Ann. Inst. Statist. Math.
29 (1977), Part A, 131-152

TWO ERRORS IN STATISTICAL MODEL FITTING

NOBUO INAGAKI

(Received Apr. 26, 1976)

1. Introduction

Almost all methods of statistical estimation are carried out under the assumption that the number of parameters to be estimated is fixed and known although their values are unknown. However, in problems of statistical model fitting, it is one of the most important and difficult problems to suitably determine the number of unknown parameters. Generally, problems of statistical model fitting are treated as those of statistical test of goodness of fit and not discussed enough from the point of view of statistical estimation. That is, it is, first, tested whether or not the unknown parameters satisfy certain restrictions and next properties of estimation are discussed under either one of the null hypothesis and the alternative. (For example, Aitchison and Silvey [1]).

Recently Akaike [2] investigates problems of model fitting in the time series analysis from the point of view of statistical prediction and obtains a marvellous and reasonable criterion for deciding the numbers of unknown parameters. This criterion is based on the expectation of the Kullback-Leibler (K-L) information measure of an estimated density and the true density. Therefore we shall call it "Akaike's Information Criterion (AIC)". It is very interesting that the AIC statistic is equal to the C_p statistic due to Mallows [11] in the case of linear regression models. But it is noted that the latter is applied only in the case of linear regression models although the former is done in more general cases. Lindley [9] shows that the C_p statistic is obtained as a Bayes solution under a cost given by the number of explaining vectors to be utilized and a uniform prior distribution. But we should, indeed, discuss what cost to be adopted.

In the present paper we shall introduce two errors, one of which, K^M (say), is caused on account of "modelling" and the other, K^E (say), is done on account of "estimation". Those are based on K-L information measures and so, may be regarded to be equivalent. With this reason we shall define the risk of a pair of modelling and estimation by the sum of the above two errors: $R=K^M+K^E$ (say). (See Section

2.) In the normal linear regression case, the AIC statistic (which is equal to the C_p statistic, there,) is shown in Section 3 to be an estimator of the above risk of model of the number of parameters and estimation of their values, and furthermore a "ridge estimator" (see Hoerl and Kennard [3] and Lindley and Smith [11]) is, also, shown in Section 4 to be derived from the risk of model of the norm of parameters and estimation of their values. In Section 5 we shall propose a definition of statistical model fitting and its loss function, and obtain that the AIC statistic and a ridge estimator are Bayes solutions under the loss functions of the corresponding models and estimations stated in Sections 3 and 4, respectively. It is remarkable that this definition of statistical model fitting and its loss function are intended to enable us to test goodness of model fitting, and at the same time to estimate of parameters.

Asymptotic properties of the AIC statistic are discussed by Inagaki and Ogata [7]. The theory of the weak convergence of likelihood ratio random fields is very useful for this sake. (See LeCam [8], Ibragimov and Khas'minskii [4], [5], and Inagaki and Ogata [6], [7].)

2. Errors of modelling and estimation based on modelling

Let Θ be a parameter space and

$$\mathcal{F}=\{f(y,\theta):\ \theta\in\Theta\} \tag{2.1}$$

be a family of probability density functions (p.d.f.'s) indexed by the elements of Θ. Let observation Y be distributed according to a p.d.f. $f(y,\theta_0)\in\mathcal{F}$ (that is, $\theta_0\in\Theta$).

Let $\mathcal{A}$ be an index space and C_α be a parameter space for each $\alpha\in\mathcal{A}$. Let

$$\mathcal{G}_\alpha=\{g_\alpha(y,\zeta):\ \zeta\in C_\alpha\} \tag{2.2}$$

be a family of p.d.f.'s for every $\alpha\in\mathcal{A}$ and let's call it a "model" for $\mathcal{F}$. Suppose that for each $\theta\in\Theta$ and $\alpha\in\mathcal{A}$ there exists an element $\zeta_\alpha(\theta)\in C_\alpha$ satisfying the following:

$$\int \log\{f(y,\theta)/g_\alpha(y,\zeta_\alpha(\theta))\}f(y,\theta)dy = \min_{\zeta\in C_\alpha}\int \log\{f(y,\theta)/g_\alpha(y,\zeta)\}f(y,\theta)dy\ . \tag{2.3}$$

Then we define the error of model $\mathcal{G}_\alpha$ for the true state $f(y,\theta)$, $K^M(\alpha|\theta)$ (say), by the Kullback-Leibler (K-L) information of $g_\alpha(y,\zeta_\alpha(\theta))\ (\in\mathcal{G}_\alpha)$ under $f(y,\theta)\ (\in\mathcal{F})$:

$$K^M(\alpha|\theta)=\int \log\{f(y,\theta)/g_\alpha(y,\zeta_\alpha(\theta))\}f(y,\theta)dy\ . \tag{2.4}$$

Hence it follows that the above definition of the error of model $\mathcal{G}_\alpha$, $K^M(\alpha|\theta)$, depends only on a subset $\tilde{\mathcal{G}}_\alpha$ of $\mathcal{G}_\alpha$ such that

$$\tilde{\mathcal{G}}_\alpha=\{g_\alpha(y,\zeta_\alpha(\theta)):\ \theta\in\Theta\} \qquad \text{(say)} \tag{2.5}$$

which has the common parameter space Θ for every $\alpha\in\mathcal{A}$.

Now, letting $T=T(Y)$ be any estimator for $\theta\in\Theta$, we shall call $\zeta_\alpha T=\zeta_\alpha(T(Y))$ (say) an estimator for $\zeta_\alpha(\theta)$ based on model $\mathcal{G}_\alpha$ and define the error of $\zeta_\alpha T$ by the expectation of the K-L information of $g_\alpha(z,\zeta_\alpha T(Y))$ under $g_\alpha(z,\zeta_\alpha(\theta))$:

$$K^E(T,\theta|\alpha)=\mathrm{E}\left[\int \log\{g_\alpha(z,\zeta_\alpha(\theta))/g_\alpha(z,\zeta_\alpha T(Y))\}g_\alpha(z,\zeta_\alpha(\theta))dz\right] \tag{2.6}$$

(say)

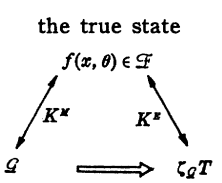
$$=\int\left[\int \log\{g_\alpha(z,\zeta_\alpha(\theta))/g_\alpha(z,\zeta_\alpha T(y))\}g_\alpha(z,\zeta_\alpha(\theta))dz\right]\times f(y,\theta)dy\ .$$

That is, $K^E(T,\theta|\alpha)$ does not mean a usual risk of estimator T of θ but a measurement of the effect of T based on modelling.

In the present paper we intend to approximate the family $\mathcal{F}$ (which includes the true p.d.f. $f(y,\theta_0)$) by a model $\mathcal{G}$ (one of given models $\{\mathcal{G}_\alpha\}$ $\alpha\in\mathcal{A}$) and at the same time, to estimate the true parameter θ_0 based on $\mathcal{G}$. In general we may face the following contradiction in model fitting:

(2.7) The larger is a fitting model $\mathcal{G}$, the smaller is the error of modelling, K^M, but the larger is the error of estimation based on $\mathcal{G}$, K^E, (because the vaguer is the information based on $\mathcal{G}$ with respect to parameter). The reverse is also true.

Note that two errors, K^M and K^E, can be regarded to be equivalent in the sense of the K-L information measure. Therefore we shall define the risk of model and estimator $(\mathcal{G}_\alpha, T)$ (simply, denote it by (α, T)



a fitting model an estimator based on $\mathcal{G}$

Fig. 2.1.

unless confused) by

$$R(\alpha, T|\theta)=K^M(\alpha|\theta)+K^E(T, \theta|\alpha) \quad \text{(say)} \tag{2.8}$$

when θ is true. Thus, the problem in modelling and estimation based on modelling becomes to choose model and estimator (α, T) that minimize the risk $R(\alpha, T|\theta)$. Section 3 is related to AIC statistic and C_p statistic which are decision functions for dimension of parameter. Section 4 is related to ridge estimation which can be regarded as a decision function for norm of parameter.

