

J.D. Opsomer · M. Francisco-Fernández

Finding Local Departures from a Parametric Model Using Nonparametric Regression

the date of receipt and acceptance should be inserted later

Abstract Goodness-of-fit evaluation of a parametric regression model is often done through hypothesis testing, where the fit of the model of interest is compared statistically to that obtained under a broader class of models. Nonparametric regression models are frequently used as the latter type of model, because of their flexibility and wide applicability. To date, this type

J.D. Opsomer

Colorado State University

Fort Collins, CO 80523

USA

E-mail: jopsomer@stat.colostate.edu

M. Francisco-Fernández

Universidad de A Coruña

A Coruña

Spain

E-mail: mariofr@udc.es

of tests has generally been performed globally, by comparing the parametric and nonparametric fits over the whole range of the data. However, in some instances it might be of interest to test for deviations from the parametric model that are localized to a subset of the data. In this case, a global test will have low power and hence can miss important local deviations. Alternatively, a naive testing approach that discards all observations outside the local interval will suffer from reduced sample size and potential overfitting. We therefore propose a new local goodness-of-fit test for parametric regression models that can be applied to a subset of the data but relies on global model fits, and propose a bootstrap-based approach for obtaining the distribution of the test statistic. We compare the new approach with the global and the naive tests, both theoretically and through simulations, and illustrate its practical behavior in an application. We find that the local test has a better ability to detect local deviations than the other two tests.

Key words: Cramér-von Mises test, wild bootstrap, local polynomial regression.

1 Introduction

A common task for statisticians is to determine whether a certain parametric model is an appropriate representation for a dataset. As part of this determination, it is often desirable to formally test the model, by treating the model as a null hypothesis against an alternative (broader) model and evaluating the probability of obtaining the observed data under the null hypothesis. The choice of the alternative hypothesis model is critical in this determination,

since only deviations from the null hypothesis model that are explicitly allowed by the alternative can be assessed. Nonparametric models are therefore a conceptually appealing choice for the alternative hypothesis, since they only require weak model assumptions such as continuity and/or differentiability.

A number of authors have developed goodness-of-fit tests for parametric models that rely on a smooth alternative estimated by a nonparametric regression method, including [3], [21], [4], [2] and [7]. [19] proposed test statistics based on kernel-weighted log likelihood fits to detect global lack of fit, and derived their asymptotic distributions. [11], [22], [20], [1] and [6] described tests based on overall distance between the parametric and nonparametric fits. [12] discussed a similar approach and compared it with tests based on smoothing the residuals from a parametric fit, similar to the one of [23]. [5] used the difference between estimates of the variance obtained under a parametric and nonparametric fit as a test statistic. [9] proposed a test that measures whether higher derivatives of the unknown function, as estimated by nonparametric regression, are statistically significant. An issue common in these procedures is that they require the selection of a tuning constant for the nonparametric regression under the alternative, or equivalently, a decision on the degree of smoothness of the alternative model. A number of methods have been proposed that automatically adapt to varying degrees of smoothness, including the testing approaches of [18], [13] and [10].

While the methods covered in those articles are quite different, the deviations between the parametric fit and the (nonparametric) fit under the alternative model are all measured *globally*, i.e. they measure the cumulative

discrepancies between both models over the whole range of the data. However, there are cases in which we might be more interested in discovering *local* deviations from the null model. Such a deviation would be caused by a departure from a parametric model for a subset of the data only, while the fit for the rest of the data might be well-represented by the parametric model. This type of local departures from the null model specification should not be confused with the term “local alternative,” which refers to deviations between the null and the alternative model that become smaller as the sample size increases.

The idea of testing for local deviations was motivated by a collaboration of the first author with U.S. Bureau of Labor Statistics staff, to evaluate the appropriateness of a no-intercept linear (ratio) model used as a component in the procedure for producing monthly employment estimates. In this model, establishment-level employment for the current month is assumed to be related through the ratio model to the employment in the previous month. This appears to be a reasonable overall model, but there is some concern that the overall relationship might not hold uniformly well for all establishment sizes (especially for the smallest or largest establishments), and hence lead to poor predictions for some of the establishments. A local test for deviations from the overall parametric model could be useful for testing for local deviations from this model, even in cases where the overall model fits well.

Because of confidentiality issues, the data from the application described above are not available to the public. We therefore consider a different application in the current article, which will be used to illustrate the local testing

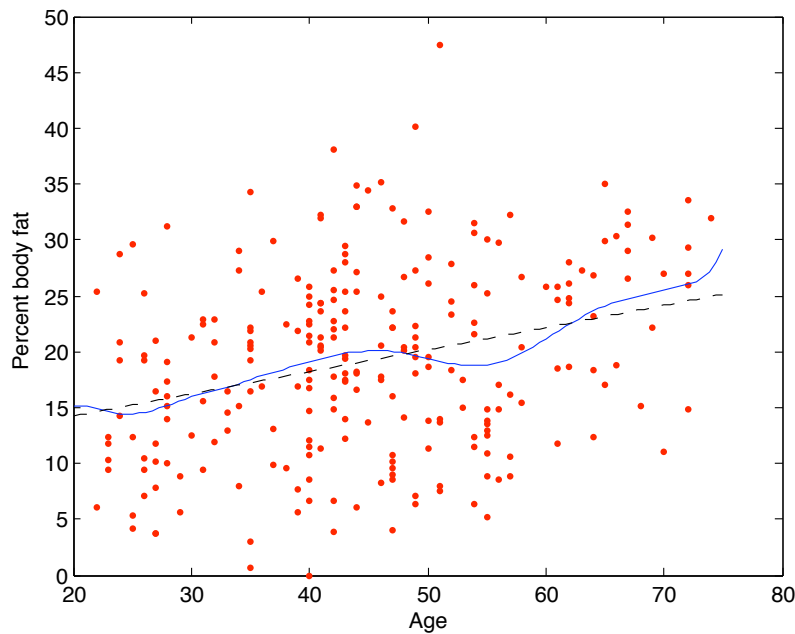


Fig. 1 Plot of percentage bodyfat versus age for 252 men. Dashed line represents linear regression fit, solid line represents local linear regression fit.

