Semiparametric Testing of the Link Function in Models For Binary Outcomes

Isabel Proença† and Christian Ritter‡

†Institut de Statistique[1]

and

‡Center for Operations Research and Econometrics

Université Catholique de Louvain

Belgium

December 11, 1994

Abstract

This work stands in the larger context of tests of parametric models against semiparametric alternatives. A semiparametric statistic proposed by Horowitz and Härdle (1994), which we call the H-H statistic, can be used to construct a test for unknown deviations from a hypothesized link function in a parametric single index model which rejects for large values of the statistic. Here we study empirically the finite sample performance of this statistic for the important special case of logistic regression for binary data. We show that its asymptotic distribution is not a good approximation for the finite sample distribution up to sample sizes of several thousand observations. For the chosen examples the value of the test statistic for finite sample sizes tends to be smaller than

[1]voie du Roman Pays 34, 1348 Louvain-la-Neuve, Belgium. We thank W. Härdle, L. Simar, M. Müller and A. Kneip for helpful comments on this paper. A part of the research for this paper was carried out within the Sonderforschungsbereich 373 at Humboldt University Berlin; the paper was printed using funds made available by the Deutsche Forschungsgemeinschaft. Isabel Proença was also funded by J.N.I.C.T under grant 423/90-RM. Christian Ritter was also funded by the European Commission under grant ERBCHBGCT920167.

what would be expected under the asymptotic approximation, thus reducing the power of the test based on asymptotic critical values. Moreover, the finite sample variance of the statistic depends on the bandwidth, which also render asymptotic critical values unsuitable. We propose a modified statistic that reduces considerably both the bias and bandwidth dependency.

Keywords: Kernel smoothing, logistic regression, finite sample properties, simulation.

1 Introduction

This paper focuses on the problem of checking the adequacy of the logit link in binary logistic regression. Horowitz and Härdle (1994) introduced a new tool to perform such specification testing which we call the H-H statistic from now on. Here we study its finite sample properties by simulations and analytic calculations. We find a finite sample bias of the statistic and a bandwidth dependency of its variance. Thus, critical values derived from asymptotic calculations are unsuitable and lead to unsatisfactory empirical power. Based on these results, we propose a modified statistic which is less biased and bandwidth dependent under the null-hypothesis and more powerful for detecting misspecifications of the link function.

Our study has a narrow focus but far-reaching implications. First, the chosen application is important by itself; binary logistic regression is one of the most used statistical methods, especially in the fields of biology, medicine, and economics. Second, there is a large number of similar techniques to which our results extend naturally, such as probit regression and other forms of parametric binary models with fixed link functions. Finally, the test studied here belongs to the large class of moment condition tests (Newey, 1985) whose critical values are calculated from asymptotic approximations. Our results underline the importance of thorough investigations of the finite sample properties of such statistics and of possibly required small sample adjustments.

The paper proceeds as follows. In Section 2, we give some examples in which binary data arise naturally, define the binary logistic regression model, discuss its underlying assumptions, and present ways of testing for misspecification of the link function. In Section 3, we explain how the H-H statistic is constructed, how it can be computed in practice, what its asymptotic distribution is, and how it can be used to detect deviations from a specified link function in a single index model. We also point out, why problems with using asymptotic critical values in finite sample situations can be expected. In Sections 4 and 5, we report on simulation studies of the finite sample properties of the H-H statistic for known and unknown linear predictors, display their results, and interpret them in the context of type-I error probability and power. First, we inspect the case of known linear predictors under the null-hypothesis. We show that for small bandwidths the asymptotic approximation takes hold rapidly, while for large bandwidths (oversmoothing), the variance of the standardized statistic increases quickly with the sample size. Then, we study the statistic for the case of estimated linear predictors under the null hypothesis, and under the alternative of a complementary log-log link. We show that the estimation of the linear predictors severely changes the finite sample distribution of the H-H statistic. In particular, its empirical critical values can become much smaller than the asymptotic critical values, which causes severely reduced type-I error probabilities and power if the asymptotic critical values are used. We also note that the alternative produces a clear positive shift of the test statistic, implying that the power can be restored if a suitable small sample correction can be found. In addition, we find that the variance of the test statistic under estimated linear predictors is less sensitive to the choice of the bandwidth than under known linear predictors. These ideas are used in Section 6, where we derive a modified, but asymptotically equivalent statistic whose finite sample mean and variance are corrected to eliminate the distorting effect of estimating the linear predictors. In Section 7, we repeat the simulations of Section 5 and show that the type-I error probabilities achieved with the modified statistic using asymptotic critical values are nearly correct and bandwidth independent. We show also that the modified statistic has greatly improved power when compared with the original version. In Section 8, we summarize our results and suggest further research.

2 Logistic Regression For Binary Outcomes

2.1 Motivation

In many fields of research, including micro-economics, biology and medicine, binary outcomes arise very naturally and their analysis is conducted in a routinely manner, assisted by commercial software and a large body of experience. Typical examples are: the choice of public or individual transportation to go to work, the decision to buy a given product or not, the affliction of a patient with a certain disease, and the presence or absence of a gene in the offspring of a breeding experiment. Commonly, one wishes to know how the outcome is related to a number of explanatory variables and how it can be predicted from them. For a given individual or experimental unit i, we call the outcome y_i and the explanatory variables $x_{i1}, ..., x_{ip}$; y_i assumes the value one if the specified condition occurs and zero otherwise.

Binary regression, whose aim is to describe $pr(y_i = 1|x_{i1}, ..., x_{ip})$, can be motivated in at least two ways. First, we can interpret the outcome as the combination of two factors. An unobserved random disposition μ_i, and an exposure function $v(x_{i1}, ..., x_{ip})$. If the exposure exceeds the disposition, $v(x_{i1}, ..., x_{ip}) \geq \mu_i$, the outcome is one, otherwise it is zero. The function $v(_{i1}, ..., x_{ip})$ is called the index function. If one assumes that all μ_i come from the same distribution F, then the above probability is equal to $pr(y_i = 1|x_{i1}, ..., x_{ip}) = F(\eta_i)$ with $\eta_i = v(x_{i1}, ..., x_{ip})$. Second, we can imagine the outcomes y_i as the results of Bernoulli experiments with success probabilities π_i which relate to the explanatory variables via the model $\pi_i = F[v(x_{i1}, ..., x_{ip})]$. In this formulation, F is called the link function. Usually, the index function is assumed linear, i.e., $v(x_{i1}, ..., x_{ip}) = \alpha + x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p$, with unknown parameters α, β_1, $\ldots$, β_p to be estimated from the data. If the link function is the logit, i.e., $F(\eta) = [1 + \exp(-\eta)]$ one speaks of logistic regression, if it is the standard normal cumulative distribution function, one speaks of probit regression. Here, we only consider the logistic regression.

The analysis of a data set consists in estimating the parameters of the model, i.e., α, β_1, ... , β_p (by iteratively weighted least squares, for example) and in assessing whether the form of the model is suitable for the data. The appeal of the logistic regression lies in its simple treatment and in the ease of subsequent interpretations. For example, an estimate of 1 for β_1 implies that increases of one unit in the explanatory variable x_1, while keeping the other variables constant, generates an e fold increase of the odds ratio $\pi_i/(1-\pi_i)$.

Safeguards are needed to assure that the three major assumptions governing the logistic regression model are not violated. First, we assume that the outcomes can be regarded as the results of independent Bernoulli trials. For example, if the outcomes occur in a time order or if they are nested within individuals (e.g., vision problems on left and right eye), this may be inappropriate and modifications may be required. Second, we assume that the index function may be linear; third, that the link function is of the logistic form. In practice, all three assumptions must be verified to assure valid interpretations. For example, misspecification of the link function can induce a large increase in the bias and the mean squared error of the estimates as documented in Czado (1989). Here, we focus entirely on tests of the adequacy of the link function.

2.2 Tests of the Adequacy of The Logit Link in Binary Models

There are many ways of constructing tests of the adequacy of the logit link ranging from fully parametric to fully nonparametric approaches. In general, there are two main strategies. First, one can attempt to estimate a model based on a family (parametric or nonparametric) of link functions which includes the logit and test whether the logit is rejected. Second, one can estimate the model under the logit and check *a posteriori* whether this choice was appropriate. The first approach leads to methods such as the technique based on parametric families of link function described by Newton, Czado and Chappell (1993), or the fully nonparametric technique described by Azzalini, Bowman and Härdle (1989) and Härdle and Mammen (1993). Examples of the second approach are the goodness-of-fit test introduced by le Cessie and van Houwelingen (1991), and the H-H test. A special framework for developing tests such as the one sought here is given by Newey (1985). Their idea is to find suitable moment conditions which hold under the null-hypothesis and which are violated under the alternative. In this context, Bierens (1990) deduces a statistic to test a parametric model against a nonparametric alternative. The H-H test also belongs to this class of moment condition tests (Horowitz and Härdle, 1994).

Each approach has its merits, and it depends on the properties of the problem under study, such as sample size and feared alternatives, which one is the most appropriate. In general, more flexibility in the choice of alternatives has to be paid for by a loss of power. This loss of power can be considerable for fully nonparametric treatments, since they suffer from the curse of dimensionality (meaning that asymptotic convergence is slower as the dimensionality of the problem increases, see Härdle (1990)). To achieve acceptable power, nonparametric tests can easily require several thousand observations, especially for binary data. Parametric tests can be powerful even for small data sets, but only against a limited selection of alternatives. Therefore, semiparametric tests, such as the one studied here, can play a very important role in the middleground between small and large sample situations. They can be constructed to avoid the curse of dimensionality, while still permitting a much larger alternative than is available in the parametric framework. In order to play this role they must have known properties for small and medium sized samples. So far, the development of semiparametric tests has been driven by asymptotic results and the practitioner's needs, i.e., controlled small sample behavior, have not been sufficiently taken into account. The aim of this article is to fill this void by appropriately modifying the H-H test and the statistic it is based on.

3 The H-H Statistic

The H-H Statistic (Horowitz and Härdle, 1994) is constructed from a special kernel estimate of the link function given the linear predictors, $\eta_i, i = 1, .., n$. We denote this estimator by $\hat{F}^h$, where h stands for the bandwidth used for the kernel. The test statistic is of the form

$$T^h = h^{\frac{1}{2}} \sum_{i=1}^{n} u(\hat{\eta}_i)[\hat{F}^h(\hat{\eta}_i) - F(\hat{\eta}_i)][y_i - F(\hat{\eta}_i)], \tag{1}$$

where $u(\eta_i)$ is a weight function, which downweights extreme η_i's, and where the $\hat{\eta}_i$ are the estimated linear predictors. Those estimated linear predictors are obtained from the estimates of the model parameters $\beta = (\alpha, \beta_1, ..., \beta_p)$ under the presumed link function.

In the following paragraphs we build the H-H statistic from its elementary components and give the heuristic reasons underlying its construction. We begin with the simplified case in which the null-hypothesis holds and the linear predictors are known, i.e., in which we have the real η_i instead of the estimated ones.

3.1 Known Linear Predictors

Let us denote the simple kernel estimator proposed by Nadaraya (1964) and Watson (1964) (called Nadaraya–Watson estimator in the remainder of this work) by

$$\hat{M}^h(\eta_i) = \frac{\sum_j^n y_j K[(\eta_j - \eta_i)/h]}{\sum_j^n K[(\eta_j - \eta_i)/h]}, \tag{2}$$

where K denotes the kernel function and h the bandwidth. In the following we assume that the kernel is of order 2; generalizations to kernels of order r are straightforward. If this simple estimator is used to obtain $\hat{F}^h$ in formula (1), and if the bandwidth tends to zero, the sum $\sum_{i=1}^n u(\eta_i)[\hat{M}^h(\eta_i) - F(\eta_i)][y_i - F(\eta_i)]$ approaches $\sum_{i=1}^n u(\eta_i)[y_i - F(\eta_i)][y_i - F(\eta_i)]$, which has a finite, positive expectation. In practice, this implies that the mean of this version of the statistic depends strongly on the bandwidth, which is undesirable.

