A paradox in least squares estimation of linear regression models

Z. D. Bai

Dept. of Applied Math., National Sun Yat-sen Univ., Kaohsiung, Taiwan and

Dept. of Stat., National Univ. of Singapore, Singapore

Meihui Guo[1]

Dept. of Applied Math., National Sun Yat-sen Univ., Kaohsiung, Taiwan

Abstract: This note considers a paradox arising in the least squares estimation of linear regression models in which the error terms are assumed to be iid and possess finite r-th moment, for $r \in [1, 2)$. We give a concrete example to show that the least squares estimator of the slope parameter is inconsistent when the intercept parameter of the model is given. However, surprisingly this estimator is consistent when the intercept parameter is intendedly assumed to be unknown and reestimated simultaneously with the slope parameter.

Keywords: consistency; least squares estimate; r-th moment.

[1] Supported in part by the grant NSC 88-2118-M-110-003 from the National Science

Council of Taiwan.

1 Introduction

Consider the multiple regression model:

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, 2, \cdots$$

where x_{ij} are given explanatory variables, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$ is the vector of unknown regression coefficients and ϵ_i's are iid random errors. The least squares estimate (LSE) $\mathbf{b}_n = (b_{n1}, \cdots, b_{np})'$ of the vector $\boldsymbol{\beta}$ based on the design matrix $X_n = (x_{ij})_{1 \le i \le n, 1 \le j \le p}$ and the response vector $\mathbf{y}_n = (y_1, \cdots, y_n)'$ is given by $\mathbf{b}_n = (X_n' X_n)^{-1} X_n' \mathbf{y}_n$. When $X_n' X_n$ is nonsingular and $Var(\epsilon_i) < \infty$, $\mathbf{b}_n$ is the best linear unbiased estimator of $\boldsymbol{\beta}$ (Gauss-Markov theorem), and is a consistent estimator of $\boldsymbol{\beta}$ if and only if $\lambda_{\mathbf{min}}(X_n' X_n)$ (the smallest positive eigenvalue of $X_n' X_n$) converges to infinity as n tends to infinity (Drygas, 1976). There are also a lot of research works done on the strong consistency of $\mathbf{b}_n$, references are made to, among others, Lai, Robbins and Wei (1979) and Chen, Lai and Wei (1981, 1983). Since the LSE possesses good properties for both small and large sample theory when the variance of the ϵ_i is finite, it is the most commonly used estimator of $\boldsymbol{\beta}$ in the literature. When the error have finite r-th moment only for some $r < 2$, some properties of the LSE can be found in Bai (1982).

If some of the regression parameters are known, say $(\beta_{q+1}, \cdots, \beta_p)$, then one can reset the above linear model as

$$z_n = y_n - \beta_{q+1} x_{n,q+1} - \cdots - \beta_p x_{np} = \beta_1 x_{n1} + \cdots + \beta_q x_{nq} + \epsilon_n.$$

This new model has a smaller number of parameters. In such case, there are two ways to estimate the parameters $(\beta_1, \cdots, \beta_q)$ by using the two models respectively. Which model will give a better estimator? It is a common belief that a model of a smaller number of parameters should provide better estimators (more efficient or of smaller MSE). In most cases, this belief has been confirmed. However, Chen and An

(1997) raised a challenge to this belief. They constructed a linear regression model as follows:

$$y_i = \beta x_i + \epsilon_i, \quad 1 \le i \le n, \tag{1}$$

for which the LSE $\hat{\beta}_n$ of β was shown to be inconsistent, where the errors have finite absolute first moment. On the other hand, it was shown surprisingly that the LSE $\tilde{\beta}_n$ of β is weakly consistent when intendedly adding an intercept term to the model (1), i.e.,

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad 1 \le i \le n. \tag{2}$$

It should be noted that gaining the consistency of the LSE $\tilde{\beta}_n$ of β does not cause any serious trade off to the estimation of α. That means, both the estimators of α and β are consistent (i.e., $\tilde{\alpha}_n \to 0$, in prob.). This paradox strongly challenges the unjustified belief and shows that reduction of parameter dimensions does not always benefit the statistical inferences.