procedure. The data are from the “bodyfat” dataset available on StatLib (<http://lib.stat.cmu.edu/>). They were originally described in [16] and made available to the public in [14]. In this application, an estimate of the percentage bodyfat is obtained for 252 men through underwater weighting, together with a set of associated variables such as age and various physiological measurements. Figure 1 shows the fits for percentage bodyfat as a function of age, using both linear regression and nonparametric regression. While the linear fit appears broadly reasonable as an overall model, a question of interest might be whether the “dip” observed between 50 and 60 years of age represents a statistically significant deviation from the linear fit. We will discuss this question further in Section 4.

In situations like those described above, three nonparametric testing approaches are in principle possible: (1) apply a global test for goodness of fit and attempt to capture the local deviation, (2) apply a global test after discarding the data outside the range of the potential local deviation, or (3) adapt a global test so that it specifically targets the local deviation. The third type of test does not exist in the statistical literature and will be developed later in this article, but we will describe some of the drawbacks with the first two testing approaches in detecting local deviations.

Because of the cumulative or “integrated” nature of most global measures of discrepancy between null and alternative hypotheses, a global test is likely to be unable to differentiate between one relatively large but localized discrepancy and many smaller discrepancies between the fits. The former type might be due to a significant local departure from the null model, while the latter could be due to the presence of noise around a correctly specified null model. As statistical tests are designed to accept the null hypothesis if the deviation could reasonably be explained by noise, this inability to differentiate between both types of deviation results in global methods having low power in situations with localized deviations.

The second type of tests consists of applying a global test on a restricted dataset obtained by discarding all data outside the range of interest. This approach, which we will refer to as a “naive” test, has the intuitive advantage of specifically targeting the area of interest, but has a number of important drawbacks. First, by reducing the sample size, the test loses power, since both the null and the alternative fits are now computed on a subset of the

data. Second, this procedure suffer from potentially severe “over-fitting bias,” in the sense that the parametric fit in the local interval might deviate from the overall parametric model fit, with a resulting further loss of power. Finally, if multiple local intervals are to be tested, the naive procedure requires that the null and alternative fit be recomputed for each interval, which adds to the computational complexity of the procedure.

The third type, to be further developed in what follows, is a local testing procedure that can be applied when a global parametric fit is obtained, but the presence of local deviations is suspected in a subset of the data. By relying on global fits but restricting the testing interval to a specific area, this testing approach can be expected to combine the good features of global and naive tests while avoiding most of their disadvantages.

It is clear that the idea of restricting a test for departure from a parametric shape to a local interval entails the risk of “data snooping,” in which a feature is first identified visually and then a test is constructed around it. In our theoretical discussion of the local test, we will be assuming an a priori specified interval and hence do not account for possible data snooping. In the application of the test to the bodyfat dataset, we will discuss a Bonferroni adjustment to the local test.

The remainder of the article is as follows. In Section 2, we generalize the global Cramér-von Mises-based test of [1], derive the asymptotic properties of the proposed local test, and compute the asymptotic relative efficiencies for the local test in comparison with the global and the naive tests. We also propose a bootstrap-based method to compute the critical value for the local

test without relying on its asymptotic distribution. In Section 3, we compare the behavior of the three testing procedures in simulation experiments. Section 4 discusses the application of the testing procedures to the bodyfat dataset.

2 Local Test

We describe the local hypothesis test for the situation in which the assumed null model is a polynomial in a single continuous covariate, as was done by [1]. The approach continues to hold, with some modifications, if a different parametric form is considered, and other local testing procedures such as those mentioned in Section 1 could also be “localized” in a manner similar to what we will be describing here.

Suppose that we observe data $(x_i, Y_i), i = 1, \dots, n$, where $x_i \in [a, b]$, and we are interested in fitting a regression model to these data. We believe that the model

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i$$

with $\mathbf{x} = (1, x, \dots, x^p)^T$ for some fixed $p \geq 0$, is a reasonable overall specification for the data, but we are concerned that there might be local departures from that model. Therefore, we would like to test the hypothesis

$$H_0 : \text{E}(Y|x) = \mathbf{x}^T \boldsymbol{\beta}$$

against the alternative

$$H_A : \text{E}(Y|x) = m(x) \neq \mathbf{x}^T \boldsymbol{\beta}$$

where $m(\cdot)$ is a smooth but otherwise unspecified function. Because we are interested in a local test, we would like to develop a procedure that can be applied to a specific interval within $[a, b]$ and determine whether observed discrepancies in that interval are statistically significant. If such a local discrepancy is identified and found to be sufficiently large to warrant modifying the model, then the overall parametric model could be extended in such a way that it captures the observed deviation.