We can arrange that T^h has zero mean regardless of the bandwidth by assuring that the two factors $\hat{F}^h(\eta_i) - F(\eta_i)$ and $y_i - F(\eta_i)$ are independent for all i. This can be accomplished, for example, by using the leave-one-out estimator of the form

$$\hat{G}^h(\eta_i) = \frac{\sum_{j \neq i} y_j K[(\eta_j - \eta_i)/h]}{\sum_{j \neq i} K[(\eta_j - \eta_i)/h]} \tag{3}$$

instead, which is, by construction, independent of y_i. Furthermore, to achieve asymptotic unbiasedness while preserving the optimality of the integrated mean-square error Horowitz and Härdle (1994) use the correction suggested by Bierens (1987) which combines two leave-one-out Nadaraya–Watson estimators

$$\hat{G}^h(\eta_i) = \frac{\sum_{j \neq i} y_j K[(\eta_j - \eta_i)/h]}{\sum_{j \neq i} K[(\eta_j - \eta_i)/h]}$$

and

$$\hat{G}^s(\eta_i) = \frac{\sum\limits_{j \neq i} y_j K[(\eta_j - \eta_i)/s]}{\sum\limits_{j \neq i} K[(\eta_j - \eta_i)/s]}$$

with different bandwidths h and $s = hn^{(1-\delta)/(2r+1)}$. Here δ is a constant between zero and one. In the remainder, we are only working with kernels of order $r = 2$ for which this formula becomes $s = hn^{(1-\delta)/5}$. The combined estimator is given by

$$\hat{F}^h(\eta_i) = \left[\hat{G}_i^h - \left(\frac{h}{s}\right)^2 \hat{G}_i^s\right] / \left[1 - \left(\frac{h}{s}\right)^2\right]. \tag{4}$$

We refer to the leave-one-out rule as the independence correction and to the elimination of the asymptotic bias as the Bierens correction.

Using the properties of independence and asymptotic unbiasedness, Horowitz and Härdle prove that as n tends to infinity and h tends to zero proportionally to $n^{-1/5}$ the statistic T^h approaches a normal distribution with mean zero and variance

$$\sigma_T^2 = 2C_K \int u(\eta)[\sigma^2(\eta)]^2 d\eta, \tag{5}$$

where

$$C_K = \int_{-\infty}^{\infty} K(\eta)^2 d\eta.$$

and

$$\sigma^2(\eta) = V[Y|\eta] = F(\eta)[1 - F(\eta)]. \tag{6}$$

3.2 Unknown Linear Predictors

The independence correction works as long as the linear predictors η_i are known. If the linear predictors are estimated, $\hat{F}^h(\hat{\eta}_i) - F(\hat{\eta}_i)$ and $y_i - F(\hat{\eta}_i)$ become dependent, since the $F(\hat{\eta}_i)$ and $\hat{F}^h(\hat{\eta}_i)$ depend on y_i via the estimation of the model parameters. This produces a non-zero finite sample expectation of the test statistic under the null hypothesis; as our simulations suggest, this bias remains important even for samples of several thousand observations.

Asymptotically, the estimator of the model parameters is $\sqrt{n}$ consistent, and converges therefore faster than the nonparametric estimate of the link function. Thus, for very large n we return to the situation of known linear predictors.

3.3 Violated Null Hypothesis

Under many alternatives, there will be systematic differences between $\hat{F}^h(\eta_i)$ and the assumed $F(\eta_i)$ even if the (wrong) logit link is used for estimating the model. T^h has positive mean when $E[y_i|\eta_i]$ is not $F(\eta_i)$ (Horowitz and Härdle, 1994) and can therefore T^h be used to construct a one-sided test for the adequacy of a specified link function which rejects for large values. For large enough sample sizes this test rejects if $T^h/\sigma_T > z_{1-\alpha}$, where $z_{1-\alpha}$ is the $1-\alpha$ percentile of a standard normal distribution.

As mentioned above, the estimation of the linear predictors leads to a bias in the statistic under the null-hypothesis which carries over to mild alternatives. Thus, for small samples, the asymptotic critical value T^h/σ_T may not produce the desired type-I error probabilities and may affect the power. Under most alternatives, the exact effect cannot be traced analytically and simulations are required to elucidate the situation.

4 Simulations For Known Linear Predictors: The Null-Hypothesis

4.1 Versions of the Statistic

Here we study the small and moderate sample properties of T^h under the null-hypothesis for known linear predictors. We focus on the the effect of changing the bandwidth and on the effect of ignoring the independence and Bierens corrections in the construction of T^h. That is, we study four variations of the test statistic,

$$T^h = h^{\frac{1}{2}} \sum_{i=1}^{n} u(\eta_i)[\hat{M}^h(\eta_i) - F(\eta_i)][y_i - F(\eta_i)], \qquad \text{(V1)}$$

with $\hat{M}^h$ as in (2),

$$T^h = h^{\frac{1}{2}} \sum_{i=1}^{n} u(\eta_i)[\hat{G}^h(\eta_i) - F(\eta_i)][y_i - F(\eta_i)], \qquad \text{(V2)}$$

with $\hat{G}^h$ as in (3),

$$T^h = h^{\frac{1}{2}} \sum_{i=1}^{n} u(\eta_i)[\hat{H}^h(\eta_i) - F(\eta_i)][y_i - F(\eta_i)], \qquad \text{(V3)}$$

where

$$\hat{H}^h(\eta_i) = \left[\hat{M}^h(\eta_i) - \left(\frac{h}{s}\right)^2 \hat{M}^s(\eta_i)\right] / \left[1 - \left(\frac{h}{s}\right)^2\right],$$

and

$$T^h = h^{\frac{1}{2}} \sum_{i=1}^{n} u(\eta_i)[\hat{F}^h(\eta_i) - F(\eta_i)][y_i - F(\eta_i)], \qquad \text{(V4)}$$

with $\hat{F}^h$ as in (4). These versions correspond to an uncorrected Nadaraya–Watson estimator (V1), a leave-one-out Nadaraya–Watson estimator (V2), a Bierens corrected Nadaraya–Watson estimator (V3), and a Bierens corrected leave-one-out Nadaraya–Watson estimator (the full H-H statistic), respectively.

4.2 Design of Simulations

For the parametric model we assume that $\eta_i = \beta_0 + \beta_1 x_i$ with known $\beta_0 = 0$ and $\beta_1 = 1$, i.e., with known linear predictors. From this model we simulate data sets of various sample sizes and calculate each time the standardized statistic T^h/σ_T. That is, for given sample size n and smoothing scheme (V1, V2, V3, or V4), we

1. generate n random linear predictors x_i from a standard normal distribution,

2. calculate $\eta_i = x_i$,

3. generate y_i from Bernoulli distributions with success probabilities

$$P(y_i = y_i|\eta_i) = 1/[1 + \exp(-\eta_i)],$$

4. for each h in $0.1, 0.2, \ldots, 1.0$ we

   - calculate T^h and its asymptotic variance, σ_T^2,
   - report the studendized statistic T^h/σ_T.

For our calculation of T^h we use a weight function $u(\eta)$ equal to one for all η_i except for 5 percent at either extreme, where it is zero.

Remark: Using a standard normal distribution to generate the linear predictors implies that most of them are close to zero and only a few trace the outer shape of the link function. This "design" is not optimal for detecting deviations from the logit link function and low power can be expected. It is, however, realistic, since in many observational studies, one finds that most linear predictors are close to zero and that their spread is limited to a few units.

4.3 Results

For the full statistic given by (V4), the lower right plots in Figures 1, 2, and 3 show the 5th, 25th, 50th, 75th, and 95th percentiles of the standardized statistic for the bandwidths $h = 0.1, 0.2, \ldots, 1.0$ and the sample sizes 50, 500, and 1000, which were chosen to extend beyond the sample sizes used in an earlier study by Proença (1993).

The picture is most stable for 50 observations. There the medians are close to zero, the variances close to one, and the quantiles reveal that the distribution of T^h/σ_T is slightly skewed to the left. For 500 and 1000 observations, one notices a fanning out of the variance for bandwidth larger than 0.4. This is an effect of oversmoothing. The upper right plot in Figure 4 shows an example for 1000 data points and $h = 0.5$. This bandwidth is already large enough to make the kernel estimate flatter than the logit curve. While 50 observations may not be sufficient to reveal this feature, it becomes more prominent for the sample sizes 500 and 1000. For small bandwidths ($h < 0.4$), the simulations with 500 and 1000 observations result in percentiles which approach quickly the asymptotic values suggesting rapid convergence in distribution.

Dropping Bierens' correction, i.e., using variation (V3), means using a single leave-one-out kernel estimator. The result of the simulation are shown in the the upper right plots in Figures 1, 2, and 3. It is hard to say whether dropping Bierens' correction makes any difference. There is weak indication that without the correction and for small h the variance is reduced. For larger h and sample sizes there is a vague suggestion of a negative bias.

Dropping the independence correction, i.e., using variations (V2) or (V1), has a much stronger effect than dropping Bierens' correction. This is shown in the plots on the left sides of Figures 1, 2, and T^h is now positively biased and this bias increases with decreasing bandwidth.

The top plot of Figure 8 shows the rejection rates for the full statistic (V4) using a critical value of 1.645 corresponding to an asymptotic type-I error probability of 5%. The empirical size tends to be larger than the nominal size of 5%. The empirical size increases for large bandwidths and this effect becomes worse with increasing sample size. For sample size 50 the most important deviations from the nominal size occur at small bandwidths.

Our study suggests that the independence correction is essential, while the Bierens correction, although required for theoretical reasons, has little practical importance for small to medium sample sizes. Convergence to the asymptotic distribution is rapid for small bandwidth; large bandwidth cause problems.

5 Simulation For Estimated Linear Predictors: The Null Hypothesis And The Complementary Log-Log Link as Alternative

In this section, we make a small change to the implementation of the simulation to reflect what we do in practice when we apply the test. We now estimate β_0 and β_1 under the null-hypothesis, the logit link, calculate $\hat{\eta}_i$ from $\hat{\beta}_0 + \hat{\beta}_1 x_i$, and construct the test using the $\hat{\eta}_i$. We simulate under two scenarios.

$$pr(y_i|X_i = x_i) = \frac{1}{[1 + \exp(-\eta_i)]}, \qquad \text{(S1)}$$

$$pr(y_i|X_i = x_i) = 1 - \exp(-\exp(\eta_i)). \qquad \text{(S2)}$$

The first scenario corresponds to the null-hypothesis, the second one to a complementary log-log alternative.

A comparison of Figures 5, 6, and 7 with Figures 1, 2, and 3 shows that the estimation of the linear predictors distorts the distribution of the test statistic under the null-hypothesis. This distortion consists in an increased skewness (to the left) and a shrinking of the variance for large bandwidths. As a consequence, the empirical type-I error is much smaller than the asymptotic critical value suggests. The alternative (S2) produces a positive shift in the distribution. Yet, this shift hardly overcomes the beforementioned distortion and the rejection rate hardly exceeds 20 percent even for a samples of 1000 observations.

The middle and bottom plots of Figure 8 shows the rejection rates for the full statistic using a critical value of 1.645 corresponding to an asymptotic type-I error probability of 5%. The middle plot is for the null-hypothesis, the bottom plot for the alternative of data originating from a complementary log-log model. This can be compared with the top plot which shows the rejection rates for the null-hypothesis for known linear predictors. One sees that the empirical type-I error probabilities for estimated linear predictors are much smaller than 5% and that consequently the power is very low.

