AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
http://ageconsearch.umn.edu
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

Discussion Paper

CHUNG-HUA INSTITUTION FOR ECONOMIC RESEARCH
75 Chang Hsing St., Taipei, Taiwan
Republic of China

Discussion Paper

CHUNG-HUA INSTITUTION FOR ECONOMIC RESEARCH
75 Chang Hsing St., Taipei, Taiwan
Republic of China

DISCUSSION PAPER

No: 8401

SECOND-ORDER EFFICIENCY IN THE ESTIMATION OF HETEROSCEDASTIC REGRESSION MODELS

Chihwa Kao*

Chung-Hua Institution for Economic Research

*The author wishes to thank John Hause, Gary Simon and John Winn for their comments on the early drafts of this paper.

SECOND-ORDER EFFICIENCY IN THE ESTIMATION OF HETEROSCEDASTIC REGRESSION MODELS

Chihwa Kao

Chung-Hua Institution for Economic Research

This paper provides a theoretical comparison of two asymptotically efficient estimators for the heteroscedastic regression models: multiplicative heteroscedasticity (Harvey, 1976) and constant coefficient of variation (Amemiya, 1973). This paper demonstrates the optimality of the maximum likelihood esitmate in the sense of second-order efficiency. I find that the Efron's statistical curvature depends on the kurtosis and the skewness of the regressors, and the degree of the heteroscedasticity. The second-order result I found in this paper could be incorporated into the estimation procedure to improve the asymptotic estimate of the variance of the MLE and thus alter the confidence interval in finite samples.

1. Introduction

One criticism of the maximum likelihood estimate is that the computation in some cases is very laborious. A practical alternative is to search for other estimates which are asymptotically as efficient as the MLE but computationally simpler. This alternative can be seen in the econometrics literature in the examples of the weighted least squares and the second-round estimate. The problem of computational difficulty has, however, become less important in the era of high speed computers.

Asymptotic efficiency is a large-sample property, and there is no guarantee that it is practical when the sample size is small. The issue that should be addressed is the justification of asymptotically efficient estimates in small samples. Is the MLE still the champion for small samples? I believe that a promising way to distinguish among asymptotically efficient estimates is to examine their convergence rates to the asymptotic variance. It is known that estimates which differ from the MLE by $o_p(n^{-1/2})$ will also be asymptotically efficient. Therefore, we should differentiate among estimates at $o_p(n^{-1})$.

Fisher (1925) and Rao (1961, 1962, 1963) introduced second-order efficiency to distinguish among different asymptotically efficient estimates. Second-order efficiency has rarely been used in the field of econometrics, where the usual investigations seems to stop when a first-order efficient estimate is found. At most, some simulation results are reported but are limited to special cases. Few satisfactory analytic results have been given in the five decades since Fisher's 1925 paper.

The main purpose of this paper is to examine two equivalent first-order efficient estimates discussed in Harvey (1976) and Amemiya (1973). They are the MLE and the second-round estimate. This paper demonstrates the optimality of the MLE in the sense of second-order efficiency. The difference in second-order efficiency (or the information loss) is found in a closed form. In section 2, I first discuss the equivalence of second-order efficiency and information loss; also introduced a measure of the second-order efficiency based on the asymptotic expansion of the variance. The information loss of the maximum likelihood estimate for the multiplicative heteroscedasticity and known coefficient of variation is given in section 3. I found that the Efron's curvature of the MLE depends on the kurtosis and skewness of the regressors and the degree of heteroscedasticity for those special cases. In section 4, the information loss of the second-round estimate is discussed. It is shown that the information loss of the MLE is less than the information loss of the second-round estimate or any other first-order efficient estimate or any other first-order efficient estimate. The conclusion is given in section 5.

2. Asymptotic Information Loss

Fisher emphasized that efficiency of a statistic is to be judged by its closeness to the efficient score $\cdot d\log L/d\theta$. Here L is the likelihood function, and θ is the vector of parameters. Rao (1961) gave a new definition of asymptotic efficiency, that is, a consistent estimator T_n of θ is said to be first order efficient if

$$|n^{-\frac{1}{2}}(d\log L/d\theta) - i_\theta n^{\frac{1}{2}}(T_n-\theta)| \to 0 \quad \text{in probability,} \qquad (1)$$

where i_θ is Fisher information per observation. Rao's definition states that the asymptotic correlation between $n^{-1/2}(d\log L/d\theta)$ and $n^{1/2}(T_n-\theta)$ is unity. More precisely, it says that $i_\theta n^{1/2}(T_n-\theta)$ has the same limiting distribution as $(1/n^{1/2})(d\log L/d\theta)$, which is $N(0, i_\theta)$, of course. There are many estimators which are first-order efficient. To distinguish among them, it is natural to examine the rapidity of convergency of (1). Consider the random variable

$$|(d\log L/d\theta) - i_\theta n(T_n-\theta)|, \qquad (2)$$

which is $n^{1/2}$ times the random variable in (1). The asymptotic variance of (2) was defined as the index for second-order efficiency.

This new definition corresponds to Fisher's definition of asymptotic efficiency, that is $\lim\{[I(T_n)][I(X_n)]^{-1}\} = \underline{1}$, where $I(T_n)$ and $I(X_n)$ are Fisher's information on θ in T_n and whole samples X_n and $\underline{1}$ is the identity matrix. There are, however, many estimates which are first-order efficient or satisfy $\lim\{[I(T_n)][I(X_n)]^{-1}\} = \underline{1}$. To distinguish among them, Fisher suggested the use of $\lim[I(X_n)-I(T_n)]$ as a measure for comparing the efficiency of estimators.

