BIOMETRICS 47, 1267–1282
December 1991

A Goodness-of-Fit Test for Binary Regression Models, Based on Smoothing Methods

S. le Cessie and J. C. van Houwelingen

Department of Medical Statistics, University of Leiden,
P. O. Box 9512, 2300 RA Leiden, The Netherlands

SUMMARY

A new global test statistic for models with continuous covariates and binary response is introduced. The test statistic is based on nonparametric kernel methods. Explicit expressions are given for the mean and variance of the test statistic. Asymptotic properties are considered and approximate corrections due to parameter estimation are presented.

Properties of the test statistic are studied by simulation. The goodness-of-fit method is illustrated on data from a Dutch follow-up study on preterm infants. Recommendations for practitioners are given.

1. Introduction

In biostatistics, models with binary response are often used, for example, to model the probability of getting a disease. The formal description is as follows. Let (X_i, Y_i), $i = 1, \ldots, n$, be a sequence of observations, where Y_i are independent binary (1/0) outcome random variables and X_i are row vectors of covariates. For the probability that $Y = 1$, conditional on $X = x$, we write

$$g(x) = \Pr(Y = 1 \mid X = x).$$

Parametric methods to model this probability are well developed. Well known are the logistic regression model (Cox, 1970) and the probit regression model (Finney, 1947). In order to safely apply a parametric model one needs to know that that particular model is reasonable. Here, we consider the testing of the hypothesis that the real regression function $g(x)$ is equal to some model function $g_0(x)$. Both functions $g(x)$ and $g_0(x)$ are assumed to be continuous.

In this paper, a test statistic is defined to check this hypothesis, making use of nonparametric kernel methods. In this way, it is possible to deal properly with continuous covariates.

This paper is organized as follows. In Section 2 we will discuss some of the drawbacks of current goodness-of-fit methods for the logistic model. Some methods discussed are valid only for the logistic model; other methods can also be applied to other binary models. In Section 3 the test statistic itself is defined, its mean and variance are given, and its performance for finite samples is determined in two simulation examples. Theoretical properties of the test statistic and relations of the theory and the results of the simulations are given in Sections 4 and 5. In Section 4 the link is made with chi-squared test statistics designed for discrete covariates and in Section 5 asymptotic properties of the test statistic T, such as the asymptotic mean and variance and the asymptotic distribution under the null hypothesis, are examined. The problems arising when the function g_0 is estimated are considered in Section 6. In Section 7 recommendations for the use of goodness-of-fit test

Key words: Binary data; Goodness of fit; Kernel estimation; Logistic regression.

in practical applications are given. The test proposed in this paper is applied to a real data set as an illustration. Concluding remarks are made in Section 8.

2. Goodness-of-Fit Methods for the Logistic Regression Model

In this section some methods for assessing the goodness of fit of the logistic model are discussed. For the logistic model the probability that $Y = 1$, conditional on $X = x$, is

$$g_0(x) = \frac{e^{\alpha + x\beta}}{1 + e^{\alpha + x\beta}},$$

or, equivalently,

$$\operatorname{logit}(g_0(x)) = \log[g_0(x)/(1 - g_0(x))] = \alpha + x\beta.$$

If there are a limited number of different covariate patterns and replicated measurements for each covariate pattern, goodness of fit can be examined by the methods intended for categorical data, such as the Pearson chi-squared statistic or the likelihood ratio statistic. The asymptotic distributions of these test statistics are derived under the condition that the number of observations in each cell tends to infinity. These methods therefore fail if there are no replicated measurements.

Many current methods, designed for models with continuous covariates, are based on pooling the observations according to the model probabilities $g_0(x)$.

An example of this approach is the popular method due to Hosmer and Lemeshow (1980). They pool the data according to the value of $g_0(x)$ in a number of groups and calculate a chi-squared-type statistic. A second example is the method due to Brown (1982). Here, the logistic model is embedded in a family of models with two additional parameters. If these parameters are zero the logistic model is obtained.

Kernel methods have been used by Copas (1983) to examine the fit graphically. A kernel estimation $\tilde{g}(z)$ of the function $g(z)$ is calculated, where $z = \alpha + x\beta$, the logit of $g_0(x)$. Then $\operatorname{logit}(\tilde{g}(z))$ is plotted against z. If the model is valid one obtains a straight line at 45° through the origin.

Since these three methods look only in the direction where $g_0(x)$ [or, equivalently, $\operatorname{logit}(g_0(x))$] varies for discrepancies, orthogonal deviations of the model are not detected. These methods are therefore completely insensitive to differences in $g(x)$ within the pooled groups. In Section 3, a simulation example is shown in which the Hosmer–Lemeshow test statistic behaves badly for this reason.

Partitioning of the covariate space to group the data is a way to deal with these problems. However, the problems of how to group the data and how many groups to choose remain. Tsiatis (1980) introduced a score test for the effect of indicator functions of subsets of the covariate space but did not specify the number of subsets and the way of grouping. Landwehr, Pregibon, and Shoemaker (1984) proposed a graphical method to examine the fit of the model using clusters of neighboring points. Their method is based on a partition of the deviance in a pure-error component and a lack-of-fit component. Fowlkes (1987) suggested replacing the clustering method by a nearest-neighbor method. However, in this graphical approach the problem of how to define precisely when to accept and when to reject the null hypothesis remains.

3. The Test Statistic

Smoothing methods can be used to define a goodness-of-fit test that does not suffer from the problems mentioned in Section 2. Generalizing the one-dimensional approach of Copas (1983), one can compare the function $g_0(x)$ with a kernel estimate $\tilde{g}(x)$ of $g(x)$. This has been done recently by Azzalini, Bowman, and Härdle (1989). By simulation, they obtained confidence bands for the nonparametric curve under the null hypothesis. Furthermore they compared the function $g_0(x)$ with $\tilde{g}(x)$ formally, by defining a pseudo-likelihood ratio statistic.

However, a problem occurring in this approach is the bias in the nonparametric estimate $\tilde{g}(x)$. This bias disappears when we consider a test statistic that is based on a kernel estimate of the standardized residuals, as these residuals have expectation 0.

We first discuss the situation where there is a single covariate X. The residuals of the model, $Y_i - g_0(X_i)$, are standardized by dividing them by their standard deviation under the null hypothesis,

$$r(X_i) = \frac{Y_i - g_0(X_i)}{\sqrt{g_0(X_i)(1 - g_0(X_i))}},$$

so that their variance under the null hypothesis is equal to 1. Note that the range of g_0 should be a subset of (0, 1) to avoid division by 0.

