LARGE SAMPLE PROPERTIES OF A SERIES ESTIMATIONS OF CROSS-VALIDATION IN LINEAR REGRESSION MODELS

by

Lin Chen	and	Yongguang Zhang
Purdue University		Institute of Systems Science
U.S.A		Beijing, P.R. China

Technical Report #96-19

Department of Statistics
Purdue University
West Lafayette, IN USA

May 1996

*

LARGE SAMPLE PROPERTIES OF A SERIES ESTIMATIONS OF CROSS-VALIDATION IN LINEAR REGRESSION MODELS

Lin Chen

Purdue University, West Lafayette, IN-47907

Yongguang Zhang

Institute of Systems Science, Beijing 100080, P.R. China

ABSTRACT

Based on asymptotic unbiasedness of prediction, consistency and stability of a series estimations of cross-validation in linear biased and unbiased regression models are described in detail.[1]

1. INTRODUCTION

Consider linear regression model

$$y = \sum_{i=1}^{p} \beta_j x_j + \epsilon, \tag{1.1}$$

let $X = (x_1, ..., x_p)^\tau$; ϵ and X are random variable and vector, muturally independent; EX =**0**, $E\epsilon = 0$; $var\epsilon = \sigma^2$, VAR X=V>0; suppose the joint distribution of (X, y) is $F(X, y)$; β_j, $j = 1...p$ are parameters. Now we get a group of data $(X_i, y_i), i = 1...n$ iid, $(X_1, y_1) \sim F(X, y)$. Then

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i. \tag{1.2}$$

The classical least square estimate of $\beta = (\beta_1, ..., \beta_p)^\tau$ is

$$\hat{\beta}(n, p) = [X(n, p)^\tau X(n, p)]^{-1} X(n, p)^\tau Y(n), \tag{1.3}$$

[1] *Key Words: cross-validation(CV) , mean square error(MSE), conditional mean square error of prediction (CMSEP), v-fold-CV (CV_{nv}), modified v-fold-CV(CV^*_{nv}), basic symmetric condition.*

where $Y(n) = (y_1, ..., y_n)^\tau; X(n,p) = (x_{ij})_{n\times p}$. Let $F_n(X,y)$ be the empirical distribution on $[(X_i, y_i),\ i = 1...n]$. Then as an estimate of goodness of model fittness , the mean square error–MSE (noted by s_e)

$$s_e = \frac{1}{n}\sum_{i=1}^{n}(y_i - X_i^\tau \hat{\beta}(n,p))^2 \tag{1.4}$$

can be formulated :

$$\int (y - X^\tau \hat{\beta}(n,p))^2 dF_n(X,y). \tag{1.5}$$

In fact we get $\hat{\beta}(n,p)$ from $\min_\beta \int (y - X^\tau\beta) dF_n(X,y)$. So we replace it by $\beta_{|F_n}$ for the convenience of following discussion. So

$$s_e = \int (y - X^\tau \beta_{|F_n})^2 dF_n(X,y). \tag{1.6}$$

It is a computable assessment on given data. From view of prediction Allen gave the quantity MSEP(Allen 1971)

$$MSEP = E_F(y - X^\tau \beta_{|F_n})^2, \tag{1.7}$$

and used it value the goodness of the estimate-$\beta_{|F_n}$. Because the distribution F is unknown another more practical quantity valuing the goodness of the estimate-$\beta_{|F_n}$ should be

$$s_n = E_F[(y - X^\tau \beta_{|F_n})^2 \mid ((X_1, y_1), ..., (X_n, y_n))]. \tag{1.8}$$

For convenience of following reductions it is formulated as

$$s_n = \int (y - X^\tau \beta_{|F_n})^2 dF(X,y). \tag{1.9}$$

We call it conditional mean square error of prediction(CMSEP). Here the integration don't have effect on $\beta_{|F_n}$. Clearly s_e can be thought as an estimate of s_n.

From the following analysis we know s_e is a poor estimator of s_n. Look the series estimates of CV.

$$CV_{nn} = \frac{1}{n}\sum_{i=1}^{n}\int (y - X^\tau \beta_{|F^{-i}}) dF^i(X,y), \tag{1.10}$$

where $F^{-i}(X,y)$ is the empirical distribution on$[(X_j,y_j), j=1...n, j\neq i]$. $F^i(X,y)$ is the degenerate distribution on (X_i,y_i).

And the more general CV_{nv}: Divide the data $[(X_i,y_i), i=1...n]$ into v group, make the size of each group as equal as possible. i.e.

$$[\frac{n}{v}] \leq m_\alpha \leq [\frac{n}{v}]+1 \quad \forall\alpha. \tag{1.11}$$

F_{n_α} is the empirical distribution on all data except $\alpha-th$ group data $(m_\alpha+n_\alpha=n)$ so $v-fold-CV$:

$$CV_{nv}=\sum_{\alpha=1}^{v}\rho_\alpha\int(y-X^\tau\beta_{|F_{n_\alpha}})^2dF_{m_\alpha}, \tag{1.12}$$

where $\rho_\alpha=\frac{m_\alpha}{n}, clearly \;\; \rho_\alpha \overset{n\to\infty}{\longrightarrow} \frac{1}{v} \quad \forall\alpha$

$$CV_{nv}=\frac{1}{n_\alpha}\sum_{\alpha=1}^{v}\sum_{i_\alpha=1}^{m_\alpha}(y_{i_\alpha}-X_{i_\alpha}^\tau\beta_{|F_{n_\alpha}})^2. \tag{1.13}$$

