The Canadian Journal of Statistics
Vol. 32, No. 4, 2004, Pages 373–386
La revue canadienne de statistique

Goodness-of-fit methods for matched case-control studies

Patrick J. ARBOGAST and Danyu Y. LIN

Key words and phrases: Conditional likelihood; cumulative residual; link function; logistic regression; model misspecification; regression diagnostic; residual.

MSC 2000: Primary 62J20; secondary 62M07.

Abstract. The authors propose graphical and numerical methods for checking the adequacy of the logistic regression model for matched case-control data. Their approach is based on the cumulative sum of residuals over the covariate or linear predictor. Under the assumed model, the cumulative residual process converges weakly to a centered Gaussian limit whose distribution can be approximated via computer simulation. The observed cumulative residual pattern can then be compared both visually and analytically to a certain number of simulated realizations of the approximate limiting process under the null hypothesis. The proposed techniques allow one to check the functional form of each covariate, the logistic link function as well as the overall model adequacy. The authors assess the performance of the proposed methods through simulation studies and illustrate them using data from a cardiovascular study.

Méthodes d'adéquation pour les études cas-contrôle appariées

Résumé : Les auteurs proposent des méthodes graphiques et numériques pouvant servir à juger de l'adéquation d'un modèle de régression logistique pour des données cas-contrôle appariées. Leur approche fait intervenir le cumul des résidus par rapport à la covariable ou au prédicteur linéaire. Lorsque le modèle est bon, le processus des résidus cumulés converge faiblement vers une limite gaussienne centrée dont la loi peut être approchée par voie de simulation. Le patron observé des résidus cumulés peut alors être comparé visuellement et analytiquement à un certain nombre de réalisations simulées du processus limite approximatif sous l'hypothèse nulle. Les techniques proposées permettent de vérifier la forme fonctionnelle de chaque covariable, le lien logistique, ainsi que l'adéquation générale du modèle. Les auteurs en évaluent la performance au moyen de simulations et en illustrent l'application sur des données issues d'une étude cardiovasculaire.

1. INTRODUCTION

Case-control studies are commonly used to investigate the relationship between disease and exposure. Often, there exist risk factors that confound this association. The effects of confounding can be adjusted by multiple regression during the analysis or via stratification at the design stage. The advantage of the latter is that inefficiencies due to too many or too few subjects per stratum are avoided (Breslow & Day 1980, ch. 5). The matched case-control study design employs very fine stratification in which one or more controls are matched to each case according to the case's values of the matching variables.

Normally, matched case-control data are formulated by a highly stratified logistic regression model in which the differences among the matched sets are represented by the stratum-specific intercept terms while the effects of exposure and nonmatching confounders are represented by a common set of slope parameters. The conditional likelihood is then used to eliminate the intercept terms. The resulting estimators of the slope parameters are consistent and asymptotically normal (Breslow & Day 1980, ch. 7).

Diagnostic measures for detecting outliers and influential subjects on matched sets have been studied by Pregibon (1984), Moolgavkar, Lustbader & Venzon (1984, 1985), Bedrick & Hill (1996), and Hosmer & Lemeshow (2000, ch. 7). However, there does not exist any method for assessing the adequacy of the functional form of a covariate or the logistic link function, and

neither is there any omnibus goodness-of-fit test for determining the overall model adequacy.

Our interest in the model checking techniques was partly motivated by the Cardiac Arrest Blood Study (CABS), a population-based matched case-control study investigating the effects of various exposures on the the risk of primary cardiac arrest (Siscovick et al. 1995). One exposure of interest is alcohol consumption. The medical literature suggests a nonlinear relationship between alcohol consumption and risk of heart disease. Specifically, compared to nondrinkers, low-to-moderate drinking is associated with a reduced risk of heart disease, whereas heavy drinkers may experience an increased risk (e.g., Goldberg, Hahn & Parkes 1995). We wish to determine the true functional form for the relationship between alcohol consumption and risk of primary cardiac arrest on the basis of the CABS data.

In this paper, we develop a class of graphical and numerical methods for assessing the adequacy of individual components of the logistic regression model (e.g., the functional form of a covariate or the logistic link function) as well as the overall model adequacy for matched case-control data. Our methods are based on the cumulative residual process, and are similar to residual process methods that have been developed in other regression settings (Su & Wei 1991; Lin, Wei & Ying 1993, 2002). The proposed methods are presented in the next section. In Section 3, simulation results on the performance of the proposed methods are reported. In Section 4, the proposed methods are applied to the aforementioned Cardiac Arrest Blood Study.

2. METHODS

2.1. Preliminaries.

Let Y denote disease status, taking values 1 for cases and 0 for controls, and let $X \equiv (X_1, \ldots, X_p)^\top$ be a $p \times 1$ vector of covariates. Suppose that N matched sets are sampled, the ith set consisting of one case and M_i controls ($i = 1, \ldots, N$). Assume that, for the ith matched set, Y given X has a Bernoulli distribution with success probability

$$\mathrm{P}(Y = 1 \mid X; \alpha_i, \beta) = \frac{e^{\alpha_i + \beta^\top X}}{1 + e^{\alpha_i + \beta^\top X}}. \tag{1}$$

Without loss of generality, let X_{i0} denote the value of X for the case and X_{ij} ($j = 1, \ldots, M_i$) the values of X for the controls in the ith matched set. Suppose that it is known that the $M_i + 1$ covariate vectors X_{ij} ($j = 0, 1, \ldots, M_i$) are observed in the ith matched set, but it is not known which of these corresponds to the case. Under model (1), the conditional probability that X_{i0} corresponds to the case, as observed, and X_{ij} ($j = 1, \ldots, M_i$) corresponds to controls is

$$\frac{e^{\beta^\top X_{i0}}}{\sum_{\ell=0}^{M_i} e^{\beta^\top X_{i\ell}}}.$$

The conditional likelihood is

$$L(\beta) = \prod_{i=1}^{N} \frac{e^{\beta^\top X_{i0}}}{\sum_{\ell=0}^{M_i} e^{\beta^\top X_{i\ell}}},$$

which can be expressed as

$$L(\beta) = \prod_{i=1}^{N} \prod_{j=0}^{M_i} \mu_{ij}^{Y_{ij}},$$

where Y_{ij} denotes the disease status, i.e., the value of Y, for the jth subject in the ith matched set, and

$$\mu_{ij} = \frac{e^{\beta^\top X_{ij}}}{\sum_{\ell=0}^{M_i} e^{\beta^\top X_{i\ell}}}.$$

The score function and information matrix can be written, respectively, as

$$U(\beta) = \sum_{i=1}^{N}\sum_{j=0}^{M_i}(Y_{ij} - \mu_{ij})X_{ij}, \tag{2}$$

and

$$\mathcal{I}(\beta) = \sum_{i=1}^{N}\sum_{j=0}^{M_i}\mu_{ij}\widetilde{X}_{ij}^{\otimes 2},$$

where

$$\widetilde{X}_{ij} = X_{ij} - \sum_{\ell=0}^{M_i}\mu_{i\ell}X_{i\ell}$$

and $a^{\otimes 2} = aa^{\top}$. Denote the solution to $U(\beta) = 0$ by $\widehat{\beta}$.