A smoothing function of these standardized residuals is obtained by the kernel estimate of Nadaraya (1964) and Watson (1964), which is defined by

$$\tilde{r}(x) = \frac{\sum_j r(X_j)K[(x - X_j)/h_n]}{\sum_j K[(x - X_j)/h_n]}.$$

Here the parameter h_n is the bandwidth, which controls the amount of smoothing. It depends in some way on the sample size n. The function K is a nonnegative symmetric bounded kernel function, zero outside a closed interval $[-a, a]$, and it is normalized according to $\int K(z)\,dz = 1$ and $\int K(z)^2\,dz = 1$.

The smoothed residual $\tilde{r}(x)$ is a weighted average of the residuals in the neighborhood of x, where the bandwidth determines the size of the region over which the residuals are averaged and the kernel function determines the weighting. There is extensive literature on kernel smoothing of regression functions. Two relevant references are Gasser and Müller (1979) and Collomb (1981).

It is not difficult to show that for each x the mean of $\tilde{r}(x)$ conditional on the observed X's is

$$\mathrm{E}(\tilde{r}(x)) = 0,$$

and since the observations are independent,

$$\mathrm{var}(\tilde{r}(x)) = \frac{\sum_j K[(x - X_j)/h_n]^2}{\{\sum_j K[(x - X_j)/h_n]\}^2}.$$

In this paper a weighted sum of the smoothed standardized residuals is used as the goodness-of-fit measure. The test statistic T is defined as

$$T = n^{-1} \sum_i \tilde{r}(X_i)^2 v(X_i),$$

where

$$v(X_i) = \frac{\{\sum_j K[(X_i - X_j)/h_n]\}^2}{\sum_j K[(X_i - X_j)/h_n]^2}$$

is the inverse of the variance of the smoothed standardized residual in X_i. Each term $\tilde{r}(X_i)^2$ is multiplied by $v(X_i)$. Therefore each observation can be expected to yield the same contribution to the statistic T under the null hypothesis. The multiplication of $\tilde{r}(X_i)^2$ by $v(X_i)$ is essential for obtaining decent asymptotic results. Otherwise the asymptotic mean and variance of T depend on the range of X and tend to infinity if the range of X tends to infinity.

If $g(x)$ is equal to $g_0(x)$, the mean of T, conditional on $X_1, \ldots, X_n$, is

$$\mathrm{E}(T) = n^{-1} \sum_i \operatorname{var}(\tilde{r}(X_i))v(X_i) = 1.$$

The variance of T has complex form but can be calculated exactly. This is done in the Appendix. There it is proven that

$$\operatorname{var}(T) = n^{-2} \sum_i \sum_j \left(\sum_k w_{ik}^2 \sum_k w_{jk}^2 \right)^{-1} \times \left(\sum_k \frac{w_{ik}^2 w_{jk}^2 (6g_0(X_k)^2 - 6g_0(X_k) + 1)}{g_0(X_k)(1 - g_0(X_k))} + 2\left(\sum_k w_{ik} w_{jk} \right)^2 \right), \tag{3.1}$$

where $w_{ij} = K[(X_i - X_j)/h_n]$.

It is not difficult to extend this approach to higher dimensions. In d dimensions we choose to use the multiplicative kernel function $\mathbf{K}(z)$, which is defined for $z = (z_1, \ldots, z_d)$ as

$$\mathbf{K}(z) = \prod_{l=1}^{d} K(z_l),$$

where K is a one-dimensional kernel function as defined above. The parameter h_l is the bandwidth for the lth covariate. By defining $h_l = h_n s_l$, with s_l the standard deviation of the lth covariate, the number of bandwidth parameters is reduced to 1. In this situation we will refer to h_n as the standardized bandwidth.

All the results achieved in this and the following sections for the one-dimensional situation can be extended to higher dimensions by replacing the kernel function K in the one-dimensional case by the multiplicative kernel function. To obtain decent asymptotic results, nh_n^d must tend to infinity. Since this is all quite straightforward, only the situation with a one-dimensional covariate is considered in detail in the rest of this paper.

Example 1 In this example the distribution of T under the null hypothesis that g is equal to a known function g_0 is examined. The logistic model $\operatorname{logit}(g(x)) = -3 + 6x$ is used in the first study with x a one-dimensional covariate.

To simplify calculations a uniform kernel is used, defined by $K(z) = 1$ if $-\frac{1}{2} \leqslant z \leqslant \frac{1}{2}$ and $K(z) = 0$ elsewhere. In the first series of simulations, 500 replicates of 100 observations were generated, where x is equally spaced on the interval [0, 1], so $g(x)$ varies between .047 and .953. The exact mean and variance of T were calculated and the cutoff points of T were compared with the cutoff points of the normal distribution and with the cutoff points of a $c\chi^2_{(\nu)}$ distribution, where $\chi^2_{(\nu)}$ has a chi-squared distribution with ν degrees of freedom. This second distribution is a commonly used approximation for distributions of nonnegative random variables [see, for example, Cox and Hinkley (1974, p. 463)]. The constants c and ν are obtained by equating the mean and variance of T and $c\chi^2_{(\nu)}$, yielding $c = \operatorname{var}(T)/(2\mathrm{E}(T))$ and $\nu = 2\mathrm{E}(T)^2/\operatorname{var}(T)$. The performance is determined for different choices of bandwidth. The results are summarized in Table 1.

The data presented in Table 1 show that for all bandwidths the approximation of the cutoff points of T by its asymptotic normal distribution is quite good for significance levels of .10 and .05 but that the distribution of T has heavier tails, since for $\alpha = .01$, the proportion of rejected null hypotheses is too high. When the cutoff points of T are compared with a $c\chi^2_{(\nu)}$distribution, the approximations for the smaller significance levels are improved.

Table 1
Null hypothesis case, 500 *simulations. Comparison of the cutoff points of T with the normal and the* $c\chi^2_{(\nu)}$ *distributions.*

	Normal distribution: Proportion rejected null hypotheses				$c\chi^2_{(\nu)}$ distribution: Proportion rejected null hypotheses			
Bandwidth	$\alpha = .10$	$\alpha = .05$	$\alpha = .025$	$\alpha = .01$	$\alpha = .10$	$\alpha = .05$	$\alpha = .025$	$\alpha = .01$
.015	.112	.056	.038	.014	.104	.050	.022	.004
.105	.110	.076	.044	.028	.110	.052	.028	.016
.255	.106	.064	.038	.022	.096	.040	.022	.008
.505	.088	.052	.040	.028	.080	.040	.022	.014
.755	.074	.048	.034	.030	.074	.034	.024	.010

Example 2 In the second simulation series the power of T is examined. Observations were generated from the model $\text{logit}(g(x_1, x_2)) = -3 + 3x_1 + (3x_2 - 1.5)^2$ and compared with a logistic model lacking the effect of x_2, $\text{logit}(g_0(x_1)) = \alpha + \beta x_1$. We avoid the effects of estimating the parameters by using in all simulations for g_0 the function $\text{logit}(g_0(x_1)) = -2.03 + 2.72x_1$. The values of -2.03 and 2.72 are the maximum likelihood estimates for α and β, obtained from a very large simulated data set.

