Journal of Modern Applied Statistical Methods

Volume 20 | Issue 2 Article 3

Goodness-of-Fit Tests in Logistic Regression with Continuous Covariates

Justin Shang

Timothy J. Robinson

Shaun S. Wulff

Recommended Citation

Justin Shang, Timothy J. Robinson, Shaun S. Wulff (2021). Goodness-of-Fit Tests in Logistic Regression with Continuous Covariates. Journal of Modern Applied Statistical Methods, 20(2), https://doi.org/10.56801/Jmasm.V20.i2.3

Journal of Modern Applied Statistical Methods
2021, Vol. 20, No. 2,
Doi: 10.56801/Jmasm.V20.i2.3

ISSN 1538 – 9472

Goodness-of-Fit Tests in Logistic Regression with Continuous Covariates

Justin Shang **Timothy J. Robinson** **Shaun S. Wulff**

Goodness-of-fit tests are used to assess model adequacy. There are a number of ways in which a fitted model can be inadequate (Xie et al. (2008)). For instance, the linear systematic component of the model may be incorrectly specified, a covariate may not be specified in the appropriate functional form, some important covariates may have been omitted from the model, or the link function may be misspecified. All these model misspecifications could affect consistency of the coefficient estimation and can lead to biased estimates of treatment effects (Gail et al. (1988); Hauck et al. (1991)).

Keywords: Goodness-of-fit, Logistic Regression, Covariates.

1. Introduction

Goodness-of-fit tests are used to assess model adequacy. There are a number of ways in which a fitted model can be inadequate (Xie et al. (2008)). For instance, the linear systematic component of the model may be incorrectly specified, a covariate may not be given in the appropriate functional form, some important covariates may have been omitted from the model, or the link function may be misspecified. All these model misspecifications could affect consistency of the coefficient estimation, and can lead to biased estimates of treatment effects (Gail et al. (1988); Hauck et al. (1991)).

Many approaches have been developed for goodness-of-fit tests in logistic regression. These approaches can be generally summarized into the following categories: (1) Chi- square based tests. The chi-square statistic was introduced by Pearson (1900) and the theory and applications have been subsequently expanded by Fisher, Yates, and others (see Agresti (1990)). Later, several modified chi-square tests have been proposed, such as those by Hosmer and Lemeshow (1980), Tsiatis (1980), Pulkstenis and Robinson (2002), and Xie et al. (2008). Chi-square based goodness-of-fit tests are the primary focus of this study and more details are provided in later sections. (2) Information matrix approaches. White (1982) proposes a test to detect model misspecification based on the information matrix. Newey (1985) proposes a calculation procedure to employ the "outer product of the gradient" (OPG) covariance matrix estimator of the information matrix test statistic. Orme (1988) proposes a simple calculation procedure for the information matrix test statistic for general models of binary data by employing the maximum likelihood covariance

matrix estimator rather than the OPG estimator. (3) Residual methods. Copas (1989) conducts a study on the unweighted residual sum of squares, and propose a test based on scaled chi-square statistics. Hosmer et al. (1997) study the unweighted sum-of-squares (le Cessie- van Houwelingen-Copas-Hosmer) goodness-of-fit test and find this test to be superior to the other tests they examined. Copas (1983) introduces a nonparametric kernel method to examine the smoothed residuals. This approach has been extended by Azzalini et al. (1989), who propose a pseudo-likelihood ratio statistic based on nonparametric kernel method on the residuals. Later, le Cessie and van Houwelingen (1991) further refine the approach of Azzalini et al. (1989) using an unbiased estimator for the test statistic. (4) R^2. Mittlbock and Schemper (1996) study the properties of 12 different R^2 measures and recommend a R^2 coefficient based on the log-likelihood. (5) Other methods. Stukel (1988) incorporates two shape parameters to extend the formulation of the logistic model and improve the model fit. Hosmer et al. (1997) suggest that the Stukel test has higher power than other tests they examined for misspecified link functions and comparable power to other tests. Osius and Rojek (1992) derive asymptotic moments for a general class of goodness-of-fit statistics for multinomial models based on the weighted deviations of observed and expected frequencies and then conduct a standardized test statistic using the normal distribution. Qin and Zhang (1997) propose a Kolmogorov-Smirnov statistic to test the validity of the logistic link function.

The chi-square based goodness-of-fit tests are most commonly used in logistic regression. Chi-square based goodness-of-fit tests depend on the number of covariate patterns of the predictors compared to the number of positive responses. The covariate pattern represents a single set of values for the predictors in a model (Hosmer and Lemeshow (2000)). When all predictors are categorical, the Pearson chi-square test and the deviance test can be applied. The Hosmer and Lemeshow (1980) test can be used when there are continuous predictors. When there are continuous and categorical predictors, the Pulkstenis and Robinson (2002) tests and the Xie et al. (2008) tests are applicable.

However, it has been found that these chi-square based tests have similar deficiencies (Hosmer et al. (1997), Xie et al. (2008)). For instance, these tests are considered conservative as they can have low power to detect specific types of lack-of-fit, these methods are highly dependent on how the observations are grouped, there is uncertainty in the degrees of freedom for the tests, and the nature of the lack-of-fit can be difficult to identify. Hence, the purpose of this research is to explore these chi-square based goodness-of-fit tests of Hosmer and Lemeshow (1980), Pulkstenis and Robinson (2002), and Xie et al. (2008). Specifically, the aims of this study are to (1) study the reasons that lead to low power in these tests, (2) propose new modifications of these testing procedures to improve goodness-of-fit assessment, (3) assess the size and power of the proposed test statistics through simulations, and (4) apply the proposed tests on a clinical trial dataset for illustration purposes.

2. Goodness-of-fit tests

Consider a $N \times 1$ vector of binary responses $\mathbf{Y} = (\mathrm{Y}_1, \mathrm{Y}_2, \ldots, \mathrm{Y}_N)'$, where Y_i is coded as 1 or 0 for $i = 1, 2, \ldots, N$. For convenience, the coding of 1 is referred to as a positive outcome, and 0 is referred to as a negative outcome. Assume the responses $\mathbf{Y}$ are independent from the *Bernoulli* distribution with *true probabilities* of positive outcome $\dot{\boldsymbol{\pi}} = (\dot{\pi}_1, \dot{\pi}_2, \ldots, \dot{\pi}_N)'$. The true probabilities ($\dot{\boldsymbol{\pi}}$) are unknown in practice, but can be assessed using a logistic regression model. The observed $p \times 1$ vector of regressors for observation i is denoted $\boldsymbol{x}_i = (1, x_{1i}, x_{2i}, \ldots, x_{ki})'$ with $p = k + 1$ and a corresponding $p \times 1$ vector of regression coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)'$. The *model-based* probability of a positive outcome (π_i) is allowed to depend upon the vector of regressors ($\boldsymbol{x}_i$) and regression coefficients ($\boldsymbol{\beta}$) with a link function. Using a logit link function, the logistic regression model is defined as

$$\text{logit}(\pi(\boldsymbol{x}_i, \boldsymbol{\beta})) = \log\left\{\frac{\pi(\boldsymbol{x}_i, \boldsymbol{\beta})}{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\beta})}\right\} = \boldsymbol{x}_i'\boldsymbol{\beta} \ . \tag{1}$$

The estimated probability of a positive outcome can be obtained using the maximum likelihood estimator (MLE). The MLE ($\widehat{\mathbf{B}}$) is the value of $\boldsymbol{\beta}$ which maximizes the likelihood function

$$\mathrm{L} = \mathrm{L}(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{n} \pi_i(\boldsymbol{x}_i, \boldsymbol{\beta})^{y_i} \left(1 - \pi_i(\boldsymbol{x}_i, \boldsymbol{\beta})\right)^{1-y_i} \ . \tag{2}$$

The resulting estimator of the probability of a positive outcome is $\widehat{\Pi}_i = \pi(\boldsymbol{x}_i, \widehat{\mathbf{B}})$. The estimates $\widehat{\boldsymbol{\beta}}$ and $\hat{\pi}_i = \pi(\boldsymbol{x}_i, \widehat{\boldsymbol{\beta}})$ can be obtained from the observed data $\boldsymbol{y}$.