From the fallowing discussion we know the asymptotic ratio of CV_{nv} to s_n is slower than that of CV_{nn}. Thus Burman(1989) modified CV_{nv} and get CV_{nv}^*

$$\begin{aligned} CV_{nv}^* &= CV_{nv}+s_e-\sum_{\alpha=1}^{v}\rho_\alpha\int(y-X^\tau\beta_{|F_{n_\alpha}})^2dF_n(X,y) \\ &= CV_{nv}+s_e-\frac{1}{n}\sum_{\alpha=1}^{v}\rho_\alpha\sum_{i=1}^{n}(y_i-X_i^\tau\beta_{|F_{n_\alpha}})^2. \end{aligned} \tag{1.14}$$

(Put v=2 we get learning-testing estimate, we consider it as a special case of our CV approach)

Specially

$$\begin{aligned} CV_{nn}^* &= CV_{nn}+s_e-\frac{1}{n^2}\sum_{\alpha=1}^{n}\sum_{i=1}^{n}(y_i-X_i^\tau\beta_{|F^{-\alpha}})^2 \\ &= (1-\frac{1}{n})CV_{nn}+s_e-\frac{n-1}{n^2}\sum_{\alpha=1}^{n}s_e(\alpha). \end{aligned} \tag{1.15}$$

$s_e(\alpha)$ is the MSE on data $[(y_j,X_j)\; j=1...n\; j\neq\alpha]$.

Asyptotical properties of these statistics are describled in detail in section 2. All proofs are given in appendix.

2 Large Sample Properties of A Series Estimates of CV Under Moment Estimation

2.1. The Models Are Unbiased.

We call (1.1) unbiased model.

[Theorem1] Given conditions of model (1.1) let E_F be the expectation with respect to $F(X, y)$ then

$$E_F s_e = \sigma^2(1 - \frac{p}{n}). \tag{2.1.1}$$

if there exist constants $\delta_1 > 0 \quad \delta_2 \geq 0$ such that

$$E_F[\lambda_{min}(n)]^{-(2+\delta_1)} \leq K n^{\delta_2} \tag{2.1.2}$$

where $\lambda_{min}(n)$ is the minimum eigenvalue of random coefficient matrix

$$\frac{1}{n} X(n,p)^r X(n,p).$$

K may have relation to $\delta_1, \ \delta_2$. Then

$$E_F s_n \geq \sigma^2(1 + \frac{p}{n}) \tag{2.1.3}$$

$$E_F s_n = \sigma^2(1 + \frac{p}{n}) + O(n^{-2}) \tag{2.1.4}$$

[Corollary1] If $X \sim N_p(O, V)$ then $\forall \ \delta_1 > 0$ there exists K (may have relation to δ_1,) such that

$$E_F[\lambda_{min}(n)]^{-(2+\delta_1)} \leq K. \tag{2.1.5}$$

[Corollary2] If $X \sim N_p(O, V) \quad \epsilon \sim N(0, \sigma^2)$ then we have

$$E_F s_n = \sigma^2(1 + \frac{p}{n - p - 1}). \tag{2.1.6}$$

From theorem1 we get

$$E_F(s_n - s_e) \sim \frac{2p}{n}\sigma^2. \tag{2.1.7}$$

If we fix p then

$$E_F(s_n - s_e) \overset{n\to\infty}{\longrightarrow} 0,$$

so we can say s_e is an asymptotic unbiased estimate of s_n. But in the view of it's asymptotic ratio it's a poor estimator of s_n.

[Theorem2] Under conditions of model (1.1) and definitions above if there exist constants $\delta_{1\alpha} > 0 \quad \delta_{2\alpha} \geq 0$ such that

$$E_F[\lambda_{min}(n_\alpha)]^{-(2+\delta_{1\alpha})} \leq K_\alpha n^{\delta_{2\alpha}} \qquad \forall \alpha = 1...v \tag{2.1.8}$$

where $\lambda_{min}(n_\alpha)$ is the minimum eigenvalue of random coefficient matrix

$$\frac{1}{n_\alpha} X(n_\alpha, p)^\tau X(n_\alpha, p)$$

K_α may have relation to $\delta_{1\alpha},\ \delta_{2\alpha}$

Then

$$E_F(cv_{nv} - s_n) = \frac{\sigma^2 P}{(v-1)n} + O(n^{-2}) \tag{2.1.9}$$

$$E_F(cv_{nn} - s_n) = \frac{\sigma^2 P}{(n-1)n} + O(n^{-3}) \tag{2.1.10}$$

[Corollary3] In theorem2 if $X \sim N_p(O, V)$ then condition (2.1.8) can be omitted and the results are still hold. Additionally if $\epsilon \sim N(0, \sigma^2)$, the we have

$$E_F(cv_{nv} - s_n) = \sigma^2 p \sum_{\alpha=1}^{v} \rho_\alpha \left(\frac{1}{n_\alpha - p - 1} - \frac{1}{n - p - 1}\right)$$

$$\overset{n\to\infty}{\longrightarrow} \frac{\sigma^2 p}{(v-1)(n-p-1)} \tag{2.1.11}$$

$$E_F(cv_{nn} - s_n) \overset{n\to\infty}{\longrightarrow} \frac{\sigma^2 p}{(n-p-1)(n-p-2)} \tag{2.1.12}$$

