

DEPARTMENT OF STATISTICS
North Carolina State University
2501 Founders Drive, Campus Box 8203
Raleigh, NC 27695-8203

Institute of Statistics Mimeo Series No. 2591

Testing goodness-of-fit in logistic case-control studies

Howard D. Bondell

Department of Statistics, North Carolina State University

Raleigh, NC 27695-8203, U.S.A.

(email: bondell@stat.ncsu.edu)

Testing goodness-of-fit in logistic case-control studies

Howard D. Bondell

Department of Statistics, North Carolina State University

Raleigh, NC 27695-8203, U.S.A.

(email: bondell@stat.ncsu.edu)

Abstract

This paper presents a goodness-of-fit test for the logistic regression model under case-control sampling. The test statistic is constructed via a discrepancy between two competing kernel density estimators of the underlying conditional distributions given case-control status. The proposed goodness-of-fit test is shown to compare very favorably with previously proposed tests for case-control sampling in terms of power. The test statistic can be easily computed as a quadratic form in the residuals from a prospective logistic regression maximum likelihood fit. In addition, the proposed test is affine invariant and has an alternative representation in terms of empirical characteristic functions.

Keywords: Biased sampling; Case-control data; Goodness-of-fit; Kernel density; Logistic regression; Retrospective sampling.

1 Introduction

Given a binary variable Y and a $p \times 1$ vector X of covariates, the logistic regression model states that

$$\text{pr}(Y = 1|X = x) = \frac{\exp(\alpha^* + x^T\beta)}{1 + \exp(\alpha^* + x^T\beta)}, \quad (1.1)$$

where α^* is a scalar parameter and β is a $p \times 1$ vector of parameters. Under this model, the marginal distribution of X is left completely unspecified.

One approach to examining the relation between Y and X is via an observational study, or prospective sampling, where the sampling is either a simple random sample from the joint distribution, F_{XY} , or from the conditional distribution, $Y|X$. In using the logistic regression model under this sampling scheme, it is implicitly assumed that this model is appropriate for the data at hand. To test this assumption, a number of goodness-of-fit tests have been proposed. This prospective sampling situation is the setting in which the typical goodness-of-fit procedures are constructed and distribution theory discussed. Hosmer et al. (1997) compare the properties of some of these tests via a simulation study.

Alternatively, one can employ a case-control sampling scheme to examine the relation between the variables. Case-control sampling, also known as retrospective sampling, refers to independently sampling from F_0 , the conditional distribution of $X|(Y = 0)$, and F_1 , defined similarly. Goodness-of-fit testing under this sampling plan has received much less attention despite the widespread use of the logistic regression assumption in case-control studies.

Under case-control sampling the logistic regression assumption is equivalent to a two-sample semiparametric biased sampling model. Specifically, letting f_0 and f_1 denote the corresponding conditional densities and using (1.1), an application of Bayes'

rule yields the equivalent model:

$$\begin{aligned} u_1, \dots, u_{n_0} & \text{ is a random sample with density } f_0(x), \\ z_1, \dots, z_{n_1} & \text{ is a random sample with density } f_1(x) = \exp(\alpha + x^T \beta) f_0(x), \end{aligned} \tag{1.2}$$

where $\alpha = \alpha^* + \log\{(1 - \pi)/\pi\}$ and $\pi = \text{pr}(Y = 1)$. Model (1.2) is a particular case of the biased sampling model discussed by Vardi (1985).

The equivalent model (1.2) is the starting point for the goodness-of-fit tests derived under the case-control point of view in Qin and Zhang (1997) and Zhang (1999, 2001). Gilbert (2004) generalizes the goodness-of-fit approach of Qin and Zhang (1997) to the multiple sample version of (1.2). Bondell (2005) uses this model formulation to derive robust estimators of the parameters in the logistic regression model.

