Walter Jahn; Marko Riedel

Reduction of the dimension in the linear model with stochastic regressors

Commentationes Mathematicae Universitatis Carolinae, Vol. 25 (1984), No. 4, 747--761

Persistent URL: http://dml.cz/dmlcz/106340

COMMENTATIONES MATHEMATICAE UNIVERSITATIS CAROLINAE

25,4 (1984)

REDUCTION OF THE DIMENSION IN THE LINEAR MODEL WITH STOCHASTIC REGRESSORS

W. JAHN and M. RIEDEL

ABSTRACT: First of all we introduce the linear model with stochastic regressors. The estimates of the parameter $\mathfrak{B}$ and $\sigma^2_{Y|X}$ of this model are influenced by multicollinearity. As one of the possibilities to reduce the degree of multicollinearity subset regression is proposed. As a criteria for the selection of a model for the best extrapolation we use the mean square error of extrapolation. Some important properties of the estimates of the selected model will be shown.

KEY WORDS: Linear model with stochastic regressors, multicollinearity, mean square error of extrapolation, subset regression.

AMS: 62 J 99

1. INTRODUCTION

First of all we will give a short introduction to the linear model with stochastic regressors. It will be shown that the estimates for the parameters of this model such as the vector of regression coefficients and the conditional variance possess the usual properties as unbiasedness and consistency. In this model the multicollinearity plays an important role. Its effect on the estimates is also shortly demonstrated and by an example illustrated. To correct the estimates from this effect it is necessary to reduce the degree of multicollinearity. One of the possibilities for this is the subset regression which can be considered as a kind of the reduction of the dimension of the parameter space. As a criteria for the selection of a model for the best extrapolation of the regressand by all or a subset of the regressands we use the mean square error of extrapolation which will be stated in theorem 3. For all selections k we show that $\|\mathfrak{B} - \hat{\mathfrak{B}}(k)\|^2$ converges uniformly for all k from a certain set K to $\|\mathfrak{B} - \mathfrak{B}(k)\|^2$.

2. THE LINEAR MODEL WITH STOCHASTIC REGRESSORS

Consider an 1 x (n+1) random vector Z with the expectation μ and the covariance matrix Σ . Z, μ and Σ are partitioned as

$$Z = (Y, X), \quad \mu = (\mu_Y, \mu_X)$$

$$\begin{pmatrix} \sigma_Y^2 & \sigma_{Y.X} \\ \sigma_{X.Y} & \Sigma_{XX} \end{pmatrix} = \Sigma$$

where Y and μ_Y are 1 x 1, X and μ_X are 1 x n, and Σ_{XX} is n x n.

The problem is to determine the regressand Y by the regressors X. For convenience we will let NV_{n+1} denote the class of 1x(n+1) random vectors Z having the $N_{n+1}(0,\Sigma)$ distribution with positive definite matrices Σ . It is well known that for $Z \in NV_{n+1}$

$$E(Y/X) = X \Sigma_{XX}^{-1} \sigma_{X.Y} =: X \mathcal{B}$$

and

$$\text{var } (Y/X) = \sigma_Y^2 - \sigma_{Y.X} \Sigma_{XX}^{-1} \sigma_{X.Y} =: \sigma_{Y/X}^2 .$$

Moreover, the random variable $\varepsilon := Y - X\mathcal{B}$ and the 1 x n random vector X are independent and $\varepsilon \sim N_1(0, \sigma_{Y/X}^2)$. In other words, X and ε determine Y in a linear manner, as

$$Y = X \mathcal{B} + \varepsilon \quad . \tag{1}$$

In order to obtain the maximum likelihood estimators of $\mathcal{B}$ and $\sigma_{Y/X}^2$ it is not necessary to restrict ourselves to normally distributed regressors. Therefore, we introduce a generalized parametric family F instead of NV.

As suggested by (1), we now consider random vectors Z which are defined by X and ε according to

$$Z = (X \mathcal{B} + \varepsilon, X) = (Y, X)$$

Let F be the class of 1 x (n+1) random vectors Z possessing following properties:

(i) X and ε are independent

(ii) $\varepsilon \sim N(0, \sigma^2)$ for some $\sigma^2 > 0$

(iii) $X \sim G_\vartheta$ for some $\vartheta \in \Theta$, where $\mathcal{G} = \{ G_\vartheta : \vartheta \in \Theta \}$ is an arbitrary family of distribu-

tions on $\mathbb{R}^n$ with the parametric space Θ and positive definite covariance matrices Σ_{XX}.

Note that $NV_{n+1} \in F$ iff $N_n(\mathcal{O}, \Sigma_{XX}) \in \mathfrak{C}$ for all Σ_{XX}.

