Detecting heteroskedasticity in nonparametric regression using weighted empirical processes

Justin Chown[1] and Ursula U. Müller[2]

Abstract. Heteroskedastic errors can lead to inaccurate statistical conclusions if they are not properly handled. We introduce a test for heteroskedasticity for the nonparametric regression model with multiple covariates. It is based on a suitable residual-based empirical distribution function. The residuals are constructed using local polynomial smoothing. Our test statistic involves a "detection function" that can verify heteroskedasticity by exploiting just the independence-dependence structure between the detection function and model errors, i.e. we do not require a specific model of the variance function. The procedure is asymptotically distribution free: inferences made from it do not depend on unknown parameters. It is consistent at the parametric (root-n) rate of convergence. Our results are extended to the case of missing responses and illustrated with simulations.

Keywords: heteroskedastic nonparametric regression, local polynomial smoother, missing at random, transfer principle, weighted empirical process

2010 AMS Subject Classifications: Primary: 62G08, 62G10; Secondary: 62G20, 62G30.

1. Introduction

When analysing data, it is common practice to first explore available options using various plotting techniques. For regression models, a key tool is to construct a plot of the model residuals in absolute value against fitted values. If there is only one covariate, we can use a plot of the residuals in absolute value against that covariate. This technique helps determine if theoretical requirements for certain statistical procedures are satisfied, in particular whether or not the variation in the errors remain constant across values of the covariate. This is an important assumption that we want to examine more closely. We will therefore consider the model with constant error variance σ_0^2, the *homoskedastic model*, which we can write in the form

$$Y = r(X) + \varepsilon, \quad \varepsilon = \sigma_0 e.$$

The function r is the regression function and σ_0 a positive constant. We consider a response variable Y, a covariate *vector* X and assume that X and the random variable e are independent, where e has mean equal to zero and variance equal to one.

Corresponding author: Justin Chown (justin.chown@ruhr-uni-bochum.de)
[1] *Ruhr-Universität Bochum, Fakultät für Mathematik, Lehrstuhl für Stochastik, 44780 Bochum, Germany.*
[2] *Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA.*

When the variation in the data is not constant across the covariate values, the *heteroskedastic model* is adequate:

$$Y = r(X) + \varepsilon, \quad \varepsilon = \sigma(X)e. \tag{1.1}$$

Here $\sigma(\cdot)$ is a scale function with $E[\sigma^2(X)] = \sigma_0^2$. Model (1.1) contains the homoskedastic regression model as a (degenerate) special case with $\sigma \equiv \sigma_0$, a constant function. In order to discriminate between both models we assume that $\sigma(\cdot)$ is non-constant in the heteroskedastic case, i.e. it varies with the values of the covariates X.

Testing for heteroskedasticity is of great importance: many procedures lead to inconsistent and inaccurate results if the heteroskedastic model is appropriate but not properly handled. Consider model (1.1) with a parametric regression function, e.g. linear regression with $r(X) = \vartheta^\top X$. The ordinary least squares estimator $\hat{\vartheta}$ of the parameter vector ϑ, which is constructed for the homoskedastic model, will still be consistent under heteroskedasticity. However it will be less accurate than an estimator that puts more weight on observations (X, Y) with small variance $\sigma^2(X)$ (and less weight when the variance is large). The estimated variance of $\hat{\vartheta}$ will be biased if the model is in fact heteroskedastic, so testing hypotheses based on $\hat{\vartheta}$ may lead to invalid conclusions.

The relationship between the homoskedastic and heteroskedastic models can be expressed in terms of statistical hypotheses:

$$H_0 : \exists\ \sigma_0 > 0,\ \sigma(\cdot) = \sigma_0 \quad a.e.\ (G) \qquad \text{vs.} \qquad H_a :\ \sigma(\cdot) \in \Sigma.$$

Here G is the distribution function of the covariates X and $\Sigma = \{\sigma \in L_2(G) : \sigma(\cdot) > 0 \text{ and non-constant } a.e.(G)\}$ is a space of scale functions. The null hypothesis corresponds to the homoskedastic model and the alternative hypothesis to the heteroskedastic model.

The tests introduced in this article are inspired by Stute, Xu and Zhu (2008), who propose tests for a parametric regression model with high-dimensional covariates against nonparametric alternatives, and by Koul, Müller and Schick (2012), who develop tests for linearity of a semiparametric regression function for fully observed data and for a missing data model. These approaches are in the spirit of Stute (1997), who introduces marked empirical processes to test parametric models for the regression function in nonparametric regression with univariate covariates.

Our test statistics are based on weighted empirical distribution functions of residuals. The form of these statistics is strikingly simple and their associated limiting behaviour is obtained by considering the related weighted empirical process. We will show that our test statistic converges with root-n rate to a Brownian bridge. Hence it is asymptotically distribution free and quantiles are available. We consider detecting heteroskedasticity (represented by the non-constant scale function $\sigma(\cdot)$) by using some (non-constant) "detection function" $\omega(\cdot)$ in the space Σ. To explain the idea, we consider the weighted error distribution function

$$E\big[\omega(X)\mathbf{1}[\varepsilon \le t]\big] = E\big[\omega(X)\mathbf{1}[\sigma(X)e \le t]\big], \qquad t \in \mathbb{R}.$$

If the null hypothesis is true $\varepsilon = \sigma_0 e$ and we can write

$$E\big[\omega(X)\mathbf{1}[\varepsilon \le t]\big] = E\big[\omega(X)\mathbf{1}[\sigma_0 e \le t]\big] = E\big[E\big[\omega(X)\big]\mathbf{1}[\sigma_0 e \le t]\big] = E\big[E\big[\omega(X)\big]\mathbf{1}[\varepsilon \le t]\big],$$

$t \in \mathbb{R}$. Here we have used that under the null hypothesis the covariates X and the errors $\varepsilon = \sigma_0 e$ are independent. This motivates a test based on the difference between the two

quantities, i.e. on

$$E\big[\{\omega(X) - E\big[\omega(X)\big]\}\mathbf{1}[\varepsilon \le t]\big], \qquad t \in \mathbb{R}, \tag{1.2}$$

which is zero under H_0, but typically not under H_a (see Remark 1 in Section 2 for further details). We can estimate the outer expectation by its empirical version, which yields a test based on

$$U_n(t) = n^{-1/2} \sum_{j=1}^{n} \Big\{\omega(X_j) - E\big[\omega(X_j)\big]\Big\}\mathbf{1}[\varepsilon_j \le t], \qquad t \in \mathbb{R}.$$

This is a process in the Skorohod space $D(-\infty, \infty)$. To move this process to the more convenient space $D[-\infty, \infty]$, we define the familiar limit $U_n(-\infty) = 0$ and the limit

$$U_n(\infty) = n^{-1/2} \sum_{j=1}^{n} \Big\{\omega(X_j) - E\big[\omega(X_j)\big]\Big\}.$$

Since the variance of $U_n(\infty)$ equals the variance of $\omega(X)$ it is clear the asymptotic distribution of $\sup_{t\in\mathbb{R}} |U_n(t)|$ will depend on $\mathrm{Var}\{\omega(X)\}$, which is not desirable for obtaining a standard distribution useful for statistical inference. We therefore standardise $U_n(t)$ and obtain the weighted empirical process

$$S_n(t) = n^{-1/2} \sum_{j=1}^{n} W_j \mathbf{1}[\varepsilon_j \le t], \qquad t \in \mathbb{R},$$

with weights

$$W_j = \frac{\omega(X_j) - E\big[\omega(X_j)\big]}{\mathrm{Var}[\omega(X_j)]^{1/2}} = \frac{\omega(X_j) - E\big[\omega(X_j)\big]}{E\big[\{\omega(X_j) - E\big[\omega(X_j)\big]\}^2\big]^{1/2}}, \qquad j = 1, \ldots, n. \tag{1.3}$$

The weights W_j are centred to guarantee that the tests are asymptotically distribution free. Related research on (unweighted) residual-based empirical distribution functions typically provides uniform expansions involving a non-negligible stochastic drift parameter that includes the error density as a parameter of the underlying distribution, i.e. the statistics are not distribution free (see e.g. Akritas and Van Keilegom, 2001; Müller, Schick and Wefelmeyer, 2007, 2009). This is in contrast to our case where the mean zero weights ensure that no drift emerges (see the discussion before Lemma 2 in Section 5 for more details).

The process S_n cannot be used for testing because it depends on unknown quantities. Our final test statistic T_n will therefore be based on an estimated version of S_n with the errors estimated by residuals $\hat{\varepsilon}_j = Y_j - \hat{r}(X_j)$, $j = 1, \ldots, n$, from a sample of n i.i.d. random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$. Here $\hat{r}$ is a suitable estimator of the regression function. In this article we assume a nonparametric regression model and estimate the unknown smooth regression function r using a nonparametric function estimator; see Section 2 for details.

When $\sigma(\cdot) \equiv \sigma_0$ is a constant function (the null hypothesis is true), we expect the estimated process to behave like $S_n(t)$ and exhibit a standard limiting behaviour. However, if $\sigma(\cdot)$ is non-constant (the alternative hypothesis is true), the residuals $\hat{\varepsilon}_j$ will estimate $\varepsilon_j = \sigma(X_j)e_j \neq \sigma_0 e_j$ (and the weights W_j and the errors $\varepsilon_j = \sigma(X_j)e_j$ will not be independent). We expect the estimated process will show a different limiting behaviour in this case. Note that our test exploits just the independence–dependence structure between the covariates and the errors. For this reason it is also clear that it will only work in our model, which

specifically assumes $\varepsilon = \sigma(X)e$, to test H_0 "Var$[Y|X]$ is constant", and not in models $Y = r(X) + \varepsilon$, where ε may depend on X in a more general way.

The choice of the weights, i.e. of the detection function ω, is important to guarantee that the tests are powerful: it is clear that ω must be non-constant to detect heteroskedasticity. If the alternative hypothesis is true, it will be advantageous to have weights that are highly correlated with the scale function σ to increase the power of the test. We explain this at the end of Section 2, where we also construct weights based on an estimate $\hat{\sigma}(\cdot)$ of $\sigma(\cdot)$.

Tests for heteroskedasticity are well studied for various regression models. Glejser (1969) forms a test using the absolute values of the residuals of a linear regression fitted by ordinary least-squares. White (1980) constructs an estimator of the covariance matrix of the ordinary least-squares estimator in linear regression and proposes a test based on this estimator. Cook and Weisberg (1983) derive a score test for a parametric form of the scale function of the errors in a linear regression. Eubank and Thomas (1993) study a test for heteroskedasticity, which is related to the score test, for the nonparametric regression model with normal errors. Although the last article studies a nonparametric regression function, all these tests are parametric tests since the heteroskedacticity is modelled parametrically. It is therefore possible that the tests have no power if those models are misspecified.

More recent papers that test for a parametric scale function σ_θ are Dette, Neumeyer and Van Keilegom (2007) and Dette and Hetzler (2009). Dette et al. (2007) construct a test for nonparametric regression with univariate covariates based on the Kolomogorov–Smirnov and the Cramer–von–Mises statistic. They recommend a bootstrap approach to implement the tests. Dette and Hetzler (2009) construct a test for a (univariate) fixed design nonparametric regression model using an empirical process. The tests proposed in both papers converge at the root-n rate.

Our approach is completely nonparametric (and thus more robust than parametric approaches). Let us therefore look more closely at some competing nonparametric approaches. Dette and Munk (1998) and Dette (2002) both consider nonparametric regression with fixed design (i.e. random univariate or multivariate covariates are not treated). Dette and Munk create tests based on the L_2 distance between an approximation of the variance function and σ_0; in the 2002 paper Dette proposes a residual-based test using kernel estimators. In both papers the test statistics are asymptotically normal with variance to be estimated. The convergence rate is slower than the parametric root-n rate (which is the rate of our test), and local alternatives of order $n^{-1/4}$ (Dette and Munk, 1998) or of order $n^{-1/2}h_n^{-1/4}$ (with bandwidth $h_n \to 0$) can be detected (Dette, 2002).

