SEQUENTIAL METHODS IN STATISTICS
BANACH CENTER PUBLICATIONS, VOLUME 16
PWN POLISH SCIENTIFIC PUBLISHERS
WARSAW 1985

POSSIBILITIES OF THE DETERMINATION OF SAMPLE SIZE IN REGRESSION ANALYSIS: A REVIEW

W. JAHN

Section of Mathematics, Karl-Marx University of Leipzig, Leipzig, G.D.R.

1. Introduction

We consider the following problem: X is a random vector with $X \sim N_{n+1}(\mu, \Sigma)$,

$$X^T = (X_{n+1}, X^{(1)T}), \qquad X^{(1)T} = (X_1, \ldots, X_n), \qquad \mu^T = (\mu_{n+1}, \mu^{(1)T})$$

$$\Sigma = \begin{pmatrix} \sigma_{n+1}^2 & \sigma_{n+1\cdot(1)}^T \\ \sigma_{n+1\cdot(1)} & \Sigma_{11} \end{pmatrix}$$

where $\sigma_{n+1\cdot(1)}$ is the vector of covariances between X_{n+1} and the subvector $X^{(1)}$ and Σ_{11} is the covariance matrix of $X^{(1)}$. We are interested in the conditional moments of X_{n+1} under the condition of $X^{(1)} = x^{(1)}$:

$$E(X_{n+1}|X^{(1)} = x^{(1)}) = \mu_{n+1} + \sigma_{n+1\cdot(1)}^T \Sigma_{11}^{-1}(x^{(1)} - \mu^{(1)}) =: \mu_{n+1/(1)} \tag{1.1}$$

and

$$\operatorname{var}(X_{n+1}|X^{(1)}) = \sigma_{n+1}^2 - \sigma_{n+1\cdot(1)}^T \Sigma_{11}^{-1} \sigma_{n+1\cdot(1)} =: \sigma_{n+1/(1)}^2 \tag{1.2}$$

These moments should be estimated on the basis of a (mathematical) sample

$$X_{(N)}^T = (X_{n+1}^{(N)}, X_{(N)}^{(1)T})$$

of size N. The substitution $\sigma_{n+1\cdot(1)}^T \Sigma_{11}^{-1} =: \boldsymbol{B}$ leads to the following model:

$$X_{n+1}^{(N)} = \boldsymbol{B} x_{(N)}^{(1)} + \varepsilon^{(N)} \tag{1.3}$$

where

$$E(\varepsilon | X^{(1)} = x^{(1)}) = 0 \tag{A1}$$

$$E(\varepsilon^{(N)T} \varepsilon^{(N)} | X^{(1)}_{(N)} = x^{(1)}_{(N)}) = \sigma^2_{n+1/(1)} I_N,$$

$$\varepsilon^{(N)} \sim N(\mathbf{0}, \sigma^2_{n+1/(1)} I_N).$$

This model is an ordinary linear model in which the matrix $X^{(1)}_{(N)}$ is no longer assumed constant. Since the X_i $(n+1)$-dimensional i.i.d. random vectors such that $E(X_i) = \mu < \infty$, $\operatorname{cov}(X_i) = \Sigma$, and X_i are independent of ε for all i and N and since the distribution of X_i is absolutely continuous with respect to the Lebesgue measure, it follows that

$$\Sigma > 0 \tag{A2}$$

The assumption (A2) implies that the rank of $X^{(1)}_{(N)}$ is n with probability one, and hence $X^{(1)}_{(N)} X^{(1)T}_{(N)}$ is invertible with probability one. Conditionally on $X^{(1)}_1, \ldots, X^{(1)}_N$ this model is an ordinary linear model as discussed, for instance, in HUMAK (1977). In the following sections of the paper the following possibilities will be given

1) of calculating the properties of the estimators for the model parameters if N and n are fixed,

2) of improving the properties of these estimators by eliminating of variables of $X^{(1)}$, i.e., where N is fixed and n is variable,

3) of determining a suitably large sample size ensuring a certain degree of precision for estimating by means of inequalities by Tong (1980) or by cross validation and

4) of sequential procedures for the estimation of the conditional moments of a multinormal distribution.

2. Properties of the estimators ${}_s\hat{B}^{(N)}$ and ${}_sS^2_{n+1/(1)}$

If assumptions (A1) and (A2) are satisfied, then the following well-known results are valid:

$${}_s\hat{B}^{(N)} = (X^{(1)}_{(N)} X^{(1)T}_{(N)})^{-1} X^{(1)}_{(N)} X^{(N)T}_{n+1}, \tag{2.1}$$

$${}_sS^{(N)2}_{n+1/(1)} = \frac{1}{N-n} \hat{\varepsilon}^{(N)} \hat{\varepsilon}^{(N)T} \tag{2.2}$$

where

$$\hat{\varepsilon}^{(N)} = X^{(N)}_{n+1} - {}_s\hat{B}^{(N)} x^{(1)}_{(N)},$$

${}_s\hat{B}^{(N)}$ and ${}_sS^{(N)2}_{n+1/(1)}$ are unbiased for B and $\sigma^2_{n+1/(1)}$, respectively, if $E(X^{(1)}_{(N)} X^{(1)T}_{(N)})^{-1}$ exists. The covariance matrix of ${}_s\hat{B}^{(N)}$ is given by $\operatorname{cov}({}_s\hat{B}^{(N)}) = (N-n-1)^{-1} \sigma^2_{n+1/(1)} \Sigma_{11}^{-1}$.

Moreover we can prove the following theorems:

THEOREM 2.1.

$$\boldsymbol{L}(\hat{\boldsymbol{B}}^{(N)}, S^{(N)2}_{n+1/(1)}) = \boldsymbol{L}({}_s\hat{\boldsymbol{B}}^{(N)}, {}_sS^{(N)2}_{n+1/(1)} | \boldsymbol{a}_{11})$$

where $\boldsymbol{L}(\cdot)$ *means the distribution of* $(\cdot)$, $\hat{\boldsymbol{B}}^{(N)}$ *and* $S^{(N)2}_{n+1/(1)}$ *are the usual estimators in the ordinary linear model with fixed regressors* (KLM) *and*

$$\boldsymbol{a}_{11} = \sum_{i=1}^{N} (x_i^{(1)} - \bar{x}^{(1)})(x_i^{(1)} - \bar{x}^{(1)})^T.$$

THEOREM 2.2. 1. $\boldsymbol{L}({}_sA^2_{n+1/(1)}) = \sigma^2_{n+1/(1)} \chi^2_{N-n-1}$;

2. $\boldsymbol{L}({}_s\hat{\boldsymbol{B}}^{(N)} / \boldsymbol{A}_{11} = \boldsymbol{a}_{11}) = N(\boldsymbol{B}, \sigma^2_{n+1/(1)} \boldsymbol{a}_{11}^{-1})$;

3. ${}_sS^{(N)2}_{n+1/(1)}$, ${}_s\hat{\boldsymbol{B}}^{(N)}$ *and* $\boldsymbol{A}_{11}$ *are independent*;

4. $$f_{s\hat{\boldsymbol{B}}}(\lambda) = \frac{\Gamma(\frac{1}{2}N)|\Sigma_{11}|^{1/2}}{\Gamma(\frac{1}{2}(N-n))\pi^{n/2}\sigma^n_{n+1/(1)}\{1+(\lambda-\boldsymbol{B})^T\Sigma_{11}(\lambda-\boldsymbol{B})\sigma^{-2}_{n+1/(1)}\}^{N/2}}$$

for $-\infty < \lambda_j < \infty$, $j = 1, \ldots, n$.

