ETLA ELINKEINOELÄMÄN TUTKIMUSLAITOS
THE RESEARCH INSTITUTE OF THE FINNISH ECONOMY
Lönnrotinkatu 4 B, 00120 Helsinki 12, Finland, tel. 601322

Keskusteluaiheita
Discussion papers

TIMO TERÄSVIRTA

PRIOR INFORMATION AND BIASED ESTIMATION IN LINEAR MODELS*

No. 69 17.11.1980

* This is a slightly revised version of a paper prepared for Ekonometridagen in Stockholm, May 6, 1980. I am grateful to Ilkka Mellin for his comments on the earlier version.

Abstract

The paper contains a general but rather brief discussion of biased estimators for linear models based on prior information. It stresses the unfavourable consequences of model misspecification to some favourable properties of certain widely discussed biased estimators.

Keywords. biased estimation; mixed estimator; minimax estimator; model misspecification; restricted least squares; ridge estimator.

1. INTRODUCTION

During the last few years the biased estimation of linear models has received much attention in the statistical literature. It has been pointed out that by rejecting the notion of unbiasedness it is sometimes possible to reduce the mean square error in the estimation of these models as compared to the least squares (maximum likelihood) estimation. It has also been shown that under certain loss criteria the maximum likelihood estimator is inadmissible and an estimator which dominates the maximum likelihood estimator (the James-Stein estimator) has been constructed.

In econometrics biased estimation has been a commonplace technique for a long time. A widely applied feature have been the linear restrictions between parameters. Although the restrictions have been imposed as exact, they may not have been valid: in this case the restricted least squares estimators are biased. A good example of such restrictions are the polynomial lag restrictions (Almon, 1965) assuming that the regression coefficients lie on a polynomial of low degree. These restrictions which are very rarely exactly true when they are non-trivial can be written as a set of linear restrictions as demonstrated by Teräsvirta (1970) and Shiller (1973).

Another example of biased restrictions are the zero restrictions: a variable is excluded from the model although its regression coefficient does not equal zero.

Theil and Goldberger (1961) introduced the use of stochastic linear restrictions. They were originally considered as unbiased but it is fully possibly to assume that they are biased and extend the notion of biased estimation to include also the so-called mixed estimation.

These restrictions, deterministic or stochastic, are often introduced because of the multicollinearity among the expalanatory variables which leads to low estimation accuracy. Very often they reflect the personal views of the model builder and can be considered,at least implicitly, to be prior information. On the other hand, of course, the Bayesian approach offers a useful strategy for combining prior views and sample information. However, later on we shall be largely interested in the circumstances under which the least squares estimation can be improved (risk in estimation reduced) by the use of possibly biased prior information. Since this problem is not relevant to a Bayesian, the classical framework is adopted for purposes of this paper.

A wide area of biased estimation in econometrics is the estimation of the structural parameters of linear simultaneous equation models. Those estimators will not be considered here, however. For a discussion on comparisons between various structural form estimators the reader is referred to Theil (1971).

2. BIASED ESTIMATION WITHOUT PRIOR INFORMÀTION

If there is multicollinearity present in the linear model

$$y = X\beta + \varepsilon, \; E\varepsilon = 0, \; \mathrm{cov}(\varepsilon) = \sigma^2 I \tag{2.1}$$

where y and ε are stochastic $n \times 1$ vectors, X is an $n \times p$ matrix with rank p and independent of ε, β is the $p \times 1$ parameter vector and the number of observations is fairly low, the least squares estimator $b = UX'y$ where $U = (X'X)^{-1}$ may not be very reliable. The estimates may have "wrong signs" and be too large in absolute value and thus unacceptable to the model builder in the light of his (implicit) prior information. To improve the estimation, Hoerl and Kennard (1970) suggested the ridge estimator

$$b_I^*(k) = (X'X + kI)^{-1}X'y, \; k > 0. \tag{2.2}$$

When k grows from zero to infinity the values of (2.2) shrink toward zero. As a theoretical fact speaking for (2.2), Hoerl and Kennard pointed out that for any model of type (2.1) there always exists a $k > 0$ such that $\mathrm{MSE}(b_I^*(k)) < \mathrm{MSE}(b)$. This property has nothing to do with possible multicollinearity. Whenever $\beta \neq 0$, the ridge estimator is biased.