In the following lemma, I will verify a proposition not proved in Rao (1961). The lemma states that the asymptotic variance of

$$(d\log L/d\theta) - i_\theta n(T_n-\theta)$$

is the limit of the average conditional variance of $d\log L/d\theta$ given T_n, i.e., $\underline{E}[\underline{Var}(d\log L/d\theta|T_n)]$. In other words, Rao's second-order efficiency is equivalent to Fisher's information

loss and moreover it is equal to $\underline{E}[\underline{Var}(d\log L/d\theta|T_n)]$, where symbols $\underline{E}$ and $\underline{Var}$ denote the asymptotic expectation and conditional variance or the arguments respectively; see, e.g., Hosoya (1980).

Definition: Asymptotic information loss (Harvey 1979)

Estimation can be regarded as a reduction of data as well as point estimation of parameters. From the former point of view, any desirable estimation procedure must be such that the reduction of data to an estimate involves some loss of information. The amount of information contained in a statistic is usually measured by Fisher information. Let $p(T_n|\theta)$ be the probability density of a statistic T_n, then Fisher information in T_n is defined as $Var[\dot{\ell}_T]$, where $\dot{\ell}_T = d[\log p(T_n|\theta)]/dT_n$. Therefore the information loss suffered by the reduction of data $X_n = (x_1, x_2, \ldots, x_n)$ to an estimate T_n can be expressed as

$$I(X_n) - I(T_n) = Var[\dot{\ell}_\theta] - Var[\dot{\ell}_T],$$

where $I(X_n)$ is the Fisher information in whole sample X_n.

Lemma 1:

$$Var[d\log L/d\theta - i_\theta n(T_n-\theta)]$$
$$= \underline{E}[\underline{Var}(d\log L/d\theta|T_n)]$$
$$= I(X_n) - I(T_n).$$

Proof:

Let $\dot{\ell}_\theta = d\log L/d\theta$ and $\dot{\ell}_T = d\log[p(T_n|\theta)]/dT_n$, where logL is the log-likelihood function and $p(T_n|\theta)$ is the probability density of T_n.

Since $i_\theta n(T_n-\theta)$ is asymptotically equivalent to $\dot{\ell}_T$,

$$\begin{aligned}
&\mathrm{Var}[d\log L/d\theta - i_\theta n(T_n-\theta)] \\
&= \mathrm{Var}[\dot{\ell}_\theta - \dot{\ell}_T] \\
&= \mathrm{Var}[\dot{\ell}_\theta] + \mathrm{Var}[\dot{\ell}_T] - 2\mathrm{Cov}[\dot{\ell}_T, \dot{\ell}_\theta] \\
&= I(X_n) + I(T_n) - 2I(T_n) \\
&= I(X_n) - I(T_n).
\end{aligned}$$

This uses the fact that

$$\begin{aligned}
&\mathrm{Cov}[\dot{\ell}_T, \dot{\ell}_\theta] \\
&= E[\dot{\ell}_T\dot{\ell}_\theta] \\
&= \underline{E}\{\dot{\ell}_T\, \underline{E}[\dot{\ell}_\theta|T_n]\} \quad \text{using} \quad \underline{E}[\dot{\ell}_T\dot{\ell}_\theta|T_n] = \dot{\ell}_T\, \underline{E}(\dot{\ell}_\theta|T_n) \\
&= \underline{E}[\dot{\ell}_T\dot{\ell}_T] \quad \text{using} \quad \underline{E}(\dot{\ell}_\theta|T_n) = \dot{\ell}_T \text{ (Rao 1961)} \\
&= I(T_n).
\end{aligned}$$

It is known that

$$\begin{aligned}
&\underline{E}[\underline{\mathrm{Var}}(\dot{\ell}_\theta|T_n)] \\
&= \mathrm{Var}(\dot{\ell}_\theta) - \underline{\mathrm{Var}}[\underline{E}(\dot{\ell}_\theta|T_n)].
\end{aligned}$$

Hence

$$\begin{aligned}
&\underline{E}[\underline{\mathrm{Var}}(\dot{\ell}_\theta|T_n)] \\
&= \mathrm{Var}[\dot{\ell}_\theta] - \mathrm{Var}[\dot{\ell}_T] \quad \text{using} \quad E(\dot{\ell}_\theta|T_n) = \dot{\ell}_T \\
&= I(X_n) - I(T_n) \\
&= \underline{E}[\underline{\mathrm{Var}}(d\log L/d\theta|T_n)].
\end{aligned}$$

This completes the lemma. Q.E.D.

Rao (1963) gave another definition of second-order efficiency based on the expansion (under some conditions) of the variance of T_n after correction for bias, that is,

$$\mathrm{Var}(T_n) = [I(X_n)]^{-1} + [\psi(\theta)]/n^2 + o(n^{-2}).$$

The quantity $\psi(\theta)$ is a measure of the information loss. On the other hand, Efron (1975) pointed out that by modifying the MLE we can obtain an estimator with the same bias structure and smaller variance than any other consistent, efficient estimator.

3. Asymptotic Information Loss of the MLE

Fisher (1925) and Hinkley (1980) pointed out that important small sample questions for the maximum likelihood estimation include the following: (1) How much information did the MLE lose? (2) How can it be recovered? (3) How can one use the recovered information? For our purpose here, I only examine the first question: the asymptotic information loss of the maximum likelihood estimate.

