1) + g_2(Z_2) + g_3(Z_3)$. For identifiability, we assume that for the "anchor" covariate X_1, $\tau\beta_1 \neq 0$; in addition, $Pf_i = Pg_i = 0$, where P is the expectation. Motivated by Figure 1, we assume $\epsilon \sim N(0, \sigma^2)$.

In (3) and (2), α, β, τ, and σ are the unknown parametric parameters. f and g are the unknown nonparametric covariate effects. Motivated by McClelland et al. (2006), we assume that f and g are smooth functions.

3. PENALIZED ESTIMATION AND DETERMINATION OF PROPORTIONALITY

We propose a penalized estimation approach and use penalized splines for nonparametric covariate effects. The equivalence between penalized spline models and mixed models has been well established (Speed 1991; Wang 1998; Wand 2003). We take advantage of this equivalence and transform the hypothesis testing on proportionality to one on fixed parameters and variance components in the corresponding mixed models. Inference is then made through the marginal likelihood of the semi-continuous data. Hypothesis testing on variance components or smoothing parameters, or more generally on nonparametric functions in semiparametric regression, has been investigated (Hardle et al. 1998; Zhang and Lin 2003; Claeskens 2004; Liu et al. 2005; Crainiceanu et al. 2005; Fan and Jiang 2007; Jose Lombardia and Sperlich 2008; Kauermann et al. 2009). We choose the likelihood ratio based test, which has been shown to be more powerful in the literature. The parametric bootstrap is used to obtain approximated null distributions.

3.1 Penalized estimation

For an observation with covariate (X, Z) and response Y, the log-likelihood function is

$$\begin{aligned}(5)\quad & l(\alpha, \beta, \tau, \sigma, f, g|X, Z) \\ &= I(Y>0)\Big\{ -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) \\ &\quad - \frac{(Y - \tau(\beta'\tilde{X} + f(Z)) - \alpha'\tilde{\tilde{X}} - g(Z))^2}{2\sigma^2}\Big\} \\ &\quad + I(Y>0)\log(\phi(\beta'\tilde{X} + f(Z))) \\ &\quad + I(Y=0)\log(1 - \phi(\beta'\tilde{X} + f(Z))).\end{aligned}$$

In this study, we set ϕ as the logit link function. Assume n iid observations. With smooth f and g, we consider the penalized maximum likelihood estimate (PMLE)

$$(6)\quad (\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g}) = argmax\left\{P_n l - \lambda_f^2 J^2(f) - \lambda_g^2 J^2(g)\right\}.$$

Here P_n is the empirical measure, λ_f and λ_g are the tuning parameters, $J^2(f) = \sum_{i=1}^3 J^2(f_i) = \sum_{i=1}^3 \int (f_i^{(s)})^2 dZ_i$ is the penalty on smoothness, and $f_i^{(s)}$ is the s^{th} derivative of f_i. In this study, we set $s = 2$.

3.2 Estimation with thin plate splines

Under the assumptions described in Appendix, we limit $\hat{f}$ and $\hat{g}$ to be spline functions. In general, the penalty in (6) not necessarily leads to a thin plate spline solution. However, when the regression function is one dimensional, the thin plate penalty (equation (4.48) on page 135 of Gu (2002)) is the same as the integrated squared second derivative penalty. This is demonstrated in Example 4.1 of Gu (2002). In our study, the functions f_is and g_is are one dimensional and we use $s = 2$. Thus, we use thin plate splines with K knots for estimation of the nonparametric covariate effects. The penalized splines we use include smoothing splines as a special case when the knots are the design points. For the development of asymptotic properties, the full basis function space (with knots at the design points) is needed. In computation, we follow common practice and take the number of knots to be smaller than the number of design points. As a limitation of this study, we do not provide theoretical justification for the validity of this approach. Of note, even though quite a few studies have used a smaller number of knots, only Kim and Gu (2004) provides a rigorous development.

For a generic function $m(x)$, its thin plate spline representation is

$$(7)\quad m(x) \approx d_0 + d_1 x + \sum_{k=1}^{K} c_k |x - p_k|^3,$$

where d_0, d_1 and c_ks are the unknown regression coefficients and p_ks are the fixed knots.

For $i = 1, 2, 3$, at the design points, we have $f_i(Z_i) = T_i d_{fi} + \Sigma_i c_{fi}$, $g_i(Z_i) = T_i d_{gi} + \Sigma_i c_{gi}$, where $T_i = (1, Z_i)$, $\Sigma_i = (|Z_i - p_{i1}|^3, \ldots, |Z_i - p_{iK}|^3)$, p_{ik}s are the knots, and $d_{fi} = (d_{0fi}, d_{1fi})'$, $d_{gi} = (d_{0gi}, d_{1gi})'$, $c_{fi} = (c_{1fi}, \ldots, c_{Kfi})'$, $c_{gi} = (c_{1gi}, \ldots, c_{Kgi})'$ are the regression coefficients. Denote $\theta = (\alpha_0,\ \alpha_2,\ \alpha_3,\ \beta_0,\ \beta_1,\ \beta_2,\ \beta_3,\ \tau,\ \sigma,\ d'_{f1},\ d'_{f2},\ d'_{f3},\ d'_{g1},\ d'_{g2},\ d'_{g3})'$ and $b = (c'_{f1},\ c'_{f2},\ c'_{f3},\ c'_{g1},\ c'_{g2},\ c'_{g3})'$. Once the knots are chosen following Wahba (1990), penalization on the smoothness is equivalent to penalization on the coefficient b. To allow further flexibility, instead of using unified λ_f and λ_g for all components of f and g, we can use different λ_{fi} and λ_{gi} for $i = 1, 2, 3$. With these notations, the penalized log-likelihood function defined in (6) can be rewritten as

$$(8)\qquad P_n l(Y|\theta, b, \sigma^2) - \sum_{i=1}^{3} \lambda_{fi}^2 c'_{fi} D_i c_{fi} - \sum_{i=1}^{3} \lambda_{gi}^2 c'_{gi} D_i c_{gi},$$

with $l(Y|\theta, b, \sigma^2) = I(Y > 0)\eta_1 - \log(1 + \exp(\eta_1)) - I(Y > 0)\left(\frac{1}{2}\log(2\pi) + \frac{1}{2}\log\sigma^2 + \frac{(Y-\eta_2)^2}{2\sigma^2}\right)$, $\eta_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \sum_{i=1}^{3}(T_i d_{fi} + \Sigma_i c_{fi})$, and $\eta_2 = \tau\eta_1 + \alpha_0 + \alpha_2 X_2 + \alpha_3 X_3 + \sum_{i=1}^{3}(T_i d_{gi} + \Sigma_i c_{gi})$. D_i is a $K \times K$ matrix with its kth row equal to $(|p_{ik} - p_{i1}|^3, \ldots, |p_{ik} - p_{iK}|^3)$.