[Corollary4] Under conditions of model (1.1) if $\epsilon \sim N(0, \sigma^2), \quad X \sim N_p(0, V)$ then

$$E_F(cv^*_{nv} - s_n) \sim \frac{p^2\sigma^2}{n^2(v-1)}$$
$$E_F(cv^*_{nn} - s_n) \sim \frac{p^2\sigma^2}{n^2(n-1)}$$

[Theorem3] In theorem1 If $E(X^\tau X) < \infty, E\epsilon^4 < \infty, M = E\epsilon^4 - \sigma^4$ We have

$$var(CV_{nv} - s_n) \sim \frac{M}{n} + O(n^{-2}) \tag{2.1.13}$$
$$var(CV_{nn} - s_n) \sim \frac{M}{n} + O(n^{-2}). \tag{2.1.14}$$

If $\epsilon \sim N(0, \sigma^2), X \sim N_p(0, V)$ We have:

$$var(CV^*_{nv} - s_n) \sim \frac{M}{n} + O(n^{-2}) \tag{2.1.15}$$
$$var(CV^*_{nn} - s_n) \sim \frac{M}{n} + O(n^{-2}). \tag{2.1.16}$$

[Corollary5] In theorem1 If $E\epsilon^4 < \infty, E(X^\tau X) < \infty$ and $M = E\epsilon^4 - \sigma^4$ we have

$$var(s_n - s_e) \sim \frac{M}{n} + O(n^{-2}) \tag{2.1.17}.$$

From above we know that modified $CV - CV^*_{nv}$ and CV itself have the same stability as s_e.

[Theorem4] Under conditions of model(1.1) if

$$\lambda_{min}(n) \xrightarrow{a.s.} \infty \tag{2.1.18}$$
$$log(\lambda_{max}(n)) \overset{a.s.}{=} O(\lambda_{min}(n)) \tag{2.1.19}$$

Where $\lambda_{max}(n)$ $\lambda_{min}(n)$ are maximum and minimum eigenvalue of random matrix:

$$X(n,p)^\tau X(n,p)$$

we have

$$s_n \xrightarrow{a.s.} \sigma^2 \tag{2.1.20}$$

$$s_e - s_n \xrightarrow{a.s.} 0 \tag{2.1.21}$$

[Theorem5] Under conditions of theorem4 if the Basic Symmetric Condition (BSC) is hold:

$$\lim_{n\to\infty} \max_{1\le i\le n} A_{ii} \xrightarrow{a.s.} 0 \tag{2.1.22}$$

where A_{ii} $i = 1...n$ are diagonal element of

$$P_{X(n,p)} = X(n,p)(X(n,p)^\tau X(n,p))^{-1})X(n,p)^\tau$$

Then

$$CV_{nv} - s_n \xrightarrow{a.s.} 0. \tag{2.1.23}$$

Specially

$$CV_{nn} - s_n \xrightarrow{a.s.} 0. \tag{2.1.24}$$

.

2.2 The Models Are Biased

Here 'biased' mean that the true model isn't (1.1), dimentions of variable X may be (1) $t > p$ (2) $t < p$ (3) $t = \infty$. This is nature in practice. In practice we don't know the true model and we use the model (1.1) on given data to get estimates, therefor we should value the goodness of these estimates on the true model instead of model (1.1).

To distinguish we call model (1.1) as practical model. All notes are kept except X_p for X and $\beta(p)$ for β. In this section we assumed $\epsilon \sim N(0,\sigma^2)$, $X \sim N_p(O,V)$

Suppose the true model is

$$y = X_t^\tau \beta(t) + \epsilon, \tag{2.2.1}$$

where $X_t^\tau = (x_1,...,x_t), \beta(t)^\tau = (\beta_1,...,\beta_t), \epsilon, X_t$ independent.

$\epsilon \sim N(0,\sigma^2), X_t \sim N_t(\widetilde{0_t}, V_t)$. $\widetilde{0_t^\tau} = (0,...,0)_t, V_t > 0.$

Firstly let's explain in model (1.1) and (2.2.1) we may assumed V and V_t are I and I_t separately. For example in (2.2.1)

$$\begin{aligned} y &= X_t^\tau \beta(t) + \epsilon \\ &= X_t^\tau (V_t^{-\frac{1}{2}})^\tau (V_t^{\frac{1}{2}}) \beta(t) + \epsilon \\ &= \widetilde{X}_t{}^\tau \widetilde{\beta}(t) + \epsilon, \end{aligned}$$

where

$$\widetilde{\beta}(t) = (V_t^{\frac{1}{2}})\beta(t)$$

$$\widetilde{X}_t = (V_t^{-\frac{1}{2}})X_t$$

$$VAR\widetilde{X}_t = V_t^{-\frac{1}{2}} V_t V_t^{-\frac{1}{2}} = I_t.$$

[Theorem6] Given conditions and assumptions above we have

$t > p$ *or* $t = \infty$:

$$E_F s_n = (\sigma^2 + \sigma_p^2)(1 + \frac{p}{n-p-1}) \tag{2.2.2}$$

where $\sigma_p^2 = \beta_r^\tau \beta_r < \infty, \beta(r) = (\beta_{p+1}, ..., \beta_t), r = t - p.$

