Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

S.A. van de Geer

Asymptotic normality of minimum L_1-norm estimators in linear regression

Department of Mathematical Statistics Report MS-R8806 June

The Centre for Mathematics and Computer Science is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

Asymptotic Normality of Minimum L_1-Norm Estimators in Linear Regression

Sara van de Geer
School of Mathematics
University of Bristol
University Walk
Bristol BS8 1TW

It is shown that the minimum L_1-norm estimator in linear regression is asymptotically normal with covariance matrix proportional to the inverse of the covariance matrix of the regressors. The method of proof makes use of asymptotic equicontinuity of VC-graph classes of functions.

1980 Classification Numbers: 62F12, 62J05.
Keywords & Phrases: asymptotic normality, empirical process, linear regression, minimum L_1-norm estimator, Vapnik-Chervonenkis class.

1. Introduction and main theorem

Consider the regression

$$y_k = \theta_0^T x_k + e_k,\ k = 1, \cdots, n, \cdots,$$

where $\{e_k\}$ is a sequence of i.i.d. random variables with median zero and density h, $\{x_k\}$ is an i.i.d. sequence of random vectors in $\mathbb{R}^d$, independent of $\{e_k\}$, and with $\mathbb{E}(x_1 x_1^T) = \Sigma, 0 < \Sigma < \infty$, and where $\theta_0 \in \mathbb{R}^d$ is unknown. The minimum L_1-norm estimator $\hat{\theta}_n$ is defined by

$$\sum_{k=1}^{n} |y_k - \hat{\theta}_n^T x_k| = \min_{\theta} \sum_{k=1}^{n} |y_k - \theta^T x_k|.$$

In van de Geer (1988), conditions for the $\mathcal{O}_p(n^{-1/2})$-rate of convergence for $\hat{\theta}_n$ are presented. In fact, in that paper the regressors and disturbances are not required to be i.i.d. sequences. The i.i.d. assumption is also not crucial here, but we maintain it to simplify the exposition.

Now, given the recent theory on empirical processes, asymptotic normality of $\hat{\theta}_n$ can be proved relatively easily, once the $\mathcal{O}_p(n^{-1/2})$-rate is already established. For this purpose, we introduce the process

$$l_n(\tau) = \sum_{k=1}^{n} |y_k - (\theta_0 + n^{-1/2}\tau)^T x_k| - \sum_{k=1}^{n} |y_k - \theta_0^T x_k|,\ \tau \in \mathbb{R}^d.$$

We shall show in Lemma 3 that under a regularity condition on h

$$l_n(\tau) = 2\tau^T W_n + \tau^T \Sigma \tau h(0) + o_P(1), \tag{1}$$

uniformly for all $|\tau| \leq K$. Here, W_n is a random vector that does not depend on τ, and that converges in law to a normal distribution with mean zero and covariance matrix $\Sigma/4$:

$$W_n \xrightarrow{\mathcal{L}} N(0, \Sigma/4).$$

From (1), one sees that the minimum of $\{l_n(\tau): |\tau| \leq K\}$ is attained at

$$\hat{\tau}_n = \Sigma^{-1} W_n / h(0) + o_P(1),$$

Report MS-R8806
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

which converges in distribution to a $N(0,\Sigma^{-1}/[4h(0)^2])$ law. The consequence will be asymptotic normality for $\hat{\theta}_n$.

THEOREM 1. *Suppose that for some* $\eta>0$, $M<\infty$

$$\eta \leqslant h(\epsilon) \leqslant M \text{ for all } -\eta \leqslant \epsilon \leqslant \eta. \tag{2}$$

Assume that

$$\max_{1\leqslant k\leqslant n} |x_k| = o_P(n^{1/2}). \tag{3}$$

Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N(0,\Sigma^{-1}/[4h(0)^2]).$$

This theorem follows from (1), since conditions (2) and (3) ensure that $|\hat{\theta}_n - \theta_0| = \mathcal{O}_p(n^{-1/2})$ (see VAN DE GEER (1988)). Equation (1) in turn is proved using empirical process theory - see Lemma 3.

REMARK. In POLLARD (1985) an alternative but closely related method for proving asymptotic normality in general cases, is presented. One of his examples is estimation of the median.

