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 the options available using various data 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 whether or not theoretical requirements for certain statistical procedures are satisfied, in particular whether or not the variation in the errors remains constant across values of the covariate. This is an important assumption, which we want to examine more closely. We will therefore consider the model with constant error variance σ_0^2, the *homoskedastic model*

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

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, DE*
[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:

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

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 be able 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 a version 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 and testing hypotheses based on $\hat{\vartheta}$ may lead to invalid conclusions.

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

$$\begin{aligned} H_0 &: \exists\ \sigma_0 > 0,\ \sigma(\cdot) = \sigma_0 \qquad a.e.\ (G), \\ H_a &: \sigma(\cdot) \in \Sigma. \end{aligned}$$

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. Rejection of the null hypothesis would imply that sufficient statistical evidence is gathered in the data to declare the homoskedastic model inappropriate.

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. Dette and Munk (1998) and Dette (2002) also consider nonparametric regression. Dette and Munk create tests based on an approximation of the variance function; in the 2002 paper Dette proposes a residual-based test using kernel estimators. This approach is extended to the case of a partially linear regression by You and Chen (2005) and Lin and Qu (2012). Dette, Neumeyer and Van Keilegom (2007) construct a test for a parametric form of a scale function of the errors from a nonparametric regression using a bootstrap approach. Dette and Hetzler (2009) construct a test for a parametric form of a scale function of a heteroskedastic nonparametric regression using an empirical process. These approaches either require strict modelling of the scale function, but are consistent at the "parametric" root-n rate of convergence, or favour more parsimonious modelling conditions but converge at a slower rate. This means a trade-off between modelling conditions and how much data is required for meaningful statistical inference.

In contrast to some of the above articles we will not require a specific (parametric) model for the unknown variance function. Our approach is new in that our proposed test statistic for the nonparametric regression model converges at the root-n rate. Moreover, we allow X to be multivariate, which is also new. The tests proposed by Dette (2002) for the nonparametric regression model are probably closest to our approach. However, Dette only considers the case where X is univariate and his tests converge with rates slower than root-n.

The tests introduced in this article are inspired 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 studies a test for a parametric regression against nonparametric alternatives and, in particular, of Stute, Xu and Zhu (2008), who propose a related test suitable for high-dimensional covariates. The test statistics are based on weighted empirical distribution functions. The form of these statistics is strikingly simple and their associated limiting behaviour is obtained by considering the related weighted empirical process.

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

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

If the null hypothesis is true, we can write

$$E\big[\omega(X)\mathbf{1}[\sigma_0 e \le t]\big] = E\big[E\big[\omega(X)\big]\mathbf{1}[\sigma_0 e \le t]\big], \qquad t \in \mathbb{R}.$$

Here we have also used that under the null hyptothesis the covariates X and the errors $\sigma(X)e = \sigma_0 e$ are independent. This motivates a test based on the difference between the two quantities (which is zero only under H_0), i.e. on

$$E\big[\big\{\omega(X) - E\big[\omega(X)\big]\big\}\mathbf{1}[\sigma_0 e \le t]\big], \qquad t \in \mathbb{R}.$$

We can estimate the outer expectation by its empirical version, which yields a test based on

$$U_n(t) = n^{-1/2}\sum_{j=1}^{n}\Big\{\omega(X_j) - E\big[\omega(X_j)\big]\Big\}\mathbf{1}[\sigma_0 e_j \le t], \qquad t \in \mathbb{R}.$$

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. Therefore, we standardise $U_n(t)$ and obtain the weighted empirical process

$$S_n(t) = n^{-1/2}\sum_{j=1}^{n} W_j\mathbf{1}[\sigma_0 e_j \le t], \qquad t \in \mathbb{R},$$

with weights

$$W_j = \frac{\omega(X_j) - E\big[\omega(X_j)\big]}{\sqrt{\mathrm{Var}[\omega(X_j)]}} = \frac{\omega(X_j) - E\big[\omega(X_j)\big]}{\sqrt{E\big[\big\{\omega(X_j) - E\big[\omega(X_j)\big]\big\}^2\big]}}, \qquad j = 1, \ldots, n. \tag{1.2}$$

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 $\sigma(X_j)e_j \neq \sigma_0 e_j$ (and the weights W_j and the errors $\sigma(X_j)e_j$ 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.

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 give reasons for this behaviour at the end of Section 2, where we also construct weights based on an estimate $\hat{\sigma}(\cdot)$ of $\sigma(\cdot)$.

We are interested in both the case when all data are completely observed, the "full model", and the case when responses Y are missing at random (MAR), the "MAR model". Here the observed data 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 $\pi(\cdot) \equiv 1$.

In this article, we will show that our test statistics T_n, defined in (2.1) 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 (given in 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. This means that inference based on the limiting distribution of the test statistic does not depend on parameters of the underlying distribution. 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), i.e. 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.

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

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

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 densities and c_n is a bandwidth. 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 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

$$T_n = \sup_{t\in\mathbb{R}} \left| n^{-1/2} \sum_{j=1}^{n} \hat{W}_j \mathbf{1}\big[\hat{\varepsilon}_j \le t\big] \right| \tag{2.1}$$

with

$$\hat{W}_j = \left\{ \omega(X_j) - \frac{1}{n}\sum_{k=1}^{n} \omega(X_k) \right\} \Bigg/ \left[\frac{1}{n}\sum_{m=1}^{n} \left\{ \omega(X_m) - \frac{1}{n}\sum_{k=1}^{n} \omega(X_k) \right\}^2 \right]^{1/2}, \quad \omega \in \Sigma, \tag{2.2}$$

for $j = 1, \ldots, n$. The term in absolute brackets of (2.1) is an approximation (under H_0) of the process $S_n(t)$ from the Introduction, now with the standardised weights W_j from (1.2) 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.

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 densities $w_1, \ldots, w_m$ 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*

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

with $\hat{W}_j$ specified in (2.2) above, converges in distribution to $\sup_{t\in[0,1]} |B_0(t)|$, where B_0 denotes the standard Brownian bridge.

