한국통계학회논문집 제2권 2호
1995년 12월 pp. 146-154

A Simple Bias-Correction Rule for the Apparent Prediction Error[1)]

Beong-Soo So[2)]

Abstract

By using simple Taylor expansion, we derive an easy bias-correction rule for the apparent prediction error of the predictor defined by the general M-estimators with respect to an arbitrary measure of prediction error. Our method has a considerable computational advantage over the previous methods based on the resampling technique such as Cross-validaton and Boothtrap. Connections with AIC, Cross-Validation and Boothtrap are discussed too.

1. Introduction

In a pioneering paper in the statistical model identification problem , Akaike (1973) proposed a new criterion for the model choice which is equivalent to the following : If k indexes the model , choose the model k to maximize the quantity ;

$$AIC(k) = L(\widehat{\beta_k} ; k) - p_k \tag{1.1}$$

where $L(\widehat{\beta_k} ; k)$ is the maximized log-likelihood function of the model k, $\widehat{\beta_k}$ is the MLE(Maximum Likelihood Estimator) of the parameter β_k and p_k is the dimensionality of the parameter β_k. Akaike's criteria , which is better known as AIC (Akaike Information Criterion) in the literature , stemmed from the clear recognition that unreserved maximization of the likelihood provides an unsatisfactory method of choice between models that differ appreciably in their dimensionality.

On the other hand , Efron (1983),(1986) considered the problem of the downward bias of the apparent prediction error in the GLM (Generalized Linear Model) and compared the performances of several bias correction methods ,including computer-intensive resampling methods such as Cross-Validation and Boothtrap , for the apparent prediction error and noted

---

1) This research was supported by the research fund from Ehwa Womans University 1993.
2) Department of Statistics, Ewha Womans University, Seodaemun, Seoul, 127-750, KOREA.

incidentally that one of the method coincides with AIC for the special deviance-type loss functon.

Our main objective in this paper is to unify bias-correction techniques of Akake (1973) and Efron (1986) and to derive a simple bias-correction rule for the apparent prediction error which is applicable not only to the arbitrary predictor based on the general M-estimator and but also to any measure of prediction eror.

In addition to the generality of the method , our method has a cosiderable computational advantage over the other computer-intensive resampling methods such as Cross-Validation and Boothtrap . When used as a model identification criterion, our method can be considered as a non-parametric alternative to AIC . As a versitile data-analytic tool, it can be applied effectively not only to the problem of assesing predictive powers of the familiar likelihood based predictors in the general regression set-up such as GLM but also to the problem of discriminating various non-linear predictors in the multivariate regression and the discriminant analyses.

This paper is organized as follows ; In section 2 we derive a key lemma which will provide a simple useful bias-corrected estimate of the expected prediction error of an arbitrary non-linear predictor based on the general M-type estimator . Then we consider the relationship of our method and other alternative non-parametric methods such as Cross-Validation and Boothtrap and show that they are all asymptotically equivalent . In section 3 we give several examples which illustrate versitility of our simple bias-correction rule in evaluating predictive powers of the various predictors including ones based on the ridge-type estimator occuring in the linear and logistic regression models with respect to arbitrary prediction errors.

2. Main Result

Let (X_1,Y_1) ,…, (X_n,Y_n) , (X,Y) be a random sample from the common distrbution P defined on the sample space $S=S_x\times S_y$. Let $f(X;\beta)$ $\beta\in\theta$ be the class of possible predictors of Y of given functional form containing unknown parameter vector β which represents the possible choice available to the statistician. Suppose we have a goodness of fit measure $L(Y,\ f(X,\hat{\beta}))$ which reflects the prediction error of the predictor $f(X\ ,\hat{\beta})$ derived from the estimator $\hat{\beta}$ selected by the statistician. Ideally the best choice for the statistician will be the β^* which is defined by :

$$\beta^* = \arg\min_{\beta\in\theta} E_P[L(Y, f(X;\beta))] \ . \tag{2.1}$$

Because we do not know the true underlying distribution P of (X,Y) in practice , we are forced to use some empirical estmate $\hat{\beta}$ of β^*. For example we may substitute β^* by its empirical version $\hat{\beta}$ which is formally defined by :

$$\hat{\beta} = \arg\min_{\beta \in \Theta} \sum_{i=1}^{n} L(Y_i,\ f(X_i;\beta)/n\ . \tag{2.2}$$

In accessing goodness of the possible choice $\hat{\beta}$, two different notions of prediction errors are relevent :
First , we have the *conditional* prediction error defined by ;

$$PE(\hat{\beta}) = E_P(L(Y, f(X;\hat{\beta})) \tag{2.3}$$

and secondly we may consider *unconditional expected* prediction error defined by ;

$$E_n[PE(\hat{\beta})] = E_n[E_P(L(Y, f(X;\hat{\beta}))] \tag{2.4}$$

where expectaion in (2.3) is taken with respect to the *new* observation (X,Y) only and the double expectations in (2.4) are taken with respect to the *trainning* data $(X_1,Y_1),\cdots,(X_n,Y_n)$ and *new* data (X,Y) simultaneously.