In summary, we find that the asymptotic approximation is not satisfactory when the linear predictors have to be estimated and when the sample size does not exceed a few thousand observations (some additional explorations were conducted for selected bandwidths and 2000 and 4000 observations and only little improvement towards the asymptotic distribution was noted). Corrections are therefore necessary.

6 Correcting for Bias and Bandwidth Dependency Under the Null-Hypothesis

In this section, we propose a modified statistic to test semiparametrically the specification of the link in single index models. We construct a simplification of the H-H statistic in which the above mentioned bias and bandwidth dependency can be calculated directly and use these results to devise appropriate modifications.

We proceed in two stages. At first we return to the case where the linear predictors are known and present approximations of the small sample expectation and variance of the modified statistic. Then, we show what happens when the linear predictors are estimated. We use the results to design adjustments which extend the validity of asymptotic critical values to samples as small as 50 observations.

6.1 The modified statistic

First, suppose that the linear predictors are known. Define the residuals $r(\eta_i)$ as the differences between the observations y_i and the assumed model $F(\eta_i)$. Then, the H-H statistic can be written as

$$T^h = h^{\frac{1}{2}} \sum_{i=1}^{n} u(\eta_i)[\hat{F}^h(\eta_i) - F(\eta_i)]r(\eta_i). \tag{7}$$

If the bandwidth is large, the smoothed version of the link, $\hat{F}^h(\eta_i)$, deviates considerably from $F(\eta_i)$ and makes the expectation of $\hat{F}^h(\eta_i) - F(\eta_i)$ nonzero. This leads to a distortion of the distribution of T^h which, for fixed bandwidth, becomes more pronounced as the sample sizes increases (see Figures 5, 6 and 7).

To avoid this distortion, we propose to replace the difference $\hat{F}^h(\eta_i) - F(\eta_i)$ by the smoothed residuals

$$\tilde{r}(\eta_i) = \frac{\sum_{j \neq i} r(\eta_j) K[(\eta_j - \eta_i)/h]}{\sum_{j \neq i} K[(\eta_j - \eta_i)/h]}$$

instead. This corresponds to smoothing the outcomes y_i and the values $F(\hat{\eta}_i)$ simultaneously. This provides a first modification of H-H statistic given by

$$MT^h = \sqrt{h} \sum_{i=1}^{n} u(\eta_i)\tilde{r}(\eta_i)r(\eta_i),$$

with

$$r(\eta_i) = y_i - F(\eta_i).$$

Here $u(\eta_i)$ is again a weight function that downweights extreme observations. The leave-one-out estimate in the calculation of the smoothed residuals (independence correction) is used in order to achieve independence between $r(\eta_i)$ and $\tilde{r}(\eta_i)$.

Define $\tilde{\mathbf{r}}$ as the vector containing the smoothed residuals $\tilde{r}_1, \cdots, \tilde{r}_n$ with $\tilde{r}_i = \tilde{r}(\eta_i)$. Note that $\tilde{\mathbf{r}}$ can be written as $\tilde{\mathbf{r}} = W\ \mathbf{r}$, the product between the matrix that contains the smoothing weights (usually called the smoothing matrix) and the vector of residuals $\mathbf{r} = (r_1, \cdots, r_n)'$, where $r_i = r(\eta_i)$. For the leave-one-out Nadaraya–Watson estimator the, smoothing matrix has elements w_{ij} defined as

$$w_{ij} = \begin{cases} \frac{K[(\eta_j - \eta_i)/h]}{\sum_{j \neq i} K[(\eta_j - \eta_i)/h]} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \tag{8}$$

The modified H-H statistic can therefore be written as

$$MT^h = \sqrt{h}\, \mathbf{r}^T U W \ \mathbf{r},$$

where the matrix U is a diagonal matrix with elements $u_i = u(\eta_i)$. The matrix UW is not symmetric. In the following derivations, it is more convenient to work with a symmetrized version given by $W^\star = (1/2)(UW + W^T U)$ which enables us to write MT^h as the quadratic form

$$MT^h = \sqrt{h}\, \mathbf{r}^T W^\star \ \mathbf{r}.$$

Since the $r(\eta_i)$ are mutually independent and have zero expectation conditional on $\mathbf{x}_i$ under the null, and since the elements in the diagonal of $W^\star$ are zero the expectation of MT^h conditional on $\mathbf{x}_i$ is zero under the null. Straightforward calculation shows that its variance is

$$\begin{aligned}\sigma^2_{MT} &= 2h \sum_i \sum_{j \neq i} \sigma^2(\eta_i)\sigma^2(\eta_j) {w^\star_{ij}}^2 \\ \sigma^2_{MT} &= h \sum_i \sum_{j \neq i} \sigma^2(\eta_i)\sigma^2(\eta_j)(u(\eta_i) w^2_{ij} + u(\eta_i) w_{ij} u(\eta_j) w_{ji}).\end{aligned}$$

with $\sigma^2(\eta) = F(\eta)[1 - F(\eta)]$. Moreover, the modified version has the same asymptotic distribution under the null as the H-H statistic as it is stated in the next theorem.

Theorem 1 *Under H_0 and under the assumptions in Horowitz and Härdle (1994) the modified H-H statistic is asymptotically distributed as $N(0, \sigma^2_T)$ with σ^2_T as in equation (5).*

The proof of this theorem follows Horowitz and Härdle (1994) and is given in the appendix.

As an aside, note that $r(\eta_i)$ has expectation zero conditional on $\mathbf{x}_i$ with $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$. Therefore, under the null hypothesis, the smoothing operation on the residuals is trying to find a function which is identically equal to zero. This means that the Nadaraya-Watson estimator does not produce a smoothing bias and consequently Bierens' correction is no longer needed. Moreover, the conditions forcing the bandwidth to converge to zero at the rate $n^{-1/5}$ can be dropped. It was introduced in order to trade off bias and variance in the kernel estimation of F. We think that tractable asymptotic approximations can in fact be obtained for any fixed h. If this can be established, the test can act like a parametric test if the bandwidth is kept large, and like a semiparametric test if the bandwidth tends to zero with increasing sample size.

6.2 Effect of Estimating the Linear Predictors

In practice, the true $r(\eta_i)$ have to be substituted by the fitted residuals,

$$r(\hat{\eta}_i) = y_i - F(\hat{\eta}_i).$$

with $\hat{\eta}_i = \mathbf{x}_i\hat{\beta}$. To analyze the effect of substituting the true residuals by their fitted counterparts we use the approximation employed in le Cessie and van Houwelingen (1991). Following these authors, the vector of fitted residuals $\hat{\mathbf{r}} = (\hat{r}_1 \cdots \hat{r}_n)'$, where $\hat{r}_i = r(\hat{\eta}_i)$, is related to the vector of true residuals by

$$\hat{\mathbf{r}} = (I - H)\ \mathbf{r} + o_p(n^{-1/2})$$

where H is the matrix

$$H = VX(X^TVX)^{-1}X^T.$$

Here X is the design matrix with rows $\mathbf{x}_i$, $i = 1, ..., n$ and V is a diagonal matrix with the weights $v_i = \sigma^2(\eta_i)$. Thus, we can approximate MT^h by

$$AMT^h = \sqrt{h}\, \mathbf{r}^T(I - H)^T W^*(I - H)\ \mathbf{r} = \sqrt{h}\, \mathbf{r}^T D\ \mathbf{r}.$$

with $D = (I - H)^T W^*(I - H)$. The expectation of AMT^h conditional on $\mathbf{x}_i$ is given by

$$\begin{aligned} E[AMT^h] &= \sqrt{h}\ \sum_i d_{ii}\sigma^2(\eta_i) \\ &= -tr(V^{\frac{1}{2}}W^*V^{\frac{1}{2}}H^*), \end{aligned}$$

where $H^* = V^{\frac{1}{2}}X(X'VX)^{-1}X'V^{\frac{1}{2}}$. In general, this expectation is nonzero and we conjecture that it is negative in most cases. Note that W^* is symmetric, but not always positive definite. Nevertheless, the expression $tr(V^{\frac{1}{2}}W^*V^{\frac{1}{2}}H^*)$ can be rewritten as $tr[(X'VW^*VX)(X'VX)^{-1}]$ and one sees that it suffices that the trace of matrix $X^{*\prime}W^*X^*$ with the "standardized" design $X^* = VX(X'VX)^{-\frac{1}{2}}$ is positive to make the expectation negative. This was the case in all situations we studied.

The variance of AMT^h conditional on $\mathbf{x}_i$ is equal to

$$\begin{aligned}
\sigma^2_{AMT} &= h\sum_i d^2_{ii}\{E[r(\eta_i)^4] - [\sigma^2(\eta_i)]^2\} + 2\sum_i\sum_{j\neq i}\sigma^2(\eta_i)\sigma^2(\eta_j)d^2_{ij} \\
&= h\sum_i d^2_{ii}\{\sigma^2(\eta_i) - 4[\sigma^2(\eta_i)]^2\} + 2\sum_i\sum_{j\neq i}\sigma^2(\eta_i)\sigma^2(\eta_j)d^2_{ij} \\
&= 2tr(DVDV) + \sum_{i=1}^{n} d^2_{ii}[\sigma^2(\eta_i) - 6\sigma^4(\eta_i)].
\end{aligned}$$

The conditional expectation and variance of AMT^h can be taken as approximations of the conditional expectation and variance of the estimated MT^h. They can be easily estimated under the null hypothesis because the only unknowns are the conditional variances of the residuals $\sigma^2(\eta_i)$. The expressions for $E[AMT^h]$ and σ^2_{AMT} can be used to adjust and studentize the statistic by

$$\frac{MT^h - E[AMT^h]}{\sigma_{AMT}} \tag{9}$$

6.3 Alternative Method for Bias Correction

Obviously, the bias of the statistic stemming from estimating the linear predictors can be corrected for by adjusting by the quantity $tr(W^*H)$. In practice, this adjustment can be computed with reasonable efficiency using smoothing methods. Alternatively, a second approach can be used which promises to have a larger range of applicability and which will be discussed in the following.

The bias in the estimated MT^h is caused by the dependence among the fitted residuals introduced by the estimation of the parameter vector β. A first order Taylor expansion of $r(\hat{\eta}_i) = y_i - F(\hat{\eta}_i)$ about its value in $r(\eta_i) = y_i - F(\eta_i)$, the residual for the assumed model, (with $\hat{\eta}_i = \mathbf{x}_i^T\hat{\beta}$) and $\eta_i = \mathbf{x}_i^T\beta_0$), gives

$$r(\hat{\eta}_i) = r(\eta_i) - (\hat{\beta} - \beta_0)\mathbf{x}_i\ \sigma^2(\eta_i) + o_p(n^{-1/2})$$

On the other hand, $\hat{\beta}$ is linear in $\mathbf{y}$ and can be written as $\hat{\beta} = C + AX^T\mathbf{y}$. After combining both expressions and simplifying the results, we have for $i < j$

$$E[r(\hat{\eta}_i)r(\hat{\eta}_j)] = -2\sigma^2(\eta_i)\sigma^2(\eta_j)\mathbf{x}_i^T A\mathbf{x}_j + \sigma^2(\eta_i)\sigma^2(\eta_j)\mathbf{x}_i^T AX^T VXA\mathbf{x}_j + o_p(1).$$

Suppose now that the i-th residual is obtained from the fit without the i–th observation. That is, $r_{-i}(\hat{\eta}_i)$ is calculated using the fitted coefficients $\hat{\beta}_{-i} = C + AX_{-i}^T\mathbf{y}_{-i}$ where X_{-i} and $\mathbf{y}_{-i}$ denote the X matrix and the $\mathbf{y}$ vector without row i respectively. This yields

$$E[r_{-i}(\hat{\eta}_i)r_{-j}(\hat{\eta}_j)] = \sigma^2(\eta_i)\sigma^2(\eta_j)\mathbf{x}_i^T AX_{-i}^T V_0 X_{-j} A\mathbf{x}_j + o_p(1).$$

with V_0 a diagonal matrix of dimension $n-1$ whose elements are

$$v_{0;k} = \begin{cases} v_k & \text{for } 1 \leq k < i \\ 0 & \text{for } i \leq k < j \\ v_{k+1} & \text{for } j \leq k \leq n-1 \end{cases} .$$

This correction reduces the covariance between the residuals since the first term in the preceding expression has vanished and the second is of smaller size. Our practical studies suggest that this method is as effective as the direct adjustment.

