The Korean Communications
in Statistics Vol. 8, No. 1, 2001
pp. 291-297

On the Residual Empirical Distribution Function of Stochastic Regression with Correlated Errors[1]

Issa Fakhre-Zakeri[2] and Sangyeol Lee[3]

Abstract

For a stochastic regression model in which the errors are assumed to form a stationary linear process, we show that the difference between the empirical distribution functions of the errors and the estimates of those errors converges uniformly in probability to zero at the rate of $o_p(n^{-1/2})$ as the sample size n increases.

Keywords : Residual empirical process; stochastic regression model; linear process.

1. Introduction

Consider a stochastic regression model

$$Y_t = \beta' X_t + Z_t, \quad t = 1,2,\cdots \tag{1}$$

where Y_t is an observed scalar dependent variable, $\beta = (\beta_1, \cdots, \beta_p)'$ is a $p\times 1$ vector of unknown regression coefficients, X_t is an observable p-dimensional stationary process with $EX_t = 0$ and $E\|X_t\|^2 < \infty$, Z_t is an unobservable random disturbance which is independent of X_t and is assumed to be a stationary linear process of the form $Z_t = \sum_{j=0}^{\infty} a_j V_{t-j}$ in which V_t are independent and identically distributed (i.i.d.) random variables with zero mean and finite variance, $|a_j| \le cj^{-q}$, $c>0$, and $q>5$. Here $\| \ \|$ denotes the Euclidean norm. The setting covers a broad range of stochastic process and time series models; it includes

1) This work was supported by the Korea Research Foundation (1998-015-D00047) made in the year of 1998.

2) Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260, U.S.A.
E-mail: ifakhrez@email.unc.edu

3) Department of Statistics, Seol National University, Seoul, 151-742 korea
E-mail : sylee@stats.snu.ac.kr

finite-parameter distributed lag models (or transfer function models) with serially correlated errors.

Let $\widehat{\beta_n}$ denote any estimate of β - such as the least squares or least absolute deviations estimate - based on the observations $(Y_1, X_1), \cdots, (Y_n, X_n)$, and let $\widehat{Z_t}$ denote the t-th residual (estimated disturbance from the regression) defined by

$$\widehat{Z_t} = Y_t - \widehat{\beta_n}' X_t, \quad t = 1, 2, \cdots, n \tag{2}$$

Let $F_n(x) = \frac{1}{n} \sum_{t=1}^{n} I(Z_t \le x)$ denote the empirical distribution function of Z_t based on the n observations. The corresponding residual empirical distribution function is defined by

$$\widehat{F_n}(x) = \frac{1}{n} \sum_{t=1}^{n} I(\widehat{Z_t} \le x) \tag{3}$$

An important application of (3) includes tests of goodness-of-fit of models based on the sample distribution function. For a review of earlier works on empirical processes, see Shorack and Wellner (1986). Asymptotic properties of residual empirical processes have been investigated in regression models with fixed design by Koul (1984) and in AR(p) models by Boldin (1982). For a detailed exposition of residual empirical process see Koul (1992), Koul and Surgailis (1997) and Lee and Wei (1999). The purpose of this paper is to establish the following theorem.

Theorem. *Assume that* $\widehat{\beta_n} - \beta = O_p(n^{-1/2})$, *and that* F, *the distribution function of* Z_t *on the real line* R, *is twice differentiable with* $\sup_{x \in R} |F'(x)| < \infty$ *and* $\sup_{x \in R} |F''(x)| < \infty$. *Further, assume that*

$$n^{-1/2} \sum_{t=1}^{n} X_t = O_p(1) \quad \text{and} \quad \sum_{h=0}^{\infty} |Cov(W_t, W_{t+h})| < \infty,$$

where $W_t = \|X_t\| - E\|X_t\|$. *Then*

$$\sup_{x \in R} n^{1/2} | \widehat{F_n}(x) - F_n(x) | = o_p(1) . \tag{4}$$

2. Proof

The following lemmas are used in the proof of the theorem.

Lemma 1. *For each positive integer* m, *let* F_m *denote the distribution function of* $\sum_{j=0}^{m-1} a_j V_{t-j}$. *Suppose that* F *satisfies* $M := \sup_{x \in R} |F'(x)| < \infty$. *Then there exist* $0 < \nu < 1/2$ *and* $\alpha > 1/2$, *such that* $\sup_{x \in R} |F(x) - F_{m_n}(x)| = O(n^{-\alpha})$, *where* $m_n = n^{\nu}$.

