Journal of Econometrics 37 (1988) 277–292. North-Holland

ESTIMATORS OF THE DISTURBANCE VARIANCE IN ECONOMETRIC MODELS
Small-sample Bias and the Existence of Moments

Jean-Marie DUFOUR*

Université de Montréal, Montréal, Canada H3C 3J7

Received August 1986, final version received June 1987

For general linear and non-linear models with stochastic regressors, we give simple exact bounds on the expected value of standard least-squares estimators of the disturbance variance. The bounds are valid for any correlation structure between the disturbances. We give simple conditions for residuals and variance estimators to have finite moments. In particular, for normal disturbances, all the moments exist. We also present analogous results for generalized least squares, simple and weighted L_p estimation, and maximum likelihood. In the latter case, we give an information inequality related to the estimation of the entropy of a distribution.

1. Introduction

In the context of the classical linear model, it is well known that the usual variance estimator $s^2 = \hat{u}'\hat{u}/(n-k)$ is unbiased while the maximum-likelihood estimator (assuming the normality of the disturbances) $\hat{\sigma}^2 = \hat{u}'\hat{u}/n$ is biased downward. On the other hand, when disturbances are correlated, both s^2 and $\hat{\sigma}^2$ are usually biased.

In this context, an important question is the direction of the bias. Several authors studied the expected value of s^2 when the disturbances are correlated. Watson (1955), Sathe and Vinod (1974) and Neudecker (1977, 1978) obtained bounds for $\mathrm{E}(s^2)$ in terms of the eigenvalues of the covariance matrix of the disturbances. Neudecker (1977, 1978) also provided numerical evaluations of these bounds for the case where the disturbances follow a first-order autoregressive [AR(1)] process. Again for AR(1) disturbances, Theil (1971, pp. 256–257) gave a simple approximation of $\mathrm{E}(s^2)$ for a model with an

*The author thanks Takeshi Amemiya, Jacques Drèze, Marc Gaudry, James MacKinnon, Michel Mouchart, Heinz Neudecker, Pierre Perron, Jean-François Richard, Jean-Marie Rolin and two anonymous referees for several useful comments. This work was supported by CORE (Université Catholique de Louvain), the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, Fonds FCAR (Government of Québec) and the Centre de Recherche et Dévelopment en Economique (Université de Montréal).

intercept and one regressor, while Johnston (1972, pp. 248–249) and Maddala (1977, pp. 282) gave similar results for a model with one regressor only; the last three also assume that the non-constant regressor is generated by an AR(1) scheme. More recently, for the case of observations with constant mean and variance, David (1985) gave a simple upper bound for $\mathrm{E}(s^2)$ which is valid irrespective of the covariance structure between the observations, while Dufour (1986) gave a similar result for linear models with fixed full-column-rank regressor matrix.

The above work generally concludes that the bias of s^2 can be positive or negative, though a negative bias appears somewhat more frequent. This also suggests that the same conclusion holds for $\hat{\sigma}^2$ since $\hat{\sigma}^2 = [(n-k)/n]s^2$. However, the bias of $\hat{\sigma}^2$ is not explicitly discussed.

All these results are limited to linear models with fixed regressors. They are based on studying the properties of the matrix $[I - X(X'X)^{-1}X']V$, where X is the matrix of regressors and V is the covariance matrix of the disturbances (up to a scale factor). Very specific assumptions on the covariance matrix of the disturbances are needed to compute bounds on $\mathrm{E}(s^2)$. For the case where the disturbances follow an AR(1) process, with autocorrelation coefficient ρ, Neudecker (1977, p. 1258) indicates: 'For higher values of ρ the interval grows rapidly, given n. Here the practical use of the method becomes limited.' Further, in non-linear models, it seems that similar methods would be difficult to apply.

In this paper, we use a different approach based on exploiting the extremal properties of least squares and other estimation methods. We derive in a simple manner general small-sample results on the bias, the existence of moments and the distribution function of variance estimators in *linear* and *non-linear* models, with *fixed* or *stochastic explanatory variables*. We present four types of results.

First, we give exact general upper bounds for the expected values of $\hat{\sigma}^2$ and s^2, which are valid whenever a model is estimated by simple least squares. In particular, when the disturbances of the model have the same variance σ^2, we find that $\mathrm{E}(s^2) \leq [n/(n-k)]\sigma^2$ or, equivalently, $\mathrm{E}(\hat{\sigma}^2) \leq \sigma^2$ in *all* cases. This holds for both linear and non-linear models, possibly with linear or non-linear restrictions on the parameters, for *any correlation structure* between the disturbances and without any assumption on the form of the distribution (except finite second moments). The model may contain arbitrary stochastic explanatory variables, such as lagged dependent variables (e.g., autoregressive models) or endogenous variables (e.g., structural equations). We show also that a related property holds for *misspecified* regression and time-series models. Even though the proof of these results is remarkably simple, the latter do not seem to be known. In several cases, the upper bounds or the approximate expressions for $\mathrm{E}(s^2)$ reported by previous authors [Johnston (1972), Maddala (1977), Neudecker (1977), Theil (1971)] are very close and may even exceed the general bounds given here, despite the fact that they rely on much more stringent assumptions.

Second, we consider the problem of the existence of finite moments for variance estimators in non-linear models with arbitrary stochastic regressors. No result seems currently available on this problem. We give a general sufficient condition which guarantees the existence of the moments of each estimated residual as well as those of variance estimators up to a given order: the condition depends only on the distribution of the disturbances, not on the form of the model or the nature of the stochastic explanatory variables.

Third, when the disturbances are normally distributed, we show that the distribution function of the sum of squared residuals is bounded downward by a chi-square distribution or by the distribution of a linear combination of independent chi-square variables. Consequently, estimated residuals and variance estimators have moments of all orders.