This is the form of a multivariate central t distribution in n variables and $(N-n)$ degrees of freedom.

THEOREM 2.3. *The sequence of estimators* $\{{}_s\hat{\boldsymbol{B}}^{(N)}\}_{N\in\boldsymbol{N}}$ *and* $\{{}_sS^{(N)2}_{n+1/(1)}\}_{N\in\boldsymbol{N}}$ *of the linear models with stochastic regressors* (SLM) *are consistent for* $\boldsymbol{B}$ *and* $\sigma^2_{n+1/(1)}$. [1]

THEOREM 2.4. *Let* $\{\tilde{\varepsilon}^{(N)}\}_{N\in\boldsymbol{N}}$ *be a sequence of i.i.d. random variables belonging to the sequence of models*

$$\{X^{(N)}_{n+1} = {}_s\hat{\boldsymbol{B}}^{(N)} x^{(1)}_{(N)} + \tilde{\varepsilon}^{(N)}\}_{N\in\boldsymbol{N}} \qquad \text{where } \tilde{\varepsilon}^{(N)} \sim F,\ F \text{ is independent of } N.$$

Under the following conditions:

1) *the distribution of* $X^{(1)}$ *has finite second moments*,

2) $m\left(\frac{1}{\sqrt{N}} X^{(1)}_{(N)}\right) \xrightarrow[N\to\infty]{\text{a.e.}} \boldsymbol{0}$ *where* $m(\boldsymbol{A}) = \max\limits_{\substack{1\leqslant i\leqslant N\\ 1\leqslant j\leqslant n}} (a_{ij})$,

3) *there exists an invertible matrix* Σ_{11} *with the property that*

$$\frac{1}{N}(X^{(1)}_{(N)} X^{(1)T}_{(N)}) \xrightarrow[N\to\infty]{\text{a.s.}} \Sigma_{11};$$

then we get

$$\boldsymbol{L}(\sqrt{N}({}_s\hat{\boldsymbol{B}}^{(N)} - \boldsymbol{B})) \xrightarrow[N\to\infty]{} N(0, \sigma^2_{n+1/(1)} \Sigma^{-1}_{11}).$$

THEOREM 2.5.

$$\boldsymbol{L}\left\{\frac{({}_s\hat{\boldsymbol{B}}^{(N)} - \boldsymbol{B}) X^{(1)}_{(N)} X^{(1)T}_{(N)} ({}_s\hat{\boldsymbol{B}}^{(N)} - \boldsymbol{B})^T}{{}_sS^{(N)2}_{n+1/(1)}}\right\} \xrightarrow[N\to\infty]{\text{w}} \boldsymbol{L}\left(\frac{Y\Sigma_{11}Y^T}{\sigma^2_{n+1/(1)}}\right) = \chi^2_n.$$

---

[1] If (A1) is satisfied.

Remarks. 1. It can be shown under fairly general conditions that the F-test is asymptotically valid in the sense that the size of the F-test is not affected asymptotically by non-normality (Arnold (1980)). This means that the Hubers condition (the largest diagonal element of $X_{(N)}^{(1)T}(X_{(N)}^{(1)} X_{(N)}^{(1)T})^{-1} X_{(N)}^{(1)}$ tends to 0 with probability 1) is satisfied and therefore we have

$$F^{(N)} = \frac{(C_s\hat{B}^{(N)} - CB)[C^T(X_{(N)}^{(1)} X_{(N)}^{(1)T})^{-1} C]^{-1}(C_s\hat{B}^{(N)} - B)^T}{k \cdot {}_sS_{n+1/(1)}^{(N)2}} \xrightarrow{v} \frac{\chi^2}{k}$$

for all $(k \times n)$ matrices C that have rank k. (If we consider testing that $CB = d$ where C is a specified $(k \times n)$ matrix of rank k and d is a specified vector, then we get the usual F test.)

2. The maximum-likelihood estimates of the regression coefficients and the conditional variance are identical for the two models under normality assumptions, but their distributions differ considerably. The residuals for the two fitted models also have different distributions. Further, the joint distribution of $\hat{X}_{n+1} = \hat{B}^{(N)} x_{(N)}^{(1)}$ is multivariate normal for the fixed model but unknown for the random case.

If there are no other abnormalities, such as multicollinearity or autocorrelation of the ε_i, then the Theorems 2.1 to 2.5 are able to justify the use of the sample $X_{(N)}$ for the determination of the estimators.

If there are abnormalities, for instance a high degree of multicollinearity, then one can expect incorrect estimators, unreliable test statistics, confidence intervals which are too wide and so on. Consequently, it is necessary to examine the assumption of the model before one can analyse if N is large enough. ([2])

3. Determination of an optimal n

The aim of regression analysis is to find an extrapolation equation for X_{n+1}. It is well known (Hocking (1976)) that when a regression equation with parameters estimated from given data is used for extrapolation beyond the range of this data, the accuracy of such extrapolation (prediction) can be very poor. One of the reasons is that the statistical relationship between X_{n+1} and $X^{(1)}$ within the sample does not necessarily remain the same outside the range of the sample. The consequence is that the variances of the estimates can be large and can contribute a large amount to the expected mean square error of extrapolation. Reducing the mean square error, while preserving the usual maximum-like-lihood estimators, one can include in the extrapolation equation only a subset $X^{(2)}$ of the whole set $X^{(1)}$ of regressors in the model. But the expected mean squared error can be large if the number of variables in the extrapolation equation is sufficiently small.

([2]) New results you can find in Jahn, Riedel (1984).

It can be shown (Oliker (1978)) that there exists a certain optimal number of variables among $X^{(1)}$ in the model, such that the expected mean square error corresponding to the extrapolation equation with these variables is minimal.

THEOREM 3.1 (Oliker (1978)). *The expected mean square error is*

$$E[E\{(X_{n+1}-\hat{X}^{(N)}_{n+1})^2|X^{(1)}_{(N)}\,X^{(1)}\}] = \sigma^2_{n+1/(1)}\frac{N+1}{N}\left(1+\frac{n}{N-n-2}\right)$$

where

$$\hat{X}^{(N)}_{n+1} = {}_s\hat{B}^{(N)}\,x^{(2)}_{(N)}, \qquad X^{(1)T}_{(N)} = (X^{(2)T}_{(N)},\ X^{(3)T}_{(N)}).$$

CONCLUSIONS. 1. $c(n) = 1+\dfrac{n}{N-n-2}$ *is a monotonically increasing function of* n, $c(0)=1$, $c(n)\to\infty$ *when* $n\to N-2$.