The objective function defined in (8) is concave in both θ and b, and can be maximized using the Newton-Raphson approach.

3.2.1 Tuning parameter selection

We use a Generalized Maximum Likelihood (GML) smoothing parameter selection approach, which is built on a close connection between the penalized log-likelihood (8) and log-likelihood of a mixed model. Although the connection between penalized smoothing and mixed models has been previously observed (Speed 1991; Wang 1998; Wand 2003), we may be the first to explore this connection with semi-continuous data. The main challenge is that the likelihood function of such mixed models involves high dimensional integration, and the standard practice of using the Laplace approximation leads to biased estimates of the variance components.

First we note that the penalized log-likelihood in (8) is equivalent to the log joint likelihood of the response Y and the following random effects:

$$(9)\quad c_{fi} \sim N(0, D_i^+/\lambda_{fi}^2),\ c_{gi} \sim N(0, D_i^+/\lambda_{gi}^2),\ i = 1, 2, 3,$$

where D_i^+ is the Moore-Penrose inverse of D_i (Graybill, 2001). The equivalence is due to $c'_{hi} D_i c_{hi} = c'_{hi}(D_i^+)^+ c_{hi}$ for $h = f, g$. When the distribution of Y belongs to the exponential family, the log joint likelihood is exactly the penalized quasi-likelihood (PQL) in Breslow and Clayton (1993), which also discusses singular variance matrices of random effects and recommends the use of Moore-Penrose inverse.

If we assume a flat prior on θ, then the GML criterion estimates the smoothing parameters and σ^2 from the marginal density of Y, which is

(10)

$$\begin{aligned} &L(Y|\lambda_{f1}, \lambda_{f2}, \lambda_{f3}, \lambda_{g1}, \lambda_{g2}, \lambda_{g3}, \sigma^2) \\ &\quad = \int \exp\left(P_n l(Y|\theta, b, \sigma^2) - \sum_{i=1}^{3} l(c_{fi}) - \sum_{i=1}^{3} l(c_{gi})\right) \\ &\qquad \times d\theta dc_{f1} \cdots dc_{g3}, \end{aligned}$$

where $l(c_{fi})$ and $l(c_{gi})$ are the log-likelihood functions of the normal distributions in (9).

If $l(Y|\theta, b, \sigma^2)$ were a normal likelihood, the GML criterion would give the REML estimates of the tuning parameters, which are the inverse of the variance components in a mixed-effects model with c_{fi}s and c_{gi}s as the random effects. Under this mixed-effects model framework, alternatively, we can use a full marginal likelihood (ML) approach, which allows us to estimate the fixed effect θ together with the variance components. Here, the full marginal likelihood of Y is

(11)

$$\begin{aligned} &L(Y|\theta, \lambda_{f1}, \lambda_{f2}, \lambda_{f3}, \lambda_{g1}, \lambda_{g2}, \lambda_{g3}, \sigma^2) \\ &\quad = \int \exp\left(P_n l(Y|\theta, b, \sigma^2) - \sum_{i=1}^{3} l(c_{fi}) - \sum_{i=1}^{3} l(c_{gi})\right) \\ &\qquad \times dc_{f1} \cdots dc_{g3}. \end{aligned}$$

The REML and ML approaches are asymptotically equivalent, with the former more efficient for estimating variance components and the latter more convenient for inferences. In this study, since estimation and testing of both fixed effects and tuning parameters are of interest, we adopt the ML approach and carry out the multivariate integration in (11) using the spherical-radial quadrature algorithm (Monohan and Genz 1997).

For a generic multivariate integration $\int f(u)du$ with integration dimension d, the spherical-radial quadrature algorithm involves two steps. First the integrand $f(u)$ is transformed into an approximate spherically symmetrical function $f^*(x)$ through $f^*(x) = |B|^{-1} f(\hat{u} + B^{-1}x)$ where $\hat{u}$ and $H = B'B$ are the mode and the hessian matrix of the integrand. For the integration in (11), the integrand is a concave function with close-form gradient and hessian. The Newton-Raphson algorithm can be used to find the mode $\hat{u}$ rather quickly. After the transformation, a change of variable is performed so that the multivariate integration is now in terms of a scalar radius and a vector of length d. The second step involves evaluating the transformed integrand at predefined radial and spherical quadrature points. With the 7 point Gauss-Kronrod rule for the radius and the simplex rule by

Monahan and Genz (1997), this step needs $7(d+1)$ integrand evaluations to obtain the integral approximation. As argued by Clarkson and Zhan (2002), since our purpose is to obtain the maximum likelihood estimates, we do not need to approximate the likelihood with very high accuracy. Similar to Clarkson and Zhan (2002), we find that one application of the simplex rule (as opposed to multiple applications with rotations) is sufficient. We conduct the integration (11) on a typical desk PC and find that it takes about 0.2 second with sample size 1,000 and $d = 60$ (i.e, 10 knots for each nonparametric function). We use the *nlm* function in R (which uses a Newton-type algorithm) for optimization of (11) and find that it takes about 6 minutes in the same setting. For estimation or inference that only involves the smoothing parameters, optimizing (10) is computationally more efficient than (11) since (10) is a function of much lower dimensionality, although the integration dimension is higher.

3.3 Determination of proportionality

Determination of proportionality with respect to X_i is equivalent to testing $H_0 : \alpha_i = 0$ *vs* $H_1 : \alpha_i \neq 0, i = 2, 3$. With Z_i, determination of proportionality amounts to testing $H_0 : d_{gi} = 0, \lambda_{gi} = \infty$ *vs* $H_1 : d_{gi} \neq 0$ or $\lambda_{gi} \neq \infty, i = 1, 2, 3$.

Motivated by studies on simple linear models (Wahba 1990) and generalized linear models (Liu et al. 2005) as well as Guo (2002) and Crainiceanu et al. (2005), for both parametric and nonparametric covariate effects, we propose using the following likelihood ratio test statistic based on the ML defined in (11):

(12)

$$T_{ML} = \frac{\sup_{H_0} L(Y|\theta, \lambda_{f1}, \lambda_{f2}, \lambda_{f3}, \lambda_{g1}, \lambda_{g2}, \lambda_{g3}, \sigma^2)}{\sup_{H_0 \cup H_1} L(Y|\theta, \lambda_{f1}, \lambda_{f2}, \lambda_{f3}, \lambda_{g1}, \lambda_{g2}, \lambda_{g3}, \sigma^2)}.$$

In our study, there are multiple covariates and multiple scenarios of partial proportionality. To fully determine the proportionality structure, we use a step-wise approach. Denote A, A_P, and A_N as the index sets of all covariates, covariates with proportional effects, and covariates with nonproportional effects, respectively. Denote $|A_P|$ as the cardinality of A_P.

