The Annals of Statistics
1998, Vol. 26, No. 5, 1916–1934

MODEL CHECKS FOR REGRESSION: AN INNOVATION PROCESS APPROACH

BY WINFRIED STUTE, SILKE THIES AND LI-XING ZHU[1]

University of Giessen, University of Giessen and Chinese Academy of Sciences

In the context of regression analysis it is known that the residual cusum process may serve as a basis for the construction of various omnibus, smooth and directional goodness-of-fit tests. Since a deeper analysis requires the decomposition of the cusums into their principal components and this is difficult to obtain, we propose to replace this process by its innovation martingale. It turns out that the resulting tests are (asymptotically) distribution free under composite null models and may be readily performed. A simulation study is included which indicates that the distributional approximations already work for small to moderate sample sizes.

1. Introduction and main results. It is the purpose of the present paper to provide some further methodology for model checks in regression. As noted by Stute (1997), the cusum process of the residuals may serve as a basis for various goodness-of-fit tests in this field. In particular, it was demonstrated that for power investigations as well as for the derivation of smooth and directional tests it is necessary to compute the principal components of the residual cusum process. While this is possible in principle, some numerical work is needed when it comes down to checking a model for a given data set.

In this paper we propose to replace the cusum process by its innovation martingale. For this, note that the cusum process (in theory) admits a decomposition into a martingale and a compensator. As will be shown, the martingale part converges, for large sample size, to a Brownian motion in transformed time. For the new processes, principal components are readily available and no extra numerical work is needed so that smooth and directional tests, for example, may be easily performed.

To be more specific, let (X, Y) be a random observation in some Euclidean space $\mathbb{R}^{d+1}$ such that the random variable Y has a finite expectation. To simplify the notation, we shall assume throughout that X is univariate. Denote with

$$m(x) = \mathbb{E}\{Y|X = x\}$$

the regression function of Y with respect to X. We let

$$\mathscr{M} = \{m(\cdot, \theta)\colon \theta \in \Theta\}$$

denote a family of functions parameterized by some p-dimensional vector $\theta \in \Theta \subset \mathbb{R}^p$. Assuming that m belongs to $\mathscr{M}$, the main emphasis in the lit-

Received October 1996; revised May 1998.
[1]Supported in part by a fellowship of the Max-Planck Gesellschaft.
AMS 1991 *subject classifications*. Primary 62G30, 60G44; secondary 62G10.
Key words and phrases. Residual cusum process, innovation process, goodness-of-fit tests.

erature then has been on estimation of or testing hypotheses about the "true parameter θ_0" satisfying $m = m(\cdot, \theta_0)$. For example, $\mathscr{M}$ may consist of all functions

$$m(x, \theta) = \theta_1 g_1(x) + \cdots + \theta_p g_p(x), \qquad \theta = (\theta_1, \ldots, \theta_p)^t, \tag{1.1}$$

spanned by a given basis $g_1, \ldots, g_p$ and where t denotes transpose. Clearly, each investigation of θ_0 should be accompanied with a proper check whether the (composite) model $\mathscr{M}$, that is, the hypothesis

$$H_0\colon m \in \mathscr{M}$$

is at all satisfied. Stute (1997) contains a fairly comprehensive list of references on nonparametric model checks for regression. The main emphasis of that paper was to point out that many goodness-of-fit tests could be based on the cusum process of the residuals

$$\tilde{R}_n(x) = n^{-1/2} \sum_{i=1}^{n} 1_{\{X_i \le x\}} [Y_i - m(X_i, \theta_n)]. \tag{1.2}$$

Here (X_i, Y_i), $1 \le i \le n$, is a sample of independent observations with the same distribution as (X, Y), and θ_n is, under H_0, a square-root consistent estimator of θ_0. In a sense, $\tilde{R}_n$ is a marked empirical process, the marks being given by the residuals obtained from fitting the data to the hypothetical model. Among other things, it was shown in Stute (1997) that under mild regularity assumptions on $\mathscr{M}$ and θ_n, the process $\tilde{R}_n$ has a Gaussian limit under H_0. Furthermore, the paper contains a principal component analysis of $\tilde{R}_n$ to the effect, that:

(a) The power of the Cramér–von Mises test associated with $\tilde{R}_n$ could be investigated in detail.
(b) The local power of directional tests based on $\tilde{R}_n$ could be analyzed.
(c) One was able to derive Neyman smooth tests for regression when the hypothesis is composite.
(d) Optimal Neyman-Pearson tests could be derived when the alternative is specified.

As a conclusion we thus see that there is good reason to base model diagnostics on the cusum process of the residuals rather than the residuals themselves, as is done in a noninferential way in many textbooks.

Unfortunately, as with many other nonparametric procedures involving estimated parameters, the processes $\tilde{R}_n$ and their limit are typically not distribution free in that their distributions may depend on model characteristics, the choice of θ_n or even the unknown parameter θ_0. These facts seriously limit the applicability of the procedures in goodness-of-fit testing, since, for example, critical values will not be available from existing tables. Consider Durbin (1973), who investigated the ordinary empirical process with estimated parameters rather than the residual cusum process. As a way out of this dilemma, Khmaladze (1981), for univariate X's, utilized an innovation process approach

to construct a linear operator T such that the estimated empirical process transformed by T is at least asymptotically distribution free. In the context of regression, distributional feasibility may also be achieved by the Bootstrap; see Stute, González Manteiga and Presedo Quindïmil (1998). As we have noted above, however, it is not only the distributional character of the underlying processes but their decomposition into principal components which enhances the statistical applications and which makes our approach attractive.

Determining the innovation martingale requires constructing a transformation T such that $T\tilde{R}_n$ (approximately) is a martingale. In the limit it will be a Brownian motion in transformed time. Since T depends on quantities which in practice are unknown, it needs to be replaced by an empirical substitute T_n. As it will turn out, replacement of T by T_n will not change the limit so that modulo a transformation in time, $T_n\tilde{R}_n$ is in fact asymptotically distribution free. Extensions to the multivariate case are possible in the spirit of Khmaladze (1988).

This section will provide the main results of the paper. Section 2 presents some simulation results for finite sample size, while proofs are deferred to Section 3.