The model was simulated 500 times and for each simulation 500 observations were generated, where $x_1 = 0(\frac{1}{9})1$ and $x_2 = 0(\frac{1}{49})1$. The results of the simulations are summarized in Table 2. The performance of T depends heavily on the bandwidth. If it is chosen too small the statistic has no power and if it is chosen too big all local deviations are smoothed away.

The Hosmer–Lemeshow statistic was also calculated in each simulation, with the observations pooled according to their value $g_0(x)$ in 10 equal sized groups. The cutoff points of the Hosmer–Lemeshow statistic are compared with a $\chi^2_{(10)}$ distribution. The number of degrees of freedom of this distribution is adjusted for the fact that the function g_0 is not re-estimated in each simulation. The results are given in Table 3.

Table 2
Alternative model $\text{logit}(g(x_1, x_2)) = -3 + 3x_1 + (3x_2 - 1.5)^2$, 500 *simulations. Comparison of the cutoff points of T with the normal and the* $c\chi^2_{(\nu)}$ *distributions.*

	Normal distribution: Proportion rejected null hypotheses				$c\chi^2_{(\nu)}$ distribution: Proportion rejected null hypotheses			
Bandwidth	$\alpha = .10$	$\alpha = .05$	$\alpha = .025$	$\alpha = .01$	$\alpha = .10$	$\alpha = .05$	$\alpha = .025$	$\alpha = .01$
.05	.690	.562	.458	.318	.690	.556	.428	.292
.15	.910	.856	.796	.692	.910	.838	.748	.638
.25	.988	.984	.984	.984	.988	.984	.984	.966
.35	.990	.984	.980	.970	.988	.982	.976	.942
.50	.988	.974	.948	.924	.986	.964	.916	.856
.75	.620	.502	.412	.342	.590	.416	.312	.210

Table 3
Comparison of the cutoff points of the Hosmer–Lemeshow statistic and a χ^2_{10} *distribution under the alternative model*

Proportion rejected null hypotheses			
$\alpha = .10$	$\alpha = .05$	$\alpha = .025$	$\alpha = .01$
.058	.028	.014	.008

In the calculation of the Hosmer–Lemeshow statistic, the observations are grouped such that the local fluctuations are cancelled out in each pooled cell. Hence the Hosmer–Lemeshow statistic is not able to detect the deviations of the model but behaves as if g is equal to g_0.

4. Discrete Covariates

Although the test statistic T is designed for continuous covariates, it is interesting to see the links between T and ordinary chi-squared statistics, intended for discrete covariates. Consider the situation where the covariates are discrete and the number of different covariate patterns is finite. The observations can be grouped into m cells, corresponding to the m different covariate patterns. If the bandwidth h_n is chosen so small that for each two different covariate patterns X_i and X_j, $K[(X_i - X_j)/h_n] = 0$, then T becomes

$$T = n^{-1} \sum_{\text{cells}} n_i \tilde{r}(X_i)^2 v(X_i),$$

where n_i is the number of observations in cell i and X_i is the covariate pattern in cell i. For this situation $\tilde{r}(X_i)^2$ can be rewritten as

$$\tilde{r}(X_i)^2 = \frac{((\sum Y - n_i g_0(X_i))K(0))^2}{g_0(X_i)(1 - g_0(X_i))} \times \frac{1}{(n_i K(0))^2},$$

where the sum of the Y's is taken within cell i. Therefore

$$\tilde{r}(X_i)^2 = \frac{(O_i - E_i)^2}{n_i^2 g_0(X_i)(1 - g_0(X_i))}.$$

Here O_i is the observed number of Y's in cell i with outcome 1 and E_i the expected number. So in this situation the terms $n_i \tilde{r}(X_i)^2$ are equal to ordinary chi-squared statistics. Since $v(X_i)$ is in this situation equal to n_i, T will become

$$T = n^{-1} \sum_{\text{cells}} n_i \frac{(O_i - E_i)^2}{n_i g_0(X_i)(1 - g_0(X_i))}.$$

Thus T is a weighted sum of chi-squared statistics with weights $v(X_i) = n_i$. In our approach it is not possible to define a statistic that is equal to the ordinary Pearson chi-squared statistic for discrete covariates, since the weight factors $v(X_i)$ are necessary in the definition of T to deal with unbounded covariates.

This connection between T and chi-squared statistics is an argument to compare the distribution of T under the null hypothesis with a function of chi-squared distributions. As we saw in the previous section, the choice of estimating the distribution of T by a $c\chi^2_{(\nu)}$ distribution was in the first simulation series of Section 3 quite satisfying.

5. Asymptotic Properties

In this section some asymptotic properties of $\tilde{r}(x)$ and the test statistic T will be discussed in order to examine the large-sample behavior of the test statistic. Therefore we assume some regularity conditions. The design points $(X_1, \ldots, X_n)$ must be densely distributed in the limit over the interval of interest. The empirical distribution of the design tends to a distribution with bounded density f and the function g_0 is bounded away from 0 and 1 to achieve bounded standardized residuals.

Bickel and Rosenblatt (1973) and Rosenblatt (1975) considered the asymptotic distribution of different measures of kernel density estimates. Rosenblatt proved the asymptotic normality for certain quadratic forms of the density estimate by partitioning the form in a sum of independent big blocks of length Δ and small strips of length h_n, with $h_n = o(\Delta)$. In this way asymptotic normality for quadratic forms of kernel estimates of regression functions can also be proved. If the design is chosen such that the points are reasonably spread on the interval of interest with an empirical distribution with no extreme peaks, asymptotically the small strips will have no influence on the distribution. The central limit theorem then yields that the asymptotic distribution of the sum of the independent big blocks is normal.

The asymptotic variance of T under the null hypothesis is examined in the Appendix. If $nh_n \to \infty$ and $h_n \to 0$ for $n \to \infty$, then a first-order approximation of the variance of T is

$$\operatorname{var}(T) = 2h_n \int_x f(x)^2\, dx \int_\alpha \left(\int_z K(z)K(z + \alpha)\, dz \right)^2 d\alpha.$$

This approximation for the variance can be useful if there are many data points and the bandwidth is large, since the calculation of the exact variance is in this case very time-consuming. The integral $\int f(x)^2\, dx$ can be approximated by $n^{-1} \sum \hat{f}(X_i)$, with $\hat{f}(X_i) = (nh_n)^{-1} \sum_j K[(X_i - X_j)/h_n]$, the kernel estimate of the density. To see the effects of using the asymptotic variance, we return in the following example to the situation described in Example 1.

