R. de Econometria	Rio de Janeiro	v. VII nº 2	p. 5 - 28	novembro 1987
-------------------	----------------	-------------	-----------	---------------

ADAPTIVE ESTIMATION OF HETEROSKEDASTIC ECONOMETRIC MODELS*

P. M. Robinson**

ABSTRACT

We focus on the linear model with conditional heteroskedasticity of unknown form. "Adaptive" estimators of the coefficients of the linear model, based on no rigid parameterisation of the heteroskedasticity, but having the same asymptotic efficiency as estimators which do use such information, are surveyed. A small Monte Carlo study of their performance is reported. We describe a modification of the popular paradigm in which the variance is a function of the mean, allowing this function to be of unknown form. We describe a modification also of the autoregressive conditional heteroskedasticity (ARCH) model, in which the heteroskedasticity function is of unknown form.

* This article is based on research funded by the Economic and Social Research Council (ESRC) reference number: B00232156. The Monte Carlo simulations in Section 3 were carried out by Miguel Delgado.

** London School of Economics.

1. INTRODUCTION

Much econometric methodology concerns parametric inference on the expectation of one set of economic variables, Y , conditional on another set of economic variables, X . This conditional expectation is assumed to be a given function of X and of finitely many unknown parameters, β . For simplicity we shall throughout take the dimensionality of Y to be one, and the conditional expectation to be linear both in X and β , thus

$$E(Y|X) = \beta'X, \quad (1)$$

where X and β are $p \times 1$ column vectors and the prime indicates transposition. Given n observations (X_i, Y_i) , $i = 1, \dots, n$, many estimators of β are available, but their properties depend critically on other features of the conditional distribution of Y given X , besides (1). In particular, the properties of the ordinary least squares (OLS) estimator $\hat{\beta}$ of β depend on the conditional variance of Y (equivalently the conditional variance of the "residual" variable $U = Y - \beta'X$), namely

$$V(Y|X) = E((Y - \beta'X)^2|X) = \sigma^2(X), \quad (2)$$

to the extent that under regularity conditions the limiting covariance matrix of $n^{1/2}(\hat{\beta} - \beta)$ is a non-trivial function of $\sigma^2(X)$. This means that if $\sigma^2(X)$ is not believed constant over X , and if standard large-sample inferences are to be based on the central

limit theorem for β , proper account must be taken of $\sigma^2(X)$'s presence. A method of doing this was proposed and developed by Eicker (1963, 1967), Hinkley (1977), White (1980), and others. However, while such methods afford consistent interval estimators and asymptotically valid hypothesis tests, such interval estimators are not necessarily asymptotically the smallest possible, and such tests are not asymptotically locally most powerful, unless $\sigma^2(X)$ is in fact constant. This is because there can exist more efficient estimators of β , of the generalized least squares (GLS) type, which can form the basis for asymptotically more efficient inferences.

The GLS estimators of β are of the form

$$\hat{\beta} = \left(\sum_{i=1}^n X_i X_i' \hat{\sigma}_i^{-2} \right)^{-1} \sum_{i=1}^n X_i Y_i \hat{\sigma}_i^{-2}, \quad (3)$$

where $\hat{\sigma}_i^2$ is a predictor of $\sigma_i^2 = \sigma^2(X_i)$. In textbooks on regression analysis, and in the sample survey literature, $\sigma^2(X)$ is often assumed to be known up to a factor of proportionality, $\sigma^2(X) = \gamma g(X)$, where g is known but γ is unknown, when we may simply insert $g_i = g(X_i)$ in place of $\hat{\sigma}_i^2$ in (3). Notice that this assumption constitutes a statement that we have even more detailed knowledge of the conditional variance (2) than of the conditional mean (1). To reflect a comparable degree of knowledge of the conditional mean and variance, one postulates a parametric function $\sigma^2(X) = g(X; \theta)$, where g is now a known function of X and of an unknown, finite-dimensional parameter vector θ . The statistical and econometric literature is replete with proposals for functional forms g . For given g , one finds an estimator $\hat{\theta}$ of θ , for example by (possibly nonlinearly) regressing the squares of the OLS residuals $\hat{U}_i = Y_i - \hat{\beta}'X_i$ on $g(X_i; \theta)$, and inserts the $\hat{\sigma}_i^2 = g(X_i; \hat{\theta})$ in (3). Under regularity conditions, the resulting $\hat{\beta}$ will have the same limiting distribution as if $\hat{\sigma}_i^2$ in (3) were the actual variance.