$t < p$:

$$E_F s_n = \sigma^2(1 + \frac{p}{n-p-1}) \tag{2.2.3}$$

where $\beta(r) = (\beta_{t+1}, ..., \beta_p), \quad r = p - t.$

[Corollary6] Under conditions and assumptions in theorem6 we have

$$t > p \qquad Es_e = (1 - \frac{p}{n})(\sigma^2 + \sigma_p^2) \tag{2.2.4}$$

$$t < p \qquad Es_e = (1 - \frac{p}{n})\sigma^2 \tag{2.2.5}$$

$$t = \infty \qquad Es_e \sim (1 - \frac{p}{n})(\sigma^2 + \sigma_p^2) \tag{2.2.6}$$

[Theorem7] Given conditions and assumptions in theorem6 we have

$t > p \quad or \quad t = \infty :$

$$E(CV_{nv} - s_n) \sim \frac{p(\sigma^2 + \sigma_p^2)}{n-p-1}\frac{1}{v-1} \tag{2.2.7}$$

$$E(CV_{nv} - s_n) \sim \frac{p(\sigma^2 + \sigma_p^2)}{n-p-1}\frac{1}{n-p-2}. \tag{2.2.8}$$

$t < p :$

$$E(CV_{nv} - s_n) \sim \frac{p\sigma^2}{n-p-1}\frac{1}{v-1} \tag{2.2.9}$$

$$E(cv_{nn} - s_n) \sim \frac{p\sigma^2}{n-p-1}\frac{1}{n-p-2}. \tag{2.2.10}$$

[Corollary7] Given conditions and assumptions of theorem6

We have

$t > p \quad or \quad t = \infty :$

$$E(CV^*_{nv} - s_n) \sim \frac{p^2(\sigma^2 + \sigma_p^2)}{n^2}\frac{1}{v-1} \tag{2.2.11}$$

$$E(CV^*_{nn} - s_n) \sim \frac{p^2(\sigma^2 + \sigma_p^2)}{n^2}\frac{1}{n-1}. \tag{2.2.12}$$

$t < p :$

$$E(CV^*_{nv} - s_n) \sim \frac{p^2\sigma^2}{n^2}\frac{1}{v-1} \tag{2.2.13}$$

$$E(CV^*_{nn} - s_n) \sim \frac{p^2\sigma^2}{n^2}\frac{1}{n-1}. \tag{2.2.14}$$

[Theorem8] Whenever $t > p, t = \infty \quad or \quad t < p$ we have

(Under the conditions and assumptions of theorem6)

$$var(CV_{nv} - s_n) \sim \frac{M + (r+1)\sigma_p^4}{n} + O(n^{-2}) \tag{2.2.15}$$

$$var(CV_{nn} - s_n) \sim \frac{M + (r+1)\sigma_p^4}{n} + O(n^{-2}) \tag{2.2.16}$$

$$var(CV^*_{nv} - s_n) \sim \frac{M + (r+1)\sigma_p^4}{n} + O(n^{-2}) \tag{2.2.17}$$

$$var(CV^*_{nn} - s_n) \sim \frac{M + (r+1)\sigma_p^4}{n} + O(n^{-2}). \tag{2.2.18}$$

Consistency of the series estimations of CV with s_n can be similarly obtained.

Appendix

[Lemma1] **(Theorem of Moment Analysis)**. See [6]

$A = (a_{ij})_{p\times p}$ $C = (c_{ij})_{p\times p}$ are matrixes. A is a function of C: $A = A(C)$. The estimation of C based on a group of samples $x_1.....x_n$ is $\hat{C}$. If

(1) The function A has second order continuous partial derivative in a neighbor of C

(2) There exist constants $\delta_1 > 0$, $\delta_2 \geq 0$, $K > 0$(*may has relation to* δ_1 δ_2) such that elements of $A(\hat{C})$ satisfy

$$E_F \mid a_{ij}(\hat{C}) \mid^{2+\delta_1} \leq Kn^{\delta_2}$$

(3) There exists nature number $k_0 \geq (3 + 1.5\delta_1 + 2\delta_2)\delta^{-1}(or \quad k_0 = \infty)$ such that elements of $\widetilde{C}$: $\widetilde{c}_{ij}$ $(\widetilde{C} = \hat{C} - C)$ satisfy

$$E_F \widetilde{c}_{ij} = O(n^{-1})$$
$$E_F \widetilde{c}_{ij}^{2k} = O(n^{-k}) \quad k \leq k_0 + 1$$

Then as an estimation of A(C) $A(\hat{C})$ satisfies:

$$E_F \mid A(\hat{C}) - A(C) \mid = O(n^{-1}).$$

Theorem1

proof:

$$\begin{aligned}
E_F s_e &= E_F \frac{1}{n}\sum_{i=1}^{n}(y_i - X_i^\tau \beta_{|F_n})^2 \\
&= E_F(E_F \frac{1}{n}\sum_{i=1}^{n}(y_i - X_i^\tau \beta_{|F_n})^2 \mid (X_1, ..., X_n)) \\
&= \frac{1}{n}E_F(E_F Y(n)^\tau P^{\perp}_{X(n,p)} Y(n) \mid (X_1, ..., X_n)) \\
&= \frac{1}{n} tr E_F(\sigma^2 P^{\perp}_{X(n,p)}) \\
&= \sigma^2(1 - \frac{p}{n}).
\end{aligned}$$