1. Initialize $A_P = A$;
2. For $a \in A_P$, fit an intermediate model with covariates in $A_P - \{a\}$ having proportional effects and covariates in $A_N \cup \{a\}$ having non-proportional effects. Compute the p-value for proportionality using the bootstrap approach described below.
3. Repeat Step 2 over all $a \in A_P$ and compare the $|A_P|$ p-values so obtained. Denote a^* as index of the covariate with the smallest p-value. If the smallest p-value is not significant, abort loop. Otherwise, update $A_P = A_P - \{a^*\}$ and $A_N = A_N \cup \{a^*\}$.
4. If $|A_P| = 0$, abort loop. Otherwise, repeat Steps 2 and 3.

This approach starts with all covariate effects being proportional. In Step 2, we determine the significance of proportionality of each covariate effect. In Step 3, the proportionality constraint on one covariate effect is released. Iteration is terminated once A_P cannot be further reduced. Motivated by Liu et al. (2005), we propose the following bootstrap approach to compute the significance of proportionality.

1. Fit the null and full models;
2. Generate random errors from the normal distribution with mean zero and variance $\hat{\sigma}^2$ estimated from the full model;
3. Under the null, compute the probability of $Y > 0$ from model (3) and generate the binary $I(Y > 0)$. For those with $Y > 0$, generate the continuous Y values. Here Ys are equal to the sum of the null model evaluated at the design points and the normal random errors;
4. With the generated responses, estimate the null and full models again. Compute the statistic T_{ML};
5. Repeat Steps 2 to 5 B (e.g. 500) times. An empirical p-value can then be computed.

A byproduct of the above procedure is the bootstrap confidence intervals for both the parametric and nonparametric parameters, which can serve as the basis for inference.

The likelihood ratio test and the bootstrap procedure can be computationally expensive. To calculate the likelihood ratio test statistic, we need to fit the null and full models. When fitting the full model, we suggest setting initial values as the estimates based on the null model, which may speed up the computation. The bootstrap procedure is highly parallel, which makes it computationally affordable.

3.4 Asymptotic properties

Although many intermediate models are needed in order to determine the proportionality structure, we are most interested in the final models, i.e., models with proportionality properly determined. For those models, we establish asymptotic properties of the PMLE. Sufficient conditions are provided in Appendix. Denote the true value of $(\alpha, \beta, \tau, \sigma, f, g)$ as $(\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T)$. Define $d^2((\alpha, \beta, \tau, \sigma, f, g), (\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T)) = (\alpha - \alpha_T)^2 + (\beta - \beta_T)^2 + (\tau - \tau_T)^2 + (\sigma - \sigma_T)^2 + \int (f - f_T)^2 dP_Z + \int (g - g_T)^2 dP_Z$, with P_Z denoting the distribution function of Z.

Lemma 1. *Under assumptions A1-A4 provided in Appendix,*

$$d((\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g}), (\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T)) = O_p(n^{-s/(2s+1)});\ J(\hat{f}), J(\hat{g}) = O_p(1).$$

The estimates of nonparametric covariate effects are consistent and have the optimal convergence rate. Lemma 1 also establishes that $J(\hat{f}), J(\hat{g}) = O_p(1)$. That is, $\hat{f}$ and $\hat{g}$ have the "right" order of smoothness. The L_2 consistency, together with the smoothness and compactness conditions described in Appendix, can lead to the uniform consistency of $\hat{f}$ and $\hat{g}$, i.e., $\sup|\hat{f}-f_T| = o_P(1)$ and $\sup|\hat{g}-g_T| = o_P(1)$. For the estimates of parametric parameters, we have the following results.

Lemma 2. *With assumptions and Σ specified in Appendix,*

$$\sqrt{n}\{(\hat{\alpha},\hat{\beta},\hat{\tau},\hat{\sigma}) - (\alpha_T,\beta_T,\tau_T,\sigma_T)\} \to_D N(0,\Sigma).$$

Despite the slow convergence rate of $\hat{f}$ and $\hat{g}$, the estimates of parametric parameters are still $\sqrt{n}$ consistent and asymptotically normally distributed.

4. SIMULATION

In simulation, we generate data from

$$\begin{aligned} &Pr(Y>0|X,Z) = logit(\eta_1), \\ &\text{and for } Y>0, Y|X,Z = \eta_2 + \epsilon, \end{aligned} \tag{13}$$

where $\eta_1 = -4 + 5X_1 - 2.5X_2 + 1.5X_3 + 8\sin(6Z_1) + 7Z_2 - 20(Z_2-0.5)^2$, $\tau = 0.2$, and $\sigma = 0.5$. We assume that $X_1 = 0$ or 1 with probability 0.5; $X_2 = 1, 2, 3$, or 4 with probability 0.25; $X_3 \sim N(0,1)$; Z_1 is equally spaced between 0 and 1; and $Z_2 \sim Unif[0,1]$. We set the sample size $n = 1,000$. We define the "difference function" as $\eta_2 - \tau\eta_1$. Determination of proportionality then amounts to testing if components of the difference function are equal to zero. As shown in Table 1, ten difference functions are considered. For a clear view, we omit the intercepts in Table 1, which are needed to satisfy the identifiability assumption of $Pf_i = Pg_i = 0$. In simulation, X_1 is chosen as the anchor.

We first investigate the determination of proportionality. In Table 1, we present the power of detecting non-proportionality computed based on 1,000 replicates. We can see that in general, the proposed approach can correctly identify the proportionality structure. When proportionality holds for a specific covariate, the power is usually close to 0.05, the nominal significance level. In contrast, when proportionality does not hold, the proposed approach can identify the non-proportionality with a high probability. Consider, for example, difference function $0.1X_3 + Z_2$. With probabilities 0.80 and 0.99, the non-proportionality with respect to X_3 and Z_2 can be identified. The error rates of mistakenly identifying non-proportionality with respect to X_2 and Z_1 are 0.066 and 0.032, respectively. In addition, when the regression coefficients in difference functions increase, the power increases. Consider for example difference functions $0.33X_3 + 0.5Z_2$ and $0.33X_3 + Z_2$. When the regression coefficient of Z_2 increases from 0.5 to 1, the power increases from 0.40 to 0.95.

Table 1. Simulation study: power of testing non-proportionality with various difference functions

	Power			
Difference function	X_2	X_3	Z_1	Z_2
0	0.040	0.046	0.043	0.078
$0.33X_3$	0.045	1	0.021	0.054
$5Z_1 + Z_1^2 + 0.9Z_2$	0.051	0.064	0.635	0.806
$0.8X_2 + 5Z_1 + 2Z_1^2 + 0.8Z_2$	0.900	0.046	0.820	0.620
$0.33X_3 + 5Z_1 + 10Z_1^2 + Z_2$	0.076	0.980	1	0.920
$0.33X_3 + 0.5Z_2$	0.062	1	0.033	0.400
$0.33X_3 + Z_2$	0.079	1	0.048	0.950
$0.3X_2 + 0.33X_3$	0.220	1	0.042	0.051
$0.5X_2 + 0.33X_3$	0.560	1	0.035	0.050
$0.05X_3 + Z_2$	0.045	0.160	0.038	0.960
$0.1X_3 + Z_2$	0.066	0.800	0.032	0.990

For the final models, we evaluate performance of the penalized estimation and bootstrap inference. We show a representative example of the estimation results in Figure 2, where data is generated with difference function $0.33X_3 + 5Z_1 + 10Z_1^2 + Z_2$. For the covariates with nonparametric effects, the mean estimates fit the unknown true functions well. The 95% confidence intervals provide satisfactory coverage. As expected, the confidence intervals become wider, when it is closer to the boundaries and there are fewer observations. Note that, for identifiability, we have assumed $Pf_i = Pg_i = 0$. We omit the intercepts in Table 1. The intercepts have been added back in Figure 2. We have examined estimation results for parametric parameters and found negligible biases, satisfactory convergence rates, marginal distributions close to normal, and satisfactory bootstrap coverage.