The approach in Dette (2002) is extended to the case of a partially linear regression by You and Chen (2005) and Lin and Qu (2012). The same idea is also used in Zheng (2009), who proposes a local smoothing test for nonparametric regression, now with multivariate covariates, which is also our scenario. The test statistic is again asymptotically normally distributed and requires a consistent estimator of the variance. The test is able to detect local alternatives of order $n^{-1/2}h_n^{-m/4}$, where m is the dimension of the covariate vector, which is in agreement with the order in Dette (2002) for the case of a univariate X, $m = 1$. The test statistic is quite involved, using multivariate Nadaraya–Watson estimators. A wild bootstrap approach is used to implement the test since the normal approximation "may not be accurate (...) in finite samples" (Zheng, 2009, Section 5). Zheng's approach was used by Zhu et al. (2016). These authors use single-index type models (involving linear combinations

of covariates) for the mean and the variance functions in order to handle high-dimensional covariates. Again, these tests converge with a rate slower than root-n and can only detect local alternatives of order $n^{-1/2}h_n^{-q/4}$, where now q is the number of linear combinations.

Let us finally have a closer look at two articles which, like our paper, use weighted empirical processes, namely Zhu et al. (2001) and Koul and Song (2010). Zhu et al. use a Cramer–von–Mises type statistic of a marked empirical process of (multivariate) *covariates* X instead of univariate *residuals* $\hat{\varepsilon}$, i.e. involving $\mathbf{1}(X \leq x)$ instead of $\mathbf{1}(\hat{\varepsilon} \leq t)$. Their test can detect local alternatives "up to" order $n^{-1/2}$ and, for the case of univariate covariates, may be asymptotically distribution free ("under some condition"). The authors point directly to the bootstrap to obtain suitable quantiles since the limiting distribution of their statistic has a complicated covariance structure. However, the approach using marked empirical processes of *covariates* seems to be problematic when the covariate X is multivariate because of possible dependencies between the components of X, as pointed out in Stute et al. (2008, p. 454). Koul and Song (2010) check parametric models for the variance function and therefore are not directly comparable with our approach, but their tests have some desirable properties as well: they are distribution free (converging to a Brownian motion), and able to detect alternatives of order $n^{-1/2}$. Koul and Song's tests are based on a Khmaladze type transform of a marked empirical process of univariate *covariates*. Therefore a generalisation to the case of a multivariate X does not seem to be advisable.

Summing up, our approach is new in that we are the first to use a completely nonparametric approach based on weighted "marked" empirical process of (univariate) residuals to test for heteroskedasticity. Our tests achieve the parametric rate root-n, which, so far, could only be achieved in Zhu et al. (2001), or if the tests involve some parametric component, e.g. a parametric model for the variance function. Another advantage of our method is that the proposed tests are asymptotically distribution free (and quantiles are readily available), while most of the competing approaches require bootstrap to implement the test.

In this article we are also interested in the case when responses Y are missing at random (MAR), which we call the "MAR model", in order to distinguish it from the "full model", when all data are completely observed. In the MAR model the observations can be written as independent copies $(X_1, \delta_1 Y_1, \delta_1), \ldots, (X_n, \delta_n Y_n, \delta_n)$ of a base observation $(X, \delta Y, \delta)$, where δ is an indicator which equals one if Y is observed and zero otherwise. Assuming that responses are *missing at random* means the distribution of δ given the pair (X, Y) depends only on the covariates X (which are always observed), i.e.

$$P(\delta = 1|X, Y) = P(\delta = 1|X) = \pi(X).$$

This implies that Y and δ are conditionally independent given X. Assuming that responses are missing at random is often reasonable; see Little and Rubin (2002, Chapter 1). Working with this missing data model is advantageous because the missingness mechanism is ignorable, i.e. $\pi(\cdot)$ can be estimated. It is therefore possible to draw valid statistical conclusions without auxiliary information, in contrast to the model with data that are "not missing at random" (NMAR). Note how the MAR model covers the full model as a special case with all indicators δ equal to 1, hence $\pi(\cdot) \equiv 1$.

We will show that our test statistics T_n, defined in (2.2) for the full model, and $T_{n,c}$, defined in (3.1) for the MAR model, may be used to test for the presence of heteroskedasticity. The subscript "c" indicates that our test statistic $T_{n,c}$ uses only the completely observed

data; i.e. we use only observations (X, Y) where δ equals one, called the *complete cases*. In particular, we use only the available residuals $\hat{\varepsilon}_{j,c} = Y_j - \hat{r}_c(X_j)$, where $\hat{r}_c$ is a suitable complete case estimator of the regression function r. Demonstrating this will require two steps. First, we study the full model and provide the limiting distribution of the test statistic T_n under the null hypothesis in Theorem 1. Then we apply the *transfer principle* for complete case statistics (Koul et al., 2012) to adapt the results of Theorem 1 to the MAR model.

Since residuals can only be computed for data (X, Y) that are completely observed, the transfer principle is useful for developing residual-based statistical procedures in MAR regression models. Our proposed (residual-based) tests are asymptotically distribution free. The transfer principle guarantees, in this case, that the test statistic and its complete case version have the same limiting distribution (under a mild condition). This means that one can simply omit the incomplete cases and work with the same quantiles as in the full model, which is desirable due to its simplicity.

This article is structured as follows. Section 2 contains the statement of the test statistic and the asymptotic results for the full model. Section 3 extends the results of the full model to the MAR model. Simulations in Section 4 investigate the performance of these tests. Technical arguments supporting the results in Section 2 are given in Section 5. Section 6 concludes the article with a discussion of possible extensions of the proposed methodology.

2. Completely observed data

We begin with the full model and require the following standard condition (which guarantees good performance of nonparametric function estimators):

ASSUMPTION 1. The covariate vector X is quasi-uniform on the cube $[0,1]^m$; i.e. X has a density that is bounded and bounded away from zero on $[0,1]^m$.

As in Müller, Schick and Wefelmeyer (2009), we require the regression function to be in the Hölder space $H(d, \gamma)$, i.e. it has continuous partial derivatives of order d (or higher) and the partial d-th derivatives are Hölder with exponent $\gamma \in (0, 1]$. We estimate the regression function r by a local polynomial smoother $\hat{r}$ of degree d. The choice of d will not only depend on the number of derivatives of r, but also on the dimension m of the covariate vector. (We will need more smoothness if m is large.) We write F and f for the distribution function and the density of the errors $\sigma_0 e$ which will have to satisfy certain smoothness and moment conditions.

In order to describe the local polynomial smoother, let $i = (i_1, \ldots, i_m)$ be a multi-index and $I(d)$ be the set of multi-indices that satisfy $i_1 + \ldots + i_m \le d$. Then $\hat{r}$ is defined as the component $\hat{\beta}_0$ corresponding to the multi-index $0 = (0, \ldots, 0)$ of a minimiser

$$\hat{\beta} = \operatorname*{arg\,min}_{\beta=(\beta_i)_{i\in I(d)}} \sum_{j=1}^{n} \Big\{ Y_j - \sum_{i\in I(d)} \beta_i \psi_i\Big(\frac{X_j - x}{c_n}\Big) \Big\}^2 w\Big(\frac{X_j - x}{c_n}\Big), \tag{2.1}$$

where

$$\psi_i(x) = \frac{x_1^{i_1}}{i_1!} \cdots \frac{x_m^{i_m}}{i_m!}, \qquad x = (x_1, \ldots, x_m) \in [0,1]^m,$$

$w(x) = w_1(x_1) \cdots w_m(x_m)$ is a product of probability densities with compact support, and c_n is a bandwidth. A typical choice for w_i would be the Epanechnikov or the tricube kernel.

The estimator $\hat{r}$ was studied in Müller et al. (2009), who provide a uniform expansion of an empirical distribution function based on residuals

$$\hat{\varepsilon}_j = Y_j - \hat{r}(X_j), \quad j = 1, \ldots, n.$$

The proof uses results from a crucial technical lemma, Lemma 1 in that article (written here as Lemma 1 in Section 5), which gives important asymptotic properties of $\hat{r}$. We will use these properties in Section 5 to derive the limiting distribution of our test statistic, which is based on a weighted version of the empirical distribution function proposed by Müller et al. (2009).

For the full model, the test statistic is given as

$$T_n = \sup_{t \in \mathbb{R}} \Big| n^{-1/2} \sum_{j=1}^{n} \hat{W}_j \mathbf{1}\big[\hat{\varepsilon}_j \le t\big] \Big| \tag{2.2}$$

with

$$\hat{W}_j = \Big\{ \omega(X_j) - \frac{1}{n} \sum_{k=1}^{n} \omega(X_k) \Big\} \Big/ \Big[\frac{1}{n} \sum_{m=1}^{n} \Big\{ \omega(X_m) - \frac{1}{n} \sum_{k=1}^{n} \omega(X_k) \Big\}^2 \Big]^{1/2}, \quad \omega \in \Sigma, \tag{2.3}$$

for $j = 1, \ldots, n$. The term in absolute brackets of (2.2) is an approximation (under H_0) of the process $S_n(t)$ from the Introduction, now with the standardised weights W_j from (1.3) replaced by empirically estimated weights $\hat{W}_j$. Recall that ω must be a non-constant function, i.e. $\omega \in \Sigma$, which is crucial to guarantee that the test is able to detect heteroskedasticity.

We arrive at our main result, the limiting distribution for the test statistic T_n in the fully observed model. The proof is given in Section 5.

THEOREM 1. *Let the distribution G of the covariates X satisfy Assumption 1. Suppose the regression function r belongs to the Hölder space $H(d, \gamma)$ with $s = d + \gamma > 3m/2$; the distribution F of the error variable $\sigma_0 e$ has mean zero, a finite moment of order $\zeta > 4s/(2s - m)$ and a Lebesgue density f that is both uniformly continuous and bounded; the kernel functions $w_1, \ldots, w_m$ used in the local polynomial smoother (2.1) are $(m+2)-$times continuously differentiable and have compact support $[-1, 1]$. Let $c_n \sim \{n \log(n)\}^{-1/(2s)}$. Let the null hypothesis hold. Then*

$$T_n = \sup_{t \in \mathbb{R}} \Big| n^{-1/2} \sum_{j=1}^{n} \hat{W}_j \mathbf{1}\big[\hat{\varepsilon}_j \le t\big] \Big|$$

with $\hat{W}_j$ specified in (2.3) above, converges in distribution to $\sup_{t \in [0,1]} |B_0(t)|$, where B_0 denotes the standard Brownian bridge.

The distribution of $\sup_{t \in [0,1]} |B_0(t)|$ is a standard distribution, whose upper α-quantiles b_α can be approximately calculated using formula (12) on page 34 of Shorack and Wellner (1986), i.e.

$$P\Big(\sup_{t \in [0,1]} \big|B_0(t)\big| \le b \Big) = \sqrt{\frac{2\pi}{b}} \sum_{k=1}^{\infty} \exp\Big(- \frac{(2k-1)^2 \pi^2}{8 b^2} \Big), \qquad b > 0.$$

We calculate that $b_{0.05} = 1.1779$ is an appropriate quantile for a 5% level test.