Goodness-of-fit (GOF) of a model refers to how well an assumed model approximates the true data generating process. Thus, a good fitting model should produce predicted values that are close to the observed values from the process (Hosmer et al. (2013), p. 154). In logistic regression models, goodness-of-fit is generally thought of within the context of comparing an observed binomial count for a particular grouping to the model expected count for the same grouping. Suppose there are g groups denoted as rows in the table. The rows are indexed by g with n_g independent Bernoulli trials per group. Let $\mathrm{O}_{g,1} = \sum_{i=1}^{n_g} \mathrm{Y}_i$ denote the observed number of presences in group g and let $\mathrm{O}_{g,0} = n_g - \mathrm{O}_{g,1}$ denote the observed number of absences in group g. The responses Y_i are assumed to be generated from a true unknown model with corresponding probability $\dot{\pi}_i$. Thus, $\dot{e}_{g,1} = \mathrm{E}[\mathrm{O}_{g,1}] = \sum_{i=1}^{n_g} \dot{\pi}_i$ and $\dot{e}_{g,0} = \mathrm{E}[\mathrm{O}_{g,0}] = n_g - \dot{e}_{g,1}$ are the true expected counts for positives and negatives, respectively, based upon that true model. When a model is assumed as in (1), the associated unknown model-based probabilities are π_i. The model-based expected counts for positive and negative outcomes are $e_{g,1} = \sum_{i=1}^{n_g} \pi_i$ and $e_{g,0} = n_g - e_{g,1}$, respectively. The MLE of the model-based expected counts are $\widehat{\mathrm{E}}_{g,1} = \sum_{i=1}^{n_g} \widehat{\Pi}_i$ and $\widehat{\mathrm{E}}_{g,0} = n_g - \widehat{\mathrm{E}}_{g,1}$.

Formally, a test of goodness-of-fit consists of the hypotheses (Agresti, 1990, pp. 42-43)

H_0: $e_{g,1} = \dot{e}_{g,1}$ and $e_{g,0} = \dot{e}_{g,0}$ for all g ,

H_1: $e_{g,1} \neq \dot{e}_{g,1}$ or $e_{g,0} \neq \dot{e}_{g,0}$ for some g . (3)

The null hypothesis in (3) indicates that the model-based expected counts are the same as the expected counts from the true unknown model. A model "fits" when H_0 holds and does not "fit" when H_1 holds. The quantities in (3) must be estimated to carry out the goodness-of-fit test. The observed counts $\mathrm{O}_{g,1}$ and $\mathrm{O}_{g,0}$ are used in place of $\dot{e}_{g,1}$ and $\dot{e}_{g,0}$, respectively. The estimated expected counts $\widehat{\mathrm{E}}_{g,1}$ and $\widehat{\mathrm{E}}_{g,0}$ are used in place of the model-based expected counts $e_{g,1}$ and $e_{g,0}$, respectively. The layout of these quantities is shown in Table 1. A test statistic is needed to measure the closeness within the combinations $(\mathrm{O}_{g,1}, \widehat{\mathrm{E}}_{g,1})$ and $(\mathrm{O}_{g,0}, \widehat{\mathrm{E}}_{g,0})$ for $g = 1, \ldots G$.

Many goodness of fit statistics are based upon the Pearson chi-square statistic. Differences among these statistics amounts to selection of the groups in Table 1. The Pearson Chi-Square statistic can be expressed as (Hosmer et al., 2013, p. 158)

$$\widehat{\mathrm{C}}_P = \sum_{g=1}^{G} \frac{\left(\mathrm{O}_{g,1} - \widehat{\mathrm{E}}_{g,1}\right)^2}{\widehat{\mathrm{E}}_{g,1}} + \sum_{g=1}^{G} \frac{\left(\mathrm{O}_{g,0} - \widehat{\mathrm{E}}_{g,0}\right)^2}{\widehat{\mathrm{E}}_{g,0}} =$$

$$\sum_{g=1}^{G} \frac{\left(\mathrm{O}_{g,1} - n_g\bar{\widehat{\Pi}}_g\right)^2}{n_g\bar{\widehat{\Pi}}_g} + \sum_{g=1}^{G} \frac{\left(n_g - \mathrm{O}_{g,1} - n_g\left(1-\bar{\widehat{\Pi}}_g\right)\right)^2}{n_g\left(1-\bar{\widehat{\Pi}}_g\right)} \tag{4}$$

$$= \sum_{g=1}^{G} \frac{\left(\mathrm{O}_{g,1} - n_g\bar{\widehat{\Pi}}_g\right)^2}{n_g\bar{\widehat{\Pi}}_g\left(1-\bar{\widehat{\Pi}}_g\right)}\left(1 - \bar{\widehat{\Pi}}_g + \bar{\widehat{\Pi}}_g\right),$$

noting $\widehat{\mathrm{E}}_{g,1} = \sum_{i=1}^{n_g} \widehat{\Pi}_i = n_g\bar{\widehat{\Pi}}_g$ and $\widehat{\mathrm{E}}_{g,0} = n_g - \sum_{i=1}^{n_g} \bar{\widehat{\Pi}}_g = n_g\left(1 - \bar{\widehat{\Pi}}_g\right)$. For example, $\widehat{\mathrm{C}}_P$ could be used when each group denotes a unique covariate pattern among c categorical regressors. It is recommended that observed values of $\widehat{\mathrm{E}}_{g,1}$ exceed 5 for all g in order to assume that the null distribution of $\widehat{\mathrm{C}}_P$ is χ^2_{G-c-1}. Thus, this approach would not be appropriate when the model contains numerous categorical covariates or continuous covariates since these settings may result in numerous groups with observed values of $\widehat{\mathrm{E}}_{g,1}$ less than 5.

Table 1. Observed and expected counts by groups.

	Positive outcome		Negative outcome		
Grou	Observe	Expecte	Observe	Expecte	Tot
1	$\mathrm{O}_{1,1}$	$\widehat{\mathrm{E}}_{1,1}$	$\mathrm{O}_{1,0}$	$\widehat{\mathrm{E}}_{1,0}$	n_1
⋮	⋮	⋮	⋮	⋮	⋮
g	$\mathrm{O}_{g,1}$	$\widehat{\mathrm{E}}_{g,1}$	$\mathrm{O}_{g,0}$	$\widehat{\mathrm{E}}_{g,0}$	n_g
⋮	⋮	⋮	⋮	⋮	⋮
G	$\mathrm{O}_{G,1}$	$\widehat{\mathrm{E}}_{G,1}$	$\mathrm{O}_{G,0}$	$\widehat{\mathrm{E}}_{G,0}$	n_G

In order to perform goodness-of-fit testing for a model containing continuous covariates, Hosmer and Lemeshow (1980) (denoted here as HL) propose a test which first orders all responses according to their estimated probabilities, and then classifies them into ten deciles. This approach essentially provides a substitute set of covariate patterns with sufficient cell sizes (Pulkstenis and Robinson, 2002). In this case, $G = 10$, so that there are 10 rows in the classification table. The first group contains roughly 10% of the smallest estimated probabilities while the last group contains roughly 10% of the largest estimated probabilities. The observed and expected cell counts are obtained for each of these groups, and the statistic in (4) is calculated (Pulkstenis and Robinson, 2002, equation (3)). Denote the HL statistic as $\hat{\mathrm{C}}_{HL}$. The value of this statistic can differ according to how the deciles are calculated. The limiting null distribution of $\hat{\mathrm{C}}_{HL}$ is taken to be χ^2_{G-2} from which the p-value (HL.p) can be calculated (Hosmer and Lemeshow, 2013, Section 5.2.2). The HL test can be used with continuous and categorical predictors in the specified model. However, Hosmer and Lemeshow (2013, p. 161) warn that this test should be used when there is an adequate number of covariate patterns. Otherwise, there will be numerous ties in the rankings of the estimated probabilities which could affect selection of the deciles.