Fourth, we give analogous results for a number of alternative estimation methods: weighted (or generalized) least squares where parameters of the covariance matrix may be estimated jointly with the other parameters of the model, weighted and unweighted L_p estimation (including least absolute deviations) and maximum likelihood. The results on generalized least squares apply to variance estimators based on several important non-linear estimation methods in regressions with autocorrelated errors (e.g., the Cochrane–Orcutt method and various extensions of it). For the maximum-likelihood method, we give an information inequality showing that a very natural estimator of the *entropy* of a distribution always tends to underestimate the true entropy if it is based on maximum likelihood.

In section 2, we state our basic assumptions and, for the sake of mathematical convenience, start by proving general propositions on weighted and unweighted L_p estimation. Then, in section 3, we study in detail least squares methods (simple and weighted) and present the results on the estimation of the disturbance variance. Finally, in section 4, we consider estimation by maximum likelihood and give the information inequality.[1]

2. Framework and basic propositions

In the sequel, we shall consider models of the following form:

Assumption A. Let Z be a matrix of observations in a space S. Z satisfies the equation

$$q(Z,\beta) = u \tag{1}$$

[1] In this paper, we do not study the variance of regression coefficient estimators or the effect of autocorrelation on test statistics. Under the general assumptions adopted here, deriving small-sample properties is a much more difficult task. All available results on this question are either approximate or require very specific assumptions.

where β is a $k \times 1$ vector of parameters in a parameter space Ω, $q(Z, \beta) = [q_1(Z, \beta), \dots, q_n(Z, \beta)]'$ is an $n \times 1$ vector function of (Z, β) and $u = (u_1, \dots, u_n)'$ is an $n \times 1$ vector of random disturbances $(n \geq 1, k \geq 1)$.

The matrix Z includes stochastic variables but may also contain non-stochastic variables (e.g., fixed regressors). Frequently, Z has the form $Z = [y, X]$, where y is a vector of observations on a 'dependent variable' and X is a matrix of 'explanatory variables'. Clearly, models of the form

$$y = F(X, \beta) + u, \tag{2}$$

where $F(X, \beta)$ is some non-linear function of (X, β), or

$$y = X\beta + u, \tag{3}$$

satisfy Assumption A. The matrix X may be fixed or random. In particular, X may include lagged dependent variables or endogenous variables. Note also that we put no restriction on the covariance matrix of u. More generally, regression models, structural equations and various dynamic models, either linear or non-linear, usually satisfy Assumption A.

When $\operatorname{var}(u_i) = \sigma^2$, $i = 1, \dots, n$, we may wish to estimate σ^2. If $\hat{\beta}$ is an estimator of β, it is natural to estimate σ^2 with $\hat{\sigma}^2 = \hat{u}'\hat{u}/n$, where $\hat{u} = q(Z, \hat{\beta})$, or by the corresponding estimator corrected for degrees of freedom $s^2 = \hat{u}'\hat{u}/(n-k)$. Below, we give simple sufficient conditions for the existence of the moments of $\hat{u}$, $\hat{\sigma}^2$ and s^2, and derive general bounds for the expected values of $\hat{\sigma}^2$ and s^2, when β is estimated by least squares. However, before studying estimators of the disturbance variance, it will be useful to derive more general results.

Several estimation methods for β (e.g., simple least squares, weighted least squares, least absolute deviations, L_p estimation) are based on minimizing a function of $q(Z, \beta)$. Most of these objective functions can be viewed as special cases of a 'weighted L_p criterion' (defined in Proposition 1). L_p estimators, such as those obtained by minimizing the sum of absolute deviations ($p = 1$), are especially useful for robust estimation [see Judge et al. (1985, pp. 836–837)]. Further, the minimized value of the objective function typically yields an estimate of a parameter characterizing the dispersion of the disturbance distribution (e.g., the absolute moment of order p of u_i). In the following proposition, we present some general results on weighted L_p estimation: simple conditions for the existence of the moments of estimated residuals and bounds for the expected value of the moment estimator based on the minimized L_p criterion. Note that, in the proposition below as well as the rest of the paper, β refers to the true parameter value while $\tilde{\beta}$ refers to any possible value in the parameter space Ω. Further, in each proposition, we suppose without statement (whenever required) that the relevant estimators ($\hat{\beta}$, $\hat{\sigma}$ or $\hat{\gamma}_p$) and residuals ($\hat{u}$ or $\hat{v}$) are measurable functions of Z.

Proposition 1 (weighted L_p estimation). Let Assumption A hold, let ω be some subset of the parameter space Ω such that $\beta \in \omega$, let

$$v(Z, \tilde{\beta}) = P(\tilde{\beta})q(Z, \tilde{\beta}) = [v_1(Z, \tilde{\beta}), \dots, v_n(Z, \tilde{\beta})]',$$

where $P(\tilde{\beta})$ is an $n \times n$ matrix (which may depend on $\tilde{\beta}$) and suppose that the function

$$S_p(\tilde{\beta}; Z) = \sum_{i=1}^{n} |v_i(Z, \tilde{\beta})|^p, \qquad p > 0, \tag{4}$$

has a minimum over the subset ω for each $Z \in S$. Let also $\hat{\beta} = \hat{\beta}(Z)$ be an estimate of β which minimizes $S_p(\tilde{\beta}; Z)$ with respect to $\tilde{\beta} \in \omega$, $\hat{v} = v(Z, \hat{\beta}) = (\hat{v}_1, \dots, \hat{v}_n)'$ and $\hat{\gamma}_p = (\sum_{i=1}^{n} |\hat{v}_i|^p)/n$. If $\mathrm{E}(|u_i|^m) < \infty$, $i = 1, \dots, n$, where $m > 0$, then

$$\mathrm{E}(|\hat{v}_i|^r) < \infty, \qquad i = 1, \dots, n, \quad \textit{for} \quad 0 < r \leq m, \tag{5}$$

$$\mathrm{E}(\hat{\gamma}_p^r) < \infty \quad \textit{for} \quad 0 < r \leq m/p. \tag{6}$$