Remark 1 (power under fixed alternatives). It is possible that the test has no power if the detection function ω is not properly chosen. To see this consider the difference

$$E\big[\{\omega(X) - E\big[\omega(X)\big]\}\mathbf{1}[\varepsilon \le t]\big]$$

from equation (1.2) in the introduction, which is zero under H_0. The test has no power if the difference is also zero under H_a, i.e. if

$$E\big[\omega(X)\mathbf{1}[\varepsilon \le t]\big] = E\Big[\omega(X)F\Big(\frac{t}{\sigma(X)}\Big)\Big] = E\{\omega(X)\}E\Big[F\Big(\frac{t}{\sigma(X)}\Big)\Big] = E\{\omega(X)\}E\big[\mathbf{1}[\varepsilon \le t]\big],$$

which means that $\omega(X)$ and $F(t/\sigma(X))$ are uncorrelated. This happens, for example, if X and ε are both uniformly distributed on $[0,1]$ and if $\sigma(X) = [1+\sin(2\pi X)]^{-1}$ and $\omega(X) = 1+\cos(2\pi X)$. Then $E[\omega(X)F(t/\sigma(X))] = tE[\omega(X)/\sigma(X)]$. It is easy to check that this indeed equals $tE[\omega(X)]E[1/\sigma(X)]$ since $E[\omega(X)/\sigma(X)] = E[\omega(X)] = E[1/\sigma(X)] = 1$.

Remark 2 (power under local alternatives). To derive the power of the test under local alternatives of the form $\sigma = \sigma_{n\Delta} = \sigma_0 + n^{-1/2}\Delta$ with $\sigma_{n\Delta} \in \Sigma$, we use Le Cam's third lemma. This result states that a local shift Δ away from the null hypothesis results in an additive shift of the asymptotic distribution of T_n; see e.g. page 90 of van der Vaart (1998). The shift is calculated as the covariance between T_n and $\log(dF_{n\Delta}/dF)$ under H_0 where $\varepsilon = \sigma_0 e$. Here $F_{n\Delta}(t) = P(\{\sigma_0 + n^{-1/2}\Delta(X)\}e \le t \,|\, X)$. A brief sketch shows $E[T_n \log(dF_{n\Delta}/dF)]$ is equal to

$$\begin{aligned}
&E\Big[\Big\{n^{-1/2}\sum_{j=1}^{n} W_j\mathbf{1}\big[\sigma_0 e_j \le t\big]\Big\}\Big\{n^{-1/2}\sum_{j=1}^{n}\Delta(X_j) + n^{-1/2}\sum_{j=1}^{n}\Delta(X_j)\sigma_0 e_j\frac{f'(\sigma_0 e_j)}{f(\sigma_0 e_j)}\Big\}\Big] + o(1)\\
&= E[W\Delta]F(t) + E[W\Delta]\int_{-\infty}^{t} s\,\frac{f'(s)}{f(s)}\,F(ds) + o(1)\\
&= t\,f(t)E[W\Delta] + o(1).
\end{aligned}$$

Here we have, for simplicity, assumed that F is differentiable with finite Fisher information for location and scale. Hence, under a contiguous alternative H_a, the distribution of the test statistic T_n limits to $\sup_{t\in[0,1]}|B_0(t) + F^{-1}(t)\{f\circ F^{-1}(t)\}E[W\Delta]|$, writing F^{-1} for the quantile function of F.

Since the weights in our test statistic are standardised, only the shape of ω may have an effect on the behaviour of the statistic – location and scale have no influence. From Remark 2 we know that the test has no power under local alternatives if $E[W\Delta] = 0$, i.e. by definition of W, if the detection function $\omega(X)$ and the scale function $\sigma(X) = \sigma_0 + n^{-1/2}\Delta(X)$ are uncorrelated. This happens, for example, if X has a standard uniform distribution, if $\Delta(X) = 1+\sin(2\pi X)$ and if we choose $\omega(X) = 1+\cos(2\pi X)$; cf. Remark 1.

From Remark 2 it is also clear that the power of the test increases with $E(W\Delta)$, i.e. if $\omega(X)$ and $\Delta(X)$ are highly correlated. So it can be expected that the test will perform best when ω is a linear transformation of the scale function σ. This suggests choosing ω similar in shape to σ, in order to obtain a powerful test. We propose using $\omega = \hat{\sigma}$, where $\hat{\sigma}$ is a consistent estimator of σ. Assume for simplicity that the regression function r and the second conditional moment $r_2(\cdot) = E[Y^2 \,|\, X = (\cdot)]$ of Y given X belong to the same Hölder class $H(d,\gamma)$ with $s = d+\gamma$. Then r_2 can be estimated by a local polynomial smoother $\hat{r}_2$,

which is defined as $\hat{r}$ in (2.1) but now with Y_j^2 in place of Y_j, $j = 1, \ldots, n$. This leads to an estimator of σ using $\hat{\sigma} = \{\hat{r}_2 - \hat{r}^2\}^{1/2}$. The estimated weights are then given by

$$\tilde{W}_j = \Big\{\hat{\sigma}(X_j) - \frac{1}{n}\sum_{k=1}^{n}\hat{\sigma}(X_k)\Big\}\Big/\Big[\frac{1}{n}\sum_{m=1}^{n}\Big\{\hat{\sigma}(X_m) - \frac{1}{n}\sum_{k=1}^{n}\hat{\sigma}(X_k)\Big\}^2\Big]^{1/2} \tag{2.4}$$

for $j = 1, \ldots, n$.

The formal result for this choice of weights is given in Theorem 2 below.

Note that the weights in (2.4) are non-degenerate under the null hypothesis because the terms in the numerator and in the denominator have the same asymptotic order. Since the statistic converges weakly to a Gaussian process, which is determined by its mean and covariance functions, and since our weights are scaled and centred, they do not affect the asymptotic distribution.

THEOREM 2. *Suppose the assumptions of Theorem 1 are satisfied with the error variable $\sigma_0 e$ having a finite moment of order larger than 8. Assume that r and r_2 belong to the Hölder space $H(d, \gamma)$ with $s = d + \gamma > 3m/2$. Then under the null hypothesis*

$$\tilde{T}_n = \sup_{t\in\mathbb{R}}\Big|n^{-1/2}\sum_{j=1}^{n}\tilde{W}_j\mathbf{1}\big[\hat{\varepsilon}_j \le t\big]\Big|$$

with $\tilde{W}_j$ specified in (2.4) above, converges in distribution to $\sup_{t\in\mathbb{R}}|B_0(t)|$, where B_0 denotes the standard Brownian bridge. In addition, $\tilde{T}_n$ consistently detects alternative hypotheses $\sigma \in H(d, \gamma)$, and $\tilde{T}_n$ is asymptotically most powerful for detecting local alternative hypotheses of the form $\sigma_0 + n^{-1/2}\Delta(\cdot)$, where $\Delta \in H(d, \gamma)$.

The first part of Theorem 2 is verified in the supplementary online materials. The last statement concerning the power of the test follows from Remark 2 and the discussion afterwards, combined with the fact that we can consistently estimate the scale provided $\Delta \in H(d, \gamma)$.

3. Responses missing at random

We now consider the MAR model. The complete case test statistic is given by

$$T_{n,c} = \sup_{t\in\mathbb{R}}\Big|N^{-1/2}\sum_{j=1}^{n}\delta_j\hat{W}_{j,c}\mathbf{1}\big[\hat{\varepsilon}_{j,c} \le t\big]\Big|, \quad \text{with } \hat{\varepsilon}_{j,c} = Y_j - \hat{r}_c(X_j). \tag{3.1}$$

Here $N = \sum_{j=1}^{n}\delta_j$ is the number of complete cases and $\hat{W}_{j,c}$ denotes the weights from equation (2.3) in the previous section, which are now constructed using only the complete cases. The estimator $\hat{r}_c$ is the complete case version of $\hat{r}$; i.e. the component $\hat{\beta}_{c,0}$ corresponding to the multi-index $0 = (0, \ldots, 0)$ of a minimiser

$$\hat{\beta}_c = \operatorname*{arg\,min}_{\beta=(\beta_i)_{i\in I(d)}}\sum_{j=1}^{n}\delta_j\Big\{Y_j - \sum_{i\in I(d)}\beta_i\psi_i\Big(\frac{X_j - x}{c_n}\Big)\Big\}^2 w\Big(\frac{X_j - x}{c_n}\Big),$$

which is defined as in the previous section, but now also involves the indicator δ_j.

The transfer principle for complete case statistics (Koul et al., 2012) states that if the limiting distribution of a statistic in the full model is $\mathcal{L}(Q)$, with Q the joint distribution of (X, Y), then the distribution of its complete case version in the MAR model will be

$\mathcal{L}(Q_1)$, where Q_1 is the conditional distribution of (X, Y) given $\delta = 1$. The implication holds provided Q_1 satisfies the same model assumptions as Q. For our problem this means that Q_1 must meet the assumptions imposed on Q by Theorem 1. It is easy to see how this affects only the covariate distribution G. Due to the independence of X and e, the distribution Q of (X, Y) factors into the marginal distribution G of X and the conditional distribution of Y given X, i.e. the distribution F of the errors $\sigma_0 e$. This means we can write $Q = G \otimes F$. The MAR assumption implies that e and δ are independent. Hence the distribution F of the errors remains unaffected when we move from Q to the conditional distribution Q_1 given $\delta = 1$, and we have $Q_1 = G_1 \otimes F$, where G_1 is the distribution of X given $\delta = 1$. Thus, Assumption 1 about G must be restated; we also have to assume the detection function ω is square-integrable with respect to G_1.

ASSUMPTION 2. The conditional distribution G_1 of the covariate vector X given $\delta = 1$ is quasi-uniform on the cube $[0, 1]^m$; i.e. it has a density that is bounded and bounded away from zero on $[0, 1]^m$.

The limiting distribution $\mathcal{L}(Q)$ of the test statistic in the full model in Theorem 1 is given by $\sup_{t\in[0,1]} |B_0(t)|$. Hence it does *not* depend on the joint distribution Q of (X, Y) (or on unknown parameters). This makes the test particularly interesting for the MAR model, since the limiting distribution of the complete case statistic $\mathcal{L}(Q_1)$ is the same as the distribution of the full model statistic, $\mathcal{L}(Q_1) = \mathcal{L}(Q)$, i.e. it is also given by $\sup_{t\in[0,1]} |B_0(t)|$. Combining these arguments already provides proof for the main result for the MAR model.

THEOREM 3. *Let the null hypothesis hold. Suppose the assumptions of Theorem 1 are satisfied, with Assumption 2 in place of Assumption 1, and let $\omega \in L_2(G_1)$ be positive and non-constant. Write*

$$\hat{W}_{j,c} = \Big\{\delta_j\omega(X_j) - \frac{1}{N}\sum_{k=1}^{n}\delta_k\omega(X_k)\Big\}\Big/\Big[\frac{1}{N}\sum_{m=1}^{n}\Big\{\delta_m\omega(X_m) - \frac{1}{N}\sum_{k=1}^{n}\delta_k\omega(X_k)\Big\}^2\Big]^{1/2}$$

and $\hat{\varepsilon}_{j,c} = Y_j - \hat{r}_c(X_j)$. Then

$$T_{n,c} = \sup_{t\in\mathbb{R}}\Big|N^{-1/2}\sum_{j=1}^{n}\delta_j\hat{W}_{j,c}\mathbf{1}\big[\hat{\varepsilon}_{j,c} \le t\big]\Big|$$

converges in distribution to $\sup_{t\in[0,1]} |B_0(t)|$, where B_0 denotes the standard Brownian bridge.

This result is very useful: if the assumptions of the MAR model are satisfied it allows us to simply delete the incomplete cases and implement the test for the full model; i.e. we may use the same quantiles.

REMARK 3. Following the discussions above and those preceding Theorem 2 in the previous section we can construct estimated weights based on complete cases as follows. The first and second conditional moments of Y given X can be estimated by complete case versions $\hat{r}_c$ and $\hat{r}_{2,c}$ of the local polynomial smoothers $\hat{r}$ and $\hat{r}_2$. Hence, $\hat{\sigma}_c(\cdot) = \{\hat{r}_{2,c}(\cdot) - \hat{r}_c^2(\cdot)\}^{1/2}$ is a consistent complete case estimator of $\omega(\cdot) = \sigma(\cdot)$ (which optimises the power of the test). The complete case version of the test statistic $\tilde{T}_n$ is

$$\tilde{T}_{n,c} = \sup_{t\in\mathbb{R}}\Big|N^{-1/2}\sum_{j=1}^{n}\delta_j\tilde{W}_{j,c}\mathbf{1}\big[\hat{\varepsilon}_{j,c} \le t\big]\Big|,$$

where the weights $\tilde{W}_{j,c}$ are complete case versions of $\tilde{W}_j$; see (2.4). The transfer principle then implies that the results of Theorem 2 continue to hold for $\tilde{T}_{n,c}$, i.e. $\tilde{T}_{n,c}$ tends under the null hypothesis in distribution to $\sup_{t\in[0,1]} |B_0(t)|$ and is asymptotically most powerful for detecting smooth local alternatives.