The logic behind the majority of these approaches stem from the idea that based on the data, the baseline distribution F_0 can be estimated in two ways. One is to use the empirical distribution of the first sample (the controls), and the other is to use both samples by exploiting the model (1.2) and use the nonparametric maximum likelihood estimator (NPMLE). If the model were true, the latter estimate of the distribution F_0 should be used. However, if the model were not true, the empirical distribution of the controls is the only information regarding F_0 contained in the combined sample. Intuitively, if the model were at least approximately true, the two estimates of F_0 should be close; hence a measure of discrepancy between them could be used as a measure of fit.

Most typical measures of discrepancy are based on the cumulative distribution function (Qin and Zhang, 1997; Gilbert, 2004; Bondell, 2005). However, the goodness-of-fit procedures based on the distribution function lack the property of affine invariance, in that an affine transformation of the covariate vector will lead to a different value of the test statistic and even possibly a different conclusion regarding the fit of the model. This lack of affine invariance can lead to misleading results in some common practical

situations.

An affine invariant test statistic is proposed in this paper, based on the case-control sampling scheme. The statistic is proposed as an L^2 distance between two competing kernel density estimates, and has an alternative representation as a discrepancy based on the empirical characteristic function. Additionally, the proposed test statistic is easily computable and interpretable as a quadratic form in the residuals obtained from a prospective logistic regression fit. In contrast to chi-squared type goodness-of-fit tests, the proposed test also does not require the space to be partitioned into a finite number of categories.

The remainder of the paper is organized as follows. In the next section, the test statistic is constructed from the biased sampling perspective via a discrepancy between two kernel density estimators. The connection between the formulation in terms of a kernel density estimate and the corresponding formulation via the characteristic function is discussed in Section 3. The asymptotic distribution theory is given in Section 4. In Section 5, the statistic is shown to be a quadratic form in the residuals from a prospective fit, thus greatly simplifying computation. In addition, the test statistic is shown to be affine invariant, and a simple example of problems associated with a lack of affine invariance is discussed. The goodness-of-fit test is then applied to the multiple logistic regression model in the analysis of a real data set in Section 6. Finally, the power of the test is shown to compare very favorably to the existing procedures via a simulation study.

2 Construction of the test statistic

Letting $(x_1, \dots, x_n) = (u_1, \dots, u_{n_0}, z_1, \dots, z_{n_1})$ denote the combined sample, explicitly, the baseline distribution, F_0 , from model (1.2) can be estimated in the following two

ways:

$$\begin{aligned}\tilde{F}_0(t; \tilde{\alpha}, \tilde{\beta}) &= \sum_{i=1}^n \tilde{p}_i(\tilde{\alpha}, \tilde{\beta}) I(x_i \leq t), \\ \hat{F}_0(t) &= \sum_{i=1}^{n_0} \hat{p}_i I(u_i \leq t),\end{aligned}\tag{2.1}$$

where $\hat{p}_i = n_0^{-1}$, and $\tilde{p}_i(\tilde{\alpha}, \tilde{\beta}) = n_0^{-1} \{1 + \rho \exp(\tilde{\alpha} + x_i^T \tilde{\beta})\}^{-1}$, with $\rho \equiv n_1/n_0$. The former is the NPMLE under the model (1.2), while the latter is under no model assumption. The estimator $(\tilde{\alpha}, \tilde{\beta})$ is the maximum likelihood estimator whose distribution under case-control sampling is discussed by Prentice and Pyke (1979) and Qin and Zhang (1997). By the usual definition of the cumulative distribution function, for two p -dimensional vectors, $(s \leq t) \equiv (s_1 \leq t_1, \dots, s_p \leq t_p)$.

