Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Goodman and Li, J Biomet Biostat 2012, 3:2
DOI: 10.4172/2155-6180.1000136

Research Article Open Access

Nonparametric Diagnostic Test for Conditional Logistic Regression

Melody S. Goodman* and Yi Li

[1]*Division of Public Health Sciences, Department of Surgery, Washington University in St. Louis School of Medicine, St. Louis, MO, USA*
[2]*Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, USA*

Abstract

The use of conditional logistic regression models to analyze matched case-control data has become standard in statistical analysis. However, methods to test the fit of these models has primarily focused on influential observations and the presence of outliers, while little attention has been given to the functional form of the covariates. In this paper we present methods to test the functional form of the covariates in the conditional logistic regression model, these methods are based on nonparametric smoothers. We assess the performance of the proposed methods via simulation studies and illustrate an example of their use on data from a community based intervention.

Keywords: Conditional logistic regression; Kernel smoother; Model diagnostics

Introduction

Conditional logistic regression, an important extension of the logistic regression model, allows for the analysis of data with stratified samples [1]. Stratified samples are often encountered in epidemiological research. Stratification can occur in the study design (e.g., when the data are collected from different sites) or during analysis (e.g., controlling for a covariate). A commonly used stratified study design is the matched case-control study. Conditional logistic regression is often used to investigate the relationship between an outcome and a set of prognostic factors in matched case-control studies, as it is designed for the analysis of data with small stratum-specific sample sizes. In such a setting the outcome of interest is whether a subject is a case or a control [2]. Like other regression models, conditional logistic regression models allow for multiple variables, continuous exposures, confounding, and effect-modifying variables to be handled appropriately [3].

Although most would agree that assessing model fit is an important step in data analysis, it is often skipped when diagnostic tests are not readily available or easily implemented. We are not aware of a standard statistical package that has the option to calculate diagnostic statistics for any matched design. Thus the only option for many analysts is an ad-hoc approach that is done by creating a data set containing the difference variables and using standard logistic regression diagnostics [2]. The approaches currently available in the literature mainly focus on diagnostic methods that test for influential pairs and outliers [2,4-7], and do not have the ability to test the overall model adequacy. We are interested in testing whether the functional form of the covariates is correctly specified. The lack of existing methods leads us to develop a diagnostic test for conditional logistic regression based on nonparametric smoothing.

In this paper we propose a nonparametric diagnostic test for matched case-control conditional logistic regression to test the functional form of the covariates. We briefly review the conditional logistic regression model and existing diagnostic test; a more detailed review can be found elsewhere (for example, see [2]). Next, we review a method proposed by Hart [8] for nonparametric diagnostic tests in a linear model. We extend this methodology to a matched case-control conditional logistic regression setting and show the Type I error and power of the test statistic via simulations. We illustrate an application of this methodology using the Healthy Directions data [9,10].

Conditional Logistic Regression Model

Consider a matched case-control study with K matched sets. The sets are determined by values of the matching variables, in this case there are K distinct possible matching groups. Suppose there are n_k subjects in stratum k, k =1,2…,K of which n_{1k} are cases and n_{0k} are controls. The stratum specific logistic regression model has the form

$$\pi_k(\mathbf{x}) = \frac{e^{\alpha_k + \beta'\mathbf{x}}}{1 + e^{\alpha_k + \beta'\mathbf{x}}},$$

where α$_k$ denotes the contribution to the logit of all constant terms within the k^{th} stratum. $\beta' = (\beta_1, \beta_2, \ldots \beta_p)$ is the vector of coefficients, where β_q is the change in log-odds for a one unit increase in the covariate x_q holding all other covariates constant in every stratum [2].