The proof of Theorem 1 is given in Section 5. We note the distribution of $\sup_{t\in[0,1]} |B_0(t)|$ is a standard distribution, whose upper α-quantiles b_α can be calculated using the formula

$$\alpha = P\bigg(\sup_{t\in[0,1]} \big|B_0(t)\big| > b_\alpha \bigg) = \exp\big(-2b_\alpha^2\big),$$

i.e. $b_\alpha = (\log \alpha^{-1/2})^{1/2}$; see page 37 of Shorack and Wellner (2009). For example, the critical value of a 5% level test is approximately 1.224.

REMARK 1 (POWER OF THE TEST). *To derive the power of the test under local alternatives of the form $\sigma_{n\Delta} = \sigma_0 + n^{-1/2}\Delta$, $\Delta \in \Sigma$ 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. Here $F_{n\Delta}(t) = P\big(\{\sigma_0 + n^{-1/2}\Delta(X)\}e \le t\big)$. A brief sketch gives*

$$E\left[T_n \log\left(\frac{dF_{n\Delta}}{dF} \right) \right] = E\left[n^{-1/2} \sum_{j=1}^{n} \left\{ W_j \mathbf{1}\big[\sigma_0 e_j \le t\big] \right\} \left\{ n^{-1/2} \Delta(X_j) \frac{f'(\sigma_0 e_j)}{f(\sigma_0 e_j)} \right\} \right] + o_p(1)$$

$$= E(W\Delta)\int_{-\infty}^{t}\frac{f'(s)}{f(s)}\,F(ds) + o_p(1)$$
$$= f(t)E(W\Delta) + o_p(1).$$

Hence, under a contiguous alternative H_a, the distribution of the test statistic T_n limits to $\sup_{t\in[0,1]}|B_0(t) + \{f\circ F^{-1}(t)\}E(W\Delta)|$, 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 1, we find the power of our test increases with $E(W\Delta)$. So it can be expected that our test will perform best when ω is a linear transformation of the scale function σ. This suggests simply using an estimator $\hat{\sigma}$ of the scale function σ in order to obtain a powerful test. We expect that this will not change the asymptotic distribution of the test statistic under the null hypothesis.

We have studied this more closely, assuming that σ is in the same Hölder class as r, that is, $\sigma \in H(d,\gamma)$. Then we can estimate σ by a local polynomial estimator $\hat{\sigma}(x) = \{\hat{r}_2(x) - \hat{r}^2(x)\}^{1/2}$. Here $\hat{r}_2$ is a local polynomial estimate of the second conditional moment $E(Y^2|X)$ of Y given X, which is defined in the same way as $\hat{r}$, but with Y_j replaced by Y_j^2. Our estimated weights are then given by

$$\tilde{W}_j = \left\{\hat{\sigma}(X_j) - \frac{1}{n}\sum_{k=1}^{n}\hat{\sigma}(X_k)\right\}\Bigg/\left[\frac{1}{n}\sum_{m=1}^{n}\left\{\hat{\sigma}(X_m) - \frac{1}{n}\sum_{k=1}^{n}\hat{\sigma}(X_k)\right\}^2\right]^{1/2} \tag{2.3}$$

for $j = 1,\ldots,n$. Using similar Donsker class arguments as in the proofs of Theorem 1 and Lemma 2 in Section 5, it is straightforward but lengthy to verify that the asymptotic statements from Theorem 1 continue to hold for this choice of weights. We therefore omit the proofs. The formal result is given in Theorem 2 below. The last part of Theorem 2 concerning the power of the test follows from Remark 1.

THEOREM 2. *Suppose the assumptions of Theorem 1 are satisfied with, additionally, the error variable $\sigma_0 e$ having a finite moment of order greater than $8s/(2s-m)$. Assume the alternative hypothesis restricts $\sigma(\cdot)$ to the Hölder class $H(d,\gamma)$, with $H(d,\gamma)$ as in Theorem 1. Then, under the null hypothesis,*

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

with $\tilde{W}_j$ specified in (2.3) above, converges in distribution to $\sup_{t\in\mathbb{R}}|B_0(t)|$, where B_0 denotes the standard Brownian bridge, and $\tilde{T}_n$ is asymptotically most powerful for detecting alternative hypotheses of the form $\sigma_0 + n^{-1/2}\Delta_d(\cdot)$, where $\Delta_d \in H(d,\gamma)$.

3. Responses missing at random

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

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

Here $N = \sum_{j=1}^{n}\delta_j$ is the number of complete cases and $\hat{W}_{j,c}$ denotes the weights from equation (2.2) 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

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

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 $\sup_{t\in[0,1]} |B_0(t)|$, i.e. 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 2. *Following the discussions above and preceding Theorem 2 in the previous section, we can construct estimated weights based on complete cases as follows. The second conditional moment of Y given X can be estimated by a complete case estimator $\hat{r}_{2,c}$, which is computed in the same way as $\hat{r}_c$, but now with Y_j replaced by Y_j^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*

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

where the weights $\tilde{W}_{j,c}$ are complete case versions of $\tilde{W}_j$; see (2.3). 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.

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 $j = 1,\dots,n$. 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, sample sizes 50, 100, 200 and 1000, and bandwidths $c_n \sim \{n\log(n)\}^{-1/4}$.

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:

$$\sigma_0(x) = 1, \qquad \sigma_1(x) = 0.4 + 4x^2,$$
$$\sigma_2(x) = \frac{e^2-5}{e^2-1} + 4\frac{e^x}{e-e^{-1}} \quad \text{and} \quad \sigma_3(x) = 1 + 15\frac{|x|}{\sqrt{n}}.$$

The constant scale function σ_0 allows for the (5%) level of the test to be checked. 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 of the test is maximised if ω equals the scale function σ (or is a linear transformation of σ), as explained at the end of Section 2. We constructed estimated weights based on the assumption that σ_1, σ_2 and σ_3 are all continuously differentiable and their derivatives satisfy a Hölder condition. This allows us to construct suitable local-constant estimators, which we use in our weights. The bandwidth is chosen automatically by the function *loess* in *R*. We chose $\hat{\sigma}$ as the the square root of the variance estimate at each value of the covariate;