The function $\sigma^2(X)$ is essentially a "nuisance" function, the focus of interest being β in equation (1), and a functional form for $\sigma^2(X)$, even one involving unknown parameters, is liable to be somewhat arbitrary. If $g(X;\theta)$ is mis-specified, then $\hat{\beta}$ will still typically be consistent, but the estimator of the covariance matrix of $n^{1/2}(\hat{\beta} - \beta)$, namely

$$\left(\frac{1}{n} \sum_{i=1}^n X_i X_i' \hat{\sigma}_i^{-2} \right)^{-1}, \quad (4)$$

will typically be inconsistent, so its use will lead to inconsistent interval estimators and asymptotically invalid hypothesis tests. There is thus motivation for avoiding the imposition of any fixed functional form on $\sigma^2(X)$, in which case the model formed by (1) and (2) may be termed "semiparametric". It is possible to construct estimators $\hat{\beta}$ of the form (3) which use estimators $\hat{\sigma}_i^2$ that are not based on any specific functional form, in the sense that under a suitable asymptotic regime this $\hat{\beta}$ has the same asymptotic efficiency as one based on a $\sigma^2(X)$ that is known up to scale or up to finitely many unknown parameters, so its limiting covariance matrix can be consistently estimated by (4); such a $\hat{\beta}$ may be termed "adaptive". We emphasize at the outset that this "adaptive" property is really an artifact not so much of the estimator itself but of the nature of the asymptotic theory, because in reality σ^2 is "parameterised" in an unconventional fashion, and the number of "parameters" is regarded as increasing slowly as n increases, not held fixed as in standard parametric theory. Thus the appeal of the "semiparametric" theory over the "parametric" theory is to some extent contrived and the mystique that sometimes surrounds "semiparametric" estimators is not really deserved. Nevertheless, the semiparametric theory does reflect a somewhat different, more flexible, philosophy from the parametric one, and the estimators of $\sigma^2(X)$ that have been employed are of a different character, based on nonparametric regression. The idea of using nonparametric regression to estimate $\sigma^2(X)$, and thence β , seems to have originated

whit Rose (1978), and the methodology and theory was subsequently developed and adapted by Carroll (1982) and Robinson (1985). In this paper we first review the work of these authors, then report results of a Monte Carlo study of one of these estimators. In subsequent sections we discuss modifications not previously treated in the literature, to maximum likelihood estimation in the case that the conditional variance $\sigma^2(X)$ is an unknown function only of the conditional mean, $\beta'X$; and to the autogressive conditional heteroskedasticity (ARCH) model introduced by Engle (1982), where time series dynamics are involved in the sense that instead of conditioning in (2) on exogenous variables one is conditioning on either lagged dependent variables or innovations at previous time points - we relax previous assumptions that the conditional variance has known functional form.

2. ADAPTIVE GLS ESTIMATORS

The conditional variance (2) may be viewed as a regression function, of $(Y - \beta'X)^2$ on X , or as a functional of regression functions, namely $\sigma^2(X) = E(Y^2|X) - (E(Y|X))^2$. Either of these representations lead to nonparametric estimators of σ^2 , via nonparametric regression. A number of methods of nonparametric regression have been proposed, and this is not the place to review them; a recent review is Collomb (1985). The "adaptive" estimators of β of Rose (1978) and Carroll (1982) use kernel estimators of $\sigma^2(X)$, while those of Robinson (1985) use nearest neighbour estimators of $\sigma^2(X)$, and it will suffice to discuss these methods. In order to add some necessary flexibility we shall introduce an observable $q \times 1$ vector Z such that $\sigma^2(\cdot)$ is actually known to depend only on Z . The vector Z consists only of bona fide "variables", excluding any constant intercept term present in X .

More generally, Z can be a subset of X , or a functional of X . Indeed, one could also modify the specification in (1) and (2) such that the conditioning is on X and Z respectively, where neither X nor Z is completely functionally dependent on the other. Thus we behave as if we have observables $Z_i, i = 1, \dots, n$, as well as the observables X_i and $Y_i, i = 1, \dots, n$.