Pulkstenis and Robinson (2002) (here denoted as PR) propose a goodness-of-fit test statistic which accommodates both continuous covariates and M covariate patterns from c categorical regressors. The responses are sorted by the corresponding estimated probabilities within each of the M categories. Each category is then split into two sub-categories where one subcategory has estimated probabilities below the median, and the other subcategory has estimated probabilities above the median. The observed and expected cell counts can be computed for each of these $G = 2M$ groups, and the statistic in (4) is calculated (Pulkstenis and Robinson, 2002, equation (4)). Denote the PR test statistic as $\hat{\mathrm{C}}_{PR}$. Pulkstenis and Robinson (2002) suggest that the null distribution of $\hat{\mathrm{C}}_{PR}$ is χ^2_{2M-c-2} from which the p-value (PR.p) can be calculated. The value of $\hat{\mathrm{C}}_{PR}$ can vary depending upon how the median is calculated in the presence of ties.

Xie et al. (2008) (here denoted as XIE) propose a goodness-of-fit test in which groups are formed by partitioning the covariate space. They utilize cluster analysis to obtain regions of similarity with respect to Euclidean distance among the k regressors. The resulting clusters form the covariate patterns which determine the groups. Xie et al. (2008) propose using $G = 10$ clusters when $k < 5$ and $G = k + 5$ clusters when $k \geq 5$. The observed and expected cell counts can be computed for each of these G groups formed from the clusters and the statistic in (4) is calculated (Xie et al., 2008, p 2706). Denote the XIE test statistic as $\hat{\mathrm{C}}_{\mathrm{XIE}}$. Xie et al. (2008) suggest the null distribution of $\hat{\mathrm{C}}_{\mathrm{XIE}}$ is $\chi^2_{G-k/2-2}$ from which the p-value (XIE.p) can be calculated. These authors utilize the Ward method of clustering (Rencher and Christensen, 2012, pp. 520-521). Their test is appropriate for any combination of continuous and categorical regressors such that the number of covariate patterns

exceeds G. The clustering is based upon all regressors in the model, including interaction terms.

The HL, PR, XIE tests are similar in that they are all chi-square based tests. The major differences exist in the grouping algorithms. For instance, Hosmer and Lemeshow (1980) use deciles of the sorted estimated probabilities to form the groups, Pulkstenis and Robinson (2002) use the median of the sorted estimated probabilities within each covariate pattern (as determined by categorical regressors) to form the groups, and Xie et al. (2008) used the covariate space and then apply clustering to obtain the groups. However, these groupings may not be appropriate if the specified model contains lack-of-fit. For instance, if a first-order term, an interaction term, or a quadratic term is omitted from the model, then the estimated probabilities $(\hat{\boldsymbol{\pi}})$ could have large discrepancies from the true probabilities $(\dot{\boldsymbol{\pi}})$, which could then lead to poor groupings in these tests.

3. Proposed tests

As discussed in the previous section, if continuous predictors exist in a model, then the Pearson chi-square test cannot be applied directly. Some grouping algorithms are needed to compute the modified chi-square test statistics (HL, PR, and XIE). Misspecifications of the covariates in an assumed model could cause large discrepancies between the estimated probabilities and the true probabilities that would adversely affect the groupings in the modified chi-square test statistics. In particular, this problem may lead to low power in detecting some types of lack-of-fit.

These concerns suggest new approaches for improving these goodness-of-fit tests. Specifically, it might be possible to achieve better test performance by forming the groupings based upon estimated probabilities from an over-fit logistic regression model. For instance, this over-fit model could be obtained using all of the observed covariates, interaction terms, as well as nonlinear functions of the covariates. Rather than having to specify all of these terms, a more convenient approach would be to use a Generalized Additive Model (GAM) (Hastie and Tibshirani (1986)). The GAM formulation of the logistic regression model is given by

$$\text{logit}(\pi(\boldsymbol{x}_i)) = \log\left\{\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right\} = s_0 + \sum_{j=1}^{p} s_j(x_{i,j}) \qquad (5)$$

where s_j is the smooth function for variable j. The GAM is a generalization of the standard linear model, and it allows easier interpretations of the contributions of each variable. Hastie and Tibshirani (1986) suggest a mixture of the generalized linear models (Nelder and Wedderburn (1972)) and the GAM to be used in practice, which is the approach adopted in this study for the proposed tests. The resulting model is

$$\text{logit}(\pi(\boldsymbol{x}_i, \boldsymbol{\beta})) = \log\left\{\frac{\pi(\boldsymbol{x}_i, \boldsymbol{\beta})}{1-\pi(\boldsymbol{x}_i, \boldsymbol{\beta})}\right\} = s_0 + \sum_{j=1}^{p} \beta_j x_{i,j} + \sum_{j=p+1}^{q} s_j(x_{i,j}) \ . \qquad (6)$$

where $x_{i,j}$, $j = 1, \ldots, p$, are the categorical variables, and $x_{i,j}$, $j = p+1, \ldots, q$, are the continuous variables. The smooth functions (s_j) can be estimated by the

scatterplot smoother which is an iterative procedure called the local smoothing algorithm. The local smoothing algorithm utilizes scatter plot smoothers to generalize the usual Fisher scoring procedure for computing maximum likelihood estimates. Hastie and Tibshirani (1986) use the local average estimates based on symmetric nearest neighborhoods. Associated with a neighborhood is the span or window size, w, which is the proportion of total points contained in each neighborhood. Different scatterplot smoothers can be used such as a running mean, running median, running least squares line, kernel estimate, or spline (Reinsch (1967), Cleveland (1979)). The running mean is not a satisfactory smoother because it creates large biases at the end points (Hastie and Tibshirani (1986)). The kernel or spline smoother could be expected to work well, but requires an increased cost of computation. So, Hastie and Tibshirani (1986) suggest using the running lines smoother, which produces reasonable results and has the advantage that the estimate in a neighborhood can be found by updating the estimate of the previous neighborhood. In summary, GAM proves to be useful in uncovering nonlinear covariate effects. It has the advantage of being completely automatic. Hence, the GAM provides a convenient tool to develop an over-fit model that can be used for grouping in the newly proposed goodness-of-fit tests. The proposed approach for the goodness-of-fit tests is to develop a mixture GAM from (6) with all of the available covariates as well as two-way interaction terms to obtain an overfit model. The estimator of the mixture GAM probabilities are denoted as $\widehat{\mathbf{\Pi}}_{GAM}$, which is used to construct the groupings for the previously mentioned goodness-of-fit tests. The smooth functions for s consist of a maximum order of 3 polynomial components for each of the continuous variables. The *gam* function from the *gam* package in R is used fit the GAM. The associated modified tests are denoted as HL.GAM, PR.GAM, and XIE.GAM. The degrees of freedom for these proposed tests are obtained following the same rules as specified previously. Simulation studies and a clinical trial example will be performed in later sections to illustrate and evaluate the proposed testing procedures.