4. Simulation results

A brief simulation study demonstrates the effectiveness of a hypothesis test using the test statistics given above for the full model and the MAR model.

The test statistics for the full model and the MAR model are based on the nonparametric estimator $\hat{r}$ (see (2.1)), which involves a bandwidth $c_n = c\,\{n\log(n)\}^{-1/(2s)}$, with proportionality constant c that has to be suitably chosen. We recommend selecting c_n (and thus c) by cross-validation, i.e. c_n is the bandwidth that minimises the leave-one-out cross-validated estimate of the mean squared prediction error (see e.g. Härdle and Marron, 1985). This procedure is easy to implement, is fully data-driven and performed well in our study. It can also be used in the scale function estimator $\hat{\sigma}$ for a test based on $\tilde{T}_n$, which is what we did in the examples below.

The test level is $\alpha = 5\%$ in the following and the asymptotic quantile (introduced after Theorem 1) is therefore $b_{0.05} \approx 1.1779$. For small to moderate sample sizes we recommend the smooth residual bootstrap approach by Neumeyer (2009) for estimating the quantiles, which worked well in our simulation study. In particular, when the sample sizes were small (50 or less) the results using the asymptotic quantile $b_{0.05}$ were not satisfactory in general and the results using the bootstrap quantile were more plausible. At moderate and larger sample sizes the bootstrap quantiles and the asymptotic quantile $b_{0.05}$ produced similar results.

The bootstrap method is suitable for our setting as it makes it possible to produce a smooth bootstrap distribution that satisfies our model assumptions. Note that the smooth bootstrap approach is based on residuals $\hat{\varepsilon} = Y - \hat{r}(X)$ and the estimated weights $\tilde{W}_j$ from (2.4), i.e. it also involves the bandwidths selected by cross-validation that were introduced above.

Example 1: testing for heteroskedasticity with one covariate. For the simulations we chose the regression function as

$$r(x) = 2x + 3\cos(\pi x)$$

to preserve the nonparametric nature of the model. The covariates were generated from a uniform distribution and errors from a standard normal distribution: $X_j \sim U(-1,1)$ and $e_j \sim N(0,1)$ for $j = 1,\ldots,n$. Finally, the indicators δ_j have a Bernoulli($\pi(x)$) distribution, with $\pi(x) = P(\delta = 1|X = x)$. In this study we use a logistic distribution function for $\pi(x)$ with a mean of 0 and a scale parameter of 1. As a consequence the average amount of missing data is around 50%, ranging between 27% and 73%. We work with $d = 1$, the locally linear smoother and sample sizes 50, 100, 200 and 300.

In order to investigate the level and power of the test in the full model and in the MAR model we consider the following scale functions:

$$\sigma_0(x) = 1, \qquad \sigma_1(x) = 0.4 + 4x^2,$$
$$\sigma_2(x) = 2e^x - 0.5, \qquad \sigma_3(x) = 1 + 15n^{-1/2}x^2.$$

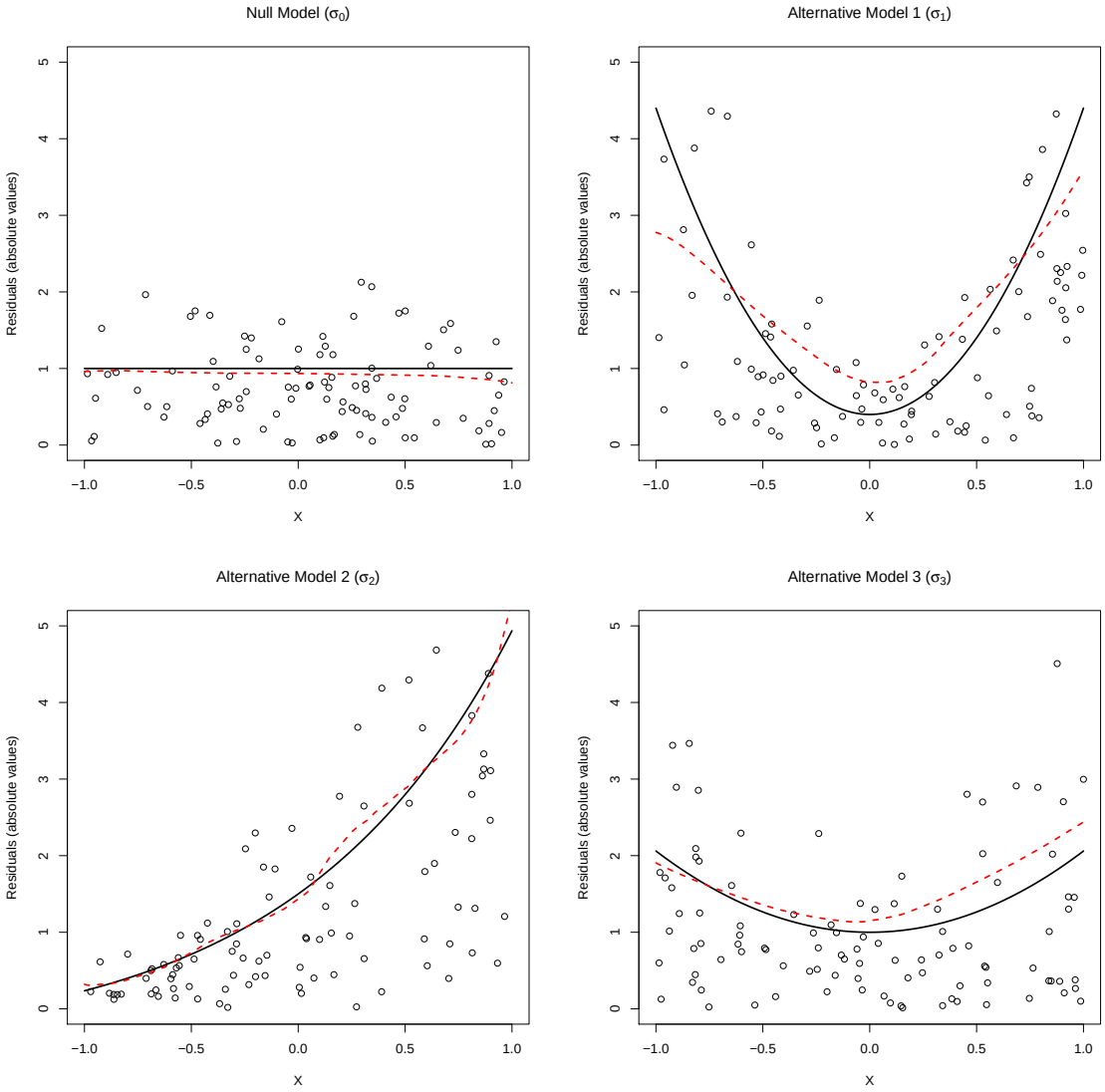

FIGURE 1. Scatter plots of absolute valued residuals. Each plot also shows the underlying scale function in black (solid line) and a kernel smoothed estimate of the scale function in red (dashed line).

The constant scale function σ_0 allows for the (5%) level of the test to be checked. As an illustration, we generated a random dataset of size 100 for each scenario. A scatter plot of the residuals (in absolute value) from the nonparametric regression is given for each dataset (Figure 1).

The simulations based on (non-constant) scale functions σ_1, σ_2 and σ_3 give an indication of the power of the test in different scenarios. In particular, we consider the power of the test against the local alternative σ_3. The power is maximised if ω equals the scale function σ (or is a linear transformation of σ); see Remark 2 in Section 2 and the discussion following it. The formula for the weights based on an estimator $\hat{\sigma}(x) = \{\hat{r}_2(x) - \hat{r}^2(x)\}^{1/2}$ of σ is given in (2.4).

σ \ n	50	100	200	300
σ_0	0.016 (0.058)	0.019 (0.056)	0.033 (0.039)	0.039
σ_1	0.426 (0.477)	0.939 (0.945)	1.000 (1.000)	1.000
σ_2	0.487 (0.631)	0.971 (0.957)	1.000 (0.996)	1.000
σ_3	0.127 (0.176)	0.299 (0.387)	0.500 (0.576)	0.668

TABLE 1. Example 1: Simulated level (σ_0 figures) and power for fully observed data ($\tilde{T}_n$). Figures in parentheses are obtained from bootstrap quantiles.

σ \ n	50	100	200	300
σ_0	0.009 (0.054)	0.015 (0.051)	0.029 (0.048)	0.037
σ_1	0.097 (0.173)	0.482 (0.573)	0.957 (0.953)	0.998
σ_2	0.112 (0.223)	0.443 (0.550)	0.945 (0.913)	1.000
σ_3	0.032 (0.080)	0.097 (0.148)	0.197 (0.283)	0.304

TABLE 2. Example 1: Simulated level (σ_0 figures) and power for missing data ($\tilde{T}_{n,c}$). Figures in parentheses are obtained from bootstrap quantiles.

To check the performance of our test we conducted simulations of 1000 runs. Table 1 shows the test results for fully observed data ($\tilde{T}_n$). Similar results are given for missing data ($\tilde{T}_{n,c}$) in Table 2.

The figures corresponding to the null hypothesis (σ_0) show that test levels for fully observed data ($\tilde{T}_n$) are near the desired 5% at larger sample sizes (3.9% at sample size 300) but more conservative at smaller sample sizes (1.6% at sample size 50). The results for missing data ($\tilde{T}_{n,c}$) are further away from 5% when the sample size is small and the asymptotic quantile $b_{0.05}$ is used. Both testing procedures appear to have levels near the desired 5% when the smooth bootstrap quantiles are used, which is expected.

We now consider the power of each test beginning with σ_1. The procedure for fully observed data ($\tilde{T}_n$) performs very well at moderate and larger sample sizes. For example, we rejected the null hypothesis 93.9% of the time at the moderate sample size of 100 using the asymptotic quantile. In this case, results using the smooth bootstrap quantile were almost identical (rejecting 94.5% of the time). Similar results were obtained for missing data ($\tilde{T}_{n,c}$), but they are (as expected) less impressive. Note that the smooth bootstrap quantiles do not (in general) perform dramatically better than the asymptotic quantile $b_{0.05}$.

The figures corresponding to σ_2 and σ_3 show that both tests have difficulty rejecting when samples are small. The procedure for fully observed data ($\tilde{T}_n$) only rejected the null hypothesis 48.7% (σ_2) and 12.7% (σ_3) of the time for samples of size 50 and less often when data were missing. Here the smooth bootstrap quantiles show improved performance over the asymptotic quantile $b_{0.05}$ and reject the null hypothesis 63.1% (σ_2) and 17.6% (σ_3) of the time. The results are similar for missing data but with reduced performance (as expected).

In conclusion, each test performs well and the procedures $\tilde{T}_n$ and $\tilde{T}_{n,c}$ proposed in this article appear promising for detecting heteroskedasticity. When sample sizes are small the smooth bootstrap quantiles appear to be helpful.

σ \ n	50	100	200	300
σ_0	0.017	0.015	0.016	0.019
σ_1	0.153	0.749	0.996	1.000
σ_2	0.027	0.023	0.025	0.026

TABLE 3. Example 2: Simulated level (σ_0 figures) and power for $\tilde{T}_n$ using detection function $\omega_1 = \sigma_1$.