Where $P^{\perp}_{X(n,p)} = I - X(n,p)(X(n,p)^{\tau}X(n,p))^{-1}X(n,p)^{\tau}$

$$\begin{aligned} s_n &= \int (y - X^{\tau}\beta_{|F_n})^2 dF(X,y) \\ &= \int (\epsilon + X^{\tau}(\beta - \beta_{|F_n}))^2 dF(X,y) \\ &= \sigma^2 + (\beta - \beta_{|F_n})^{\tau} \int XX^{\tau} dF(X,y)(\beta - \beta_{|F_n}) \end{aligned}$$

$$\begin{aligned} E_F s_n &= \sigma^2 + tr \int XX^{\tau} dF(X,y) E_F(\beta - \beta_{|F_n})(\beta - \beta_{|F_n})^{\tau} \\ &= \sigma^2 + tr \int XX^{\tau} dF(X,y) E_F(E_F(\beta - \beta_{|F_n})(\beta - \beta_{|F_n})^{\tau} \mid (X_1, ..., X_n)) \\ &= \sigma^2 + tr \int XX^{\tau} dF(X,y) E_F[(X(n,p)^{\tau}X(n,p))^{-1}\sigma^2)] \end{aligned}$$

$$\begin{aligned} &= \sigma^2 + \sigma^2 tr \int XX^{\tau} \frac{1}{n} E_F(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^{\tau})^{-1} dF(X,y) \\ &= \sigma^2 + \frac{\sigma^2}{n} \int tr[XX^{\tau} E_F(\frac{1}{n}\sum X_i X_i^{\tau})^{-1}] dF(X,y) \\ &= \sigma^2 + \frac{\sigma^2}{n} \int X^{\tau}[E_F(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^{\tau})^{-1}] X dF(X,y) \\ &\overset{*}{\geq} \sigma^2 + \frac{\sigma^2}{n} \int X^{\tau} V^{-1} X dF(X.y) \\ &= \sigma^2(1 + \frac{p}{n}). \end{aligned}$$

*is according to Jessen inequality.

From conditions of model(1.1) and condition (2.1.2) we know according to lemma1:

$$E_F S_n = \sigma^2(1 + \frac{p}{n}) + O(n^{-2})$$

. #

Theorem2

Proof: Note

$$F_n(X,y) - F(X,y) = D_n(X,y)$$

$$F_{m_\alpha}(X,y) - F(X,y) = D_{m_\alpha}(X,y)$$

$$CV_{nv} - s_n = \sum_{\alpha=1}^{v} \rho_\alpha \int (y - X^\tau \beta_{|Fn_\alpha})^2 dF_{m_\alpha} - \int (y - X^\tau \beta_{|F_n})^2 dF$$
$$= I_1 + I_2 + I_3.$$

Here:

$$I_1 = \int (y - X^\tau \beta_{|F})^2 dD_n(X, y)$$
$$I_2 = \sum_{\alpha=1}^{v} \rho_\alpha \int [(y - X^\tau \beta_{|F_{n_\alpha}})^2 - (y - X^\tau \beta_{|f_n})^2] dD_{m_\alpha}(X, y)$$
$$I_3 = \sum_{\alpha=1}^{v} \rho_\alpha \int [(y - X^\tau \beta_{|F_{n_\alpha}})^2 - (y - X^\tau \beta_{|F_n})^2] dF(x, y)$$
$$(\beta_{|F} = \beta).$$

Clearly $E_F I_1 = E_F I_2 = 0$

$$E_F(CV_{nv} - s_n)$$
$$= E_F I_3$$
$$= E_F \sum_{\alpha=1}^{v} \rho_\alpha \int [(y - X^\tau \beta_{|F_{n_\alpha}})^2 - (y - X^\tau \beta_{|F_n})^2] dF(X, y)$$
$$= E_F \sum_{\alpha=1}^{v} \rho_\alpha \int [(\beta - \beta_{|F_{n_\alpha}})^\tau X X^\tau (\beta - \beta_{|F_{n_\alpha}})$$
$$- (\beta - \beta_{|F_n})^\tau X X^\tau (\beta - \beta_{F_n})] dF(X, y)$$

$$= E_F \sum_{\alpha=1}^{v} \rho_\alpha [(\beta - \beta_{|F_{n_\alpha}})^\tau V (\beta - \beta_{|F_{n_\alpha}}) - (\beta - \beta_{|F_n})^\tau V (\beta - \beta_{|F_n})]$$
$$= E_F (\sum_{\alpha=1}^{v} \rho_\alpha E_F ([(\beta - \beta_{|F_{n_\alpha}})^\tau V (\beta - \beta_{F_{n_\alpha}})$$
$$- (\beta - \beta_{|F_n})^\tau V (\beta - \beta_{|F_n})] \mid (X_1, ..., X_n)))$$
$$= E_F \sum_{\alpha=1}^{v} \rho_\alpha tr(V(X(n_\alpha, p)^\tau X(n_\alpha, p))^{-1} - V(X(n, p)^\tau X(n, p))^{-1}) \sigma^2,$$

From conditions of model (1.1) and (2.1.8) according to Lemma1 we get

$$E_F(CV_{nv} - s_n) = \sigma^2 p \sum_{\alpha=1}^{v} \rho_\alpha (\frac{1}{n_\alpha} - \frac{1}{n}) + O(n^{-2})$$
$$= \frac{\sigma^2 p}{n} \frac{1}{v-1} + O(n^{-2}).$$