Note that μ_{ij} is the conditional probability of $Y_{ij} = 1$ given that there is only one case among the $M_i + 1$ subjects in the ith matched set. Thus, the right-hand side of (2) is a sum of N independent zero-mean random vectors. It then follows from standard asymptotic arguments that $N^{1/2}(\widehat{\beta}-\beta)$ converges in distribution to a zero-mean normal random vector with covariance matrix Ω^{-1}, where $\Omega = \lim_{N\to\infty} N^{-1}\mathcal{I}(\beta)$.

2.2. Residuals.

Residuals normally take the form of the observed values minus the predicted values of the response. Let us consider

$$r_{ij} = Y_{ij} - \mu_{ij}, \quad j = 0, 1, \ldots, M_i;\ i = 1, \ldots, N.$$

As mentioned above, μ_{ij} is the expected value of Y_{ij} under the matched case-control sampling. Clearly, $\mathrm{E}(r_{ij}) = 0$ and $\mathrm{cov}(r_{ij}, r_{\ell m}) = 0$ for $i \neq \ell$. Thus, we define the residuals as

$$\widehat{r}_{ij} = Y_{ij} - \widehat{\mu}_{ij}, \quad j = 0, 1, \ldots, M_i;\ i = 1, \ldots, N,$$

where $\widehat{\mu}_{ij}$ is μ_{ij} with β replaced by $\widehat{\beta}$. The residual $\widehat{r}_{ij}$ is the difference between the observed disease status and the estimated conditional probability of disease. The $\widehat{r}_{ij}$ behave like ordinary residuals in that

$$\sum_{i=1}^{N}\sum_{j=0}^{M_i}\widehat{r}_{ij} = 0,$$

and, for large N, $\mathrm{E}(\widehat{r}_{ij}) \approx 0$, and $\mathrm{cov}(\widehat{r}_{ij}, \widehat{r}_{\ell m}) \approx 0$ for $i \neq \ell$. We intend to develop model-checking techniques on the basis of these residuals.

2.3. Functional forms of covariates.

A common approach to assessing the functional form of a covariate in other statistical models is to plot the residuals versus the covariate (e.g., McCullagh & Nelder 1989, ch. 12). A similar plot can be constructed for model (1) on the basis of the $\widehat{r}_{ij}$. However, this approach is exploratory and subject to interpretation: it is difficult to determine whether a seemingly unusual residual pattern reflects a faulty functional form or natural variation. Furthermore, such plots are uninformative for binary data because all the points lie on one of the two curves according as $Y = 0$ or $Y = 1$. To avoid these problems, we propose to use the cumulative sum of the $\widehat{r}_{ij}$ over the covariate of interest to check its functional form.

Let X_{ijk} denote the kth component of X for the jth subject in the ith matched set. Consider the following stochastic process:

$$W_k(t;\widehat{\beta}) = N^{-1/2}\sum_{i=1}^{N}\sum_{j=0}^{M_i}\widehat{r}_{ij}I(X_{ijk} \le t),$$

which is the cumulative sum of the residuals $\widehat{r}_{ij}$ over the values of the kth covariate component X_k. Under the null hypothesis $\mathcal{H}_0$ that (1) is correct, $W_k(t;\widehat{\beta})$ fluctuates about 0 as t varies. We show in the Appendix that, under $\mathcal{H}_0$, $W_k(t;\widehat{\beta})$ converges weakly to a zero-mean Gaussian process whose distribution can be approximated by

$$\widehat{W}_k(t;\widehat{\beta}) = N^{-1/2}\sum_{i=1}^{N} Z_i \sum_{j=0}^{M_i} \widehat{r}_{ij}\left[I(X_{ijk} \le t) + \widehat{\eta}_k^{\top}(t;\widehat{\beta})\{N^{-1}\mathcal{I}(\widehat{\beta})\}^{-1}\widehat{X}_{ij}\right],$$

where

$$\widehat{\eta}_k(t;\beta) = N^{-1/2}\partial W_k(t;\beta)/\partial\beta = -N^{-1}\sum_{i=1}^{N}\sum_{j=0}^{M_i}\mu_{ij}\widetilde{X}_{ij}I(X_{ijk} \le t),$$

$Z_1, \ldots, Z_N$ are independent standard normal random variables, and

$$\widehat{X}_{ij} = X_{ij} - \sum_{\ell=0}^{M_i}\widehat{\mu}_{i\ell}X_{i\ell}.$$

To assess whether the observed pattern of $W_k(t;\widehat{\beta})$ is abnormal, we plot it along with a few, say 20, realizations of $\widehat{W}_k(t;\widehat{\beta})$. The process $\widehat{W}_k(t;\widehat{\beta})$ can be generated by taking repeated random samples of $(Z_1, \ldots, Z_N)$ while holding the observed data (Y_{ij}, X_{ij}) $(j = 0, 1, \ldots, M_i;$ $i = 1, \ldots, N)$ fixed.

Numerical tests can be constructed as well. Since $W_k(t;\widehat{\beta})$ fluctuates about zero under $\mathcal{H}_0$, we consider the supremum statistic

$$G_k \equiv \sup_{t\in\mathbb{R}}|W_k(t;\widehat{\beta})|.$$

Let g_k be the observed value of G_k. An unusually large value of g_k would suggest that the functional form of X_k is inappropriate. To determine if g_k is too large, we compute $\mathrm{P}(G_k \ge g_k)$, which can be approximated with $\mathrm{P}(\widehat{G}_k \ge g_k)$, where $\widehat{G}_k = \sup_{t\in\mathbb{R}}|\widehat{W}_k(t;\widehat{\beta})|$. In turn, $\mathrm{P}(\widehat{G}_k \ge g_k)$ can be estimated via simulation by generating a large number (e.g., 1000 or 10000) of realizations of $\widehat{W}_k(t;\widehat{\beta})$. In the Appendix, we show that the test based on G_k is generally consistent against misspecification of the functional form of X_k.