2. RESULTS FROM EMPIRICAL PROCESS THEORY

Let G be the distribution of x_1 and H the distribution of e_1. Define $P=G\times H$. Let $\mathcal{F}$ be a class of functions on $\mathbb{R}^{d+1}$ with

$$\|f\|^2 = \int |f|^2 \, dP < \infty, \; f \in \mathcal{F}. \tag{4}$$

Denote the envelope of $\mathcal{F}$ by

$$F = \sup_{f\in\mathcal{F}} |f|.$$

DEFINITION. The *graph* of a function $f:\mathbb{R}^{d+1}\rightarrow\mathbb{R}$ is defined as the set

$$A = \{(z,t): t \leqslant f(z) \leqslant 0 \text{ or } 0 \leqslant f(z) \leqslant t\}.$$

A class $\mathcal{Q}$ of sets $A\subset\mathbb{R}^{d+2}$ is called a *Vapnik-Chervonenkis class* if for some r and some N_0, and for all $N \geqslant N_0$, any set S of N points in $\mathbb{R}^{d+2}$ is such that there are at most N^r distinct sets of the form $S\cap A$ with $A\in\mathcal{Q}$.

A class $\mathcal{F}$ of functions is called a *VC-graph class* if the graphs of the functions in $\mathcal{F}$ form a Vapnik-Chervonenkis class.

We refer to DUDLEY (1984, Ch. 9) and POLLARD (1984, Ch. 2) thorough treatment of Vapnik-Chervonenkis and VC-graph classes. All we need here is their result that each class of sets that are expressible as finite unions of finite intersections of halfspaces, is a Vapnik-Chervonenkis class. E.g.

$$\mathcal{F} = \{f(z) = \gamma^T z 1_{\{\alpha^T z \geqslant \beta\}}(z): \gamma,\alpha\in\mathbb{R}^{d+1}, \beta\in\mathbb{R}\}$$

is a VC-graph class, since its graphs are of the form

$$(\{t\leqslant\gamma^T z\leqslant 0\}\cap\{\alpha^T z\geqslant\beta\})\cup(\{0\leqslant\gamma^T z\leqslant t\}\cap\{\alpha^T z\geqslant\beta\})\cup(\{\alpha^T z<\beta\}\cap\{t=0\})$$

i.e. the union of 3 sets each of which is the intersection of 3 halfspaces.

Now, let P_n be the empirical measure based on $(x_1,e_1), \cdots, (x_n,e_n)$. Define

$$\nu_n(f) = \sqrt{n} \int f d(P_n - P), \; f\in\mathcal{F}.$$

We shall consider ν_n as stochastic process on $\mathcal{F}$, and therefore we need to impose measurability conditions. To avoid digressions, we leave these unspecified and assume thoughout the paper that $\mathcal{F}$ is *permissible* in the sense of POLLARD (1984).

THEOREM 2. *Suppose that* $\mathcal{F}$ *is a VC-graph class with envelope F satisfying* $\|F\|<\infty$. *Then the asymptotic equicontinuity criterion holds: for all* $\delta>0$ *there is an* n_δ *such that*

$$\text{Prob}\Big(\sup_{\substack{f,\tilde{f}\in\mathcal{F}\\ \|f-\tilde{f}\|<\delta}} |\nu_n(f)-\nu_n(\tilde{f})| >\delta\Big)<\delta \tag{5}$$

for all $n\geq n_\delta$.

PROOF. This is a special case of a more general result for *Donsker* classes. See e.g. POLLARD (1982, 1984) □

We shall use Theorem 2 in e.g. the following form: if $\mathcal{F}$ is a VC-graph class with envelope F satisfying $\|F\|<\infty$, and if $\{\mathcal{F}_n\}$ is a sequence of subclasses with $\|f\|=o(1)$ uniformly in $f\in\mathcal{F}_n$, then $\nu_n(f)=o_P(1)$ uniformly in $f\in\mathcal{F}_n$. Here, we notify the abuse of notation that the subscript "P" in the stochastic order symbol refers to the probability measure for the whole sequence $(x_1,e_1),(x_2,e_2),\cdots$.

3. PROOF OF EQUALITY (1)

LEMMA 3. *Suppose that for some* $\eta>0$, $M<\infty$, $h(\epsilon)\leq M$ *for all* $\epsilon\in[-\eta,\eta]$, *then*

$$l_n(\tau) = 2\tau^T W_n + \tau^T\Sigma\tau h(0) + o_P(1),$$

uniformly for all $|\tau|\leq K$, *where* W_n *does not depend on* τ *and*

$$W_n \xrightarrow{\mathcal{L}} N(0,\Sigma/4).$$

PROOF. We have

$$\begin{aligned} l_n(\tau) &= \sum_{k=1}^n |e_k - n^{-1/2}\tau^T x_k| - \sum_{k=1}^n |e_k| \\ &= \frac{1}{\sqrt{n}}\sum_{e_k\leq n^{-1/2}\tau^T x_k} \tau^T x_k - \frac{1}{\sqrt{n}}\sum_{e_k> n^{-1/2}\tau^T x_k} \tau^T x_k \\ &\quad -2\sum_{\substack{\tau^T x_k>0\\ 0<e_k\leq n^{-1/2}\tau^T x_k}} e_k + 2\sum_{\substack{\tau^T x_k<0\\ n^{-1/2}\tau^T x_k<e_k\leq 0}} e_k. \end{aligned} \tag{6}$$