The conditional probability for the k^{th} stratum is obtained as the probability of the observed data conditional on the stratum total and the number of cases observed:

$$l_k(\beta) = \frac{\prod_{i=1}^{n_{1k}} P(\mathbf{x_i} \mid y_i = 1) \prod_{i=n_{1k}+1}^{n_k} P(\mathbf{x_i} \mid y_i = 0)}{\sum_{j=1}^{c_k} \left[\prod_{i_j=1}^{n_{1k}} P(\mathbf{x}_{ji_j} | y_{i_j} = 1) \prod_{i_j=n_{1k}+1}^{n_k} P(\mathbf{x}_{ji_j} | y_{i_j} = 0) \right]}$$

$$= \frac{\prod_{i=1}^{n_{1k}} e^{\beta'\mathbf{x}_i}}{\sum_{j=1}^{c_k} \prod_{i_j=1}^{n_{1k}} e^{\beta'\mathbf{x}_{ji_j}}},$$

where c_k is the number of possible assignments of case status to n_{1k} subjects among the n_k subjects, and j denotes any one of the c_k assignments. For any assignment we let subjects 1 to n_{1k} correspond to cases and subjects n_{1k} +1 to n_k to the controls. This is indexed by i for the observed data and by i_j for the jth possible assignment. The full conditional likelihood is the product of $l_k(\beta)$ over the K strata [2].

When the matching is 1:1, the conditional likelihood for the k^{th} stratum is given by

*Corresponding author: Melody S. Goodman,Division of Public Health Sciences, Department of Surgery, Washington University in St. Louis School of Medicine, St. Louis, MO, USA, E-mail: goodmanm@wustl.edu

Received November 30, 2011; **Accepted** March 27, 2012; **Published** March 28, 2012

Citation: Goodman MS, Li Y (2012) Nonparametric Diagnostic Test for Conditional Logistic Regression. J Biomet Biostat 3:136. doi:10.4172/2155-6180.1000136

$$l_k(\beta) = \frac{e^{x_{1k}\beta}}{e^{x_{1k}\beta} + e^{x_{0k}\beta}},$$

where x_{1k} is the data vector for the case and x_{0k} is the data vector for the control. For matched case-control studies with one case per match set, this likelihood function reduces to that of the cox model for the continuous time scale [11].

Conditional logistic regression diagnostics

As with all other statistical models, the common practice of using conditional logistic regression models raises questions about model fit and the stability of parameter estimates. Pregibon [12] proposes model diagnostic methods for logistic regression and more specifically for conditional logistic regression [5]. Two of these methods are ANOVA-like tables and one-degree-of-freedom test for model adequacy. ANOVA-like tables use deviance measures broken down into three categories, namely total deviance, unexplained deviance, and explained deviance, similar in spirit to the way the analysis of variance table breaks down the variability in a linear model into explainable (fitted effects) and unexplainable (residual) components. These deviance measures can be used to develop summary measures such as R^2; that is, if the explained deviance accounts for a large portion of the total deviance, then some of the variables included in the fitted model are important. For matched case-control studies R^2 measures the strength of the linear association between the log odds ratio and exposure variables. Since R^2 cannot adequately quantify nonlinear associations, and it measures the strength but not the adequacy of a linear association, it cannot be thought of as a goodness-of-fit statistic; and therefore it should be used with caution in this setting [5]. If the data analyst has an idea of the type of model variation being experienced then one-degree-of-freedom test for model adequacy can be used to test the hypothesis that the model is deviant in some way. This is usually done by testing an augmented model (including the hypothesized deviant) to the current model. This method is not useful if the type of model deviation is unknown or the model is deviant in more than one way [5]. A major disadvantage of both of these methods is that they are highly influenced by the presence of outliers. Outliers can have dramatic effects on model fit and parameter estimates even when there are large sample sizes [4,12].

Diagnostic tests for detecting the influence of outliers or influential pairs on matched case-control analysis as proposed by Pregibon [5], Moolgavar et al. [4,6], Bedrick and Hill [7], and Hosmer and Lemeshow [2] do not have the ability to test overall model adequacy. Arbogast and Lin [13] have developed methodology for assessing the adequacy of the functional form of the covariates, the logistic link function, as well as the overall model fit for matched case-control data. The methods they propose are based on the cumulative residual process. Disadvantages of the methods they propose is that they are computationally intensive (can be done in minutes with the power of computers available now), were implemented in Fortran, and provide yes/no results which does not provide any insight about how the model is deviant when the null hypothesis is rejected.

