STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:229–247
Published online 28 October 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.1932

Model-checking techniques for stratified case-control studies

Patrick G. Arbogast[1,*,†] and D. Y. Lin[2]

[1]*Department of Biostatistics, Vanderbilt University, S-2323 Medical Center North, Nashville, TN 37232-2158, U.S.A.*
[2]*Department of Biostatistics, CB #7420, McGavran-Greenberg Hall, University of North Carolina, Chapel Hill, NC 27599-7420, U.S.A.*

SUMMARY

We present graphical and numerical methods for assessing the adequacy of the logistic regression model for stratified case-control data. The proposed methods are derived from the cumulative sum of residuals over the covariate or linear predictor. Under the assumed model, the cumulative residual process converges weakly to a zero-mean Gaussian process whose distribution can be approximated via Monte Carlo simulation. The observed cumulative residual pattern can then be compared both visually and analytically to a number of simulated realizations from the approximate null distribution. These comparisons enable one to examine the functional form of each covariate, the logistic link function as well as the overall model adequacy. Simulation studies demonstrate that the proposed methods perform well in practical settings. Illustration with an oesophageal cancer study is provided. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: goodness of fit; link function; logistic regression; model misspecification; regression diagnostic; residual

1. INTRODUCTION

Case-control studies are routinely conducted to investigate the relationship between exposure and disease. The statistical analysis of case-control data is typically based on the logistic regression model, which relates the probability of developing the disease to the exposure of interest and other risk factors. If one ignores the feature of unequal selection probabilities of the case-control design and proceeds as if the observations came from a random sample of the entire population, the standard maximum likelihood method will provide valid inference for the slope parameters (i.e. log odds ratios), but not for the intercept term [1].

There exist various diagnostic methods for the logistic regression; see Reference [2] for a review. Specifically, Pregibon [3] studied methods for detecting outliers and influential

*Correspondence to: Patrick G. Arbogast, Department of Biostatistics, Vanderbilt University, S-2323 Medical Center North, Nashville, TN 37232-2158, U.S.A.
†E-mail: patrick.arbogast@vanderbilt.edu

Received May 2003
Accepted May 2004

subjects. Hosmer and Lemeshow [4] developed an overall goodness-of-fit test that employs a grouping method based on the deciles of the risk. Osius and Rojek [5] derived a large-sample normal approximation to the Pearson chi-square statistic. Stukel [6] developed a two degree-of-freedom test that assesses the tails of the logistic regression model.

The above tests were developed for independent and identically distributed data, but they can also be applied to case-control data [7]. Recently, two goodness-of-fit tests have been derived specifically for case-control studies: Qin and Zhang [8] proposed a Kolmogorov–Smirnov-type statistic, and Zhang [9] proposed a chi-square-type statistic.

Although the existing tests are useful for assessing the goodness-of-fit of the model, they do not provide insights into the nature of model misspecification. Furthermore, there does not exist any method for assessing the adequacy of the functional form of a covariate, which is a common form of model misspecification.

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 stratified case-control data. 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 data taken from the Ille-et-Vilaine oesophageal cancer study previously considered by Breslow and Day [10].

2. METHODS

2.1. Logistic regression

Let Y denote the disease status, taking values 1 for cases and 0 for controls, and let $X^\dagger = (X_1, \ldots, X_p)'$ be a $p \times 1$ vector of covariates. Suppose that stratified case-control sampling is undertaken in which n_{ij} subjects are sampled from the ith $(i = 0, 1)$ disease category and jth $(j = 1, \ldots, J)$ stratum, where strata are formed on the basis of, for example, gender or age groups. Let $n_i = \sum_{j=1}^{J} n_{ij}$ and $n = n_0 + n_1$. Assume that, in the jth stratum, the conditional distribution of Y given $X^\dagger$ is Bernoulli with success probability

$$\Pr(Y = 1 | X^\dagger; \alpha_j, \beta) = \frac{e^{\alpha_j + \beta' X^\dagger}}{1 + e^{\alpha_j + \beta' X^\dagger}} \tag{1}$$

where α_j is the intercept term for the jth stratum, and $\beta = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector of regression coefficients.

In stratified case-control studies, subjects are sampled conditionally on their disease status and on their values of the stratification variables so that the resulting sample is not a random sample from the whole population. If one ignores the feature of unequal selection probabilities of the case-control design and proceeds as if the observations came from a random sample within each stratum of the entire population, the standard maximum likelihood method will provide valid inference for β but not for the α_j's [1]. Scott and Wild [11] showed that the estimators for the intercept terms are biased by o_j, where

$$o_j = \log \left\{ \frac{n_{1j} \Pr(Y = 0 | \text{stratum} = j)}{n_{0j} \Pr(Y = 1 | \text{stratum} = j)} \right\}$$

Let $\delta_j = o_j + \alpha_j$, $\theta = (\delta_1, \ldots, \delta_J, \beta')'$, and $X = (S', X^{\dagger\prime})'$, where S is a $J \times 1$ vector of stratum indicator variables, i.e. the jth component of S equals 1, denoting the jth stratum, and the other components equal 0. Let X_{ijl} and Y_{ijl} denote the values of X and Y, respectively, for the lth subject ($l = 1, \ldots, n_{ij}$) in the ith disease category and jth stratum. Under model (1), the score function for θ is

$$\mathrm{U}(\theta) = \sum_{i=0}^{1} \sum_{j=1}^{J} \sum_{l=1}^{n_{ij}} \{Y_{ijl} - p_j(X_{ijl}; \theta)\} X_{ijl}$$

where

$$p_j(X; \theta) = \frac{\mathrm{e}^{\delta_j + \beta' X^{\dagger}}}{1 + \mathrm{e}^{\delta_j + \beta' X^{\dagger}}} \tag{2}$$

Note that $p_j(X; \theta)$ is the conditional probability that a member of the case-control sample from stratum j with regression variables X is a case. Let $\widehat{\theta}$ denote the solution to the score equation $\mathrm{U}(\theta) = 0$, and let $\mathscr{I}(\theta)$ denote the observed information matrix, i.e.

$$\mathscr{I}(\theta) = \sum_{i=0}^{1} \sum_{j=1}^{J} \sum_{l=1}^{n_{ij}} p_j(X_{ijl}; \theta)\{1 - p_j(X_{ijl}; \theta)\} X_{ijl} X_{ijl}'$$

2.2. Residuals

Residuals normally take the form of the observed minus the predicted values of the response. Consider

$$r_{ijl} = Y_{ijl} - p_j(X_{ijl}; \theta), \quad i = 0, 1, \ \ j = 1, \ldots, J, \ \ l = 1, \ldots, n_{ij}$$

As mentioned above, $p_j(X_{ijl}; \theta)$ is the expected value of Y_{ijl} conditional on X_{ijl} and on being selected into the jth stratum of the case-control sample. Clearly, $E(r_{ijl}) = 0$, and $\mathrm{cov}(r_{ijl}, r_{i'j'l'}) = 0$ for $(ijl) \neq (i'j'l')$. Thus, we define the residuals as

$$\widehat{r}_{ijl} = Y_{ijl} - p_j(X_{ijl}; \widehat{\theta}), \quad i = 0, 1, \ \ j = 1, \ldots, J, \ \ l = 1, \ldots, n_{ij}$$

The residual $\widehat{r}_{ijl}$ is the difference between the observed disease status and the estimated probability of disease conditional on the stratified case-control sample. The $\widehat{r}_{ijl}$'s behave like ordinary residuals in the familiar linear regression in that $\sum_{i=0}^{1} \sum_{j=1}^{J} \sum_{l=1}^{n_{ij}} \widehat{r}_{ijl} = 0$, and for large n, $E(\widehat{r}_{ijl}) \approx 0$, and $\mathrm{cov}(\widehat{r}_{ijl}, \widehat{r}_{i'j'l'}) \approx 0$ for $(ijl) \neq (i'j'l')$. These residuals are the building blocks for the proposed model checking techniques.

2.3. Functional forms of covariates

A common approach to assessing the functional form of a covariate in linear regression is to plot the residuals versus the covariate [12]. A similar plot can be constructed for model (1) on the basis of the $\widehat{r}_{ijl}$'s. However, this approach is highly subjective: 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 to $Y = 0$ or 1. To avoid these problems, we propose to use the cumulative sum of the $\widehat{r}_{ijl}$'s over the covariate of interest to check the functional form. The resulting residual analysis is objective and informative.