The distributional theory becomes much simpler if we replace the residuals by the true errors. For this, define

$$R_n(x) = n^{-1/2} \sum_{i=1}^{n} 1_{\{X_i \le x\}} [Y_i - m(X_i)].$$

Assuming $\mathbb{E}Y^2 < \infty$, let

$$\sigma^2(u) = \mathrm{Var}(Y|X = u)$$

denote the conditional variance of Y given $X = u$, and put

$$\psi(x) = \int_{-\infty}^{x} \sigma^2(u) F(du) \quad \text{where } X \sim F.$$

Clearly, ψ is a nondecreasing nonnegative function. In the homoscedastic case, $\sigma^2(u) \equiv \sigma^2$ is a constant, whence $\psi(x) = \sigma^2 F(x)$. R_n and $\tilde{R}_n$ are random elements in the Skorohod space $D[-\infty, \infty]$ endowed with the topology of weak convergence. It is readily seen that

$$\mathrm{Cov}[R_n(x_1), R_n(x_2)] = \psi(x_1 \wedge x_2),$$

that is, R_n has the same covariance structure as $B \circ \psi$, with B denoting a standard Brownian motion. Moreover, the finite-dimensional distributions of R_n weakly converge to those of $B \circ \psi$. An argument showing tightness, compare Stute (1997), therefore yields

$$R_n \to B \circ \psi = R_\infty. \tag{1.3}$$

For composite model checks, the function m needs to be replaced by its parametric fit m_{θ_n} so that we come up with $\tilde{R}_n$ as defined in (1.2). Note that under

H_0 we have $m = m_{\theta_0}$ and therefore

$$\tilde{R}_n(x) = R_n(x) - n^{-1/2} \sum_{i=1}^{n} 1_{\{X_i \le x\}} [m(X_i, \theta_n) - m(X_i, \theta_0)].$$

We now state some regularity assumptions on θ_n and the model $\mathscr{M}$ under which $\tilde{R}_n$ admits a weak limit.

(A) Under H_0, that is, $m = m(\cdot, \theta_0)$ for some $\theta_0 \in \Theta$, we have

$$n^{1/2}(\theta_n - \theta_0) = n^{-1/2} \sum_{i=1}^{n} l(X_i, Y_i, \theta_0) + o_{\mathbb{P}}(1),$$

where l is a vector-valued function such that:

(i) $\mathbb{E}\{l(X, Y, \theta_0)\} = 0$;
(ii) $L(\theta_0) = \mathbb{E}\{l(X, Y, \theta_0) l^t(X, Y, \theta_0)\}$ exists.

Assumption (B) is concerned with the regularity of $\mathscr{M}$:

(B)(i) $m(x, \theta)$ is continuously differentiable with respect to θ in the interior set of Θ. Put

$$g(x, \theta) = \operatorname{grad}_\theta(m(x, \theta)) = (g_1(x, \theta), \ldots, g_p(x, \theta))^t$$

and assume that

(B)(ii) $$|g_i(x, \theta)| \le M(x) \quad \text{for all } \theta \in \Theta \text{ and } 1 \le i \le p$$

for an F-integrable function M.

Set

$$G(x, \theta) = \int_{-\infty}^{x} g(u, \theta) F(du).$$

Note that in the linear model (1.1), $g_i(x, \theta) \equiv g_i(x)$ does not depend on θ and similarly for $G(x, \theta)$. Under (B), condition (A) is satisfied for the least squares estimator and (under further regularity assumptions) its robust modifications. See, for example, Maronna and Yohai (1981).

Under (A) and (B), the second sum in $\tilde{R}_n(x)$ converges in distribution to $G^t(x)V$, where V is a centered normal vector with covariance $L(\theta_0)$ and $G(x) = G(x, \theta_0)$. Corollary 1.3 in Stute (1997) asserts that an invariance principle similar to (1.3) also holds for $\tilde{R}_n$:

$$\tilde{R}_n \to B \circ \psi - G^t V \equiv \tilde{R}_\infty.$$

Note that typically V depends on B, since both components of $\tilde{R}_\infty$ are limits of terms computed from the same set of data.

We now introduce scale invariant versions of R_n and $\tilde{R}_n$, namely,

$$R_n^0(x) = n^{-1/2} \sum_{i=1}^{n} 1_{\{X_i \le x\}} \sigma^{-1}(X_i)[Y_i - m(X_i)]$$

and

$$\tilde{R}_n^0(x) = n^{-1/2} \sum_{i=1}^{n} 1_{\{X_i \le x\}} \sigma^{-1}(X_i)[Y_i - m(X_i, \theta_n)].$$

Replacing (B)(ii) by

(B)(iii) $$\sigma^{-1}(x)|g_i(x, \theta)| \le M(x) \quad \text{for all } \theta \in \Theta \text{ and } 1 \le i \le p,$$

we obtain

$$R_n^0 \to B \circ F = R_\infty^0$$

and

$$\tilde{R}_n^0 \to B \circ F - G_0^t V_0 = \tilde{R}_\infty^0, \tag{1.4}$$

with

$$G_0(x) = G_0(x, \theta_0) = \int_{-\infty}^{x} \sigma^{-1}(u) g(u, \theta_0) F(du).$$

Note that the time transformation in the limiting process is now F rather than ψ, so that the part involving the Brownian motion no longer depends on σ.

The function G_0 contains the main features of the underlying model:

1. The marginal d.f. F of the X's.
2. The (local) structure of $\mathscr{M}$, given in terms of g.
3. The (conditional) variances $\sigma^2(u)$ of the errors.