Nonparametric Diagnostic Test for the Linear Model

Hart [8] proposes a method to test the lack of fit in a linear model using a linear smoother. Consider the model

$$Y_i = h(x_i) + \epsilon_i, \qquad i = 1,\ldots,n,$$

in which $\epsilon_1,\ldots,\epsilon_n$ are zero mean, independent random variables with constant variance $\sigma^2 < \infty$. The principal aim is to learn about the relationship between x and Y as it is expressed through the regression function h. In a parametric approach to inferring h, one assumes that

$$h \in S_\Theta \equiv \{h(\cdot;\theta) : \theta \in \Theta\},$$

where Θ is some subset of p-dimensional Euclidean space [8].

The lack-of-fit test is based on smoothing methodology, with an interest in testing the null hypothesis that h is in some parametric class of functions S_Θ against the alternative that h is not in S_Θ. The idea behind this methodology is that one computes a smooth curve and compares it with a curve that is "expected" under the null hypothesis. If the smooth curve differs sufficiently from the expected curve, then there is evidence that the null hypothesis is false [8]. Smoothing based tests have several advantages; they are omnibus in the sense of being consistent against each member of a very large class of alternative hypotheses, they tend to be more powerful than some of the well-known omnibus tests, and they come with a smoother [8].

A linear smoother can be used to test the fit of a model, or formally the null hypothesis,

$$H_0 : h \in S_\Theta = \{h(\cdot;\theta) : \theta \in \Theta\}.$$

When applied to the residuals, a linear smoother at the point x has the form:

$$\hat{g}(x;S) = \sum_{i=1}^{n} w_i(x;S)\hat{e}_i, \tag{1}$$

where the weights $w_i(x; S)$, i =1,..,n are constants that do not depend on the data $Y_1,\ldots,Y_n$, or any unknown parameters, S denotes the value of a smoothing parameter, and e_i are the residuals [8].

Suppose that $\hat{h}(\cdot;S)$ is a nonparametric estimate of h based on a linear smooth of $Y_1,\ldots,Y_n$, $\hat{\theta}$ denotes our estimate of θ based on the assumption that the null model is true, and $\hat{g}(x;S)$ is defined in equation 1. It can be shown [8] that,

$$\hat{h}(x;S) - h(x;\hat{\theta}) = \hat{g}(x;S) + \text{Bias}\{\hat{h}(x;S),\hat{\theta}\}, \tag{2}$$

where Bias $\{\hat{h}(x;S),\hat{\theta}\}$ denotes the bias of $\hat{h}(x;S)$ when the null is true and θ is the true parameter value.

If the null hypothesis is true, the residuals should behave like zero mean, uncorrelated random variables. Therefore, the linear smoother in equation (1) should be relatively flat and centered around zero. A subjective diagnostic is to plot the estimate $\hat{g}(\cdot;S)$ and see how much it differs from the zero function. Often a pattern will emerge in the smooth that was not evident in a plot of the residuals. However, looks can be deceiving so it is important to have a statistic that objectively measures the difference between $\hat{g}(\cdot;S)$ and zero. Hart [8] proposes

$$R_n = \frac{n^{-1}\sum_{i=1}^{n} \hat{g}^2(x_i;S)}{\hat{\sigma}^2},$$

where the numerator measures the "size" of the function g and $\hat{\sigma}^2$ is a model free estimator of the variance σ^2. It can be shown [8] that

$$\frac{R_n - E(R_n)}{\sqrt{var(R_n)}} \xrightarrow{D} N(0,1).$$

Extension of Nonparametric Diagnostic test to Conditional Logistic Models

Consider the model

$$z(\mu(x)) = f(x_i) + \eta_i,\ i = 1,\ldots,n$$

where $\mu(x) = E(Y \mid x)$, $z(\cdot)$ is the logit function, and x_i, $i = 1,\ldots n$ are fixed design points. We assume that $\eta_1,\ldots,\eta_n$ are independent random variables with $E(\eta_i) = 0$ and $Var(\eta_i) = \sigma^2 < \infty$. As in the case of simple regression, the principal aim is to learn about the relationship between x and Y as it is expressed through the regression function f. In a parametric approach to inferring f, one assumes that

$$f \in S_\Theta \equiv \{f(\cdot;\theta) : \theta \in \Theta\},$$

where Θ is some subset of p-dimensional Euclidean space [8].