Example 3 Consider the situation of Example 1. It is easy to calculate the asymptotic variance of T in this situation, assuming that the covariate x is uniformly distributed on $[0, 1]$, since

$$2h_n \int_x f(x)^2\, dx \int_\alpha \left(\int_z K(z)K(z + \alpha)\, dz \right)^2 d\alpha = 4h_n \int_0^1 \left(\int_{-1/2}^{1/2-\alpha} 1\, dz \right)^2 d\alpha = \frac{4h_n}{3}.$$

The same simulated data are used as in Example 1 but now T is compared with the normal and $c\chi^2_{(\nu)}$ distributions, using the asymptotic variance instead of the exact variance. In Table 4 the results are summarized.

For $h_n = .015$ the exact and asymptotic variances are different and the estimation by the asymptotic variance gives bad results but for larger bandwidths the size of T is still well controlled. Comparing this table with Table 1, we see that for large bandwidths the effect of replacing the exact variance by the asymptotic variance is small. Since the computational effort needed to calculate the variance of T increases with bandwidth, using the asymptotic variance instead of the exact variance for the larger bandwidths will save a lot of time and does not influence the results too much.

Table 4

Data of Example 1, *using the asymptotic variance instead of the exact variance. Comparison of the cutoff points of T with the normal and the* $c\chi^2_{(\nu)}$ *distributions.*

			Normal distribution: Proportion rejected null hypotheses				$c\chi^2_{(\nu)}$ distribution: Proportion rejected null hypotheses			
Bandwidth	Exact variance	Asymptotic variance	$\alpha = .10$	$\alpha = .05$	$\alpha = .025$	$\alpha = .01$	$\alpha = .10$	$\alpha = .05$	$\alpha = .025$	$\alpha = .01$
.015	.048	.020	.198	.152	.114	.074	.198	.144	.096	.056
.105	.174	.140	.122	.086	.062	.034	.116	.078	.036	.020
.255	.353	.340	.108	.064	.040	.022	.102	.042	.022	.008
.505	.653	.673	.084	.050	.036	.024	.078	.040	.022	.012
.755	.969	1.006	.074	.046	.032	.030	.072	.032	.022	.010

6. The Effect of Estimation

In the previous sections the function g_0 was known. Before T can be used as a real goodness-of-fit statistic, we also must consider the situation where g_0 is estimated. Suppose g_0 can be written as

$$g_0(x) = g(x, \beta_0),$$

with g a known function (for example, the logistic function) and β_0 an unknown parameter column-vector of fixed length. We write $\hat{\beta}$ for the maximum likelihood estimate of β_0 and $\hat{g}(x)$ for $g(x, \hat{\beta})$. Under mild regularity conditions $\hat{\beta}$ satisfies $n^{1/2}(\hat{\beta} - \beta_0) = O_p(1)$. The test statistic we achieve when the true parameters are replaced by the estimated parameters is

$$\hat{T} = n^{-1} \sum_i \tilde{r}_e(X_i)^2 v(X_i),$$

where $\tilde{r}_e(X_i)$ are the smoothed standardized residuals with the true function value $g_0(X_i)$ replaced by the estimated value $\hat{g}(X_i)$. It can be shown that, with some regularity conditions on the design and the kernel function, $\hat{T}$ behaves asymptotically as T.

However, although asymptotically the effect of estimation is negligible, for a finite sample this effect can be quite large. The estimated function $\hat{g}$ tends to be closer to the data points, so the mean and variance of $\hat{T}$ will be smaller than the mean and variance of T. This effect will increase if the bandwidth increases and the residuals are averaged over a larger region. For the logistic model, correcting for this effect can be done in the following way. For convenience we change to matrix notation and write the logistic model in matrix form,

$$\operatorname{logit}(\hat{g}(X)) = 1\hat{\beta}_0 + X\hat{\beta}_1, = X^*\hat{\beta}, \tag{6.1}$$

with $X^* = (1, X)$, the concatenation of the matrix of covariates X with a vector of ones to handle the constant term. A Taylor expansion of $Y_i - \hat{g}(X_i)$ about its value in β_0, the real parameter value, gives

$$Y_i - \hat{g}(X_i) = Y_i - g_0(X_i) - (\hat{\beta} - \beta_0)' \left.\frac{\partial g(X_i, \beta)}{\partial \beta}\right|_{\beta=\beta_0} + o_p(n^{-1/2}), \tag{6.2}$$

where a prime indicates the transpose of a matrix. Expanding the first derivative of the log likelihood around β_0 yields

$$\left.\frac{\partial \log L(\beta)}{\partial \beta}\right|_{\beta=\hat{\beta}} = \left.\frac{\partial \log L(\beta)}{\partial \beta}\right|_{\beta=\beta_0} + (\hat{\beta} - \beta_0)' \left.\frac{\partial^2 \log L(\beta)}{\partial \beta^2}\right|_{\beta=\beta_0} + o_p(n^{-1/2}). \tag{6.3}$$

Combining (6.2) and (6.3) and calculating all the derivatives in these equations, we can write

$$Y - \hat{g}(X) = (I - H)(Y - g_0(X)) + o_p(n^{-1/2}),$$

where H is the matrix

$$H = VX^*[(X^*)'VX^*]^{-1}(X^*)'$$

and with V the diagonal matrix with the weights $v_i = g_0(X_i)(1 - g_0(X_i))$ on the diagonal.

Using the expression $(I - H)(Y - g_0(X))$ as an approximation for $Y - \hat{g}(X)$, estimates for the mean and variance of the test statistic with estimated parameters can be made. Therefore we write $\hat{T}$ as

$$\hat{T} = n^{-1} \sum_i [W_i'\hat{V}^{-1/2}(Y - \hat{g}(X))]^2, \tag{6.4}$$

with W_i the column vector with jth element the coefficient $w_{ij}/(\sum_k w_{ik}^2)^{1/2}$ and $\hat{V}$ the diagonal matrix with the weights $\hat{v}_i = \hat{g}(X_i)(1 - \hat{g}(X_i))$ on the diagonal. By replacing $Y - \hat{g}(X)$ by $(I - H)(Y - g_0(X))$ and $\hat{V}$ by V in (6.4), the mean of $\hat{T}$ under the null hypothesis, conditional on $X_1, \ldots, X_n$ can now be approximated by

$$\begin{aligned} \mathrm{E}(\hat{T}) &= n^{-1} \sum_i W_i' V^{-1/2}(I - H)V(I - H)' V^{-1/2} W_i \\ &= n^{-1} \sum_i W_i'(I - V^{-1/2}HV^{1/2})W_i \\ &= n^{-1} \sum_i W_i'(I - V^{1/2}X^*[(X^*)' VX^*]^{-1}(X^*)' V^{1/2})W_i \\ &= \mathrm{E}(T) - n^{-1} \sum_i W_i' V^{1/2}X^*[(X^*)' VX^*]^{-1}(X^*)' V^{1/2} W_i. \end{aligned} \tag{6.5}$$

