TESTING LACK-OF-FIT OF GENERALIZED LINEAR MODELS VIA LAPLACE APPROXIMATION

A Dissertation

by

DANIEL LAURENCE GLAB

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2011

Major Subject: Statistics

TESTING LACK-OF-FIT OF GENERALIZED LINEAR MODELS VIA LAPLACE APPROXIMATION

A Dissertation

by

DANIEL LAURENCE GLAB

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Thomas E. Wehrly
Committee Members,	Jeffrey D. Hart
	Francis J. Narcowich
	Simon J. Sheather
Head of Department,	Simon J. Sheather

May 2011

Major Subject: Statistics

ABSTRACT

Testing Lack-of-Fit of Generalized Linear Models
via Laplace Approximation.
(May 2011)
Daniel Laurence Glab, B.S., University of Wisconsin-Madison;
M.S., Texas A&M University
Chair of Advisory Committee: Dr. Thomas E. Wehrly

In this study we develop a new method for testing the null hypothesis that the predictor function in a canonical link regression model has a prescribed linear form. The class of models, which we will refer to as canonical link regression models, constitutes arguably the most important subclass of generalized linear models and includes several of the most popular generalized linear models. In addition to the primary contribution of this study, we will revisit several other tests in the existing literature. The common feature among the proposed test, as well as the existing tests, is that they are all based on orthogonal series estimators and used to detect departures from a null model.

Our proposal for a new lack-of-fit test is inspired by the recent contribution of Hart and is based on a Laplace approximation to the posterior probability of the null hypothesis. Despite having a Bayesian construction, the resulting statistic is implemented in a frequentist fashion. The formulation of the statistic is based on characterizing departures from the predictor function in terms of Fourier coefficients, and subsequent testing that all of these coefficients are 0. The resulting test statistic can be characterized as a weighted sum of exponentiated squared Fourier coefficient estimators, whereas the weights depend on user-specified prior probabilities. The prior probabilities provide the investigator the flexibility to examine specific departures from the prescribed model. Alternatively, the use of noninformative priors produces a new omnibus lack-of-fit statistic.

We present a thorough numerical study of the proposed test and the various existing orthogonal series-based tests in the context of the logistic regression model. Simulation studies demonstrate that the test statistics under consideration possess desirable power properties against alternatives that have been identified in the existing literature as being important.

To Mom and Dad

ACKNOWLEDGMENTS

Throughout my pursuit of a Ph.D., I have benefitted from the generosity of a number of people. I would like to take this opportunity to express my gratitude to several of those who have helped me the most.

Certainly, I must start by thanking my advisor, Dr. Wehrly, for his thoughtful and patient guidance. I cannot imagine that there is a more approachable and supportive advisor than Dr. Wehrly. It is impossible to cite everything he has done for me throughout this process. He has endured countless burdens and inconveniences in guiding me to completion of my dissertation. However, I would not have grown or learned as much without his insight and input.

In addition to Dr. Wehrly, I want to thank the rest of my committee for the time they took out of their schedule to review drafts of my dissertation, as well as attend my prelim and defense. Anyone who takes the time to skim this dissertation (if not read it) will note that Dr. Hart is cited often. Indeed, his work inspired the problem upon which this dissertation is based. Furthermore, I have benefitted greatly from his considerable insight into this topic. He has reviewed several drafts of this dissertation and has provided invaluable guidance on the theoretical results presented herein. I would also like to thank Dr. Sheather and Dr. Narcowich for their comments, suggestions and bringing some valuable references and research resources to my attention.

I am indebted to the other members of the faculty of the Texas A&M. I have been privileged to have had many outstanding teachers throughout my higher education; however, I found the faculty in the Statistics Department to be as dedicated to teaching and authoritative in regards to subject matter as one could ever hope. Not only did they impart a great deal of knowledge, but they also set a high standard for mastery of the field of

statistics. Indeed, I hope to one day attain this standard. I would especially like to thank Dr. Michael Longnecker for giving me an opportunity to teach and for helping me resolve the occasional crisis.

I also thank Dr. Edward Murguia for giving me a valuable opportunity to pursue collaborative work with The Mexican American and U.S. Latino Research Center (MALRC), as well as all the encouragement and guidance he provided. Dr. Murguia always treated me with dignity and took the time to make sure that I was doing well, and for this I am grateful. Moreover, through my experience at MALRC, I have gained a great deal of exposure to some rather profound social science concepts. These ideas have certainly opened my eyes to different ways to view the world. I would also like to thank Darlene Flores for creating a friendly work environment, as well as for the kindness that she and her husband, Danny, showed me during my time at MALRC.

I would be remiss if I failed to acknowledge the support and guidance provided by my fellow graduate students throughout my studies at A&M. In particular, I indebted to Michele Olds and John Wagaman–I probably would not have been able to survive graduate school without their help and encouragement.

Finally, I would like to give special thanks to my family for their support: Debbie, Tony and Uncle Bob for regularly checking-in with me and trying to encourage me through several frustrating and disheartening moments encountered throughout graduate study. Last but not least, I want to thank Mom and Dad for making it possible for me to pursue meaningful and engaging work. It goes without saying that I would not be where I am today if they had not put my well-being ahead of their own. Their sacrifices have always provided me with the motivation to persist in pursuing my goals.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

TABLE Page

TABLE Page

TABLE Page

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Overview

In this chapter we will define the class of models known as generalized linear models (Section 1.2) as well as the canonical link regression model (Section 1.2.3), which may arguably be the most familiar, if not most important, subclass within generalized linear models. In addition to defining these models we present several examples (Sections 1.2.3, 1.2.3 and 1.2.3) and selected theoretical results which are utilized in the subsequent chapters (Section 1.3). We will also formally present the problem of testing lack of fit for such classes of models. Finally, we will conclude this chapter by providing an overview of the remainder of this dissertation.

1.2 The Generalized Linear Model

The class of linear models for which the distribution of the response variable is in the general exponential family is called the *generalized linear model* (GLM). GLMs constitute one of the most important model classes for data analysis since most of the nonnormal regression models used in practice are members of this class, see, e.g. McCullagh and Nelder (1989), Fahrmeir and Tutz (2001), Shao (2003).

More to make the definition more formal and precise, we will assume the following throughout the remainder of this chapter: Suppose the data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are observed where, for $i = 1, \ldots, n$, $\mathbf{x}_i$ is a fixed vector of covariates and y_i is a scalar response for the ith subject.

This dissertation follows the style of *Biometrics.*

1.2.1 Response Distribution

A GLM consists of a response Y with independent observations $y_1, \ldots, y_n$, each of which have an exponential family probability density function or mass function of form

$$f(y_i; \theta_i, \psi) = \exp\{[y_i\theta_i - b(\theta_i)]/a(\psi) + c(y_i, \psi)\}, \tag{1.1}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions; in practice, $c(\cdot)$ need not be specified explicitly. The parameter θ_i is called the canonical parameter and, in the context of a GLM, the value of θ_i may vary for $i = 1, \ldots, n$ as a function of covariates; that is, $\theta_i = \theta(\mathbf{x}_i)$. The parameter ψ is the unknown dispersion parameter that allows a more flexible relationship between the mean and variance than the traditional least squares regression model. In particular, for Y_i with a probability density or mass function given by (1.1) we have

$$E(Y_i) = b'(\theta_i) \equiv \mu_i \tag{1.2}$$

and

$$\text{var}(Y_i) = a(\psi)b''(\theta_i) \equiv a(\psi)v(\mu_i) \tag{1.3}$$

where the variance function $v(\mu)$ is uniquely determined by the specific exponential family through the relation $v(\mu) = b''(\theta)$. Several important distributions are special cases of (1.1), including the Poisson and binomial.

1.2.2 Components of the Exponential Family Regression Model

Using the exponential family for a regression analysis requires three specifications. First, we need to specify the *random component* of the model; that is, we have to choose which member of (1.1) will be taken as the response distribution. Since $b(\theta_i)$ uniquely deter-

mines each member of (1.1), specifying the random component amounts to selecting $b(\cdot)$ to produce a distribution which reflects the observed data type for the response.

The *systematic component* of an exponential family regression model refers to the specification of the function used to obtain an estimate the unknown regression function, which is denoted by η. Since η is a function of the covariates, we will write $\eta_i = \eta(\mathbf{x}_i)$, $i = 1, \ldots, n$. In general, η can be estimated using either a parametric or nonparametric regression model. Examples of nonparametric regression models include the generalized additive model (Hastie and Tibshirani, 1986, 1987, 1990) and the single index model (Ichimura, 1993). A GLM estimates a vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^{\mathrm{T}}$ as a function of the explanatory variables through a linear model

$$\eta(\mathbf{x}_i; \boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j \gamma_j(\mathbf{x}_i), \quad i = 1, \ldots, n. \tag{1.4}$$

where $\gamma_1, \ldots, \gamma_p$ are known functions, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ an unknown parameter vector. This linear combination of explanatory variables is called the *linear predictor*. Usually, $\gamma_1(\mathbf{x}_i) = 1$ for all i, in which case β_1 will be regarded as the coefficient of an intercept in the model. The linearity distinguishes GLMs from other exponential family regression models such as generalized additive models or single index models, which use more general regression functions.

The third component of an exponential family regression model is a *link function* that connects the random and systematic components. The model links μ_i to η_i through the formula

$$h(\mu_i) = \eta_i, \tag{1.5}$$

where the link function h is a monotonic, differentiable function. Hence the link function "links" the mean response to the explanatory variables. Since $\mu_i = b'(\theta_i)$, there is an

implied relationship

$$g(\theta_i) = \eta_i \tag{1.6}$$

between θ_i and $\boldsymbol{\beta}$.

The link function $h(\mu) = \mu$, called the *identity link*, has $\eta_i = \mu_i$. It specifies a linear model for the mean itself. The choice of $h(\mu_i)$ such that $\theta_i = h(\mu_i)$ or

$$\theta_i = \eta_i \tag{1.7}$$

is called the *canonical-link function*.

In a canonical link regression model, the analyst has more control over specification of the systematic component than any other component comprising the model. We have noted that specification of the response distribution is dictated primarily by the observed data type. Furthermore, since the canonical link requires that $\theta_i = h(\mu_i)$ while $\mu_i = b'(\theta_i)$, specification of $h(\cdot)$ depends upon the specification of $b(\cdot)$. In other words, for the canonical link regression model, the link is implicitly specified upon specification of the response distribution. A few examples will be discussed in the following subsection.

1.2.3 Canonical Link Models

Canonical link models constitute some of the most commonly used models within the class of GLMs. In this section we present canonical link models corresponding to several of the most familiar response distributions.

The normal regression model

Suppose y_i, $i = 1, \ldots, n$ have been observed from a normally distributed response, i.e.,

$$Y_i = \mu_i + \epsilon_i, \quad i = 1, \ldots, n,$$

$\mu_i = \mu(\mathbf{x}_i)$, $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $N(0, 1)$. We can write the response density as

$$\varphi(y_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y_i - \mu_i)^2}{2\sigma^2}\right\} = \exp\left\{\frac{y_i\mu_i - \mu_i^2/2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma)\right\} \quad (1.8)$$

Thus, taking

$$\theta_i = \mu_i \quad (1.9)$$

it is clear that Y_i has an exponential family distribution as in (1.1) with

$$a(\psi) = \sigma^2, \quad b(\eta(\mathbf{x}_i)) = [\eta(\mathbf{x}_i)]^2/2, \quad c(y_i, \psi) = -y_i^2/(2\sigma^2) - \log(\sqrt{2\pi}\sigma)$$

where we set $\psi = \sigma$. From (1.7) and (1.12) we see that the canonical link model for normal data is obtained by taking h in (1.5) to be the identity link, $h(\mu_i) = \mu_i$, so that

$$\eta(\mathbf{x}_i) = h(\mu_i) = \mu_i = \theta_i. \quad (1.10)$$

The logistic regression model

Suppose now that the observed response is binary, that is, $y_i = 0, 1$, $i = 1, \ldots, n$. In this case one would be inclined to proceed as if y_i is a realization of the $\mathrm{Bernoulli}(\mu_i)$ random variable Y_i where $\mu_i = \mu(\mathbf{x}_i) \in (0, 1)$, $i = 1, \ldots, n$ denotes the probability of "success" for the ith individual. The probability mass function of Y_i is given by

$$P(Y_i = y_i) = \mu_i^{y_i}(1-\mu_i)^{1-y_i} = \exp\left\{y_i \log\left(\frac{\mu_i}{1-\mu_i}\right) - \log(1-\mu_i)\right\}, \;\; y_i = 0, 1 \tag{1.11}$$

Thus, taking

$$\theta_i = \log\left(\frac{\mu_i}{1-\mu_i}\right) \tag{1.12}$$

it is clear that Y_i has an exponential family distribution as in (1.1) with

$$a(\psi) = 1, \quad b(\theta_i) = \log(1-\mu_i), \quad c(y, \psi) \equiv 0. \tag{1.13}$$

Note the absence of a nuisance parameter ψ for this response distribution. From (1.7) and (1.12) we see that the canonical link model for Bernoulli data is obtained by taking h in (1.5) to be the logistic link, $\log\{\mu_i/(1-\mu_i)\}$, so that

$$\eta(\mathbf{x}_i) = h(\mu_i) = \log\{\mu_i/(1-\mu_i)\} = \theta_i. \tag{1.14}$$

This in turn implies that the probability of success can be expressed as follows

$$\mu(\mathbf{x}_i) = \frac{\exp\{\eta(\mathbf{x}_i)\}}{1+\exp\{\eta(\mathbf{x}_i)\}}. \tag{1.15}$$

The Poisson regression model

Suppose y_i, $i = 1, \ldots, n$ denote counts so that the response values are nonnegative integers. In this case a reasonable distribution for Y_i is the Poisson distribution. Then the probability mass function of Y_i is given by

$$P(Y_i = y_i) = \frac{e^{\mu_i}\mu_i^{y_i}}{y_i!} = \exp\left\{y_i \log(\mu_i) + \mu_i - \log(y_i!)\right\}, \;\; y_i = 0, 1, 2, \ldots \tag{1.16}$$

Thus, taking

$$\theta_i = \log(\mu_i) \tag{1.17}$$

it is clear that Y_i has an exponential family distribution as in (1.1) with

$$a(\psi) = 1, \quad b(\theta_i) = \mu_i, \quad c(y, \psi) \equiv 0. \tag{1.18}$$

As in the case of logistic regression, the response distribution does not have a nuisance parameter ψ. From (1.7) and (1.17) we see that the canonical link model for Poisson data is obtained by taking h in (1.5) to be the log link, $h(\mu_i) = \log\mu_i$, so that

$$\eta(\mathbf{x}_i) = h(\mu_i) = \log\mu_i = \theta_i. \tag{1.19}$$

1.3 Selected Estimation and Inference Results

To make our discussion of generalized linear models more complete, we feel that it is beneficial to briefly review of some fundamental results on estimation and inference applied to generalized linear models (particularly, canonical link regression models). While the results described in this section will be familiar to the reader, this discussion provides an opportunity to further establish terminology and notation conventions that will be used repeatedly throughout the remainder of this dissertation. The results summarized in this section both provide the impetus for pursuing the method we propose as well as justification for its use.

In the context of generalized linear models, both estimation and inference are based on (log-)likelihoods. In light of (1.7), we write the log-likelihood function for a canonical link model explicitly in terms of a specified regression function as follows

$$\begin{aligned} l(\boldsymbol{\beta}, \psi) &= \sum_{i=1}^{n} \log f(y_i; \eta(\mathbf{x}_i; \boldsymbol{\beta}), \psi) \\ &= \sum_{i=1}^{n} \{[y_i \eta(\mathbf{x}_i; \boldsymbol{\beta}) - b(\eta(\mathbf{x}_i; \boldsymbol{\beta}))]/a(\psi) + c(y_i, \psi)\}, \end{aligned} \tag{1.20}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are defined as in 1.1.

1.3.1 Regularity Conditions

Several assumptions are required to ensure that inference results we will review in this section actually hold and that the parameter estimators possess certain desirable properties. While there is no unique collection of assumptions, there are several generally accepted conditions that have been adopted in the literature. These assumptions are often referred to as ("standard" or "general") *regularity conditions*. Regularity conditions mainly relate to identifiability of the parameters; existence and behavior derivatives of the response density with respect to the parameters; existence of third moments of y; and convergence of the Hessian of the log-likelihood scaled by the sample size, $l(\boldsymbol{\beta}, \psi)$, to a positive definite limit as the sample size tends to infinity. We will defer a formal presentation of regularity conditions until Chapter IV where we will use them explicitly. A collection of regularity conditions that are appropriate for application to GLMs is presented in Shao (2003).

1.3.2 Estimation

In a generalized linear model, the parameter of interest is $\boldsymbol{\beta}$. This parameter is usually estimated by maximum likelihood estimation. Possible candidates for the maximum likelihood estimates are the roots of the score function

$$s_n(\widehat{\boldsymbol{\beta}}) = 0 \tag{1.21}$$

where $\widehat{\boldsymbol{\beta}}^{\mathrm{T}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)$ denotes the estimated value of $\boldsymbol{\beta}$ and

$$\begin{aligned} s_n(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta}, \psi)}{\partial \boldsymbol{\beta}} &= \frac{1}{a(\psi)} \sum_{i=1}^{n} \left\{ [y_i - b'(\eta(\mathbf{x}_i; \boldsymbol{\beta}))] \frac{\partial \eta(\mathbf{x}_i; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right\} \\ &= \frac{1}{a(\psi)} \sum_{i=1}^{n} \left\{ [y_i - b'(\eta(\mathbf{x}_i; \boldsymbol{\beta}))] \Gamma_i \right\}. \end{aligned} \tag{1.22}$$

Note that an MLE of $\boldsymbol{\beta}$ can be obtained without estimating ψ. Obtaining an estimate of ψ by maximum likelihood estimation is generally difficult in practice, so several other alternative estimators have been suggested (Fahrmeir and Tutz, 2001; McCullagh and Nelder, 1989).

A closed form solution of the MLE of $\widehat{\boldsymbol{\beta}}$ is not available for most GLMs. Thus, numerical techniques such as the Newton-Raphson or the Fisher-scoring methods are required to obtain estimates. These two methods are identical for canonical link regression models and are summarized by the following iteration procedure:

$$\widehat{\boldsymbol{\beta}}_n^{(t+1)} = \widehat{\boldsymbol{\beta}}_n^{(t)} - [J(\widehat{\boldsymbol{\beta}}_n^{(t)})]^{-1} s_n(\widehat{\boldsymbol{\beta}}_n^{(t)}), \quad t = 0, 1, \ldots, \tag{1.23}$$

where $J(\widehat{\boldsymbol{\beta}}_{mk}) = J(\boldsymbol{\beta})|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}$ and $J(\boldsymbol{\beta}_{mk})$ is the Hessian of the log-likelihood

$$J(\boldsymbol{\beta}) = -\frac{\partial^2 l(\boldsymbol{\beta}, \psi)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} = (\Gamma^{\mathrm{T}} W(\boldsymbol{\beta}) \Gamma) \tag{1.24}$$

where $W(\boldsymbol{\beta}) = \mathrm{diag}\{w_1(\boldsymbol{\beta}), \ldots, w_n(\boldsymbol{\beta})\}$, and $w_j^{-1}(\boldsymbol{\beta}) = h'(\mu_j(\boldsymbol{\beta}))^2 v(\mu_j(\boldsymbol{\beta}))$ (McCullagh and Nelder, 1989). The matrix $J(\boldsymbol{\beta})$ is often called the *information matrix*.

A numerical approximation to an MLE that we will find particularly useful is called the "one-step" MLE and can be obtained by taking the first iteration of the Newton-Raphson procedure (i.e., $t = 0$)

$$\widehat{\boldsymbol{\beta}}_n^{(1)} = \widehat{\boldsymbol{\beta}}_n^{(0)} - [J(\widehat{\boldsymbol{\beta}}_n^{(0)})]^{-1} s_n(\widehat{\boldsymbol{\beta}}_n^{(0)})$$

where $\widehat{\boldsymbol{\beta}}_n^{(0)}$ is the initial value.

1.3.3 Statistical Properties of Estimators

Under general regularity conditions the following hold for $\widehat{\boldsymbol{\beta}}$ (Shao, 2003)

(i) There is a unique sequence $\{\widehat{\boldsymbol{\beta}}_n\}$ such that

$$P(s_n(\widehat{\boldsymbol{\beta}}_n) = 0) \to 1 \text{ and } \widehat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta},$$

(ii) Let $\mathcal{I}_n(\boldsymbol{\beta}) = \mathrm{Var}(s_n(\boldsymbol{\beta}))$. Then

$$[\mathcal{I}_n(\boldsymbol{\beta})]^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} N_p(\mathbf{0}, I_p).$$

where I_p is a $p \times p$ identity matrix.

(iii) If ψ in (1.1) is known or the p.d.f. in (1.1) indexed by $(\boldsymbol{\beta}, \psi)$ satisfies the suitable conditions, then $\mathcal{I}_n(\boldsymbol{\beta}) = J(\boldsymbol{\beta})$, that is, $\widehat{\boldsymbol{\beta}}_n$ is asymptotically efficient.

If an initial estimate, $\widehat{\boldsymbol{\beta}}_n^{(0)}$, is $\sqrt{n}$-consistent for $\boldsymbol{\beta}$, then the one-step MLE $\widehat{\boldsymbol{\beta}}_n^{(1)}$ is asymptotically efficient under the same conditions (Shao, 2003).

1.3.4 Likelihood-based Inference on Regression Coefficients

In our subsequent discussion, we will often encounter basic hypothesis testing techniques designed to assess the contribution of a subset of the covariates to the linear predictor of a GLM. While such methods of testing are well-known (particularly in the context of traditional linear model theory), we will take this opportunity to briefly review two test statistics that can be used to carry out such tests and highlight their properties as well as establish notational conventions that will be utilized throughout the remainder of this dissertation.

Let $\boldsymbol{\beta}$ be partitioned as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$. Suppose, without loss of generality, that we wish to investigate the contribution of the subset of covariates corresponding to $\boldsymbol{\beta}_2$. Thus, it is of interest to test

$$H_0 : \ \boldsymbol{\beta}_2 = 0, \ \boldsymbol{\beta}_1, \ \psi \ \text{unrestricted} \tag{1.25}$$

against

$$H_1 : \ \boldsymbol{\beta}_1, \ \boldsymbol{\beta}_2, \ \psi \ \text{unrestricted}. \tag{1.26}$$

We will distinguish the parameter estimates obtained from the models corresponding to these hypotheses by denoting the unrestricted MLE under H_1 as $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2)$ and the MLE under the restriction of H_0 as $\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\beta}}_1, \mathbf{0})$. Correspondingly, $\widehat{\psi}$ and $\widetilde{\psi}$ are consistent estimators of ψ under H_1 and H_0, respectively.

The *likelihood ratio statistic*

$$\mathcal{L} = -2\{l(\widetilde{\boldsymbol{\beta}}) - l(\widehat{\boldsymbol{\beta}})\}$$

compares the unrestricted maximum $l(\widehat{\boldsymbol{\beta}}) = l(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\psi})$ of the (log-)likelihood with the maximum $l(\widetilde{\boldsymbol{\beta}}) = l(\widetilde{\boldsymbol{\beta}}_1, \mathbf{0}, \widetilde{\psi})$ obtained for the restricted MLE $\widetilde{\boldsymbol{\beta}}$, computed under H_0 (note that the dependence upon the dispersion parameter ψ is suppressed for convenience). H_0 will be rejected in favor of H_1 if the unrestricted maximum $l(\widehat{\boldsymbol{\beta}})$ is significantly larger than $l(\widetilde{\boldsymbol{\beta}})$, implying that $\mathcal{L}$ is large.

Alternatively, a test based on the *score statistic* rejects H_0 when the value of

$$\mathcal{S} = [s(\widetilde{\boldsymbol{\beta}})]^{\mathrm{T}} J^{-1}(\widetilde{\boldsymbol{\beta}}) s(\widetilde{\boldsymbol{\beta}}) \tag{1.27}$$

is large, where $s(\boldsymbol{\beta})$ is the score function and $J(\boldsymbol{\beta})$ is the information matrix. The score test is based on the rationale that the statistic $\mathcal{S}$ measures the distance between the score

function evaluated at the estimates obtained under the null hypothesis and the zero vector in a fashion similar to that of the Mahalanobis distance. However, by (1.21), $s(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$ and hence $\mathcal{S}$ measures the discrepancy between $s(\widehat{\boldsymbol{\beta}})$ and $s(\widetilde{\boldsymbol{\beta}})$. If H_0 is not true, the estimates of $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ will differ so that $s(\widetilde{\boldsymbol{\beta}})$ will be significantly different from $\mathbf{0}$, which in turn leads to a large value of $\mathcal{S}$ and rejection of H_0. It is worth noting that (1.27) can be simplified to an expression involving a subvector of the score vector and submatrix of the inverse information matrix each constructed to conform with the partitioning of $\boldsymbol{\beta}$.

Under H_0, both $\mathcal{L}$ and $\mathcal{S}$ are asymptotically $\chi^2(r)$, $r = \dim(\boldsymbol{\beta}_2)$, provided that regularity conditions such as those described in Section 1.3.1 hold. Under such regularity conditions, the Taylor series expansion can be used to express the likelihood ratio in terms of the score statistic as follows

$$-2\{l(\widetilde{\boldsymbol{\beta}}) - l(\widehat{\boldsymbol{\beta}})\} = [s(\widetilde{\boldsymbol{\beta}})]^{\mathrm{T}} J^{-1}(\widetilde{\boldsymbol{\beta}}) s(\widetilde{\boldsymbol{\beta}}) + O_p(n^{-1/2}).$$

Thus, the score statistic is asymptotically equivalent to the likelihood ratio statistic. In fact, this representation along with the asymptotic normality of $s(\boldsymbol{\beta})$ is a key step in demonstrating that $\mathcal{L}$ has a chi-square limiting distribution (Shao, 2003).

Generally, the likelihood ratio statistic $\mathcal{L}$ is preferred for moderate sample sizes (Fahrmeir and Tutz, 2001). However, by the asymptotic equivalence of $\mathcal{L}$ and $\mathcal{S}$ cited above, it is reasonable to use $\mathcal{S}$ for larger sample sizes since the two test statistics will tend to agree closely. In such cases $\mathcal{S}$ is often preferred because its computation only requires an estimate from the null (i.e., constrained) model. This property of the score statistic is especially convenient in situations where there are multiple tests under consideration.

1.4 Testing the Fit of a Parametric Model

In order for a model such as a GLM to be of practical use, we must have some assurance that it provides a reasonably accurate description of the data. Tests of fit provide a means of assessing how well a statistical model fits the observed data. Some of the more familiar tests of fit achieve this objective rather explicitly by directly measuring the discrepancy between observed values and the values expected under the proposed model.

In our discussion we will assume that the model under consideration can be thought of as a "final model" in that it represents a regression model determined via a formal model building analysis. In particular, we assume that to the best of our knowledge, the model contains those variables that should be in the model and that the variables have been entered in the correct functional form.

In general, tests of fit fall mainly into two categories: (a) parametric methods designed to detect specific types of departures from the prescribed model, and (b) nonparametric methods. Parametric tests embed the model under consideration in a wider class of (parametric) models and check if the data can be better described by the more general model. If not, we stay with our fitted model. On the other hand, a nonparametric test of a parametric hypothesis does not evaluate specific parametric alternatives, but rather tests unspecific hypotheses of the form 'the model fits' versus the alternative 'the model does not fit'. Such tests are appealing in that, for a sufficiently large sample size, they are able to detect virtually any departure from the hypothesized parametric model. While we will discuss tests from each of the two categories cited, we will focus primarily on nonparametric tests.

One approach for constructing nonparametric tests is to embed the null model in a much wider class of parametric models that increases without bound as $n \to \infty$. The class of parametric models constitutes a collection of alternative models against which the null is tested. This strategy for constructing tests will be studied extensively in this

dissertation. While this approach may sound similar to the parametric methods described above, it is distinct from parametric tests in that these are not designed to detect specific departures. That is, the class of alternatives is formulated in order to approximate any possible departure from the null model and the departures approximated by the collection of alternative models grows as $n \to \infty$.

1.4.1 The Lack-of-Fit Test

For the exponential regression model described in Section 1.2 consider a test of whether η belongs to a specified parametric family,

$$H_0 : \ \eta(\cdot) \in \{\eta(\cdot; \boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathcal{B}\} \tag{1.28}$$

where the parameter space $\mathcal{B}$ is a subset of $\mathbb{R}^p$ with p a finite positive integer.

In this context, our interest is in tests that are sensitive to essentially any departure from a proposed parametric model for η. Stated more precisely, H_0 is contrasted with the nonparametric alternative

$$H_a : \ \eta(\cdot) \notin \{\eta(\cdot; \boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathcal{B}\}. \tag{1.29}$$

Since this formulation leaves the functional form of η unspecified, the test we have just described constitutes a nonparametric test of the regression function.

In the context of regression, such a test is usually referred to as a *lack-of-fit* test. An analogous procedure for testing whether a set of independent identically distributed observations arises from a given class of probability distributions is more often called a *goodness-of-fit* test; however, several authors working in the regression context have used the term “goodness-of-fit” in reference to their proposed method. Consequently, we will use the terms lack-of-fit and goodness-of-fit somewhat interchangeably, but we will often

refer to such testing methods generically “test of fit”.

1.4.2 Smoothing-based Tests of Fit

Lack-of-fit tests may be constructed by means of nonparametric smoothers. The motivation for the use of nonparametric function estimates as a means to validate parametric models comes from the notion that a well-constructed nonparametric estimate will be free of any unjustified restrictions which may be imposed by specification of a parametric model. That is, by imposing minimum structure on the regression function, a nonparametric curve estimate is designed to reflect *only* the evidence which is available from the data. This notion is summed up in the familiar expression, “nonparametric methods let the data speak for themselves.” Thus, if one accepts the notion that the nonparametric estimate depicts the data well, then a fitted parametric model which produces a meaningful departure from a given nonparametric estimate may be viewed as an inadequate fit for the observed data. This is the fundamental premise for pursuing lack-of-fit tests based on nonparametric smoothing.

Smoothing-based tests have the following desirable characteristics, (Hart, 1997):

1. They are omnibus in the sense of being consistent against each member of a very large class of alternative hypotheses.

2. They tend to be more powerful than competing omnibus tests.

3. The corresponding nonparametric function estimate provides insight to the nature of the lack of fit.

There are two ways to utilize smoothing methods in testing the fit of a parametric model: either compare a nonparametric estimate of $\eta(\mathbf{x})$ to the parametric model or else examine a smoothed version of the residuals (obtained from the parametric model) for

departures from zero. In the former approach one obtains both a parametric and a nonparametric estimate of the regression function and then proceeds to examine some measure of discrepancy (e.g., Kullback-Leibler difference) between the two estimates. Following the logic discussed above, one would reject the parametric model if a nontrivial discrepancy were observed. In the latter approach one obtains residuals for the parametric model and subsequently obtains a nonparametric estimate of the underlying residual function. In this case, if the parametric model holds, one would expect that the underlying residual function is identically 0. Thus, nontrivial departures from 0 in the nonparametric estimate of the residual function would lead to rejection of the parametric model. Smoothed residual methods are generally easier to implement and possess desirable theoretical properties (Hart, 1997). Thus, as we will see in the subsequent chapters, residual-based methods have received more attention in the literature.

1.5 Discussion

In Chapters II and III we will review and discuss two distinct strands of research which provide various means of testing the fit of GLMs. In Chapter IV propose and examine a test for canonical link regression models which is inspired by a recent contribution to the series based lack-of-fit testing literature. In Chapter V we study the power properties of the test proposed in Chapter IV in a logistic regression setting via simulation and compare the test's performance with some of the more widely accepted tests discussed in Chapter II. In Chapter VI we will discuss our final conclusions based on findings from the previous chapters and identify future directions for research.

In Chapter II we will review the first of two collections, which focuses on testing the fit of an important special case of generalized linear models, the logistic regression model. As we will describe in greater detail in Chapter II, examining the fit of a logistic regression

model presents some practical problems which are unique to this specific model. Hence, while some of the methods discussed in Chapter II apply to any GLM, most have been developed specifically for logistic regression models and address the binary nature of the response. A few of these methods have gained rather wide acceptance and are viewed as the preferred means to the test the logistic regression model.

Next, in Chapter III we will discuss the second strand of development which provides an alternative approach to testing the fit of a regression model by utilizing orthogonal series estimators to detect departures from a proposed null model. Several of these tests have been applied in a canonical link regression setting and thus can be used to test the fit of a logistic regression model. This line of development has emerged out of the literature for nonparametric tests of fit which have been inspired by the Kolmogorov-Smirnov and Cramér-von Mises tests of goodness-of-fit.

With our review we intend to provide answers to the following questions for the two collections of literature identified above in an effort to justify further examination of our proposal:

1. What are the contributions which have been made to the lack-of-fit testing literature within each of the two collections of literature?

2. How do these two collections of literature differ each other?

3. How do the contributions within each collection differ from each other and in what ways are they similar?

4. What has the literature revealed in regards to the relative performance of these tests in settings often encountered in practice?

With regard to the last question, we note that while tests based on parametric families will be discussed briefly in Section 2.6, we will primarily focus on omnibus tests of fit. Hart

(2009) explains that no one omnibus test will ever be superior (in terms of power) to every other omnibus test. Thus, we will compare the tests on the basis of the "overall" power properties reported in the literature and other factors, such as simplicity and how widely they can be applied.

Finally, it is interesting to note that the bibliographies of these two collections of research contain few common references. Most of these common references are devoted to general issues such as GLMs and their properties rather than testing methodology. Furthermore, it is even more rare that articles in one collection research cites articles from the other. Consequently, there has been no resolution to the issue of the relative performance of the methods from these two collections of research. It is our hope that the findings of Chapter V will provide some measure of resolution to this question.

CHAPTER II

TESTS OF FIT FOR LOGISTIC REGRESSION MODELS

2.1 Overview

Given the wide use and applicability of the logistic regression model to analyze data from essentially every field of applied science, finding means to validate fitted models is a rather urgent issue. As a consequence, there has been a great deal of research devoted to development of methods to assess the adequacy a fitted logistic regression model. These techniques include residual analysis, diagnostic measures such as pseudo R^2 measures as well as formal tests of overall fit of the model. While residual analysis and diagnostic measures can provide useful insight, their interpretation is open to subjectivity. On the other hand, formal tests of fit are equipped with p-values which can be more objectively interpreted. Furthermore, while residual analysis can be extremely valuable for assessing the fit of a model, it can only provide insight on a case by case basis. Tests of fit, however, combine all evidence existing in the data into a single indicator of overall fit. The focus of the remainder of our discussion will be on tests of fit.

The literature on tests of fit for logistic regression is vast. To help prioritize the topics of this chapter, we will primarily concentrate on tests which have been studied in two review articles. These articles are Hosmer, Hosmer, Le Cessie, and Lemeshow (1997) and Kuss (2002) and are devoted to comparison of several well-known tests of fit for the logistic regression model. The focus of these papers was mainly on global tests of fit with special attention given to the efficacy of these tests in the presence of sparse data, an issue which we will discuss later in this chapter.

The rest of this chapter is organized as follows. In Section 2.2, we will discuss some basic notation and terminology conventions which are applicable to all of the tests studied

in this chapter. In Section 2.3, we will discuss several tests related to Pearson's chi-square test. In Section 2.4, we will discuss tests which make direct use of nonparametric function estimation in order to assess model fit. In Section 2.5, we will discuss a specification test which has been utilized to evaluate model fit. In Section 2.6 the goodness-of-link test for logistic regression will be reviewed. In Section 2.7, we conclude the chapter with a summary of progress made in lack-of-fit tests for logistic regression models and present an approach to testing the fit of a logistic regression model which is often overlooked in the literature and was not considered among the tests studied in Hosmer et al. (1997).

2.2 Background and Fundamental Concepts

In Section 1.2.3, we reviewed the logistic regression model and discussed its use to estimate the probability of an event of interest for binary response data. In this section we revisit the binary response setting and introduce some notational conventions and commonly observed features of such data.

Suppose that our fitted model contains p independent variables, $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$, and let J denote the number of distinct values of $\mathbf{x}$ observed. We will refer to each of these distinct values of $\mathbf{x}$ as a "covariate pattern", while the collection of individuals sharing a covariate pattern are referred to as a "covariate class". For $j = 1, 2, \ldots, J(n)$ we denote the jth covariate class by $\mathbf{x}_j^*$, the size of the covariate class with $\mathbf{x} = \mathbf{x}_j^*$ by m_j, and the number of individuals for which $y_j = 1$ by $y_j^* = \sum_{i=1}^{n} y_i I_{\{\mathbf{x}_i = \mathbf{x}_j^*\}}$. It follows that $\sum m_j = n$. Note that in general, $J(n)$ is a function of n since increasing the sample could lead to new covariate patterns. The distinction between $J(n)$ and n is important in our subsequent discussion because most goodness-of-fit tests for generalized linear models (and hence logistic regression models) are assessed over the distinct fitted values of the model (of which there are $J(n)$–each corresponding to a distinct covariate pattern) and not

the individual observations (Hosmer and Lemeshow, 2000; Fahrmeir and Tutz, 2001).

