Semiparametric inference in general heteroscedastic regression models

Cédric Heuchenne[1], Ingrid Van Keilegom[2]

Abstract

Suppose the random vector (X, Y) satisfies the regression model $Y = m(X) + \sigma(X)\varepsilon$, where $m(\cdot) = E(Y|\cdot)$, $\sigma^2(\cdot) = \mathrm{Var}(Y|\cdot)$ and ε is independent of X. The covariate X is d-dimensional ($d \geq 1$), the response Y is one-dimensional, and m and σ are unknown but smooth functions. New goodness-of-fit testing procedures for parametric forms of the residuals distribution are proposed. They can be considered as a first step providing information that can be largely used for further inference on $m(\cdot)$ and $\sigma(\cdot)$. The methodology is described in practice and the asymptotic properties of the corresponding statistics are developed.

1 Introduction

Suppose the d-dimensional random vector X and the random variable Y satisfy the following heteroscedastic regression model :

$$Y = m(X) + \sigma(X)\varepsilon, \tag{1}$$

where ε (with $F_\varepsilon(\cdot) = P(\varepsilon \leq \cdot)$) is independent of the random vector X, $m(X) = E[Y|X]$ and $\sigma^2(X) = \mathrm{Var}[Y|X]$.

If many efforts were achieved to test hypotheses concerning $m(\cdot)$, $\sigma^2(\cdot)$ and $F_\varepsilon(\cdot)$, usual behavior when the aim is to completely identify model (1) seems to stay more classical and is summarized in the sequel. First, parametric curves are fitted on $m(\cdot)$ and $\sigma(\cdot)$ and second, shape of the so-obtained residuals distribution and independency between X and ε are tested. Some improvements of this methodology are straigthforward. On one side, its robustness could be increased since it involves parametric conditional moments when inferring on the residuals and, on the other side, related inferential procedures on $m(\cdot)$ and $\sigma(\cdot)$ could be more efficient when preliminarily assessing a particular parametric residuals distribution.

Therefore, the scope of this paper is to suggest inversion of complete identification of model (1) by starting inference from the residuals distribution. In

[1]University of Liège, E-mail: C.Heuchenne@ulg.ac.be
[2]Catholic University of Louvain, E-mail: ingrid.vankeilegom@uclouvain.be

this framework, we propose to consider the combined estimation and testing procedure for a parametric residuals distribution without assuming any parametric information for $m(\cdot)$ and $\sigma(\cdot)$. Thus, we aim at testing the hypothesis

$$H_0 : F_\varepsilon(y) \in \mathcal{F} \text{ versus } H_1 : F_\varepsilon(y) \notin \mathcal{F}, \tag{2}$$

where $\mathcal{F} = \{F_{\varepsilon\theta} : \theta \in \Theta\}$ is a class of parametric distributions, Θ is a compact subset of $I\!R^k$ and k is a positive integer. We denote the true value of θ by θ_0. In this test, parameters assumed under H_0 are estimated by a maximum likelihood approach before being introduced into specific statistics. Under the null hypothesis, estimators of those parameters and the derived process are shown to reach the same optimal rate of convergence as in the usual case where $m(\cdot)$ and $\sigma(\cdot)$ are parametric functions. In practice, the resulting tests seem to be sensitive to the bandwidthes selection procedure since power largely varies according to the choice of the smoothing parameters for $m(\cdot)$ and $\sigma(\cdot)$. In particular, power looks higher when loss functions using (a cross validation version of) the residuals themselves are involved. Moreover, direct applications in this context of the important theoretical extension of residuals distribution estimation to the multiple regression case (recently studied by Neumeyer and Van Keilegom, 2008) enable to study its behavior in practice.

The paper is organized as follows. In the methodological part (Section 2), the estimator of θ_0 and the testing procedure are described in detail and in the theoretical part (Section 3), the main asymptotic results are summarized including the asymptotic normality of the estimator for θ_0 and the weak convergence of the proposed test statistics under H_0.