Further, if $\mathrm{E}(|v_i(Z, \beta)|^p) = \gamma_{pi}$, $i = 1, \dots, n$, then

$$\mathrm{E}(\hat{\gamma}_p) \leq \sum_{i=1}^{n} \gamma_{pi}/n, \tag{7}$$

and, when $\mathrm{E}(|v_i(Z, \beta)|^p) = \gamma_p$, $i = 1, \dots, n$,

$$\mathrm{E}(\hat{\gamma}_p) \leq \gamma_p. \tag{8}$$

Proof. Since $\hat{\beta}$ minimizes $S_p(\tilde{\beta}; Z)$, we have $S_p(\tilde{\beta}; Z) \leq S_p(\hat{\beta}; Z)$ for all $\tilde{\beta} \in \omega$; in particular, $S_p(\hat{\beta}; Z) \leq S_p(\beta; Z)$ where $\beta \in \omega$ is the true parameter value. Hence

$$|\hat{v}_i|^p \leq \sum_{i=1}^{n} |\hat{v}_i|^p \leq \sum_{i=1}^{n} |v_i|^p, \tag{9}$$

where $v \equiv v(Z, \beta) \equiv (v_1, \dots, v_n)'$. Using Hölder's inequality [see Mitrinovic

1970, pp. 50–54)], we see easily that

$$\left[\sum_{i=1}^{n}|v_i|\right]^r \leq C(n,r)\sum_{i=1}^{n}|v_i|^r \quad \text{for} \quad r>0, \tag{10}$$

where $C(n,r)=n^{r-1}$, if $r>1$, and $C(n,r)=1$, if $0\leq r\leq 1$. From (9) and (10), we get

$$\begin{aligned}|\hat{v}_i|^r = \left[|\hat{v}_i|^p\right]^{r/p} &\leq \left[\sum_{i=1}^{n}|v_i|^p\right]^{r/p} \\ &\leq C(n,r/p)\sum_{i=1}^{n}|v_i|^r \quad \text{for} \quad r>0.\end{aligned} \tag{11}$$

Further, since $v_i = p_i(\beta)'u$ where $p_i(\beta)' = [p_{i1}(\beta),\dots,p_{in}(\beta)]'$ is the ith row of $P(\beta)$, we have

$$\begin{aligned}|v_i|^m &\leq \left[\sum_{j=1}^{n}p_{ij}(\beta)^2\right]^{m/2}\left[\sum_{j=1}^{n}u_j^2\right]^{m/2} \\ &\leq C(n,m/2)\left[\sum_{j=1}^{n}p_{ij}(\beta)^2\right]^{m/2}\left[\sum_{j=1}^{n}|u_j|^m\right], \qquad i=1,\dots,n;\end{aligned} \tag{12}$$

hence $\mathrm{E}(|v_i|^m)<\infty$, $i=1,\dots,n$, because $\mathrm{E}(|u_j|^m)<\infty$, $j=1,\dots,n$ [see Loève (1977, p. 121)]. Thus $\mathrm{E}(|v_i|^r)<\infty$ for $0<r\leq m$, $i=1,\dots,n$, and, by (11),

$$\mathrm{E}(|\hat{v}_i|^r)<\infty, \qquad i=1,\dots,n, \quad 0<r\leq m.$$

Using (10) with v_i replaced by $\hat{v}_i$, we also have

$$\left[\sum_{i=1}^{n}|\hat{v}_i|^p\right]^{r/p} \leq C(n,r/p)\sum_{i=1}^{n}|\hat{v}_i|^r \quad \text{for} \quad r>0.$$

Hence

$$\mathrm{E}(\hat{\gamma}_p^r) = \frac{1}{n^r}\mathrm{E}\left\{\left[\sum_{i=1}^{n}|\hat{v}_i|^p\right]^r\right\} < \infty \quad \text{for} \quad 0<r\leq m/p.$$

Finally, when $\mathrm{E}(|v_i|^p) = \gamma_{pi} < \infty$, $i = 1,\dots,n$, we get from (9)

$$\mathrm{E}(\hat{\gamma}_p) \le \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} |v_i|^p\right) = \frac{1}{n}\sum_{i=1}^{n} \gamma_{pi},$$

which then reduces to (8) when $\gamma_{pi} = \gamma_p$, $i = 1,\dots,n$. □

In the above proposition, β may be estimated under a set of linear or non-linear restrictions, which are represented by the subset $\omega \subseteq \Omega$, and the estimate $\hat{\beta}$ need not be unique. We simply require that the restrictions be true ($\beta \in \omega$) and at least one value $\tilde{\beta} \in \omega$ minimize $S(\tilde{\beta}, Z)$. The result thus holds if there is exact multicollinearity (in which case several values of $\tilde{\beta}$ may minimize the objective function). The matrix $P(\tilde{\beta})$ can be a transformation matrix making corrections for heteroskedastic or autocorrelated disturbances. When $P(\tilde{\beta}) = I_n$, we get the standard (unweighted) L_p criterion. For example, with $p = 1$, this yields the sum of absolute deviations. For the standard L_p criterion, we get the following important corollary of Proposition 1.