Further suppose that for all $\vartheta \in \Theta$ there exists a density g_ϑ of G_ϑ and denote the density of $N_1(0, \sigma^2)$ by f_{σ^2}. Then the density of Z with the parameter (B, σ^2, ϑ) is given by

$$f(y,x) = g_\vartheta(x)\, f_{\sigma^2}(y - x\mathbb{B}) \tag{2}$$

where, as before X is a row vector and $\mathbb{B}$ is a column vector.

Estimating $\mathbb{B}$ and σ^2 we take a sample of size $N > n$ of Z and denote it by

$$\mathbb{Z} = (\mathbb{Y}, \mathbb{X})$$

where the results of the i-th trial $Z_i = (Y_i, X_i)$ are written in the i-th row of $\mathbb{Z}$.

Obviously, from (1) we get the representation

$$\mathbb{Y} = \mathbb{X}\mathbb{B} + \mathfrak{E} \tag{1'}$$

with N x 1 random vector $\mathfrak{E} \sim N_N(0, \sigma^2 I_N)$ where I_N is the N x N identity matrix. Using now (2) we get the logarithmic likelihood function

$$l(B, \sigma^2, \vartheta; z) = \sum_{i=1}^{N} \log f_{\sigma^2}(y_i - x_i B) + \sum_{i=1}^{N} \log g_\vartheta(x_i)$$

$$=: l_1(B, \sigma^2; z) + l_2(\vartheta; z).$$

By the property (ii) of F we obtain

$$l_1(B, \sigma^2; \mathbb{Z}) = -\frac{1}{2\sigma^2}(\mathbb{y} - \mathbb{x}B)^T(\mathbb{y} - \mathbb{x}B) + \frac{N}{2}\log(2\pi\sigma^2).$$

A result of Okamoto [1973] yields that for all $\vartheta \in \Theta$

$$G_\vartheta\{|\mathbb{X}^T\mathbb{X}| > 0\} = 1;$$

here $|\mathbb{A}|$ denotes the determinant of a matrix $\mathbb{A}$. Hence, there exist the maximum likelihood estimates $\hat{\mathbb{B}}$ of $\mathbb{B}$ and $\hat{\sigma}^2$ of σ^2 and they have a similar structure as in the linear model (with non-random regressors).

Note that

$$\mathbb{B} = (\mathbb{X}^T\mathbb{X})^{-1}\,\mathbb{X}^T\,\mathbb{Y} = \mathbb{X}^+\,\mathbb{Y}$$

where $\mathbb{X}^+$ is the Moore-Penrose inverse of $\mathbb{X}$. As in the classical case instead of $\hat{\sigma}^2$ we use the estimate

$$S^2 = \frac{N-n}{n}\,\hat{\sigma}^2 = \frac{1}{N-n}\,(\mathbb{Y} - \mathbb{X}\hat{\mathbb{B}})^T(\mathbb{Y} - \mathbb{X}\hat{\mathbb{B}})$$

which is unbiased (see theorem 1).

The following theorem gives some properties of the estimators $\hat{\mathbb{B}}$ and S^2.

Theorem 1: For $Z \in P$ we take a sample of size $N > n$. Suppose that the expectation of $(\mathbb{X}^T\mathbb{X})^{-1}$ exists then

$$E\,(\hat{\mathbb{B}}/\mathbb{X}) = \mathbb{B} \tag{4}$$

$$\operatorname{cov}\,(\hat{\mathbb{B}}/\mathbb{X}) = \sigma^2(\mathbb{X}^T\mathbb{X})^{-1} \tag{5}$$

and

$$E(S^2/\mathbb{X}) = \sigma^2 \tag{6}$$

In particular, $\hat{\mathbb{B}}$ and S^2 are unbiased.

Remark 1: If $Z \in N_{n+1}$ then the n x n random matrix $\mathbb{X}^T\mathbb{X}$ from a sample of size N and the expectation of $(\mathbb{X}^T\mathbb{X})^{-1}$ exists if $N > n+1$; moreover (see Kshirsagar [1972])

$$E\,(\mathbb{X}^T\mathbb{X})^{-1} = \frac{1}{N-n-1}\,\Sigma_{XX}^{-1} \tag{7}$$

Proof: Clearly

$$\mathbb{X}^+(\mathbb{X}^+)^T = (\mathbb{X}^T\mathbb{X})^{-1}$$

and the existence of the expectation of $(\mathbb{X}^T\mathbb{X})^{-1}$ implies the existence of the expectation of $\mathbb{X}^+$. Using (1') and (ii) of $\mathcal{F}$ we conclude

$$E\,(\hat{\mathbb{B}}/\mathbb{X}) = E(\mathbb{B} + \mathbb{X}^+\mathbb{E}/\mathbb{X}) = \mathbb{B}\ ; \quad \text{i.e. (4).}$$