We will use the Cramér-von Mises test statistic proposed in [1], adjusted for the fact that we are only testing for deviations in an domain $D_n \subset [a, b]$ (the subscript n is added as we will be considering the asymptotic properties of the test and want to allow the domain to shrink with n). We first define the parametric and nonparametric estimators that will be used to compute estimates under H_0 and H_A . Under H_0 , the mean function is assumed to be a polynomial in x , and the parameter β is estimated by least squares regression. Let $m_0(x; \hat{\beta})$ denote the estimator for $m(x)$ in that case. Under H_A , the function m is no longer assumed to be a polynomial function, but it is still assumed to be smooth function of x . In that case, $m(x)$ is estimated by local polynomial regression. Specifically,

$$\hat{m}(x; h_n) = \mathbf{e}'_1 (\mathbf{X}'_x \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{X}'_x \mathbf{W}_x \mathbf{Y} \quad (1)$$

where $\mathbf{e}_1 = (1, 0, \dots, 0)'$, $\mathbf{Y} = (Y_1, \dots, Y_n)'$,

$$\mathbf{X}_x = \begin{bmatrix} 1 & x_1 - x & \dots & (x_1 - x)^q \\ \vdots & \vdots & & \vdots \\ 1 & x_n - x & \dots & (x_n - x)^q \end{bmatrix},$$

and $\mathbf{W}_x = \text{diag}\{K((x_1 - x)/h_n), \dots, K((x_n - x)/h_n)\}$, with $K(\cdot)$ a kernel function, q the degree of the local polynomial and h_n the bandwidth (conditions on both will be discussed below).

The local Cramér-von Mises test statistic is defined as

$$T_{n,L} = \int_{D_n} \left(\hat{m}(x; h_n) - m_0(x; \hat{\boldsymbol{\beta}}) \right)^2 w(x) dx, \quad (2)$$

where $w(\cdot)$ is a fixed weight function. It is clear that the statistic $T_{n,L}$ will be large when the difference between the parametric and nonparametric fits, evaluated on the domain D_n , is large, and small otherwise. Specifically, the types of model deviations that can be captured by this test are of the form $m(x) - m_0(x; \boldsymbol{\beta}) \equiv g_n(x)$, where $g_n(\cdot)$ is a non-zero bounded function orthogonal to polynomials of degree p . Since we are interested in local deviations, we let $g_n(x) = c_n g(x) I_{\{x \in D_n\}}$, with $g(\cdot)$ a fixed non-zero function, and $I_{\{A\}} = 1$ if A is true and 0 otherwise. This formulation will allow for fixed alternatives, for which $c_n \equiv c$, as well as local alternatives, where $c_n \rightarrow 0$.

The following assumptions are needed to derive the asymptotic distribution of $T_{n,L}$.

A 1 *Model assumptions: The variable x has compact support $[a, b]$, and its density $f(\cdot)$ is bounded away from 0 and is twice continuously differentiable. The function $m(\cdot)$ is bounded on $[a, b]$ and has $q+1$ bounded derivatives. The variance function $\sigma^2(x) = \text{Var}(Y|x)$ is continuous and bounded away from 0 and ∞ .*

A 2 *Parametric estimator assumptions: The parameter $\boldsymbol{\beta}$ is estimated by $\hat{\boldsymbol{\beta}}$, with $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = O_p(1/\sqrt{n})$.*

A 3 *Nonparametric estimator assumptions:* The kernel K is symmetric, bounded and has compact support. The local polynomial degree $q \geq p$, where p is the degree of the null model polynomial. The bandwidth satisfies $h_n \rightarrow 0$, $nh_n^{q+1} \rightarrow \infty$.

A 4 *Local deviation assumptions:* The domain D_n is of size $\|D_n\|$, and if $\|D_n\| \rightarrow 0$ then we require $h_n/\|D_n\| \rightarrow 0$. The size of the local deviation, $|c_n|$, is bounded, and if $c_n \rightarrow 0$, then $h_n/c_n \rightarrow 0$.

Under those assumptions, we obtain the following result, which generalizes that of [1] for the case of testing in a pre-specified domain D_n . The proof is in the Appendix.

Theorem 1 *Under assumptions A1–A4, the test statistic $T_{n,L}$ defined in (2) has the asymptotic distribution*

$$V_{Ln}^{-1/2}(T_{n,L} - b_{0n} - b_{1n}) \rightarrow_{\mathcal{L}} N(0, 1),$$

where

$$b_{0n} = \frac{1}{nh_n} \tilde{K}_q^{(2)}(0) \int_{D_n} \frac{\sigma^2(x)w(x)}{f(x)} dx,$$

$$b_{1n} = c_n^2 \int_{D_n} \left(\tilde{K}_{q,h} * g(x) \right)^2 w(x) dx,$$

and

$$V_{Ln} = 2 \frac{1}{n^2 h_n} \tilde{K}_q^{(4)}(0) \int_{D_n} \left(\frac{\sigma^2(x)w(x)}{f(x)} \right)^2 dx,$$

where $\tilde{K}_q(\cdot)$ is the equivalent kernel of order q , $\tilde{K}_q^{(r)}(\cdot)$ denotes its r -fold convolution with itself and $\tilde{K}_{q,h}(x) = 1/h_n \tilde{K}_q(x/h_n)$.

For the definition of the equivalent kernel, see [8, p.64]. While the exact expressions for the equivalent kernels and convolutions are relatively involved,

they can be computed for any particular choice of $K(\cdot)$. The asymptotic distribution in Theorem 1 is valid under both the null and the alternative hypothesis, regardless of whether the latter is fixed ($c_n = c$) or local ($c_n = o(1)$). If the null model is true, i.e. $c_n = 0$, then $b_{1n} = 0$, so that this result can be used to derive an asymptotic test as long as b_{0n} is known or can be estimated consistently.