Corollary 1.1 (L_p estimation). Let Assumption A hold, $\beta \in \omega \subseteq \Omega$, and suppose that the function $S_p(\tilde{\beta}; Z) = \sum_{i=1}^{n} |q_i(Z, \tilde{\beta})|^p$, $p > 0$, has a minimum over the subset ω, for each $Z \in S$. Let $\hat{\beta} = \hat{\beta}(Z)$ be an estimate of β obtained by minimizing $S_p(\tilde{\beta}, Z)$ with respect to $\tilde{\beta} \in \omega$, $\hat{u} = q(Z, \hat{\beta})$ and $\hat{\gamma}_p = (\sum_{i=1}^{n} |\hat{u}_i|^p)/n$. If $\mathrm{E}(|u_i|^m) < \infty$, $i = 1,\dots,n$, where $m > 0$, then

$$\mathrm{E}(|\hat{u}_i|^r) < \infty, \qquad i = 1,\dots,n, \quad \textit{for} \quad 0 < r \le m, \tag{13}$$

$$\mathrm{E}(\hat{\gamma}_p^r) < \infty \quad \textit{for} \quad 0 < r \le m/p. \tag{14}$$

Further, if $\mathrm{E}(|u_i|^p) = \gamma_{pi}$, $i = 1,\dots,n$, then

$$\mathrm{E}(\hat{\gamma}_p) \le \sum_{i=1}^{n} \gamma_{pi}/n, \tag{15}$$

and, when $\mathrm{E}(|u_i|^p) = \gamma_p$, $i = 1,\dots,n$,

$$\mathrm{E}(\hat{\gamma}_p) \le \gamma_p. \tag{16}$$

3. Least-squares estimation

Since least squares (simple or weighted) are probably the most widely used estimation method in econometrics, it is worthwhile to study the latter in detail. In this case, the objective function is a quadratic form in $q(Z, \tilde{\beta})$ and the minimized value of the objective function typically yields an estimator of the disturbance variance. We first consider weighted least squares (or generalized least squares) where the weights may be functions of model parameters.

Proposition 2 (generalized least squares). Let Assumption A hold, let ω be some subset of the parameter space such that $\beta \in \omega$, $A(\tilde{\beta}) = P(\tilde{\beta})'P(\tilde{\beta})$, where $P(\tilde{\beta})$ is an $n \times n$ matrix, and suppose that the weighted sum-of-squares function

$$S(\tilde{\beta}; Z) = q(Z, \tilde{\beta})' A(\tilde{\beta}) q(Z, \tilde{\beta})$$

has a minimum over the subset ω, for each $Z \in S$. Let $\hat{\beta} = \hat{\beta}(Z)$ be an estimate of β obtainined by minimizing $S(\tilde{\beta}, Z)$ with respect to $\tilde{\beta} \in \omega$, $\hat{u} = q(Z, \hat{\beta})$, $\hat{v} = P(\hat{\beta})\hat{u}$ and $\hat{\sigma}^2 = \hat{u}'A(\hat{\beta})\hat{u}/n$. If $\mathrm{E}(|u_i|^m) < \infty$, $i = 1, \ldots, n$, where m is a positive real number, then $\mathrm{E}(|\hat{v}_i|^r) < \infty$, $i = 1, \ldots, n$, for $0 < r \leq m$, and $\mathrm{E}(\hat{\sigma}^{2r}) < \infty$ for $0 < r \leq m/2$. If $\mathrm{E}(uu') = \sigma^2 V(\beta)$, where σ^2 and $V(\beta)$ are scaled so that $\mathrm{tr}[V(\beta)] = n$, then

$$\mathrm{E}(\hat{\sigma}^2) \leq \frac{\sigma^2}{n} \mathrm{tr}[A(\beta)V(\beta)]. \tag{17}$$

Further, when $\mathrm{V}(\hat{\beta})$ is non-singular for all $\tilde{\beta} \in \omega$ and $\mathrm{A}(\beta) = \mathrm{V}(\beta)^{-1}$,

$$\mathrm{E}(\hat{\sigma}^2) \leq \sigma^2. \tag{18}$$

Proof. The results on the finiteness of the moments $\mathrm{E}(|\hat{v}_i|^r)$ and $\mathrm{E}(\hat{\sigma}^{2r})$ follow directly from Propostion 1 with $p = 2$ and $S(\tilde{\beta}; Z) = S_2(\tilde{\beta}; Z)$. Further, since $\hat{\beta}$ minimizes $S(\tilde{\beta}; Z)$, we have

$$\hat{u}'A(\hat{\beta})\hat{u} \leq u'A(\beta)u. \tag{19}$$

Thus, when $\mathrm{E}(uu') = \sigma^2 V(\beta)$,

$$\begin{aligned} \mathrm{E}(\hat{\sigma}^2) &= \mathrm{E}\left[\hat{u}'A(\hat{\beta})\hat{u}/n\right] \\ &\leq \frac{1}{n}\mathrm{E}[u'A(\beta)u] \\ &= \frac{\sigma^2}{n}\mathrm{tr}[A(\beta)V(\beta)]. \end{aligned}$$

Finally, if we let $A(\beta) = V(\beta)^{-1}$, we find (18). □

Note here that the condition $\mathrm{tr}(V) = n$ is introduced to identify σ^2 and thus involves no loss of generality: when the variances of the disturbances are all equal, we have $\mathrm{var}(u_i) = \sigma^2$, $i = 1, \ldots, n$; when the disturbances are heteroskedastic, we have $\sigma^2 = \sum_{i=1}^n \mathrm{var}(u_i)/n$.

Usually, the functions $q(\cdot)$ and $A(\cdot)$ depend on different vectors of parameters: $q(Z, \beta) = \bar{q}(Z, \beta_1)$ and $A(\beta) = \bar{A}(\beta_2)$, where $\beta = (\beta_1', \beta_2')'$. Several important estimation methods in econometrics are based on minimizing a weighted sum of squares of the form $S(\tilde{\beta}; Z) = \bar{q}(Z, \tilde{\beta}_1)'\bar{A}(\tilde{\beta}_2)q(Z, \tilde{\beta}_1)$ by an iterative or a search method, e.g., the Cochrane–Orcutt and the iterative Prais–Winsten methods in linear regressions with AR(1) errors, the non-linear methods described by Pagan (1974) for regressions with AR(p) errors, etc.; for a general discussion, see Judge et al. (1985, ch. 8).