4. Simulation results

In this section, several simulation scenarios are considered with the proposed test procedures. These scenarios include assumed models containing lack-of-fit from omitting an interaction term, a quadratic term, or a first-order term. These simulation scenarios are used to evaluate the size and the power of the proposed tests. The simulations involve randomly generating N binary values from the Bernoulli distribution with true model probabilities $\dot{\pi}_i$ as

$$\mathrm{Y}_i^{(t)} \text{ i. i. d. } Bernoulli(\dot{\pi}_i)\ ,\ \ i = 1,2,\dots,N\ . \tag{7}$$

The simulation of N binary values is repeated 5000 times with $t = 1,2,\dots,5000$. For each t, the goodness of fit tests are conducted at level $\alpha = 0.05$. The number of

rejections out of 5000 are used to assess the size of the tests under the correctly specified model and the power of the tests under the incorrectly specified models.

4.1 Omitting an interaction term

Consider the following model modified from Xie et al. (2008). The simulated true probabilities $\dot{\pi}_i$ are randomly generated using

$$\log\left(\frac{\dot{\pi}_i(\boldsymbol{x}_i)}{1-\dot{\pi}_i(\boldsymbol{x}_i)}\right) = -1.7918 + x_{i,1} + x_{i,2} + 0.1352x_{i,3} + 1.7918x_{i,4} + 0.5973x_{i,3}x_{i,4}\ , \quad (8)$$

where $x_{i,1}$ i.i.d $Bernoulli(0.5)$, $x_{i,2}$ i.i.d $Bernoulli(0.15)$, and $x_{i,3}$ i.i.d $U(-3,3)$. The sample size was chosen to be $N = 500$ so that it would be possible to detect lack-of-fit when it exists. Three scenarios are developed to study the influence of covariates on the model fit, where $x_{i,4}$ is randomly sampled from $N(0,4)$, $U(-3,3)$, and $Beta(4,2) \times 6 - 3$, respectively. The latter two distribution are specified to have supports on $[-3,3]$. The specified normal distribution has about 86.6% of the values within $[-3,3]$. The support values for $x_{i,4}$ are important since it determines the magnitude of the interaction $(x_{i,3}x_{i,4})$ in (8) relative to the other covariates. Three models are considered:

(a) the full model that includes the interaction between $x_{i,3}$ and $x_{i,4}$,

$$\log\left(\frac{\pi(\boldsymbol{x}_i,\boldsymbol{\beta})}{1-\pi(\boldsymbol{x}_i,\boldsymbol{\beta})}\right) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \beta_5 x_{i,3}x_{i,4}, \quad (9)$$

with estimated probabilities $\underline{\hat{\pi}}_s$;

(b) the reduced model that omits the interaction between $x_{i,3}$ and $x_{i,4}$,

$$\log\left(\frac{\pi(\boldsymbol{x}_i,\boldsymbol{\beta})}{1-\pi(\boldsymbol{x}_i,\boldsymbol{\beta})}\right) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4}, \quad (10)$$

with estimated probabilities $\underline{\hat{\pi}}_r$;

(c) the mixture GAM to be used in the proposed tests,

$$\log\left(\frac{\pi(\boldsymbol{x}_i,\boldsymbol{\beta})}{1-\pi(\boldsymbol{x}_i,\boldsymbol{\beta})}\right) = s_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,1}x_{i,2} + s_1(x_{i,3}) + s_2(x_{i,4}) + s_3(x_{i,1}x_{i,3}) + s_4(x_{i,1}x_{i,4}) + s_5(x_{i,2}x_{i,3}) + s_6(x_{i,2}x_{i,4}) + s_7(x_{i,3}x_{i,4}), \quad (11)$$

with estimated probabilities $\underline{\hat{\pi}}_{GAM}$.

Test results from 5000 iterations are summarized in Table 2. The sizes of all the tests appear to be around the 0.05 level. However, the power of the tests vary. The Hosmer-Lemeshow test (HL) has relatively high power with about 75% rejection rate on the reduced model in scenarios 1 and 2, and a 43% rejection rate in scenario 3. The Pulkstenis and Robinson chi-square (PR) test has low power with rejection rate <10% in all three scenarios. The Xie chi-square test (XIE) has a rejection rate at 26% in scenario 1, 81% in scenario 2, and 39% in scenario 3. However, the modified tests using the over-fit GAM in (11) has increased the power for all tests and scenarios. The modified Hosmer and Lemeshow test (HL+GAM) has a rejection rate at around 90% in the three scenarios. The power from the modified

Xie test (XIE+GAM) has been increased with rejection rates at greater than 90% in scenarios 1 and 2, and 74% in scenario 3. The new grouping algorithm also improved the Pulkstenis and Robinson chi-square test as the rejection rates have increased to 79% in scenario 1, 57% in scenario 2, and 62% in scenario 3.

Table 2. Goodness-of-fit tests applied to the full model (9) and to the lack-of-fit model (10) that omits an interaction term.

	Full model				Reduced			
$x_{i,4}$ i. i. d $N(0,4)$	Mea	Var	d	Siz	Mea	Var	d	Powe
HL	7.72	48.40	8	6.60	29.6	411.7	8	75.18
PR	3.66	5.28	4	2.44	4.94	7.862	4	6.80
Xie	3.02	3.48	6.	0.14	11.4	23.83	7	25.80
HL + GAM	7.79	17.18	8	2.98	27.6	56.19	8	97.18
PR + GAM	4.87	5.89	4	4.44	14.9	41.03	4	79.28
Xie + GAM	4.69	7.38	6	1.06	26.7	170.11	6.	94.38
$x_{i,4}$ i. i. d $U(-3,3)$	Mea	Var	d	Siz	Mea	Var	d	Powe
HL	7.56	47.65	8	5.98	27.9	298.24	8	76.18
PR	3.72	5.44	4	2.48	4.30	6.501	4	4.44
Xie	4.48	7.09	6.	1.18	20.1	47.80	7	80.52
HL + GAM	7.58	17.44	8	2.94	25.0	43.25	8	94.40
PR + GAM	4.74	5.73	4	4.44	11.2	29.93	4	56.52
Xie + GAM	5.75	10.39	6	3.48	28.2	120.31	6.	95.90
$x_{i,4}$ i. i. d $Beta(4,2) \times 6$	Mea	Var	d	Siz	Mea	Var	d	Powe
HL	7.76	18.68	8	4.96	15.5	54.43	8	43.12
PR	3.83	6.36	4	3.62	4.18	7.138	4	4.50
Xie	5.13	26.11	6.	5.44	13.2	25.74	7	39.48
HL + GAM	8.85	14.15	8	5.32	23.1	43.22	8	88.60
PR + GAM	4.11	5.31	4	2.92	11.3	20.14	4	62.26
Xie + GAM	6.03	15.39	6	5.36	18.1	51.49	6.	74.06

Figure 1 shows the groupings formed by each goodness-of-fit test based upon the true probabilities ($\boldsymbol{\pi}$) in (8) versus those based upon the estimated probabilites ($\hat{\boldsymbol{\pi}}_s$) from the full model in (9), the estimated probabilites ($\hat{\boldsymbol{\pi}}_r$) from the reduced model in (10), and the estimated probabilities ($\hat{\boldsymbol{\pi}}_{GAM}$) from the over-fit GAM. Hence, a plot of these groupings between the true and estimated probabilities illustrates discrepancies in the grouping results. For example, consider the Hosmer and Lemeshow test in scenario 1 (Figure 1a, b, c). With $G = 10$ deciles, the counts $n_{i,j} = 10$ from cell (i, j) in the plot (displayed in different color) provides the average number of observations that have been grouped into decile i based on the true probabilities ($\boldsymbol{\pi}$) from (8), but are grouped into decile j based on the estimated probabilities ($\hat{\boldsymbol{\pi}}_s, \hat{\boldsymbol{\pi}}_r, \hat{\boldsymbol{\pi}}_{GAM}$) obtained from (9), (10) and (11), respectively, with $\sum_{i=1}^{10}\sum_{j=1}^{10} n_{i,j} = 500$. Only those cells with $n_{i,j} \geq 5$ observations are displayed in the plots.