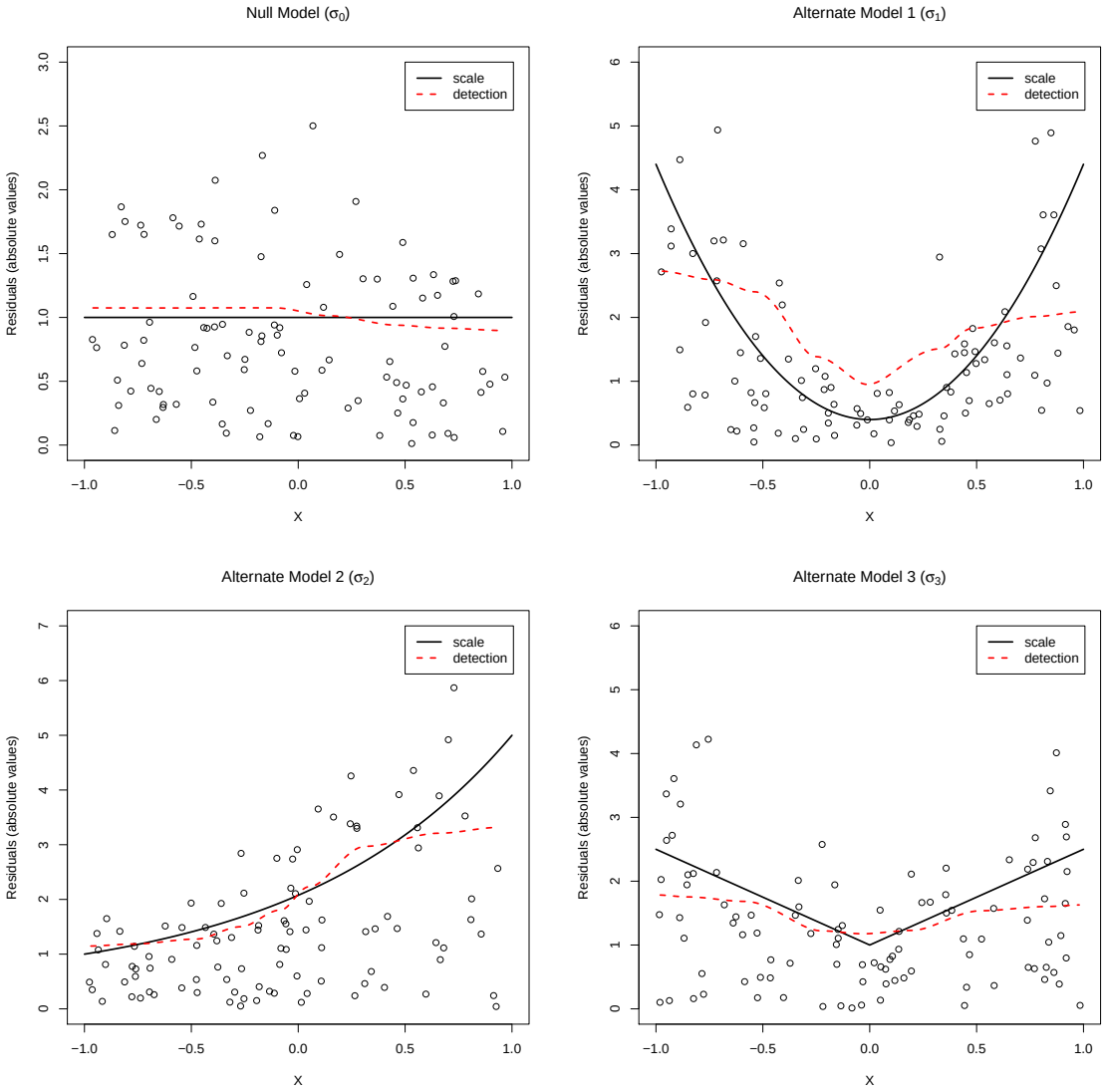

FIGURE 1. Scatter plots of absolute valued residuals. Each plot also shows the underlying scale function and a kernel smoothed estimate of the scale function.

see Remark 1 in Section 2 and the discussion following it. The critical value for the 5% level of each test is approximately 1.224.

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 plots also show the underlying scale functions in black (solid line) and estimated scale functions in red (dashed line).

To check the performance of our test we conducted simulations of 1000 runs. Table 1 shows the test results using T_n (fully observed data) and $T_{n,c}$ (missing data). The figures corresponding to the null hypothesis (σ_0) show Type I error rates near the desired 5%. The results for the test using $T_{n,c}$ are more conservative than the results for full model based on T_n.

Test for heteroskedastic errors								
	T_n				$T_{n,c}$			
n	50	100	200	1000	50	100	200	1000
σ_0	0.054	0.050	0.047	0.033	0.046	0.045	0.037	0.023
σ_1	0.565	0.979	1.000	1.000	0.098	0.511	0.969	1.000
σ_2	0.141	0.643	0.993	1.000	0.031	0.160	0.595	1.000
σ_3	0.079	0.190	0.364	0.724	0.007	0.047	0.128	0.287

TABLE 1. Simulated level (σ_0 figures) and power for T_n and $T_{n,c}$.

We now consider the power of each test. The figures corresponding to σ_1 and σ_2 each show the procedure T_n to perform well at moderate and larger sample sizes. Using T_n, we rejected the null hypothesis 100% (σ_1) and 99.3% (σ_2) of the time for samples of size 200. Similar results were obtained for the test using $T_{n,c}$, but they are (as expected) less impressive. For the figures corresponding to σ_3, both test procedures have difficulty rejecting in small samples. Using T_n, we rejected the null hypothesis 72.4% of the time for samples of size 1000, but only 7.9% of the time for samples of size 50. Again, the results are similar for missing data. In conclusion, each test performs well and the procedures T_n and $T_{n,c}$ proposed in this article appear particularly promising for detecting heteroskedasticity.