Further we need a result for conditional expectations of random matrices. Let $\mathbb{A}_j$ be u_j x v_j random matrices for j=1,2,3,4 and suppose that $\mathbb{A}_1$ and $\mathbb{A}_3$ are measurable with respect to the σ-algebra generated by $\mathbb{A}_4$ and $v_1 = u_2$, $v_2 = u_3$.
Then

$$E(\mathbb{A}_1\,\mathbb{A}_2\,\mathbb{A}_3/\mathbb{A}_4) = \mathbb{A}_1\,E(\mathbb{A}_2/\mathbb{A}_4)\,\mathbb{A}_3 \tag{8}$$

provided that the expectation of $\mathbb{A}_2$ exists.
It is easy to see that

$$(\hat{\mathbb{B}} - \mathbb{B})\,(\hat{\mathbb{B}} - \mathbb{B})^T = \mathbb{X}^+ \mathfrak{E}\,\mathfrak{E}^T\,(\mathbb{X}^+)^T. \qquad (9)$$

Using now (3) and (8). With $\mathbb{A}_1 = \mathbb{A}_3 = \mathbb{X}^+$, $\mathbb{A}_2 = \mathfrak{E}\,\mathfrak{E}^T$, $\mathbb{A}_4 = \mathbb{X}$ the statement (5) is an immediate consequence. Putting

$$\mathbb{M} = I_N - \mathbb{X}\,\mathbb{X}^+$$

we can write

$$(N-n)\,S^2 = \mathfrak{E}^T\,\mathbb{M}\,\mathfrak{E} = \operatorname{tr}(\mathbb{M}\,\mathfrak{E}\,\mathfrak{E}^T).$$

Applying again (8) with $\mathbb{A}_1 = \mathbb{M}$, $\mathbb{A}_2 = \mathfrak{E}\,\mathfrak{E}^T$, $\mathbb{A}_3 = I_N$ and $\mathbb{A}_4 = \mathbb{X}$ we get

$$(N-n)\,E(S^2/\mathbb{X}) = \operatorname{tr}\,(\mathbb{M}\,E(\mathfrak{E}\,\mathfrak{E}^T/\mathbb{X}) = (N-n)\sigma^2$$

as $\mathbb{M}$ is idempotent and (6) is established. For our next purpose the maximum likelihood estimate of $\mathbb{B}$ of a sample of size N is written as $\hat{\mathbb{B}}^{(N)}$.

Theorem 2: For $Z \in NV_{n+1}$ we take a sample of size N.
Then the sequence of estimates $\{\hat{\mathbb{B}}^{(N)}\}$ is consistent to $\mathbb{B}$.

Proof: From theorem 1 and remark 1 the estimates $\hat{\mathbb{B}}^{(N)}$ are unbiased. Then for the consistence of $\{\mathbb{B}^{(N)}\}$ it suffices to show that

$$\lim_{N\to\infty} \operatorname{tr}\{\operatorname{cov}\,(\hat{\mathbb{B}}^{(N)})\} = 0. \qquad (10)$$

From theorem 1 and (7) it follows

$$\operatorname{tr}\{\operatorname{cov}\,(\hat{\mathbb{B}})\} = \frac{\sigma^2 n}{N-n-1}$$

hence (10) is valid.

3. MULTICOLLINEARITY AND ITS CONSEQUENCES

As a measure of the dependence of the regressors X we use the determinant of the correlation matrix $\mathbb{R}_{XX}$ of X, namely, the regressors X are said to be multicollinear of degree δ, $1 \le \delta$, if

$$\delta = \frac{1}{|\mathbb{R}_{XX}|} \quad .$$

In application there are no possibilities giving a bound δ_0 for the degree of multicollinearity in such a way that for $\delta < \delta_0$ the properties of the estimate $\hat{\mathfrak{B}}$ are scarcely influenced by the multicollinearity but for $\delta \geq \delta_0$ this estimate is not useful. The only way to study the effect of δ is to investigate its influence on the estimate. From theorem 1 we see that cov $(\hat{\mathfrak{B}})$ depends on the degree of multicollinearity. Note that also the statistics t_j for testing the hypothesis H_0: $B_j = 0$ are dependent of δ . The larger the degree of multicollinearity the smaller are t_j.

The complicated dependence also of parameters like $\sigma^2_{Y/X}$ of the multicollinearity is now studied in the following simple example.