Let us consider first the kernel method. Introduce an integrable, real-valued "kernel" function K , defined on q -dimensional Euclidean space, and a $q \times q$ positive definite "bandwidth" matrix h . A kernel nonparametric estimator of the regression function of a variate V on Z is

$$\hat{E}(V|Z) = \frac{\sum_{j=1}^n V_j K(h^{-1}(Z - Z_j))}{\sum_{j=1}^n K(h^{-1}(Z - Z_j))}, \quad (5)$$

given observables $V_i, i = 1, \dots, n$, see Nadaraya (1964), Watson (1964). Thus, one estimator of σ_i^2 is

$$\begin{aligned} \hat{\sigma}_i^2 &= \frac{\sum_{j=1}^n Y_j^2 K(h^{-1}(Z_i - Z_j))}{\sum_{j=1}^n K(h^{-1}(Z_i - Z_j))} \\ &\quad - \left(\frac{\sum_{j=1}^n Y_j K(h^{-1}(Z_i - Z_j))}{\sum_{j=1}^n K(h^{-1}(Z_i - Z_j))} \right)^2. \end{aligned} \quad (6)$$

An alternative estimator is based on the OLS residuals $\hat{U}_i^2$,

$$\hat{\sigma}_i^2 = \frac{\sum_{j=1}^n \hat{U}_j^2 K(h^{-1}(Z_i - Z_j))}{\sum_{j=1}^n K(h^{-1}(Z_i - Z_j))}. \quad (7)$$

The estimators (6) and (7) were studied by Rose (1978), Carroll (1982), Matloff, Rose and Thai (1984), with emphasis on the case $q = 1$. The function K , chosen by the practitioner, is typically heavily concentrated round the origin, and popular choices are the

multivariate normal density, and the uniform density, though K need not be non-negative everywhere. The matrix h is also chosen by the practitioner and is regarded in the asymptotic statistical theory as approaching 0 as $n \rightarrow \infty$, but at a slower rate. Carroll (1982) showed that $\hat{\sigma}$ using (7) is asymptotically efficient for $q = 1$. When $q > 1$ it is unlikely that an h proportional to the identity matrix will be a good choice, as the scales of elements of Z will generally vary. Thus h may be a diagonal matrix or a more general positive definite matrix: when K is the multivariate normal density, a natural choice for h^2 would be the sample covariance matrix of the Z_i . The choice of h is quite crucial, and a possible automatic method of choosing h is cross-validation, i.e. define $\hat{\sigma}_{i(i)}^2$ as (6) or (7) after omitting the i th summands from the sums, and find the value of h that minimises the weighted residual sum of squares

$$\sum_{i=1}^n (Y_i - \hat{\beta}'_{(i)} X_i)^2 \hat{\sigma}_{i(i)}^{-2}, \quad (8)$$

where $\hat{\beta}$ is defined like (3), but with $\hat{\sigma}_i^2$ replaced by $\hat{\sigma}_{i(i)}^2$. The estimators (6) or (7) can be used in the GLS formula (3), and the covariance matrix formula (4).

The second type of approach, nearest neighbours, weights observations not according to the distance of their Z from the Z at which the nonparametric function $\sigma^2(Z)$ is being estimated, but according to ranks. The nearest neighbour method was introduced by Fix and Hodges (1951), a more recent reference is Stone (1977), and there is a large literature concerning its properties, but its first use in adaptive estimation is in Robinson (1985). If $z_{i\ell}$ is the ℓ th element of Z_i , define

$$S_\ell = \frac{1}{n-1} \sum_{i=1}^n \{z_{i\ell} - (\frac{1}{n} \sum_{i=1}^n z_{i\ell})\}^2, \quad \ell = 1, \dots, q,$$

$$\rho_{ij} = \sum_{\ell=1}^q \frac{(z_{i\ell} - z_{j\ell})^2}{S_\ell}, \quad i, j = 1, \dots, n, \quad i \neq j.$$

For an integer $k < n$, the constants c_i are positive for $i=1, \dots, k$, zero for $i = k+1, \dots, n$, and sum to 1, for example

$$\begin{aligned} \text{Uniform } c_i &: c_i = k^{-1} & , \quad i = 1, \dots, k , \\ \text{Triangular } c_i &: c_i \propto k-i+1 & , \quad i = 1, \dots, k , \\ \text{Quadratic } c_i &: c_i \propto \{k^2 - (i-1)^2\} , & i = 1, \dots, k . \end{aligned} \quad (9)$$