Binary response data are often referred to as "sparse" when a sizeable proportion of m_j's are small with the extreme case occurring when $m_i = 1$ for each i (i.e., each covariate pattern is observed only once). McCullagh and Nelder (1989, p.120) explain that sparseness should not be misinterpreted as an indication that the data contain little information about the underlying model. As we will discuss later in this section, sparse data presents problems for two of the more commonly used classical global tests of fit for logistic regression models. Methods which evaluate the fit for sparse data models are of particular interest because, as Kuss (2002) states, sparseness appears to be "more the rule than the exception in today's data sets". This is due to the fact that sparse data is generally a consequence of the inclusion of continuous covariates in the set of candidate explanatory variables used to fit the model (Hosmer et al., 1997). Sparseness can also result in data sets consisting of a relatively large number of explanatory variables.

2.3 Chi-square Tests of Fit

Two summary statistics often used to assess the adequacy of generalized linear models are the residual deviance (likelihood ratio) and Pearson chi-square. The Pearson statistic has the following general formula

$$X^2 = \sum_{i=1}^{J(n)} \widehat{r}_j^2 \tag{2.1}$$

where

$$\widehat{r}_j = \frac{\widehat{e}_j}{\sqrt{v(\widehat{\mu}_i)}}, \;\; j = 1, \ldots, J(n), \tag{2.2}$$

in which $\widehat{e}_j = y_j^* - \widehat{y}_j$, $\widehat{y}_j$ is the fitted value corresponding to the jth covariate pattern and $v(\widehat{\mu}_i)$ is as defined in Section 1.2.1. For binary response data, $\widehat{y}_j = m_j\widehat{\pi}_j$ and $v(\widehat{\mu}_i) = m_j\widehat{\pi}_j(1 - \widehat{\pi}_j)$. $\widehat{e}_j$ is often called a "response residual" and is simply an applica-

tion of the usual residual definition for the Gaussian linear model. $\widehat{r}_j$ is called a Pearson residual and is clearly obtained by rescaling $\widehat{e}_j$. While we will not pursue a discussion of direct examination of residuals to assess model adequacy, we will observe that several other test statistics can be expressed in terms of residuals. Recognizing how these statistics depend on residuals will help us compare how the statistics operate and simplify some of our notation. These expressions also help simplify arguments for justifying fundamental theoretical results.

The (residual) deviance is the GLM analog of the residual sum of squares in the linear regression and is defined as follows

$$D = 2\phi \sum_{j=1}^{J(n)} \{l_i(\widehat{\mu}_i) - l_i(y_i)\}\,, \tag{2.3}$$

where $\widehat{\mu}_i$, $v(\widehat{\mu}_i)$ are the estimated mean and variance function, respectively, and $l_i(y_i)$ is the individual log-likelihood where $\widehat{\mu}_i$ is replaced by y_i (the maximum likelihood achievable). For binary response data we can write (2.3) in terms of Bernoulli log-likelihood functions to obtain

$$D = -2 \sum_{j=1}^{J(n)} \left\{ y_j^* \log\left(\frac{y_j^*}{m_j \widehat{\pi}_j}\right) + (m_j - y_j^*) \log\left(\frac{m_j - y_j^*}{m_j(1-\widehat{\pi}_j)}\right) \right\}, \tag{2.4}$$

where ψ is omitted from the notation since the response distribution in this case does not depend on a dispersion parameter.

Large values of X^2 and D typically indicate lack-of-fit. For binary response data, significance can be assessed by comparing these statistics with the $\chi^2(J(n)-p-1)$ distribution for large n provided that certain conditions hold. In particular, one must be able to assume that $m_j \widehat{\pi}_j\,(1-\widehat{\pi}_j) \to \infty$ for each $j = 1, \ldots, J(n)$ if n were permitted to approach ∞ while $J(n)$ remains constant (McCullagh and Nelder, 1989, p.118). Hence $m_j \to \infty$ for $j = 1, \ldots, J(n)$ while $J(n)$ must remain constant in order for X^2 and D to have an

asymptotic chi-square distribution.

McCullagh and Nelder (1989, p.120) explain that when the data are sparse, the deviance function and Pearson's statistic fail to satisfy conditions required in order to attain the asymptotic chi-square distribution used to evaluate significance of tests based on these statistics. Obviously, when $m_j = 1$ for a sizable portion of the data, it is unreasonable to assume the conditions described above hold. Consequently, large values of X^2 or D cannot necessarily provide evidence for lack of fit. Furthermore, in extreme cases these statistics can fail to measure a discrepancy between the fitted model and observed data.

To illustrate their last point, McCullagh and Nelder (1989) consider (2.4) in the strictly sparse case for which $m_j = 1$, $j = 1, \ldots, J(n)$. Note that $y\text{log}y = (1-y)\text{log}(1-y) = 0$ when $y = 0$ or 1 and according to Section 2.2, $y_j^* = y_j$ for $m_j = 1$. Further, $\widehat{\eta}_j = \text{log}(\widehat{\pi}_j/(1-\widehat{\pi}_j)) = \mathbf{x}_j^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$. Noting that $J(n) = n$ we see that (2.4) simplifies to

$$\begin{aligned} D &= -2\sum_{j=1}^{n}\left\{y_j\text{log}\left(\frac{y_j}{\widehat{\pi}_j}\right) + (1-y_j)\text{log}\left(\frac{1-y_j}{1-\widehat{\pi}_j}\right)\right\} \\ &= -2\sum_{j=1}^{n}\left\{\widehat{\pi}_j\text{log}\left(\frac{\widehat{\pi}_j}{1-\widehat{\pi}_j}\right) + \text{log}\left(1-\widehat{\pi}_j\right)\right\} \\ &= -2\widehat{\boldsymbol{\beta}}^{\mathrm{T}}X^{\mathrm{T}}\mathbf{Y} - 2\sum\text{log}(1-\widehat{\pi}_j) \\ &= -2\widehat{\boldsymbol{\eta}}^{\mathrm{T}}\widehat{\boldsymbol{\pi}} - 2\sum\text{log}(1-\widehat{\pi}_j) \end{aligned} \tag{2.5}$$

since $X^{\mathrm{T}}Y = X^{\mathrm{T}}\widehat{\boldsymbol{\mu}}$ is the maximum-likelihood equation. Inspection of (2.5) reveals that for $m_j = 1$, D is a function of $\mathbf{Y}$ only through $\widehat{\boldsymbol{\beta}}$. Hence, given $\widehat{\boldsymbol{\beta}}$, D has a conditionally degenerate distribution. Consequently, D is incapable of measuring the discrepancy between the fitted values from the model and the observed response values when the data is strictly sparse and hence cannot be used to test the fit of the logistic regression model.

For the remainder of this section we will review several proposals which have been offered in the existing literature to overcome the shortcomings of the traditional chi-square tests of fit for binary response models described above. There are three approaches that have been used in an effort to resolve these problems:

1. modify the reference distribution for assessing significance as was considered in McCullagh (1985, 1986) and Osius and Rojek (1992);

2. consider modifying existing test statistics as was proposed in Farrington (1996) and Copas (1989);

3. group observations as has been suggested by Hosmer and Lemeshow (1980) and Tsiatis (1980).

2.3.1 Tests Based on Modified Limiting Distributions for X^2 and D

McCullagh and Nelder (1989) assert that when the m_j are small but mostly greater than one, either D or X^2 may be used to test the fit of a logistic regression model. However, it is apparent from the above discussion that the χ^2 distribution cannot be used to assess significance for these statistics. There have been two basic approaches presented in the literature for obtaining an appropriate reference distribution for D and X^2 when the assumption of large m_j is not reasonable.

Tests based on the conditional distributions of X^2 and D given $\widehat{\beta}$

McCullagh (1985, 1986) argued that goodness of fit should be assessed using the conditional distribution of the statistic rather than its marginal distribution in the case where the observed data are extensive but sparse (i.e., large n, small m_i). McCullagh proposes standardizing X^2 or D by their conditional asymptotic moments given the parameter estimates

$\widehat{\boldsymbol{\beta}}$. Statistical significance is then assessed using the standard normal distribution as the reference distribution for the standardized statistic.

McCullagh (1985) obtained approximations to the first three moments of the unconditional and conditional distributions of the Pearson X^2-statistic for canonical link regression models models which, as demonstrated in Section 1.2.3, includes the logistic regression model. By conditioning on a sufficient statistic of the parameter estimates, the dependence upon $\widehat{\boldsymbol{\beta}}$ is removed from X^2. Consequently, this method accounts for the fact that the parameters from the logistic regression model have been estimated rather than fixed in advance. Approximate formulae for the conditional mean and variance of X^2 for logistic regression models can be found in McCullagh and Nelder (1989),

$$E(X^2|\widehat{\boldsymbol{\beta}}) \simeq (n-p-1) - \frac{1}{2}\sum_{i=1}^{n}(1-6\widehat{v}_i)\widehat{V}_{ii} + \frac{1}{2}\sum_{ij} m_i\widehat{v}_i(1-2\widehat{\pi}_j)\widehat{V}_{ii}\widehat{V}_{ij}(1-2\widehat{\pi}_j), \quad (2.6)$$

and

$$\mathrm{var}(X^2|\widehat{\boldsymbol{\beta}}) \simeq (n-p-1) - \frac{1}{2}\sum_{i=1}^{n}\left\{2n + n\widehat{\rho}_4 - \sum_{ij}(1-2\widehat{\pi}_i)(1-2\widehat{\pi}_j)\widehat{V}_{ij}\right\}, \quad (2.7)$$

where V_{ij} are the elements of $\mathbf{V} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$, the approximate covariance matrix of $\widehat{\boldsymbol{\eta}}$ and $\rho_4 = n^{-1}\sum[\kappa_4^i/(\kappa_2^i)^2]$ with κ_2^i, κ_4^i being the second and fourth cumulants of Y_i, respectively. Although similar results could be derived for D, they are too complex for practical use (McCullagh, 1986; McCullagh and Nelder, 1989).

Tests based on the marginal distributions of X^2 and D

Osius and Rojek (1992) derived tests are based on first-order normal approximations of the power-divergence statistics of Cressie and Read (1984). This class of tests, denoted

SD_λ, is indexed by a real number $\lambda \in \mathbb{R}$ and includes both X^2 and D as special cases corresponding to $\lambda = 1, 0$, respectively. A statistical test can be performed by standardizing SD_λ by estimated values of the large sample approximations of the mean and variance for SD_λ and comparing the resulting value to the standard normal distribution. Osius and Rojek (1992) derived asymptotic moments for SD_λ in logistic regression under sparseness assumptions ($N, M \to \infty$ where $M = \sum_i^N m_i$), however, moments in closed form can only be calculated for $\lambda = 1$, that is, X^2. For strictly binary data, Osius and Rojek show that the conditional and unconditional moments are asymptotically equivalent, at least to first order. Thus, one would expect similar conclusions to be reached in using Osius and Rojek's and McCullagh's tests (see Section 2.3.1).

For logistic regression, Osius and Rojek's moment approximations yield the following estimator of the mean for which no calculation is necessary

$$\widehat{E}(X^2) = J(n). \tag{2.8}$$

The variance may be estimated by

$$\widehat{\mathrm{var}}(X^2) = RSS \tag{2.9}$$

which is the residual sum of squares of an ordinary weighted linear regression of the variable $c_j = (1 - 2\widehat{\pi}_j)/v_j$, $j = 1, \ldots, J(n)$ on the covariates with weights $v_j = m_j\widehat{\pi}_j(1 - \widehat{\pi}_j)$, $j = 1, \ldots, J(n)$. Hosmer et al. (1997) found that, for small samples of strictly binary response data, better distributional results can be obtained if an estimate of the conditional mean and variance obtained by McCullagh (see (2.6) and (2.6)) were used instead of (2.8) and (2.9). This finding makes sense given earlier comments regarding the relationship between conditional and unconditional moments for binary data (i.e., the magnitude of error of the first order approximation should increase with smaller samples).

2.3.2 Modified Test Statistics

Farrington test

Revisiting the conditioning principle cited in Section 2.3.1, Farrington (1996) extended the results of McCullagh (1985) to models with non-canonical links. However, rather than using the Pearson statistic, Farrington used an estimating equations approach proposed in Moore (1986) to obtain a modification to Pearson's statistic by the addition of a first-order component. The statistic (expressed in terms suitable for logistic regression)

$$X_F^2 = X^2 + \sum_{j=1}^{J(n)} \frac{-(1-2\widehat{\pi}_j)}{m_j\widehat{\pi}_j(1-\widehat{\pi}_j)}(y_j^* - m_j\widehat{\pi}_j) \tag{2.10}$$

is shown to have minimum variance within the family considered, where X^2 is the Pearson statistic discussed earlier in this chapter. Significance can be assessed using the standard normal distribution with the standardized statistic. The standardized statistic can be obtained using approximate moments for X_F^2 which can be calculated in closed form.

X_F^2 is shown to induce local orthogonality with the regression parameters (Farrington, 1996). That is, the Farrington statistic removes the dependence upon $\widehat{\boldsymbol{\beta}}$ from the distribution of X^2, which produces substantial simplifications of the moment approximations and increased power. Consequently, Farrington's statistic can be considered as an improvement of the McCullagh method. However, the Farrington test has the structural deficiency that when $m_i \equiv 1$, then $X_F^2 \approx N$. In this case, the test will never reject the null hypothesis of a good fit.

Extensions of Farrington's method have been considered by Paul and Deng (2000) who examined analogous modifications to the deviance statistic and Paul and Deng (2002) who introduced a score statistic inspired by Farrington (1996). We will not pursue these tests further since they were not studied in any of the comparison articles cited in the

Overview of this chapter: Hosmer et al. (1997), Hosmer and Hjort (2002), or Kuss (2002).

Copas unweighted sum of squares test

Copas (1989) has proposed using the unweighted residual sum-of-squares to test equality of proportions in a $2 \times C$ contingency table. Hosmer et al. (1997) have studied a modified version of this statistic in order to assess the adequacy of logistic regression model. In the context of binary response regression, the statistic is written as

$$S = \sum_{j=1}^{J(n)} (y_j^* - m_j \widehat{\pi}_j)^2 \tag{2.11}$$

where $\widehat{\pi}_j$ is the predicted response probability for the jth covariate pattern. Copas argued that, for large samples, significance for his test may be assessed using a chi-square distribution. Consequently, Hosmer et al. (1997) use a chi-square distribution as the reference distribution for the modified statistic.

Note the similarity between S and the Pearson statistic for binary response data, X^2. S is the sum of squared residual values which *resemble* the "response" residuals discussed earlier in this chapter, while X^2 is the sum of squared Pearson residuals. On the surface this may not appear to be a profound difference, however, Copas (1989) explains that by dropping the denominator in each component of the sum, less weight is given to covariate patterns for which the value of m_i is small.

Hosmer et al. (1997) and Kuss (2002) studied this test in the context of logistic regression. In this case Hosmer et al. (1997) argue that statistical significance can be assessed using the following $z-$statistic:

$$z = \frac{S - \text{trace}(V)}{\sqrt{Var[S - \text{trace}(V)]}}, \tag{2.12}$$

which has a large sample standard normal distribution. In (2.12) $\mathrm{trace}(V)$ is a large sample approximation of the mean of S. Hosmer et al. (1997) derived the following approximations for the asymptotic moments for their version of S:

$$E[S - \mathrm{trace}(V)] \approx 0,$$

$$Var[S - \mathrm{trace}(V)] \approx d'(I - M)Vd,$$

in which $d_i = 1 - 2\widehat{\pi}_i$, $i = 1, \ldots, n$; V is given by $V = \mathrm{diag}[\widehat{v}_i : i = 1, \ldots, n]$ where $\widehat{v}_i = \widehat{\pi}_i(1 - \widehat{\pi}_i)$ and $M = VX(X^{\mathrm{T}}VX)^{-1}X^{\mathrm{T}}$. In practice, an estimate of the variance can be obtained from the residual sum-of-squares from the regression of $\widehat{\boldsymbol{d}}$ on X with weights $\widehat{V}$.

2.3.3 Tests Based on Grouping Observations

Hosmer-Lemeshow tests

Hosmer and Lemeshow (1980) devised a chi-square-inspired test of fit for binary response models which imposes a grouping strategy to create the conditions which permit use of the standard large sample theory discussed at the beginning of this section. More precisely, their approach is to aggregate the n observations into a fixed number of groups, g, which effectively produces $2 \times g$ contingency table. Then a Pearson-like statistic that compares the observed and expected cell frequencies of the resulting table can be calculated using

$$X^2_{HL} = \sum_{l=1}^{g} \frac{(o_l - n_l\bar{\pi}_l)^2}{n_l\bar{\pi}_l(1 - \bar{\pi}_l)}. \tag{2.13}$$

In the above formula n_l denotes the number of observations in the lth group, o_l is the

number of successes in the lth group and $\bar{\pi}_l$ is the average probability

$$\bar{\pi}_l = \frac{1}{n_l} \sum_{k=1}^{m(n_l)} m_k \widehat{p}_k, \tag{2.14}$$

where $m(n_l)$ is the number of unique patterns in the lth group.

The rationale for this approach is borrowed from goodness-of-fit literature for sparse contingency tables (Hosmer and Lemeshow, 1980). This logic permits them to conjecture that X^2_{HL} will have a chi-square distribution since $n_l \to \infty$ as $n \to \infty$ while g is taken to be fixed. While they do not justify this claim rigorously, Hosmer and Lemeshow (1980) use simulation results to argue that the distribution of X^2 is roughly approximated by chi-squared with $df = g - 2$ (Hosmer and Lemeshow, 1980).

Some comments clarifying the construction of the g groups are in order. To this end, it is useful to view the data in terms of $2 \times J(n)$ contingency table. The two rows correspond to the values possible values of the binary outcome variable y and the $J(n)$ columns correspond to the assumed number of distinct values observed for the covariates in the model. The observed cell frequencies are the number of "successes" or "failures" recorded for each covariate pattern. Collapsing the columns of the $2 \times J(n)$ table into a $2 \times g$ table to which formulas (2.13) and (2.14) can be applied. The columns of the collapsed table (i.e., the groups) are determined by dividing the sorted predicted probabilities into g partitions and subsequently assigning observations to groups corresponding these partitions.

Hosmer and Lemeshow (1980) proposed two approaches for creating the partitions for the (sorted) predicted probabilities. In the first strategy, observations corresponding to the sorted estimated logistic probabilities are partitioned into g groups with approximately n/g observations in each group. An alternative approach is based on dividing the interval $[0, 1]$ into g fixed subintervals and assigning observations to a group when its corresponding predicted probability falls into that group's subinterval. The first grouping strategy is

generally preferred over the fixed intervals approach because it is possible for the number of observations differ greatly across the groups when fixed subintervals are used. Typically g is taken to be 10 in either strategy (Hosmer and Lemeshow, 2000).

The Hosmer-Lemeshow test is generally regarded as the standard test for assessing the fit of logistic regression models. This is evident in that it has been implemented in all major statistical packages. However, the test has some noteworthy deficiencies that have been revealed in the literature (Hosmer et al., 1997).

In general, test statistics which are constructed from fixed groups, such as the Hosmer-Lemeshow statistic, have been shown to be dependent upon the choice of the groups (Hosmer et al., 1997). According to Bertolini, D'Amico, Nardi, Tinazzi, and Apolone (2000), such problems arise when test is applied to a data set which is not strictly sparse (i.e., ties exist). In this situation it is often unclear to which group a given observation should be assigned. Hence algorithms which calculate the test statistic using different methods for grouping observations will often lead to different conclusions. This fact is noteworthy given that various statistical packages implement the test utilizing different algorithms making it possible for conflicting conclusions to be reached regarding a model's adequacy for a given dataset when multiple software packages are utilized (Pigeon and Heyse, 1999). Hosmer et al. (1997) reported results from fitting the same data set in several statistical packages, obtaining identical values for the estimated parameters but six different values for the p-value of the Hosmer-Lemeshow test ranging from 0.02 to 0.16. An even more alarming discrepancy was reported in Pigeon and Heyse (1999) who found p-values ranging from 0.02 to 0.45 for a single data set.

In addition to the deficiency described above, there have been other concerns about the Hosmer-Lemeshow test raised in the literature. Pigeon and Heyse (1999) reveal problems with the validity of the χ^2-distribution in assessing significance of the Hosmer-Lemeshow statistic which they argue results from the constructing groups based on ranked probabil-

ity estimates. Moreover, le Cessie and van Houwelingen (1991) argue that the Hosmer-Lemeshow strategy of grouping observations based on ranked probabilities produces tests that lack power to detect departures from the model in regions of the '$\mathbf{x}$' space that yield the same estimated probabilities. Pulksenis and Robinson (2002) have proposed a solution to the problem revealed by le Cessie and van Houwelingen using a two-stage modification of the Hosmer-Lemeshow test; however, the conditions required to implement this approach have been criticized as being limited (Kuss, 2002).

Finally, it should be noted that Tsiatis (1980) proposed a chi-square lack-of-fit test for the logistic regression model which differs from the approach developed by Hosmer and Lemeshow in that it is based on partitioning the space of covariates into g distinct, fixed groups. While Tsiatis' method appears to have been cited often by researchers who have applied it in their work, it was not studied in either of the review papers used to guide our discussion. Thus, we will not discuss this method further.

2.3.4 Remarks

It is interesting to note that both the McCullagh and Farrington tests are derived conditionally on sufficient statistics of the parameter estimates. Hence they both account for the additional error resulting from estimating the parameters of the logistic regression model. However, recall that we noted in Section 2.3.1 that Osius and Rojek (1992) demonstrated that conditional moments of the Pearson statistic can be approximated, at least to first order, by the unconditional moments. Thus, it seems reasonable to anticipate similar conclusions would be reached using both tests to assess the fit of a model.

2.4 Smoothing-based Tests of Fit

An alternative to trying to amend the deficiencies with the chi-square tests is to consider tests based on nonparametric smoothers. The rationale for pursuing tests based on nonparametric smoothing techniques was discussed in Section 1.4.2. These approaches include both tests based on smoothed residuals and tests which compare a nonparametric estimate to a parametric estimate. Tests of the latter type have been developed for logistic regression in Xiang and Wahba (1995) and for generalized linear models in Azzalini, Bowman, and Härdle (1989). The approach taken in Xiang and Wahba (1995) uses symmetrized Kullback-Leibler distance between smoothing spline and parametric estimates of the model, while Azzalini et al. (1989) base their approach on comparing the parametric and nonparametric estimates using a pseudo-likelihood ratio test statistic. However, we will focus our discussion on the smoothed residuals method for binary response models introduced in le Cessie and van Houwelingen (1991) since it was studied in the review paper of Hosmer et al. (1997). Furthermore, a residual smoothing method would generally be preferable for reasons cited in Section 1.4.2.

In addition to le Cessie and van Houwelingen's test cited above, we will also discuss a test proposed in Royston (1992) which was also studied in Hosmer et al. (1997) and is based on residual cusums. We find it appropriate to include a discussion of these methods in our section smoothing-based tests of fit because Eubank and Hart (1993) demonstrated that cusum-based tests are special cases of a wider class of tests based on nonparametric function estimation ideas. However, cusum-based tests differ from the residual smoothing approach of le Cessie and van Houwelingen (1991) in that cusum-based tests do not require specification of a smoothing parameter. Tests with this feature are often called "nonadaptive". We will discuss the distinction between adaptive and nonadaptive tests further in the next chapter.

2.4.1 Kernel-smoothed Residual Tests

le Cessie and van Houwelingen (1991) proposed testing the fit of a binary response model by applying the procedure of Nadaraya (1964) and Watson (1964) to the standardized residuals as follows

$$\widehat{r}_S(\mathbf{x}) = \frac{\sum_{j=1}^{n} \widehat{r}_j K_p[\mathbf{H}^{-1}(\mathbf{x} - \mathbf{x}_j)]}{\sum_{j=1}^{n} K_p[\mathbf{H}^{-1}(\mathbf{x} - \mathbf{x}_j)]}, \tag{2.15}$$

where $\widehat{r}_j$ denotes the jth standardized residual (see Section 2.3), $\mathbf{H} = \mathrm{diag}[h_1, \ldots, h_p]$ is a diagonal matrix of bandwidths, and $K_p(\cdot)$ denotes the multiplicative kernel

$$K_p(\mathbf{x}) = \prod_{l=1}^{p} K(x_l), \tag{2.16}$$

in which K is a symmetric, nonnegative, univariate kernel function with finite support $[-a, a]$ satisfying $\int_{-a}^{a} K(x)dx = 0$ and $\int_{-a}^{a} K^2(x)dx = 1$. Note that $K_p(\cdot)$ is defined so that the same univariate kernel is applied to each covariate; however, covariate-specific bandwidths, h_l are used. The bandwidth parameter controls the degree of smoothing and, in general, depends on the kernel, the sample size, n, and the number of covariates as well as the unknown model (Hart, 1997). To reduce the number of bandwidth parameters to 1 le Cessie and van Houwelingen (1991) define $h_l = h_n s_l$, where h_n is a global bandwidth parameter depending on n and s_l is the standard deviation of the lth covariate. le Cessie and van Houwelingen (1991) recommend choosing h_l so that roughly $\sqrt{n}$ observations contribute to the calculation of each r_S.

As we discussed in Section 1.4.2, the rationale for examining smoothed standardized residuals for insight into model fit is motivated by the recognition that under the null hypothesis of correct model specification, the smoothed standardized residuals $\widehat{r}_S(\mathbf{x}_i)$,

$i = 1, \ldots, n$, can each be considered an estimator of zero. This observation motivates the test statistic proposed by le Cessie and van Houwelingen (1991) which is given by the formula

$$T = \frac{1}{n} \sum_{i=1}^{n} \widehat{r}_S^2(\mathbf{x}_i) w(\mathbf{x}_i) \tag{2.17}$$

where

$$w(\mathbf{x}_i) = \frac{\{\sum_{j=1}^{n} K_p[\mathbf{H}^{-1}(\mathbf{x}_i - \mathbf{x}_j)]\}^2}{\sum_{j=1}^{n} K_p^2[\mathbf{H}^{-1}(\mathbf{x}_i - \mathbf{x}_j)]}. \tag{2.18}$$

The test resulting from T is reasoned to circumvent problems cited for the Hosmer-Lemeshow test and they have been found to have better (though not uniformly better) power properties than several other commonly used tests for logistic regression (Hosmer et al., 1997).

le Cessie and van Houwelingen demonstrate that for large samples, the distribution of T can be approximated by $b\chi_v^2$, where χ_v^2 is a chi-square random variable with v degrees of freedom and b a constant. The values of v and c are determined by $b = 2\widehat{E}(T)/\widehat{\mathrm{var}}(T)$, $v = 2\widehat{E}^2(T)/\widehat{\mathrm{var}}(T)$ and $\widehat{E}(T)$ and $\widehat{\mathrm{var}}(T)$ are the estimated mean and variance of T. Clearly, evaluation of significance for le Cessie and van Houwelingen's test requires a means to obtain estimates of moments of T.

Hosmer et al. (1997) provide simplified approximations of $E(T)$ and $\mathrm{var}(T)$ for the special case of logistic regression. Hosmer et al. obtained these approximations by means of a first-order approximation of T, $T \cong \mathbf{e}^{\mathrm{T}} A_r \mathbf{e}$, where $\mathbf{e}$ is the column vector of residuals, $A_r = (I - M)^{\mathrm{T}} Q_r (I - M)$, $Q_r = V^{-1/2}(W^{\mathrm{T}} D_r^{-1} W)V^{-1/2}$, $M = VX(X^{\mathrm{T}}VX)^{-1}X^{\mathrm{T}}$ is the logistic regression version of the hat matrix, W is the $n \times n$ matrix of weights whose (i, l)th element is $w_{i,l} = K_m(x_i - x_l)$, D_r is an $n \times n$ diagonal matrix that contains the diagonal elements of the matrix WW^{T}, and V defined as in 2.29. Well-known results for

moments of quadratic forms, Seber (1977), yield that

$$E(T) = \mathrm{trace}(A_r V) \tag{2.19}$$

and

$$\mathrm{var}(T) = \sum_{i=1}^{n} a_{rii}^2 v_i (1 - 6v_i) + 2\mathrm{trace}(A_r V A_r V). \tag{2.20}$$

Hosmer et al. (1997) reported problems with calculating $\mathrm{var}(T)$ in their simulations, so they used a more computationally efficient approximation of $\mathrm{var}(T)$ which was presented in le Cessie and van Houwelingen (1991)

$$\widehat{\mathrm{var}}(T) \cong 2\left(\frac{2}{3}\right)^p \frac{\mathrm{trace}(WW^{\mathrm{T}})}{n^2} \tag{2.21}$$

for the $\widehat{\mathrm{var}}(T)$. This approximation leads to a reduction of the order of computation to evaluate the matrix A_r from n^4 to n^2.

2.4.2 Residual Cusum Tests

Several contributions have been made to literature on tests of fit that utilize a cumulative sum (cusum) of residuals from the estimated model. The motivation for pursuing tests based on cusum processes is that if the fitted model is the correct model, then partial sums should vary in an unsystematic manner about zero as the index of the process varies.

Su and Wei (1991), Beran and Millar (1992), and Royston (1992) have proposed lack-of-fit tests based on cusums with applications to logistic regression. More recently Stute and Zhu (2002) introduced a residual cusum-based test statistic to assess the validity of a generalized linear model which was inspired by a collection of work Stute developed with various colleagues: Stute (1997), Stute, Gonzalez Mantiega, and Presedo Quindimil

(1998), and Stute, Thies, and Zhu (1998). We will limit our attention to the method of Royston (1992) which was studied in Hosmer et al. (1997). Hosmer et al. cite the convenient large sample results presented for the statistics studied in Royston (1992) as being advantageous over competing methods. By contrast, Su and Wei (1991) and Beran and Millar (1992) required computationally intensive bootstrap procedures to implement their proposals.

Royston (1992) proposed two statistics designed to detect monotonic and quadratic departures from linearity in the logit. Both statistics are based on the cumulative sum of residuals

$$q_l = -\sum_{i=1}^{l}(y_{(i)} - \widehat{\pi}_{(i)}) \tag{2.22}$$

where $\widehat{\pi}_{(i)}$ is the ith largest estimated logistic probability and $y_{(i)}$ is the associated value of the outcome variable. Royston (1992) assumed that the observations had been sorted according to a specific covariate of interest, while Hosmer et al. (1997) studied Royston's test by sorting according to estimated probabilities (note that Royston's original assumptions are consistent with typical assumptions in the related literature). The statistic for detecting monotone departures is

$$C_1 = \max_{1\leq l\leq n} |q_l|. \tag{2.23}$$

The statistic for detecting quadratic departures is

$$C_2 = \max_{1\leq l\leq n/2} |q_l - q_{n-l}|. \tag{2.24}$$

The monotone test is a special case of Su and Wei's test in the case of a single covariate,

while both monotone and the quadratic tests are a special case of the test statistics derived in Beran and Miller.

Royston presented his method primarily as a means to graphically determine whether or not a fitted model adequately represents the relationship between the predicted probability and a single covariate of interest (this is why Royston assumed observations ordered according to the covariate of interest). Royston did not specifically advocate the use of these statistics as global tests of fit, however as we stated above, he provides easily computed transformations of the two statistics that allow calculation of p-values using the standard normal distribution. Furthermore, Hosmer et al. (1997) argue that since the tests are designed to be sensitive to monotonic and quadratic departures in the logit, Royston's statistic seems potentially beneficial.

Inlow (2001) has criticized the large sample approximations proposed for these statistics. In particular, he has argued that Royston's formulas do not take into account whether or not the model is specified *a priori* or estimated from the data. Moreover, simulation studies ultimately revealed some power deficiencies of this test (Hosmer et al., 1997).

2.5 The Information Matrix Specification Test

A general specification test that has been used to assess the adequacy of the logistic regression model is the information matrix (IM) test. This test was originally proposed by White (1982) for general testing of likelihood specification. Lechner (1991), Thomas (1993) and Aparicio and Villanua (2001) have all considered tests based on a special case of White's statistic for binary response models which was first presented in Orme (1988). The IM test is based on the well-known information-matrix equivalence theorem which essentially states that when the model is correctly specified, the following expression holds

$$-E\left(\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\right) = E\left(\frac{\partial l}{\partial\boldsymbol{\beta}}\frac{\partial l}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\right) \tag{2.25}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients and l is the log-likelihood for the logistic regression model (see Sections 1.2.3 and 1.3). In words, (2.25) states that the information matrix can be expressed as either the expected value of the Hessian of the likelihood or the expected value of the outer product of first-order partial derivatives of the likelihood. A test statistic is formed by comparing the elements of the two different estimators of the information matrix obtained by utilizing each of these expressions. These two estimators should give comparable results under a satisfactory model fit. The IM test has been criticized for being difficult to compute in practice (Hosmer and Lemeshow, 2000); however, it has been shown to possess reasonable power even with strictly sparse data (Kuss, 2002).

Kuss (2002) presents an explicit expression of the IM statistic for logistic regression models which evaluates the difference of the diagonal elements of the two estimators results in the $((p+1)\times 1)$-vector

$$\widehat{d} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\pi}_i)(1 - 2\widehat{\pi}_i)z_i \tag{2.26}$$

with $z_i = (1, x_{i1}^2, \ldots, x_{ip}^2)^{\mathrm{T}}$ where the components of $\widehat{d}$ sum to 0 in the case of a good model fit. After standardization with an appropriate variance, the test statistic can be compared to a χ^2_{p+1}-distribution. Note that the IM-test is calculated for the individual and not for the grouped observations so we do not expect problems with sparse data.

2.6 Goodness of Link Tests

As mentioned in Section 1.2.2, a link function must be specified in order to completely specify a GLM. For binary response data, the logistic link function is the most commonly used; however, there are alternatives to the logistic link and using different links can have a profound impact on the specification of the linear predictor (Collett, 1991). That is, the choice of link function and the structure of the predictor function are interdependent. Consequently, examination of the adequacy of a given link function has to be made on the basis of a final model. In this section we will discuss some selected approaches for assessing the adequacy of the logistic link function.

2.6.1 Tests Based on Parametric Families of Link Functions

The goodness-of-link test differs from the tests of fit described above in that it utilizes a parametric family to assess the adequacy of the specified link; that is, there is a pre-specified family of alternatives against which the fitted model is tested. Pregibon (1980) proposed a general approach for testing adequacy of link specification based on a parametric family of link functions. Several families of link functions have been proposed in the literature and are typically formulated as parametric generalizations of the logit and probit models. Generally, these families have been proposed in order to more adequately model binary response data. However, these families have proven useful in detecting possible inadequacy of logit and probit models. We will now discuss an approach for using families of link functions to test the adequacy of a specified link which was introduced by Pregibon (1980).

Let g denote the correct, though unknown link function. As an alternative to the logistic link, consider a function $g(\pi;\boldsymbol{\alpha})$ with $\boldsymbol{\alpha} \in \mathcal{A} \subset \mathbb{R}^d$ such that $g(\pi;\boldsymbol{\alpha}_0) = \log\left(\frac{\pi}{1-\pi}\right)$ for some $\boldsymbol{\alpha}_0 \in \mathcal{A}$. It is assumed that $g \in \mathcal{G}\{g(\pi;\boldsymbol{\alpha}) : \boldsymbol{\alpha} \in \mathcal{A}\}$. Thus, an estimate of g can be obtained by means of a maximum likelihood estimate of $\boldsymbol{\alpha}$.

In order to utilize the collection of link functions defined by $g(\pi;\boldsymbol{\alpha})$ to detect departures from the logistic link, $g(\pi;\boldsymbol{\alpha}_0)$, Pregibon (1980) proposed a first-order Taylor series expansion of $g(\pi;\boldsymbol{\alpha})$ about $\boldsymbol{\alpha}_0$

$$g(\pi;\boldsymbol{\alpha}) \approx g(\pi;\boldsymbol{\alpha}_0) + (\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T \frac{\partial g(\pi;\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\bigg|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_0}. \tag{2.27}$$

Let $\boldsymbol{\alpha}^* \in \mathcal{A}$ be the value which yields the true model; that is, $g = g(\pi;\boldsymbol{\alpha}^*) = X\boldsymbol{\beta}$. Then assuming that $\boldsymbol{\alpha}^*$ and $\boldsymbol{\alpha}_0$ are sufficiently close, then g can be approximated by $g(\pi;\boldsymbol{\alpha}_0)$ so that

$$\begin{aligned} g(\pi;\boldsymbol{\alpha}_0) &= g(\pi;\boldsymbol{\alpha}^*) + [g(\pi;\boldsymbol{\alpha}_0) - g(\pi;\boldsymbol{\alpha}^*)] \\ &= X\boldsymbol{\beta} + \boldsymbol{\gamma}\mathbf{z} \end{aligned} \tag{2.28}$$

where $\boldsymbol{\gamma} = \boldsymbol{\alpha}^* - \boldsymbol{\alpha}_0$ and $\mathbf{z} = \partial g(\pi;\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_0}$.

The variable $\mathbf{z}$ is often referred to as a "constructed variable" and can be viewed as compensating for the departure of the logistic link function hypothesized in the null model from the true link function. Thus, a test of the hypothesis $H_0 : \gamma = 0$ in equation (2.28) provides a test of the adequacy of the logistic link function. In principle, this hypothesis can be tested by means of a likelihood ratio, score, or Wald test. Note that $\mathbf{z}$ depends on π, which are unknown. Consequently, in practice the fitted response probabilities obtained from fitting a logistic regression model are used to construct $\mathbf{z}_i$'s corresponding to each observation. Of course, the procedure described above can be modified in order to test the adequacy of other link functions, such as the probit link.