Let X_{ijlk} denote the kth component of X for the lth subject in the jth stratum and ith disease group. Consider the following stochastic process:

$$W_k(t;\widehat{\theta}) = n^{-1/2} \sum_{i=0}^{1} \sum_{j=1}^{J} \sum_{l=1}^{n_{ij}} \widehat{r}_{ijl} I(X_{ijlk} \leqslant t)$$

which is the cumulative sum of the residuals $\widehat{r}_{ijl}$ over the values of the kth covariate component X_k. This process compares the observed and predicted values of the response over all possible values of X_k and is thus informative about the misspecification of the functional form of X_k. Under the null hypothesis H_0 that model (1) is correct, $W_k(t;\widehat{\theta})$ fluctuates around 0 as t varies. We show in Appendix A that, under H_0, $W_k(t;\widehat{\theta})$ converges weakly to a zero-mean Gaussian process whose distribution can be approximated by that of

$$\widehat{W}_k(t;\widehat{\theta}) = n^{-1/2} \sum_{i=0}^{1} \sum_{j=1}^{J} \sum_{l=1}^{n_{ij}} Z_{ijl} \widehat{r}_{ijl} [I(X_{ijlk} \leqslant t) + \widehat{\eta}_k'(t;\widehat{\theta}) \{n^{-1} \mathscr{I}(\widehat{\theta})\}^{-1} X_{ijl}]$$

where

$$\widehat{\eta}_k(t;\theta) = n^{-1/2} \partial W_k(t;\theta)/\partial\theta = -n^{-1} \sum_{i=0}^{1} \sum_{j=1}^{J} \sum_{l=1}^{n_{ij}} p_j(X_{ijl};\theta)\{1 - p_j(X_{ijl};\theta)\} X_{ijl} I(X_{ijlk} \leqslant t)$$

and Z_{ijl} $(i = 0, 1;\ j = 1, \ldots, J;\ l = 1, \ldots, n_{ij})$ are independent standard normal random variables.

To assess whether the observed pattern of $W_k(t;\widehat{\theta})$ is abnormal or not, we plot it along with a few, say 20, realizations of $\widehat{W}_k(t;\widehat{\theta})$. The process $\widehat{W}_k(t;\widehat{\theta})$ can be generated by taking repeated random samples of $\{Z_{ijl}\}$ while fixing the data (Y_{ijl}, X_{ijl}) $(i = 0, 1;\ j = 1, \ldots, J;\ l = 1, \ldots, n_{ij})$ at their observed values.

Numerical tests can be constructed as well. Since $W_k(t;\widehat{\theta})$ fluctuates around zero under H_0, it is natural to consider the Kolmogorov-type supremum statistic $G_k \equiv \sup_{t \in \mathscr{R}} |W_k(t;\widehat{\theta})|$, where $\mathscr{R}$ denotes the real line. Let g_k denote 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 the statistical significance of the test, we compute the probability $\Pr(G_k \geqslant g_k)$, which can be approximated by $\Pr(\widehat{G}_k \geqslant g_k)$, where $\widehat{G}_k = \sup_{t \in \mathscr{R}} |\widehat{W}_k(t;\widehat{\theta})|$. In turn, $\Pr(\widehat{G}_k \geqslant g_k)$ can be estimated by generating a large number (e.g. 1000 or 10 000) of realizations of $\widehat{W}_k(t;\widehat{\theta})$. In Appendix A, 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. To assess this aspect of the model, we consider

$$W_\rho(t;\widehat{\theta}) = n^{-1/2} \sum_{i=0}^{1} \sum_{j=1}^{J} \sum_{l=1}^{n_{ij}} \widehat{r}_{ijl} I(\widehat{\theta}' X_{ijl} \leqslant t)$$

which is the same as $W_k(t;\widehat{\theta})$ except that the residuals are summed over the values of $\widehat{\theta}' X$ instead of X_k. This process compares the observed and predicted values of the response over all possible values of the linear predictor and is thus informative about the misspecification

of the linear predictor or the link function. We show in Appendix A that, under H_0, $W_\rho(t;\widehat{\theta})$ converges weakly to the same limiting zero-mean Gaussian process as

$$\widehat{W}_\rho(t;\widehat{\theta}) = n^{-1/2}\sum_{i=0}^{1}\sum_{j=1}^{J}\sum_{l=1}^{n_{ij}} Z_{ijl}\widehat{r}_{ijl}[I(\widehat{\theta}'X_{ijl}\leqslant t) + \widehat{\eta}'_\rho(t;\widehat{\theta})\{n^{-1}\mathscr{I}(\widehat{\theta})\}^{-1}X_{ijl}]$$

where $\widehat{\eta}_\rho(t;\theta)$ is $\widehat{\eta}_k(t;\theta)$ with $I(X_{ijlk}\leqslant t)$ replaced with $I(\theta'X_{ijl}\leqslant t)$. As in Section 2.3, graphical and numerical procedures can be constructed to assess whether the observed pattern of $W_\rho(t;\widehat{\theta})$ is abnormal or not. The supremum test statistic $G_\rho \equiv \sup_{t\in\mathscr{R}}|W_\rho(t;\widehat{\theta})|$ is shown in Appendix A to be generally consistent against misspecification of the logistic link function.

2.5. Overall model adequacy

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

$$W_o(x;\widehat{\theta}) = n^{-1/2}\sum_{i=0}^{1}\sum_{j=1}^{J}\sum_{l=1}^{n_{ij}}\widehat{r}_{ijl}I(X_{ijl}\leqslant x)$$

where $x=(s_1,\ldots,s_J,x_1,\ldots,x_p)'$, and $I(X_{ijl}\leqslant x)$ is the indicator function for the event that all components of X_{ijl} are no larger than the corresponding components of x. This process compares the observed and predicted values of the response over all possible combinations of the covariates and thus is informative about the misspecification of any aspect of the model. We show in Appendix A that, under H_0, $W_o(x;\widehat{\theta})$ converges weakly to the same zero-mean Gaussian process as

$$\widehat{W}_o(x;\widehat{\theta}) = n^{-1/2}\sum_{i=0}^{1}\sum_{j=1}^{J}\sum_{l=1}^{n_{ij}} Z_{ijl}\widehat{r}_{ijl}[I(X_{ijl}\leqslant x) + \widehat{\eta}'_o(x;\widehat{\theta})\{n^{-1}\mathscr{I}(\widehat{\theta})\}^{-1}X_{ijl}]$$

where $\widehat{\eta}_o(x;\theta)$ is $\widehat{\eta}_k(t;\theta)$ with $I(X_{ijlk}\leqslant t)$ replaced with $I(X_{ijl}\leqslant x)$. Since $W_o(x;\widehat{\theta})$ is a multiparameter process, it is difficult to graphically assess whether the observed pattern of $W_o(x;\widehat{\theta})$ is unusual. However, the supremum test statistic $G_o \equiv \sup_{x\in\mathscr{R}^{J+p}}|W_o(x;\widehat{\theta})|$ can be used. The p-value can again be estimated via simulation. In Appendix A, 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). We defined two strata by $Q=0$ versus 1, where Q is Bernoulli with success probability 0.25. We sampled $n_{ij}=50$, 100 and 200 cases and controls from each stratum. For each simulation setting, 1000 case-control samples were generated. For each sample, we performed supremum goodness-of-fit tests based on $W_k(t;\widehat{\theta})$, $W_\rho(t;\widehat{\theta})$, and $W_o(x;\widehat{\theta})$. The nominal significance level for each test was set at 0.05, and the empirical size and power were estimated. The null hypothesis H_0 was rejected if the observed supremum statistic exceeded the 95th percentile of the supremum of the approximating null distribution. The percentile was estimated from 1000 realizations.