Let r_{ij} be 1 plus the number of Z's closer to Z_i than Z_j , and s_{ij} be 1 plus the number of Z's equally far from Z_i as Z_j , and define

$$\begin{aligned} W_{ij} &= \frac{1}{s_{ij}} \sum_{\ell=r_{ij}}^{r_{ij} + s_{ij} - 1} c_\ell , \quad i \neq j ; \\ &= 0 & , \quad i=j . \end{aligned}$$

When there are no "ties", so far as rankings of Z_j from Z_i are concerned, $s_{ij} = 1$ for all j , so that $W_{ij} = c_{r_{ij}}$. The nearest neighbour estimators of σ_i^2 are given by

$$\hat{\sigma}_i^2 = \sum_{j=1}^n Y_j^2 W_{ij} - \left(\sum_{j=1}^n Y_j W_{ij} \right)^2 , \quad (10)$$

or

$$\hat{\sigma}_i^2 = \sum_{j=1}^n U_j^2 W_{ij} . \quad (11)$$

Both (10) and (11) and modifications were mentioned by Robinson (1985), who emphasised (11). The practitioner chooses the pattern of weights c_i , and also the number of nearest neighbours, k , where k is here the "bandwidth" parameter, which in the asymptotic theory increases slowly as $n \rightarrow \infty$. Robinson (1985) established asymptotic efficiency of $\hat{\beta}$ using (11), under milder regularity conditions on the random variables than Carroll (1982), and the consistency of the covariance matrix estimator (4). Note that the definition of ρ_{ij} automatically accounts for variations in scale of Z 's elements. The choice of k is crucial, and a possible automatic method of choosing k is by the value that minimises (8), where we note that our definitions of $\hat{\sigma}_i^2$ in (10) and (11) already "leave out" the i th observation so that $\hat{\sigma}_{i(i)}^2 = \hat{\sigma}_i^2$ here.

If either the kernel or nearest neighbour method is used but one wishes to protect oneself against a poor choice of "bandwidth", a "heteroskedasticity-robust" covariance matrix estimator for $n^{1/2} (\hat{\beta} - \beta)$, in place of (4), is

$$\left(\frac{1}{n} \sum_{i=1}^n x_i x_i' \hat{\sigma}_i^{-2} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i x_i' \hat{U}_i^2 \hat{\sigma}_i^{-4} \right) \left(\frac{1}{n} \sum_{i=1}^n x_i x_i' \hat{\sigma}_i^{-2} \right)^{-1}$$

3. MONTE CARLO SIMULATIONS

Simulations examining adaptive estimators using the kernel estimators (6) and (7) have been presented by Carroll (1982) and Matloff, Rose and Thai (1984), so we focus instead on the estimator proposed by Robinson (1985) that uses the nearest neighbour estimator (11). Many simulations have been carried out by Miguel Delgado on the University of London's CRAY-1S/1000 computer, and we present only a tiny selection of these; further details of the

simulations, and of the computational algorithm employed, may be obtained from him at LSE's Economics Department. The regression simulated was

$$y_i = \beta_1 + \beta_2 z_{i1} + \beta_3 z_{i2} + U_i ,$$

where $\beta_1 = \beta_2 = \beta_3 = 1$ and U_i and z_{i1} are both $N(0,1)$ while z_{i2} is Poisson with mean 4, z_{i1} , z_{i2} and U_i being independent for each i . The regression model thus contains one continuous explanatory variable, and one that is discrete with infinite support. We took

$$\sigma^2(z_i) = 0.5\{1 + 0.5(1 + 0.6z_{i2} + 0.5z_{i2}^2)\} ,$$

so $\sigma^2(z)$ actually varies with only one of the z 's, but we proceeded as if the econometrician is ignorant of this fact and took $q = 2$ when computing the nearest neighbour estimators (11), using both z_{i1} and z_{i2} . We report results only for $n = 20$, but observe that this sample size is really too small to warrant taking seriously the asymptotic properties described in the previous section. However we do report variations in the number of nearest neighbours k in (11), taking $k = 2, 5$ and 8 in order to study the sensitivity of our adaptive estimators to degree of smoothing, and for each of these k we also tried each of the three weights in (9). One thousand replicates were generated, and summary statistics reported in Tables 1-3. Table 1 gives results for estimating β_i , $i = 1, 2, 3$. The figures given under the heading BIAS, SD (standard deviation) and RMSE (root mean squared error) are actually not raw figures but ratios of these quantities (after averaging over the replications) to those for the infeasible GLS estimator computed from full knowledge of $\sigma^2(z_i)$. In each of the tables we report results for the OLS estimator and for each of the nine combinations of the three choices of k and three weight functions, $2, U$ meaning

"k=2, Uniform", T meaning "Triangular" and Q meaning "Quadratic".

