I N S T I T U T D E

S T A T I S T I Q U E

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

D I S C U S S I O N

P A P E R

0405

GOODNESS-OF-FIT TESTS IN PARAMETRIC REGRESSION BASED ON THE ESTIMATION OF THE ERROR DISTRIBUTION

I. VAN KEILEGOM, W. GONZALEZ MANTEIGA and C. SANCHEZ SELLERO

http://www.stat.ucl.ac.be

Goodness-of-fit tests in parametric regression based on the estimation of the error distribution

Ingrid Van Keilegom[1] Wenceslao González Manteiga[2]

Université catholique de Louvain *Universidad de Santiago de Compostela*

César Sánchez Sellero[2]

Universidad de Santiago de Compostela

March 3, 2004

Abstract

Consider a heteroscedastic regression model $Y = m(X) + \sigma(X)\varepsilon$, where $m(X) = E(Y|X)$ and $\sigma^2(X) = \mathrm{Var}(Y|X)$ are unknown and the error ε is independent of the covariate X. We propose a new type of test statistic for testing whether the regression curve $m(\cdot)$ belongs to some parametric family of regression functions. The proposed test statistic measures the distance between the empirical distribution function of the parametric and of the nonparametric residuals. The asymptotic theory of the proposed test is developed and the proposed testing procedure is illustrated by means of a small simulation study and the analysis of a data set.

Key Words: Bootstrap; Goodness-of-fit; Heteroscedastic regression; Model check; Nonlinear regression; Nonparametric regression; Residual distribution.

[1] Financial support from the IAP research network nr. P5/24 of the Belgian government (Belgian Science Policy) is gratefully acknowledged.
[2] Financial support from the Spanish Ministry of Science and Technology (with additional European FEDER support) through project BFM2002-03213 and from Xunta de Galicia through project PGIDIT03PXIC20702PN are gratefully acknowledged.

1 Introduction

Let (X, Y) be a random vector, where Y denotes a possible transformation of the variable of interest and X is a covariate, and assume X and Y satisfy the following heteroscedastic regression model :

$$Y = m(X) + \sigma(X)\varepsilon, \tag{1.1}$$

where the error variable ε is independent of X, $m(X) = E(Y|X)$ is the unknown regression function and $\sigma^2(X) = \text{Var}(Y|X)$ is the conditional variance function. Finally, let (X_i, Y_i), $i = 1, \ldots, n$, denote independent replications of (X, Y).

We like to test the hypothesis

$$H_0 : m \in \mathcal{M} \text{ versus } H_1 : m \notin \mathcal{M}, \tag{1.2}$$

where $\mathcal{M} = \{m_\theta : \theta \in \Theta\}$ is some parametric class of regression functions and $\Theta \subset I\!R^p$.

Over the last ten years the literature related to problem (1.2) has been growing rapidly. Many of the different proposals are based on estimating

$$D^2 = \min_{\theta \in \Theta} \int (m(x) - m_\theta(x))^2 \, dF_X(x),$$

where F_X is the distribution function of X. This idea was carried out either by estimating D^2 directly (see Dette and Munk (1998) or Dette, Munk and Wagner (2000)), or by estimating related quantities, like the difference of the variances under the null and under the alternative (see Dette (1999), for example).

Alternatively, other authors estimate D^2 by

$$\hat{D}^2 = \int (\hat{m}(x) - m_{\hat{\theta}}(x))^2 \, d\hat{F}_X(x),$$

where $\hat{m}$ is a nonparametric estimator of the regression function, $\hat{\theta}$ is an estimator of the parameter and $\hat{F}_X$ is the empirical distribution of X. See, for example, Härdle and Mammen (1993) where $\hat{m}$ is taken to be the Nadaraya-Watson estimator, and Alcalá et al. (1999) where $\hat{m}$ is taken to be a local linear estimator.

Simultaneously, another family of tests was developed on the basis of estimating the integrated regression function, $M(x) = \int_{-\infty}^{x} m(u) \, dF_X(u) = E(Y I_{\{X \le x\}})$, and then dealing with functionals of Kolmogorov-Smirnov or Cramér-von Mises type, defined over the empirical regression process (see Stute (1997)):

$$\sqrt{n}\left(\hat{M}(x) - M_{\hat{\theta}}(x)\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Y_i - m_{\hat{\theta}}(X_i)) \, I_{\{X_i \le x\}}.$$

Other tests based on empirical likelihood ratio have been considered recently (see Fan et al. (2001), Chen et al. (2003)).

One way to compare the different approaches is given by the capacity of the tests to detect contiguous alternatives, $H_0 : m(\cdot) = m_{\theta_0}(\cdot) + b_n r(\cdot)$, where r is a function orthogonal to the space of functions $\mathcal{M} = \{m_\theta : \theta \in \Theta\}$ and $b_n \to 0$.