2 Description of the method

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be an i.i.d. random sample generated from model (1), where the components of X_i are denoted by $(X_{i1}, \ldots, X_{id})$ $(i = 1, \ldots, n)$. The distribution of ε and X are denoted by F_ε and F_X respectively, and their probability density functions by f_ε and f_X (in the same way, $f_{\varepsilon\theta}(y) = dF_{\varepsilon\theta}(y)/dy$). We start by estimating the regression function $m(x)$ and the variance function $\sigma^2(x)$ at an arbitrary point $x = (x_1, \ldots, x_d)$ in the support R_X of X, which we suppose to be a compact subset of $I\!R^d$. We estimate $m(x)$ by a local polynomial estimator of degree p, i.e. $\hat{m}(x) = \hat{\beta}_0$, where $\hat{\beta}_0$ is the first component of $\hat{\beta}$, which is the solution of the local minimization problem

$$\min_\beta \sum_{i=1}^{n} \{Y_i - P_i(\beta, x, p)\}^2 K_h(X_i - x), \tag{3}$$

where $P_i(\beta, x, p)$ is a polynomial of order p built up with all $0 \le k \le p$ products of factors of the form $X_{ij} - x_j$ $(j = 1, \ldots, d)$, and β is the vector of all coefficients of this polynomial. Here, for $u = (u_1, \ldots, u_d) \in I\!R^d$, $K(u) = \prod_{j=1}^{d} k(u_j)$ is a $d-$dimensional product kernel, k is a univariate kernel function, $h = (h_1, \ldots, h_d)$ is a $d-$dimensional bandwidth vector converging to zero when n tends to infinity and

$K_h(u) = \prod_{j=1}^d k(u_j/h_j)/h_j$. In the same way, $\hat{\sigma}^2(x) = \hat{\gamma}_0$ is the first component of $\hat{\gamma}$, which is the solution of the local minimization problem

$$\min_{\gamma} \sum_{i=1}^{n} \{(Y_i - \hat{m}_l(X_i))^2 - P_i(\gamma, x, q)\}^2 K_g(X_i - x), \tag{4}$$

where $\hat{m}_l(X_i)$, $i = 1, \ldots, n$, is a local polynomial estimator obtained from (3) (where h is replaced by l) and $P_i(\gamma, x, q)$, γ, $K_g(u)$ and $g = (g_1, \ldots, g_d)$ are defined in a similar way as $P_i(\beta, x, p)$, β, $K_h(u)$ and h. An estimator for $\sigma(x)$, $\hat{\sigma}(x)$, will be simply obtained by taking the square root of $\hat{\gamma}_0$.

The nonparametric residuals can then be introduced into the likelihood function which will provide a vector of parameter estimators θ_n for θ_0 by solving the maximization problem

$$\max_{\theta \in \Theta} \sum_{i=1}^{n} \log f_{\varepsilon\theta}(\hat{\varepsilon}_i), \tag{5}$$

where $\hat{\varepsilon}_i = (Y_i - \hat{m}(X_i))/\hat{\sigma}(X_i)$ $(i = 1, \ldots, n)$.

Remark 2.1 (Choice of the smoothing parameters) The objective is to provide an easy and data-driven way to select the smoothing parameters in (3) and (4). To this end, we propose (with $h_1 = \cdots = h_g$ and $h = g = l$ in (3) and (4))

$$h_n = \text{argmin}_h \sum_{j=1}^{n} \frac{(Y_j - \hat{m}_{h,-j}(X_j))^2}{\hat{\sigma}^2_{h,-j}(X_j)}, \tag{6}$$