In one series of studies, we let $X^\dagger = (X_1, X_1^2)'$, where X_1 has a normal distribution with unit variance and mean of 4 if $Q = 0$ and 5 if $Q = 1$. The dependence of the mean of X_1 on Q creates a confounding effect. We set $\beta = (-0.25, \beta_2)'$, where $\beta_2 = 0.05$, 0.1, 0.15, 0.20 or 0.25. We set α_0 and α_1 so that 0.1–0.2 per cent of the simulated population were cases. This represents the type of population in which a case-control study is likely to be conducted. To estimate the size of each test, the data were fit using $X^\dagger$. To estimate the power, the data were fit with X_1^2 omitted. For comparisons, the Wald statistic for testing $\beta_2 = 0$ as well as Hosmer and Lemeshow's and Stukel's goodness-of-fit tests were evaluated. In Hosmer and Lemeshow's test, we partitioned the case-control sample into 10 groups according to the deciles of the estimated probabilities $p_j(X_{ij}; \widehat{\theta})$. Stukel's goodness-of-fit method is a two degree-of-freedom test evaluating $\zeta_1 = 0$ and $\zeta_2 = 0$ in the revised logistic model $\text{logit}\{\Pr(Y = 1|X; \theta, \zeta_1, \zeta_2)\} = \theta' X + \frac{1}{2}(\theta' X)^2\{\zeta_1 I(\theta' X \geqslant 0) - \zeta_2 I(\theta' X < 0)\}$. The two additional parameters allow the revised logistic model to be either symmetric or asymmetric with tails either lighter or heavier than in the original model.

The simulation results are summarized in Table I, where G_1 denotes the supremum test assessing the functional form of X_1. The supremum tests have proper sizes and good powers. The proposed tests are more powerful than Hosmer and Lemeshow's test, and have power similar to Stukel's test. As previously mentioned, the latter two tests do not provide insights into the nature of model misspecification, which in this case lies in the functional form of X_1. The Wald test is optimal in testing extra parameters in embedded parametric models. However, unlike the supremum tests, the Wald test cannot be used to test against non-nested alternatives, such as which functional form of X_1 is more appropriate, or whether the chosen functional form is satisfactory.

To illustrate the graphical procedures, we consider a simulated data set with $n_{ij} = 100$ generated from $X^\dagger$ in which $\beta = (-0.25, 0.25)'$, but the data are fit with X_1^2 omitted. Figure 1 displays the 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_1 is less than 0.001, indicating that modelling X_1 on a linear scale is inappropriate. When the simulated data set is refit including X_1^2, the p-value for the supremum test for the functional form of X_1 jumps to 0.491. Figure 2 contains the plot of the observed cumulative residual process under the true functional form of X_1. The observed process now appear to randomly fluctuate about zero as expected.

When the true functional form of a covariate is unknown, the observed pattern of the cumulative residual process, such as that depicted in Figure 1, can provide useful hint about the correct functional form. Specifically, the pattern of the cumulative residual process seen in Figure 1 occurs when a quadratic covariate is misspecified as a linear term.

To illustrate other patterns of observed cumulative residual processes, we consider a covariate whose correct functional form is logarithmic. We simulated a data set with $n_{ij} = 100$ in which $X^\dagger = (\log(X_1))$, where $\log(X_1)$ has the standard normal distribution, and $\beta_1 = 1$. Figure 3 contains a plot of the observed cumulative residual process when X_1 is modelled as a linear term and demonstrates the pattern observed when the correct functional form is logarithmic. Another illustration is when the relationship between a continuous covariate and disease incidence is a threshold effect. We simulated a data set with $n_{ij} = 100$ generated from $X^\dagger = (I(X_1 \geqslant 0))$, where X_1 is a standard normal variable, and $\beta = 1$. Figure 4 contains a plot of the observed cumulative residual process when X_1 is expressed as a linear term.

Table I. Simulation results for evaluating the sizes and powers of the supremum tests under $X^{\dagger} = (X_1, X_1^2)$.

		$n_{ij} = 50$		$n_{ij} = 100$		$n_{ij} = 200$	
β_2		Size	Power	Size	Power	Size	Power
0.05	G_1	0.057	0.070	0.057	0.086	0.068	0.132
	G_ρ	0.040	0.070	0.052	0.086	0.060	0.096
	G_o	0.062	0.068	0.057	0.083	0.065	0.118
	HL*	0.030	0.059	0.037	0.068	0.028	0.077
	Stukel	0.063	0.084	0.047	0.104	0.052	0.138
	Wald	—	0.095	—	0.161	—	0.262
0.1	G_1	0.064	0.135	0.057	0.227	0.069	0.396
	G_ρ	0.058	0.135	0.053	0.151	0.066	0.275
	G_o	0.061	0.136	0.054	0.211	0.064	0.392
	HL*	0.033	0.085	0.039	0.123	0.038	0.189
	Stukel	0.055	0.168	0.043	0.253	0.052	0.440
	Wald	—	0.261	—	0.451	—	0.716
0.15	G_1	0.056	0.270	0.046	0.475	0.059	0.805
	G_ρ	0.062	0.195	0.055	0.311	0.066	0.549
	G_o	0.058	0.263	0.046	0.455	0.060	0.799
	HL*	0.038	0.114	0.045	0.235	0.035	0.460
	Stukel	0.058	0.310	0.052	0.489	0.051	0.791
	Wald	—	0.536	—	0.802	—	0.988
0.2	G_1	0.061	0.481	0.043	0.790	0.056	0.979
	G_ρ	0.060	0.320	0.053	0.545	0.045	0.853
	G_o	0.064	0.468	0.044	0.789	0.064	0.975
	HL*	0.049	0.223	0.042	0.440	0.044	0.776
	Stukel	0.056	0.484	0.049	0.795	0.056	0.971
	Wald	—	0.814	—	0.980	—	1.0
0.25	G_1	0.065	0.743	0.057	0.972	0.061	1.0
	G_ρ	0.069	0.468	0.058	0.775	0.054	0.979
	G_o	0.053	0.736	0.055	0.970	0.059	1.0
	HL*	0.043	0.347	0.046	0.672	0.042	0.958
	Stukel	0.069	0.684	0.051	0.947	0.053	0.999
	Wald	—	0.963	—	0.999	—	1.0

*Hosmer–Lemeshow's goodness-of-fit test.

This demonstrates the pattern observed when the true functional form is a threshold effect at $X_1 = 0$.

4. ILLE-ET-VILAINE OESOPHAGEAL CANCER STUDY

The proposed goodness-of-fit methods were applied to data taken from the Ille-et-Vilaine study of oesophageal cancer [10]. The cases consisted of 200 males diagnosed with oesophageal cancer at a regional hospital in the Ille-et-Vilaine district of Brittany between January 1972 and