Example: Consider $Z \in NV$ with $n = 2$ and var (Y) = var (X_1) = var $(X_2) = 1$. Then

$$\Sigma = \begin{pmatrix} 1 & \varrho_{12} & \varrho_{13} \\ \varrho_{21} & 1 & \varrho_{23} \\ \varrho_{31} & \varrho_{32} & 1 \end{pmatrix}$$

and $X = (X_1, X_2)$ is multicollinear of degree δ if $\varrho^2_{23} = \frac{\delta - 1}{\delta}$.
In this case it follows for $\sigma^2(\delta) = \sigma^2_{Y/X}$

$$\sigma^2(\delta) = \begin{cases} 1 - (\varrho^2_{12} + \varrho^2_{13})\delta + 2\,\varrho_{12}\,\varrho_{13}\sqrt{(1-\delta)\delta} & \text{if } \varrho_{23} \geq 0 \\ 1 - (\varrho^2_{12} + \varrho^2_{13})\delta - 2\,\varrho_{12}\,\varrho_{13}\sqrt{(1-\delta)\delta} & \text{if } \varrho_{23} < 0 \,. \end{cases}$$

Because Σ is positive definite we have $\varrho^2_{12} < 1$, $\varrho^2_{13} < 1$ and $\varrho_{23} \in (a,b)$

with $a,b = \varrho_{12}\,\varrho_{13} \pm \sqrt{1 - \varrho^2_{12} - \varrho^2_{13} + \varrho^2_{13}\,\varrho^2_{12}}$.

Only for $\varrho_{12} = \varrho_{13}$ we get $b = 1$. Further, put

$$A = \begin{cases} \frac{1}{1-a^2} & \text{for } a \geq 0 \\ 1 & \text{for } a < 0 \end{cases}$$

and

$$B = \begin{cases} \frac{1}{1-b^2} & \text{for } \varrho_{12} \neq \varrho_{13} \\ \infty & \text{for } \varrho_{12} = \varrho_{13} \end{cases}$$

It is easy to see that

$$\sigma^2(A+) = \begin{cases} 0 & \text{for } a \geq 0 \\ 1-(\varrho_{12}^2 + \varrho_{13}^2) & \text{for } a < 0 \end{cases}$$

and

$$\sigma^2(B-) = \begin{cases} 0 & \text{for } \varrho_{12} \neq \varrho_{13} \\ 1-\varrho_{12}^2 & \text{for } \varrho_{12} = \varrho_{13} \end{cases}$$

In order to study the behaviour of $\sigma^2(\delta)$ we have to distinguish three cases: (i) $a \geq 0$, (ii) $b \leq 0$, (iii) $a < 0 < b$.
As the transformation $\tilde{\varrho}_{12} = -\varrho_{12}$, $\tilde{\varrho}_{13} = \varrho_{13}$, $\tilde{\varrho}_{23} = -\varrho_{23}$ is invariant for $\sigma^2_{Y/X}$ we only have to consider the case $a \geq 0$ or $a < 0 < b$ and $\varrho_{23} \geq 0$. Then the function $\sigma^2(\delta)$ is monotonously increasing in (A, δ_0) and monotonously decreasing in (δ_0, B) where

$$\delta_0 = \begin{cases} \frac{\max(\varrho_{12}^2, \varrho_{13}^2)}{|\varrho_{12}^2 - \varrho_{13}^2|} & \text{for } \varrho_{12} \neq \varrho_{13} \\ \infty & \text{for } \varrho_{12} = \varrho_{13}. \end{cases}$$

The function $\sigma^2(\delta)$ reaches its maximum $1 - \max(\varrho_{12}^2, \varrho_{13}^2)$ at δ_0.
This example shows that $\sigma^2_{Y/X}$ depends on the degree of multicollinearity as well as on the correlation structure. If $\varrho_{12}/\varrho_{13}$ goes to one then δ_0 tends to infinity and consequently, high degree of multicollinearity may be combined with great $\sigma^2_{Y/X}$.
A sequence of simulation examples of more complicated structure have shown us the same effect, see Jahn [1984].

From these examples we get the intention that the mean square error of the extrapolation of the regressand by n regressors would be reduced using only $m < n$ regressors with a greater determinant of correlation matrix than the one of the original regressors. In this way the subset regression is a method to reduce the degree of multicollinearity and therefore to improve the estimate $\hat{\mathbb{B}}$.

With problems like this have among others also delt Oliker [1978], Akaike [1970, 1973, 1974, 1977, 1978], Bierens [1980], Mallows [1973], Hooking [1976], Shibata [1981].

4. DETERMINATION OF THE DIMENSION

In this section we study the subset regression for the linear model with stochastic regressors (2.1) with $Z \in NV_{n+1}$. Suppose that we select the regressors $X_{k_1},\ldots,X_{k_m}$; $1 \leq m \leq n$, and remove the regressors $X_{h_1},\ldots,X_{h_{n-m}}$ and put $k = (k_1,\ldots,k_m)$, $k_1 < \ldots < k_m$ and $h = (h_1,\ldots,h_{n-m})$, $h_1 < \ldots < h_{n-m}$. Moreover, we set $X(k) = (X_{k_1},\ldots,X_{k_m})$, $\mathcal{B}(k) = (B_{k_1},\ldots,B_{k_m})^T$.