It seems the test is affected by the amount of smoothing used to construct the regression function estimator. If the data are under-smoothed, the regression estimate is attempting to explain too much of the model (the errors as well as the underlying regression function). In this case the estimate shows a large variation and residuals will have smaller than expected magnitudes. The test will then be conservative because the residual-based empirical distribution function will have lighter tails than if more smoothing were used. However, if the data are over-smoothed then the regression estimate does not explain enough of the model. In this case the estimate has a large bias and residuals will have larger than expected magnitudes. The test will then be liberal because the residual-based empirical distribution function will have heavier tails than it would have if less smoothing were used. Since the bandwidth is $c_n = c\{n\log(n)\}^{-1/4}$, we considered constants c between 1.5 (chosen so that $c_{100} \approx 1/3$) and 2.5. Using only a multiplier of 1.5 for each sample size produced similar results to those of Table 1 above, but they were more conservative (due to under-smoothing the regression function). Using only a multiplier of 2.5 for each sample size yielded figures similar to those of Table 1, but they were more liberal (due to over-smoothing the regression function).

Example 2: testing for heteroskedasticity with two covariates. Throughout this example we work with the regression function

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

which again preserves the nonparametric nature of the study. The covariates X_1 and X_2 are each independently generated from a uniform distribution on the interval $[-1, 1]$. As above we generate the model errors from a standard normal distribution. We do not consider missing data, because we expect the conclusions to mirror those of the first simulation study. Here we are interested in the differences in performance of our test T_n, for the full model, when we select different weights. We work with $d = 3$, the locally cubic smoother, and bandwidths $c_n = c\{n\log(n)\}^{-1/8}$. The level of the test is 5% as in Example 1 above.

Test for heteroskedastic errors									
	σ_0			σ_1			σ_2		
n	ω_1	ω_2	ω_3	ω_1	ω_2	ω_3	ω_1	ω_2	ω_3
100	0.002	0.008	0.005	0.707	0.146	0.240	0.808	0.179	0.287
200	0.003	0.009	0.006	1.000	0.520	0.963	1.000	0.524	0.971
500	0.009	0.016	0.021	1.000	0.977	1.000	1.000	0.989	1.000
1000	0.036	0.040	0.045	1.000	1.000	1.000	1.000	1.000	1.000

TABLE 2. Simulated level (σ_0 figures) and power for T_n regressing on two covariates.

For the simulations we use three scale functions: $\sigma_0 \equiv 1$, $\sigma_1(x_1, x_2) = 0.5 + 5x_1^2 + 5x_2^2$ and $\sigma_2(x_1, x_2) = 0.5 + 5x_1^2 + 5x_2^2 + 2.5(x_1x_2)^2$. Our weights are constructed based on detection functions $\omega_1 = \sigma_1$, $\omega_2(x_1, x_2) = 1 + \cos\{(\pi/2)(x_1 + x_2)\}$ and ω_3 is an estimated scale function similar to the procedure in the first example. We choose the constant in the bandwidth c_n of the locally cubic smoother to be $c = 5$ for the two cases of known (fixed) weights (ω_1 and ω_2) and $c = 4$ for the case of estimated weights (ω_3). The estimated weights are based on a kernel smoothing of the squared residuals of each nonparametric regression and also require choosing a bandwidth (d_n). A practical choice is one that minimises the asymptotic mean squared error. We choose a product of tricubic kernels with a single bandwidth $d_n = 3n^{-1/6}$ (see, for example, Härdle and Müller, 2000), which is different from the bandwidths used for the locally cubic smoother.

From the discussion above it is clear that $\omega_1 = \sigma_1$ will provide the largest power for detecting σ_1 but not necessarily for detecting σ_2. The choices ω_2 and ω_3 will then illustrate the test performance when we choose (or guess) some reasonable non-constant detection function and when we use an estimator of the scale function to increase the power of the test.

We conducted simulations consisting of 1000 runs, now using sample sizes 100, 200, 500 and 1000. The results are displayed in Table 2. The figures in the left panel ($\sigma_0 \equiv 1$) corresponding to the test level (5%) show the tests are all highly conservative and only reach adequate levels at samples of size 1000. Nevertheless, when we consider the figures in the remaining panels, corresponding to the powers of each test, we find considerable differences between the tests. It is clear that testing using $\omega_1 = \sigma_1$ provides the best results in the second column referring to σ_1 (best weights). Since σ_1 and σ_2 are similar in shape, the results for the test based on $\omega_1 = \sigma_1$ are also quite convincing when σ_2 is the underlying scale function (third column). Testing using ω_3 (estimated weights) provides comparable results to those of ω_1. The only notable difference between the two procedures occurs when the sample size is small (100 observations). Here we find the test using ω_1 gives powers 0.707 (σ_1) and 0.808 (σ_2) while the test using ω_3 only gives powers 0.240 (σ_1) and 0.287 (σ_2). This difference in behavior is expected.

When we consider the test using ω_2, we see a considerable decrease in power at smaller sample sizes. At 200 observations, the tests using ω_1 and ω_3 already have powers at or near 1.000, but the test using ω_2 only gives powers 0.520 (σ_1) and 0.524 (σ_2). Only at very large sample sizes are all three testing procedures similar. In conclusion, we find the test using an arbitrary non-constant weight function is useful but will normally be outperformed by the test using estimated weights that attempt to optimise the power of the test. All three test procedures are fairly conservative.

5. Technical details

In this section we present the proof of Theorem 1 and some auxiliary results. As explained earlier, 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 (i.e. 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

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

where $\sigma_0 e_j$ is the unobserved "model error" from the null hypothesis and $W_1, \dots, W_n$ are the standardised weights given in (1.2). We will now demonstrate that 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

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

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 (1.2).

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*

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

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

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

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. Note: an unweighted version of that difference is considered in the proof of Theorem 2.2 of Müller, Schick and Wefelmeyer (2007).