5. ANALYSIS OF MESA

The MESA is a population based, multi-center study of subclinical cardiovascular diseases. The study cohort consists of 6,814 subjects with age ranging from 45 to 84 at the baseline. Subjects with missing measurements are removed, leading to a sample size of 6,658 for downstream analysis. The CAC has a mixture distribution, with about half of the CAC scores equal to zero and the rest continuously distributed. We adopt the two-part model. In the first part, we assume the logit link function. In the second part, we study $\log(1 + CAC)$, which has a distribution close to normal.

Following McClelland et al. (2006), we consider the following predictors: gender (female is used as the reference group), race (Caucasian, African-American, Chinese, and Hispanic; Caucasian is used as the reference group), former smoker (binary indicator), current smoker (binary indicator), diabetes (binary indicator), SBP (systolic blood pressure), DBP (diastolic blood pressure), age, BMI (body mass index), LDL cholesterol, and HDL cholesterol. Among the 13 covariates, 7 are binary, which naturally correspond to parametric covariate effects. In addition, our preliminary analysis suggests linear effects for SBP and DBP. Thus, in

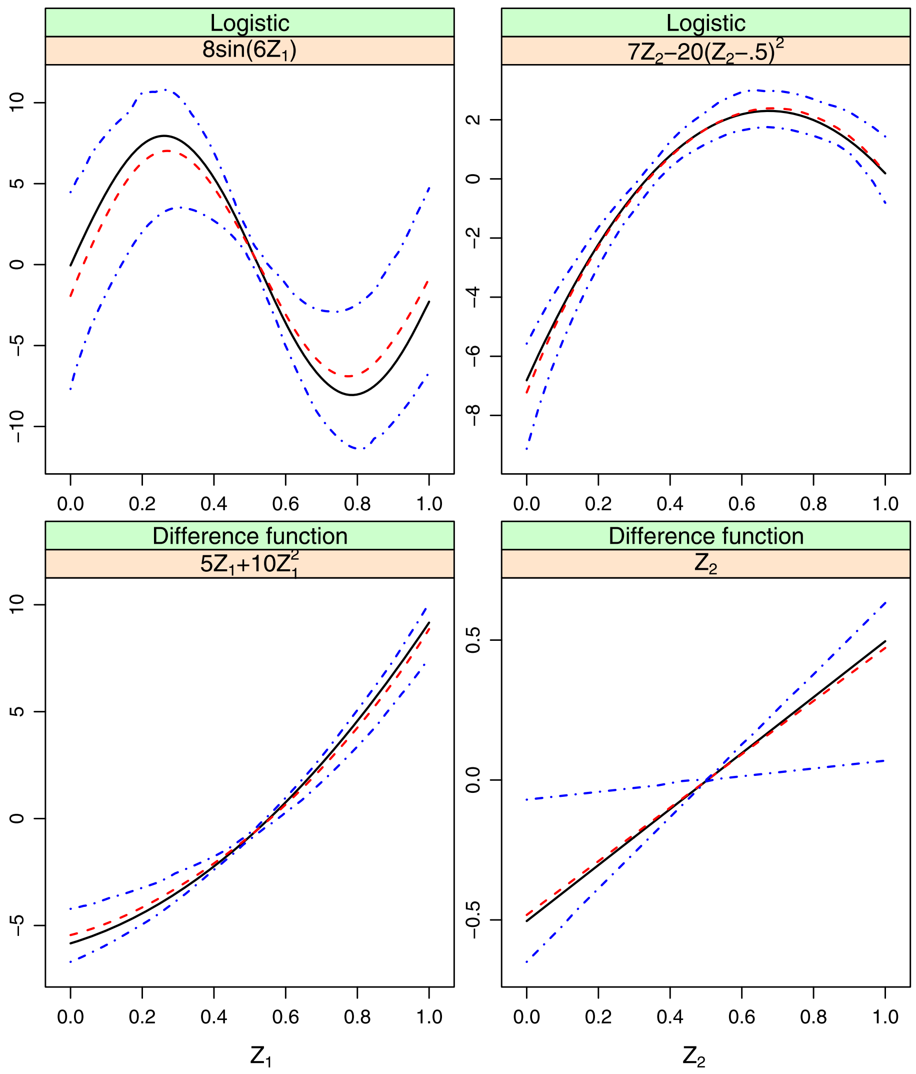

Figure 2. Simulation with difference function $0.33X_3 + 5Z_1 + 10Z_1^2 + Z_2$: estimation and inference results for nonparametric covariate effects. Solid black line: true covariate effect; Red dashed line: mean estimates; Blue dash-dotted lines: mean 95% confidence intervals.

the semiparametric models, there are 9 parametric covariate effects and 4 nonparametric ones. Following Han and Kronmal (2006), X_3 is selected as the anchor.

We use the step-wise approach to determine proportionality. In the first step, we find that the proportionality of LDL effect has the smallest p-value (< 0.001). Thus we release the proportionality constraint on LDL. In the second step, we find that the HDL effect has the smallest p-value (0.012). We then fit a model with the proportionality constraints on LDL and HDL released. In the third step, for covariates other than LDL and HDL, we find that releasing the proportionality constraints leads to insignificant p-values. We thus conclude that proportionality holds for all covariates except LDL and HDL.

For the final model with proportionality constraints on all covariates expect LDL and HDL, we present the estimates of parametric regression coefficients in Table 2 and estimates of nonparametric covariate effects in Figure 3. We find that the following risk factors are significantly associated with a higher level of CAC: being male, being Caucasian, being a smoker (both former and current), having diabetes, and having a higher level of SBP. Those findings are consistent with the literature.