Proof. Let

$$Z_{t,m} = \sum_{j=0}^{m-1} a_j V_{t-j} \quad \text{and} \quad Z^*_{t,m} = \sum_{j=m}^{\infty} a_j V_{t-j},$$

where m is an integer. For arbitrary $\varepsilon > 0$, we write $\phi(m, \varepsilon) = P(\bigcup_{t=1}^{n} |Z^*_{t,m}| > \varepsilon)$.

By elementary inequalities

$$\begin{aligned}\phi(m,\varepsilon) &\leq \varepsilon^{-2} E \max_{1\leq t\leq n} Z^{*2}_{t,m} \\ &\leq \varepsilon^{-2}\Big(\sum_{j=m}^{\infty} (E\max_{1\leq t\leq n} V^2_{t-j})^{1/2}\Big)^2 \leq C\varepsilon^{-2} n \Big(\sum_{j=m}^{\infty} |a_j|\Big)^2, \quad C > 0.\end{aligned}$$

Let us denote $\varepsilon_n = n^{-\alpha}$ for some $\alpha > 1/2$, $\eta_n = \max_{1\leq t\leq n} |Z^*_{t,m_n}|$. Since by assumption $|a_j| \leq cj^{-q}$, $c > 0$, $q > 5$, we have

$$\phi_n := \phi(m_n, \varepsilon_n) \leq Cn^{1+2\alpha}\Big(\sum_{j=m_n}^{\infty} |a_j|\Big)^2 \leq n^{1+2\alpha} O(n^{-2\nu(q-1)}).$$

Consequently, we can find $\nu \in (0, 1/2)$ and $\alpha > 1/2$ such that

$$\begin{aligned}\phi_n &= O(n^{-\rho}) \quad \text{for some } \rho > 0 \\ \eta_n &= O_p(n^{-3/2+x}) \quad \text{for some } x > 0, \qquad (7) \\ \delta_n &:= 2\phi(m_n, \varepsilon_n) + 6M\varepsilon_n = O(n^{-\alpha}).\end{aligned}$$

Using (1.6) and (1.7) of Chanda and Ruymgaart (1990), hereafter referred to as (CR), it is immediate that

$$\sup_{x\in R} |F(x) - F_{m_n}(x)| \leq \delta_n = O(n^{-\alpha}).$$

This completes the proof. □

***Lemma** 2. Suppose that F has bounded first and second derivatives. Let $\lambda > 0$. Define $U_n(x, y) : R^2 \to R$ by*

$$U_n(x,y) = n^{-1} \sum_{t=1}^{n} [I(Z_t \leq x) - F(x) + F(y) - I(Z_t \leq y)].$$

Assume that $a_n < b_n$ satisfy

$$\begin{aligned} n^{1/2}(F(b_n) - F(a_n)) &\to 0 \quad \text{as} \to \infty, \\ n^{\alpha}(F(b_n) - F(a_n)) &\to \infty \quad \text{as} \to \infty. \end{aligned}$$

where α is the number in (5). Then,

$$P(\sup_{a_n \leq x < y \leq b_n} |U_n(x,y)| \geq \lambda) \leq C_1 n^{\nu} \exp(-K\lambda^{1/2-\nu}) + C_2 n^{1-\rho},$$

for sufficiently large n, where the positive constants C_1, C_2, and K are independent of λ, a_n, b_n.

Proof. Define

$$\psi(\lambda) = 2\lambda^{-2}\int_0^{\lambda} \log(1+x)\,dx, \quad \lambda>0\ ;\ \psi(0)=1.$$

Using the integration by parts, we see that

$$\psi(\lambda) = 2\lambda^{-2}(\lambda\log(1+\lambda) - \lambda + \log(1+\lambda)).$$