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. In the following we will write, for any function g from $[0,1]^m$ to $\mathbb{R}$, $\|g\|_{x,\infty} = \sup_{x\in[0,1]^m} |g(x)|$ and, for any function h from $\mathbb{R}$ to $\mathbb{R}$, $\|h\|_{t,\infty} = \sup_{t\in\mathbb{R}} |h(t)|$. We begin by noting that the assumptions of Lemma 1 are satisfied. Hence, by property (5.3) of Lemma 1, there is a random function $\hat{a}$ which approximates $\hat{r} - r$ and therefore still depends on the data $\mathbb{D} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. Also there is an $\alpha > 0$ such that $P(\hat{a} \in H_1(m, \alpha)) \to 1$. We will first show an auxiliary statement, namely that the (simpler) class of functions

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

is $G \otimes F$–Donsker. To prove this, it suffices to verify Dudley's entropy integral condition:

$$\int_0^\infty \sqrt{\log N_{[]}\big((\epsilon, \mathfrak{F}, L_2(G \otimes F)\big)}\, d\epsilon < \infty$$

(see Theorem 2.5.6 of van der Vaart and Wellner, 1996). Here $N_{[]}(\epsilon, \mathfrak{F}, L_2(G \otimes F))$ is the number of brackets of length no greater than ϵ required to cover $\mathfrak{F}$ and $L_2(G \otimes F)$ is the L_2–norm with respect to the measure $G \otimes F$. Bracketing numbers are a measure of the amount of entropy residing in the class $\mathfrak{F}$.

We will proceed similarly to the proof of Lemma A.1 of Van Keilegom and Akritas (1999) and find a suitable function of ϵ that satisfies the above integral condition concerning the the bracketing numbers $N_{[]}(\epsilon, \mathfrak{F}, L_2(G \otimes F))$. Let $\epsilon > 0$. Since functions in $\mathfrak{F}$ are sums of two terms, we will show the first space $\mathfrak{F}_1 = \{(X, \sigma_0 e) \mapsto W\mathbf{1}[\sigma_0 e \le t + a(X)] \,:\, t \in \mathbb{R},\ a \in H_1(m, \alpha)\}$ satisfies the integral condition above. The proof for the second space $\mathfrak{F}_2$ is similar and therefore omitted. By Theorem 2.7.1 of van der Vaart and Wellner (1996) it follows, for some positive constant K, that $n_r = N_{[]}(\epsilon^2/(2\|f\|_{t,\infty} E[W^2]), H_1(m, \alpha), \|\cdot\|_{x,\infty}) \le \exp\{K\epsilon^{-(2m)/(m+\alpha)}\}$. Let $a_{l1} \le a_{u1}, \ldots, a_{ln_r} \le a_{un_r}$ be the functions defining the n_r brackets for $H_1(m, \alpha)$ based on the $\|\cdot\|_{x,\infty}$–norm. It then follows for the brackets $a_{li} \le a_{ui}$ to satisfy $E[W^2\{a_{ui}(X) - a_{li}(X)\}] \le \epsilon^2/(2\|f\|_{t,\infty})$, $i = 1, \ldots, n_r$.

The random variable W can be either positive–valued or negative–valued. We can then construct brackets for $\mathbb{R}$ based on this information and the n_r brackets for a as follows. For every fixed $t \in \mathbb{R}$ and $i = 1, \ldots, n_r$, we have, writing $W^- = W\mathbf{1}[W < 0]$ and $W^+ = W\mathbf{1}[W \ge 0]$,

$$W^-\mathbf{1}\big[\sigma_0 e \le t + a_{ui}(X)\big] \le W^-\mathbf{1}\big[\sigma_0 e \le t + a(X)\big] \le W^-\mathbf{1}\big[\sigma_0 e \le t + a_{li}(X)\big]$$

and

$$W^+\mathbf{1}\big[\sigma_0 e \le t + a_{li}(X)\big] \le W^+\mathbf{1}\big[\sigma_0 e \le t + a(X)\big] \le W^+\mathbf{1}\big[\sigma_0 e \le t + a_{ui}(X)\big].$$

Now define $F_{li}(t) = F(t + a_{li}(X))$ for the conditional probability that $\sigma_0 e$ is at most $t + a_{li}(X)$ given the covariates X. Further, let t_{lij_1}, for $j_1 = 1, \ldots, O(\epsilon^{-2})$, partition $\mathbb{R} \cup \{-\infty, \infty\}$ into segments having F_{li}–probability at most $\epsilon^2/(4E[W^2])$. Similarly, define $F_{ui}(t) = F(t + a_{ui}(X))$ and let t_{uij_2}, for $j_2 = 1, \ldots, O(\epsilon^{-2})$, partition $\mathbb{R} \cup \{-\infty, \infty\}$ into segments having F_{ui}–probability at most $\epsilon^2/(4E[W^2])$. Hence, we obtain the following bracket for t: $t^-_{lij_1} \leq t \leq t^+_{uij_2}$, where $t^-_{lij_1}$ is the largest t_{lij_1} that is less than or equal to t and $t^+_{uij_2}$ is the smallest t_{uij_2} that is larger than or equal to t.

We will now show the brackets for $\mathfrak{F}_1$ are given by

$$\begin{aligned}
&W^-\mathbf{1}\big[\sigma_0 e \leq t^+_{uij_1} + a_{ui}(X)\big] + W^+\mathbf{1}\big[\sigma_0 e \leq t^-_{lij_1} + a_{li}(X)\big] \\
&\leq W\mathbf{1}\big[\sigma_0 e \leq t + a(X)\big] \\
&\leq W^-\mathbf{1}\big[\sigma_0 e \leq t^-_{lij_1} + a_{li}(X)\big] + W^+\mathbf{1}\big[\sigma_0 e \leq t^+_{uij_2} + a_{ui}(X)\big].
\end{aligned}$$