We will use a linear smoother, one that is linear in the residuals, to test the fit of our model; in this case the null hypothesis is

$$H_0 : f(\mathbf{x}) = \mathbf{x}'\beta.$$

Our linear smoother will be similar to that of equation (1), however **x** is now a vector of length n , the total number of observation in strata i, instead of a scalar. The smoother will have the form

$$\hat{m}(\mathbf{x};S) = \sum_{i=1}^{n} \mathbf{w_i}(\mathbf{x};S)\mathbf{\hat{r}_i},$$

where the weights $\mathbf{w}_i(\mathbf{x}; \mathrm{S})$ $i=1,\ldots,n$, are constant vectors for the i^{th} strata that do not depend on the data $Y_1,\ldots,Y_n$, or any unknown parameters, S denotes the value of a smoothing parameter and $\mathbf{r_i}$ is the vector of residuals for the i^{th} strata.

Formulation of the test statistic for 1:1 matched case-control study

Since each strata has two observations, let $\mathbf{x_i} = [x_{i,1}\ x_{i,2}]'$ and $\mathbf{\hat{r}_i} = \left[\hat{r}_{i,1} \quad \hat{r}_{i,2}\right]'$. For any vector $\mathbf{x} = [x_1\ x_2]$ our smoother will have the form

$$\hat{m}(\mathbf{x};S) = \sum_{i=1}^{n} \mathbf{w_i}(\mathbf{x};S)'\mathbf{\hat{r}}_i, \tag{3}$$

Where $\mathbf{w_i}(\mathbf{x}; S)$ is a vector valued function defined by $\mathbf{w_i}(\mathbf{x};S)' = [w_i(x_1)\ w_i(x_2)]$, and $w_{i,j} = w_i(x_j)$. Thus our smoother is basically a weighted sum of the residuals. We use a Nadaraya-Watson type kernel smoother to define the weights

$$\mathbf{w_i}(\mathbf{x};S) = \mathbf{K_i}\left(\frac{\mathbf{x}-\mathbf{x_i}}{S}\right),$$

where $\mathbf{K_i} = \left[K_i\left(\frac{x_1-x_{i,1}}{S}\right) \quad K_i\left(\frac{x_2-x_{i,2}}{S}\right)\right]'$, and K is the Epanechnikov kernel with the form

$$K(u) = \frac{3}{4}(1-u^2)I_{(-1,1)}(u).$$

Thus our weights are defined by

$$w_{i,j} = \frac{3}{4}\left[1-\left(\frac{x_j - x_{i,j}}{S}\right)^2\right] I_{(-1,1)}\left(\frac{x_j - x_{i,j}}{S}\right).$$

Our lack-of-fit test statistic will have the form

$$R = \frac{n^{-1}\sum_{k=1}^{n} \hat{m}^2(\mathbf{x_k};S)}{\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{i}^{n} \mathbf{w_i}(\mathbf{x_k})' var(\mathbf{r_i} \mid Y_{i,1}+Y_{i,2}=1)\mathbf{w_i}(\mathbf{x_k})\right]}. \tag{4}$$

Using the central limit theorem of quadratic forms, as proposed by De Jong [14], it can be shown that

$$\frac{R-E(R)}{\sqrt{var(R)}} \rightarrow N(0,1)$$

when the difference between R and $W(n)$ is negligible asymptotically, where $W(n) = \Sigma_{1\le i\le n}\Sigma_{1\le j\le n}\ a_{ij}X_iX_j$. Since the $var(R)$ is not a trivial calculation, we use bootstrapping to approximate the distribution of our test statistic. Using arguments from Hall and Hart [15], Hart [8] suggests that under certain conditions the bootstrap approach can yield a better approximation to the sampling distribution of a test statistic than the normal distribution.

Choice of smoothing parameter

The smoother in equation 3 and test statistic in equation 4 are dependent on a smoothing parameter, S. Choosing the appropriate value of the smoothing parameter is imperative as different values of S correspond to different tests of the null hypothesis. Ideally one would want to choose a value that maximizes power. However, often times the most powerful parameter will depend on some unknown function [16]. There are different types of techniques that can be used to determine the appropriate value of the smoothing parameter, "by-eye", data driven, and reasonable idea (i.e., the experimenter has a reasonable idea of the type of alternative to expect and chooses a smoothing parameter that is optimal over that class of alternatives). There have been some suggestions [8,16] about the use of data driven methods such as cross validation in choosing the appropriate value. King et al. [16] suggest that there are some drawbacks to these methods as they were not designed to maximize the power of a test and they add randomness to the test, effecting the distributional theory. Hart [8] suggest that the technique used to identify the appropriate value of the smoothing parameter should depend on the data analyst's reasons for fitting a nonparametric smooth curve.