For example, in linear regressions with stationary AR(1) disturbances, the covariance matrix is $\mathrm{E}(uu') = \sigma^2[C(\rho)'C(\rho)]^{-1}$, where $|\rho| < 1$ and $C(\rho)$ is the $n \times n$ matrix

$$C(\rho) = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & \ldots & 0 & 0 \\ -\rho & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \ldots & -\rho & 1 \end{bmatrix}.$$

For this case, the Cochrane–Orcutt method takes

$$A(\rho) = [BC(\rho)]'[BC(\rho)], \qquad B = \begin{bmatrix} 0 & 0' \\ 0 & I_{n-1} \end{bmatrix};$$

hence, by Proposition 2,

$$\begin{aligned} \mathrm{E}(\hat{\sigma}^2) &= \mathrm{E}\big[S(\hat{\beta}; Z)/n\big] \\ &\leq \frac{\sigma^2}{n}\mathrm{tr}[A(\rho)V(\rho)] \\ &= \frac{n-1}{n}\sigma^2, \end{aligned}$$

or, equivalently, $\mathrm{E}(\hat{\sigma}_1^2) \leq \sigma^2$, where $\hat{\sigma}_1^2 = S(\hat{\beta}; Z)/(n-1)$. Similarly, if we take the iterative Prais–Winsten method [which retains the term $(1-\rho^2)u_1^2$ in the sum of squares], it is easy to see that $\mathrm{E}(\hat{\sigma}^2) \leq \sigma^2$. The same results also hold if the regression is non-linear. Note, on the other hand, that the property given by Proposition 2 does not necessarily hold if ρ (more generally, the parameter vector of the error covariance matrix) is estimated separately by an arbitrary consistent method.

If we take $P(\tilde{\beta}) = I_n$ in Proposition 2, the sum $S(\tilde{\beta}; Z)$ becomes a simple (unweighted) sum of squares. The following corollary thus follows in a straightforward way from it:

Corollary 2.1 (simple least squares). Let Assumption A hold and let ω be some subset of the parameter space Ω such that $\beta \in \omega$. Suppose that the sum-of-squares function $S(\tilde{\beta}; Z) = q(Z, \tilde{\beta})'q(Z, \tilde{\beta})$ has a minimum over the subset ω for each $Z \in S$, let $\hat{\beta} = \hat{\beta}(Z)$ be an estimate of β obtained by minimizing $S(\tilde{\beta}; Z)$ with respect to $\tilde{\beta} \in \omega$, and let $\hat{u} = q(Z, \hat{\beta})$. If $\mathrm{E}(|u_i|^m) < \infty$, $i = 1, \ldots, n$, *where m is a positive real number, then*

$$\mathrm{E}(|\hat{u}_i|^r) < \infty, \qquad i = 1, \ldots, n, \quad \textit{for} \quad 0 < r \leq m, \tag{20}$$

and the statistic $\hat{\sigma}^2 = \hat{u}'\hat{u}/n$ has finite moments $\mathrm{E}(\hat{\sigma}^{2r})$ *up to order $m/2$ $(0 < r \leq m/2)$, and similarly for the statistic $s^2 = \hat{u}'\hat{u}/(n-k)$. If* $\mathrm{E}(uu') = \sigma^2 V$, *where σ^2 and V are scaled so that* $\mathrm{tr}(V) = n$, *then*

$$0 \leq \mathrm{E}(\hat{\sigma}^2) \leq \sigma^2, \qquad 0 \leq \mathrm{E}(s^2) \leq \frac{n}{n-k}\sigma^2. \tag{21}$$

The sufficient conditions given above for the existence of the moments of $\hat{u}$ and $\hat{\sigma}^2$ do not depend on the structure of the Z matrix or the form of the $q(\cdot)$ function. It is remarkable that the upper bounds on $\mathrm{E}(\hat{\sigma}^2)$ and $\mathrm{E}(s^2)$ hold exactly for highly non-linear models, irrespective whether the model contains lagged dependent variables or endogenous explanatory variables and for any correlation structure between the disturbances. In the simple least-squares case, the estimator $\hat{\sigma}^2$ always tends to underestimate σ^2. The expected value of s^2 is bounded by $[n/(n-k)]\sigma^2$: this does not preclude an upward bias but, for k/n small, the bound is very close to σ^2. The bias of s^2, if it is positive, can never be greater than $[k/(n-k)]\sigma^2$. In absolute value, the bias of s^2 is never greater than $\sigma^2 \max\{1, [k/(n-k)]\}$.

Given the practical importance of linear models, it may be useful to state the correpsonding result for the linear case (unrestricted):

Corollary 2.2 (linear models). Let y be an $n \times 1$ vector of observations such that

$$y = X\beta + u,$$

where X is an $n \times k$ matrix and u is a vector of disturbances. Let $\hat{\beta}$ be any least-squares estimate of β, $\hat{u} = y - X\hat{\beta}$, $\hat{\sigma}^2 = \hat{u}'\hat{u}/n$ and $s_0^2 = \hat{u}'\hat{u}/[n - \mathrm{rank}(X)]$. If $\mathrm{E}(|u_i|^m) < \infty$, $i = 1, \ldots, n$, *where m is a positive real number, then* $\mathrm{E}(|\hat{u}_i|^r) < \infty$, $i = 1, \ldots, n$, *for $0 < r \leq m$, and*

$$\mathrm{E}(\hat{\sigma}^{2r}) < \infty, \quad \mathrm{E}\left(s_0^{2r}\right) < \infty \quad \textit{for} \quad 0 < r \leq m/2.$$