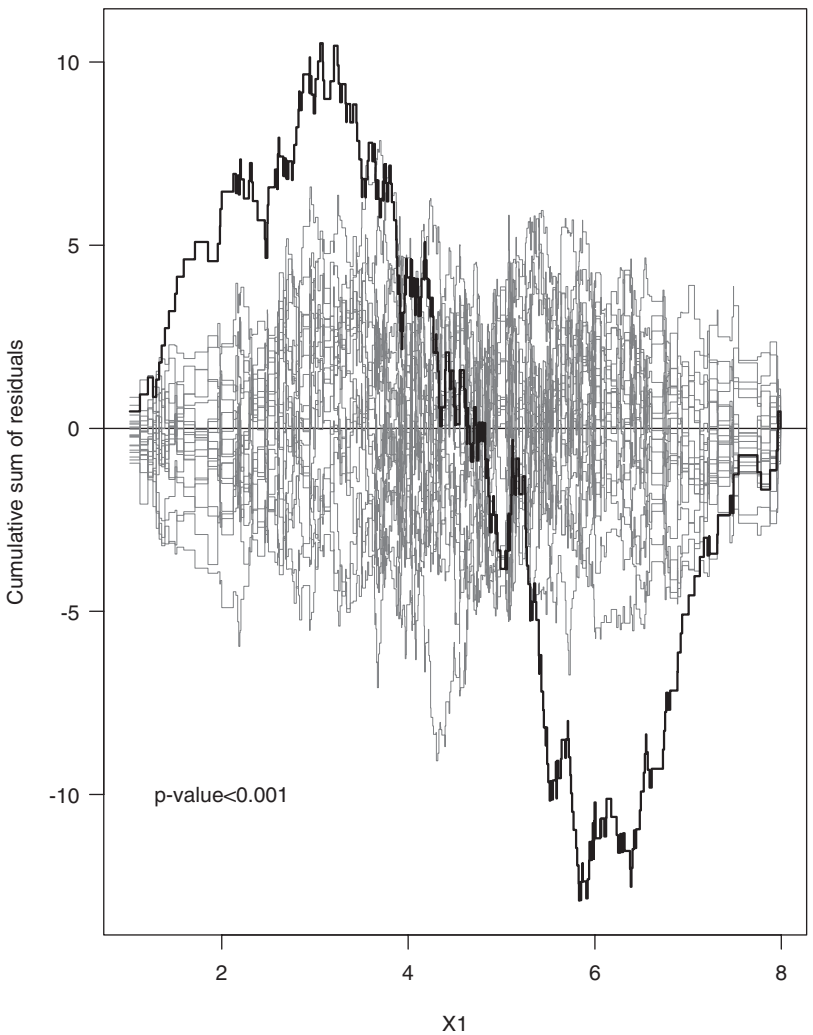

Figure 1. Plot of cumulative residuals versus X_1 in the misspecified logistic model for a simulated data set. True model is (X_1, X_1^2), and the fitted model omits X_1^2. The black line indicates the observed process, and the grey lines indicate 20 simulated realizations.

April 1974. There were 775 adult male controls drawn from electoral lists in each commune who provided sufficient data for analysis. Subjects were administered a dietary interview containing questions about tobacco use and alcohol consumption as well as other dietary risk factors. We applied the proposed methods to the analysis relating tobacco use and alcohol consumption to risk of oesophageal cancer. Since the controls tended to be younger than the cases, age was included in all of the models.

We first considered age, alcohol consumption, and tobacco use as linear terms. Note that tobacco use was recorded as a discrete variable having nine levels. However, to analyse

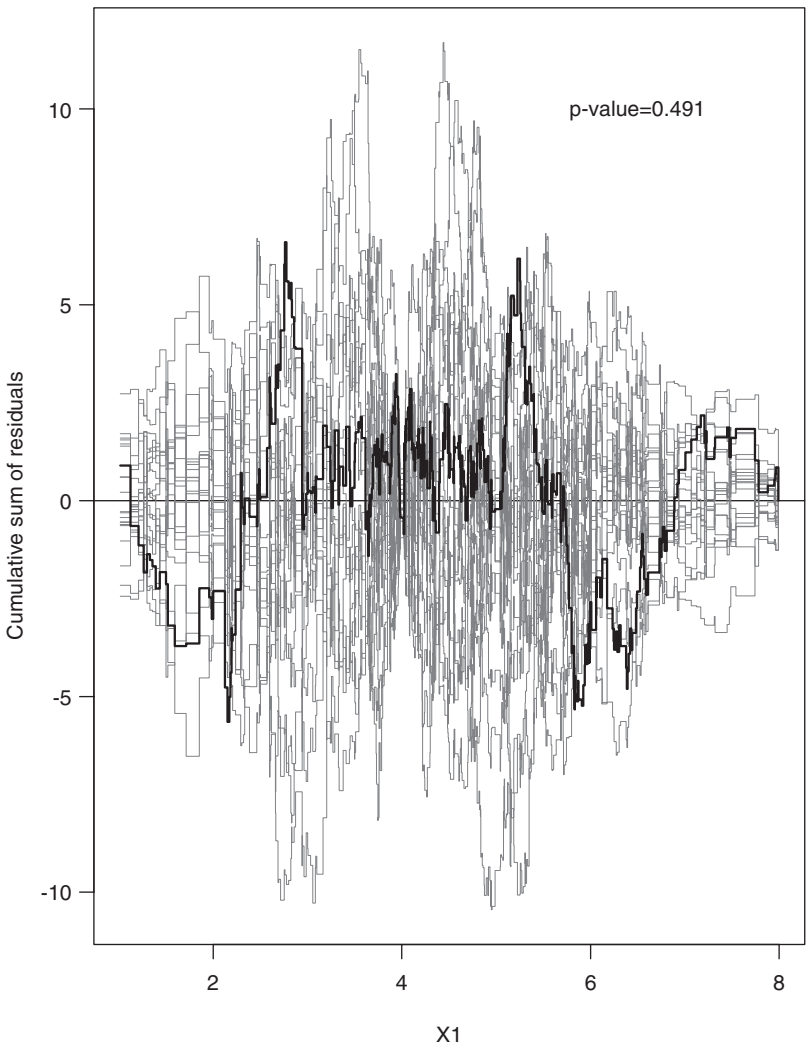

Figure 2. Plot of cumulative residuals versus X_1 under the true logistic model (X_1, X_1^2) for a simulated data set. The black line indicates the observed process, and the grey lines indicate 20 simulated realizations.

it quantitatively, Breslow and Day assigned values to each level and treated the data as continuous. We used the same quantitative values here. The p-values for the supremum tests assessing the functional forms of age, alcohol, and tobacco were 0.001, 0.043, and 0.001, respectively. Figure 5 contains the cumulative residual plot for age. The pattern of the observed process resembles the pattern when a quadratic term is omitted from the model and the sign of the regression coefficient for the quadratic term is negative. The data were refit adding age^2, and the p-value for the supremum test for age increased to 0.123. Figure 6 contains the

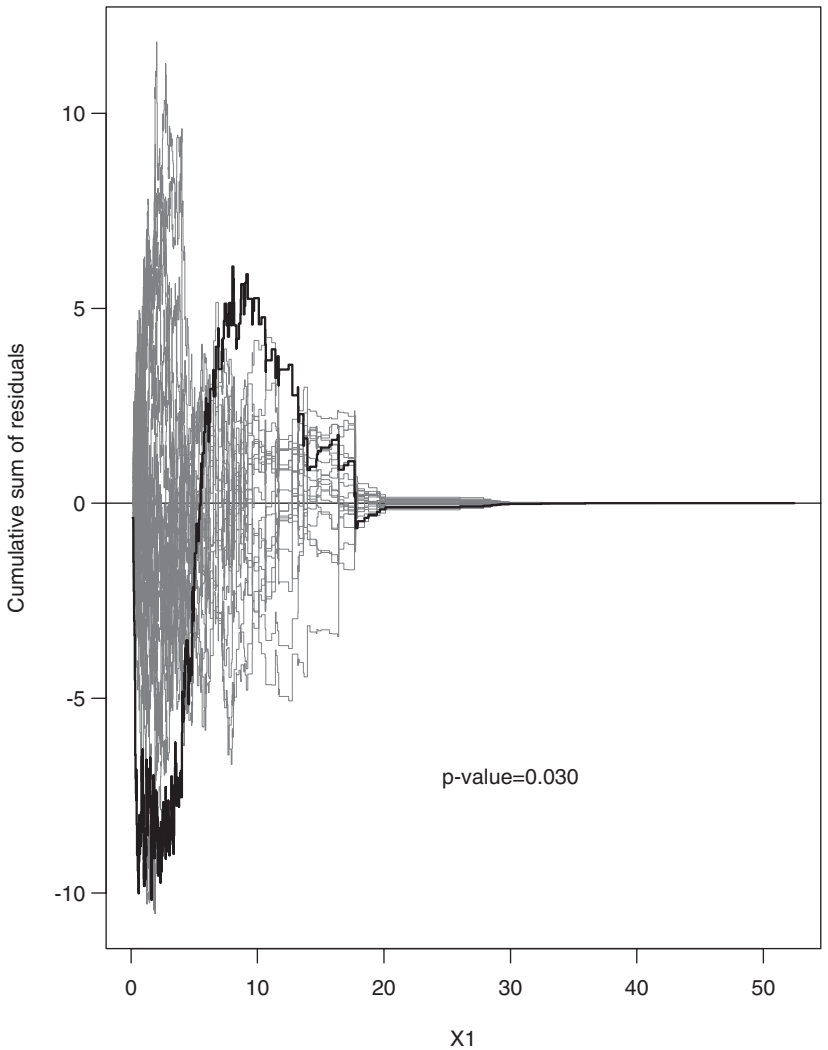

Figure 3. Plot of cumulative residuals versus X_1 in the misspecified logistic model for a simulated data set. The true functional form of X_1 is $\log(X_1)$ and the fitted model uses X_1. The black line indicates the observed process, and the grey lines indicate 20 simulated realizations.

cumulative residual plot for age in the quadratic model. The p-value for the supremum test for alcohol was increased to 0.146, whereas the p-value for the supremum test for tobacco use remained at 0.001.