Hence $\lambda\psi(\lambda)\geq K,\ K>0$, for sufficiently large λ. Now it follows from our assumptions and Theorem 2.1 of (CR), after substituting m_n and ϕ_n in (5) and (7) for m and ϕ, that

$$P(\sup_{a_n\leq x<y\leq b_n}|U_n(x,y)|\geq\lambda) \leq m_n C\exp(-K\lambda n^{1/2}/m_n) + n\phi_n$$
$$\leq C_1 n^{\nu}\exp(-K\lambda n^{1/2-\nu}) + C_2 n^{1-\rho}$$

for sufficiently large n. □

Proof of the Theorem. We only deal with the case where the dimension p of X_t is equal to 1, since the other cases can be handled similarly. Let $N_n = n^{\theta}$, where $1/2 < \theta < \min\{1,\alpha\}$, and let $x_r = \sup\{x: F(x) = r/N_n\}$, $r=1,\cdots,N_n$. For convenience, hereafter we refer to m_n as m. Set $b_{n,t} = (\widehat{\beta}_n - \beta)X_t$. Observe that

$$\sup_{x\in R}\sqrt{n}|\widehat{F_{n(x)}} - F_n(x)|$$
$$= \sup_{x\in R}|n^{-1/2}\sum_{t=1}^{n}[I(Z_t\leq x+b_{n,t}) - I(Z_t\leq x)]|$$
$$\leq \Lambda_1+\Lambda_2+\Lambda_3,$$

where,

$$\Lambda_1 = \sup_{x\in R}|n^{-1/2}\sum_{t=1}^{n}\{I(Z_t\leq x) - I(Z_{t,m}\leq x\pm\eta_n)\}|$$
$$\Lambda_2 = \sup_{x\in R}|n^{-1/2}\sum_{t=1}^{n}\{F(x\pm\eta_n+b_{n,t}) - F(x\pm\eta_n)\}|$$
$$\Lambda_3 = \sup_{x\in R}|n^{-1/2}\sum_{t=1}^{n}[I(Z_{t,m}\leq x\pm\eta_n+b_{n,t}) - F(x\pm\eta_n+b_{n,t}) - F(x\pm\eta_n) - I(Z_{t,m}\leq x\pm\eta_n)]|.$$

Recall from (7) that η_n defined in (6) is $O_p(n^{-3/2+\kappa})$. By this and the mean value theorem, we have that $\Lambda_2 = o_p(1)$. To prove $\Lambda_1 = o_p(1)$, it suffices to show that

$$A_1 := \sup_{x\in R}|n^{-1/2}\sum_{t=1}^{n}[I(Z_t\leq x+n^{-3/2}) - I(Z_t\leq x)]| = o_p(1).$$

For $x\in(x_r, x_{r+1}]$, $I(Z_t\leq x+n^{-3/2}) - I(Z_t\leq x)$ is no more than

$$I(Z_t\leq x_{r+1}+n^{-3/2}) - F(x_{r+1}+n^{-3/2}) + F(x_{r+1}) - I(Z_t\leq x_{r+1})$$
$$+F(x_{r+1}+n^{-3/2}) - F(x_{r+1}) + I(Z_t\leq x_{r+1}) - I(Z_t\leq x_r),$$

and a similar argument applies to a lower bound as well. Thus, we have that for sufficiently large n,

$$A_1 \le \sup_r |n^{-1/2}\sum_{t=1}^{n}(I(Z_t \le x_r + n^{-3/2}) - F(x_r + n^{-3/2}) + F(x_r) - I(Z_t \le x_r))|$$

$$+ \sup_r |(n^{-1/2})\sum_{t=1}^{n}(I(Z_t \le x_r) - F(x_{r+1}) + F(x_r) - I(Z_t \le x_r))| + o_p(1)$$

$$\le 2 \sup_r \sup_{x_r \le x < y \le x_{r+1}} |U_{n(x,y)}| + o_p(1)$$

because $x_r + n^{-3/2} \le x_{r+1}$ for sufficiently large n. Since by Lemma 2, for any $\lambda > 0$,

$$P(\sup_{r,\ x_r \le x < y \le x_{r+1}} |U_n(x,y)| \ge \lambda) \le C_2 N_n n^{\nu} \exp(-K\lambda n^{1/2-\nu}) + (1+N_n)C_2 n^{1-\rho} \to 0, \tag{8}$$

as $n \to \infty$, we have $A_1 = o_p(1)$. This proves $\Lambda_1 = o_p(1)$.