3. Dimension and values of parameter

Let the parameter space Θ be a subset of the k-dimensional Euclidean space R^k. Suppose that we are given such a prior information that $\mathcal{F}$ in (2.1) is known and the true parameter vector $\theta_0=(\theta_0^{(1)}, \cdots, \theta_0^{(k)})^t$ (say) satisfies

$$\theta_0=(\theta_0^{(1)}, \cdots, \theta_0^{(r_0)}, \theta_{00}^{(r_0+1)}, \cdots, \theta_{00}^{(k)})^t$$

where $\theta_{00}=(\theta_{00}^{(1)}, \cdots, \theta_{00}^{(k)})^t$ (say) is a given interior point of Θ but r_0 is unknown except that $1\leqq r_0\leqq k$. Without any loss of generality we may assume that

$$\theta_{00}=0=(0, \cdots, 0)^t \in \Theta .$$

For $\theta=(\theta^{(1)}, \cdots, \theta^{(r)}, \theta^{(r+1)}, \cdots, \theta^{(k)})^t \in \Theta$, set

$$ {}_r\theta=(\theta^{(1)}, \cdots, \theta^{(r)}, \overbrace{0, \cdots, 0}^{k-r})^t , \tag{3.1}$$

and assume

$$ {}_r\theta \in \Theta , \qquad 1\leqq r\leqq k . \tag{3.2}$$

Then the prior information is restated as follows:

$$\begin{aligned} &\theta_0={}_{r_0}\theta_0=(\theta_0^{(1)}, \cdots, \theta_0^{(r_0)}, \overbrace{0, \cdots, 0}^{k-r_0})^t , \\ &\theta_0^{(r_0)}\neq 0 , \\ &1\leqq r_0\leqq k , \qquad \text{but unknown} . \end{aligned} \tag{3.3}$$

In this case, what matters is how to estimate r_0 and $\theta_0^{(1)}, \cdots, \theta_0^{(r_0)}$, so called "dimension and values" of parameter θ_0, respectively. Applying the concepts in Section 2, we shall, reasonably, make modelling in the following forms:

$$\mathcal{A}=\{r \text{ integer}: 1\leqq r\leqq k\}\ ,$$

$$C_r={}_r\Theta=\{{}_r\zeta:\ \zeta\in\Theta\} \qquad \text{(say)}\ , \tag{3.4}$$

$$\mathcal{G}_r={}_r\mathcal{F}=\{f(y,\ {}_r\zeta):\ \zeta\in\Theta\} \qquad \text{(say)}\ .$$

The error of model ${}_r\mathcal{F}$, $K^M(r|\theta)$ becomes to be

$$K^M(r|\theta)=\int \log\{f(y,\theta)/f(y,\zeta_r(\theta))\}f(y,\theta)dy \tag{3.5}$$

$$=\min_{\zeta\in\Theta}\int \log\{f(y,\theta)/f(y,{}_r\zeta)\}f(y,\theta)dy\ .$$

Note

$${}_1\mathcal{F}\subset{}_2\mathcal{F}\subset\cdots\subset{}_k\mathcal{F}=\mathcal{F}\ . \tag{3.6}$$

The error of estimator T based on ${}_r\mathcal{F}$ is given by

$$K^E(T,\theta|r)=\int\left[\int \log\{f(z,\zeta_r(\theta))/f(z,\zeta_r T(y))\}f(z,\zeta_r(\theta))dz\right] \tag{3.7}$$

$$\times f(y,\theta)dy\ .$$

3.1. *Normal linear regression case*

Let

$$x_r=(x_{1r},\cdots,x_{nr})^t\ , \qquad 1\leqq r\leqq k$$

be explaining vectors and set

$$X_r=(x_1,\cdots,x_r)\ , \qquad n\times r\text{-matrix},\ 1\leqq r\leqq k\ , \tag{3.8}$$

$$X=X_k=(x_1,\cdots,x_k)\ .$$

Suppose that observations $Y=(Y_1,\cdots,Y_n)^t$ are distributed according to n-dimensional normal distribution $N_n(X\theta,\sigma^2 I)$ where $\sigma^2>0$ is known, $I=\begin{pmatrix}1 & & 0\\ & \ddots & \\ 0 & & 1\end{pmatrix}$, the n-dimensional unit matrix, and $\theta=(\theta^{(1)},\cdots,\theta^{(k)})^t\in R^k$ is the unknown parameter. The problem of choosing the first r explaining vectors, $x_1,\cdots,x_r$ (but r is unknown) in the normal linear regression model may be regarded as a special case of modelling (3.4) with

$$f(y,\theta)=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}(y-X\theta)^t(y-X\theta)\right\} \tag{3.9}$$

for $y=(y_1,\cdots,y_n)^t$. Since

$$\int \log\{f(y,\theta)/f(y,{}_r\zeta)\}f(y,\theta)dy=\frac{1}{2\sigma^2}(X\theta-X\cdot{}_r\zeta)^t(X\theta-X\cdot{}_r\zeta)\ , \tag{3.10}$$

we obtain

$$\zeta_r(\theta)=\tilde{P}_r\cdot X\theta \qquad X\cdot\zeta_r(\theta)=P_r\cdot X\theta \tag{3.11}$$

where

$$\begin{aligned} P_r&=X_r(X_r^tX_r)^-X_r^t\,, \qquad n\times n\text{-matrix}\,,\\ \tilde{P}_r&=\begin{bmatrix}(X_r^tX_r)^-X_r^t\\ 0\end{bmatrix}\,, \qquad k\times n\text{-matrix}\,, \end{aligned} \tag{3.12}$$

and $(X_r^tX_r)^-$ is a generalized inverse matrix of $X_r^tX_r$. Then, P_r is the projection matrix from the n-dimensional Euclidean space R^n to the vector space generated by $\{x_1,\cdots,x_r\}$, $\mathcal{V}(x_1,\cdots,x_r)$ (say). Thus we have the following lemma.

LEMMA 3.1.

(i) $\zeta_r(\theta)=\tilde{P}_rX\theta$ *and* $X\zeta_r(\theta)=P_rX\theta$.

(ii) $K^M(r\,|\,\theta)=(1/2\sigma^2)(X\theta)^t(P_k-P_r)(X\theta)$, *and*

$$\begin{aligned} K^E(T,\theta\,|\,r)&=\mathrm{E}\,\frac{1}{2\sigma^2}(T-\theta)^tX^tP_rX(T-\theta)\\ &=\int\frac{1}{2\sigma^2}(\theta-T(y))^tX^tP_rX(\theta-T(y))f(y,\theta)dy\,. \end{aligned} \tag{3.13}$$

Now, we obtain from (3.9) that the maximum likelihood estimator of θ, $\hat{\theta}$ (say), is given by

$$\hat{\theta}=\tilde{P}_ky\,, \tag{3.14}$$

and hence from (3.11) and (3.14) that

$$\zeta_r(\hat{\theta})=\tilde{P}_rX\hat{\theta}=\tilde{P}_rP_ky=\tilde{P}_ry\,, \qquad X\zeta_r(\hat{\theta})=P_ry\,. \tag{3.15}$$

Note that $\zeta_r(\hat{\theta})=\tilde{P}_ry$ is the maximum likelihood estimator of $\zeta_r(\theta)$. Further we have from (3.13) that

$$\begin{aligned} K^E(\hat{\theta},\theta\,|\,r)&=\mathrm{E}\left\{\frac{1}{2\sigma^2}(X\theta-P_ky)^tP_r(X\theta-P_ky)\right\}\\ &=\mathrm{E}\left\{\frac{1}{2\sigma^2}(X\theta-y)^tP_r(X\theta-y)\right\}=\frac{1}{2}\,\mathrm{trace}\,P_r\\ &=\frac{r}{2}\,, \qquad \text{if rank } X_r=r\,. \end{aligned} \tag{3.16}$$

LEMMA 3.2. *The risk of model* $_r\mathcal{F}$ *under the maximum likelihood estimation is as follows*:

$$R(r, \hat{\theta}\,|\,\theta)=\frac{1}{2\sigma^2}(X\theta)^t(P_k-P_r)(X\theta)+\frac{1}{2}\operatorname{trace} P_r \tag{3.17}$$

$$=\frac{1}{2}\left\{\frac{1}{2\sigma^2}(X\theta)^t(P_k-P_r)(X\theta)+r\right\}, \qquad \textit{if } \operatorname{rank} X_r=r .$$

THEOREM 3.1. *The C_p statistic due to Mallows* [11],

$$C(r)=\frac{1}{\sigma^2}y^t(P_k-P_r)y+2\operatorname{trace} P_r-\operatorname{trace} P_k \qquad (say) \tag{3.18}$$

is an unbiased estimator of $2R(r, \hat{\theta}\,|\,\theta)$.