For a given kernel, $K(x) \geq 0$, with $\int K^2(x) dx < \infty$, and an empirical distribution, F_n , the convolution density,

$$\int K(x - y) dF_n(y),$$

defines a kernel density estimate. Under the model assumptions in (1.2), two competing empirical distributions for F_0 are available, \tilde{F}_0 or \hat{F}_0 , as in (2.1), so that for a kernel, K , a density estimate can be constructed in two ways

$$\tilde{f}(x) \equiv \int K(x - y) d\tilde{F}_0(y)\tag{2.2}$$

or

$$\hat{f}(x) \equiv \int K(x - y) d\hat{F}_0(y).\tag{2.3}$$

To measure the discrepancy between the two distributions, one can base the measure on the density instead of the cumulative distribution function. A common measure of discrepancy between densities is the integrated squared error (Hall, 1984; Anderson et al., 1994). For any two densities f and g , define the integrated squared error as

$$I = \int \{f(x) - g(x)\}^2 dx.\tag{2.4}$$

For a given kernel, K , the test statistic proposed in this paper for the case-control logistic regression model is the scaled integrated squared error

$$I_n \equiv n \int \left\{ \tilde{f}(x) - \hat{f}(x) \right\}^2 dx, \quad (2.5)$$

with $\tilde{f}(x)$ and $\hat{f}(x)$ defined by (2.2) and (2.3). Note that $\tilde{f}(x)$ contains the parameter estimates $(\tilde{\alpha}, \tilde{\beta})$, while both $\tilde{f}(x)$ and $\hat{f}(x)$ involve the data.

3 Connection to characteristic function

In this section, an alternative interpretation of the test statistic in terms of the empirical characteristic function is given. Based on the semi-parametric model (1.2), the empirical characteristic function can also be constructed in two ways using either empirical distribution. The characteristic functions of these two distributions are given by

$$\psi_{\tilde{F}_0}(t) = \int \exp(it^T x) d\tilde{F}_0(x) \quad (3.1)$$

and

$$\psi_{\hat{F}_0}(t) = \int \exp(it^T x) d\hat{F}_0(x). \quad (3.2)$$

As opposed to the distribution function or density, a measure of discrepancy based on the characteristic function can be used. For a discrepancy between two characteristic functions, following Heathcote (1977) and Henze et al. (2003) one can define an integrated weighted squared difference. Using the two characteristic functions in (3.1) and (3.2), an integrated weighted squared difference is defined by

$$T = n \int |\psi_{\tilde{F}_0}(t) - \psi_{\hat{F}_0}(t)|^2 w(t) dt. \quad (3.3)$$

Using the representation of the kernel density estimate as the convolution in (2.2) and (2.3), by Parseval's identity, one obtains that the test statistic (2.5) can be equivalently expressed as

$$I_n = n (2\pi)^{-p} \int \left| \hat{K}(t) \{ \psi_{\hat{F}_0}(t) - \psi_{\tilde{F}_0}(t) \} \right|^2 dt, \quad (3.4)$$

where $\hat{K}(t)$ denotes the fourier transform of the square-integrable kernel K .

Clearly, for a given kernel K , setting $w(t) \equiv (2\pi)^{-p} |\hat{K}(t)|^2$ in (3.3) creates an equivalence between the two choices of discrepancy measures. The representation in terms of the kernel density estimate is used in the remainder of this paper, although the construction of the test statistic can be done from either point of view.

4 Asymptotic distribution

The asymptotic distribution of I_n , or the equivalently defined statistic via the characteristic function, is now derived under the null hypothesis of model (1.2). The case of $p = 1$ is used for simplicity of notation, although the results can be naturally generalized to the case of $p > 1$.