Table 1

β_1 RATIOS

ESTIMATOR	BIAS	SD	RMSE
OLS	39.072	2.916	2.916
2,U	31.418	2.855	2.856
2,T	18.265	2.865	2.865
2,Q	25.862	2.856	2.856
5,U	33.028	1.987	1.988
5,T	31.056	1.986	1.986
5,Q	33.193	1.972	1.973
8,U	27.450	2.051	2.051
8,T	20.255	1.943	1.943
8,Q	23.207	1.960	1.960

Table 2

β_2 RATIOS

ESTIMATOR	BIAS	SD	RMSE
OLS	32.776	2.181	2.184
2,U	23.393	2.353	2.354
2,T	19.414	2.363	2.364
2,Q	21.619	2.353	2.354
5,U	9.741	1.654	1.655
5,T	12.885	1.670	1.670
5,Q	11.264	1.652	1.653
8,U	6.114	1.746	1.747
8,T	8.137	1.654	1.654
8,Q	7.277	1.676	1.676

Table 3
 β_3 RATIOS

ESTIMATOR	BIAS	SD	RMSE
OLS	6.721	2.432	2.432
2,U	8.089	2.327	2.328
2,T	4.749	2.358	2.358
2,Q	6.720	2.336	2.336
5,U	6.620	1.771	1.772
5,T	8.214	1.771	1.772
5,Q	8.108	1.761	1.762
8,U	8.368	1.842	1.842
8,T	5.767	1.733	1.734
8,Q	6.619	1.755	1.755

The bias ratios for the adaptive method are quite poor, though surprisingly in most cases slightly less than those for OLS; the biases do not seem to vary very systematically with k or with the weights c_i . For β_1 , the adaptive estimators perform very disappointingly in that they are actually more variable than OLS, but for β_2 (with $k=5,8$) and for β_3 this situation is reversed. As expected, variability varies inversely with k . Generally the adaptive estimators' performance is far inferior to its asymptotic performance, but we repeat that $n = 20$ is a small sample.

4. VARIANCE-A-FUNCTION-OF-THE-MEAN

Some empirical evidence suggests that the conditional variance depends on X only through the conditional mean, thus $\sigma^2(X) = g(\beta'X)$ for some positive-valued function g . For example, $g(x) \propto x$ (when $\beta'X$ is positive-valued) as for Poisson data, or $g(x) \propto x^2$, as in Amemiya

(1973). In most of the literature, g has been assumed known, at least up to a factor of proportionality (see e.g. Amemiya, 1973, Jobson and Fuller, 1980, and references therein). However, it is natural to consider the case g unknown, and here Carroll (1982) considered the GLS estimator (3) when prior information that $\sigma^2(X)$ depends on X only via $\beta'X$ is used: the vector variable Z in the argument of the kernel function K in (7) is replaced by the scalar $\hat{\beta}'X$,

reduction of dimensionality in kernel estimation is valuable computationally, and is likely associated with improved finite-sample statistical properties, no gain in asymptotic efficiency is achieved over the estimators described in Section 2 which do not restrict σ^2 : the limiting covariance matrix is still

$$\{E(XX'\sigma^{-2}(X))\}^{-1}. \quad (12)$$

As noted by Robinson (1985), efficiency gains are achievable via a semiparametric maximum likelihood approach, which we now detail.

If the conditional distribution of Y given X is normal, specifically

$$Y \sim N(\beta'X, g(\beta'X)) \quad (13)$$

the log likelihood is given by

$$Q(\beta) = -\frac{1}{2n} \sum_{i=1}^n \{\log 2\pi + \log g(\beta'X_i) + (Y_i - \beta'X_i)^2 g(\beta'X_i)^{-1}\}, \quad (14)$$

where, at risk of possible confusion, β represents any admissible

value, not necessarily the true one in (13). Pretend first that $g(x)$ is of known, differentiable, form, and write $G(x) = (d/dx)g(x)$. Then

$$\frac{\partial Q(\beta)}{\partial \beta} = \frac{1}{n} \sum_{i=1}^n \frac{x_i}{g(\beta'x_i)} [U_i + \frac{1}{2} G(\beta'x_i) \{ \frac{U_i^2}{g(\beta'x_i)} - 1 \}].$$