where $\hat{m}_{h,-j}(X_j)$ is a local polynomial estimator obtained by an expression of the type (3) for $x = X_j$ but based on a sample for which the j^{th} data point has been removed. Moreover, $\hat{\sigma}^2_{h,-j}(X_j)$ is a local polynomial estimator obtained from an expression of the type (4) for $x = X_j$, but based on the couples $(X_i, (Y_i - \hat{m}_{h,-i}(X_i))^2)$ $(i = 1, \ldots, j-1, j+1, \ldots, n)$. In practice, this technique gives nice power for the testing procedures discussed in this paper. Indeed, even though least squares cross-validation in this context usually lead to overfitting, smoothing is here improved thanks to the particular cross-validation version of the estimated conditional variance at the denominator of (6). Simulations results highlighting this feature can be obtained by request to the authors.

Remark 2.2 (Order of local polynomials) Apart from the practical choice h_n discussed above, it is important to mention here the dependency between h (g, l) and the dimension d of X. Indeed, the sample size n should increase exponentially with d to preserve the convergence rates (curse of dimensionality). Consequently, for fixed sample size n, in order to compensate for this curse of dimensionality, the bandwidthes h_j, g_j and l_j $(j = 1, \ldots, d)$ should increase exponentially with $1/d$ (see condition (C4) in the next section). An indirect consequence of this is that the degree of the polynomials $P_i(\beta, x, p)$ and $P_i(\gamma, x, q)$ should increase when d increases (this follows from condition (C4)). For example, we will have to choose

p and q at least equal to 2 when $d = 2$ (local quadratic estimators), and at least equal to 4 when $d = 3$ (order 4 local polynomial estimators).

Next, the test statistics are constructed from the difference between $F_{\varepsilon\theta_n}(y)$ and the nonparametric estimator of $F_\varepsilon(y)$:

$$\hat{F}_\varepsilon(y) = \frac{1}{n}\sum_{i=1}^{n} I(\hat{\varepsilon}_i \leq y). \tag{7}$$

This estimator was first studied by Akritas and Van Keilegom (2001) and then extended to the case where X is $d-$dimensional by Neumeyer and Van Keilegom (2008). Consider the process

$$W_n(y) = n^{1/2}(\hat{F}_\varepsilon(y) - F_{\varepsilon\theta_n}(y)), \quad -\infty < y < \infty, \tag{8}$$

and define the following test statistics of the Kolmogorov-Smirnov and Cramér-von Mises types :

$$T_{KS} = n^{1/2} \sup_{-\infty<y<\infty} |\hat{F}_\varepsilon(y) - F_{\varepsilon\theta_n}(y)|, \tag{9}$$

and

$$T_{CM} = n \int (\hat{F}_\varepsilon(y) - F_{\varepsilon\theta_n}(y))^2 d\hat{F}_\varepsilon(y). \tag{10}$$

3 Asymptotic results

We now turn to the analysis of the asymptotic properties of the estimator θ_n and of the test statistics T_{KS} and T_{CM}. Proofs of the results of this section can be obtained by request to the authors.

In the assumptions in the sequel, we denote for simplicity the true location (resp. scale) function by m_0 (resp. σ_0). For an arbitrary $\theta = (\theta_1 \ \cdots \ \theta_k)^t$, let denote $\dot{f}_{\varepsilon\theta}(y) = (\frac{\partial}{\partial\theta_1} f_{\varepsilon\theta}(y) \ \cdots \ \frac{\partial}{\partial\theta_k} f_{\varepsilon\theta}(y))^t$, $f'_{\varepsilon\theta}(y) = \frac{d}{dy} f_{\varepsilon\theta}(y)$ and

$$\Omega = E\Big[\frac{\dot{f}_{\varepsilon\theta_0}(\varepsilon)\dot{f}^t_{\varepsilon\theta_0}(\varepsilon)}{f^2_{\varepsilon\theta_0}(\varepsilon)}\Big].$$

For arbitrary functions m and $\sigma > 0$ defined on R_X, let

$$G(\theta, m, \sigma) = E\left[\frac{\dot{f}_{\varepsilon\theta}\Big(\frac{Y-m(X)}{\sigma(X)}\Big)}{f_{\varepsilon\theta}\Big(\frac{Y-m(X)}{\sigma(X)}\Big)}\right].$$

(C1) For all $\delta > 0$, there exists an $\varepsilon > 0$ such that $\inf_{\|\theta-\theta_0\|>\delta} \|G(\theta, m_0, \sigma_0)\| > \varepsilon$.