2. $\sigma^2_{n+1/(1)} = \sigma^2_{n+1} - \sigma^T_{n+1\cdot(1)}\Sigma^{-1}_{11}\sigma_{n+1\cdot(1)}$ *is a nonincreasing function of* n, *since* Σ_{11} *is positive definite. Therefore, there exists a certain* n_0, $0\leqslant n_0 \leqslant N-2$, *such that the expected mean square error is minimal.*

3. n_0 *depends on the sample size.*

4. $$\frac{N-n-2}{N-m-2}\leqslant\frac{1-R^2_{n+1\cdot(1)}}{1-R^2_{n+1\cdot(2)}}\leqslant 1-R^2_{n+1\cdot(3)/(2)}$$

where $X^{(1)T} = (X^{(2)T},\ X^{(3)T})$, *the dimension of* $X^{(2)}$ *is* m *and* $R_{n+1\cdot(1)}$, $R_{n+1\cdot(2)}$ *are multiple correlation coefficients and* $R_{n+1\cdot(3)/(2)}$ *is the partial-multiple correlation coefficient.*

Theorem 3.1 can be used for the calculation of n_0. But with n_0 we do not have those variables which can be eliminated.

The following lemmas enable us to construct a simple sequential procedure for the elimination of variables.

LEMMA 3.1.

$$(N-n)\,{}_sS^{(N)2}_{n+1/(1)} = X^{(N)}_{n+1}\left(I - X^{(1)T}_{(N)}\left(X^{(1)}_{(N)}\,X^{(1)T}_{(N)}\right)^{-1}X^{(1)}_{(N)}\right)X^{(N)T}_{n+1}.$$

LEMMA 3.2.

$$(N-m)\,{}_sS^{(N)2}_{n+1/(2)} - (N-n)\,{}_sS^{(N)2}_{n+1/(1)} = \left({}_s\hat{B}^{(N)}_{(1)}\,X^{(1)}_{(N)} - {}_s\hat{B}^{(N)}_{(2)}\,X^{(2)}_{(N)}\right)X^{(N)}_{n+1} = \mathrm{Red}_m.$$

CONCLUSIONS. 1. $\mathrm{Red}_m + (N-m)\,{}_sS^{(N)2}_{n+1/(1)} = (N-m)\,{}_sS^{(N)2}_{n+1/(2)}$.

2. *The reduction of the sum of squares of deviations of the components of* $X^{(N)}_{n+1}$ *from the regression hyperplane* ${}_s\hat{B}^{(N)}_{(2)}\,X^{(2)}_{(N)}$ *by means of addition of regressors of* $X^{(3)}_{(N)}$ *is*

$$\left(N-(n-m)\right){}_sS^{(N)2}_{n+1/(3)} - (N-n)\,{}_sS^{(N)2}_{n+1/(1)} = \left({}_s\hat{B}^{(N)}_{(1)}\,X^{(1)}_{(N)} - {}_s\hat{B}^{(N)}_{(3)}\,X^{(3)}_{(N)}\right)X^{(N)T}_{n+1}.$$

3. *Let* $S_n := \dfrac{(N+1)(N-2)}{N(N-n-2)}\,{}_sS^{(N)2}_{n+1/(1)}$ *be the estimator of the expected*

mean squared error of Theorem 3.1. *Then we get*

$$S_m = \frac{[\mathrm{Red}_m - (N-n)\,{}_sS^{(N)2}_{n+1/(1)}]\,[(N+1)(N-2)}{N(N-m-2)(N-m)}.$$

4. $\mathrm{Red}_1 = \tau_j = \dfrac{{}_sS^{(N)2}_{n+1/(1)}\,{}_s\hat{B}_j^2}{S^2_{{}_s\hat{B}_j}}$.

With these lemmas we can construct a sequential procedure for the determination of n_0, see Thompson (1978).

4. Determination of sample size

The sample size problem in the linear model with stochastic regressors is very difficult. First, it is not known what specific $\{X_{ij}\}$, $i = 1,\ldots,N$, $j = 1,\ldots,n$ will be sampled, or, in some cases, what range of values is reasonable. Second, the distribution of $\hat{X}_{n+1}$ is not known.

4.1. Estimation of the mean of a multinormal distribution – fixed sample size procedure. We consider first the n-dimensional multinormal distribution with mean μ and covariance matrix $\Sigma^1 = \sigma^2\Sigma$, where σ^2 is a scalar and Σ is a known $n \times n$ matrix (positive definite). First, we suppose that σ^2 is known. If we use $\bar{X}^{(N)} = \dfrac{1}{N}\sum_{i=1}^{N} X_i$, the expected cost for fixed N is (Wang (1980))

$$E_N(L) = \frac{A\sigma^\eta}{N^{\eta/2}}\,E\left[\frac{N(\bar{X}^{(N)}-\mu)^T\,\Sigma^{-1}\,(\bar{X}^{(N)}-\mu)^{\eta/2}}{\sigma^2}\right] + N$$

$$= A(n,\eta)\,\sigma^\eta/N^{\eta/2} + N \tag{4.1}$$

where

$$A(n,\eta) = \frac{A\cdot 2^{\eta/2}\,\Gamma\left(\frac{1}{2}(n+\eta)\right)}{\Gamma\left(\frac{1}{2}n\right)},$$

A and η are positive constants, $[A(\mu-\hat{\mu})(\Sigma^1)^{-1}(\mu-\hat{\mu})]^{\eta/2} + N$ the cost function. Treating N as a continuous variable, we differentiate $E_N(L)$ with respect to N and set the derivative equal to zero, obtaining the relation

$$-A(n,\eta)\,\sigma^\eta\,\frac{\eta}{2N^{(1+\eta)/2}} + 1 = 0.$$

From this we get the optimal sample size N_0

$$N_0 = [A(n,\eta)\,\sigma^\eta\cdot\tfrac{1}{2}\eta]^{2/(\eta+2)}. \tag{4.2}$$

4.2. Determination of sample size with probability inequalities. Let $X_1, \ldots, X_N$ be a sequence of n-dimensional random vectors ($n \geqslant 1$) possibly mutually dependent with joint distribution belonging to the family $\{P_\theta\colon \theta \in \Theta\}$ and let T_N be a consistent estimator of θ, which is a function of $X_1, \ldots, X_N$. Then, for any Borel-measurable set $A \subseteq R^k$ containing a certain neighbourhood of the origin and any α, $0 < \alpha < 1$, there exists a smallest natural number N_0 such that

$$P_N := P_\theta(\{T_N - \theta \in A\}) \geqslant \alpha$$

for $N \geqslant N_0$. We call the number $N_0 = N_0(A, \alpha)$ the *necessary sample size* (n.s.s.) corresponding to the precision A and the significance level α with respect to $T_N - \theta$. We consider the precision of the shape $A = \bigtimes_{j=1}^{k} [g_j, r_j]$, where $-\infty \leqslant g_j < r_j \leqslant \infty$. The n.s.s. depends essentially on the distribution of T_N, which is a function of θ. Hence, the n.s.s. can be determined only if θ is known. By means of the probability inequality it is possible to give bounds for P_N (Bergmann, Fritzsche, and Riedel (1984)). If we have a lower bound $Q_N \leqslant P_N$, then from the definition of the n.s.s. it follows that $Q_{N_0-1} < \alpha$; hence $N_0 - 1 \leqslant \sup\{N\colon Q_N < \alpha\}$.

We note that the right-hand side is a maximum of a finite set if $\varliminf_{N\to\infty} Q_N > \alpha$. In many cases the condition $Q_N < \alpha$ can be transformed into a computable form. In particular cases it is possible to obtain the supremum mentioned above.

4.2.1. *Upper bounds of the n.s.s. if T_N has a known covariance matrix.* We suppose that the covariance matrix of T_N, $\Sigma^{(N)}$ exists and the precision

$$A_0 = \bigtimes_{j=1}^{k} [-E_j, E_j]$$

with positive real numbers E_j is given. Using modified Chebyshev inequalities, Bergmann et al. (1984) derived from Corollary 2 of Theorem 7.2.2 in Tong (1980) the bounds of the n.s.s.