2.4. Link function.

Another source of model misspecification is the logistic link function relating $\{\mathrm{P}(Y = 1 \mid X;\alpha_i,\beta)\}$ to $\alpha_i + \beta^{\top}X$. For instance, the true link may be the complementary log-log function. To assess the adequacy of the logistic link function, we consider

$$W_\rho(t;\widehat{\beta}) = N^{-1/2}\sum_{i=1}^{N}\sum_{j=0}^{M_i}\widehat{r}_{ij}I(\widehat{\beta}^{\top}X_{ij} \le t),$$

which is the same as $W_k(t;\widehat{\beta})$ except that the residuals are summed over the values of $\widehat{\beta}^{\top}X$ instead of X_k. We show in the Appendix that, under $\mathcal{H}_0$, $W_\rho(t;\widehat{\beta})$ converges weakly to the same limiting zero-mean Gaussian process as

$$\widehat{W}_\rho(t;\widehat{\beta}) = N^{-1/2}\sum_{i=1}^{N} Z_i \sum_{j=0}^{M_i}\widehat{r}_{ij}\left[I(\widehat{\beta}^{\top}X_{ij} \le t) + \widehat{\eta}_\rho^{\top}(t;\widehat{\beta})\{N^{-1}\mathcal{I}(\widehat{\beta})\}^{-1}\widehat{X}_{ij}\right],$$

where

$$\widehat{\eta}_\rho(t;\beta) = -N^{-1}\sum_{i=1}^{N}\sum_{j=0}^{M_i}\mu_{ij}\widetilde{X}_{ij}I(\beta^{\top}X_{ij} \le t).$$

As in Section 2.3, graphical and numerical procedures can be constructed to assess whether the observed pattern of $W_\rho(t;\widehat{\beta})$ is unusual. The supremum test statistic $G_\rho \equiv \sup_{t\in\mathbb{R}} |W_\rho(t;\widehat{\beta})|$ is shown in the Appendix to be generally consistent against misspecification of the logistic link function.

2.5. Overall model adequacy.

To assess the overall adequacy of model (1), we consider

$$W_0(x;\widehat{\beta}) = N^{-1/2}\sum_{i=1}^{N}\sum_{j=0}^{M_i}\widehat{r}_{ij}I(X_{ij}\le x),$$

where $x = (x_1,\ldots,x_p)^\top$, and $I(X_{ij}\le x)$ is the indicator function for the event that all p components of X_{ij} are no larger than the corresponding components of x.

Note that $W_0(x;\widehat{\beta})$ is a multiparameter process. It is shown in the Appendix that, under $\mathcal{H}_0$, $W_0(x;\widehat{\beta})$ converges weakly to the same zero-mean Gaussian process as

$$\widehat{W}_0(x;\widehat{\beta}) = N^{-1/2}\sum_{i=1}^{N} Z_i \sum_{j=0}^{M_i}\widehat{r}_{ij}\left[I(X_{ij}\le x) + \widehat{\eta}_0^\top(x;\widehat{\beta})\{N^{-1}\mathcal{I}(\widehat{\beta})\}^{-1}\widehat{X}_{ij}\right],$$

where

$$\widehat{\eta}_0(x;\beta) = -N^{-1}\sum_{i=1}^{N}\sum_{j=0}^{M_i}\mu_{ij}\widetilde{X}_{ij}I(X_{ij}\le x).$$

Since $W_0(x;\widehat{\beta})$ is multiparameter, it is difficult to graphically assess whether the observed pattern of $W_0(x;\widehat{\beta})$ is unusual. However, the supremum test statistic $G_0 \equiv \sup_{x\in\mathbb{R}^p}|W_0(x;\widehat{\beta})|$ can be used. The P-value can again be estimated via simulation. In the Appendix, we show that this test is consistent against any departures from model (1).

3. SIMULATION STUDIES

Extensive simulation studies were conducted to evaluate the performance of the goodness-of-fit methods described in Section 2. Disease incidence was generated from model (1). Matching was created by a Uniform(20,70) random variable grouped into intervals of length 5 (i.e., 20–25, . . ., 65–70). This mimics age, a risk factor commonly used for matching. A control was matched to a case if he/she belonged to the same (age) group. We sampled N= 25, 50, 100, and 300 matched sets, each set consisting of one case and three controls. For each simulation setting, 1000 matched case-control samples were generated. For each sample, we performed supremum goodness-of-fit tests based on $W_k(t;\widehat{\beta})$, $W_\rho(t;\widehat{\beta})$ and $W_0(x;\widehat{\beta})$. The nominal significance level for each test was set at 0.05, and the empirical size and power were estimated. The null hypothesis $\mathcal{H}_0$ was rejected if the observed supremum statistic exceeded the 95th percentile for the supremum of the approximating distribution. The percentile was estimated from 1000 realizations.

In one series of studies, we set $X = (X_1, X_2, X_2^2)$, where X_1 is Bernoulli with success probability 0.4, and X_2 has a normal distribution with unit variance and mean of 5 if $X_1 = 1$, and 4 if $X_1 = 0$. The dependence of the mean of X_2 on X_1 creates a confounding effect in that X_1 is related to X_2 and independently related to Y; this reflects the common situation in medical studies in which the relationship between an exposure of interest and outcome is influenced by the presence of a confounder. We set $\beta = (0.5, -0.25, \beta_3)$, where $\beta_3 = (0.05, 0.1, 0.20, 0.25)$. This generates models where the lack of linearity becomes progressively more pronounced. The α_i were chosen so that between 0.1% and 0.2% of the simulated population were cases. This represents the type of population in which case-control studies are typically conducted. To estimate the size of each test, the data were fit using X. To estimate the power, the data were fit

omitting X_2^2. For comparison, the Wald statistic for testing $\beta_3 = 0$ was also evaluated. The simulation results are summarized in Table 1. Note that G_2 is the test statistic for assessing the functional form of X_2.

TABLE 1: Simulation results for the sizes and powers of the supremum tests under $X = (X_1, X_2, X_2^2)$.

		Wald	G_2		G_ρ		G_0	
β_3	N	power	size	power	size	power	size	power
0.05	25	0.056	0.055	0.050	0.038	0.050	0.051	0.049
	50	0.086	0.054	0.060	0.049	0.064	0.055	0.055
	100	0.120	0.043	0.055	0.043	0.066	0.040	0.060
	300	0.253	0.043	0.110	0.052	0.059	0.046	0.110
0.1	25	0.107	0.053	0.064	0.046	0.049	0.051	0.065
	50	0.195	0.060	0.088	0.046	0.079	0.056	0.084
	100	0.355	0.044	0.128	0.050	0.086	0.041	0.129
	300	0.769	0.047	0.393	0.055	0.160	0.045	0.386
0.2	25	0.370	0.049	0.165	0.046	0.121	0.044	0.158
	50	0.677	0.054	0.351	0.036	0.233	0.056	0.322
	100	0.932	0.053	0.625	0.061	0.442	0.049	0.611
	300	1.0	0.033	0.986	0.048	0.840	0.033	0.986
0.25	25	0.599	0.049	0.265	0.036	0.179	0.042	0.268
	50	0.892	0.056	0.555	0.046	0.434	0.053	0.547
	100	0.993	0.047	0.830	0.052	0.749	0.043	0.870
	300	1.0	0.034	1.0	0.046	0.987	0.035	1.0

The supremum tests have proper sizes and good powers. Note that the Wald test is optimal in testing extra parameters in embedded parametric models. Unlike the supremum tests, however, the Wald procedure cannot be used to assess which functional form of X_2 is more appropriate or whether the chosen functional form is satisfactory.

To illustrate the graphical method, we consider simulated data of 300 matched sets generated from X in which $\beta = (0.5, -0.25, 0.25)$, but the data are fit omitting X_2^2. Figure 1 contains a plot of the observed cumulative residuals as well as 20 realizations from the approximating null distribution.