(C2) Uniformly for all $\theta \in \Theta$, $G(\theta, m, \sigma)$ is continuous with respect to the supremum norm in (m, σ) at $(m, \sigma) = (m_0, \sigma_0)$. Moreover, Ω is non-singular.

(C3) k is a symmetric probability density function supported on $[-1,1]$, k is d times continuously differentiable, and $k^{(j)}(\pm 1) = 0$ for $j = 0, \ldots, d-1$.

(C4) h_j, g_j and l_j are of the same order ($j = 1, \ldots, d$) and satisfy $h_j/h^* \to c_j$, $g_j/h^* \to d_j$ and $l_j/h^* \to e_j$ for some $0 < c_j, d_j, e_j < \infty$ and some baseline bandwidth h^*. Moreover, for $r = p$ or q, h^* satisfies $n{h^*}^{2r+4} \to 0$ when r is even, $n{h^*}^{2r+2} \to 0$ when r is odd and $n{h^*}^{3d+\delta} \to \infty$ for some small $\delta > 0$.

(C5) All partial derivatives of F_X up to order $2d+1$ exist on the interior of R_X, they are uniformly continuous and $\inf_{x \in R_X} f_X(x) > 0$.

(C6) All partial derivatives of m_0 and σ_0 up to order $p+2$ exist on the interior of R_X, they are uniformly continuous and $\inf_{x \in R_X} \sigma_0(x) > 0$.

(C7) All (mixed) derivatives up to order 3 of $F_{\varepsilon\theta}(y)$ with respect to y and the components of θ exist and are continuous. Moreover, $\sup_y |y^2 f'_{\varepsilon\theta}(y)| < \infty$ and $E|Y|^6 < \infty$.

Theorem 1. *Assume (C1)-(C7). Then, under H_0,*

$$\theta_n - \theta_0 = -\Omega^{-1}\, n^{-1} \sum_{i=1}^{n} \xi(\varepsilon_i) + o_P(n^{-1/2}),$$

where

$$\xi(t) = \frac{\dot{f}_{\varepsilon\theta_0}(t)}{f_{\varepsilon\theta_0}(t)} + \int \frac{\dot{f}_{\varepsilon\theta_0}(y) f'_{\varepsilon\theta_0}(y)}{f_{\varepsilon\theta_0}(y)} \Big\{ t + \frac{y}{2}(t^2 - 1) \Big\}\, dy.$$

Moreover,

$$n^{1/2}(\theta_n - \theta_0) \xrightarrow{d} N(0, \Omega^{-1} V \Omega^{-1}),$$

where $V = E[\xi(\varepsilon)\xi^t(\varepsilon)]$.

Theorem 2. *Assume (C1)-(C7). Then, under H_0,*

$$\begin{aligned} &\hat{F}_\varepsilon(y) - F_{\varepsilon\theta_n}(y) \\ &= n^{-1} \sum_{i=1}^{n} \Big[I(\varepsilon_i \le y) - F_\varepsilon(y) + \varphi(\varepsilon_i, y) + \dot{F}^t_{\varepsilon\theta_0}(y)\, \Omega^{-1}\, \xi(\varepsilon_i) \Big] + R_n(y), \end{aligned}$$

where $\sup_{-\infty < y < \infty} |R_n(y)| = o_P(n^{-1/2})$, $\dot{F}_{\varepsilon\theta}(y) = (\frac{\partial}{\partial\theta_1} F_{\varepsilon\theta}(y) \;\cdots\; \frac{\partial}{\partial\theta_k} F_{\varepsilon\theta}(y))^t$, and