There is an obvious shrinkage effect of the mean in this approximation, since the second term of this expression is always positive. The calculation of the variance is more complex but can be done in the same way as is shown in the Appendix in the situation where the true function g_0 is known. This yields

$$\begin{aligned} \mathrm{var}(\hat{T}) = n^{-2} \sum_i \sum_j \left(\sum_k w_{ik}^2 \sum_k w_{jk}^2 \right)^{-1} \\ \times \left(\sum_k \left(c_{ik}^2 c_{jk}^2 (6g_0(X_k)^2 - 6g_0(X_k) + 1) g_0(X_k)(1 - g_0(X_k)) \right) \right. \\ \left. + 2\left(\sum_k c_{ik}c_{jk}\, g_0(X_k)(1 - g_0(X_k)) \right)^2 \right), \end{aligned} \tag{6.6}$$

with

$$c_{ik} = \frac{w_{ik}}{\sqrt{g_0(X_k)(1 - g_0(X_k))}} - \sum_s \frac{w_{is} h_{sk}}{\sqrt{g_0(X_s)(1 - g_0(X_s))}}.$$

Substituting for the unknown function values of g_0 the values of the estimated function $\hat{g}$ yields estimates of the mean and variance of $\hat{T}$ for which the size of the test statistic is well controlled, as is shown in the following example.

Example 4 Finally, the effect of estimating the parameters is considered. In each simulation series 100 observations were generated from the model $\mathrm{logit}(g(x)) = -3 + 6x$, $x = 0(\frac{1}{99})1$, but this time for each replication the maximum likelihood estimates of the parameters were calculated by a Newton–Raphson iterative procedure. The test statistic was calculated, using the uniform kernel and as bandwidth $h_n = .15$. The mean and variance of the test statistic were computed, using the corrections (6.5) and (6.6). The cutoff points are given in Table 5. From this it can be seen that when the parameters are estimated, the

Table 5
The performance of T under the null hypothesis with estimated parameters, 100 *simulations, bandwidth* = .15. *Comparison of the cutoff points of T with the normal and the* $c\chi^2_{(\nu)}$ *distributions.*

Normal distribution: Proportion rejected null hypotheses			$c\chi^2_{(\nu)}$ distribution: Proportion rejected null hypotheses		
$\alpha = .10$	$\alpha = .05$	$\alpha = .025$	$\alpha = .10$	$\alpha = .05$	$\alpha = .025$
.08	.05	.02	.08	.03	.00

size of the test statistic is also well controlled. As in the situation when the true parameters were known, the comparison with the $c\chi^2_{(\nu)}$ distribution yields more conservative results.

7. Practical Considerations

The test statistic and its adjusted mean can be calculated easily. The adjusted variance involves far more computational effort, since the correlation between the squared residuals has to be calculated for each pair of observations. The asymptotic variance requires far less effort.

In the literature of kernel regression and density estimation, it is suggested that the choice of the kernel function is not so important. The simplest kernel to deal with is the uniform kernel on $[-\frac{1}{2}, \frac{1}{2}]$, which has been used throughout this paper. There is no evidence that a different type of kernel function would alter the results significantly.

The choice of the bandwidth is crucial. It depends on the number of observations, the number of variables used in the smoothing procedure, and the kind of alternatives expected. Furthermore, when the parameters are estimated, the mean and variance of $\hat{T}$ tend toward 0 when the bandwidth increases. Hence, if the asymptotic variance is used, the bandwidth must not be chosen too large to prevent too conservative a test.

Based on our experiences, we suggest a bandwidth such that each region over which the residuals are averaged contains approximately $\sqrt{n}$ observations. In the simulation Examples 1, 3, and 4, this yields a bandwidth of about .10 and in Example 2, this suggestion corresponds to a bandwidth of about .22. As can be seen from the examples, this is a sensible bandwidth, though on the small side in the cases considered here. But as remarked in the previous paragraph, it is quite important that the bandwidth be not too large when the parameters are estimated.

Azzalini et al. (1989) proposed the use of a cross-validated bandwidth, such that the kernel estimate $\tilde{g}$ of g is optimal. However, cross-validation is a time-consuming procedure. Furthermore, it is not clear that the bandwidth yielding the best kernel estimate $\tilde{g}$ is also the bandwidth corresponding to the greatest power of the test statistic.

When there are many different covariates included in the model, power can be gained if the residuals are not smoothed according to all variables but to a carefully selected subset of covariates. For example, if higher-order terms are included in the model, only the variables corresponding to the linear terms may be used to smooth the residuals. In this case the matrix X^* in the formulas (6.5) and (6.6) must contain all variables in the model.

In the simulation studies, both the normal and the scaled chi-squared distributions controlled the size of the test statistic well. The scaled chi-squared distribution is slightly favored, especially if the significance level is small, because it yields somewhat more conservative results.

The individual components of the test statistic are very helpful in case analysis. These standardized smoothed residuals are a diagnostic tool in detecting outliers and indicating directions in which the model can be improved. They can be examined and plotted in the same way as residuals in linear regression. For example, the residuals can be plotted against explanatory variables or fitted values. An extensive treatment of residual analysis is given in Cook and Weisberg (1982). Note that the use of these smoothed residuals is preferable to an approach based on a direct comparison of the model-based estimate versus a nonparametric kernel estimate of the probability function since the kernel estimate can be biased.

Using these recommendations, one obtains an easy-to-perform method to assess the goodness of fit of a logistic model. To use the test statistic on a computer, we have written a SAS/IML macro, FITTEST, which can be implemented directly in SAS programs. It assesses the goodness of fit of a logistic model, fitted by the SAS procedure PROC LOGIST.

The macro computes the test statistic, the adjusted mean, and the asymptotic variance and compares the test statistic with a scaled chi-squared distribution. As input parameters the macro requires the set of variables, according to which the residuals are smoothed, and a value for the bandwidth. If no bandwidth is specified the default value is chosen such that each smoothing region contains approximately $\sqrt{n}$ points. This macro is available upon request from the authors.