The asymptotic results in Theorem 1 allow us to compare the power of the proposed local test with that of the global test of [1], and with that of the naive test which ignores any data outside of the interval D_n in both model fitting and the computation of the test statistic. The asymptotic distribution of the global test statistic, denoted by $T_{n,G}$ in what follows, was derived in [1]. In the case of a local model deviation $g_n(\cdot)$ of the form $c_n g(x) I_{\{x \in D_n\}}$ and satisfying assumption A4, the asymptotic distribution of $T_{n,G}$ is similar to that shown in Theorem 1 for $T_{n,L}$, except that the D_n are replaced by the full interval $[a, b]$ in b_{0n} and V_{Ln} (but not in b_{1n}). The asymptotic distribution of the naive local test, denoted $T_{n,N}$, is the same as in Theorem 1 except that n needs to be replaced by n^* , the sample size for the observations that fall in D_n , in both b_{0n} and V_{Ln} . Note that the asymptotic bias term b_{1n} is the same as in Theorem 1 for all three tests. We will denote by V_{tn} , $t = L, G, N$ the asymptotic variance of the three statistics.

The asymptotic power of all three tests depends only on the ratio $b_{1n}/\sqrt{V_{tn}}$. We are assuming that we can estimate the bias term b_{0n} for each of them, and denote by $T_{n,t}^\dagger$ the test statistic adjusted for this bias. The asymptotic

power of the adjusted tests is defined as

$$P_{\alpha,t} \equiv \Pr \left(\frac{T_{n,t}^\dagger}{\sqrt{V_{tn}}} > z_{1-\alpha} \right),$$

where $z_{1-\alpha}$ is the $1 - \alpha$ quantile of the standard normal distribution. A common approach to measure the efficiency of statistical tests is through the asymptotic relative efficiency (ARE), also called Pitman efficiency ([15]), which compares the sample sizes of different tests required to achieve a given power. In order to make the tests comparable, we will use the sample sizes for the whole interval $[a, b]$ for all three tests, and denote those by n_L, n_G and n_N , respectively. In the case of the naive test, this means that the sample size is set at n_N , but that only n^* observations are actually used in the construction of the test over D_n . Therefore, we make the following additional assumption.

A 5 *Sample size assumption:* The sample size n^* in the interval D_n satisfies $n^* = n \int_{D_n} f(x) dx$.

Assumption A5 gives an explicit relationship between the overall sample size and that in D_n . If the x_i are randomly generated from the density $f(\cdot)$, it will hold in probability. The following theorem describes the asymptotic relative efficiency (ARE) of the three adjusted test statistics and is proven in the Appendix.

Theorem 2 *Under the assumptions A1–A5, the ARE for the adjusted test statistics $T_{n,L}^\dagger, T_{n,G}^\dagger$ and $T_{n,N}^\dagger$ are*

$$\begin{aligned}\frac{n_L}{n_G} &= \sqrt{\frac{\int_{D_n} r(x) dx}{\int_{[a,b]} r(x) dx}} (1 + o(1)) = O(\|D_n\|^{1/2}) \\ \frac{n_L}{n_N} &= \int_{D_n} f(x) dx (1 + o(1)) = O(\|D_n\|) \\ \frac{n_G}{n_N} &= \sqrt{\frac{\int_{[a,b]} r(x) dx}{\int_{D_n} r(x) dx}} \int_{D_n} f(x) dx (1 + o(1)) = O(\|D_n\|^{1/2})\end{aligned}$$

with $r(x) = \left(\frac{\sigma^2(x)w(x)}{f(x)}\right)^2$.

In the case of $\|D_n\| \rightarrow 0$, each of these ARE converge to 0, so that the local test is more efficient than both the global and the naive test, and the global test is more efficient than the naive test. If D_n is treated as a fixed interval, the first two ARE are strictly smaller than 1 unless $D_n = [a, b]$, showing that the local test again dominates both the global and the naive test. However, the comparison between the global and the naive test is generally ambiguous when D_n is fixed. These results hold under the fixed as well as the local alternative, as c_n does not have an effect on the leading terms of the AREs.

As noted in other nonparametric testing contexts, the asymptotic distribution of Theorem 1 is often not sufficiently precise for constructing a practical test in small-to-medium sample size situation. Hence, following [1], we instead construct a bootstrap-based distribution function for $T_{n,L}$, from which relevant probabilities can be estimated. The procedure of [1] can be applied directly to obtain a bootstrap distribution for $T_{n,G}$ and $T_{n,L}$.

The procedure is based on the *wild bootstrap* procedure of [11]. An important advantage of the wild bootstrap is that the bootstrap distribution remains valid for any variance function specification satisfying assumption A1, because the first three moments of each bootstrap error (over the bootstrap distribution) are constructed to match those of the corresponding sample residual under the null model. Specifically, let the sample residuals be denoted as $\hat{\xi}_i = Y_i - m_0(x_i; \hat{\beta})$, $i = 1, 2, \dots, n$. For each $i = 1, 2, \dots, n$, the bootstrap error ξ_i^* is selected from a two-point distribution satisfying

$$E^*(\xi_i^*) = 0, \quad E^*(\xi_i^{*2}) = \hat{\xi}_i^2, \quad E^*(\xi_i^{*3}) = \hat{\xi}_i^3, \quad (3)$$

where E^* denotes the expectation operator over the bootstrap distribution. The three requirements in (3) induce the set of constraints

$$\begin{aligned} \gamma e_{1i} + (1 - \gamma)e_{2i} &= 0 \\ \gamma e_{1i}^2 + (1 - \gamma)e_{2i}^2 &= \hat{\xi}_i^2 \\ \gamma e_{1i}^3 + (1 - \gamma)e_{2i}^3 &= \hat{\xi}_i^3, \end{aligned}$$

where e_{1i}, e_{2i} denote the two possible point values for $\hat{\xi}_i^*$ and γ the probability of obtaining e_{1i} . A solution to these constraints is given by $\gamma = (5 + 5^{1/2})/10$, $e_{1i} = \hat{\xi}_i(1 - 5^{1/2})/2$ and $e_{2i} = \hat{\xi}_i(1 + 5^{1/2})/2$.