2.6.2 Stukel Generalized Logistic Link Function

The family of link functions presented in Stukel (1988) has gained the widest acceptance for testing adequacy of the logistic link function. Furthermore, the test resulting from applying the method of Pregibon to Stukel's link family was recommended by Hosmer et al. (1997) because of its superior power relative to other tests considered. The Stukel model extends the standard logit link function with two additional parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ and is defined in terms of the CDF, as follows:

$$F_S(\eta; \alpha) = \frac{e^{h_\alpha(\eta)}}{1 + e^{h_\alpha(\eta)}} \tag{2.29}$$

in which, for $\eta \leq 0$, h_α is defined as

$$h_\alpha(\eta) = \begin{cases} \alpha_1^{-1}(\exp(\alpha_1|\eta|) - 1), & \alpha_1 > 0 \\ \eta & \alpha_1 = 0 \\ -\alpha_1^{-1}\log(1 - \alpha_1|\eta|), & \alpha_1 < 0. \end{cases} \tag{2.30}$$

For $\eta \leq 0$, h_α is defined as

$$h_\alpha(\eta) = \begin{cases} -\alpha_2^{-1}(\exp(\alpha_2|\eta|) - 1), & \alpha_2 > 0 \\ \eta & \alpha_2 = 0 \\ \alpha_2^{-1}\log(1 - \alpha_2|\eta|), & \alpha_2 < 0. \end{cases} \tag{2.31}$$

The parameters α_1 and α_2 control both the symmetry and tail weight of the generalized link function. If $\alpha_1 = \alpha_2$, the corresponding probability curve $F_S(\eta; \alpha)$ is symmetric. Tail weight is dictated by the particular values of α_1 and α_2. For instance, the generalized logistic link function approximates the probit link when $\alpha_1 = \alpha_2 \approx 0.165$ while this link

Table 1. Coefficients for Stukel's generalized logistic model.

Link	α_1	α_2
logistic	0	0
probit	0.165	0.165
complimentary log-log	0.620	-0.037
Laplace	-0.077	-0.077

function reduces to the usual linear logistic model when $\alpha_1 = 0$ and $\alpha_2 = 0$. Table 1 lists several well-known link functions approximated by $h_\alpha(\cdot)$ along with the corresponding values of α_1 and α_2 that yield the approximation.

Applying the approach described in Section 2.6.1, a two degree-of-freedom test of the hypothesis that both parameters are equal to zero can be obtained. Recall that the test can be implemented by means of a score, Wald or likelihood ratio test of the coefficients corresponding to the constructed variables resulting from a Taylor's expansion. For the Stukel link function these constructed variables can be written as $z_1 = \frac{1}{2}\widehat{\eta}^2 I(\widehat{\eta} \geq 0)$ and $z_2 = -\frac{1}{2}\widehat{\eta}^2 I(\widehat{\eta} < 0)$, $\widehat{\eta} = \mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$, where $I(\cdot)$ is the usual indicator function.

Alternative general families of link functions to that given by Stukel have been suggested by Pregibon (1980), Prentice (1976), Aranda-Ordaz (1981), and Guerrero and Johnson (1982). Among the more popular of these alternatives is the link family introduced by Prentice for which Brown (1982) developed a different two-parameter score test. While the Prentice model offers the same level of flexibility in modeling departures from the logistic link as the Stukel family, the test utilizing the Stukel family is more direct and easier to implement (Stukel, 1988).

2.7 Discussion

It should be noted that despite the superior power properties cited for some of the tests discussed in this chapter, none of these tests had uniformly good power against all departures

from the null model considered in Hosmer et al. (1997) and Kuss (2002). That is, all of the tests studied in these two papers have weaknesses and power properties reported therein.

Recognizing the apparent absence of a test with desirable power properties against a wide variety of alternatives, several authors who have written on the subject of evaluating the fit of a logistic regression model ultimately conclude that is advisable to use a combination of residual analysis, diagnostic measures and *multiple* tests. In other words, one should avoid the temptation to naively accept results from a single model assessment. See, for instance, McCullagh and Nelder (1989), Hosmer and Lemeshow (2000) and Agresti (2002).

Two of the more authoritative sources on the theory and application of the generalized linear model have offered a particularly intriguing recommendation for evaluating the fit of a logistic regression model. After noting the deficiencies of D or X^2 as well as their modifications (see Section 2.3), Agresti (2002, p.177) and McCullagh and Nelder (1989, p.122) discussed alternative approaches to evaluating the fit of a logistic regression model which can be used to supplement assessments determined by means of global tests of fit such as the ones discussed in this chapter. They point out that lack-of-fit can be detected by comparing the working model with more complex models in which nonlinear effects (such as quadratic terms) or interactions are added to the working model and observing the reduction in deviance. If a more complex model does not result in a better fit of the data, then we have assurance that working model is reasonable.

Agresti (2002) and McCullagh and Nelder (1989) both advocate examining complex models which reflect the scientific context of the model. While this advice sounds appealing, unfortunately, it is often the case that there is not a clear scientific motivation for additional terms. Furthermore, it is very plausible that deficiencies in the fit of the model cannot be explained by scientific reasoning. In the absence of scientifically relevant additions to the working model, one is relegated to examining arbitrary departures from the

model or simply accepting the working model without any further scrutiny.

To conclude, we view the lack of a uniformly powerful test of fit for the logistic regression model as an indication that there is still a need for additional research in the testing literature. Furthermore, we present the sentiments conveyed in the recommendations of Agresti (2002) and McCullagh and Nelder (1989) as evidence that our proposed direction of research is well-founded and provides an intuitively desirable means of evaluating the fit of a logistic regression model. More specifically, in the next chapter we discuss a systematic, unambiguous way of considering general departures from a working model. This approach is based on well-developed theoretical principles and ultimately provides direction toward a new test of fit that can be applied to the logistic regression model.

CHAPTER III

SERIES-BASED LACK-OF-FIT TESTS

3.1 Overview

In the previous chapter we presented a review of some widely-used methods for testing the fit of a logistic regression model. In this chapter we will discuss the existing literature on lack-of-fit tests which makes use of orthogonal series to detect departures from a parametric model. Though we will a focus on applications to GLMs and closely related problems, it should be noted that the principles upon which these methods are based extend beyond the generalized linear model setting. These concepts have been applied to a variety of modeling scenarios such as spectral analysis and testing the goodness of fit of a probability distribution (Aerts, Claeskens, and Hart, 1999).

The literature on series-based methods on testing the fit of probability models is vast to say the least, and hence, our review will not be exhaustive. Rather, our survey of the literature on series-based lack-of-fit tests will be focussed on a relatively small number of references. However, given that our proposed method has been inspired by this literature, we will discuss these methods in greater detail. Thus, we intend to impart an understanding of how these tests work as well as document the benefits and drawbacks associated with the various tests. In doing so, we will present a collection of terminology and concepts which we believe the reader may find useful in our subsequent discussion. Ultimately, we will to reveal a new direction for research as well as motivation for our pursuit of this direction.

The rest of this chapter will be organized as follows. In Section 3.2, we will discuss some basic concepts which are applicable to all of the tests studied in this chapter. In Section 3.3, we will discuss series-based tests of fit for generalized linear models. In Section 3.4, we will discuss a recently developed method which makes use of the Laplace approxi-

mation in the derivation of the test statistic. In Section 3.6, we conclude the chapter with a summary of progress made in series-based lack-of-fit tests and present a new direction for research.

3.2 Background and Fundamental Concepts

3.2.1 Model Assumptions

In this chapter we will limit our attention to the class of generalized linear models which has been studied in relevant literature. In particular, we will focus on the canonical link regression model described in Section 1.2.3. However, as we explained in Section 1.2.3 this covers arguably the most important models within the larger class of GLMs, including the logistic regression model.

Recall from Section 1.2 that we assume the data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are observed where, for $i = 1, \ldots, n$, $\mathbf{x}_i$ is a fixed vector of covariates and y_i is a scalar response for the ith subject. In light of the definitions given in Section 1.2, the response distribution for the canonical link regression model can be expressed directly in terms of the covariates as follows

$$f(y; \eta(\mathbf{x}_i), \boldsymbol{\psi}) = \exp\{[y\eta(\mathbf{x}_i) - b(\eta(\mathbf{x}_i))]/a(\psi) + c(y, \psi)\}, \tag{3.1}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions, $\eta(\cdot)$ is an unknown function and ψ an unknown, real-valued dispersion parameter (i.e., $k = 1$).

3.2.2 Inference Problem

Some comments regarding series-based tests of fit for canonical link models are in order. Recall that in Section 1.2.2 construction of a generalized linear model requires specifying three components. In general, specification of any one of these components of the model

may influence the other two components. Hence, given the limited flexibility in the specification of the components of a canonical link regression model (see Section 1.2.2), the null hypothesis for a global test of the fit of a canonical link regression model can be stated directly as follows

$$H_0 : \eta(\mathbf{x}) = \sum_{j=1}^{p} \beta_j \gamma_j(\mathbf{x}) \equiv \eta(\mathbf{x}; \boldsymbol{\beta}), \quad \forall \mathbf{x}, \tag{3.2}$$

where, as described in Chapter I, $\eta(\mathbf{x}; \boldsymbol{\beta})$ is a parametric model proposed for $\eta(\cdot)$ in which $\gamma_1, \ldots, \gamma_p$ are known functions and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ an unknown parameter vector. Recall that an intercept term can be accommodated by defining $\gamma_1(\mathbf{x}) \equiv 1, \ \forall \mathbf{x}$. A useful and widely studied special case of (3.2) for which $\eta(\mathbf{x}; \boldsymbol{\beta}) = \beta_1$ for all $\mathbf{x}$ is referred to as the "no-effect" hypothesis. The resulting hypothesis test is called a "test of no-effect".

3.2.3 Series Expansion

We start by briefly discussing some basic ideas regarding how series representations of functions satisfying general conditions can be utilized to formulate nonparametric estimators with properties which prove useful in testing the fit of a proposed function. In the interest of convenience and clarity, we will assume for the remainder of this section that $\mathbf{x} \in \mathbb{R}$ and use x to denote this continuous, real-valued covariate. Observe that if η and $\gamma_1, \ldots, \gamma_p$ satisfy general conditions, then we can express the unknown regression function $\eta(\cdot)$ in terms of its departure from $\eta(\cdot; \boldsymbol{\beta})$ as

$$\eta(x) = \eta(x; \boldsymbol{\beta}) + \Delta(x), \tag{3.3}$$

where Δ has the series representation

$$\Delta(x) = \sum_{k=0}^{\infty} \phi_k u_k(x), \ \ x \in [a, b], \ \ a, b \in \mathbb{R} \tag{3.4}$$

for constants $\phi_0, \phi_1, \ldots$, where $\{u_1(\cdot), u_2(\cdot), \ldots\}$ is a collection of known functions that are continuous on the range of x and that span a "large" space of functions. It is not required that the u_j's be orthogonal; however, orthogonality is desirable in that it simplifies practical computation and enhances clarity of proofs of theoretical results. In the GLM setting specialized orthogonality conditions are required to attain such simplification (see Section **??**). Furthermore, it is understood that each u_k is not a linear combination $\gamma_1, \ldots, \gamma_p$. We will refer to the u_k's as the basis functions of a series representation of Δ. Popular examples for basis functions include trigonometric functions, wavelets and orthogonal Legendre or Hermite polynomials.

A reasonable approach to approximating the function of interest would be to truncate the sum in (3.4) after the first j terms. The representation specified by (3.3) and (3.4) motivates approximations of $\eta(\cdot)$ constructed by considering only finite contributions to the sum as follows:

$$\begin{aligned} \eta(x; \theta_1, \ldots, \theta_{p+j}) = &\quad \eta(x; \theta_1, \ldots, \theta_p) + \textstyle\sum_{k=1}^{j} \theta_{p+k} u_k(x) \\ = &\quad \eta(x; \boldsymbol{\beta}) + \textstyle\sum_{k=1}^{j} \phi_k u_k(x) =: \eta_j(x), \quad j = 1, 2, \ldots \end{aligned} \tag{3.5}$$

where we define $\theta_k = \beta_k$ for $k = 1, \ldots, p$ and $\theta_k = \phi_{k-p}$ for $k = p+1, \ldots, p+j$. The above formulation produces a sequence $\{\eta(\cdot; \theta_1, \ldots, \theta_{p+j}) : j = 1, 2, \ldots\}$ of approximators of η with the property that $\eta(\cdot; \theta_1, \ldots, \theta_{p+j}) \equiv \eta(\cdot; \theta_1, \ldots, \theta_{p+j}, \mathbf{0})$ for each $j = 1, 2, \ldots$ and all allowable parameter values $\theta_1, \ldots, \theta_{p+j}$. In other words, the models for η are nested in that a model of a given order contains all terms contained in every model of a smaller order and they become increasingly complex as j increases. Furthermore, as $j \to \infty$, functions of the form $\eta(\cdot; \theta_1, \ldots, \theta_{p+j})$ span the space of all functions of interest. For convenience, we will denote the jth alternative specification of η by $\eta_j(x)$. That is, in the context of series-based

alternatives formulated above we have $\eta_j(x) \equiv \eta(x; \theta_1, \ldots, \theta_{p+j})$.

Now observe that the null hypothesis (3.2) is equivalent to

$$H_0 : \phi_1 = \phi_2 = \cdots = 0. \tag{3.6}$$

Thus, an omnibus test of (3.2) can be obtained by using the series-based function approximators to construct alternatives to $\eta(\cdot; \boldsymbol{\beta})$. The maximum likelihood estimator of $\eta_j(x)$ is

$$\widehat{\eta}_j(x) = \quad \eta(x; \widehat{\theta}_1, \ldots, \widehat{\theta}_p) + \textstyle\sum_{k=1}^{j} \widehat{\theta}_{p+k} u_k(x), \quad j = 0, 1, \ldots, K_n, \tag{3.7}$$

where $\widehat{\theta}_k$, $k = 1, \ldots, p+j$ are maximum likelihood estimators with $K_n \leq n$. Note that the finite sum would capture any portion of the residual deviance which is left "unexplained" by the proposed (null) model (assuming without loss of generality that there are no redundant terms contained in the linear predictor and the finite series; if this were the case, the redundant basis function is simply discarded from the collection of basis functions used in the finite series).

The truncated series described above can be viewed as a nonparametric estimator of an unknown regression function. Since the order of the sum varies across the candidate models, the order of the series-based regression estimator plays the role of smoothing parameter. The order of the truncated series estimator is sometimes referred to as a truncation point. In many settings this property is enough to ensure that there exist tests based on the models $\eta_1, \ldots, \eta_{K_n}$ that are consistent against any continuous alternative to H_0, so long as K_n tends to ∞ at an appropriate rate with the sample size (Aerts et al., 1999; Aerts, Claeskens, and Hart, 2004).

3.3 Tests of Fit for Generalized Linear Models

Lack-of-fit tests applicable to GLMs have been developed in the collection of papers by Aerts et al. (1999), (2000) and (2004). Aerts et al. (1999) and (2000) approach the problem using the concepts of an order selection-based test, while Aerts et al. (2004) uses a Bayesian rationale to motivate a test statistic formulation which ultimately leads to a statistic that explicitly depends on squared Fourier coefficients in a similar fashion to the cusum or Neyman smooth tests. It is worth emphasizing that, in the context of generalized linear regression, the methods proposed in Aerts et al. (1999), (2000) and (2004) require that the regression model be a member of the subclass of GLMs known as canonical link models described in Sections 1.2.2 and 1.2.3.

All of the tests proposed in this collection of papers can be viewed as generalizations of existing methods for testing the fit of Gaussian-based regression models. A comprehensive discussion of Gaussian-based methodology developed prior to 1997 can be found in Hart (1997). Much of the work we will review in this chapter utilizes model selection criteria and appears to be directly inspired by Eubank and Hart (1992) which developed many of these techniques for Gaussian-based models.

3.3.1 Testing the Fit of a GLM with a Single Regressor via Automated Order Selection

In this case several alternatives are formulated in terms of departures from the proposed null model in which the departure is modeled in terms of finite series approximations described above. The idea underlying the method presented in Aerts et al. (1999) is to use a model selection criterion such as AIC, BIC, etc. to select the "best" model for $\eta(\cdot)$ from the estimated models $\widehat{\eta}_0, \widehat{\eta}_1, \ldots, \widehat{\eta}_{K_n}$. The null model is rejected if it is not selected by the model criterion. Moreover, while it is of primary interest to evaluate the fit of the null model, the approach just described somewhat serendipitously provides an estimate of $\eta(\cdot)$

in the event that the null model is found to be inadequate.

Several different selection criteria are proposed and examined, including a criterion inspired by the Akaike information criterion (AIC) and others based on various score statistics. Aerts et al. (1999) demonstrate that their tests are consistent against essentially any alternative hypothesis. Furthermore, they demonstrate via simulation that their test possesses competitive power properties.

AIC-inspired criteria

In a likelihood context, a popular method of model selection is the AIC. Aerts et al. (1999) define the modified AIC by

$$\mathrm{MAIC}(r; C_n) = \mathcal{L}_r - C_n r, \;\; r = 0, 1, \ldots, R_n, \tag{3.8}$$

where C_n is some constant larger than 1, R_n could be either fixed or tending to infinity with n and $\mathcal{L}_r = 2(l_r - l_0)$, $r = 0, 1, \ldots, R_n$, is the loglikelihood ratio corresponding to the approximator $\eta_r(\cdot)$, in which the loglikelihood $l_r = l(\eta_r, \psi)$ can be written explicitly as follows

$$l(\eta_r, \psi) = \sum_{i=1}^{n} \{[y_i \eta_r(x_i) - b(\eta_r(x_i))]/a(\psi) + c(y_i, \psi)\}, \tag{3.9}$$

Note that the maximizer of AIC and BIC is equal to the maximizer of $\mathrm{MAIC}(r)$ when $C_n = 2$ and $C_n = \log(n)$, respectively. Now let $\widehat{r}_{C_n}$ be the maximizer of $\mathrm{MAIC}(r; C_n)$. A possible test of H_0 against a general alternative is to reject H_0 if the maximizer, $\widehat{r}_{C_n}$, of $\mathrm{MAIC}(r; C_n)$ is larger than 0. By appropriate choice of C_n the asymptotic type I error probability of the test,

$$\text{reject } H_0 \text{ when } \widehat{r}_{C_n} > 0, \tag{3.10}$$

can be any number between 0 and 1. Under certain regularity conditions given in Theorem 1 of Aerts et al. (1999), the limiting level of this test (as $n \to \infty$) is about $.29$ when the AIC penalty constant, $C_n = 2$, is used. Values of C_n yielding other test levels can be obtained following a proposal of Eubank and Hart (1992). For example, a test of asymptotic level $.05$ is obtained by using $C_n = 4.18$. (See Hart 1997, p. 178, for values of C_n leading to other test levels.)

Score-based criteria

The proposed AIC-based tests can be written in terms of the likelihood ratio statistic $\mathcal{L}_r$ for testing hypothesis (3.2) against the alternative that $\eta(\cdot)$ has the form $\eta_r(\cdot)$. The score statistic provides a computationally attractive approximation of the likelihood ratio statistic which only requires fitting the null model. Aerts et al. (1999) identify this feature of the score statistic as being particularly advantageous for application of their method since it is plausible that a large number of alternative models may be required to carry out the test in some circumstances. Aerts et al. (1999) explain that the Wald statistic can also be used as an approximation to the likelihood ratio statistic. However, the authors cite need to obtain "unrestricted" maximum likelihood estimators and the Wald statistic's lack of invariance under equivalent reparameterizations of nonlinear restrictions as being drawbacks to using it as an approximation in their setting.

Analogous to the definition of MAIC, Aerts et al. (1999) and Aerts et al. (2000) define the *score information criteria* (SIC),

$$\mathrm{SIC}(r; C_n) = \mathcal{S}_r - C_n r, \;\; r = 0, 1, \ldots, \tag{3.11}$$

where C_n, R_n are as defined above and $\mathcal{S}_r$ is the score statistic described in Section 1.3.4 applied to the null hypothesis (3.6). For canonical link regression models (Section 1.2.3),

the score statistic can be written as

$$\mathcal{S}_r = \sum_{j=1}^{r} \frac{n\widehat{\phi}_j^2}{a(\widehat{\psi}_0)}, \tag{3.12}$$

where

$$\widehat{\phi}_k = \frac{1}{n}\sum_{i=1}^{n}[y_i - b'(\eta(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}^0))]\widehat{u}_k(\mathbf{x}_i), \;\; k = 1,\ldots,K. \tag{3.13}$$

Aerts et al. (2000) point out that expression (3.12) has essentially the same form as the statistic of Neyman's classical smooth test (Hart, 1997).

As in the modified information criteria discussed in the previous section, an apparently sensible test of (3.2) is one that rejects H_0 when the maximizer, $\widetilde{r}_{C_n}$, of $\mathrm{SIC}(r, C_n)$ is larger than 0. Theorem 3 of Aerts et al. (1999) asserts that under H_0, $\widetilde{r}_{C_n}$ and $\widehat{r}_{C_n}$ have the same limiting distribution.

Tests based on order selection

While one may conduct a test based directly on the order selected by MAIC or SIC, other related statistics have been proposed. We will review several of these statistics in this section. For convenience we will discuss these tests in the context of SIC though, in principle, analogous tests can be constructed for the MAIC.

Aerts et al. (1999) present alternate, equivalent expressions for the test statistics reviewed in the previous sections. For example, observe that the SIC test rejects H_0 if and only if $\mathrm{SIC}(r; C_n) > 0$ for some r in $\{0, 1, \ldots, R_n\}$, which is equivalent to rejecting H_0 when $T_{\mathrm{OS}} > C_n$, with

$$T_{\mathrm{OS}} = \max_{1\le r\le R_n}\{\mathcal{S}_r/r\}. \tag{3.14}$$

Note that C_n acts as both the penalty constant in SIC and the critical value of T_{OS}. Thus,

taking $C_n = 4.18$, the results in a test with limiting size of 0.05 as noted for the order selection test (3.10). This test has been studied in the context of Gaussian response models in Eubank and Hart (1992). Recalling expression (3.12), we can see that (3.14) can be written explicitly in terms of sample Fourier coefficients which closely resembles the well-known data-driven Neyman smooth-type statistic in the context of Gaussian response models (Hart, 1997).

Among the other statistics, Aerts et al. (2000) studied the score statistics corresponding to models chosen by the score analogs of AIC and BIC:

$$S_a = \mathcal{S}_{\widehat{r}_a}$$

where $\widehat{r}_a = \arg\max_{0 \le r \le R_n} \mathrm{SIC}(r; 2)$;

$$S_b = \mathcal{S}_{\widehat{r}_b}$$

where $\widehat{r}_b = \arg\max_{1 \le r \le R_n} \mathrm{SIC}(r; \log(n))$. Observe that $\widehat{r}_b$ maximizes $\mathrm{SIC}(r; \log n)$ over $1, \ldots, R_n$ rather than $0, 1, \ldots, R_n$. This definition accounts for a consistency property of BIC-type order selection criteria (Aerts et al., 2000).

Another statistic studied in Aerts et al. (2000) is a standardized version of S_a

$$T_a = \frac{\mathcal{S}_{\widehat{r}_a} - \widehat{r}_a}{\max(1, \widehat{r}_a^{1/2})}; \tag{3.15}$$

Aerts et al. claim that standardizing S_a greatly stabilizes the null distribution of the statistic, which leads to a meaningful improvement in the power for T_a. It is further claimed that the null distribution of S_b is already quite stable which makes standardization of S_b unnecessary.

Finally, Aerts et al. (2000) considered using the AIC-type score criterion evaluated at its maximum as lack-of-fit statistic

$$T_{\max} = \mathrm{SIC}(\widehat{r}_a; 2). \tag{3.16}$$

Use of this statistic was first considered by Parzen (1977) to test lack-of-fit of time series models.

Aerts et al. (2000) provide large sample approximations for the null distributions of the statistics reviewed in this section. Let $Z_1, Z_2, \ldots$ be a sequence of independent and identically distributed standard normal random variables, define $V_0 = 0$, $V_r = Z_1^2 + Z_2^2 + \cdots + Z_r^2$, for $r = 1, 2, \ldots$, and $\widetilde{r}$ to be the value of r that maximizes $V_r - 2r$ over $r = 0, 1, \ldots$. If the null hypothesis (3.6) and the assumptions of Theorem 1 of Aerts et al. (2000) hold, then the aforementioned theorem ensures that S_a, S_b, T_a, T_{OS}, and $T_{\max}$ converge in distribution to $V_{\widetilde{r}}$, V_1, $(V_{\widetilde{r}} - \widetilde{r})/\max(1, \widetilde{r}^{1/2})$, $\max_{r \geq 1}(V_r/r)$, and $V_{\widetilde{r}} - 2\widetilde{r}$, respectively as $n \to \infty$.

3.3.2 Extension to Multiple Regression

Aerts et al. (2000) extend the proposal of Aerts et al. (1999) described in Section 3.3.1 to multiple regression. Aerts et al. (2000) explain that care must be taken with how one constructs the sequence of alternatives to test the adequacy of $\eta(\cdot; \boldsymbol{\beta})$ in order to ensure that the resulting test will possess desirable power properties. To demonstrate how this works, we follow the example described in Aerts et al. (2000) and consider the case in which η is an unknown function of the covariates x_1 and x_2. In this context, the null hypothesis (3.2) can be written as

$$H_0 : \eta \in \{\eta(\cdot, \cdot; \boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathcal{B}\}. \tag{3.17}$$

In analogy to the case of only one covariate, an alternative model obtained from a series expansion which uses basis functions u_j may be expressed as

$$\eta(x_1, x_2) = \eta(x_1, x_2; \boldsymbol{\beta}) + \sum_{j,k \in \Lambda} \sum \phi_{jk} u_j(x_1) u_k(x_2) \tag{3.18}$$

where Λ is the index set for a given alternative model. It is evident from (3.18) that the index set Λ uniquely determines a given alternative since it specifies the particular subset of basis functions which compose that alternative. Furthermore, the definition of Λ will, in general, depend on the specification of null model. For example, suppose we wish to test the null model $\eta(x_1, x_2; \boldsymbol{\beta}) = \beta_0 + \beta_1 u_1(x_1) + \beta_2 u_2(x_2)$, then it is obvious that neither $u_1(x_1)$ nor $u_2(x_2)$ should be included in the sequence representing the alternative model. In light of this dependence upon the null model, Aerts et al. (2000) limited their discussion to the situation where the function $\eta(x_1, x_2; \boldsymbol{\beta})$ is constant in order to make notation simpler. Under the no-effect null hypothesis, Λ is a subset of $\{(j,k) : 0 \leq j, k < n, j + k > 0\}$.

Aerts et al. (2000) present tests which generalize the score-based model selection criteria in Section 3.3.1 using multivariate alternatives specified by (3.18). However, likelihood-based model selection criteria presented in Section 3.3.1 can be generalized in the same manner. The resulting model selection criteria corresponding to the log-likelihood ratio and score statistics are given by

$$\begin{aligned} MAIC(\Lambda; C_n) &= \mathcal{L}_{\Lambda,n} - C_n N(\Lambda), \\ SIC(\Lambda; C_n) &= \mathcal{S}_{\Lambda,n} - C_n N(\Lambda), \end{aligned} \tag{3.19}$$

respectively, where $N(\Lambda)$ denotes the number of elements in Λ. Critical points and p-values of the lack-of-fit tests can be obtained via asymptotic distribution theory or by use of the bootstrap.

To carry out this test in practice, $\mathrm{SIC}(\Lambda; C_n)$ must be maximized over some collection of subsets $\Lambda_1, \Lambda_2, \ldots, \Lambda_{m_n}$. Aerts et al. (2000) require that this collection of Λ_j's satisfy the following assumptions

1. $\Lambda_1 \subset \Lambda_2 \subset \cdots \subset \Lambda_{m_n}$ and

2. $N(\Lambda_{m_n}) \to \infty$ in such a way that, for each $(j,k) \neq (0,0)$ $(j,k \geq 0)$, (j,k) is in Λ_{m_n} for all n sufficiently large.

The first assumption imposed on the index sets is required so that corresponding models emulate the hierarchical (i.e., nested) fashion in which model sequences are constructed in the single covariate setting. Without this assumption the distributions of the resultant statistics will, in general, depend on parameters of the null model, even when $n \to \infty$ (Aerts et al., 2000). The second assumption is needed in order to ensure that the test is consistent against virtually any alternative to H_0, Aerts et al..

Figure 1 shows four possible model sequences Aerts et al. (2000) discussed for two covariate setting described above. The first few models in the sequences are graphically represented by plotting the number of the step in which the basis elements enter the model for each index (j,k). For the model sequence depicted in Figure 1 (a), $u_1(x_1)$, $u_1(x_2)$ and the interaction $u_1(x_1)u_1(x_2)$ terms are added in step 1; that is, $\Lambda_1 = \{(0,1),(1,0),(1,1)\}$. In step 2 the following terms are added: $u_1(x_1)$, $u_1(x_2)$, $u_2(x_1)u_1(x_2)$, $u_1(x_1)u_2(x_2)$ and $u_2(x_1)u_2(x_2)$ so that $\Lambda_2 = \{(0,1),(1,0),(1,1),(0,2),(2,0),(1,2),(2,1),(2,2)\}$. Note that $\Lambda_1 \subset \Lambda_2$. This model sequence adds $2j+1$ terms to the previous model at step j. Inspection of (3.19) reveals that penalization against a given model is linearly related to the number of parameters in the model which, in turn, grows rather fast as the number of models considered increases. This clearly limits the number of models from this sequence which can be compared with the null model and, consequently, tests based on this sequence will possess undesirable power properties. This problem is less severe in the sequence depicted in Figure 1 (b), where only $j+1$ terms are added at each step. Figures 1 (c) and (d) are even more parsimonious. The sequence illustrated in Figure 1 (c) includes the main effects corresponding to frequency j at step $2j-1$ and j interaction terms at step $2j$. The

sequence described in Figure 1 (d) is clearly the most parsimonious in that no more than two new terms are added at each step. Aerts et al. (2000) state that there exist other model sequences leading to omnibus tests.

The large sample results of the statistics reviewed in Section 3.3.1 can be generalized to accommodate the multiple regression approach described in this section. Let Z_{jk}, for $k = 1, \ldots, N_j$ and $j = 1, 2, \ldots$, be independent and identically distributed standard normal random variables where $N_j = N(\Lambda_j) - N(\Lambda_{j-1})$, for $j = 1, 2, \ldots$ with $\Lambda_0 = \emptyset$ and Λ_j corresponding to a suitable sequence of alternatives such as those described above for $j = 1, 2, \ldots$. Then Theorem 1 of Aerts et al. (2000) generalizes with V_r and $\widetilde{r}$ defined as follows:

$$V_0 = 0, \;\; V_r = \sum_{j=1}^{r} \sum_{k=1}^{N_j} Z_{jk}^2 \;\; (r = 1, 2, \ldots), \tag{3.20}$$

and $\widetilde{r} = \arg\max\{V_r - 2N(\Lambda_r) : r = 0, 1, \ldots\}$.

Since the techniques described in this section rely on nonparametric smoothers, one would expect that these methods are vulnerable to the curse of dimensionality. Aerts et al. (2000) explain that that for an omnibus test that places the same emphasis on all p covariates, the upper bound on the order of the series-based alternatives, R_n, must not exceed $n^{1/p}$. The consequence is that higher order alternatives cannot be included in the model sequence and hence the ability of these tests to detect higher frequency departures from the null can quickly diminish as the dimension of the x-space increases. This limitation can be circumvented to an extent by formulating model sequences with the ability to detect specific departures from the null model.

Aerts et al. (2000) explain how to choose a path in a way to detect specific departures

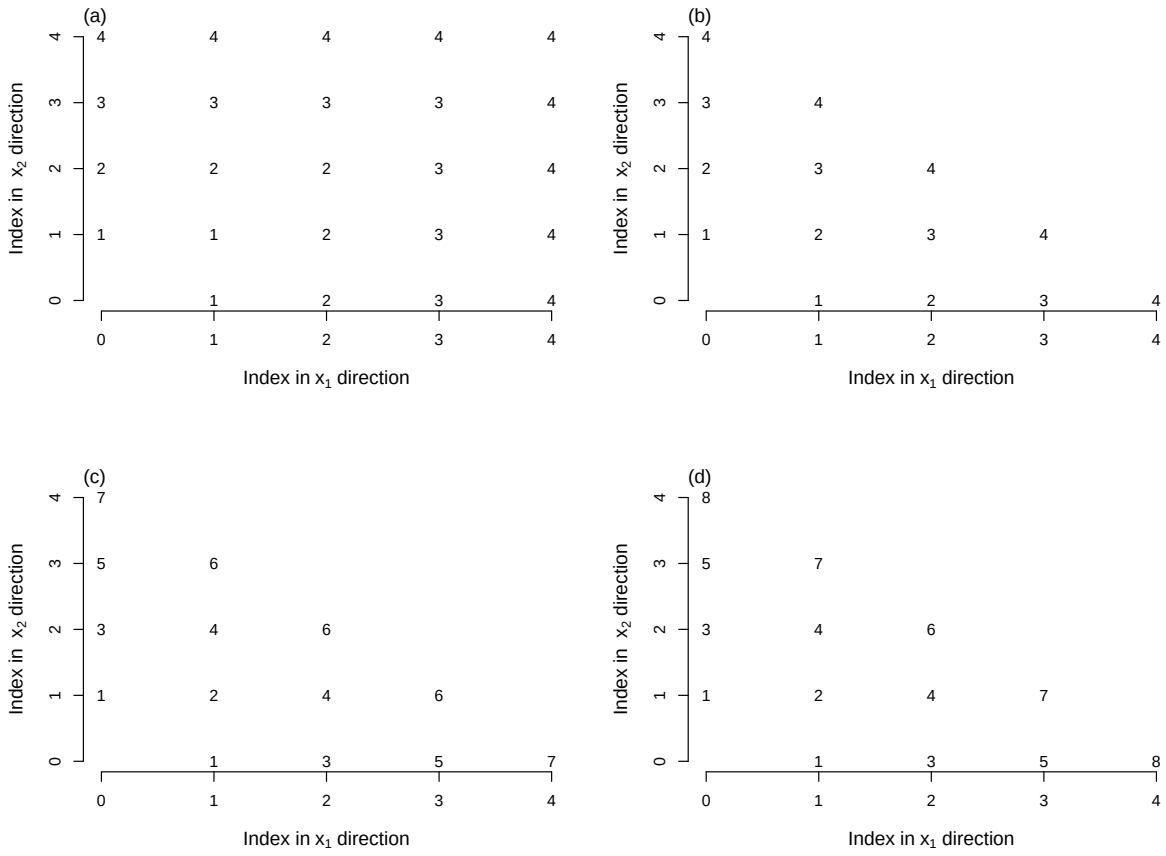

Fig. 1. *Four examples of model sequences in two dimensions.*

of interest. For example, one can specify model sequences in order to test the adequacy of the specified link function in a generalized linear model (see Section 3.3.2) or the presence of interaction when $\eta(\cdot;\boldsymbol{\beta})$ is specified to be an additive model (see Section 3.3.2).

Additive models

Additive models are a well-known tool for circumventing the curse of dimensionality. Additive models are formulated to provide an estimate of the marginal effect of a covariate on the response. Thus, if one assumes an absence of interaction effects (or at least presumes that such effects are negligible), then alternatives to the null model (3.2) can be constructed using additive models. For example, in the two-covariate setting alternatives to (3.17) can be written as

$$\eta(x_1, x_2) = \eta(x_1, x_2; \boldsymbol{\beta}) + \sum_{j=1}^{k_r} \phi_j u_j(x_1) + \sum_{j=1}^{l_r} \phi_j u_j(x_2), \tag{3.21}$$

where $k_r \geq k_{r-1}$ and $l_r \geq l_{r-1}$ for $r = 2, 3, \ldots$.

Aerts et al. (2000) refer to a type of test as a *diagonal test* which is based on sequences of nested models constructed from these alternatives. For this test Aerts et al. insist that $k_r = l_r$ while letting k_r increase by 1 at each step, so that $\Lambda_j = \{(1,0),(0,1),(2,0),(0,2),\ldots,(j,0),(0,j)\}$. The resulting path $\{(k_r, l_r) : r \geq 1\}$ corresponding to this test proceeds along the diagonal $\{(k_r, k_r) : r \geq 1\}$ and hence the name "diagonal" test. Note that in this strategy only two terms are added to each subsequent model; however, as we noted in Section 3.3.2, the number of additional terms may grow without bound. The asymptotic distribution theory of Aerts et al. (2000) yields $V_r = (Z_1^2 + Z_2^2) + \cdots + (Z_{2r-1}^2 + Z_{2r}^2)$. This approach can be extended in a fairly direct way to models with more than two covariates.

A goodness-of-link test

Aerts et al. (2000) describe how their method can be utilized to test the adequacy of the specified link. In this case the hypothesized model is contrasted with alternative models of the form

$$\eta(x_1, x_2) = \eta(x_1, x_2; \boldsymbol{\beta}) + \sum_{j \in \Lambda} \phi_j u_j\{\eta(x_1, x_2; \boldsymbol{\beta})\}. \tag{3.22}$$

This construction provides an alternative approach to the tests discussed in Section 2.6. However, it seems worth noting that this formulation bears some resemblance to the traditional goodness-of-link methods discussed in the previous chapter. In particular, the u_j's play a similar role to the constructed variables utilized in the classical goodness-of-link test reviewed in Section 2.6.1. Aerts et al.'s proposal clearly differs from the technique reviewed earlier in that it utilizes nonparametric methods to detect departures from the link function, while the traditional goodness-of-link test utilizes generalized parametric link models to detect departures from a proposed link.