As we have pointed out in our introductory remarks, a detailed study of $\tilde{R}_\infty^0$ is difficult. Our strategy will therefore be to first transform $\tilde{R}_\infty^0$ into its martingale part $B \circ F$. In view of (1.4) it suffices to construct a linear transformation T satisfying

$$TR_\infty^0 = R_\infty^0 \quad \text{in distribution} \tag{1.5}$$

and

$$T(G_0^t V) \equiv 0. \tag{1.6}$$

To explicitly obtain T, put $g(u) = g(u, \theta_0)$ and set

$$A(x) = \int_x^\infty g(u) g^t(u) \sigma^{-2}(u) F(du),$$

a nonnegative definite $p \times p$-matrix, ignoring its dependence on θ_0 for a moment. Assuming that $A(x)$ is nonsingular, we define

$$(Tf)(x) = f(x) - \int_{-\infty}^{x} \sigma^{-1}(y) g^t(y) A^{-1}(y) \left[\int_y^\infty \sigma^{-1}(z) g(z) f(dz) \right] F(dy).$$

We will apply T to functions f which are either of bounded variation or Brownian motion $B \circ F$. In the latter case, the inner integral needs to be interpreted as a stochastic integral. Now, with the above T, it is not difficult

to see that (1.6) is satisfied. Since T is a linear operator, TR_∞^0 is a centered Gaussian process. For (1.5), it thus remains to show that

$$\text{Cov}[TR_\infty^0(r), TR_\infty^0(s)] = F(r \wedge s). \tag{1.7}$$

A proof of (1.7) will be deferred to Section 3 (Lemma 3.1). Altogether we thus have the following result.

THEOREM 1.1. *Let $\tilde{R}_\infty^0$ be defined by* (1.4) *and assume that $A(x)$ is nonsingular for all x. Then we have in distribution*

$$T\tilde{R}_\infty^0 = TR_\infty^0 = R_\infty^0 = B \circ F. \tag{1.8}$$

To motivate the next result, we may expect that there exist finite sample analogs of (1.5), (1.6) and (1.8). For this, recall

$$R_n^0(x) = n^{-1/2} \sum_{i=1}^{n} 1_{\{X_i \le x\}} \sigma^{-1}(X_i)[Y_i - m(X_i)],$$

the scale invariant modification of R_n. We shall show that, under H_0,

$$T\tilde{R}_n^0 = TR_n^0 + o_{\mathbb{P}}(1). \tag{1.9}$$

This constitutes the finite sample analog of the first equation in (1.8). Furthermore, the second equation suggests that

$$TR_n^0 \to B \circ F \quad \text{in distribution.} \tag{1.10}$$

Assertions (1.9) and (1.10) will be verified in Section 3 (Lemmas 3.2 and 3.3). Together they yield the following theorem.

THEOREM 1.2. *Under* (A) *and* (B), *assume that $A(x)$ is nonsingular for all x. Then, under H_0, we have*

$$T\tilde{R}_n^0 \to B \circ F \text{ in distribution}$$

in the Skorohod space $D[-\infty, -\infty)$.

Note that convergence in $D[-\infty, \infty)$ means convergence in $D[-\infty, x_0]$ for each finite x_0. See Pollard (1984). If $A(x)$ is nonsingular only for a restricted set of x's, all processes likewise need to be restricted to proper subsets of the real line. Further comments on why we have to restrict ourselves to finite x_0 are postponed to the end of this section.

Theorem 1.2 is fine from a probabilistic viewpoint. For statistical applications such as goodness-of-fit testing, it is still inappropriate since both $\tilde{R}_n^0$ and T involve unknown quantities like $\sigma^2(u)$, θ_0 and F. To apply our method to a given set of data, the transformation T, for example, needs to be replaced by an empirical analog T_n. We then need to show that the resulting processes have the same limit as $T\tilde{R}_n^0$. Actually, this will be achieved by showing that their difference goes to zero in probability uniformly on compacta.

In the homoscedastic case we simply have to replace $\tilde{R}_n^0$ by $\sigma_n^{-1}\tilde{R}_n$, where σ_n^2 is the (normalized) residual sum of squares and similarly in T. In the general heteroscedastic case, however, it is the function $\sigma^2(u)$ rather than the constant σ^2 which needs to be estimated from the data. In view of

$$\sigma^2(u) = \mathbb{E}\{Y^2|X = u\} - m^2(u),$$

any consistent nonparametric regression curve estimator may serve as an empirical substitute for the conditional second moment. Under H_0, the second part, $m^2(u)$, may be estimated by $m^2(u, \theta_n)$. As it turns out, this procedure works in principle, under some restrictive smoothness assumptions on $\sigma^2(u)$. A somewhat different approach which works under much weaker conditions is the following: split the whole sample into two parts, (X_i, Y_i), $1 \le i \le n_1$, and (X_i, Y_i), $n_1+1 \le i \le n$, where $n_1 \to \infty$ and $n - n_1 \to \infty$ as $n \to \infty$. Then estimate $\sigma^2(u)$ from the first part, say by $\sigma_{n_1}^2(u)$, and let the cusum process be based on the second half. This leads to the two processes

$$R_n^1(x) = (n - n_1)^{-1/2} \sum_{i=n_1+1}^{n} 1_{\{X_i \le x\}} \sigma_{n_1}^{-1}(X_i)[Y_i - m(X_i)]$$

and

$$\tilde{R}_n^1(x) = (n - n_1)^{-1/2} \sum_{i=n_1+1}^{n} 1_{\{X_i \le x\}} \sigma_{n_1}^{-1}(X_i)[Y_i - m(X_i, \theta_{n_1})].$$

Finally, the transformation T_n is defined by

$$\begin{aligned}(T_n f)(x) = f(x) - \int_{-\infty}^{x} \sigma_{n_1}^{-1}(y) g^t(y, \theta_{n_1}) A_{n_1}^{-1}(y) \\ \times \left[\int_y^{\infty} \sigma_{n_1}^{-1}(z) g(z, \theta_{n_1}) f(dz)\right] F_{n_1}(dy).\end{aligned} \tag{1.11}$$

Here F_{n_1} is the empirical d.f. of $X_{n_1+1}, \ldots, X_n$, the estimator θ_{n_1} is computed from (X_i, Y_i), $n_1 + 1 \le i \le n$, and

$$A_{n_1}(y) = \int_y^{\infty} g(u, \theta_{n_1}) g^t(u, \theta_{n_1}) \sigma_{n_1}^{-2}(u) F_{n_1}(du).$$

