Ann. Inst. Statist. Math.
39 (1987), Part A, 533-548

AN EFFECTIVE SELECTION OF REGRESSION VARIABLES WHEN THE ERROR DISTRIBUTION IS INCORRECTLY SPECIFIED*

WOLFGANG HÄRDLE

(Received Sept. 9, 1985; revised Apr. 24, 1986)

Summary

An asymptotically efficient selection of regression variables is considered in the situation where the statistician estimates regression parameters by the maximum likelihood method but fails to choose a likelihood function matching the true error distribution. The proposed procedure is useful when a robust regression technique is applied but the data in fact do not require that treatment. Examples and a Monte Carlo study are presented and relationships to other selectors such as Mallows' C_p are investigated.

1. Introduction and results

Suppose that $Y=(Y_1,\cdots,Y_n)'$ is a random vector of n observations with mean $\mu=(\mu_1,\cdots,\mu_n)'$ and assume that each component μ_i is associated with a covariate x_i, such that $\mu_i=\langle x_i,\beta\rangle$. Assume that the parameter vector is infinite dimensional; then at most n elements of β can be estimated on the basis of the observations. Suppose that a certain likelihood function, not necessarily matching the true error distribution, has been selected by the statistician, and that parameter estimates $\hat{\beta}(p)$ in a finite dimensional submodel p have been obtained by the maximum likelihood principle. The regression curve μ_i at x_i is then estimated by $\hat{\mu}_i(p)=\langle x_i,\hat{\beta}(p)\rangle$ and a loss $L_n(p)=\|\mu-\hat{\mu}(p)\|^2$ is suffered. We shall consider an efficient model selection procedure that asymptotically minimizes the loss $L_n(p)$ over a certain class of finite dimensional models of increasing dimension.

This paper completes earlier papers in various ways. Breiman and Freedman [3], Shibata [12] considered the problem of selecting regres-

* Research supported by Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 123 "Stochastische Mathematische Modelle" and AFOSR Contract No. F49620 82 C 0009.
Key words and phrases: Variable selection, regression analysis, robust regression, model choice.

sion variables when the true error distribution is known to be Gaussian and derived selectors that are equivalent to ours in this case. In the setting of least squares estimation Li ([8]) gave conditions for asymptotic efficiency of model choice procedures based on cross validation, FPE and other means.

Schrader and Hettmansperger ([11]) considered a robust analysis of variance based on Huber's M-estimates and propose a likelihood ratio type of test for testing between finite dimensional submodels. This viewpoint was also taken by Ronchetti ([10]) who derived a "robust model selection" procedure that is related to ours.

A mismatch of a chosen likelihood function and of a true error distribution can happen in the case when the statistician applies a robust regression estimation technique (Huber [6]) but the data is in fact Gaussian. One may also think of the reverse situation that a Gaussian maximum likelihood estimate (i.e., the least squares estimate) is computed but the true error distribution is different, possibly a long tailed outlier generating distribution.

The general idea of regression model selection procedures is to minimize a penalized form of the residual sum of squares. For instance Akaike's AIC ([1]) penalizes the dimensionality of the model with the penalty constant 2. The AIC-score is asymptotically optimal in the case of Gaussian errors and the least square estimation technique as was shown by Shibata ([12]). In the case of a mismatch between the true error distribution and the chosen likelihood function the proposed regression model procedure has a similar structure but the penalty constant is changed depending on the type of mismatch. This can be heuristically described as follows. If there are outliers in the data generated by a long tailed error distribution and AIC is applied based on a Gaussian maximum likelihood estimate the data will be overfitted since the model selection procedure will fit the outliers. The model selection procedure to be presented below penalizes more a high dimensional model since the penalty constant is bigger than 2. On the other hand if the data is indeed Gaussian and a robust regression technique is applied, the penalty constant will be less than 2. An example of this kind is considered in Section 5 where a simulation study is presented.

In the simple case that the data is Gaussian and the statistician chooses a Gaussian likelihood function, then our model selection procedure is equivalent to Mallows' C_p (see [9], Section 4). This entails equivalence to many other selectors such as FPE, AIC, GCV, as was shown by Li ([8]).

We will assume that the control variables $x_i=(x_{i1}, x_{i2}, \cdots)'$, $i=1, \cdots, n$ and the parameter vector $\beta=(\beta_1, \beta_2, \cdots)'$ are in l_2. The model

can then be written as

$$Y=X\beta+e=\mu+e$$

where $e=(e_1,\cdots,e_n)'$ is the vector of the independent observation errors having distribution F with density f and $X'=(x_1' x_2'\cdots x_n')$ is considered as a linear operator from l_2 to R^n. By $p=(p_1, p_2,\cdots, p_{k(p)})$ we denote a finite dimensional submodel with parameter

$$\beta'(p)=(0,\cdots,\beta_{p_1},0,\cdots,\beta_{p_2},0,\cdots,\beta_{p_{k(p)}},0,\cdots)\ .$$