THEOREM 4.1. *The n.s.s. $N_0 = N_0(A_0, \alpha)$ satisfies*

$$1 - \alpha < \frac{\sqrt{V^{N_0-1}} + \sqrt{(k-1)(k \cdot w^{(N_0-1)} - v^{(N-1)})^2}}{k^2} \leqslant w^{N_0-1},$$

where

$$V^{(N)} = V(\Sigma_+^{(N)}) = \sum_{i,j=1}^{k} \sigma'_{ij}, \qquad w^{(N)} = \operatorname{tr}(\Sigma_+^{(N)}),$$

$$\Sigma_+^{(N)} = D\Sigma^{(N)} D = (\sigma'_{ij}), \qquad D = \operatorname{diag}\left(\frac{1}{E_j}\right).$$

COROLLARY. *If in addition for all N the matrix $\Sigma^{(N)}$ has the representation*

$$\Sigma^{(N)} = \frac{\Sigma}{N}$$

with a certain nonnegative definite and symmetric matrix, then

$$N_0 - 1 < \frac{\sqrt{V} + \sqrt{(k-1)(k\cdot w - V)^2}}{(1-\alpha)k^2} \leqslant \frac{w}{1-\alpha}$$

where $\Sigma_+ = D\Sigma D$, $V = V(\Sigma_+)$, $w = \operatorname{tr}(\Sigma_+)$.

(*The equality sign holds if T_N has uncorrelated components.*)

4.2.2. *Upper bounds of the n.s.s. if T_N has a normal distribution.*

We suppose that $T_N \sim N_k(\theta, \Sigma^{(N)})$ and denote

$$\Phi_0(x) = \int_0^x e^{-t^2/2}\, dt$$

and

$$\sigma^{(N)} = \max_{j=1,\ldots,k} \sigma_{jj}^{(N)},$$

and denote by $\chi^2_{r,\alpha}$ the α-quantile of the chi-square distribution with r degrees of freedom.

From Theorem 2.2.2. in Tong (1980) we get

THEOREM 4.2. *For the n.s.s. $N_0 = N_0(A_0, \alpha)$ the following relation holds*

$$\prod_{j=1}^{k} \Phi_0\left(\frac{1}{\sqrt{\sigma_{jj}^{(N_0-1)}}}\right) < \frac{\alpha}{2^k}.$$

This inequality implies

$$1 < \chi^2_{1,\alpha^{1/k}} \sigma^{(N_0-1)}.$$

THEOREM 4.3. *The n.s.s. satisfies*

$$1 < H_{\lambda_1^{(N_0-1)},\ldots,\lambda_k^{(N_0-1)},\alpha},$$

where $H_{\lambda_1^{(N_0-1)},\ldots,\lambda_k^{(N_0-1)},\alpha}$ is the α fractile of the distribution function as a mixture of chi-square distribution functions and $\lambda_j^{(N)}$, $0 \leqslant \lambda_1^{(N)} \leqslant \ldots \leqslant \lambda_k^{(N)}$ are the eigenvalues of Σ_+. If $Y_1, \ldots, Y_r$ are independent chi-square distributed random variables with one degree of freedom, then $H_{\gamma_1,\ldots,\gamma_k}(x)$ is the distribution function of $\sum_{j=1}^{k} \gamma_j Y_i$ where γ_j are real numbers.

Remark. The simple inequality

$$H_{\lambda_1^{(N_0-1)},\ldots,\lambda_k^{(N_0-1)},\alpha} \leqslant \lambda_k^{(N_0-1)} \chi^2_{k,\alpha}$$

implies a greater bound then Theorem 4.2 because

$$\lambda_k^{(N_0-1)} = \max_{x \in R^k} \frac{x^T \Sigma_+^{(N_0-1)} x}{x^T x} \geqslant \sigma^{(N_0-1)}$$

and

$$\chi^2_{k,\alpha} \geqslant \chi^2_{1,\alpha^{1/k}}.$$

THEOREM 4.4. *Suppose that the matrices $\Sigma^{(N)}$ are positive definite for all natural numbers N. The n.s.s. $N_0 = N_0(A_0, \alpha)$ satisfies*

$$\int_0^1 (\cosh \tfrac{1}{2} x^2 m^{(N_0-1)} - \sinh \tfrac{1}{2} x^2 M^{(N_0-1)}) x^{k-1}\, dx < \frac{\alpha}{k} |\Sigma_+^{(N_0-1)}| \left(\frac{\pi}{2}\right)^{k/2}$$

where $m^{(N)} := \dfrac{1}{\sigma^{(N)}}$, $M^{(N)} := \max\limits_{x \in A_0} x^T (\Sigma^{(N)})^{-1} x$.

In a similar way it is possible to find lower bounds of the n.s.s. if T_N has a normal distribution. The lower bounds are useful for the estimation of the goodness of upper bounds.

4.3. Cross validation approach to sample size. The aim of cross validation is to assess the statistical procedure which is based on a statistical model

$$P = \{P_\theta\colon\ \theta \in \Theta\}, \qquad \Theta \subseteq R^{n+1}$$

for the data $x_i \in D_N := \{(x_{n+1\cdot i}, x_i^{(1)T}),\ i = 1, \ldots, N\}$, whose conceptual ingredients are probability and parameter. For this assessment we introduce the decision function

$$\delta = \delta(X, D_N)$$

and the loss function $L(X_{n+1}, d)$, where $d \in D_X$, the set of all possible decisions taken for $(X_{n+1}, X^{(1)T})$. $L(x_{n+1}, d)$ is observable. The question is now, how is δ to be assessed when only D_N is available?

"The philosophy (Stone (1977)) of cross validation is to decline to make the assessment of $\delta(X^{(1)}, D_N)$ on the basis of the quantities $L(x_{n+1\cdot i}, \delta(x_i, D_N))$, $i = 1, \ldots, N$, on the ground that $\delta(x_i^{(1)}, D_N)$ is influenced by $x_{n+1\cdot i}$, thereby affecting the honesty of evaluating $\delta(x_i^{(1)}, D_N)$ as a decision for the $(x_i^{(1)}, x_{n+1\cdot i})$."

One of the cross-validatory assessment criteria is

$$C = \sum_{i=1}^{N} w_i L(x_{n+1\cdot i}, \delta(x_i^{(1)}, D_{N\setminus i}))$$

where $D_{N\setminus i}$ denote D_N with $(x_i^{(1)}, x_{n+1\cdot i})$ omitted, and $w_i = w(x_i^{(1)}, D_{N\setminus i})$, $i = 1, \ldots, N$, $\sum_{i=1}^{N} w_i = 1$ are nonnegative weights.

4.3.1. *A cross-validation approach to sample size determination for regression models.* We only want to use cross-validation for the determination of sample size.