Clearly, $\Lambda_3 = o_p(1)$ if

$$\tilde{\Lambda}_3 = \sup_{x \in R} |n^{-1/2}\sum_{t=1}^{n} I(Z_{t,m} \le x + b_{n,t}) - F(x + b_{n,t}) + F(x) - I(Z_{t,m} \le x)| = o_p(1). \tag{9}$$

The first step in establishing (9) is to observe that

$$\tilde{\Lambda}_3 \le \Lambda_{31} + \Lambda_{32} + \Lambda_{33} + \Lambda_{34},$$

where

$$\Lambda_{31} = \sup_r |n^{-1/2}\sum_{t=1}^{n}\{I(Z_{t,m} \le x + b_{n,t}) - F_m(x_r + b_{n,t}) + F_m(x_r) - I(Z_{t,m} \le x_r)\}|,$$

$$\Lambda_{32} = \sup_r |n^{-1/2}\sum_{t=1}^{n}\{F(x_{r+1} + b_{n,t}) - F(x_r + b_{n,t})\}|,$$

$$\Lambda_{33} = \sup_r |n^{-1/2}\sum_{t=1}^{n}\{I(Z_{t,m} \le x_{r+1}) - F_m(x_{r+1}) + F_m(x_r) - I(Z_{t,m} \le x_r)\}|,$$

$$\Lambda_{34} = 4\sqrt{n} \sup_{x \in R} |F(x) - F_m(x)|.$$

It is clear that $\Lambda_{32} = o_p(1)$, and $\Lambda_{34} = o_p(1)$ due to Lemma 1. Since by Lemma 1,

$$n^{1/2}[F_m(x_{x+1}) - F_m(x_r)] = n^{1/2}O(n^{-\alpha}) + n^{1/2}N_n^{-1} \to 0 \quad \text{as} \quad n \to \infty,$$

it follows from (2.13) of (CR) that

$$P(|\Lambda_{33}| \ge \lambda) \le (N_n + 1)n^{1-\nu} C \exp(-K\lambda n^{1/2-\nu}),$$

for sufficiently large n. Hence $\Lambda_{33} = o_p(1)$.

Now it remains to prove $\Lambda_{31} = o_p(1)$. Toward this end, let

$$\xi_{n,t} = X_t / \sqrt{n}.$$

Since by assumption $n^{-1/2}\max_{1 \le t \le n}|X_t| = o_p(1)$ and $\hat{\beta}_n - \beta = O_p(n^{-1/2})$, $\Lambda_{31} = o_p(1)$ if for any $B > 0$,

$$\begin{aligned} A_2 &:= \sup_{r,|s|\le B} |n^{-1/2}\sum_{t=1}^{n}[I(Z_{t,m}\le x_r+s\xi_{n,t})-F_m(x_r+s\xi_{n,t})+F_m(x_r)-I(Z_{t,m}\le x_r)] \\ &= o_p(1). \end{aligned} \tag{10}$$

Set

$$s_i = -B+2Bi/n, \quad i=0,1,\cdots,n,$$

$$\tau^+ = \sup_{s_i\le s\le s_{i+1}} s\xi_{n,t} \text{ and } \tau^- = \inf_{s_i\le s\le s_{i+1}} s\xi_{n,t}.$$

Define

$$d_t^{\pm} = I(Z_{t,m}\le x_r+\tau_{it}^{\pm})-F_m(x_r+\tau_{it}^{\pm})+F_m(x_r)-I(Z_{t,m}\le x_r).$$

Using Taylor's series expansion, we see that A_2 is bounded by

$$A_{31}^{*} = \sup_{r,i} |n^{-1/2}\sum_{t=1}^{n} d_t^{\pm}| + o_p(1). \tag{11}$$

Write $n=mu+v$. For simplicity, assume $v=0$. Note that

$$\max_{1\le k\le m} \sum_{l=0}^{u-1}|X_{ml+k}| = O_p(n^{1-\nu}), \tag{12}$$

because by assumption on X_t

$$\begin{aligned} E(\max_{1\le k\le m} \sum_{l=0}^{u-1}(|X_{ml+k}|-E|X_1|))^2 &\le \sum_{k=1}^{m} E\Big(\sum_{l=0}^{u-1}(|X_{ml+k}|-E|X_1|)\Big)^2 \\ &\le 2\sum_{k=1}^{m} u\cdot \sum_{h=0}^{u-1}|Cov(W_k, W_{k+mh})| = O(n). \end{aligned}$$