When one is unsure about the appropriate value of the smoothing parameter the significance trace approach plots the p-values for a range of smoothing parameters [17]. This allows for the assessment of the effect of choice of smoothing parameter on the results of the test. When the resulting p-values are all above or below α the result of the test is clear as conclusions made are independent of the smoothing parameter. Azzalini and Bowman [17] suggest a range of values for which the smoothing parameter should be examined and provide suggestions on how to handle cases where an interpretation of the significance trace is not completely conclusive since it fluctuates around the significance level.

A Simulation Study

We conducted a simulation study to investigate the Type I error rate and power of our proposed test statistic in a 1:1 matched conditional logistic regression model. These studies were conducted as follows. We simulated data from a matched case control conditional logistic regression model of the form:

$$logit(Y_{ij}|Y_{i1}+Y_{i2}=1) = \beta x_{ij},$$

$i = 1,..,n$ $j = 1,2$ where x_{ij} is a uniform random variable and $\beta = 3.5$. For each sample size considered $n = 25$, 50, and 100, where n the number of matched sets, 500 independent sets of data is were generated. For a given set of data, the test statistic R (equation 4) was calculated. 5,000 bootstrap samples of size n were generated from each of the 500 data sets by drawing with replacement from the residuals $\mathbf{r}_1,\ldots,\mathbf{r_n}$. Bootstrap test of the nominal level $\alpha = 0.05$ were conducted for various values of the smoothing parameter.

The Type I error rate seems to improve with increases in sample size, and seems to be affected by the value of the smoothing parameter, however, for certain values of the smoothing parameter the Type I error rates are consistently on target. In Table 1 we can see that regardless of sample size, with a smoothing parameter of 0.2 (what we consider a medium size parameter), we simulate a type I error equivalent to the level of the test.

When the sample size is low (i.e. 25 pairs) the type I error fluctuates a bit as the smoothing parameter changes. This is most likely a sample size issue with the application of the central limit theorem. However, with a larger sample size (i.e. 50 or 100 pairs), the CLT works well and the validity of the test does not depend on the value of the smoothing parameter under the null hypothesis. Hence, our type I error is well controlled at 0.05 for various choices of smoothing parameters in larger samples.

We tested the power of our test statistic against two common alternatives. First we tested the power of our test statistic against a quadratic alternative, where the null model is missing the quadratic term. Power simulations were conducted in a similar manner as the Type I error simulations. The null and alternative models for the quadratic alternative are specified as follows:

$$H_0 : logit(Y_{ij}|Y_{i1}+Y_{i2} =1) = \beta_1 x_{ij}$$

$$H_1 : logit(Y_{ij} \mid Y_{i1} + Y_{i2} = 1) = \beta_1 x_{ij} + \beta_2 x_{ij}^2,$$

where β_1 = 3.5 and β_2 = 2.5. Since the null model is nested in the alternative we also calculated the power of the Likelihood Ratio Test (LRT). The power studies for the LRT were conducted as follows. For each sample size, n = 25, 50, and 100, we simulated 5,000 independent data sets under the alternative model and fit conditional logistic regression models under both the null and alternative hypotheses and computed a likelihood ratio test statistic for each data set.

Table 2 presents the simulation results; we see that the power of our test statistic increases with increases in sample size. Samples as small as 50 have very good power against this alternative. Since the power of the LRT is independent of the value of the smoothing parameter we compare the power of the LRT for a given sample size to the power of our test statistic at various values of the smoothing parameter for that sample size. Against the missing quadratic alternative the LRT has more power than our test statistic, but our test statistic performs fairly well compared to the LRT which is most powerful in this setting.

#of Matched Sets	Smoothing Parameter	Type I Error
25	0.02	0.068
	0.2	0.050
	2.0	0.064
50	0.02	0.058
	0.2	0.050
	2.0	0.048
100	0.02	0.052
	0.2	0.050
	2.0	0.054

Table 1: Empirical Level of Test Statistic in Bootstrap Simulation Study.