Thus,

$$C(r^*)=\min\{C(r):\ 1\leqq r\leqq k\} \qquad \text{(say)} \tag{3.19}$$

can be regarded as an estimator of

$$2R(r_1, \hat{\theta}\,|\,\theta)=\min\{2R(r, \hat{\theta}\,|\,\theta):\ 1\leqq r\leqq k\} \qquad \text{(say)} . \tag{3.20}$$

That is, we shall be able to consider that r^* is an estimator of r_1 (i.e. model ${}_{r_1}\mathcal{F}$) which minimizes $R(r, \hat{\theta}\,|\,\theta)$, the risk of modelling under the maximum likelihood estimation. Set

$$v_r(X\theta)=\frac{1}{\sigma^2}(X\theta)^t(P_r-P_{r-1})(X\theta) , \qquad 2\leqq r\leqq k$$

$$v_1(X\theta)=\frac{1}{\sigma^2}(X\theta)^tP_1(X\theta) . \tag{3.21}$$

Then,

$$2R(r, \hat{\theta}\,|\,\theta)=(1-v_1(X\theta))+\cdots+(1-v_r(X\theta))+\frac{1}{\sigma^2}(X\theta)^t(X\theta) . \tag{3.22}$$

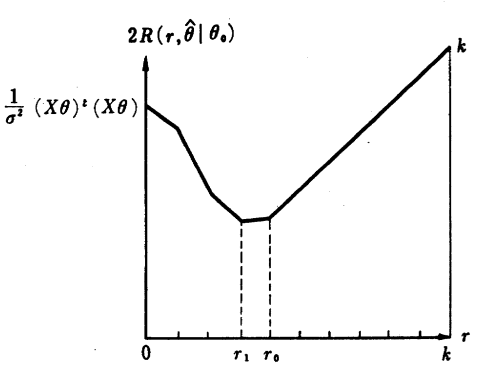


Fig. 2.2.

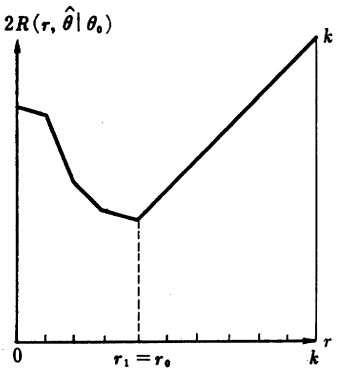


Fig. 2.3.

From the definition, (3.12), of projection P_r,

$$v_r(X\theta_0)=0\,, \qquad \text{for } r\geqq r_0 \text{ with } {}_{r_0}\theta_0=\theta_0\,. \tag{3.23}$$

See Fig. 2.2. Further, if $v_r(X\theta_0)\leqq 1$, for all r such that $2\leqq r\leqq r_0$, we have $r_1=r_0$ in (3.20). See Fig. 2.3.

3.2. *Large sample (i.i.d.) case*

Let $\mathcal{F}$ be a family of n-product densities such that

$$f_n(y,\theta)=\prod_{i=1}^{n} f(y_i,\theta)\ \text{(say)}\,, \qquad y=(y_1,\cdots,y_n)^t\,.$$

Then

$${}_r\mathcal{F}=\{f_n(y,{}_r\zeta)\colon \zeta\in\Theta\}\,.$$

Inagaki and Ogata [6] obtain the following results about the likelihood ratio random field and its related statistics.

LEMMA 3.3 (Theorem 3.4 in [6]). *The likelihood ratio random field*

$$h\precsim Z_n(h)=f_n(y,\theta_0+h/\sqrt{n})/f_n(y,\theta_0) \qquad (say) \tag{3.24}$$

converges weakly to a degenerated normal random field

$$h\precsim Z(h)=\exp\left\{h^t\Gamma^{1/2}\xi-\frac{1}{2}h^t\Gamma h\right\} \qquad (say) \tag{3.25}$$

as $n\to\infty$, where $\Gamma=\Gamma(\theta_0)$ is the Fisher information matrix at θ_0 and ξ is a k-dimensional standard normal random vector $N_k(0, I)$.

This lemma suggests that asymptotic properties of the likelihood ratio, Z_n, and its related statistics are ascribable to properties of Z and functionals on it which are functions of a k-dimensional normal random vector ξ. Now, let

$$\begin{aligned}
&\Gamma^{1/2}=(\gamma_1^{1/2},\cdots,\gamma_k^{1/2})\,,\\
&{}_r\Gamma^{1/2}=(\gamma_1^{1/2},\cdots,\gamma_r^{1/2},\overbrace{0,\cdots,0}^{k-r})\,,\\
&\mathcal{V}({}_r\Gamma^{1/2})=\text{the vector space generated with vector } \gamma_1^{1/2},\cdots,\gamma_r^{1/2} \text{ of } {}_r\Gamma^{1/2}\,,\\
&P_r : R^k\to\mathcal{V}({}_r\Gamma^{1/2})\,, \qquad \text{projection}\,.
\end{aligned} \tag{3.26}$$

Since ${}_r(\theta_0+h/\sqrt{n})=\theta_0+{}_rh/\sqrt{n}$ for $r\geqq r_0$ with ${}_{r_0}\theta_0=\theta_0$, we may define

$${}_r\zeta(\theta_0+h/\sqrt{n})=\theta_0+{}_r\zeta_n(h)/\sqrt{n} \qquad \text{(say)} \tag{3.27}$$

for $r\geqq r_0$ in (3.5):

$$K_n^M(r|h)=\int \log\{f_n(y,\theta_0+h/\sqrt{n})/f_n(y,\theta_0+{}_r\zeta_n(h)/\sqrt{n})\} \times f_n(y,\theta_0+h/\sqrt{n})dy \quad \text{(say)}. \tag{3.28}$$

Denote the distribution of a random variable Z under P_θ by $\mathcal{L}[Z|\theta]$.

LEMMA 3.4 (*Lemma* 5.5 *in* [6]). *For* $r\geqq r_0$,

$$K_n^M(r|h)\to\frac{1}{2}h^t\Gamma^{1/2}(I-P_r)\Gamma^{1/2}h\,,\qquad {}_r\zeta_n(h)\to{}_r\zeta(h)\,, \tag{3.29}$$

as $n\to\infty$, *where convergences are uniform on* $|h|\leqq M$ *and* ${}_r\zeta(h)$ *satisfies*

$$\Gamma^{1/2}{}_r\zeta(h)=P_r\Gamma^{1/2}h\,. \tag{3.30}$$

LEMMA 3.5 (*see* (i) *of Theorem* 2.1 *and Application* 4.3 *in* [6]). *For maximum likelihood estimators* $\hat\theta_{kn}=\hat\theta_{kn}(y)$ *for* θ *and* $\hat\theta_{rn}=\hat\theta_{rn}(y)$ *for* ${}_r\zeta$ *and under* $r\geqq r_0$

$$\mathcal{L}[\log\{f_n(y,\hat\theta_{kn})/f_n(y,\hat\theta_{rn})\}|\theta_0]\to\frac{1}{2}\xi^t(I-P_r)\xi\,, \tag{3.31}$$

and hence by the contiguity of $\{P_{n\theta_0}\}$ *and* $\{P_{n,\theta_0+h/\sqrt{n}}\}$

$$\mathcal{L}\Big[\log\{f_n(y,\hat\theta_{kn})/f_n(y,\hat\theta_{rn})\} -\frac{1}{2}h^t\Gamma^{1/2}(I-P_r)\Gamma^{1/2}h\,|\,\theta_0+h/\sqrt{n}\Big]\to\frac{1}{2}\xi^t(I-P_r)\xi\,. \tag{3.32}$$