The statistician chooses a likelihood function ρ of which he believes to represent the true error distribution, and estimates the parameters in a submodel p by maximizing the approximate likelihood function

$$\prod_{i=1}^{n}\rho(Y_i-x_i'(p)\beta(p))$$

where $x_i'(p)=(0,\cdots,x_{ip_1},0,\cdots,x_{ip_2},0,\cdots,x_{ip_{k(p)}},0,\cdots)$. Call this maximum likelihood estimate $\hat{\beta}(p)$, and define $\psi(u)=-(d/du)\log\rho(u)$, $\gamma=\mathrm{E}_F\,\psi^2(e)/(\mathrm{E}_F\,\psi'(e))^2$, $B_n=X'X$ and let P_n be a family of models p. A possible selection rule for choosing a model $p\in P_n$ could be defined by $W_n^{(1)}(p)=-\|\hat{\mu}(p)\|^2+2\gamma k(p)+\|\mu\|^2$, since

$$\begin{aligned}(1.1)\qquad W_n^{(1)}(p)-L_n(p)&=-\|\hat{\mu}(p)\|^2+2\gamma k(p)+\|\mu\|^2-\|\hat{\mu}(p)-\mu\|^2\\&=2\{\gamma k(p)-\langle\hat{\beta}(p)-\beta,\hat{\beta}(p)\rangle_{B_n}\}\end{aligned}$$

where $\langle u, v\rangle_{B_n}$ denotes the bilinear form $u'B_nv$ for vectors $u, v\in l_2$. It will be shown that the last term in (1.1) is tending to a constant uniformly over the model class P_n. Then minimizing $W_n^{(1)}(p)$ over P_n will be the same task, at least asymptotically, as minimizing $L_n(p)$. However, $W_n^{(1)}$ cannot be computed directly from the data since it depends on the unknown regression curve μ. But note that the last term in $W_n^{(1)}(p)$ is independent of the model p. We will therefore define

$$W_n(p)=-\|\hat{\mu}(p)\|^2+2\gamma k(p)$$

as the score function that is to be minimized over P_n. The problem of simultaneously estimating γ from the data, in order to make W_n completely data driven is considered in Section 4.

Remark 1. If the statistician is in the happy situation of knowing f, then he will choose $\rho\equiv f$. If f is symmetric, then by partial integration

$$I(F)=\mathrm{E}_F\,\psi^2=\int\psi^2 f=\int(f'/f)^2 f=\int f'\psi=\int\psi' f=\mathrm{E}_F\,\psi'$$

and therefore the constant γ reduces to $(\mathrm{E}_F\,\psi')^{-1}=I(F)^{-1}$, the Fisher-

information number in a location family with density f.

We will use the concept of asymptotic efficiency as in Li [8], Shibata [12] and Stone [13]: A selected $\hat{p}$ is called *asymptotically optimal* if, as $n \to \infty$

$$\frac{L_n(\hat{p})}{\inf_{p \in P_n} L_n(p)} \xrightarrow{p} 1 . \tag{1.2}$$

The following condition on ψ will be needed.

CONDITION 1. The function ψ is centered i.e., $\mathrm{E}_F \psi(e)=0$ and twice differentiable with bounded second derivative. We furthermore assume that $\mathrm{E}_F [q^{-1}(\psi'(e)-q)]^{2N}<\infty$ for some positive integer N and $q=\mathrm{E}_F \psi'(e)$ >0.

The estimates $\hat{\beta}(p)$ will be compared with the Gauss-Markov estimates in the model p based on the (unobservable) pseudodata $\tilde{Y}_i=\mu_i+\tilde{e}_i$, $\tilde{e}_i=\psi(e_i)/q$. Define $X(p)$ as the (n, p) matrix containing the nonzero control variables in model p and assume that $B_n(p)=X'(p)X(p)$ has full rank $k(p)$. Then the Gauss-Markov estimate of μ based on the pseudodata $\tilde{Y}=(\tilde{Y}_1, \cdots, \tilde{Y}_n)'$ is defined as $\tilde{\mu}(p)=H_n(p)\tilde{Y}$, where $H_n(p)=X(p)\cdot B_n^{-1}(p)X'(p)$ denotes the hat matrix in model p. The loss for the Gauss-Markov estimate is $\tilde{L}_n(p)=\|\tilde{\mu}(p)-\mu\|^2$ which will be approximately $L_n(p)$ as will be seen later on. The speed at which the cardinality of P_n is allowed to grow is controlled by

CONDITION 2. There exists a positive integer N such that with $\tilde{R}_n(p)=\mathrm{E}_F \tilde{L}_n(p)$

$$\sum_{p \in P_n} \tilde{R}_n(p)^{-N} \to 0 , \qquad \text{as} \quad n \to \infty .$$

Let $h(p)$ be the largest diagonal element of the hat matrix $H_n(p)$. The speed of $h(p)$ relative to $\tilde{R}_n(p)$ is controlled by

CONDITION 3.

$$\sup_{p \in P_n} h(p)\tilde{R}_n(p) \to 0 , \qquad \text{as} \quad n \to \infty .$$

Remark 2. It follows from Condition 3 that

$$k^2(p)/n \to 0 , \qquad \text{as} \quad n \to \infty ,$$

since $\tilde{R}_n(p)=\gamma k(p)+\|\mu-\mu(p)\|^2$, $\mu(p)=H_n(p)\mu$. This should be seen as an analogue of the necessary condition, $p^2/n \to 0$, that can be found in Huber ([7], p. 166). Conditions 2 and 3 imply also

$$\sum_{p \in P_n} h(p)^N \to 0 . \tag{1.3}$$

Remark 3. If ψ is bounded, as is assumed in a robust regression analysis, Condition 2 can be weakened. It is seen from the proofs that in this case Bernstein's inequality could be used instead of Whittle's ([14], Theorem 2). Condition 2 could be weakened to $\sum_{p\in P_n} \exp(-C\tilde{R}_n(p)) \to 0$, for some $C>0$. In the robust estimation of location so-called redescending ψ-functions have been introduced (see Andrews et al. [2]). A direct application of such a ψ-function which is zero outside some interval is not possible, since points close to infinity also solve the likelihood equation. The usual approach is to couple such estimators to consistent estimators with monotone ψ-functions as is described for instance in Härdle [5], p. 173. A similar procedure seems possible in the setting described here but we did not investigate it.

Condition 1 could be weakened to piecewise twice differentiable ψ-functions, but as Huber [7] we decided to state a stronger condition in order to have a simpler outline of the proof.