In this paper a new type of test is introduced based on the distance between the empirical distribution of the parametric residuals (under the null) and that of the nonparametric residuals (in the general situation). We use Kolmogorov-Smirnov and Cramér-von Mises type of statistics to measure the distance between the two empirical distributions. The proposed test can detect alternatives at the rate $n^{-1/2}$, typical for the tests based on empirical regression processes.

The paper is organised as follows. In the next section we will explain the proposed testing procedure. Section 3 contains the theoretical results. In Section 4 we show some simulations to assess the behaviour of the test in several examples, and we apply it to a real data set. Finally, the Appendix contains the proofs of the main results.

2 Proposed test

Consider the random vector (X, Y) satisfying model (1.1). Let $F_\varepsilon(y) = P(\varepsilon \leq y)$, $F(y|x) = P(Y \leq y|X = x)$ and $F_X(x) = P(X \leq x)$. The probability density functions of $F_\varepsilon(y)$ and $F_X(x)$ will be denoted respectively by $f_\varepsilon(y)$ and $f_X(x)$.

We first estimate the distribution of ε in a nonparametric way. Define for x in the support $[a, b]$ of X,

$$\hat{m}(x) = \sum_{i=1}^{n} W_i(x, h_n) Y_i, \quad \hat{\sigma}^2(x) = \sum_{i=1}^{n} W_i(x, h_n) Y_i^2 - \hat{m}^2(x), \tag{2.1}$$

where

$$W_i(x, h_n) = \frac{K\left(\frac{x - X_i}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{x - X_j}{h_n}\right)}$$

are Nadaraya-Watson weights, K is a known probability density function (kernel) and $\{h_n\}$ is an appropriate bandwidth sequence. This leads to

$$\hat{F}_\varepsilon(y) = n^{-1} \sum_{i=1}^{n} I(\hat{\varepsilon}_i \leq y), \tag{2.2}$$

where $\hat{\varepsilon}_i = (Y_i - \hat{m}(X_i))/\hat{\sigma}(X_i)$ are the nonparametric residuals. This estimator has been proposed and studied by Akritas and Van Keilegom (2001).

Next, under H_0 we estimate $m(x)$ by the least squares method for nonlinear regression, i.e. we estimate $m(x)$ by $m_{\hat\theta}(x)$, where $\hat\theta$ is a minimizer (over $\theta \in \Theta$) of the expression

$$S_n(\theta) = n^{-1} \sum_{i=1}^{n} (Y_i - m_\theta(X_i))^2. \tag{2.3}$$

Basic properties of this estimator have been studied by Wu (1981), White (1981, 1982), Seber and Wild (1989) among others. This leads to

$$\hat F_{\varepsilon 0}(y) = n^{-1} \sum_{i=1}^{n} I(\hat\varepsilon_{i0} \le y), \tag{2.4}$$

where $\hat\varepsilon_{i0} = (Y_i - m_{\hat\theta}(X_i))/\hat\sigma(X_i)$ are the residuals estimated under H_0.

The test statistics that we will use are the Kolmogorov-Smirnov type statistic

$$T_{KS} = n^{1/2} \sup_{-\infty<y<\infty} |\hat F_\varepsilon(y) - \hat F_{\varepsilon 0}(y)|$$

and the Cramér-von Mises type statistic

$$T_{CM} = n \int [\hat F_\varepsilon(y) - \hat F_{\varepsilon 0}(y)]^2 \, d\hat F_\varepsilon(y).$$

To study the local power of these statistics, consider the local alternative

$$H_1 : m(\cdot) \equiv m_{\theta_0}(\cdot) + n^{-1/2} r(\cdot),$$

for some function r, and where θ_0 is the true value of θ under H_0.

Define the following additional notation :

$$S_0(\theta) = E[\sigma^2(X)] + E[(m_\theta(X) - m_{\theta_0}(X))^2],$$
$$\Omega = \left\{ E\left[\frac{\partial m_{\theta_0}(X)}{\partial \theta_r} \frac{\partial m_{\theta_0}(X)}{\partial \theta_s} \right] \right\}_{r,s=1,\ldots,p},$$
$$\eta_\theta(x,y) = \Omega^{-1} \frac{\partial m_\theta(x)}{\partial \theta} (y - m_\theta(x)),$$

where

$$\frac{\partial m_\theta(x)}{\partial \theta} = \left(\frac{\partial m_\theta(x)}{\partial \theta_r} \right)_{r=1,\ldots,p}$$

is a $(p \times 1)$-vector and $\theta = (\theta_1, \ldots, \theta_p)'$. Note that $S_0(\theta)$ equals the expected value of $S_n(\theta)$ under H_0.

Remark 2.1 An alternative estimator for $\sigma^2(x)$ under H_0 is given by

$$\hat\sigma^2_{alt}(x) = \sum_{i=1}^{n} W_i(x, h_n) Y_i^2 - m^2_{\hat\theta}(x),$$

which, in contrast to $\hat{\sigma}^2(x)$, makes explicitly use of the form of $m(x)$ under H_0. However, this estimator has the drawback that it can attain negative values. We therefore prefer to work with $\hat{\sigma}^2(x)$.