If $\mathrm{E}(uu') = \sigma^2 V$, *where* σ^2 *and* V *are scaled so that* $\mathrm{tr}(V) = n$, *then*

$$\mathrm{E}(\hat{\sigma}^2) \leq \sigma^2,$$

and, provided $\mathrm{rank}(X)$ *is a fixed integer with probability 1,*

$$\mathrm{E}(s_0^2) \leq \{n/[n - \mathrm{rank}(X)]\}\sigma^2.$$

To define s_0^2, we use $\mathrm{rank}(X)$ rather than k because the matrix X may not have full column rank. In this case, several values of $\hat{\beta}$ minimize the residual sum of squares but the minimal value of $(y - X\tilde{\beta})'(y - X\tilde{\beta})$ is unique. By a 'least-squares estimate' of β, we mean any of these values (each of which may be defined by using a different generalized inverse of $X'X$). There is always at least one value $\hat{\beta}$ which minimizes the sum of squares. For further discussion, see Rao (1973, ch. 4).

It is interesting to compare the bound on $\mathrm{E}(s^2)$ with other available results on $\mathrm{E}(s^2)$ under dependence. Since the upper bound $n/(n-k)$ for $\mathrm{E}(s^2)/\sigma^2$ holds for all correlation structures, one should be able to obtain tighter bounds by assuming fixed regressors and a specific process on the disturbances. Neudecker (1977, 1978) give such bounds for an AR(1) process on the errors. As expected, the upper bounds reported in table I of Neudecker (1977) are smaller than $n/(n-k)$. However, in many cases, they are surprisingly close to $n/(n-k)$: for example, for $\rho = 0.8$, $n = 10$ and $k = 3$, the upper bound reported by Neudecker is 1.375 while $n/(n-k) = 1.429$, etc. Neudecker (1977, table II) also gives approximate bounds based on Anderson (1948, eq. 39): some of these bounds actually exceed the maximum possible value for all possible correlation structures. Theil (1971, pp. 256–257) supplies an approximate formula for $\mathrm{E}(s^2)$ when the model has an intercept, one regressor generated by an AR(1) scheme (with coefficient r) and AR(1) disturbances ($k = 2$):

$$\mathrm{E}(s^2) \simeq \frac{\sigma^2}{n-2}\left[n - \frac{2}{1-\rho} - 2\rho r\right], \qquad |\rho| < 1, \quad |r| < 1.$$

When $\rho \to -1$ and $r \to +1$, this yields $\mathrm{E}(s^2)/\sigma^2 \to (n+1)/(n-2) > n/(n-2)$, e.g., with $\rho = -0.99$ and $r = 0.99$, $\mathrm{E}(s^2)/\sigma^2 \simeq (n + 0.955)/(n-2)$. Clearly, the latter is not a possible value of $\mathrm{E}(s^2)$, irrespective of the sample size.[2] Finally, the approximate expression given by Maddala (1977, p. 282) for a similar model with no intercept ($k = 1$) is always less than the upper bound $\sigma^2[n/(n-1)]$ for $|\rho| < 1$ and $|r| < 1$, but can be as close to it as one wishes (e.g., by letting $\rho \to 1$ and $r \to 1$).

[2] Note also that, as $\rho \to 1$, the same approximation yields a negative value for $\mathrm{E}(s^2)$. Theil's approximation is appropriate only when $|\rho|$ is relatively small.

Another useful implication of Corollary 2.1 deals with possibly 'misspecified' regression and time-series models:

Corollary 2.3 (misspecified regression and time-series models). *Let* $\{Z_t\colon t \in Z\}$ *be an* $m \times 1$ *second-order stationary process,* Y_t *a given component of* Z_t, *and* X_t *a* $(k-1)\times 1$ *vector made up of components of* $(Z'_{t+p}, \ldots, Z'_{t+q})'$, *where* $p \leq q$ *and* $k \geq 1$. *Consider the best linear predictor (in the mean-square sense) of* Y_t *given* X_t,

$$\hat{Y}_t \equiv P(Y_t|X_t) = \gamma_0 + X_t'\gamma,$$

and the corresponding mean-square prediction error

$$\sigma^2_{Y|X} = \mathrm{E}\left[(Y_t - \hat{Y}_t)^2\right].$$

Then, if we estimate the model

$$Y_t = \gamma_0 + X_t'\gamma + u_t, \qquad t = 1, \ldots, n,$$

by ordinary least squares (OLS), we have

$$0 \leq \mathrm{E}(\hat{\sigma}^2) \leq \sigma^2_{Y|X}, \qquad 0 \leq \mathrm{E}(s^2) \leq \frac{n}{n-k}\sigma^2_{Y|X},$$

where $\hat{\sigma}^2 = \hat{u}'\hat{u}/n$, $s^2 = \hat{u}'\hat{u}/(n-k)$ *and* $\hat{u}$ *is the vector of OLS residuals.*

Suppose that a variable Y_t follows an AR(2) stationary process but we fit an AR(1) model using n observations $Y_1, \ldots, Y_n$. Then the disturbances of the 'misspecified' model are serially correlated. Despite this complication, we can state from the latter corollary that $\hat{\sigma}^2$ tends to underestimate the mean-square prediction error of the linear projection of Y_t on Y_{t-1}.

If we assume that u has a multinormal distribution, it is possible to bound the cumulative distribution function of the residual sum of squares. The bounds are distributions of chi-square variables or linear combinations of independent chi-square variables. Consequently, all the moments of $\hat{\sigma}^2$ and s^2 are finite. In the following proposition, we prove this result for weighted least squares.