Denote by $\hat{p}$ a model $p \in P_n$ that minimizes $W_n(p)$ over P_n. The main result is as follows.

THEOREM. *Under Conditions* 1–3, $\hat{p}$ *is asymptotically optimal.*

The rest of the paper is organized in five sections. In Section 2 the theorem above is shown, in Section 3 we give a variety of examples that satisfy our Conditions 1–3, and in Section 4 the estimation of γ and the relation to other model selection procedures is investigated. In Section 5 a Monte Carlo example of the lemmas that are needed in showing the asymptotic optimality.

2. Proof of Theorem

In the proof of Theorem, the following lemmas will be used.

LEMMA 2.1. *Under the conditions of Theorem, for all* $\varepsilon>0$

$$P\{\sup_{p\in P_n} \|\hat{\mu}(p)-\tilde{\mu}(p)\|^2/\tilde{R}_n(p)>\varepsilon\}\to 0\,, \qquad as \quad n\to\infty\,.$$

LEMMA 2.2. *Under the conditions of Theorem, for all* $\varepsilon>0$

$$P\{\sup_{p\in P_n} |\tilde{L}_n(p)-\tilde{R}_n(p)|/\tilde{R}_n(p)>\varepsilon\}\to 0\,, \qquad as \quad n\to\infty\,.$$

LEMMA 2.3. *Under the conditions of Theorem, for all* $\varepsilon>0$

$$P\{\sup_{p\in P_n} |\gamma k(p)-(\tilde{\mu}(p)-\mu)'\tilde{\mu}(p)+\mu'\tilde{e}|/\tilde{R}_n(p)>\varepsilon\}\to 0\,, \qquad as \quad n\to\infty\,.$$

Recall that the Gauss-Markov estimate based on the pseudodata $\tilde{Y}$

is $\tilde{\beta}(p)=B_n^{-1}(p)X'(p)\tilde{Y}$. The crossterm in (1.1) will be approximated by a corresponding crossterm based on the linearized estimates $\tilde{\beta}(p)$.

$$(2.1)\qquad \begin{aligned}&\langle\hat{\beta}(p)-\beta,\hat{\beta}(p)\rangle_{B_n}-\langle\tilde{\beta}(p)-\beta,\tilde{\beta}(p)\rangle_{B_n}\\&=\|\tilde{\mu}(p)-\hat{\mu}(p)\|^2+\langle\hat{\beta}(p)-\tilde{\beta}(p),\tilde{\beta}(p)-\beta(p)\rangle_{B_n}\\&\quad+\langle\hat{\beta}(p)-\tilde{\beta}(p),\beta(p)\rangle_{B_n}+\langle\tilde{\beta}(p)-\beta,\hat{\beta}(p)-\tilde{\beta}(p)\rangle_{B_n}.\end{aligned}$$

By Lemma 2.1, the first term is of lower order than $\tilde{R}_n(p)$ uniformly over P_n, the second term is bounded by the Cauchy-Schwarz inequality and then Lemmas 2.1 and 2.2 are applied. The third term is handled by formula (6.2), given in the proof of Lemma 2.1, by setting $\alpha=\beta(p)$, $\eta=\hat{\beta}(p)$. The fourth term is handled as the second term. Suppose that

$$(2.2)\qquad \sup_{p,p'\in P_n}\left|\frac{(W_n(p)-W_n(p'))-(L_n(p)-L_n(p'))}{L_n(p)+L_n(p')}\right|=o_p(1),$$

and let p^* denote a minimizer of $L_n(p)$ over P_n. Then by (2.2) with probability greater than $1-\varepsilon$,

$$\frac{W_n(\hat{p})-W_n(p^*)-(L_n(\hat{p})-L_n(p^*))}{L_n(\hat{p})+L_n(p^*)}\geqq-\varepsilon.$$

By the definition of $\hat{p}$, $W_n(\hat{p})-W_n(p^*)\leqq 0$, therefore,

$$-(L_n(\hat{p})-L_n(p^*))\geqq-\varepsilon(L_n(\hat{p})+L_n(p^*))$$

$$L_n(p^*)(1+\varepsilon)\geqq L_n(\hat{p})(1-\varepsilon)$$

$$1\geqq\frac{L_n(p^*)}{L_n(\hat{p})}\geqq\frac{1-\varepsilon}{1+\varepsilon}$$

which shows that (1.2) holds, i.e., $\hat{p}$ is asymptotically optimal. Formula (2.2) follows by observing that

$$\begin{aligned}&\frac{(W_n'(p)+\mu'\tilde{e}-L_n(p))}{L_n(p)}\\&=\frac{2(\gamma k(p)-\langle\hat{\beta}(p)-\beta,\hat{\beta}(p)\rangle_{B_n}+\mu'\tilde{e})}{\tilde{R}_n(p)}\cdot\frac{\tilde{L}_n(p)}{L_n(p)}\cdot\frac{\tilde{R}_n(p)}{\tilde{L}_n(p)}.\end{aligned}$$

The first factor is tending to zero in probability, uniformly over P_n by Lemma 2.3 and formula (2.1). The two other factors tend to one in probability, uniformly over P_n, by Lemmas 2.1 and 2.2.