The main results below require the following regularity conditions :

(A1)(i) $nh_n^4 \to 0$ and $nh_n^{3+2\delta}(\log h_n^{-1})^{-1} \to \infty$ for some $\delta > 0$.
(ii) K has compact support, $\int uK(u)\,du = 0$ and K is twice continuously differentiable.

(A2)(i) F_X is three times continuously differentiable and $\inf_{a \le x \le b} f_X(x) > 0$.
(ii) m and σ are twice continuously differentiable and $\inf_{a \le x \le b} \sigma(x) > 0$.

(A3) $F'(y|x)$ is continuous in (x, y) and $\sup_{x,y} |y^2 F'(y|x)| < \infty$, and the same holds for all other partial derivatives of $F(y|x)$ with respect to x and y up to order two.

(A4)(i) Θ is a compact subspace of $I\!R^p$.
(ii) $m_\theta(x)$ is continuous in θ for all x.
(iii) For all $\theta \in \Theta$, $m_\theta^2(x) \le M(x)$ for some integrable function M.
(iv) For all $\varepsilon > 0$, $\inf_{\|\theta - \theta_0\| > \varepsilon} E[(m_\theta(X) - m_{\theta_0}(X))^2] > 0$ ($\|\cdot\|$ denotes the Euclidean norm).
(v) $E[\sigma^2(X)] < \infty$.

(A5)(i) θ_0 is an interior point of Θ.
(ii) Ω is non-singular.
(iii) $m_\theta(x)$ is twice differentiable with respect to θ for all x.
(iv) $E\Big(\Big\|\frac{\partial m_{\theta_0}(X)}{\partial \theta}\Big\|^2\Big) < \infty$ and $E\Big(\Big\|\frac{\partial^2 m_{\theta_0}(X)}{\partial \theta \partial \theta'}\Big\|^2\Big) < \infty$.
(v) For $k = 0, 1, 2$, $\Big\|\frac{\partial^k m_{\theta_1}(x)}{\partial \theta^k} - \frac{\partial^k m_{\theta_0}(x)}{\partial \theta^k}\Big\| \le h_k(x)\|\theta_1 - \theta_0\|$ for all $\theta_1 \in \Theta$, where $Eh_k^2(X) < \infty$ ($\frac{\partial^2}{\partial \theta^2}$ stands for $\frac{\partial^2}{\partial \theta \partial \theta'}$).

(A6) $E[r^2(X)] < \infty$ and $E\Big[r(X)\frac{\partial^2 m_{\theta_0}(X)}{\partial \theta \partial \theta'}\Big] < \infty$.

3 Main results

Theorem 3.1 *Assume (A1)-(A5). Then, under H_0,*

$$\begin{aligned}
&\hat{F}_\varepsilon(y) - \hat{F}_{\varepsilon 0}(y) \\
&= f_\varepsilon(y) n^{-1} \sum_{i=1}^n \left\{ \sigma^{-1}(X_i)[Y_i - m(X_i)] - \int \sigma^{-1}(x) \left(\frac{\partial m_{\theta_0}(x)}{\partial \theta} \right)' dF_X(x)\, \eta_{\theta_0}(X_i, Y_i) \right\} \\
&\quad + R_n(y),
\end{aligned}$$

where $\sup_{-\infty<y<\infty} |R_n(y)| = o_P(n^{-1/2})$.

Corollary 3.2 *Assume (A1)-(A5). Then, under* H_0*, the process* $n^{1/2}(\hat{F}_\varepsilon(y) - \hat{F}_{\varepsilon 0}(y))$ $(-\infty < y < \infty)$ *converges weakly to* $f_\varepsilon(y)W$*, where* W *is a zero-mean normal random variable with variance*

$$Var(W) = E\left[\sigma^{-1}(X)[Y - m(X)] - \int \sigma^{-1}(x)\left(\frac{\partial m_{\theta_0}(x)}{\partial\theta}\right)' dF_X(x)\,\eta_{\theta_0}(X,Y)\right]^2.$$

Note that the limiting process has an extremely simple structure, as it factorizes in a deterministic function only depending on y and a random variable independent of y.