In Chen and An (1997), the errors are constructed to have the finite first moment only, that is, for any $r > 1$, their r-th moment is infinite. It is natural to ask, what will happen when the errors have finite r-th moment, for $1 < r < 2$. Does there exist a number $r_0 \in [1, 2)$ to ensure that the consistency of the LSE for the over-parameterized model implies the consistency of the LSE for the true model, provided the error have finite $r > r_0$ moment? To gain a deeper insight into this problem, we continue the study on the paradox for the LSE in linear regression models. Through a universal example, we show that for any $1 \le r < 2$, there is a linear regression model (1), for which ϵ_i has finite absolute r-th moment, the LSE is inconsistent whereas the LSE under over-parameterized model (2) is consistent. Furthermore, the construction is simpler than those given in Chen and An (1997).

It is well known that the LSE of the regression coefficient β for the model (2) is

$$\tilde{\beta}_n = \sum_{i=1}^{n}(x_i - \bar{x}_n)y_i / \sum_{i=1}^{n}(x_i - \bar{x}_n)^2, \text{where } \ \bar{x}_n = \sum_{i=1}^{n} x_i/n,$$

while that for the model (1) is

$$\hat{\beta}_n = \sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} x_i^2.$$

Intuitively, estimating α would lead to lower efficiency. Thus, we expect $\hat{\beta}_n$ to be a better estimator. However, this is not always true when $Var(\epsilon_i)$ is infinite. In fact, this is the case when ϵ_i's are iid with finite non-zero variance, i.e., if $\tilde{\beta}_n$ is consistent, so must be $\hat{\beta}_n$. And, based on the fact that $\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$ may have a much lower increasing rate than $\sum_{i=1}^{n} x_i^2$, it is easy to construct an example showing that the converse is not true. However, the superiority of $\hat{\beta}_n$ to $\tilde{\beta}_n$ may no longer hold when $Var(\epsilon_i)$ does not exist.

This abnormal behavior of the LSE when the ϵ_i's possess finite r-th moment, where $1 \le r < 2$ is just the main point of the stusy of Chen and An (1997) and the present paper. In section 2, an example is constructed for which $E|\epsilon_i|^r < \infty$, $1 \le r < 2$ and the main proposition shows that $\tilde{\beta}_n \to \beta$ whereas $\limsup |\hat{\beta}_n - \beta| = \infty$. Some discussions and conclusions are given in Section 3.

2 Construction of the Example and Proofs of the Main Proposition

In this section, for the model (1), we define a design sequence $x_1, x_2, \cdots$ and a sequence of iid random errors $\{\epsilon_i\}$ so that $E(\epsilon_i) = 0$ and $E|\epsilon_i|^r < \infty$, for $1 \le r < 2$. We comment herewith that the distribution of the errors can be discrete, singularly continuous or absolutely continuous, although that we constructed is discrete. This can be seen from the fact that the results given below remain true if the error sequence is replaced by $\tilde{e}_i = \epsilon_i + \eta_i$, where η_i is absolutely or singularly continuously distributed and has finite second moment. The next step, we show that $\tilde{\beta}_n$ converges weakly to β but not $\hat{\beta}_n$. Substituting (1) into the definitions of $\hat{\beta}_n$ and $\tilde{\beta}_n$, we get

$$\hat{b}_n = \hat{\beta}_n - \beta = \sum_{i=1}^{n} x_i \varepsilon_i / \sum_{i=1}^{n} x_i^2, \qquad \tilde{b}_n = \tilde{\beta}_n - \beta = \sum_{i=1}^{n} (x_i - \bar{x}_n)\varepsilon_i / \sum_{i=1}^{n} (x_i - \bar{x}_n)^2,$$

where $\varepsilon, \varepsilon_1, \varepsilon_2, \cdots$ are iid with $E(\varepsilon) = 0$. Let $1 \leq r < 2$. We shall show the following proposition.