Figure 7 shows the cumulative residual plot for tobacco use. The pattern of the observed process resembles the pattern when a logarithmic scale is misspecified as a linear scale. Note that when Breslow and Day modelled tobacco use as a continuous covariate, their final model expressed tobacco on a logarithmic scale. We refit the data using log(tobacco). However,

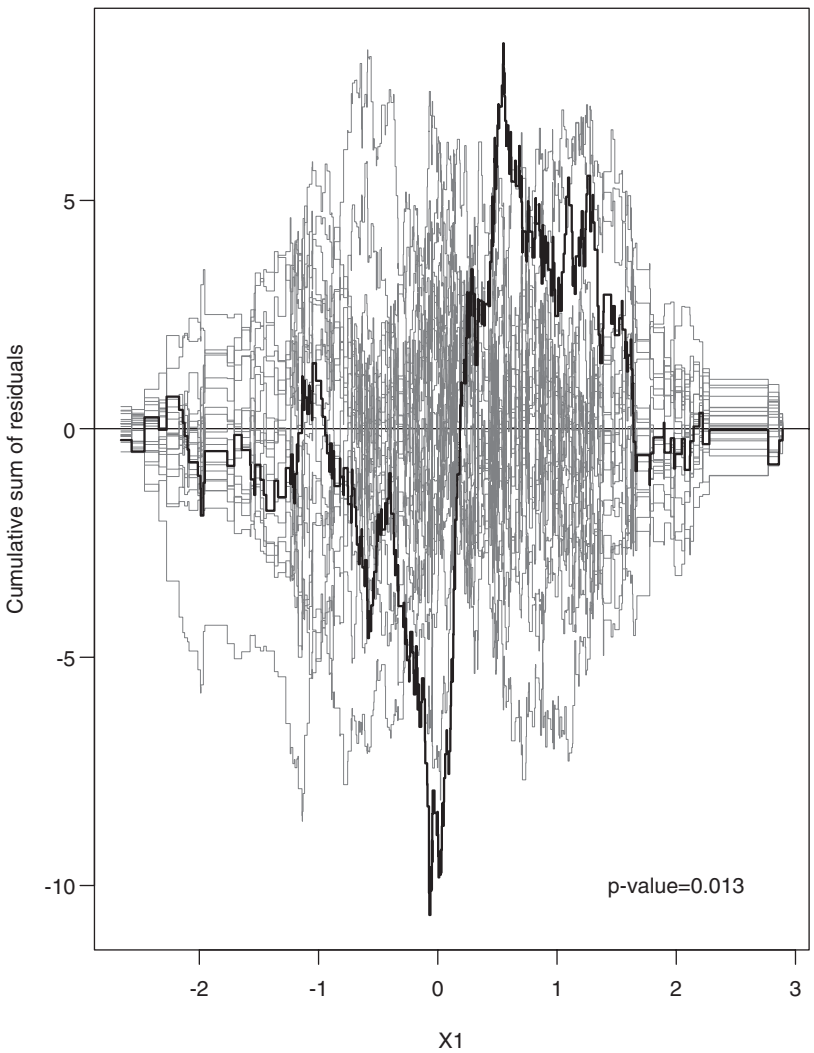

Figure 4. Plot of cumulative residuals versus X_1 in the true logistic model for a simulated data set. The true functional form of X_1 is $I(X_1 \geqslant 0)$ and the fitted model uses X_1. The black line indicates the observed process, and the grey lines indicate 20 simulated realizations.

the p-value for log(tobacco) was 0.003, still indicating misspecification. Figure 8 contains a cumulative residual plot for log(tobacco). The sharp drop at zero tobacco use is similar to the pattern observed when there is a threshold effect at zero, i.e. there is an association for any tobacco use. When we refit the data adding an indicator for tobacco users (a variable taking value 1 for tobacco users and 0 for non-users), the p-value for tobacco use increased to 0.067, indicating improvement. However, there was still some misspecification. In addition to using the natural logarithm of tobacco use, Breslow and Day expressed tobacco as four

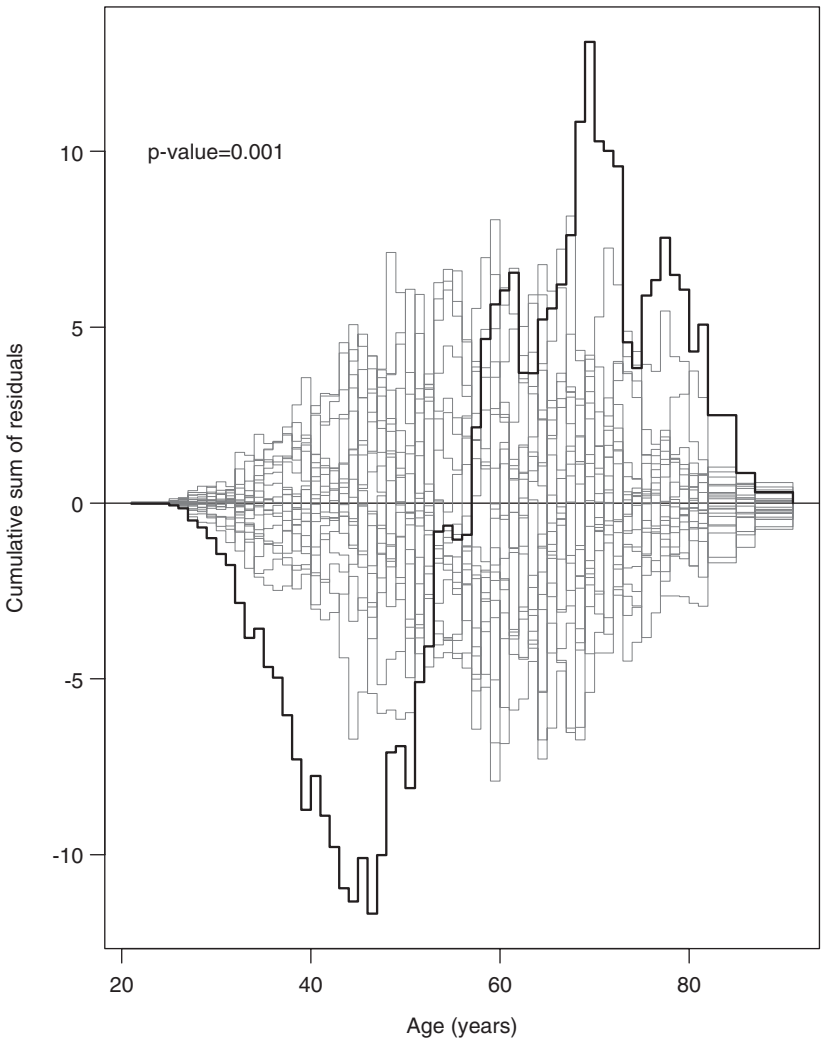

Figure 5. Plot of cumulative residuals versus age (years) in the logistic model with age, alcohol, and tobacco for the Ille-et-Vilaine oesophageal cancer data. The black line indicates the observed process, and the grey lines indicate 20 simulated realizations.

categories: 0–9, 10–19, 20–29, 30 + g/day. In light of this, we applied our method to this qualitative scale for tobacco use. However, based on our earlier cumulative residual plot, we split the lowest category into two groups to distinguish non-tobacco users from users of tobacco, i.e. we split the 0–9 g/day category into 0 and 1–9 g/day. Since tobacco use is now a categorical variable, the functional form of tobacco use is no longer an issue. The p-values for the supremum tests for assessing the overall model adequacy and the logistic link function were 0.297 and 0.104. Table II summarizes the results based on the final model.