As an immediate consequence, we obtain the limiting distribution of the Kolmogorov-Smirnov and Cramér-von Mises statistic :

Corollary 3.3 *Assume (A1)-(A5). Then, under* H_0,

$$T_{KS} \xrightarrow{d} \sup_{-\infty<y<\infty} |f_\varepsilon(y)|\,|W|$$

$$T_{CM} \xrightarrow{d} \int f_\varepsilon^2(y)\,dF_\varepsilon(y)\,W^2$$

We consider now the limiting behavior of the two test statistics under the local alternative H_1 :

Theorem 3.4 *Assume (A1)-(A6). Then, under* H_1,

$$T_{KS} \xrightarrow{d} \sup_{-\infty<y<\infty} |f_\varepsilon(y)|\,|W+b|$$

$$T_{CM} \xrightarrow{d} \int f_\varepsilon^2(y)\,dF_\varepsilon(y)\,(W+b)^2$$

where

$$b = -\int \sigma^{-1}(x)\left(\frac{\partial m_{\theta_0}(x)}{\partial\theta}\right)' dF(x)\,\Omega^{-1}\int r(x)\frac{\partial m_{\theta_0}(x)}{\partial\theta}dF(x) + \int \sigma^{-1}(x)r(x)dF(x).$$

4 Simulations and data analysis

We have carried out some simulations to assess the behaviour of the proposed test with small to moderate sample sizes.

To apply the test in practice, we use a bootstrap procedure based on the residual distribution (see e.g. González-Manteiga et al. (1994) and Mammen (2000)) to approximate the distribution of the test statistics. To this end, let $\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_n$ be obtained by standardizing the residuals $\hat{\varepsilon}_i$ $(i = 1, \ldots, n)$, and let $\tilde{F}_\varepsilon(y)$ be the (slightly) smoothed empirical distribution of $\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_n$ (since the representation in Theorem 3.1 depends on the density $f_\varepsilon(y)$, slight smoothing is necessary here. In other related problems smoothed resampling has been proved useful, see e.g. Hall et al. (1989) and González-Manteiga et al. (1994)). Next, draw independent random variables $\varepsilon_1^*, \ldots, \varepsilon_n^*$ from $\tilde{F}_\varepsilon$ and define

$$Y_i^* = m_{\hat{\theta}}\,(X_i) + \hat{\sigma}(X_i)\varepsilon_i^* \qquad i \in \{1, \ldots, n\}.$$

From the bootstrap sample a bootstrap version of the test statistic is obtained. In practice, the critical values of the original test statistic are approximated from B replicates of its bootstrap version.

The simulated model was taken to be

$$Y_i = \theta X_i + a\,(X_i) + \frac{1}{2}\,(1 + X_i)\,\varepsilon_i \qquad i \in \{1, \ldots, n\}$$

where the null hypothesis consists in the parametric model

$$H_0 : m(x) = \theta x \;\text{ with }\; \theta \in R,$$

the term $a(X_i)$ represents the deviation from the null hypothesis, $X_1, \ldots, X_n$ are independent and uniformly distributed on the unit interval, and $\varepsilon_1, \ldots, \varepsilon_n$ are independent standard normal random variables. The true parameter was taken to be $\theta_0 = 5$, and three alternatives were considered: $a(x) = 0.6(x - 0.5)$, $a(x) = 2x^2$ and $a(x) = 0.5\sin(4\pi x)$.

Data sharpening methods, proposed by Choi et al. (2000), were employed to improve the nonparametric estimators of the regression function and of the conditional variance, preventing from border effects.

The following table shows the percentage of rejections obtained with the test statistics T_{KS} and T_{CM} for one thousand samples, and approximating the distribution of the test statistic by bootstrap of the standardized residuals as described above and on the basis of five thousand replicates. The nominal level was 5%. The table contains the percentages of rejections under the null hypothesis and under different values of the deviation (represented by the function a), and for different values of the sample size and the bandwidth.

These results show a good approximation to the nominal level under the null hypothesis. Under the alternative, the power increases with the sample size, and it is not much influenced by the bandwidth. The results obtained with Kolmogorov-Smirnov and Cramér-von Mises type of statistics are also quite similar.

		$a(x)=0$		$a(x)=0.6(x-0.5)$		$a(x)=2x^2$		$a(x)=0.5\sin(4\pi x)$	
n	h	T_{KS}	T_{CM}	T_{KS}	T_{CM}	T_{KS}	T_{CM}	T_{KS}	T_{CM}
50	0.25	6.7	5.0	19.1	10.5	28.0	17.1	21.4	25.5
50	0.30	5.3	4.4	21.8	14.1	25.9	15.1	20.7	24.8
50	0.35	4.3	2.9	18.5	10.2	22.9	12.5	23.9	27.8
100	0.20	6.1	4.7	44.9	40.4	53.5	48.9	49.5	55.2
100	0.25	5.0	3.8	44.3	38.2	53.2	46.9	43.8	48.9
100	0.30	5.4	4.7	40.7	35.2	49.1	42.8	37.0	44.9
200	0.20	5.7	4.7	79.3	79.8	88.3	88.3	82.3	85.8
200	0.25	6.1	6.3	76.7	75.4	85.8	85.8	77.0	80.5
200	0.30	4.8	4.0	71.0	70.3	81.8	82.6	68.8	75.2