3. Examples

We start with a reformulation of Condition 2 in the case of hier-

archical model sequences, i.e., $P_n=\{(1), (1, 2), \cdots, (1, 2, \cdots, p_n)\}$ with p_n tending to infinity. In this case Condition 2 follows for $N=2$ from

CONDITION 2′.

$$\inf_{p\in P_n} \tilde{R}_n(p)\to\infty , \qquad \text{as} \qquad n\to\infty .$$

We slightly abuse notation by writing j for $(1,\cdots, j)$. Then Condition 2 follows from $\tilde{R}_n(j)=\gamma j+\|\mu(p)-\mu\|^2$ and

$$\sum_{p\in P_n} \tilde{R}_n(p)^{-2}=\sum_{j=1}^{J_n} \tilde{R}_n(j)^{-2}+\sum_{j=J_n+1}^{P_n} \tilde{R}_n(j)^{-2}$$

$$\leqq J_n\{\inf_{p\in P_n} \tilde{R}_n(p)\}^{-2}+\gamma^{-2}\sum_{j=J_n+1}^{\infty} j^{-2}\to 0 ,$$

if J_n tends to infinity slowly enough. In the following examples we assume that P_n represents a hierarchical model sequence. The following lemma, which is due to Shibata [12], is useful in checking Condition 2′.

LEMMA 3.1. *Assume that with a positive divergent sequence* $\{c_n\}$ *the linear operator* $c_n^{-1}B_n$ *converges weakly to a nonsingular operator* $B: l_2\to l_2$, *such that every* $p\times p$ *principal submatrix* $B(p)$ *has full rank* p *for all* $p>0$. *If* β *has infinitely many nonzero coordinates, then Condition 2′ holds and* p^* *diverges to infinity, as* $n\to\infty$.

Are the conditions of Theorem fulfilled for typical examples? We check conditions in examples given by Shibata [12].

Example 1. Consider the polynomial regression on the interval $[0, 1)$. Here

$$X_{ij}=\left(\frac{i-1}{n}\right)^{j-1}, \qquad i=1,\cdots, n, \qquad j=1, 2,\cdots$$

and

$$Y_i=\sum_{j=1}^{\infty}\left(\frac{i-1}{n}\right)^{j-1}\beta_j+e_i , \qquad i=1,\cdots, n$$

are observed.

Condition 1 is model independent and is an assumption about the error distribution. Condition 2′ is satisfied via Lemma 3.1 (set $c_n=n$). It remains to check Condition 3. The symmetric matrix $B_n^{-1}(p)$ has a spectral decomposition

$$B_n^{-1}(p)=\Gamma_n\Lambda_n\Gamma_n' ,$$

where $\Lambda_n=\text{diag}(\lambda_1(p),\cdots,\lambda_p(p))$ and $\Gamma_n=(\gamma_1,\cdots,\gamma_p)$, $\gamma_j=(\gamma_{1j},\cdots,\gamma_{pj})'$ the i-th normalized eigenvector of $B_n^{-1}(p)$. Lemma 3.1 insures that $\lambda_{\min}(p)$ the smallest eigenvalue of $n^{-1}B_n(p)$ is bounded above zero by a constant C. Therefore each diagonal element h_i of $H_n(p)$ can be estimated by

$$h_i=\sum_{j=1}^{p}\sum_{k=1}^{p}x_{ij}x_{ik}\left(\sum_{l=1}^{p}\gamma_{lj}\gamma_{lk}\lambda_l(p)\right)\leqq\sum_{l=1}^{p}\lambda_l(p)\left(\sum_{j=1}^{p}x_{ij}^2\right)\left(\sum_{j=1}^{p}\gamma_{lj}^2\right)$$

$$\leqq p\lambda_{\max}(p)\sum_{j=1}^{p}x_{ij}^2\leqq p^2\lambda_{\min}(p)^{-1}\leqq C^{-1}p^2/n\ .$$

So Condition 3 is fulfilled if we ask for

$$\sup_{1\leq p\leq p_n}p^2\tilde{R}_n(p)/n\to 0\ . \tag{3.1}$$

A necessary condition is $p^3/n\to 0$ which is slightly stronger than Huber's conditions ([7]).

Example 2. Consider the following representation of the regression curve

$$\mu_i=\sum_{j=1}^{\infty}\beta_j\cos(\pi(j-1)(i-1)/nj)\ .$$

Here the observations are taken at $x=0, n^{-1},\cdots,((n-1)/n)$. As in the example above Condition 2 is satisfied by Lemma 3.1, setting $c_n=n/2$. Condition 3 is satisfied by similar arguments as above if we assume that (3.1) holds.

Example 3. Consider the robust M-estimation of location at different units x_j. Observations are taken repeatedly at p_n different units and n/p_n observations are taken at the point x_j, $j=1,\cdots,p_n$. Assume that $\mathrm{E}_F\,\psi(e)=0$, then Condition 1 is satisfied if ψ, ψ' are bounded. Shibata ([12], p. 51) shows that Condition 2 is satisfied if the vectors of the control-variables $(x_1,\cdots,x_{p_n})$ are linearly independent. Condition 3 can be checked as above.

4. Other methods and estimation of γ

There are a variety of other model selection methods, most of which were shown to be equivalent to Mallows' C_p. We therefore compare our method with C_p only. For simplicity, we work with the linearized estimate $\tilde{\mu}(p)$ based on the pseudodata $\tilde{Y}$. Mallows' score function ([9]) reads

$$C_p(p)=\|\tilde{Y}-\tilde{\mu}(p)\|^2+2\gamma k(p)$$

$$=\|\tilde{e}\|^2+\tilde{L}_n(p)+2\tilde{e}'(I_n-H_n(p))\mu+2\{\gamma k(p)-\tilde{e}'H_n(p)\tilde{e}\}\ .$$

The first term is independent of p, the third and the last term vanish uniformly over model classes P_n, as can be seen in the next section. This shows that a model selected by C_p is asymptotically optimal.