In the same way we can see by Lemma 5.5 in [6] that, for estimators $T_n=T_n(y)$ such that $\{\mathcal{L}(\sqrt{n}(T_n-\theta_0)|\theta_0)\}$ are relatively compact,

$$\int \log\{f_n(z,\theta_0+{}_r\zeta_n(h)/\sqrt{n})/f_n(z,\theta_0+{}_r\zeta_n(\sqrt{n}(T_n-\theta_0))/\sqrt{n})\} \times f_n(z,\theta_0+{}_r\zeta_n(h)/\sqrt{n})dz -\frac{1}{2}(\sqrt{n}(T_n-\theta_0)-h)^t\Gamma^{1/2}P_r\Gamma^{1/2}(\sqrt{n}(T_n-\theta_0)-h)\to 0$$

in P_{θ_0}. Hence we have the following lemma for the maximum likelihood estimator $\hat\theta_n=\hat\theta_{kn}(y)$ (say) of θ_0.

LEMMA 3.6. *For* $r\geqq r_0$,

$$K_n^E(\hat\theta_n,\theta_0+h/\sqrt{n}\,|\,r) =\int\Big[\int \log\{f_n(z,\theta_0+{}_r\zeta_n(h)/\sqrt{n}) /f_n(z,\theta_0+{}_r\zeta_n(\sqrt{n}(\hat\theta_n(y)-\theta_0))/\sqrt{n})\} \tag{3.33}$$

$$\times f_n(z, \theta_0+{}_r\zeta_n(h)/\sqrt{n})dz\Big] f_n(y, \theta_0+h/\sqrt{n})dy \qquad (say)$$

$$\to r/2 .$$

After all, we conclude that the AIC statistic due to Akaike [2],

$$\text{(3.34)} \qquad \mathrm{AIC}_n(r)=2\log\{f_n(y, \hat{\theta}_{kn}(y))/f_n(y, \hat{\theta}_{rn}(y))\}+2r-k \qquad \text{(say)}$$

is an estimator of

$$\text{(3.35)} \qquad 2R_n(r, \hat{\theta}_n|h)=2\{K_n^M(r|h)+K_n^E(\hat{\theta}_n, \theta_0+h/\sqrt{n}\,|r)\} \qquad \text{(say)}$$

where $R_n(r, \hat{\theta}_n|h)$ is the risk of model ${}_r\mathcal{F}$ under the maximum likelihood estimation and further, have the following.

THEOREM 3.2 (*See* 4.3 *of Section* 4 *in* [6]).

$$\text{(3.36)} \qquad \mathrm{AIC}_n(r)\to C(r)=\xi^t(I-P_r)\xi+2r-k , \qquad \textit{as } n\to\infty \textit{ for } r\geqq r_0$$

where $C(r)$ *is the Mallow's* C_p *statistic in* (3.18), *and for* $1\leqq r<r_0$,

$$\text{(3.37)} \qquad \mathrm{AIC}_n(r)\to\infty , \qquad \textit{as } n\to\infty .$$

Note that $R_n(r, \hat{\theta}_n|h)\to\infty$ for $1\leqq r<r_0$. Therefore, for all sufficiently large n, r_n^* such that

$$\text{(3.38)} \qquad \mathrm{AIC}_n(r_n^*)=\min\{\mathrm{AIC}_n(r): 1\leqq r\leqq k\}$$

can be regarded as an estimator r_0 in (3.3) which minimizes $R_n(r, \hat{\theta}_n|h)$ for all sufficiently large n (in the same way as in (3.23)).

It is very interesting that the AIC is exactly equal to the C_p statistic in the case of linear regression models. But it is remarkable that the latter applied only in the case of linear regression models although the former is done in more general cases.

4. Norm and values of parameter

We return to the normal linear regression case discussed in 3.1. For symmetric matrix X^tX, there is an orthogonal matrix S satisfying

$$\text{(4.1)} \qquad S^t(X^tX)S=\Lambda=\begin{pmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_k\end{pmatrix} \qquad \text{(say)}$$

where $\lambda_1, \cdots, \lambda_k\geqq 0$. Assume

$$\text{(4.2)} \qquad \lambda_1, \cdots, \lambda_k>0 , \qquad \text{i.e. rank } X=k .$$

We shall consider another modelling in the normal linear regression case

such that

$$\begin{aligned} &\mathcal{A}=R^+=\{\alpha\text{: real and positive}\}\,, \\ &C_\alpha=\Theta_\alpha=\{\zeta\in\Theta\text{: }|\zeta|\leqq\alpha\}\,, \qquad \text{(say)} \\ &\mathcal{G}_\alpha=\mathcal{F}_\alpha=\{f(y,\zeta)\text{: }|\zeta|\leqq\alpha\}\,, \end{aligned} \tag{4.3}$$

where $|\zeta|$ is a norm of ζ so that

$$\begin{aligned} &|\zeta|_0=\max\{|\zeta^{(1)}|,\cdots,|\zeta^{(k)}|\}\,, \\ &|\zeta|_I^2=\zeta^t\zeta\,, \\ &|\zeta|_X^2=\zeta^t(X^tX)\zeta\,, \\ &|\zeta|_A^2=\zeta^tA\zeta\,, \qquad \text{(say)} \end{aligned} \tag{4.4}$$

for a positive definite matrix A. However, we shall discuss the case of Euclidean norm $|\zeta|_I$ in the present paper. Remark

$$\mathcal{F}_{\alpha_1}\subset\mathcal{F}_{\alpha_2}\,, \qquad \text{if } \alpha_1<\alpha_2\,. \tag{4.5}$$

By (3.10), $\zeta_\alpha(\theta)$ (defined in (2.3)) is given as the solution of ζ that minimizes

$$(\zeta-\theta)^tX^tX(\zeta-\theta)\,, \qquad \text{subject to } \zeta^t\zeta\leqq\alpha^2\,. \tag{4.6}$$

Set

$$F(\zeta,\nu)=(\zeta-\theta)^tX^tX(\zeta-\theta)+\nu(\zeta^t\zeta-\alpha^2)\,. \tag{4.7}$$

Kuhn-Tucker condition is as follows: there exist $\nu^0\geqq0$ and ζ^0 such that

$$\left(\frac{\partial F}{\partial\zeta}\right)_{\zeta^0}=0\,, \qquad \left(\frac{\partial F}{\partial\nu}\right)_{\nu^0}\leqq0\,, \qquad \nu^0\left(\frac{\partial F}{\partial\nu}\right)_{\nu^0}=0\,. \tag{4.8}$$

That is,

$$\begin{aligned} &(\zeta^0-\theta)^tX^tX+\nu^0\zeta^{0t}=0\,, \\ &\zeta^{0t}\zeta^0-\alpha^2=0\,, \qquad \text{if } |\theta|>\alpha\,, \\ &\nu^0=0\,, \qquad \text{if } |\theta|\leqq\alpha\,. \end{aligned} \tag{4.9}$$