Example 5 The data for this example come from the Project on Preterm and Small-for-Gestational-Age Infants in The Netherlands (POPS), a Dutch follow-up study on preterm infants (Verloove and Verwey, 1988). Data were collected on 1,338 infants, born in 1983 in The Netherlands, with a gestational age of less than 32 completed weeks and/or a birthweight of less than 1,500 g. In this example gestational age and birthweight are used as covariates. The outcome considers the situation after 2 years. The dependent variable Y is 1 if an infant has died within 2 years after birth or survived with a major handicap. After deleting the observations with missing data, a data set of 1,310 infants remains.

A logistic model was fitted on these data with both covariates linear. This yields as model $\text{logit}(g(x_1, x_2)) = \alpha + \beta_1 x_1 + \beta_2 x_2$, where x_1 represents gestational age in weeks and x_2 represents birthweight in 100g. The maximum likelihood estimates of the unknown parameters are $\hat{\alpha} = 9.482$ (s.e. = .807), $\hat{\beta}_1 = -.292$ (s.e. = .030), and $\hat{\beta}_2 = -.120$ (s.e. = .024) and the deviance has the value 1,403.21.

To calculate the test statistic T, a uniform kernel is used and the residuals are smoothed according to the two explanatory variables. The value of the default standardized bandwidth is .665. To obtain the bandwidths for each covariate, this bandwidth is multiplied by the standard deviation of the covariate. The value of T is 1.83; its mean, adjusted for estimation of the parameters, is .90 and the asymptotic variance is .0385. Comparing T with the normal distribution yields $P = 5 \times 10^{-7}$ and comparing T with the scaled chi-squared distribution yields $P = 8 \times 10^{-5}$. Hence, the departures of the linear model are very significant.

Two other goodness-of-fit statistics, the Hosmer–Lemeshow statistic, with the observations pooled in 10 equally sized groups, according to the predicted probability, and the Brown statistic, are calculated. Both statistics are also very significant, with P-values less than .001.

Information about why the model does not fit can be found by plotting the individual contribution of each observation to the test statistic. In Figure 1 these terms $\tilde{r}(X_i)^2 v(X_i)$, multiplied by the sign of $\tilde{r}(X_i)$, are drawn. Observations with a large positive term are indicated with a "★", observations with a large negative term with a "●". To show the distribution of the data, the points with a small contribution to the test statistic are also plotted in Figure 1. We see that in the upper left-hand corner the residuals are very large. These residuals correspond to children with a normal gestational age but with a very low birthweight. These children are growth-retarded and the simple model predicts far too small a probability of failure for these children. Note that in the calculation of the Hosmer–Lemeshow statistic and the Brown statistic, this group of observations is pooled with a totally different group of observations about infants who have a high predicted probability of being healthy—those with a low gestational age and a high birthweight.

Since the model predicts too low a risk for the smallest and largest gestational ages and too high a risk for the observations in the center, quadratic terms are included in the model. In this model, $\text{logit}(g(x_1, x_2)) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 - 30)^2 + \beta_4 (x_2 - 12)^2$, the maximum likelihood estimates are $\hat{\alpha} = 9.943$ (s.e. = .838), $\hat{\beta}_1 = -.319$ (s.e = .031), $\hat{\beta}_2 = -.124$ (s.e = .026), $\hat{\beta}_3 = .0357$ (s.e = .0068), and $\hat{\beta}_4 = .0189$ (s.e = .0042) with a

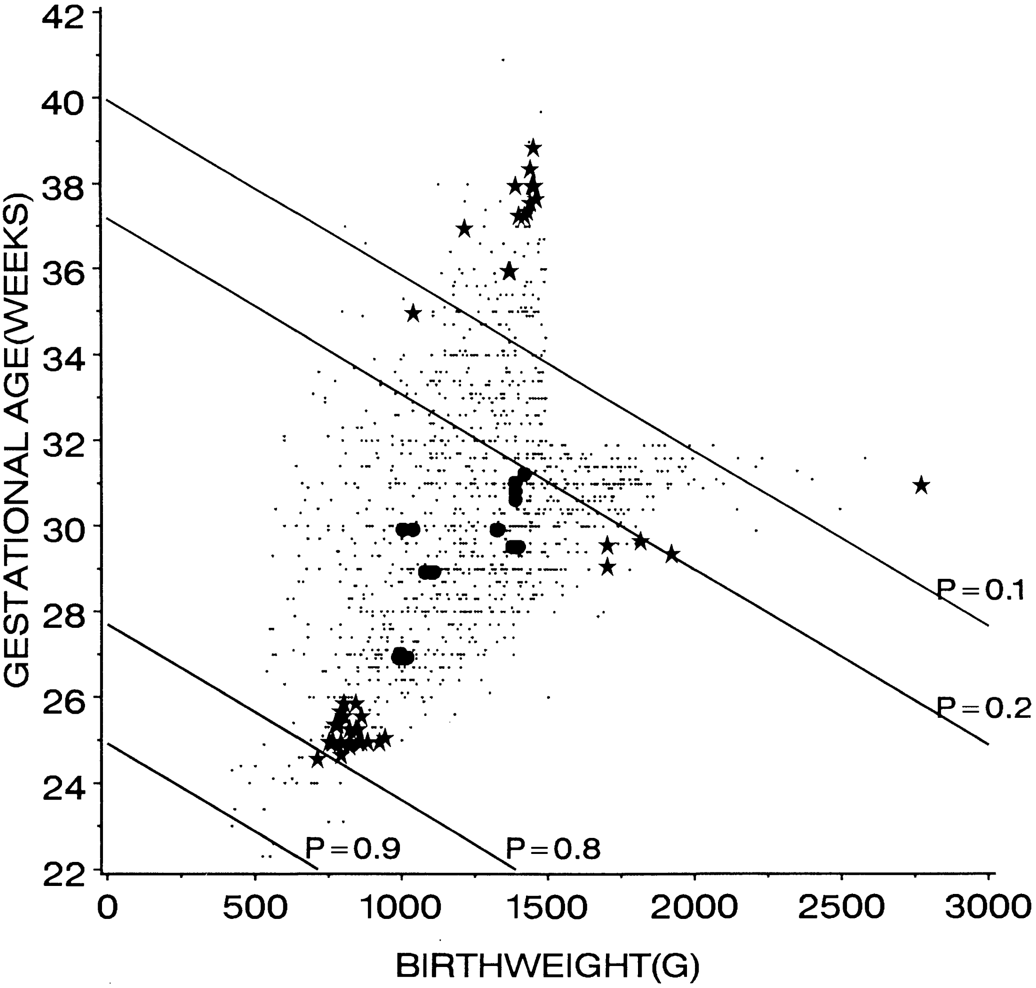

Figure 1. Plot of the contribution of the individual observations to the test statistic. The observations with a high contribution to the test statistic (i.e., $v(X_i)\tilde{r}(X_i)^2 \geqslant 6.25$) are indicated with a "★" if they have a positive smoothed standardized residual and otherwise with a "●". Small dots indicate the observations with a small contribution. The contour lines for $\hat{p} = .1, .2, .8, .9$ are also drawn.

deviance of 1,357.28. Both the quadratic term in gestational age and the quadratic term in birthweight contribute significantly to the model.