Then the model (2.1) can be written as

$$Y = X(k)\,\mathcal{B}(k) + X(h)\,\mathcal{B}(h) + \varepsilon . \qquad (1)$$

Taking a sample of size $N > n$ we denote its result by

$$\mathcal{Z} = (\mathcal{Y}, \mathcal{X}(k), \mathcal{X}(h)) ;$$

hence we have

$$\mathcal{Y} = \mathcal{X}(k)\,\mathcal{B}(k) + \mathcal{X}(h)\,\mathcal{B}(h) + \mathcal{E} . \qquad (2)$$

The main object of regression analysis is to extrapolate $Y(E)$ by a random vector $X(E)$ which is independent of $\mathcal{X}$ and $Z(E)$ and each row of $\mathcal{Z}$ are identically distributed. As above the restriction of $X(E)$ to the variables k is denoted by $X(E,k)$.

The maximum likelihood extrapolation of the future observation on $Y(E)$ at $X(E)$ is given by

$$\hat{Y}(E,k) = X(E,k)\,\hat{\mathcal{B}}(k).$$

For fixed k and h the covariance matrix Σ_{XX} is partitioned as

$$\Sigma_{XX} = \begin{pmatrix} \Sigma_{kk} & \Sigma_{kh} \\ \Sigma_{hk} & \Sigma_{hh} \end{pmatrix}$$

where Σ_{kk} is $m \times m$ and Σ_{hh} is $(n-m) \times (n-m)$.

For simplicity put

$$\Sigma_{hh/k} = \Sigma_{hh} - \Sigma_{hk}\,\Sigma_{kk}^{-1}\,\Sigma_{kh}$$

$$B_{h/k} = \Sigma_{kk}^{-1}\,\Sigma_{kh} \quad .$$

Theorem 3: If $N > n+1$ then

$$E\left[(Y(E) - \hat{Y}(E,k))^2 / X(k), X(E,k)\right]$$

$$= X(E,k)(X(k)^T X(k))^{-1} X^T(E,k)\left[B(h)^T \Sigma_{hh/k} B(h) + \sigma^2_{Y/X}\right] + \sigma^2_{Y/X}$$

and

$$E\,(Y(E) - \hat{Y}(E,k))^2 = \left(1 + \frac{n}{N-n-1}\right)\left[\sigma^2_{Y/X} + B(h)^T \Sigma_{hh/k} B(h)\right] \quad (4)$$

In particular,

$$\lim_{N\to\infty} E\,(Y(E) - \hat{Y}(E,k))^2 = B(h)^T \Sigma_{hh/k} B(h) + \sigma^2_{Y/X} \quad (4')$$

Proof: From (1) and (2) we conclude

$$Y(E) - \hat{Y}(E,k) = X(E,k)\,B(h) + \varepsilon - X(E,k)X^+(k)\,X(h)\,B(h)$$

$$- X(E,k)\,X^+(k)\,\xi .$$

As $X(E)$ and ε, ξ and ε, as well as $X(E)$ and ε are independent we then obtain

$$E\left[(Y(E) - \hat{Y}(E,k))^2 / X, X(E)\right]$$

$$= B(h)^T X(E,h)^T X(E,h) B(h) + \sigma^2_{Y/X}\left(1 + X(E,k)(X(k)^T X(k))^{-1} X(E,k)^T\right)$$

$$+ \left(X(h)\,B(h)\right)^T (X(E,k)\,X(k)^+)^T (X(E,k)\,X(k)^+)(X(h)\,B(h))$$

$$- 2\,B(h)^T X(E,h)^T X(E,h)\,X(k)^+\,X(h)\,B(h) \quad (5)$$

$$=: I_1 + I_2 + I_3 - 2I_4 .$$

From the usual normal theory we can derive the following conditional expectations:

$$E\left[X(E,h)^T X(E,h) / X(E,k)\right] = \Sigma_{hh/k} + B_{h\,k}^T X(E,k)^T X(E,k)\,B_{h/k} \quad (6)$$

$$E\left[X(E,h) / X(E,k)\right] = X(E,k)\,B_{h/k} \quad (7)$$

$$E\left[X(h) / X(k)\right] = X(k)\,B_{h/k} \quad (8)$$

and

$$E\left[\mathcal{Y}(h)\,\mathcal{B}(h)\,\mathcal{B}(h)^T\,\mathcal{Y}(h)^T/\,\mathcal{Y}(k)\right]$$

$$= \mathcal{B}(h)^T \Sigma_{hh/k}\,\mathcal{B}(h)\,I_N + \mathcal{Y}(k)\,\mathcal{B}(h)\,\mathcal{B}(h)^T\mathcal{Y}(k)^T \quad . \tag{9}$$