In the following section, we show how the modified statistic performs in practice for the scenarios defined in Section 5.

7 Performance of the Modified Statistic

In this section, we report on simulation results obtained using the modified statistic for the two scenarios (S1) and (S2) discussed in Section 5. However, since computing the bias and the adjusted variance 500 and 1000 observation is still very time demanding, a shortcut was used. That is, the bias and the adjusted variance were only computed for a few replicates of the data sets and their averages were used to adjust the statistic for the simulations.

Figure 9 shows the 5th, 25th, 50th, 75th, and 95th percentiles of the standardized statistic plotted versus the bandwidth for both scenarios and for the three sample sizes 50, 500 and 1000 observation. Also the corresponding percentiles of a standard normal distribution are shown. One finds that under the null hypothesis the finite sample percentiles of the standardized statistic are rather close to the percentiles of the asymptotic approximation. In particular, the empirical upper critical values of the statistic are very close to the corresponding asymptotic values. This means that the type-I error probability of the test based on the modified statistic is nearly independent of the bandwidth used for smoothing.

Figure 10 shows the picture for the complementary log-log alternative. The shift is clearly visible and becomes stronger with increasing sample size. Figure 11 shows the corresponding rejection rates under null-hypothesis and under the alternative. A comparison with Figure 8 shows a clear improvement. The power obtained with the modified statistic for 1000 observations is about three times higher than before and attains 60 to 70 percent for large bandwidth. It is quite natural that the power depends on the bandwidth even though the type-I error probability is nearly constant, for exactly how the alternative distinguishes itself from the null-hypothesis determines at which bandwidth the expectation of the test statistic will be the most positive. Therefore, an examination of a trace of the test statistic versus the bandwidth which indicates "rejection" in some part and "acceptance" in another is by no means contradictory. On the contrary, it provides information about the way the alternative differs from the null-hypothesis.

Note that the distribution of MT^h/σ^2_{AMT} is skewed for larger sample sizes and bandwidths. One reason for this may be that averages of the bias and of the adjusted variance were used to calculate the adjusted statistic instead of individual ones.

8 Summary, Discussion, Further Research

8.1 Summary

Because of its simplicity and its ease of handling, the H-H statistic is an appealing tool for testing deviations from a specified link. Yet, in its original form, it can have low power if asymptotic critical values are used for samples of less than several thousand observations. Moreover, the choice of the bandwidth used for smoothing affects both the type-I error and the power of the test. For example, the larger the bandwidth the larger the variance of the test statistic reflecting oversmoothing (clearly visible for sample size 1000). The modifications introduced in Section 6 permit restoring the type-I error to nearly its asymptotic value and making it nearly independent to the choice of the bandwidth. Moreover, they improve the power of the test (at least in the situations presented here). The power cannot be made independent of the bandwidth. In slight violation of the common paradigm of significant tests, this can be considered as a feature instead of a shortcoming. One can argue that it is best to calculate the statistic over a large range of bandwidth and to check for which bandwidth the test rejects (if any) and for which it does not. The trace of the statistic with respect to the bandwidth is characteristic for the alternative. For example, wiggly alternatives produce rejections at small bandwidths, larger bumps at medium bandwidth, and overall distortions, such as the complementary log-log, at large bandwidths.

8.2 Discussion

This work shows that although asymptotic derivations play an essential role in the construction of the statistic presented here, thorough small sample studies are necessary to certify its usefulness in practice. Here, the study of the behavior for small and medium sized samples led to a a statistic which is asymptotically equivalent to the original version, but which has better properties in its practice. The achieved decoupling of the modified statistic from the choice of bandwidth under the null-hypothesis allows using it not only as mere testing tool but also as an instrument for identifying alternatives. We therefore recommend to always calculate and graph the trace of the standardized test statistic over a large range of bandwidths (from 0.1 to 1.0, say) as a diagnostic plot.

Unfortunately, even with the modified statistic the power observed in our study remains little satisfactory ($60-70\%$ against the complementary log-log alternative for 1000 observations), and this suggest that without further improvement, the modified test is only sensitive for sample sizes of several thousand observations.

8.3 Further Research

There are several directions in which further research can be beneficial. First, further modifications of the statistic should be aimed at improving the power versus a large collection of alternatives without destroying the nearly bandwidth-independent type-

I error probability. This can be approached by employing better smoothing methods, or by attempting to estimate the index function under a nonparametric link function. As a first step, the smoothed estimate of the link function could be used to re-estimate the parameters and this operation could be repeated several times. As alternative a semiparametric estimator as presented in Klein and Spady (1993) or Bonneu, Delecroix and Malin (1993) can also be used. The results presented in this paper should allow to determine easily if such a modification produces an appreciable improvement. Further research is also needed on the interpretation of the traces of the test statistic as a function of the bandwidth, when rejection is observed in some areas but not in others. A major practical problem which still needs to be studied is the efficient calculation of the variance of the statistic for medium sample sizes. Currently, this calculation uses matrices of dimension $n \times n$, which become unpleasant for more than a few hundred observations. It appears, however, that for fixed bandwidth the variance converges rather rapidly to an asymptotic variance. This asymptotic variance, once derived, can provide the necessary gain in computational efficiency needed for samples of several thousand observations.

In summary, a statistic has been constructed which has a well-controlled behavior for small and medium sized samples. However, much work remains before the statistic or a variation of it will have reached the maturity to make it a reliable and user-friendly tool in the verification of link functions.

A Proof of Theorem 1

The proof of theorem 1 follows the proof in Horowitz and Härdle (1994) closely, although fewer approximations are needed for the modified statistic than for the original one. The assumptions, mainly conditions on boundedness, and smoothness can be found in Horowitz and Härdle (1994) pp. 25–26. We begin with a lemma that shows that we can substitute the statistic by an approximation with a deterministic denominator. The asymptotic distribution is then derived from a central limit theorem for degenerate U-statistics of Hall (1984). In addition to the notation introduced in Section 6, we use $\bar{\tilde{r}}(\eta_i) = \tilde{r}(\eta_i)\hat{p}^h(\eta_i)/p(\eta_i)$ with $p(\eta_i)$, the "true" density of the linear predictors, and its estimate

$$\hat{p}^h(\eta_i) = \frac{1}{nh}\sum_{j\neq i} K[(\eta_j - \eta_i)/h].$$

We remind the reader that η_i equals $\mathbf{x}_i^T\beta_0$.

Lemma 1 *Define* $MT_1^h = \sqrt{h}\sum_i u(\eta_i)r(\eta_i)\bar{\tilde{r}}(\eta_i)$. *Under the assumptions of Horowitz and Härdle (1994)*

$$MT^h = MT_1^h + o_p(1)$$

Proof:

Notice that $MT^h = MT_1^h + R_1$ with $R_1 = \sqrt{h}\sum_i u(\eta_i)[\tilde{r}(\eta_i) - \bar{\tilde{r}}(\eta_i)]$. By lemma 1 of Horowitz and Härdle (1994) $|p(\eta) - \hat{p}^h(\eta)| = O((\log n)^{1/2}/(nh)^{1/2})$ almost surely and uniformly over η. That allows us to write

$$\tilde{r}(\eta_i) - \bar{\tilde{r}}(\eta_i) = \frac{\bar{\tilde{r}}(\eta_i)[p(\eta_i) - \hat{p}^h(\eta_i)]}{p(\eta_i)} + O((\log n)^{1/2}/(nh)^{1/2})$$

almost surely and $R_1 = \sqrt{h}\sum_i u(\eta_i)r(\eta_i)\bar{\tilde{r}}(\eta_i)[p(\eta_i) - \hat{p}^h(\eta_i)]/p(\eta_i) + o_p(1)$. $E(R_1) = 0$ because $\bar{\tilde{r}}(\eta_i)$ does not depend on $r(\eta_i)$. Next we shall prove that $V(R_1) = o(1)$ where $V(\bullet)$ stands for variance.

$$\begin{aligned} V(R_1) &= hE\left\{\sum_i u^2(\eta_i)r^2(\eta_i)\bar{\tilde{r}}(\eta_i)^2[p(\eta_i) - p^h(\eta_i)]^2/p(\eta_i)^2\right. \\ &+ \left.\sum_i\sum_{j\neq i} u(\eta_i)u(\eta_j)r(\eta_i)r(\eta_j)\bar{\tilde{r}}(\eta_i)\bar{\tilde{r}}(\eta_j)\frac{[p(\eta_i) - \hat{p}^h(\eta_i)]}{p(\eta_i)}\frac{[p(\eta_j) - \hat{p}^h(\eta_j)]}{p(\eta_j)}\right\} \end{aligned}$$

and

$$
\begin{aligned}
& E\{u(\eta_i)r(\eta_i)\bar{\tilde{r}}(\eta_j)[p(\eta_j) - \hat{p}^h(\eta_j)]/p(\eta_i)\} \\
& \quad = u(\eta_i)\sigma^2(\eta_i)E\{K[(\eta_i - \eta_j)/h](1/nh)[1/p(\eta_j)^2][p(\eta_j) - \hat{p}^h(\eta_j)]\} \\
& \quad = o(n^{-1})
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
V(R_1) &= hE\left\{u(\eta_i)^2 r(\eta_i)^2 \bar{\tilde{r}}(\eta_i)^2 [p(\eta_i) - \hat{p}^h(\eta_i)]^2/p(\eta_i)^2\right\} \\
&\leq [h/p(\eta_i)^2]E\left\{\sum_i u(\eta_i)^2 r(\eta_i)^2 \bar{\tilde{r}}(\eta_i)^2 O((\log n)/(nh))\right\}
\end{aligned}
$$

almost surely. Following Bierens (1987) (pp 104) $E[\bar{\tilde{r}}(\eta_i)^2] = O(1/(nh))$ uniformly over η (and given that by assumption $u(\bullet)$ is bounded) consequently $V(R_1) = o(1)$. From Chebyshev's inequality $R_1 = o_p(1)$. Q.E.D.