The squared length of the proposed brackets above is equal to

$$\begin{aligned}
&E\bigg[\Big(W^-\Big\{\mathbf{1}\big[\sigma_0 e \leq t^-_{lij_1} + a_{li}(X)\big] - \mathbf{1}\big[\sigma_0 e \leq t^+_{uij_2} + a_{ui}(X)\big]\Big\} \\
&\qquad + W^+\Big\{\mathbf{1}\big[\sigma_0 e \leq t^+_{uij_2} + a_{ui}(X)\big] - \mathbf{1}\big[\sigma_0 e \leq t^-_{lij_1} + a_{li}(X)\big]\Big\}\Big)^2\bigg] \\
&= E\Big[W^2\Big\{\mathbf{1}[W < 0] + \mathbf{1}[W \geq 0]\Big\}\Big\{\mathbf{1}\big[\sigma_0 e \leq t^+_{uij_2} + a_{ui}(X)\big] - \mathbf{1}\big[\sigma_0 e \leq t^-_{lij_1} + a_{li}(X)\big]\Big\}\Big] \\
&= E\Big[W^2\Big\{F_{ui}\big(t^+_{uij_2}\big) - F_{li}\big(t^-_{lij_1}\big)\Big\}\Big],
\end{aligned}$$

which is bounded by

$$E\big[W^2\big\{F_{ui}(t) - F_{li}(t)\big\}\big] + \frac{\epsilon^2}{2}.$$

Consider the first term. Since the distribution function F has a bounded density function f, the inequality above for $E[W^2\{a_{ui}(X) - a_{li}(X)\}^2]$ implies

$$E\big[W^2\{F_{ui}(t) - F_{li}(t)\}\big] \leq \|f\|_{t,\infty} E\big[W^2\{a_{ui}(X) - a_{li}(X)\}\big] \leq \frac{\epsilon^2}{2}.$$

Hence, the $L_2(G \otimes F)$–lengths of our proposed brackets are bounded by ϵ as desired. It then follows, for every $\epsilon > 0$, that the number of brackets is at most $O(\epsilon^{-2} \exp\{K\epsilon^{-(2m)/(m+\alpha)}\})$ and, for $\epsilon > 1$, one bracket suffices. Therefore, we can choose appropriate positive constants C_1 and C_2 to find

$$\int_0^\infty \sqrt{\log N_{[]}\big(\epsilon, \mathfrak{F}_1, L_2(G \otimes F)\big)}\, d\epsilon = \int_0^1 \sqrt{\log N_{[]}\big(\epsilon, \mathfrak{F}_1, L_2(G \otimes F)\big)}\, d\epsilon \leq C_1 + C_2 \frac{m + \alpha}{\alpha}.$$

Since $\alpha > 0$, the bound above is finite and Dudley's entropy integral condition holds. This shows the class $\mathfrak{F}_1$ is $G \otimes F$–Donsker, which combined with the statement for $\mathfrak{F}_2$ implies the class $\mathfrak{F}$ is $G \otimes F$–Donsker.

It 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 $\varphi > 0$,

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

We are interested in the case that involves the approximation $\hat{a}$ in place of a, where 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 the following negligibility condition on the variance guarantees that the extended class of processes involving $\hat{a}$ is also equicontinuous. To prove that variance condition, we fix the function $\hat{a}$ by conditioning on the observed data $\mathbb{D}$. The variation of a function from the extension of $\mathfrak{F}$, i.e. now involving $\hat{a}$ instead of a, is equal to

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

which is bounded by

$$
\begin{aligned}
&E\Big[W^2\Big\{F\big(\max\{t,\, t + \hat{a}(X)\}\big) - F\big(\min\{t,\, t + \hat{a}(X)\}\big)\Big\}\,\Big|\,\mathbb{D}\Big] \\
&\le \|f\|_{t,\infty} E\big[W^2\big(\max\{t,\, t + \hat{a}(X)\} - \min\{t,\, t + \hat{a}(X)\}\big)\,\big|\,\mathbb{D}\big] \\
&= \|f\|_{t,\infty} E\big[W^2\big|\hat{a}(X)\big|\,\big|\,\mathbb{D}\big] \\
&\le \|f\|_{t,\infty} E\big[W^2\big]\,\|\hat{a}\|_{x,\infty} \\
&= o_p(1),
\end{aligned}
$$

i.e. the variance is asymptotically negligible. Here we used $\|\hat{a}\|_{x,\infty} = o_p(1)$; see page 961 of the proof of Lemma 1 in Müller et al. (2009). Hence we have asymptotic equicontinuity and therefore

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

We now decompose R_1 from the first assertion as the sum of

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

and

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

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

We have already shown the first term is $o_p(n^{-1/2})$, uniformly in $t \in \mathbb{R}$. By property (5.4) of Lemma 1, $A_n = \|\hat{r}-r-\hat{a}\|_{x,\infty} = o_p(n^{-1/2})$. The decomposition of W_j into $W_j^- + W_j^+$, $j = 1,\dots,n$, yields the following bounds for the weighted indicator functions, for each $j = 1,\dots,n$:

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

and

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

This implies that we can find a bound for (5.6) by calculating

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

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

Hence, the first assertion follows from showing

$$
\begin{aligned}
\sup_{t\in\mathbb{R}}\Bigg|&\frac{1}{n}\sum_{j=1}^{n} \big|W_j\big|\Big\{\mathbf{1}\big[\sigma_0 e_j \le t + A_n + \hat a(X_j)\big] - F\big(t + A_n + \hat a(X_j)\big)\Big\}\\
&-\frac{1}{n}\sum_{j=1}^{n} \big|W_j\big|\Big\{\mathbf{1}\big[\sigma_0 e_j \le t - A_n + \hat a(X_j)\big] - F\big(t - A_n + \hat a(X_j)\big)\Big\}\Bigg| = o_p(n^{-1/2}),
\end{aligned}
\tag{5.7}
$$

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

and

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

Beginning with (5.7), since the random variables $|W_1|,\ldots,|W_n|$ are square integrable, the class of functions

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

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 consider the variation condition under the norm in (5.5), now for the function $(X,\sigma_0 e)\mapsto|W|\{\mathbf{1}[\sigma_0 e\le t_n+a(X)\}-\mathbf{1}[\sigma_0 e\le s_n+a(X)]-F(t_n+a(X))+F(s_n+a(X))$, which is equal to

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

This is bounded by

$$\begin{aligned}
&E\left[W^2\Big\{F\big(\max\{t_n+a(X),\,s_n+a(X)\}\big)-F\big(\min\{t_n+a(X),\,s_n+a(X)\}\big)\Big\}\right]\\
&\le\|f\|_{t,\infty}E\big[W^2\big]\big|t_n-s_n\big|.
\end{aligned}$$

Since the bound above is $o(1)$, equicontinuity now implies, 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 $|t_n-s_n|=o(1)$,