Proposition. *There exists a d.f. F of ε and a design sequence $x_1, x_2, \cdots$, such that*

$$\begin{aligned} E(\varepsilon) &= 0, \quad E|\varepsilon|^r < \infty, \\ \hat{b}_n &\not\to 0, \quad \text{in pr.} \\ \text{whereas} \quad \tilde{b}_n &\to 0, \quad \text{in pr..} \end{aligned}$$

Proof. We split the proof into several pieces.

(a) The construction of the distribution and some inequalities.

Select $0 < \lambda < 1$ and m (large) such that

$$m\lambda(2 - r) > 4, \quad \text{and} \quad m(1 - \lambda)(r - 1) < 1.$$

This is possible since $2 - r > 0$. For $k \geq 2$, set

$$p_k = P(\varepsilon = k^{mk}) = k^{-mrk-2}.$$

Since $\sum_{k=2}^{\infty} p_k \leq \sum_{k=2}^{\infty} k^{-4} < 1$, we have

$$p_1 = P(\varepsilon = -a) = 1 - \sum_{k=2}^{\infty} p_k > 0.$$

By choosing

$$a = p_1^{-1} \sum_{k=2}^{\infty} k^{mk} p_k = p_1^{-1} \sum_{k=2}^{\infty} k^{-mk(r-1)-2} < \infty,$$

we have

$$E(\varepsilon) = 0$$

$$E(|\varepsilon|^r) = p_1 a^r + \sum_{k=2} k^{-2} < \infty$$

and

$$nP(|\varepsilon| \geq n^{1/r}) \leq E(|\varepsilon|^r I(|\varepsilon| \geq n^{1/r})) \to 0. \tag{3}$$

Set $n_k = \left[k^{mrk}\right]$. Then, we have

$$k \sim \frac{\log n_k}{mr \log\log n_k} \sim \frac{\log n_{k+1}}{mr \log\log n_{k+1}}$$

and

$$\frac{n_{k+1}}{n_k} \sim (ek)^{mr} \sim \left(\frac{e \log n_k}{mr \log\log n_k}\right)^{mr}.$$

Here and in the sequel, $a_k \sim b_k$ means $a_k/b_k \to 1$. For $n_k < n \leq n_{k+1}$, we also have

$$\begin{aligned} E\varepsilon I(|\varepsilon| \geq n^{1/r}) &= \sum_{j=k+1}^{\infty} j^{-mj(r-1)-2} \\ &\sim (k+1)^{-m(k+1)(r-1)-2} \\ &\simeq n_{k+1}^{-(r-1)/r} \left(\frac{\log\log n_{k+1}}{\log n_{k+1}}\right)^2, \quad (1 < r < 2), \end{aligned} \tag{4}$$

$$E\varepsilon I(|\varepsilon| \geq n) = \sum_{j=k+1}^{\infty} j^{-2} \simeq (k+1)^{-1} \simeq \left(\frac{\log\log n_{k+1}}{\log n_{k+1}}\right), \quad (r = 1) \tag{4'}$$

and

$$\begin{aligned} E\varepsilon^2 I(|\varepsilon| < n^{1/r}) &= p_1 a^2 + \sum_{j=2}^{k} j^{mj(2-r)-2} \\ &\sim k^{mk(2-r)-2} \\ &\simeq n_k^{(2-r)/r} \left(\frac{\log\log n_k}{\log n_k}\right)^2, \quad (1 \leq r < 2). \end{aligned} \tag{5}$$

Here and in the sequel, $a_k \simeq b_k$ means $0 < c < a_k/b_k < C < \infty$ for some $0 < c < C$.