Lemma 2: The asymptotic information loss of the MLE is given in the case of scalar θ by

$$I(X_n) - I(\text{MLE}) = [1/I(X_n)]\{\text{Var}(\ddot{\ell}_\theta) - [\text{Cov}(\ddot{\ell}_\theta, \dot{\ell}_\theta)]^2/\text{Var}(\dot{\ell}_\theta)\},$$

where $\dot{\ell}_\theta = d\log L/d\theta$, $\ddot{\ell}_\theta = d^2\log L/d\theta^2$ and $\hat{\theta} = \text{MLE}$ represents the maximum likelihood estimate.

Proof:

The Taylor expansion of MLE at true parameter θ gives

$$\dot{\ell}_\theta \doteq \dot{\ell}_\theta + \ddot{\ell}_\theta(\text{MLE}-\theta).$$

We know that $\dot{\ell}_{\hat{\theta}} = 0$ by the definition of the MLE.

Then

$$\dot{\ell}_{\hat{\theta}} \doteq \ddot{\ell}_\theta(\theta-\text{MLE}).$$

Since for maximum likelihood the condition MLE = t implies $\dot{\ell}_{\hat{\theta}} = \dot{\ell}_t = 0$, and taking the conditional variance given MLE on both sides we get the following result:

$$\underline{\text{Var}}[\dot{\ell}_\theta | \text{MLE} = t]$$
$$= (\theta - t)^2 \{\underline{\text{Var}}(\ddot{\ell}_\theta | \text{MLE} = t)\}.$$

Since $\dot{\ell}_\theta$ is a first-order approximation to $\dot{\ell}_{\hat{\theta}}$, then

$$\underline{\text{Var}}[\dot{\ell}_\theta | \text{MLE} = t]$$
$$= (\theta - t)^2 \text{ Var}[\ddot{\ell}_\theta | \dot{\ell}_\theta = 0].$$

We can use the properties of the joint asymptotic normal distribution of $\ddot{\ell}_\theta$ and $\dot{\ell}_\theta$, e.g., Hosoya (1979), to get the following asymptotic result:

$$\underline{\text{Var}}[\ddot{\ell}_\theta | \dot{\ell}_\theta = 0]$$
$$= \text{Var}(\ddot{\ell}_\theta) - [\text{Cov}(\ddot{\ell}_\theta, \dot{\ell}_\theta)]^2 / \text{Var}(\dot{\ell}_\theta).$$

We know that $I(X_n) - I(\text{MLE}) = \underline{E}[\underline{\text{Var}}(\dot{\ell}_\theta | \text{MLE})]$ from lemma 1, and the asymptotic variance of $(\theta - \text{MLE})$ is $1/I(X_n)$, then

$$I(X_n) - I(\text{MLE})$$
$$= [1/I(X_n)]\{\text{Var}(\ddot{\ell}_\theta) - [\text{Cov}(\ddot{\ell}_\theta, \dot{\ell}_\theta)]^2 / \text{Var}(\dot{\ell}_\theta)\}.$$

Q.E.D.

Lemma 3: (Efron 1975)

$I(X_n) - I(\text{MLE}) = I(X_n)(\text{Efron's curvature})$,

Efron's curvature is a measure of how quickly the efficient score is changing as θ changes. In other words, the variance of the MLE exceeds the Cramer-Rao lower bound in approximate proportion to Efron's curvature. Efron showed that the statistical curvature at θ is given by $(|M|/[I(X_n)]^3)^{1/2}$ where M is the covariance matrix of $\ddot{\ell}_\theta$ and $\dot{\ell}_\theta$ given by

$$M = \begin{pmatrix} \mathrm{Var}[\dot{\ell}_\theta] & \mathrm{Cov}[\ddot{\ell}_\theta, \dot{\ell}_\theta] \\ \mathrm{Cov}[\dot{\ell}_\theta, \ddot{\ell}_\theta] & \mathrm{Var}[\ddot{\ell}_\theta] \end{pmatrix}$$

The relationship between the asymptotic variance and the statistical curvature of the MLE can then be stated as in the following lemma.

Lemma 3.a:

$\mathrm{Var}(\mathrm{MLE}) = [1/I(X_n)][1 + (\text{Efron's curvature})/n] + o(n^{-2})$.

From this we expect that this second-order result will preform better in small samples approximation than the asymptotic result $1/I(X_n)$.

Example 1:

Consider Harvey's (1976) multiplicative heteroscedastic regression model

$$Y_t = \beta' X_t + \mu_t$$

$$\text{and } \sigma_t^2 = \exp(\alpha' Z_t),$$

where X_t is a $k \times 1$ vector of regressors, β is a $k \times 1$ vector of parameters, Z_t is a $p \times 1$ vector of observations on a set of variables which are usually, though not necessary, related to the regressors and α is a $p \times 1$ vector of parameters. The μ_t's are disturbance terms which are independently and normally distributed with zero means. The following example is a special case of Harvey's model by assuming β is known and α and Z_t are one-dimensional and Z_t has mean zero, i.e., Z_t denotes deviation from mean.