The 'max' tests in models with any number of covariates

The 'max' test described in Aerts et al. (2000) provides a way of constructing an omnibus test from multiple specialized tests. To clarify, consider the two-covariate case in which specialized alternative models are constructed as follows

$$\eta(x_1, x_2) = \eta(x_1, x_2; \boldsymbol{\beta}) + \sum_{j \in \Lambda} \phi_j u_j(x_k) \;\; (k = 1, 2), \tag{3.23}$$

where departures from the null model are investigated for only one of the covariates. Clearly, such an alternative would useful only if one presumes that x_k alone is respon-

sible for lack of fit of the null model. However, by taking the maximum of the test statistic values obtained by using this sequence of alternatives for each of the covariates separately, one can obtain a test that is sensitive to departures from the null model caused by either of the two covariates.

Aerts et al. (2000) explain how the idea described above can be applied to models with $p > 2$ covariates using sequences for alternative models described in Section 3.3.2. In this case, one would use the following alternative for each pair of covariates separately

$$\eta(x_1, \ldots, x_p) = \eta(x_1, \ldots, x_p; \boldsymbol{\beta}) + \sum_{(j,k)\in\Lambda} \phi_{j,k} u_j(x_r) u_k(x_s), \tag{3.24}$$

where $1 \leq r \neq s \leq p$ and Λ is an index set formulated to follow one of the paths reviewed in Section 3.3.2. The test statistic is then taken to be the maximum of all $d(d-1)/2$ test statistics.

Finally, the level of a test constructed in a manner described above can be controlled by using either Bonferroni's inequality or by a bootstrap method. Aerts et al. (2000) explain that in situations where the number of covariates p is large, one might find a bootstrap procedure preferable since application of Bonferroni's inequality will result in a very conservative test.

3.3.3 Bayesian-Motivated Tests of Function Fit

Aerts et al. (2004) propose a Bayes-inspired nonparametric test of (3.2). In particular, they use the BIC approximation to the posterior probability of H_0 as a criterion for detecting departures from the proposed null model. The motivation for this approach is that one would generally interpret a sufficiently small value of this probability as evidence refuting the null model and would consequently be inclined to reject H_0. A Bayesian would directly

use the estimated value of the posterior probability approximation to assess the plausibility of H_0. However, Aerts et al. (2004) provide the asymptotic distribution for a statistic they derived from the posterior distribution, so that one may assess significance in a traditional frequentist fashion. Aerts et al. (2004) to be the first peer-reviewed article to study lack-of-fit tests based on posterior probabilities; however, a test based on this premise was studied by Hart (1997) in the special case of Gaussian-based regression models.

Test statistic and distribution theory

Formulation of the posterior probability requires consideration of a collection of alternative models denoted $M_1, \ldots, M_K$, where each M_j corresponds to a different parametric specification for the function η. The model which assumes that H_0 is true will be called M_0. Let $\mathbf{y} = (y_1, \ldots, y_n)$ denote the observed response values. With this notation we can apply Bayes' Theorem to express the posterior probability of the null model as

$$\begin{aligned} P(M_0|\mathbf{y}) &= \frac{\mathrm{pr}(\mathbf{y}|M_0)\mathrm{pr}(M_0)}{\sum_{j=0}^{K}\mathrm{pr}(\mathbf{y}|M_j)\mathrm{pr}(M_j)} = \left\{1 + \sum_{j=1}^{K}\frac{\mathrm{pr}(M_j)}{\mathrm{pr}(M_0)}\frac{\mathrm{pr}(\mathbf{y}|M_j)}{\mathrm{pr}(\mathbf{y}|M_0)}\right\}^{-1} \\ &= \left\{1 + \sum_{j=1}^{K}\frac{\mathrm{pr}(M_j)}{\mathrm{pr}(M_0)}\exp\left[\log\left(\mathrm{pr}(\mathbf{y}|M_j)\right) - \log\left(\mathrm{pr}(\mathbf{y}|M_0)\right)\right]\right\}^{-1} \end{aligned} \tag{3.25}$$

where $\mathrm{pr}(M_j)$, $j = 0, 1, \ldots, K$ denotes the prior probability of the jth model and $\mathrm{pr}(\mathbf{y}|M_j)$ represents the marginal likelihood of the under the jth model. Evaluating $\mathrm{pr}(\mathbf{y}|M_j)$ is often difficult in practice. Aerts et al. use the following version of the BIC which is a well-known and easily computed approximation of $\mathrm{pr}(\mathbf{y}|M_j)$:

$$\begin{aligned}\log(\mathrm{pr}(\mathbf{y}|M_j)) &= \log\left(\int_{\Theta_j}\mathrm{pr}(\mathbf{y}|\boldsymbol{\theta}_j,M_j)\pi(\boldsymbol{\theta}_j|M_j)\boldsymbol{\theta}_j\right)\\ &\approx \log(\mathrm{pr}(\mathbf{y}|M_j,\widehat{\boldsymbol{\theta}}_j)) - \frac{1}{2}m_j\log n =: \mathrm{BIC}_j\end{aligned} \tag{3.26}$$

where $\mathrm{pr}(\mathbf{y}|M_j,\widehat{\boldsymbol{\theta}}_j)$ is the likelihood corresponding to model M_j and m_j is the dimension of model M_j. Using noninformative priors for the model probabilities, that is, $\mathrm{pr}(M_j) = \mathrm{pr}(M_0)$, $j = 1,\ldots,K$, (3.25) can be reexpressed in terms of (3.26) as follows

$$\begin{aligned}P(M_0|\mathbf{y}) &\approx \left\{1+\textstyle\sum_{j=1}^{K}\exp\left[\mathrm{BIC}_j-\mathrm{BIC}_0\right]\right\}^{-1}\\ &= \left\{1+\textstyle\sum_{j=1}^{K}n^{-(1/2)(m_j-m_0)}\exp[\mathcal{L}_j/2]\right\}^{-1} =: \pi_{\mathrm{BIC}}\end{aligned} \tag{3.27}$$

where $\mathcal{L}_j = \log(L_j/L_0)$, i.e., the log-likelihood ratio of the M_j and M_0 models.

Clearly from (3.27), small values of $P(M_0|\mathbf{y})$ are evident in small values of π_{BIC}. Furthermore, small values of π_{BIC} clearly correspond to large values of

$$\sqrt{n}(1-\pi_{BIC}) = \frac{\widetilde{S}_n}{1+\widetilde{S}_n/\sqrt{n}} \tag{3.28}$$

where

$$\widetilde{S}_n = \sum_{k=1}^{K}\exp\left\{\mathcal{L}_k/2\right\}. \tag{3.29}$$

It can be shown that $\mathcal{L}_k$ can be approximated by the score statistic $n\widehat{\phi}_j^2/a(\widehat{\psi}_0)$ written explicitly in terms of Fourier coefficient estimators

$$\widehat{\phi}_k = \frac{1}{n}\sum_{i=1}^{n}[y_i - b'(\eta(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}^0))]\widehat{u}_k(\mathbf{x}_i),\ \ k=1,\ldots,K \tag{3.30}$$

in which $\widehat{u}_k(\cdot)$, $k = 1, \ldots, K$ denote basis functions that have been scaled to produce the required orthogonality conditions (this approximation will be discussed further in Chapter IV). Thus, we may in turn approximate $\widetilde{S}_n$ by

$$S_{n,BIC} = \sum_{k=1}^{K} \exp\left\{ \frac{n\widehat{\phi}_j^2}{2a(\widehat{\psi}_0)} \right\}. \tag{3.31}$$

The quantity $n\widehat{\phi}_j^2/a(\widehat{\psi}_0)$ is known to have the same limiting distribution as the log-likelihood ratio $\mathcal{L}_{1k}$ under the null hypothesis and general regularity conditions, which suggests that under general conditions the limiting distribution of $\widetilde{S}_K$ is the same as that of S_n (Aerts et al., 2004).

Tests based on (3.27) may be implemented in either a Bayesian or frequentist fashion. A Bayesian would directly use the estimated value of the posterior probability approximation to assess the plausibility of H_0, while a frequentist would determine the null distribution of π_n, and then reject H_0 at level of significance α if and only if π_n is smaller than an quantile of this distribution. Methods of the latter type which are derived from Bayesian principles, but used in frequentist fashion are referred to as *frequentist-Bayes* (Hart, 2009).

The issue of choosing alternative models $M_1, M_2, \ldots$ requires special consideration for this test. As with all series-based tests, the alternative models used to construct the test are extremely important in ensuring consistency of the test against virtually any departure from the null model (see Section 3.2.3). Aerts et al. (2004) consider two main types of alternative models.

Nested alternatives

The first type of alternative model considered is the class of nested alternatives discussed in Section 3.2.3. Recall that as $j \to \infty$, functions of the form (3.5) span the space of all functions that are continuous on $[0, 1]$. Aerts et al. explain that as long as K tends to ∞

at an appropriate rate with the sample size, this property is generally enough to ensure that there exist tests based on the models $M_1, \ldots, M_K$ that are consistent against any continuous alternative to H_0.

Applying Theorem 2 of Aerts et al. (2004) to canonical link regression models implies that when π_{BIC} is constructed using nested models, then there exists a sequence $\{K_n\}$ tending to infinity such that under H_0, we have

$$n^{1/2}\left[1-\left\{1+\sum_{j=1}^{K_n}\exp\left(\mathrm{BIC}_j-\mathrm{BIC}_0\right)\right\}^{-1}\right] \xrightarrow{d} \exp\left(\frac{1}{2}\chi_1^2\right) \tag{3.32}$$

as $n \to \infty$.

This means that the power of a test based on (3.27) and constructed from nested alternatives will depend solely on the alternative of smallest dimension. This conclusion effectively defeats the purpose of applying π_{BIC} using the nested alternatives of Section 3.2.3. Clearly, there is no added benefit to considering a series of dimension larger than $p+1$ when constructing π_{BIC}.

Singleton alternatives

In an effort to rectify the apparent shortcoming of nested alternatives noted above, Aerts et al. (2004) formulated a class of alternative models which they refer to as singletons. Singletons contain only one more parameter than the null model, $\eta(\cdot;\boldsymbol{\beta})$ and hence are not nested within each other. To illustrate this class of alternatives in the case where η is a function defined on $[0,1]$, a candidate for η_j is

$$\eta(x;\boldsymbol{\beta})+\phi_j\cos(\pi j x). \tag{3.33}$$

Note that this collection of alternatives does not necessarily contain η, even in the limit. However, Aerts et al. (2004) argue that the resulting test will usually be consistent as long as M_0 is not the best approximation to η among the $M_0, M_1, M_2, \ldots$. In the case where η is continuous and $\eta \notin \mathcal{N}$, there will exist a k such that the MLE of ϕ_k in $\eta(x;\boldsymbol{\beta})+\phi_k \cos(\pi k x)$ consistently estimates a nonzero quantity. Aerts et al. claim that such a property implies the existence of a consistent test.

Under H_0 and regularity conditions presented in Aerts et al. (2004), the authors show that when $S_{n,BIC}$ is constructed using singleton models, then

$$\frac{S_{n,BIC} - a_K}{b_K} \xrightarrow{d} S \tag{3.34}$$

as n and K tend to infinity, where in the notation of Samorodnitsky and Taqqu (1994), S has the stable distribution $S_1(1,1,0)$ and

$$a_K = \frac{\sqrt{\pi}}{2} \cdot \frac{K}{\sqrt{\log K}} \text{ and } b_K = \frac{K a_K}{\sqrt{\pi}} \int_1^\infty \frac{\sin(x/a_K)}{x^2\sqrt{\log x}} dx, \quad K = 1, 2, \ldots. \tag{3.35}$$

The method used by Aerts et al. to prove the result quoted above requires that the number of alternatives, K, approach infinity at a rate no faster than $o(n^{1/8})$.

Comments

In principle, the alternative models considered need not be limited to the two classes cited above. In practice, however, there are limitations. For example, one intuitively appealing class of alternatives is the collection of all models of the form

$$\eta(x;\boldsymbol{\beta}) + \sum_{j \in \mathcal{A}} \phi_j u_j(x) \tag{3.36}$$

where $\mathcal{A}$ is an arbitrary subset of $0, 1, \ldots, K$ for some K. Note that this collection of alternatives contains both collections of nested and singleton alternatives as well as a vast collection of other possible alternatives. Unfortunately, such alternatives are problematic if K grows with sample size. In particular, this collection of alternatives requires that 2^{K+1} models must be fitted, which becomes prohibitively large very quickly.

The approach presented in Aerts et al. (2004) does have a couple of drawbacks. First, the asymptotic distribution for the test statistic is relatively complex making the test potentially difficult to implement in practice. Also, this test has some undesirable power properties including an inability of the test to detect $1/\sqrt{n}$-local alternatives.

3.4 Lack-of-Fit Tests Based on Laplace Approximations

Hart (2009) revisits the notion of testing lack-of-fit using statistics based on approximation of the posterior probability of the null hypothesis in a frequentist fashion. A key difference in the proposal made in Hart (2009) from the approach introduced in Aerts et al. (2004) is that the former uses the method of Laplace to approximate posterior probabilities whereas the latter uses BIC. The motivation for pursuing a test statistic based on the Laplace method is that it is known to yield a more refined, accurate approximation of the posterior probability (Kass and Raftery, 1995; Raftery, 1996). Consequently, one would presume that the resulting test statistic will possess improved power properties over tests based on the BIC approximation.

3.4.1 Model Assumptions

Another noteworthy difference between Hart (2009) and Aerts et al. (2004) is that Hart (2009) assumes a special case of the model conditions presented in Section 3.2.1 in which the response is normally distributed. Consequently, the method as presented in Hart (2009)

is not justified for use in all models described in Section 3.2.1. That is, the observations $Y_1, \ldots, Y_n$ are assumed to be generated from the model

$$Y_i = \eta(\mathbf{x}_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3.37}$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are fixed, d-dimensional design points, and the unobserved errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed as $N(0, \sigma^2)$.

3.4.2 Test Statistic and Distribution Theory

Under the assumption of normal response data, the null hypothesis (3.6) is tested using the following statistic

$$B_n = \sum_{j=1}^{n} \rho_j \exp\left(\frac{n\widehat{\phi}_j^2}{2\widehat{\sigma}^2}\right), \tag{3.38}$$

where $\rho_j = \pi_j/(1-\pi_j)$. It is assumed that $\phi_1, \ldots, \phi_n$ are a priori independent with

$$P(\phi_j = 0) = 1 - \pi_j, \ j = 1, \ldots, n, \tag{3.39}$$

where $\pi_j < 1$ for all j and, given that $\phi_j \neq 0$, ϕ_j has density η, $j = 1, \ldots, n$. Furthermore, given that $\pi_j \neq 0$, π_j has density g, $j = 1, \ldots, n$.

Now suppose that g is Lipschitz continuous around 0 and there exists $\delta < 1$ such that $\sum_{j=1}^{\infty} \pi_j^{\delta} < \infty$. Hart (2009) proceeds to demonstrate that under $\sqrt{n}$-local alternatives defined as $\phi_j = \frac{1}{\sqrt{n}}\lambda_j$, $n = 1, 2, \ldots$, $j = 1, \ldots, n$ where it is assumed that $\lambda_j \to \infty$ as $j \to \infty$, B_n converges in distribution to

$$g(0) \sum_{j=1}^{\infty} \frac{\pi_j}{1-\pi_j} \exp\left[(Z_j + \lambda_j/\sigma)^2/2\right], \tag{3.40}$$

which is an almost surely convergent series (Cline, 1983) where $Z_1, Z_2, \ldots$ are i.i.d. standard normal random variables.

3.4.3 Comments

Simulation studies presented in Hart (2009) demonstrate that the frequentist-Bayes tests presented therein have good power against a wide variety of departures from the null model. In comparison to other well-known omnibus tests, the simulation results reveal that this test has superior power against high frequency alternatives while performing competitively against low frequency alternatives. The omnibus tests against which the Laplace based test was compared included two particularly relevant nonadaptive tests as well as an adaptive test which utilizes a selection criteria based on a compromise of the AIC and BIC proposed by Inglot and Ledwina (1996). Such power properties are remarkable given that examination of (3.38) reveals that B_n is nonadaptive.

Hart (2009) notes that modification of the statistic and its limiting distribution are required in order to ensure that use of this statistic is valid for more general models. As we will demonstrate explicitly in Chapter IV, when applying Laplace approximations to more general models, the statistic analogous to B_n will be a weighted sum of likelihood ratios $\widehat{L}_0/\widehat{L}_j$, where $\widehat{L}_0$ and $\widehat{L}_j$, $j = 1, \ldots, K$, are maximized likelihoods of the null and K alternative models, respectively. Writing $\mathcal{L}_j = 2\log(\widehat{L}_0/\widehat{L}_j)$, we have

$$\frac{\widehat{L}_j}{\widehat{L}_0} = \exp\left(\frac{\mathcal{L}_j}{2}\right), \tag{3.41}$$

and if the null is nested within each alternative, then under standard regularity conditions each $\mathcal{L}_j$ will have an asymptotic χ^2 distribution under the null hypothesis. In short, the "sum of exponentials" phenomenon can be attributed to two factors: (i) the use of a posterior probability to test H_0, and (ii) consideration of more than two models. When the null

is compared to just one other model, our frequentist-Bayes test is essentially the same as a likelihood ratio test.

3.5 Nonadaptive Tests

To this point we have neglected to address a key characteristic that distinguishes the various tests described above. Examination of the tests derived from posterior probabilities reveals that neither requires explicit specification of a truncation point. It is worth noting that the proposals of Aerts et al. (2004) impose a a condition on the rate of growth of the order of the sum relative to the sample size. A consequence of the assumption of a Gaussian response in Hart (2009) is that no such constraint is required.

This differs from the approach of Aerts et al. (1999) and Aerts et al. (2000), which are both based on data-driven selection of the truncation point. Tests utilizing data selected values of a smoothing parameter (in our case, the truncation point) are often referred to as “adaptive” tests, while tests that do not use data-driven values are, of course, called “nonadaptive”. Hart (2009) explains that nonadaptive tests have been generally dismissed in favor of well-constructed adaptive tests because the former tend to have good power only against certain types of alternatives. This has been demonstrated in simulation studies as well as in formal analysis of power under local alternatives (Eubank and Hart, 1993; Hart, 1997). Consequently, the power properties reported for the method introduced in Hart (2009) are particularly striking in that they defy the generally accepted notions regarding the relative performance of adaptive and nonadaptive tests.

To clarify which aspects of his proposed method are responsible for the superior power properties, Hart (2009) compared the formulae of three nonadaptive tests in the case of Gaussian response models. In the remainder of this section we will review some of conclusions reached in that comparison. Table 2 shows the formulae for the statistics of Hart

(2009), B_n, to the BIC-based statistic of Aerts et al. (2004), $S_{n,BIC}$, and a cusum-based statistic similar to those discussed in Section 2.4.2 expressed in terms of Fourier coefficients (see Hart (1997)). Brief inspection of Table 2 reveals rather obvious similarities in the forms of these statistics. This resemblance is particularly interesting because both of these statistics have been reported to have unsatisfactory power properties; however, Hart (2009) explains that the ways in which B_n differs from the BIC and cusum statistics actually lead to improved power. Consequently, this comparison leads to new insights regarding the apparent deficiencies of the BIC and cusum statistics.

Table 2. Three nonadaptive lack-of-fit statistics.

Aerts, et al. (2004):	$S_{n,BIC} = \sum_{k=1}^{K} \exp\left\{n\widehat{\phi}_k^2/2\widehat{\sigma}^2\right\}$
Hart (2009):	$B_n = \sum_{k=1}^{K} \rho_k \exp\left\{n\widehat{\phi}_k^2/2\widehat{\sigma}^2\right\}$
cusum approximation:	$C_n = 2\sum_{k=1}^{K} w_{k,n}\widehat{\phi}_k^2/\widehat{\sigma}^2$

In comparing the form of the Laplace based statistic in (3.38) to that of the BIC based statistic in (3.29), one sees that the latter is sum of exponentiated squared, normalized Fourier coefficients while the former is composed of a *weighted* sum of exponentiated squared, normalized Fourier coefficients. It turns out that the superior power reported for the Laplace based statistic is a consequence a stabilizing effect of these prior weights (Hart, 2009). These prior weights have the added benefit of allowing the investigator to adapt the test in order to detect specific departures from the null model. Alternatively, using noninformative priors yields omnibus lack-of-fit statistics. Finally, a consequence which is evident in the derivations presented in Hart (2009) is that B_n arises naturally from a posterior probability constructed from generally formulated Bayesian model averages. On

the other hand, the formulation of the posterior probability used in Aerts et al. (2004) was limited to singleton models. Hart (2009) comments that such models would rarely be used in function estimation, which in turn makes the formulation from which the BIC based test statistic was derived seem somewhat contrived by comparison.

A similar comparison of the form of B_n to that of C_n reveals that C_n is a weighted sum of squared normalized Fourier coefficients while is composed of a weighted sum of *exponentiated* squared, normalized Fourier coefficients. Hart (2009) found that in the special case where $\rho_j = w_{j,n} = j^{-2}$, $j = 1, \ldots, n$, B_n has better overall power than C_n. The superior power properties of B_n against higher frequency alternatives have been attributed to the exponentiation of the Fourier coefficients (Hart, 2009). This discovery led Hart (2009) to conclude that the deficiencies reported for the cusum statistic are a consequence of using a relatively ineffective function of each Fourier coefficient rather than excessive downweighting.

3.6 Discussion

In reviewing the literature on series-based tests of fit we have found that adaptive methods share the following features:

1. they are generally easy to implement;

2. most possess desirable power properties;

Furthermore, we noted that nonadaptive tests typically do not share these properties with the notable exception of the Laplace based statistic reviewed in Section 3.4.

One point that we touched on briefly in this chapter is that series-based lack-of-fit tests for generalized linear models have been inspired by well-established tests for Gaussian-based models. In the next chapter we will pursue this line of reasoning and revisit the recent

proposal of Hart (2009) in the context of generalized linear models. Given the relative ease of implementation of this method as well as its desirable power properties reported in Hart (2009), we contend that such a development presents a promising direction for further research. Thus, we intend to propose an analogous statistic which is suitable for testing the fit of generalized linear models as well as provide justification that has not yet been presented in the existing literature for such a proposal.

CHAPTER IV

A LACK-OF-FIT TEST FOR GENERALIZED LINEAR MODELS BASED ON LAPLACE APPROXIMATION

4.1 Overview

In Chapter III we reviewed several series-based lack-of-fit tests. Among the methods reviewed, we discussed two that share a special distinction in that they are both derived from approximations of the posterior probability of a hypothesized model (i.e., null model). The method of Aerts et al. (2004) applies to a rather general class of models, but as we reported previously, it has several shortcomings. Hart (2009) presents a lack-of-fit test that overcomes the shortcomings of Aerts et al., but the class of models for which the method was formulated was limited in comparison to the model assumptions considered by Aerts et al. In this chapter we will apply the ideas from Hart (2009) to the generalized linear model conditions addressed in Aerts et al. (2004). Thus, borrowing concepts from both sources we will obtain a lack-of-fit test for generalized linear models that retains the desirable properties cited for Hart's method.

In Section 4.2 we will state the general model assumptions and discuss suitable alternative models for developing our test (see Section 4.2.1) and the appropriate orthogonality conditions (see Section 4.2.3). In Section 4.3 we will formulate the posterior probability of a hypothesized model and subsequently derive the test statistic under the assumptions of Section 4.2. In Section 4.4 the properties of the statistics based on likelihood ratios and their null-equivalent score statistics will be studied. This examination will include identifying the appropriate limiting distribution for each statistic and examination of the score-based statistic's power against local alternatives (see Section 4.4.2).

4.2 Model Assumptions and Inference Problem

Suppose the data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are observed, where $\mathbf{x}_i$ is a vector of covariates and y_i is a scalar response. Our focus will be on canonical link regression models which were introduced in Section 1.2.3. Assuming the covariates to be fixed and the observations to be independent, the log-likelihood function can be written as

$$l(\eta, \psi) = \sum_{i=1}^{n} \{[y_i \eta(\mathbf{x}_i) - b(\eta(\mathbf{x}_i))]/a(\psi) + c(y_i, \psi)\}, \tag{4.1}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions, η is an unknown function and ψ an unknown dispersion parameter (see Section 1.2). We consider testing the null hypothesis

$$H_0 : \eta(\mathbf{x}) = \sum_{j=1}^{p} \beta_j \gamma_j(\mathbf{x}) \equiv \eta(\mathbf{x}; \boldsymbol{\beta}), \tag{4.2}$$

where $\gamma_1, \ldots, \gamma_p$ are known functions, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is an unknown parameter vector. Note that an intercept term can be accommodated by defining $\gamma_1(\mathbf{x}) \equiv 1, \ \forall \mathbf{x}$. The asymptotic maximizer of the expected log-likelihood

$$\frac{1}{n} \sum_{i=1}^{n} [b'(\eta(\mathbf{x}_i))\eta(\mathbf{x}_i; \boldsymbol{\beta}) - b(\eta(\mathbf{x}_i; \boldsymbol{\beta}))], \tag{4.3}$$

with respect to $\boldsymbol{\beta}$ is denoted $\boldsymbol{\beta}^0 = (\beta_1^0, \ldots, \beta_p^0)^{\mathrm{T}}$, which is the true parameter vector when H_0 is true and provides a best null approximation to η when H_0 is false.

4.2.1 Alternative Models

We will pursue an omnibus test of (4.2). Put simply, this requires that the test we develop has the ability to detect departures from (4.2) within a very wide class of alternative models. To this end, we will consider a collection of alternative models which differ only in their specification of $\eta(\mathbf{x})$ and ψ; that is, for each alternative, the data will be assumed to have

log-likelihood given by (4.1). Furthermore, these alternative formulations of $\eta(\mathbf{x})$ need not be nested within each other.

The specific forms of our alternatives will be based on Fourier-type regression models. Let $\mathcal{U} = \{u_1, u_2, \ldots, u_K\}$ where $K < n - p$ is a fixed, user-specified integer and $u_1, u_2, \ldots, u_K$ denote basis functions such as cosines, wavelets, or orthogonal polynomials. Now for $m = 0, 1, \ldots, K$, define $n_m = \binom{K}{m}$ and let $S_{m1}, \ldots, S_{mn_m}$ be the n_m subsets of $\{1, \ldots, K\}$ of size m. For each m and k, let $\overline{S}_{mk} = \{1, \ldots, K\} \setminus S_{mk}$. The alternatives considered will be of the form

$$\eta_{mk}(\mathbf{x}) = \eta(\mathbf{x}; \boldsymbol{\beta}) + \sum_{j \in S_{mk}} \phi_j u_j(\mathbf{x});\ k = 1, \ldots, n_m,\ m = 1, \ldots, K, \tag{4.4}$$

where for $j \in S_{mk}$, we have $u_j \in \mathcal{U}$; that is, u_j's used to estimate each alternative will be limited to the pre-specified collection $\mathcal{U}$. Inspection of (4.4) reveals that the null hypothesis (4.2) is "nested" within each of the alternative models. In fact by definition of S_{mk}, for $m = 0$, we have $S_{01} = \emptyset$ and $\eta_{01}(\mathbf{x}) = \eta(\mathbf{x}; \boldsymbol{\beta})$, where $\eta(\mathbf{x}; \boldsymbol{\beta})$ denotes the null model defined in (4.2). This is an important feature of these alternatives which will be utilized in the development of the test statistic and its sampling distribution.

For notational convenience, let M_{mk} denote the probability model corresponding to η_{mk} for $k = 1, \ldots, n_m,\ m = 1, \ldots, K$ and let M_0 denote the model corresponding to the null hypothesis (4.2). Furthermore, the log-likelihood corresponding to each model will be written as follows

$$l(\eta_{mk}, \psi) = \sum_{i=1}^{n} \{[y_i \eta_{mk}(\mathbf{x}_i) - b(\eta_{mk}(\mathbf{x}_i))]/a(\psi) + c(y_i, \psi)\}, \tag{4.5}$$

which reflects our desire to have the competing specifications of (4.1) differ only in their specification of the linear predictor, and ψ.

We now discuss some issues that must be considered regarding K. From the above

formulation, K can be regarded as the highest frequency considered in the Fourier-type alternatives characterized by (4.4). Thus, one would typically be inclined to prefer that K be fairly "large" so that we have some assurance that the union of $M_0,\ M_{mk},\ k = 1, \ldots, n_m,\ m = 1, \ldots, K$ should come close to spanning the space of all possibilities for η.

It would appear that a test statistic based on the alternatives specified in (4.4) provides a means of detecting a wide range of departures from H_0 and would hence provide assurance of an omnibus test. Unfortunately, alternative models specified by (4.4) have a major defect. Such alternatives become problematic because the number of models that must be fitted is $\sum_{m=0}^{K} n_m = 2^K$, which will clearly be large if K is chosen to be large (as suggested above). The number of alternatives which must be fitted to the data becomes prohibitively large even for relatively small samples. Despite the fact that the collection of alternatives defined by (4.4) is impractical, this definition provides a conceptually useful starting point for developing a test statistic. During the development of the test statistic we will revisit this issue and address it as we derive the statistic.

4.2.2 Notation

We will now introduce some notation in order to obtain more convenient and concise expressions for the models defined thus far. We start by noting that the problem we have described so far bears a striking resemblance to a variable selection problem presented in Wang and George (2007), so some of the notation which follows has been inspired by that reference.

For $k = 1, \ldots, n_m,\ m = 1, \ldots, K$, let $T_{mk} = [\Gamma\ U_{mk}]$ be an $n \times (p + m)$ matrix, where $\Gamma = [\boldsymbol{\gamma}_1\ \boldsymbol{\gamma}_2\ \cdots\ \boldsymbol{\gamma}_p]$, with $\boldsymbol{\gamma}_j = (\gamma_j(\mathbf{x}_1), \gamma_j(\mathbf{x}_2), \ldots, \gamma_j(\mathbf{x}_n))^{\mathrm{T}}$ for $j = 1, \ldots, p$ and similarly, $U_{mk} = [\mathbf{u}_j]_{j \in S_{mk}}$ with $\mathbf{u}_j = (u_j(\mathbf{x}_1), u_j(\mathbf{x}_2), \ldots, u_j(\mathbf{x}_n))^{\mathrm{T}}$ for $j = 1, \ldots, K$. Accordingly define $\boldsymbol{\theta}_{mk} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\phi}_{mk}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, and $\boldsymbol{\phi}_{mk} = (\phi_j)_{j \in S_{mk}}^{\mathrm{T}}$.

The vector $\boldsymbol{\theta}_{m0}$ is $(\boldsymbol{\beta}^{\mathrm{T}}, \mathbf{0}_m^{\mathrm{T}})^{\mathrm{T}}$, while $\boldsymbol{\theta}_m^0 = ((\boldsymbol{\beta}^0)^{\mathrm{T}}, \mathbf{0}_m^{\mathrm{T}})^{\mathrm{T}}$ denotes the maximizer of (4.3) when parameterized as a function of $\boldsymbol{\theta}$. With this notation we express η under model M_{mk} for $k = 1, \ldots, n_m,\ m = 1, \ldots, K$ as

$$\eta_{mk}(\mathbf{x}) = \eta(\mathbf{x}; \boldsymbol{\beta}) + \sum_{j \in S_{mk}} \phi_j u_j(\mathbf{x}). \tag{4.6}$$

Now note that maximum likelihood estimators of these parameters will depend on the specific model fit to the data. In order to distinguish parameter estimates from various models, let $\widehat{\boldsymbol{\theta}}_{mk} = (\widehat{\boldsymbol{\beta}}_{mk}^{\mathrm{T}}, \widehat{\boldsymbol{\phi}}_{mk}^{\mathrm{T}})^{\mathrm{T}}$, where $\widehat{\boldsymbol{\beta}}_{mk} = (\widehat{\beta}_{mk1}, \ldots, \widehat{\beta}_{mkp})^{\mathrm{T}}$, and $\widehat{\boldsymbol{\phi}}_{mk} = (\widehat{\phi}_j)_{j \in S_{mk}}^{\mathrm{T}}$ denote the corresponding values estimated for model M_{mk} from the sample data and $\widehat{\eta}_{mk}$ denotes the value of η_{mk} estimated by substituting $\widehat{\boldsymbol{\beta}}_{mk}$ and $\widehat{\boldsymbol{\phi}}_{mk}$ into (4.6). Finally, the estimated value of the dispersion parameter under model M_{mk} is denoted by $a(\widehat{\psi}_{mk})$.

4.2.3 Orthogonality Conditions

To produce test statistics that are meaningful and powerful, we impose the following orthonormality conditions which were introduced in Aerts et al. (2004) to accommodate models such as the type described in Section 4.2:

$$\sum_{i=1}^{n} \gamma_j(\mathbf{x}_i) u_k(\mathbf{x}_i) b''(\eta(\mathbf{x}_i; \boldsymbol{\beta}^0)) = 0, \quad j = 1, 2, \ldots, p,\ k = 1, 2, \ldots, K, \tag{4.7}$$

and

$$\frac{1}{n} \sum_{i=1}^{n} b''(\eta(\mathbf{x}_i; \boldsymbol{\beta}^0)) u_j(\mathbf{x}_i) u_k(\mathbf{x}_i) = \begin{cases} 1 & \text{if } j = k, \\ 0 & \text{if } j \neq k. \end{cases} \tag{4.8}$$

In practice, an approximation to (4.7) and (4.8) can be obtained as follows. First, obtain $(\widehat{\boldsymbol{\beta}}^0, \widehat{\psi}^0)$, the maximizer of the null likelihood function, and let $\widehat{W}$ be the $n \times n$ diagonal matrix with diagonal elements $b''(\eta(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}^0))$, $i = 1, \ldots, n$. We assume that $\widehat{\boldsymbol{\beta}}^0$ converges in

probability to $\boldsymbol{\beta}^0$. Now, choose a set of functions $v_1, v_2, \ldots$ that are a basis for all functions of interest and let $V = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_K]$ where $\mathbf{v}_j = (v_j(\mathbf{x}_1), v_j(\mathbf{x}_2), \ldots, v_j(\mathbf{x}_n))^{\mathrm{T}}$ for $j = 1, \ldots, K$. Then apply a Gram-Schmidt procedure to the columns of the matrix $[\widehat{W}^{1/2}\Gamma\ V]$ to obtain a collection of vectors $\widehat{\mathbf{v}}_1, \widehat{\mathbf{v}}_2, \ldots, \widehat{\mathbf{v}}_K$. Finally, taking $\widehat{\mathbf{u}}_j = \sqrt{n} \times \widehat{W}^{-1/2}\widehat{\mathbf{v}}_j$ produces a collection of vectors $\widehat{\mathbf{u}}_j = (\widehat{u}_j(\mathbf{x}_1), \widehat{u}_j(\mathbf{x}_2), \ldots, \widehat{u}_j(\mathbf{x}_n))^{\mathrm{T}}$ for $j = 1, \ldots, K$ with components which possess the desired properties.

4.3 Derivation of Test Statistics

To test the null hypothesis (4.2), we shall consider the approach presented in Hart (2009). This requires that we first propose a prior distribution for the Fourier coefficients $\phi_1, \ldots, \phi_K$, and then compute the posterior probability, $P(M_0|\mathbf{D})$, of the null model which is denoted by M_0. One would be inclined to reject H_0 when the statistic $P(M_0|\mathbf{D})$ is sufficiently small. A frequentist would determine the cutoff point for rejection by deriving the frequency distribution of $P(M_0|\mathbf{D})$ under H_0 and then choosing an appropriate Type I error probability.

In order to simplify our subsequent discussion and derivations, we will not consider assigning a prior to the dispersion parameter, ψ. However, since H_0 can be characterized in terms of $\phi_1, \ldots, \phi_K$ (see Chapter III), the Fourier coefficients are the parameters of primary interest. Thus, imposing priors on these parameters is essential in formulating the posterior probability.

4.3.1 Applying the Laplace Approximation

We now consider the general recommendations made in Hart (2009) for modifying the Laplace approximation approach presented therein so that it is appropriate for the conditions assumed in Section 4.2. It will be assumed that $\phi_1, \ldots, \phi_K$ are a priori independent

with

$$P(\phi_j = 0) = 1 - \pi_j,\ j = 1, \ldots, K, \tag{4.9}$$

where $\pi_j < 1$ for all j and, given that $\phi_j \neq 0$, ϕ_j has density g, $j = 1, \ldots, K$.

We will assess the validity of H_0 in (4.2) by calculating its posterior probability. From Bayes Theorem we can express the posterior probability of M_0 as follows

$$P(M_0|\mathbf{D}) = \frac{p_0(\mathbf{D})}{p_{\text{marg}}(\mathbf{D})} = \left\{1 + \sum_{m=1}^{n}\sum_{k=1}^{n_m}\prod_{j\in S_{mk}}\left(\frac{\pi_j}{1-\pi_j}\right)B_{mk}\right\}^{-1}, \tag{4.10}$$

where

$$p_0(\mathbf{D}) = P(\mathbf{D}|M_0)\prod_{j=1}^{n}(1-\pi_j), \tag{4.11}$$

$$p_{\text{marg}}(\mathbf{D}) = p_0(\mathbf{D}) + \sum_{m=1}^{K}\sum_{k=1}^{n_m}P(\mathbf{D}|M_{mk})\prod_{j\in S_{mk}}\pi_j\prod_{j\in\overline{S}_{mk}}(1-\pi_j), \tag{4.12}$$

and B_{mk} denotes the Bayes factor defined as follows

$$B_{mk} = \frac{P(\mathbf{D}|M_{mk})}{P(\mathbf{D}|M_0)}, \tag{4.13}$$

with

$$P(\mathbf{D}|M_{mk}) = \int_{\boldsymbol{\theta}_{mk}} \text{pr}(\mathbf{D}|\boldsymbol{\theta}_{mk}, M_{mk}) \times \prod_{j\in S_{mk}} g(\phi_j)d\boldsymbol{\theta}_{mk}. \tag{4.14}$$

In the above expressions the $\mathbf{D}$ denotes "data" and $P(\mathbf{D}|M)$ denotes the marginal likelihood for the data under model M while $\text{pr}(\mathbf{D}|\boldsymbol{\theta}, M)$ denotes the conditional distribution of

the data, given model M and its parameter. Note that when viewed as a function of $\boldsymbol{\theta}$ one typically refers to $\mathrm{pr}(\mathbf{D}|\boldsymbol{\theta}, M)$ as the "likelihood" of $\boldsymbol{\theta}$. In our case $\mathrm{pr}(\mathbf{D}|\boldsymbol{\theta}_{mk}, M_{mk}) \propto \exp\{l(\eta_{mk}, \psi)\}$ where $l(\eta_{mk}, \psi)$ was defined in (4.5).