Proof of theorem 1:

Define $Z_i = (\eta_i, r(\eta_i))$ and

$$
\begin{aligned}
A_n(Z_i, Z_j) &= \frac{1}{\sqrt{hn}} u(\eta_i)r(\eta_i)r(\eta_j)\frac{K[(\eta_j - \eta_i)/h]}{p(\eta_i)} \\
H_n(Z_i, Z_j) &= A_n(Z_i, Z_j) + A_n(Z_j, Z_i) \\
&= \frac{1}{\sqrt{hn}} r(\eta_i)r(\eta_j)\left[\frac{u(\eta_i)}{p(\eta_i)} + \frac{u(\eta_j)}{p(\eta_j)}\right] K[(\eta_i - \eta_j)/h] \\
\mu(Z_i) &= E[H_n(Z_i, Z_j)|Z_i] = 0 \\
Q_n(Z_i, Z_j) &= E[H_n(Z_l, Z_i)H_n(Z_l, Z_j)|Z_i, Z_j]
\end{aligned}
$$

Given Lemma 1 and after some algebra the modified H-H statistic can be written

$$
MT^h = \sum_i \sum_{j<i} H_n(Z_i, Z_j) + o_p(1)
$$

The first term in the formula above is a U-statistic and given that $\mu(Z_i) = 0$ this is a degenerate U-statistic. According to Theorem 1 of Hall (1984) this U-statistic will converge in distribution to a $N(0, \sigma_T^2)$ if

$$\{E[Q_n(Z_i, Z_j)^2] - n^{-1}E[H_n(Z_i, Z_j)^4]\}/\{E[H_n(Z_i, Z_j)^2]\}^2 \to 0, \tag{10}$$

$$(1/2)n^2 E[H_n(Z_i, Z_j)^2] \to \sigma_T^2 \tag{11}$$

as $n \to \infty$

$$\begin{aligned} Q_n(Z_i, Z_j) &= \frac{1}{n^2 h} E_{l|i,j}\left\{ r(\eta_l) r(\eta_i) \left[\frac{u(\eta_l)}{p(\eta_l)} + \frac{u(\eta_i}{p(\eta_i)} \right] \right. \\ &\times \left. K[(\eta_i - \eta_l)/h] K[(\eta_j - \eta_l)/h] r(\eta_l) r(\eta_i) \left[\frac{u(\eta_l)}{p(\eta_l)} + \frac{u(\eta_i}{p(\eta_i)} \right] \right\} \\ &= \frac{1}{n^2 h} \int r(\eta_i) r(\eta_j) R(\eta_l, \eta_i, \eta_j) K[(\eta_i - \eta_l)/h] K[(\eta_j - \eta_l)/h] p(\eta_l) d\eta_l \end{aligned}$$

and $R(\bullet)$ is a bounded continuous function. Making a change of variable $\xi = (\eta_l - \eta_i)/h$ gives,

$$Q_n(Z_i, Z_j) = \int r(h\xi + \eta_i, \eta_i, \eta_j) K(\xi) K[\xi + (\eta_i - \eta_j)/h] p(\eta_i + h\xi) d\xi$$

and

$$\begin{aligned} E[Q_n(Z_i, Z_j)^2] &= cn^{-4} \int R_n(h\xi_1 + \eta_i, h\xi_2 + \eta_j, \eta_i, \eta_j) K(\xi_1) K(\xi_2) \\ &\times K[\xi_1 + (\eta_i - \eta_j)/h] K[\xi_2 + (\eta_i - \eta_j)/h] p(h\xi_1 + \eta_i) \\ &\times p(\xi_2 + \eta_j) p(\eta_i) p(\eta_j) d\xi_1 d\xi_2 d\eta_i d\eta_j \\ &= O(n^{-4}) O(h) = o(1) \end{aligned} \tag{12}$$

Regarding $E[H_n(\bullet)^4]$ note that,

$$H_n(Z_i, Z_j)^4 = n^{-4} h^{-2} \bar{R}_n(Z_i, Z_j)$$

and $\bar{R}_n(Z_i, Z_j)$ is bounded uniformly over n. Therefore

$$n^{-1} E[H_n(Z_i, Z_j)^4] = O(1/(n^5 h^2)) \tag{13}$$

Considering $E[H_n(\bullet)^2]$ we have

$$\begin{aligned}
H_n(Z_i, Z_j)^2 &= A_n(Z_i, Z_j)^2 + A_n(Z_j, Z_i)^2 + 2A_n(Z_i, Z_j)A_n(Z_j, Z_i) \\
A_n(Z_i, Z_j)^2 &= \frac{1}{n^2 h p(\eta_i)^2} u(\eta_i)^2 r(\eta_i)^2 r(\eta_j)^2 K[(\eta_j - \eta_i)/h]^2 \\
n^2 E[A_n(Z_i, Z_j)^2] &= \int \frac{u(\eta_i)^2}{p(\eta_i)^2} \sigma^2(\eta_i)\sigma^2(h\xi + \eta_i)K(\xi)^2 p(\eta_i)p(h\xi + \eta_i)d\xi d\eta_i \\
&= C_K \int u(\eta)^2[\sigma^2(\eta)]^2 d\eta + o(1) \qquad (14)
\end{aligned}$$

By arguments similar to those used to obtain (14)

$$n^2 E[A_n(Z_i, Z_j)A_n(Z_j, Z_i)] = C_K \int u(\eta)^2[\sigma^2(\eta)]^2 d\eta + o(1)$$

Therefore,

$$n^2 E[H_n(Z_i, Z_j)^2] = 4C_K \int u(\eta)^2[\sigma^2(\eta)]^2 d\eta + o(1) \qquad (15)$$

Given results (12), (13) and (15) conditions (10) and (11) are verified. Q.E.D.

References

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

Bierens, H. J. (1987). Kernel estimators of regression functions, *in* T. Bewley (ed.), *Advances in Econometrics: Fifth World Congress*, Cambridge University Press, New York, pp. 99–144.

Bierens, H. J. (1990). A consistent conditional moment test of functional form, *Econometrica* **58**: 1443–1458.

Bonneu, M., Delecroix, M. and Malin, E. (1993). Semiparametric versus nonparametric estimation in single index regression model: A computational approach, *Computational Statistics* (8): 207–222.

Czado, C. (1989). *Link misspecification and data selected transformations in binary regression models*, PhD thesis, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, N.Y.

Hall, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators, *Journal of multivariate analysis* **14**(1): 1–16.

Härdle, W. (1990). *Applied Nonparametric Regression*, Econometric Society Monographs, 19, Cambridge University Press.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric *versus* parametric regression fits, *Annals of Statistics* **21**(4): 1926–1947.

Horowitz, J. and Härdle, W. (1994). Testing a parametric model against a semiparametric alternative, *Econometric theory*.

Klein, R. and Spady, R. (1993). An efficient semiparametric estimator for binary response models, *Econometrica* **61**(2): 387–421.

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.

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

Newey, W. (1985). Maximum likelihood specification testing and conditional moment test, *Econometrica* **53**: 1047–1070.

Newton, M., Czado, C. and Chappell, R. (1993). Semiparametric bayesian inference for binary regression, *Technical report*, Department of Statistics, University of Wisconsin–Madison, 1210 W Dayton Street, Madison, WI 53706.

Proença, I. (1993). On the performance of the H-H test, *Technical Report 9310*, Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin, Spandauerstr. 1, 10178 Berlin, Germany.

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

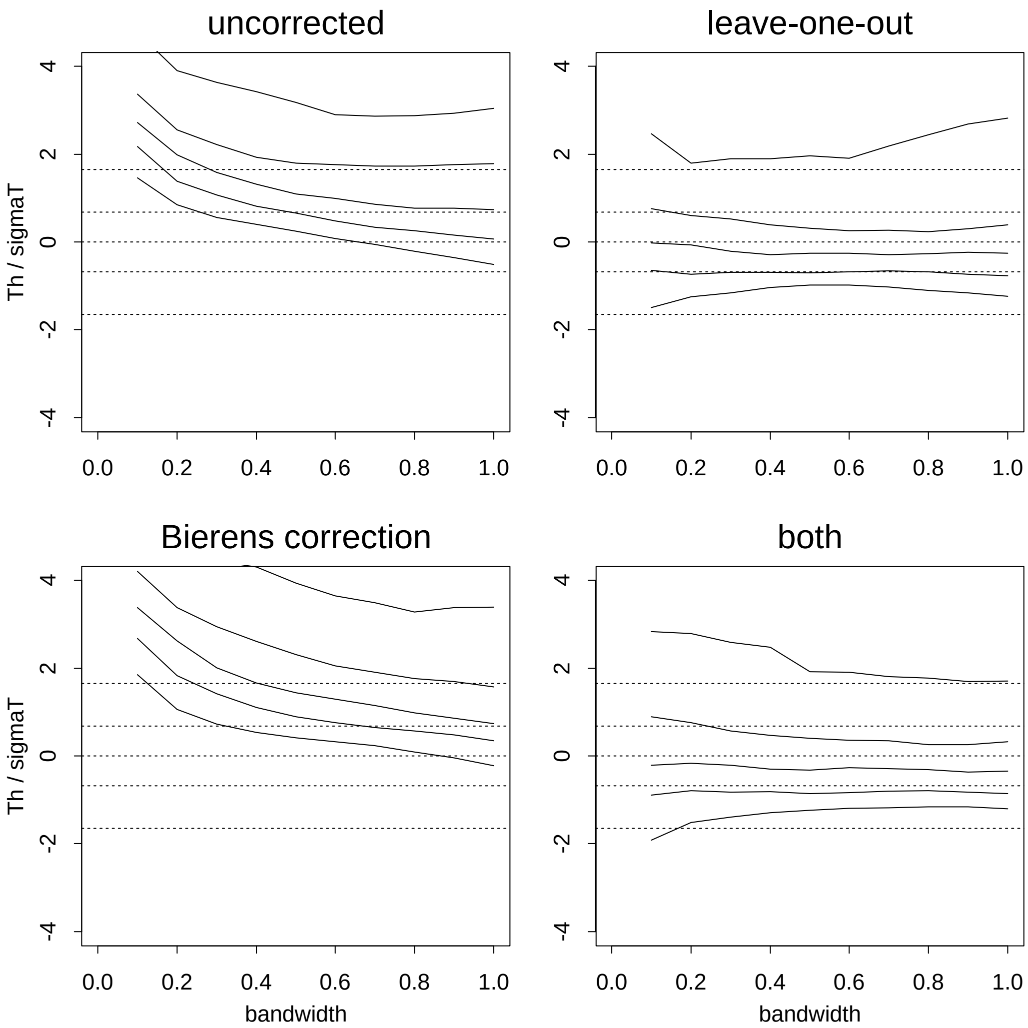

Figure 1: 50 observations, 200 simulations. Upper left: simple kernel smoother; upper right: kernel smoother with independence correction; lower left: kernel smoother with Bierens' correction; lower right: kernel smoother with both independence and Bierens' correction (standard). Solid lines: 5th, 25th, 50th, 75th, and 95th empirical percentiles; dotted lines: 5th, 25th, 50th, 75th, and 95th percentiles of standard normal distribution.

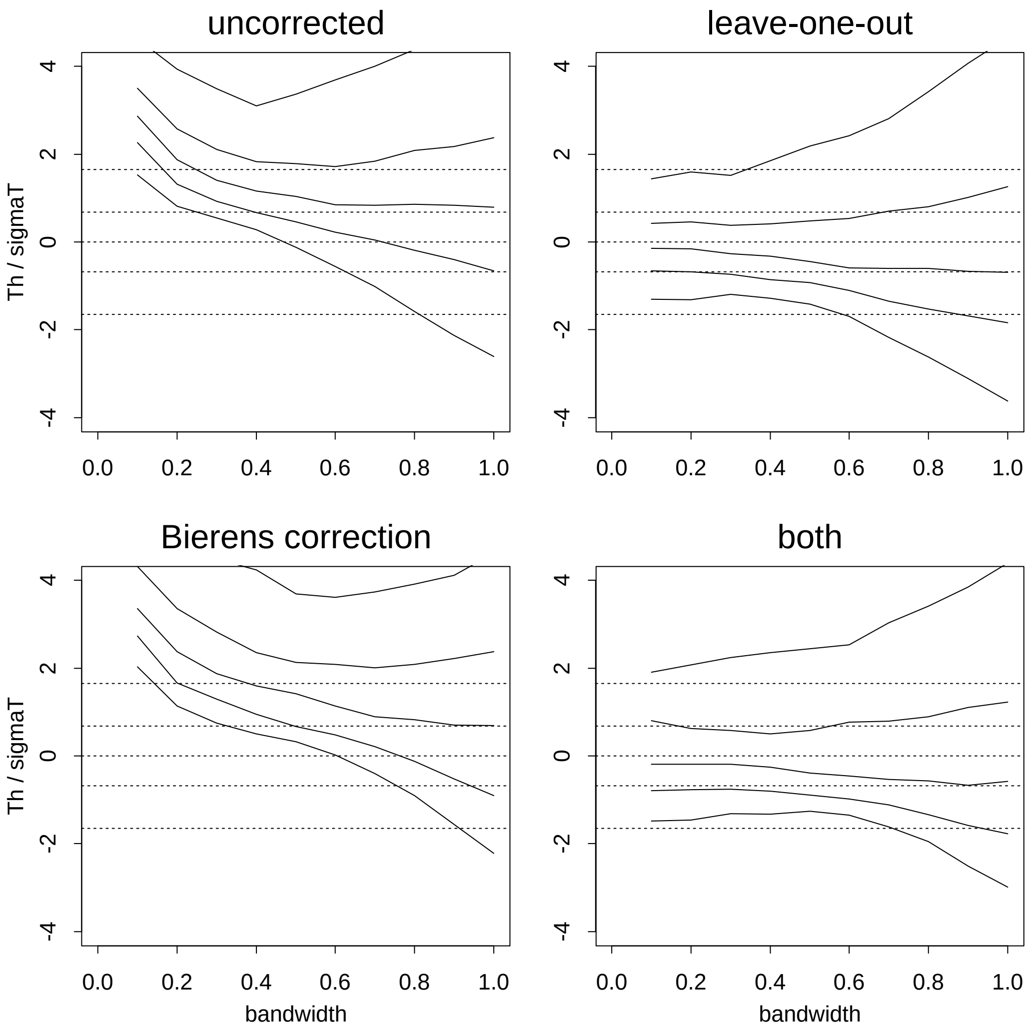

Figure 2: 500 observations, 200 simulations. Upper left: simple kernel smoother; upper right: kernel smoother with independence correction; lower left: kernel smoother with Bierens' correction; lower right: kernel smoother with both independence and Bierens' correction (standard). Solid lines: 5th, 25th, 50th, 75th, and 95th empirical percentiles; dotted lines: 5th, 25th, 50th, 75th, and 95th percentiles of standard normal distribution.