In this framework our main objective in this paper is to find a good estimate of the expected prediction error (2.4) of an arbitrary predictor $f(X;\hat{\beta})$ defined by the statistician with respect to an arbitrary prediction error. The most natural and widely used estimate is the *naive resubstitution* estimate of the prediction error of the predictor which is often called *apparent prediction error* in the literature and is defined by the formula :

$$\widehat{PE}(\hat{\beta}) = (1/n)\sum_{i=1}^{n} L(Y_i, f(X_i;\hat{\beta})) \tag{2.5}$$

One of the most serious drawback of the naive estimate (2.5) is that it *underestimates* the true prediction error in most cases. As is well-known in the variables-selection problem in the regression and discriminant analyses , this may cause a serious problem of *overfit* when we have several alternative predictors which may have widely differnt number of parameters.

In order to remove the systematic negative bias of the naive resubstituion estimate and to get the better estimate of the prediction error , we first introduce the notion of *excess error (or optimism)* of the naive estimate (2.5) by ;

$$\Delta(\widehat{\beta_k}) = PE(\widehat{\beta_k}) - \widehat{PE}(\widehat{\beta_k})\ . \tag{2.6}$$

Next lemma , which is easy to prove but very useful , will be the basis for the derivation of the right bias-corection rule and finally will provide better estimate of the true prediction error of the arbitrary prediction rule defined by the general M-estimator with respect to arbitrary measures of prediction errors.

Lemma. Let $L(y,f(x;\beta))$ be a function which is twice differentiable with respect to $\beta\in\theta\subset R^k$ for any $(x,y)\in S$. Suppose we can interchange the integral and differentiation signs in the following . Then we have :

$$\begin{aligned}\Delta(\hat{\beta}) &= PE(\hat{\beta}) - \widehat{PE}(\hat{\beta})\\ &= E_P(L(Y,f(X;\beta)) - \sum_{i=1}^{n} L(Y_i,f(X_i;\beta))/n\\ &\quad + E_P[L_\beta(Y,f(X;\beta))\cdot(\hat{\beta}-\beta)] - \sum_{i=1}^{n} L_\beta(Y_i,f(X_i,\beta))/n\cdot(\hat{\beta}-\beta) \qquad (2.7)\\ &\quad + (1/2)(\hat{\beta}-\beta)'\,[E_P L_{\beta\beta}(Y,f(X;\bar{\beta})) - \sum_{i=1}^{n} L_{\beta\beta}(Y_i,f(X_i;\bar{\beta}))/n](\hat{\beta}-\beta)\end{aligned}$$

where $L_\beta=[\partial L/\partial\beta_i]$ is a $1\times k$ gradient vector , $L_{\beta\beta}=[\partial^2 L/\partial\beta_i\partial\beta_j]$ is a $k\times k$ Hesseian matrix and $\bar{\beta}=\lambda\beta+(1-\lambda)\hat{\beta}$, $0\le\lambda\le 1$.

Proof . If we expand $\Delta(\hat{\beta})$ around β upto second order terms by the Taylor series , we get the result ;

$$\Delta(\hat{\beta})=\Delta(\beta)+\Delta_\beta(\beta)(\hat{\beta}-\beta)+(\hat{\beta}-\beta)'\,\Delta_{\beta\beta}(\bar{\beta})(\hat{\beta}-\beta)/2$$

immediately where $\Delta_\beta=[\partial\Delta/\partial\beta_i]$ is a $1\times k$ vector and $\Delta_{\beta\beta}=[\partial^2\Delta/\partial\beta_i\partial\beta_j]$ is a $k\times k$ matrix . This completes the proof.