We now apply a common variant of the "pure" Laplace approximation in which the prior is evaluated at the maximum likelihood estimate of $\widehat{\boldsymbol{\theta}}$ rather than at the posterior mode; see Kass and Raftery (1995). This yields

$$P(\mathbf{D}|M_{mk}) \approx (2\pi)^{(m+p)/2}|J_{mk}^{-1}(\widehat{\boldsymbol{\theta}}_{mk})|^{1/2}\mathrm{pr}(\mathbf{D}|\widehat{\boldsymbol{\theta}}_{mk}, M_{mk}) \times \textstyle\prod_{j\in S_{mk}} g(\widehat{\phi}_j), \quad (4.15)$$

where $J_{mk}(\widehat{\boldsymbol{\theta}}_{mk})$ denotes the Hessian matrix $-\partial l(\eta_{mk}, \psi)/\partial\boldsymbol{\theta}_{mk}\partial\boldsymbol{\theta}_{mk}^{\mathrm{T}}$ evaluated at $\widehat{\boldsymbol{\theta}}_{mk}$; $J_0(\widehat{\boldsymbol{\theta}}_0)$ is defined similarly. That is, $J_{mk}(\widehat{\boldsymbol{\theta}}_{mk})$ and $J_0(\widehat{\boldsymbol{\theta}}_0)$ represent the information matrices for models M_{mk} and M_0, respectively. The approximation given in (4.15) implies

$$\begin{aligned}\widehat{B}_{mk} &\approx \frac{(2\pi)^{(m+p)/2}|J_{mk}^{-1}(\widehat{\boldsymbol{\theta}}_{mk})|^{1/2}\mathrm{pr}(\mathbf{D}|\widehat{\boldsymbol{\theta}}_{mk}, M_{mk})\prod_{j\in S_{mk}} g(\widehat{\phi}_j)}{(2\pi)^{p/2}|J_0^{-1}(\widehat{\boldsymbol{\theta}}_0)|^{1/2}\mathrm{pr}(\mathbf{D}|\widehat{\boldsymbol{\theta}}_0, M_0)} \\ &= (2\pi)^{m/2}\left(\frac{|J_{mk}^{-1}(\widehat{\boldsymbol{\theta}}_{mk})|}{|J_0^{-1}(\widehat{\boldsymbol{\theta}}_0)|}\right)^{1/2}\exp\left\{\frac{\mathcal{L}_{mk}}{2}\right\}\prod_{j\in S_{mk}} g(\widehat{\phi}_j).\end{aligned} \quad (4.16)$$

In equation (4.16) $\mathcal{L}_{mk} = 2\log(\mathrm{pr}(\mathbf{D}|\widehat{\boldsymbol{\theta}}_{mk}, M_{mk})/\mathrm{pr}(\mathbf{D}|\widehat{\boldsymbol{\theta}}_0, M_0))$ where $\mathrm{pr}(\mathbf{D}|\widehat{\boldsymbol{\theta}}_0, M_0)$ and $\mathrm{pr}(\mathbf{D}|\widehat{\boldsymbol{\theta}}_{mk}, M_{mk})$, are the maximized likelihoods of the null model, M_0, and the alternative model M_{mk}, respectively. Note that $\mathcal{L}_{mk}$ is the standard likelihood-ratio test statistic which will have an asymptotic χ^2 distribution under the null hypothesis when M_0 is nested within M_{mk} and standard regularity conditions hold. Models as defined in Section 4.2 are nested and are known to satisfy standard regularity conditions.

From inspection of (4.10) it is clear that the frequentist test that rejects H_0 for small values of $\widehat{P}(M_0|\mathbf{D})$ (i.e., $P(M_0|\mathbf{D})$ evaluated at $\widehat{B}_{mk}$ $k = 1, \ldots, n_m,\ m = 1, \ldots, K$) is

equivalent to one that rejects for large values of

$$(2\pi)^{-1/2} \sum_{m=1}^{K} \sum_{k=1}^{n_m} \left(\prod_{j \in S_{mk}} \frac{\pi_j}{1-\pi_j} \right) \widehat{B}_{mk}. \tag{4.17}$$

Thus, applying the approximation from (4.16) leads to rejection of H_0 for large values of

$$E_{n,K} := \sum_{m=1}^{K} \sum_{k=1}^{n_m} (2\pi)^{(m-1)/2} \left(\prod_{j \in S_{mk}} \frac{\pi_j}{1-\pi_j} g(\widehat{\phi}_j) \right) \frac{|J_{mk}^{-1}(\widehat{\boldsymbol{\theta}}_{mk})|^{1/2}}{|J_0^{-1}(\widehat{\boldsymbol{\theta}}_0)|^{1/2}} \exp\left\{ \frac{\mathcal{L}_{mk}}{2} \right\}. \tag{4.18}$$

We now examine how the limiting distribution of $E_{n,K}$ depends on the collection of alternative models used to construct it. The following theorem was inspired by Theorem 1 of Aerts et al. (2004).

The following are assumptions needed in our proofs of theoretical results presented throughout this chapter:

A1. The design points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are fixed and confined to a compact subset $\mathcal{S}$ of $\mathbb{R}^d$ for all n.

A2. The functions $\gamma_1, \ldots, \gamma_p, u_1, u_2, \ldots$ satisfy the following conditions:

(i) There exists $B_1^* < \infty$ such that

$$\sup_{1 \le j \le p,\ \mathbf{x} \in \mathcal{S}} |\gamma_j(\mathbf{x})| < B_1^* \text{ and}$$

(ii) there exists a sequence of positive constants $\{B_j : j = 1, 2, \ldots\}$ such that

$$\sup_{1 \le j \le K,\ \mathbf{x} \in \mathcal{S}} |u_j(\mathbf{x})| < B_K, \quad K = 1, 2, \ldots.$$

A3. The functions $u_1, u_2, \ldots$ satisfy (4.7) and (4.8) and $\widehat{u}_1, \ldots, \widehat{u}_K$ are constructed from $\gamma_1, \ldots, \gamma_p, v_1, v_2, \ldots$ as described in Section 4.2.3.

A4. The dispersion parameter $a(\psi_0)$ is positive, and for $k = 1, \ldots, n_m$, $m = 1, \ldots, K$ the MLEs $\widehat{\boldsymbol{\psi}}_{mk}$ and $\widehat{\boldsymbol{\theta}}_{mk}$ of ψ_0 and $\boldsymbol{\theta}_0$, respectively, are such that $E(a(\widehat{\boldsymbol{\psi}}_{mk}) - a(\psi_0))^2$ and $E\|\widehat{\boldsymbol{\theta}}_{mk} - \boldsymbol{\theta}_0\|^2$ exist and are each $O(n^{-1})$.

A5. Let $\mathcal{B}$ be the parameter space for $\boldsymbol{\beta}$. There exists a compact, connected subset $\mathcal{N}$ of $\mathcal{B}$ such that $\boldsymbol{\beta}^0 \in \mathcal{N}$ and, for each $\mathbf{x} \in \mathcal{S}$, $\eta(\mathbf{x}; \boldsymbol{\beta})$ is a continuous function of $\boldsymbol{\beta}$ on $\mathcal{N}$.

A6. The function b is such that b'' is nonnegative and b''' exists and is bounded by a constant B_2^* for all $\boldsymbol{\beta}$ and all $\mathbf{x} \in \mathcal{S}$.

A7. The prior density of ϕ_k, g, is bounded and Lipschitz continuous for $k = 1, \ldots, K$.

A8. $n^{-1} J_0(\boldsymbol{\theta}_0) \to J_0^*$ as $n \to \infty$ where J_0^* is some $p \times p$ positive definite matrix.

Assumptions A1.-A6. are based on conditions imposed in Aerts et al. (2004), A7. is a condition used in Hart (2009), and A8. is a necessary assumption which is discussed in Fahrmeir and Tutz (2001).

Theorem 4.3.1. Let $\mathcal{A}$ be a set containing only the finite collection of models, M_{mk}, $k = 1, \ldots, n_m$, $m = 1, \ldots, K$ defined in Section 4.2.1. Then under H_0, we have

$$n^{1/2} E_{n,K} \xrightarrow{d} g(0) a^{1/2}(\psi_0) \sum_{k=1}^{K} \frac{\pi_k}{1 - \pi_k} \exp(V_k/2) \text{ as } n \to \infty \tag{4.19}$$

where $V_1, \ldots, V_K$ are independently distributed random variables each having the χ_1^2 distribution.

Proof. Throughout the proof let $C_1, C_2, \ldots$ denote positive constants that depend on neither n nor K. We will make use of the decomposition $E_{n,K} = \Delta_1 + \Delta_2$ where

$$\Delta_1 = \sum_{k=1}^{K} \frac{\pi_k}{1 - \pi_k} g(\widehat{\phi}_k) \left(\frac{|J_{1k}^{-1}(\widehat{\boldsymbol{\theta}}_{1k})|}{|J_0^{-1}(\widehat{\boldsymbol{\theta}}_0)|} \right)^{1/2} \exp\left\{ \frac{\mathcal{L}_{1k}}{2} \right\} \tag{4.20}$$

and

$$\Delta_2 = \sum_{m=2}^{K}\sum_{k=1}^{n_m}(2\pi)^{(m-1)/2}\left(\prod_{j\in S_{mk}}\frac{\pi_j}{1-\pi_j}g(\widehat{\phi}_j)\right)\frac{|J_{mk}^{-1}(\widehat{\boldsymbol{\theta}}_{mk})|^{1/2}}{|J_0^{-1}(\widehat{\boldsymbol{\theta}}_0)|^{1/2}}\exp\left\{\frac{\mathcal{L}_{mk}}{2}\right\}. \quad (4.21)$$

Defining

$$D_n = \left(\frac{|n^{-1}J_{mk}(\widehat{\boldsymbol{\theta}}_{mk})|}{|n^{-1}J_0(\widehat{\boldsymbol{\theta}}_0)|}\right)^{-1/2}\exp\left\{\frac{\mathcal{L}_{mk}}{2}\right\},$$

we now show that $\Delta_2 = o_p(1)$:

$$\begin{aligned}
|\Delta_2| &= \sum_{m=2}^{K}\sum_{k=1}^{n_m}(2\pi)^{(m-1)/2}\left(\prod_{j\in S_{mk}}\frac{\pi_j}{1-\pi_j}g(\widehat{\phi}_j)\right)n^{-m/2}D_n \\
&\leq n^{-1}\sum_{m=2}^{K}\sum_{k=1}^{n_m}(2\pi)^{(m-1)/2}\left(\prod_{j\in S_{mk}}\frac{\pi_j}{1-\pi_j}g(\widehat{\phi}_j)\right)D_n \\
&\leq n^{-1}\sum_{m=2}^{K}C_1^m(2\pi)^{(m-1)/2}\sum_{k=1}^{n_m}\left(\prod_{j\in S_{mk}}\frac{\pi_j}{1-\pi_j}\right)D_n \\
&= C_2 n^{-1}O_p(1) = O_p(n^{-1}).
\end{aligned}$$

In the above calculations, the first inequality follows from the fact that $n^{-m/2} \leq n^{-1}$ for $m \geq 2$, while the second inequality follows from A7.; that is, the fact that g is bounded by a constant, call it C_1. The concluding equality is a consequence of $\exp\{\mathcal{L}_{mk}/2\} = O_p(1)$ and $|n^{-1}J_{mk}(\widehat{\boldsymbol{\theta}}_{mk})|/|n^{-1}J_0(\widehat{\boldsymbol{\theta}}_0)| = O_p(1)$, while $\exp\{\mathcal{L}_{mk}/2\} = O_p(1)$ follows from the fact that if a sequence of random variables converges in distribution, then it must also be bounded in probability (Serfling, 1980). Under H_0 we are assured by McCullagh and Nelder (1989) that $\mathcal{L}_{mk}$ has a limiting chi-square distribution with m degrees

of freedom, which leads us to conclude that the sequence is $O_p(1)$. The assertion that $|n^{-1}J_{mk}(\widehat{\boldsymbol{\theta}}_{mk})|/|n^{-1}J_0(\widehat{\boldsymbol{\theta}}_0)| = O_p(1)$ will be addressed in the subsequent discussion.

We will now examine the limit of the ratio $|n^{-1}J_{mk}(\widehat{\boldsymbol{\theta}}_{mk})|/|n^{-1}J_0(\widehat{\boldsymbol{\theta}}_0)|$. First, under the assumed orthogonality conditions (4.7) and (4.8), the information matrix simplifies to

$$J_{mk}(\boldsymbol{\theta}_{mk}) = -\frac{\partial^2 l(\eta_{mk}, \psi)}{\partial \boldsymbol{\theta}_{mk} \partial \boldsymbol{\theta}_{mk}^{\mathrm{T}}} = \begin{bmatrix} -\dfrac{\partial^2 l(\eta_{mk}, \psi)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} & \mathbf{0}_{m \times p} \\ \\ \mathbf{0}_{p \times m} & (n/a(\psi))\, I_m \end{bmatrix} \tag{4.22}$$

where

$$\frac{\partial^2 l(\eta_{mk}, \psi)}{\partial \beta_q \partial \beta_r} = -\frac{1}{a(\psi)} \sum_{i=1}^{n} \{\gamma_r(\mathbf{x}_i)\gamma_q(\mathbf{x}_i) b''(\eta_{mk}(\mathbf{x}_i))\} \text{ for all } q,\ r = 1, \ldots, p.$$

and I_m is the $m \times m$ identity matrix. According to Theorem 13.3.8 of Harville and from examination of (4.22), for $k = 0, 1, \ldots, n_m$, $m = 1, \ldots, K$ we may write

$$\left| n^{-1} J_{mk}(\widehat{\boldsymbol{\theta}}_{mk}) \right| = \left(a(\widehat{\psi}_{mk}) \right)^{-m} \left| -\frac{1}{n} \frac{\partial^2 l(\widehat{\eta}_{mk}, \widehat{\psi}_{mk})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} \right|. \tag{4.23}$$

Now observe that since $\widehat{\boldsymbol{\theta}}_{mk}$ is consistent for $\boldsymbol{\theta}_0$ by assumption A4. and that $|n^{-1}J_{mk}(\widehat{\boldsymbol{\theta}}_{mk})|$ is a continuous function of $\widehat{\boldsymbol{\theta}}_{mk}$, we may apply Theorem 1.7.ii on p. 24 of Serfling (1980) and assumption A8. (with some algebra) to (4.23) and get

$$\left| n^{-1} J_{mk}(\widehat{\boldsymbol{\theta}}_{mk}) \right| \xrightarrow{p} \left(a(\psi_0) \right)^{-m} |J_0^*|. \tag{4.24}$$

Finally, J_0^* is assumed to be positive definite, so that $|J_0^*| > 0$. Thus, noting that $|n^{-1}J_{mk}(\widehat{\boldsymbol{\theta}}_{mk})|$ converges to a degenerate random variable, we may apply Slutsky's theorem to conclude that

$$|n^{-1}J_{mk}(\widehat{\boldsymbol{\theta}}_{mk})|/|n^{-1}J_0(\widehat{\boldsymbol{\theta}}_0)| \xrightarrow{p} (a(\psi_0))^{-m}\,. \tag{4.25}$$

Moreover, it is now obvious that $|n^{-1}J_{mk}(\widehat{\boldsymbol{\theta}}_{mk})|/|n^{-1}J_0(\widehat{\boldsymbol{\theta}}_0)| = O_p(1)$.

To conclude this proof, we now argue that

$$\sqrt{n}\Delta_1 \xrightarrow{d} g(0)a^{1/2}(\psi_0)\sum_{k=1}^{K}\frac{\pi_k}{1-\pi_k}\exp(V_k/2) \text{ as } n\to\infty.$$

Rewrite $\sqrt{n}\Delta_1$ as

$$\sqrt{n}\Delta_1 = A_n + \Delta_{n11} + \Delta_{n12} \tag{4.26}$$

where

$$A_n = g(0)a^{1/2}(\psi_0)\sum_{k=1}^{K}\frac{\pi_k}{1-\pi_k}\exp\left\{\frac{\mathcal{L}_{1k}}{2}\right\}, \tag{4.27}$$

$$\Delta_{n11} = \sum_{k=1}^{K}\frac{\pi_k}{1-\pi_k}\left[g(\widehat{\phi}_k)-g(0)\right]\left(\frac{|n^{-1}J_{1k}(\widehat{\boldsymbol{\theta}}_{1k})|}{|n^{-1}J_0(\widehat{\boldsymbol{\theta}}_0)|}\right)^{-1/2}\exp\left\{\frac{\mathcal{L}_{1k}}{2}\right\} \tag{4.28}$$

and

$$\Delta_{n12} = g(0)a^{1/2}(\psi_0)\sum_{k=1}^{K}\frac{\pi_k}{1-\pi_k}\left[\left(a(\psi_0)\frac{|n^{-1}J_{1k}(\widehat{\boldsymbol{\theta}}_{1k})|}{|n^{-1}J_0(\widehat{\boldsymbol{\theta}}_0)|}\right)^{-1/2}-1\right]\exp\left\{\frac{\mathcal{L}_{1k}}{2}\right\}. \tag{4.29}$$

First note that g is Lipschitz continuous by A8. so that for some constant $C_4>0$, $|g(\widehat{\phi}_k)-g(0)| \le C_4|\widehat{\phi}_k| = O_p(n^{-1/2})$. Hence $\Delta_{n11}\xrightarrow{p}0$. Furthermore, from (4.25), we have $|n^{-1}J_{1k}(\widehat{\boldsymbol{\theta}}_{1k})|/|n^{-1}J_0(\widehat{\boldsymbol{\theta}}_0)| \xrightarrow{p} (a(\psi_0))^{-1}$ as $n\to\infty$, and so $\Delta_{n12}\xrightarrow{p}0$. Finally, $\mathcal{L}_{1k}\xrightarrow{d}V_k$ as $n\to\infty$ by the more general result on the convergence of $\mathcal{L}_{mk}$ cited above. The desired

convergence follows from Slutsky's Theorem. □

Theorem 4.3.1 shows that $E_{n,K}$ (or equivalently, the Laplace-based approximation of the posterior probability of H_0, $P(M_0|\mathbf{D})$) generally depends only on the models having the smallest number of elements. This is the same conclusion reached in Aerts et al. (2004) for the test statistic considered therein. So while $E_{n,K}$ provides a model average over a wide collection of possible alternative models, one may limit attention to a specific subclass of models and produce an asymptotically equivalent statistic. This observation motivates consideration of the following test statistic

$$\widetilde{S}_K^L = a^{1/2}(\widehat{\psi}_0) \sum_{k=1}^{K} \frac{\pi_k}{1-\pi_k} g(\widehat{\phi}_k) \exp\left\{\frac{\mathcal{L}_{1k}}{2}\right\} \tag{4.30}$$

where $\mathcal{L}_{1k} = 2\log(\mathrm{pr}(\mathbf{D}|\widehat{\boldsymbol{\theta}}_{1k}, M_{1k})/\mathrm{pr}(\mathbf{D}|\widehat{\boldsymbol{\theta}}_0, M_0))$. From inspection of (4.30) it is apparent that $\widetilde{S}_K^L$ is composed of likelihood ratio test statistics comparing M_0 to M_{1k}, $k = 1, \ldots, K$. This fact was noted in Hart (2009), however, its application to models described in Section 4.2 was not pursued in that paper.

Referring to Theorem 4.3.1, we see that the prior density, g, results in a multiplicative constant, $g(0)$, in the limiting distribution of $\widetilde{S}_K^L$. Thus, g appears to be of little benefit, which in turn leads us to take g be a constant (i.e., the improper uniform prior). Likewise, we observe that $a^{1/2}(\widehat{\psi}_0)$ also results in a multiplicative constant. Furthermore, $a^{1/2}(\widehat{\psi}_0)$ does not account for the influence of the added terms corresponding to singleton alternatives. Ultimately, we may drop the multipliers $a^{1/2}(\widehat{\psi}_0)$ and $g(\widehat{\phi}_k)$, $k = 1, \ldots, K$ from $\widetilde{S}_K^L$ to further simplify the test statistic. This leads us to the following statistic, which we will refer to as the "Bayes sum" statistic

$$S_K^L = \sum_{k=1}^{K} \frac{\pi_k}{1-\pi_k} \exp\left\{\frac{\mathcal{L}_{1k}}{2}\right\}. \tag{4.31}$$

Given the definition of M_{mk}, we can see that each alternative M_{1k} has the form

$$\eta(\mathbf{x};\boldsymbol{\beta}) + \phi_k u_k(\mathbf{x}), \quad k = 1, \ldots, K. \tag{4.32}$$

These types of alternatives were considered extensively in Aerts et al. (2004) and are called "singletons". They have the noteworthy feature that they contain only one more parameter than M_0 (i.e., a single Fourier-type coefficient). Furthermore, since each singleton contains M_0, the $\mathcal{L}_j$'s can be viewed as valid likelihood ratio test statistics (assuming the necessary regularity conditions hold).

Further reassurance that tests based on S_K^L will be effective (i.e., sensitive to departures from M_0) is provided by Aerts et al. (2004). Aerts et al. (2004) note that for tests based on singletons to be consistent, it is usually enough that the best approximation to η among the models entertained is not the null model.

4.3.2 Score-based Test Statistic

In practice, applying S_K^L involves computing K likelihood ratios, which in turn requires fitting the null model plus each of the K singleton alternatives under consideration. Since we wish our test of H_0 to be nonparametric, K should be fairly "large" and consequently the computational demands of fitting $K+1$ models could prohibit use of S_K^L. To circumvent this potential obstacle, Aerts et al. (2004) proposed a score analog of their BIC statistic. The score analog is obtained by replacing each of the K likelihood ratio statistics with its corresponding score approximation (i.e., score statistic), which is known to have the same limiting distribution under the null hypothesis and general regularity conditions. The score statistic is computationally preferable to the likelihood ratio statistic in that it only requires estimation of the null model.

In order to see how the rationale described above can be applied to S_K^L we start by noting that for $k = 1, \ldots, K$, $\mathcal{L}_{1k}$ can be viewed as the likelihood ratio test statistic for

testing the following hypothesis

$$H_0 : \phi_k = 0. \tag{4.33}$$

Recall from the discussion of Section 1.3.3 that under H_0 and general regularity conditions $\mathcal{L}_{1k}$ can be approximated by a quadratic form composed of the information matrix and the score function evaluated at the MLE of the coefficients from the null model (see Section 1.3.4). The asymptotically equivalent score statistic can be expressed as follows

$$\begin{aligned} \mathcal{S}_k &= [s(\widehat{\boldsymbol{\theta}}_{1k}^{(0)})]^{\mathrm{T}} J(\widehat{\boldsymbol{\theta}}_{1k}^{(0)}) s(\widehat{\boldsymbol{\theta}}_{1k}^{(0)}) \\ &= \frac{n}{a(\widehat{\psi}^0)} \left[\frac{1}{n} \sum_{i=1}^{n} [y_i - b'(\eta(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}^0))] \widehat{u}_k(\mathbf{x}_i) \right]^2, \end{aligned} \tag{4.34}$$

where $s(\cdot)$ is the score function and J is the Hessian matrix. Note that $\mathcal{S}_k$ simplifies since the first p elements of $\partial l(\widehat{\eta}_{1k}, \widehat{\psi}^0)/\partial \boldsymbol{\theta}_{1k}$ are 0 by definition of the MLE.

Now recognize that $\dfrac{1}{n} \sum_{i=1}^{n} [y_i - b'(\eta(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}^0))] \widehat{u}_k(\mathbf{x}_i)$ is a one-step estimator of ϕ_k obtained by taking the initial value of ϕ_k to be 0 (as specified by H_0) with the estimate of $\boldsymbol{\beta}$ computed assuming H_0 to be true. Applying the definition presented in Section 1.3

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{1k}^{(1)} &= \widehat{\boldsymbol{\theta}}_{1k}^{(0)} - [J(\widehat{\boldsymbol{\theta}}_{1k}^{(0)})]^{-1} s(\widehat{\boldsymbol{\theta}}_{1k}^{(0)}) \\ &= \begin{bmatrix} \left(a(\widehat{\psi}^0) \Gamma^{\mathrm{T}} W(\widehat{\boldsymbol{\beta}}^0) \Gamma \right)^{-1} \sum_{i=1}^{n} [y_i - b'(\eta(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}^0))] \Gamma_i \\ \\ \dfrac{1}{n} \sum_{i=1}^{n} [y_i - b'(\eta(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}^0))] \widehat{u}_k(\mathbf{x}_i) \end{bmatrix} \end{aligned} \tag{4.35}$$

where

$$-\frac{\partial^2 l(\eta_{1k}, \psi)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} = (\Gamma^{\mathrm{T}} W(\boldsymbol{\beta}) \Gamma) \tag{4.36}$$

with $W(\boldsymbol{\beta}) = \text{diag}\{w_1(\boldsymbol{\beta}), \ldots, w_n(\boldsymbol{\beta})\}$, and $w_j(\boldsymbol{\beta}) = b''(\eta_{1k}(\mathbf{x}_i))$.

Hence we define

$$\widehat{\phi}_k = \frac{1}{n} \sum_{i=1}^{n} [y_i - b'(\eta(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}^0))] \widehat{u}_k(\mathbf{x}_i), \quad k = 1, \ldots, K. \tag{4.37}$$

Finally, the above observations lead us to define the statistic S_K by

$$S_K = \sum_{k=1}^{K} \frac{\pi_k}{1 - \pi_k} \exp \left\{ \frac{n \widehat{\phi}_k^2}{2a(\widehat{\psi}^0)} \right\}. \tag{4.38}$$

From the above discussion, it is clear that the quantity $n\widehat{\phi}_k^2 / a(\widehat{\psi}^0)$ has the same limiting distribution as the log-likelihood ratio $\mathcal{L}_{1k}$ under the null hypothesis, which suggests that under general conditions the limiting distribution of S_K^L is the same as that of S_K. Thus, in the following discussion, we limit our attention to S_K recognizing that the same result will hold for S_K^L.

4.4 Statistical Properties, Asymptotic Distribution Theory

Our study of the large sample distribution theory for the Bayes sum statistic is divided into two parts. First, we will examine the asymptotic properties of the Fourier-type coefficients under local alternatives. We will then examine the implications of the observed properties on S_K^L and S_K.

4.4.1 Asymptotic Behavior of Fourier Coefficients

We now examine the limiting behavior of the statistics presented above. We will need the following assumption for our subsequent theorem:

A9. $E(\widehat{\xi}_j - \xi_j)^2 = O(n^{-1})$, $j = 1, \ldots, p + K$, where the ξ_js and $\widehat{\xi}_j$s are the coefficients arising from application of the Gram-Schmidt process for known and estimated β, respectively.

Theorem 4.4.1. For $k = 1, \ldots, K$, where K is any positive integer, define $\widehat{\phi}_k$ as in (4.37). Assume that the function η in our generalized linear model has the form

$$\eta_n(\mathbf{x}) = \eta(\mathbf{x}; \boldsymbol{\beta}^0) + \sum_{j=1}^{K} \phi_j u_j(\mathbf{x}) \tag{4.39}$$

where

$$\phi_k = \frac{\phi_k^*}{\sqrt{n}},\ n = 1, 2, \ldots,\ k = 1, \ldots, K, \tag{4.40}$$

with $|\phi_k^*| < \infty$, $k = 1, \ldots, K$. Suppose also that $\max_{1 \le i \le n} E[y_i - b'(\eta_n(\mathbf{x}_i))]^4 < \infty$ uniformly in n. Then we have

$$\frac{\sqrt{n}}{a^{1/2}(\widehat{\psi}^0)}(\widehat{\phi}_1, \ldots, \widehat{\phi}_K)^{\mathrm{T}} \xrightarrow{d} N\left(\frac{1}{a^{1/2}(\psi^0)}\boldsymbol{\phi}^*, I_K\right) \tag{4.41}$$

as $n \to \infty$ where $\boldsymbol{\phi}^* = (\phi_1^*, \ldots, \phi_K^*)^{\mathrm{T}}$ and I_K is the $K \times K$ identity matrix.

Proof. Throughout the proof let $C_1, C_2, \ldots$ denote positive constants that depend on neither n nor K. Our approach will be based on examining the components of the following decomposition of the Fourier coefficient estimators:

$$\widehat{\phi}_k = \widetilde{\phi}_k + e_{1k} + e_{2k} + e_{3k} \tag{4.42}$$

where for $k = 1, \ldots, K$,

$$\begin{aligned}
\widetilde{\phi}_k &= \frac{1}{n}\sum_{i=1}^{n}[y_i - b'(\eta_n(\mathbf{x}_i))]u_k(\mathbf{x}_i), \\
e_{1k} &= \frac{1}{n}\sum_{i=1}^{n}[b'(\eta_n(\mathbf{x}_i)) - b'(\eta(\mathbf{x}_i;\boldsymbol{\beta}^0))]u_k(\mathbf{x}_i), \\
e_{2k} &= \frac{1}{n}\sum_{i=1}^{n}[b'(\eta(\mathbf{x}_i;\boldsymbol{\beta}^0)) - b'(\eta(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}^0))]u_k(\mathbf{x}_i), \\
e_{3k} &= \frac{1}{n}\sum_{i=1}^{n}[y_i - b'(\eta(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}^0))][\widehat{u}_k(\mathbf{x}_i) - u_k(\mathbf{x}_i)].
\end{aligned} \tag{4.43}$$

The following proof will be organized into two main parts:

(a) showing that the estimated coefficients $\widehat{\phi}_k$ can be approximated by $\widetilde{\phi}_k$ (i.e., e_{1k}, e_{2k}, and e_{3k} are negligible relative to $\widetilde{\phi}_k$); and

(b) obtaining the joint large-sample distribution of the $\widetilde{\phi}_k$'s.

In addressing part (a), we start with e_{1k} and observe that under local alternatives, for $k = 1, \ldots, K$ we have:

$$\begin{aligned}
b'(\eta_n(\mathbf{x}_i)) - b'(\eta(\mathbf{x}_i;\boldsymbol{\beta}^0)) &= (\eta_n(\mathbf{x}_i) - \eta(\mathbf{x}_i;\boldsymbol{\beta}^0))b''(\eta(\mathbf{x}_i;\boldsymbol{\beta}^0)) \\
&\quad + \frac{1}{2}(\eta_n(\mathbf{x}_i) - \eta(\mathbf{x}_i;\boldsymbol{\beta}^0))^2 b'''(\widetilde{\eta}_i^{(1)}) \\
&= b''(\eta(\mathbf{x}_i;\boldsymbol{\beta}^0))\sum_{j=1}^{K}\phi_j u_j(\mathbf{x}_i) \\
&\quad + \frac{1}{2}b'''(\widetilde{\eta}_i^{(1)})\left[\sum_{j=1}^{K}\phi_j u_j(\mathbf{x}_i)\right]^2,
\end{aligned} \tag{4.44}$$

where $\widetilde{\eta}_i^{(1)}$ is some point interior to the interval joining $\eta(\mathbf{x}_i;\boldsymbol{\beta}^0)$ and $\eta_n(\mathbf{x}_i)$. Thus,

$$e_{1k} \quad = \phi_k + O(n^{-1}), \tag{4.45}$$

where we have used A2 (ii), A6, (4.8), and (4.40).

We now examine e_{2k}. Applying a Taylor's expansion, we obtain

$$\begin{aligned} e_{2k} \quad = -\frac{1}{n}\sum_{i=1}^{n}\Big[&(\eta(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}^0)-\eta(\mathbf{x}_i;\boldsymbol{\beta}^0))b''(\eta(\mathbf{x}_i;\boldsymbol{\beta}^0)) \\ &+\frac{1}{2}\,(\eta(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}^0)-\eta(\mathbf{x}_i;\boldsymbol{\beta}^0))^2 b'''(\widetilde{\eta}_i^{(2)})\Big]\,u_k(\mathbf{x}_i) \end{aligned} \tag{4.46}$$

where $\widetilde{\eta}_i^{(2)}$ is some point interior to the interval joining $\eta(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}^0)$ and $\eta(\mathbf{x}_i;\boldsymbol{\beta}^0)$. The assumed orthogonality conditions (4.7) imply that (4.46) simplifies to

$$e_{2k} = -\frac{1}{2n}\sum_{i=1}^{n}(\eta(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}^0)-\eta(\mathbf{x}_i;\boldsymbol{\beta}^0))^2 \times b'''(\widetilde{\eta}_i^{(2)})u_k(\mathbf{x}_i), \tag{4.47}$$

and we have

$$\begin{aligned} |e_{2k}| \quad &\leq \left(\frac{1}{n}\sum_{i=1}^{n} b'''(\widetilde{\eta}_i^{(2)})^2 u_k^2(\mathbf{x}_i)\right)^{1/2} \max_{1\leq i\leq n}(\eta(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}^0)-\eta(\mathbf{x}_i;\boldsymbol{\beta}^0))^2/2 \\ &\leq C_2\|\widehat{\boldsymbol{\beta}}^0-\boldsymbol{\beta}^0\|^2 = O_p(n^{-1}) \end{aligned} \tag{4.48}$$

where the last line follows from A4 and A6.

Now by assumption A9, we may write e_{3k} in terms of coefficients obtained through the Gram-Schmidt process and find:

$$e_{3k} \quad = \textstyle\sum_{j=1}^{p+k}(\widehat{\xi}_{jk}-\xi_{jk})\frac{1}{n}\sum_{i=1}^{n}[y_i - b'(\eta(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}^0))]v_j(\mathbf{x}_i) = O_p(n^{-1}). \tag{4.49}$$

The limiting behavior follows from the fact that $\frac{1}{n}\sum_{i=1}^{n}[y_i - b'(\eta(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}^0))]v_j(\mathbf{x}_i) = O_p(n^{-1/2})$, which in turn follows from decomposing $\frac{1}{n}\sum_{i=1}^{n}[y_i - b'(\eta(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}^0))]v_j(\mathbf{x}_i)$ in a manner similar to our decomposition of $\widehat{\phi}_k$ in (4.42) and examining its components in essentially the same way that we have analyzed e_{1k} and e_{2k}.