(b) The construction of the designs and related inequalities

Let

$$x_1 = x_2 = 1, \quad x_3 = -1$$

and for $i \geq 4$,

$$x_i = \begin{cases} \theta(i) i^{(1-r)/r} \left(\frac{\log\log i}{\log i}\right)^3, & \text{if for some } k,\ n_k^{1-\eta} n_{k+1}^{\eta} \leq i \leq n_k^{1-\lambda} n_{k+1}^{\lambda} \\ \theta(i) i^{(1-r)/r} \left(\frac{\log\log i}{\log i}\right)^{1/2}, & \text{otherwise} \end{cases}$$

where $0 < \eta < \lambda$ such that $m(1-\lambda)(r-1) < 1$ and $\theta(i) = -1$ for $i \equiv 0(\text{mod}(3))$ and 1 otherwise. Then we have

$$\sum_{i=1}^{n} x_i^2 >> \begin{cases} n_k^{(2-r)/r} \left(\frac{\log\log n_k}{\log n_k}\right)^{6-m\lambda(2-r)}, & \text{if } n_k^{1-\lambda} n_{k+1}^{\lambda} \leq n \leq n_{k+1} \\ n_k^{(2-r)/r} \left(\frac{\log\log n_k}{\log n_k}\right), & \text{if } n_k < n < n_k^{1-\lambda} n_{k+1}^{\lambda}. \end{cases} \tag{6}$$

Here, $a_k >> b_k$ means $a_k > cb_k$ for some $c > 0$ and all k. For the first case, applying Lemma A1 (see Appendix), we have

$$\begin{aligned} \sum_{i=1}^{n} x_i^2 &>> \sum_{n_k^{1-\eta} n_{k+1}^{\eta} \leq i \leq n_k^{1-\lambda} n_{k+1}^{\lambda}} i^{2(1-r)/r} \left(\frac{\log\log i}{\log i}\right)^6 \\ &>> \left(\frac{\log\log(n_k^{1-\lambda} n_{k+1}^{\lambda})}{\log(n_k^{1-\lambda} n_{k+1}^{\lambda})}\right)^6 \left[(n_k^{1-\lambda} n_{k+1}^{\lambda})^{(2-r)/r} - (n_k^{1-\eta} n_{k+1}^{\eta})^{(2-r)/r}\right] \\ &>> n_k^{(2-r)/r} \left(\frac{\log\log n_k}{\log n_k}\right)^6 \left(\frac{n_{k+1}}{n_k}\right)^{\lambda(2-r)/r} \\ &>> n_k^{(2-r)/r} \left(\frac{\log\log n_k}{\log n_k}\right)^{6-m\lambda(2-r)}. \end{aligned}$$

For the second case, we have

$$\begin{aligned} \sum_{i=1}^{n} x_i^2 &>> \sum_{n_{k-1}^{1-\lambda} n_k^{\lambda} \leq i \leq n_k} i^{(2-2r)/r} \left(\frac{\log\log i}{\log i}\right) \\ &>> \left(n_k^{(2-r)/r} - (n_{k-1}^{1-\lambda} n_k^{\lambda})^{(2-r)/r}\right) \left(\frac{\log\log n_k}{\log n_k}\right) \\ &= n_k^{(2-r)/r} \left(\frac{\log\log n_k}{\log n_k}\right)(1 + o(1)). \end{aligned}$$

¿From the definition of the designs, it is easy to see that

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} |x_i|} = 1/3.$$

This simply implies that

$$\limsup_{n\to\infty} \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}{\sum_{i=1}^{n} x_i^2} = 1 - \liminf_{n\to\infty} \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n \sum_{i=1}^{n} x_i^2}$$

$$= 1 - \liminf_{n\to\infty} \frac{\left(\sum_{i=1}^n |x_i|\right)^2}{9n\sum_{i=1}^n x_i^2} \geq 8/9. \tag{7}$$

When $n = [n_k^{1-\lambda} n_{k+1}^{\lambda} - 2]$, we have

$$\sum_{i=1}^n x_i \simeq n^{1/r} \left(\frac{\log\log n}{\log n}\right)^3$$

and

$$\sum_{i=1}^n x_i^2 \simeq n^{(2-r)/r} \left(\frac{\log\log n}{\log n}\right)^6.$$