The θ corresponds to symbol α in the following,

The log-likelihood function is

$$\log L = -\frac{1}{2}\sum_t \alpha Z_t - \frac{1}{2}\sum_t \exp(-\alpha Z_t)(Y_t-\beta X_t)^2.$$

The first and second derivatives are:

$$d(\log L)/d\alpha = -\frac{1}{2}\sum_t Z_t + \frac{1}{2}\sum_t \exp(-\alpha Z_t) Z_t (Y_t-\beta X_t)^2$$

and

$$d^2(\log L)/d\alpha^2 = -\frac{1}{2}\sum_t \exp(-\alpha Z_t) Z_t^2 (Y_t-\beta X_t)^2.$$

The information matrix $I(X_n)$ is given by

$$\mathrm{Var}[\dot{\ell}_\theta] = -E[\ddot{\ell}_\theta] = \frac{1}{2}\sum_t Z_t^2.$$

Also

$$\begin{aligned}
&\mathrm{Cov}[\ddot{\ell}_\theta, \dot{\ell}_\theta] \\
&= E[\ddot{\ell}_\theta \dot{\ell}_\theta] \\
&= \frac{1}{4} E\Big[\sum_t Z_t^3 \exp(-\alpha Z_t)\mu_t^2 + \sum_{t\neq s} Z_t Z_s^2 \exp(-\alpha Z_s)\mu_s^2 \\
&\qquad - \sum_t \{\exp(-\alpha Z_t)\mu_t^2\}^2 Z_t^3 - \sum_{t\neq s} \exp(-\alpha Z_t)\exp(-\alpha Z_s) Z_t Z_s^2 \mu_t^2 \mu_s^2\Big] \\
&= \frac{1}{4}\Big(\sum_t Z_t + \sum_{t\neq s} Z_t Z_s^2 - 3\sum_t Z_t^3 - \sum_{t\neq s} Z_t Z_s^2\Big) \\
&= -\frac{1}{2}\sum_t Z_t^3,
\end{aligned}$$

and

$$\mathrm{Var}[\ddot{\ell}_\theta] = E[\ddot{\ell}_\theta]^2 - E^2[\ddot{\ell}_\theta] = \frac{1}{2}\sum_t Z_t^4,$$

because $E[\ddot{\ell}_\theta]^2 = \frac{3}{4}\sum_t Z_t^4 + \frac{1}{4}\sum_{t\neq s} Z_t^2 Z_s^2$ and $E^2[\ddot{\ell}_\theta] = \frac{1}{4}\sum_t Z_t^4 + \frac{1}{4}\sum_{t\neq s} Z_t^2 Z_s^2$.

Therefore the information loss of the MLE of α in this example is:

$$I(X_n) - I(MLE)$$

$$= [1/I(X_n)]\{Var(\ddot{\ell}_\theta) - [Cov(\ddot{\ell}_\theta, \dot{\ell}_\theta)]^2/Var[\dot{\ell}_\theta]\}$$

$$= \{Var[\ddot{\ell}_\theta]/I(X_n)\} - \{Cov[\ddot{\ell}_\theta, \dot{\ell}_\theta]\}^2/[I(X_n)]^2$$

$$= [\sum_t z_t^4/\sum_t z_t^2] - [(\sum_t z_t^3)^2/(\sum_t z_t^2)^2].$$

Efron's statistical curvature of the MLE is given by

$$= [1/I(X_n)][I(x_n) - I(MLE)]$$

$$= (2/\sum_t z_t^2)\{[(\sum_t z_t^4)/(\sum_t z_t^2)] - [(\sum_t z_t^3)^2/(\sum_t z_t^2)^2]\}$$

$$= 2\{[(\sum_t z_t^4)/(\sum_t z_t^2)^2] - [(\sum_t z_t^3)^2/(\sum_t z_t^2)^3]\}.$$

We know that $(\sum_t z_t^3)/(\sum_t z_t^2)^{3/2}$ is the skewness of Z_t and $(\sum_t z_t^4)/(\sum_t z_t^2)^2$ is the kurtosis of Z_t. Therefore Efron's curvature of the MLE in this example can be rewritten as

Efron's curvature

$$= 2[(\text{Kurtosis of } Z_t) - (\text{Skewness of } Z_t)^2].$$

This result suggests that Efron's curvature depends on the kurtosis and skewness of Z_t in this example.

The following example, log chi-squared regression, which is a special case of multiplicative heteroscedasticity, is taken from Hause (1980). In many cases it might be reasonable to assume that the variance is constant within subgroups of observations but that it varies from group to group. Suppose that there are m groups with p observations in each group. Suppose that each group i has its distinct mean U_i, and its own values for the vector of variance-determining covariates Z_i. Let Y_{ij} denote

the j-th observation from group i. The variance of Y_{ij} is set to equal $\exp(\alpha'Z_i)$, where i = 1, 2, ... , m and j = 1, 2, ... , p.

Example 2: Log chi-squared regression

The log-likelihood function is:

$$k - \frac{1}{2}\sum_{i}^{m}[p(\alpha'Z_i) + \sum_{j}^{p}\exp(-\alpha'Z_i)(Y_{ij}-U_i)^2].$$

The corresponding first-order conditions for the MLE of the U's and α are:

$$\partial \log L/\partial U_k = \sum_{j}^{p}\exp(-\alpha'Z_k)(Y_{kj}-U_k) = 0$$

and

$$\partial \log L/\partial \alpha_j = -\frac{1}{2}p\sum_{i}^{m}(1-S_{ii}[\exp(-\alpha'Z_i)]Z_{ij}) = 0,$$

where $S_{ii} = \sum_{j}^{p}(Y_{ij}-\overline{Y_i})^2/p$.

From above we know that the MLE of U_k is $\overline{Y_k}$, the sample mean of Y's of the k-th group. Hause (1980) showed that an equivalent formulation of this special case is given by the following log chi-squared regression:

$$\log(S_{ii}) = \alpha'Z_i + V_i,$$

where S_{ii} is the MLE of the variance of Y_{ij}, and V_i is distributed log chi-squared with k = p - 1 degrees of freedom.