σ \ n	50	100	200	300
σ_0	0.003	0.010	0.018	0.019
σ_1	0.085	0.415	0.904	0.989
σ_2	0.015	0.037	0.118	0.208

TABLE 4. Example 2: Simulated level (σ_0 figures) and power for $\tilde{T}_n$ using detection function $\omega_2 = 1 + \cos((\pi/2)(x_1 + x_2))$.

σ \ n	50	100	200	300
σ_0	0.063	0.049	0.060	0.069
σ_1	0.186	0.617	0.971	1.000
σ_2	0.477	0.934	1.000	1.000

TABLE 5. Example 2: Simulated level (σ_0 figures) and power for $\tilde{T}_n$ using detection function ω_3 as an estimated scale function.

Example 2: testing for heteroskedasticity with two covariates. Here we work with the regression function

$$r(x_1, x_2) = 2x_1 - x_2 + 3e^{x_1+x_2}.$$

The covariates X_1 and X_2 are generated from a joint normal distribution, with component variances 1 and correlation coefficient 1/2, restricted to the interval $[-1, 1]^2$ by rejection sampling. As above we generate the model errors from a standard normal distribution. In this example we do not consider missing data because we expect the conclusions to mirror those of the first simulation study. We are interested in the performance of our testing procedure when we select different weights. We work with $d = 3$, the locally cubic smoother, and sample sizes 50, 100, 200 and 300. The level of the test is 5%, as in Example 1. In Example 1 the bootstrap and the test based on the asymptotic quantile $b_{0.05}$ produced similar results for sample sizes 100 and larger. We therefore only consider the latter method in this example.

For the simulations we use three scale functions:

$$\sigma_0 \equiv 1, \quad \sigma_1(x_1, x_2) = 0.5 + 5x_1^2 + 5x_2^2, \quad \sigma_2(x_1, x_2) = 4 + 3.5 \sin\big((\pi/2)(x_1 + x_2)\big)$$

Our weights are constructed based on detection functions:

$$\omega_1 = \sigma_1, \quad \omega_2(x_1, x_2) = 1 + \cos\big((\pi/2)(x_1 + x_2)\big).$$

and ω_3 is an estimated scale function as in Example 1. We expect that the choice $\omega_1 = \sigma_1$ will provide the largest power for detecting σ_1. We also consider the choice ω_2 to illustrate the test performance when we choose some reasonable non-constant detection function. The detection function ω_3 is based on a locally cubic estimator for the scale function.

We conducted simulations consisting of 1000 runs. The results for ω_1, ω_2 and ω_3 are given in Tables 3, 4 and 5, respectively. The figures corresponding to the test level $\alpha = 5\%$ ($\sigma_0 \equiv 1$) and the fixed detection functions ω_1 and ω_2 show the tests using the asymptotic quantile are conservative, which mirrors the results from Example 1. At sample size 300, the rejection rates for the tests using ω_1 and ω_2 are both near 2%. The test using ω_3, an estimated scale function, generally produces higher rejection rates (but still near the nominal level). At the sample size 300 the rejection rate for the test using ω_3 is about 7%.

When we consider the remaining figures corresponding to the powers of each test, we find considerable differences. The test using $\omega_1 = \sigma_1$ (Table 3) provides, as expected, the best results when σ_1 is in fact the underlying scale function. The corresponding figures for the test that uses the estimated scale function ω_3 (Table 5) are similar. The results in Table 4 indicate that the test using $\omega_2 = 1 + \cos((\pi/2)(x_1 + x_2))$ is less effective for detecting σ_1, but still quite good for the larger sample sizes 200 and 300. Comparing the three tests in Tables 3-5 for detecting σ_2, we see that only the test with the estimated scale function ω_3 appears to be powerful.

Only at very large sample sizes can we expect that all three testing procedures will provide similar results. In conclusion, we find the test using an arbitrary non-constant weight function is useful but will normally be outperformed by a test using estimated weights.

Remark 4 (curse of dimensionality). The simulation results in Tables 3-5 suggest that our proposed tests, which are based on local polynomial smoothers, may not be very reliable when the dimension of the covariate vector is large. In this case the smoother (as well as other nonparametric function estimators) will be affected by the "curse of dimensionality", which is implicated by the entropy results in Section 5. To meet the situation with many covariates in practice, we recommend working with dimension-reducing transformations: choose, for example, a transformation ξ of the covariate vector X such that $V = \xi(X)$ is just one covariate (and the function estimator is not affected by the dimensionality problem). A popular example would be the single-index model, where ξ is a linear combination of the components of X. Working with such transformations will not change the independence-dependence structure between the detection function and the errors, which is key for our procedure to work.

5. Technical details

In this section we present the proof of Theorem 1, the limiting distribution of T_n under the null hypothesis, and some auxiliary results. As explained in Section 3, we do not have to prove Theorem 3 for the MAR model: it suffices to consider the full model and the test statistic T_n. Our approach consists of two steps. Our first step will be to use Theorem 2.2.4 in Koul's 2002 book on weighted empirical processes to obtain the limiting distribution of an asymptotically linear statistic (a sum of i.i.d. random variables) that is related to T_n. Then we review some results from Müller, Schick and Wefelmeyer (2009), who propose local polynomial smoothers to estimate a regression function of many covariates. Using

these results, we will show that the statistic T_n and the asymptotically linear statistic are indistinguishable for large samples, i.e. they have the same limiting distribution.

The asymptotically linear statistic, which is an empirical process related to T_n, is defined similarly to T_n as

$$\sup_{t\in\mathbb{R}}\Big|n^{-1/2}\sum_{j=1}^{n}W_j\Big\{\mathbf{1}[\varepsilon_j\le t]-F(t)\Big\}\Big| = \sup_{t\in\mathbb{R}}\Big|n^{-1/2}\sum_{j=1}^{n}W_j\Big\{\mathbf{1}[\sigma_0 e_j\le t]-F(t)\Big\}\Big|, \tag{5.1}$$

where $\varepsilon_j = \sigma_0 e_j$ is the unobserved "model error" from the null hypothesis and $W_1,\dots,W_n$ are the standardised weights given in (1.3). We will now demonstrate that (under H_0) the requirements for Koul's theorem are satisfied. The asymptotic statement is given afterwards in Corollary 1.

Theorem 2.2.4 of Koul (2002) states that

$$\zeta_n(t) = n^{-1/2}\sum_{j=1}^{n}D_j\Big\{\mathbf{1}\big[C_j\le t\big]-K(t)\Big\}\xrightarrow{D}\xi\Big\{B_0\circ K(t)\Big\},\qquad t\in\mathbb{R},\ \text{as } n\to\infty,$$

where B_0 is the standard Brownian bridge in the Skorohod space $D[0,1]$, independent of a random variable ξ. The roles of his random variable C_j and the square integrable random variable D_j, which are assumed to be independent, are now played by $\sigma_0 e_j$ and W_j, $j = 1,\dots,n$. The distribution function K corresponds to our error distribution function F and is assumed to have a uniformly continuous Lebesgue density. The random variable ξ from above comes from Koul's requirement that

$$\Big|\frac{1}{n}\sum_{j=1}^{n}D_j^2\Big|^{1/2} = \xi + o_p(1)\qquad \text{for some positive r.v. } \xi.$$

Here we work with W_j, in place of D_j, with $E(W_j^2) = 1$. Therefore, by the law of large numbers, $n^{-1}\sum_{j=1}^{n}W_j^2 = 1 + o_p(1)$ and, using the continuous mapping theorem, $|n^{-1}\sum_{j=1}^{n}W_j^2|^{1/2} = 1 + o_p(1)$, i.e. $\xi\equiv 1$. Hence we have

$$n^{-1/2}\sum_{j=1}^{n}W_j\Big\{\mathbf{1}[\sigma_0 e_j\le t]-F(t)\Big\}\xrightarrow{D}B_0\circ F(t),\qquad t\in\mathbb{R},\ \text{as } n\to\infty.$$

Taking the supremum with respect to $t\in\mathbb{R}$, the right-hand side becomes $\sup_{t\in\mathbb{R}}|B_0\circ F(t)| = \sup_{t\in[0,1]}|B_0(t)|$, which specifies the asymptotic distribution of the asymptotically linear statistic (5.1). Note that Koul's theorem also provides the limiting distribution for a shifted version $\hat{\zeta}_n$ of ζ_n that involves random variables $Z_1,\dots,Z_n$. Since we only need the simpler result for ζ_n, we do not need to verify the more complicated assumptions regarding the Z_j's. This shows the conditions of Theorem 2.2.4 in Koul (2002) are indeed satisfied. We will formulate this result as a corollary. Since we only require the weights to be square-integrable functions of X_j with $E(W_j^2) = 1$, we will not require the explicit form (1.3).

Corollary 1. *Consider the homoskedastic nonparametric regression model $Y = r(X) + \sigma_0 e$. Assume the distribution function F of the errors has a uniformly continuous Lebesgue density f that is positive almost everywhere. Further, let W_j be a square integrable function*

of X_j satisfying $E(W_j^2) = 1$, $j = 1, \ldots, n$. Then

$$\sup_{t\in\mathbb{R}} \Big| n^{-1/2} \sum_{j=1}^{n} W_j \Big\{ \mathbf{1}[\sigma_0 e_j \le t] - F(t) \Big\} \Big| \xrightarrow{D} \sup_{t\in[0,1]} |B_0(t)|, \qquad \text{as } n \to \infty, \tag{5.2}$$

where B_0 denotes the standard Brownian bridge.

For our second step, we will show that T_n and the asymptotically linear statistic (5.1) are asymptotically equivalent. To begin we rewrite T_n, using the identity (under H_0) $\hat{\varepsilon} = Y - \hat{r}(X) = \sigma_0 e - \hat{r}(X) + r(X)$, as

$$\sup_{t\in\mathbb{R}} \Big| n^{-1/2} \sum_{j=1}^{n} \hat{W}_j \mathbf{1}\big[\hat{\varepsilon}_j \le t\big] \Big| = \sup_{t\in\mathbb{R}} \Big| n^{-1/2} \sum_{j=1}^{n} \hat{W}_j \mathbf{1}\big[\sigma_0 e_j \le t + \hat{r}(X_j) - r(X_j)\big] \Big|.$$

We will first consider the shift in the indicator function from t to $t + \hat{r} - r$, which comes in because T_n involves an estimator $\hat{r}$ of the regression function.

Consider now the Hölder space $H(d, \gamma)$ from Section 2, i.e. the space of functions that have partial derivatives of order d that are Hölder with exponent $\gamma \in (0, 1]$. For these functions we define the norm

$$\|h\|_{d,\gamma} = \max_{i\in I(d)} \sup_{x\in[0,1]^m} \big| D^i h(x) \big| + \max_{i\in I(d)} \sup_{x,\, y\in[0,1]^m,\ x\neq y} \frac{|D^i h(y) - D^i h(x)|}{\|x - y\|^\gamma},$$

where $\|v\|$ is the Euclidean norm of a real-valued vector v and

$$D^i h(x) = \frac{\partial^{i_1+\cdots+i_m}}{\partial x_1^{i_1} \cdots \partial x_m^{i_m}} h(x), \qquad x = (x_1, \ldots, x_m) \in [0, 1]^m.$$

Write $H_1(d, \gamma)$ for the unit ball of $H(d, \gamma)$ using this norm. These function spaces are particularly useful for studying local polynomial smoothers $\hat{r}$ as defined in Section 2. Müller et al. (2009) make use of these spaces to derive many useful facts concerning regression function estimation using local polynomials. We will use some of their results to prove Theorem 1; see Lemma 1 below.