Now, we define mappings as follows:

$$\begin{aligned} &\tilde{Q}_\alpha=(X^tX+\nu_\alpha I)^{-1}X^t\,, \qquad k\times n\text{-matrix}\,, \\ &Q_\alpha=X\tilde{Q}_\theta=X(X^tX+\nu_\alpha I)^{-1}X^t\,, \qquad n\times n\text{-matrix} \end{aligned} \tag{4.10}$$

where $\nu_\alpha=\nu(\alpha,\theta)$ is a positive number satisfying

$$(X\theta)^t\tilde{Q}_\alpha^t\tilde{Q}_\alpha(X\theta)=\alpha^2 \quad \text{if } |\theta|>\alpha\,, \qquad \nu_\alpha=0 \quad \text{if } |\theta|\leqq\alpha\,, \tag{4.11}$$

for $\alpha>0$ and $\theta\in\Theta$. Hence, we see from (4.6) and (4.9)–(4.11) that

$$\zeta_\alpha(\theta)=\tilde{Q}_\alpha X\theta\,. \tag{4.12}$$

Given and fixed θ in (4.11). According to

$$\alpha^2=\theta^t(\tilde{Q}_\alpha X)^t(\tilde{Q}_\alpha X)\theta=\sum_{i=1}^{k}\frac{\beta_i^2\lambda_i^2}{(\lambda_i+\nu)^2}$$

where

$$\beta=(\beta_1,\cdots,\beta_k)^t=S^t\theta\,, \quad \text{say, (see (4.1))}\,. \tag{4.13}$$

$\nu=\nu_\alpha=\nu(\alpha,\theta)$ has such a derivative that

$$\frac{d\nu}{d\alpha}=-\alpha\Big/\left\{\sum_{i=1}^{k}\frac{\beta_i^2\lambda_i^2}{(\lambda_i+\nu)^3}\right\}<0\,, \quad \text{for } \nu\geqq 0\,. \tag{4.14}$$

That is, $\nu=\nu_\alpha=\nu(\alpha,\theta)$ is a decreasing function of α with

$$\nu\to 0 \quad \text{as } \alpha\to|\theta|\,, \qquad \nu\to\infty \quad \text{as } \alpha\to 0+\,. \tag{4.15}$$

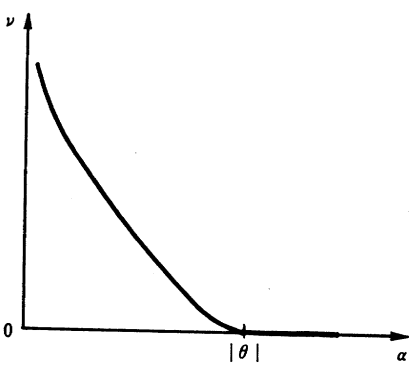


Fig. 4.1.

See Fig. 4.1.

For an estimator $T=T(Y)$ for θ,

$$\zeta_\alpha T=\zeta_\alpha(T(Y))=\tilde{Q}_{\alpha T}XT(Y) \tag{4.16}$$

is an estimator for $\zeta_\alpha(\theta)$, where $\tilde{Q}_{\alpha T}$ is $\tilde{Q}_\alpha$ with $\nu_\alpha=\nu(\alpha,T)$ in (4.11) and hence

$$|\zeta_\alpha T|=\alpha \quad \text{if } |T|>\alpha \text{ (from (4.11))}\,. \tag{4.17}$$

See Fig. 4.2. We shall introduce another estimator for $\zeta_\alpha(\theta)$,

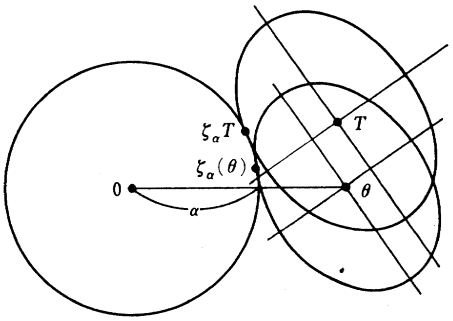
$$\tilde{\zeta}_\alpha T=\tilde{Q}_\alpha XT(Y)\,, \quad \text{with } \nu_\alpha=\nu(\alpha,\theta)\,, \tag{4.18}$$

which is so called a "ridge estimator" for θ (see Hoerl and Kennard [3]). See Fig. 4.3.

The following lemma is easily seen from (3.10) and (4.12).

LEMMA 4.1.

(i) *The error of model $\mathcal{F}_\alpha$ is*



ellipse $\theta \propto$ ellipse T

Fig. 4.2.

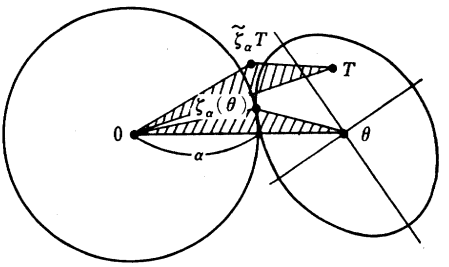


Fig. 4.3.

$$K^M(\alpha|\theta)=\frac{1}{2\sigma^2}(X\theta)^t(P_k-Q_\alpha)^2(X\theta)\ . \tag{4.19}$$

(ii) *The errors of estimators $\zeta_\alpha T$ and $\tilde{\zeta}_\alpha T$ are*

$$K^E(T,\theta|\alpha)=\mathrm{E}\left\{\frac{1}{2\sigma^2}(Q_\alpha X\theta-Q_{\alpha T}XT)^t(Q_\alpha X\theta-Q_{\alpha T}XT)\right\}\ ,$$

$$\tilde{K}^E(T,\theta|\alpha)=\mathrm{E}\left\{\frac{1}{2\sigma^2}(X\theta-XT)^tQ_\alpha^2(X\theta-XT)\right\}\ , \tag{4.20}$$

respectively.

In the sequel, we shall treat only the ridge case.

LEMMA 4.2. *For the maximum likelihood estimator $\hat{\theta}=\tilde{P}_kY$ in* (3.11), *the error of $\tilde{\zeta}_\alpha\hat{\theta}$ is*

$$\tilde{K}^E(\hat{\theta},\theta|\alpha)=\frac{1}{2}\operatorname{trace}Q_\alpha^2=\frac{1}{2}\sum_{i=1}^{k}\left(\frac{\lambda_i}{\lambda_i+\nu_\alpha}\right)^2\ . \tag{4.21}$$

Further, the risk of model α under the maximum likelihood estimation, $\tilde{R}(\alpha,\hat{\theta}|\theta)$ (say), is

$$\begin{aligned}\tilde{R}(\alpha,\hat{\theta}|\theta)&=\frac{1}{2\sigma^2}(X\theta)^t(P_k-Q_\alpha)^2(X\theta)+\frac{1}{2}\operatorname{trace}Q_\alpha^2\\&=\frac{1}{2}\left\{\frac{\nu_\alpha^2}{\sigma^2}\sum_{i=1}^{k}\frac{\beta_i^2\lambda_i}{(\lambda_i+\nu_\alpha)^2}+\sum_{i=1}^{k}\frac{\lambda_i^2}{(\lambda_i+\nu_\alpha)^2}\right\}\\&=\mathrm{E}\left\{\frac{1}{2\sigma^2}(Q_\alpha Y-X\theta)^t(Q_\alpha Y-X\theta)\right\}\end{aligned} \tag{4.22}$$

for $\beta=(\beta_1,\cdots,\beta_k)^t=S^t\theta$ in (4.13).

PROOF. By the virtue of

$$\tilde{Q}_\alpha P_k=(X^tX+\nu_\alpha I)^{-1}X^t\cdot X(X^tX)^{-1}X^t=\tilde{Q}_\alpha\,, \tag{4.23}$$

we have

$$\begin{aligned}\tilde{K}^E(\hat{\theta},\theta|\alpha)&=\mathrm{E}\left\{\frac{1}{2\sigma^2}(X\theta-P_ky)^tQ_\alpha^2(X\theta-P_ky)\right\}\\&=\frac{1}{2}\operatorname{trace}Q_\alpha^2P_k=\frac{1}{2}\operatorname{trace}Q_\alpha^2\\&=\frac{1}{2}\operatorname{trace}\{(X^tX+\nu_\alpha I)^{-1}X^tX\}^2\end{aligned}$$

and so from (4.1)

$$=\frac{1}{2}\sum_{i=1}^{k}\left(\frac{\lambda_i}{\lambda_i+\nu_\alpha^2}\right)^2.$$

It follows from (4.19) and (4.21) that

$$\tilde{R}(\alpha,\hat{\theta}|\theta)=\frac{1}{2}\left\{\frac{1}{\sigma^2}\theta^t(X-Q_\alpha X)^t(X-Q_\alpha X)\theta+\operatorname{trace}Q_\alpha^2\right\}.$$

Now, from (4.10)

$$\begin{aligned}X-Q_\alpha X&=X-X(X^tX+\nu_\alpha I)^{-1}(X^tX+\nu_\alpha I-\nu_\alpha I)\\&=\nu_\alpha X(X^tX+\nu_\alpha I)=\nu_\alpha\tilde{Q}^t\end{aligned}$$

and hence

$$(X-Q_\alpha X)^t(X-Q_\alpha X)=\nu_\alpha^2(X^tX+\nu_\alpha I)^{-1}X^tX(X^tX+\nu_\alpha I)^{-1}.$$

This leads to (4.22) by using (4.13).