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

As before we may assume that $\hat{a}$ belongs to $H_1(m,\alpha)$. Now conditioning on $\mathbb{D}$ and letting $T_n=t+A_n$ and $S_n=t-A_n$, which now satisfies $|T_n-S_n|=2A_n=o_p(1)$, we find

$$\begin{aligned}
&\mathrm{Var}\Big[|W|\Big\{\mathbf{1}\big[\sigma_0 e\le T_n+\hat{a}(X)\big]-\mathbf{1}\big[\sigma_0 e\le S_n+\hat{a}(X)\big]\\
&\qquad\qquad-F\big(T_n+\hat{a}(X)\big)-F\big(S_n+\hat{a}(X)\big)\Big\}\,\Big|\,\mathbb{D}\Big]\\
&\le 2\|f\|_{t,\infty}E\big[W^2\big]A_n.
\end{aligned}$$

Since $A_n = o_p(1)$, in the same way as before, with T_n and S_n conditional on $\mathbb{D}$ playing the roles of t_n and s_n above, the negligibility condition on the variance is satisfied. This combined with the fact that $\mathfrak{F}^+$ is a Donsker class (and the corresponding class of processes is equicontinuous) therefore implies that (5.7) is satisfied.

Turning our attention now to (5.8), we find that $n^{-1}\sum_{j=1}^n |W_j|$ is consistent for $E|W|$, and we have $E|W| \leq E^{1/2}[W^2] < \infty$. Since both $|W_j|$ and $F(t+A_n+\hat{a}(X_j)) - F(t-A_n+\hat{a}(X_j))$ are positive-valued for each $j = 1, \ldots, n$, we can find a bound for (5.8) by calculating

$$\sup_{t\in\mathbb{R}} \frac{1}{n}\sum_{j=1}^n \big|W_j\big|\Big\{F\big(t+A_n+\hat{a}(X_j)\big) - F\big(t-A_n+\hat{a}(X_j)\big)\Big\} \leq 2\|f\|_{t,\infty}A_n\frac{1}{n}\sum_{j=1}^n \big|W_j\big|.$$

Since $A_n = o_p(n^{-1/2})$ by Lemma 1 and since $n^{-1}\sum_{j=1}^n |W_j|$ is consistent for $E|W|$, we obtain for the bound above to also be $o_p(n^{-1/2})$, i.e. (5.8) holds.

We can bound the left-hand side of (5.9) by $\|f\|_{t,\infty}A_n n^{-1}\sum_{j=1}^n |W_j|$. Using again $A_n = o_p(n^{-1/2})$ and the consistency of $n^{-1}\sum_{j=1}^n |W_j|$, this bound is $o_p(n^{-1/2})$. Therefore (5.9) holds, which concludes the proof of the first assertion that $\|R_1\|_{t,\infty} = o_p(n^{-1/2})$.

We will now prove the second assertion that $\|R_2\|_{t,\infty} = o_p(n^{-1/2})$. In the same way as above, we will incorporate the approximation $\hat{a}$ to separate the stochastic process into two parts and then argue each part is negligible at the $n^{-1/2}$ rate of convergence. The main difference between the proof technique used above, where we additionally required separating the process using the decomposition of the weights into their respective positive and negative components, and the technique used now comes from one remainder term: we now require that the random variables $W_1, \ldots, W_n$ each have mean zero, which allows us to use the central limit theorem in combination with the result $\|\hat{a}\|_{x,\infty} = o_p(1)$. This means we can write R_2 as

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

This shows that $\|R_2\|_{t,\infty}$ is bounded by three terms:

$$\sup_{t\in\mathbb{R}}\left|\frac{1}{n}\sum_{j=1}^n W_j\Big\{F\big(t+\hat{a}(X_j)\big) - F(t) - E\Big[F\big(t+\hat{a}(X)\big) - F(t)\,\Big|\,\mathbb{D}\Big]\Big\}\right|, \tag{5.10}$$

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

and the third term is the left–hand side of (5.9), which we have already shown is $o_p(n^{-1/2})$. From the arguments above, it follows for the class of functions

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

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

$$2E\big[W^2\big]\Big\{F\big(t+\|\hat{a}\|_{x,\infty}\big)-F(t)\Big\}^2\le 2\|f\|_{t,\infty}^2E\big[W^2\big]\|\hat{a}\|_{x,\infty}^2.$$

Since we have already used that $\|\hat{a}\|_{x,\infty}=o_p(1)$, the bound above is $o_p(1)$, i.e. the variance is asymptotically negligible. This combined with equicontinuity implies that the term in (5.10) has rate $o_p(n^{-1/2})$, as desired.

Finally we can bound (5.11) by

$$\|f\|_{t,\infty}\|\hat{a}\|_{x,\infty}\left|\frac{1}{n}\sum_{j=1}^{n}W_j\right|.$$

The central limit theorem combined with $E[W_j]=0$ $j=1,\dots,n$, gives $|\frac{1}{n}\sum_{j=1}^{n}W_j|=O_p(n^{-1/2})$. Since $\|\hat{a}\|_{x,\infty}=o_p(1)$ this shows that the bound above, and also (5.11), is of order $o_p(n^{-1/2})$. This completes the proof of the second assertion that $\|R_2\|_{t,\infty}=o_p(n^{-1/2})$. □

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),

$$n^{-1/2}\sum_{j=1}^{n}W_j\Big\{\mathbf{1}[\sigma_0e_j\le t]-F(t)\Big\},$$

with W_j given in (1.2). 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

$$\sup_{t\in\mathbb{R}}\left|\frac{1}{n}\sum_{j=1}^{n}\hat{W}_j\mathbf{1}\big[\hat{\varepsilon}_j\le t\big]-\frac{1}{n}\sum_{j=1}^{n}W_j\Big\{\mathbf{1}[\sigma_0e_j\le t]-F(t)\Big\}\right|=o_p(n^{-1/2}). \tag{5.12}$$

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

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

It follows from $\sum_{j=1}^n \hat{W}_j = 0$ that we can decompose the difference in (5.12) 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,