Stein (1956) showed that in estimating the mean of the $p \times 1$ vector $y \sim N(\mu, I)$, where $p \geqq 3$, the unbiased estimator is inadmissible if the quadratic loss function $q = (\hat{\mu} - \mu)'(\hat{\mu} - \mu)$ is used. Instead, the estimator

$$\hat{\mu}_{JS} = (1 - \alpha/y'y)y \tag{2.3}$$

where $0 < \alpha < 2(p-2)$, dominates the least squares estimator under quadratic loss as shown by James and Stein (1961). This basic result has been extended to the linear models (2.1) when $\varepsilon \sim N(0,\sigma^2 I)$, see Sclove (1968). The estimator

$$b_{JS} = (1 - \frac{(p-2)\nu s^2}{(\nu + 2)b'X'Xb})\, b$$

where s^2 is the usual estimator of σ^2 with ν degrees of freedom, dominates b for $p \geqq 3$ under the quadratic risk

$$R(\tilde{b}, \beta, X'X) = E(\tilde{b} - \beta)'X'X(\tilde{b} - \beta). \tag{2.4}$$

The James-Stein estimators have been extended and generalized in many ways, cf. e.g. Draper and Van Nostrand (1979), Judge and Bock (1978) and Zellner and Vandaele (1975) for references. Among the few econometric applications of these estimators a study by Aigner and Judge (1977), and Judge and Bock (1978, Ch. 13) can be mentioned.

A problem with the ridge estimator is that the values of k for which $\mathrm{MSE}(b_I^*(k)) < \mathrm{MSE}(b)$ are unknown, and therefore in practice k must somehow be estimated. A vast amount of estimators for k have been suggested in the literature and the properties of the corresponding ridge estimators have been investigated by simulation experiments; for references and general discussion see Vinod (1978), Draper and Van Nostrand (1979) and Smith and Campbell (1980). Some analytical results of interest in this context can be found in Teräsvirta (1981).

3. LINEAR PRIOR INFORMATION AND BIASED ESTIMATION

Even if some Bayesians have objected, see e.g. Kiefer and Richard (1979), the ridge estimator (2.2) has sometimes been given a Bayesian interpretation. In fact, (2.2) is the posterior mean of β if normality of errors in (2.1) and a normal-gamma prior distribution for (β,σ^2) are assumed so that $\beta|\sigma^2 \sim N(0,(\sigma^2/k)I)$ and $\sigma^{-2} \sim \Gamma(\delta^2,0)$. From the prior information point of view the interpretation is illuminating since it shows the fact, also stressed by Kiefer and Richard, that the "prior" seldom reflects the actual prior information possessed by the researcher. He should thus be able to choose an estimator, based on his prior knowledge, which outperforms the ridge estimator.

One popular alternative in economic applications have been the linear restrictions

$$r_0 = R\beta \tag{3.1}$$

where R is an $m \times p$ matrix with rank m. These restrictions may not fully reflect the researcher's prior information either but may come closer than the ridge-type restrictions. The restricted least squares estimator of β is

$$b_R = b - UR'(RUR')^{-1}(r_0 - Rb). \tag{3.2}$$

If the restrictions (3.1) hold exactly, b_R is unbiased and has smaller mean square error than b. Generalize the concept of quadratic risk (2.4) of $\tilde{b}$ by defining

$$R(\tilde{b},\beta,A) = E(\tilde{b}-\beta)'A(\tilde{b}-\beta) \qquad (3.3)$$

where $A \geqq 0$ (non-negative definite). Then a necessary and sufficient condition for

$$R(b,\beta,A) \geqq R(b_R,\beta,A)$$

for all $A \geqq 0$ is (Toro-Vizcarrondo and Wallace, 1968)

$$q(s_0) = (1/\sigma^2)s_0'(RUR')^{-1}s_0 \leqq 1 \qquad (3.4)$$

where $s_0 = r_0 - R\beta$. Inequality (3.4) shows that even if (3.1) is not valid so that (3.2) is biased, the restricted least squares estimator can be superior to the least squares estimator. If (2.4) is used for the comparison, the corresponding necessary and sufficient condition for the superiority of b_R over b is $q(s_0) \leqq m$, see Wallace (1972). There is no general result implying the existence of such a positive k that $MSE(b_I^*(k)) \leqq MSE(b_R)$. Note, however, that <u>if</u> (3.1) is valid, then the estimator

$$b_R^* = b_I^*(k) - UR'(RUR')^{-1}(r_0 - Rb_I^*(k))$$

has smaller risk than b_R for some $k > 0$, see Teräsvirta (1981).