Adding to the model both quadratic terms yields $T = .98$, with an adjusted mean of .87. To compute T for this model, the same bandwidth was used and smoothing of the residuals in the test statistic was done according to the variables corresponding to the linear terms. The asymptotic variance does not depend on the estimation of the parameters and it remains the same as in the linear model. Comparing T with a normal distribution and with the scaled chi-squared distribution gave no evidence against the null hypothesis (P-values .29 and .28, respectively). This is confirmed by the fact that a likelihood ratio test showed that neither one of the third-order terms nor a first-order interaction term contributes significantly to the model. The Hosmer–Lemeshow statistic was also not significant ($P = .19$). Only the Brown statistic is still slightly significant ($P = .06$). However, the Brown statistic is not an overall goodness-of-fit test but only a specific check if the link function is logistic. It suggests that change of the link function would improve the model slightly.

8. Discussion

The test statistic proposed in this paper solves many of the problems with the methods mentioned in Section 2. The test statistic T detects deviations of the model in all directions, in contrast to the methods of Hosmer and Lemeshow (1980), Brown (1982), and Copas (1983). There is no need to partition the data in subsets as is done, for example, in the method of Tsiatis (1980). Each observation is treated in the same manner: A weighted average of the standardized residuals in its neighborhood is calculated. Furthermore, it can be formally specified when the null hypothesis should be accepted and when rejected, since the cutoff points of this test statistic are well approximated under the null hypothesis by its asymptotic normal distribution and even better by the scaled chi-squared distribution. The individual contribution of the observations to the test statistic can be used as a diagnostic tool to detect the parts of the data where the model does not fit.

The choice of the bandwidth is crucial and the suggestion made in Section 7 is quite ad hoc. At the moment we are studying a more sophisticated approach, in which the test statistic $T(h)$ is computed for a range of different bandwidths h and significance testing is based on the asymptotic distribution of the process $T(h)$.

This paper has considered only the case when the model contains continuous covariates. In the case of both continuous and categorical variables a different treatment is required. We distinguish three different approaches.

The first method is based on stratification by the categorical variables, and can be used if the number of categories is small compared to the number of observations. Consider $\sum T_s$, where T_s is the test statistic in the sth category. The mean and asymptotic variance of $\sum T_s$ are just the sum of the mean and asymptotic variance of the separate T_s's. The distribution of $\sum T_s$ can be approximated by a scaled chi-squared distribution $c\chi^2_{(\nu)}$, where c and ν may be approximated as described in Example 1.

A more complicated situation occurs when there are several categorical variables, and the number of observations per category is small. One approach is to smooth the residuals only in the directions of the continuous covariates and ignore the effect of the categorical variables on the fit of the model. If the lack of fit is caused only by incorrect modelling of the continuous variables, this would be a good approach. In this situation one has to assume that there is no mutual interaction between the categorical variables in the model, and none between the categorical and continuous variables.

If incorrect modelling of the categorical covariates is suspected, a different approach has to be taken. In the following we give some thoughts of a third approach, although more research is needed before it can be used by practitioners. We are thinking of an "overlap" kernel, a kernel function based on the number of variables on which two observations coincide.

One way to realize this is to associate each value of a categorical variable with a separate dummy variable. The test statistic can then be calculated as if the dummy variables were continuous. The number of categorical variables on which two observations differ is in this way measured by the distance between their sets of dummy variables.

However, not all results obtained in the previous sections will hold if this approach is taken. The exact expressions for the mean and variance of T are still valid, but the asymptotic expression for the variance of T and the statements about the asymptotic normal distribution will not hold. In obtaining the asymptotic expressions, we have used that in the limit the X's are densely distributed. The recommendations for the choice of the bandwidth also must be reconsidered.

In summary, when categorical variables are included in the model, the first and second approaches can be used safely. Further research is needed before the third approach can be applied.

Résumé

Un nouveau test pour des modèles avec réponse binaire et variables explicatives continues est introduit. Le test repose sur la méthode non paramétrique du noyau. Des expressions explicites sont fournies pour la moyenne et la variance de la statistique de test. Les propriétés asymptotiques sont étudiées et des formules de corrections liées à l'estimation des paramètres sont présentées.

Les propriétés du test sont étudiées par simulation. Le test d'adéquation est appliqué à des données provenant d'une étude néerlandaise sur les enfants prématurés. Des recommendations à usage des praticiens sont données.

References

Azzalini, A., Bowman, A. W., and Härdle, W. (1989). On the use of nonparametric regression for model checking. *Biometrika* **76,** 1–12.

Bickel, P. J. and Rosenblatt, M. (1973). On some global measure of the deviations of density function estimates. *Annals of Statistics* **1,** 1071–1095.

Brown, C. C. (1982). On a goodness-of-fit test for the logistic model based on score statistics. *Communications in Statistics—Theory and Methods* **11,** 1087–1105.

Collomb, G. (1981). Estimation non paramétrique de la régression: Revue bibliographique. *International Statistical Review* **49,** 75–93.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* London: Chapman and Hall.

Copas, J. B. (1983). Plotting *p* against *x*. *Applied Statistics* **32,** 25–31.

Cox, D. R. (1970). *The Analysis of Binary Data.* London: Methuen.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics.* London: Chapman and Hall.

Finney, D. J. (1947). *Probit Analysis.* Cambridge: Cambridge University Press.

Fowlkes, E. B. (1987). Some diagnostics for binary logistic regression via smoothing. *Biometrika* **74,** 503–515.

Gasser, Th. and Müller, H. G. (1979). Kernel estimations of regression functions. In *Smoothing Techniques for Curve Estimation,* Th. Gasser and M. Rosenblatt (eds), 23–68. Berlin: Springer-Verlag.

Hosmer, D. W. and Lemeshow, S. (1980). Goodness-of-fit tests for the multiple logistic regression model. *Communications in Statistics—Theory and Methods* **9,** 1043–1069.

Landwehr, J. M., Pregibon, D., and Shoemaker, A. C. (1984). Graphical methods for assessing logistic regression models (with Discussion). *Journal of the American Statistical Association* **79,** 61–83.

Nadaraya, E. A. (1964). On estimation regression. *Theory of Probability and Its Applications* **9,** 141–142.

Rosenblatt, M. (1975). A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *Annals of Statistics* **3,** 1–14.