# of Matched Sets	Power of LRT	Smoothing Parameter	Power
25	0.847	0.2	0.70
		0.8	0.77
		2.0	0.79
50	0.985	0.2	0.94
		0.8	0.96
		2.0	0.97
100	0.996	0.2	0.98
		0.8	0.99
		2.0	0.99

Table 2: Power of Test Statistic against Quadratic Alternative.

Next we tested the power of our test statistic against the log alternative, where the null model is linear in the covariates and the alternative model is log linear in the covariates. The null and alternative models are specified as follows:

$$H_0 : logit(Y_{ij} \mid Y_{i1}+Y_{i2} =1) = \beta_1 x_{ij},$$

$$H_1 : logit(Y_{ij} \mid Y_{i1}+Y_{i2} =1) = \beta_1 \log (x_{ij}),$$

where β_1 = 3.5. Here we are assessing the power of the likelihood ratio test against model misspecification. To calculate power of the LRT we generated data under the alternative that the covariates are log linear (H_A) and calculated the likelihood ratio test statistic as if the alternative model was the linear model (H_1) against the null model (H_0):

$$H_A : logit(Y_{ij} \mid Y_{i1}+Y_{i2} =1) = \beta_1 log(x_{ij}),$$

$$H_0 : logit(Y_{ij} \mid Y_{i1}+Y_{i2} =1) = \beta_1$$

$$H_1: logit(Y_{ij} \mid Y_{i1}+Y_{i2} =1) = \beta_1 x_{ij}$$

The results of the simulation are presented in Table 3, the power of our test statistic improves with increases in the number of matched sets. Samples sizes as small as 25 have power between 70 and 80 percent, and sample sizes greater than 50 have power of over 89 percent. The LRT does not perform well against model misspecification especially for small sample sizes and our test statistic has more power in all cases.

In summary, we can see from simulation that there is some sensitivity of the results to the choice in smoothing parameter. For small sample sizes (n=25) our test statistic did reasonably well in the Type I error and power analysis. For samples of size 50 or larger our test statistic appears to have good properties, irrespective of the value of the smoothing parameter.

Application to the Healthy Directions Data: An Example

We illustrate the proposed methods by applying them to the data from the Harvard Cancer Prevention Program Project (HCPPP) Healthy Directions, which is composed of two randomized controlled trials, one in health centers (HC)[10], and another in small businesses (SB) [9]. The overarching goal of the HCPPP was to create a new generation of cancer prevention interventions that would be effective among working class, multi-ethnic populations. The study aims and sampling strategies are published in greater detail elsewhere [9,10].

The primary goals of the intervention were to have participants modify their behavioral risk factors for Cancer, namely increase fruit and vegetable intake, decrease red meat consumption, increase physical activity levels, and increase multivitamin intake. The investigators were interested in a summary measure to determine how well an individual did on the intervention as a whole. We developed a multiple risk factor summary score based on the assumption that the summary measure would be a linear combination of the four behavioral risk factors. We used a conditional logistic regression model and analyzed the data as if it came from a matched case-control study, where each individual is a control, pre-intervention, and a case, post-intervention. The outcome of the conditional logistic regression model is a subjects' intervention status (pre or post-intervention). The coefficients for the score are the parameter estimates from the conditional logistic regression model. The development of the multiple risk factor summary measure can be found in greater detail elsewhere [18].

After developing our summary score we wanted to formally test the assumption that the relationship between the individual continuous risk factors and the logit of the intervention status is linear. To do so we used a linear smoother (equation 3) and test statistic (equation 4). The outcome Y is an indicator of intervention time, pre-intervention or post-intervention, x is the number of fruits and vegetables consumed per week by the study participant. Approximations of the P-values were estimated using bootstrap, and determined by values of our test statistic from bootstrap samples that were more extreme than the value of our test statistic from our observed sample. Five hundred bootstrap samples were generated each with a sample size equivalent to that of the Healthy Directions data (n=1,209). Since we are unsure as the appropriate value of the smoothing parameter we used the significance trace method, where one computes P-values at several different values of the smoothing parameter. The question as to the most appropriate value of the smoothing parameter is irrelevant if all the P-values are greater than or less than the level of significance of the test. In these two cases the resulting conclusion drawn on the hypothesis is independent of the smoothing parameter.