Table 1. *Percentage of rejections (nominal level 5%).*

Application to the milk yield data. Now the new test is applied to the data as given in table 4 of Weihrather (1993). These data present the average daily milk yields per week (in liters) for a cow for a period of 41 weeks. A common model to describe lactation curves is Wood's curve

$$L(t) = \beta_0 t^{\beta_1} \exp(\beta_2 t),$$

where $L(t)$ denotes the theoretical milk yield at time t. Taking logarithms, we arrive at the linear model

$$m(t) = \theta_0 + \theta_1 \log t + \theta_2 t.$$

The goodness-of-fit of this model was checked by using the test proposed in this paper, both with the Kolmogorov-Smirnov and the Cramér-von Mises type statistics. The covariate 'time' is transformed to the interval [0,1]. Two bandwidths were employed, $h = 0.15$ and $h = 0.20$. For each bandwidth, the test statistics were calculated, and critical values for the significance level of 5% were approximated by the bootstrap of the standard nonparametric residuals. For the bootstrap approximation 100,000 replicates were taken, and also used to obtain the p-values. All these results are shown in Table 2. As a conclusion, Wood's curve should be rejected as a model for this type of data. This conclusion was also attained by Weihrather (1993) and Stute et al. (1998), from the application of their respective tests to this data set.

	T_{KS}			T_{CM}		
	Statistic	Critical value	p-value	Statistic	Critical value	p-value
$h = 0.15$	1.4056	1.0932	0.00063	0.6538	0.1648	0.00001
$h = 0.20$	1.4056	1.0932	0.00055	0.5967	0.1856	0.00001

Table 2. *Results of the proposed test with the milk yield data.*

Appendix : Proofs

We start with two lemmas that state the consistency and an asymptotic representation for $\hat{\theta}$ under H_1. To obtain the analogues of these results under H_0 it suffices to take $r \equiv 0$ and to drop condition $(A6)$. Note that Wu (1981), Seber and Wild (1989) among others, obtained similar results for the special case of a homoscedastic model that satisfies H_0.

Lemma A.1 *Assume (A4) and let $E[r^2(X)] < \infty$. Then, under H_1,*

$$\hat{\theta} - \theta_0 \xrightarrow{P} 0.$$

Proof. First note that by Lemma 2 in Jennrich (1969), $\hat{\theta}$ is measurable. We prove the result by verifying the conditions of Theorem 5.7 in van der Vaart (1998, p. 45) for $M_n = -S_n$ and $M = -S_0$. From the definition of $\hat{\theta}$ and condition $(A4)(iv)$, it follows that it suffices to show that $\sup_\theta |S_n(\theta) - S_0(\theta)| \to_P 0$. The latter can be proved by making use of Theorem 2 in Jennrich (1969).

Lemma A.2 *Assume* $(A4)$–$(A6)$. *Then, under H_1,*

$$\hat{\theta} - \theta_0 = n^{-1} \sum_{i=1}^{n} \eta_{\tilde{\theta}_{0n}}(X_i, Y_i) + \Omega^{-1} n^{-1/2} \int r(x) \frac{\partial m_{\theta_0}(x)}{\partial \theta} dF(x) + o_P(n^{-1/2}),$$

where $\tilde{\theta}_{0n}$ is the minimizer of

$$E[(m(X) - m_\theta(X))^2].$$

Proof. First note that $\hat{\theta} - \theta_0 = o_P(1)$ (using Lemma A.1) and $\tilde{\theta}_{0n} - \theta_0 = o(1)$. The latter can be shown in a similar way as in Lemma A.1, by using however $\tilde{S}_{0n}(\theta) =$

$E[\sigma^2(X)] + E[(m_\theta(X) - m(X))^2]$ instead of $S_n(\theta)$. Hence, it follows from assumption $(A5)(i)$ that $\hat{\theta}$ and $\tilde{\theta}_{0n}$ are interior points of Θ for n large enough. Now write

$$\frac{\partial S_n(\hat{\theta})}{\partial \theta} - \frac{\partial S_n(\tilde{\theta}_{0n})}{\partial \theta} = \frac{\partial^2 S_n(\theta_{1n})}{\partial \theta \partial \theta'}(\hat{\theta} - \tilde{\theta}_{0n}),$$

for some θ_{1n} between $\hat{\theta}$ and $\tilde{\theta}_{0n}$. Then,

$$\hat{\theta} - \tilde{\theta}_{0n} = -\left(\frac{\partial^2 S_n(\theta_{1n})}{\partial \theta \partial \theta'}\right)^{-1} \frac{\partial S_n(\tilde{\theta}_{0n})}{\partial \theta}. \tag{A.1}$$

First,

$$\frac{\partial S_n(\tilde{\theta}_{0n})}{\partial \theta} = -2n^{-1} \sum_{i=1}^n (Y_i - m_{\tilde{\theta}_{0n}}(X_i)) \frac{\partial m_{\tilde{\theta}_{0n}}(X_i)}{\partial \theta}.$$