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

Verloove, S. P. and Verwey, R. Y. (1988). Project on preterm and small for gestational age infants in The Netherlands, 1983 (Thesis, University of Leiden). Ann Arbor, Michigan: University Microfilm International. No. 8807276.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā, Series A* **26,** 359–372.

Received May 1989; *revised January and May* 1990; *accepted July* 1990.

Appendix

Variance of T, Exact and Asymptotic

We calculate the exact and asymptotic variance, conditional on the observed X's, of the test statistic T under the hypothesis that $g_0(x) = g(x)$. Recall that T is defined as

$$T = n^{-1} \sum_i \tilde{r}(X_i)^2 v(X_i).$$

For brevity we will write

$$w_{ij} = K\left(\frac{X_i - X_j}{h_n}\right).$$

The variance will be

$$\operatorname{var}(T) = n^{-2} \sum_i \sum_j v(X_i)v(X_j)\operatorname{cov}(\tilde{r}(X_i)^2, \tilde{r}(X_j)^2). \tag{A.1}$$

The covariance between $\tilde{r}(X_i)^2$ and $\tilde{r}(X_j)^2$ can be written as

$$\begin{aligned}\operatorname{cov}(\tilde{r}(X_i)^2, \tilde{r}(X_j)^2) &= \left(\sum_k w_{ik} \sum_k w_{jk}\right)^{-2} \\ &\quad \times \operatorname{cov}\left(\left(\sum_k \frac{(Y_k - g_0(X_k))w_{ik}}{\sqrt{g_0(X_k)(1 - g_0(X_k))}}\right)^2, \left(\sum_k \frac{(Y_k - g_0(X_k))w_{jk}}{\sqrt{g_0(X_k)(1 - g_0(X_k))}}\right)^2\right) \\ &= \left(\sum_k w_{ik} \sum_k w_{jk}\right)^{-2} \operatorname{cov}\left(\left(\sum_k Z_k c_k\right)^2, \left(\sum_k Z_k d_k\right)^2\right),\end{aligned} \tag{A.2}$$

with $Z_k = Y_k - g_0(X_k)$,

$$c_k = \frac{w_{ik}}{\sqrt{g_0(X_k)(1 - g_0(X_k))}}, \quad \text{and} \quad d_k = \frac{w_{jk}}{\sqrt{g_0(X_k)(1 - g_0(X_k))}}.$$

Then, using the fact that Z_k has mean 0, straightforward calculations yield

$$\begin{aligned}\operatorname{cov}\left(\left(\sum_k Z_k c_k\right)^2, \left(\sum_k Z_k d_k\right)^2\right) &= \sum_k c_k^2 d_k^2 \mathrm{E}(Z_k^4) + \sum_{k \neq l} c_k^2 d_l^2 \mathrm{E}(Z_k^2)\mathrm{E}(Z_l^2) \\ &\quad + 2 \sum_{k \neq l} c_k c_l d_k d_l \mathrm{E}(Z_k^2)\mathrm{E}(Z_l^2) - \left(\sum_k c_k^2 \mathrm{E}(Z_k^2)\right)\left(\sum_k d_k^2 \mathrm{E}(Z_k^2)\right) \\ &= \sum_k c_k^2 d_k^2 \mathrm{E}(Z_k^4) + 2\left(\sum_k c_k d_k \mathrm{E}(Z_k^2)\right)^2 - 3 \sum_k c_k^2 d_k^2 (\mathrm{E}(Z_k^2))^2.\end{aligned}$$

The variable Z_k is equal to $Y_k - g_0(X_k)$ and Y_k is binomial 0–1 distributed, with probability $p_k = g_0(X_k)$ so the moments of Z_k can be calculated. Then

$$\begin{aligned}&\operatorname{cov}\left(\left(\sum_k Z_k c_k\right)^2, \left(\sum_k Z_k d_k\right)^2\right) \\ &\quad = \sum_k c_k^2 d_k^2 p_k(1 - p_k)((1 - p_k)^3 + p_k^3) + 2\left(\sum_k c_k d_k p_k(1 - p_k)\right)^2 - 3 \sum_k c_k^2 d_k^2 p_k^2(1 - p_k)^2 \\ &\quad = \sum_k c_k^2 d_k^2 p_k(1 - p_k)(6p_k^2 - 6p_k + 1) + 2\left(\sum_k c_k d_k p_k(1 - p_k)\right)^2.\end{aligned} \tag{A.3}$$

Substituting (A.3) back into (A.2) and (A.2) back into (A.1) gives the required expression (3.1) for the variance of T:

$$\begin{aligned}\operatorname{var}(T) &= n^{-2} \sum_i \sum_j \left(\sum_k w_{ik}^2 \sum_k w_{jk}^2\right)^{-1} \\ &\quad \times \left(\sum_k \frac{w_{ik}^2 w_{jk}^2 (6g_0(X_k)^2 - 6g_0(X_k) + 1)}{g_0(X_k)(1 - g_0(X_k))} + 2\left(\sum_k w_{ik} w_{jk}\right)^2\right).\end{aligned}$$

We can make some asymptotic calculations. Note that the variance can be rewritten as

$$\operatorname{var}(T) = \int \int \left[\int nK\left(\frac{x-t}{h_n}\right)^2 dF_n(t) \int nK\left(\frac{y-t}{h_n}\right)^2 dF_n(t) \right]^{-1}$$

$$\times \left[n \int \frac{K[(x-t)/h_n]^2 K[(y-t)/h_n]^2 (6g_0(t)^2 - 6g_0(t) + 1)}{g_0(t)(1-g_0(t))} dF_n(t) \right.$$

$$\left. + 2n^2 \left(\int K\left(\frac{x-t}{h_n}\right) K\left(\frac{y-t}{h_n}\right) dF_n(t) \right)^2 \right] dF_n(y)\, dF_n(x).$$

Replacing $dF_n(t)$ by its limit $f(t)\,dt$, substituting for $y = x + \alpha h_n$, $t = x + zh_n$, and taking some first-order approximations yields

$$\operatorname{var}(T) = n^{-1} \int_x \frac{6g_0(x)^2 - 6g_0(x) + 1}{g_0(x)(1-g_0(x))} f(x)\, dx$$

$$+ 2h_n \int f(x)^2\, dx \int_\alpha \left(\int_z K(z)K(z+\alpha)\, dz \right)^2 d\alpha.$$

Since $nh_n \rightarrow \infty$, the second term dominates. Therefore a first-order approximation of the variance is

$$\operatorname{var}(T) \sim 2h_n \int f(x)^2\, dx \int \left(\int K(z)K(z+\alpha)\, dz \right)^2 d\alpha.$$