4.3.1.1. The fixed model. The regression function estimate is

$$\hat{X}_{n+1} = \hat{\boldsymbol{B}} x_{(N)}^{(1)}, \qquad \text{where} \quad \hat{\boldsymbol{B}} = (x_{(N)}^{(1)} x_{(N)}^{(1)T})^{-1} x_{(N)}^{(1)} X_{n+1}^{(N)T}$$

and

$$\operatorname{var}(\hat{X}_{n+1} | X^{(1)} = x^{(1)}) = \operatorname{var}(\hat{\boldsymbol{B}} x_{(N)}^{(1)}) = \sigma_{n+1/(1)}^2 x_{(N)}^{(1)T} (x_{(N)}^{(1)} x_{(N)}^{(1)T})^{-1} x_{(N)}^{(1)}.$$

In order to estimate, on an a priori basis, the necessary sample size, such that for any arbitrary chosen constant δ

$$P(\{|\hat{X}_{n+1} - \mu_{n+1/(1)}| < \delta\}) = 1 - \alpha,$$

the experimenter has only to estimate $\sigma_{n+1/(1)}^2$ and then solve

$$(\sigma_{n+1/(1)} t_{N-1,(1-\alpha)/2})/N^{1/2} = \delta \tag{4.4}$$

for N.

4.3.2.2. The random model. The sample size problem in this model is much more complex because it is not known what specific $\{x_i^{(1)}\}$ will be sampled or, in some cases, what range of values are reasonable, and the distribution of $\hat{X}_{n+1}$ is not known.

By Park and Dudycha (1974) a solution to the sample size problem was developed to the extent that it reduces the problem to the evaluation of a probability density with one unknown parameter, namely the squared multiple correlation coefficient R^2. For the estimation of the conditional expectation value we can write (we use our model (1.3)):

$$\hat{E}(X_{n+1} | X^{(1)} = x^{(1)}) = \hat{\boldsymbol{B}}_0 + {}_s\hat{\boldsymbol{B}} x_{(N)}^{(1)}. \tag{4.5}$$

It can be shown that

$$E({}_s\hat{\boldsymbol{B}}\, {}_s\hat{\boldsymbol{B}}^T) = \boldsymbol{B}\boldsymbol{B}^T + \sigma_{n+1/(1)}^2 \Sigma_{11}^{-1}/(N-n-2) \tag{4.6}$$

and

$$E({}_s\hat{\boldsymbol{B}}) = \boldsymbol{B}.$$

From this follows

$$E({}_s\hat{\boldsymbol{B}}\, {}_s\hat{\boldsymbol{B}}^T) - E({}_s\hat{\boldsymbol{B}})\, E({}_s\hat{\boldsymbol{B}}^T) = \Sigma_{{}_s\hat{\boldsymbol{B}}} = \sigma_{n+1/(1)}^2 \Sigma_{11}^{-1}/(N-n-2). \tag{4.7}$$

Cross validation as a test of an estimated regression function consists of drawing the second random sample $(X_{n+1\cdot k}, X_{1k}, \ldots, X_{nk})$, $k = 1, \ldots, N$, and

correlating the observed dependent variable $\{X_{n+1\cdot k}\}$ with its estimated values obtained from the original equation. The resultant correlation $\hat{R}_c({}_s\hat{\boldsymbol{B}})$ is an estimate of the validity of the derived equation in the population. $R_c({}_s\hat{\boldsymbol{B}})$ estimates

$$R_c({}_s\hat{\boldsymbol{B}}) = \frac{E(X_{n+1}\,\hat{X}_{n+1})}{[E(X_{n+1}^2)\,E(\hat{X}_{n+1}^2)]^{1/2}},$$

$$R_c({}_s\hat{\boldsymbol{B}}) = \frac{E(X_{n+1\,s}\hat{\boldsymbol{B}}x^{(1)})}{\{\sigma_{n+1}^2\,E[{}_s\hat{\boldsymbol{B}}x_{(N)}^{(1)}\,x_{(N)}^{(1)T}\,{}_s\hat{\boldsymbol{B}}^T]\}^{1/2}} = \frac{\sigma_{n+1\cdot(1)\,s}^T\hat{\boldsymbol{B}}^T}{[\sigma_{n+1\,s}^2\hat{\boldsymbol{B}}\,\Sigma_{11\,s}\hat{\boldsymbol{B}}^T]^{1/2}}. \tag{4.8}$$

$R_c({}_s\hat{\boldsymbol{B}})$ is a population parameter for a given set ${}_s\hat{\boldsymbol{B}}$, but is a random variable over the possible vector of ${}_s\hat{\boldsymbol{B}}$.

LEMMA 4.1. *$R_c({}_s\hat{\boldsymbol{B}}) \leqslant R_{n+1\cdot(1)}$ where $R_{n+1\cdot(1)}$ is the maximum correlation possible in the population.*

THEOREM 4.5.

$$P(\{R_c({}_s\hat{\boldsymbol{B}}) \leqslant \lambda\}) =: F_{R_c^2}(\lambda), \qquad F_{R_c^2}(\lambda) = F\left(\lambda\left(\frac{n-1}{R_{n+1\cdot(1)}^2 - \lambda}\right)\right)$$

where F represents the noncentral F distribution.

Proof. See Park and Dudycha (1974). ∎

COROLLARY.
$$R_c^2({}_s\hat{\boldsymbol{B}}) = \frac{R_{n+1\cdot(1)}^2}{1+\dfrac{n-1}{F_{1,n-1,\delta_1}}}$$

where $F_{1,n-1,\delta_1}$ is distributed as a noncentral F with the noncentrality parameter

$$\delta_1 = \left|(N-n-2)^{1/2}\frac{R_{n+1\cdot(1)}}{(1-R_{n+1\cdot(1)}^2)^{1/2}}\right|.$$

From this corollary we get for any positive ε

$$P(\{R_{n+1\cdot(1)} - R_c({}_s\hat{\boldsymbol{B}}) \leqslant \varepsilon\}) = P\left(\left\{-(n-1)^{1/2}\left[\frac{R_{n+1\cdot(1)}^2}{\varepsilon} - 1\right]^{1/2} \leqslant t_{n-1,\delta_1} \leqslant (n-1)^{1/2}\left[\frac{R_{n+1\cdot(1)}^2}{\varepsilon} - 1\right]^{1/2}\right\}\right).$$

Using the noncentral t distribution, we get

$$N^* = (1-R_{n+1\cdot(1)}^2)\,\delta_1^2/R_{n+1\cdot(1)}^2 + (n+2)$$

as a function of $R_{n+1\cdot(1)}$, ε and n.

5. Sequential sampling procedures

5.1. The sequential decision problem. Using the terminology of Ferguson (1967), we shall formulate our problem as a sequential decision problem. The elements are as follows ([3])

1. $\Theta = \{\theta = (\mu_{n+1/(1)}, \sigma^2_{n+1/(1)})^T\colon\ -\infty < \mu_{n+1/(1)} < \infty, 0 < \sigma^2_{n+1/(1)} < \infty\}$ is the parameter space.

2. $A = \Theta$ is the space of terminal actions. The statistician chooses the parameter $\hat{\theta} \in \Theta$ if he stops sampling, $\hat{\theta} = ({}_s\hat{\boldsymbol{B}}, {}_sS^2_{n+1/(1)})^T$. Let the true parameter be θ_0, then we have to consider two hypotheses: H_0: $\theta = \theta_0$ against H_1: $\theta = \theta_1$, $\theta_0 \neq \theta_1$.