The fact that the index n_1 is used to indicate computation from the first part of the sample in the case of σ^2 and from the second part for the other quantities should not cause any confusion. To demonstrate the effect of splitting the data into two parts, note that conditionally on the first n_1 data, R_n^1 is a sum of independent centered processes with covariance function

$$K_{n_1}(r, s) = \int_{-\infty}^{r \wedge s} \sigma^2(u)/\sigma_{n_1}^2(u) F(du).$$

We shall see that under appropriate conditions

$$\sup_{r,s} \mathbb{E}|K_{n_1}(r, s) - F(r \wedge s)| \to 0, \tag{1.12}$$

which together with the above-mentioned independence of the summands yields

$$R_n^1 \to B \circ F \quad \text{in distribution.} \tag{1.13}$$

For (1.11) we recall that under no conditions other than square-integrability of Y do there exist universally consistent estimators of $\sigma^2(u)$ satisfying

$$\mathbb{E} \int |\sigma_{n_1}^2(u) - \sigma^2(u)| F(du) \to 0 \tag{1.14}$$

as $n_1 \to \infty$. Compare Stone (1977), Devroye and Wagner (1980), Spiegelman and Sacks (1980), or Stute (1994) who showed that nearest neighbor and kernel-type estimators are universally consistent under broad assumptions on the smoothing parameter.

For the convergence of K_{n_1} we need to assume that $\sigma^2(u)$ is bounded away from zero:

$$\sigma^2(u) \geq a > 0 \quad \text{for some } a. \tag{1.15}$$

For theoretical purposes we also want to guarantee that the $\sigma_{n_1}^2$ are bounded away from zero. This may be achieved without disturbing the uniform consistency (1.13) by just taking the maximum of $\sigma_{n_1}^2$ and a very small positive number. It is then easy to see that (1.13) implies (1.11). Therefore the inequality in Condition (B)(iii) is also satisfied if we replace $\sigma^2(u)$ by $\sigma_{n_1}^2(u)$.

THEOREM 1.3. *Under the assumptions of Theorem* 1.2, *let* (1.14) *be satisfied. Let* $\sigma_{n_1}^2$ *be a universally consistent estimator of* σ^2 *bounded away from zero. Then, under* H_0,

$$T_n \tilde{R}_n^1 \to B \circ F \quad \textit{in distribution in the space } D[-\infty, \infty).$$

The process $T_n \tilde{R}_n^1$ may be completely computed from a given set of data. The assumptions in Theorem 1.3 are trivially satisfied for homoscedastic linear models whenever $g_1, \ldots, g_p$ are F-integrable.

For small to moderate sample size n, the accuracy of the approximation in Theorem 1.3 gets worse for large values of x. This is because for T_n we have to replace $A^{-1}(x)$ by $A_n^{-1}(x)$. These matrices are unbounded on the whole real line and often not uniformly continuous in the underlying parameter θ. Consequently the underlying processes may become, for given sample size n, very unstable in the extreme right tails. Hence test statistics based on all of $T_n \tilde{R}_n^1$ may not attain the given level. Consequently, for a given n, we have to restrict $T_n \tilde{R}_n^1$ to compact intervals $[-\infty, x_0]$.

In the next section we will report on several simulation results which are designed to demonstrate, for finite sample size n, the accuracy of the distributional approximations and the power of the tests. Both homoscedastic and heteroscedastic models will be investigated. Power will be studied under fixed and local alternatives. Of course a practical choice of x_0 should depend on the data. In our study we will consider for x_0 the 99% quantile of the X-data.

We shall first consider the Cramér–von Mises (CvM) test associated with $T_n\tilde{R}_n^1$. This is an omnibus test and not designed to detect any particular deviation from the null model. Therefore we shall also study linear statistics based on $T_n\tilde{R}_n^1$. They constitute approximations to the Neyman–Pearson test statistic when the alternative approaches the hypothetical model from a given direction. Some robustness properties of this test will be addressed only briefly.

2. A simulation study. In this section we will show that the asymptotic results provide good approximations for small sample sizes. We first consider the homoscedastic case. The constant σ^2 has been estimated from the whole sample by the normalized residual sum of squares

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}[Y_i - m(X_i, \theta_n)]^2.$$

As a test we consider the Cramér–von Mises test associated with $T_n\tilde{R}_n$. Put

$$\begin{aligned} W_n^2 &\equiv W_n^2(x_0) \\ &= \sigma_n^{-2}\int_{-\infty}^{x_0}[T_n\tilde{R}_n(x)]^2 F_n(dx) = \sigma_n^{-2}n^{-1}\sum_{i=1}^{n} 1_{\{X_i \le x_0\}}[T_n\tilde{R}_n(X_i)]^2. \end{aligned}$$

For a continuous F, Theorem 1.3 together with the continuous mapping theorem imply that, in distribution,

$$\begin{aligned} W_n^2 &\to \int_{-\infty}^{x_0}[B \circ F(x)]^2 F(dx) \\ &= \int_0^{F(x_0)} B^2(u)\,du = F^2(x_0)\int_0^1 B^2(u)\,du, \end{aligned}$$

where the last equality follows from the scaling property of B. Thus to get an asymptotically distribution free test statistic, we have to set

$$\tilde{W}_n^2 = \sigma_n^{-2}F_n^{-2}(x_0)\int_{-\infty}^{x_0}[T_n\tilde{R}_n(x)]^2 F_n(dx).$$

From the above,

$$\tilde{W}_n^2 \to \int_0^1 B^2(u)\,du \quad \text{in distribution.}$$

In our simulation study we let

$$\mathscr{M} = \{m(\cdot, \theta)\colon m(x, \theta) = \theta x\}$$

be the family of linear functions through the origin. We generated $n = 200$ data (X_i, Y_i) according to

$$Y_i = 5X_i + aX_i^2 + \varepsilon_i, \qquad 1 \le i \le n.$$

Hence H_0: $m \in \mathscr{M}$ holds with $\theta_0 = 5$ if and only if $a = 0$. The regressor X_i is uniformly distributed on the unit interval, while ε_i is taken independently from a normal $\mathscr{N}(0, \sigma^2)$ distribution. For x_0 we always chose the 99% quantile of the X-data. The preassigned significance levels were $\alpha = 0.05$ and $\alpha = 0.01$.