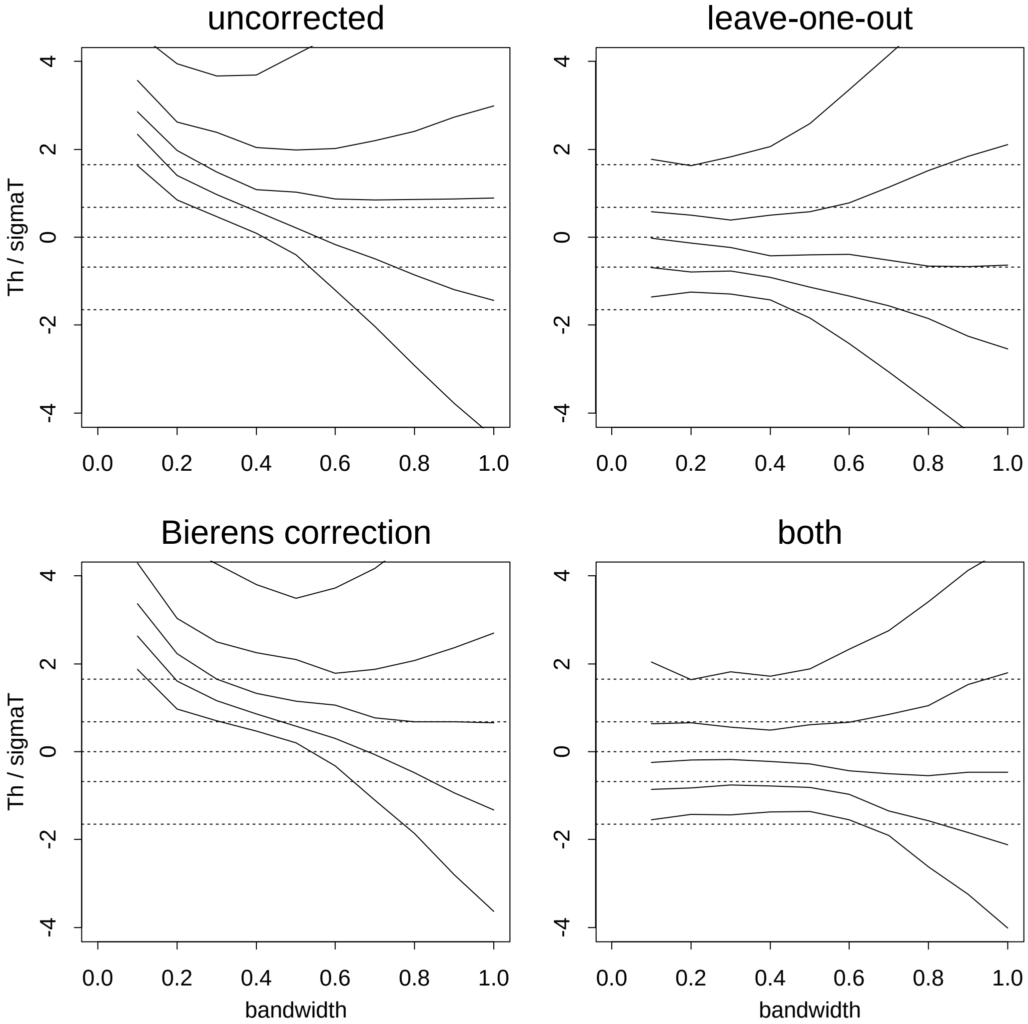

Figure 3: 1000 observations, 200 simulations. Upper left: simple kernel smoother; upper right: kernel smoother with independence correction; lower left: kernel smoother with Bierens' correction; lower right: kernel smoother with both independence and Bierens' correction (standard). Solid lines: 5th, 25th, 50th, 75th, and 95th empirical percentiles; dotted lines: 5th, 25th, 50th, 75th, and 95th percentiles of standard normal distribution.

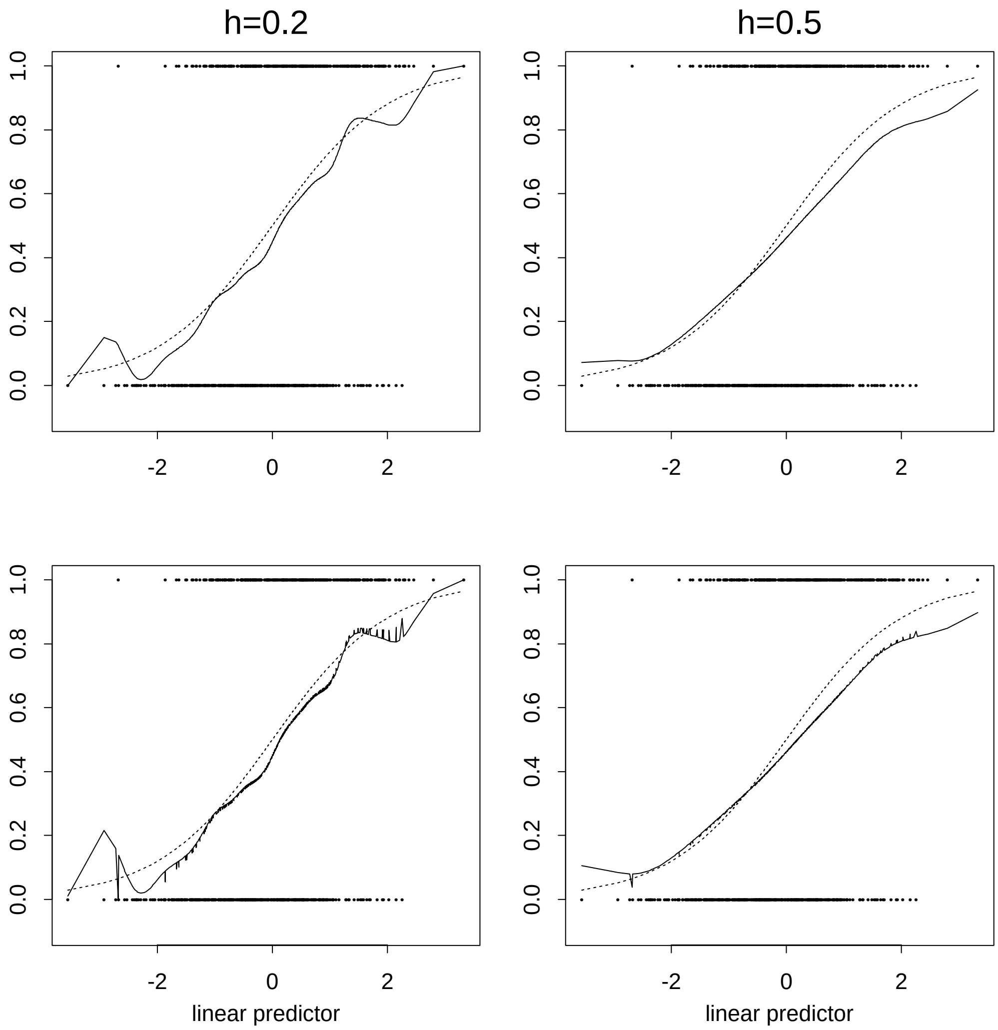

Figure 4: 1000 observations. Solid lines: kernel estimate; dotted line: logit. Top: simple kernel estimates, bottom: kernel estimates using the leave-one-out rule.

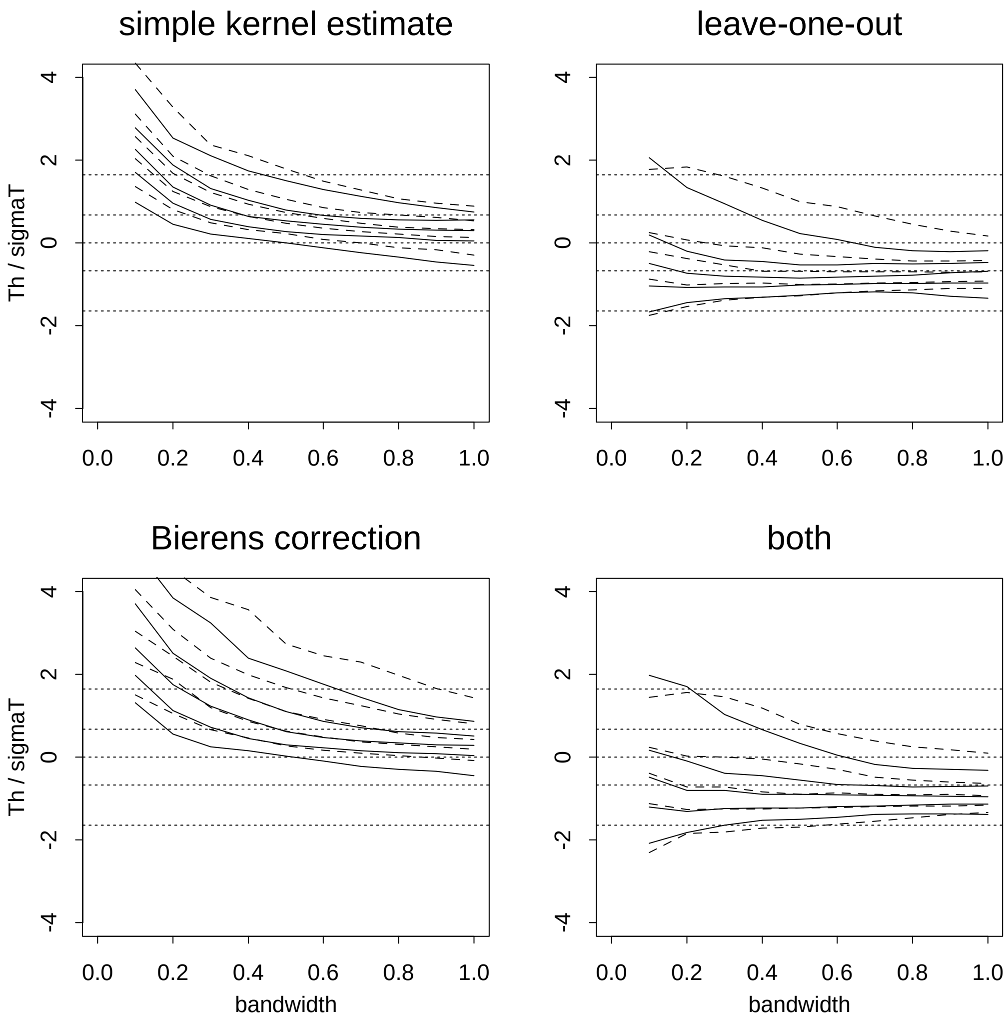

Figure 5: 50 observations, 200 simulations. Upper left: simple kernel smoother; upper right: kernel smoother with independence correction; lower left: kernel smoother with Bierens' correction; lower right: kernel smoother with both independence and Bierens' correction (standard). Solid lines: 5th, 25th, 50th, 75th, and 95th empirical percentiles under the null-hypothesis; dashed lines: 5th, 25th, 50th, 75th percentiles under the alternative; dotted lines: 5th, 25th, 50th, 75th, and 95th percentiles of standard normal distribution.