# of Matched Sets	Power of LRT	Smoothing Parameter	Power
25	0.34	0.2	0.70
		0.8	0.73
		2.0	0.76
50	0.61	0.2	0.89
		0.8	0.95
		2.0	0.95
100	0.88	0.2	0.99
		0.8	0.99
		2.0	0.99

Table 3: Power of Test Statistic against Log Alternative.

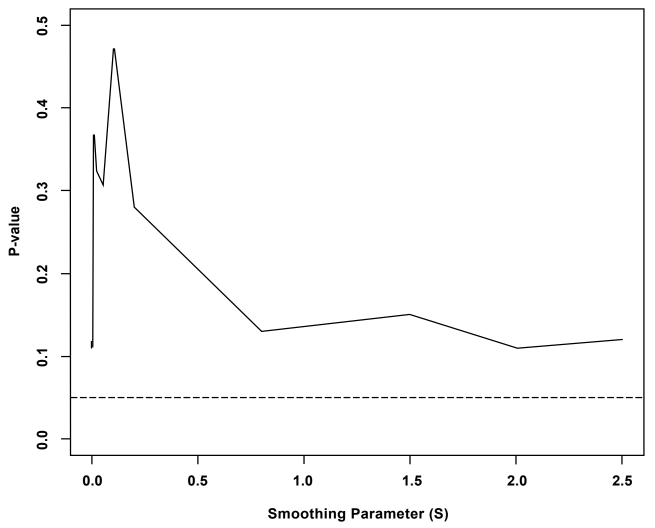

Figure 1: Significance Trace for the Healthy Directions Data.

Figure 1 shows the significance trace from the Healthy Directions data. The test statistic, R (equation 4), was computed at 15 different values of the smoothing parameter S, for each bootstrap sample. With a significance level of 0.05, we can see that we would fail to reject the null hypothesis regardless of the value of the smoothing parameter. Therefore, we can conclude that the relationship between x and Y in our nonparametric smooth function is not significantly different than that in our conditional logistic regression model. There is no evidence that our assumption of linear relationship between fruit and vegetable intake and the logit of intervention status has been violated.

Discussion

The use of a nonparametric smoother to test the linearity assumption was adopted from the methods proposed by Hart [8] and expanded to fit our conditional logistic regression model by extending the current methodology from one dimension to higher dimensions. The use of nonparametric smoothing methodology has several advantages. One of the most attractive advantages is that the test comes with a smoother [8]. Other methods provide a yes or no decision but do not provide any insight about the underlying function. The bias free nature of smoothed residuals is another advantage of our methodology. If one were to plot the left hand side of equation 2, $\hat{r}(x;S)-r(x;\hat{\theta})$ versus x, a systematic pattern would not be unusual even if the null hypothesis were true, due to the bias in the smoother $\hat{r}(\cdot;S)$. However, a pattern in the graph of the smoothed residuals is not expected unless the regression function actually differs from the null model [8]. From equation 2 we can see that when the bias is negligible, the residual smoother, $\hat{g}(x;S)$ is equal to a smooth of the data, $\hat{r}(x;S)$ minus the truth under the null hypothesis, $r(x;\hat{\theta})$. The unbiased nature of the residual smoother makes it a good tool to determine whether or not the null hypothesis should be rejected. Although often bias, the smooth of the data is a good tool to provide insight into the data when the null hypothesis is rejected. It can be used to determine how the model is deviant from the null hypothesis.

We bootstrapped the residuals in the simulations because this approach approximates the distribution of our test statistic relatively well [8]. This also allowed us to save computation time, cutting the computation time by almost half. However, the fact that our proposed methodology is still computationally intensive is a major limitation. For a data set containing 50 matched sets it takes approximately two minutes to produce a p-value based on 500 bootstrap samples. Despite the computation time, we have developed an R program that is simple to use. The user needs only to input the value of the smoothing parameter, the number of bootstrap samples desired, and the data in the specified format. The program returns the value of the test statistic for the original sample, the value of the test statistic for each of the bootstrap samples, and a p-value based on this information. It also produces a plot of the density of the bootstrapped test statistics indicating the value of the test statistic from the observed data and the p-value.