Specially v=n then

$$E_F(CV_{nn} - s_n) = \frac{\sigma^2 p}{n}\frac{1}{n-1} + O(n^{-3}).$$

#

Theorem3

Proof: From theorem2 we know

$$E_F(CV_{nv} - s_n) \sim \frac{p\sigma^2}{n}\frac{1}{v-1}.$$

Then

$$(E_F(CV_{nv} - s_n))^2 \sim \frac{p^2\sigma^4}{n^2}\frac{1}{(v-1)^2}.$$

From (2.1.6) we have $CV_{nv} - s_n = I_1 + I_2 + I_3$ So

$$E_F(CV_{nv} - s_n)^2 = E_F I_1^2 + E_F I_2^2 + E_F I_3^2 + 2E_F I_1 I_2 + 2E_F I_2 I_3 + 2E_F I_1 I_3$$

$$\begin{aligned} E_F I_1^2 &= E_F(\int (y - X^\tau \beta_{|F})^2 dD_n(X,y))^2 \\ &= E_F(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 - \sigma^2)^2 \\ &= \frac{1}{n}(E\epsilon^4 - \sigma^4) \\ &= \frac{M}{n}. \end{aligned}$$

After complicated induction we know all

$$E_F I_2^2, E_F I_3^2, 2E_F I_1 I_2, 2E_F I_1 I_3, 2E_F I_2 I_3 \ \ are\ O(n^{-2}).$$

So

$$\begin{aligned} var(CV_{nv} - s_n) &= E_F(CV_{nv} - s_n)^2 - (E_F(CV_{nv} - s_n))^2 \\ &= \frac{M}{n} + O(n^{-2}). \end{aligned}$$

If $\epsilon \sim N_p(0, V)$ then from corollary3 we know

$$E_F(CV^*_{nv} - s_n) \sim \frac{p^2\sigma^2}{n^2}\frac{1}{v-1},$$

then

$$(E_F(CV^*_{nv} - s_n))^2 \sim \frac{p^4\sigma^4}{n^4}\frac{1}{(v-1)^2}.$$

It's easy to see $CV^*_{nv} - s_n = I_1 + I_2 + I_4$ where

$$I_4 = \sum_{\alpha=1}^{v} \rho_\alpha \int [(y - X^\tau \beta_{|F_{n_\alpha}})^2 - (y - X^\tau \beta_{|F_n})^2] dD_n(X, y)$$

Since

$$E_F I_4^2 \sim O(n^{-3}); 2E_F I_1 I_4 \sim O(n^{-2}); 2E_F I_2 I_4 \sim O(n^{-3}).$$

Then

$$var(CV^*_{nv} - s_n) \sim \frac{M}{n} + O(n^{-2})$$

Specially v=n,(2.1.14),(2.1.16) are hold.

. #

Theorem4

Proof: $s_n = \int (y - X^\tau \beta_{|f_n})^2 dF(X, y) = \sigma^2 + (\beta - \beta_{|F_n})^\tau V(\beta - \beta_{|F_n})$

From [21] we know $\beta_{F_n} \xrightarrow{a.s.} \beta$. So $s_n \xrightarrow{a.s.} \sigma^2$.

From Glesser(1966) we also know

$$s_e \xrightarrow{as} \sigma^2.$$

. #

Theorem5

Proof: From (2.12) we can get:

$$CV_{nv} = \frac{1}{n}\sum_{\alpha=1}^{v} \hat{\epsilon}_{m_\alpha}(I - P_{m_\alpha})^{-2}\hat{\epsilon}_{m_\alpha}. \qquad (AP1)$$

Here $\hat{\epsilon}_{m_\alpha}$ is the subvector of $\hat{\epsilon} = Y(n) - X(n,p)^\tau \beta_{|F_n}$ corresponding to the α-th partion and

$$P_{m_\alpha} = X^\tau_{m_\alpha}(X(n,p)^\tau X(n,p))^{-1} X_{m_\alpha}$$

X_{m_α} is the submatrix of $X(n,p)$ corresponding to α-th partion.

From the condition BSC we have

$$\lim_{n\to\infty} A_{i_\alpha i_\alpha} \xrightarrow{a.s.} 0 \Rightarrow \lim_{n\to\infty} P_{m_\alpha} \xrightarrow{a.s.} 0 \ \forall\alpha$$

therefor $CV_{nv} - s_e \xrightarrow{a.s.} 0$ From theorem4 $s_e - s_n \xrightarrow{a.s.} 0$ so $CV_{nv} - s_n \xrightarrow{a.s.} 0$.

. #

Theorem6

Proof:

$t > p$ *or* $t = \infty$:

$$\begin{aligned} s_n &= \int (y - X_p^\tau \beta_{F_n}(p))^2 dF(X_t, y) \\ &= \int (\epsilon + X_t^\tau \beta(t) - X_p^\tau E(\beta_{|F_n}(p) \mid (X_1(P), ..., X_n(p))) \\ &+ X_p^\tau E(\beta_{|F_n}(p) \mid (X_1(p), ..., X_n(p))) - X_p^\tau \beta_{|F_n}(p))^2 dF(X_t, y) \\ &= \sigma^2 + \beta(r)^\tau \beta(r) + (\bar{\beta}(p) - \beta_{|F_n}(p))^\tau (\bar{\beta}(p) - \beta_{|F_n}(p)) \end{aligned}$$

where $\bar{\beta}(p) = E(\beta_{|F_n}(p) \mid (X_1(p), ..., X_n(p)) = \beta(p) + V_p^{-1} V_{pr} \beta(r)$