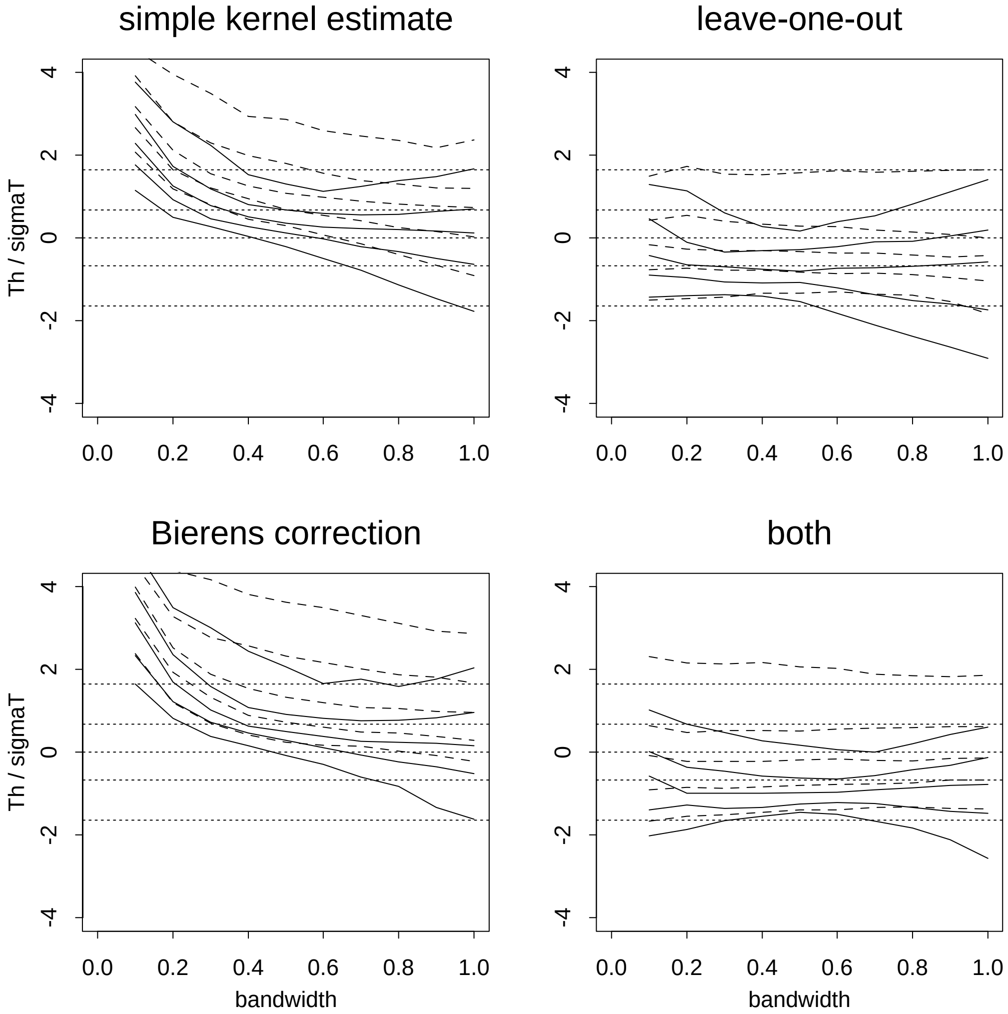

Figure 6: 500 observations, 200 simulations. Upper left: simple kernel smoother; upper right: kernel smoother with independence correction; lower left: kernel smoother with Bierens' correction; lower right: kernel smoother with both independence and Bierens' correction (standard). Solid lines: 5th, 25th, 50th, 75th, and 95th empirical percentiles under the null-hypothesis; dashed lines: 5th, 25th, 50th, 75th percentiles under the alternative; dotted lines: 5th, 25th, 50th, 75th, and 95th percentiles of standard normal distribution.

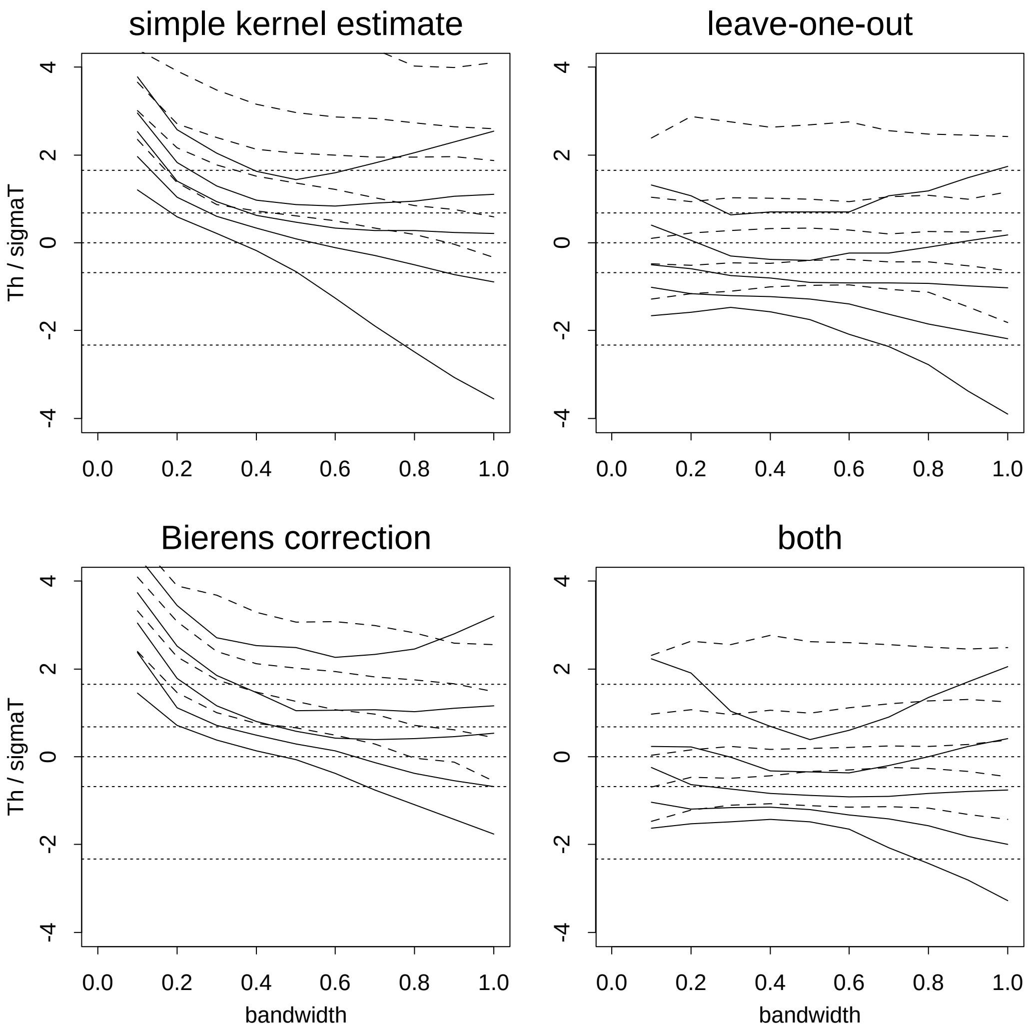

Figure 7: 1000 observations, 200 simulations. Upper left: simple kernel smoother; upper right: kernel smoother with independence correction; lower left: kernel smoother with Bierens' correction; lower right: kernel smoother with both independence and Bierens' correction (standard). Solid lines: 5th, 25th, 50th, 75th, and 95th empirical percentiles under the null-hypothesis; dashed lines: 5th, 25th, 50th, 75th percentiles under the alternative; dotted lines: 5th, 25th, 50th, 75th, and 95th percentiles of standard normal distribution.

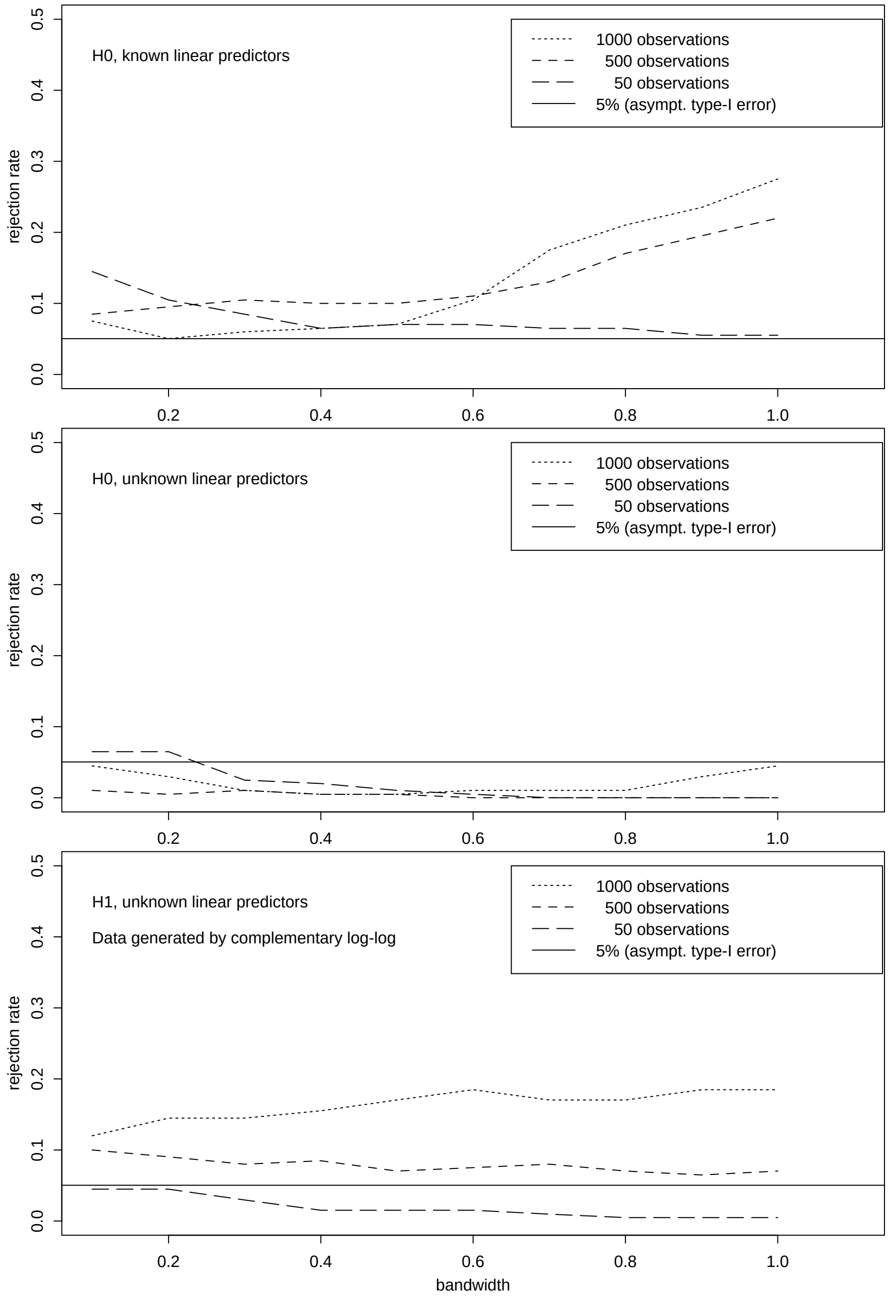

Figure 8: Rejection rates under the full H-H statistic.