Next,

$$\begin{aligned}
&\frac{\partial^2 S_n(\theta_{1n})}{\partial \theta \partial \theta'} \\
&= 2n^{-1} \sum_{i=1}^n \frac{\partial m_{\theta_{1n}}(X_i)}{\partial \theta} \left(\frac{\partial m_{\theta_{1n}}(X_i)}{\partial \theta}\right)' - 2n^{-1} \sum_{i=1}^n (Y_i - m_{\theta_{1n}}(X_i)) \frac{\partial^2 m_{\theta_{1n}}(X_i)}{\partial \theta \partial \theta'} \\
&= 2n^{-1} \sum_{i=1}^n \frac{\partial m_{\theta_0}(X_i)}{\partial \theta} \left(\frac{\partial m_{\theta_0}(X_i)}{\partial \theta}\right)' - 2n^{-1} \sum_{i=1}^n (Y_i - m_{\theta_0}(X_i)) \frac{\partial^2 m_{\theta_0}(X_i)}{\partial \theta \partial \theta'} + o_P(1),
\end{aligned}$$

using a similar derivation as in the proof of Theorem 3.15 p. 139 in Sánchez Sellero (2001), and using the fact that $\hat{\theta} - \theta_0 = o_P(1)$ (see Lemma A.1). The first term on the right hand side equals

$$2E\left[\frac{\partial m_{\theta_0}(X)}{\partial \theta} \left(\frac{\partial m_{\theta_0}(X)}{\partial \theta}\right)'\right] + o_P(1) = 2\Omega + o_P(1),$$

while the second term can be written as

$$-2E\left[(Y - m_{\theta_0}(X)) \frac{\partial^2 m_{\theta_0}(X)}{\partial \theta \partial \theta'}\right] + o_P(1) = o_P(1),$$

since $m(X) - m_{\theta_0}(X) = n^{-1/2} r(X)$. Hence,

$$\hat{\theta} - \tilde{\theta}_{0n} = n^{-1} \sum_{i=1}^n \eta_{\tilde{\theta}_{0n}}(X_i, Y_i) + o_P(n^{-1/2}).$$

(Note that $E[(m(X) - m_{\tilde{\theta}_{0n}}(X)) \frac{\partial m_{\tilde{\theta}_{0n}}(X)}{\partial \theta}] = 0$, which implies that the numerator on the right hand side of (A.1) is $O_P(n^{-1/2})$.) Next, since $\tilde{\theta}_{0n}$ minimizes $\tilde{S}_{0n}(\theta)$, we have for some θ_{2n} between θ_0 and $\tilde{\theta}_{0n}$,

$$\begin{aligned}
\tilde{\theta}_{0n} - \theta_0 &= -\left(\frac{\partial^2 \tilde{S}_{0n}(\theta_{2n})}{\partial \theta \partial \theta'}\right)^{-1} \frac{\partial \tilde{S}_{0n}(\theta_0)}{\partial \theta} \\
&= \Omega^{-1} n^{-1/2} \int r(x) \frac{\partial m_{\theta_0}(x)}{\partial \theta} dF(x) + o(n^{-1/2}),
\end{aligned}$$

where the last equality follows in a similar way as for $\hat{\theta} - \tilde{\theta}_{0n}$. This gives the result.

Proof of Theorem 3.1. It follows from the proof of Theorem 1 in Akritas and Van Keilegom (2001) (hereafter called AVK) that

$$\hat{F}_\varepsilon(y) - F_\varepsilon(y) = n^{-1}\sum_{i=1}^n I(\varepsilon_i \le y) - F_\varepsilon(y)$$
$$+f_\varepsilon(y)\int \sigma^{-1}(x)[y\{\hat{\sigma}(x) - \sigma(x)\} + \hat{m}(x) - m(x)]\, dF_X(x) + R_{n1}(y),$$

where $\sup_y |R_{n1}(y)| = o_P(n^{-1/2})$. Note that AVK assume that m and σ are L-functionals that depend on a certain score function J. It is straightforward to show that the results of AVK can be extended to the choice $J \equiv 1$, which leads to the conditional mean and variance that we consider in this paper (it suffices to replace Propositions 3–5 in AVK by their analogues for the estimators of the conditional mean and variance).