Remark 1. The above formula (2.7) represents the canonical decomposition of the excess error $\Delta(\hat{\beta})$ of the naive resubtitution estimate $\widehat{PE}(\hat{\beta})$ into three parts :

$$\Delta(\hat{\beta})=A_n+B_n+r_n \qquad (2.8)$$

where $A_n=\Delta(\beta)$ is a *random* part with zero expectation and $B_n=\Delta_\beta(\beta)(\hat{\beta}-\beta)$ represents the *systematic bias* term and finally $r_n=o(1/n)$ is a small error term which is negligible in most cases.

One immediate consequence of the above lemma is the simple representation of the expected excess error of the arbitrary predictor defined by the M-estimator . Suppose that we have an estimator $\hat{\beta}$ of β which is *asymptotically linear* in the sense that :

$$\hat{\beta} - \beta = \sum_{i=1}^{n} M(X_i, Y_i; \beta, P)/n + o(1/\sqrt{n}) \tag{2.9}$$

where $M(X,Y;\beta,P)$ is the $k \times 1$ vector of influence function of the estimator $\hat{\beta}$ such that $E_P[M(X,Y;\beta,P)] = 0$.

Then we note that ;

$$E_n[\Delta(\hat{\beta})] = -[L_\beta, M]/n + o(1/n) \tag{2.10}$$

where $[L_\beta, M] = E_p(L_\beta \cdot M) = E_P[\sum_{i=1}^{k}(\partial L/\partial \beta_i) \cdot M_i]$.

Remark 2. Above expression (2.10) for the expected optimism show the average amount of under-estimation of the naive resubstitution estimate of prediction error of the predictor defined by the M-estimator $\hat{\beta}$.

In practice we have to use some empirical estimate of the bias term $[L, M]$. For example we can use the estimate ;

$$\overline{[L,M]} = \sum_{i=1}^{n} L_\beta(Y_i, X_i; \hat{\beta}) \cdot M(X_i, Y_i; \hat{\beta}, \widehat{P_n})/n \tag{2.11}$$

where $\widehat{P_n}$ is the empirical distribution of the random sample $S = \{(X_i, Y_i)\}_{i=1}^{n}$.

Motivated by the above result, we now introduce the following bias-corrected estimate of the expected prediction error of the predictor $f(X;\hat{\beta})$.

Definition. We define the bias-corrected estimate $PE_A(\hat{\beta})$ of the expected prediction error $E_n[PE(\hat{\beta})]$ by ;

$$PE_A(\hat{\beta}) = \widehat{PE}(\hat{\beta}) - \overline{[L,M]}/n . \tag{2.12}$$

Remark 3. Suppose that $L(Y,X;\beta) = -\log f(Y|X,\beta)$ for some parametric family of conditional probability density functions of Y given X and assume $\hat{\beta} = \arg\max_\beta \prod_{i=1}^{n} f(Y_i | X_i, \beta)$ is the conditional MLE of the parameter β . Then , under usual regularity conditions , we have typically ;

$$\hat{\beta} - \beta = -[E_P L_{\beta\beta}(Y,X;\beta)]^{-1} \sum_{i=1}^{n} L_\beta(Y_i,X_i;\beta)/n + o(1/\sqrt{n}) . \tag{2.13}$$

Thus our expression for the bias-corrected estimate of the expected prediction error reduces to the trace-type criteria which is sometimes called TIC (Trace Information Criterion) in the literature ;

$$TIC(\hat{\beta}) = -\sum_{i=1}^{n} \log f(Y_i|X_i, \hat{\beta}) + Tr[(\overline{L_{\beta\beta}})^{-1} \cdot \overline{(L_\beta^t L_\beta)}] \tag{2.14}$$

where

$$\overline{L_{\beta\beta}} = \sum_{i=1}^{n} L_{\beta\beta}(Y_i, X_i; \hat{\beta})/n$$

$$\overline{(L_\beta^t, L_\beta)} = \sum_{i=1}^{n} L_\beta^{'} (Y_i, X_i; \hat{\beta}) L_\beta(Y_i, X_i; \hat{\beta})/n .$$

See the Appendix of Linhart and Zucchini (1986) for more detailed reguraity conditions in this special case.