A bootstrap sample under the null model is constructed by letting $Y_i^* = m_0(x_i; \hat{\beta}) + \xi_i^*$, $i = 1, \dots, n$, and the bootstrap test statistics $T_{n,L}^*$ is computed as in (2). We prove the following result in the Appendix, with similar results holding for bootstrapped statistics $T_{n,G}^*$ and $T_{n,N}^*$ by [1].

Theorem 3 *Under assumptions A1–A4, the bootstrap test statistic $T_{n,L}^*$ has the same asymptotic distribution as $T_{n,L}$ under H_0 :*

$$V_{Ln}^{-1/2}(T_{n,L}^* - b_{0n}) \rightarrow_{\mathcal{L}} N(0, 1).$$

3 Simulation Study

In this Section, we describe a simulation experiment to compare the behavior of the three tests. We generated 1000 samples of sample size $n = 100$ for the regression model $Y_i = m(x_i) + \sigma\varepsilon_i$, with

$$m(x) = (2x - 1) + 0.5 \exp(-50(2x - 1.5)^2), \quad (4)$$

$\sigma^2 = 0.25$, a fixed and equally-spaced design in the interval $[0, 1]$, and errors generated from independent and identically distributed $N(0, 1)$ random variables. The wild bootstrap resampling was performed $B = 1000$ times for each sample. The weight function used in all three tests was taken as $w(x) = 1$. The interval $[0, 1]$ was split in six sub-intervals, $[0, 0.2]$, $[0.2, 0.4]$, $[0.4, 0.6]$, $[0.6, 0.8]$, $[0.8, 1]$ as well as $[0.15, 0.35]$, where the deviation from a linear model is localized. Figure 2 shows the regression function (4), a random sample from this model and the six sub-intervals considered.

In this model, the parametric estimator under the null hypothesis was calculated using the Ordinary Least Squares estimator, while the model under the alternative hypothesis was fitted using local linear regression with an Epanechnikov kernel and bandwidth values $h = 0.05, 0.075$ and 0.1 . In Table 1, the simulated rejection probabilities obtained for $T_{n,L}$ are presented

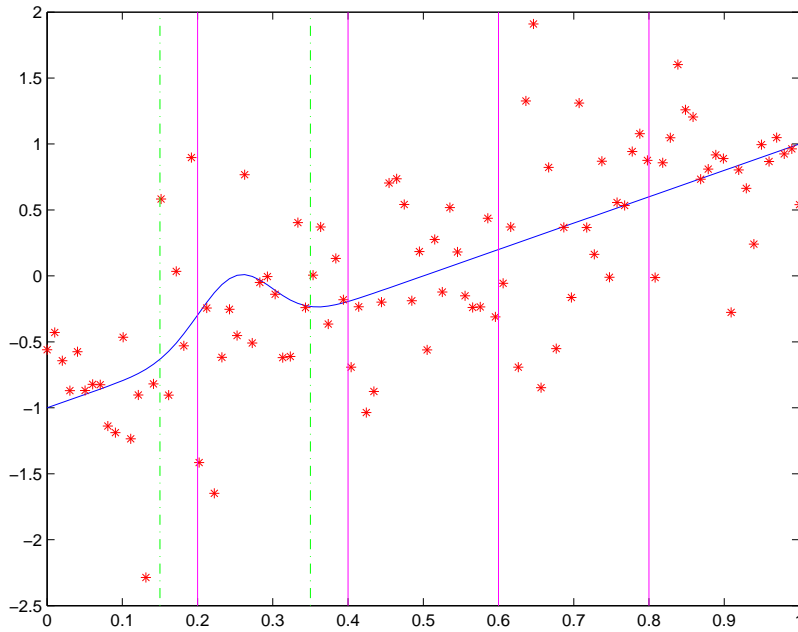


Fig. 2 Regression model (4), a random sample from this model and the six sub-intervals considered.

for significance levels $\alpha = 0.05$ and 0.1 , as well as the mean and standard deviation of the p -values over the 1000 trials. Tables 2 and 3 show the same results for $T_{n,G}$ and for $T_{n,N}$, respectively.

When comparing the results for the tests, it is clear that the probability of (correctly) rejecting the null hypothesis as well as the average p -value are substantially higher in the local test on the intervals $[0.2, 0.4]$ and $[0.15, 0.35]$ than in the global test, even though the overall deviation in the mean model is the same. The naive test performed the worst, almost completely missing the model deviation. This is likely due primarily to the overfitting bias described earlier, which causes a line fitted inside each of the intervals to be close to the

Interval	[0, 0.2]			[0.2, 0.4]			[0.4, 0.6]		
h	0.05	0.075	0.1	0.05	0.075	0.1	0.05	0.075	0.1
$\alpha = 0.05$	0.099	0.140	0.176	0.393	0.466	0.509	0.048	0.067	0.089
$\alpha = 0.1$	0.209	0.271	0.284	0.541	0.636	0.652	0.116	0.138	0.153
$\mu_{p\text{-value}}$	0.333	0.309	0.295	0.163	0.138	0.128	0.435	0.431	0.425
$\sigma_{p\text{-value}}$	0.249	0.254	0.253	0.196	0.188	0.181	0.274	0.282	0.285

Interval	[0.6, 0.8]			[0.8, 1]			[0.15, 0.35]		
h	0.05	0.075	0.1	0.05	0.075	0.1	0.05	0.075	0.1
$\alpha = 0.05$	0.036	0.045	0.055	0.037	0.045	0.052	0.414	0.483	0.546
$\alpha = 0.1$	0.102	0.120	0.121	0.102	0.115	0.115	0.567	0.652	0.698
$\mu_{p\text{-value}}$	0.462	0.462	0.462	0.443	0.450	0.456	0.155	0.128	0.117
$\sigma_{p\text{-value}}$	0.275	0.285	0.289	0.265	0.273	0.278	0.189	0.180	0.176

Table 1 Rejection probabilities, and average and standard deviation of the p -values of the local test $T_{n,L}$ for simulation experiment on regression model (4).