THEOREM 1 *For a fixed kernel K , under model (1.2) with true parameter (α_0, β_0) and suitable regularity conditions, as $n \rightarrow \infty$, the process $n^{1/2}\{\tilde{f}(t) - \hat{f}(t)\}$ converges to $W(t)$, a Gaussian process with mean 0 and covariance function given by*

$$V(s, t) \equiv E\{W(s)W(t)\} = \rho(1 + \rho) \left\{ A_2(s, t) - (A_0(s), A_1(s)) A^{-1} \begin{pmatrix} A_0(t) \\ A_1(t) \end{pmatrix} \right\},$$

where

$$A_0(t) = \int K(t - y)q(\alpha_0 + \beta_0 y) dG(y), \quad A_1(t) = \int K(t - y)q(\alpha_0 + \beta_0 y) y dG(y),$$

$$A_2(s, t) = \int K(s - y)K(t - y)q(\alpha_0 + \beta_0 y) dG(y),$$

$$A = \begin{pmatrix} \int q(\alpha_0 + \beta_0 y) dG(y) & \int q(\alpha_0 + \beta_0 y) y dG(y) \\ \int q(\alpha_0 + \beta_0 y) y dG(y) & \int q(\alpha_0 + \beta_0 y) y^2 dG(y) \end{pmatrix}$$

and $q(t) \equiv \exp(t) / \{1 + \rho \exp(t)\}$.

In Theorem 1 of Qin and Zhang (1997), the asymptotic null distribution of the process

$$n^{1/2} \left\{ \tilde{F}(t) - \hat{F}(t) \right\} = n^{1/2} \int I(x \leq t) \{d\tilde{F}(x) - d\hat{F}(x)\}$$

is derived. The distribution of the process

$$n^{1/2} \left\{ \tilde{f}(t) - \hat{f}(t) \right\} = n^{1/2} \int K(t-x) \{d\tilde{F}(x) - d\hat{F}(x)\}$$

is derived in a parallel manner, hence the details of the proof are omitted here and the reader is referred to that paper.

REMARK: A minor correction to the covariance function for the theorem in Qin and Zhang (1997) is noted. The covariance should read

$$E\{W(s)W(t)\} = \rho(1 + \rho) \left\{ A_0(\min\{s, t\}) - (A_0(s), A_1(s)) A^{-1} \begin{pmatrix} A_0(t) \\ A_1(t) \end{pmatrix} \right\},$$

where $A_0(t)$ and $A_1(t)$ are as defined there. In the derivation, it was implicitly assumed that $s \leq t$, and thus $\min\{s, t\}$ was replaced by s in their theorem.

Now consider the random variable,

$$\xi_L = \int_{-L}^L W^2(t) dt \tag{4.1}$$

for any L . On any closed interval, the Gaussian process $W(t)$ can be represented in terms of its eigenfunction expansion. Specifically, let the sequence $\{\lambda_j, \phi_j\}_{j=1}^{\infty}$ solve the system of eigenvalue equations

$$\lambda_j \phi_j(t) = \int_{-L}^L \phi_j(s) V(s, t) ds, \tag{4.2}$$

with the functions $\{\phi_j\}$ taken to be orthonormal. Then

$$W(t) = \sum_{j=1}^{\infty} Z_j \phi_j(t), \quad (4.3)$$

where the Z_j are independent normal random variables with mean zero and variance λ_j . Hence the random variable ξ_L in (4.1) has the representation

$$\xi_L = \sum_{j=1}^{\infty} Z_j^2, \quad (4.4)$$

a sum of independent $\lambda_j \chi_1^2$ variables, and thus the following corollary holds.

COROLLARY 1 *Under the conditions of Theorem 1, as $n \rightarrow \infty$, the random variable*

$$I_n^L \equiv n \int_{-L}^L \{\tilde{f}(t) - \hat{f}(t)\}^2 dt$$

tends to a sum of independent $\lambda_j \chi_1^2$ variables, where the sequence $\{\lambda_j\}_{j=1}^{\infty}$ is defined via (4.2).

REMARK: The corollary to Theorem 1 is not completely satisfactory in that, ideally, the desired result is for $L = \infty$, but the eigenfunction expansion is valid only on a bounded interval. In real data, this will not be of practical concern if one chooses L far enough outside of the range of the data.