The P-value for the supremum test for the functional form of X_2 is less than 0.001, indicating that modelling X_2 as a linear term is inappropriate. When the simulated data set is refit including X_2^2, the P-value for the supremum test for the functional form of X_2 jumps to 0.976. When the true functional form of a covariate is not known, the pattern of the observed cumulative residual process, such as depicted in Figure 1, can provide insight into the correct functional form. Lin, Wei & Ying (2002) discuss interpretating cumulative residual plots for generalized linear models. Our cumulative residual plots can be interpreted in a similar manner.

4. CARDIAC ARREST BLOOD STUDY

The proposed goodness-of-fit methods were applied to data from the Cardiac Arrest Blood Study (CABS) mentioned in Section 1. The study was conducted by the Cardiovascular Health Research Unit at the University of Washington. Cases were out-of-hospital primary cardiac arrests (PCAs) in the King County of the state of Washington between October 1988 and June 1994. Controls residing in King County were selected by random digit dialing. Cases and controls

were required to be married and without a history of clinically recognized heart disease or life-threatening comorbidity. Between one and three controls were matched to each case based on gender and age within 7 years. In-person interviews were used to collect data from spouses of 362 eligible cases and 581 eligible controls. Subjects were interviewed on their habitual alcohol consumption during the prior year. Data were collected on the usual frequency and quantity of the following types of alcohol consumed: beer, light beer, wine, and distilled spirits. The average number of grams of alcohol consumed per day was calculated.

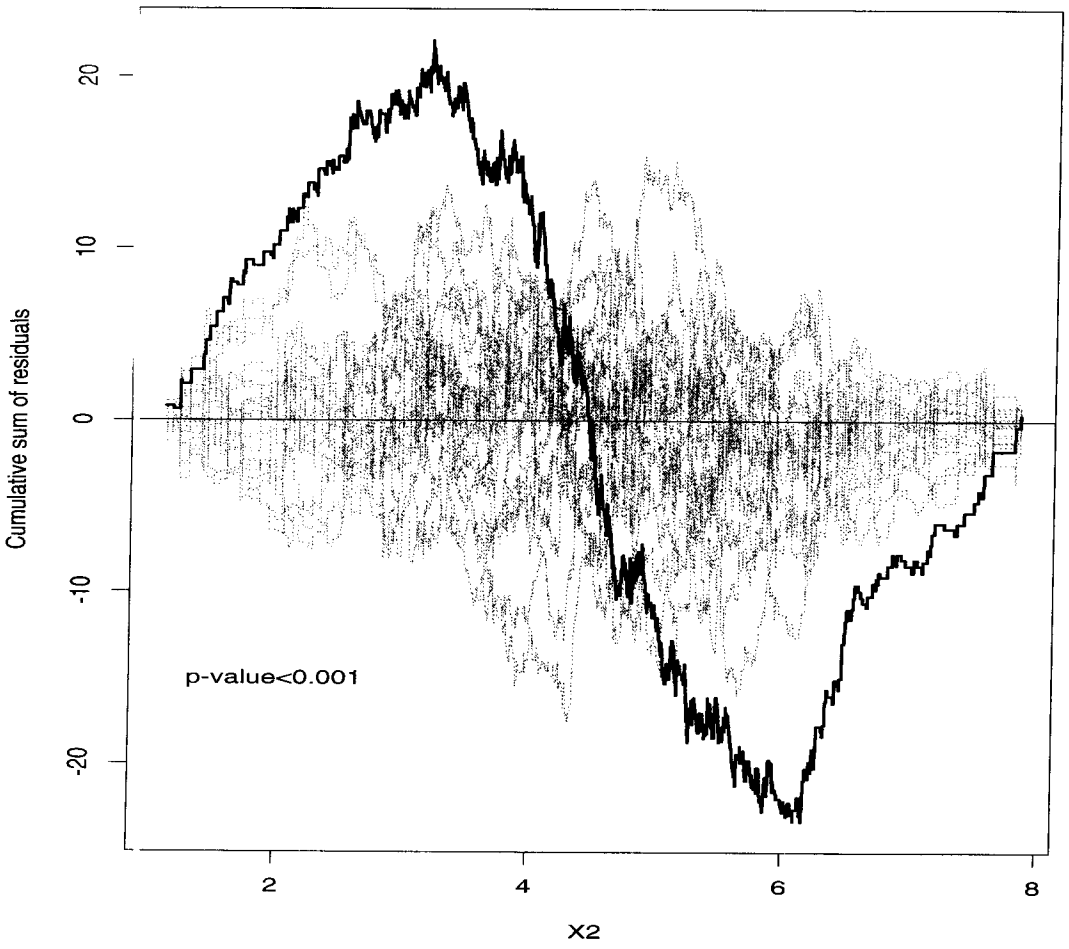

FIGURE : Plot of cumulative residuals versus X_2 in the misspecified logistic model for a simulated data set. The true model is (X_1, X_2, X_2^2), and the fitted model omits X_2^2. The black line indicates the observed process and the gray lines indicate 20 simulated realizations.

To investigate the functional form of alcohol consumption, we first consider alcohol consumption as a linear term in the model which adjusts only for residual confounding by age. Figure 2 contains a plot of the observed cumulative residual process and 20 realizations from the null distribution versus alcohol consumption (g/day).

The P-value for the supremum test for the functional form of alcohol is less than 0.001, indicating that the linear term is inappropriate. The alcohol consumption data are heavily skewed. For the cases, the median alcohol consumption is 1.6 g/day with range 0–236.8 g/day. For the

controls, the median alcohol consumption is 3.8 g/day with range 0–249.9 g/day. In view of this phenomenon, the alcohol consumption data are refit using the logarithmic transform. Figure 3 contains a plot of the observed cumulative residual process and 20 realizations from the null distribution versus log(alcohol).