Under regularity conditions

$$n^{1/2} \frac{\partial Q(\beta)}{\partial \beta} \xrightarrow{d} N(0, \Phi)$$

at the true β , where

$$\Phi = E\left\{ \frac{XX'}{g(\beta'X)} \left(1 + \frac{1}{2} \frac{G(\beta'X)^2}{g(\beta'X)} \right) \right\} \quad (15)$$

is also the probability limit of $-(\partial^2 / \partial \beta \partial \beta') Q(\beta)$. Denote by β^* a value of β maximising (14). Then

$$n^{1/2} (\beta^* - \beta) \xrightarrow{d} N(0, \Phi^{-1}). \quad (16)$$

Because $\Phi^{-1} \leq (12)$, β^* is at least as efficient asymptotically as the adaptive $\hat{\beta}$ estimators discussed in Section 2 (equality is attained when there is no conditional heteroskedasticity), but we emphasize that this result assumes the structure (13), in the absence of which β^* can be inconsistent. Notice that when g is known a more realistic specification is $\sigma^2(X) = \gamma g(\beta'X)$, where γ is unknown, (indeed Jobson and Fuller (1980) allowed for a more general dependence of $\sigma^2(X)$ on $\beta'X$, X and other parameters). The effect of this specification is to add a positive term to the

limiting covariance matrix ϕ^{-1} in (16). However, since we want to treat $g(x)$ as unknown at any x -value we cannot identify γ and so we aim for an estimator that is adaptive relative to β^* as defined above, and thus even more efficient than the estimators of Amemiya (1973), Jobson and Fuller (1980), despite their assumption of a known functional form.

It remains to construct this adaptive estimator. Let us use kernel estimation. For given β , $g(\beta'X_i)$ can be predicted by

$$\hat{g}(\beta'X_i) = \frac{\sum_{j=1}^n \hat{U}_j^2 K_{ij}(\beta)}{\sum_{j=1}^n K_{ij}(\beta)} \quad (17)$$

for given β , where K is here real-valued and

$$K_{ij}(\beta) = K(\beta'(X_i - X_j)/h).$$

On inserting (17) in (14), we might optimise over β . We still have to construct an estimator of (15) in order to conduct statistical inferences, however, and we propose a "one-step" estimator which simultaneously achieves the same asymptotic efficiency without nonlinear optimisation, and entails as a by-product an estimator of (15). Assume that a differentiable kernel K is used, for example the standard normal density, and write $L(x) = (d/dx)K(x)$, for example $L(x) = -xK(x)$ in the normal case. We need to estimate also G , and nonparametric estimation of derivatives of regression curves has been studied in the statistical literature (e.g. Schuster and Yakowitz, 1979). The obvious estimator, for given β , is

$$\hat{G}(\beta'X_i) = \frac{d}{d(\beta'X_i)} \hat{g}(\beta'X_i)$$

$$= \frac{\sum_j \hat{U}_j^2 L_{ij}(\beta)}{\sum_j K_{ij}(\beta)} - \frac{\sum_j \hat{U}_j^2 K_{ij}(\beta) \sum_j L_{ij}(\beta)}{(\sum_j K_{ij}(\beta))^2},$$

where the $\hat{U}_j$ are still the OLS residuals and

$$L_{ij}(\beta) = L(\beta'(X_i - X_j)/h).$$

Then take one Gauss-Newton step from the OLS estimator $\hat{\beta}$,

$$\beta^+ = \tilde{\beta} - \tilde{\Phi}^{-1}(\tilde{\beta}) \hat{\Phi}(\tilde{\beta}) \quad (18)$$

where

$$\hat{\Phi}(\beta) = \frac{1}{n} \sum_{i=1}^n \frac{X_i - X'_1}{g(\beta'X_i)} \left(1 + \frac{1}{2} \frac{\tilde{G}(\beta'X_i)^2}{\tilde{g}(\beta'X_i)} \right),$$

$$\tilde{\Phi}(\beta) = \frac{1}{n} \sum_{i=1}^n \frac{X_i}{\tilde{g}(\beta'X_i)} \left[\hat{U}_i + \frac{1}{2} \frac{\tilde{G}(\beta'X_i)}{\tilde{g}(\beta'X_i)} \left(\frac{\hat{U}_i}{\tilde{g}(\beta'X_i)} - 1 \right) \right].$$

Under regularity conditions, β^+ should have the same limiting distribution as β^* , see (16), with ϕ^{-1} consistently estimated by $\tilde{\Phi}^{-1}(\tilde{\beta})$. If desired, further iterations can be conducted in the manner of (18), but with no further improvement in efficiency.