It can now be seen that $W_n(p)$ has a similar structure.

$$\begin{aligned} W_n(p) &= -\|\tilde{\mu}(p)\|^2 + 2\gamma k(p) \\ &= -\|\tilde{\mu}(p)-\mu\|^2 - \|\mu\|^2 - 2(\tilde{\mu}-\mu)'(\mu-\tilde{\mu}) - 2(\tilde{\mu}-\mu)'\tilde{\mu} + 2\gamma k(p) \\ &= \tilde{L}_n(p) + 2\gamma k(p) - 2\tilde{e}' H_n(p)\tilde{e} + 2\tilde{e}'(I-H_n(p))\mu + 2\mu'\tilde{e} - \|\mu\|^2 . \end{aligned}$$

Here the last two terms are independent of the model. The remaining terms are identical to those in Mallows' C_p, which shows that $W_n(p)$ is equivalent to C_p.

It could be argued that the score function that is proposed here is not so reasonable in a practical application since the constant γ is unknown to the statistician. However, if the constant γ can be consistently estimated (independent of p) then the score function based on an estimated γ is also asymptotically optimal. A consistent estimate $\hat{\gamma}_n$ of γ is provided, for instance, by

$$n^{-1}\sum_{i=1}^{n}\psi^2(\hat{e}_i(p_n)) \Big/ \left(n^{-1}\sum_{i=1}^{n}\psi'(\hat{e}_i(p_n))\right)^2$$

where $\hat{e}_i(p_n)$ denote residuals from a fit with a deterministic model p_n, increasing in magnitude as $n\to\infty$. A Taylor expansion and the Cauchy-Schwarz inequality show that $\hat{\gamma}_n \xrightarrow{p} \gamma$, as $n\to\infty$.

5. A simulation study

A small Monte Carlo study was carried out to study the behavior of $W_n(p)$ when applied to some real data. The data were generated according to Example 1 (Section 3) with $\mu_i=\sin(z_i)$, $z_i=-\pi+2((i-1)/n)\pi$, $n=100$, 200 and normal Gaussian error. The original data for $n=100$ is shown in Figure 1. The data do not directly suggest a certain type of model, to a model selection procedure seems to be appropriate. Some of the observations (around $x\approx 1$) look a little bit isolated so that an applied statistician might want to apply a robust regression technique. In this example we have chosen a ψ-function that is linear in $[-2, 2]$ and a constant outside. Such a ψ-function does not satisfy Condition 2 as it stands but as it was argued in Remark 3 the results also hold for this specific choice of a nondifferentiable ψ-function. Straightforward calculations show that $\gamma=1.274$. The values of $L_n(p)$ and $W_n(p)=\|\mu(p)\|^2 + 2\cdot 1.274\cdot p$ (in the hierarchical model case) are presented in Table 1 for $n=100$ and $n=200$. For both sample sizes the p that minimizes $L_n(p)$ is 4. The selected $\hat{p}$ for $n=100$ was $\hat{p}=4$ and for $n=200$ it was

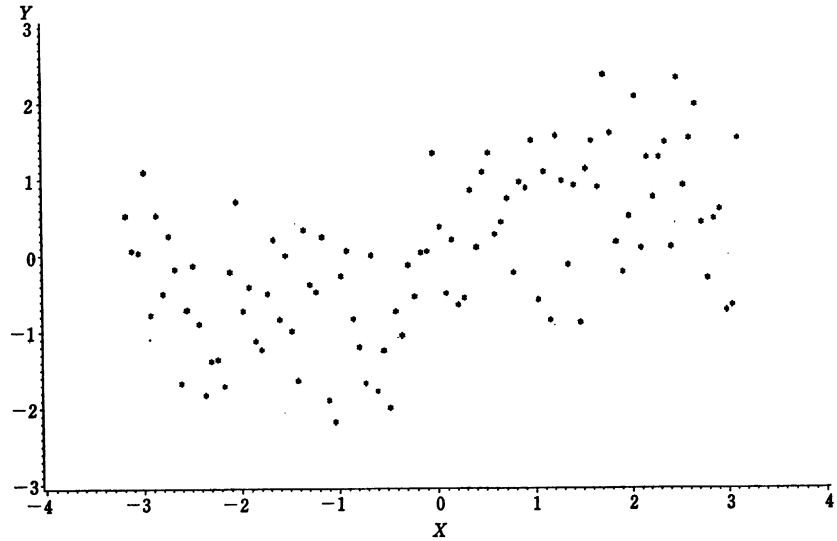

Fig. 1. Original data, $n=100$.

Table 1.

$n=100$			$n=200$		
p	$L_n(p)$	$W_n(p)$	p	$L_n(p)$	$W_n(p)$
2	20.5891	−24.715	2	43.0369	−71.85
3	22.6782	−24.977	3	43.0157	−69.34
4	3.0268	−37.416	4	5.3095	−122.71
5	3.1176	−35.006	5	6.1518	−120.93
6	3.8345	−32.620	6	6.9489	−123.51
7	3.9246	−30.165	7	7.4540	−121.48
8	5.0409	−28.835	8	8.6594	−120.01
9	5.0816	−26.327	9	12.2348	−121.03
10	11.3078	−30.014	10	12.7236	−118.97
11	11.3094	−27.467	11	12.7411	−116.44
12	11.4650	−25.075	12	12.8033	−113.95
13	14.2398	−25.302	13	12.9057	−111.51
14	14.9177	−23.432	14	12.9229	−108.98
15	15.4464	−21.412	15	18.3032	−111.81
16	15.4861	−18.904	16	18.3735	−109.33
17	15.5287	−16.399	17	20.0331	−108.44
18	15.5287	−13.851	18	20.0331	−105.89
19	15.5287	−11.303	19	20.0331	−103.35
20	15.5288	−8.755	20	24.0869	−104.85