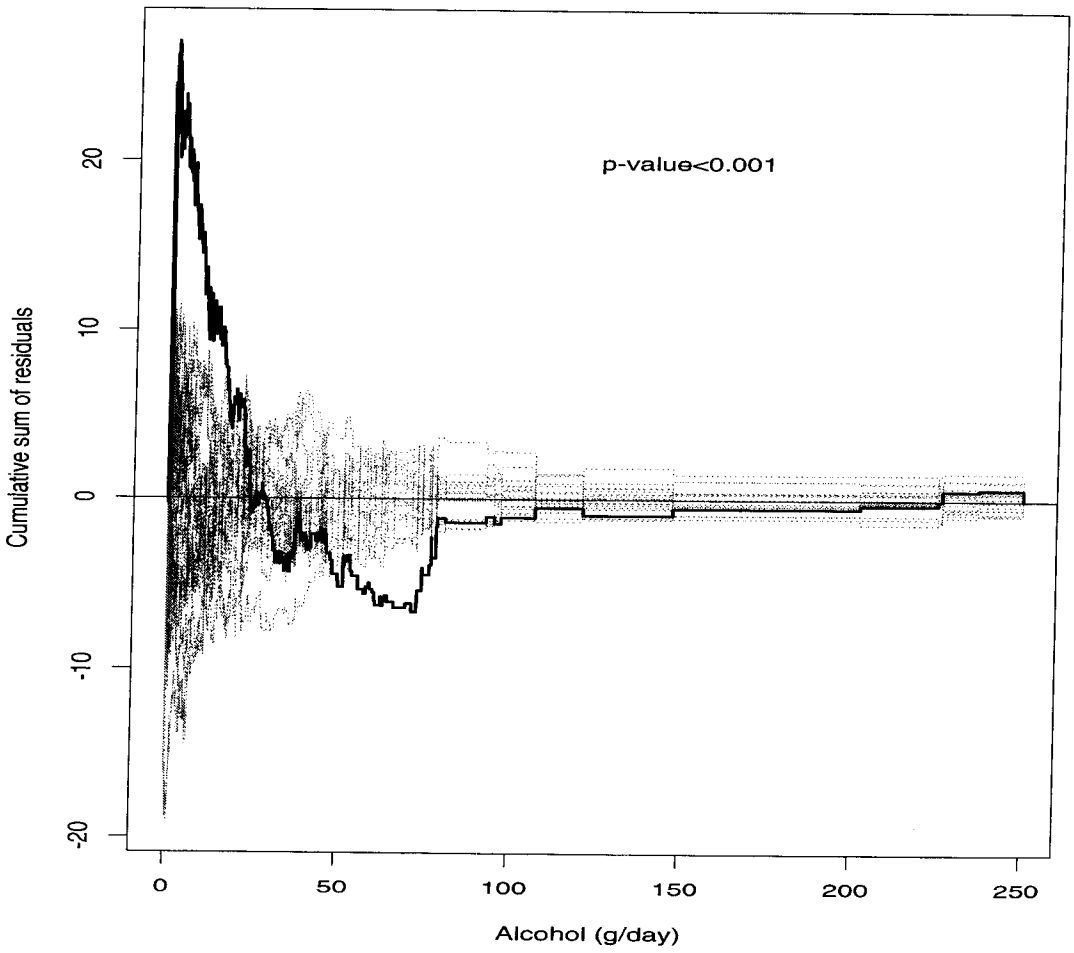

FIGURE 2: Plot of cumulative residuals versus alcohol consumption (g/day) in the logistic model with alcohol and age for the CABS data. The black line indicates the observed process and the gray lines indicate 20 simulated realizations.

Though log(alcohol) alone is inappropriate (the P-value for the supremum test is 0.012), the pattern of the observed process is similar to the pattern observed in our simulation studies when a quadratic term is omitted from the model. The data are refit adding $\log^2$(alcohol), and the P-value for the supremum test increased to 0.455, indicating that this functional form for alcohol consumption is reasonable. Since the relationship between alcohol consumption and risk of primary cardiac arrest may be confounded by other risk factors, the data are refit adjusting for the following covariates: smoking status (never, former, current), diagnosis of diabetes mellitus, diagnosis of hypertension, family history of myocardial infarction or sudden cardiac death, and education. The P-values for the supremum tests assessing the functional form of alcohol consumption, the logistic link function, and the overall adequacy of the model are 0.544, 0.701, and 0.492, re-

Cumulative sum of residuals

spectively. Since our model includes log(alcohol) and $\log^2$(alcohol), a plot of odds ratios versus levels of alcohol consumption is a useful means to describe the relationship between alcohol intake and risk of primary cardiac arrest. Recall that for rare events, odds ratios approximate relative risks. Figure 4 contains a plot of odds ratios and 95% confidence intervals for average daily alcohol intake ranging from 0 to 250 g/day. The reference level is 0 g/day, i.e., nondrinkers. Compared to nondrinkers, low-to-moderate drinking (1–25 g/day) is associated with a reduced risk of primary cardiac arrest. Note that 1 g/day and 25 g/day correspond to roughly 1 drink per 2 weeks and 2 drinks/day, respectively. For higher levels of alcohol consumption, the protective effect of alcohol is no longer present.

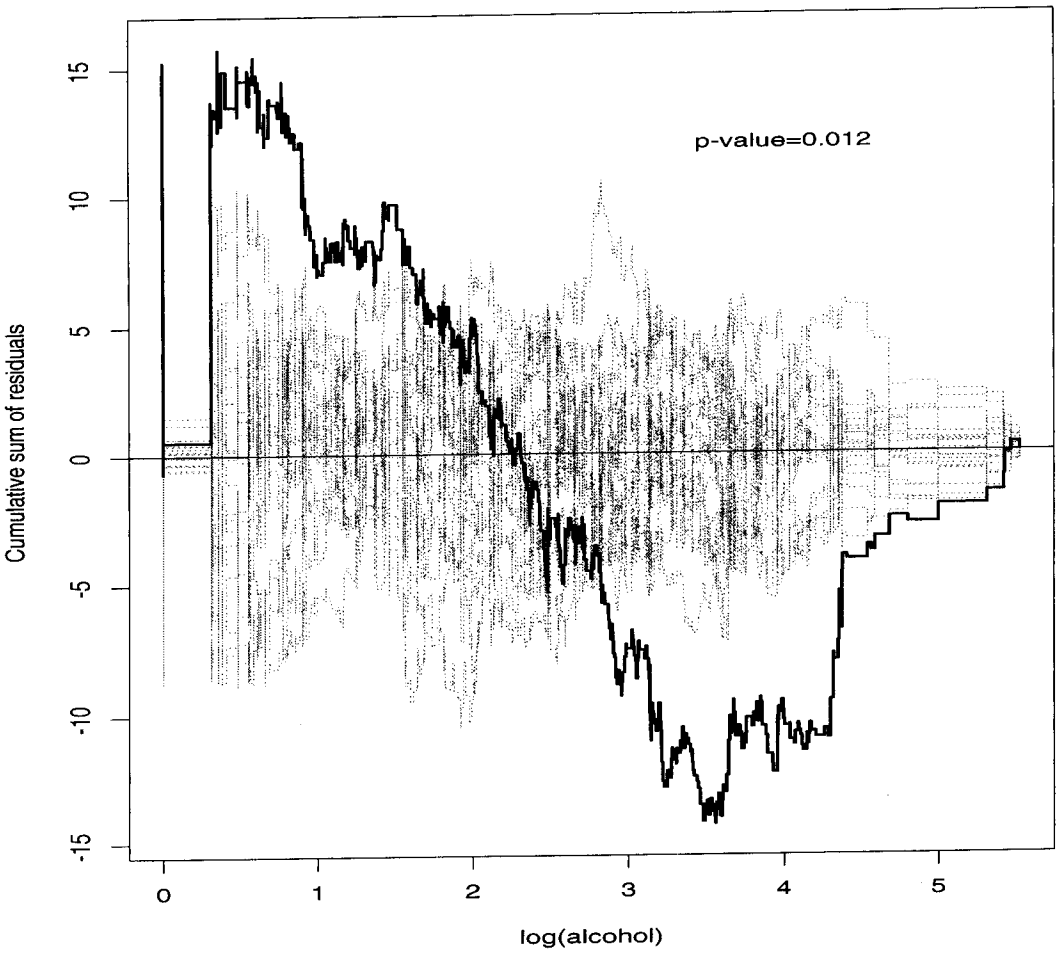

FIGURE 3: Plot of cumulative residuals versus log(alcohol) in the logistic model with log(alcohol) and age for the CABS data. The black line indicates the observed process and the gray lines indicate 20 simulated realizations.