The following lemma is an immediate result of the last lemma.

LEMMA 4.3. *$\tilde{R}(\nu)=\tilde{R}(\alpha,\hat{\theta}|\theta)$ with $\nu=\nu_\alpha$ is a function of ν as follows: for a given and fixed θ,*

$$\tilde{R}(\nu)=\frac{1}{2\sigma^2}\left\{\nu^2\sum_{i=1}^{k}\frac{\beta_i^2\lambda_i}{(\lambda_i+\nu)^2}+\sigma^2\sum_{i=1}^{k}\frac{\lambda_i^2}{(\lambda_i+\nu)^2}\right\},$$

$$\tilde{R}(0+)=\frac{k}{2}\,,\qquad \tilde{R}(+\infty)=\sum_{i=1}^{k}\beta_i^2\lambda_i=|X\theta|^2$$

$$\frac{\partial}{\partial\nu}\tilde{R}(\nu)=\frac{\nu}{\sigma^2}\sum_{i=1}^{k}\frac{\beta_i^2\lambda_i^2}{(\lambda_i+\nu)^3}-\sum_{i=1}^{k}\frac{\lambda_i^2}{(\lambda_i+\nu)^3}\,, \tag{4.24}$$

$$\frac{\partial}{\partial\nu}\tilde{R}(\nu)<0\quad \textit{if}\ \nu<\sigma^2/\beta_{\max}^2\,,\qquad \frac{\partial}{\partial\nu}\tilde{R}(\nu)>0\quad \textit{if}\ \nu>\sigma^2/\beta_{\min}^2\,,$$

where

(4.25) $\beta^2_{\max}=\max\{\beta_1^2,\cdots,\beta_k^2\}$, $\beta^2_{\min}=\min\{\beta_1^2,\cdots,\beta_k^2\}$, *(say)* .

On the other hand, ridge estimator $\tilde{\zeta}_a\hat{\theta}=\tilde{Q}_aY$ (recall (4.23)) has the following properties (see Hoerl and Kennard [3]): for a given and fixed θ,

$$L(\nu)=L(\nu_a)=\mathrm{E}\,(\tilde{Q}_aY-\theta)'(\tilde{Q}_aY-\theta) \qquad \text{(say)}$$
$$=\nu^2\sum_{i=1}^{k}\frac{\beta_i^2}{(\lambda_i+\nu)^2}+\sigma^2\sum_{i=1}^{k}\frac{\lambda_i}{(\lambda_i+\nu)^2} ,$$

$$L(0+)=\sigma^2\sum_{i=1}^{k}\frac{1}{\lambda_i} , \qquad L(+\infty)=\sum_{i=1}^{\infty}\beta_i^2=|\theta|^2$$

(4.26)

$$\frac{\partial}{\partial\nu}L(\nu)=2\left\{\sum_{i=1}^{k}\frac{\beta_i^2\lambda_i}{(\lambda_i+\nu)^3}-\sigma^2\sum_{i=1}^{k}\frac{\lambda_i}{(\lambda_i+\nu)^3}\right\} ,$$

$$\frac{\partial}{\partial\nu}L(\nu)<0 \quad \text{if } \nu<\sigma^2/\beta^2_{\max} , \qquad \frac{\partial}{\partial\nu}L(\nu)>0 \quad \text{if } \nu>\sigma^2/\beta^2_{\min} .$$

Fig. 4.5 is due to Hoerl and Kennard [3] (Fig. 1).

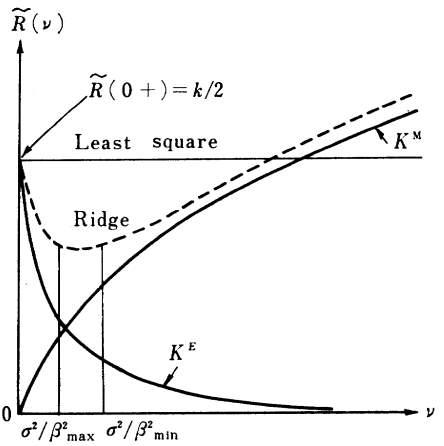


Fig. 4.4.

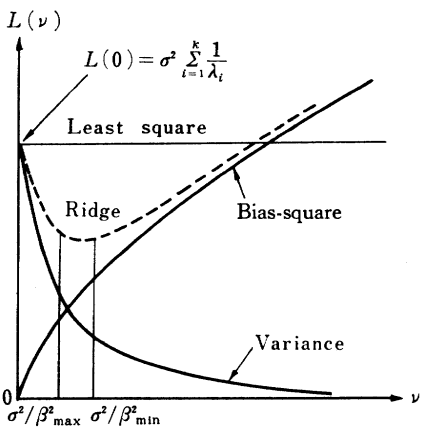


Fig. 4.5.

THEOREM 4.1. *There always exists a positive number $\nu>0$ such that*

$$\tilde{R}(\nu)<\tilde{R}(0+)=k/2 .$$

Further, if $\nu<\sigma^2/\beta^2_{\max}$,

(4.27) $\tilde{R}(\nu)<\tilde{R}(0+)$ *and* $L(\nu)<L(0+)$.

Now, we shall estimate $K^M(\nu)=K^M(\nu_a)=K^M(\alpha,\hat{\theta}\,|\,\theta)$ by an unbiased estimator,

$$\frac{1}{2\sigma^2}Y^t(P_k-Q)^2Y-\frac{1}{2}\operatorname{trace}(P_k-Q)^2 .$$

Then, we see

$$B_1(\nu)=\frac{1}{\sigma^2}Y^t(P_k-Q_a)^2Y+2\operatorname{trace}Q_a-k \qquad \text{(say)} \tag{4.28}$$

is an unbiased estimator of $2\tilde{R}(\nu)$. The following lemma is proved similarly as Lemma 4.3.

LEMMA 4.4.

$$B_1(\nu)=\frac{\nu^2}{\sigma^2}\sum_{i=1}^{k}\frac{z_i^2\lambda_i}{(\lambda_i+\nu)^2}+2\sum_{i=1}^{k}\frac{\lambda_i}{\lambda_i+\nu}-k$$

$$B_1(0+)=k\ , \qquad B_1(+\infty)=\frac{1}{\sigma^2}\sum_{i=1}^{k}z_i^2\lambda_i-k=\frac{1}{\sigma^2}|P_ky|^2-k$$

(4.29)

$$\frac{\partial}{\partial\nu}B_1(\nu)=2\left\{\frac{\nu}{\sigma^2}\sum_{i=1}^{k}\frac{z_i^2\lambda_i^2}{(\lambda_i+\nu)^3}-\sum_{i=1}^{k}\frac{\lambda_i}{(\lambda_i+\nu)^2}\right\}$$

$$\frac{\partial}{\partial\nu}B_1(\nu)<0 \quad \textit{if } \nu<\sigma^2/(z^2\lambda)_{\max}\ , \qquad \frac{\partial}{\partial\nu}B_1(\nu)>0 \quad \textit{if } \nu>\sigma^2/(z^2\lambda)_{\min}$$

where

$$z=(z_1,\cdots,z_k)^t=S^t\tilde{P}_kY\ ,$$

(4.30)

$$(z^2\lambda)_{\max}=\max\{z_1^2\lambda_1,\cdots,z_k^2\lambda_k\}\ , \qquad (z^2\lambda)_{\min}=\min\{z_1^2\lambda_1,\cdots,z_k^2\lambda_k\}\ .$$

Thus, we can conclude that ν_1^* and ν_0, $0<\nu^*$, $\nu_0<\infty$, exist which satisfy

$$B_1(\nu_1^*)=\min\{B(\nu):\ 0<\nu<\infty\}\ ,$$

(4.31)

$$\tilde{R}(\nu_0)=\min\{R(\nu):\ 0<\nu<\infty\}\ ,$$

and that model α_1^* obtained by ν_1^* and the equation

$$(P_kY)^t\tilde{Q}_a^t\tilde{Q}_a(P_kY)=\alpha^2 \tag{4.32}$$

is an estimator α_0 obtained by ν_0 and (4.11). After all, we have a "ridge estimator"

$$\tilde{\zeta}_{\alpha_1^*}\hat{\theta}=\tilde{Q}_{\alpha_1^*}Y=(X^tX+\nu_1^*I)^{-1}X^tY$$

for θ.