Remark 4. If we further assume that the conditional distribution of Y given X has the probability density function $f(Y|X;\beta)$ for some β with respect to a dominating measure μ_y in S_y , we get

$$Tr[(E_P L_{\beta\beta})^{-1}(E_P(L_\beta^{'}, L_\beta))] = Tr(I_k) = k$$

and our criterion reduces to the simpler criterion AIC ;

$$-AIC(\hat{\beta}) = -\sum_{i=1}^{n} \log f(Y_i|X_I, \hat{\beta}) + k . \tag{2.15}$$

Remark 5. As is noted by Efron (1983) , there are two well-known non-parametric estimates of the expected prediction errors , Cross-Validation and Boothtrap estimates , which are defined respectively by ;

$$PE_{CV} = \sum_{i=1}^{n} L(Y_i, X_i; \hat{\beta}_{-i})/n \tag{2.16}$$

$$PE_{Boot} = \widehat{PE}(\hat{\beta}) + E^*_{\hat{P}_n}[\widehat{PE}(\hat{\beta}^*) - \widehat{PE}^*(\hat{\beta}^*)] \tag{2.17}$$

where $\hat{\beta}_{-i}$ is the estimate of β computed from the deleted data set $S_{-i} = \{(X_1,Y_1), \cdots, (X_{i-1},Y_{i-1}), (X_{i+1},Y_{i+1}), \cdots (X_n,Y_n)\}$, $\widehat{P_n}$ is the empirical distribution of the data set $S=\{(X_i,Y_i)\}_{i=1}^n$, $S^*=\{(X_i^*,Y_i^*)\}_{i=1}^n$ is the Boothtrap sample of size n drawn from the empirical distribution $\widehat{P_n}$ and $\widehat{\beta^*}$ is the estimate of β computed from the Boothtrap sample S^* and the expectation $E^*_{\widehat{P_n}}[\cdot]$ in (2.17) is taken with respect to the Boothtrap sample S^*.

Stone (1977) demonstrated that PE_A and PE_{CV} are asymptotically equivalent when the condition (2.13) holds. On the other hand, if we apply (2,10) to the Boothtrap sample S^* drawn from the empirical distribution $\widehat{P_n}$, we obtain the result ;

$$\widehat{\beta^*} - \hat{\beta} = \sum_{i=1}^n M(X_i^*,Y_i^*;\hat{\beta},\ \hat{P}_n)/n + o(1/\sqrt{n}) . \tag{2.18}$$

This in turn implies that ;

$$\begin{aligned} PE_{Boot} &= \widehat{PE}(\hat{\beta})+E^*_{\widehat{P_n}}[\Delta(\widehat{\beta^*})] \\ &= \widehat{PE}(\hat{\beta}) - E^*_{\widehat{P_n}}[L(Y^*,X^*;\hat{\beta})M(Y^*,X^*;\hat{\beta},\ \hat{P}_n)]/n + o(1/n) \\ &= \widehat{PE}(\hat{\beta}) - \overline{[L,M]}/n + o(1/n) \\ &= PE_A(\hat{\beta}) + o(1/n) \end{aligned} \tag{2.19}$$

where we have used the lemma applied to the Boothtrap sample S^* drawn from the empirical distribution $\hat{P}_n$. This establishes the asymptotic equivalence of PE_A and PE_{Boot}.

3. Examples and Discussions

In this section we give several examples which illustrate the computation of the appropriate bias-correction terms for the apparent prediction errors of the various non-linear predictors with respect to different measures of prediction errors .

Example 1. (Linear Regression) Here we consider the linear predictor :

$$\hat{y}(x)=f(x;\hat{\beta})=\sum_{i=1}^k \widehat{\beta_i}x_i \tag{3.1}$$

based on the OLS (Ordinary Least Squares) estmator $\hat{\beta}$ given by :

$$\hat{\beta}=(S_{XX})^{-1}S_{XY}$$

where $S_{XX}=[\sum_{i=1}^{n} X_{ij}X_{ik}/n]$ is a $k\times k$ matrix and $S_{XY}=[\sum_{i=1}^{n} Y_i X_{ij}/n]$ is a $k\times 1$ vector. If we use the usual square-error loss function $L(y,\hat{y})=(y-\hat{y})^2$, we obtain, as a bias-corrected estimate of the prediction error of the linear predictor (3.1), the expression ;

$$PE_A(\hat{\beta}) = SSE(\hat{\beta})/n + 2\cdot Tr(S_{XX}^{-1}S_{XX}^{*})/n \tag{3.2}$$

immediately from (2.12) where $S_{XX}^{*}=[\sum_{i=1}^{n} e_i^2 X_{ij}X_{ik}/n]$, $SSE(\hat{\beta})=\sum_{i=1}^{n} e_i^2$ and $e_i = Y_i-\hat{\beta}' X_i$.