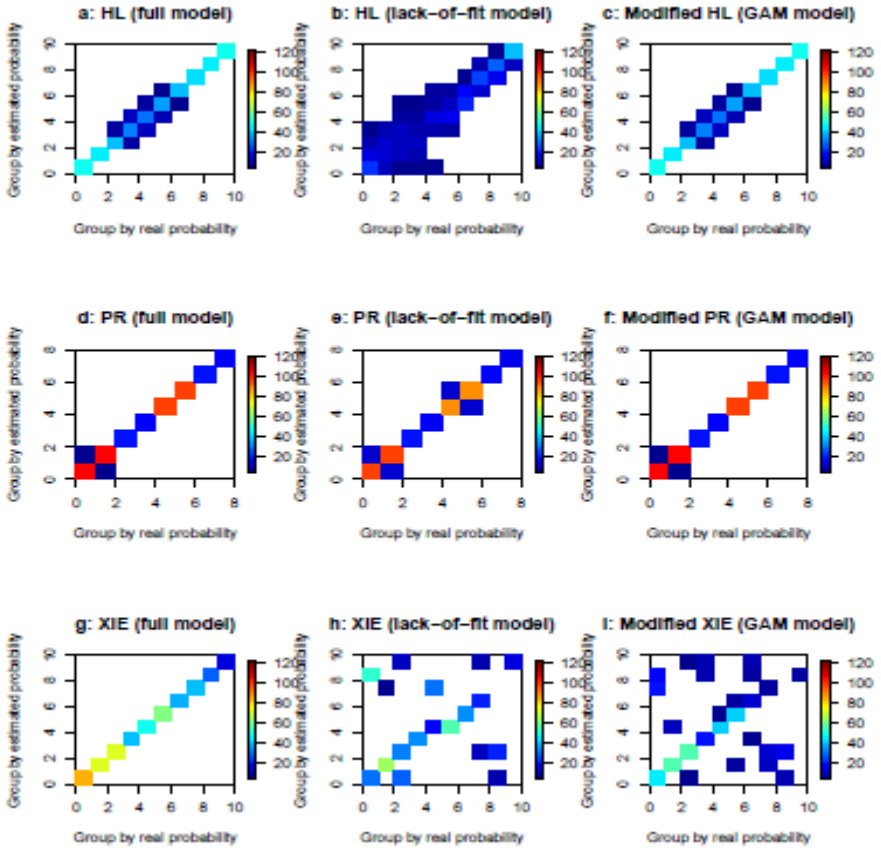

Figure 1: Group ID based on simulated true probabilities from (8) compared to (a) group ID based on the estimated probabilities from the full model (9) used for the original Hosmer and Lemeshow test; (b) group ID based on the estimated probabilities from the reduced model (10) used for the original Hosmer and Lemeshow test; (c) group ID based on the estimated probabilities from the GAM (11) used for the modified Hosmer and Lemeshow test; (d) group ID based on the estimated probabilities from the full model (9) used for the original Pulkstenis and Robinson tests; (e) group ID based on the estimated probabilities from the reduced model (10) used for the original Pulkstenis and Robinson tests; (f) group ID based on the estimated probabilities from the GAM (11) used for the modified Pulkstenis and Robinson tests; (g) group ID based on the estimated probabilities from the full model (9) used for the original Xie test; (h) group ID based on the estimated probabilities from the reduced model (10) used for the original Xie test; and (i) group ID based on the estimated probabilities from the GAM (11) used for the modified Xie test.

Specifically, the grouping results based on the estimated upon probabilities $(\hat{\boldsymbol{\pi}}_s)$ from the full model in (9) match well with grouping results based the simulations of the true probabilities $(\dot{\boldsymbol{\pi}})$. This holds for the Hosmer and Lemeshow test (Figure 1a),

the Pulkstenis and Robinson test (Figure 1d), and the Xie test (Figure 1g). Consider the tests using the groupings based upon the reduced model in (10). For the Hosmer and Lemeshow test (Figure 1b), the groupings based on the estimated probabilities from the reduced model ($\hat{\boldsymbol{\pi}}_r$) are different from the groupings based on simulations of the true probabilities ($\dot{\boldsymbol{\pi}}$). These grouping results are better in the Pulkstenis and Robinson test (see Figure 1e). In the Xie test, omitting an interaction term in the reduced model causes large discrepancies in the groupings (see Figure 1h). More importantly, after using the probabilities ($\hat{\boldsymbol{\pi}}_{GAM}$) from the GAM (11), the grouping results have been improved, especially for the modified Hosmer-Lemeshow test. As shown in Figure 1c, the grouping results based on the estimated probabilities from the GAM generally match well with grouping results based on simulated true probabilities ($\dot{\boldsymbol{\pi}}$). The grouping in the modified Pulkstenis and Robinson tests (Figure 1f) is better than the reduced model. However, the modified Xie test (Figure 1i) does not show much improvement in the grouping results after incorporating the estimated probabilities ($\hat{\boldsymbol{\pi}}_{GAM}$) from the GAM in the cluster analysis.

Table 3. Summaries of R^2 between true probabilities and estimated probabilities.

	Full model (9)		Reduced model		GAM model	
	Mean	SD	Mean	SD	Mean	SD
$x_{i,4}$ i. i. d $N(0,4)$	0.9940	0.00369	0.9305	0.00520	0.9798	0.00655
$x_{i,4}$ i. i. d $U(-3,3)$	0.9937	0.00386	0.9397	0.00644	0.9792	0.00674
$x_{i,4}$ i. i. d $Beta(4,2) \times 6$	0.9926	0.00472	0.9260	0.00510	0.9747	0.00803

The simulation also allows for comparison of the simulated true probabilities and the estimated probabilities. The squared correlation R^2 is computed for each simulation and the mean and standard deviations are given from the 5000 iterations. In general, the squared correlation between true probabilities and the estimated probabilities from the full model is high (> 0.99) in the three scenarios. The estimated probabilities from the reduced model show some differences from the simulated true probabilities with an average R^2 close to 0.93 in the three scenarios. The over-fit GAM seems to approximate the true probabilities well, with an average R^2 greather than 0.97 for all 3 scenarios.

4.2 Omitting a quadratic term

Consider the following model modified from Xie et al. (2008). The simulated true probabilities $\dot{\pi}_i$ are randomly generated using

$$\log\left(\frac{\dot{\pi}_i(\boldsymbol{x}_i)}{1-\dot{\pi}_i(\boldsymbol{x}_i)}\right) = -3.2324 + x_{i,1} + x_{i,2} + x_{i,3} + 0.5583x_{i,4} + 0.5002x_{i,4}^2, \quad (12)$$

where $x_{i,1}$ i.i.d $Bernoulli(0.5)$, $x_{i,2}$ i.i.d $Bernoulli(0.15)$, $x_{i,3}$ i.i.d $U(-3,3)$. The sample size was chosen to be $N = 500$. Again, three scenarios are developed where $x_{i,4}$ is randomly sampled from $N(0,4)$, $U(-3,3)$, and $Beta(4,2) \times 6 - 3$, respectively. Three models are considered:

(a) the full model that includes the interaction between $x_{i,3}$ and $x_{i,4}$,

$$\log\left(\frac{\pi(x_i,\boldsymbol{\beta})}{1-\pi(x_i,\boldsymbol{\beta})}\right) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \beta_5 x_{i,4}^2, \quad (13)$$

with estimated probabilities $\hat{\boldsymbol{\pi}}_s$;

(b) the reduced model that omits the quadratic term $x_{i,4}^2$,

$$\log\left(\frac{\pi(x_i,\boldsymbol{\beta})}{1-\pi(x_i,\boldsymbol{\beta})}\right) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4}, \quad (14)$$

with estimated probabilities $\hat{\boldsymbol{\pi}}_r$;

(c) the mixture GAM to be used in the proposed tests,

$$\log\left(\frac{\pi(x_i,\boldsymbol{\beta})}{1-\pi(x_i,\boldsymbol{\beta})}\right) = s_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,1}x_{i,2} + s_1(x_{i,3}) + s_2(x_{i,4})$$
$$+ s_3(x_{i,1}x_{i,3}) + s_4(x_{i,1}x_{i,4}) + s_5(x_{i,2}x_{i,3}) + s_6(x_{i,2}x_{i,4}) + s_7(x_{i,3}x_{i,4}), \quad (15)$$

with estimated probabilities $\hat{\boldsymbol{\pi}}_{GAM}$.