3. Let $L(\boldsymbol{B}, {}_s\hat{\boldsymbol{B}})$ be the loss function. It is a real-valued function on $\Theta \times A$ representing the loss when the statistician makes his decision. L can be defined by

$$
\begin{aligned}
&L(\boldsymbol{B}, {}_s\hat{\boldsymbol{B}}) \\
&= \int [B_0 + B_1 x_1 + \ldots + B_n x_n - (\hat{B}_0 + \hat{B}_1 x_1 + \ldots + \hat{B}_n x_n)]^2 \, dF(x_1, \ldots, x_n) \\
&= \sum_{j=1}^{n} \sum_{j'=1}^{n} \mu_{jj'} (B_j - \hat{B}_j)(B_{j'} - \hat{B}_{j'}) = (\boldsymbol{B} - {}_s\hat{\boldsymbol{B}})\, \boldsymbol{M} (\boldsymbol{B} - {}_s\hat{\boldsymbol{B}})^T
\end{aligned}
$$

where

$$
\boldsymbol{M} = \begin{pmatrix} 1 & \mu_{01} & \cdots & \mu_{0n} \\ \mu_{10} & \mu_{11} & \cdots & \mu_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ \mu_{n0} & \mu_{n1} & \cdots & \mu_{nn} \end{pmatrix}
$$

and $\mu_{jj'} = E(X_j X_{j'})$, $j, j' = 0, 1, \ldots, n$; $X_0 \equiv 1$, F is the distribution function of $X^{(1)} = (X_1, \ldots, X_n)^T$, $\boldsymbol{B} = (B_0, B_1, \ldots, B_n)$ are the unknown parameters of the conditional expectation $E(X_{n+1} | X^{(1)} = x^{(1)})$ and $\hat{\boldsymbol{B}} = (\hat{B}_0, \hat{B}_1, \ldots, \hat{B}_n)$ are the estimators for these coefficients. $\boldsymbol{M}$ is positive definite since

a) for any $(a_0, a_1, \ldots, a_n)$ not equal to zero we have $P(a + a_1 x_1 + \ldots + a_n x_n = 0) = 0$,

b) $E(X_j^2) < \infty$, $j = 1, \ldots, n$.

Denoting $\boldsymbol{B}$ by $\boldsymbol{B}_0$ under hypothesis H_0 and by $\boldsymbol{B}_1$ under H_1, we can suppose that

$$
L(\boldsymbol{B}_0, \boldsymbol{B}_0) = L(\boldsymbol{B}_1, \boldsymbol{B}_1) = -r, \qquad L(\boldsymbol{B}_0, \boldsymbol{B}_1) = L(\boldsymbol{B}_1, \boldsymbol{B}_0) = l.
$$

4. $X_1, \ldots, X_N, \ldots$ is a sequence of random variables available to the statistician. They are independent and distributed as $X_i \sim N_{n+1}(\mu, \Sigma)$.

Under the condition $X^{(1)} = x^{(1)}$, X_{n+1} has the density

$$
f_{n+1/(1)}(x) = f_{n+1,(1)}(x_{n+1}, x^{(1)}) / f_{(1)}(x^{(1)}).
$$

([3]) See also Hee and Hordijk (1975).

For brevity we write

$$P_{n+1/(1)} = f_{n+1/(1)} = N_1\left(\mu_{n+1} + \sigma^T_{n+1\cdot(1)}\Sigma_{11}^{-1}(x^{(1)} - \mu^{(1)});\ \sigma^2_{n+1/(1)}\right).$$

The statistician will choose a decision rule, which may be divided into two parts:

- a stopping rule for sampling;
- a terminal decision rule.

5. A stopping rule ψ is a sequence of functions

$$\psi = \{\psi_0,\ \psi_1(x_1),\ \psi_2(x_1, x_2),\ \psi_3(x_1, x_2, x_3), \ldots\}$$

where x_i is the realization of $X_i = (X_{1i}, \ldots, X_{n+1\cdot i})^T$ and $\psi(x_1, \ldots, x_i) = 0$ or 1.

If $\psi_i(x_1, \ldots, x_i) = 1$ then $\psi_N(x_1, \ldots, x_N) = 1$ for all $N \geqslant i$, i.e., if $x_1, \ldots, x_i$ are observed, the statistician will stop sampling if $\psi_i(x_1, \ldots, x_i) = 1$ and he will make another observation if $\psi_i(x_1, \ldots, x_i) = 0$. If $\psi_0 \equiv 1$, then he makes his decision without sampling. Using ψ, we can define a random variable τ, called a *stopping time*, by

$$\tau = \min\{i:\ \psi_i(x_1, \ldots, x_i) = 1\}.$$

Hence τ is the time stopping the sampling if the stopping rule ψ is used.

6. A terminal decision rule δ is a sequence of functions

$$\delta = \{\delta_0,\ \delta_1(x_1),\ \delta_2(x_1, x_2),\ \delta_3(x_1, x_2, x_3), \ldots\}$$

where $\delta_i(x_1, \ldots, x_i)$ is a function with values in A. For each sequence of realizations $x_1, \ldots, x_N$ the statistician specifies by δ_i the decision he will make under the condition that he stops sampling at the time t. So a decision rule is a pair (ψ, δ).

7. The risk function $R(\theta, (\psi, \delta))$ of a decision rule (ψ, δ) is the expected loss plus the expected sampling costs when θ is the true parameter:

$$R(\theta(\psi, \delta)) = E_\theta[L(\theta, \delta_\tau(x_1, \ldots, x_\tau) + \tau c]$$

where τ is the stopping time defined by ψ. In general, and also in our case, there exists no decision rule (ψ^*, δ^*) such that

$$R(\theta, (\psi^*, \delta^*)) \leqslant R(\theta, (\psi, \delta));\ \forall\,\theta \in \Theta,\ \forall(\psi, \delta).$$

Hence we have to choose other criteria to determine the best decision rule.

Such criteria are:

a) the minimax rule (ψ^*, δ^*), for which

$$\inf_{(\psi,\delta)} \max_{\theta\in\Theta} R(\theta, (\psi, \delta)) = \max_{\theta\in\Theta} R(\theta, (\psi^*, \delta^*));$$

b) the Bayes rule with respect to the distribution π over Θ if

$$\sum_{\theta\in\Theta} R(\theta, (\psi^*, \delta^*))\pi(\theta) = \inf_{(\psi,\delta)} \sum_{\theta\in\Theta} R(\theta, (\psi, \delta))\pi(\theta).$$

5.2. The Bayesian approach. We assume that $\hat{\theta}$ is a random variable which takes on the value θ_0 with probability $\pi(\theta_0)$ and the value θ_1 with probability $\pi(\theta_1)$, $[\pi(\theta_0) = \pi,\ \pi(\theta_1) = 1-\pi,\ 0 \leqslant \pi \leqslant 1]$. Note that the simultaneous distribution of $\hat{\theta}, x_1, x_2, \ldots$ is completely determined by the assumption in 4°, i.e., by the distribution assumption. The distribution of $\hat{\theta}$ is called the *prior distribution* and it is completely determined by π.