For Age and BMI, their nonparametric covariate effects are proportional (Figure 3). It is interesting that their effects are almost linear, which suggests that it may be possible to further simplify the model by assuming parametric Age and BMI effects. The bootstrap confidence intervals suggest that both the Age and BMI effects are significant. Increases in Age and BMI are associated with a higher

Table 2. Analysis of MESA. Parametric regression coefficients in the full model (with no proportionality constraint) and the final model (with proportionality properly determined). Estimates (bootstrap standard errors) in the logistic (η_1) and linear (η_2) models

Predictor	Full model		Final model	
	η_1	η_2	η_1	η_2
Gender: Male (X_1)	0.945 (0.092)	0.618 (0.099)	0.960 (0.078)	0.651 (0.053)
Race: Chinese (X_2)	−0.119 (0.070)	−0.285 (0.081)	−0.211 (0.078)	−0.143 (0.053)
Race: African-American (X_3)	−0.787 (0.071)	−0.398 (0.085)	−0.727 (0.063)	−0.493 (0.047)
Race: Hispanic (X_4)	−0.628 (0.074)	−0.358 (0.073)	−0.594 (0.063)	−0.402 (0.045)
Former smoker (X_5)	0.370 (0.072)	0.213 (0.071)	0.354 (0.052)	0.240 (0.036)
Current smoker (X_6)	0.609 (0.094)	0.328 (0.096)	0.573 (0.078)	0.388 (0.052)
Diabetes (X_7)	0.243 (0.070)	0.275 (0.068)	0.299 (0.055)	0.203 (0.038)
SBP (X_8)	0.009 (0.002)	0.004 (0.002)	0.008 (0.002)	0.005 (0.001)
DBP (X_9)	−0.0034 (0.004)	0.0032 (0.004)	−0.0009 (0.004)	−0.0006 (0.002)
τ			0.678 (0.037)	
σ	1.677 (0.021)		1.680 (0.021)	

level of CAC, which is consistent with findings in the literature. For LDL and HDL, the proportionality does not hold. The shapes of covariate effects are significantly different in the two parts of the model. For HDL, its covariate effects have an "U" shape. In the literature, nonparametric modeling of HDL has not been well investigated. This study is among the first to find this interesting relationship between HDL and CAC. Implications of this finding need to be pursued in future biomedical studies. For LDL, it is interesting that the covariate effects are close to linear. Increase in LDL is associated with a higher probability of nonzero CAC, which is consistent with findings in the literature. The bootstrap confidence intervals suggest the significance of LDL effect. For nonzero CAC values, the LDL effect is negligible.

To complement the above analysis, we also fit the full model with no proportionality constraint. Estimation results are shown in Table 2 and Figure 4. Comparing the full and final models, we find that estimates in the two models are reasonably close. This is expected since estimates under both models are asymptotically consistent. An important finding is that in general, estimates in the final model have smaller variances. In Table 2, all bootstrap standard errors (except for that of X_2 in η_1) in the full model are larger than or equal to their counterparts in the final model. The improved efficiency is consistent with studies on parametric models in Han and Kronmal (2006) and others.

6. CONCLUSION

In this article, we study the semiparametric two-part modeling of the CAC in MESA. We use a penalized maximum likelihood approach for estimation and a step-wise hypothesis testing approach for determination of proportionality. Our numerical and theoretical studies show that the proposed method can properly identify the proportionality structure, and the estimation results are satisfactory.

We conduct detailed analysis of the CAC in MESA. By adopting the flexible semiparametric two-part model, this study can provide a deeper understanding of the development of CAC. Specifically, this study is among the first to find the interesting "U" shape for the effects of HDL in both parts of the model and the different shapes of the LDL effects. Assuming parametric models, Han and Kronmal (2006) conclude that the effects of HDL, LDL, diabetes, and race-Chinese are not proportional. In contrast, with the semiparametric model, we only conclude nonproportionality for HDL and LDL. Our analysis disproves the hypothesis that the change from a zero to a positive Agaston score and the change from a lower to a higher Agaston score share the same biological process. Instead, we find that risk factors affect the CAC level via at least two different mechanisms, with the cholesterol having a different mechanism from the other risk factors.

In our models, to be consistent with previous studies such as Han and Kronmal (2006) and McClelland et al. (2006), we assume additive covariate effects. We note that it is possible to extend the proposed method, accommodate interactions, and conduct analysis with transformed covariates (for example the ratio HDL/LDL). The proposed model and method have no "built-in" robustness. We suspect that the performance of the proposed method can be unsatisfactory under model misspecification. The proposed tuning parameter selection method has been motivated by several published studies. Our numerical studies show that the tuning parameters so selected have satisfactory performance. In theoretical investigation, we provide the asymptotic rate for the tuning. However, as in many other studies, it is not completely clear whether the tuning parameters selected using the proposed approach match the asymptotics. The proposed method demands an anchor covariate. The anchor is needed in many other studies that have an identifiability constraint. In theory, as long as the correspond-