Now we are able to study the conditional expectation of I_1, I_3 and I_4 under $\mathcal{Y}(k)$ and $X(E,k)$. By virtue of (6) we get

$$E\left[I_1/X(E,k)\right] = \mathcal{B}(h)^T \mathcal{B}_{h/k}{}^T\, X(E,k)^T\, X(E,k)\,\mathcal{B}_{h/k}\,\mathcal{B}(h)$$

$$+ \mathcal{B}(h)^T \Sigma_{hh/k}\,\mathcal{B}(h) \quad . \tag{10}$$

Note that

$$E\left[I_4/\mathcal{Y}(k),X(E,k)\right] = \mathcal{B}(h)^T\, E\left[X(E,h)/X(E,k)\right]$$

$$\cdot X(E,k)\,\mathcal{Y}(k)^+\, E\left[\mathcal{Y}(h)/\mathcal{Y}(k)\right]\mathcal{B}(h) \quad .$$

In accoording to (7) and (8) we derive

$$E\left[I_4/\mathcal{Y}(k),X(E,k)\right] = \mathcal{B}(h)^T \mathcal{B}_{hk}{}^T\, X(E,k)^T\, X(E,k)\,\mathcal{B}_{n/k}\,\mathcal{B}(h) \tag{11}$$

Finally using (9) we see

$$E\left[I_3/\mathcal{Y}(k)\right] = E\left[\operatorname{tr}\left\{X(E,k)\,\mathcal{Y}(k)^+\right\}^T \left\{X(E,k)\,\mathcal{Y}(k)^+\right\}\right.$$

$$\left.\cdot\,\mathcal{Y}(h)\,\mathcal{B}(h)\,\mathcal{B}(h)^T\,\mathcal{Y}(h)^T/\,\mathcal{Y}(k)\right]$$

$$= X(E,k)\,(\mathcal{Y}(k)^T\mathcal{Y}(k))^{-1}\,X(E,k)^T\,\mathcal{B}(h)^T\Sigma_{hh/k}\,\mathcal{B}(h)$$

$$+ \mathcal{B}(h)\,\mathcal{B}_{h/k}{}^T\,X(E,k)^T\,X(E,k)\,\mathcal{B}_{h/k}\,\mathcal{B}(h) \quad . \tag{12}$$

From (5), (10), (11), (12) the statement (3) follows immediately.

Now we show (4). Using remark 1 we conclude

$E\left[X(E,k)\,(\mathcal{Y}(k)^T\,\mathcal{Y}(k))^{-1}\,X^T(E,k)\right] = \frac{m}{N-m-1}$ and statement (4) follows.

Next we give the expected extrapolation error if all n regressors are applied in regression.

Corollary 1: If $N > n+1$ then

$$E\left[(Y(E) - \hat{Y}(E,n))^2/\,\mathcal{Y},\,X(E)\right]$$
$$= \sigma^2_{Y/X}\left[1 + X(E)(\mathcal{Y}^T\mathcal{Y})^{-1}\,X(E)^T\right] \tag{13}$$

and

$$E\,(Y(E) - \hat{Y}(E,h))^2 = \sigma^2_{Y/X}\,(1 + \frac{n}{N-n-1}) \,. \tag{14}$$

In order to consider a sequence of linear models with stochastic regressors we start with a sequence $(X_j,\ j \in \mathbb{N})$ of regressors and suppose that we select from it m, namely $X(k) = (X_{k_1},\dots,X_{k_m})$ with $k = (k_1,\dots,k_m)$ and $m > 1$. As before we assume $Z(k) = (Y,X(k)) \in NV_{m+1}$ and

$$Y = X(k)\,\mathcal{B}(k) + \mathcal{E}(k) \tag{15}$$

with $\mathcal{B}(k) = (B_{k_1},\dots,B_{k_m})^T \in \mathbb{R}^m$.