5. DISCUSSION

We have developed graphical and numerical methods for assessing the adequacy of the logistic regression model for matched case-control studies by using the cumulative sums of the residuals. Similar methods were developed in other regression settings. Su & Wei (1991) proposed a

numerical test of overall model adequacy for the generalized linear model, and Lin, Wei & Ying (1993, 2002) developed graphical and numerical methods for the proportional hazards model and the generalized linear model. Although the basic approaches are similar, the unique sampling scheme for the matched case-control study entails new challenges. The forms of the residuals, the cumulative sums, and the asymptotic approximations used here differ substantially from those of Su & Wei (1991), and Lin, Wei & Ying (1993, 2002). Furthermore, we demonstrated that the proposed tests based on the cumulative sums are generally consistent against misspecification of model (1) for matched case-control studies.

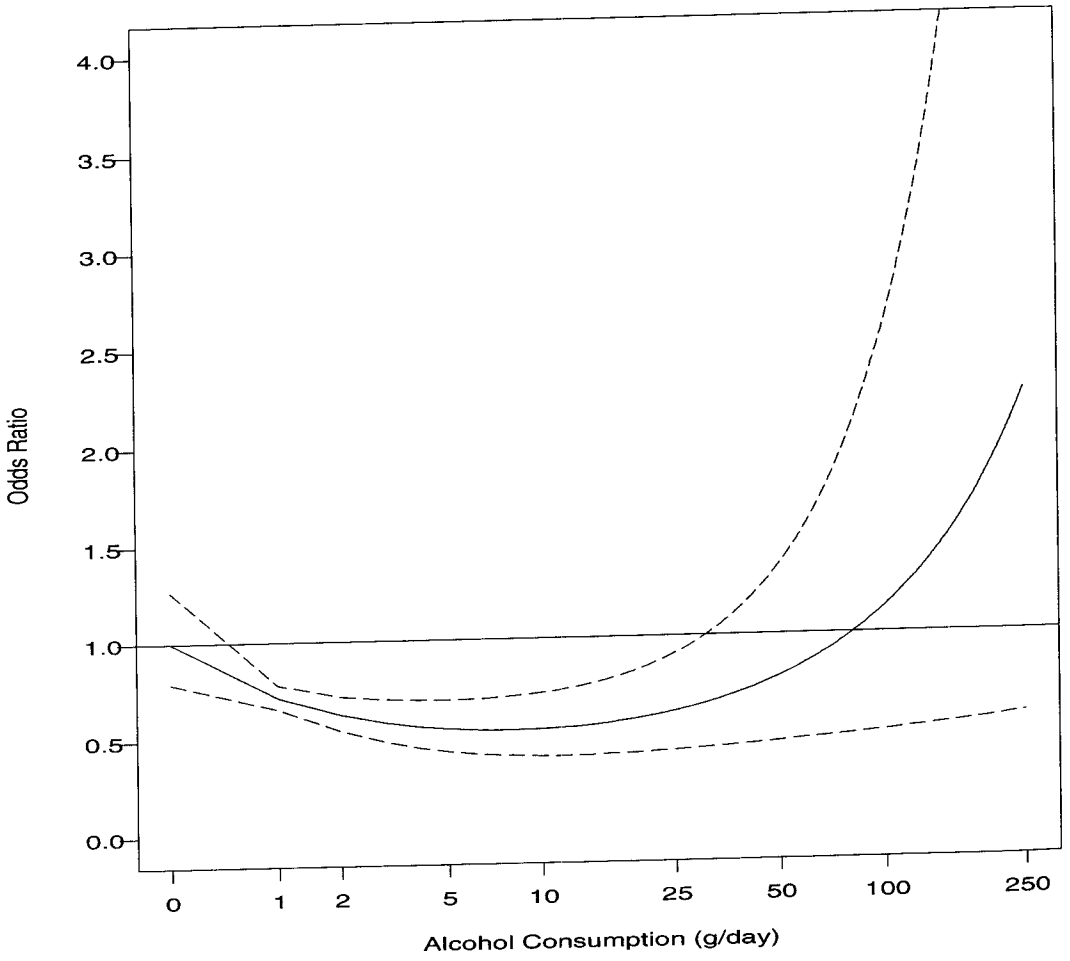

FIGURE 4: Plot of odds ratios and 95% confidence intervals for alcohol consumption (g/day) for the CABS data in the logistic model adjusting for age, smoking, diabetes, hypertension, family history of myocardial infarction or sudden cardiac death, and education. The solid line indicates odds ratios, and the dashed lines indicate 95% confidence intervals.

Royston & Altman (1994) and Royston, Ambler & Sauerbrei (1999) proposed fractional polynomials as a means to investigate the functional forms of continuous covariates. Hosmer & Lemeshow (2000, ch. 7) discussed their application to matched case-control data. Fractional polynomials can be used to reveal functional forms which improve the model fit. This approach is subjective, however, and cannot be used to assess the adequacy of a given functional form.

Our proposed methods are objective and can assess model adequacy.

The proposed methods were implemented in FORTRAN. Though they are numerically intensive, this is not an issue given the power of today's personal computers. In our implementation of these methods, assessing the functional form of a covariate took at most a few seconds. Assessing overall model adequacy in a model consisting of several covariates took at most one minute.

Matching is not always employed in case-control studies. We are currently developing goodness-of-fit methods for unmatched case-control studies. The results will be communicated in a separate report.

APPENDIX

A.1. Weak convergence of W_k, W_ρ, and W_0.

We first establish the weak convergence of $W_0(x;\widehat{\beta})$ under model (1). Consider the one-term Taylor series expansion of $W_0(x;\widehat{\beta})$ at β, namely

$$W_0(x;\widehat{\beta}) = W_0(x;\beta) + \widehat{\eta}_0^\top(x;\beta^*)N^{1/2}(\widehat{\beta}-\beta), \tag{A1}$$

where β^* is on the line segment between $\widehat{\beta}$ and β. Note that

$$W_0(x;\beta) = N^{-1/2}\sum_{i=1}^{N}\xi_i(x) \quad \text{with} \quad \xi_i(x) = \sum_{j=0}^{M_i} r_{ij}I(X_{ij}\le x).$$

Each $\xi_i(x)$ is the difference of two monotone functions in x. Therefore, the processes $\{\xi_i(x); i = 1,\ldots,N\}$ are "manageable" (Pollard 1990, p. 38; Billias, Gu & Ying 1997, proof of Th. 2.1). It then follows from the functional central limit theorem (Pollard 1990, p. 53) that $W_0(x;\beta)$ is tight. Let $\eta_0(x;\beta) = \lim_{N\to\infty}\widehat{\eta}_0(x;\beta)$. Since $\widehat{\eta}_0(t;\beta^*)$ converges almost surely to $\eta_0(t;\beta)$ and $N^{1/2}(\widehat{\beta}-\beta)$ converges in distribution, the second term on the right-hand side of (A1) is also tight. Therefore, $W_0(x;\widehat{\beta})$ is tight.