Notice that the quadratic term $(x_{i,4}^2)$ and the cubic term $(x_{i,4}^3)$ are included in this overfit GAM since the smoothing function has maximum order of 3 polynomial components.

The test results from 5000 iterations are given in Table 4. The size of all tests are close to the 0.05 level. The Hosmer-Lemeshow test has low power with rejection rates at 15% in scenario 1, 6% in scenario 2, and 22% in scenario 3. The Pulkstenis and Robinson chi-square test has low power with rejection rates less than 10% in all scenarios. The Xie chi-square test behaves differently with rejection rate of 100% in scenario 1, 68% in scenario 2, and 11% in scenario 3. However, the estimated power of test statistics has been dramatically increased to almost 100% in the proposed tests (HL+ GAM, PR + GAM, XIE + GAM) in scenarios 1 and 2. For scenario 3, the power from the proposed tests (HL + GAM, PR + GAM, XIE + GAM) are also higher than the original tests (HL, PR, XIE).Table 5 gives summaries for R^2 to measure the squared correlation between the simulated true probabilities ($\hat{\boldsymbol{\pi}}$) and the estimated probabilities for the full model ($\hat{\boldsymbol{\pi}}_s$), reduced model ($\hat{\boldsymbol{\pi}}_r$), and GAM ($\hat{\boldsymbol{\pi}}_{GAM}$). In general, the correlations between true probabilities versus the estimated probabilities from the full model are very high (>0.99) in the three scenarios. The GAM also seems to approximate the true probabilities very well with an average R^2 > 0.97 for all three scenarios. The estimated probabilities from the reduced model show differences from the simulated true probabilities with an average R^2 of 0.63 in scenario 1, 0.81 in scenario 2, and 0.97 in scenario 3, respectively. The power from the proposed tests does not provide much improvement for scenario 3 as the estimated probabilities do not improve the average squared correlation to the true probabilities in scenario 3 from the reduced model.

Table 4. Goodness-of-fit tests applied to the full model (13) and to the lack-of-fit model (14) that omits an interaction term.

	Full model				Reduced			
$x_{i,4}$ i. i. d $N(0,4)$	Mea	Var	d	Siz	Mea	Var	d	Pow
HL	8.01	22.13	8	6.50	10.4	23.13	8	14.58
PR	3.87	6.05	4	3.14	3.88	6.084	4	3.42
Xie	4.50	33.51	6.	3.06	50.7	80.61	7	100.0
HL + GAM	8.23	12.88	8	4.30	127.4	236.70	8	100.0
PR + GAM	5.28	9.21	4	9.74	93.3	268.08	4	100.0
Xie + GAM	5.87	71.58	6	4.10	119.5	617.49	6.	100.0
$x_{i,4}$ i. i. d $U(-3,3)$	Mea	Var	d	Siz	Mea	Var	d	Pow
HL	7.95	17.08	8	5.26	8.37	16.27	8	5.82
PR	3.94	6.27	4	3.28	4.04	5.905	4	2.98
Xie	4.04	5.26	6.	0.36	17.4	36.57	7	68.40
HL + GAM	8.53	12.04	8	3.96	51.0	132.69	8	100.0
PR + GAM	4.53	6.13	4	4.26	27.5	67.06	4	99.74
Xie + GAM	5.46	8.24	6	2.02	34.3	160.29	6.	98.12
$x_{i,4}$ i. i. d $Beta(4,2) \times 6$	Mea	Var	d	Siz	Mea	Var	d	Pow
HL	7.96	16.31	8	5.26	11.8	61.77	8	21.90
PR	3.88	6.12	4	3.34	4.27	7.380	4	5.10
Xie	4.19	6.20	6.	0.82	8.41	18.67	7	10.80
HL + GAM	8.40	12.13	8	4.02	13.3	21.84	8	28.28
PR + GAM	4.54	5.97	4	4.26	6.94	10.79	4	20.26
Xie + GAM	5.61	8.53	6	2.90	9.47	28.71	6.	18.58

4.3 Omitting a first-order term

For this scenario, consider the model in (8) which is modified by replacing the interaction term $(x_{i,3}x_{i,4})$ with a new varible $(x_{i,5})$. The simulated true probabilities $(\dot{\pi}_i)$ are randomly generated using

$$\log\left(\frac{\dot{\pi}_i(\boldsymbol{x}_i)}{1-\dot{\pi}_i(\boldsymbol{x}_i)}\right) = -1.7918 + x_{i,1} + x_{i,2} + 0.1352x_{i,3} + 1.7918x_{i,4} + 0.5973x_{i,5}, \quad (16)$$

where $x_{i,1}$ i.i.d $Bernoulli(0.5)$, $x_{i,2}$ i.i.d $Bernoulli(0.15)$, $x_{i,3}$ i.i.d $U(-3,3)$, and $x_{i,4}$ i.i.d $N(0,4)$ with $N = 500$ and $T = 5000$. Three scenarios are developed where $x_{i,5}$ is randomly sampled from $N(0,4)$, $U(-3,3)$, and $Beta(4,2) \times 6 - 3$, respectively. Three models are considered:

(a) the full model that includes the first order term '$x_{i,5}$',

$$\log\left(\frac{\pi(\boldsymbol{x}_i,\boldsymbol{\beta})}{1-\pi(\boldsymbol{x}_i,\boldsymbol{\beta})}\right) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \beta_5 x_{i,5}, \quad (17)$$

with estimated probabilities $\underline{\hat{\pi}}_s$;

(b) the reduced model that omits the first-order term $x_{i,5}$,

$$\log\left(\frac{\pi(\boldsymbol{x}_i,\boldsymbol{\beta})}{1-\pi(\boldsymbol{x}_i,\boldsymbol{\beta})}\right) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4}, \quad (18)$$

with estimated probabilities $\underline{\widehat{\pi}}_r$;

(c) the mixture GAM to be used in the proposed tests,

$$\log\left(\frac{\pi(x_i,\boldsymbol{\beta})}{1-\pi(x_i,\boldsymbol{\beta})}\right) = s_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,1}x_{i,2} + s_1(x_{i,3}) + s_2(x_{i,4}) + s_3(x_{i,5}) + s_4(x_{i,1}x_{i,3}) + s_5(x_{i,1}x_{i,4}) + s_6(x_{i,2}x_{i,3}) + s_7(x_{i,2}x_{i,4}) + s_8(x_{i,3}x_{i,4}) \quad (19)$$

with estimated probabilities $\underline{\widehat{\pi}}_{GAM}$.

Table 5: Summaries of R^2 between true probabilities and estimated probabilities.