$\hat{p}=6$. The shape of the functions $L_n(p)$ and $W_n(p)$ are given in Figures 2 and 3. The parameters for $n=100$ were $\hat{\beta}_1(\hat{p})=0.1017$, $\hat{\beta}_2(\hat{p})=0.9979$, $\hat{\beta}_3(\hat{p})=-0.0006$, $\hat{\beta}_4(\hat{p})=-0.1108$ and thus quite close to the true parameters $\beta_1=\beta_3=0$, $\beta_2=1$, $\beta_4=1/6$. Although for $n=200$ score $W_n(p)$ misses the order of the model that minimizes $L_n(p)$ the fit will not be too bad as the difference of $L_n(4)$ and $L_n(6)$ suggested. In Figure 4 the true curve μ_i and the fitted model curve $\hat{\beta}_i(\hat{p})$, $p=4$ are shown. The

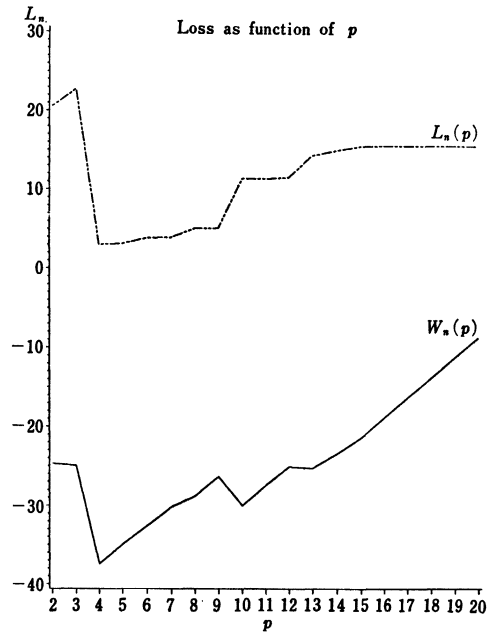

Fig. 2. L_n and W_n.

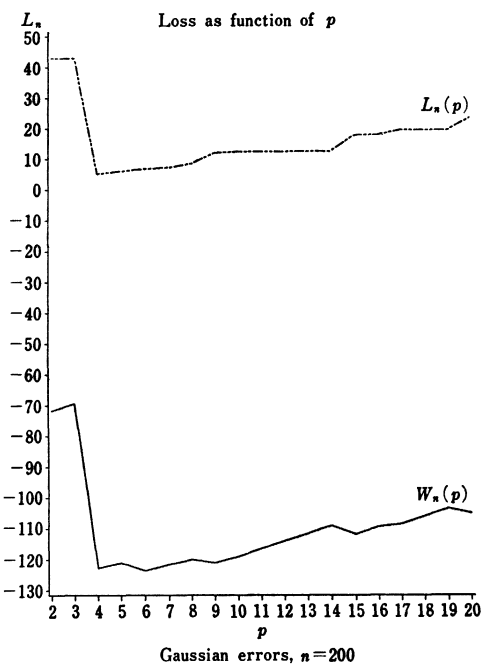

Fig. 3. L_n and W_n.

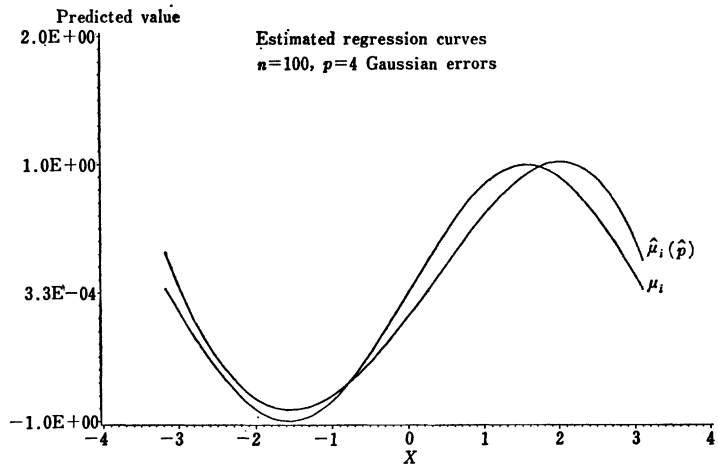

Fig. 4. μ_i and $\hat{\mu}_i(\hat{p})$.

fit was constructed for $n=100$ and with the parameters $\hat{\beta}_j(\hat{p})$, $j=1, \cdots, 4$ as given above. We also studied the case $n=100$, the minimum p for L_n was 6 and this $\hat{p}$ was also selected by W_n. This supports the theory that suggests an increasing $\hat{p}$ as n tends to infinity.

6. Proofs

In this section we give the proofs of Lemmas 2.1–2.3. The proof of Lemma 2.1 follows a related proof in Huber [7], Section 7.4. Similar ideas were used by Cox ([4]) who considered M-type smoothing splines. In order to simplify notation we will consider the hierarchical case only, i.e., the model "p" is identified with "$(1, 2, \cdots, k(p))$, $k(p)=p$". Furthermore we assume without loss of generality that the coordinate system in the p-dimensional subspace of the first p components has been chosen so that $X'(p)X(p)=I_p$. Consider the mapping $\Phi: R^p \to R^p$, $\Phi_k(\eta) = -q^{-1}\sum_{i=1}^{n} \psi\left(Y_i - \sum_{j=1}^{p} x_{ij}\eta_j\right)x_{ik}$, $k=1, \cdots, p$ where $\eta=(\eta_1, \cdots, \eta_p)' \in R^p$. A zero (with respect to η) of Φ will be compared with a zero of $\psi_k(\eta)= \eta_k - \sum_{i=1}^{n}(\mu_i+\tilde{e}_i)x_{ik}$, where $\tilde{e}_i=\psi(e_i)/q$, $q=\mathrm{E}_F\,\psi'(e)$. The zero of $\psi_k(\eta)$ is the least squares estimate $\tilde{\beta}(p)=X(p)\tilde{Y}$ based on the pseudodata $\tilde{Y}$. Consider an arbitrary normalized vector $a \in R^p$, $\|a\|=1$. A Taylor expansion of Φ, using Condition 1, leads to