The above restrictions are deterministic although the prior information of the researcher is often more vague. One way of taking this into account is testing the validity of the restrictions first and using b_R only if the restrictions are accepted. This reasoning leads to the pre-test estimator, cf. e.g. Judge and Bock (1978), which can be written as

$$
\begin{aligned}
t_\alpha(b_R,b) &= I_\alpha(Q)b + [1 - I_\alpha(Q)]b_R \\
&= b - [1 - I_\alpha(Q)]UR'(RUR')^{-1}(r_0 - Rb). \qquad (3.5)
\end{aligned}
$$

Indicator function $I_\alpha(Q)$ has value one if the null hypothesis (3.1) is rejected at a given significance level α in favour of $r_0 \neq R\beta$ and zero otherwise. Estimator (3.5) is biased unless $s_0 = 0$, and from Cohen (1965) it can be inferred that if the risk criterion (2.4) is used, it is inadmissible as well. Judge and Bock (1978) contains a good discussion on the performance of (3.5) vis-à-vis the least squares estimator.

Another way of allowing for uncertain prior information in the classical framework consists of making (3.1) stochastic by "observing" r and R which are related as

$$r = R\beta + \varphi \qquad (3.6)$$

where r and φ are stochastic $m \times 1$ vectors, $Er|R\beta = r_0$, $E\varphi = 0$, $\text{cov}(\varphi) = (\sigma^2/k)I$ and $\text{cov}(\varepsilon,\varphi) = 0$. This additional information is then incorporated into the sample and the resulting mixed estimator is (Theil and Goldberger, 1961)

$$b_R(k) = (X'X + kR'R)^{-1}(X'y + kR'r). \qquad (3.7)$$

Estimator (3.7) is biased unless $r_0 = R\beta$. It has smaller risk than b for all $A \geqq 0$ if and only if

$$(1/\sigma^2)s_0'S_k s_0 \leqq 1 \qquad (3.8)$$

where $S_k = (k^{-1}I + RUR')^{-1}$. Note that the ridge estimator is not exactly a special case of (3.7) with $R = I$ and $r_0 = 0$ since r is stochastic and its components have positive variances. However, if r is observed as zero and $R = I$, (2.2) and (3.7) yield identical estimates.

The left-hand side of (3.8) is a monotonously increasing function of k and approaches zero as $k \rightarrow 0+$. When $k \rightarrow \infty$, it converges towards (3.4). Thus for given X, R, r_0, β and σ^2 we have the theoretical result that there is always such a $k > 0$ that (3.8) holds, cf. Teräsvirta (1981), but in practice nothing is known about the range of the values of k satisfying (3.8) and the size of the maximal reduction in risk.

Condition (3.8) has no direct practical significance because it depends on unknown parameters β and σ^2. However, (3.8) is a testable proposition. Under H_0: $(1/2\sigma^2)s_0'S_k s_0 \leqq 1/2$ the compatibility test statistic

$$Q_k = \hat{s}_0'S_k\hat{s}_0/m\hat{\sigma}^2 \qquad (3.9)$$

where $\hat{s} = r - Rb$ and $\hat{\sigma}^2 = (n-p)^{-1}y'(I - XUX')y$ is the unbiased estimator of σ^2, follows a non-central $F(m,n-p,1/2)$ distribution, see Yancey _et al_. (1974) and Bock and Judge (1978), and H_0 is rejected if (3.9) exceeds the critical value.

If k were known or if we assumed $k = \infty$, we could form another pre-test estimator on the basis of the result of the test. The difference is that here the test is not for the unbiasedness of (3.6) (or truth of (3.1)) but rather for the hypothesis that the risk of estimation is reduced through the use of the mixed estimator $b_R(k)$ (or restricted estimator b_R) instead of b.

A similar procedure can be developed for the case A = X'X. However, for $k < \infty$, we can only test the validity of a sufficient condition for the risk reduction, cf. Teräsvirta (1980a), whereas $q(s_0) \leqq m$ is a necessary and sufficient condition for b_R to be superior to b.

4. NON-LINEAR PRIOR INFORMATION

The prior distribution discussed in the preceding section is linear. It is perhaps less known that the sampling-theoretic framework is also suitable for handling non-linear prior information. Assume that

$$\beta' R' R\beta \leqq \sigma^2/k \tag{4.1}$$

i.e. the regression coefficients are located within or on an ellipsoid in the subspace of the coefficient space. The ellipsoid need not be centred in the origin of the subspace but any other point will do as well. If we use $A = aa'$ where $a \neq 0$ is a p x 1 vector, minimizing the supremum of the quadratic risk under (4.1) yields the minimax estimator

$$b_R^*(k) = (X'X + kR'R)^{-1}X'y. \tag{4.2}$$

see for instance Kuks and Olman (1972), Bibby and Toutenburg (1977) and Peele and Ryan (1980). This biased estimator is identical to the ridge estimator when R = I. In this case the ellipsoid in (4.1) becomes a sphere. It has smallest minimax risk for all $a \neq 0$ if (4.1) is valid, cf. Bunke (1975).