$$V_t = \begin{pmatrix} V_p & V_{pr} \\ V_{rp} & V_r \end{pmatrix}.$$

According to assumption $V_t = I_t$ so $V_{pr} = 0$ then $\bar{\beta}(p) = \beta(p)$

$$s_n = \sigma^2 + \sigma_p^2 + (\beta(p) - \beta_{|F_n}(p))^\tau (\beta(p) - \beta_{|F_n}(p))$$

$$\begin{aligned} s_n &= \sigma^2 + \sigma_p^2 + (\sigma^2 + \sigma_p^2) \mid\mid (X(n,p)^\tau X(n,p))^{-1} X(n,p)^\tau \varphi \mid\mid \\ &= (\sigma^2 + \sigma_p^2)(1 + \eta^\tau (X(n,p)^\tau X(n,p))^{-1} \eta) \\ E_F s_n &= (\sigma^2 + \sigma_p^2)(1 + \frac{p}{n-p-1}) \end{aligned}$$

Where $\varphi = (\varphi_1, ..., \varphi_p)^\tau$ and

$$\varphi_i = [(\sum_{j=p+1}^{t} \beta_j x_{ij}) + \epsilon_i] / (\sigma^2 + \sigma_p^2)^{\frac{1}{2}}$$

$\eta = S^{-1} X(n,p)^\tau \varphi$ S is the square root of $X(n,p)^\tau X(n,p)$

$t < p$

$$\begin{aligned} s_n &= \int (y - X_p^\tau \beta_{|F_n}(p))^2 dF(X_p, y) \\ &= \int (X_t^\tau \beta(t) + \epsilon - X_p^\tau \beta_{|F_n}(p))^2 dF(X_p, y) \\ &= \int (X_t^\tau \beta(t) - X_t^\tau \beta_{|F_n}(t) - X_r^\tau \beta_{|F_n}(r) + \epsilon)^2 dF(X_p, y) \\ &= \sigma^2 + \beta_{|F_n}(r)^\tau \beta_{|F_n}(r) + (\beta(t) - \beta_{|F_n}(t))^\tau (\beta(t) - \beta_{|F_n}(t)) \\ &- 2\beta_{|F_n}(r)^\tau V_{rt} (\beta(t) - \beta_{|F_n}(t)). \end{aligned}$$

Similarly $V_{rt} = 0$

$$s_n = \sigma^2 + \sigma^2 \parallel (X(n,p)^\tau X(n,p))^{-1} X(n,p)^\tau \varphi \parallel$$
$$= \sigma^2(1 + \eta^\tau (X(n,p)^\tau X(n,p))^{-1}\eta)$$
$$E_F s_n = \sigma^2(1 + \frac{p}{n-p-1})$$

Where $\varphi = (\varphi_1, ..., \varphi_p)^\tau$ and

$$\varphi_i = [(\sum_{j=t+1}^{p} \beta_j x_{ij}) + \epsilon_i]/\sigma$$

$\eta = S^{-1}X(n,p)^\tau\varphi$ S is the square root of $X(n,p)^\tau X(n,p)$

$$X(n,p) = [X(n,t), X(n,r)]$$

#

Theorem7

Proof: It's similar to theorem6.

Theorem8

Proof: We only give the proof of (2.1.16)

Since

$$(E(CV_{nv} - s_n))^2 \sim O(n^{-2})$$
$$E(CV_{nv} - s_e)^2 = E(I_1 + I_2 + I_3)^2$$

$$EI_1^2 = E(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 - \sigma^2 + \beta(r)^\tau(\frac{1}{n}\sum_{i=1}^{n} X_r(i)X_r(i)^\tau - I)\beta(r))^2$$
$$= E(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 - \sigma^2)^2 + E(\beta(r)^\tau(\frac{1}{n}\sum_{i=1}^{n} X_r(i)X_r(i)^\tau - I)\beta(r))^2$$
$$= \frac{M}{n} + \frac{1}{n}(r+1)(\beta(r)^\tau\beta(r))$$
$$= \frac{1}{n}(M + (r+1)\sigma_p^4)$$

and $EI_2^2, EI_1I_2, EI_1I_3, EI_2I_3, EI_3^2$ all are$O(n^{-2})$ *or* $O(n^{-3})$. So

$$var(CV_{nv} - s_n) \sim \frac{1}{n}(M + (r+1)\sigma_p^4) + O(n^{-2}).$$

#

REFERENCES

[1]. Akaike, H. (1974). A New Look at the Statistical Identification Model. IEEE Trans. Automat Control 19, 716–723.

[2]. Akaike, H. (1978). A Bayesian Analysis of the Minimum AIC Procedure. Ann. Inst. Statist. Math. A 30, 9–14.

[3]. Akaike, H. (1985). Prediction and Entropy. in *a celebration of statistics.* The ISI centenary volume eds. A, Atkinson and Fienberg, New York: Springer-verlag, 1-24.

[4]. Allen, D. M. (1971). Mean Square Error of Prediction as a Criterion for Selecting Variables. Technometrics 13. 469–475.

[5]. Allen, D. M. (1974). The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction. Technometrics 16, 125–127.

[6]. An Hongzhi, Cheng Ping (1980). Asymptotic MSE of Autoregressive Parameter Estimation. ACTA MATHEMATICAE APPLICATAE SICINA Vol. 3, No. 1.

[7]. Bickel, P. and Zhang Ping(1992). Variable Selection in Nonparametric Regression with Categorical Covariates. JASA 417.