TABLE 1
Percentages of times H_0 was rejected: $F = U(0, 1)$

		$\tilde{W}_n^2$-test	
a	σ^2	$\alpha = 0.05$ (%)	$\alpha = 0.01$ (%)
0	1	5.7%	1.5%
	2	5.2%	0.8%
	3	4.9%	0.8%
1	1	30.0%	12.7%
	2	19.8%	7.5%
	3	15.4%	4.5%
2	1	81.8%	58.8%
	2	52.2%	28.8%
	3	39.0%	18.6%

Various values for σ^2 and a were considered. In each case we list the percentages of times H_0 was rejected. The asymptotic critical values for $\int B^2$ were taken from Shorack and Wellner [(1986), page 748]. All reported values are based on 1000 replications of $\tilde{W}_n^2$. Here and in the following, the computations were done with MATLAB.

In Table 1, we see that under H_0 the actual percentages of times H_0 was rejected is close to the nominal level. Under the alternative, of course, the distribution of $\tilde{W}_n^2$ depends on the ingredients of the model, namely a, σ^2 and F. The power decreases as σ^2 increases, while it gets larger with a. These effects have been discussed in detail for the original process $\tilde{R}_n$, from both the theoretical and applied viewpoint, in Stute (1997). To further discuss the role of F, and for forthcoming discussions, we investigate $\tilde{W}_n^2$ under the local alternatives $a = n^{-1/2}$. Table 2 exhibits the power of the $\tilde{W}_n^2$-test when F is

TABLE 2
Percentages of times H_0 was rejected: $F = U(0, 1)$, $a = n^{-1/2}$

		$\tilde{W}_n^2$-test	
n	σ^2	$\alpha = 0.05$ (%)	$\alpha = 0.01$ (%)
50	1	6.6%	1.5%
	2	6.3%	1.0%
	3	5.7%	0.9%
100	1	6.5%	1.3%
	2	5.2%	0.9%
	3	5.2%	0.8%
200	1	6.2%	1.3%
	2	5.3%	1.0%
	3	5.1%	0.9%

TABLE 3
Percentages of times H_0 was rejected: $F = U(-2, 2)$, $a = n^{-1/2}$

		$\tilde{W}_n^2$-test	
n	σ^2	$\alpha = 0.05$ (%)	$\alpha = 0.01$ (%)
50	1	27.5%	9.5%
	2	15.9%	3.5%
	3	11.9%	2.0%
100	1	26.1%	9.0%
	2	15.6%	4.4%
	3	12.1%	3.3%
200	1	24.5%	9.2%
	2	16.5%	4.9%
	3	12.1%	4.0%

again the uniform distribution on $(0, 1)$, while in Table 3 the underlying F is the uniform distribution on $(-2, 2)$. In the first case, the power of the $\tilde{W}_n^2$-test is poor, since on the support of F, the unit interval (0,1), the true regression function $5x + n^{-1/2}x^2$ is almost indistinguishable from $5x$, particularly when noise is inherent. In the second case, we also have information on $(-2, 0)$. Compared to $(0, 2)$ the two functions are now more apart resulting in a larger power of the test.

Table 4 exhibits the results for a Crámer–von Mises test when the errors are heteroscedastic. The model considered was

$$Y_i = 5X_i + aX_i^2 + \left(1 + \frac{X_i^2}{2}\right)\varepsilon_i,$$

where $\varepsilon_i \sim \mathcal{N}(0, 1)$. The total sample size was $2n$. Half of the sample was used for estimating the conditional variance. As an estimator we took the

TABLE 4
Percentages of times H_0 was rejected: $F = U(0, 1)$; $\sigma^2(x) = 1 + x^2/2$

		$\tilde{W}_n^2$-test	
n	a	$\alpha = 0.05$ (%)	$\alpha = 0.01$ (%)
100	0	4.3%	0.6%
	1	9.0%	1.0%
	2	16.4%	2.4%
200	0	5.6%	1.1%
	1	11.3%	3.6%
	2	27.2%	12.3%

Nadaraya–Watson estimator with Gaussian kernel. The bandwidth was $h = cn^{-1/2}$ with $c = 0.5$. We only mention in passing that in the present context, when it is not our primary goal to estimate $\sigma^2(x)$ but to use it as part of an empirical integral, h's of the order $n^{-1/2}$ are preferable to the usual $n^{-1/5}$ order which traditionally occurs in curve estimation. See also Stute and González Manteiga (1990) for related phenomena in the context of smoothed linear regression estimators. For θ_n we took the weighted LSE.

Note that in the last example noise increases as we let x tend to 1. This makes the problem of checking the validity of H_0 more difficult, since it is the neighborhood of 1 where the deviation from the null model is largest but this is blurred by increasing noise.

We now demonstrate how Theorem 1.3 may be utilized to derive optimal tests for H_0 versus

$$H_1\colon m(x) = m(x, \theta_0) + n^{-1/2} r(x) \quad \text{for some } \theta_0,$$

that is, the deviation from the null model is specified through a direction r. Only the homoscedastic case is dealt with. As in our first example the test statistic is based on $\sigma_n^{-1} T_n \tilde{R}_n$ and therefore asymptotically distribution-free under H_0 modulo the time transformation F. The alternative H_1 is local as we approach H_0 at (the unknown) θ_0 from the direction r at the rate $n^{-1/2}$. Again θ_n will be the least squares estimator.

For the original processes $\tilde{R}_n$ rather than $T_n \tilde{R}_n$, the distributional theory was studied in detail in Section 2 of Stute (1997). In particular it was pointed out there that under H_1 the limit of $\tilde{R}_n$ equals, when $\sigma^2 = 1$,

$$B \circ F - G^t V + \int_{-\infty}^{\bullet} r(u) F(du) - G^t v.$$

Here v is a deterministic vector depending on F, g and r. In view of (1.5) and (1.6) the limit of $T_n \tilde{R}_n$ under H_1 becomes

$$B \circ F + T\left[\int_{-\infty}^{\bullet} r(u) F(du)\right]. \tag{2.1}$$

We thus see that, when applying the transformation T, the random part is mapped into Brownian motion $B \circ F$, while the deterministic part is replaced by $T[\int^{\bullet} r dF]$.