In practice, if we want to apply (4.2) in a situation where information of type $\beta' R' R\beta \leqq c$ exists, the unknown variance σ^2 has to be estimated so that k can be determined. An interesting suggestion has been made by Toutenburg and Roeder (1978) who consider the case where the researcher has non-linear information of type $\underline{a}_j \leqq \beta_j \leqq \bar{a}_j$, j = 1,...,p. There is no closed form estimator based upon these inequality constraints although the estimates can be found using non-linear programming methods. The authors suggest that the restrictions be approximated by the smallest ellipsoid containing the cuboid defined by the above inequalities, whereafter (4.2) can be applied using an estimate of σ^2.

Even here the prior information can be incorrect and the question is how this affects the optimality properties of (4.2). The model builder may have thought that (4.1) holds and used (4.2). Later on, he may obtain more accurate information indicating that, in fact,

$$(\beta - \beta_0)'R'R(\beta - \beta_0) \leqq \sigma^2/d. \qquad (4.3)$$

It is obvious that if we choose c small enough in (4.1) so that (4.1) contains (4.3), then (4.1) is also correct and the estimator (4.2) has smaller minimax risk than b. But then, if c is small, the minimax estimator is already close to the least squares estimator. We do not elaborate further here but refer to Teräsvirta (1980b) for a more detailed treatment of the properties of (4.2) when (4.3) is valid.

5. LINEAR RESTRICTIONS AND MODEL MISSPECIFICATION

All the above considerations have been based upon correctly specified models. It will be demonstrated in this section that the misspecification of models deprives the mixed estimators of the theoretical properties discussed above and sometimes used as arguments in favour of them. Write (2.1) as

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon \qquad (5.1)$$

where X_j is an $n \times p_j$ matrix, $j = 1,2$; $p_1 + p_2 = p$. Suppose that the researcher incorrectly sets $\beta_2 = 0$ and estimates

$$y = X_1\beta_1 + \varepsilon^{(1)}$$

subject to biased stochastic prior information

$$r_1 = R_1\beta_1 + \varphi_1$$

where $Er_1|R_1\beta_1 = r_1^0$, $E\varphi_1 = 0$ and $\text{cov}(\varphi_1) = (\sigma^2/k_1)I$.

It can be shown (Teräsvirta, 1981) that the mixed estimator

$$b_R(k_1) = (X_1'X_1 + k_1R_1'R_1)^{-1}(X_1'y + k_1R_1'r_1) \qquad (5.2)$$

is superior to b, the least squares estimator of the correctly specified model, for all $A \geqq 0$ if and only if

$$q_{12}(s_1) = (1/\sigma^2)[\beta_2'G^{-1}\beta_2 + (s_1 - R_1U_1X_1'X_2\beta_2)'S_1(s_1 - R_1U_1X_1'X_2\beta_2)] \leqq 1 \qquad (5.3)$$

where

$$G = (X_2'X_2 - X_2'X_1U_1X_1'X_2)^{-1},\ U_1 = (X_1'X_1)^{-1} \text{ and } s_1 = r_1^0 - R_1\beta_1.$$

A necessary condition for (5.3) to hold is that

$$(1/\sigma^2)\beta_2'G^{-1}\beta_2 \leqq 1 \qquad (5.4)$$

which is a necessary and sufficient condition for the risk of the least squares estimator of the misspecified model to be smaller than that of b in the estimation of $\beta = (\beta_1', \beta_2')$. There is a $k_1 > 0$ such that (5.3) holds only if (5.4) is valid.

If we apply the weaker superiority condition with the risk (2.4), our superiority condition is $q_{12}(s_1) \leqq (m_1 + p_2)$. The condition corresponding to (5.4),

$$(1/\sigma^2)\beta_2'G^{-1}\beta_2 \leqq p_2 \qquad (5.5)$$

is no longer necessary for $q_{12}(s_1) \leqq m_1 + p_2$.

Similar conditions could be obtained for the superiority of ridge estimators over b. Thus the argument of Hoerl and Kennard (1970) for the ridge estimator mentioned above does not extend to the estimation of parameters in misspecified linear models.

The above discussion shows that the potential benefits of prior information heavily depend on the specification of the model. A correct specification and ordinary least squares without restrictions may often be a better combination than an incorrent specification with biased stochastic or deterministic prior restrictions. The message thus is that careful specification of the model remains a most crucial thing and that more or less artificial prior restrictions should not be used as a substitute for the specification effort.