Lemma 1 (Lemma 1 of Müller, Schick and Wefelmeyer, 2009). *Let the local polynomial smoother $\hat{r}$, the regression function r, the covariate distribution G and the error distribution F satisfy the assumptions of Theorem 1. Then there is a random function $\hat{a}$ such that, for some $\alpha > 0$,*

$$P(\hat{a} \in H_1(m, \alpha)) \to 1, \tag{5.3}$$

$$\sup_{x\in[0,1]^m} \big| \hat{r}(x) - r(x) - \hat{a}(x) \big| = o_p(n^{-1/2}). \tag{5.4}$$

We now use these results to show the difference between the asymptotically linear statistic (5.1) and an empirical process related to the shifted version of T_n (called R_1 in Lemma 2 below) are asymptotically negligible. An unweighted version of that difference (with $W_j = 1$) is considered in Theorem 2.2 of Müller, Schick and Wefelmeyer (2007). Since that statistic does not involve centred weights, the second part of R_1 (called R_2 in the lemma) is not asymptotically negligible: it becomes a stochastic drift parameter that depends on the error density f ($f(t) \int \hat{a}\, dQ$ in that article) and is therefore *not* distribution free. This is in contrast to our case where we have mean zero weights, so R_2 does not affect the limiting distribution.

LEMMA 2. *Let the null hypothesis hold. Suppose the assumptions of Theorem 1 on $\hat{r}$, r, G and F are satisfied. Let W_j be a square integrable function of X_j satisfying $E[W_j^2] < \infty$, $j = 1, \ldots, n$. Then $\sup_{t\in\mathbb{R}} |R_1| = o_p(n^{-1/2})$, where*

$$R_1 = \frac{1}{n}\sum_{j=1}^{n} W_j\Big\{\mathbf{1}\big[\sigma_0 e_j \le t + \hat{r}(X_j) - r(X_j)\big] - \mathbf{1}[\sigma_0 e_j \le t] - F\big(t + \hat{r}(X_j) - r(X_j)\big) + F(t)\Big\}.$$

If, additionally, $E[W_j] = 0$, $j = 1, \ldots, n$, then $\sup_{t\in\mathbb{R}} |R_2| = o_p(n^{-1/2})$, where

$$R_2 = \frac{1}{n}\sum_{j=1}^{n} W_j\Big\{F\big(t + \hat{r}(X_j) - r(X_j)\big) - F(t)\Big\}.$$

PROOF. Observe that the class of functions

$$\mathfrak{F} = \Big\{(X, \sigma_0 e) \mapsto W\Big\{\mathbf{1}\big[\sigma_0 e \le t + a(X)\big] - F\big(t + a(X)\big)\Big\} \,:\, t \in \mathbb{R},\ a \in H_1(m, \alpha)\Big\}$$

is $G\otimes F$-Donsker, which follows from the fact that W_j is a fixed, square-integrable function of X_j and the class of indicator functions in the definition of $\mathfrak{F}$ is $G\otimes F$-Donsker from Theorem 2.1 of Müller et al. (2007). It then follows from Corollary 2.3.12 of van der Vaart and Wellner (1996) that empirical processes ranging over the Donsker class $\mathfrak{F}$ are asymptotically equicontinuous, i.e. we have, for any $\varphi > 0$,

$$\lim_{\kappa\downarrow 0} \limsup_{n\to\infty} P\Big(\sup_{\{\mathfrak{f}_1, \mathfrak{f}_2 \in \mathfrak{F} \,:\, \mathrm{Var}(\mathfrak{f}_1 - \mathfrak{f}_2) < \kappa\}} n^{-1/2} \Big| \sum_{j=1}^{n} \Big\{\mathfrak{f}_1(X_j, \sigma_0 e_j) - \mathfrak{f}_2(X_j, \sigma_0 e_j)\Big\}\Big| > \varphi\Big) = 0. \tag{5.5}$$

We are interested in the case that involves the approximation $\hat{a}$ of $\hat{r} - r$ in place of a (see Lemma 1). Then the corresponding class of functions is, in general, no longer Donsker (and the equicontinuity property does not hold). However, we can assume that $\hat{a}$ is in $H_1(m, \alpha)$, which holds on an event that has probability tending to one. This together with a negligibility condition on the variance guarantees that the extended class of processes involving $\hat{a}$ is also equicontinuous.

The term R_1 from the first assertion can be written as the sum of

$$\frac{1}{n}\sum_{j=1}^{n} W_j\Big\{\mathbf{1}\big[\sigma_0 e_j \le t + \hat{a}(X_j)\big] - \mathbf{1}[\sigma_0 e_j \le t] - F\big(t + \hat{a}(X_j)\big) + F(t)\Big\} \tag{5.6}$$

and

$$\begin{aligned} &\frac{1}{n}\sum_{j=1}^{n} W_j\Big\{\mathbf{1}\big[\sigma_0 e_j \le t + \hat{r}(X_j) - r(X_j)\big] - F\big(t + \hat{r}(X_j) - r(X_j)\big)\Big\} \\ &- \frac{1}{n}\sum_{j=1}^{n} W_j\Big\{\mathbf{1}\big[\sigma_0 e_j \le t + \hat{a}(X_j)\big] - F\big(t + \hat{a}(X_j)\big)\Big\}. \end{aligned} \tag{5.7}$$

The first assertion, $\|R_1\|_\infty = o_p(n^{-1/2})$, will follow if we show this separately for the two terms in (5.6) and (5.7). Consider (5.6) first. We fix the function $\hat{a}$ by conditioning on the observed data $\mathbb{D} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. The variance of a function from the extension

of $\mathfrak{F}$ that involves $\hat{a}$ instead of a is

$$\begin{aligned}
&\operatorname{Var}\Big[W\Big\{\mathbf{1}\big[\sigma_0 e \le t+\hat{a}(X)\big]-\mathbf{1}[\sigma_0 e\le t]-F\big(t+\hat{a}(X)\big)+F(t)\Big\}\,\Big|\,\mathbb{D}\Big]\\
&=E\Big[W^2\Big\{F\big(\max\{t,\,t+\hat{a}(X)\}\big)-F\big(\min\{t,\,t+\hat{a}(X)\}\big)\\
&\qquad\qquad-\Big\{F\big(\max\{t,\,t+\hat{a}(X)\}\big)-F\big(\min\{t,\,t+\hat{a}(X)\}\big)\Big\}^2\Big\}\,\Big|\,\mathbb{D}\Big].
\end{aligned}$$

If this variance is $o_p(1)$, then the extended class of processes is equicontinuous, and the term in (5.6) has the desired order $o_p(n^{-1/2})$, uniformly in $t\in\mathbb{R}$. That the variance condition holds true is easy to see: the last term is bounded by $\|f\|_\infty E\big[W^2\big]\|\hat{a}\|_\infty = o_p(1)$ with $\|\hat{a}\|_\infty = o_p(1)$ (see page 961 of the proof of Lemma 1 in Müller et al., 2009).

Turning our attention now to the second term (5.7), we have that $\hat{r}-r=(\hat{r}-r-\hat{a})+\hat{a}$, and, by property (5.4) of Lemma 1, $A_n=\|\hat{r}-r-\hat{a}\|_\infty=o_p(n^{-1/2})$. Write $W_j^-=W_j\mathbf{1}[W_j<0]$ and $W_j^+=W_j\mathbf{1}[W_j\ge 0]$ for the negative and the positive part of W_j, i.e. $W_j=W_j^-+W_j^+$, $j=1,\dots,n$. This yields the following bounds for the weighted indicator functions:

$$\begin{aligned}
W_j^-\mathbf{1}\big[\sigma_0 e_j\le t+\hat{r}(X_j)-r(X_j)\big]&\le W_j^-\mathbf{1}\big[\sigma_0 e_j\le t-A_n+\hat{a}(X_j)\big],\\
W_j^-\mathbf{1}\big[\sigma_0 e_j\le t+\hat{a}(X_j)\big]&\ge W_j^-\mathbf{1}\big[\sigma_0 e_j\le t+A_n+\hat{a}(X_j)\big],\\
W_j^+\mathbf{1}\big[\sigma_0 e_j\le t+\hat{r}(X_j)-r(X_j)\big]&\le W_j^+\mathbf{1}\big[\sigma_0 e_j\le t+A_n+\hat{a}(X_j)\big].
\end{aligned}$$

and

$$W_j^+\mathbf{1}\big[\sigma_0 e_j\le t+\hat{a}(X_j)\big]\ge W_j^+\mathbf{1}\big[\sigma_0 e_j\le t-A_n+\hat{a}(X_j)\big].$$

Straightforward calculations show that (5.7) is bounded by

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^n\Big\{W_j^+-W_j^-\Big\}\Big\{\mathbf{1}\big[\sigma_0 e_j\le t+A_n+\hat{a}(X_j)\big]-F\big(t+A_n+\hat{a}(X_j)\big)\Big\}\\
&\quad-\frac{1}{n}\sum_{j=1}^n\Big\{W_j^+-W_j^-\Big\}\Big\{\mathbf{1}\big[\sigma_0 e_j\le t-A_n+\hat{a}(X_j)\big]-F\big(t-A_n+\hat{a}(X_j)\big)\Big\}\\
&\quad+\frac{1}{n}\sum_{j=1}^n\Big\{W_j^+-W_j^-\Big\}\Big\{F\big(t+A_n+\hat{a}(X_j)\big)-F\big(t-A_n+\hat{a}(X_j)\big)\Big\}\\
&\quad-\frac{1}{n}\sum_{j=1}^n W_j\Big\{F\big(t+\hat{r}(X_j)-r(X_j)\big)-F\big(t+\hat{a}(X_j)\big)\Big\}\\
&=\frac{1}{n}\sum_{j=1}^n\big|W_j\big|\Big\{\mathbf{1}\big[\sigma_0 e_j\le t+A_n+\hat{a}(X_j)\big]-F\big(t+A_n+\hat{a}(X_j)\big)\Big\}\\
&\quad-\frac{1}{n}\sum_{j=1}^n\big|W_j\big|\Big\{\mathbf{1}\big[\sigma_0 e_j\le t-A_n+\hat{a}(X_j)\big]-F\big(t-A_n+\hat{a}(X_j)\big)\Big\}\\
&\quad+\frac{1}{n}\sum_{j=1}^n\big|W_j\big|\Big\{F\big(t+A_n+\hat{a}(X_j)\big)-F\big(t-A_n+\hat{a}(X_j)\big)\Big\}\\
&\quad-\frac{1}{n}\sum_{j=1}^n W_j\Big\{F\big(t+\hat{r}(X_j)-r(X_j)\big)-F\big(t+\hat{a}(X_j)\big)\Big\}.
\end{aligned}$$

Hence (5.7) is $o_p(n^{-1/2})$ uniformly in $t \in \mathbb{R}$ holds if we show

$$\sup_{t\in\mathbb{R}} \Big| \frac{1}{n} \sum_{j=1}^{n} |W_j| \Big\{ \mathbf{1}\big[\sigma_0 e_j \le t + A_n + \hat{a}(X_j)\big] - F\big(t + A_n + \hat{a}(X_j)\big) \Big\} - \frac{1}{n} \sum_{j=1}^{n} |W_j| \Big\{ \mathbf{1}\big[\sigma_0 e_j \le t - A_n + \hat{a}(X_j)\big] - F\big(t - A_n + \hat{a}(X_j)\big) \Big\} \Big| = o_p(n^{-1/2}), \tag{5.8}$$

$$\sup_{t\in\mathbb{R}} \frac{1}{n} \sum_{j=1}^{n} |W_j| \Big\{ F\big(t + A_n + \hat{a}(X_j)\big) - F\big(t - A_n + \hat{a}(X_j)\big) \Big\} = o_p(n^{-1/2}) \tag{5.9}$$

and

$$\sup_{t\in\mathbb{R}} \Big| \frac{1}{n} \sum_{j=1}^{n} W_j \Big\{ F\big(t + \hat{r}(X_j) - r(X_j)\big) - F\big(t + \hat{a}(X_j)\big) \Big\} \Big| = o_p(n^{-1/2}). \tag{5.10}$$