The Bayes risk of a decision rule (ψ, δ) depends on π and is defined analogously to 7° as the expected loss plus the expected sampling costs:

$$R(\pi, (\psi, \delta)) = E\left[L(\hat{\theta}, \delta_\tau(x_1, \ldots, x_\tau)) + \tau c\right].$$

Conditioning on $\hat{\theta} = \theta$ gives

$$R(\pi, (\psi, \delta)) = \sum_{\theta \in \Theta} E\left[L(\theta, \delta_\tau(x_1, \ldots, x_\tau)) + \tau c | \hat{\theta} = \theta\right] \pi(\theta).$$

Compare this formula with (7°b) – hence the Bayes decision rule (ψ^*, δ^*) satisfies

$$R(\pi, (\psi^*, \delta^*)) = \inf_{(\psi, \delta)} R(\pi, (\psi, \delta)).$$

The choice of the prior distribution of $\hat{\theta}$ (or equivalently the choice of π) depends on the (subjective) opinion of the decision maker. Sometimes he has prior information about the two hypotheses, which he can translate in a prior distribution.

5.3. The sequential decision process. We now introduce the posterior distribution of $\hat{\theta}$, given the sequence of observations

$$X_1 = x_1, \ldots, X_N = x_N \qquad \text{where} \quad X_i^T = (X_{n+1}, X^{(1)T}),\ i = 1, \ldots, N,$$

$$y_N = P(\{\hat{\theta} = \theta_0 | X_1 = x_1, \ldots, X_N = x_N\}), \qquad N = 1, 2, 3, \ldots$$

With the rule of Bayes it is easy to verify that

$$y_N = \frac{P(\{X_1 = x_1, \ldots, X_N = x_N | \hat{\theta} = \theta_0\})\, P(\{\hat{\theta} = \theta_0\})}{\sum_{\theta \in \Theta} P(\{X_1 = x_1, \ldots, X_N = x_N | \hat{\theta} = \theta\})\, P(\{\hat{\theta} = \theta\})}, \qquad N = 1, 2, \ldots$$

If we use the distribution from 4°, then we get

$$y_N = \frac{\prod_{i=1}^{N} f_{n+1/(1)}(x; \theta_0)\, P(\{\hat{\theta} = \theta_0\})}{\sum_{\theta \in \Theta} \prod_{i=1}^{N} f_{n+1/(1)}(x; \theta)\, P(\{\hat{\theta} = \theta\})}$$

$$= \frac{\pi \prod_{i=1}^{N} (2\pi)^{-(n+1)/2} (\sigma^2_{n+1/(1)})^{-1/2} \exp\{-\frac{1}{2}(x_{n+1\cdot i} - \boldsymbol{B}_0 x_i^{(1)})^2 \sigma^{-2}_{n+1/(1)}\}}{\sum_{\theta \in \Theta} \pi \prod_{i=1}^{N} (2\pi)^{-(n+1)/2} (\sigma^2_{n+1/(1)})^{-1/2} \exp\{-\frac{1}{2}(x_{n+1\cdot i} - \boldsymbol{B} x_i^{(1)})^2 \sigma^{-2}_{n+1/(1)}\}}$$

$$= \frac{\pi (2\pi)^{-N(n+1)/2} (\sigma^2_{n+1/(1)})^{-N/2} \exp\{-2\sigma^{-2}_{n+1/(1)} \sum_{i=1}^{N} (x_{n+1\cdot i} - \boldsymbol{B}_0 x_i^{(1)})^2\}}{\sum_{\theta \in \Theta} \pi (2\pi)^{-N(n+1)/2} (\sigma^2_{n+1/(1)})^{-N/2} \exp\{-2\sigma^{-2}_{(n+1)/(1)} \sum_{i=1}^{N} (x_{n+1\cdot i} - \boldsymbol{B} x_i^{(1)})^2\}}$$

where $\theta = (\boldsymbol{B}, \sigma^2_{n+1/(1)})$.

With the two hypotheses,

$$H_0\colon\ \theta = \theta_0;\ \sigma^2_{n+1/(1)} > 0; \qquad H_1\colon\ \theta = \theta_1;\ \sigma^2_{n+1/(1)} > 0,\ \theta \neq \theta_0,$$

we get

$$y_N = \pi \exp\{-\sigma^{-2}_{n+1/(1)} \sum_{i=1}^{N} (x_{n+1\cdot i} - \boldsymbol{B}_0 x_i^{(1)})^2\} \times$$

$$\times \Big[\pi \exp\{-\sigma^{-2}_{n+1/(1)} \sum_{n+1/(1)}^{N} (x_{n+1\cdot i} - \boldsymbol{B}_0 x_i^{(1)})^2 +$$

$$+ (1-\pi) \exp\{-2\sigma^{-2}_{(n+1)/(1)} \sum_{i=1}^{N} (x_{n+1\cdot i} - \boldsymbol{B}_1 x_i^{(1)})^2\}\Big]^{-1}.$$

We now put $y_0 = \pi$,

$$y_N = \frac{\exp\{-2\sigma^{-2}_{n+1/(1)} \sum_{i=1}^{N} (x_{n+1\cdot i} - \boldsymbol{B}_0 x_i^{(1)})^2\}}{\exp\{-2\sigma^{-2}_{n+1/(1)} \sum_{i=1}^{N} (\sqrt{2}\, x_{n+1\cdot i} - (\boldsymbol{B}_0 + \boldsymbol{B}_1) x_i^{(1)})^2\}}.$$

Now the value y_{N+1} can be computed from y_N and x_{N+1} by

$$y_{N+1} = \frac{\exp\{-2\sigma^{-2}_{n+1/(1)} (x_{n+1\cdot N+1} - \boldsymbol{B}_0 x^{(1)}_{N+1})^2\}}{\exp\{-2\sigma^{-2}_{n+1/(1)} (\sqrt{2}\, x_{n+1\cdot N+1} - (\boldsymbol{B}_0 + \boldsymbol{B}_1) x^{(1)}_{N+1})^2\}}, \qquad N = 1, 2, \ldots$$

So y_{N+1} is the prior probability of $\hat{\theta} = \theta$ for the $(N+1)$th experiment.

Considering the sequence $y_1, y_2, \ldots$ as functions of the random variables instead of their realizations, we get a sequence of random variables $Y_0, Y_1, Y_2, \ldots$ recursively defined by $P(\{Y_0 = \pi\}) = 1$,

$$Y_{n+1} = \frac{\exp\{-2\sigma^{-2}_{n+1/(1)} (X_{n+1\cdot N+1} - \boldsymbol{B}_0 X^{(1)}_{N+1})^2\}}{\exp\{-2\sigma^{-2}_{n+1/(1)} (\sqrt{2}\, X_{n+1\cdot N+1} - (\boldsymbol{B}_0 + \boldsymbol{B}_1) X^{(1)}_{N+1})^2\}}$$

$$= \exp\{-2\sigma^{-2}_{n+1/(1)} [(X_{n+1\cdot N+1} - \boldsymbol{B}_0 X^{(1)}_{N+1})^2 -$$

$$- (\sqrt{2}\, X_{n+1\cdot N+1} - (\boldsymbol{B}_0 + \boldsymbol{B}_1) X^{(1)}_{N+1})^2]\}.$$

Without proof we state some well-known facts from statistical decision theory (Ferguson (1927)).

a) Consider the sequence $Y_0, Y_1, \ldots$ For each sequential decision rule (ψ, δ) based on $X_1, X_2, \ldots$ there is a sequential decision rule (ψ, δ_0) based on $Y_0, Y_1, \ldots$ which is as good as (ψ, δ) since there is a one-one correspondence between the two sequences. So we have only to consider $Y_0, Y_1, \ldots$ when we are searching for the Bayes rule.

b) If for $N = 1, 2, \ldots, \delta_N(Y_1, \ldots, Y_N)$ is a Bayes rule with respect to π for the decision problem based on the fixed sample size of N observations, then for any stopping rule ψ the risk $R(\pi, (\psi^*, \delta^*))$ is minimized by $\delta = (\delta_0, \delta_1, \delta_2, \ldots)$

c) When we consider also randomized action rules and stopping rules there exists no randomized pair (ψ, δ) with a lower risk if the loss functions is bounded and if there exists for each N a fixed sample size Bayes rule.

d) It is easy to verify that for a fixed sample size of N observations the Bayes risk with respect to π is

$$E_\pi[\min\{-rY_N + l(1-Y_N),\ -r(1-Y_N) + lY_N\}] + Nc,$$

where the subscript π indicates the dependence of the prior distribution and r and l is the reward or the penalty $(r \geqslant l \geqslant 0)$.