5. Statistical model fitting

In this section we shall define the concept of "Statistical model fitting" and its risk. Further, we shall show that the AIC estimator and the ridge estimator are obtained as Bayes solutions of the above risks under uniform prior distribution, respectively.

Denote an estimator for $\alpha \in \mathcal{A}$ by

$$\tau = \tau(Y) = \sum_{\alpha \in \mathcal{A}} \alpha \chi_{H_\alpha}(Y) \tag{5.1}$$

where for each α

$$\begin{aligned} &H_\alpha = \{y;\ \tau(y) = \alpha\}, \text{ the contour of } \tau, \\ &\chi_{H_\alpha} = \text{the indicator function of set } H_\alpha. \end{aligned} \tag{5.2}$$

That is, τ is a multiple decision function for models $\mathcal{G}_\alpha$, $\alpha \in \mathcal{A}$ in (2.2). Define a loss function of estimator τ by

$$J(\tau(y), \theta) = \log \{f(y, \theta)/g_{\tau(y)}(y, \zeta_{\tau(y)}(\theta))\}, \tag{5.3}$$

using ζ_α in (2.3). It is remarkable that $J(\tau(y), \theta)$ is not always nonnegative, but it is decomposed into the sum of the following two parts, J_0 and J_1 (say), J_0 being common to all $\alpha \in \mathcal{A}$ (and so independent of τ), and J_1 nonnegative:

$$\begin{aligned} J_0(\theta) &= \log \{f(y, \theta)/g^0(y, \theta)\} \\ J_1(\tau(y), \theta) &= \log \{g^0(y, \theta)/g_{\tau(y)}(y, \zeta_{\tau(y)}(\theta))\} \end{aligned} \tag{5.4}$$

where

$$g^0(y, \theta) = \sup_{\alpha \in \mathcal{A}} g_\alpha(y, \zeta_\alpha(\theta)). \tag{5.5}$$

We call a pair of estimators for model and parameter, (τ, T), a "statistical model fitting" and define a loss and a risk of model fitting (τ, T) by

$$\begin{aligned} W(\tau(y), T(y) \mid \theta) = J(\tau(y), \theta) + \int &\log \{f(z, \zeta_{\tau(y)}(\theta))/f(z, \zeta_{\tau(y)} T(y))\} \\ &\times f(z, \zeta_{\tau(y)}(\theta)) dz. \end{aligned} \tag{5.6}$$

and

$$\mathcal{R}(\tau, T \mid \theta) = \mathrm{E}\, W(\tau(Y), T(Y) \mid \theta) = \int W(\tau(y), T(y) \mid \theta) f(y, \theta) dy, \tag{5.7}$$

respectively.

5.1. AIC *statistic*

Under the same situations as in 3.1 (normal linear regression), we shall consider a model fitting for dimension and values of parameter. Then, an estimator for r, $1 \leqq r \leqq k$, is

$$\tau = \tau(Y) = \sum_{r=1}^{k} r\, \chi_{H_r}(Y) \,. \tag{5.8}$$

It is easy to see by (3.9), (3.11) and (5.3) that

$$J(\tau(y), \theta) = \frac{1}{2\sigma^2}(y - P_{\tau(y)}X\theta)^t(y - P_{\tau(y)}X\theta) - \frac{1}{2\sigma^2}(y - X\theta)^t(y - X\theta) \,. \tag{5.9}$$

Therefore, from (3.13) and (5.6) through (5.9) we have the following lemma.

LEMMA 5.1. *The loss and the risk of a model fitting* (τ, T) *are given by*

$$W(\tau, T|\theta) = \frac{1}{2\sigma^2}(y - P_\tau X\theta)^t(y - P_\tau X\theta) - \frac{1}{2\sigma^2}(y - X\theta)^t(y - X\theta) + \frac{1}{2\sigma^2}(XT - X\theta)^t P_\tau (XT - X\theta) \,, \tag{5.10}$$

and

$$\mathcal{R}(\tau, Y|\theta) = \int \frac{1}{2\sigma^2}\{(y - P_{\tau(y)}X\theta)^t(y - P_{\tau(y)}X\theta) + (XT(y) - X\theta)^t P_{\tau(y)}(XT(y) - X\theta)\} f(y, \theta)dy - n/2 \,, \tag{5.11}$$

respectively.

Now, let's take the Lebesgue measure on R^k as a prior distribution of θ. Then the posterior distribution, $\Psi(\theta|y)$ (say), becomes to be the k-dimensional normal distribution $N_k(\tilde{P}_k y, \sigma^2(X^tX)^-)$, that is,

$$\Psi(\theta|y) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^k |X^tX|^{1/2} \exp\left\{-\frac{1}{2\sigma^2}(\theta - \tilde{P}_k y)^t(X^tX)(\theta - \tilde{P}_k y)\right\} \,, \tag{5.12}$$

if rank $X = k$. Hence, the posterior risk is calculated as follows:

$$\begin{aligned}\rho(\tau, T|y) &= \int W(\tau, T|\theta)\Psi(\theta|y)d\theta \qquad \text{(say)} \\ &= \frac{1}{2\sigma^2} y^t(P_k - P_\tau)y + \frac{1}{2\sigma^2}\int \{(\theta - \tilde{P}_k y)^t X^t P_\tau X(\theta - \tilde{P}_k y) \\ &\quad - (\theta - \tilde{P}_k y)^t X^t X(\theta - \tilde{P}_k y) + (\theta - T)^t X^t P_\tau X(\theta - T)\} \\ &\quad \times \Psi(\theta|y)d\theta\end{aligned} \tag{5.13}$$

$$=\frac{1}{2\sigma^2}y^t(P_k-P_\tau)y+\frac{1}{2}\operatorname{trace}P_\tau-\frac{1}{2}\operatorname{trace}P_k$$

$$+\frac{1}{2\sigma^2}(T-\tilde{P}_k y)^t X^t P_\tau X(T-\tilde{P}_k y)+\frac{1}{2}\operatorname{trace}P_\tau$$

$$=\frac{1}{2}\left\{C(\tau)+\frac{1}{\sigma^2}(T-\tilde{P}_k y)^t X^t P_\tau X(T-\tilde{P}_k y)\right\} \qquad \text{(recall (3.18))}.$$

Note, again, that the AIC statistic is the same as the C_p statistic in the normal linear regression case. We shall call a model fitting $(r^*, \hat{\theta})$ with r^* in (3.19) and $\hat{\theta}$ in (3.14), the "AIC model fitting". (5.13) leads to the following theorem.

THEOREM 5.1. *The* AIC *model fitting* $(r^*, \hat{\theta})$ *is obtained as the (generalized) Bayes solution when the loss function is* $W(\tau, T|\theta)$ *in* (5.10) *and the prior distribution is the Lebesgue measure on* R^k.

In the i.i.d. large sample case, Inagaki and Ogata [6] obtain the following theorem.