Proposition 3 (bound on the distribution of a weighted sum of squares). *Let the assumptions of Proposition 2 hold and suppose that* $u \sim \mathrm{N}_n[0, \sigma^2 V(\beta)]$, *where* $\mathrm{rank}[V(\beta)] = \nu$, $1 \leq \nu \leq n$ *and* $\sigma^2 > 0$. *Let* $\lambda_1, \ldots, \lambda_n$ *be the eigenvalues of* $V(\beta)$ *numbered so that* $\lambda_i > 0$, $i = 1, \ldots, \nu$, $\Lambda_1 = \mathrm{diag}(\lambda_1, \ldots, \lambda_\nu)$ *the matrix of the non-zero eigenvalues and* Q_1 *the matrix whose columns are corresponding eigen-*

vectors (in the same order). Then $\mathrm{E}(|\hat{v}_i|^r) < \infty$ *and* $\mathrm{E}(\hat{\sigma}^{2r}) < \infty$ *for all* $r > 0$, *and*

$$P\left[\frac{\hat{u}'A(\hat{\beta})\hat{u}}{\sigma^2} \leq x\right] \geq P\left[\sum_{i=1}^{\nu} \mu_i X_i^2 \leq x\right], \quad \textit{for all } x, \tag{23}$$

where $X_1^2, \dots, X_n^2$ *are independent random variables each following a* $\chi^2(1)$ *distribution, and* $\mu_1, \dots, \mu_\nu$ *are the eigenvalues of the matrix* $\Lambda_1^{1/2} Q_1' A(\beta) Q_1 \Lambda_1^{1/2}$.

Proof. As in the proof of Proposition 2, we have $0 \leq \hat{v}'\hat{v} = \hat{u}'\mathrm{A}(\hat{\beta})\hat{u} \leq u'A(\beta)u$. Since u is multinormal, $u'A(\beta)u$ has moments of all orders, hence, for any $r > 0$,

$$\mathrm{E}[(\hat{v}'\hat{v})^r] \equiv \mathrm{E}\{[\hat{u}'A(\hat{\beta})\hat{u}]^r\} \leq \mathrm{E}\{[u'A(\beta)u]^r\},$$

and

$$\mathrm{E}(\hat{\sigma}^{2r}) < \infty, \quad \mathrm{E}(|\hat{v}_i|^r) < \infty, \qquad i = 1, \dots, n.$$

Since $\operatorname{rank}(V) = \nu$, $1 \leq \nu \leq n$, we can write $V = Q\Lambda Q'$, where $Q = [Q_1, Q_2]$ is an orthogonal matrix and

$$\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n) = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{bmatrix};$$

when $\nu = n$, we set $\Lambda = \Lambda_1$. Let $\bar{u} = Q'u$ and $\Lambda_1^{-1/2} = \operatorname{diag}(\lambda_1^{-1/2}, \dots, \lambda_\nu^{-1/2})$. Then $\bar{u} = (w', 0')'$, where $w = (v_1, \dots, v_\nu)'$ is a $\nu \times 1$ vector such that $w \sim \mathrm{N}_\nu[0, \sigma^2 \Lambda_1]$, and

$$u'A(\beta)u = u'QQ'A(\beta)QQ'u = w'Q_1'A(\beta)Q_1 w = z'Dz,$$

where $D = \Lambda_1^{1/2} Q_1' A(\beta) Q_1 \Lambda_1^{1/2}$ and $z = \Lambda_1^{-1/2} w \sim \mathrm{N}_\nu[0, \sigma^2 I_\nu]$. We can write $D = RNR'$, where $N = \operatorname{diag}(\mu_1, \dots, \mu_\nu)$, $\mu_1, \dots, \mu_\nu$ are the eigenvalues of D and $R'R = RR' = I_\nu$. Then Rz and z have the same distribution and

$$u'A(\beta)u = z'RNR'z = \sigma^2 \sum_{i=1}^{\nu} \mu_i X_i^2,$$

where $X_1^2, \dots, X_\nu^2$ are independent $\chi^2(1)$ random variables. □

For unweighted least squares, an appreciably simpler result holds. The latter follows by considering the special case $P(\tilde{\beta}) = A(\tilde{\beta}) = I_n$ in Proposition 3.

Corollary 3.1 (*bound on the distribution of the sum of squares*). *Let the assumptions of Corollary 2.1 hold and suppose that* $u \sim \mathrm{N}_n[0, \sigma^2 V]$, *where* $\mathrm{tr}(V) = n$ *and* $\sigma^2 > 0$. *Then* $\mathrm{E}(|\hat{u}_i|^r) < \infty$, $i = 1, \dots, n$, $\mathrm{E}(\hat{\sigma}^{2r}) < \infty$ *for all* $r > 0$, *and*

$$P\left[\frac{\hat{u}'\hat{u}}{\sigma^2} \leq x\right] \geq P\left[\sum_{i=1}^{\nu} \lambda_i X_i^2 \leq x\right], \quad \textit{for all } x, \tag{24}$$

where $X_1^2, \dots, X_\nu^2$ *are independent random variables each following a* $\chi^2(1)$ *distribution, and* $\lambda_1, \dots, \lambda_\nu$ *are the non-zero eigenvalues of* V. *Further if* $V = I_n$,

$$P\left[\frac{\hat{u}'\hat{u}}{\sigma^2} \leq x\right] \geq P\left[X^2(n) \leq x\right], \quad \textit{for all} \quad x, \tag{25}$$

where $X^2(n)$ *follows a* $\chi^2(n)$ *distribution.*

Again Proposition 3 and its corollary hold exactly for both linear or non-linear models, with stochastic and non-stochastic regressors. In the simple least-squares case with $V = I_n$, it is easy to see that (25) can be used to obtain conservative one-sided confidence intervals for σ^2:

$$P\left[\sigma^2 \geq \frac{\hat{u}'\hat{u}}{\chi_\alpha^2(n)}\right] \geq 1 - \alpha,$$

where $P[X^2(n) \geq \chi_\alpha^2(n)] = \alpha$ and $0 < \alpha < 1$. Note also that Proposition 3 implies bounds on all the moments of $\hat{\sigma}^2$.