$$\sum_{k=1}^{p} a_k(\Phi_k(\eta)-\psi_k(\eta)) = -q^{-1}\sum_{k=1}^{p} a_k \sum_{i=1}^{n}(\psi'(e_i)-q)\left(\sum_{j=p+1}^{\infty} x_{ij}\beta_j\right)x_{ik}$$
$$-q^{-1}\sum_{k=1}^{p} a_k \sum_{i=1}^{n}(\psi'(e_i)-q)\sum_{j=1}^{p} x_{ij}x_{ik}(\beta_j-\eta_j)$$

$$-\frac{1}{2}q^{-1}\sum_{k=1}^{p}a_k\sum_{i=1}^{n}\phi''\left(e_i+\nu\left(\sum_{j=1}^{\infty}x_{ij}\beta_j-\sum_{j=1}^{p}x_{ij}\eta_j\right)\right)$$
$$\cdot\left(\sum_{j=1}^{\infty}x_{ij}\beta_j-\sum_{j=1}^{p}x_{ij}\eta_j\right)^2x_{ik}$$
$$=T_{1,n}(p)+T_{2,n}(p)+T_{3,n}(p)\,,\qquad \nu\in(-1,1)\,.$$

We will now show that each of these terms uniformly vanishes over P_n, in the sense that

$$\sup_{p\in P_n}T_{\alpha,n}(p)/\tilde{R}_n^{1/2}(p)\xrightarrow{p}0\,,\qquad \alpha=1,2,3 \tag{6.1}$$

for all (η, a) in the set

$$\mathcal{F}_n=\bigcap_{p\in P_n}\left\{(\eta,a):\ \sum_{i=1}^{n}\left(\sum_{j=1}^{p}x_{ij}(\beta_j-\eta_j)\right)^2\leqq K\tilde{R}_n(p),\ \|a\|=1\right\}\,.$$

Define for $i=1,\cdots,n$

$$V_i=q^{-1}(\phi'(e_i)-q)\,,$$

$$B_{i,n}(p)=\sum_{j=p+1}^{\infty}x_{ij}\beta_j\,,$$

$$s_i=\sum_{k=1}^{p}a_kx_{ik}\,,$$

$$\Delta_{i,n}(p)=\sum_{j=1}^{p}x_{ij}(\beta_j-\eta_j)\,.$$

Note that

$$\|s\|^2=\sum_{i=1}^{n}s_i^2=\sum_{i=1}^{n}\left(\sum_{k=1}^{p}a_kx_{ik}\right)^2=\|X(p)a\|^2=\|a\|^2=1\,.$$

The first term $T_{1,n}(p)$ is estimated as follows.

$$\mathrm{P}\{\sup_{p\in P_n}|T_{1,n}(p)|/\tilde{R}_n^{1/2}(p)>\varepsilon\}$$
$$\leqq\sum_{p\in P_n}\varepsilon^{-2N}\,\mathrm{E}\,\{|T_{1,n}(p)|^{2N}/\tilde{R}_n^N(p)\}$$
$$=\sum_{p\in P_n}\varepsilon^{-2N}\,\mathrm{E}\left\{\left|\sum_{i=1}^{n}s_iB_{i,n}(p)V_i\right|^{2N}\Big/\tilde{R}_n^N(p)\right\}\,.$$

Applying Condition 1 and Whittle's inequality ([14], Theorem 2), this term is bounded by

$$\sum_{p\in P_n}C_1\varepsilon^{-2N}\left(\sum_{i=1}^{n}s_i^2B_{i,n}^2(p)\right)^N\Big/\tilde{R}_n^N(p)$$
$$\leqq C_1\varepsilon^{-2N}\sum_{p\in P_n}\left(\max_{i=1}^{n}S_i^2\right)^N\left(\sum_{i=1}^{n}B_{i,n}^2(p)\right)^N\Big/\tilde{R}_n^N(p)\,,\qquad C_1>0\,.$$

Recall that $\tilde{R}_n(p)=\gamma p+\|(I_n-H_n(p))\mu\|^2\geqq\sum_{i=1}^{n}B_{i,n}^2(p)$. Conditions 2, 3 (see

formula (1.3) in Remark 2) and the simple inequality $s_i^2 \leqq \sum_{j=1}^{p} x_{ij}^2 \sum_{k=1}^{p} a_k^2 = h_i(p)$, where $h_i(p)$ denotes the i-th diagonal element of the hat matrix $H_n(p)$, imply that (6.1) holds for $\alpha=1$.

The second term is estimated similarly. We omit some details. If $(\eta, a) \in \mathcal{F}_n$, then as above

$$\begin{aligned} &\mathrm{P}\{\sup_{p\in P_n} |T_{2,n}(p)|/\tilde{R}_n^{1/2}(p) > \varepsilon\} \\ &\quad \leqq C_1 \varepsilon^{-2N} \sum_{p\in P_n} h(p)^N \Big(\sum_{i=1}^{n} \Delta_{i,n}^2(p)\Big)^N \Big/ \tilde{R}_n(p)^N \\ &\quad \leqq C_1 \gamma^{-N} K^N \varepsilon^{-2N} \sum_{p\in P_n} h(p)^N . \end{aligned}$$

Now apply Conditions 2 and 3 as described in formula (1.3) in Remark 2.