THEOREM 5.2 (*Theorem* 5.2 *in* [6]). *The* AIC *model fitting* (r_n^*, θ_n^*) *with* r_n^* *defined in* (3.38) *and* $\theta^*=\theta_{r_n^*,n}$ (*see* (3.31)), *is asymptotically equivalent to a Bayes solution under a uniform prior distribution.*

5.2. *Ridge estimators*

We shall discuss a model fitting for norm and values of parameter under the same situations as in Section 4. Let

$$\tau=\tau(Y)=\sum_{\alpha\in\mathcal{A}}\alpha\chi_{H_\alpha}(Y) \tag{5.14}$$

be an estimator for α, $0<\alpha<\infty$. Then it follows from (3.9), (4.12) and (5.3) that

$$J(\tau(y), \theta)=\frac{1}{2\sigma^2}(y-Q_{\tau(y)}X\theta)^t(y-Q_{\tau(y)}X\theta)-\frac{1}{2\sigma^2}(y-X\theta)^t(y-X\theta)\ . \tag{5.15}$$

When not $\zeta_\alpha T$ but $\tilde{\zeta}_\alpha T$ is utilized in Lemma 4.1, we have the following lemma from (4.20), (5.6), (5.7) and (5.15).

LEMMA 5.2. *The loss and the risk of a model fitting* (τ, T) *are given by*

$$\tilde{W}(\tau, T|\theta)=\frac{1}{2\sigma^2}(y-Q_\tau X\theta)^t(y-Q_\tau X\theta)-\frac{1}{2\sigma^2}(y-X\theta)^t(y-X\theta) \tag{5.16}$$

$$+\frac{1}{2\sigma^2}(XT-X\theta)^t Q_\tau^2(XT-X\theta)\ ,$$

and

$$\text{(5.17)}\qquad \tilde{\mathcal{P}}(\tau, T|\theta)=\int \frac{1}{2\sigma^2}\{(y-Q_{\tau(y)}X\theta)^t(y-Q_{\tau(y)}X\theta) +(XT(y)-X\theta)^tQ^2_{\tau(y)}(XT(y)-X\theta)\}f(y,\theta)dy-n/2\ ,$$

respectively.

Let's take the Lebesgue measure on R^k as a prior distribution, again. As in (5.13), we see (recalling (4.23))

$$\begin{aligned}\text{(5.18)}\qquad \tilde{\rho}(\tau, T|y)&=\int \tilde{W}(\tau, T|\theta)\Psi(\theta|y)d\theta \qquad \text{(say)}\\ &=\frac{1}{2\sigma^2}\int\{(\theta-\tilde{P}_k)^tX^tQ^2_\tau X(\theta-\tilde{P}_ky)-2y^tQ_\tau X\theta\\ &\quad+2y^tQ^2_\tau X\theta-y^tQ^2_\tau y-(\theta-\tilde{P}_ky)^tX^tX(\theta-\tilde{P}_ky)\\ &\quad+y^tP_ky+(\theta-T)X^tQ^2_\tau X(\theta-T)\}\Psi(\theta|y)d\theta\\ &=\frac{1}{2}\left\{\frac{1}{\sigma^2}y^t(P_k-Q_\tau)^2y+2\ \text{trace}\ Q^2_\tau-k\right.\\ &\quad\left.+(T-\tilde{P}_ky)^tX^tQ^2_\tau X(T-\tilde{P}_ky)\right\}\ .\end{aligned}$$

Set

$$\text{(5.19)}\qquad \nu=\nu_\tau(y)=\nu(\tau(y),\tilde{P}_ky)\qquad \text{(say)}$$

(see (4.10) and (4.11)), and

$$\text{(5.20)}\qquad B_2(\nu)=\frac{1}{\sigma^2}y^t(P_k-Q_\tau)^2y+2\ \text{trace}\ Q^2_\tau-k\ .$$

In the same way as Lemmas 4.3 and 4.4, it holds that

$$\text{(5.21)}\qquad \begin{aligned}&B_2(\nu)=\frac{\nu^2}{\sigma^2}\sum_{i=1}^k\frac{z_i^2\lambda_i}{(\lambda_i+\nu)^2}+2\sum_{i=1}^k\frac{\lambda_i^2}{(\lambda_i+\nu)^2}-k\\ &B_2(0+)=k\ ,\qquad B_2(+\infty)=\frac{1}{\sigma^2}|P_ky|^2-k\\ &\frac{\partial}{\partial\nu}B_2(\nu)=2\left\{\frac{\nu}{\sigma^2}\sum_{i=1}^k\frac{z_i^2\lambda_i^2}{(\lambda_i+\nu)^3}-2\sum_{i=1}^k\frac{\lambda_i^2}{(\lambda_i+\nu)^3}\right\}\\ &\frac{\partial}{\partial\nu}B_2(\nu)<0\ ,\quad \text{if}\ \nu<2\sigma^2/z^2_{\max}\ ,\qquad \frac{\partial}{\partial\nu}B_2(\nu)>0\quad \text{if}\ \nu>2\sigma^2/z^2_{\min}\ ,\end{aligned}$$

where $z^2_{\max}=\max_{i=1,\cdots,k}z_i^2$ and $z^2_{\min}=\min_{i=1,\cdots,k}z_i^2$ (see (4.30)). Therefore there exists $\nu_2^*>0$ such that

(5.22) $$B_2(\nu_2^*)=\min\{B_2(\nu): 0<\nu<\infty\}$$

and hence, we can obtain α_2^* by the virtue of ν_2^* and the equation (4.32). Consequently, we have the following theorem from (5.18) and (5.22).

THEOREM 5.3. *The model fitting* $(\alpha_2^*, \hat{\theta})$, *(recall* $\hat{\theta}=\tilde{P}_k y$*), based on ridge estimation, is the Bayes solution under loss function* $W(\tau, T|\theta)$ *of* (5.16) *and Lebesgue prior distribution.*

6. Remarks

(a) It is apparent that ridge estimators obtained by using norms $|\zeta|=|\zeta|_0$ and $|\zeta|_I$ in (4.4) correspond to a general ridge estimator and a so called shrinkage one, respectively.

(b) It may be necessary that properties of estimators with decision rule (for example, AIC, C_p, $B(\nu)$ and so on) are investigated from the view of estimation theory.

(c) We might have to argue, first of all, what is statistical model fitting.

Acknowledgements

The author wishes to thank Professor H. Akaike for his suggestions and kind guidance of the problems with respect to the Akaike's Information Criterion. The author also wishes to thank Professor M. Okamoto and Mr. Y. Takada for several discussions and valuable comments.

OSAKA UNIVERSITY

REFERENCES

[1] Aitchison, J. and Silvey, S. D. (1958). Maximum-likelihood estimation of parameters subject to restraints, *Ann. Math. Statist.* **29**, 813-828.

[2] Akaike, H. (1972). Information theory and an extension of the maximum likelihood principle, *Proc. 2nd Intl. Symp. on Information Theory*, Supplement to Problems of Control and Information Theory, 267-281.

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

[4] Ibragimov, I. A. and Khas'minskii, R. Z. (1972). Asymptotic behavior of statistical estimators in the smooth case, *Theory Prob. Appl.*, **17**, 443-460.

[5] Ibragimov, I. A. and Khas'minskii, R. Z. (1973). Asymptotic behavior of some statistical estimators II. Limit theorems for the a posteriori density and Bayes' estimators, *Theory Prob. Appl.*, **18**, 76-91.

[6] Inagaki, N. and Ogata, Y. (1975). The weak convergence of likelihood ratio random fields and its applications, *Ann. Inst. Statist. Math.*, **27**, 391-419.

[7] Inagaki, N. and Ogata, Y. (1975). The weak convergence of likelihood ratio random

fields for Markov observation, *Research Memorandum*, No. 79, The Institute of Statistical Mathematics.

[8] LeCam, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates, *Ann. Math. Statist.*, **41**, 802-828.

[9] Lindley, D. V. (1968). The choice of variables in multiple regression, *J. R. Statist. Soc.* B, **30**, 31-53.

[10] Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for linear model, *J. R. Statist. Soc.* B, **34**, 1-18.

[11] Mallows, C. L. (1973). Some comments on C_p, *Technometics*, **15**, 661-675.