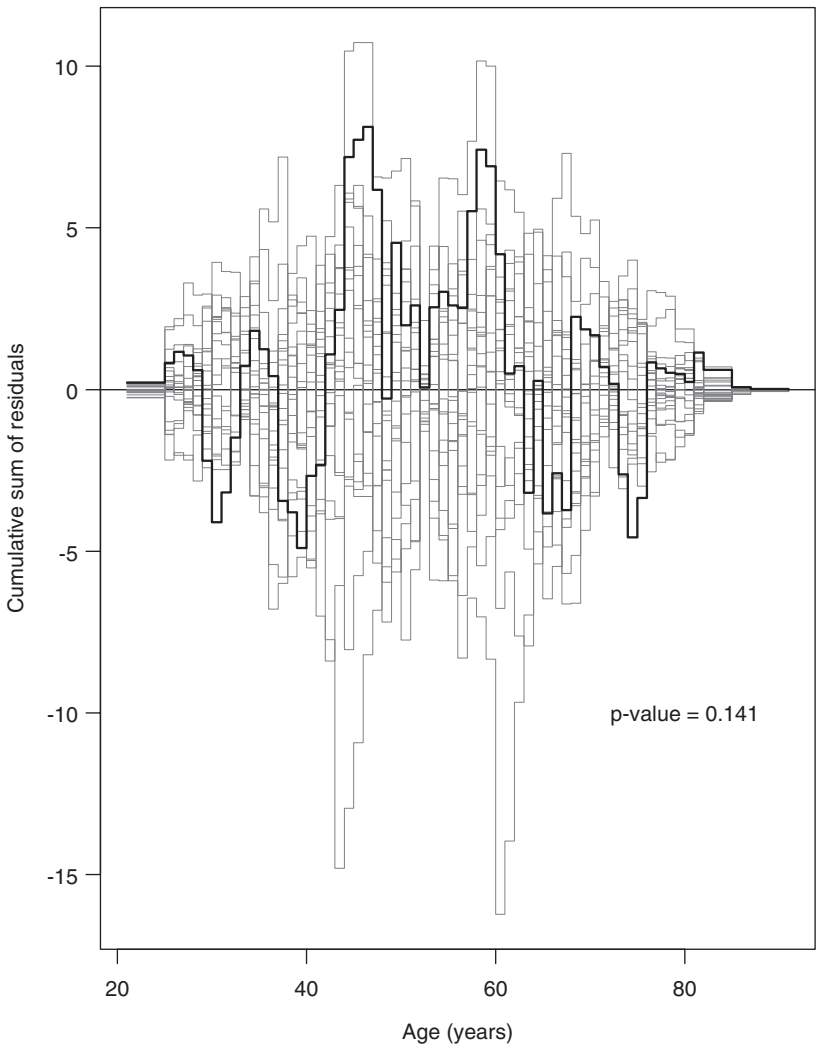

Figure 6. Plot of cumulative residuals versus age (years) in the logistic model with age, age^2, alcohol, and tobacco for the Ille-et-Vilaine oesophageal cancer data. The black line indicates the observed process, and the grey lines indicate 20 simulated realizations.

Based on the odds ratio estimates for tobacco use, it is apparent why log(tobacco) plus an indicator for any tobacco use improved the model fit although some misspecification was still present. Compared to non-tobacco users, any tobacco use is associated with an immediate increase in risk of oesophageal cancer, and then this risk plateaus as tobacco use increases. However, the inadequacy of the log(tobacco) manifests itself at $30+$ g/day of tobacco use, which is associated with a threefold increase in risk compared to the other non-zero tobacco use categories.

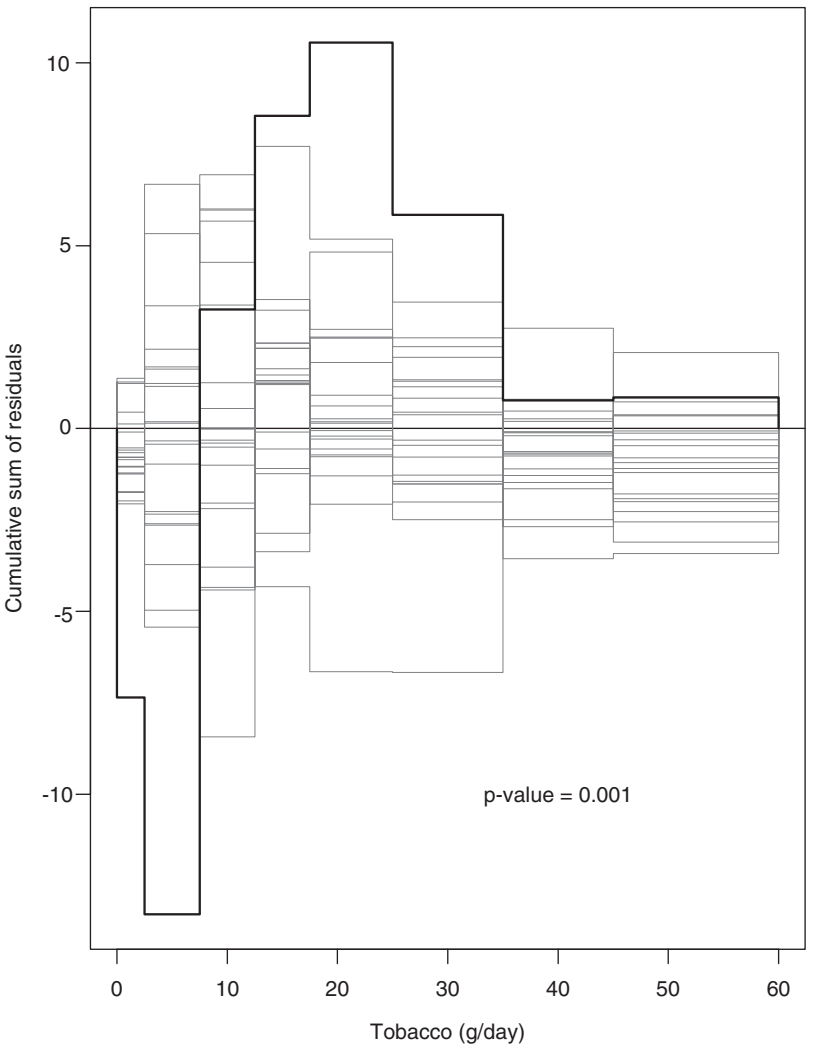

Figure 7. Plot of cumulative residuals versus tobacco use (g/day) in the logistic model with age, age^2, alcohol, and tobacco for the Ille-et-Vilaine oesophageal cancer data. The black line indicates the observed process, and the grey lines indicate 20 simulated realizations.

5. DISCUSSION

We have developed graphical and numerical methods for assessing the adequacy of the logistic regression model for stratified case-control studies using the cumulative sums of residuals. Similar methods have previously been developed for generalized linear models [13, 14] and the proportional hazards model [15, 16]. The outcome-based sampling scheme for the case-control study entails new challenges. The forms of the residuals, the cumulative sums, and the asymptotic approximations used for the stratified case-control design differ from those of