5. ARCH MODEL

Now let us suppose the (X_i, Y_i) , $i=1,2,\dots$ form a time series, so they are not independent across i , indeed

$$Y_i | \Omega_i \sim N(\beta'X_i, g_i), \quad g_i = g(U_{i-1}, \dots, U_{i-q}), \quad (19)$$

where Ω_i is the information set,

$\Omega_i = \{X_i, X_{i-1}, \dots; Y_{i-1}, Y_{i-2}, \dots\}$, $U_i = Y_i - \beta'X_i$ as before, g is a non-trivial function of $U_{i-1}, \dots, U_{i-q}$, and X_i might include lagged Y_i . Notice that (19) implies the U_i are uncorrelated with constant variance $\sigma^2 = E(g_i)$, but they are not independent in general. The OLS estimator $\hat{\beta}$ of β is still consistent for β in general, indeed when X_i includes no lagged Y_i the usual OLS standard errors are still valid. However, under (19) a more efficient, maximum likelihood, estimator is available. The model (19) is a version of the ARCH regression model considered by Engle (1982), and developed by a number of subsequent authors. It has been assumed that g is a known parametric function, specifically

$$g(u_1, \dots, u_q) = \alpha_0 + \alpha_1 u_1^2 + \dots + \alpha_q u_q^2, \quad (20)$$

for unknown coefficients $\alpha_0, \dots, \alpha_q$, but there seems to be no compelling reason in general to assert (20) or any other rigid functional form.

The log likelihood on the basis of (19) is

$$Q(\beta) = \frac{1}{2n} \sum_{i=q+1}^n \left\{ \log 2\pi + \log g_i + \frac{(Y_i - \beta'X_i)^2}{g_i} \right\}. \quad (21)$$

Taking g to be known for the time being (but not necessarily (20)),

$$G_i = \frac{\partial g_i}{\partial \beta} = - \sum_{\ell=1}^m X_{i-\ell} g^\ell(u_{i-1}, \dots, u_{i-q}) ,$$

where

$$g^\ell(u) = \frac{\partial g(u)}{\partial u_\ell} , \quad \ell=1, \dots, q ,$$

and u_ℓ is the ℓ th element of u . Thus

$$\frac{\partial Q(\beta)}{\partial \beta} = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{X_i U_i}{g_i} + \frac{1}{2} \frac{G_i}{g_i} \left(\frac{U_i^2}{g_i} - 1 \right) \right\} .$$

It is easily seen that under regularity conditions, if β^* is the estimator maximising (21),

$$n^{1/2} (\beta^* - \beta) \xrightarrow{d} N(0, \Psi^{-1}) , \quad (22)$$

where

$$\Psi = E \left\{ \frac{X_i X_i'}{g_i} + \frac{1}{2} \frac{G_i G_i'}{g_i^2} \right\} . \quad (23)$$

When X_i contains no lagged Y_i ,

$$E \left(\frac{X_i X_i'}{g_i} \right) = E(X_i X_i') E \left(\frac{1}{g_i} \right) \geq \frac{E(X_i X_i')}{\sigma^2}$$

by Jensen's inequality so that β^* is indeed more efficient than OLS. When X_i does contain lagged y_i the limiting covariance matrix of OLS is

$$\{E(X_i X_i')\}^{-1} E(X_i X_i' g_i) \{E(X_i X_i')\}^{-1}$$

and again this dominates $\{E(X_i X_i' / g_i)\}^{-1}$, so that again β^* is more efficient. Notice that (23) reduces to expression (24) of Engle (1982) in the special case (20). If we now take g to be unknown, we can construct an adaptive estimator as follows. Let

$$\hat{g}_i = \frac{n}{j=q+1} \hat{u}_j^2 K_{ij} / \sum_{j=q+1}^n K_{ij},$$

where

$$K_{ij} = K(h^{-1}(\hat{u}_{i-1} - \hat{u}_{j-1}, \dots, \hat{u}_{i-q} - \hat{u}_{j-q}))$$

where, as in Section 2, K is a kernel function of q variables and h is a $q \times q$ matrix. Let