REFERENCES

Aigner, D.J. and G.G. Judge (1977). Application of pre-test and Stein estimators to economic data. *Econometrica 45*, 1279-1288.

Almon, S. (1965). The distributed lag between capital appropriation and expenditures. *Econometrica 33,* 178-196.

Bibby, J. and H. Toutenburg (1977). *Prediction and improved estimation in linear models.* New York: John Wiley.

Bunke, O. (1975). Minimax linear, ridge and shrunken estimators for linear parameters. *Mathematische Operationsforschung und Statistik 6,* 697-701.

Cohen, A. (1965). Estimates of linear combination of the parameters in the mean vector of a multivariate distribution. *Annals of Mathematical Statistics 36,* 78-87.

Draper, N.R. and R.C. Van Nostrand (1979). Ridge regression and James-Stein estimation: Review and comments. *Technometrics 21,* 451-465.

Hoerl, A.E. and R.W. Kennard (1970). Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics 12,* 55-67.

James, W. and C. Stein (1961). Estimation with quadratic loss, 361-379 in J. Neyman (ed.): *Proceedings of the Fourth Berkeley Symposium, Vol. 1.* Berkeley and Los Angeles: University of California Press.

Judge, G.G. and M.E. Bock (1978). *The statistical implications of pre-test and Stein-rule estimators in econometrics.* Amsterdam: North-Holland.

Kiefer, N.M. and J.-F. Richard (1979). A Bayesian approach to hypothesis testing and evaluating estimation strategies. Université Catholique de Louvain, C.O.R.E. Discussion Paper No. 7927.

Kuks, J. and V. Olman (1972). Minimaksnaja linejnaja ocenka koefficientov regressii. *Eesti NSV teaduste akadeemia toimetised 21,* 66-72.

Peele, L. and T.P. Ryan (1980). Minimax linear regression estimators with application to ridge regression. *Technometrics* (forthcoming).

Sclove, S.L. (1968). Improved estimators for coefficients in linear regression. *Journal of the American Statistical Association 63,* 596-606.

Shiller, R.J. (1973). A distributed lag estimator derived from smoothness priors. *Econometrica 41,* 775-788.

Smith, G. and F. Campbell (1980). A critique of some ridge regression methods. *Journal of the American Statistical Association 75*, 74-103 (with discussion).

Stein, C. (1956). Inadmissibility of the usual estimation for the mean of a multivariate normal distribution, 197-206 in J. Neyman (ed.): *Proceedings of the Third Berkeley Symposium.* Berkeley and Los Angeles: University of California Press.

Teräsvirta, T. (1970). *On stepwise regression and economic forecasting.* Helsinki: Kansantaloudellinen yhdistys.

Teräsvirta, T. (1980a). Linear restrictions in misspecified linear models and polynomial distributed lag estimation. University of Helsinki, Department of Statistics, Research Report No. 16.

Teräsvirta, T. (1980b). The minimax estimation of linear models under incorrect prior restrictions. University of Helsinki, Department of Statistics, Research Report No. 15.

Teräsvirta, T. (1981). Some results on improving the least squares estimation of linear models by mixed estimation. *Scandinavian Journal of Statistics 8*, 33-38.

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

Theil, H. and A.S. Goldberger (1961). On pure and mixed statistical estimation in economics. *International Economic Review 2,* 65-78.

Toro-Vizcarrondo, C. and T.D. Wallace (1968). A test of the mean square error criterion for restrictions in linear regression. *Journal of the American Statistical Association 63,* 558-572.

Toutenburg, H. and B. Roeder (1978). Minimax-linear and Theil estimator for restrained regression coefficients. *Mathematische Operationsforschung und Statistik, Series Statistics, 9,* 499-505.

Vinod, H.D. (1978). A survey of ridge regression and related techniques for improvements over ordinary least squares. *Review of Economics and Statistics 60,* 121-131.

Wallace, T.D. (1972). Weaker criteria and tests for linear restrictions in regression. *Econometrica 40,* 689-698.

Yancey, T.A., G.G. Judge and M.E. Bock (1974). A mean square error test when stochastic restrictions are used in regression. *Communications in Statistics 3,* 755-768.

Zellner, A. and W. Vandaele (1975). Bayes-Stein estimators for *k*-means, regression and simultaneous equation models, 627-653 in S.E. Fienberg and A. Zellner (eds.): *Studies in Bayesian econometrics and statistics in honor of Leonard J. Savage.* Amsterdam: North-Holland.