In other words we obtain this model by putting in (1) $\mathcal{B}(h) = 0$ and $n = \infty$. This means that the selection of regressors k is strongly connected with parameter $\mathcal{B}$. Further put

$$\sigma^2(k) = \sigma^2_{Y/X(k)} \,.$$

Taking a sample of size $N > n$ we denote its result by $\mathcal{Z}(k) = (\mathcal{Y},\mathcal{X}(k))$, i.e. we have

$$\mathcal{Y} = \mathcal{X}(k)\,\mathcal{B}(k) + \mathcal{E}(k). \tag{16}$$

For an m x 1 random or not random vector c put

$$\|c\|^2 = c^T\,\mathcal{X}(k)^T\mathcal{X}(k)\;\; c$$

$$\|c\|^2_k = c^T\,\Sigma(k)\;\; c \,.$$

Using remark 1 we get

$$E\,\|\mathcal{B}-\hat{\mathcal{B}}(k)\|^2 = \|\mathcal{B} - \mathcal{B}(k)\|^2_k + \frac{m}{N-m-1}\,\sigma^2(k) =: M(k,N). \tag{17}$$

Obviously, for all selections k, $\|\mathcal{B} - \hat{\mathcal{B}}(k)\|^2$ converges in probability to $\|\mathcal{B} - \mathcal{B}(k)\|^2_k$ as $N \longrightarrow \infty$. Now we show even that this is valid uniformly for all k from the set

$$K_{m,N,\xi} = \left\{ k : k_m < N^{\frac{1}{2}-\xi} \right\}$$

where $0 < \xi < \frac{1}{2}$.

Theorem 4: For all $0 < \tau < 1$ we have

$$\lim_{N \to \infty} P\left\{ \| \mathbb{B} - \hat{\mathbb{B}}(k) \|^2 > (1 - \tau)\, M(k,N) \;\forall k \in K_{m,N,\xi} \right\} = 1$$

provided that for all N

$$R_N := \sup_{k \in K_{m,N,\xi}} \frac{\| \mathbb{B} - \mathbb{B}(k) \|_k^2}{\sigma^2(k)} < \infty$$

For the proof of theorem 4 we need following result.

Lemma 1: Let χ_n^2 be a χ^2 random variable with n degrees of freedom. Then for any $0 < \eta < n$, we get

$$F_n(n-\eta) := P(\chi_n^2 < n - \eta) \leq e^{\eta/2} \left(1 - \frac{\eta}{n}\right)^{n/2} \qquad (19)$$

Proof: Using the moment generating function of χ_n^2 we see for $t \leq 0$

$$(1 - 2t)^{-\frac{n}{2}} = \int_0^\infty e^{tx} F_n(dx)$$

$$\geq \int_0^{n-\eta} e^{tx} F_n(dx)$$

$$> e^{t(n-\eta)} F_n(n - \eta) ;$$

i.e. $F_n(n-\eta) < e^{-t(n-\eta)} (1-2t)^{-n/2}$

For

$$t = \frac{-\eta}{2(n-\eta)}$$

the upper bound of the last inequality is minimal and (19) is established.

Proof of theorem 4: Obviously, we have

$$a_N := P(\|\mathbb{B} - \hat{\mathbb{B}}(k)\|^2 < (1-\tau) M(k,N) \exists k \in K_{m,N,\xi}) \leq$$

$$\leq \sum_{k \in K_{m,N,\xi}} P(\|\mathbb{B} - \hat{\mathbb{B}}(k)\|^2 \leq (1-\tau) M(k,N)) \leq$$

$$\leq \sum_{k \in K_{m,N,\xi}} P\left(\frac{\|\hat{\mathbb{B}}(k) - \mathbb{B}(k)\|^2}{\sigma^2(k)} \leq \frac{m}{N-m-1} - \frac{\tau M(k,N)}{\sigma^2(k)} \right)$$

It is easy to see that $\|B(k) - \hat{B}(k)\|^2 / \delta^2(k)$ is χ^2 distributed with m degrees of freedom. Hence, it follows

$$a_N \leq \sum_{k \in K'_{m,N,\xi}} F_m \left(\frac{m}{N-m-1} - \tau \frac{M(k,N)}{\delta^2(k)} \right)$$

where

$$K'_{m,N,\xi} = \left\{ k \in K_{m,N,\xi} : \frac{m}{N-m-1} - \tau \frac{M(k,N)}{\delta^2(k)} > 0 \right\} .$$

Applying now lemma 1 for $n = m$ and

$$\eta = - \frac{m}{N-m-1} + \tau \frac{M(k,N)}{\delta^2(k)} + m =$$

$$= m\left(1 - \frac{1-\tau}{N-m-1}\right) + \tau \frac{\|B - B(k)\|_k^2}{\delta^2(k)}$$

we see that for $k \in K'_{m,N,\xi}$

$$F_m \left(\frac{m}{N-m-1} - \tau \frac{M(k,N)}{\delta^2(k)} \right)$$

$$\leq e^{\eta/2} \left(\frac{1-\tau}{N-m-1} - \tau \frac{\|B - B(k)\|_k^2}{m\,\delta^2(k)} \right)^{m/2} \leq e^{\eta/2} \left(\frac{1-\tau}{N-m-1}\right)^{\eta/2} .$$