$$\hat{G}_i = - \sum_{\ell=1}^q x_{i-\ell} \hat{g}_i^\ell$$

where

$$\hat{g}_i^\ell = \frac{\sum_j \hat{u}_j^2 L_{ij}^\ell}{\sum_j K_{ij}} - \frac{\sum_j \hat{u}_j^2 K_{ij} \sum_j L_{ij}}{(\sum_j K_{ij})^2}$$

$$L_{ij}^{\ell} = L^{\ell}(h^{-1}(\hat{U}_{i-1}-\hat{U}_{j-1}, \dots, \hat{U}_{i-q}-\hat{U}_{j-q}))$$

and $L^{\ell}(u) = (\partial/\partial u_{\ell})L(u)$, so that when K is a product of q standard normals, $L^{\ell}(u) = -u_{\ell}L(u)$. Using the one-step approach of Section 4,

$$\beta^{+} = \tilde{\beta} - \tilde{\Psi}^{-1}\tilde{\psi},$$

where

$$\hat{\Psi} = \frac{1}{n} \sum_{i=q+1}^n \left(\frac{X_i X_i'}{g_i} + \frac{1}{2} \frac{\hat{G}_i \hat{G}_i'}{\hat{g}_i^2} \right),$$

$$\hat{\psi} = \frac{1}{n} \sum_{i=q+1}^n \left\{ \frac{X_i \hat{U}_i}{\hat{g}_i} + \frac{\hat{G}_i}{g_i} \left(\frac{\hat{U}_i^2}{\hat{g}_i} - 1 \right) \right\}$$

Under regularity conditions, β^* should have the same limiting distribution as β^* , see (22), with ψ^{-1} estimated by $\hat{\Psi}^{-1}$. Thus we have constructed an adaptive version of Engle's (1982) estimator, that relaxes his assumption that the form of heteroskedasticity is known with no loss of efficiency.

REFERENCES

- AMEMIYA, T. "Regression Analysis is Proportional to the Square of its Expectation". Journal of the American Statistical Association 68: 928-934, 1973.
- CARROLL, R. J. "Adapting for Heteroskedasticity in Linear Models". Annals of Statistics 10: 1224-1233, 1982.
- COLLOMB, G. "Nonparametric Regression: An Up-To-Date Bibliography". Statistics 2: 309-324, 1985.
- EICKER, F. "Asymptotic Normality and Consistency of the Least Squares Estimator for Families of Linear Regressions". Annals of Mathematical Statistics 34: 447-456, 1963.
- _____. "Limit Theorems for Regressors in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, ed. by L. LeCam and J. Neyman. Berkeley, University of California Press, 59-82, 1967.
- ENGLE, R. F. "Autoregressive Conditional the Variance of United Kingdom Inflation". Econometrica 50: 987-1007, 1982.
- FIX, E. and J. L. HODGES. "Discriminatory Analysis, Nonparametric Discrimination, Consistency Properties". Randolph Field, Texas, Project 21-49-004, Report N° 4, 1951.
- HINKLEY, D. V. "Jackknifing in Unbalanced Situations". Technometrics 19: 285-292, 1977.
- JOBSON, J. D. and W. A. FULLER. "Least Squares Estimation When the Covariance Matrix and Parameter Vector are Functionally Related". Journal of the American Statistical Association 78: 176-181, 1980.
- MATLOFF, N., R. ROSE and R. TAI. "A Comparison of Two Methods For Estimating Optimal Weights in Regression Analysis". Journal of Statistical Computation and Simulation 19: 265-274, 1984.
- NADARAYA, E. A. "On Estimating Regression". Theory of Probability and its Applications 9: 141-142, 1964.

- ROBINSON, P. M. "Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form". Forthcoming Econometrica, 1985.
- ROSE, R. L. Nonparametric Estimation of Weights in Least-Squares Regression Analysis. Doctoral Dissertation, University of California at Davis, 1978.
- SCHUSTER, E. and S. YAKOWITZ. "Contributions to the Theory of Nonparametric Regression, with Application to System Identification". Annals of Statistics 7: 139-149, 1979.
- STONE, C. J. "Consistent Nonparametric Regression" (with discussion). Annals of Statistics 5: 595-645, 1977.
- WATSON, G. S. "Smooth Regression Analysis". Sankhya, Series A 26: 359-372, 1964.
- WHITE, H. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity". Econometrica 48: 817-838, 1980.