	Full model (13)		Reduced model		GAM model	
	Mean	SD	Mean	SD	Mean	SD
$x_{i,4}$ i. i. d $N(0,4)$	0.9935	0.00406	0.6333	0.00468	0.9759	0.00708
$x_{i,4}$ i. i. d $U(-3,3)$	0.9926	0.00474	0.8144	0.00429	0.9730	0.00840
$x_{i,4}$ i. i. d $Beta(4,2) \times 6$	0.9924	0.00477	0.9660	0.00582	0.9745	0.00799

The GAM in (19) does not contain interactions involving $x_{i,5}$. This was necessary to inusre convergence of the GAM fit which could be problematic in the presence of numerous correlated predictors.

Goodness-of-fit tests are known to have low power in detecting a missing first-order term (Xie et al., 2008). The proposed modified tests are also likely to have low power to detect a missing first-order term if that missing term is not incorporated into the GAM. However, if the necessary data is available, the proposed approach provides a way to incorporate this missing model information into the grouping mechanism in order to better detect lack-of-fit. Thus, the proposed GAM in (19) incorporates the missing predictor $(x_{i,5})$ assuming such data is available. The simulation results are shown in Table 6.

Table 6: Goodness-of-fit tests applied to the full model (17) and to the lack-of-fit model (18) that omits a first-order term.

	Full model				Reduced			
$x_{i,4}$ i. i. d $N(0,4)$	Mea	Var	d	Siz	Mea	Var	d	Pow
HL	7.92	60.61	8	6.76	6.96	15.82	8	
PR	3.62	5.34	4	2.60	3.66	5.490	4	2.36
Xie	4.06	6.890	6.	1.26	4.25	27.11	7	
HL + GAM	7.45	10.23	8	1.98	26.71	47.108	8	96.48
PR + GAM	4.57	5.18	4	3.40	15.1	30.409	4	85.24
Xie + GAM	5.72	9.666	6	2.70	16.47	35.933	6.	66.74
$x_{i,4}$ i. i. d $U(-3,3)$	Mea	Var	d	Siz	Mea	Var	d	Pow
HL	7.84	50.20	8	6.12	7.34	17.38	8	4.20
PR	3.72	5.93	4	2.74	3.45	4.967	4	1.90
Xie	3.45	12.4	6.	1.60	4.16	22.75	7	2.62

HL + GAM	7.69	10.79	8	2.44	21.41	38.912	8	83.28
PR + GAM	4.33	5.00	4	2.90	12.23	25.36	4	67.78
Xie + GAM	5.85	7.89	6	2.38	12.65	25.086	6.	38.94
$x_{i,4}$ i. i. d $Beta(4,2) \times 6$	Mea	Var	d	Siz	Mea	Var	d	Pow
HL	7.83	31.02	8	7.26	7.62	19.83	8	
PR	3.67	6.66	4	3.40	3.55	4.902	4	2.18
Xie	3.28	3.34	6.	0.06	4.52	24.37	7	
HL + GAM	7.46	9.056	8	1.88	11.71	16.27	8	16.26
PR + GAM	4.02	5.11	4	2.48	6.47	10.14	4	16.82
Xie + GAM	5.79	9.45	6	2.98	7.69	12.35	6.	

The tests results from a simulation with 5000 iterations are shown in Table 6. The rejection rate of these tests on the full model are around 5%, which suggests the size of these test remains at level α. HL, PR and XIE tests all have low power to detect lack-of-fit associated with the omission of a first-order term. However, if the first-order term $x_{i,5}$ is included in the GAM (19), the power can be increased to 96% and 83% in the modified Hosmer and Lemeshow test, 85% and 68% in the modified Pulkstenis and Robinson chi-square test, and 67% and 39% in the modified Xie chi-square test. The proposed tests did not have very high power for the third scenario, though there is some improvement over the original tests. The proposed tests also did not provide increased power when $x_{i,5}$ was not included in the GAM (19) as might be expected.

The results are confirmed by the R^2 between the simulated true probabilities ($\dot{\boldsymbol{\pi}}$) and the estimated probabilities. For instance, the squared correlations between $\dot{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\pi}}_{GAM}$ in (19) are relatively close to those between $\dot{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\pi}}_{s}$. On the other hand, the squared correlations between $\dot{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\pi}}_{GAM}$ without $x_{i,5}$ are closer to those between $\dot{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\pi}}_{r}$. Thus, the incorporation of the omitted first-order term in the GAM would be expected to produce groupings consistent with those of the full model.

Table 7: Summaries of R^2 between true probabilities and estimated probabilities.

	Full model (17)		Reduced model (18)		GAM (19)	
					GAM w/o $x_{i,5}$	
	Mean	SD	Mean	SD	Mean	SD
$x_{i,5}$ i. i. d $N(0,4)$	0.9947	0.00321	0.9303	0.00265	0.9813	0.00576
					0.9244	0.00496
$x_{i,5}$ i. i. d $U(-3,3)$	0.9946	0.00331	0.9451	0.00295	0.9810	0.00578
					0.9380	0.00509
$x_{i,5}$ i. i. d $Beta(4,2) \times 6$	0.9944	0.00347	0.9760	0.00291	0.9809	0.00611
					0.9656	0.00545

4.4 Clinical trial example

The proposed tests are applied to a clinical trial example described by Barat et al. (2011). The data was collected at two urban clinics and two suburban clinics by Johns Hopkins University in an effort to identify characteristics of young female patients who successfully complete the three-injection sequence of the Gardasil quadrivalent human papillomavirus vaccine. The data consists of measurements taken from 1413 cases of young female patients aged 11-26 years. Original predictors include *age* (continuous variable), *race* (white, black, Hispanic, unknown, or other), *insurance* (0 represents that the patient received 'Medical Assistance', 1 represents 'Private Payer', 2 represents 'Hospital Based', and 3 represents 'Military'), *location* (0 represents suburban and 1 represents urban), and *practice* (0 represents 'Pediatrics', 1 represents 'Family Practice', and 2 represents 'OBGYN'). All subjects are classified into 10 age cohorts, where the first age cohort contains all subjects from 11 to 13 years old, and the last age cohort contains all subjects from 25 to 26 years old. There are indications of a non-linear relationship involving age based upon the empirical logits computed by each age cohort. Eight models have been posited which are listed in Table 8. An overfit mixture GAM is developed to include all first-order terms and polynomial terms as

$$\log\left(\frac{\pi(\boldsymbol{x}_i,\boldsymbol{\beta})}{1-\pi(\boldsymbol{x}_i,\boldsymbol{\beta})}\right) = s_0 + \beta_1 insurance_i + \beta_2 race_i + \beta_3 location_i + \beta_4 practice_i + \beta_5 insurance_i \times race_i + \beta_6 race_i \times location_i + \beta_7 race_i \times practice_i + \beta_8 insurance_i \times location_i + \beta_9 insurance_i \times practice_i + \beta_{10} location_i \times practice_i + s_1(age_i) + s_2(age_i \times insurance_i) + s_3(age_i \times race_i) + s_4(age_i \times location_i) + s_5(age_i \times practice_i)\,. \quad (20)$$

Table 8. Eight models for the clinical trial data.

Mo	Main terms	Interactions
1	age, age^2, age^3, insurance, race, location, practice	age × race, age × insurance, age × location, age × practice, race × insurance, race × location, race × practice,
2	age, age^2, age^3, insurance, race, location, practice	age × practice, race × practice, insurance × location, location×practice
3	age, age^2, age^3, insurance, race, location, practice	age × practice, insurance × location, location×practice
4	age, age^2, age^3, insurance, race, location, practice	age × practice, insurance × location
5	age, age^2, age^3, insurance, race, location, practice	age × practice
6	age, age^2, age^3, insurance, race, location, practice	
7	age, age^2, insurance, race, location, practice	

8	age, insurance, race, location, practice	

Table 9 provides the AIC as well as the p-values from the goodness-of-fit tests. The p-values from these test results show no evidence against the claim that that Models 1, 2, and 3 provide an adequate fit. When additional interaction terms are eliminated as in Models 4, 5, and 6, or cubic and quadratic term of *age* are eliminated as in Models 7 and 8, the Hosmer-Lemeshow test shows no evidence against the claim that all models provides an adequate fit. However, the Pulkstenis and Robinson test, the Xie test, as well as the modified Hosmer and Lemeshow test, the modified Pulkstenis and Robinson test, and the modified Xie test show strong evidence against the claim that Models 4 to 8 fit provide an adequate fit. Based upon all these test results, only Models 1, 2 and 3 appear to provide an adequate fit. Since Model 3 has the smallest AIC as shown in Table 9, this model could likely be chosen to be the final model.