The third term, involving the second derivative of ψ is bounded by

$$\frac{1}{2} q^{-1} \sup \psi'' \max_{i=1}^{n} |s_i| \sum_{i=1}^{n} (B_{i,n}(p) + \Delta_{i,n}(p))^2 .$$

If $(\eta, a) \in \mathcal{F}_n$ we obtain that with a constant C_2

$$|T_{3,n}(p)|/\tilde{R}_n^{1/2}(p) \leqq C_2 h(p)^{1/2} \tilde{R}_n(p)^{1/2} ,$$

which tends to zero by Condition 3.

Altogether we have shown that

$$\sup_{p\in P_n} \Big| \sum_{k=1}^{p} a_k (\Phi_k(\eta) - \Psi_k(\eta)) \Big| \Big/ \tilde{R}_n^{1/2}(p) \xrightarrow{p} 0 \tag{6.2}$$

for all $(\eta, a) \in \mathcal{F}_n$. This entails that for all η in the set

$$\begin{aligned} &\mathcal{G}_n = \Big\{ \eta \in R^p : \sup_{p\in P_n} \sum_{i=1}^{n} \Big(\sum_{k=1}^{p} (\eta_k - \beta_k) X_{ik} \Big)^2 \Big/ \tilde{R}_n(p) \leqq K \Big\} , \\ &\sup_{p\in P_n} \|\Phi(\eta) - \Psi(\eta)\| / \tilde{R}_n^{1/2}(p) \xrightarrow{p} 0 . \end{aligned} \tag{6.3}$$

Condition 2 and bounds on higher moments as above imply that with probability greater than $1-\delta$,

$$\sup_{p\in P_n} \|\tilde{\mu}(p) - \mu(p)\|^2 / \tilde{R}_n(p) < \gamma + \varepsilon . \tag{6.4}$$

This shows that $\tilde{\beta}(p) \in \mathcal{G}_n$ with high probability. Note that

$$\begin{aligned} \|\Phi(\eta) - \eta\| &= \|\Phi(\eta) - \Psi(\eta) + (\beta(p) - \tilde{\beta}(p)) + \beta(p)\| \\ &\leqq \|\Phi(\eta) - \Psi(\eta)\| + \|\beta(p) - \tilde{\beta}(p)\| + \|\beta(p)\| . \end{aligned} \tag{6.5}$$

From formula (6.3) we know that the first term vanishes asymptotically. From (6.4) we conclude that for K big enough

$$\sup_{p\in P_n} \|\beta(p)-\tilde{\beta}(p)\|/\tilde{R}_n^{1/2}(p) \leqq \frac{1}{2}K^{1/2} .$$

Certainly the third term can be made less than $K^{1/2}p^{1/2}/2$. Thus the function $\eta \to \eta - \Phi(\eta)$ has a fixed point η^* in the compact, convex set $\mathcal{G}_n$. Since this fixed point is necessarily a zero of Φ, it is seen that $\hat{\beta}(p)$ is in $\mathcal{G}_n$ with probability greater than $1-\delta$. Substituting $\hat{\beta}(p)$ into equation (6.3) shows that Lemma 2.1 holds.

Lemma 2.2 is seen by the following equation.

$$\tilde{L}_n(p)-\tilde{R}_n(p)=\tilde{e}'H_n(p)\tilde{e}-\gamma k(p) .$$

Condition 2 implies that

$$\sup_{p\in P_n} | \|H_n(p)\tilde{e}\|^2-\gamma k(p)|/\tilde{R}_n(p) \xrightarrow{p} 0$$

which shows Lemma 2.2.

Lemma 2.3 follows similarly observing that

$$\langle \beta-\tilde{\beta}(p), \tilde{\beta}(p)\rangle_{B_n}=\tilde{e}'(I_n-H_n(p))\mu-\tilde{e}'H_n(p)\tilde{e}-\tilde{e}'\mu .$$

Acknowledgement

I would like to thank Charles J. Stone, Ritei Shibata and Johanna Behrens for helpful discussions. The suggestions of an anonymous referee helped to improve the presentation of the paper.

JOHANN-WOLFGANG-GOETHE-UNIVERSITÄT, WEST GERMANY AND
UNIVERSITY OF NORTH CAROLINA*

REFERENCES

[1] Akaike, H. (1970). Statistical predictor identification, *Ann. Inst. Statist. Math.*, **22**, 203-217.

[2] Andrews, D. F., Bickel, P. J., Hampel, F., Huber, P., Rogers, W. and Tukey, J. W. (1972). *Robust Estimation of Location*, Princeton University Press, Princeton.

[3] Breiman, L. and Freedman, D. (1983). How many variables should be entered in a regression equation, *J. Amer. Statist. Ass.*, **78**, 131-136.

[4] Cox, D. (1983). Asymptotics for *M*-type smoothing splines, *Ann. Statist.*, **11**, 530-551.

[5] Härdle, W. (1984). Robust regression function estimation, *J. Multivariate Anal.*, **14**, 169-180.

[6] Huber, P. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo, *Ann. Statist.*, **1**, 799-821.

[7] Huber, P. (1981). *Robust Statistics*, Wiley, New York.

[8] Li, K. C. (1984). Asymptotic optimality for C_p, C_l, cross-validation and generalized cross-validation: Discrete index set, Manuscript.

[9] Mallows, C. (1973). Some comments on C_p, *Technometrics*, **15**, 661-675.

* Now at Universität Bonn, West Germany.

[10] Ronchetti, E. (1985). Robust model selection in regression, *Statist. Prob. Letters*, **3**, 21-23.

[11] Schrader, R. M. and Hettmansperger, T. P. (1980). Robust analysis of variance based upon a likelihood ratio criterion, *Biometrika*, **67**, 93-101.

[12] Shibata, R. (1981). An optimal selection of regression variables, *Biometrika*, **68**, 45-54.

[13] Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates, *Ann. Statist.*, **12**, 1285-1297.

[14] Whittle, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables, *Theor. Prob. Appl.*, **3**, 302-305.