Therefore, in view of (11) and (12), (10) will follow as a consequence of the following claim: for any $D>0$,

$$A_3^{\pm} := P\Big(|n^{-1/2}\sum_{t=0}^{n} d^{\pm}|\ge\lambda, \ \max_{1\le k\le m}\sum_{l=0}^{u-1}|X_{ml+k}|\le Dn^{1-\nu}\Big) = O(n^{\nu}e^{-\zeta n^{\omega}}),$$

for some $\zeta>0$ and $\omega>0$. We consider only A_3^+, since the argument for A_3^- is similar. First note that

$$\begin{aligned} A_3^+ &\le P\Big(\sum_{k=1}^{m}|\sum_{l=0}^{u-1} d^+_{ml+k}|\ge\lambda\sqrt{u/m}\cdot m, \ \max_{1\le k\le m}\sum_{l=0}^{u-1}|X_{ml+k}|\le Dn^{1-\nu}\Big) \\ &\le \sum_{k=1}^{m} P\Big(|\sum_{l=0}^{u-1} d^{+}_{ml+k}|\ge\lambda\sqrt{u/m}, \ \sum_{l=0}^{u-1}|X_{ml+k}|\le Dn^{1-\nu}\Big) \\ &= \sum_{k=1}^{m}\int I(\sum_{l=0}^{u-1}|x_{ml+k}|\le n^{1-\nu}D)\times P\Big(|\sum_{l=0}^{u-1} d^{+}_{ml+k}|\ge\lambda\sqrt{u/m}\,|X_k=x_k,\cdots,X_{m(u-1)+k}= \\ &\qquad x_{m(u-1)+k}\Big)\times dF_{X_k,\cdots,X_{m(u-1)+k}}(x_k,\cdots,x_{m(u-1)+k}). \end{aligned}$$

Given $x_k,\cdots,x_{m(u-1)+k}$ with $\sum_{l=0}^{u-1}|x_{ml+k}|\le n^{1-\nu}D$, d^+_{ml+k}, $l=0,\cdots,u-1$, are independent, the conditional mean of d^+_{ml+k} is zero and the sum of the conditional variances is bounded by

$$\sum_{l=0}^{u-1}\{F_m(x_r+\tau^+_{i,ml+k})-F_m(x_r)\} \le uO(n^{-a})+\sum_{l=0}^{u-1}\{F(x_r+\tau^+_{i,ml+k})-F(x_r)\}$$
$$= O(n^{1-a-\nu}+n^{1/2-\nu}),$$

where we have used Lemma 1 and the mean value theorem. Now using Bernstein's inequality [cf. Pollard (1984), page 193], we see that

$$A_3^+ \le \sum_{k=1}^{m} 2\exp\left(\frac{-\lambda^2(u/m)}{O(n^{1-a-\nu}+n^{1/2-\nu})+\lambda\sqrt{u/m}/3}\right) = O(n^{\nu}e^{-\zeta n^{\omega}}),$$

for some $\zeta>0$ and $\omega>0$. Hence $\Lambda_{31}=o_p(1)$, and the proof of the theorem is now complete.

References

[1] Boldin, M. V. (1982). Estimation of the distribution of noise in an autoregression scheme. *Theory Probab. Appl.* 27, 866-871.

[2] Chanda, K. C. and Ruymgaart, F. H. (1990). General linear processes: a property of the empirical process applied to density and mode estimation. *J. Time Ser. Anal.* 11, 185-199.

[3] Koul, H. L. (1984). Tests of goodness-of-fit in linear regression. *Colloq. Math. Soc. Janos Bolyai* 45, 279-315.

[4] Koul, H. L. (1992). *Weighted Empirical and Linear Models.* IMS Lecture notes monograph series 21, Hayward, CA.

[5] Koul, H. L. and Surgailis, D. (1997). Asymptotic expansion of M-estimators with long memory errors. *Ann. Statist.* 25 818-850

[6] Lee, S. and Wei, C. Z. (1999). On residual empirical processes of stochastic regression models with applications to time series. *Ann. Statist.* 27, 237-261.

[7] Pollard, D. (1988). *Convergence of Stochastic Processes,* Springer-Verlag, New York.

[8] Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with applications in Statistics.* Wiley, New York.