Table 9. AIC and p-values from goodness-of-fit tests on the 8 models.

Mod	AIC	HL	PR	XIE	HL+GA	PR+GA	XIE+GA
1	1732.38	0.512	0.933	0.992	0.7208	0.9394	0.9155
2	1708.09	0.724	0.618	0.901	0.3215	0.6327	0.6043
3	**1699.85**	0.301	0.518	0.705	0.1719	0.5321	0.1618
4	1726.00	0.211	0.100	0.000	0.0004	0.0691	2.3e-06
5	1729.26	0.467	0.049	0.001	1.5e-	0.0498	7.8e-06
6	1732.91	0.984	0.061	0.001	2.9e-	0.0316	1.3e-05
7	1733.17	0.770	0.077	0.000	2.2e-	0.0325	5.6e-06
8	1734.95	0.967	0.033	0.000	5.9e-	0.0322	3.2e-06

In this example, the modified HL (HL+GAM) test results appear to have been improved or made more sensitive to lack-of-fit. The modified PR test (PR+GAM) did not dramatically alter the p-values from the original PR test. While the Xie test was fairly sensitive to lack-of-fit in this example, the modified Xie test (Xie+GAM) dramatically decreased the p-values.

5. Discussion

This study investigates possible causes of low power in several chi-square based goodness-of-fit tests (Hosmer and Lemeshow (1980); Pulkstenis and Robinson (2002); Pulkstenis and Robinson (2002)). In summary, for a model containing lack-of-fit, the estimated probabilities can have large discrepancies from the true probabilities, which directly affects the groupings that could lead to low power in these tests. Thus, a new grouping algorithm is proposed to develop an over-fit model that includes all the main effects terms and the interaction terms into a mixture GAM, and then utilizes the estimated probabilities from this GAM to form the

groups. The modified Hosmer-Lemeshow test, modified Pulkstenis and Robinson test, and modified Xie test are proposed based on this new grouping algorithm. Simulation results showed that this new algorithm improves the groupings, especially in the modified Hosmer-Lemeshow test.

It is found that the distribution of the covariates also affects the power of the goodness-of-fit tests. For instance, if a covariate was generated from a right skewed $Beta(4,2) \times 6 - 3$, then a lack-of-fit model could still approximate the true probabilities well. Consequently, all these goodness-of-fit tests, including the newly proposed tests, have low power in this situation. In general, the proposed tests (HL.GAM, PR.GAM, and XIE.GAM) have higher power than the original tests (HL, PR, and XIE) when testing a model that omitted an interaction term or a quadratic term. In addition, if a first-order term was omitted from the model, the original HL, PR and XIE tests all have very low power. However, if data had been collected with respect to this term and had been utilized in the GAM, the proposed tests could achieve high power. This approach might be the only option that could increase the power to detect a model that omits first-order term.

Simulation results show that the size of the proposed test statistics (HL.GAM, PR.GAM, and XIE.GAM) remain at the specified level (α). In addition, the power of the proposed tests are higher than the original tests. The proposed tests are also easy to implement and conveniently supplement the chi-square tests commonly utilized in practice. Evans and Li (2005) conclude from their simulation study, "We propose that researchers do not rely on a single goodness of fit statistic but alternatively use the statistics to compliment each other". This tests proposed here provide an important compliment to such goodness-of-fit tests. Care must also be taken to ensure that the selected GOF tests can be evaluated for all models in the candidate set in terms of the covariate patterns.

References

A. Agresti. Categorical Data Analysis. John Wiley & Sons, 1990.

A. Azzalini, A. Bowman, and A. W. Hardle. On the use of nonparametric regression for model checking. Biometrika, 76(1):1–11, 1989.

C. E. Barat, C. Wright, and B. Chou. Examining potential predictors for completion of the gardasil vaccine sequence based on data gathered at clinics of johns hopkins medical Institutions. Journal of Statistics Education, 19:1–20, 2011.

W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. J. Amer. Statist. Assoc., 74:829–836, 1979.

J. B. Copas. Plotting p against x. Appl. Statist., 32:25–31, 1983.

J. B. Copas. Unweighted sum of squares test for proportions. Appl. Statist., 38(1):71–80, 1989.

M. H. Gail,W.Y. Tan, and S. Piantadosi. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika, 75(1):57–64, 1988.

T. Hastie and R. Tibshirani. Generalized additive models. Statistical Science, 1(3):297–318, 1986.

W.W. Hauck, J. M. Neuhaus, J. D. Kalbfleisch, and S. Anderson. A consequence of omitted covariates when estimating odds ratios. J. Clin. Epidemiol., 44(1):77–81, 1991.

D. W. Hosmer and S. Lemeshow. Goodness of fit tests for the multiple logistic regression model. Commun. Statist. PartA-Theory and Methods, 9(10): 1043–69, 1980.

D. W. Hosmer and S. Lemeshow. Applied logistic regression. John Wiley & Sons, 2nd edition, 2000.

D. W. Hosmer, T. Hosmer, S. Le Cessie, and S. Lemeshow. A comparison of goodness-of fit tests for the logistic regression model. Statist. Med., 16:956–980, 1997.

D.W. Hosmer, S. Lemeshow, and R. X. Sturdivant. Applied logistic regression. John Wiley & Sons, 3th edition, 2013.

S. le Cessie and C. van Houwelingen. A goodness-of-fit test for binary regression models, based on smoothing methods. Biometrika, 47:1267–82, 1991.

M. Mittlbock and M. Schemper. Explained variation for logistic regression. Statistics in Medicine, 15:1987–97, 1996.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. Journal of the Royal Statistical Society, 135(3), 370–384, 1972.

W. K. Newey. Maximum likelihood specification testing and conditional moment tests. Econometrica, 53:1047–70, 1985.

C. Orme. The calculation of the information matrix test for binary data models. The Manchester School, 54(4):370–376, 1988.

G. Osius and D. Rojek. Normal goodness-of-fit tests for multinomial models with large degrees of freedom. Journal of the American Statistical Association, 87(420):1145–52, 1992.

K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine, 50:157–175, 1900.

E. Pulkstenis and T. J. Robinson. Two goodness-of-fit tests for logistic regression models with continuous covariates. Statist. Med., 21:79–93, 2002.

J. Qin and B. Zhang. A goodness-of-fit test for logistic regression models based on case control data. Biometrika, 84(3):609–618, 1997.

C. Reinsch. Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). J. Roy. Statist. Soc. Ser, 47(1), 1–52, 1967.

A. C. Rencher and W. F. Christensen. Methods of Multivariate Analysis. John Wiley & Sons, Inc., 3th edition, 2012.

T. A. Stukel. Generalized logistic models. Journal of the American Statistical Association, 83:426–431, 1988.

A. A. Tsiatis. A note on a goodness-of-fit test for the logistic regression model. Biometrika, 67:250–251, 1980.

H. White. Maximum likelihood estimation of misspecified models. Econometrica, 50(1):1–25, 1982.

X. J. Xie, J. Pendergast, and W. Clarke. Increasing the power: A practical approach to goodness-of-fit test for logistic regression models with coninuous predictors. Computational Statistics & Data Analysis, 52:2703–13, 2008.