Interval	[0, 1]		
h	0.05	0.075	0.1
$\alpha = 0.05$	0.224	0.275	0.286
$\alpha = 0.1$	0.358	0.426	0.442
$\mu_{p\text{-value}}$	0.237	0.216	0.212
$\sigma_{p\text{-value}}$	0.223	0.223	0.223

Table 2 Rejection probabilities, and average and standard deviation of the p -values of the global test $T_{n,G}$ for simulation experiment on regression model (4).

nonparametric fit, instead of to the overall linear fit. We can also evaluate the behavior of the local and naive tests when the null model is correct, by considering the portions of Tables 1 and 3 for the intervals $[0.6, 0.8]$ and

Interval	[0, 0.2]			[0.2, 0.4]			[0.4, 0.6]		
h	0.05	0.075	0.1	0.05	0.075	0.1	0.05	0.075	0.1
$\alpha = 0.05$	0.062	0.077	0.097	0.074	0.092	0.092	0.044	0.062	0.068
$\alpha = 0.1$	0.155	0.177	0.176	0.162	0.168	0.165	0.133	0.138	0.145
$\mu_{p\text{-value}}$	0.402	0.407	0.416	0.376	0.386	0.397	0.422	0.434	0.443
$\sigma_{p\text{-value}}$	0.264	0.274	0.284	0.266	0.277	0.279	0.266	0.277	0.285

Interval	[0.6, 0.8]			[0.8, 1]			[0.15, 0.35]		
h	0.05	0.075	0.1	0.05	0.075	0.1	0.05	0.075	0.1
$\alpha = 0.05$	0.044	0.061	0.070	0.044	0.061	0.070	0.197	0.235	0.265
$\alpha = 0.1$	0.132	0.139	0.146	0.132	0.139	0.146	0.331	0.383	0.391
$\mu_{p\text{-value}}$	0.424	0.434	0.443	0.423	0.434	0.443	0.284	0.274	0.268
$\sigma_{p\text{-value}}$	0.266	0.277	0.285	0.266	0.277	0.285	0.253	0.260	0.263

Table 3 Rejection probabilities, and average and standard deviation of the p -values of the naive test $T_{n,N}$ for simulation experiment on regression model (4).

[0.8, 1]. The results show that the tests approximately achieve the correct level of α for all three considered bandwidth values.

Finally, we evaluated the finite sample behavior of the asymptotic test of Theorem 1, by applying it to the same simulation setup as above. Under the null hypothesis, $T_{n,L}$ is approximately distributed $N(b_{0n}, V_{Ln})$, so that it requires explicit estimation of b_{0n} and V_{Ln} . Setting $f(x) = w(x) = 1$, this only requires computing the kernel convolution constants and estimating the variance parameter σ^2 based on the residuals of the parametric fit. In all the intervals considered, the asymptotic test displayed a pronounced tendency to reject the null hypothesis. This happened even when the null model was correct (as in the intervals [0.6, 0.8] and [0.8, 1]. Hence, the asymptotic test appears to be severely biased, at least at the sample size evaluated here, and

we recommend using the bootstrap-based test for practical applications of local testing.

4 Application

We now return to the bodyfat example. Prior to analyzing those data, we removed a single observation with age of 81, because it was strongly influencing both the parametric and nonparametric fits. In the plot of Figure 1, fits obtained by simple linear regression and local linear regression are shown (the latter computed with bandwidth $h = 10$). There appears to be a noticeable “dip” below the line in the 50-60 years of age region, and we will compare the different testing approaches in their ability to assess the significance of this observed pattern.

We considered the same three tests as in Section 3 and used bandwidth values of in the range $h = 5$ to 15 to ensure that the results are not determined by the bandwidth choice. We also considered a range of testing intervals, to illustrate the effect of the interval choice on the results. The wild bootstrap approach with 1000 replicates was used to compute the p -values.

The first four rows of Table 4 show the bootstrap p -values for the proposed local test with bandwidths $h = 5, 10, 15$ for intervals centered around the dip in $[50, 60]$ but increasing in width from 10 to 40. The middle two rows labeled “Global” display the results for two versions of the global test: the interval $[20, 75]$ encompasses all the data, and the interval $[25, 70]$ corresponds to using a weight function $w(\cdot)$ in (2) to remove the boundary effect of the nonparametric fit. The final four rows show the results for the naive tests

Tests	Interval	Bandwidth		
		$h = 5$	$h = 10$	$h = 15$
Local	[50, 60]	0.014	0.007	0.012
	[45, 65]	0.068	0.030	0.052
	[40, 70]	0.080	0.053	0.089
	[35, 75]	0.265	0.245	0.275
Global	[20, 75]	0.577	0.512	0.458
	[25, 70]	0.139	0.084	0.108
Naive	[50, 60]	0.119	0.197	0.231
	[45, 65]	0.278	0.141	0.097
	[40, 70]	0.144	0.118	0.076
	[35, 75]	0.281	0.223	0.213

Table 4 Bootstrap p -values obtained by nonparametric goodness-of-fit testing procedures for the linear model for the bodyfat-age relationship.

obtained by first discarding all the observations outside the interval and performing a global test on the remaining data.

The local test assigns a high degree of significance to the observed discrepancy between both curves when the interval is exactly [50,60] for all bandwidth values. The significance level decreases as the interval increases, but a significant departure (at the 5% confidence level) is still identified with $h = 10$ for the interval [40,70]. The global tests are unable to reject the hypothesis of linearity for this dataset at the 5% significance level, even though the test that discards the highly variable boundary fits performs significantly better than the one that uses the full interval. In fact, the large increase in p -values observed for the local test on the interval [35,75] relative to the nar-

rower ones is likely also partly due to the increase in the variability of the boundary fits in the interval [70,75].

