Erasmus University Rotterdam
ECONOMETRIC INSTITUTE
Nederlandse School of economics

Report 7323 ES

NOTE ON A LARGE-SAMPLE RESULT IN SPECIFICATION ANALYSIS

by T. Kloek

November 1973

NOTE ON A LARGE-SAMPLE RESULT IN SPECIFICATION ANALYSIS

by T. Kloek[1]

This note deals with specification analysis in the linear model, as introduced by Theil [5, Section 6.2.4]. If two linear models have different sents of explanatory variables and the same variable to be explained, the residual variance of the correct model has a smaller mean value than that of the incorrect one. The interest of this result is moderate as, in particular when the sample size is small, the analyst who adopts the decision rule of choosing the model with the smaller residual variance, may make wrong decisions with a substantial probability. This was emphasized by Koerts and Abrahamse [3], who computed some probability distribution functions for squared correlation coefficients. Their numerical examples were based on 15 observations and 3 to 5 explanatory variables, so that their distributions showed substantial dispersion. In the present note it is shown that the probability of adopting the "wrong" model on the basis of the decision rule mentioned above converges to zero as the sample size increases. This result seems to be intuitively obvious, but the proof is not trivial.

More precisely, we consider a linear model with n observations and k explanatory variables, which is formally described by $y_n = X_n\beta + \varepsilon_n$. It is assumed that the disturbances are independently and identically distributed with zero mean and variance σ^2; also that the $n \times k$ matrix X_n is nonstochastic. Let B_n be any[2] $k \times n$ matrix such that $X_nB_nX_n = X_n$ and X_nB_n is symmetric, then X_nB_n is idempotent and its trace equals its rank. We define the residual variance by

$$s_n^2 = y_n'(I_n - X_nB_n)y_n/q \tag{1}$$

where $q = n - \text{rank}\,(X_n)$.

---

[1] The author is indebted to dr. R. Harkema for some useful comments.

[2] Note that the Moore-Penrose generalized inverse satisfies these conditions, so that the existence of B_n is guaranteed. If X_n has rank k it can be shown that B_n is uniquely defined and given by $B_n = (X_n'X_n)^{-1}X_n'$.

Suppose that $y_n = X_n\beta + \varepsilon_n$ is the correctly specified model and that the alternative set of regressors is represented by an $n \times h$ matrix Z_n which satisfies the condition that the space spanned by its columns does not contain $X_n\beta$. Otherwise, a vector γ would exist such that $Z_n\gamma = X_n\beta$ so that both specifications would correctly describe $E(y_n)$. Hence, a specification which contains all the variables of the correct specification plus some additional irrelevant variables is not incorrect in our sense, though it may produce an inefficient estimator for β. Let C_n be an $h \times n$ matrix such that $Z_nC_nZ_n = Z_n$ and Z_nC_n is symmetric. Then the residual variance of the "wrong" model is given by

$$t_n^2 = y_n'(I_n - Z_nC_n)y_n/m \tag{2}$$

where $m = n - \text{rank}\,(Z_n)$. Substituting $y_n = X_n\beta + \varepsilon_n$ we obtain

$$t_n^2 = \beta'X_n'H_nX_n\beta/m + 2\beta'X_n'H_n\varepsilon_n/m + \varepsilon_n'H_n\varepsilon_n/m \tag{3}$$

where H_n is the projection matrix

$$H_n = I_n - Z_nC_n \tag{4}$$

which satisfies $H_n' = H_n$, $H_n^2 = H_n$, $H_nZ_n = 0$. Note that the first right-hand term of (3) can be interpreted as the residual variance of the hypothetical regression of $X_n\beta$ on Z_n, which is positive by our condition on Z_n. In view of our large-sample theorem we modify this condition by assuming a positive lower bound

$$\theta_n^2 \equiv \beta'X_n'H_nX_n\beta/m > g^2 \tag{5}$$

for a certain number $g > 0$ and from a certain value of n onward. So we exclude the possibility that θ_n^2 would converge to zero, which would mean that the specification error would vanish in the long run. Obviously, the sequence $\{\theta_n\}_{n=1}^{\infty}$ may either converge to a certain limit θ^2, say, or diverge. In the former, more restrictive, case our main result can be proved in a very simple way; see (8).

For further reference we recall that under the conditions mentioned above

(6) $$\operatorname{plim}_{n\to\infty} s_n^2 = \sigma^2$$

as was shown[3] in [3] for the case rank $(X_n) = k$ and in [1] for the more general case rank $(X_n) \leq k$. Along the same lines one can show that

(7) $$\operatorname{plim}_{n\to\infty} \varepsilon_n' H_n \varepsilon_n / m = \sigma^2$$

Now we can state and prove the following simple result . If $\lim_{n\to\infty} \theta_n^2 = \theta^2 \geq g^2$, then

(8) $$\operatorname{plim}_{n\to\infty} t_n^2 = \sigma^2 + \theta^2$$

Proof: The first right-hand term in (3) converges to θ^2 by assumption, the second has a second moment $4\theta_n^2\sigma^2/m$ so that it converges to zero in the squared mean, and for the third term we refer to (7).