It can be shown easily that the information loss of the MLE of α in the special case in which α and Z_i are one-dimensional is

$$I(X_n) - I(\text{MLE})$$

$$= (\sum_i z_i^4)/(\sum_i z_i^2) - (\sum_i z_i^3)^2/(\sum_i z_i^2)^2,$$

and the statistical curvature of MLE of α is given by:

Efron's curvature

$$= \frac{2}{k} [(\sum_i z_i^4)/(\sum_i z_i^2)^2 - (\sum_i z_i^3)^2/(\sum_i z_i^2)^3].$$

We note that the information loss is independent of k and statistical curvature goes to zero as the k goes to infinity. This means that if the number of observations in each class increase , then Efron's curvature decreases. In the limiting case, the Efron's curvature goes to zero as the number of observations in each class goes to infinity. This result is not surprising at all. It is known that the log chi-squared distribution converges to a normal distribution when the degrees of freedom goes to infinity. We also note that the information loss or Efron's curvature of the variance parameter α depends on the second, third and fourth moments of the regressors.

Example 3: Regression Models with Constant Coefficient of Variation

The heteroscedastic regression models when the variance of dependent variable is proportional to the square of its mean value has appeared in many econometric works, e.g., Amemiya (1973).

Consider the random variables Y_t with

$$E\, Y_t = \beta' X_t$$

and $(t = 1, \ldots, T)$

$$\text{Var}\, Y_t = \frac{1}{\lambda} (\beta' X_t)^2.$$

Thus, the variance of the dependent variable is proportional to the square of its expected value. That is, the coefficient of the variance of Y_t, $1/\lambda = (\text{Var } Y_t)^{1/2}/E\ Y_t$, is constant regardless of the values of X_t and Y_t.

As Amemiya (1973) pointed it out that the weighted least squares estimator:

$$\hat{\beta}_{WLS} = [\Sigma_t \frac{X_t X_t'}{t(\tilde{\beta}X_t)^2}]^{-1} \Sigma_t \frac{X_t Y_t}{(\tilde{\beta}X_t)^2}$$

with variance-covariance matrix

$$\text{Var } \hat{\beta}_{WLS} = \frac{1}{\lambda} [\Sigma_t \frac{X_t X_t'}{(\tilde{\beta}' X_t)^2}],$$

is asymptotically efficient when the Y_t is distributed as gamma $(\lambda, \lambda/\beta X_t)$, where $\tilde{\beta}$ is the least squares estimator.

That is,

$Y_t \sim$ gamma $(\lambda, \lambda/\beta X_t)$, $\lambda > 0$ and $\beta' X_t > 0$, with density

$$f(Y_t) = \frac{1}{\Gamma(\lambda)} [\frac{\lambda}{\beta' X_t}]^{\lambda} Y_t^{\lambda-1} \exp[-\frac{\lambda}{\beta X_t}] Y_t .$$

The log-likelihood function is given by

$$\log L = -n\log\Gamma(\lambda) + n\lambda\log\lambda - \lambda\Sigma_t \log\beta' X_t + (\lambda-1)\Sigma_t \log Y_t - \lambda\Sigma_t \frac{Y_t}{\beta X_t} .$$

The first and second derivatives of the log-likelihood function are,

$$\frac{\partial \log L}{\partial \beta} = -\lambda\Sigma_t \frac{1}{\beta' X_t} X_t + \lambda\Sigma_t \frac{Y_t}{(\beta' X_t)^2} X_t$$

and

$$\frac{\partial \log L}{\partial \beta \partial \beta'} = \lambda \sum_t \frac{1}{(\beta' X_t)^2} X_t X_t' - 2\lambda \sum_t \frac{Y_t}{(\beta' X_t)^3} X_t X_t' ,$$

where λ is a known constant by assumption. Therefore the asympotitic variance of the maximum likelihood estimator $\hat{\beta}_{MLE}$ is

$$\text{Var } \hat{\beta}_{MLE} = [- E(\frac{\partial^2 \log L}{\partial \beta \partial \beta'})]^{-1}$$

$$= \frac{1}{\lambda} [\sum_t \frac{1}{(\beta' X_t)^2} X_t X_t']^{-1} .$$

Of course, we note that the asymptotic variance-covariance maxtrix of the weighted least squares estimator of β is the same as the maximum likelihood estimator.

Again, let's assume λ is known and β and X_t are one-dimensional for simplicity. The θ corresponding to symbol β in the example. Therefore

$$\dot{\ell}_\theta = \frac{\partial \log L}{\partial \beta} = -\lambda \sum_t \frac{X_t}{\beta X_t} + \lambda \sum_t \frac{Y_t}{(\beta X_t)^2} X_t$$

$$\ddot{\ell}_\theta = \frac{\partial^2 \log L}{\partial \beta^2} = \lambda \sum_t \frac{X_t^2}{(\beta X_t)^2} - 2\lambda \sum_t \frac{Y_t}{(\beta X_t)^3} X_t^2$$

and

$$\dddot{\ell}_\theta = \frac{\partial^3 \log L}{\partial \beta^3} = -2\lambda \sum_t \frac{X_t^3}{(\beta X_t)^3} - 2\lambda \sum_t \frac{X_t^3}{(\beta X_t)^4} (\beta X_t - 3Y_t).$$

The information $I(X_n)$ is given by

$$\text{Var } \dot{\ell}_\theta = -E\ddot{\ell}_\theta = -\lambda \sum_t \frac{X_t^2}{(\beta X_t)^2} .$$

From the regularity conditions of the likelihood function we have the following result,

$$E(\dddot{\ell}_\theta) + E(\ddot{\ell}_\theta \dot{\ell}_\theta) = - \frac{\partial E \dot{\ell}_\theta^2}{\partial \theta} = \frac{\partial E \ddot{\ell}_\theta}{\partial \theta} . \quad \text{(e.g., Peers, 1971)}$$

This result simplies the calculation of $E(\ddot{\ell}_\theta \dot{\ell}_\theta)$.