[8]. Bowman, A. W. (1984). An Alternative Method of Cross-Validation for the Smoothing of Density Estimators. Biometrika 71, 353–60.

[9]. Breiman, L. J and Freedman, D. (1983). How Many Variables Should Be Entered in a Regression Equation. JASA 78, 131–136.

[10]. Breiman, L. J. , Friedman, J. , Olshen, R. and Stone, C. (1984). Classification and Regression Trees. Beiment California: Wadsworth.

[11]. Burman, P. (1989). A Comparative Study of Ordinary Cross-Validation, V- Fold Cross-Validation and The Repeated Learning-Testing Methods. Biometrika 76, 503–514.

[12]. Efron, B. and Morris, C. (1973). Combining Possibly Related Estimation Problems (with discussion). J. R. Statist. Soc. B, 35. 379.

[13]. Efron, B. (1983). Estimating the Error Rate of Prediction Rule-Improvement on Cross-Validation. JASA 382.

[14]. Efron, B. (1992). Jackknife-after -Bootstrap standarded Errors and Influence Function. JRSS vol 54, no.1.

[15]. Friedman, J. (1991). Multivariate Adaptive Regression Spline . Ann. vol 19. No.1.

[16]. Geisser, S. (1974). A Predictive Approach to the Random Effect Model. Biometrika 61, 101–107.

[17]. Geisser, S. (1975). The Predictive Sample Reuse Method with Applications. JASA. 70, 320–328.

[18]. Glesser, L. J. (1966). Correction to "on the asymptotic theory of fixed size sequential confidence bounds for linear regression parameters". Ann. Math. Statist. 37, 1053–5.

[19]. Hall, P. (1982). Cross-Validation in Density Estimation. Biometrika 68, 287-94.

[20]. Hall, P. (1984). Cross-Validation in Nonparametric Estimation of Probability and Probability Density. Biometrika 71, 341-51.

[21]. Helms, R. W. (1974). The Average Estimated Variance Criterion for the Selection of Variables Problem in General Linear Models. Technometrics 16, 216–273.

[22]. Hocking, R. R. (1976). The Analysis and Selection of Variables in Linear Regression. Biometrics 62, 1–49.

[23]. Hurvich, C. M. and TASI, C. L. (1989). Regression and Times Series Model Selection in Small Samples. Biometrika 76, 297–307.

[24]. Horst, P. (1941). Prediction of Personal Adjustment.New York: Social Science Research Council(Bulletin 48).

[25]. Jensen, L. and Stone, M. (1976). Cross-Validatory Choice of Weights for Inte- Intrablock Estimation in Balanced Incompleted Designs. Biometrics 32, 677–681.

[26]. Larson, S. C. (1931). The Shrinkage of the Coefficient of Multiple Correlation. J. Educ. Psychol. 22, 25–55.

[27]. Li, K. C. (1987). Asymptotic Optimality for C_p, C_l, Cross -Validation and Generalized Cross-Validation: Discrete Index Set. Ann. Statist. 15, 958–75.

[28]. Li, K. C. (1985). From Stein's Unbiased Risk Estimation to the Method of Generalized Cross-Validation. Ann. Statist. 13, 1352–1377.

[29]. Mallows, C. L. (1973). it Some Comments on C_p. Technometrics 15, 661–75.

[30]. Nishii, R. (1984). Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression. Ann. Statist. 12, 758–765.

[31]. Rao, C. R and Wu, Y. (1989). A Strong Consistent Procedure For Model Selection in a Regression Problem. Biometrika 76, 369–374.

[32]. Schwarz, G. (1978). Estimating the Dimension of a Model. Ann. Statist. 6, 461–164.

[33]. Shibata, R. (1981). An Optimal Selection of Regression Variables. Biometrika 68, 45–54.

[34]. Sims, C. A. (1971). Distributed Lag Estimation When the Parameter Space Is Explicitly Infinite-Dimensional. Ann. Math. Statist. 42, 1622 —1636.

[35]. Stone, M. (1974 a). Cross-Validatory Choice and Assessment of Statistical Predictions. (with discussion) J. R. Statist. Soc. B 36,111–147.

[36]. Stone, M. (1974 b). Cross-Validation and Multinomial prediction Biometrika 61, 509–515.

[37]. Stone, M. (1977). Asymptotics for and against Cross-Validation. Biometrika 64, 29–35.

[38]. Stone, M. (1977). An Asymptotic Equivalence of Choice of Model by Cross- Validation and Akaike's criterion . J. R. Statist. Soc. B 39, 44 —47.

[39]. Stone, M. (1977). Cross-Validation: A Review. Math Oper. Statist. Ser. Statist. 9, 127–139.

[40]. Sugiura, N. (1978). Further Analysis of the Data by Akaike's Information Criterion and the Finite Corrections. Commun. Statist. A 7. 13–26.

[41]. Thompson, M. (1978). Selection of Variables in Multiple Regression. Internal. Statist. Rev. 46 1–20, 129–146.

[42]. Whaba, G. (1977). Practical Approximate Solutions to Linear Operator Equations When the Data are Noisy. SIAM J. Number. Anal. 14, 651– 667.

[43]. Whaba, G. (1985). A Comparison of GCV and GML for Choosing the Smoothing Parameter in the Generalized Spline Smoothing Problem. Ann. Statist. 13, 1378–1402.

[44]. Zhang Ping.(1991). Variable Selection in Nonparametric Regression with Continuous Covariates. Ann. vol 19. No. 4.