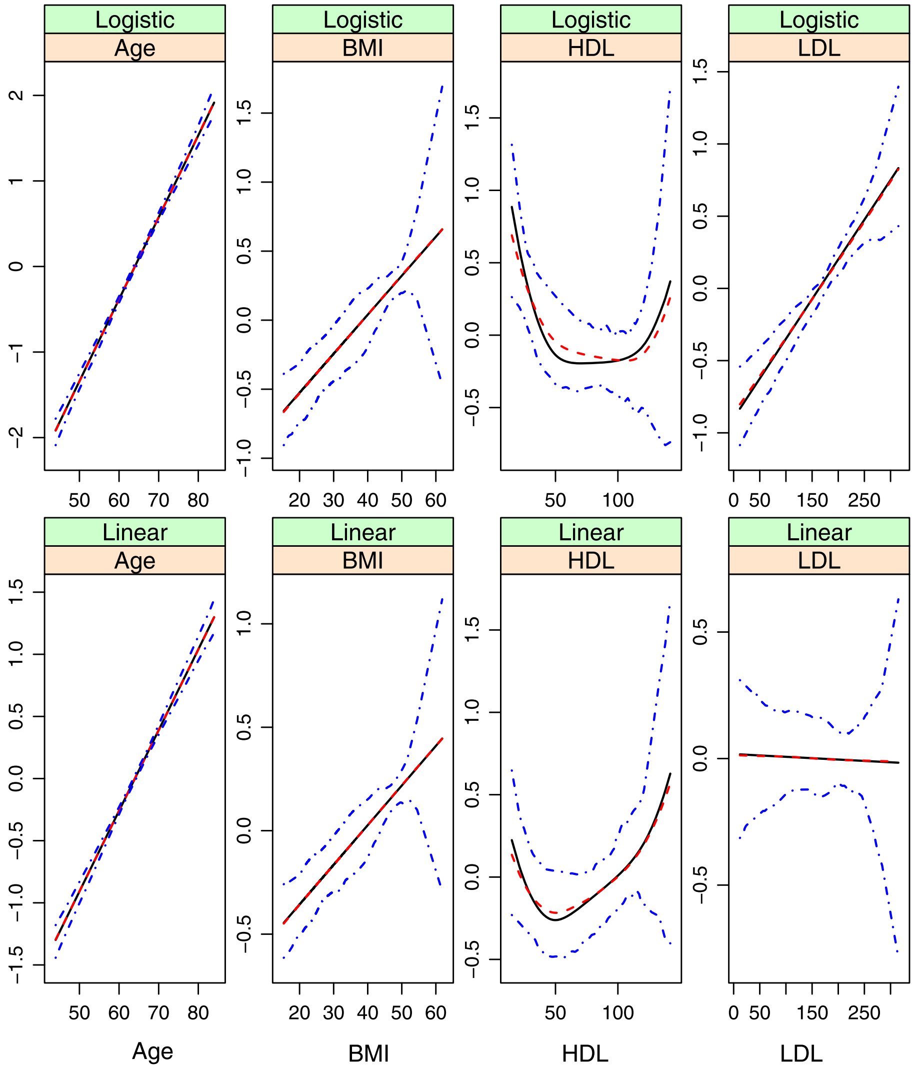

Figure 3. Analysis of MESA. Estimated nonparametric covariate effects in the **final model** *with proportionality properly determined. Solid black line: estimate; Red dashed line: mean estimate from bootstrap samples; Blue dash-dotted lines: 95% confidence intervals.*

ing covariate effect is nonzero, it does not matter which covariate is selected as the anchor; In practice, we propose selecting a covariate with a "strong" effect, which can be parametric or nonparametric. We chose a parametric covariate effect partly to follow Han and Kronmal (2006) and partly to simplify the computation. In assumption A1 (Appendix), we assume that the true value of τ is bounded away from zero. In addition, we expect the anchor variable to have a strong effect. In practical data analysis, if the estimated $\tau \times$ *effect of anchor covariate* is close to zero, it should raise alarm: either there should be no constraint or the choice of anchor is improper. Since it is not our focus, we refer to publications such as Ma and Huang (2007) and references therein for more detailed discussions on anchor.

ACKNOWLEDGEMENTS

We thank the editor, associate editor, and two referees for careful review and insightful comments. We thank the investigators, the staff, and the participants of MESA for their valuable contributions. A full list of participating MESA investigators and institutions can be found at http://www.mesa-nhlbi.org. This study has been supported by H98230-09-1-0044 from NSA (Liu), N01-HC95159 from NHLBI (Kronmal) and DMS 0805984 from NSF (Zhou and Ma).

APPENDIX

We provide proofs of lemmas 1 and 2. First, we make the following assumptions.

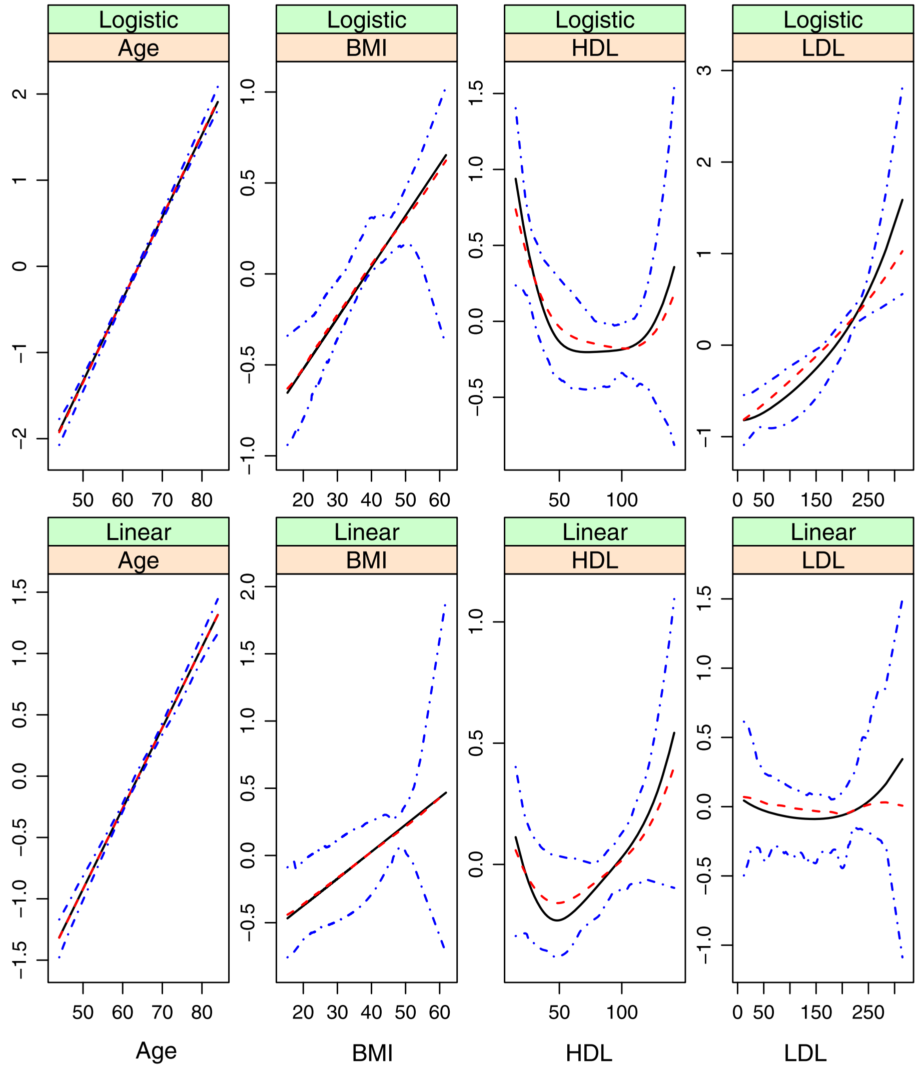

Figure 4. Analysis of MESA. Estimated nonparametric covariate effects in the **full model** *with no proportionality constraint. Solid black line: estimate; Red dashed line: mean estimate from bootstrap samples; Blue dash-dotted lines: 95% confidence intervals.*

(A1) X and Z are component-wise bounded. $(\alpha_T, \beta_T, \tau_T,\ \sigma_T)$ is an interior point of a compact set. τ_T is abounded away from 0;

(A2) Component-wise, f_T and g_T belong to the Sobolev space indexed by the order of derivative s.

(A3) $P(l(\alpha, \beta, \tau, \sigma, f, g) - l(\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T)) \leq -K_1 d^2((\alpha, \beta, \tau, \sigma, f, g), (\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T))$ with a fixed constant $K_1 > 0$.

(A4) $\lambda_f, \lambda_g = O_p(n^{-s/(2s+1)})$.

For most practical data, the compactness assumption A1 can be satisfied. We make this assumption for theoretical convenience and allow the actual bounds to remain unknown. We assume smooth nonparametric covariate effects in A2. Usually, $s = 2$. We assume that the maximizer of the likelihood function is "well-separated" in A3. This assumption can be satisfied under the compactness assumptions A1 and A2 and the differentiability of likelihood function.