After some algebra we have

$$E\dddot{\ell}_\theta = 2\lambda \sum_t \frac{X_t^3}{(\beta X_t)^3},$$

$$\frac{\partial E\ddot{\ell}_\theta}{\partial \theta} = 2\lambda \sum_t \frac{X_t^3}{(\beta X_t)^3},$$

and

$$\text{Var}\ \ddot{\ell}_\theta = 4\lambda \sum_t \frac{X_t^4}{(\beta X_t)^4}.$$

Therefore

$$E\ddot{\ell}_\theta \dot{\ell}_\theta = 0,$$

and

The information loss of the MLE of β,

$$I(X_n) - I(\text{MLE}) = 4\left(\frac{\sum_t \frac{X_t^4}{(\beta X_t)^4}}{\sum_t \frac{X_t^2}{(\beta X_t)^2}} \right).$$

The Efron's curvature

$$= \left(\frac{4}{\lambda}\right)\left(\frac{\sum_t \left(\frac{X_t}{\beta X_t}\right)^4}{\left[\sum_t \left(\frac{X_t}{\beta X_t}\right)^2\right]^2} \right).$$

We note that

$$\frac{\sum_t \left(\frac{X_t}{\beta X_t}\right)^4}{\left[\sum_t \left(\frac{X_t}{\beta X_t}\right)^2\right]^2}$$ is the kurtosis of $\frac{X_t}{\beta X_t}$.

This result can be further simplied to:

the Efron's curvature

$$= \frac{4}{\lambda} ,$$

in this special case.

This result suggests that the Efron's curvature depends on the degree of heteroscedasticity, i.e., the value of λ . Since

$$\text{Var } Y_t = \frac{1}{\lambda} (\beta' X_t)^2 ,$$

therefore Var Y_t and hence the Efron's curvature decreases as λ increases.

From lemma 3.a we know that Efron's curvature can be used to improve the asymptotic variance of the MLE. Let us demonstrate this lemma using example 2 in the following:

This lemma says that the variance of MLE can be estimated by [1/I(X_n)](1 + [Efron's curvature]/n), where n is the sample size.

We know that

$$I(X_n) = \frac{2}{k} (\sum_t z_t^2), \text{ calculable from data,}$$

and

Efron's curvature

$$= \frac{2}{k}\{[(\sum_t z_t^4)/(\sum_t z_t^2)^2] - [(\sum_t z_t^3)^2/(\sum_t z_t^2)^3]\},$$

from previous calculations, also calculable from data.

Therefore

Var(MLE of α)

= [1/I(X_n)][1 + Efron's curvature/n]

$$= (\frac{2}{k})/(\sum_t z_t^2) + (\frac{1}{n})(\frac{2}{k})^2\{[(\sum_t z_t^4)/(\sum_t z_t^2)^3] - [(\sum_t z_t^3)^2/(\sum_t z_t^2)^4]\}.$$

which can be calculated easily from data and may give a better

approximation in fixed sample size. In other words, we can use this information to adjust the estimated variance and thus alter a confidence interval.

4. Information Loss of Second-Round Estimator

The asymptotic information loss of the second-round estimator or any computationally complicated first-order efficient estimator is very difficult to find analytically. But it is possible to derive a relation which compares the asymptotic information loss of the maximum likelihood estimate and another efficient estimate, as in Hosoya (1979) and Hinkley (1980).

Fisher pointed out in 1925:

In approaching the maximum likelihood solution by successive approximation we have seen that starting with an inefficient statistic, a single process of approximation will in ordinary cases give an efficient statistic differing from the maximum likelihood solution by a quantity which with increasing samples decreases as n^{-1}. The loss of information, for such efficient statistics, is therefore finite for large samples, for the additional variance of $d\log L/d\theta$ will be $[I(X_n)]^2$ Var(T_n-MLE) and $I(X_n)$ increases proportionately to the sample. If the process of approximation be repeated a statistic is obtained differing from the maximum likelihood solution only to the order of $n^{-3/2}$, and for such a statistic the loss of accuracy, beyond that suffered by the maximum likelihood solution will tend to zero for large samples. The practical procedure of fitting will thus not ordinarily require more than a second process of approximation.

Let $\hat{\theta}_2$ be the second-round estimator defined by

$$\hat{\theta}_2 = \hat{\theta}_1 - [d^2\log L/d\theta\, d\theta']^{-1}\Big|_{\hat{\theta}_1} [d\log L/d\theta]\Big|_{\hat{\theta}_1},$$

where $\hat{\theta}_1$ is a consistent estimator of θ. In the following lemma we give an intuitive proof of the first-order efficiency of the second-round estimator.

Lemma 4: The estimates that differ from the MLE by $o_p(n^{-1/2})$ will also be asymptotically efficient (first-order). (Cox and Hinkley, 1974).

Lemma 5: $\hat{\theta}_2 - \hat{\theta} = o_p(n^{-1/2})$,

where $\hat{\theta}_2$ is the second-round estimate and $\hat{\theta}$ is the MLE.