5 Computation and affine invariance

5.1 Computing the test statistic

The test statistic, I_n defined by (2.5) can be used to test the fit of the model. In order to be of practical use, it should be easily computable. A simple representation of the test statistic in terms of the residuals of a prospective logistic regression fit is now given.

Let $r_i \equiv y_i - \rho q(\tilde{\alpha} + x_i^T \tilde{\beta})$, where $q(t)$ is as in Theorem 1. This is the vector of residuals that one would obtain in a maximum likelihood fitting of a prospective logistic regression model to the data. Then the test statistic, I_n , given in (2.5), can be computed as a quadratic form in the residual vector from the following theorem.

THEOREM 2 *Let K be a spherically symmetric kernel and let $k(t) \equiv \int K(t-x)K(x)dx$, the convolution density of the kernel with itself. The test statistic, I_n admits the quadratic form representation*

$$I_n = r^T Q r,$$

with the matrix Q given by $(Q)_{ij} \equiv n_0^{-1}(1+\rho) k(x_i - x_j)$, and r is the vector of residuals from the prospective maximum likelihood fit.

The representation is particularly useful in greatly simplifying computation if the chosen kernel admits a closed form for the convolution density, $k(t)$. A simple such choice of kernel is the density of $N(0, I)$ so that $k(t)$ is then just the $N(0, 2I)$ density. This is the choice used in the examples.

5.2 Affine invariance

Along with a computational method, the representation given by Theorem 2 also allows one to see that the test statistic can be chosen so that it remains invariant under any affine transformation of the covariate vector.

Assume that the data has been standardized so that it has mean zero and identity covariance. After doing so, the residuals remain unchanged. From the representation given by Theorem 2 it is then clear that the statistic is invariant under any affine transformation of the covariate vector if the kernel is spherically symmetric. This follows

directly from the affine invariance of the residuals and the fact that the convolution is spherical if the original kernel was chosen as such.

REMARK 1: Even if one were to standardize the data, choosing a test statistic based on the cumulative distribution function as in Qin and Zhang (1997), Gilbert (2004) and Bondell (2005) can yield location-scale invariance, but not full affine invariance. This is due to the fact that the multivariate ordering of the data imposed by the cumulative distribution function is likely to change after rotation.

REMARK 2: Although the covariance matrix must be estimated in order to standardize the data, any root- n consistent scale-equivariant estimator suffices, and the asymptotic distribution is identical to the situation where the true covariance were known.

A simple affine transformation of the covariate vector is the result of changing the sign of one component. In practice, a covariate is often derived as a difference in two quantities. This particular affine transformation then is the result of switching the order of the quantities in taking the difference. Clearly, if one were to claim a reasonable model fit (or lack thereof), this claim should remain consistent after this simple interchange of ordering. A small simulation shows that a lack of affine invariance can give quite surprising results.

A bivariate model with 20 controls given by $N(0, I)$ and 20 cases given by $N(\mu, \sigma I)$ with $\mu = (1, 1)$ and $\sigma = 2$ results in quadratic terms needed for both covariates. A linear logistic model was fit so that the test should detect the lack of fit. After generating 100 samples, matched samples were constructed by changing the sign of only the first component. The test statistic proposed in this paper rejects the logistic model at the $\alpha = .05$ level in 83 out of the 100 samples and gives identical results on both sets of samples, as it is affine invariant. In comparison, the statistic of Qin &

Zhang (1997) rejects only 54 times with the original samples and 49 times after the sign change. This substantial difference in power is explored further in §7. Perhaps even more problematic is the fact that the two conclusions given by this test statistic actually disagree in 29 out of the 100 paired samples. This almost 30% disagreement after a simple sign change should lead one to consider affine invariance as an important characteristic of a test procedure.