Proof of Lemma 1

Definition (Bracketing number)**.** Let $(\mathbb{F}, ||\cdot||)$ be a subset of a normed space of real function h on some set. Given two functions h_1 and h_2, the bracket $[h_1, h_2]$ is the set of all functions h with $h_1 \leq h \leq h_2$. An ϵ bracket is a bracket $[h_1, h_2]$ with $||h_1 - h_2|| \leq \epsilon$. The bracketing number $N_{[]}(\epsilon, \mathbb{F}, ||\cdot||)$ is the minimum number of ϵ brackets needed to cover $\mathbb{F}$. The entropy with bracketing is the logarithm of the bracketing number.

van de Geer (2002) proves that, for the functional class

$$\tilde{\mathbb{H}} = \left\{ h : [0,1] \to [0,1], \int (h^{(s)}(x))^2 dx < 1 \right\},$$

$\log N_{[]}(\epsilon, \tilde{\mathbb{H}}, L_2(P)) \leq K_2 \epsilon^{-1/s}$, for fixed K_2 and s and all ϵ.

Under the boundedness assumptions A1 and A2 and the differentiability of the log-likelihood function, we have

$$\log N_{[]}(\epsilon, l(\alpha, \beta, \tau, \sigma, f, g), L_2(P)) \leq K_3 \epsilon^{-1/s}, \tag{14}$$

for a fixed constant K_3.

Examination of the log-likelihood suggests that if $\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma} \to \infty$, then $P_n l \to -\infty$. Thus, we are able to focus on the set of bounded $\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}$, although the actual bound remains unknown. As we optimize in the Sobolev space (indexed by the order of derivative s), $\hat{f}$ and $\hat{g}$ are smoothing splines. The proof follows Theorem 1.3.1 of Wahba (1990; p. 11).

From the definition of PMLE, we have

$$\begin{aligned} & P_n l(\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g}) - \lambda_f^2 J^2(\hat{f}) - \lambda_g^2 J^2(\hat{g}) \\ & \quad \geq P_n l(\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T) - \lambda_f^2 J^2(f_T) - \lambda_g^2 J^2(g_T). \end{aligned}$$

From the properties of likelihood function, we have

$$P l(\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g}) \leq P l(\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T).$$

Combining the above two equations, we get

$$\begin{aligned} & \lambda_f^2 J^2(\hat{f}) + \lambda_g^2 J^2(\hat{g}) + P(l(\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T) \\ & \quad - l(\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g})) \leq \lambda_f^2 J^2(f_T) + \lambda_g^2 J^2(g_T) \\ & \quad + (P_n - P)(l(\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g}) - l(\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T)). \end{aligned} \tag{15}$$

In addition, the entropy result in (14) implies that

$$\begin{aligned} & (P_n - P)(l(\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T) - l(\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g})) \\ & \quad = o_P(n^{-1/2})(1 + J(f_T) + J(g_T) + J(\hat{f}) + J(\hat{g})). \end{aligned} \tag{16}$$

Combining equations (15) and (16) with assumption A4, we have

$$\lambda_f J(\hat{f}) = o_P(1) \quad \text{and} \quad \lambda_g J(\hat{g}) = o_P(1). \tag{17}$$

Under assumption A3, equations (15) and (16) imply that

$$\begin{aligned} & K_1 d^2((\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T), (\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g})) \\ & \quad \leq o_P(1) + o_P(n^{-1/2})(1 + J(f_T) + J(g_T) + J(\hat{f}) + J(\hat{g})). \end{aligned}$$

This equation and equation (17) lead to the consistency of PMLE. To prove the rate of convergence, we use the following result.

van de Geer (2000) consider a uniformly bounded class of functions Γ, with $\sup_{\gamma \in \Gamma} |\gamma - \gamma_0|_\infty < \infty$ and a fixed $\gamma_0 \in \Gamma$, and $\log N_{[]}(\epsilon, \Gamma, P) \leq K_4 \epsilon^{-b}$ for all $\epsilon > 0$, where $b \in (0, 2)$ and K_4 is a fixed constant. Then for $\delta_n = n^{-1/(2+b)}$,

$$\sup_{\gamma \in \Gamma} \frac{|(P_n - P)(\gamma - \gamma_0)|}{||\gamma - \gamma_0||_2^{1-b/2} \vee \sqrt{n} \delta_n^2} = O_p(n^{-1/2}), \tag{18}$$

where $x \vee y = max(x, y)$.

Under the compactness assumptions A1 and A2 and considering the differentiability of log-likelihood function, we have

$$\begin{aligned} & K_1 d^2((\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g}), (\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T)) \\ & \quad \leq P(l(\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T) - l(\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g})) \\ & \quad \leq K_5 d^2((\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g}), (\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T)), \end{aligned} \tag{19}$$

where K_5 is a fixed constant. Combining equations (18) with (19) and (15), we have

$$\begin{aligned} & \lambda_f^2 J^2(\hat{f}) + \lambda_g^2 J^2(\hat{g}) \\ & \quad + K_1 d^2((\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g}), (\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T)) \\ & \quad \leq \lambda_f^2 J^2(f_T) + \lambda_g^2 J^2(g_T) + O_P(n^{-1/2})(1 + J(f_T) \\ & \quad + J(\hat{f}) + J(g_T) + J(\hat{g})) \\ & \quad \times \{ d^{1-1/2s}((\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g}), (\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T)) \\ & \quad \vee n^{\frac{1-2s}{2(2s+1)}} \}. \end{aligned} \tag{20}$$

Note that all the three terms on the left-hand side are positive. Compare each term with the right-hand side. Simple calculations give that

$$\begin{aligned} & J(\hat{f}) = O_P(1) \quad \text{and} \quad J(\hat{g}) = O_P(1), \\ & d((\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g}), (\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T)) \\ & \quad = O_P(n^{-s/(2s+1)}). \end{aligned}$$

Proof of Lemma 2

To prove the $\sqrt{n}$ consistency and asymptotic normality, we apply Theorem 1 in Ma and Kosorok (2005). Application of this theorem requires the following conditions to hold: (a) consistency and rate of convergence, which has been established in Lemma 1; (b) finite asymptotic variance, which is shown below; (c) stochastic equicontinuity, which can be established using the entropy result and the consistency result; and (d) smoothness of the model, which holds with the differentiability of likelihood function.

Thus, to prove Lemma 2, we only need to establish the non-singularity of the information matrix. Denote $\dot{l}_\alpha, \dot{l}_\beta, \dot{l}_\tau, \dot{l}_\sigma$ as the partial derivatives of the log-likelihood function with respect to $\alpha, \beta, \tau, \sigma$, respectively. For $t_f, t_g \sim 0$, consider $f_t = f + t_f \xi_f$ and $g_t = g + t_g \xi_g$, such that

f_t, g_t still satisfy assumption A2. Denote the space generated by $\xi_f \otimes \xi_g$ as $\mathbb{B}$. The score operators for f and g are $\dot{l}_f[\xi_f] = \lim_{t_f \to 0} \frac{l(\alpha,\beta,\tau,\sigma,f_t,g) - l(\alpha,\beta,\tau,\sigma,f,g)}{t_f}$ and $\dot{l}_g[\xi_g] = \lim_{t_g \to 0} \frac{l(\alpha,\beta,\tau,\sigma,f,g_t) - l(\alpha,\beta,\tau,\sigma,f,g)}{t_g}$. Denote $\dot{l}_1 = (\dot{l}_\alpha, \dot{l}_\beta, \dot{l}_\tau, \dot{l}_\sigma)'$ as the score function for the parametric parameters and $\dot{l}_{f,g}[\xi_f, \xi_g] = (\dot{l}_f[\xi_f], \dot{l}_g[\xi_g])$ as the score operator for the nonparametric parameters.