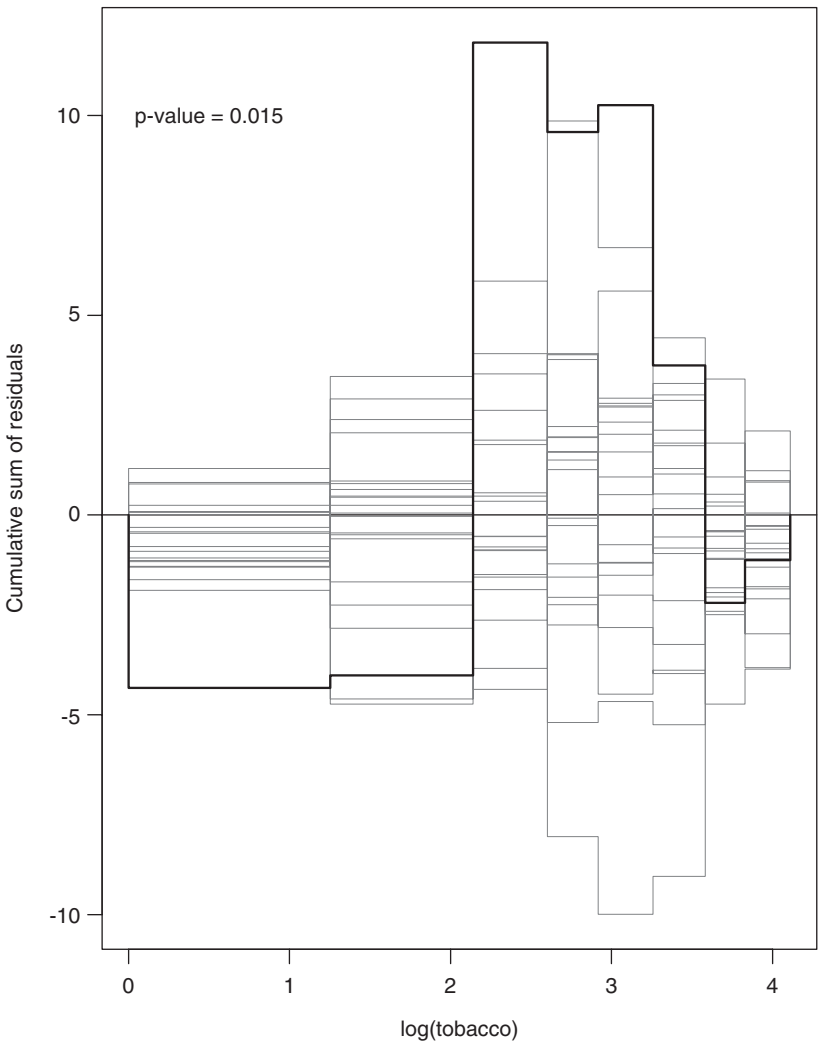

Figure 8. Plot of cumulative residuals versus log(tobacco) in the logistic model with age, age^2, alcohol, and log(tobacco) for the Ille-et-Vilaine oesophageal cancer data. The black line indicates the observed process, and the grey lines indicate 20 simulated realizations.

the other models. Furthermore, we have established the consistency of the supremum tests for stratified case-control studies.

All the proposed tests are testing the null hypothesis that the entire model is correct, although they are designed against different alternatives. Specifically, G_k is most sensitive to the misspecification of the functional form of X_k and G_ρ is most sensitive to the misspecification of the link function. Although we refer to G_ρ as a link function test, it actually tests both the link function and the linear predictor.

Table II. Final model for the Ille-et-Vilaine study data.

Parameter	Estimate	SE	OR	95 per cent CI
Age (year)	0.10	0.014	—	—
Age^2 (year)	−0.0034	0.00076	—	—
Alcohol* (g/day)	0.025	0.0026	1.03	1.02–1.03
Tobacco use†				
1–9 (g/day)	1.81	0.39	6.10	2.82–13.18
10–19 (g/day)	1.72	0.40	5.59	2.56–12.21
20–29(g/day)	1.81	0.43	6.11	2.65–14.07
30+ (g/day)	2.90	0.47	18.26	7.26–45.97

*Non-drinkers are the reference group.
†Non-smokers are the reference group.

Royston and Altman [17] and Royston *et al.* [18] proposed fractional polynomials as a means to investigate the functional form of continuous covariates, and Hosmer and Lemeshow [7] discussed their application to case-control data. Fractional polynomials can be used to reveal functional forms which improve the model fit. This approach is subjective, and cannot be used to assess the adequacy of a given functional form.

Recently, Pulkstnis and Robinson [19] developed two goodness-of-fit tests for logistic regression models analogous to the deviance and Pearson chi-squared tests as well as Hosmer and Lemeshow's goodness-of-fit test. Their tests are useful for detecting interactions, and their approach can be applied to case-control data. However, their tests tend to have low power when the functional form of a continuous covariate is misspecified. Furthermore, their tests require the logistic regression model to include both categorical and continuous covariates.

FORTRAN programs that implement the proposed methods are available from the authors. Future work consists of developing macros in commercial statistical software packages, such as SAS, R, and Strata, that implement our propose graphical and numerical techniques.

APPENDIX A

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

We first establish the weak convergence of $W_o(x;\widehat{\theta})$ under model (1). Consider the one-term Taylor series expansion of $W_o(x;\widehat{\theta})$ at θ:

$$W_o(x;\widehat{\theta}) = W_o(x;\theta) + \widehat{\eta}_o'(x;\theta^*)n^{1/2}(\widehat{\theta}-\theta) \tag{A1}$$

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

$$W_o(x;\theta) = n^{-1/2}\sum_{i=0}^{1}\sum_{j=1}^{J}\sum_{l=1}^{n_{ij}} r_{ijl}I(X_{ijl}\leqslant x)$$

Since each term $r_{ijl}I(X_{ijl}\leqslant x)$ is the difference of two monotone functions in x, the processes $\{r_{ijl}I(X_{ijl}\leqslant x);\ i=0,1;\ j=1,\ldots,J;\ l=1,\ldots,n_{ij}\}$ are 'manageable' [20, 21]. It then follows from the functional central limit theorem [20] that $W_o(x;\theta)$ is tight. Let $\eta_o(x;\theta)=$

$\lim_{n\to\infty}\widehat{\eta}_o(x;\theta)$. Since $\widehat{\eta}_o(x;\theta)$ converges almost surely to $\eta_o(x;\theta)$ and $n^{1/2}(\widehat{\theta}-\theta)$ converges in distribution, the second term on the right-hand side of (A1) is tight. Therefore, $W_o(x;\widehat{\theta})$ is tight.

Since $n^{1/2}(\widehat{\theta}-\theta)$ is asymptotically equivalent to $n^{-1/2}\Omega^{-1}\mathrm{U}(\theta)$, where $\Omega=\lim_{n\to\infty} n^{-1}\mathscr{I}(\theta)$, $W_o(x;\widehat{\theta})$ is asymptotically equivalent to

$$\widetilde{W}_o(x;\theta)=n^{-1/2}\sum_{i=0}^{1}\sum_{j=1}^{J}\sum_{l=1}^{n_{ij}} r_{ijl}\{I(X_{ijl}\leqslant x)+\eta_o'(x;\theta)\Omega^{-1}X_{ijl}\}$$

For fixed x, the finite-dimensional distributions of $\widetilde{W}_o(x;\theta)$ are asymptotically zero-mean multivariate normal, implying the same for $W_o(x;\widehat{\theta})$. This fact, together with the tightness of $W_o(x;\widehat{\theta})$, implies that $W_o(x;\widehat{\theta})$ converges weakly to a zero-mean Gaussian process with covariance function

$$\xi(s,t)=\sum_{i=0}^{1}\sum_{j=1}^{J} v_{ij}E\{\Psi_{ij1}(s)\Psi_{ij1}(t)'\}$$

at (s,t) as $n\to\infty$, where $v_{ij}=\lim_{n\to\infty}(n_{ij}n^{-1})$ and $\Psi_{ijl}(x)=r_{ijl}\{I(X_{ijl}\leqslant x)+\eta_o'(x;\theta)\Omega^{-1}X_{ijl}\}$.

The process $W_k(t;\widehat{\theta})$ is a special case of $W_o(x;\widehat{\theta})$ with $x_m=\infty$ for all $m\neq k$. Hence, the weak convergence of $W_k(t;\widehat{\theta})$ follows from the above result.

To establish the weak convergence of $W_\rho(t;\widehat{\theta})$, let $B_\varepsilon(\theta)=\{b:\|b-\theta\|\leqslant\varepsilon\}$, and suppose that for some $\varepsilon>0$, the function $\Pr(b'X\leqslant t)$ is continuous in $(b,t)\in B_\varepsilon(\theta)\times[t_1,t_2]$. If follows from the earlier arguments for $W_o(t;\widehat{\theta})$ that $W_\rho(t;\widehat{\theta})=\widetilde{W}_\rho(t;\widehat{\theta})+o_p(1)$, where

$$\widetilde{W}_\rho(t;b)=n^{-1/2}\sum_{i=0}^{1}\sum_{j=1}^{J}\sum_{l=1}^{n_{ij}} r_{ijl}\{I(b'X_{ijl}\leqslant t)+\eta_\rho'(t;b)\Omega^{-1}X_{ijl}\}$$

and $\eta_\rho(t;b)=\lim_{n\to\infty}\widehat{\eta}_\rho(t;b)$. Furthermore, $\widetilde{W}_\rho(t;b)$ converges weakly on $B_\varepsilon(\theta)\times[t_1,t_2]$ to a zero-mean Gaussian process and is stochastically equicontinuous [20]. In particular, $W_\rho(t;\widehat{\theta})$ and $\widetilde{W}_\rho(t;\theta)$ are asymptotically equivalent and thus converge to the same limiting Gaussian process.