Example 2. (**Ridge Regression**) Suppose we use the ridge-regression estimator $\hat{\beta}(\alpha)$ of β ;

$$\hat{\beta}(\alpha)=(S_{XX}+\alpha I_k)^{-1}S_{XY} \quad , \alpha>0 \tag{3.3}$$

instead of the OLS estimator $\hat{\beta}$. Then we obtain the following bias-corrected estimate of the prediction error of the corresponding predictor ; $f(X, \hat{\beta}(\alpha)) = \hat{\beta}(\alpha)' X$.

$$PE_A(\hat{\beta}(\alpha)) = SSE(\hat{\beta}(\alpha))/n + 2Tr[(S_{XX}+\alpha I_k)^{-1}S_{XX}^{**}] \tag{3.4}$$

where $SSE(\hat{\beta}(\alpha))=\sum_{i=1}^{n}(Y_i-\hat{\beta}(\alpha)' X_i)^2$ and $S_{XX}^{**}=[\sum_{i=1}^{n}(e_iX_i+\alpha\beta)(e_iX_i+\alpha\beta)'/n]$. Note that we can use the minimizer of the expression (3.4) as an alternative estimator of the smoothness parameter α in the ridge-regression estimator .

Example 3. (**Logistic Regression**) Here we assume that we have a binary response variable Y with a vector X of several predictor variables and the conditionnal distribution of Y given X is a Bernoulli distribution with success probability $P(Y=1|X)=p(X)$. Suppose also that we have a simple logistic model ;

$$\log(p(X_i)/(1-p(X_i))) = \beta' X_i = \sum_{j=1}^{k}\beta_j X_{ij} \quad , i=1,\cdots,n$$

where $X_i=[X_{ij}]$ is a $k\times 1$ vector of regressor variables i-th subject.

Then the MLE $\hat{\beta}$ of β , which is defined implicitly as the unique solution of the likelihood equation, satisfies the following relation ;

$$\hat{\beta} - \beta = \left(\sum_{i=1}^{n} p_i(1-p_i)X_iX_i'/n\right)^{-1}\left(\sum_{i=1}^{n}(Y_i-p_i)X_i/n\right) + o(1/\sqrt{n})$$

where $p_i = p(X_i;\hat{\beta}) = \widehat{\beta' X_i}$. If we use the loss function $L(y,p)=(y-p)^2$, then we obtain the bias-corrected estimate of the prediction error of the predictor $p(X;\hat{\beta})$;

$$PE_A(\hat{\beta}) = SSE/n + (2/n)\,Tr\left[\sum_{i=1}^{n} p_i(1-p_i)X_iX_i'/n\right]^{-1}\left(\sum_{i=1}^{n} e_i^2 p_i(1-p_i)X_iX_i'/n\right) \quad (3.5)$$

where $SSE = \sum_{i=1}^{n} e_i^2$, $e_i = Y_i - p(X_i;\hat{\beta})$. On the other hand, if we use the minus log-likelihood as a loss function : $L(y,p) = -ylog(p/(1-p) + \log(1-p)$, we get the following result ;

$$PE_A(\hat{\beta}) = \sum_{i=1}^{n} L(Y_i, p(X_i,\hat{\beta}))/n + Tr\left[\left(\sum_{i=1}^{n} p_i(1-p_i)X_iX_i'/n\right)^{-1}\left(\sum_{i=1}^{n} e_i^2 X_iX_i'/n\right)\right] . \quad (3.6)$$

Remark 6. If we consider a *fixed*-regressor regression model and use the appropriate definition of prediction errors as in Efron (1986), we may derive an analogue of the lemma for the fixed-regressor case and obtain the similar bias-correcton rule for the apparent prediction errors. This possibility and other modifications will be considered in a seperate paper.

References

[1] Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Princple, In B.N. Petrov and F. (saki ceds.), Second International Symposium on Informaton Theory. Budapest 1,Akademiai Kiado, 649-660.

[2] Efron, B. (1983). Estimating the Error Rate of a Prediction Rule ; Improvements on Cross-Validation, *Journal of American Statistical Association*, Vol. 78, 316-331.

[3] Efron, B. (1986). How Biased is the Apparent Error Rate of a Prediction Rule ?, *Journal of the American Statistical Association*, Vol. 81, 461-470.

[4] Linhart, H. and Zucchini, W. (1986). *Model Selection,* Wiley, New-York.

[5] Stone, M. (1977). An Asymptotic equivalence of Choice of Model by Cross-Validation and Akaike's Criterion, *Journal of the Royal Statistical Society,* B 39, 447-47.