Beginning with (5.8), since the random variables $|W_1|, \ldots, |W_n|$ are square integrable, the class of functions

$$\mathfrak{F}^+ = \Big\{ (X, \sigma_0 e) \mapsto |W| \Big\{ \mathbf{1}\big[\sigma_0 e \le t + a(X)\big] - F\big(t + a(X)\big) \Big\} \,:\, t \in \mathbb{R},\ a \in H_1(m, \alpha) \Big\}$$

is also $G \otimes F$-Donsker. Therefore the asymptotic equicontinuity property holds for empirical processes ranging over $\mathfrak{F}^+$, i.e. (5.5) holds with $\mathfrak{F}^+$ in place of $\mathfrak{F}$. However, rather than investigating the situation where $\hat{a}$ is limiting toward zero, as we did above, we will consider two sequences of real numbers $\{s_n\}_{n=1}^{\infty}$ and $\{t_n\}_{n=1}^{\infty}$ satisfying $|t_n - s_n| = o(1)$, which corresponds to the case of random sequences $t \pm A_n$ conditional on the data $\mathbb{D}$. Analogously to the calculations following (5.5), we now prove the variance condition for the function $(X, \sigma_0 e) \mapsto |W|\{\mathbf{1}[\sigma_0 e \le t_n + a(X)] - \mathbf{1}[\sigma_0 e \le s_n + a(X)] - F(t_n + a(X)) + F(s_n + a(X))\}$. The variance is

$$\begin{aligned} &\mathrm{Var}\Big[|W| \Big\{ \mathbf{1}\big[\sigma_0 e \le t_n + a(X)\big] - \mathbf{1}\big[\sigma_0 e \le s_n + a(X)\big] - F\big(t_n + a(X)\big) - F\big(s_n + a(X)\big) \Big\} \Big] \\ &= E\Big[W^2 \Big\{ F\big(\max\{t_n + a(X),\, s_n + a(X)\}\big) - F\big(\min\{t_n + a(X),\, s_n + a(X)\}\big) \\ &\qquad - \Big\{ F\big(\max\{t_n + a(X),\, s_n + a(X)\}\big) - F\big(\min\{t_n + a(X),\, s_n + a(X)\}\big) \Big\}^2 \Big\} \Big]. \end{aligned}$$

and bounded by $\|f\|_\infty E[W^2]|t_n - s_n| = o(1)$. Hence we have equicontinuity, and therefore, for any $a \in H_1(m, \alpha)$ and sequences of real numbers $\{s_n\}_{n=1}^{\infty}$ and $\{t_n\}_{n=1}^{\infty}$ satisfying $|t_n - s_n| = o(1)$,

$$\sup_{t\in\mathbb{R}} \Big| \frac{1}{n} \sum_{j=1}^{n} |W_j| \Big\{ \mathbf{1}\big[\sigma_0 e_j \le t_n + a(X_j)\big] - F\big(t_n + a(X_j)\big) \Big\} - \frac{1}{n} \sum_{j=1}^{n} |W_j| \Big\{ \mathbf{1}\big[\sigma_0 e_j \le s_n + a(X_j)\big] - F\big(s_n + a(X_j)\big) \Big\} \Big| = o_p(n^{-1/2}).$$

Equation (5.8) follows analogously, with t_n replaced by $t + A_n$, s_n by $t - A_n$ and a by $\hat{a}$.

Now consider (5.9) and (5.10). Since $E|W| \le E^{1/2}[W^2] < \infty$, $n^{-1} \sum_{j=1}^{n} |W_j|$ is consistent for $E|W|$. The left-hand side of (5.9) is bounded by $2\|f\|_\infty A_n n^{-1} \sum_{j=1}^{n} |W_j|$ and (5.10) is bounded by $\|f\|_\infty A_n n^{-1} \sum_{j=1}^{n} |W_j|$. Since $A_n = o_p(n^{-1/2})$, these bounds are also $o_p(n^{-1/2})$,

i.e. (5.9) and (5.10) hold. This implies that the term in (5.7) has order $o_p(n^{-1/2})$, uniformly in $t \in \mathbb{R}$, which completes the proof of $\|R_1\|_\infty = o_p(n^{-1/2})$.

We will now prove the second assertion that $\|R_2\|_\infty = o_p(n^{-1/2})$. The proof is simpler than the proof of the first assertion since we now require that the random variables $W_1, \dots, W_n$ have mean zero, which allows us to use the central limit theorem. Write R_2 as

$$\begin{aligned} R_2 = \frac{1}{n}\sum_{j=1}^n W_j\Big\{F\big(t+\hat a(X_j)\big) - F(t) - E\Big[F\big(t+\hat a(X)\big) - F(t)\,\Big|\,\mathbb{D}\Big]\Big\} \\ + E\Big[F\big(t+\hat a(X)\big) - F(t)\,\Big|\,\mathbb{D}\Big]\Big(\frac{1}{n}\sum_{j=1}^n W_j\Big) \\ + \frac{1}{n}\sum_{j=1}^n W_j\Big\{F\big(t+\hat r(X_j)-r(X_j)\big) - F\big(t+\hat a(X_j)\big)\Big\}. \end{aligned}$$

Then $\|R_2\|_\infty$ is bounded by three terms:

$$\sup_{t\in\mathbb{R}}\Big|\frac{1}{n}\sum_{j=1}^n W_j\Big\{F\big(t+\hat a(X_j)\big) - F(t) - E\Big[F\big(t+\hat a(X)\big) - F(t)\,\Big|\,\mathbb{D}\Big]\Big\}\Big|, \tag{5.11}$$

$$\sup_{t\in\mathbb{R}}\Big|E\Big[F\big(t+\hat a(X)\big) - F(t)\,\Big|\,\mathbb{D}\Big]\Big|\Big|\frac{1}{n}\sum_{j=1}^n W_j\Big|, \tag{5.12}$$

and the third term is the left-hand side of (5.10), which we have already shown is $o_p(n^{-1/2})$. From the arguments above, it follows for the class of functions

$$\mathfrak{F}_2 = \Big\{X \mapsto W\Big\{F\big(t+a(X)\big) - E\big[F\big(t+a(X)\big)\big]\Big\} \,:\, t\in\mathbb{R},\ a\in H_1(m,\alpha)\Big\}$$

to be G-Donsker. Therefore, empirical processes ranging over $\mathfrak{F}_2$ are asymptotically equicontinuous as in (5.5), but now without $\sigma_0 e$ and with $\mathfrak{F}_2$ in place of $\mathfrak{F}$.

As before, we can assume that $\hat a$ belongs to $H_1(m,\alpha)$. We will now show the variance condition is satisfied for the function $X \mapsto W\{F(t+\hat a(X)) - F(t) - E[F(t+\hat a(X)) - F(t)\,|\,\mathbb{D}]\}$. This variance is equal to

$$\begin{aligned} E\Big[W^2\Big\{F\big(t+\hat a(X)\big) - F(t)\Big\}^2\,\Big|\,\mathbb{D}\Big] + E\big[W^2\big]E^2\Big[F\big(t+\hat a(X)\big) - F(t)\,\Big|\,\mathbb{D}\Big] \\ - 2E\Big[W^2\Big\{F\big(t+\hat a(X)\big) - F(t)\Big\}\,\Big|\,\mathbb{D}\Big]E\Big[F\big(t+\hat a(X)\big) - F(t)\,\Big|\,\mathbb{D}\Big], \end{aligned}$$

and is bounded by $2\|f\|_\infty^2 E[W^2]\|\hat a\|_\infty^2$. Since $\|\hat a\|_\infty = o_p(1)$, the bound above is $o_p(1)$ and the variance is asymptotically negligible. Hence we have equicontinuity and (5.11) is $o_p(n^{-1/2})$, as desired.

Finally we can bound (5.12) by $\|f\|_\infty\|\hat a\|_\infty|n^{-1}\sum_{j=1}^n W_j|$. The central limit theorem combined with $E[W_j] = 0$, $j = 1,\dots,n$, gives $|n^{-1}\sum_{j=1}^n W_j| = O_p(n^{-1/2})$. Since $\|\hat a\|_\infty = o_p(1)$, both the bound above and (5.12) are of the order $o_p(n^{-1/2})$. This completes the proof of the second assertion that $\|R_2\|_\infty = o_p(n^{-1/2})$. □

Using the results of Lemma 2, we will now show that the test statistic T_n and the asymptotically linear statistic above are asymptotically equivalent. This will imply the

limiting distribution of T_n is the same as that of the asymptotically linear statistic (5.1), which we have already investigated; see Corollary 1.

Proof of Theorem 1. Consider the asymptotically linear statistic from (5.1),

$$n^{-1/2}\sum_{j=1}^{n} W_j\Big\{\mathbf{1}[\sigma_0 e_j \le t] - F(t)\Big\},$$

with W_j given in (1.3). It follows, by the arguments preceding Corollary 1, for this statistic to have the limiting distribution $B_0 \circ F(t)$, where B_0 is the Brownian bridge. We will now show that

$$\sup_{t\in\mathbb{R}}\Big|\frac{1}{n}\sum_{j=1}^{n}\hat{W}_j\mathbf{1}\big[\hat{\varepsilon}_j \le t\big] - \frac{1}{n}\sum_{j=1}^{n} W_j\Big\{\mathbf{1}[\sigma_0 e_j \le t] - F(t)\Big\}\Big| = o_p(n^{-1/2}). \tag{5.13}$$

Combining the above, the desired statement of Theorem 1 concerning the limiting distribution of the test statistic T_n follows, i.e.

$$T_n = \sup_{t\in\mathbb{R}}\Big|n^{-1/2}\sum_{j=1}^{n}\hat{W}_j\mathbf{1}\big[\hat{\varepsilon}_j \le t\big]\Big| \xrightarrow{D} \sup_{t\in[0,1]}|B_0(t)|.$$

It follows from $\sum_{j=1}^{n}\hat{W}_j = 0$ that we can decompose the difference in (5.13) into the following sum of five remainder terms: $R_1 + R_3 + R_4 - R_5 - R_6$, where R_1 and R_2 (which is part of R_3) are the remainder terms of Lemma 2, and where the other terms are defined as follows,

$$R_3 = \hat{V}R_2, \qquad \hat{V} = \Big(\frac{\mathrm{Var}[\omega(X_1)]}{\frac{1}{n}\sum_{j=1}^{n}\{\omega(X_j) - \frac{1}{n}\sum_{k=1}^{n}\omega(X_k)\}^2}\Big)^{1/2},$$

$$R_4 = (\hat{V}-1)\Big(\frac{1}{n}\sum_{j=1}^{n} W_j\Big\{\mathbf{1}\big[\sigma_0 e \le t + \hat{r}(X_j) - r(X_j)\big] - F\big(t + \hat{r}(X_j) - r(X_j)\big)\Big\}\Big),$$

$$R_5 = \hat{V}\Big(\frac{1}{n}\sum_{j=1}^{n} W_j\Big)\Big(\frac{1}{n}\sum_{j=1}^{n}\Big\{\mathbf{1}\big[\sigma_0 e \le t + \hat{r}(X_j) - r(X_j)\big] - F\big(t + \hat{r}(X_j) - r(X_j)\big)\Big\}\Big),$$

$$R_6 = \hat{V}\Big(\frac{1}{n}\sum_{j=1}^{n} W_j\Big)\Big(\frac{1}{n}\sum_{j=1}^{n}\Big\{F\big(t + \hat{r}(X_j) - r(X_j)\big) - F(t)\Big\}\Big).$$

Showing $\sup_{t\in\mathbb{R}}|R_i| = o_p(n^{-1/2})$, $i = 1,\dots,6$, will conclude the proof. The statement for $i = 1$ holds true by the first part of Lemma 2 and the statement for $i = 2$ holds true by the second part of the same lemma. Note that the assumptions of both statements of Lemma 2 are satisfied for our choice of weights $W_1,\dots,W_n$. The statement for $i = 3$ follows from the second statement of the same lemma regarding R_2 and from the fact that the first quantity of R_3, $\hat{V}$, is a consistent estimator of one.