Proof:

We know that

$$d\log L/d\theta = (d\log L/d\theta)\Big|_{\hat{\theta}_1} + (\theta-\hat{\theta}_1)(d^2\log L/d\theta\, d\theta')\Big|_{\hat{\theta}_1} + o|\theta-\hat{\theta}_1|$$

and $\theta-\theta_1 = o_p(n^{-1/2})$. This implies that $o|\theta-\hat{\theta}_1| = o_p(n^{-1/2})$.

Since MLE($\hat{\theta}$) is the solution of $d\log L/d\theta = 0$ without $o_p(n^{-1/2})$ and $\hat{\theta}_2$ is the solution of $d\log L/d\theta = 0$ with $o_p(n^{-1/2})$. It is clear that $\hat{\theta}_2$ - MLE $= o_p(n^{-1/2})$.

The information loss of the second-round estimator was noted in 1925 by Fisher. He stated that

$$I(X_n) - I(T_n) = I(X_n) - I(\text{MLE}) + [I(x_n)]^2 \text{Var}(T_n - \text{MLE}),$$

where T_n is any first-order efficient estimator other than the MLE, but without proof. Hosoya (1979) and Hinkley (1980) gave a proof that the asymptotic information loss of any efficient

(first-order) estimate will be greater than the asymptotic information loss of the maximum likelihood estimate; thus $I(MLE) > I(T_n)$.

The following theorem is given by Hosoya (1979). From lemma 1, we know that the information loss of a first-order efficient estimate can be expressed as $\underline{E}[\underline{Var}(\dot{\ell}_\theta | T_n)]$. However, Hosoya pointed out that

> If T_n is consistent, this information loss $\underline{E}[\underline{Var}(\dot{\ell}_\theta | T_n)]$ is meaningless asymptotically. Thus, in order to investigate the asymptotic behaviour of T_n, it is more convenient to consider the quantity: $I(X_n) - I(T_n) = \underline{E}\{\underline{Var}[\dot{\ell}_\theta | n^{1/2}(T_n-\theta)]\}$, where underlined expectation and conditional variance are understood to be taken according to the joint asymptotic distribution of $\dot{\ell}_\theta$ and $n^{1/2}(T_n-\theta)$.

Lemma 6: $n^{1/2}(T_n-\theta)$ and $n(T_n-MLE)$ are asymptotically independent,

where T_n is a first-order efficient estimator of θ and $MLE(\hat{\theta})$ is the maximum likelihood estimate. (Hosoya 1979)

We assume T_n and $\hat{\theta}$ are one-dimensional, where T_n is a first-order efficient estimator of θ and $MLE(\hat{\theta})$ is the maximum likelihood estimate.

Theorem 1: (Fisher-Rao Theorem)

$$I(X_n) - I(T_n) = I(X_n) - I(MLE) + Var[n(MLE-T_n)]/[V(\theta)]^2,$$

where $V(\theta)$ is the asymptotic variance of $n^{1/2}(T_n-\theta)$ and the variance is taken with respect to the asymptotic distribution of $n(MLE-T_n)$.

Proof:

Let T_n be an first-order efficient estimate of θ and $MLE(\hat{\theta})$ be the maximum likelihood estimate, then

$$0 = \dot{\ell}_{\hat{\theta}} \doteq \dot{\ell}_{\theta} + (MLE-\theta)\ddot{\ell}_{\theta} = \dot{\ell}_{\theta} + (MLE-T_n+T_n-\theta)\ddot{\ell}_{\theta}.$$

Therefore

$$\dot{\ell}_{\theta} = n(MLE-T_n)(1/n)(-\ddot{\ell}_{\theta}) - n^{1/2}(T_n-\theta)(n^{-1/2}\ddot{\ell}_{\theta})$$

$$\doteq n(MLE-T_n)/V(\theta) - n^{1/2}(T_n-\theta)(n^{-1/2}\ddot{\ell}_{\theta}),$$

because $(1/n)(-\ddot{\ell}_{\theta}) \to 1/V(\theta)$.

Consider

$$I(X_n) - I(T_n)$$

$$= \underline{E}[\underline{Var}\{\dot{\ell}_{\theta} | n^{1/2}(T_n-\theta)\}]$$

$$= \underline{E}[\underline{Var}\{[n(MLE-T_n)/V(\theta) - n^{1/2}(T_n-\theta)(n^{-1/2}\ddot{\ell}_{\theta})] | n^{1/2}(T_n-\theta)\}]$$

$$= \underline{E}[\underline{Var}\{n(MLE-T_n) | n^{1/2}(T_n-\theta)\}]/[V(\theta)]^2$$

$$+ \underline{E}[\underline{Var}\{n^{1/2}(T_n-\theta)(n^{-1/2}\ddot{\ell}_{\theta}) | n^{1/2}(T_n-\theta)\}]$$

$$- 2\underline{E}[\underline{Cov}\{n(MLE-T_n)/[V(\theta)], n^{1/2}(T_n-\theta)(n^{-1/2}\ddot{\ell}_{\theta}) | n^{1/2}(T_n-\theta)\}].$$

The first term in the last expression is equal to $Var\{n(MLE-T_n)\}/[V(\theta)]^2$ since $n(MLE-T_n)$ and $n^{1/2}(T_n-\theta)$ are asymptotically independent by lemma 6. The last term in the above is equal to 0 by lemma 7. The second term in the above is the information loss of the MLE, because

$$|n^{1/2}(T_n-\theta) - n^{1/2}(MLE-\theta)| = |n^{1/2}(T_n-MLE)| \to 0,$$

in probability. Therefore

$$\underline{E}[\underline{Var}\{n^{1/2}(T_n-\theta)(\ddot{\ell}_\theta/n^{1/2})\,|\,n^{1/2}(T_n-\theta)\}]$$

$$= \underline{E}[\underline{Var}\{n^{1/2}(MLE-\theta)(\ddot{\ell}_\theta/n^{1/2})\,|\,n^{1/2}(MLE-\theta)\}]$$

$$= I(X_n) - I(MLE) \quad \text{from lemma 1.}$$

Then $I(X_n) - I(T_n) = I(X_n) - I(MLE) + Var[n(T_n-MLE)]/[V(\theta)]^2$.