We now proceed to address part (b). By applying our findings from part (a) regarding the rates of convergence for e_{1k}, e_{2k}, and e_{3k}, we find that for any arbitrary real-valued constants $b_1, \ldots, b_K$, we have

$$\begin{aligned}\sum_{k=1}^{K} b_k \frac{\sqrt{n}\widehat{\phi}_k}{a^{1/2}(\widehat{\psi}^0)} &= \sum_{k=1}^{K} b_k \frac{\sqrt{n}\widetilde{\phi}_k}{a^{1/2}(\widehat{\psi}^0)} + \sum_{k=1}^{K} b_k \frac{\sqrt{n}e_{1k}}{a^{1/2}(\widehat{\psi}^0)} \\ &\quad + \sum_{k=1}^{K} b_k \frac{\sqrt{n}e_{2k}}{a^{1/2}(\widehat{\psi}^0)} + \sum_{k=1}^{K} b_k \frac{\sqrt{n}e_{3k}}{a^{1/2}(\widehat{\psi}^0)} \\ &= \sum_{k=1}^{K} b_k \frac{\sqrt{n}\widetilde{\phi}_k}{a^{1/2}(\widehat{\psi}^0)} + \sum_{k=1}^{K} b_k \frac{\phi_k^*}{a^{1/2}(\widehat{\psi}^0)} + o_p(1).\end{aligned} \tag{4.50}$$

Furthermore, for some point $\widetilde{a}$ on the line segment connecting $a(\widehat{\psi}^0)$ and $a(\psi^0)$, we may write

$$\frac{1}{a^{1/2}(\widehat{\psi}^0)} - \frac{1}{a^{1/2}(\psi^0)} = -\frac{1}{a^{3/2}(\psi^0)}(a(\widehat{\psi}^0) - a(\psi^0)) + \frac{1}{(\widetilde{a})^{5/2}}(a(\widehat{\psi}^0) - a(\psi^0))^2, \tag{4.51}$$

so that by assumption A4, we find that

$$\begin{aligned}&\sum_{k=1}^{K} b_k \frac{\sqrt{n}\widetilde{\phi}_k}{a^{1/2}(\widehat{\psi}^0)} + \sum_{k=1}^{K} b_k \frac{\phi_k^*}{a^{1/2}(\widehat{\psi}^0)} \\ &= \sum_{k=1}^{K} b_k \frac{\sqrt{n}\widetilde{\phi}_k}{a^{1/2}(\psi^0)} + \sum_{k=1}^{K} b_k \frac{\phi_k^*}{a^{1/2}(\psi^0)} + O_p(n^{-1/2})\end{aligned} \tag{4.52}$$

Now observe that by definition of $\widetilde{\phi}_k$ we may write

$$\begin{aligned}\sum_{k=1}^{K} b_k \frac{\sqrt{n}\widetilde{\phi}_k}{a^{1/2}(\psi^0)} &= \frac{\sqrt{n}}{a^{1/2}(\psi^0)} \sum_{k=1}^{K} b_k \left[\frac{1}{n}\sum_{i=1}^{n}[y_i - b'(\eta_n(\mathbf{x}_i))]u_k(\mathbf{x}_i)\right] \\ &= \frac{n^{-1/2}}{a^{1/2}(\psi^0)} \sum_{i=1}^{n} \left[\{y_i - b'(\eta_n(\mathbf{x}_i))\}\sum_{k=1}^{K} b_k u_k(\mathbf{x}_i)\right].\end{aligned} \tag{4.53}$$

For $i = 1, \ldots, n$ define $r_i = (n \times a(\psi^0))^{-1/2}[y_i - b'(\eta_n(\mathbf{x}_i))]\sum_{k=1}^{K} b_k u_k(\mathbf{x}_i)$ and $\sigma_n^2 = \sum_{i=1}^{n} \mathrm{var}(r_i)$. Under local alternatives, we have $E(r_i) = 0$ and

$$\begin{aligned}\mathrm{var}(r_i) &= \mathrm{var}\left(\frac{n^{-1/2}}{a^{1/2}(\psi^0)}[y_i - b'(\eta_n(\mathbf{x}_i))]\sum_{k=1}^{K} b_k u_k(\mathbf{x}_i)\right) \\ &= \frac{1}{n}b''(\eta_n(\mathbf{x}_i))\left(\sum_{k=1}^{K} b_k u_k(\mathbf{x}_i)\right)^2 \\ &= \frac{1}{n}b''(\eta(\mathbf{x}_i;\boldsymbol{\beta}^0))\left(\sum_{k=1}^{K} b_k u_k(\mathbf{x}_i)\right)^2 \\ &\quad + \frac{1}{n}(\eta_n(\mathbf{x}_i) - \eta(\mathbf{x}_i;\boldsymbol{\beta}^0))b'''(\widetilde{\eta}_i^{(3)})\left(\sum_{k=1}^{K} b_k u_k(\mathbf{x}_i)\right)^2\end{aligned} \tag{4.54}$$

for some point $\widetilde{\eta}_i^{(3)}$ interior to the interval joining $\eta_n(\mathbf{x}_i)$ and $\eta(\mathbf{x}_i;\boldsymbol{\beta}^0)$. We then have

$$\begin{aligned}\sigma_n^2 &= \frac{1}{n}\sum_{i=1}^{n} b''(\eta(\mathbf{x}_i;\boldsymbol{\beta}^0))\left(\sum_{k=1}^{K} b_k u_k(\mathbf{x}_i)\right)^2 \\ &+\frac{1}{n}\sum_{i=1}^{n}(\eta_n(\mathbf{x}_i)-\eta(\mathbf{x}_i;\boldsymbol{\beta}^0))b'''(\widetilde{\eta}_i^{(3)})\left(\sum_{k=1}^{K} b_k u_k(\mathbf{x}_i)\right)^2 \\ &= \sum_{k=1}^{K} b_k^2(\mathbf{x}_i)+\Delta_n\end{aligned} \tag{4.55}$$

where

$$\begin{aligned}|\Delta_n| &= \left|\frac{1}{n}\sum_{i=1}^{n}(\eta_n(\mathbf{x}_i)-\eta(\mathbf{x}_i;\boldsymbol{\beta}^0))b'''(\widetilde{\eta}_i^{(3)})\left(\sum_{k=1}^{K} b_k u_k(\mathbf{x}_i)\right)^2\right| \\ &\leq \frac{1}{\sqrt{n}}B_K^3 B_2^*\left|\sum_{j=1}^{K}\phi_j^*\right|\left(\sum_{k=1}^{K} b_k\right)^2 = o(n^{-1/2}).\end{aligned} \tag{4.56}$$

It clearly follows that $\sigma_n^3 \to (\sum_{k=1}^{K} b_k^2)^{3/2}$ as $n \to \infty$. Since $\max_{1\leq i\leq n} E[y_i - b'(\eta_n(\mathbf{x}_i))]^4 < \infty$ uniformly in n by assumption, we have

$$\begin{aligned}\textstyle\sum_{i=1}^{n} E|r_i|^3 &= \textstyle\sum_{i=1}^{n} E\left|\frac{n^{-1/2}}{a^{1/2}(\psi^0)}[y_i - b'(\eta_n(\mathbf{x}_i))]\sum_{k=1}^{K} b_k u_k(\mathbf{x}_i)\right|^3 \\ &= \frac{n^{-3/2}}{a^{3/2}(\psi^0)}\sum_{i=1}^{n} E|[y_i - b'(\eta_n(\mathbf{x}_i))]|^3 \times \left|\sum_{k=1}^{K} b_k u_k(\mathbf{x}_i)\right|^3 \\ &\leq \frac{n^{-3/2}}{a^{3/2}(\psi^0)}B_K^3\left|\sum_{k=1}^{K} b_k\right|^3 \times C_3.\end{aligned} \tag{4.57}$$

We now check the Liapunov Condition (Resnick, 1999) and observe:

$$\frac{1}{\sigma_n^3}\sum_{i=1}^{n} E|r_i|^3 \to 0 \text{ as } n \to \infty. \tag{4.58}$$

The Lindeberg condition follows as a consequence, so that by the Lindeberg-Feller CLT along with (4.55) and (4.56), we find that

$$\sum_{k=1}^{K} b_k \frac{\sqrt{n}\widetilde{\phi}_k}{a^{1/2}(\psi^0)} \xrightarrow{d} Z \text{ as } n \to \infty, \tag{4.59}$$

where

$$Z \stackrel{D}{=} b_1 Z_1 + \cdots + b_K Z_K \sim N\left(0, \sum_{k=1}^{K} b_k^2\right), \tag{4.60}$$

with $Z_1, \ldots, Z_K$ being independent standard normal random variables. Recalling that $b_1, \ldots, b_K$ were arbitrarily chosen real-valued constants, from the Cramér-Wold theorem we conclude that

$$\frac{\sqrt{n}}{a^{1/2}(\psi^0)}(\widetilde{\phi}_1, \ldots, \widetilde{\phi}_K)^{\mathrm{T}} \xrightarrow{d} N\left(\mathbf{0}, I_K\right), \text{ as } n \to \infty. \tag{4.61}$$

where I_K is the $K \times K$ identity matrix. Thus, it follows from (4.52) and (4.50) that

$$\frac{\sqrt{n}}{a^{1/2}(\widehat{\psi}^0)}(\widehat{\phi}_1, \ldots, \widehat{\phi}_K)^{\mathrm{T}} \xrightarrow{d} N(\boldsymbol{\phi}^*, I_K), \text{ as } n \to \infty. \tag{4.62}$$

□

Note that the above result characterizes the joint asymptotic normality of score statistics corresponding to K (singleton) alternative models rather than a score statistic for a single K-dimensional parameter model, which is a well-known result. It is important to recognize that the resulting score statistics are uncorrelated and hence independent. This finding will be essential in the formulation of our next result.

4.4.2 Asymptotic Distribution Theory for S_K^L and S_K

We now consider the limiting distribution of S_K under both the null hypothesis and local alternatives that converge to the null at rate $1/\sqrt{n}$.

Corollary 4.4.1. For $k = 1, \ldots, K$, where K is any integer, define $\widehat{\phi}_k$ as in (4.37). Assume that the function η in our generalized linear model has the form

$$\eta_n(\mathbf{x}) = \eta(\mathbf{x}; \boldsymbol{\beta}^0) + \sum_{j=1}^{K} \phi_j u_j(\mathbf{x}) \tag{4.63}$$

where

$$\phi_k = \frac{\phi_k^*}{\sqrt{n}},\ n = 1, 2, \ldots,\ k = 1, \ldots, K, \tag{4.64}$$

with $|\phi_k^*| < \infty$, $k = 1, \ldots, K$. Suppose also that $\max_{1 \le i \le n} E[y_i - b'(\eta_n(\mathbf{x}_i))]^4 < \infty$ uniformly in n. Under these model assumptions and assumptions A1-A9, we have

$$S_K \xrightarrow{d} \sum_{k=1}^{K} \frac{\pi_k}{1 - \pi_k} \exp\left\{(Z_k + \phi_k^*/a^{1/2}(\psi^0))^2/2\right\} \tag{4.65}$$

as n tends to infinity where $Z_1, Z_2, \ldots, Z_K$ be i.i.d. standard normal random variables.

Proof. From Theorem 4.4.1, we have for $k = 1, \ldots, K$

$$\frac{\sqrt{n}}{a^{1/2}(\widehat{\psi}^0)} \widehat{\phi}_k \xrightarrow{d} Z_k + \phi_k^*/a^{1/2}(\psi^0) \tag{4.66}$$

as $n \to \infty$. Letting $f(t_1, \ldots, t_K) = \sum_{k=1}^{K} \frac{\pi_k}{1-\pi_k} \exp(t_k^2/2)$, the desired result follows from Theorem 1.7 (iii) of Serfling (1980) by noting that f is a Borel function.

□

Now recall that the null hypothesis (4.2) is equivalent to

$$H_0 : \phi_1 = \phi_2 = \cdots = 0. \tag{4.67}$$

Furthermore, the exponents of S_K^L and S_K are equivalent under H_0. Thus, we may use the previous result for S_K to characterize the asymptotic distribution of S_K^L under H_0.

Corollary 4.4.2. Let S_K^L be as defined in (4.31) and suppose that $\max_{1\leq i\leq n} E[y_i - b'(\eta(\mathbf{x};\boldsymbol{\beta}^0))]^4 < \infty$ uniformly in n. Under these model assumptions and assumptions A1-A9 and H_0, we have

$$S_K^L \xrightarrow{d} \sum_{k=1}^{K} \frac{\pi_k}{1-\pi_k} \exp\left\{Z_k^2/a(\psi^0)/2\right\} \tag{4.68}$$

as n tends to infinity where $Z_1, Z_2, \ldots, Z_K$ be i.i.d. standard normal random variables.

Proof. For $k = 1, \ldots, K$, where K is any integer, define $\widehat{\phi}_k$ as in (4.37). Now consider the following decomposition of S_K:

$$S_K^L = S_K + \Delta_{n1} \tag{4.69}$$

where

$$\Delta_{n1} = \sum_{k=1}^{K} \frac{\pi_k}{1-\pi_k} \exp\left\{n\widehat{\phi}_k^2/2a(\widehat{\psi}^0)\right\} [\exp(R_{kn}) - 1] \tag{4.70}$$

with $R_{kn} = (\mathcal{L}_{1k} - \mathcal{S}_k)/2$. Obviously,

$$|\Delta_{n1}| \leq \max_{1\leq k\leq K} |\exp(R_{kn}) - 1| \sum_{k=1}^{K} \frac{\pi_k}{1-\pi_k} \exp\left\{n\widehat{\phi}_k^2/2a(\widehat{\psi}^0)\right\} \tag{4.71}$$

By Taylor's theorem we have

$$\exp(R_{kn}) - 1 = R_{kn} \exp(\widetilde{R}_{kn}) \tag{4.72}$$

for $\widetilde{R}_{kn}$ such that $|\widetilde{R}_{kn}| \leq |R_{kn}|$. Define

$$\Delta_{n2} = \max_{1\leq k\leq K} |\exp(R_{kn}) - 1|, \tag{4.73}$$

which implies

$$\Delta_{n2} \leq \max_{1\leq k\leq K} |R_{kn}||\exp(R_{kn})|. \tag{4.74}$$

Noting that $R_{kn} = (\mathcal{L}_{1k} - \mathcal{S}_k)/2 = O_p(n^{-1/2})$ (Fahrmeir and Tutz, 2001), we find that $|\Delta_{n2}| = O_p(n^{-1/2})$ and hence $|\Delta_{n1}| = O_p(n^{-1/2})$.

□

Corollary 4.4.2 may appear intuitive and verifying it may initially seem pedantic. However, the result is useful and its validity is worth investigating. In particular, while the presence of chi-squared variables in the limit is predictable based on elementary asymptotic theory for parametric models, it is not as obvious that the weighted sum in the limit should be composed of *independent* exponentiated variates. One may conjecture independence based on the use of orthogonal basis functions in the construction of the test statistic, however, this would not be a sufficiently convincing observation to conclude that the individual chi-squared variates are truly independent. Furthermore, the independence of the exponentiated variates obviously makes simulating a reference distribution much more convenient. Without having established independence, one would be compelled to identify the nature of dependence among the exponentiated variates.

4.4.3 Choice of Prior Probabilities

In this section we briefly discuss some issues regarding specification of prior distributions when applying a Bayes sum statistic in practice. There are several important distinctions that need to be made between our case and the setting studied in Hart (2009) that influence the recommendations for specifying prior probabilities.

The distribution theory presented for the statistic studied in Hart (2009) was developed in a manner which permitted the upper bound of summation to increase to infinity with the

sample size. Moreover, it should be noted that each term of the weighted sum converges in distribution to $\exp\{\chi^2_1\}$ which does not have finite first moment and, thus, does not satisfy conditions required for application of the central limit theorem. However, by imposing suitable conditions on the prior probabilities Hart (2009) was able to invoke a result of Cline (1983) which ensures the convergence of infinite (weighted) sums of random variables of the type produced by the test statistic. Thus, Hart (2009) noted that taking the π_j's to be *proper* prior probabilities (a condition which is implicit in assumptions imposed on the prior probabilities) has a stabilizing effect on the statistic. Furthermore, Hart (2009) emphasized that these prior probabilities distinguish this test statistic from the BIC-based nonadaptive statistic studied in Aerts et al.(2004) and are responsible for the improved power observed in the simulation studies presented in Hart (2009).

Contrary to Hart (2009), we have assumed that K is fixed in our case. Thus, in the distribution theory we have presented in this chapter, the prior probabilities need not satisfy conditions such as those imposed in Hart (2009) to ensure convergence of the test statistic in our setting. However, there are some practical issues to keep in mind. In particular, one would typically presume that higher frequency departures from the null model would likely correspond to random noise. Thus, unless one were interested in detecting specific alternatives of interest, it would generally be advisable to specify the π_j's decrease monotonically to 0 as the frequency (i.e., j) increases.

4.5 Discussion

In this chapter we have sought to extend the ideas from Hart (2009) to the generalized linear model conditions addressed in Aerts et al. (2004). The result of this pursuit is a new lack-of-fit test for a special class of canonical link regression models. Our derivation and subsequent examination of the test statistic yielded several noteworthy theoretical findings.

Our first step in developing a test statistic was to formulate the posterior probability of a hypothesized model. The alternative models utilized in the construction of this posterior probability were based on characterizing departures from the predictor function in terms of Fourier coefficients. As we have noted, testing the hypothesis that the linear predictor has a specified parametric form is equivalent to testing that all of these coefficients are 0. A closed-form approximation of this posterior probability was then obtained by applying the Laplace approximation to the integrals that compose the marginal likelihood of the data. Rather than evaluating this probability directly, this statistic is used in frequentist fashion by means of a reference distribution. To this end, we examine the limit distribution of the statistic obtained from the posterior probability and show that, under the null hypothesis, this distribution is completely determined by the alternative models with the fewest parameters. This is noteworthy because the posterior probability is constructed from a very general, nonparametric class of alternative models. Upon recognition of the result involving the limiting distribution under the null, we propose a simplified test statistic that is a weighted sum of exponentiated likelihood ratios testing the effect of the additional Fourier terms. The weights depend on user-specified prior probabilities. From this statistic, we obtained a statistic that consists of a weighted sum of exponentiated squared Fourier coefficient estimates by substituting the likelihood ratios with their corresponding score statistics. We refer to both versions of the statistic as the “Bayes sum” statistics.

With two simplified versions of the test statistic having been identified, we turned our attention to studying the large sample properties of these statistics. Our study focused on the score-based statistic, and, in particular, we the examined asymptotic distribution of the Fourier coefficient estimators under local alternatives. To our knowledge, no such result has appeared in the literature for techniques addressing generalized linear models. We were subsequently able to characterize the limiting behavior of the score-based statistic under both the null hypothesis and local alternatives that converge to the null at rate

$1/\sqrt{n}$. Noting that under the null hypothesis, the score-based statistic provides a large sample approximation of the likelihood ratio-based test statistic, our conclusions reached for the former statistic led to a more convenient and accessible characterization of the null distribution for the latter statistic.

Finally, we offered some practical guidelines regarding the specification of prior probabilities in the test statistic. In presenting these guidelines, we noted some key differences between our statistic and the proposal of Hart (2009).

CHAPTER V

NUMERICAL RESULTS

5.1 Overview

In this chapter we will present a numerical study in order to obtain greater insight into our proposed method as well as several existing tests of fit. Our primary objective is to demonstrate the properties of the statistics presented in the previous chapter and assess their adequacy in detecting lack of fit in various situations. We will also address the need for further research on existing lack-of-fit tests based on orthogonal series, as has been noted in Aerts et al. (1999). Also, we intend to bridge a gap that exists between two parallel and distinct research efforts that have applications to testing the fit of the logistic regression model. Thus, we contend that our proposed numerical study will provide several new insights into testing the fit of the logistic regression model as well as establishing the power properties of some rather promising series-based tests against a broad range of departures from the null model.

Recognizing the trade-off between breadth and depth of such a study, we will opt for depth and pursue a thorough numerical study of one of the most widely used canonical link regression models, the logistic regression model. While the method presented in Chapter IV could be applied to any model satisfying the conditions described in Section 1.2 we will focus on investigating the performance of the statistic in the context of logistic regression because of the reported shortcomings with the better-known methods for testing the fit of such models. This approach to execution of our numerical study will serve three purposes.

First, by focusing on a single model, we will be able to examine the performance of our proposed test in detecting a wider variety of departures from the assumed model. In particular, we will study the selected test statistics in the simulation settings presented in

Hosmer et al. (1997) and Kuss (2002). Hosmer et al. and Kuss considered a wide variety of different situations that one might encounter in practice for the logistic regression model. Hence we will be able to obtain a more extensive picture of the power properties of the tests selected for our study than if we were to study a few settings for several different models. Furthermore, we will have assurance that the settings will actually be meaningful since we are duplicating situations that have been considered in studies from authoritative sources.

Second, while Aerts et al. (1999, 2000) demonstrate via simulation that their test possesses competitive power properties, they cite a need for more extensive numerical studies on existing lack-of-fit tests based on orthogonal series. To the best of our knowledge, further numerical studies have not yet been pursued. Thus, our proposed simulation study provides an opportunity to further explore the performance, applicability and limitations of these tests (particularly, the multivariate extensions) within the context of binary response regression. For example, while most of the simulation settings considered in Aerts et al. (1999), (2000) and (2004) were limited to testing no effect, we will examine a variety of departures from the logistic regression model which better reflect the types of model misspecification encountered in practice. Studying these tests will also permit us to evaluate the performance of our proposed method against the existing series-based lack-of-fit test literature. Consequently, we feel that our proposed simulation study will contribute to the understanding of existing series-based tests of fit as well as provide a deeper understanding of our proposed method of Chapter IV.

Finally, as the reader may have already noticed, much of the literature discussed in Chapter II predates the introduction of series-based techniques to accommodate generalized linear models discussed in Chapter III. Consequently, the performance of series-based tests of fit introduced in Aerts et al. (1999), (2000) and (2004) have not been compared to the performance of the tests presented in Chapter II for testing the fit of the logistic regression. By using the simulation settings which have been studied in Hosmer et al. (1997) and Kuss

(2002), we will be able to compare our findings with those which have been reported in these two sources. This will provide a means for comparison between the two collections of research.

The rest of this chapter will be organized as follows. In Section 5.2 we will use findings from numerical studies presented in the relevant literature to identify and prioritize a collection of test statistics that will be examined in our study. In Section 5.3 we study type I error and power properties of the series-based tests in the context of strictly sparse data (see Section 2.2). Power will be examined through progressively more severe departures from a hypothesized null model. In Section 5.4 power will be examined against fixed departures from the null model with gradual departures from sparsity. Section 5.5 presents an illustrative example of the method presented in Chapter IV. Finally, in Section 5.6 we will conclude with a discussion of our findings.

5.2 Test Statistics

In this section we identify the various test statistics we will use in our numerical studies and briefly discuss the rationale for examining those tests. These statistics will be limited to those that utilize orthogonal series estimators with applications to logistic regression models.

The primary statistic of interest in our study is the Bayes sum statistic studied in Chapter IV, which can be written as follows for logistic regression models:

$$S_K = \sum_{k=1}^{K} \frac{\pi_k}{1-\pi_k} \exp\left\{n\widehat{\phi}_k^2/2\right\} \tag{5.1}$$

where

$$\widehat{\phi}_k = \frac{1}{n}\sum_{i=1}^{n}[y_i - b'(\eta(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}^0))]\widehat{u}_k(\mathbf{x}_i), \;\; k = 1,\ldots,K \tag{5.2}$$

is the Fourier coefficient estimator for the singleton model. We took $\pi_k/(1-\pi_k) = k^{-2}$ and u_k's to be cosine functions.

As we discussed in Chapter III, the existing literature that addresses testing the fit of logistic regression models is essentially limited to Aerts et al. (1999, 2000, 2004), which cover methods applicable to canonical link regression models. The collection of statistics presented and studied in these papers is fairly extensive. Thus, we will economize our efforts by prioritizing statistics based on the performance and applicability of these statistics reported in the literature. Since Aerts et al. (2000) generalize the principle ideas of Aerts et al. (1999) to multivariate regression models, we will focus primarily on the findings of Aerts et al. (2000).

In their simulations, Aerts et al. (2000) studied the following statistics:

$$S_a = \mathcal{S}_{\widehat{r}_a}$$

where $\widehat{r}_a = \arg\max_{0\le r\le R_n} \mathrm{SIC}(r;2)$;

$$S_b = \mathcal{S}_{\widehat{r}_b}$$

where $\widehat{r}_b = \arg\max_{1\le r\le R_n} \mathrm{SIC}(r;\log n)$;

$$T_a = \frac{\mathcal{S}_{\widehat{r}_a} - \widehat{r}_a}{\max(1, \widehat{r}_a^{1/2})}; \quad T_{\mathrm{OS}} = \max_{1\le r\le R_n} \frac{\mathcal{S}_r}{r}; \quad T_{\max} = \mathrm{SIC}(\widehat{r}_a;2);$$

where SIC is the score information criterion and $\mathcal{S}_r$ represents the score statistic corresponding to the rth (nested) alternative model (see Chapter III). While likelihood-based criteria could have been considered, the score statistic used in constructing the criteria only requires fitting the null model which makes it particularly convenient for practical use. It is worth noting that Aerts et al. identified T_a and $T_{\max}$ as having the best overall power properties; however, they commented that more studies were required before making final

recommendations. Thus, we will include all of the above statistics.

We will exclude from our study the Bayesian-motivated statistic introduced in Aerts et al. (2004) which we reviewed in Section 3.5. This test was cited as possessing undesirable power properties in Aerts et al. (2004) and Hart (2009). Moreover, the limiting distribution for this test statistic is a stable distribution, which may be somewhat inconvenient for practical use.

Before proceeding to the simulation results, a few comments are in order regarding how we implement all of the statistics we will use in the simulation study. First, unless otherwise specified, we will present the results observed for each of the statistics cited above using cosine basis functions. The choice is somewhat arbitrary, but in order to avoid any possible confounding in the simulation results, we will use this collection of basis functions throughout the study. Second, we note that the order of the Bayes sum statistic, K, imposes an upper limit on the frequency of the cosine functions used in constructing the statistic. To ensure that cosine terms of the same frequencies are used to construct order selection tests, we impose the same upper bound on the frequency of the collection of nested alternatives used to obtain the order selection test. This is important since higher frequency basis terms might be more advantageous for detecting subtle departures from the null model and if higher frequencies were used for one statistic and not the others, then there would exist potential bias in favor of one constructed using the higher frequency basis functions. While we will use the same frequencies for all statistics we will, however, vary this common value of the upper bound in order to assess its possible impact on the frequency of rejection in the resulting statistics.

5.3 Simulation Results for Strictly Sparse Data

In this section we will present results of simulations to examine the behavior of the tests discussed in Section 5.2 when the observed data are strictly sparse (i.e., binary response without replications). The simulation settings under which these statistics will be studied are selected to replicate several of those studied in Hosmer et al. (1997). These simulation settings address each of the following main issues in the context of a logistic regression model:

1. the adequacy of the proposed null distribution of the statistics as well as the associated type I error rate;

2. power of the tests of fit to detect omission of a quadratic term;

3. power of the tests of fit to detect omission of the main effect for a dichotomous variable and its interaction with a continuous variable.

In carrying out our simulation study, 1000 random samples of sizes $n = 100$ or 500 were generated as follows. Each replicate data set is constructed by first generating the covariate values and then creating the outcome by comparing an independently generated value from the $U(0, 1)$ distribution, u, to the true logistic probability where $y = 1$ if $u \leq \pi(x)$ and $y = 0$ otherwise. Throughout our simulations we will take the simulation size to be 1000 and we will evaluate the significance at a level $\alpha = 0.05$. This yields yields a 95% margin of error of $1.96 \times \sqrt{0.05 * (1 - 0.05)/1000} = 0.014$ for the rejection probability estimate obtained in each simulation setting. A more conservative upper bound on this margin of error is, of course, given by $1.96 \times \sqrt{0.5 * (1 - 0.5)/1000} = 0.031$.

5.3.1 Evaluation of Null Distribution and Test Level

To examine the type I error rate of the tests, we consider several different situations where the data are generated from one of several different logistic regression models, each of which is based on the following general model:

$$\log\left(\frac{\pi(x_1, x_2, x_3)}{1 - \pi(x_1, x_2, x_3)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3. \tag{5.3}$$

Hosmer et al. (1997) chose the various distributions of the covariates and their corresponding coefficients to produce distributions of probabilities in the $(0, 1)$ interval that one might encounter in practice. Table 3 summarizes the various combinations of covariate distributions and true coefficient values for the logistic regression model (5.3), along with the resulting expected values for the smallest, largest, and three quartiles of the distribution of logistic probabilities for a sample of size 100. When $x_1 \sim U(-6, 6)$ with $\beta_0 = \beta_2 = \beta_3 = 0$, the resulting distribution of the probabilities is symmetric with mostly small or large probabilities. Taking $x_1 \sim U(-1, 1)$ produces a distribution with most probabilities in the center of the $(0, 1)$ interval. For $x_1 \sim \chi^2(4)$, yields a distribution mostly small probabilities and few large probabilities. The probabilities are more uniformly distributed for the other covariate distributions.

Evaluation of null distribution and test level for univariate models

The resulting per cent of times that Bayes sum and order selection-based tests lead to rejection of the null hypothesis for the univariate models summarized in Table 3 is reported in Table 4 and Table 5, respectively. Furthermore, we obtain the per cent of times each test in our study rejects the null hypothesis across several values of the truncation point, K. These values indicate that most of these tests reject the null hypothesis at a rate reasonably close to the nominal five percent level. Moreover, the per cent rejection was rather similar

Table 3. Situations used to examine the null distribution of test statistics. For each covariate distribution, $\pi_{(1)}$, $\pi_{(n)}$ and Q_1 Q_2 Q_3 are, respectively, the expected values for the smallest, largest values and the three quartiles of the distribution of logistic probabilities for a sample of size 100.

Covariate distribution	Logistic coefficients				Distribution of logistic probabilities				
	β_0	β_1	β_2	β_3	$\pi_{(1)}$	Q_1	Q_2	Q_3	$\pi_{(n)}$
$U(-6,6)$	0	0.8	0	0	0.009	0.087	0.50	0.913	0.991
$U(-4.5,4.5)$	0	0.8	0	0	0.029	0.144	0.50	0.865	0.971
$U(-3,3)$	0	0.8	0	0	0.087	0.231	0.50	0.769	0.913
$U(-1,1)$	0	0.8	0	0	0.313	0.400	0.50	0.600	0.687
$N(0,1.5)$	0	0.8	0	0	0.057	0.304	0.50	0.696	0.943
$\chi^2(4)$	-4.9	0.65	0	0	0.009	0.025	0.062	0.202	0.965
3 indep. $U(-6,6)$	0	$0.8/3$	$0.8/3$	$0.8/3$	0.028	0.234	0.50	0.767	0.972
3 indep. $N(0,1.5)$	0	$0.8/3$	$0.8/3$	$0.8/3$	0.134	0.369	0.50	0.628	0.861
Indep. $U(-6,6)$, $N(0,1.5)$ and $\chi^2(4)$	-1.3	$0.8/3$	$0.8/3$	$0.65/3$	0.052	0.204	0.386	0.608	0.928

across the various values considered for the truncation point.

Perhaps most problematic test is the one based on the score analog of the BIC, S_b. This test clearly tends to reject too often at a sample size of $n = 100$; however, the per cent of times the null hypothesis is rejected approaches the nominal $\alpha = 0.05$ when the sample size increases to 500. This reflects the findings of Aerts et al. (1999, 2000) who found that the type I error rate for this test was inflated in the settings considered therein.

The one univariate simulation setting which was uniformly problematic across all tests was the case with covariate distributed as $\chi^2(4)$. In this case none of the tests appears to reject the null hypothesis often enough. Hosmer et al. (1997) found that the statistics considered in their study tended to vary about the nominal five percent level, yet the observed departure of these statistics from $\alpha = 0.05$ level was typically no less than the departure observed for the series-based tests.

Table 4. Performance of the Bayes sum test when the correct (and fitted) logistic model is determined by (5.3) with corresponding coefficient values specified in Table 3. Simulated per cent rejection at the $\alpha = 0.05$ level is reported using sample sizes of 100 and 500 with 1000 replications. For each covariate distribution, per cent rejection was evaluated at various values of truncation point, K.

Statistic	K	covariate distribution											
		$U(-6,6)$		$U(-4.5,4.5)$		$U(-3,3)$		$U(-1,1)$		$N(0,1.5)$		$\chi^2(4)$	
		100	500	100	500	100	500	100	500	100	500	100	500
S_K	2	4.3	5.3	4.0	6.1	4.8	4.2	4.6	4.4	5.1	4.6	2.4	3.5
	3	5.5	5.7	4.9	6.2	4.9	4.5	4.7	4.3	5.9	5.6	2.5	3.3
	4	5.6	5.7	4.8	6.0	5.0	4.6	4.8	4.2	5.9	5.6	2.6	3.4
	7	5.4	5.9	4.5	5.9	4.6	4.8	4.1	4.2	6.4	5.9	2.6	3.1
	11	5.3	5.7	4.2	5.7	4.2	4.6	4.1	3.9	6.2	6.0	2.6	3.3
	15	5.4	6.0	4.4	6.0	4.3	4.6	4.2	4.2	6.7	6.0	2.6	3.2

Finally, we can see from Table 4 that the Bayes sum test exhibits behavior similar to the other order selection-based tests; however, it appears to be more conservative than these tests.

Evaluation of null distribution and test level for multivariate models

For the multivariate models, we limited the truncation point to be no greater than 4 since a truncation point greater than 4 would require inclusion of at least 125 additional terms to the alternative models, which is infeasible for a sample of size $n = 100$. Also, for the order-selection based statistics, we consider generalizations of two of the model sequences recommended by Aerts et al. (2000): sequences (c) and (d) illustrated in Figure 1 of Chapter III. For the Bayes sum statistic, we apply the following multivariate extension of formula 5.1

$$S_K = \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{l=1}^{K} \frac{\pi_{jkl}}{1-\pi_{jkl}} \exp\left\{n\widehat{\phi}^2_{jkl}/2\right\} \tag{5.4}$$

Table 5. Performance of the order selection-based test when the correct (and fitted) logistic model is determined by (5.3) with corresponding coefficient values specified in Table 3. Simulated per cent rejection at the $\alpha = 0.05$ level is reported using sample sizes of 100 and 500 with 1000 replications. For each covariate distribution, per cent rejection was evaluated at various values of truncation point, K.

Statistic	K	covariate distribution											
		$U(-6,6)$		$U(-4.5,4.5)$		$U(-3,3)$		$U(-1,1)$		$N(0,1.5)$		$\chi^2(4)$	
		100	500	100	500	100	500	100	500	100	500	100	500
S_a	2	5.3	5.2	4.8	4.6	5.0	3.7	5.9	4.0	5.1	5.4	2.5	3.0
	3	7.3	7.0	5.2	5.8	6.0	4.9	5.9	4.4	6.1	5.0	2.6	2.7
	4	7.3	8.0	5.7	5.7	6.0	5.4	5.2	5.2	7.1	6.4	2.6	2.8
	7	6.2	6.1	5.7	7.0	5.3	5.1	4.8	4.0	7.9	7.6	3.1	2.4
	11	6.3	6.5	5.2	6.1	4.4	5.3	5.1	4.3	8.4	7.5	3.3	2.4
	15	6.6	6.5	5.3	5.7	4.4	5.1	4.8	4.3	9.1	8.8	3.6	2.5
S_b	2	6.0	6.2	6.2	6.7	7.6	5.1	8.6	5.3	8.5	5.7	3.0	4.1
	3	7.2	6.6	7.4	6.9	8.8	5.6	9.3	5.8	9.0	6.3	3.1	4.3
	4	7.2	6.3	7.2	6.8	8.5	5.2	9.0	5.4	9.5	6.3	3.2	4.4
	7	7.1	6.6	7.3	6.9	8.4	5.3	9.0	5.7	10.1	6.4	3.2	4.4
	11	7.2	6.6	7.3	6.9	8.6	5.3	9.2	5.7	10.3	6.3	3.2	4.4
	15	7.2	6.8	7.3	6.9	8.7	5.7	9.2	5.9	9.3	6.3	3.3	4.4
T_a	2	4.9	5.2	4.7	6.1	5.0	3.9	4.9	3.8	5.3	5.3	2.5	3.2
	3	6.1	5.8	4.9	6.1	5.2	4.3	4.8	4.1	6.3	6.2	2.6	3.3
	4	6.1	6.1	4.7	6.1	5.3	4.4	4.8	4.2	6.8	6.2	2.7	3.1
	7	5.9	7.3	4.6	6.3	5.6	4.8	5.1	4.0	8.2	7.2	3.0	2.9
	11	6.1	6.9	4.8	6.4	5.3	4.6	5.1	4.1	8.0	7.1	3.2	3.2
	15	6.0	6.6	4.7	6.2	5.2	4.6	5.1	3.9	8.0	7.6	3.2	3.2
T_{OS}	2	4.4	5.2	4.1	5.7	5.2	4.2	5.1	4.3	5.8	4.7	2.4	3.9
	3	5.0	5.4	4.5	5.9	5.2	4.3	4.7	4.4	5.8	5.4	2.6	3.7
	4	5.1	5.5	4.6	5.9	5.3	4.4	4.7	4.3	6.0	5.2	2.7	4.0
	7	5.1	5.9	4.7	5.9	5.2	4.6	4.7	4.5	6.4	5.5	2.8	4.2
	11	5.2	5.8	4.9	5.9	5.5	4.4	5.1	4.3	6.6	5.3	2.7	4.1
	15	5.1	5.8	4.7	5.9	5.3	4.6	4.7	4.5	5.5	5.6	2.7	3.8
$T_{\max}$	2	5.1	5.1	4.7	5.8	4.8	3.8	4.9	3.6	5.0	5.5	2.7	3.1
	3	6.5	6.1	5.8	5.9	5.6	4.3	5.2	4.4	6.3	6.0	2.7	3.1
	4	6.0	6.2	4.8	5.9	5.4	4.4	5.1	4.2	6.6	6.2	2.6	3.0
	7	6.2	7.1	4.5	6.7	5.8	4.7	5.7	4.1	7.7	7.2	2.7	2.8
	11	6.3	6.8	4.7	6.4	5.4	4.3	5.6	4.1	7.8	7.1	2.8	2.9
	15	6.4	6.6	4.9	6.1	5.4	4.0	5.3	3.5	7.8	7.6	3.0	2.7

Table 6. Performance of the Bayes sum test when the correct (and fitted) logistic model is determined by (5.3) with corresponding coefficient values specified in Table 3. Simulated per cent rejection at the $\alpha = 0.05$ level is reported using sample sizes of 100 and 500 with 1000 replications. For each covariate distribution, per cent rejection was evaluated at various values of truncation point, K.

Statistic	K	covariate distribution					
		3 indep. $N(0, 1.5)$		3 indep. $N(0, 1.5)$		Indep. $U(-6, 6)$ $N(0, 1.5)$, $\chi^2(4)$	
		100	500	100	500	100	500
S_K	2	4.7	4.4	5.8	4.7	6.3	4.4
	3	6.0	4.2	5.3	4.1	6.6	6.7
	4	6.0	4.2	6.2	4.8	7.7	7.9

where $\widehat{\phi}_{jkl}$ is the Fourier coefficient estimator for the singleton model using cosine basis function. We took $\pi_{jkl}/(1 - \pi_{jkl}) = (jkl)^{-2}$.

The resulting per cent of times that Bayes sum and order selection-based tests led to rejection of the null hypothesis for the hypothesized (and true) multivariate models summarized in Table 3 is reported in Table 6 and Table 7, respectively. Furthermore, we obtain the per cent of times each test in our study rejects the null hypothesis across several values of the truncation point, K. These values indicate that most of these tests reject the null hypothesis at a rate reasonably close to the nominal five percent level. The type I error rate appears to typically be slightly inflated when $n = 100$, but not by an alarming amount. Moreover, the per cent rejection was rather similar across the various values considered for the truncation point.