Consider the first term of this expression. Define for $\xi\in\mathbb{R}^d$, $\epsilon\in\mathbb{R}$

$$f_{\tau,\alpha}(\xi,\epsilon) = \tau^T\xi\, 1_{\{\epsilon\leq\alpha^T\xi\}}(\xi,\epsilon),\quad \tau\in\mathbb{R}^d,\ \alpha\in\mathbb{R}^d.$$

Let $\mathcal{F}=\{f_{\tau,\alpha}: |\tau|\leq K,\ \alpha\in\mathbb{R}^d\}$. Clearly, $\mathcal{F}$ is a VC-graph class, and its envelope F satisfies

$$\|F\|^2 \leq \max_{|\tau|\leq K} \tau^T\Sigma\tau<\infty.$$

Moreover, $\|f_{\tau,n^{-1/2}\tau} - f_{\tau,0}\|=o(1)$, uniformly in $|\tau|\leq K$. Therefore, by the asymptotic equicontinuity criterion (5),

$$\nu_n(f_{\tau,n^{-1/2}\tau}) = \nu_n(f_{\tau,0}) + o_P(1),$$

uniformly in $|\tau|\leq K$. Thus, we can write the first term of (6) as

$$\frac{1}{\sqrt{n}}\sum_{e_k\leq n^{-1/2}\tau^T x_k} \tau^T x_k = \nu_n(f_{\tau,0}) + \sqrt{n}\int f_{\tau,n^{-1/2}\tau}\, dP + o_P(1). \tag{7}$$

Define

$$W_{1,n} = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} x_k[1_{\{e_k \leq 0\}}(e_k) - \frac{1}{2}].$$

Then $\nu_n(f_{\tau,0}) = \tau^T W_{1,n}$ since e_1 has median zero, and (7) becomes

$$\frac{1}{\sqrt{n}} \sum_{e_k \leq n^{-1/2}\tau^T x_k} \tau^T x_k = \tau^T W_{1,n} + \sqrt{n} \int f_{\tau,n^{-1/2}\tau}\, dP + o_P(1). \tag{8}$$

Similarly, for the second expression in (6) we find

$$\frac{1}{\sqrt{n}} \sum_{e_k > n^{-1/2}\tau^T x_k} \tau^T x_k = \tau^T W_{2,n} + \sqrt{n} \int \bar{f}_{\tau,n^{-1/2}\tau}\, dP + o_P(1), \tag{9}$$

uniformly in $|\tau| \leq K$, where

$$W_{2,n} = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} x_k[1_{\{e_k > 0\}}(e_k) - \frac{1}{2}],$$

and where $\bar{f}_{\tau,n^{-1/2}\tau}$ is defined as

$$\bar{f}_{\tau,n^{-1/2}\tau}(\xi,\epsilon) = \tau^T \xi 1_{\{\epsilon > n^{-1/2}\tau^T\xi\}}(\xi,\epsilon)$$

Let $H(\beta) = \int 1_{(-\infty,\beta]} dH$, $\beta \in \mathbb{R}$. For $\xi \in \mathbb{R}^d$,

$$\frac{H(n^{-1/2}\tau^T\xi) - H(0)}{n^{-1/2}\tau^T\xi} = h(0) + o(1)$$

uniformly in $|\tau| \leq K$. By assumption, $h(\epsilon) \leq M$ for all $\epsilon \in [-\eta,\eta]$. Dominated convergence now yields that

$$\sqrt{n} \int (f_{\tau,n^{-1/2}\tau} - f_{\tau,0}) dP = \tau^T \Sigma \tau h(0) + o_P(1) \tag{10}$$

and similarly

$$\sqrt{n} \int (\bar{f}_{\tau,n^{-1/2}\tau} - \bar{f}_{\tau,0})\, dP = -\tau^T \Sigma \tau h(0) + o(1), \tag{11}$$

uniformly in $|\tau| \leq K$. Furthermore, since e_1 has median zero

$$\int f_{\tau,0}\, dP = \int \bar{f}_{\tau,0}\, dP. \tag{12}$$

Combining (8), (9), (10), (11) and (12), we see that the difference of the first two terms of (6) can be written as

$$\frac{1}{\sqrt{n}} \sum_{e_k \leq n^{-1/2}\tau^T x_k} \tau^T x_k - \frac{1}{\sqrt{n}} \sum_{e_k > n^{-1/2}\tau^T x_k} \tau^T x_k = \tau^T(W_{1,n} - W_{2,n}) + 2\tau^T \Sigma \tau h(0) + o_P(1) \tag{13}$$