We next need a similar representation for $\hat{F}_{\varepsilon 0}(y) - F_\varepsilon(y)$, i.e. we need to show that Theorem 1 in AVK can be adapted to the case where $\hat{m}(x)$ is replaced by $m_{\hat{\theta}}(x)$. Careful investigation of the proof of Lemma 1 in AVK reveals that the class $C_1^{1+\delta}(R_X)$ (defined in that proof) can be replaced by the class $M_\Theta(R_X) = \{x \to (m_\theta(x) - m(x))/\sigma(x) : \theta \in \Theta\}$, since $P((m_{\hat{\theta}}(x) - m(x))/\sigma(x) \in M_\Theta(R_X)) \to 1$ as $n \to \infty$, and since the bracketing number $N_{[]}(\varepsilon^2, M_\Theta(R_X), L_2(P))$ can be easily seen to be $O(\varepsilon^{-2p})$ for any $\varepsilon > 0$ (using condition $(A5)$), which is smaller than for the class $C_1^{1+\delta}(R_X)$. (Note that by working with $M_\Theta(R_X)$ instead of with $C_1^{1+\delta}(R_X)$ we do not need to show that $\hat{m}$ in Propositions 3–5 can be replaced by $m_{\hat{\theta}}$.) It now follows that

$$\hat{F}_{\varepsilon 0}(y) - F_\varepsilon(y) = n^{-1}\sum_{i=1}^n I(\varepsilon_i \le y) - F_\varepsilon(y)$$
$$+f_\varepsilon(y)\int \sigma^{-1}(x)[y\{\hat{\sigma}(x) - \sigma(x)\} + m_{\hat{\theta}}(x) - m(x)]\, dF_X(x) + R_{n2}(y),$$

where $\sup_y |R_{n2}(y)| = o_P(n^{-1/2})$. Hence,

$$\hat{F}_\varepsilon(y) - \hat{F}_{\varepsilon 0}(y) = f_\varepsilon(y)\int \sigma^{-1}(x)[\hat{m}(x) - m_{\hat{\theta}}(x)]\, dF_X(x) + o_P(n^{-1/2}),$$

uniformly in y. Next,

$$\hat{m}(x) - m(x) = n^{-1}\sum_{i=1}^n \frac{K_h(x - X_i)}{f_X(x)}[Y_i - m(x)] + O_P((nh_n)^{-1}),$$

uniformly in x, where $K_h(\cdot) = h_n^{-1}K(\cdot/h_n)$. Since by Lemma A.2 and condition (A5),

$$\int \sigma^{-1}(x)[m_{\hat{\theta}}(x) - m(x)]\, dF_X(x)$$

$$= \int \sigma^{-1}(x) \left(\frac{\partial m_{\theta_0}(x)}{\partial \theta} \right)' dF_X(x)\, n^{-1} \sum_{i=1}^n \eta_{\theta_0}(X_i, Y_i) + o_P(n^{-1/2}),$$

we have that

$$\begin{aligned}
&\hat{F}_\varepsilon(y) - \hat{F}_{\varepsilon 0}(y) \\
&= f_\varepsilon(y) n^{-1} \sum_{i=1}^n \int \sigma^{-1}(x) K_h(x - X_i)[Y_i - m(x)]\, dx \\
&\quad - f_\varepsilon(y) \int \sigma^{-1}(x) \left(\frac{\partial m_{\theta_0}(x)}{\partial \theta} \right)' dF_X(x)\, n^{-1} \sum_{i=1}^n \eta_{\theta_0}(X_i, Y_i) + o_P(n^{-1/2}) \\
&= f_\varepsilon(y) n^{-1} \sum_{i=1}^n \left\{ \sigma^{-1}(X_i)[Y_i - m(X_i)] - \int \sigma^{-1}(x) \left(\frac{\partial m_{\theta_0}(x)}{\partial \theta} \right)' dF_X(x)\, \eta_{\theta_0}(X_i, Y_i) \right\} \\
&\quad + o_P(n^{-1/2}), \qquad \text{(A.2)}
\end{aligned}$$

uniformly in y, where the last equality follows from a Taylor expansion of order two.

Proof of Corollary 3.2. Since the expression on the right hand side of (A.2) factorizes in a deterministic component $f_\varepsilon(y)$, and a sum of i.i.d. terms not depending on y, the weak convergence of the process $n^{1/2}(\hat{F}_\varepsilon(y) - \hat{F}_{\varepsilon 0}(y))$ $(-\infty < y < \infty)$ follows immediately.

Proof of Corollary 3.3. The convergence of the Kolmogorov-Smirnov statistic follows directly from the continuous mapping theorem. For the Crámer-von Mises statistic it suffices to show that $d\hat{F}_\varepsilon(y)$ can be replaced by $dF_\varepsilon(y)$. Using the notation $W_n(y) = f_\varepsilon(y) n^{-1/2} \sum_{i=1}^n W_{ni}$, where W_{ni} is the expression between brackets on the right hand side of (A.2), we have that

$$\begin{aligned}
\left| \int W_n^2(y)\, d[\hat{F}_\varepsilon(y) - F_\varepsilon(y)] \right| &= 2 \left| \int [\hat{F}_\varepsilon(y) - F_\varepsilon(y)] W_n(y) W_n'(y)\, dy \right| \\
&\le o_P(1) \int f_\varepsilon(y) |f_\varepsilon'(y)|\, dy \\
&\le o_P(1) \sup_y |f_\varepsilon'(y)| = o_P(1),
\end{aligned}$$

since $\sup_y |\hat{F}_\varepsilon(y) - F_\varepsilon(y)| = o_P(1)$ (see Theorem 2 in AVK).