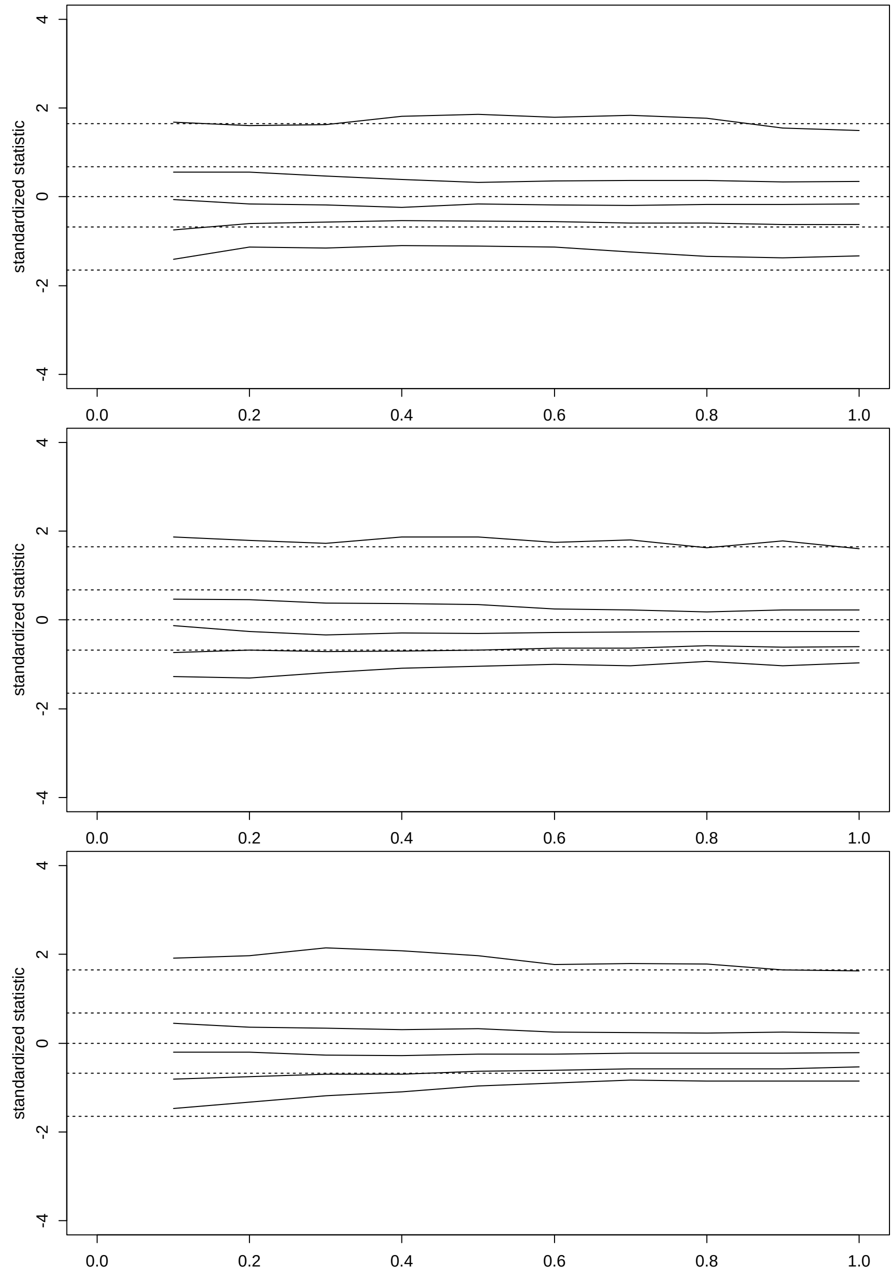

Figure 9: Standardized modified H-H statistic under the null, 5th, 25th, 50th, 75th, and 95th percentiles of empirical distribution (solid) and of standard normal (dotted). Upper plot: 50 observations, middle plot: 500 observations, lower plot: 1000 observations.

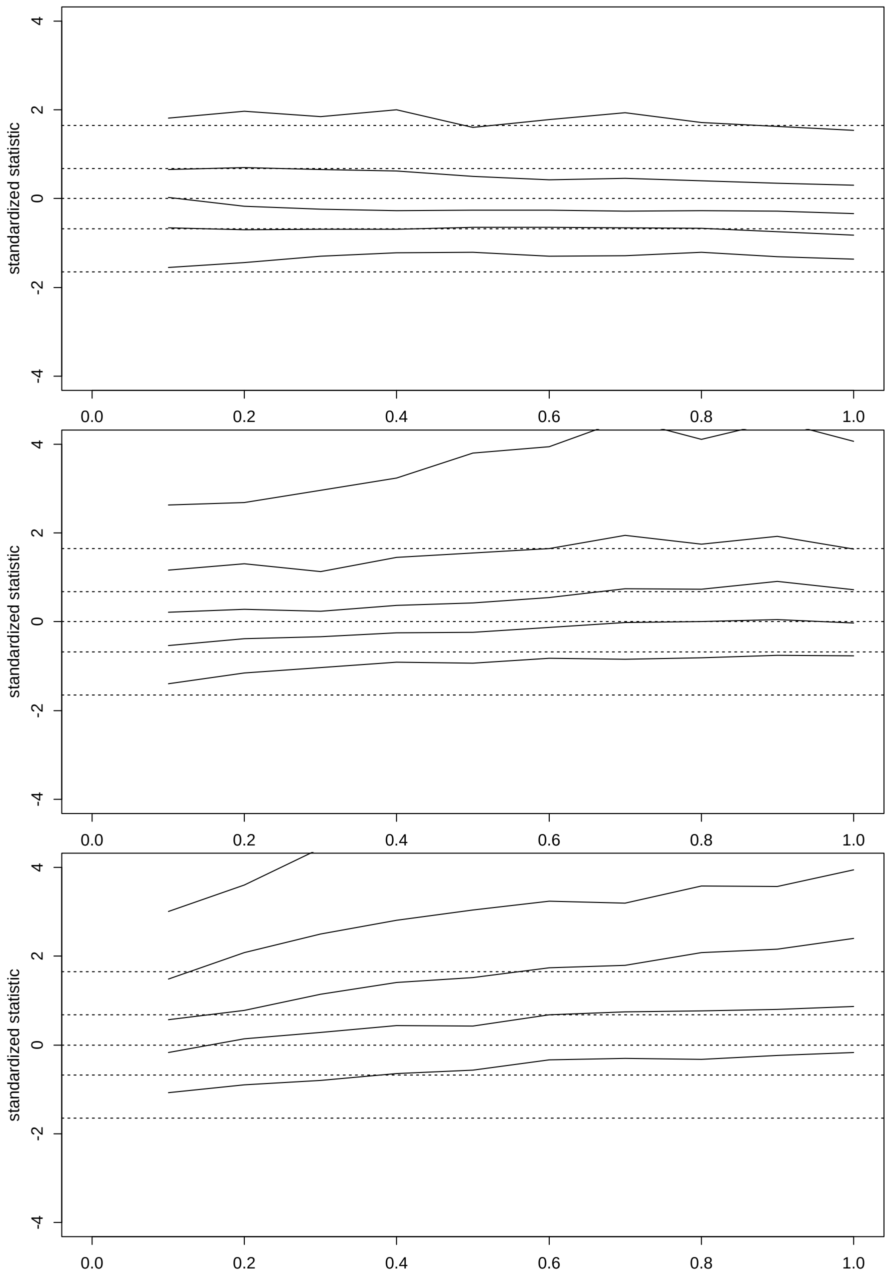

Figure 10: Standardized modified H-H statistic under the alternative complementary log-log, 5th, 25th, 50th, 75th, and 95th percentiles of empirical distribution (solid) and of standard normal (dotted). Upper plot: 50 observations, middle plot: 500 observations, lower plot: 1000 observations.

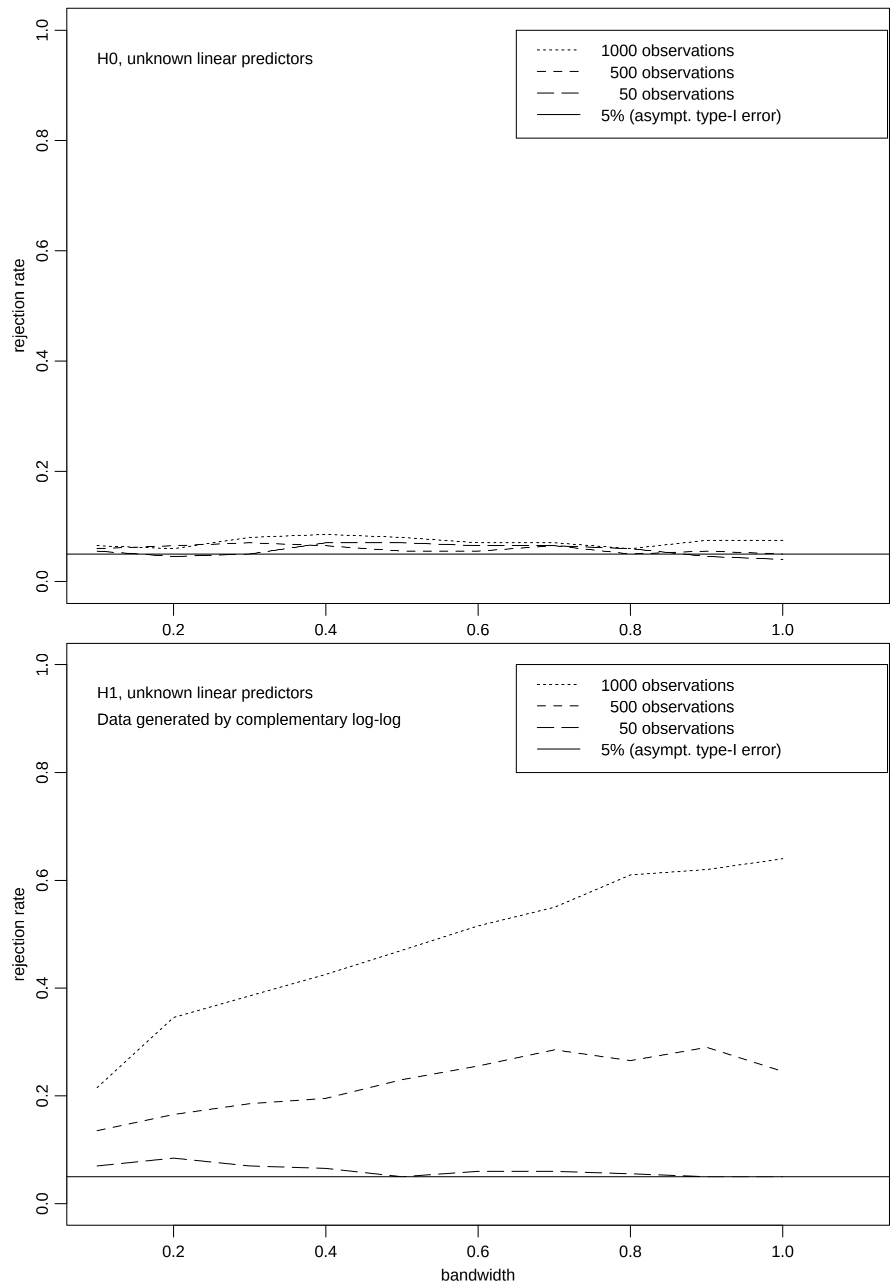

Figure 11: Rejection rates under the modified H-H statistic.