Next, we establish the weak convergence of $\widehat{W}_o(x;\widehat{\theta})$. Conditional on the data (Y_{ijl},X_{ijl}) $(i=0,1;\ j=1,\ldots,J;\ l=1,\ldots,n_{ij})$, the only random components of $\widehat{W}_o(x;\widehat{\theta})$ are the $Z_{ijl}'s$, which are standard normal. Thus it follows from the multivariate central limit theorem that, conditional on the data, the finite-dimensional distributions of $\widehat{W}_o(x;\widehat{\theta})$ are asymptotically zero-mean multivariate normal. Since $\widehat{W}_o(x;\widehat{\theta})$ consists of monotone functions in x, which are manageable, the functional central limit theorem implies that $\widehat{W}_o(x;\widehat{\theta})$ is tight. The conditional covariance function of $\widehat{W}_o(x;\widehat{\theta})$ at (s,t) is $\hat{\xi}(s,t)=n^{-1}\sum_{i=0}^{1}\sum_{j=1}^{J}\sum_{l=1}^{n_{ij}}\widehat{\Psi}_{ijl}(s)\widehat{\Psi}_{ijl}(t)'$, where

$$\widehat{\Psi}_{ijl}(x)=\widehat{r}_{ijl}[I(X_{ijl}\leqslant x)+\widehat{\eta}_o'(x;\widehat{\theta})\{n^{-1}\mathscr{I}(\widehat{\theta})\}^{-1}X_{ijl}]$$

By the uniform law of large numbers, $\hat{\xi}(s,t)$ converges uniformly to $\xi(s,t)$. Therefore, $W_o(x;\widehat{\theta})$ and $\widehat{W}_o(x;\widehat{\theta})$ converge to the same limiting zero-mean Gaussian process. Similar arguments can be used to establish the weak convergence of $\widehat{W}_\rho(t;\widehat{\theta})$, and its asymptotic equivalence to $W_\rho(t;\widehat{\theta})$.

A.2. Consistency of supremum tests

It can be shown that model (1) holds if and only if model (2) holds. Given this fact, we can establish the consistency of the proposed supremum tests on the basis of model (2). Under misspecified models, $\widehat{\theta}$ converges to a well-defined constant vector, say θ^*. Let v_{ijl} denote the true probability that a subject in the case-control sample from stratum j with covariates X_{ijl} is a case.

Consistency of $G_o = \sup_{x \in \mathscr{R}^{J+p}} |W_o(x; \widehat{\theta})|$: We claim that the test based on G_o is consistent against the general alternative H_1 that there does not exist a constant vector θ such that the true conditional probability of disease can be expressed by (2) for all possible values of X. It suffices to show that under H_1, $n^{-1/2}G_o$ is non-zero as $n \to \infty$. Under H_1,

$$n^{-1/2} W_o(x; \widehat{\theta}) \to \lim_{n \to \infty} n^{-1} \sum_{i=0}^{1} \sum_{j=1}^{J} \sum_{l=1}^{n_{ij}} \{v_{ijl} - p_j(X_{ijl}; \theta^*)\} I(X_{ijl} \leqslant x)$$

which is non-zero at least for some x. Consequently, $n^{-1/2}G_o$ converges to a non-zero constant. This establishes our claim.

Consistency of $G_\rho = \sup_{t \in \mathscr{R}} |W_\rho(t; \widehat{\theta})|$: Clearly,

$$n^{-1/2} W_\rho(t; \widehat{\theta}) \to \lim_{n \to \infty} n^{-1} \sum_{i=0}^{1} \sum_{j=1}^{J} \sum_{l=1}^{n_{ij}} \{v_{ijl} - p_j(X_{ijl}; \theta^*)\} I(\theta^{*\prime} X_{ijl} \leqslant t)$$

which is non-zero for some t unless $v_{ijl} = p_j(X_{ijl}; \theta^*)$ for all values of $\theta^{*'} X_{ijl}$. Thus, the test G_ρ is consistent against misspecification of the link function in the form of $\Pr(Y = 1 | X^\dagger) = h(\alpha_j^* + \beta^{*'} X)$, where h is not the logistic function.

Consistency of $G_k = \sup_{t \in \mathscr{R}} |W_k(t; \widehat{\theta})|$: Suppose that the true functional form of the kth covariate component is $f(X_k)$ rather than X_k, i.e.

$$v_{ijl} = \frac{\exp\{\delta_j + \beta_k f(X_{ijlk}) + \sum_{m \neq k} \beta_m x_{ijlm}\}}{1 + \exp\{\delta_j + \beta_k f(X_{ijlk}) + \sum_{m \neq k} \beta_m x_{ijlm}\}}$$

Then

$$n^{-1/2} W_k(t; \widehat{\theta}) \to \lim_{n \to \infty} n^{-1} \sum_{i=0}^{1} \sum_{j=1}^{J} \sum_{l=1}^{n_{ij}} \{v_{ijl} - p_j(X_{ijl}; \theta^*)\} I(X_{ijlk} \leqslant t)$$

which is non-zero for some t unless $v_{ijl} = p_j(X_{ijl}; \theta^*)$ for all values of X_{ijlk}. In general, $\beta_k^* \neq \beta_k$. It follows that $v_{ijl} \neq p_j(X_{ijl}; \theta^*)$ if $\beta_m^* = \beta_m$ for all $m \neq k$. 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 $p_j(X_{ijl}; \theta^*) = v_{ijl}$ for all values of X_{ijlk}. Hence, G_k is generally consistent against misspecification of the functional form.

REFERENCES

1. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika* 1979; **66**: 403–411.
2. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine* 1997; **16**:965–980.

3. Pregibon D. Logistic regression diagnostics. *The Annals of Statistics* 1981; **9**:705–724.
4. Hosmer DW, Lemeshow S. A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics* 1980; **A10**:1043–1069.
5. Osius G, Rojek D. Normal goodness-of-fit tests for multinomial models with large degrees-of-freedom. *Journal of the American Statistical Association* 1992; **87**:1145–1152.
6. Stukel TA. Generalized logistic models. *Journal of the American Statistical Association* 1988; **83**:426–431.
7. Hosmer DW, Lemeshow S. *Applied Logistic Regression* (2nd edn). Wiley: New York, 2000.
8. Qin J, Zhang B. A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika* 1997; **84**:609–618.
9. Zhang B. A chi-squared goodness-of-fit test for logistic regression models based on case-control data. *Biometrika* 1999; **86**:531–539.
10. Breslow NE, Day NE. *Statistical Methods in Cancer Research, vol. 1. The Analysis of Case-Control Studies*. International Agency for Research on Cancer: Lyon, 1980.
11. Scott AJ, Wild CJ. Fitting logistic models under case-control or choice-based sampling. *Journal of the Royal Statistical Society, Series B* 1986; **48**:170–182.
12. McCullagh P, Nelder JA. *Generalized Linear Models* (2nd edn). Chapman & Hall: London, 1989.
13. Su JQ, Wei LJ. A lack-of-fit test for the mean function in a generalized linear model. *Journal of the American Statistical Association* 1991; **86**:420–426.
14. Lin DY, Wei LJ, Ying Z. Model-checking techniques based on cumulative residuals. *Biometrics* 2002; **58**:1–12.
15. Lin DY, Wei LJ, Ying Z. Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* 1993; **80**:557–572.
16. Spiekerman CF, Lin DY. Checking the marginal Cox model for correlated failure time data. *Biometrika* 1996; **83**:143–156.
17. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Applied Statistics* 1994; **43**:429–467.
18. Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology* 1999; **28**:964–974.
19. Pulkstnis E, Robinson TJ. Two goodness-of-fit tests for logistic regression models with continuous covariates. *Statistics in Medicine* 2002; **21**:79–93.
20. Pollard D. *Empirical Processes: Theory and Applications*. Institute of Mathematical Statistics: Hayward, 1990.
21. Billias Y, Gu M, Ying Z. Towards a general asymptotic theory for Cox model with staggered entry. *The Annals of Statistics* 1997; **25**:662–682.