So we obtain by definition of R_N and η

$$a_N \leq \left(\frac{V}{m}\right) e^{\frac{1}{2}\left[m\left(1 - \frac{1-\tau}{N-m-1}\right) + \tau R_N\right]} \left[\log \frac{1-\tau}{N-m-1} + 1 \right] .$$

As $V = N^{\frac{1}{2}-\xi}$ and $\log \frac{1-\tau}{N-m-1} + 1 < 0$

for sufficiently large N, we see

$$a_N \leq \frac{1}{m!} e^{m \log V + \frac{1}{2}\left[m\left(1 - \frac{1-\tau}{N-m-1}\right)\right]} \left[-\log N + 1 + \log(1-\tau) + o(1) \right]$$

$$\leq \frac{1}{m!} e^{\log N \left[\frac{m}{2} - m\xi - \frac{1}{2} m \left(1 - \frac{1-\tau}{N-m-1}\right)\right] + \frac{m}{2}\left(1 - \frac{1-\tau}{N-m-1}\right)\left[1 + \log(1-\tau) + o(1)\right]} \leq$$

$$\leq \frac{1}{m!} N^{-m\xi + \frac{m}{2}\frac{1-\tau}{N-m-1}} e^{\frac{m}{2}\left(1 - \frac{1-\tau}{N-m-1}\right)(1 + \log(1-\tau) + o(1))} =$$

$$= N^{-m\xi + o(1)} \cdot O(1) = o(1)$$

and the statement is shown.

Acknowledgements

This paper was a lecture on the third Prague conference on asymptotic statistics. The authors thank for discussion and for the helpful comments of the reviewers of the CMUC.

REFERENCES

Anděl, J. (1982): Fitting models in time series analysis, Math. Operationsforsch, Statist., Ser. Statistics Vol.13,No.1,121-143

Akaike, H. (1970): Statistical predictor identification, Ann. Inst. Statist. Math. 22, 203-217

Akaike, H. (1973): Information theory and an extension of the m l principle in 2nd Int. Symposium on Information Theory, Eds. Petrov, Csáki. 267-281 Budapest; Akademiai Kiadó

Akaike, H. (1974): A new look at the statistical model identification. I.E.E.E. Trans. Auto Control. 19, 716-723

Akaike, H. (1977): On entropy maximation principle, Application of statistics. P.P. Krishnaiah ed. North Holland, Amsterdam 27-41

Akaike, H. (1978): A Bayesian analysis of the minimum AIC procedure. Ann. Inst. Statist. Math. A30, 9-14

Bierens, H.J. (1980): Consistent selection of explanatory variable, Stichting voor Economisch Onderzoek der Universiteit van Amsterdam, Seo Overdruck 1

Eaton, M.L. and Perleman (1973): The nonsingularity of generalized sample covariance matrices, The Annals of Statistics Vol. 1 No. 4, 710-717

Hocking, R.R. (1976): The analysis and selection of variables in linear regression. Biometrics 62, 1-49

Jahn, W. (1984): Dimensionserniedrigung von Parameterräumen im linearen Modell mit stochastischen Regressoren. To appear in: Sitzungsberichte der IGMS der Math. Gesellschaft der DDR

Kshirsagar, A.M. (1972): Multivariate Analysis, Marcel Dekker, New York

Mallows, C.L. (1973): Some comments on Cp, Technometrics 15, 213-220.

Okamoto, M. (1973): Distinctness of the eigenvalues of a quadratic form in a multivariate sample , The Annals of Statistics Vol. 1, No 4, 763-765

Olikwe, V.I. (1978): On the relationship between the sample size and the number of variables in a linear regression model, Commun. Statist. A7(6), 509-516

Park and H. Sing (1981): Collinearity and opt. restriction on regression parameters for estimating responses, Technometrics 23, 3, 289-295

Schwarz, G. (1978): Estimating the dimension of a model, The Annals of Statistics Vol.6 , No 2, 461-464

Shibata, R. (1980): Asymptotically efficient selection of the order of the model for estimating parameters of a lin. process, The Annals of Statistics Vol. 8 No 1, 45-54

Shibata, R. (1981): An optimal selection of regression variables Biometrika 68, 1, 45-54

Stein, C.M. (1969): Multivariate Analysis I, Technical report No 42, Dep. of Statist., Stanford University

Stone, M. (1979): Comments on model selection criteria of Akaike and Schwarz, J.R. Statist. Soc. B 41; 276-278

Sugiura, N. (1978): Further analysis of the data by Akaike Commun. Statist. A 7, 13-26

Karl-Marx-University Leipzig, Department of Mathematics, 701 Leipzig, Karl-Marx-Platz 10

(Oblatum 15.2. 1984)