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

$$\begin{aligned} R_4 = &\left(\left(\frac{\mathrm{Var}[\omega(X_1)]}{\frac{1}{n}\sum_{j=1}^n \{\omega(X_j) - \frac{1}{n}\sum_{k=1}^n \omega(X_k)\}^2} \right)^{1/2} - 1 \right) \\ &\times \left(\frac{1}{n}\sum_{j=1}^{n} W_j \Big\{ \mathbf{1}\big[\sigma_0 e \le t + \hat{r}(X_j) - r(X_j)\big] - F\big(t + \hat{r}(X_j) - r(X_j)\big) \Big\} \right), \end{aligned}$$

$$\begin{aligned} R_5 = &\left(\frac{\mathrm{Var}[\omega(X_1)]}{\frac{1}{n}\sum_{j=1}^n \{\omega(X_j) - \frac{1}{n}\sum_{k=1}^n \omega(X_k)\}^2} \right)^{1/2} \left(\frac{1}{n}\sum_{j=1}^{n} W_j \right) \\ &\times \left(\frac{1}{n}\sum_{j=1}^{n} \Big\{ \mathbf{1}\big[\sigma_0 e \le t + \hat{r}(X_j) - r(X_j)\big] - F\big(t + \hat{r}(X_j) - r(X_j)\big) \Big\} \right), \end{aligned}$$

and

$$\begin{aligned} R_6 = &\left(\frac{\mathrm{Var}[\omega(X_1)]}{\frac{1}{n}\sum_{j=1}^n \{\omega(X_j) - \frac{1}{n}\sum_{k=1}^n \omega(X_k)\}^2} \right)^{1/2} \left(\frac{1}{n}\sum_{j=1}^{n} W_j \right) \\ &\times \left(\frac{1}{n}\sum_{j=1}^{n} \Big\{ F\big(t + \hat{r}(X_j) - r(X_j)\big) - F(t) \Big\} \right). \end{aligned}$$

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

$$\sup_{t\in\mathbb{R}} \left| \frac{1}{n}\sum_{j=1}^{n} W_j \Big\{ \mathbf{1}\big[\sigma_0 e_j \le t + \hat{r}(X_j) - r(X_j)\big] - F\big(t + \hat{r}(X_j) - r(X_j)\big) \Big\} \right| = O_p(n^{-1/2}), \tag{5.13}$$

because the first term of R_4 both does not depend on t and is asymptotically negligible. To verify (5.13), 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. (5.13) holds with $W_j = 1$, $j = 1, \ldots, n$. 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

$$\sup_{t\in\mathbb{R}} |f(t)| \left(\frac{\mathrm{Var}[\omega(X_1)]}{\frac{1}{n}\sum_{j=1}^{n}\{\omega(X_j) - \frac{1}{n}\sum_{k=1}^{n}\omega(X_k)\}^2} \right)^{1/2}$$
$$\times \left(\sup_{x\in[0,1]^m} \left|\hat{a}(x)\right| + \sup_{x\in[0,1]^m} \left|\hat{r}(x) - r(x) - \hat{a}(x)\right| \right) \left| \frac{1}{n}\sum_{j=1}^{n} W_j \right|.$$

The second term in the bound above is 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 $\sup_{t\in\mathbb{R}} |R_6| = o_p(n^{-1/2})$. □

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

[2] Dette, H. and Munk, A. (1998). Testing heteroscedasticity in nonparametric regression. *Journal of the Royal Statistical Society, Series B*, **60**, 693-708.

[3] Dette, H. (2002). A consistent test for heteroscedasticity in nonparametric regression based on the kernel method. *Journal of Statistical Planning and Inference*, **103**, 311-329.

[4] Dette, H., Neumeyer, N. and Van Keilegom, I. (2007). A new test for the parametric form of the variance function in nonparametric regression. *Journal of the Royal Statistical Society, Series B*, **69**, 903-917.

[5] Dette, H. and Hetzler, B. (2009). A simple test for the nonparametric form of the variance function in nonparametric regression. *Annals of the Institute of Statistical Mathematics*, **61**, 861-886.

[6] Eubank, R.L. and Thomas, W. (1993). Detecting heteroscedasticity in nonparametric regression. *Journal of the Royal Statistical Society, Series B*, **55**, 145-155.

[7] Glejser, H. (1969). A new test for heteroskedasticity. *Journal of the American Statistical Association*, **64**, 316-323.

[8] Härdle, W. and Müller M. (2000). Multivariate and semiparametric kernel regression. In: *Smoothing and Regression: Approaches, Computation, and Application* (ed M.G. Schimek), John Wiley & Sons, Hoboken, NJ, USA.

[9] Koul, H.L. (2002). *Weighted Empirical Processes in Dynamic Nonlinear Models.* Lecture Notes in Statistics, Springer, New York.

[10] Koul, H.L., Müller, U.U. and Schick, A. (2012). The transfer principle: a tool for complete case analysis. *Annals of Statistics*, **40**, 3031-3049.

[11] 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.
[12] Little, R.J.A. and Rubin, D.B. (2002). *Statistical analysis with missing data.* Second edition. Wiley-Interscience.
[13] Müller, U.U., Schick A. and Wefelmeyer W. (2007). Estimating the error distribution in semiparametric regression. *Statistics and Decisions*, **25**, 1-18.
[14] Müller, U.U., Schick, A. and Wefelmeyer, W. (2009). Estimating the error distribution function in nonparametric regression with multivariate covariates. *Statistics and Probability Letters*, **79**, 957-964.
[15] Shorack, G.R. and Wellner, J.A. (2009). *Empirical Processes with Applications to Statistics.* Classics in Applied Mathematics. Society for Industrial and Applied Mathematics.
[16] Stute, W. (1997). Nonparametric model checks for regression. *Annals of Statistics*, **25**, 613-641.
[17] Stute, W., Xu, W.L. and Zhu, L.X. (2008). Model diagnosis for parametric regression in high-dimensional spaces. *Biometrika*, **95**, 451-467.
[18] van der Vaart, A.W. (1998). *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
[19] 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.
[20] Van Keilegom, I. and Akritas, M.G. (1999). Transfer of tail information in censored regression models. *Annals of Statistics*, **27**, 1745-1784.
[21] White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, **48**, 817-838.
[22] You, J. and Chen, G. (2005). Testing heteroscedasticity in partially linear regression models. *Statistics and Probability Letters*, **73**, 61-70.