If the sequence $\{\theta_n\}_{n=1}^{\infty}$ diverges, the second term in (3) may become substantially negative. So we have to prove that it is amply compensated for by the first term. Our main result is general enough to include both cases of convergence and divergence. It reads as follows: under the assumption in (5)

(9) $$\lim_{n\to\infty} P[t_n^2 < s_n^2 + \lambda g^2] = 0$$

for every λ satisfying $0 < \lambda < 1$. The proof is based on the following elementary theorem in probability theory. Let $U_1, U_2, \ldots, U_k$ be arbitrary real-valued random variables, then

(10) $$P[\sum_{i=1}^{k} U_i < 0] \leq \sum_{i=1}^{k} P[U_i < 0]$$

As some elementary exerxises will show, the inequality in (9) can be written as

(11) $$t_n^2 - \lambda g^2 - s_n^2 = \sum_{i=1}^{4} U_{in} < 0$$

with[4]

---

[3] Here, the traditional assumption that the disturbances are i.i.d. plays a role. Alternative assumptions are possible; see Révész [4], Sections 3.2, 4.2, and 6.1. I am indebted to dr. L. de Haan for drawing my attention to this reference.

[4] Note that the sequence U_{1n} $(n = 1, 2, \ldots)$ is a sequence of non-stochastic real numbers. The statement $P[U_{1n} < 0] = 0$ is trivially true if $U_{1n} > 0$.

(12a) $$U_{1n} = (\theta_n + \phi_1 g)(\theta_n - \phi_2 g) - 2\eta \qquad (\theta_n > 0)$$

(12b) $$U_{2n} = \zeta_n + (\phi_2 - \phi_1) g\theta_n$$

(12c) $$U_{3n} = \varepsilon_n' H_n \varepsilon_n / m - \sigma^2 + \eta$$

(12d) $$U_{4n} = - s_n^2 + \sigma^2 + \eta$$

(13) $$\zeta_n = 2\beta' X_n' H_n \varepsilon_n / m$$

Note that use has been made of (3) and (5). The numbers η, ϕ_1, and ϕ_2 may be chosen at will provided $0 < \phi_1\phi_2 = \lambda < 1$, but in the context of our proof we shall restrict them by

(14) $$0 < \phi_1 < \phi_2 < 1, \quad 0 < \eta < \tfrac{1}{2} g^2 (1 + \phi_1)(1 - \phi_2)$$

Now, in order to prove (9) it is - according to (10) - sufficient to prove that

(15) $$\lim_{n \to \infty} P[U_{in} < 0] = 0 \qquad i = 1, \ldots, 4$$

We shall prove these four statements in the reverse order. First, for $i = 4$ and $i = 3$ (15) immediately follows from (6) and (7), respectively. Second, for $i = 2$ use can be made of the following variant of Chebyshev's inequality

(16) $$P[|\zeta_n| > \psi_n (\operatorname{var} \zeta_n)^{\frac{1}{2}}] < \psi_n^{-2}$$

Note that $E\zeta_n = 0$ and $\operatorname{var} \zeta_n = 4\theta_n^2 \sigma^2 / m$ so that we may take $\psi_n = (\phi_2 - \phi_1) g m^{\frac{1}{2}} / 2\sigma$ which implies that ψ_n^{-2} tends to zero. Third, since the U_{1n} are fixed numbers (15) reduces to $U_{in} > 0$ for $i = 1$. To prove this is an elementary exercise which makes use of $\theta_n > g$ and the conditions in (14).

REFERENCES

[1] H. Drygas: "A Note on a Paper by T. Kloek Concerning the Consistency of Variance Estimation in the Linear Model", University of Bonn (1973), mimeographed.

[2] T. Kloek: "Note on Consistent Estimation of the Variance of the Disturbance in the Linear Model", Econometrica, Vol.40 (1972), 911-912.

[3] J. Koerts and A.P.J. Abrahamse: "The Correlation Coefficient in the General Linear Model", European Economic Review, Vol. 1 (1970), 401-427.

[4] P. Révész: The Laws of Large Numbers, Academic Press, New York and London, 1968.

[5] H. Theil: Economic Forecasts and Policy, North-Holland Publishing Company, Amsterdam (1958).

APPENDIX

This Appendix gives some details of proofs, which were left as an exercise to the reader in the main text.

Note that (16) implies

$$P[-\zeta_n > \psi_n(\text{var } \zeta_n)^{\frac{1}{2}}] < \psi_n^{-2} \tag{A.1}$$

or

$$P[\zeta_n + \psi_n(\text{var } \zeta_n)^{\frac{1}{2}} < 0] < \psi_n^{-2} \tag{A.2}$$

so that we only need to check that

$$(\phi_2 - \phi_1)g\theta_n = \psi_n(\text{var } \zeta_n)^{\frac{1}{2}} \tag{A.3}$$

compare (12b). This follows from substitution of the expressions given below (16).

Next we consider $U_{1n} > 0$; see (12a). As a starting point we use the inequalities $\theta_n > g > 0$, $0 < \phi_1 < \phi_2 < 1$. Then we obtain

$$\theta_n + \phi_1 g > g(1 + \phi_1)$$

$$\theta_n - \phi_2 g > g(1 - \phi_2)$$

If we combine this with

$$-2\eta > -g^2(1 + \phi_1)(1 - \phi_2)$$

we obtain the desired result.