Thus,

$$\frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2} \simeq n^{(r-1)/r} \left(\frac{\log\log n}{\log n}\right)^{-3} \simeq n_{k+1}^{(r-1)/r} \left(\frac{\log\log n_k}{\log n_k}\right)^{-3+m(1-\lambda)(r-1)}. \tag{8}$$

(c) The proof of the proposition

Truncation. Define $\varepsilon_{i,n} = \varepsilon_i I_{(|\varepsilon_i| < n^{1/r})}$,

$$\hat{b}_{nn} = \frac{\sum_{i=1}^n x_i \varepsilon_{i,n}}{\sum_{i=1}^n x_i^2} \quad \text{and} \quad \tilde{b}_{nn} = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)\varepsilon_{i,n}}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}.$$

By the fact $E(|\varepsilon|^r) < \infty$ and from (3), we have

$$P(\hat{b}_{nn} \neq \hat{b}_n) \leq nP(|\varepsilon| \geq n^{1/r}) \to 0$$

By the same reason, one can show that

$$P(\tilde{b}_{nn} \neq \tilde{b}_n) \to 0.$$

Convergence of the main part

Now, we proceed to show that

$$\hat{b}_{nn} - E(\hat{b}_{nn}) \to 0, \quad \text{and} \quad \tilde{b}_{nn} - E(\tilde{b}_{nn}) \to 0 \ \text{ in } pr. \tag{9}$$

For the first conclusion of (9), it is sufficient to show that

$$Var(\hat{b}_{nn}) \leq \frac{E(\varepsilon_{1,n}^2)}{\sum_{i=1}^n x_i^2} \to 0.$$

The above limit follows from (5) and (6). Then the first conclusion of (9) is proved.

Noticing the relation (7), the second conclusion of (9) can be proved in similar method.

The non- convergence of $\hat{b}_{nn}$ and the convergence of $\tilde{b}_{nn}$

The consistency of $\tilde{b}_{nn}$ follows from the second conclusion of (9) and the fact that $E(\tilde{b}_{nn}) = 0$. Thus, to complete the proof of the proposition, one needs only show that $E(\hat{b}_{nn}) \not\to 0$. In fact by (4), (4') and (8), we have

$$\begin{aligned} -E(\hat{b}_{nn}) &= \frac{-E(\varepsilon_{1,n})\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2} = \frac{E(\varepsilon I(|\varepsilon| \geq n^{1/r}))\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2} \\ &\simeq \begin{cases} \left(\frac{\log\log n_k}{\log n_k}\right)^{m(1-\lambda)(r-1)-1}, & \text{if } 1 < r < 2 \\ \left(\frac{\log\log n_k}{\log n_k}\right)^{-2}, & \text{if } r = 1. \end{cases} \\ &\to \infty. \end{aligned}$$

3 Discussion

For the multiple regression model $y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$, the LSE is obtained by minimizing the following sum of squares

$$SSE = \min_{\beta' s} \sum_{i=1}^n (y_i - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2.$$

When $Var(\epsilon_i) = \sigma^2 < \infty$, $\mathbf{y}_n = (y_1, \cdots, y_n)'$ is a member of an L_2 space. Thus, the above orthogonal projection will give a good approximation to the mean vector, which is a member of the p-dimensional sub-space spanned by the vectors $(x_{11}, \cdots, x_{n1})'$, $\cdots, (x_{1p}, \cdots, x_{np})'$. In such case, $SSE/(n-p)$ turns out to be a natural estimator of the variance σ^2 of ϵ_i's. Therefore, SSE can act as a good criterion to find a good estimator of the regression coefficient vector. Consequently, it can be proven that $\hat{\beta}_n$ is always better than $\tilde{\beta}_n$ in the sense of consistency or asymptotic normality.