5.3 A bootstrap approach

The asymptotic distribution of the test statistic under the null hypothesis derived in Section 4 involves an infinite weighted sum of chi-squared variates. The weights are computed as the eigenvalues of the covariance function given by Theorem 1. One can directly use this asymptotic approximation in the standard way by discretizing the interval $[-L, L]$ to a finite set of N points, and using the empirical estimate of the covariance function to construct an $N \times N$ covariance matrix. Then the infinite set of eigenvalues becomes a finite collection and the weighted chi-squared representation can be used. However, since the test statistic is easily computable, a resampling scheme to approximate the null distribution is instead advocated.

The bootstrap procedure to resample under the true model is straightforward as in Qin and Zhang (1997). Specifically, let $(u_1^*, \dots, u_{n_0}^*)$ be drawn independently with replacement from the complete data (x_1, \dots, x_n) with sampling probabilities given by $d\tilde{F}_0(x; \tilde{\alpha}, \tilde{\beta})$ as in (2.1). Likewise, let $(z_1^*, \dots, z_{n_1}^*)$ be drawn independently with replacement using sampling probabilities $\exp(\tilde{\alpha} + x^T \tilde{\beta}) d\tilde{F}_0(x; \tilde{\alpha}, \tilde{\beta})$. The test statistic is then calculate for this bootstrap sample based on the maximum likelihood logistic regression fit to this sample. This is then repeated a large number of times to construct the bootstrap distribution under the null hypothesis. The hypothesis test can then be conducted using the quantiles of this bootstrap distribution.

R-code to perform this goodness-of-fit test using the bootstrap procedure is available by request from the author.

6 Kyphosis Data

In this section, the proposed goodness-of-fit test is demonstrated on a real data example involving a multiple logistic regression model. Hastie and Tibshirani (1990) discuss a retrospective study of kyphosis on children who received a corrective spinal surgery. Kyphosis is defined as a forward flexion of the spine of at least 40 degrees from vertical. In this study, there are three predictors used to model the presence or absence of kyphosis in the 81 children. These predictors are: age at time of operation, location of starting vertebrae involved in the operation, and the number of vertebrae involved in the operation. Various analyses of this data agree that a linear logistic fit is not sufficient to model the effects, and that the age and start variables appear to have quadratic trends. Hastie and Tibshirani (1990) fit a generalized additive model, whose form of the functions for the predictors age and start are consistent with quadratic functions.

Fitting a linear logistic model with the three predictors, one obtains the test statistic $I_n = 4.1$. The corresponding bootstrap p-value is $p = 0.0075$, very strongly indicating a lack of fit of the simple linear logistic model.

A second model is fit with the three predictors, plus an additional quadratic term for age. For this model, the resulting test statistic is $I_n = 2.8$, with a p-value of $p = 0.0495$, still indicating a poor fit of the model.

A final model including quadratic terms for both age and start is fit and the test statistic obtained is $I_n = 1.7$, with corresponding p-value of $p = 0.3145$. Thus, agreeing with prior conclusions regarding this data set, there does not appear to be evidence to suggest that this model is not a reasonable fit to the data.

7 Simulation Study

The performance of the proposed density based test statistic, I_n , is investigated and compared with two existing goodness-of-fit tests for case-control logistic regression, the Kolmogorov-Smirnov-type statistic, Δ (Qin and Zhang, 1997), and the information matrix-based statistic, D_n (Zhang, 2001) via a small simulation study. The performance was compared by examining the power against some local alternatives to the model. The alternatives considered are generated from the alternative model as proposed in Zhang (1999, 2001).

Consider the model

$$u_1, \dots, u_{n_0} \text{ is a random sample with density } f_0(x), \tag{7.1}$$

$$z_1, \dots, z_{n_1} \text{ is a random sample with density } f_1(x, \nu) = \exp(\alpha + x^T \beta) v(x, \eta) f_0(x),$$

where $v(x, \eta)$ is a known function and there exists a unique η_0 such that $v(x, \eta_0) = 1$ for all x . Hence for alternatives of this form, testing the validity of model (1.2) is equivalent to testing $\eta = \eta_0$ in (7.1).