Since $N^{1/2}(\widehat{\beta}-\beta)$ is asymptotically equivalent to $\Omega^{-1}N^{-1/2}U(\beta)$, asymptotically $W_0(x;\widehat{\beta})$ is equivalent to $\widetilde{W}_0(x;\beta) \equiv N^{-1/2}\sum_{i=1}^{N}\Psi_i(x)$, where

$$\Psi_i(x) = \sum_{j=0}^{M_i} r_{ij}\{I(X_{ij}\le x) + \eta_0^\top(x;\beta)\Omega^{-1}\widetilde{X}_{ij}\}.$$

For fixed x, $\widetilde{W}_0(x;\beta)$ is a sum of N independent and identically distributed (i.i.d.) zero-mean random vectors. By the multivariate central limit theorem, the finite-dimensional distributions of $\widetilde{W}_0(x;\beta)$ are asymptotically zero-mean normal, implying the same for $W_0(x;\widehat{\beta})$. This fact, together with the tightness of $W_0(x;\widehat{\beta})$, implies that $W_0(x;\widehat{\beta})$ converges weakly to a zero-mean Gaussian process with covariance function $\mathrm{E}\{\Psi_1(s)\Psi_1(t)^\top\}$ at (s,t) as $N\to\infty$.

The process $W_k(t;\widehat{\beta})$ is a special case of $W_0(x;\widehat{\beta})$ with $x_\ell = \infty$ for all $\ell \ne k$. Hence, the weak convergence of $W_k(t;\widehat{\beta})$ follows from the above result.

To establish the weak convergence of $W_\rho(t;\widehat{\beta})$, we let $B_\varepsilon(\beta) = \{b : \|b-\beta\| \le \varepsilon\}$ and suppose that, for some $\varepsilon > 0$, the function $\mathrm{P}(b^\top X \le t)$ is continuous in $(b,t) \in B_\varepsilon(\beta)\times[t_1,t_2]$. By the above arguments for $W_0(t;\widehat{\beta})$, we have $W_\rho(t;\widehat{\beta}) = \widetilde{W}_\rho(t;\widehat{\beta}) + o_p(1)$, where

$$\widetilde{W}_\rho(t;b) = N^{-1/2}\sum_{i=1}^{N}\sum_{j=0}^{M_i} r_{ij}\{I(b^\top X_{ij}\le t) + \eta_\rho^\top(t;b)\Omega^{-1}\widetilde{X}_{ij}\},$$

and

$$\eta_\rho(t;b) = -\lim_{N\to\infty} N^{-1}\sum_{i=1}^{N}\sum_{j=0}^{M_i}\mu_{ij}\widetilde{X}_{ij}I(b^\top X_{ij}\le t).$$

Since $\widetilde{W}_\rho(t;b)$ is a sum of independent zero-mean terms, the above arguments for $W_0(t;\widehat{\beta})$ can be used to verify the convergence in finite-dimensional distributions and the "manageability." Therefore, $\widetilde{W}_\rho(t;b)$ converges weakly on $B_\varepsilon(\beta)\times[t_1,t_2]$ to a zero-mean Gaussian process and is stochastically equicontinuous (Pollard 1990, pp. 52–53). In particular, $\widetilde{W}_\rho(t;\widehat{\beta})$ and $\widetilde{W}_\rho(t;\beta)$ are asymptotically equivalent and thus converge to the same limiting Gaussian process.

Next, the weak convergence of $\widehat{W}_0(x;\widehat{\beta})$ is established. Conditional on the data (Y_{ij}, X_{ij}) $(i=1,\ldots,N; j=0,1,\ldots,M_i)$, the only random components in $\widehat{W}_0(x;\widehat{\beta})$ are $(Z_1,\ldots,Z_N)$. Thus, it follows from the multivariate central limit theorem that, conditional on the data, the finite-dimensional distributions of $\widehat{W}_0(x;\widehat{\beta})$ are asymptotically zero-mean normal. Since $\widehat{W}_0(x;\widehat{\beta})$ consists of monotone functions in x, which are manageable, the functional central limit theorem implies that $\widehat{W}_0(x;\widehat{\beta})$ is tight. Define

$$\widehat{\Psi}_i(t)=\sum_{j=0}^{M_i}\widehat{r}_{ij}\big[I(X_{ij}\le t)+\widehat{\eta}_0^\top(t;\widehat{\beta})\{N^{-1}\mathcal{I}(\widehat{\beta})\}^{-1}\widehat{X}_{ij}\big].$$

The conditional covariance function of $\widehat{W}_0(x;\widehat{\beta})$ at (s,t) is

$$N^{-1}\sum_{i=1}^{N}\widehat{\Psi}_i(s)\widehat{\Psi}_i(t)^\top,$$

which converges to $\mathrm{E}\{\Psi_1(s)\Psi_1(t)^\top\}$, the deterministic limiting covariance function of $W_0(x;\widehat{\beta})$. Therefore, $W_0(x;\widehat{\beta})$ and $\widehat{W}_0(x;\widehat{\beta})$ converge to the same limiting zero-mean Gaussian process. Similar arguments can be used to establish the weak convergence of $\widehat{W}_\rho(t;\widehat{\beta})$ and its asymptotic equivalence to $W_\rho(t;\widehat{\beta})$.

A.2. Consistency of supremum tests.

A.2.1. Consistency of $G_0\equiv\sup_{x\in\mathbb{R}^p}|W_0(x;\widehat{\beta})|$. We claim that the test based on G_0 is consistent against the general alternative $\mathcal{H}_1$ that there does not exist a constant vector β such that the true conditional probability of disease can be expressed by μ_{ij} for almost all X_{ij}. Under $\mathcal{H}_1$, $\widehat{\beta}\to\beta^*$ as $N\to\infty$, where β^* is some constant vector. To prove the consistency, it suffices to show that under $\mathcal{H}_1$, $N^{-1/2}G_0$ is nonzero as $N\to\infty$. Let v_{ij} denote the true conditional probability of disease for the jth subject in the ith matched set under $\mathcal{H}_1$, and let μ_{ij}^* denote μ_{ij} with β replaced by β^*. Under $\mathcal{H}_1$,

$$N^{-1/2}W_0(x;\widehat{\beta})\to\lim_{N\to\infty}N^{-1}\sum_{i=1}^{N}\sum_{j=0}^{M_i}(v_{ij}-\mu_{ij}^*)I(X_{ij}\le x),$$

which is nonzero at least for some x. Consequently, $N^{-1/2}G_0$ converges to a nonzero constant. This establishes our claim.