Proof of Theorem 3.4. First note that $E[\frac{\partial m_{\tilde{\theta}_{0n}(X)}}{\partial \theta}(m(X) - m_{\tilde{\theta}_{0n}}(X))] = 0$, and hence $E[\eta_{\tilde{\theta}_{0n}}(X, Y)] = 0$. It follows that $\hat{\theta} - \theta_0 = O_P(n^{-1/2})$ and hence, using Lemma A.2,

$$\int \sigma^{-1}(x)[m_{\hat{\theta}}(x) - m(x)] dF_X(x)$$

$$= \int \sigma^{-1}(x) \left(\frac{\partial m_{\theta_0}(x)}{\partial \theta} \right)' dF_X(x)\, n^{-1} \sum_{i=1}^{n} \eta_{\tilde{\theta}_{0n}}(X_i, Y_i)$$

$$+ \int \sigma^{-1}(x) \left(\frac{\partial m_{\theta_0}(x)}{\partial \theta} \right)' dF_X(x)\, \Omega^{-1} n^{-1/2} \int r(x) \frac{\partial m_{\theta_0}(x)}{\partial \theta} dF(x)$$

$$-n^{-1/2} \int \sigma^{-1}(x) r(x)\, dF_X(x) + o_P(n^{-1/2}).$$

It is easily seen that the asymptotic distribution of $n^{-1} \sum_{i=1}^{n} \eta_{\tilde{\theta}_{0n}}(X_i, Y_i)$ under H_1 equals the asymptotic distribution of $n^{-1} \sum_{i=1}^{n} \eta_{\theta_0}(X_i, Y_i)$ under H_0. The rest of the proof is similar to the proofs of Theorem 3.1 and Corollaries 3.2 and 3.3.

References

Akritas, M.G. and Van Keilegom, I. (2001). Nonparametric estimation of the residual distribution. *Scand. J. Statist.*, **28**, 549–568.

Alcalá, J.T., Cristóbal, J.A. and González Manteiga, W. (1999). Goodness-of-fit test for linear models based on local polynomials. *Statist. Probab. Letters*, **42**, 39–46.

Chen, S.X., Härdle, W. and Li, M. (2003). An empirical likelihood goodness-of-fit test for time series. *J. Roy. Statist. Soc. - Series B*, **65**, 663–678.

Choi, E., Hall, P. and Rousson, V. (2000). Data sharpening methods for bias reduction in nonparametric regression. *Ann. Statist.*, **28**, 1339–1355.

Dette, H. (1999). A consistent test for the functional form of a regression based on a difference of variance estimators. *Ann. Statist.*, **27**, 1012–1040.

Dette, H. and Munk, A. (1998). Validation of linear regression models. *Ann. Statist.*, **26**, 778–800.

Dette, H., Munk, A. and Wagner, T. (2000). Testing model assumptions in multivariate linear regression models. *J. Nonpar. Statist.*, **12**, 309–342.

Fan, J., Zhang, C. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.*, **29**, 153–193.

González Manteiga, W., Prada Sánchez, J.M. and Romo, J. (1994). The bootstrap - a review. *Computat. Statist.*, **9**, 165–205.

Hall, P., DiCiccio, T.J. and Romano, J.P. (1989). On smoothing and the bootstrap. *Ann. Statist.*, **17**, 692–704.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.*, **21**, 1926–1947.

Jennrich, R.I. (1969). Asymptotic properties of non-linear least-squares estimators. *Ann. Math. Statist.*, **40**, 633–643.

Mammen, E. (2000). Resampling methods for nonparametric regression. In: *Smoothing and regression*, Ed. M.G. Schimek. Wiley, New York.

Sánchez Sellero, C. (2001). *Inferencia estadística en datos con censura y/o truncamiento*, Ph.D. thesis, University of Santiago de Compostela.

Seber, G.A.F. and Wild, C.J. (1989). *Nonlinear regression.* Wiley, New York.

Stute, W. (1997). Nonparametric model checks for regression. *Ann. Statist.*, **25**, 613–641.

Stute, W., González Manteiga, W. and Presedo Quindimil, M. (1998). Bootstrap approximations in model checks for regression. *J. Amer. Statist. Assoc.*, **93**, 141–149.

van der Vaart, A.W. (1998). *Asymptotic statistics.* Cambridge Univ. Press. Cambridge.

Weihrather, G. (1993). Testing a linear regression model against nonparametric alternatives. *Metrika*, **40**, 367–379.

White, H. (1981). Consequences and detection of misspecified nonlinear regression models. *J. Amer. Statist. Assoc.*, **76**, 419–433.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1–25.

Wu, C.F. (1981). Asymptotic theory of nonlinear least-squares estimation. *Ann. Statist.*, **9**, 501–513.