Now, set

$$\lambda_j = \frac{1}{(j - \frac{1}{2})^2 \pi^2} \tag{2.2}$$

and

$$l_j(t) = \sqrt{2} \sin\left[\left(j - \tfrac{1}{2}\right)\pi t\right], \qquad j = 1, 2, \ldots, \tag{2.3}$$

the eigenvalues and eigenfunctions in the Karhunen–Loève decomposition of B. Equation (2.1) differs from B only in a nonrandom shift. Consequently (2.1) apart from a shift admits the same decomposition as B. As was pointed

out in Stute (1997), the principal components play a crucial role for deriving optimal tests for H_0 versus H_1. For the original processes, the computation of the eigenvalues and eigenfunctions is not trivial at all and requires some numerical work. One major advantage of our transformation approach is that principal components of $T_n\tilde{R}_n$ are given for free, namely by (2.2) and (2.3). The Neyman–Pearson test for H_0 versus H_1 based on $T_n\tilde{R}_n$ is defined as follows. Let

$$s(x) = T\left[\int_{-\infty}^{\bullet} r(u)F(du)\right](x)$$

be the function appearing in (2.1), and put

$$\tau_j = \int s(x)l_j(F(x))F(dx).$$

Furthermore, take

$$\rho_j = \int T\tilde{R}_n(x)l_j(F(x))F(dx).$$

The (approximate) Neyman–Pearson level α-test consists of rejecting H_0 in favor of H_1 iff

$$S = \sum_{j=1}^{\infty} \frac{\tau_j\rho_j}{\lambda_j\gamma} \geq c_{1-\alpha},$$

where $c_{1-\alpha}$ is the $1-\alpha$ quantile of a standard normal distribution and

$$\gamma^2 = \sum_{j=1}^{\infty} \frac{\tau_j^2}{\lambda_j}.$$

In practice we have to replace T by T_n and F by F_n leading to estimators $\hat{\tau}_j$ and $\hat{\rho}_j$. Also the series needs to be truncated at some finite integer j_0.

In order to stabilize $\hat{\rho}_j$, integration should again be restricted to $(-\infty, x_0]$ where x_0 is the 99%-quantile of the X-data. To obtain a properly standardized $\hat{\rho}_j$, we have to set

$$\hat{\rho}_j = \int_{-\infty}^{x_0} T_n\tilde{R}_n(x)l_j\left(\frac{F_n(x)}{p}\right)F_n(dx), \qquad p = 0.99.$$

Actually, since from the scaling property of the Brownian motion,

$$\hat{\rho}_j \to \int_0^p B(u)l_j\left(\frac{u}{p}\right)du = p^{3/2}\int_0^1 B(u)l_j(u)\,du,$$

our final test statistic is

$$\hat{S} = \sum_{j=1}^{j_0} \frac{\hat{\tau}_j\hat{\rho}_j}{p^{3/2}\lambda_j} \Bigg/ \sqrt{\sum_{j=1}^{j_0} \frac{\hat{\tau}_j^2}{\lambda_j}}.$$

TABLE 5
Percentages of times H_0 was rejected: $a = n^{-1/2}$; homoscedastic errors; $F = U(0, 1)$, $j_0 = 4$

		$\hat{S}$-test	
n	σ^2	$\alpha = 0.05$ (%)	$\alpha = 0.01$ (%)
50	1	5.4%	1.0%
	2	5.4%	0.8%
	3	4.9%	0.8%
100	1	6.3%	1.2%
	2	6.0%	1.2%
	3	5.4%	1.0%
200	1	6.4%	2.0%
	2	6.2%	1.4%
	3	5.6%	1.2%

Here $\hat{S}$ is asymptotically standard normal, large values of $\hat{S}$ being significant for a deviation from H_0 in favor of H_1. Tables 5 and 6 present the (estimated) local power of the $\hat{S}$-test in the homoscedastic case with local alternatives.

Compared to Table 3 we see from Table 6 that there is a strong evidence that the approximate Neyman–Pearson Test based on $\hat{S}$ outperforms the CvM test. The values in Tables 2 and 5 are both close to the nominal level since on $(0, 1)$, with $a = n^{-1/2}$, the alternative is very close to the null model.

Finally, we would like to mention that the test statistic becomes unstable if in $\hat{S}$ the truncation parameter j_0 is chosen too large. This is mainly because

TABLE 6
Percentages of times H_0 was rejected: $a = n^{-1/2}$; homoscedastic errors; $F = U(-2, 2)$, $j_0 = 4$

		$\hat{S}$-test	
n	σ^2	$\alpha = 0.05$ (%)	$\alpha = 0.01$ (%)
50	1	38.2%	24.9%
	2	24.1%	15.2%
	3	19.1%	10.0%
100	1	40.2%	26.8%
	2	21.6%	13.2%
	3	18.0%	9.6%
200	1	44.1%	30.6%
	2	23.4%	21.1%
	3	16.0%	7.6%

the weights λ_j tend to zero very rapidly, so that for larger j the summands are too much upweighted if $\hat{\tau}_j$ and $\hat{\rho}_j$ due to sampling errors do not match the true Fourier coefficients sufficiently well. On the other extreme, the CvM statistic $\tilde{W}_n^2$ has a series expansion in which the λ_j downweight the coefficients so that high-frequency alternatives to the null model may not be detected. As a compromise, one may base a test on a statistic

$$\hat{S}_0 = \sum_{j=1}^{j_0} \frac{\hat{\tau}_j \hat{\rho}_j}{p^{3/2} \lambda_j^{1/2}} \Bigg/ \sqrt{\sum_{j=1}^{j_0} \hat{\tau}_j^2},$$

where the new weights $\lambda_j^{1/2}$ are chosen so as to standardize $\hat{\rho}_j$. As a conclusion of our findings, we see that the residual cusums properly transformed give rise to a process which may serve as a basis for various test statistics. Asymptotically this process is distribution free modulo a transformation in time. All involved quantities are linear or at most quadratic and easy to compute. Asymptotic distributional theory provides satisfactory approximations already for small to moderate sample size.