A.2.2. Consistency of $G_\rho\equiv\sup_{t\in\mathbb{R}}|W_\rho(t;\widehat{\beta})|$. Suppose that the right-hand side of (1) is $h(\alpha_i+\beta^\top X)$. Then the conditional probability of disease for the jth subject of the ith matched set is

$$v_{ij}\equiv\frac{h_{ij}/(1-h_{ij})}{\sum_{\ell=0}^{M_i}h_{i\ell}/(1-h_{i\ell})},$$

where $h_{ij}=h(\alpha_i+\beta^\top X_{ij})$. Let β^* be the limit of $\widehat{\beta}$ and μ_{ij}^* be μ_{ij} with β replaced by β^*. Then

$$N^{-1/2}W_\rho(t;\widehat{\beta})\to\lim_{N\to\infty}N^{-1}\sum_{i=1}^{N}\sum_{j=0}^{M_i}(v_{ij}-\mu_{ij}^*)I(\beta^{*\top}X_{ij}\le t),$$

which is nonzero for some t unless $\mu_{ij}^* = v_{ij}$ for almost all $\beta^{*\top} X_{ij}$. In general, $\mu_{ij}^* = v_{ij}$ implies that $h_{ij}/(1-h_{ij}) = c_i \exp(\beta^{*\top} X_{ij})$, i.e., $h_{ij} = c_i \exp(\beta^{*\top} X_{ij})/\{1 + c_i \exp(\beta^{*\top} X_{ij})\}$, where c_i is some constant. Therefore, G_ρ is generally consistent against misspecification of the logistic link function.

A.2.3. Consistency of $G_k \equiv \sup_{t\in\mathbb{R}} |W_k(t;\widehat{\beta})|$. Under $\mathcal{H}_0$, we assume that the kth covariate component is linear, i.e., X_k. Suppose that the true functional form for the kth covariate component is $f(X_k)$ rather than X_k. Let β^* be the limit of $\widehat{\beta}$ under this misspecification. Define

$$\mu_{ij}^* = \frac{\exp\left(\sum_{m=1}^p \beta_m^* X_{ijm}\right)}{\sum_{\ell=0}^{M_i} \exp\left(\sum_{m=1}^p \beta_m^* X_{i\ell m}\right)},$$

and

$$v_{ij} = \frac{\exp\left\{\beta_k f(X_{ijk}) + \sum_{m\neq k} \beta_m X_{ijm}\right\}}{\sum_{\ell=0}^{M_i} \exp\left\{\beta_k f(X_{i\ell k}) + \sum_{m\neq k} \beta_m X_{i\ell m}\right\}}.$$

Then

$$N^{-1/2} W_k(t;\widehat{\beta}) \to \lim_{N\to\infty} N^{-1} \sum_{i=1}^{N} \sum_{j=0}^{M_i} (v_{ij} - \mu_{ij}^*) I(X_{ijk} \leq t)$$

which is nonzero for some t unless $v_{ij} = \mu_{ij}^*$ for almost all X_{ijk}. In general, $\beta_k^* \neq \beta_k$. If $\beta_m^* = \beta_m$ for all $m \neq k$, then $v_{ij} \neq \mu_{ij}^*$. In the more realistic situations in which $\beta_m^* \neq \beta_m$ $(m \neq k)$, the inequalities are unlikely to offset the misspecification of the functional form for the kth covariate component in such a way that $\mu_{ij}^* = v_{ij}$ for almost all X_{ijk}. Hence, G_k is generally consistent against misspecification of the functional form.

ACKNOWLEDGEMENTS

The authors wish to thank the Editor, the Associate Editor, and two referees for their important and constructive comments that led to significant improvement of this paper.

REFERENCES

E. J. Bedrick & J. R. Hill (1996). Assessing the fit of the logistic regression model to individual matched sets of case-control data. *Biometrics*, 52, 1–9.

Y. Billias, M. Gu & Z. Ying (1997). Towards a general asymptotic theory for Cox model with staggered entry. *The Annals of Statistics*, 25, 662–682.

N. E. Breslow & N. E. Day (1980). *Statistical Methods in Cancer Research, Volume 1: The Analysis of Case-Control Studies*. International Agency for Research on Cancer, Lyon, France.

D. M. Goldberg, S. E. Hahn & J. G. Parkes (1995). Beyond alcohol: beverage consumption and cardiovascular mortality. *Clinica Chimica Acta*, 237, 155–187.

D. W. Hosmer & S. Lemeshow (2000). *Applied Logistic Regression*. Wiley, New York.

D. Y. Lin, L. J. Wei & Z. Ying (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80, 557–572.

D. Y. Lin, L. J. Wei & Z. Ying (2002). Model-checking techniques based on cumulative residuals. *Biometrics*, 58, 1–12.

P. McCullagh & J. A. Nelder (1989). *Generalized Linear Models*. Chapman & Hall, London.

S. H. Moolgavkar, E. D. Lustbader & D. J. Venzon (1984). A geometric approach to nonlinear regression diagnostics with application to matched case-control studies. *The Annals of Statistics*, 12, 816–826.

S. H. Moolgavkar, E. D. Lustbader & D. J. Venzon (1985). Assessing the adequacy of the logistic regression model for matched case-control studies. *Statistics in Medicine*, 4, 425–435.

D. Pollard (1990). *Empirical Processes: Theory and Applications.* NSF-CBMS Regional Conference Series in Probability and Statistics 2, Institute of Mathematical Statistics, Hayward, California.

D. Pregibon (1984). Data analytic methods for matched case-control studies. *Biometrics*, 40, 639–651.

P. Royston & D. G. Altman (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Applied Statistics*, 43, 429–467.

P. Royston, G. Ambler & W. Sauerbrei (1999). The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology*, 28, 964–974.

D. Siscovick, T. Raghunathan, I. King, S. Weinmann, K. Wichlund, J. Albright, V. Bovbjerg, P. Arbogast, H. Smith, L. Kushi, L. Cobb, M. Copass, B. Psaty, R. Lematre, B. Retzlaff, M. Childs, R. Knopp (1995). Dietary intake and cell membrane levels of long-chain n-3 polyunsaturated fatty acids and the risk of primary cardiac arrest. *Journal of the American Medical Association*, 274, 1363–1367.

J. Q. Su & L. J. Wei (1991). A lack-of-fit test for the mean function in a generalized linear model. *Journal of the American Statistical Association*, 86, 420–426.

Received 18 October 2002
Accepted 12 May 2004

Patrick G. ARBOGAST: patrick.arbogast@vanderbilt.edu
Department of Biostatistics, Vanderbilt University
Nashville, TN 37232-2158, USA

Danyu Y. LIN: lin@bios.unc.edu
Department of Biostatistics, The University of North Carolina
Chapel Hill, NC 27599-7420, USA