None of naive tests are able to reject the null hypothesis at the 5% significance level, and they display a wide range of p -values for the different intervals and bandwidths. In addition to overfitting in the smaller intervals, the poor performance is again likely due to the fact that each of these naive tests is affected by a large amount of boundary variability. As the size of the intervals increase, the behavior of the naive test and the local test are increasingly similar. This behavior is not surprising, since both the parametric and nonparametric fits in the naive test become increasingly similar to the global fits for the larger intervals.

In the preceding analysis, we applied the tests after first identifying a suspicious pattern in the data. Strictly speaking, the statistical hypothesis testing framework requires that an interval be identified prior to testing, since otherwise the significance levels are not correct. However, a low p -value even after such “data snooping” is still indicative of a possible interesting local pattern in the data, which can be followed up by a more in-depth analysis to confirm or reject the observed effect.

A more rigorous (though likely very conservative) alternative is to divide the range of the observations into intervals, perform the local test in each interval and then apply a Bonferroni correction to the significance levels. For instance, we could consider the interval [50,60] as one of 5 possible 10-year intervals (ignoring the remaining 5 years for simplicity). In that case, we adjust the confidence level required to reject the null hypothesis, or equivalently we

can multiply the p -values by 5 to maintain comparison with the global test. In that case, the “corrected p -values” for the local test would be 0.07, 0.035, 0.06 for $h = 5, 10, 15$ respectively, which still provide a strong indication of a deviation from linearity in the interval of interest. These adjusted p -values remain much smaller than the corresponding p -values for the global tests.

In this example, we investigated different bandwidths and different intervals, and computed p -values for $T_{n,L}$ for all these cases in order to evaluate the overall significance of the model departure. The Bonferroni correction can be used to guard against joint inference over different testing intervals, but only at the cost of a significant loss in power. Nevertheless, in practice, the combination of dividing up the interval into sub-intervals a priori for testing purposes and applying the Bonferroni correction would appear to be a reasonable and statistically valid approach for identifying local deviations from a parametric model.

It would clearly be of interest to have a fully adaptive version of the local test, in the sense of being able to adjust both to the degree of smoothness of the alternative (through the choice of h_n) and to the local extent of the deviation (by choosing D_n). The theoretical results we described in Section 2 are non-adaptive, and require both h_n and D_n to be non-random sequences. Adaptiveness with respect to the degree of smoothness could be achieved by extending the results of [13] to this situation. Possible directions to investigate adaptiveness with respect to the local intervals might be the development of a test statistic of the form $T_{n,A} = \sup_{D \in \mathcal{D}_n} T_{n,L}$, where \mathcal{D}_n is a set of intervals of a given size. A bootstrap distribution for $T_{n,A}$ under the null hypothesis

could in principle be constructed using wild bootstrap. We do not pursue these topics further here.

Acknowledgements The research of J. Opsomer was partly supported by Personal Service Contract B9J31191 with the U.S. Bureau of Labor Statistics. The research of both authors was partly supported by MEC Grant MTM2005-00429 (ERDF included). We thank an anonymous referee for helpful comments that substantially improved the presentation of the results.

A Proofs

Proof of Theorem 1: The proof is a direct generalization of that of Theorem 2.1 in [1]. Let $\mathbf{s}'_x = \mathbf{e}'_1 (\mathbf{X}'_x \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{X}'_x \mathbf{W}_x$. We decompose $\hat{m}(x; h_n) - m_0(x; \hat{\boldsymbol{\beta}})$ into $\hat{m}(x; h_n) - m_0(x; \hat{\boldsymbol{\beta}}) = \mathbf{s}'_x \boldsymbol{\varepsilon} + \mathbf{s}'_x \mathbf{g}_n + (m_0(x; \boldsymbol{\beta}) - m_0(x; \hat{\boldsymbol{\beta}})) = A(x) + B(x) + C(x)$, with $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)'$ and $\mathbf{g}_n = (g_n(x_1), \dots, g_n(x_n))'$, and decompose $T_{n,L}$ into terms involving integrals of products of these 3 functions.

From assumption A2, it follows immediately that $\int_{D_n} C(x)^2 w(x) dx = O_p(\|D_n\| n^{-1})$.

For $B(x)$,

$$\int_{D_n} B(x)^2 w(x) dx = c_n^2 \int_{D_n} \left((\tilde{K}_h * g)(x) \right)^2 w(x) dx (1 + o_p(1))$$

follows directly from the equivalent kernel representation of the smoother and assumption A4. The leading term in this expression is $O(c_n^2 \|D_n\|)$ and corresponds to b_{1n} .

Finally, following the same steps as in [1], the quantity $\int_{D_n} A(x)^2 w(x) dx$ can be written as

$$\begin{aligned} \int_{D_n} A(x)^2 w(x) dx &= \int_{D_n} \sum_i [\mathbf{s}_x]_i^2 \varepsilon_i^2 w(x) dx + \int_{D_n} \sum_{i \neq j} [\mathbf{s}_x]_i [\mathbf{s}_x]_j \varepsilon_i \varepsilon_j w(x) dx \\ &= A_1 + A_2, \end{aligned} \tag{5}$$

with

$$A_1 = \frac{1}{nh_n} \tilde{K}^{(2)}(0) \int_{D_n} \frac{\sigma^2(x) w(x)}{f(x)} dx (1 + o_p(1)),$$

and A_2 converges in distribution to a normally distributed random variable with mean 0 and variance

$$V_{Ln} = 2 \frac{1}{n^2 h_n} K^{(4)}(0) \int_{D_n} \left(\frac{\sigma(x)^2 w(x)}{f(x)} \right)^2 dx$$

The leading term in the approximation to A_1 is b_{0n} and of order $O(\|D_n\|/nh_n)$, and the asymptotic variance V_{Ln} is $O(\|D_n\|/n^2 h_n)$.