For the simulation study, it is assumed that $f_0(x)$ is the standard normal density function and that $f_1(x, \eta)$ is the density function of a $N(\mu, \sigma_n^2)$ distribution with $\sigma_n^2 = 1/(1 - 2n^{-1/2}\theta)$ for some $\theta \in \mathbb{R}$ such that $\sigma_n^2 > 0$. It is straightforward to verify that the above model (7.1) then holds with $v(x, \eta) = \exp(\eta_1 + \eta_2 x + \eta_3 x^2)$, where $\eta = (\eta_1, \eta_2, \eta_3)^T$ and

$$\eta_1 = \frac{\mu^2}{2} \left(1 - \frac{1}{\sigma_n^2}\right) - \frac{1}{2} \log \sigma_n^2, \quad \eta_2 = \mu \left(\frac{1}{\sigma_n^2} - 1\right), \quad \eta_3 = \frac{1}{2} \left(1 - \frac{1}{\sigma_n^2}\right).$$

One then has that $\eta = \eta_0 + n^{-1/2}\gamma\{1 + o(1)\}$ as $n \rightarrow \infty$, with $\eta_0 = (0, 0, 0)^T$ and $\gamma = \theta(\mu^2 - 1, -2\mu, 1)^T$. For $\theta = 0$, the model reduces to (1.2). The performance of the goodness-of-fit tests are compared by examining the power against some local alternatives generated by $\theta \neq 0$.

The simulations consider $\theta = 0, 1.5, 3.0$ and sample sizes of $(n_0, n_1) = (40, 20)$ and $(n_0, n_1) = (60, 40)$. Furthermore, $\mu = 0.5$ is fixed, so that $\alpha = -0.125$ and $\beta = 0.5$.

Note that setting $\theta = 0, 1.5, 3.0$ yields $\sigma_n = 1.0, 1.278, 2.106$ and $\sigma_n = 1.0, 1.195, 1.581$ when $n = 60$ and $n = 100$, respectively. These are the settings used in Zhang (1999) for $n = 60$ and Zhang (2001) for $n = 100$.

For each of the six θ and sample size combinations, 1000 samples are generated from model (7.1). For each sample the critical values for each of the 1%, 5%, and 10% levels are determined (via the bootstrap for I_n and Δ , and via a chi-square approximation for D_n). By comparing the sample value of the test statistic to the corresponding critical value, a decision is made for each sample. The achieved significance levels and powers of the test statistics can then be obtained from the proportion of rejections out of the 1000 samples.

(*** TABLE 1 GOES HERE ***)

Table 1 shows the results of the simulations. The performance of the kernel-based statistic of this paper is clearly shown to dominate the two previous proposals, at least under the alternatives considered here. The achieved significance levels for all three test statistics under the null hypothesis, $\theta = 0$, are all close to the nominal level. Under the alternatives considered, the power of the newly proposed test statistic, I_n , is strictly greater than that of either of the others. In fact, in many instances considered, the power of the proposed statistic is more than double that of the non affine-invariant Kolmogorov-Smirnoff-type statistic, Δ . Other simulation setups, including skewed distributions such as the gamma, were considered and the results were similar, thus omitted.

Appendix

Proof of Theorem 2.

By definition, the test statistic,

$$\begin{aligned}
I_n &= n \int \left\{ \tilde{f}(x) - \hat{f}(x) \right\}^2 dx \\
&= n \int \left[n_0^{-1} \sum_{i=1}^n K(x - x_i) \left\{ 1 + \rho \exp(\tilde{\alpha} + x_i^T \tilde{\beta}) \right\}^{-1} - n_0^{-1} \sum_{i=1}^n (1 - y_i) K(x - x_i) \right]^2 dx.
\end{aligned}$$

After some simplification one obtains,

$$I_n = n_0^{-1} (1 + \rho) \sum_{i=1}^n \sum_{j=1}^n r_i r_j \int K(x - x_i) K(x - x_j) dx.$$

Now, since K is spherically symmetric,

$$\begin{aligned}
\int K(x - x_i) K(x - x_j) dx &= \int K(x_i - x) K(x - x_j) dx \\
&= \int K(x_i - x_j - x) K(x) dx = k(x_i - x_j).
\end{aligned}$$

This completes the proof.