5.3.2 Detecting Omission of a Quadratic Term

To evaluate the power to detect the omission of a quadratic term, Hosmer et al. (1997) utilized the following simulation strategy. For each value of the covariate, the response was generated using the following model:

Table 7. Performance of the order selection-based tests when the correct (and fitted) logistic model is determined by (5.3) with corresponding coefficient values specified in Table 3. Simulated per cent rejection at the $\alpha = 0.05$ level is reported using sample sizes of 100 and 500 with 1000 replications. For each covariate distribution, per cent rejection was evaluated at various values of truncation point, K.

Statistic	K	covariate distribution					
		3 indep. $N(0, 1.5)$		3 indep. $N(0, 1.5)$		Indep. $U(-6, 6)$ $N(0, 1.5)$, $\chi^2(4)$	
		100	500	100	500	100	500
S_a,	2	7.0	5.5	5.8	5.5	6.9	5.2
sequence (c)	3	6.7	5.7	5.7	5.0	6.8	7.1
	4	6.9	5.7	4.9	5.2	6.7	7.6
S_a,	2	7.0	5.5	5.8	5.5	6.9	5.2
sequence (d)	3	7.1	5.7	5.5	5.0	6.4	7.2
	4	7.1	5.8	4.7	5.2	6.8	7.6
S_b,	2	6.7	4.9	6.5	5.5	6.0	5.6
sequence (c)	3	6.4	5.6	6.6	5.6	6.1	5.7
	4	6.5	5.0	6.3	5.8	6.1	5.8
S_b,	2	6.7	4.9	6.5	5.5	6.0	5.6
sequence (d)	3	6.4	5.6	6.6	5.6	6.1	5.7
	4	6.5	5.0	6.3	5.8	6.1	5.8
T_a,	2	6.2	5.5	5.2	6.3	6.4	6.0
sequence (c)	3	6.3	4.9	5.6	5.4	7.2	7.4
	4	6.1	4.9	5.5	5.6	7.0	7.6
T_a,	2	6.2	5.5	5.2	6.3	6.4	6.0
sequence (d)	3	6.7	5.0	5.6	4.5	7.2	7.1
	4	6.4	5.1	5.1	5.5	6.9	7.5
T_{OS},	2	5.9	4.7	6.0	5.3	6.0	6.3
sequence (c)	3	5.9	5.1	6.1	5.4	6.4	6.8
	4	5.9	4.4	5.6	5.5	6.2	6.8
T_{OS},	2	5.9	4.7	6.0	5.3	6.0	6.3
sequence (d)	3	5.9	5.1	6.2	5.4	6.4	6.9
	4	5.9	4.4	5.7	5.5	6.3	7.0
$T_{\max}$,	2	6.2	5.1	5.6	5.8	6.4	6.1
sequence (c)	3	6.0	5.1	5.5	5.2	6.7	7.0
	4	5.9	5.0	5.2	5.6	6.8	7.3
$T_{\max}$,	2	6.2	5.1	5.6	5.8	6.4	6.1
sequence (d)	3	5.9	5.0	5.5	5.2	6.9	7.0
	4	5.8	5.1	5.3	5.7	6.9	7.4

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x + \beta_2 x^2 \tag{5.5}$$

where the distribution of the covariate, x, will be taken to be $U(-3,3)$. Hosmer et al. chose the values of the coefficients so that $\pi(-1.5) = 0.05$, $\pi(3) = 0.95$ and $\pi(-3) = J$ and $J = 0.01, 0.05, 0.1, 0.2$ and 0.4. The coefficients satisfying these conditions are presented in Table 8. This scheme produces models for which the departure from linearity becomes progressively more pronounced.

Table 8. Coefficients used in (5.5) to evaluate power to detect the omission of a quadratic term.

J	Logistic coefficients		
	β_0	β_1	β_2
0.01	-1.138	1.257	0.035
0.05	-1.963	0.981	0.218
0.10	-2.337	0.857	0.301
0.20	-2.742	0.722	0.391
0.40	-3.232	0.558	0.500

Tables 9 and 10 summarize the per cent rejected across the various departures from linearity for the Bayes sum and order selection tests, respectively. For both types of statistic, we see that the power is rather low for smaller values of J. Power does increase rather rapidly as the departure from linearity becomes more pronounced. Hosmer et al. (1997) found that all of the tests they studied, except the Royston monotone test (see Section 2.4), exhibited a similar increase in power as the departure from linearity increased. Most of the series-based tests detect an omitted quadratic term at least as well as the best performing test statistic considered in Hosmer et al. The one clear exception is the test based on the

Table 9. Performance of the Bayes sum test in detecting an omitted quadratic term. Simulated per cent rejection at the $\alpha = 0.05$ level using sample sizes of 100 and 500 with 1000 replications are reported. For each covariate distribution per cent rejection was evaluated at various values of truncation point. See text for definition of J.

Statistic	K	correct model									
		$J = 0.01$		$J = 0.05$		$J = 0.1$		$J = 0.2$		$J = 0.4$	
		100	500	100	500	100	500	100	500	100	500
S_K	2	6.6	8.0	35.6	90.5	57.4	99.6	85.2	100	97.3	100
	3	7.0	8.1	36.6	91.1	57.9	99.7	85.4	100	97.3	100
	4	7.8	7.9	36.6	90.8	58.0	99.7	85.5	100	97.3	100
	7	7.9	7.8	36.4	90.5	57.3	99.5	85.0	100	97.2	100
	11	7.6	7.7	36.0	90.1	56.9	99.4	84.5	100	97.1	100
	15	7.8	7.9	36.3	90.3	57.6	99.4	84.8	100	97.1	100

score analog to the AIC, S_a, which is still rather competitive for smaller values of the truncation point (i.e., $K = 2, 3, \text{ and } 4$).

Finally, power is enhanced markedly with an increase in sample size, however, this is to be expected. For samples of size 500 power is about 90 percent or greater for even slight departures from linearity.

5.3.3 Detecting Omission of a Dichotomous Variable and Its Interaction

To evaluate the power to detect the omission of the main effect for a dichotomous variable and its interaction with a continuous variable, the following simulation setting was studied by Hosmer et al. (1997). For each combination of realized values for the covariates, the response was generated using the following model:

$$\log\left(\frac{\pi(x,d)}{1-\pi(x,d)}\right) = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 xd \tag{5.6}$$

Table 10. Performance of the order selection-based tests in detecting an omitted quadratic term. Simulated per cent rejection at the $\alpha = 0.05$ level using sample sizes of 100 and 500 with 1000 replications are reported. For each covariate distribution per cent rejection was evaluated at various values of truncation point. See text for definition of J.

Statistic	K	correct model									
		$J = 0.01$		$J = 0.05$		$J = 0.1$		$J = 0.2$		$J = 0.4$	
		100	500	100	500	100	500	100	500	100	500
S_a	2	5.7	9.3	33.8	88.9	56.0	99.3	81.0	100	97.1	100
	3	6.6	8.1	32.4	88.2	53.6	99.3	77.7	100	96.8	100
	4	8.7	8.1	31.6	86.2	51.0	99.0	76.3	100	95.4	100
	7	8.8	7.1	31.4	79.2	47.4	98.5	73.6	100	92.7	100
	11	8.3	7.7	30.8	77.8	47.0	98.1	69.9	100	91.0	100
	15	8.4	8.1	30.5	76.4	46.7	97.6	67.6	100	89.8	100
S_b	2	9.0	9.5	39.9	90.5	62.0	99.4	86.5	100	97.7	100
	3	9.6	9.9	40.9	90.9	62.3	99.4	87.0	100	97.7	100
	4	9.7	9.5	40.8	90.7	62.2	99.4	86.9	100	97.7	100
	7	9.8	9.8	40.8	91.0	62.2	99.4	86.8	100	97.7	100
	11	9.9	9.8	41.0	90.9	62.3	99.4	87.0	100	97.7	100
	15	9.9	10.0	41.0	91.0	62.4	99.4	87.0	100	97.7	100
T_a	2	6.6	8.0	36.2	90.2	57.9	99.6	84.6	100	97.4	100
	3	7.2	8.1	36.3	91.4	58.1	99.7	83.8	100	97.5	100
	4	8.0	8.5	35.8	90.7	57.8	99.5	83.3	100	97.3	100
	7	8.3	8.9	35.8	90.4	58.0	99.4	82.6	100	97.1	100
	11	8.6	8.6	36.7	89.9	58.3	99.4	83.1	100	96.8	100
	15	8.7	8.9	36.8	89.3	58.1	99.4	82.8	100	96.9	100
T_{OS}	2	6.9	8.0	37.0	90.5	58.7	99.6	85.4	100	97.3	100
	3	6.9	8.0	37.0	90.8	58.3	99.6	85.4	100	97.3	100
	4	7.2	7.8	37.1	90.5	58.3	99.6	85.4	100	97.3	100
	7	7.5	8.2	37.3	90.9	58.4	99.6	85.5	100	97.3	100
	11	7.7	7.8	37.8	90.6	58.9	99.6	85.7	100	97.3	100
	15	7.5	8.0	37.5	90.7	58.6	99.6	85.6	100	97.3	100
$T_{\max}$	2	6.6	8.0	35.7	90.1	57.8	99.6	84.3	100	97.3	100
	3	7.1	8.0	36.3	90.9	58.0	99.6	83.5	100	97.3	100
	4	8.4	8.9	36.1	91.2	57.5	99.4	83.2	100	97.2	100
	7	8.3	9.6	35.7	90.9	58.4	99.4	81.9	100	97.0	100
	11	8.3	9.1	36.4	90.0	58.0	99.4	81.9	100	97.1	100
	15	8.5	8.7	36.8	89.7	57.8	99.4	82.0	100	97.1	100

where the distribution of the continuous covariate, x, will be taken to be $U(-3,3)$ and the dichotomous covariate, d, is generated from the Bernoulli$(1/2)$ and was independent of the continuous covariate. Hosmer et al. (1997) chose the four parameters such that $\pi(-3,0) = 0.1$, $\pi(-3,1) = 0.1$, $\pi(3,0) = 0.2$ and $\pi(3,1) = 0.2 + I$ where $I = 0.1, 0.3, 0.5$ and 0.7. The coefficients satisfying these conditions are presented in Table 11. Thus, the interaction becomes progressively more pronounced across the four models.

Table 11. Coefficients used in (5.6) to evaluate power to detect the omission of a interaction term.

J	Logistic coefficients			
	β_0	β_1	β_2	β_3
0.10	-1.792	0.135	0.269	0.090
0.30	-1.792	0.135	0.693	0.231
0.50	-1.792	0.135	1.117	0.372
0.70	-1.792	0.135	1.792	0.597

The resulting empirical power of the Bayes sum and order selection tests are reported in Table 12 and Table 13, respectively. Unfortunately, all of the series-based tests exhibit poor power properties. Despite the fact that the power of the tests does increase modestly for larger sample sizes and when the departure from the simple linear logistic regression model increases, the power is too low to assure us that these tests will be able to detect this type of misspecification in practice. In this simulation setting, the power properties of the series-based tests are somewhat similar to those reported in Hosmer et al. (1997) for the statistics studied therein. One noteworthy difference is that for $I = 0.1, 0.3, \text{ and } 0.5$ per cent rejection actually decreases as the sample size increases. In particular, the per cent of times the null hypothesis is rejected appears to converge to the nominal five percent level for several of the tests. In spite of this behavior, none of the tests of Hosmer et al. reject appreciably more than the series-based tests.

Table 12. Performance of the Bayes sum test in detecting an omitted dichotomous variable and its interaction. Simulated per cent rejection at the $\alpha = 0.05$ level using sample sizes of 100 and 500 with 1000 replications are reported. For each covariate distribution per cent rejection was evaluated at various values of truncation point. See text for definition of I.

Statistic	K	$I = 0.1$		$I = 0.3$		$I = 0.5$		$I = 0.7$	
		100	500	100	500	100	500	100	500
S_K	2	6.0	4.3	5.3	4.3	6.0	4.0	6.9	12.0
	3	6.0	4.2	5.5	4.5	6.1	4.6	6.6	11.7
	4	6.0	4.4	5.4	4.6	6.1	4.9	6.0	11.5
	7	5.3	4.0	5.2	4.4	5.7	4.8	6.0	11.4
	11	4.9	4.1	5.1	4.8	5.7	4.9	5.7	11.7
	15	5.6	4.3	5.3	4.9	6.0	4.9	6.3	11.7

To better understand the power properties in this setting, recall that the series-based test statistics assess departures from the linear predictor specified in the null hypothesis by utilizing alternatives that only involve variables included in the null model. This formulation of the test statistics suggests that the series-based tests should be at a disadvantage in detecting departures involving variables that are not included in the null model. Identifying the cause of low power is not as immediately obvious for some of the tests studied in Hosmer et al.

5.4 Simulation Results Under Departures from Sparsity

In this section we will present results of simulations to examine the behavior of the tests discussed in Section 5.2 under departures from strictly sparse data. The simulation settings under which these statistics will be studied are selected to replicate several of those studied in Kuss (2002). These simulation settings address each of the following main issues in the context of a logistic regression:

Table 13. Performance of the order selection-based tests in detecting an omitted dichotomous variable and its interaction. Simulated per cent rejection at the $\alpha = 0.05$ level using sample sizes of 100 and 500 with 1000 replications are reported. For each covariate distribution per cent rejection was evaluated at various values of truncation point. See text for definition of I.

Statistic	K	$I = 0.1$		$I = 0.3$		$I = 0.5$		$I = 0.7$	
		100	500	100	500	100	500	100	500
S_a	2	5.3	5.1	5.9	4.5	6.8	4.4	8.2	10.3
	3	5.2	5.3	6.0	4.4	5.5	4.0	7.6	8.9
	4	4.9	5.8	6.1	5.2	5.4	4.9	6.1	7.4
	7	4.2	5.6	4.6	4.7	4.8	4.3	4.4	6.0
	11	4.6	5.3	5.7	5.0	4.5	4.7	4.2	5.3
	15	4.7	5.3	6.0	4.4	4.6	4.3	4.5	5.2
S_b	2	8.8	6.1	8.6	5.2	9.5	4.6	11.5	13.5
	3	9.1	6.5	8.5	5.6	9.7	5.1	11.3	14.1
	4	9.5	6.5	9.0	5.7	10.0	5.1	11.8	14.2
	7	9.7	6.5	9.3	5.7	10.4	5.1	12.2	14.2
	11	9.7	6.3	9.4	5.3	10.6	4.8	12.4	13.7
	15	9.2	6.4	8.4	5.4	9.8	4.9	11.1	13.8
T_a	2	6.0	4.5	5.8	4.1	6.4	4.6	7.2	12.2
	3	5.6	5.1	5.3	4.3	5.9	5.1	7.0	12.1
	4	4.9	5.5	5.1	4.8	5.6	5.1	6.9	10.3
	7	4.3	5.2	5.2	5.0	5.9	4.9	6.5	10.2
	11	4.4	5.2	5.1	5.2	6.5	5.1	6.2	9.8
	15	4.2	5.1	4.4	4.5	5.6	5.1	5.6	9.3
T_{OS}	2	6.1	4.5	5.3	4.3	6.3	3.8	7.4	12.5
	3	6.2	4.7	5.0	4.5	5.9	4.2	7.0	12.5
	4	6.4	4.5	5.3	4.6	6.1	4.1	7.1	12.1
	7	6.4	4.5	5.3	4.6	6.1	4.2	7.1	12.1
	11	6.5	4.4	5.5	4.3	6.3	4.0	7.3	11.8
	15	5.8	4.5	5.1	4.6	5.4	4.2	6.3	12.2
$T_{\max}$	2	5.9	5.0	6.0	4.2	6.6	4.5	7.4	11.9
	3	5.0	5.0	5.6	4.4	6.0	4.9	7.3	11.5
	4	4.7	5.5	5.2	4.5	6.0	5.3	7.0	10.5
	7	4.4	5.3	5.3	4.7	5.2	4.4	6.1	9.1
	11	3.9	5.3	5.6	5.2	5.7	4.2	6.0	8.6
	15	3.7	5.4	4.6	5.2	5.4	4.4	5.7	8.6

1. Missing covariate.

2. Wrong functional form of the covariate.

3. Misspecified link function.

4. Overdispersion.

Each of the above issues will be studied by examining a fixed departure from a hypothesized model and varying the number of observations sharing a given covariate pattern. Recall from Section 2.2 that we denote this number by m_i. Examination of varying m_i was of interest in Kuss (2002) because several well-know test statistics that worked well for sparse data have exhibited problems in data sets that were not sparse (see Chapter II). Conversely, several other statistics that work well for non-sparse data are invalid for strictly sparse data. By construction, the order selection and Bayes sum statistics should be valid regardless of whether the data are sparse or not. However, we feel that it is of important to examine the sensitivity of the series-based tests under varying levels of sparsity in order to assess if there are appreciable changes in power across the various values of m_i.

For each simulation setting and level of sparsity, 1000 random samples of sizes, $n = 100$ or 500 were generated in a manner similar to that described earlier in this chapter for sparse data. In order to evaluate the effect of sparseness on each test, Kuss (2002) varied the number of individuals within the data set sharing the same covariate pattern. For four of his simulation settings, Kuss used values of $m_i = 1, 2, 5, 10$, where m_i denotes the number of times each covariate pattern is observed within the sample (see Section 2.2). In addition to these settings, Kuss studied a setting (labeled 1-2) in which half of the covariate patterns within each replicated sample have a single observation and the other half have two observations. Another setting (labeled 1-10) has roughly 64 per cent of the covariate patterns have a single observation, 21 per cent have two observations, 9 per cent have

five observations and 6 per cent have ten observations. To better clarify this last setting, consider the case where $n = 100$: 30 distinct covariate patterns were observed once each ($m_i = 1$), 10 distinct covariate patterns were observed two times each (i.e., 10 covariate patterns with $m_i = 2$ yielding 20 observations), 4 distinct covariate patterns were observed five times each ($m_i = 5$), and 3 distinct covariate patterns were observed 10 ten times each ($m_i = 10$), resulting in $30 + 20 + 20 + 30 = 100$ individual observations from $30 + 10 + 4 + 3 = 47$ distinct covariate patterns. Obviously, for $n = 500$, multiply all the numbers by 5. This last constellation was formulated to reflect a distribution across covariate patterns which is often encountered in practice (Kuss, 2002).

Missing covariate

To evaluate the power to detect a missing covariate across various departures from sparsity, Kuss (2002) generated data using the following model:

$$\log\left(\frac{\pi(x_1, x_2)}{1 - \pi(x_1, x_2)}\right) = 0 + 0.405x_1 + 0.223x_2; \tag{5.7}$$

where $x_i \sim U(-6, 6)$, $i = 1, 2$. We fit the following simple logistic regression model to the resulting data: $\log(\pi(x)/(1 - \pi(x))) = \beta_0 + \beta_1 x_1$. That is, we fit the model as if x_2 had been excluded in the model building process.

The resulting empirical power of the Bayes sum and order selection tests are reported in Table 14 and Table 15, respectively. It is clear that the series-based tests do not exhibit particularly good power in detecting a missing covariate. However, this is not surprising in light of the observed power of the series-based test to detect the omission of the main effect for a dichotomous variable and its interaction with a continuous variable: in both cases a variable had been suppressed. Clearly, as we move further away from strictly sparse data, the power of the series-based tests does improve. Unfortunately, these tests do not improve

Table 14. Performance of the Bayes sum test in detecting missing covariate. Simulated per cent rejection at the $\alpha = 0.05$ level using sample sizes of 100 and 500 with 1000 replications are reported. For each covariate distribution per cent rejection was evaluated at various values of truncation point. See text for explanation of the various constellations m_i.

Statistic	K	Constellation of m_i											
		1		1-2		2		1-10		5		10	
		100	500	100	500	100	500	100	500	100	500	100	500
S_K	2	6.0	4.7	4.5	5.9	6.1	6.4	8.1	11.2	12.9	9.2	19.2	18.7
	3	6.5	4.9	4.8	6.5	7.3	6.2	8.2	12.8	13.8	10.7	23.0	21.6
	4	6.5	4.7	4.6	5.8	7.5	6.2	9.3	13.6	14.2	12.0	24.6	24.4
	7	6.2	4.9	4.3	5.7	7.3	6.7	8.8	13.8	14.8	12.5	28.6	26.5
	11	5.7	5.0	4.3	5.6	6.7	6.2	8.2	14.2	14.7	13.1	30.1	29.0
	15	6.1	5.2	4.5	5.8	7.4	6.5	9.1	13.4	15.8	13.4	28.8	31.2

sufficiently to provide any assurance that these tests would be able to detect a missing covariate in practice. The power was consistent across the various values of the truncation point considered.

Kuss (2002) found that Osius-Rojek, McCullagh and Farrington tests have good power for a missing covariate when the sample size is 500 (recall from Chapter II that these three tests are somewhat similar in that they are all based on various modifications of the Pearson chi-square test). Note that no test among the collection discussed in Kuss (2002) was observed to have good power for sample size of 100.

Wrong functional form

To evaluate the power to detect a missing covariate across various departures from sparsity, Kuss (2002) generated data using the following model:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = 0.405x^2 \tag{5.8}$$

Table 15. Performance of the order selection-based tests in detecting missing covariate. Simulated per cent rejection at the $\alpha = 0.05$ level using sample sizes of 100 and 500 with 1000 replications are reported. For each covariate distribution per cent rejection was evaluated at various values of truncation point. See text for explanation of the various constellations m_i.

Statistic	K	Constellation of m_i											
		1		1-2		2		1-10		5		10	
		100	500	100	500	100	500	100	500	100	500	100	500
S_a	2	5.8	5.2	5.4	5.6	5.9	7.2	9.0	12.3	11.5	9.8	18.3	20.4
	3	6.3	5.2	5.6	5.5	7.1	6.8	11.0	13.8	14.8	11.4	23.4	24.2
	4	6.2	4.6	5.9	4.9	8.5	5.7	10.6	14.3	14.2	14.0	28.6	26.8
	7	5.4	4.5	6.1	5.7	9.4	6.9	12.5	14.7	16.5	15.9	34.5	34.6
	11	5.0	3.5	5.9	5.0	8.1	7.0	12.7	17.6	21.2	19.3	47.0	43.6
	15	4.9	3.4	6.0	5.5	7.2	7.0	12.9	18.6	23.4	21.8	62.9	50.0
S_b	2	9.6	6.0	9.2	7.2	8.7	7.3	12.6	14.3	16.9	11.0	24.7	21.5
	3	10.3	6.1	9.6	7.4	10.2	7.6	13.6	14.8	18.4	12.1	27.8	23.4
	4	10.8	6.1	10.0	7.4	10.0	7.5	13.8	14.5	18.6	12.5	29.6	23.6
	7	11.3	6.1	9.9	7.3	10.0	7.6	13.5	15.3	18.9	12.0	30.5	24.5
	11	11.4	6.0	10.5	7.7	10.4	7.6	13.7	14.1	19.0	12.9	30.3	23.4
	15	10.2	6.1	9.9	7.3	10.4	7.8	13.9	14.7	18.8	12.0	31.0	24.6
T_a	2	7.0	5.1	4.7	5.6	6.5	6.2	8.2	12.6	12.3	9.7	19.6	20.2
	3	7.3	5.5	4.2	6.2	7.6	6.0	8.8	14.0	13.7	11.0	22.6	23.1
	4	7.2	5.1	5.2	6.2	8.3	5.8	9.6	14.0	14.2	11.9	26.9	26.7
	7	6.6	5.1	5.3	5.5	8.4	6.7	10.6	16.3	16.7	14.6	32.5	32.2
	11	5.7	4.8	5.1	5.5	8.2	6.1	11.0	17.1	18.7	17.7	40.0	39.4
	15	5.0	4.7	5.6	6.2	7.7	5.9	11.9	18.7	19.8	18.5	51.8	46.7
T_{OS}	2	6.4	5.2	5.1	6.3	5.9	5.8	9.0	12.4	12.6	8.9	20.0	17.5
	3	6.2	5.1	5.0	6.2	6.4	5.9	9.3	12.8	13.3	10.2	20.5	19.5
	4	6.4	4.9	5.9	6.3	6.5	5.8	8.5	12.5	12.6	10.6	21.9	20.2
	7	6.4	4.9	4.5	5.3	6.6	6.0	8.5	13.7	13.6	11.0	23.1	21.8
	11	6.6	4.6	5.7	6.4	6.7	5.8	8.5	12.2	13.9	11.7	22.8	20.8
	15	5.7	4.9	5.3	6.2	6.7	5.8	9.1	12.8	13.2	10.8	23.6	22.3
$T_{\max}$	2	7.1	5.2	4.7	6.0	6.0	6.1	8.1	12.4	12.3	9.4	19.9	20.8
	3	7.0	5.4	4.3	6.1	7.3	5.6	9.2	14.6	13.9	11.7	22.5	24.9
	4	6.9	4.9	4.9	6.2	7.8	5.8	9.8	14.8	14.3	13.1	27.4	27.9
	7	6.3	4.8	5.2	5.9	9.0	6.6	11.9	16.2	16.4	14.9	33.0	33.8
	11	6.0	4.9	5.4	5.1	9.1	6.3	12.1	17.2	18.5	17.2	40.7	39.6
	15	5.0	4.8	5.8	5.6	9.0	6.3	12.0	17.9	19.3	17.7	51.3	45.6

Table 16. Performance of the Bayes sum test in detecting wrong functional form of the covariate. Simulated per cent rejection at the $\alpha = 0.05$ level using sample sizes of 100 and 500 with 1000 replications are reported. For each covariate distribution per cent rejection was evaluated at various values of truncation point. See text for explanation of the various constellations m_i.

Statistic	K	Constellation of m_i											
		1		1-2		2		1-10		5		10	
		100	500	100	500	100	500	100	500	100	500	100	500
S_K	2	84.2	100	87.1	100	89.7	100	88.6	100	94.6	100	97.9	99.7
	3	88.6	100	91.1	100	93.3	100	93.0	100	97.2	100	99.0	100
	4	89.1	100	91.7	100	93.8	100	94.7	100	97.2	100	98.8	100
	7	91.3	100	94.1	100	96.4	100	97.7	100	99.2	100	99.0	100
	11	91.6	100	94.4	100	97.1	100	97.8	100	99.9	100	99.1	100
	15	92.1	100	95.3	100	97.4	100	97.9	100	99.9	100	99.1	100

where $x_i \sim U(-6, 6)$, $i = 1, 2$. Again, we fit the following simple logistic regression model to the resulting data: $\log(\pi(x)/(1 - \pi(x))) = \beta_0 + \beta_1 x$.

Table 16 and Table 17 report the empirical power of the Bayes sum and order selection tests, respectively. The resulting power in this setting is, as one might expect, very similar to the power we observed for a missing quadratic term in Section 5.3.2 at the more severe departures from linearity. However, in this case, we can observe the effect of sparsity. For $n = 100$, it is clear that the power to detect the misspecified functional form of the linear predictor is enhanced noticeably for constellations providing greater degree of replication at each observed covariate value. The series-based tests lead to rejection in (almost) every dataset when $n = 500$ regardless of truncation point or degree of sparsity.

5.4.1 Evaluating Power Under Misspecification of the Link Function

To evaluate the power to detect misspecification of the link function, Kuss (2002) used the following model to generate the response:

Table 17. Performance of the order selection-based tests in detecting wrong functional form of the covariate. Simulated per cent rejection at the $\alpha = 0.05$ level using sample sizes of 100 and 500 with 1000 replications are reported. For each covariate distribution per cent rejection was evaluated at various values of truncation point. See text for explanation of the various constellations m_i.

Statistic	K	Constellation of m_i											
		1		1-2		2		1-10		5		10	
		100	500	100	500	100	500	100	500	100	500	100	500
S_a	2	91.6	100	92.8	100	93.9	100	92.2	100	96.0	100	98.7	99.9
	3	91.6	100	94.1	100	96.2	100	95.3	100	97.9	100	99.1	100
	4	91.7	100	93.8	100	94.8	100	96.5	100	97.7	100	99.2	100
	7	93.2	100	94.0	100	96.7	100	98.3	100	99.8	100	99.1	100
	11	91.6	100	94.6	100	98.1	100	99.2	100	00.0	100	99.1	100
	15	92.2	100	95.5	100	98.3	100	99.8	100	00.0	100	99.1	100
S_b	2	92.6	100	93.3	100	94.4	100	92.1	100	96.4	100	99.0	99.7
	3	94.8	100	96.2	100	97.1	100	95.9	100	97.9	100	99.3	100
	4	95.0	100	96.3	100	97.3	100	96.6	100	98.0	100	99.3	100
	7	95.8	100	97.0	100	97.7	100	97.8	100	99.3	100	99.3	100
	11	96.0	100	97.1	100	97.8	100	98.2	100	99.7	100	99.3	100
	15	96.0	100	97.1	100	97.8	100	98.2	100	99.9	100	99.3	100
T_a	2	87.1	100	89.3	100	91.4	100	89.9	100	95.3	100	98.2	99.8
	3	90.6	100	93.5	100	95.2	100	94.8	100	97.4	100	99.1	100
	4	91.6	100	94.4	100	95.3	100	95.8	100	97.8	100	99.2	100
	7	94.5	100	96.5	100	97.5	100	98.7	100	99.6	100	99.3	100
	11	95.8	100	97.1	100	98.5	100	99.4	100	100	100	99.3	100
	15	96.3	100	98.0	100	98.8	100	99.9	100	100	100	99.3	100
T_{OS}	2	81.3	100	84.2	100	87.6	100	86.8	99.8	94.1	100	97.5	99.7
	3	85.9	100	89.0	100	92.2	100	92.7	99.9	96.4	100	99.0	100
	4	86.8	100	90.3	100	92.7	100	93.8	100	97.0	100	99.0	100
	7	89.1	100	92.4	100	94.9	100	96.4	100	99.0	100	99.2	100
	11	90.1	100	92.2	100	95.5	100	96.9	100	99.5	100	99.3	100
	15	89.8	100	92.2	100	95.1	100	96.6	100	99.8	100	99.3	100
$T_{\max}$	2	88.4	100	90.0	100	92.2	100	90.6	100	95.4	100	98.5	99.8
	3	91.4	100	93.6	100	95.6	100	95.1	100	97.6	100	99.1	100
	4	92.2	100	94.8	100	95.6	100	96.2	100	97.9	100	99.2	100
	7	94.9	100	96.6	100	97.6	100	98.8	100	99.7	100	99.3	100
	11	96.0	100	97.3	100	98.7	100	99.3	100	100	100	99.3	100
	15	96.3	100	98.0	100	99.0	100	99.8	100	100	100	99.3	100

$$\log[-\log(1-\pi(x))] = 0.405x \tag{5.9}$$

where $x_i \sim U(-6, 6)$. Again, we fit the following simple logistic regression model to the resulting data: $\log\left(\pi(x)/(1-\pi(x))\right) = \beta_0 + \beta_1 x$.

Aerts et al. (2000) describe how the series-based alternatives in the order selection based tests can modified to provide a way of testing the adequacy of the link function. In the case of logistic regression the alternatives are of the form

$$\log\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + \sum_{k\in\Lambda} \phi_k u_k\{\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\} \tag{5.10}$$

These alternatives constitute a dimension-reducing nonparametric estimator often referred to as the "single-index model". Hence the test using this class of alternatives has been referred to as the "single-index test."

The resulting empirical power of the Bayes sum and order selection tests to detect link misspecification is reported in Table 18 and Table 19, respectively. The power of the series-based tests to detect misspecification of the link function appears to depend heavily on the sample size, the degree of sparsity, and the truncation point used in constructing the statistic. As in most of the settings studied previously in this chapter, there is a great deal of similarity among the various tests, however, it is worth noting that S_b and T_{OS} appear to be somewhat more conservative than the other tests across a majority of the combinations of sample size, constellation, and truncation point. This is somewhat remarkable given that S_b has consistently been the least conservative of all of the order selection tests in the other simulation settings studied. It is also worth mentioning that, for larger samples and larger values of the truncation point, the Bayes sum test appears to possess power properties that fall rather close to the midpoint between the most conservative tests (S_b, T_{OS}) and the least

Table 18. Performance of the Bayes sum test in detecting misspecified link. Simulated per cent rejection at the $\alpha = 0.05$ level using sample sizes of 100 and 500 with 1000 replications are reported. For each covariate distribution per cent rejection was evaluated at various values of truncation point. See text for explanation of the various constellations m_i.

Statistic	K	Constellation of m_i											
		1		1-2		2		1-10		5		10	
		100	500	100	500	100	500	100	500	100	500	100	500
S_K	2	9.0	7.5	9.7	9.2	13.4	10.7	18.7	12.9	25.6	16.3	38.9	33.8
	3	19.6	13.4	20.5	17.4	27.4	25.3	28.6	30.2	34.9	51.9	46.5	74.1
	4	19.8	12.6	21.7	17.4	28.2	25.4	29.6	30.4	34.5	52.5	45.4	74.7
	7	20.1	18.9	22.0	25.9	28.3	34.7	30.4	39.6	35.0	61.6	45.2	81.1
	11	19.2	23.7	20.7	30.7	27.5	39.6	28.5	46.8	33.2	65.9	44.7	84.5
	15	19.7	26.1	22.5	32.0	27.6	42.6	28.6	49.2	33.0	66.8	46.6	85.3

conservative (S_a, T_a, $T_{\max}$).

5.4.2 Overdispersion

To evaluate the power to detect misspecification of the link function, Kuss (2002) used the following model to generate the response:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = b + 0.405x \tag{5.11}$$

where $x_i \sim U(-6, 6)$, $E(b) = 0$ and $\mathrm{var}(b) = 0.323$. Again, we fit the following simple logistic regression model to the resulting data: $\log\left(\pi(x)/(1-\pi(x))\right) = \beta_0 + \beta_1 x$.

Table 20 and Table 21 report the empirical power of the Bayes sum and order selection tests, respectively. The series-based tests possess no power to detect possible presence of overdispersion. In fact, it appears that several of the tests converge to the nominal five per cent level specified for the test.

Table 19. Performance of the order selection-based tests in detecting misspecified link. Simulated per cent rejection at the $\alpha = 0.05$ level using sample sizes of 100 and 500 with 1000 replications are reported. For each covariate distribution per cent rejection was evaluated at various values of truncation point. See text for explanation of the various constellations m_i.

Statistic	K	Constellation of m_i											
		1		1-2		2		1-10		5		10	
		100	500	100	500	100	500	100	500	100	500	100	500
S_a	2	9.5	7.0	10.0	8.3	13.0	8.9	17.4	12.6	28.5	16.2	42.8	35.6
	3	27.2	16.6	31.0	23.1	38.0	29.3	39.2	36.1	46.4	58.3	51.7	78.1
	4	24.2	17.7	28.6	23.0	34.4	30.1	35.6	37.1	39.4	59.6	46.0	78.7
	7	28.3	29.8	29.8	34.3	33.0	48.7	32.5	55.1	37.1	73.7	38.7	88.2
	11	27.5	40.6	29.1	46.6	33.2	58.4	32.1	66.4	35.0	80.6	38.6	92.0
	15	25.8	45.8	29.4	53.6	31.7	64.0	30.6	71.2	32.1	84.7	49.2	93.9
S_b	2	11.4	08.1	12.8	10.1	17.6	12.1	22.3	14.4	31.7	18.0	47.9	36.5
	3	20.6	11.2	22.9	14.7	31.3	21.8	32.0	24.9	39.3	46.1	54.4	69.8
	4	20.7	11.0	23.9	14.5	31.4	22.4	32.4	25.3	40.0	46.3	55.4	70.0
	7	21.2	11.9	24.3	16.6	31.3	23.2	34.1	28.0	40.2	49.6	55.6	72.6
	11	21.3	12.0	23.7	17.0	32.0	23.2	33.4	28.4	39.1	49.1	56.2	72.8
	15	21.3	12.1	23.7	16.6	31.0	22.6	33.1	28.0	39.9	49.7	57.1	72.8
T_a	2	9.5	6.8	10.1	9.5	13.7	9.3	18.6	12.1	27.9	16.5	40.6	33.5
	3	21.0	13.5	23.6	17.9	29.5	25.6	31.9	31.6	37.1	53.8	48.2	75.6
	4	21.2	13.6	23.3	19.5	28.7	26.3	31.2	32.5	38.0	55.2	48.5	76.7
	7	24.6	24.1	27.0	29.8	32.0	40.4	34.9	48.3	40.0	68.2	48.9	85.9
	11	26.5	34.7	28.1	40.8	31.4	51.9	36.5	60.8	40.1	77.2	52.0	90.9
	15	26.6	40.4	29.1	49.0	32.2	57.9	34.5	67.1	38.1	82.8	58.2	93.6
T_{OS}	2	9.5	7.7	9.7	9.1	13.7	10.3	18.3	12.0	25.9	16.6	38.9	33.6
	3	14.8	11.3	16.0	14.4	21.2	21.3	25.1	24.9	31.2	47.5	42.5	70.3
	4	15.0	11.0	16.9	14.7	21.3	22.2	25.2	26.4	32.1	47.9	43.3	71.0
	7	15.6	14.6	17.3	19.6	21.3	26.7	26.6	32.7	33.0	55.1	43.5	76.3
	11	16.1	16.1	17.1	21.6	21.3	29.7	26.2	36.4	32.0	55.5	45.4	78.0
	15	16.0	16.6	16.5	21.7	20.9	29.3	25.2	33.7	31.5	56.7	46.6	78.0
$T_{\max}$	2	9.3	6.8	10.4	9.2	13.3	9.4	18.7	12.6	28.2	16.3	41.6	34.2
	3	22.0	13.6	24.8	19.6	31.4	26.7	33.2	33.0	39.5	54.7	49.9	76.0
	4	22.1	14.1	25.3	21.3	30.7	27.3	32.8	34.4	40.0	56.3	49.8	77.3
	7	25.9	24.6	28.1	30.6	33.8	43.0	36.0	49.4	41.7	68.7	51.4	86.3
	11	27.7	33.8	29.2	40.4	34.3	52.3	37.6	60.9	42.2	77.3	53.4	90.9
	15	27.6	39.0	29.9	47.6	35.0	57.0	36.3	65.3	40.8	81.8	58.6	93.2

Table 20. Performance of the Bayes sum test in detecting overdispersed data. Simulated per cent rejection at the $\alpha = 0.05$ level using sample sizes of 100 and 500 with 1000 replications are reported. For each covariate distribution per cent rejection was evaluated at various values of truncation point. See text for explanation of the various constellations m_i.