Consider now the third term in (6). Let

$$f_{\tau,\alpha,\beta}(\xi,\epsilon) = \beta\epsilon 1_{\{\tau^T\xi > 0\}}(\xi)\, 1_{\{0 < \epsilon \leq \alpha^T\xi\}}(\xi,\epsilon),$$

and define $\mathcal{F}^* = \{f_{\tau,\alpha,\beta}\colon \tau \in \mathbb{R}^d, \beta \in \mathbb{R}, |\alpha| \leq K/\beta\}$. Also $\mathcal{F}^*$ is a VC-graph class. Its envelope F^* satisfies

$$\|F^*\|^2 \leq \max_{|\alpha| \leq K/\beta} \beta^2 \alpha^T \Sigma \alpha < \infty.$$

But clearly, $\|f_{\tau,n^{-1/2}\tau,n^{1/2}}\| = o(1)$ uniformly in $|\tau| \leq K$, so that $\nu_n(f_{\tau,n^{-1/2}\tau,n^{1/2}}) = o_P(1)$ by the asymptotic equicontinuity criterion. Thus

$$2 \sum_{\substack{\tau^T x_k > 0 \\ 0 < e_k \leq n^{-1/2}\tau^T x_k}} e_k = 2\nu_n(f_{\tau,n^{-1/2}\tau,n^{1/2}}) + 2\sqrt{n} \int f_{\tau,n^{-1/2}\tau,n^{1/2}}\, dP \tag{14}$$

$$= 2\sqrt{n}\int f_{\tau,n^{-1/2}\tau,n^{1/2}}\, dP + o_P(1),$$

uniformly in $|\tau| \leqslant K$. Similar arguments apply to the last term in (6):

$$\sum_{\substack{\tau^T x_k<0 \\ n^{-1/2}\tau^T x_k<e_k\leqslant 0}} e_k = 2\sqrt{n}\int \bar{f}_{\tau,n^{-1/2}\tau,n^{1/2}}\, dP + o_P(1), \tag{15}$$

where

$$\bar{f}_{\tau,n^{-1/2}\tau,n^{1/2}}(\xi,\epsilon) = n^{1/2}\epsilon 1_{\{\tau^T\xi<0\}}(\xi)\, 1_{\{n^{-1/2}\tau^T\xi<\epsilon\leqslant 0\}}(\xi,\epsilon).$$

Again, by dominated convergence

$$2\sqrt{n}\int [f_{\tau,n^{-1/2}\tau,n^{1/2}} - \bar{f}_{\tau,n^{-1/2}\tau,n^{1/2}}]\, dP = \tau^T\Sigma\tau h(0) + o(1), \tag{16}$$

uniformly in $|\tau| \leqslant K$.

Hence, from (14), (15) and (16)

$$2\sum_{\substack{\tau^T x_k>0 \\ 0<e_k\leqslant n^{-1/2}\tau^T x_k}} e_k - 2\sum_{\substack{\tau^T x_k<0 \\ n^{-1/2}\tau^T x_k<e_k\leqslant 0}} e_k = \tau^T\Sigma\tau h(0) + o_P(1). \tag{17}$$

Insert (13) and (17) into (6) to obtain that

$$l_n(\tau) = \tau^T(W_{1,n} - W_{2,n}) + \tau^T\Sigma\tau h(0) + o_P(1)$$

uniformly in $|\tau| \leqslant K$. The proof is completed by writing

$$W_n = \frac{1}{2}(W_{1,n} - W_{2,n}),$$

and noting that the central limit theorem ensures that $W_n \xrightarrow{\mathcal{L}} N(0,\Sigma/4)$. □

Obviously, given that $|\hat{\theta}_n - \theta_0| = \mathcal{O}_P(n^{-1/2})$, Lemma 3 implies the asymptotic distribution as presented in Theorem 1. Moreover, it follows that

$$-2h(0)[\sum_{k=1}^{n} |y_k - \hat{\theta}_n^T x_k| - \sum_{k=1}^{n} |y_k - \theta_0^T x_k|]$$

has asymptotically a chi-square distribution with d degrees of freedom.

4. References

[1] R.M. Dudley (1984). A course on empirical processes. *Springer Lecture Notes in Math.* (Lectures given at Ecole d'Eté de Probabilités de St. Flour, 1982), 1-142.

[2] D. Pollard (1982). A central limit for empirical processes. *J. Austr. Math. Soc. (Series A) 33*, 235-248.

[3] D. Pollard (1984). *Convergence of Stochastic Processes*, Springer Series in Statistics, Springer-Verlag, New York.

[4] D. Pollard (1985). New ways to prove central limit theorems. *Economic Theory 1*, 295-314.

[5] S. van de Geer (1988). Estimating a regression function. *Report* Centre for Mathematics and Computer Science, Amsterdam.