References

- Anderson, N. H., Hall, P. and Titterton, D. M. (1994) Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *J. Multivariate Anal.* **50**, 41-54.
- Bondell, H. D. (2005) Minimum distance estimation for the logistic regression model. *Biometrika*, **92**, 724-31.
- Feuerverger, A. and McDunnough, P. (1981) On some Fourier methods for inference. *J. Am. Statist. Assoc.*, **76**, 379-87.
- Gilbert, P. B. (2004) Goodness-of-fit tests for semiparametric biased sampling models. *J. Statist. Plan. Infer.*, **118**, 51-81.
- Hall, P. (1984) Central limit theorem for integrated squared error for multivariate nonparametric density estimator. *J. Multivariate Anal.* **14**, 1-16.
- Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*, London: Chapman and Hall.
- Heathcote, C. R. (1977) The integrated squared error estimation of parameters. *Biometrika*, **64**, 255-64.

- Henze, N., Klar, B. and Meintanis, S. G. (2003) Invariant tests for symmetry about an unspecified point based on the empirical characteristic function, *J. Multivariate Anal.* **87**, 275-97.
- Hosmer, D. W., Hosmer, T., Le Cessie, S. and Lemeshow, S. (1997) A comparison of goodness-of-fit-tests for the logistic regression model, *Statist. Med.*, **16**, 965-80.
- Qin, J. and Zhang, B. (1997) A goodness-of-fit test for logistic regression models based on case-control data, *Biometrika*, **84**, 609-18.
- Vardi, Y. (1985) Empirical distributions in selection bias models, *Ann. Statist.*, **13**, 178-203.
- Zhang, B. (1999) A chi-squared goodness-of-fit test for logistic regression models based on case-control data, *Biometrika*, **86**, 531-9.
- Zhang, B. (2001) An information matrix test for logistic regression models based on case-control data, *Biometrika*, **88**, 921-32.

Table 1: *Achieved significance levels and powers of Δ , D_n and I_n under local alternatives given by model (7.1).*

θ	(n_0, n_1)	σ_n	Nominal Level (%)	Power (%)		
				Δ	D_n	I_n
0	(40, 20)	1.000	10	10.2	10.5	10.7
0	(40, 20)	1.000	5	5.1	5.4	5.5
0	(40, 20)	1.000	1	0.9	1.4	1.0
1.5	(40, 20)	1.278	10	16.9	21.5	30.5
1.5	(40, 20)	1.278	5	9.0	14.9	20.9
1.5	(40, 20)	1.278	1	1.9	6.4	7.5
3.0	(40, 20)	2.106	10	61.1	89.1	91.7
3.0	(40, 20)	2.106	5	46.3	83.0	86.4
3.0	(40, 20)	2.106	1	20.5	64.4	70.8
0	(60, 40)	1.000	10	10.5	10.4	9.8
0	(60, 40)	1.000	5	5.3	5.6	5.4
0	(60, 40)	1.000	1	1.2	1.4	1.2
1.5	(60, 40)	1.195	10	16.8	22.7	29.0
1.5	(60, 40)	1.195	5	9.9	15.1	20.0
1.5	(60, 40)	1.195	1	2.5	4.8	6.9
3.0	(60, 40)	1.581	10	52.6	76.1	85.3
3.0	(60, 40)	1.581	5	36.9	67.6	76.6
3.0	(60, 40)	1.581	1	14.0	43.8	53.3