Statistic	K	Constellation of m_i											
		1		1-2		2		1-10		5		10	
		100	500	100	500	100	500	100	500	100	500	100	500
S_K	2	3.6	4.4	4.5	4.2	4.4	4.4	4.5	5.7	4.1	6.3	6.0	5.1
	3	3.5	4.8	5.0	4.6	4.8	4.8	4.8	6.0	4.1	6.2	5.7	5.4
	4	3.2	4.9	4.7	4.9	5.4	4.8	4.6	5.4	3.7	6.3	5.7	5.2
	7	3.6	4.6	4.4	4.8	5.0	5.0	4.7	5.4	3.3	6.2	4.9	6.1
	11	3.8	4.8	4.5	5.1	5.1	4.8	4.7	4.9	3.2	6.1	5.0	5.3
	15	3.7	5.1	4.6	4.9	5.6	4.9	5.2	5.5	3.3	6.5	5.6	5.4

5.5 Examples

In this section we present examples to illustrate the application of the Bayes sum statistic developed in the previous chapter. The examples we consider focus on analyses of well-known datasets that have appeared in the literature. Furthermore, these are datasets for which adequacy of a proposed logistic regression model has been examined. The motivation in revisiting these examples is to see if our method agrees with findings that have been established and accepted in the literature as well as provide an opportunity to clarify use of the method in practice.

In each of the examples presented below, we have estimated the p-value using the proportion of times simulated replicates from the asymptotic null distribution exceeded the test statistic value calculated for the dataset. We will use 50000 replicates from the asymptotic null distribution to obtain each p-value estimate. We will also present the results of test statistics based on likelihood ratios (LRs) and their score approximation. The rationale

Table 21. Performance of the order selection-based tests in detecting overdispersed data. Simulated per cent rejection at the $\alpha = 0.05$ level using sample sizes of 100 and 500 with 1000 replications are reported. For each covariate distribution per cent rejection was evaluated at various values of truncation point. See text for explanation of the various constellations m_i.

Statistic	K	Constellation of m_i											
		1		1-2		2		1-10		5		10	
		100	500	100	500	100	500	100	500	100	500	100	500
S_a	2	3.1	4.1	4.8	4.1	4.9	4.8	5.2	6.0	4.4	6.2	5.9	5.9
	3	4.0	5.2	5.4	4.7	5.0	5.1	5.0	4.6	4.5	7.0	5.8	5.1
	4	3.3	4.8	5.4	4.9	5.7	4.5	4.4	5.0	4.8	6.4	6.3	4.2
	7	4.0	4.7	5.5	5.3	5.9	5.4	5.5	4.9	4.8	5.4	4.7	4.4
	11	4.0	5.5	5.4	4.4	6.0	5.8	5.8	4.9	4.9	4.6	5.8	4.6
	15	4.1	5.7	5.5	4.7	5.1	5.4	5.0	5.0	4.9	4.9	22.1	5.1
S_b	2	5.5	5.5	7.4	5.5	7.2	5.5	7.7	7.4	7.5	7.2	9.1	6.4
	3	5.9	5.7	7.9	5.6	7.8	5.7	8.4	7.3	7.9	7.2	10.2	6.0
	4	5.6	5.7	7.9	5.6	7.8	5.7	7.9	7.5	7.7	7.3	9.4	6.3
	7	5.9	6.3	8.3	5.2	8.0	5.6	8.4	7.0	8.1	7.2	9.4	6.0
	11	5.9	6.2	7.8	5.2	7.8	6.3	8.1	7.0	7.4	7.2	9.4	6.4
	15	5.9	6.4	7.9	5.6	8.0	5.8	8.3	7.5	7.8	7.3	9.4	6.5
T_a	2	3.2	4.9	4.4	4.0	5.0	4.6	4.9	6.1	3.9	5.6	6.6	5.3
	3	3.4	5.2	4.6	4.3	5.3	5.2	4.9	5.3	3.8	6.0	6.3	4.8
	4	3.4	4.7	4.7	4.5	5.0	5.4	4.7	5.7	3.8	6.6	6.2	4.8
	7	3.3	4.3	4.9	4.8	5.1	5.2	5.7	6.1	4.6	6.2	5.4	5.1
	11	3.4	4.6	5.0	4.4	4.6	5.2	5.8	5.7	3.9	6.4	5.7	5.1
	15	3.5	4.6	5.1	4.3	4.7	4.4	5.7	5.8	3.9	5.8	13.0	5.3
T_{OS}	2	3.4	4.1	4.9	4.6	4.2	4.3	4.7	5.6	4.3	6.0	6.5	5.4
	3	3.5	4.2	4.8	4.7	4.4	4.6	4.7	5.5	4.2	6.2	6.5	4.8
	4	3.6	4.2	5.0	4.7	4.5	4.8	4.7	5.8	4.1	6.2	6.1	4.9
	7	3.6	4.7	5.2	4.1	4.6	4.5	4.7	5.3	3.9	5.5	6.1	4.9
	11	3.6	4.7	4.9	4.1	4.5	4.9	4.5	5.3	3.9	6.2	5.9	5.4
	15	3.6	4.5	4.9	4.4	4.6	4.7	4.7	5.6	4.1	6.0	5.9	5.6
$T_{\max}$	2	3.1	4.8	4.4	4.0	5.1	4.8	5.2	6.1	3.3	5.8	6.4	5.2
	3	3.2	5.1	4.9	4.4	5.3	4.8	4.9	5.6	3.5	6.2	6.2	4.9
	4	3.0	4.9	5.0	4.5	5.1	5.0	4.7	5.4	3.8	6.4	5.2	4.7
	7	3.2	4.1	5.1	4.6	5.0	4.9	5.6	6.0	4.0	6.2	5.1	4.9
	11	3.2	4.1	5.3	4.2	5.2	5.3	5.3	5.6	4.1	6.4	5.3	5.0
	15	3.3	4.3	5.3	4.2	5.2	4.8	5.2	5.9	4.1	6.6	12.2	5.2

for presenting both test statistics is that the score-based test statistic is an asymptotic approximation of the LR-based statistic. Consequently, these statistics can lead to different conclusions if the sample size is not sufficiently large.

5.5.1 Kyphosis Data

We will apply our method in the context of the well-known and thoroughly-studied kyphosis data set presented in Hastie and Tibshirani (1990, pp. 301-303). Data were collected on 83 patients undergoing corrective spinal surgery. The objective was to determine important risk factors for kyphosis following surgery. The risk factors are age in years, the starting vertebrae level of the surgery and the number of levels involved. Two of the cases in this data set have been identified in the literature as being outliers. These cases have been removed leaving us with 81 total cases.

Hastie and Tibshirani (1990) used this dataset to exemplify how one could use a nonparametric extension of the GLM known as the generalized additive model (GAM) to guide the specification of a GLM. Upon obtaining a "final" GAM fit selected via a stepwise procedure, Hastie and Tibshirani considered several parametric approximations to the estimated GAM. After comparing these approximations, they concluded that the following model offered the best approximation since it is parsimonious yet it captures the functional form of the nonparametric fit:

$$\log\left(\frac{\pi(x_1,x_2)}{1-\pi(x_1,x_2)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 (x_2 - 12) \times I(x_2 > 12), \qquad (5.12)$$

where x_1 denotes the patients age, x_2 denotes the starting vertebrae level of the surgery and $\pi(x_1, x_2)$ denotes the probability of kyphosis at given values of x_1 and x_2.

To better clarify this parametric specification, observe Figures 2 and 3. These figures display nonparametric estimates of the marginal relationship between each of the risk

Table 22. Test statistic values and p-values for the Bayes sum statistic for the kyphosis data null model taken to be the logistic regression model given in (5.12).

K	Value of S_K	p-value
2	0.369	0.823
3	0.314	0.756
4	0.409	0.922

factors and the log odds of kyphosis of these respective relationships. Nonparametric estimates such as GAMs are advantageous for parametric model building since they can be viewed as an objective assessment of the unknown functional relationship; that is, we "let the data speak for themselves". From Figure 2, we note a clear quadratic relationship age and the estimated logit proportions of the presence of kyphosis.

In this setting we apply a multivariate extension of the Bayes sum statistic using polynomial basis functions in a manner similar to that described in Section 5.3.1, however, we will use the likelihood ratio-based statistic. The results of applying the Bayes sum test to the parametric model specified in (5.12) are presented in Table 22. The large p-values provide an indication that the proposed parametric model should not be rejected.

In the interest of examining the performance of our test when applied to real data, we will compare the findings of the test of the logistic regression model specified by (5.12) with a test of a model which is intended to constitute misspecification. That is, we will now consider a test of the following model

$$\log\left(\frac{\pi(x_1, x_2)}{1-\pi(x_1, x_2)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \tag{5.13}$$

where as in the previous model(s) π denotes the probability of kyphosis at a given value x_1, x_2 with x_1, x_2 denoting the patients age and x_2 the starting vertebrae level of the surgery, respectively.

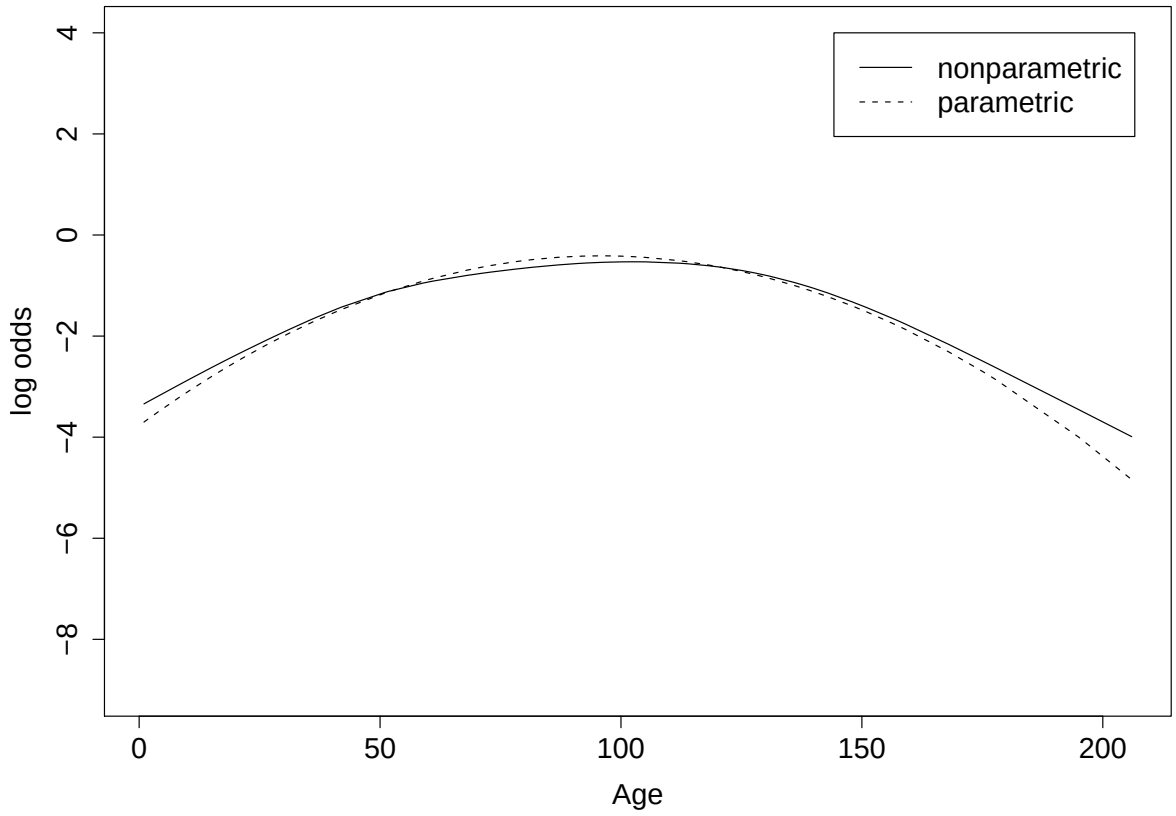

Fig. 2. *Estimated relationship between age and the log odds of kyphosis for* $n = 81$ *patients. The nonparametric curve estimate was obtained using a smoothing spline fit. The parametric model was obtained by modeling the log odds with the parametric linear predictor:* $\beta_0 + \beta_1 x + \beta_2 x^2$ *where* x *denotes the age variable.*

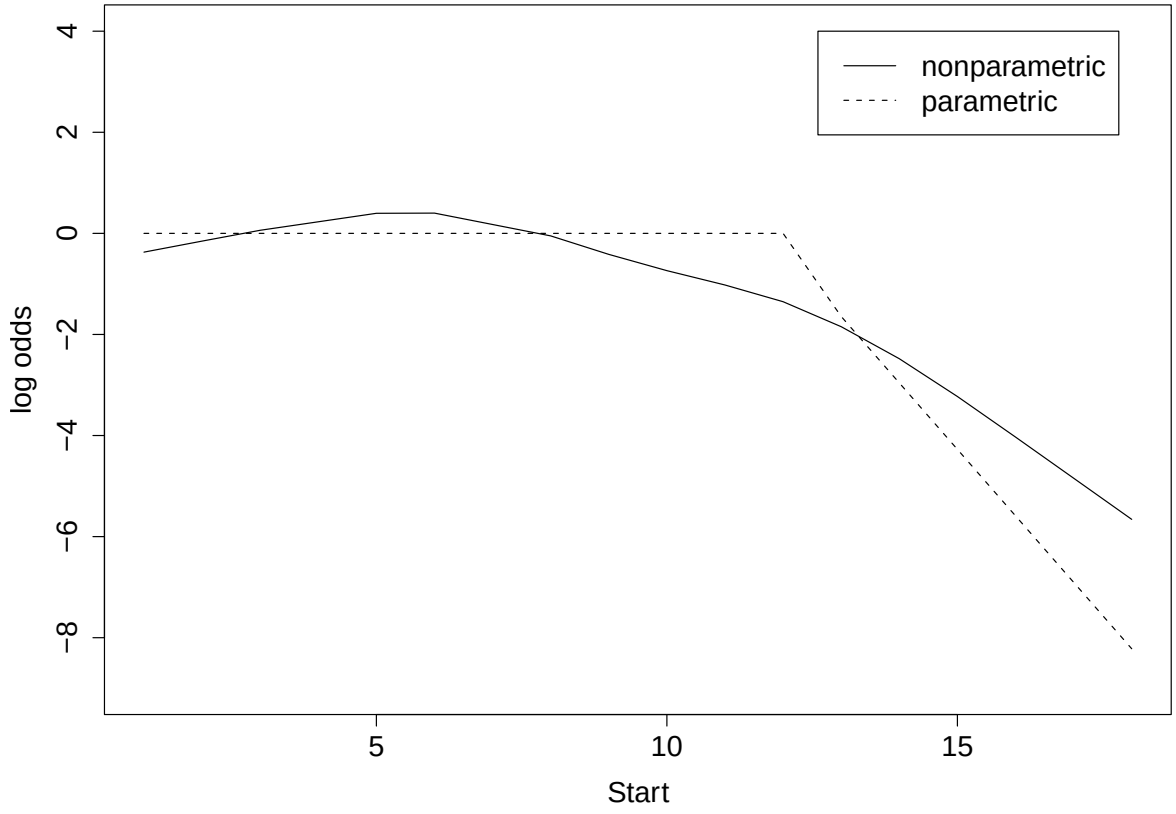

Fig. 3. *Estimated relationship between starting vertebrae level of the surgery (i.e., "start") and the log odds of kyphosis for* $n = 81$ *patients. The nonparametric curve estimate was obtained using a smoothing spline fit. The parametric model was obtained by modeling the log odds with the parametric linear predictor:* $\beta_0 + \beta_3(x - 12) \times I(x > 12)$ *where* x *denotes the start variable.*

Table 23. Test statistic values and p-values for the Bayes sum statistic for the kyphosis data null model taken to be the logistic regression model given in (5.13).

K	Value of S_K	p-value
2	2.350	0.032
3	2.538	0.054
4	2.721	0.067

Based on the results of our simulation studies of a missing quadratic term and misspecified functional form of the linear predictor presented in Sections 5.3 and 5.4, respectively, we anticipate that the Bayes sum test will detect the inadequacy of the simple logistic regression model and reject the null hypothesis that the proposed linear predictor is correct. Table 23 summarizes the test statistic values and p-values for a test of (5.13) using the likelihood ratio-based version of the Bayes sum statistic. The test clearly indicates a lack of fit for model (5.13). Examination of the results of applying the Bayes sum test to the models specified in equations (5.12) and (5.13) along with the findings of Hastie and Tibshirani (1990) leads us to conclude that our test is capable of distinguishing adequately specified models from misspecified models.

5.5.2 Coronary Artery Disease Diagnostic Data

Here we analyze a dataset which has been presented as an example in Harrell (2001). This dataset is from the Duke University Cardiovascular Disease Databank and consists of 3504 patients and 6 variables. One of the analyses conducted on this dataset involved predicting the probability of significant ($>= 75\%$ diameter narrowing in at least one important coronary artery) coronary disease. In particular, Harrell (2001) examined the adequacy of the following logit model

$$\log\left(\frac{\pi(x_1, x_2)}{1-\pi(x_1, x_2)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{5.14}$$

where π denotes the probability of significant coronary artery disease, x_1 denotes a respondent's sex ($x_1 = 0$ for males, 1 for females) and x_2 denotes a respondent's age.

As in the example discussed in Section 5.5.1, we examine a nonparametric estimate of the relationship between age and log odds of significant coronary disease for this data in order to get a visual impression of the degree of departure from the model proposed in (5.14). We estimate this relationship separately for men and women in Figure 4, noting that if the specification of (5.14) were correct, then the plot should consist of parallel straight lines. From Figure 4, we can see that while the nonparametric curve estimate for males may be adequately approximated by a straight line, there is noticeable nonlinear relationship between age and the log odds of significant coronary artery disease for women. Furthermore, there appears to be a possible interaction between the age and sex variables not accounted for in the specification in (5.14).

The results of the Bayes sum test are presented in Table 24. While not significant at the $\alpha = 0.05$ level, the p-values are rather small providing a suggestion of lack of fit. It seems noteworthy that the departure from linearity depicted in Figure 4 is not as severe as that which was observed Figures 2 and 3 for the kyphosis data. Moreover, the p-values observed in the kyphosis example are distinctly smaller than the p-values obtained in this example. In the present example, it is clear that the p-values are small enough to indicate possible lack of fit, but not excessively small to qualify as formal rejection of (5.14) at the $\alpha = 0.05$ level. Thus, it appears that the test possesses the ability reflect the severity of the departure.

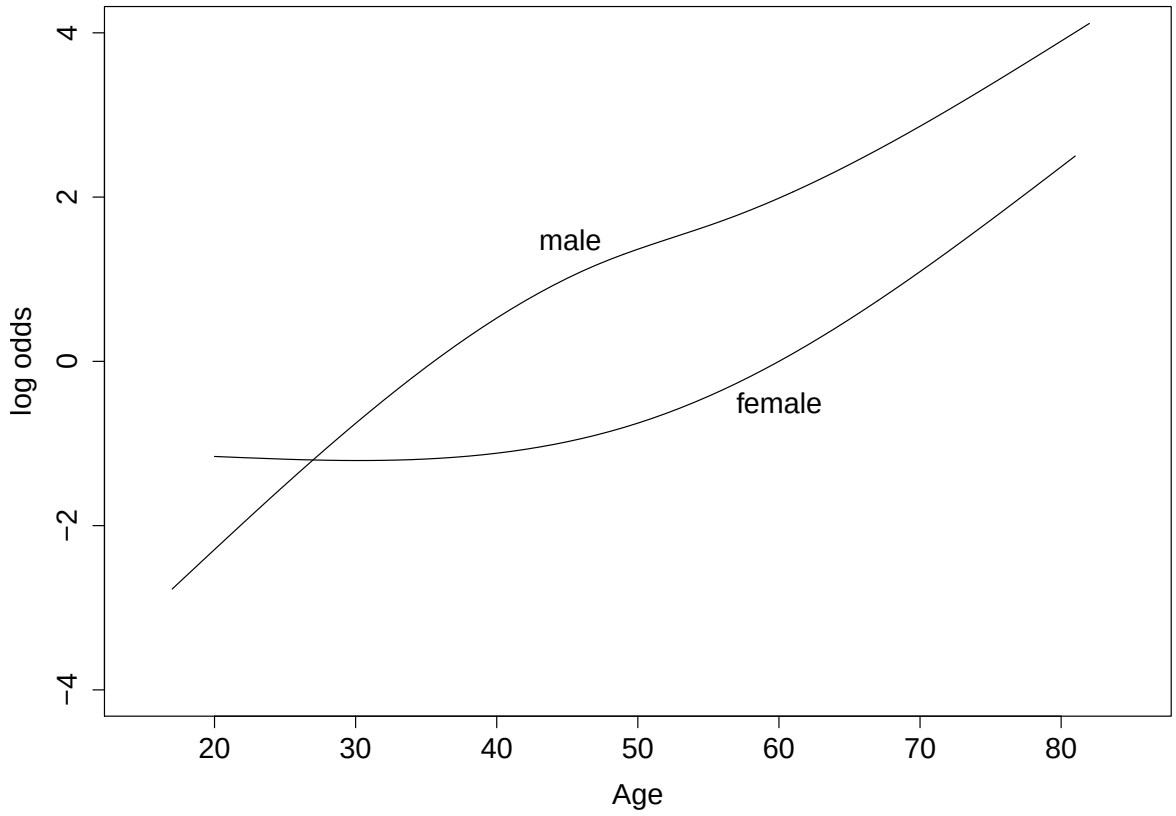

Fig. 4. *Estimated relationship between age and the log odds of significant coronary artery disease for* 2405 *male patients and* 1099 *female patients. The estimated curves were obtained using a regression spline fits. Spline fits are applied to the subsets of males and females, separately.*

Table 24. Test statistic values and p-values for the Bayes sum statistic for the Cardiac Catheterization data with null model taken to be the logistic regression model given in (5.14).

K	Value of S_K	p-value
2	5.046	0.118
3	6.734	0.089
4	7.024	0.087
7	7.130	0.090
11	7.194	0.094
15	7.234	0.092

5.6 Discussion

In this chapter we presented an extensive numerical study that revealed new insights into the performance and applicability of the series-based tests within the context of logistic regression. We observed by means of simulation that these statistics possess desirable power properties against several alternatives that have been identified in the existing literature as being important. These power properties are competitive with and in some cases superior to the properties of some of the better known tests of fit that have been studied in the literature. In addition to establishing relative performance against existing tests, our simulation results also provide a scope of the departures that the series-based statistics can detect. Finally, this simulation study has permitted us an opportunity to investigate the properties of the Bayes sum statistic developed in Chapter IV.

It is evident from the empirical power values presented for omission of a quadratic term in Section 5.3 and for misspecified functional form of a covariate in Section 5.4, the series-based tests perform best in detecting departures involving variables that have been included in the model. Indeed, we observed that none of the series-based lack-of-fit tests provide meaningful power to detect misspecification due to variables not included in the

model. This was revealed in the results of the tests in two settings: the missing covariate simulation setting of Section 5.4 and the setting for detecting omission of a dichotomous variable and its interaction of Section 5.3. Two points should be kept in mind with regards these findings. First, the two settings in which the series-based tests performed inadequately led to poor power for test statistics studied in Hosmer et al. (1997) and Kuss (2002). Second, when detecting departures in terms of variables that were actually included in the model, all of the tests performed extremely well, and several of these tests exhibited superior power properties when compared to the best performing tests studied in Hosmer et al. and Kuss.

In addition to the simulations that examined misspecification of the linear predictor, we also considered other departures from the specified model. Many of these departures were not well detected, however, most existing tests do not perform much better in these settings. Furthermore, we did find that the series-based statistics performed acceptably in detecting departures from the logistic link function under certain circumstances. Given the formulation of the series-based tests, these results are not surprising.

We found that the series-based tests can provide desirable power in detecting some types of misspecification for data with replicate covariate patterns as well as for sparse data. Furthermore, replication appears to enhance the performance of the series-based statistics. This is noteworthy that the test can be used regardless of the degree of sparsity and be expected to detect misspecification. As we discussed in Chapter II, this property is not shared by most of the existing tests of fit for logistic regression models. Consequently, we assert that the series-based tests are particularly beneficial in data sets exhibiting near sparsity (i.e., there are few replicated covariate patterns). In such situations one cannot be certain of the validity of tests designed for sparse data or tests requiring replication.

In addition to the above conclusions, we found that throughout our simulation study that the behavior of the Bayes sum statistic and the behavior of the order selection based

tests are generally similar. One noteworthy deviation from the agreement observed among these various tests was due to a test based on a score analog of the BIC criteria. This test tended reject the null hypothesis more often than the other statistics. This agreement among these tests is particularly interesting because the Bayes sum statistic is nonadaptive, and nonadaptive tests like the cusum test have been observed to have poor power properties in comparison to tests that use test statistics based on data-driven smoothing parameters (i.e., order selection-based tests). This finding is consistent with simulation results described in Hart (2009). We noticed that the value of K did not have a great deal of influence on the power of the test.

CHAPTER VI

CONCLUSION

In this dissertation we sought to contribute to the development of techniques for assessing the adequacy of generalized linear models. In particular, we have focused on lack-of-fit tests based on characterizing departures from the predictor function in terms of Fourier coefficients and subsequently testing that all of these coefficients are 0. In the pursuit of our objective, we developed a new lack-of-fit test for canonical link regression models and examined several other well-known lack-of-fit tests. Our approach for testing lack-of-fit is based on the ideas of Hart (2009). That is, we use as a test statistic a Laplace approximation to the posterior probability of the null hypothesis. Rather than evaluating this probability directly, this statistic is used in frequentist fashion by means of a reference distribution.

Our examination of the Laplace approximation-based test statistic yielded several noteworthy theoretical findings. First, we show that under the null hypothesis, the limit distribution of the statistic formulated from posterior probability is completely determined by the alternative models with the fewest parameters. In our formulation of the posterior probability, these models are the so-called singleton alternatives. This is remarkable because the posterior probability is constructed from a very general, nonparametric class of alternative models. This leads us to a test statistic that is a weighted sum of exponentiated likelihood ratios, where the weights depend on user-specified prior probabilities. Replacing the likelihood ratios with their corresponding score statistics produces a statistic that consists of a weighted sum of exponentiated squared Fourier coefficient estimates. Hence we refer to these statistics as the “Bayes sum” statistics. The prior probabilities which provide the investigator the flexibility to examine specific departures from the prescribed model. Alternatively, the use of noninformative priors produces a new omnibus lack-of-fit statistic. We then established the limiting distribution of the score version of the test statis-

tic under both the null hypothesis and local alternatives that converge to the null at rate $1/\sqrt{n}$. An interesting aspect of this result is that we obtained this result by characterizing the distribution of the coefficients under local alternatives. To our knowledge, no such result has appeared in the literature for techniques addressing generalized linear models. Under the null hypothesis, the score-based statistic provides a large sample approximation of the likelihood ratio-based test statistic. Our result also provides a null distribution for the likelihood ratio test.

Our extensive simulation study of the series-based tests within the context of logistic regression reveals that these statistics possess desirable power properties against several alternatives that have been identified in the existing literature as being important. Moreover, these power properties are competitive, and in some cases superior to, some of the better known tests of fit that have been studied in the literature. In particular, the Bayes sum and order selection-based tests perform well in detecting misspecification in the linear predictor. While we noted that other departures from the fitted model are not well detected, most existing tests do not perform much better in these settings. For the departures that the series-based tests could detect, we found that the series-based tests were less sensitive to the degree of sparsity than other existing methods. The simulation results also provide a scope of the departures that the Bayes sum statistic can detect.

Several questions have arisen that should be addressed in future research. First, is it possible to generalize the distribution theory further in order to permit the order of the sum to tend to infinity with the sample size. Robustness of prior probability selection is also of interest.

Ultimately, we conclude that the desirable properties reported in Hart (2009) apply in this generalized setting. Indeed, as we have noted above the Bayes sum statistic is easily calculated (relative to other series-based tests), it has a convenient reference distribution, and it has good power against some important departures from a proposed null model.

REFERENCES

Aerts, M., Claeskens, G., and Hart, J. D. (1999). Testing lack of fit of a parametric function. *Journal of the American Statistical Association* **94**, 869–879.

Aerts, M., Claeskens, G., and Hart, J. D. (2000). Testing lack of fit in multiple regression. *Biometrika* **87**, 405–424.

Aerts, M., Claeskens, G., and Hart, J. D. (2004). Testing lack of fit of a parametric function. *Annals of Statistics* **32**, 2580–2615.

Agresti, A. (2002). *Categorical Data Analysis*. New York: John Wiley.

Aparicio, T. and Villanua, I. (2001). The asymptotically efficient version of the information matrix test in binary choice models. a study of size and power. *Journal of Applied Statistics* **28**, 167–182.

Aranda-Ordaz, F. J. (1981). On two families of transformations to additivity for binary response data. *Biometrika* **68**, 357–363.

Azzalini, A., Bowman, A. W., and Härdle, W. (1989). Use of nonparametric regression for model checking. *Biometrika* **76**, 1–11.

Beran, R. J. and Millar, P. W. (1992). Tests of fit for logistic models. Technical Report, University of California, Berkeley.

Bertolini, G., D'Amico, R., Nardi, D., Tinazzi, A., and Apolone, G. (2000). One model, several results: The paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. *Journal of Epidemiology and Biostatistics* **5**, 251–253.

Brown, C. C. (1982). On a goodness of fit test for the logistic model based on score statistics. *Communications in Statistics, Theory & Methods* **11**, 1087–1105.

Cline, D. (1983). Infinite series of random variables with regularly varying tails. Technical Report, 83-24, University of British Columbia, Institute of Applied Mathematics and Statistics.

Collett, D. (1991). *Modelling Binary Data*. London: Chapman and Hall.

Copas, J. B. (1989). Unweighted sum of squares test for proportions. *Applied Statistics* **38**, 71–80.

Cressie, N. and Read, T. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B* **46**, 440–464.

Eubank, R. L. and Hart, J. D. (1992). Testing goodness-of-fit in regression via order selection criteria. *Annals of Statistics* **20**, 1412–1425.

Eubank, R. L. and Hart, J. D. (1993). Commonality of cusum, von Neumann and smoothing-based goodness-of-fit tests. *Biometrika* **80**, 89–98.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models, 2nd ed.* New York: Springer.

Farrington, C. P. (1996). On assessing goodness of fit of generalized linear models to sparse data. *Journal of the Royal Statistical Society, Series B* **58**, 349–360.

Guerrero, V. M. and Johnson, R. A. (1982). Use of the Box-Cox transformation with binary response models. *Biometrika* **69**, 309–314.

Harrell, F. E. (2001). *Regression Modeling Strategies*. New York: Springer.

Hart, J. D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*. New York: Springer.

Hart, J. D. (2009). Frequentist-Bayes lack-of-fit tests based on Laplace approximations. *Journal of Statistical Theory and Practice* **3**, 681–704.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science* **1**, 297–310.

Hastie, T. and Tibshirani, R. (1987). Generalized additive models: Some applications. *Journal of the American Statistical Association* **82**, 371–386.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. New York: Chapman and Hall.

Hosmer, D. W. and Hjort, N. L. (2002). Goodness-of-fit processes for logistic regression: simulation results. *Statistics in Medicine* **21**, 2723–2738.

Hosmer, D. W., Hosmer, T., Le Cessie, S., and Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine* **16**, 965–980.

Hosmer, D. W. and Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics, Theory & Methods* **A10**, 1043–1069.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression, 2nd ed.* New York: John Wiley.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics* **58**, 71–120.

Inglot, T. and Ledwina, T. (1996). Asymptotic optimality of data-driven Neyman's tests for uniformity. *Annals of Statistics* **24**, 1982–2019.

Inlow, M. (2001). On techniques for binary response modeling. Ph.D. Dissertation, Texas A&M University, Department of Statistics, College Station.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Kuss, O. (2002). Global goodness-of-fit tests in logistic regression with sparse data. *Statistics in Medicine* **21**, 3789–3801.

le Cessie, S. and van Houwelingen, J. C. (1991). A goodness-of-fit test for binary regression models, based on smoothing methods. *Biometrics* **47**, 1267–1282.

Lechner, M. (1991). Testing logit models in practice. *Empirical Economics* **16**, 177–98.

McCullagh, P. (1985). On the asymptotic distribution of Pearson's statistics in linear exponential family models. *International Statistical Review* **53**, 61–67.

McCullagh, P. (1986). The conditional distribution of goodness-of-fit statistics for discrete data. *Journal of the American Statistical Association* **81**, 104–107.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, 2nd ed.* New York: Chapman and Hall.

Moore, D. F. (1986). Asymptotic properties of moment estimators for overdispersed counts and proportions. *Biometrika* **73**, 583–588.

Orme, C. (1988). The calculation of the information matrix test for binary data models. *The Manchester School of Economic & Social Studies* **56**, 370–76.

Osius, G. and Rojek, D. (1992). Normal goodness-of-fit tests for multinomial models with large degrees of freedom. *Journal of the American Statistical Association* **87**, 1145–1152.

Parzen, M. (1977). Multiple time series: Determining the order of approximating autoregressive schemes. In *Multivariate Analysis–IV, Krishnaiah*, P. R. (ed.), pp. 283-295. Amsterdam: North-Holland.

Paul, S. R. and Deng, D. (2000). Goodness of fit of generalized linear models to sparse data. *Journal of the Royal Statistical Society, Series B* **62**, 323–333.

Paul, S. R. and Deng, D. (2002). Score test for goodness of fit of generalized linear models to sparse data. *Sankhya, Series B* **64**, 179–191.

Pigeon, J. G. and Heyse, J. F. (1999). A cautionary note about assessing the fit of logistic regression models. *Journal of Applied Statistics* **26**, 847–853.

Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Applied Statistics* **29**, 15–24.

Prentice, R. L. (1976). A generalization of the probit and logit methods for dose response curves. *Biometrics* **32**, 761–768.

Pulksenis, E. and Robinson, T. J. (2002). Two goodness-of-fit tests for regression models with continuous covariates. *Statistics in Medicine* **21**, 79–93.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika* **83**, 251–266.

Resnick, S. I. (1999). *A Probability Path*. Boston: Birkhäuser.

Royston, P. (1992). The use of cusums and other techniques in modelling continuous covariates in logistic regression. *Statistics in Medicine* **11**, 1115–1129.

Samorodnitsky, G. and Taqqu, M. S. (1994). *Stable non-Gaussian Processes: Stochastic Models with Infinite Variance*. London: Chapman and Hall.

Seber, G. A. F. (1977). *Linear Regression Analysis*. New York: John Wiley.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: John Wiley.

Shao, J. (2003). *Mathematical Statistics, 2nd ed.* New York: Springer.

Stukel, T. A. (1988). Generalized logistic regression. *Journal of the American Statistical Association* **83**, 426–431.

Stute, W., Gonzalez Mantiega, W., and Presedo Quindimil, M. (1998). Bootstrap approximations in model checks for regression. *Journal of the American Statistical Association* **93**, 141–149.

Stute, W. (1997). Nonparametric model checks for regression. *Annals of Statistics* **25**, 613–641.

Stute, W., Thies, S., and Zhu, L.-X. (1998). Model checks for regression: An innovation process approach. *Annals of Statistics* **26**, 1916–1934.

Stute, W. and Zhu, L.-X. (2002). Model checks for generalized linear models. *Scandinavian Journal of Statistics* **29**, 535–545.

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

Thomas, J. M. (1993). On testing the logistic assumption in binary dependent variable models. *Empirical Economics* **18**, 381–92.

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

Wang, X. and George, E. I. (2007). Adaptive Bayesian criteria in variable selection for generalized linear models. *Statistica Sinica* **17**, 667–690.

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

Xiang, D. and Wahba, G. (1995). Testing the generalized linear model null hypothesis versus ‘smooth’ alternatives. Technical Report, 953, University of Wisconsin.

VITA

Daniel Laurence Glab was born in Chicago, Illinois. He attended Riverside-Brookfield High School in Brookfield, Illinois. In May 2000 he received a Bachelor of Science degree in mathematics from University of Wisconsin-Madison. He continued his studies at Texas A&M, earning a Master of Science degree in statistics in December 2005 under the supervision of Dr. Randall L. Eubank and then a Doctor of Philosophy degree in statistics in May 2011 under the supervision of Dr. Thomas Wehrly. He can be reached through the following address: Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143.