There still remains a question about the appropriate choice of smoothing parameter. To obtain a test with a prescribed level of significance, the smoothing parameter should be fixed before the data are examined, by bootstrapping one can ensure approximate validity of any test based on a single smoothing parameter [8]. Therefore, one should use the smoothing parameter that corresponds to the highest level of power. However, if one does not have much knowledge about the true distribution of the data this information is unavailable. Using the significance trace method provides a partial solution to this problem but for cases where the significance trace is not definitive, the question of choosing the appropriate smoothing parameter remains unanswered, and is an area for future research.

Although most would agree that checking model adequacy is an important part of any statistical analysis, this step is often left out when appropriate methods and easily implemented software are not available to do so. Our proposed methodology and user friendly R routine provide data analyst the ability to perform nonparametric diagnostic test for the conditional logistic regression model.

Acknowledgements

The research of Dr. Goodman was supported by National Institute of Child Health and Human Development grant 5 F31 HD043695 and the Barnes-Jewish Hospital Foundation. The research of Dr. Li was supported by National Institute of Health grant R01CA95747.

References

1. Armitage P, Colton T (1998) Encyclopeida of Biostatistics. John Wiley & Sons.
2. Hosmer DW, Lemeshow S (2000) Applied Logistic Regression. (Second edn), New York: John Wiley & sons, Inc.
3. Kupper L (1998) Encyclopeida of Biostatistics. In: Armitage P, Colton T, eds. John Wiley & Sons. 2434-2437.
4. Moolgavkar SH, Lustbader ED, Venzon DJ (1985) Assessing the Adequacy of the Logistic Regression Model for Matched Case-control Studies. Stat Med 4: 425-435.
5. Pregibon D (1984) Data Analytic Methods for Matched Case-control Studies. Biometrics 40: 639-651.
6. Moolgavkar SH, Lustbader ED, Venzon DJ (1984) A Geometric Approach to Nonlinear Regression Diagnosis with Application to Matched Case-Control Studies. Ann. Statist 12: 816-826.
7. Bedrick EJ, Hill JR (1996) Assesing the Fit of the Logistic Regression Model to Individual Matched Sets of Case-Control Data. Biometrics 52: 1-9.
8. Hart JD (1997) Nonparemetric Smoothing and Lack-of-Fit Tests. New York: Springer.
9. Hunt MK, Stoddard AM, Barbeau E, Goldman R, Wallace L, et al. (2003) Cancer prevention for working class, multiethnic populations through small businesses: The Healthy Directions Study. Cancer Causes Control 14: 749-760.
10. Emmons KM, Stoddard AM, Gutheil C, Suarez EG, Lobb R, et al. (2003) Cancer prevention for working class, multi-ethnic populations through health centers: The healthy directions study. Cancer Causes Control 14: 727-737.
11. SASI (2000) SAS OnlineDoc®. (8th edn), Cary, NC, USA: SAS Institute Inc.
12. Pregibon D (1981) Logistic Regression Diagnostics. Ann. Statist 9: 705-724.
13. Arbogast PG, Lin DY (2004) Goodness-of-fit methods for matched case-control studies. Canadian Journal of Statistics 32: 373-386.
14. Jong P (1987) A Central Limit Theorem for Generalized Quadratic Froms. Probability Thoery and Related Fields 75: 261-277.
15. Hall P, Hart JD (1990) Bootstrap Test for Difference Between Means in Nonparametric Regression. J Am Stat Assoc 85: 1039-1049.
16. King E, Hart JD, Wehrly TE (1997) Testing the equality of two regression curves using linear smoothers. Statistics & Probability Letters 12: 239-247.
17. Azzalini A, Bowman A (1993) On the Use of Nonparametric Regression for Checking Linear Relationships. J R Statist Soc B 55: 549-557.
18. Goodman M, Li Y, Bennett G, Stoddard AM, Emmons KM, et al. (2006) An Evaluation of Multiple Behavioral Risk Factors for Cancer in a Working Class, Multi-Ethnic Population. Journal of Data Science 4: 291-306.