The cross-terms in $T_{n,L}$ resulting from the products of $A(x), B(x)$ and $C(x)$ are all of smaller order. \square

Proof of Theorem 2: Note first that

$$P_{\alpha,t} = \Phi \left(\frac{b_{1n}}{\sqrt{V_{tn}}} - z_{1-\alpha} \right) (1 + o(1)),$$

with $\Phi(\cdot)$ is the standard normal cumulative distribution function. To compute the ARE between two tests t and t' , we first set $P_{\alpha,t} = P_{\alpha,t'}$. By strict monotonicity and continuity of $\Phi(\cdot)$ and since b_{1n} does not depend on the test, the leading term in the ARE is obtained by equating the asymptotic variances and solving for the ratio of the sample sizes $n_t/n_{t'}$. In the case of n_L/n_G , the result is immediate. For the two ARE involving n_N , note that the asymptotic variance of $T_{n,N}^\dagger$ depends on n^* , the sample size in D_n . The ARE follow directly by applying assumption A5. \square

Proof of Theorem 3: The proof uses the method of imitation ([17, p.76]) applied to the proof of Theorem 1. Note first that by construction under H_0 , $b_{1n} = 0$ and $\text{Var}^*(\varepsilon_i^{*2}) = \hat{\varepsilon}_i^2 = \varepsilon_i^2 + O_p(n^{-1/2})$. The proof is again analogous to that of Theorem 2.1 in [1], except that when considering the terms in (5), the moments under the bootstrap model require an additional approximation step, as follows:

$$\begin{aligned} \mathbb{E}^*(A_1) &= \int_{D_n} \sum_i [\mathbf{s}_x]_i^2 \hat{\varepsilon}_i^2 w(x) dx \\ &= \int_{D_n} \mathbb{E}_{(\varepsilon,x)} \left(\sum_i [\mathbf{s}_x]_i^2 \varepsilon_i^2 \right) w(x) dx (1 + o_p(1)) \\ &= b_{0n} (1 + o_p(1)). \end{aligned}$$

The same reasoning applies to higher moments of A_1 , as well as to the derivation of the approximation for A_2 . \square

References

1. J. T. Alcalá, J. A. Cristobal, and W. González-Manteiga. Goodness-of-fit test for linear models based on local polynomials. *Statistics & Probability Letters*, 42:39–46, 1999.
2. A. Azzalini, A. W. Bowman, and W. Härdle. On the use of nonparametric regression for model checking. *Biometrika*, 76:1–12, 1989.
3. S. Bjerve, K. A. Doksum, and B. S. Yandell. Uniform confidence bounds for regression based on a simple moving average. *Scandinavian Journal of Statistics*, 12:159–169, 1985.
4. D. Cox, E. Koh, G. Wahba, and B. S. Yandell. Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. *The Annals of Statistics*, 16:113–119, 1988.
5. H. Dette. A consistent test for the functional form of a regression based on a difference of variance estimators. *The Annals of Statistics*, 27(3):1012–1040, 1999.
6. R. L. Eubank, C. S. Li, and S. Wang. Testing lack-of-fit of parametric regression models using nonparametric regression techniques. *Statistica Sinica*, 15(1):135–152, 2005.
7. R. L. Eubank and C. H. Spiegelman. Testing the goodness of fit of a linear model via nonparametric regression techniques. *Journal of the American Statistical Association*, 85:387–392, 1990.
8. J. Fan and I. Gijbels. *Local Polynomial Modelling and its Applications*. Chapman & Hall, London, 1996.
9. I. Gijbels and V. Rousson. A nonparametric least-squares test for checking a polynomial relationship. *Statistics & Probability Letters*, 51(3):253–261, 2001.
10. E. Guerre and P. Lavergne. Optimal minimax rates for nonparametric specification testing in regression models. *Econometric Theory*, 18(5):1139–1171, 2002.
11. W. Härdle and E. Mammen. Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, 21:1926–1947, 1993.

-
12. J.D. Hart. *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer Verlag, New York, 1997.
 13. J. L. Horowitz and V. G Spokoiny. An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica*, 69(3):599–631, 2001.
 14. R. W. Johnson. Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education*, 4, 1996.
 15. G. E. Noether. On a theorem of Pitman. *Annals of Mathematical Statistics*, 26:64–68, 1955.
 16. K. Penrose, A. Nelson, and A. A. Fisher. Generalized body composition prediction equation for men using simple measurement techniques (abstract). *Medicine and Science in Sports and Exercise*, 17:189, 1985.
 17. J. Shao and D. Tu. *The Jackknife and Bootstrap*. Springer, New York, 1995.
 18. V. G. Spokoiny. Adaptive hypothesis testing using wavelets. *Annals of Statistics*, 24(6):2477–2498, 1996.
 19. J. G. Staniswalis and T. A. Severini. Diagnostics for assessing regression models. *Journal of the American Statistical Association*, 86:684–692, 1991.
 20. W. Stute and W. González-Manteiga. NN goodness-of-fit tests for linear models. *Journal of Statistical Planning and Inference*, 53:75–92, 1996.
 21. R. Tibshirani and T. Hastie. Local likelihood estimation. *Journal of the American Statistical Association*, 82:559–567, 1987.
 22. G. Weirather. Testing a linear regression model against nonparametric alternatives. *Metrika*, 40:367–379, 1993.
 23. J. H. Zheng. A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics*, 75:263–289, 1996.