4. Maximum-likelihood estimation

All the above results apply when the model is estimated by a minimum-distance method (least squares, weighted least squares, minimum L_p). In some cases, the latter methods are equivalent to maximum likelihood (ML) but this is not generally the case. Does a result similar to (8), (18) or (21) hold when the model is estimated by the maximum-likelihood method?

Let $f(y)$ be a density function, where $y = (y_1, \dots, y_n)'$. A general measure of dispersion associated with the density f is the entropy

$$H(f|f) = -\int \ln[f(y)]\, f(y)\, \mathrm{d}y \equiv -\mathrm{E}_f[\ln f(Y)],$$

where Y is a random variable with density $f(y)$ If $g(y)$ is any function such that $\int g(y)\,\mathrm{d}y \leq 1$ and

$$H(g|f) = -\int \ln[g(y)]\, f(y)\, \mathrm{d}y = -\mathrm{E}_f[\ln g(Y)],$$

the following classical information inequality holds:

$$H(f|f) \leq H(g|f); \tag{26}$$

see Kullback (1959, pp. 14–15) and Rao (1973, sect. 1e.6).

When the observations are independent and identically distributed, i.e.,

$$f(y) = \prod_{i=1}^{n} h(y_i; \theta), \tag{27}$$

where θ is a vector of parameters and $h(x; \theta)$ is a density function, we have

$$\mathrm{E}_f[\ln f(Y)] = \mathrm{E}_f\left[\sum_{i=1}^{n} \ln[h(Y_i; \theta)]\right] = -nH(h|h). \tag{28}$$

If $\hat{\theta}$ is an estimate of θ, it is natural to use

$$\hat{H} = -\frac{1}{n}\sum_{i=1}^{n} \ln[h(y_i; \hat{\theta})] \tag{29}$$

as an estimate of the entropy $H(h|h)$. For the case where $\hat{\theta}$ is a ML estimate of θ, we now show that $\hat{H}$ always tends to underestimate H.

Proposition 4 (*information inequality*). *Let $Y = (Y_1, \dots, Y_n)'$ be a vector of observations with density function $f(y; \theta)$, where $\theta \in \Omega$ is a vector of parameters and $y \in S$. Suppose that $f(y; \theta)$ has a maximum with respect to $\theta \in \Omega$ for all $y \in S$, and let $\hat{\theta} = \hat{\theta}(Y)$ be a maximum-likelihood estimate of θ. Then*

$$\mathrm{E}\{\ln[f(Y; \hat{\theta})]\} \geq \mathrm{E}\{\ln[f(Y; \theta)]\}, \tag{30}$$

provided the relevant expectations exists, or equivalently

$$H(\hat{f}|f) \leq H(f|f), \tag{31}$$

where $f = f(y; \theta)$ and $\hat{f} = f(y; \hat{\theta})$.

Proof. Follows directly from the observation

$$\ln[f(y, \hat{\theta})] \geq \ln[f(y, \theta)] \quad \text{for all} \quad \theta \in \Omega, \quad y \in S. \quad \square$$

Though it may seem at first sight that (26) and (31) are incompatible, this is not the case because $\int \hat{f}(y)\,\mathrm{d}y \geq \int f(y)\,\mathrm{d}y = 1$. When (27) holds, we see immediately from Proposition 4 that

$$\mathrm{E}(\hat{H}) \leq H, \tag{32}$$

where H and $\hat{H}$ are defined by (28) and (29). Thus $\hat{H}$ always tends to underestimate the entropy of the probability density h. Finally, it is interesting to note that $H(\hat{f}|f) \leq H(f|f)$ is a converse of the inequality $H(f|f) \leq H(g|f)$.

References

Anderson, T.W., 1948, On the theory of testing serial correlation, Skandinavisk Aktuarietidskrift 31, 88–116.

David, H.A., 1985, Bias of S^2 under dependence, The American Statistician 39, 201.

Dufour, J.M., 1986, Bias of S^2 in linear regressions with dependent errors, The American Statistician 40, 284–285.

Johnston, J., 1972, Econometric methods, 2nd ed. (McGraw-Hill, New York).

Judge, G.G., W.E. Griffiths, R. Carter Hill, H. Lütkepohl and Tsoung-Chao Lee, 1985, The theory and practice of econometrics, 2nd ed. (Wiley, New York).

Kullback, S., 1959, Information theory and statistics (Wiley, New York).

Loève, M., 1977, Probability theory I, 4th ed. (Springer-Verlag, New York).

Maddala, G.S., 1977, Econometrics (McGraw-Hill, New York).

Mitrinovic, D.S., 1970, Analytic inequalities (Springer-Verlag, New York).

Neudecker, H., 1977, Bounds for the bias of the least squares estimator of σ^2 in the case of a first-order autoregressive process (positive autocorrelation), Econometrica 45, 1257–1262.

Neudecker, H., 1978, Bounds for the bias of the LS estimator in the case of a first-order (positive) autoregressive process when the regression contains a constant term, Econometrica 46, 1223–1226.

Pagan, A., 1974, A generalized approach to the treatment of autocorrelation, Australian Economic Papers 13, 267–280.

Rao, C.R., 1973, Linear statistical inference and its applications, 2nd ed. (Wiley, New York).

Sathe, S.T. and H.D. Vinod, 1974, Bounds on the variance of regression coefficients due to heteroskedastic or autoregressive errors, Econometrica, 42 333–341.

Theil, H., 1971, Principles of econometrics (Wiley, New York).

Watson, G.S., 1955, Serial correlation in regression analysis I, Biometrika 42, 327–341.