Project $\dot{l}_1$ onto the space generated by $\dot{l}_{f,g}[\xi_f, \xi_g] = (\dot{l}_f[\xi_f], \dot{l}_g[\xi_g])$. The efficient score for (α, β, τ) is $U = \dot{l}_1 - \dot{l}_{f,g}[\frac{P(\dot{l}_1)\dot{l}_{f,g}|Z}{P(\dot{l}^{\otimes}_{f,g}|Z)}]$. We further assume

(A5) $P(U'U)$ is component-wise bounded and positive definite.

Then $\Sigma = P^{-1}(U'U)$ is the asymptotic variance matrix.

Received 10 November 2010

REFERENCES

Albert, P.S., Follmann, D.A., and Barnhart, H.X. (1997). A generalized estimating equation approach for modeling random length binary vector data. *Biometrics*. **53** 1116–1124.

Bild, D.E., Bluemke, D.A., Burke, G.L., Detrano, R. et al. (2002). Multi-ethnic study of atherosclerosis: objectives and design. *American Journal of Epidemiology*. **156** 871–881.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*. **88** 9–25.

Claeskens, G. (2004). Restricted likelihood ratio lack-of-fit tests using mixed spline models. *Journal of The Royal Statistical Society, B*. **66** 909–926. MR2102472

Clarkson, D.B. and Zhan, Y. (2002). Using spherical-radial quadrature to fit generalized linear mixed effects models. *Journal of Computational and Graphical Statistics*. **11** 639–659. MR1938448

Crainiceanu, C., Ruppert, D., Claeskens, G. and Wand, M. (2005). Exact likelihood ratio tests for penalised splines. *Biometrika*. **92** 91–103. MR2158612

Fan, J. and Jiang, J. (2007). Nonparametric inference with generalized likelihood ratio tests. *TEST*. **16** 409–444. MR2365172

Ghosh, S.K., Mukhopadhyay, P. and Lu, J.C. (2006). Bayesian analysis of zero-inflated regression models. *Journal of Statistical Planning and Inference*. **136** 1360–1375. MR2253768

Graybill, F.A. (2001). *Matrices With Applications in Statistics*. Duxbury Press.

Gu, C. (2002) *Smoothing Spline ANOVA Models*. Springer. MR1876599

Guo, W. (2002). Inference in smoothing spline analysis of variance. *Journal of The Royal Statistical Society, B*. **64** 887–898. MR1979393

Han, C. and Kronmal, R.A. (2006). Two-part models for analysis of Agatston scores with possible proportionality constraints. *Communications in Statistics–Theory and Methods*. **35** 99–111. MR2274003

Hardle, W., Mammen, E. and Muller, M. (1998). Testing parametric versus semiparametric modeling in generalized linear models. *Journal of The American Statistical Association*. **93** 1461–1474. MR1666641

Jose Lombardia, M. and Sperlich, S. (2008). Semiparametric inference in generalized mixed effects models. *Journal of The Royal Statistical Society, B*. **70** 913–930. MR2530323

Kauermann, G., Claeskens, G. and Opsomer, J. D. (2009). Bootstrapping for penalized spline regression. *Journal of Computational and Graphical Statistics*. **18** 126–146. MR2649641

Kim, Y.J. and Gu, C. (2004) Smoothing spline Gaussian regression: More scalable computation via efficient approximation. *Journal of the Royal Statistical Society, Ser. B*. **66** 337–356. MR2062380

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*. **34** 1–14.

Liu, A., Meiring, W. and Wang, Y. (2005). Testing generalized linear models using smoothing spline methods. *Statistica Sinica*. **15** 235–256. MR2125730

Ma, S. (2009). Cure model with current status data. *Statistica Sinica*. **19** 233–249. MR2487887

Ma, S. and Huang, J. (2007). Combining multiple markers for classification using ROC. *Biometrics*. **63** 751–757. MR2395712

Ma, S. and Kosorok, M.R. (2005). Robust semiparametric M-estimation and the weighted bootstrap. *Journal of Multivariate Analysis*. **96** 190–217. MR2202406

McClelland, R.L., Chung, H., Detrano, R., Post, W. and Kronmal, R.A. (2006). Distribution of coronary artery calcium by race, gender, and age. Results from the Multi-Ethnic Study of Atherosclerosis (MESA). *Circulation*. **113** 30–37.

McCulloch, C.E. and Searle, S.R. (2001). *Generalized, linear, and mixed models*. New York, Chichester: John Wiley & Sons. MR1884506

Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modeling*. **5** 1–19. MR2133525

Monahan, J. and Genz, A. (1997). Spherical-radial integration rules for a Bayesian computation. *Journal of the American Statistical Association*. **92** 664–674.

Moulton, L.H., Curriero, F.C. and Barroso, P.F. (2002). Mixture models for quantitative HIV RNA data. *Statistical Methods in Medical Research*. **11** 317–325.

Ruppert, D., Wand, M.P. and Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge University Press. MR1998720

Speed, T. (1991) Comment: That BLUP is a good thing: the estimation of random effects by G.K. Robinson. *Statistical Science*. **6** 42. MR1108815

Thompson, L.A. and Chhikara, R.S. (2003). A Bayesian cure rate model for repeated measurements and interval censoring. *Proceedings of JSM 2003*.

van de Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge Series in Statistical and Probabilistic Mathematics.

Wahba, G. (1990). *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM. MR1045442

Wand, M.P. (2003). Smoothing and mixed models. *Computational Statistics*. **18** 223–249.

Wang, Y. (1998) Mixed effects smoothing spline analysis of variance. *JRSSB*. **60** 159-174. MR1625640

Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC. MR2206355

Yau, K.K. and Lee, A.H. (2001). Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. *Statistics in Medicine*. **20** 2907–2920.

Zhang, D. and Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics*. **4** 57–74.

Anna Liu
Department of Mathematics and Statistics
University of Massachusetts

Richard Kronmal
Department of Biostatistics
University of Washington

Xiaohua Zhou
Department of Biostatistics
University of Washington

Biostatistics Unit
HSR&D Center of Excellence
Veterans Affairs Puget Sound Health Care System

Shuangge Ma
School of Public Health
Yale University
E-mail address: shuangge.ma@yale.edu