$$\varphi(z, y) = f_\varepsilon(y) \Big\{ z + \frac{y}{2}(z^2 - 1) \Big\}.$$

Moreover, the process $n^{1/2}(\hat{F}_\varepsilon(y) - F_{\varepsilon\theta_n}(y))$ ($-\infty < y < \infty$) converges weakly to a zero-mean Gaussian process $W(y)$ with covariance function

$$\begin{aligned} Cov(W(y_1), W(y_2)) = E\Big[&\Big\{ I(\varepsilon \le y_1) - F_\varepsilon(y_1) + \varphi(\varepsilon, y_1) + \dot{F}^t_{\varepsilon\theta_0}(y_1)\, \Omega^{-1}\, \xi(\varepsilon) \Big\} \\ &\times \Big\{ I(\varepsilon \le y_2) - F_\varepsilon(y_2) + \varphi(\varepsilon, y_2) + \dot{F}^t_{\varepsilon\theta_0}(y_2)\, \Omega^{-1}\, \xi(\varepsilon) \Big\} \Big]. \end{aligned}$$

As a consequence of the above result, we now obtain the asymptotic limit of the test statistics T_{KS} and T_{CM} under H_0.

Corollary 1. *Assume (C1)-(C7). Then, under* H_0*,*

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

and

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

Remark 3.4 (Convergence under fixed alternatives) Note that if the error distribution F_ε is a fixed distribution (independent of the sample size n) that does not belong to the class $\mathcal{F}$, it can be easily seen that the test statistics T_{KS} and T_{CM} converge to infinity. In fact, the estimators $\hat{F}_\varepsilon$ and $F_{\varepsilon\theta_n}$ do not converge to the same distribution in that case, and hence the process $n^{1/2}(\hat{F}_\varepsilon(y) - F_{\varepsilon\theta_n}(y))$, $-\infty < y < \infty$, diverges.

Remark 3.5 (Bootstrap approximation) To estimate the distributions of the statistics T_{KS} and T_{CM} under H_0, the asymptotic result given in Corollary 1 could be used, with appropriate estimators for the unknown quantities. Alternatively, resampling techniques can provide very good precision. Here, the method we propose to use is as follows. For B fixed and for $b = 1, \ldots, B$,

1. Let $\{\varepsilon^*_{1,b}, \ldots, \varepsilon^*_{n,b}\}$ be an i.i.d. random sample from the distribution $F_{\varepsilon\theta_n}(\cdot)$.
2. Define new responses

$$Y^*_{i,b} = \hat{m}(X_i) + \hat{\sigma}(X_i)\varepsilon^*_{i,b}, \quad i = 1, \ldots, n.$$

3. Let $T^*_{KS,b}$ and $T^*_{CM,b}$ be the test statistics obtained from the bootstrap sample $\{(X_1, Y^*_{1,b}), \ldots, (X_n, Y^*_{n,b})\}$.

Then, if we denote $T^*_{KS,(b)}$ for the b-th order statistic of $T^*_{KS,1}, \ldots, T^*_{KS,B}$ and analogously for $T^*_{CM,(b)}$, then $T^*_{KS,([(1-\alpha)B]+1)}$ and $T^*_{CM,([(1-\alpha)B]+1)}$ approximate the $(1-\alpha)-$quantiles of the distributions of T_{KS} anc T_{CM} respectively (where $[\cdot]$ denotes the integer part).

References

[1] Akritas M.G., Van Keilegom I. (2001) Nonparametric estimation of the residuals distribution. *Scand. J. Statist.*, **28**, 549–567.

[2] Neumeyer N., Van Keilegom I. (2008) *Estimating the error distribution in nonparametric multiple regression with applications to model testing.* Submitted paper.