To show $\sup_{t\in\mathbb{R}}|R_4| = o_p(n^{-1/2})$, we only need to demonstrate that

$$\sup_{t\in\mathbb{R}}\Big|\frac{1}{n}\sum_{j=1}^{n} W_j\Big\{\mathbf{1}\big[\sigma_0 e_j \le t + \hat{r}(X_j) - r(X_j)\big] - F\big(t + \hat{r}(X_j) - r(X_j)\big)\Big\}\Big| = O_p(n^{-1/2}), \tag{5.14}$$

because the first term of R_4 both does not depend on t and is asymptotically negligible. To verify (5.14), combine the statement for R_1 with the limiting result (5.2) from Corollary

1 for the asymptotically linear statistic, which shows $n^{-1}\sum_{j=1}^n W_j\{\mathbf{1}[\sigma_0 e_j \le t] - F(t)\} = O_p(n^{-1/2})$, uniformly in $t \in \mathbb{R}$.

Now consider R_5 and remember that both Corollary 1 and the first statement of Lemma 2 cover the special case where all of the weights are equal to one, i.e. (5.14) holds with $W_j = 1$, $j = 1, \dots, n$. Therefore, the third term of R_5 is $O_p(n^{-1/2})$, uniformly in $t \in \mathbb{R}$. It is clear for the product of the first and second terms of R_5 to be $o_p(1)$. It then follows that $\sup_{t\in\mathbb{R}} |R_5| = o_p(n^{-1/2})$.

We find that $\sup_{t\in\mathbb{R}} |R_6|$ is bounded by

$$\sup_{t\in\mathbb{R}} |f(t)|\, \hat{V} \Big(\sup_{x\in[0,1]^m} \big|\hat{a}(x)\big| + \sup_{x\in[0,1]^m} \big|\hat{r}(x) - r(x) - \hat{a}(x)\big| \Big) \Big| \frac{1}{n}\sum_{j=1}^n W_j \Big|,$$

with $\hat{V}$ a consistent estimator of one. As in the proof of Lemma 2, we use $\sup_{x\in[0,1]^m} |\hat{a}(x)| = o_p(1)$ and $\sup_{x\in[0,1]^m} |\hat{r}(x) - r(x) - \hat{a}(x)| = o_p(1)$, e.g. see property (5.4) of Lemma 1. Hence, the third term in the bound above is $o_p(1)$. We can apply the central limit theorem to treat the fourth quantity and find it is $O_p(n^{-1/2})$. Combining these findings yields the bound above is $o_p(n^{-1/2})$. This implies $\sup_{t\in\mathbb{R}} |R_6| = o_p(n^{-1/2})$. □

6. Concluding remarks

We have introduced a completely nonparametric test to detect heteroskedasticity in a regression model with multivariate covariates that not only converges at the parametric root-n rate, but is also strikingly simple. The test has the advantage that it is asymptotically distribution free, i.e. quantiles are readily available. The same test can also be applied if responses are missing at random by simply omitting the cases that are not complete and using the same quantiles. Crucial for the performance of the test is the choice of weights: we have seen that the detection function ω should be highly correlated with the scale function σ to maximise the power of the test. If ω and σ are not or only vaguely correlated, then the test has no or almost no power. This suggests that it is best to work with a (possibly estimated) detection function ω that has the same shape as σ.

The methodology developed in this article can be easily extended to form related tests for other model conditions and/or for other regression models. We will illustrate this below by means of two examples. In Example 1 we show how we can modify our statistic (2.2) to obtain a test for the parametric form of the scale function. In Example 2 we discuss a possible test for additivity of the regression function. This example is representative for tests that are based on detecting differences in the regression function under the null and under the alternative hypothesis.

Example 1. Our method can be modified to obtain tests for the parametric form of the scale function, i.e. with null hypothesis $\sigma(\cdot) = \sigma_\theta(\cdot)$ for some $\theta \in \mathbb{R}^p$. Under the null hypothesis we have $\varepsilon = \sigma_\theta(X)e$, with e scaled and centred as before, so that the standardised residuals $Z = \varepsilon/\sigma_\theta(X) = e$ and X are independent, which is the key prerequisite for our test for heteroskedasticity. Hence we can simply use our test based on the statistic

$$\sup_{t\in\mathbb{R}} \Big| n^{-1/2} \sum_{j=1}^n \hat{W}_j \mathbf{1}\big[\hat{Z}_j \le t\big] \Big|, \quad \hat{Z}_j = \frac{\hat{\varepsilon}_j}{\sigma_{\hat{\theta}}(X_j)},$$

which is the statistic T_n from (2.2), now with estimated standardised residuals $\hat{Z}$ in place of $\hat{\varepsilon}$, where $\hat{\theta}$ is a consistent estimator of θ. We expect that the test will also be asymptotically distribution free: the standardisation will result in an asymptotically negligible drift which only affects the indicators $\mathbf{1}[\hat{Z}_j \le t]$. This can be handled using similar arguments.

Example 2. Another important application are tests about the regression function. One might, for example, want to check if certain components of the regression function are constant or irrelevant, or if the regression function has a specific structure. Suppose, for example, we assume an additive nonparametric model with two-dimensional covariate vector $X_j = (X_{1,j}, X_{2,j})^\top$, i.e. the regression function is $r(X_{1,j}, X_{2,j}) = r_1(X_{1,j}) + r_2(X_{2,j})$ under the null hypothesis. The test statistic is

$$\sup_{t\in\mathbb{R}} \Big| n^{-1/2} \sum_{j=1}^{n} \hat{W}_j \mathbf{1}\big[\hat{\varepsilon}_j \le t\big] \Big|, \quad \hat{\varepsilon}_j = Y_j - \hat{r}_1(X_{1,j}) - \hat{r}_2(X_{2,j}),$$

where $\hat{r}_1(x_1)$ and $\hat{r}_2(x_2)$ estimate $\bar{r}_1(x_1)$ and $\bar{r}_2(x_2)$, with $\bar{r}_1 = r_1$ and $\bar{r}_2 = r_2$ under H_0. For sufficiently large n we have the approximation

$$\mathbf{1}[\hat{\varepsilon} \le t] \approx \mathbf{1}[\varepsilon \le t + s(X_1, X_2)], \quad s(x_1, x_2) = \bar{r}_1(x_1) + \bar{r}_2(x_2) - r(x_1, x_2),$$

where the shift s is zero if the null hypothesis holds true. To understand the construction and the power of the test, consider again (1.2) from the introduction (cf. Remark 1 on "power under fixed alternatives"). For simplicity assume $\sigma(\cdot) = \sigma_0$, i.e. $\varepsilon = \sigma_0 e$ and X are independent. This time we have to take the shift into account and consider the difference

$$E\big[\big\{\omega(X) - E\big[\omega(X)\big]\big\}\mathbf{1}[\varepsilon \le t + s(X_1, X_2)]\big],$$

which is zero under H_0, due to the independence assumption and since $s \equiv 0$. Under the alternative hypothesis we have

$$E\big[\omega(X)\mathbf{1}[\varepsilon \le t + s(X_1, X_2)]\big] = E\Big[\omega(X) F\Big(\frac{t + s(X_1, X_2)}{\sigma_0}\Big)\Big],$$

which, in general, does not equal $E[\omega(X)]E\big(F[\{t + s(X_1, X_2)\}/\sigma_0]\big)$, i.e. the above difference is not zero. As already observed in Remark 1, we expect a good power if the detection function is suitably chosen, here in such a way that $\omega(X)$ and the shift function $s(X_1, X_2)$ are highly correlated.

Acknowledgements

Justin Chown acknowledges financial support from the contract 'Projet d'Actions de Recherche Concertées' (ARC) 11/16-039 of the 'Communauté française de Belgique', granted by the 'Académie universitaire Louvain', the IAP research network P7/06 of the Belgian Government (Belgian Science Policy) and the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823, Teilprojekt C4) of the German Research Foundation (DFG).

References

[1] Akritas, M.G. and Van Keilegom, I. (2001). Non-parametric estimation of the residual distribution. *Scand. J. Statist.*, **28**, 549-567.

[2] Cook, R.D. and Weisberg, S. (1983). Diagnostics for Heteroscedasticity in Regression. *Biometrika*, **70**, 1-10.

[3] Dette, H. and Munk, A. (1998). Testing heteroscedasticity in nonparametric regression. *J. R. Stat. Soc. Ser. B*, **60**, 693-708.
[4] Dette, H. (2002). A consistent test for heteroscedasticity in nonparametric regression based on the kernel method. *J. Stat. Plann. Inference*, **103**, 311-329.
[5] Dette, H., Neumeyer, N. and Van Keilegom, I. (2007). A new test for the parametric form of the variance function in nonparametric regression. *J. R. Stat. Soc. Ser. B*, **69**, 903-917.
[6] Dette, H. and Hetzler, B. (2009). A simple test for the nonparametric form of the variance function in nonparametric regression. *Ann. Inst. Statist. Math.*, **61**, 861-886.
[7] Eubank, R.L. and Thomas, W. (1993). Detecting heteroscedasticity in nonparametric regression. *J. R. Stat. Soc. Ser. B*, **55**, 145-155.
[8] Glejser, H. (1969). A new test for heteroskedasticity. *J. Amer. Statist. Assoc.*, **64**, 316-323.
[9] Härdle, W. and Marron, J.S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13**, 1465-1481.
[10] Koul, H.L. (2002). *Weighted Empirical Processes in Dynamic Nonlinear Models.* Lecture Notes in Statistics, Springer, New York.
[11] Koul, H.L., Müller, U.U. and Schick, A. (2012). The transfer principle: a tool for complete case analysis. *Ann. Statist.*, **40**, 3031-3049.
[12] Koul, H.L. and Song, W. (2010). Conditional variance model checking. *J. Statist. Plann. Inference*, **140**, 1056-1072.
[13] Lin, J. and Qu, X. (2012). A consistent test for heteroscedasticity in semi-parametric regression with nonparametric variance function based on the kernel method. *Statistics*, **46**, 565-576.
[14] Little, R.J.A. and Rubin, D.B. (2002). *Statistical analysis with missing data.* Second edition. Wiley-Interscience.
[15] Müller, U.U., Schick A. and Wefelmeyer W. (2007). Estimating the error distribution in semiparametric regression. *Statist. Decisions*, **25**, 1-18.
[16] Müller, U.U., Schick, A. and Wefelmeyer, W. (2009). Estimating the error distribution function in nonparametric regression with multivariate covariates. *Statist. Probab. Lett.*, **79**, 957-964.
[17] Neumeyer, N. (2009). Smooth residual bootstrap for empirical processes of non-parametric regression residuals. *Scand. J. Stat.* **36**, 204-228.
[18] Shorack, G.R. and Wellner, J.A. (1986). *Empirical Processes with Applications to Statistics.* Wiley, New York.
[19] Stute, W. (1997). Nonparametric model checks for regression. *Ann. Statist.*, **25**, 613-641.
[20] Stute, W., Xu, W.L. and Zhu, L.X. (2008). Model diagnosis for parametric regression in high-dimensional spaces. *Biometrika*, **95**, 451-467.
[21] van der Vaart, A.W. (1998). *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
[22] van der Vaart, A.W. and Wellner J.A. (1996). *Weak convergence and empirical processes. With applications to statistics.* Springer Series in Statistics. Springer-Verlag, New York.
[23] White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, **48**, 817-838.
[24] You, J. and Chen, G. (2005). Testing heteroscedasticity in partially linear regression models. *Statist. Probab. Lett.*, **73**, 61-70.
[25] Zheng, X. (2009). Testing heteroscedasticity in nonlinear and nonparametric regressions. *Canad. J. Statist.*, **37**, 282-300.
[26] Zhu, L., Fujikoshi, Y. and Naito, K. (2001). Heteroscedasticity checks for regression models. *Sci. China Ser. A*, **44**, 1236-1252.
[27] Zhu, X., Chen, F., Guo, X. and Zhu, L. (2016). Heteroscedasticity testing for regression models: a dimension reduction-based model adaptive approach. *Comput. Statist. Data Anal.*, **103**, 263-283.