3. Proofs.

LEMMA 3.1. *We have*

$$\operatorname{Cov}[TR_\infty^0(r), TR_\infty^0(s)] = F(r \wedge s), \tag{3.1}$$

that is, (1.7) *and hence Theorem* 1.1 *hold*.

PROOF. Assume $r \le s$. By definition of T, the left-hand side of (3.1) equals

$$\begin{aligned}
&\operatorname{Cov}[R_\infty^0(r), R_\infty^0(s)] \\
&\quad - \operatorname{Cov}\left[R_\infty^0(s), \int_{-\infty}^{r} \sigma^{-1}(y) g^t(y) A^{-1}(y) \int_y^\infty \sigma^{-1}(z) g(z) R_\infty^0(dz) F(dy) \right] \\
&\quad - \operatorname{Cov}\left[R_\infty^0(r), \int_{-\infty}^{s} \sigma^{-1}(y) g^t(y) A^{-1}(y) \int_y^\infty \sigma^{-1}(z) g(z) R_\infty^0(dz) F(dy) \right] \\
&\quad + \operatorname{Cov}\left[\int_{-\infty}^{r} \sigma^{-1}(y) g^t(y) A^{-1}(y) \int_y^\infty \sigma^{-1}(z) g(z) R_\infty^0(dz) F(dy), \right. \\
&\qquad \left. \int_{-\infty}^{s} \sigma^{-1}(y) g^t(y) A^{-1}(y) \int_y^\infty \sigma^{-1}(z) g(z) R_\infty^0(dz) F(dy) \right].
\end{aligned}$$

The first covariance equals $F(r)$. Upon using rules for stochastic integrals, the second covariance is seen to become

$$\int_{-\infty}^{r} \sigma^{-1}(y) g^t(y) A^{-1}(y) \int_y^s \sigma^{-1}(z) g(z) F(dz) F(dy).$$

Similarly, the third and fourth covariances equal

$$\int_{-\infty}^{r} \sigma^{-1}(y) g^t(y) A^{-1}(y) \int_y^r \sigma^{-1}(z) g(z) F(dz) F(dy)$$

and

$$\int_{-\infty}^{r}\int_{-\infty}^{s}\sigma^{-1}(y_1)g^t(y_1)A^{-1}(y_1)A(y_1\vee y_2)A^{-1}(y_2)\sigma^{-1}(y_2)g(y_2)F(dy_2)F(dy_1),$$

respectively. Summation and an application of the Fubini Theorem complete the proof. □

LEMMA 3.2. *Under* (A) *and* (B), *we have in probability and uniformly on compacta*

$$T\tilde{R}_n^0 = TR_n^0 + o_{\mathbb{P}}(1), \tag{3.2}$$

that is, (1.9) *holds*.

PROOF. Recall that

$$\begin{aligned} T\tilde{R}_n^0(x) = \tilde{R}_n^0(x) - \int_{-\infty}^{x}\sigma^{-1}(y)g^t(y)A^{-1}(y) \\ \times\left[\int_y^{\infty}\sigma^{-1}(z)g(z)\tilde{R}_n^0(dz)\right]F(dy) \end{aligned} \tag{3.3}$$

and

$$\begin{aligned} TR_n^0(x) = R_n^0(x) - \int_{-\infty}^{x}\sigma^{-1}(y)g^t(y)A^{-1}(y) \\ \times\left[\int_y^{\infty}\sigma^{-1}(z)g(z)R_n^0(dz)\right]F(dy). \end{aligned} \tag{3.4}$$

Fix some finite x_0. Under (A) and (B), we get uniformly in $x \le x_0$,

$$\tilde{R}_n^0(x) = R_n^0(x) - G_0^t(x)n^{1/2}(\theta_n - \theta_0) + o_{\mathbb{P}}(1).$$

The two integrals in (3.3) and (3.4) differ by

$$\begin{aligned} n^{-1/2}\sum_{i=1}^{n}\int_{-\infty}^{x}\sigma^{-1}(y)g^t(y)A^{-1}(y)1_{(y,\infty)}(X_i)\sigma^{-2}(X_i)g(X_i) \\ \times[m(X_i,\theta_n) - m(X_i,\theta_0)]F(dy), \end{aligned}$$

which is easily seen to be equal to

$$\int_{-\infty}^{x}\sigma^{-1}(y)g^t(y)A^{-1}(y)\int_y^{\infty}\sigma^{-2}(z)g(z)g^t(z)F(dz)F(dy)n^{1/2}(\theta_n - \theta_0) + o_{\mathbb{P}}(1).$$

By definition of A, the last double integral equals $G_0^t(x)$, however, so that (3.2) holds at least pointwise in x. Uniformity is obtained by applying a standard Glivenko–Cantelli argument. First, it suffices to consider real-valued g's only. Decompose all involved functions into their positive and negative parts so that the resulting sums and integrals are monotone in x. Together with the SLLN, uniformity in convergence is then obtained along classical lines, namely by reducing the sup over all x to a sup over a properly chosen finite grid. □

LEMMA 3.3. *Under* (A) *and* (B), *we have*

$$TR_n^0 \to B \circ F \text{ in distribution},$$

that is, (1.10) *holds*.

PROOF. TR_n^0 is a sum of i.i.d. processes. Tightness may be shown by standard arguments. Convergence of the finite-dimensional distributions follows from the multivariate CLT. Specification of the covariance structure is similar to the proof of Lemma 3.1.

For the proof of Theorem 1.3 we shall make use of Lemma 3.1 of Chang (1990), which is (properly re-)stated here just for the sake of reference.

LEMMA 3.4. *Let* V *be a relatively compact subset of* $D[-\infty, x_0]$. *Then with probability* 1

$$\int_{-\infty}^{t} v(x)[F_n(dx) - F(dx)] \to 0 \text{ as } n \to \infty,$$

uniformly in $t \le x_0$ *and* $v \in V$.