We note that $Var[n(T_n-MLE)]/[V(\theta)]^2 = [I(X_n)]^2 Var(T_n-MLE)$.

This is just what Fisher pointed out in 1925. Q.E.D.

Lemma 8: $I(X_n) - I(T_n) = I(X_n)$(Efron's curvature) + (*)

(Efron 1975),

where (*) is $Var[n(T_n-MLE)]/[V(\theta)]^2$; of course, (*) equals to zero when T_n is the MLE.

Theorem 1 actually says that the information loss of the maximum likelihood estimate is usually positive and minimal among the class of BAN (Best Asymptotically Normal) estimators. For practical application, the information loss of the second-round estimate might need computer simulation in order to evaluate the asymptotic variance of (T_n-MLE), since this usually cannot be expressed by a simple analytic formula (Hosoya 1979). In Kao (1984), I calculated $Var(T_n - MLE)$ in a Moute Carlo study of comparing several asymptotically efficient estimators.

5. Conclusion

Many econometric methods which depend heavily on asympotitic results need further examination in the actual use with fixed sample sizes. Rothenberg pointed this out in 1979:

Currently, most econometric inference procedures are justified on the basis of large sample theory. The exact sampling distributions generally are very complicated and, hence, are approximated by asymptotic distributions based on the central limit theorem. General methods for constructing estimators and tests that are optimal to a second-order of approximation have been developed by Efron (1975), Ghosh and Surbramanyam (1974) Pfanzagl (1973), and Pfanzagl and Wefelmeyer (1978a, b). This research on improved statistical procedures based on second-order asymptotic approximations has not yet been incorporated into the econometric textbooks. Furthermore, the application of general statistical theorems to the specific models used by econometricians is still in progress.

It is important in practice to know how well the maximum likelihood estimate and its large sample equivalents behave in finite samples. For example, is MLE actually superior to weighted least squares estimate or the second-round estimate in samll samples? In this paper, I showed analytically that the information loss of the second-round estimator is greater than the information loss of the MLE. The information loss of the MLE and second-round estimator could be calculated from the data. On the other hand, the second-order result I found in this paper, the quantity $2 (\Sigma_t z_t^4)/(\Sigma_t z_t^2) - (\Sigma_t z_t^3)^2/(\Sigma_t z_t^2)^3$ could be incorporated into the estimation procedure to improve the asymptotic estimate of the variance of the MLE and thus alter the confidence interval in finite samples. I believe, that this paper contributes to the studies of estimation and hypothesis testing of heteroscedastic regression models and provides a theoretical basis for choosing among asymptotically efficient estimates.

References

Amemiya, T., 1973, Regression Analysis When the Variance of the dependent variable is proportional to the square of its expectation, Journal of the American Statistical Association 68, 928-934.

Cox, D. R., 1974, Theoretical Statistics (Chapman and Hall, London).

Efron, B., 1975, Defining the Curvature of a Statistical Problem (with Application to Second-Order Efficiency), Annals of Statistics 3, 1189-1242.

Fisher, R. A., 1925, Theory of Statistical Estimation, Proc. Cambridge Philos. Soc. 22, 700-725.

Ghosh, J. K. and Subramagam, K., 1974, Second-Order Efficiency of Maximum Likelihood Estimators, Sankhya A36, 325-358.

Harvey, A. C., 1976, Estimating Regression Models with Multiplicative Heteroscedasticity, Econometrica 44, 461-465.

Hause, J. C., 1980, Parameter Estimation in the Normal Model When Mean and Variance Depend on Covariates, Mimeographed, Suny at Stony Brook.

Hinkley, D. V., 1980, Theory of Statistical Estimation: The Fishers 1925 Paper, in: Fienberg, Gani, Keifer and Krickeberg, eds., Lecture Notes in Statistics.

Hosoya, Y., 1979, Higher-Order Efficiency in the Estimation of Linear Process, Annals of Statistics 7, 516-530.

_____, 1980, Conditionality and Maximum Likelihood Estimation, in: K. Matusita, eds., Recent Developments in Statistical Inference and Data Analysis (North-Holland, Amsterdam).

Kao, C., 1984, Small Samples Properties of Estimates of Regression Models with Multiplicative Heteroscedasticity: The results of Some Monte Carlo experiments, Academia Economic papers (to appear).

Peers, H. W., 1971, Likelihood ratio and associated test criteria, Biometrika 58, 577-587.

Quandt, R. E., 1979, A Note on the Tobit Likelihood Function, Economics Letters 4, 361-362.

Rao, C. R., 1961, Asymptotic Efficiency and Limiting Information, Proc. 4th Berkeley Symp., 531-545.

___, 1962, Efficient Estimates and Optimum Inference Procedure in Large Samples, Journal of the Royal Statistical Society Ser. B, 46-72.

___, 1963, Criteria of Estimation in Large Samples, Sankhya A25, 189-206.

Rothenberg, T., 1979, Statistical Analysis of Econometric Models: A Comment, Journal of the American Statistical Association 74, 648-649.