Hence we may formulate our sequential decision problem in the following way:

Search for the stopping time τ_0 such that

$$E_\pi[\max\{rY_\tau - l(1-Y_\tau),\ r(1-Y_\tau) - lY)\} - \tau c]$$

attains for τ_0 its maximum (τ is a random variable).

Note that we are maximizing the expected return instead of minimizing the risk.

5.4. Connections with Wald's sequential probability ratio test (SPRT). Wald's sequential ratio test is useful in the same situation: two simple hypotheses H_0: $\boldsymbol{B}_0$, H_1: $\boldsymbol{B}_1$, $\boldsymbol{B}_0 \neq \boldsymbol{B}_1$. In this test the likelihood ratio test

$$\lambda_N(x_1, \ldots, X_N) = \frac{\exp\left\{-\frac{1}{2}\sigma_{n+1/(1)}^{-2} \sum_{i=1}^{N} (x_{n+1\cdot i} - \boldsymbol{B}_0 x_i^{(1)})^2\right\}}{\exp\left\{-\frac{1}{2}\sigma_{n+1/(1)}^{-2} \sum_{i=1}^{N} (x_{n+1\cdot i} - \boldsymbol{B}_1 x_i^{(1)})^2\right\}},$$

$$\log \lambda_N = z_N = -\tfrac{1}{2}\sigma_{n+1/(1)}^{-2} \sum_{i=1}^{N} [2x_{n+1\cdot i} - (\boldsymbol{B}_1 - \boldsymbol{B}_0) x_i^{(1)} + (\boldsymbol{B}_0 - \boldsymbol{B}_1) x_i^{(1)} x_i^{(1)T} (\boldsymbol{B}_0 - \boldsymbol{B}_1)^T]$$

determines, together with the real numbers A and B $0 < A \leqslant 1 \leqslant B < \infty$, the stopping time

$$\psi_N(x_1, \ldots, x_N) = \begin{cases} 0 & \text{if } A < \lambda(x_1, \ldots, x_N) < B, \\ 1 & \text{otherwise,} \end{cases}$$

and the decision rule

$$\delta_N(x_1, \ldots, x_N) = \begin{cases} \boldsymbol{B}_0 & \text{if } \lambda_N(x_1, \ldots, x_N) \geqslant B > A, \\ \boldsymbol{B}_1 & \text{if } \lambda_N(x_1, \ldots, x_N) \leqslant A < B, \\ \text{any} & \text{if } A = B = 1. \end{cases}$$

Let

$$p_1 = P(\{\text{accept } H_0 | H_1\}) \quad \text{and} \quad p_0 = P(\{\text{accept } H_1 | H_0\}),$$

denote the two error probabilities. Although in general it is tedious to compute p_0 and p_1 given A and B, here it is easily done by

$$A = \frac{p_0}{1 - p_1} \quad \text{and} \quad B = \frac{1 - p_0}{p_1}.$$

References

[1] S. F. Arnold (1980), *Asymptotic validity of F-tests for the ordinary linear model and the multiple correlation model*, JASA 75/372, 890–894.

[2] R. Bergmann, B. Fritzsche and M. Riedel (1984), *Determination of sample size in point estimation*, Math. Operationsforsch. Statist. Ser. Statist. **15**, 73–89.

[3] J. Cohen (1969), *Statistical Power Analysis for Behavioural Sciences*, Academic Press, New York.

[4] T. S. Ferguson (1967), *Mathematical Statistics, A Decision Theoretic Approach*, Academic Press, New York.

[5] D. A. Harville (1971), *On the distribution of linear combinations of non-central chi-squares*, Ann. Math. Statist. **42**, 809–811.

[6] K. M. van Hee and A. Hordijk (1975), *The technique of optimal stopping applied to a sequential sampling problem*, Preprint, Eindhoven.

[7] R. R. Hocking (1976), *The analysis and selection of variables in linear regression*, Biometrika **32**, 1–49.

[8] W. Jahn and M. Riedel (1984), *Reduction of the dimension in the linear model with stochastic regressors*, Comm. Math. Universitatis Carolinae, to appear.

[9] HUMAK (1977), *Statistische Methoden der Modellbildung*, Akademie-Verlag, Berlin.

[10] H. J. Larson and T. A. Bancroff (1963), *Sequential model building for prediction in regression analysis, II*, Ann. Math. Statist. **34**, 462–479.

[11] Odeh and Fox (1975), *Sample Size Choice Charts for Experiments with Linear Models*, Marcel Dekker, New York.

[12] V. J. Oliker (1978), *On the relationship between the sample size and the number of variables in a linear regression model*, Comm. Statist. A – Theory Methods **7**, 509–516.

[13] C. N. Park and A. C. Dudycha (1974), *A cross-validation approach to sample size determination for regression models*, JASA 69/345, 214–218.

[14] S. J. Press (1966), *Linear combinations of noncentral chi-square variates*, Ann. Math. Statist. **37**, 480–487.
[15] – (1972), *Applied Multivariate Analysis*, Holt, Rinehart and Winston, New York.
[16] C. R. Rao (1973), *Lineare statistische Methoden und ihre Anwendungen*, Akademie Verlag, Berlin.
[17] H. Robbins (1959), *Sequential estimation of the mean of a normal population*, in: H. Cramer (ed.), *Probability and Statistics*, Almquist und Wiksell, Grenander Uppsala, 235–242.
[18] K. G. Seal (1951), *On errors of estimates in various types of double sampling procedure*, Sankhyā Ser. B **11**, 125–144.
[19] R. Shibata (1981), *An optimal selection of regression variables*, Biometrika **68**, 45–54.
[20] N. Starr (1966), *The performance of a sequential procedure for the fixed width interval estimation of the mean*, Ann. Math. Statist. **37**, 36–50.
[21] M. Stone (1977), *Cross-validation: a review*, in *3rd. Int. Summer School on Problems of Model Choice and Parameter Estimation in Regression Analysis in Mühlhausen*, Thür.
[22] M. L. Thompson (1978a), *Selection of variables in multiple regression: Part I, A review and evaluation*, Internat. Statist. Rev. **46**, 1–19.
[23] – (1978b), *Selection of variables in multiple regression: Part II, Chosen procedures computations and examples*, Internat. Statist. Rev. **46**, 126–146.
[24] K. L. Tong (1980), *Probability Inequalities in Multivariate Distributions*, Academic Press, New York.
[25] Y. H. Wang (1980), *Sequential estimation of the mean of a multinormal population*, JASA 75/372, 977–983.

Presented to the semester
Sequential Methods in Statistics
September 7–December 11, 1981