We shall apply Lemma 3.4 within the following context: with F_{n_1} rather than F_n, let $\{\alpha_n\}$ be a sequence of stochastic processes which are uniformly tight, that is, for a given $\varepsilon > 0$ there exists a compact set V such that $\alpha_n \in V$ with probability at least $1 - \varepsilon$. Apply Lemma 3.4 with this V and observe that $\alpha_n \notin V$ with small probability to finally get, uniformly in t,

$$\int_{-\infty}^{t} \alpha_n(x)[F_{n_1}(dx) - F(dx)] \to 0 \quad \text{in probability.} \tag{3.5}$$

The above integrals appear as remainder terms when expanding $T_n \tilde{R}_n^1$ as a sum of independent processes to which an invariance principle applies.

PROOF OF THEOREM 1.3. First, similarly to the proof of Lemma 3.2, we obtain, upon using the fact that σ_{n_1} is bounded away from zero, that

$$T_n \tilde{R}_n^1 = T_n R_n^1 + o_{\mathbb{P}}(1),$$

uniformly in $x \le x_0$. Secondly, note that

$$\int_{-\infty}^{x} \sigma_{n_1}^{-1}(y) g^t(y, \theta_{n_1}) A_{n_1}^{-1}(y) \int_{y}^{\infty} \sigma_{n_1}^{-1}(z) g(z, \theta_{n_1}) R_n^1(dz) F_{n_1}(dy)$$
$$- \int_{-\infty}^{x} \sigma_{n_1}^{-1}(y) g^t(y) A^{-1}(y) \int_{y}^{\infty} \sigma_{n_1}^{-1}(z) g(z) R_n^1(dz) F(dy)$$

$$= \int_{-\infty}^{x} \sigma_{n_1}^{-1}(y) g^t(y) A^{-1}(y) \int_y^{\infty} \sigma_{n_1}^{-1}(z) g(z) R_n^1(dz)[F_{n_1}(dy) - F(dy)] \tag{3.6}$$

$$+ \int_{-\infty}^{x} \Bigg[\sigma_{n_1}^{-1}(y) g^t(y, \theta_{n_1}) A_{n_1}^{-1}(y) \int_y^{\infty} \sigma_{n_1}^{-1}(z) g(z, \theta_{n_1}) R_n^1(dz) - \sigma_{n_1}^{-1}(y) g^t(y) A^{-1}(y) \int_y^{\infty} \sigma_{n_1}^{-1}(z) g(z) R_n^1(dz) \Bigg] F_{n_1}(dy) \tag{3.7}$$

Putting

$$\alpha_n(y) = \sigma_{n_1}^{-1}(y) g^t(y) A^{-1}(y) \int_y^{\infty} \sigma_{n_1}^{-1}(z) g(z) R_n^1(dz),$$

it is not difficult to see that along with (1.12) and the boundedness of σ_{n_1} the sequence $\{\alpha_n\}$ is tight. Hence Lemma 3.4, respectively, (3.5) applies. We conclude that (3.6) tends to zero uniformly in $x \le x_0$.

From assumption (B) and the boundedness of $\sigma_{n_1}^{-1}$, we obtain that the processes β_n defined by

$$\beta_n(x, \theta) = \int_{-\infty}^{x} \sigma_{n_1}^{-1}(y) g^t(y, \theta) A_{n_1}^{-1}(y, \theta) \int_y^{\infty} \sigma_{n_1}^{-1}(z) g(z, \theta) R_n^1(dz) F_{n_1}(dy)$$

are uniformly tight and continuous in θ. But $\theta_{n_1} \to \theta_0$ in probability so that the integral in (3.7) tends to zero in probability as n and $n_1 \to \infty$.

Altogether we see that up to a negligible error,

$$T_n R_n^1 = T R_n^1,$$

where in the above T, σ has to be replaced by σ_{n_1}. Now mimic the proof of Theorem 1.2 to complete the proof of Theorem 1.3. □

Acknowledgment. Li-Xing Zhu's work was done at the University of Giessen while on leave from the Chinese Academy of Sciences, Beijing.

REFERENCES

CHANG, N. M. (1990). Weak convergence of a self-consistent estimator of a survival function with doubly censored data. *Ann. Statist.* **18** 391–404.

DEVROYE, L. P. and WAGNER, T. J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Ann. Statist.* **8** 231–239.

DURBIN, J. (1973). Weak convergence of the sample distribution function when parameters are estimated. *Ann. Statist.* **1** 279–290.

KHMALADZE, E. V. (1981). Martingale approach in the theory of goodness-of-fit tests. *Theory Probab. Appl.* **26** 240–257.

KHMALADZE, E. V. (1988). An innovation approach to goodness-of-fit tests in $\mathbb{R}^m$. *Ann. Statist.* **16** 1503–1516.

MARONNA, R. A. and YOHAI, V. J. (1981). Asymptotic behavior of general M-estimates for regression and scale with random carriers. *Z. Wahrsch. Verw. Gebiete* **58** 7–20.

POLLARD, D. (1984). *Convergence of Stochastic Processes.* Springer, New York.

SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York.

SPIEGELMAN, C. and SACKS, J. (1980). Consistent window estimation in nonparametric regression. *Ann. Statist.* **8** 240–246.

STONE, C. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–620.
STUTE, W. (1994). Universally consistent conditional U-statistics. *Ann. Statist.* **22** 460–473.
STUTE, W. (1997). Nonparametric model checks for regression. *Ann. Statist.* **25** 613–641.
STUTE, W. and GONZÁLEZ MANTEIGA, W. (1990). Nearest neighbor smoothing in linear regression. *J. Multivariate Anal.* **34** 61–74.
STUTE, W., GONZÁLEZ MANTEIGA, W. and PRESEDO QUINDIMIL, M. (1998). Bootstrap approximations in model checks for regression. *J. Amer. Statist. Assoc.* **93** 141–149.

W. STUTE
S. THIES
MATHEMATICS INSTITUTE
UNIVERSITY OF GIESSEN
ARNDTSTR. 2, D-35392 GIESSEN
GERMANY
E-MAIL: winfried.stute@math.uni-giessen.de

L.-X. ZHU
INSTITUTE OF APPLIED MATHEMATICS
CHINESE ACADEMY OF SCIENCES
BEIJING, 100080
CHINA