However, when $Var(\epsilon_i) = \infty$, $\mathbf{y}_n$ is no longer a member of an L_2 space. In such cases, the criterion SSE turns out to be improper for the parameter estimation and the LSE would behave weirdly. Intuitively, when $E(|\varepsilon|^r) < \infty$, $\mathbf{y}_n$ becomes a member of an L_r space, and thus the minimization

$$SrE = \min_{\beta' s} \sum_{i=1}^{n} |y_i - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}|^r$$

should be a better criterion for the parameter estimation and provide a better estimator than LSE. Of course, the centralization condition (zero mean for LSE) should be changed as $E\epsilon \mathrm{sign}(\epsilon)$. If the LrE method is employed, we believe that the estimator under the true model is always better than that under the overparametrized model.

Secondly, it is seen from the proof of the proposition that $\hat{b}_{nn} - E(\hat{b}_{nn}) \to 0$ is always implied by $\tilde{b}_{nn} = \tilde{b}_{nn} - E(\tilde{b}_{nn}) \to 0$, due to the fact that $\sum_{i=1}^n x_i^2 >> \sum_{i=1}^n (x_i - \bar{x}_n)^2$. Conversely, the later is implied by the former if $\sum_{i=1}^n (x_i - \bar{x}_n)^2$ has the same order as $\sum_{i=1}^n x_i^2$ (as in our example). Intuitively, $\hat{b}_{nn}$ always has a "smaller" random part than $\tilde{b}_{nn}$. The inconsistency of $\hat{b}_n$ is just caused by the fact that $E(\hat{b}_{nn}) \not\to 0$.

Finally, it will also be of interest to find examples of linear models for which $\tilde{b}_n$ is strongly consistent while $\hat{b}_n$ is not, when $E(|\varepsilon|^r) < \infty$.

4 Appendix

The following lemma is trivial in calculus. We cite it for ease of reference.

Lemma A1 *Assume that $a \in (0,1)$, b and c are real numbers and that A and B are integers. Then, as $A > B \to \infty$, we have*

$$\sum_{i=B}^{A} i^{-a} \log^b i \log\log^c i = \frac{1 + o(1)}{1 - a} \left(A^{1-a} \log^b A \log\log^c A - B^{1-a} \log^b B \log\log^c B \right)$$

Proof. The lemma can proved by using the trivial inequalities

$$\int_B^A x^{-a} \log^b x \log\log^c x dx < \sum_{i=B}^{A} i^{-a} \log^b i \log\log^c i < \int_{B-1}^{A} x^{-a} \log^b x \log\log^c x dx$$

and the integration by parts

$$\int_B^A x^{-a} \log^b x \log\log^c x dx = \frac{1}{1-a}\Big(A^{1-a}\log^b A \log\log^c A - B^{1-a}\log^b B \log\log^c B\Big)$$
$$-\frac{1}{1-a}\int_B^A x^{-a}(b\log\log x + c)\log^{b-1} x \log\log^{c-1} x dx.$$

References

Bai, Z. D. (1982) r-mean convergence of LS estimates of regression coefficients in linear models. *Science Exploration*, Vol. 2, No. 2, 81-84.

Chen,X and An, H. (1997) Abnormal behavior of the least squares estimate of multiple regression, *Science in China, ser. A*, **40**, 234-242.

Chen, Guijing, Lai, T. L.and Wei, C. Z. (1981) Convergence systems and strong consistency of least squares estimates in regression models. *J. Multivariate Anal.* **11**. No. 3, 319–333.

Chen, Guijing, Lai, T. L. and Wei, C. Z. (1983) Convergence systems and strong consistency of least squares estimates in regression models. *Kexue Tongbao* (English Ed.) **28,** No. 1, 16–20.

Drygas, H.(1976), Weak and Strong consistency of the least squares estimators in regression models, *Z. Wahrsch. Verw. Gebiete* **34**, 119-127.

Lai, T.L., Robbins, H. and Wei, C.Z.(1979), Strong Consistency of least squares estimates in Multiple Regression II, *J. Multivariate Anal.* **9**, 343-361.