Ann. Inst. Statist. Math.
Vol. 45, No. 1, 105–111 (1993)

ON THE ESTIMATION OF PREDICTION ERRORS IN LINEAR REGRESSION MODELS

PING ZHANG

Department of Statistics, The Wharton School of the University of Pennsylvania, 3000 Steinberg Hall-Dietrich Hall, Philadelphia, PA 19104-6302, U.S.A.

(Received September 9, 1991; revised March 6, 1992)

Abstract. Estimating the prediction error is a common practice in the statistical literature. Under a linear regression model, let e be the conditional prediction error and $\hat{e}$ be its estimate. We use $\rho(\hat{e}, e)$, the correlation coefficient between e and $\hat{e}$, to measure the performance of a particular estimation method. Reasons are given why correlation is chosen over the more popular mean squared error loss. The main results of this paper conclude that it is generally not possible to obtain good estimates of the prediction error. In particular, we show that $\rho(\hat{e}, e) = O(n^{-1/2})$ when $n \to \infty$. When the sample size is small, we argue that high values of $\rho(\hat{e}, e)$ can be achieved only when the residual error distribution has very heavy tails and when no outlier presents in the data. Finally, we show that in order for $\rho(\hat{e}, e)$ to be bounded away from zero asymptotically, $\hat{e}$ has to be biased.

Key words and phrases: Conditional prediction error, correlation.

1. Introduction

One of the ultimate goals of statistical modelling is to be able to predict future observations based on currently available information. Such is the case particularly in time series and regression analysis. One often judges the goodness of a model by looking at its prediction error. Many statistical methodologies are developed based on such a consideration. Examples include model selection and nonparametric smoothing methods, where the model and the smoothing parameter respectively are chosen so that the estimated prediction error is minimized. Proper estimation of the prediction error is of crucial importance in these areas. For references, see Linhart and Zucchini (1986), Bickel and Zhang (1991), Breiman and Freedman (1983), Härdle *et al.* (1988).

There are usually two notions of prediction error appearing in the literature, namely the conditional and unconditional errors. Suppose that we are concerned with a response variable Y that could be vector valued. Let $\tilde{Y}$ represent a future observation. Let $\hat{Y}$ be the predicted value based on the current data. Let $\mathcal{F}$ denote the σ-field generated by the sample, the conditional mean squared prediction error

is defined as $e = E[\|\tilde{Y} - \hat{Y}\|^2 | \mathcal{F}]$. The unconditional prediction error is simply $E(e)$. In the literature, when an estimate is constructed, it is often unclear which of the two prediction errors we are estimating. One often perceives the problem as that of estimating $E(e)$ simply because it fits more readily into the classical framework of parameter estimation. In practice, however, e is by all means a more honest measure of the prediction error because it measures how well one can do given the data at hand. The unconditional prediction error, on the other hand, is only an index of average performance which may or may not represent the current data. The idea of estimating an unknown random quantity has been familiar to statisticians for a long time. Bayes theory provides a good example. For non-Bayesians, the so called random effect models fit into this category (Robinson (1991)). We consider only the specific case of estimating the squared error loss which depends on both the parameters and the observations. It is without doubt that $E(e)$ would be easier to estimate than e. Härdle *et al.* (1988), in the context of bandwidth selection in kernal nonparametric regression, actually showed that the conclusion can be very different depending on which prediction error is used and the unconditional version has clear advantages. It is fair to say nonetheless that there is still substancial disagreement over the use of e or $E(e)$ in the literature.

The current work is prompted by an observation made by Härdle *et al.* (1988) which suggests that the optimal bandwidths based on $\hat{e}$ and e have a negative correlation coefficient. Some explanations of this phenomenon can be found in Johnstone (1988), Chiu and Marron (1990) and Johnstone and Hall (1991). Although not directly applicable, the results of this paper could shed some light on the above problem because the case for linear regression is much easier to understand and more insight could be obtained.

The focus of this paper is to study, for the linear regression models, the performance of $\hat{e}$ as an estimate of the conditional prediction error e. We use $\rho(\hat{e}, e)$, the correlation coefficient between $\hat{e}$ and e, to measure the performance of a particular estimation method. Traditionally, metric like error measures such as the mean squared error $E(\hat{e} - e)^2$ are considered to be more fundamental as a measure of performance. Decision theory is largely developed around such loss functions. There are two reasons that lead us to consider the correlation coefficient $\rho(\hat{e}, e)$ as an alternative to the popular mean squared loss function. First of all, many of the model selection criteria amount to minimizing $\hat{e}$. This is based on the hope that if $\hat{e}$ is small, then e should also be small, hence the selected model is good. Correlation coefficient provides a natural gauge for measuring this kind of relationship. Secondly, it has been shown by Johnstone (1988) that under the mean squared loss function, many natural estimate of e are not even admissible.

Among the main conclusions of the paper, we show that asymptotically, $\hat{e}$ fails to capture any structure of e in the sense that the correlation coefficient $\rho(\hat{e}, e) = O(n^{-1/2})$ as $n \to \infty$. This is the case whenever we require $E(\hat{e}) = E(e)$. Unlike parameter estimation, here there is no reason to believe that unbiasedness is desirable. Thus some attempt has been made to increase $\rho(\hat{e}, e)$ by relaxing the unbiasedness requirement. When the sample size is small, we argue that high values of $\rho(\hat{e}, e)$ can be achieved only when the residual error distribution has very heavy tails and when no outlier presents in the data. In general, it seems

impossible to obtain good estimates for the prediction error.

2. Regression models with fixed design

Let $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^t$ where the ϵ_i's are independent identically distributed random variables with $E\epsilon_i = 0$, $E\epsilon_i^2 = \sigma^2$ and $E\epsilon_i^4 < \infty$. Suppose that A and B are non-negative definite $n \times n$ matrices. One can easily show that

$$E(\epsilon^t A \epsilon) = \sigma^2 \operatorname{tr}(A) \tag{2.1}$$

and

$$\operatorname{cov}(\epsilon^t A\epsilon, \epsilon^t B\epsilon) = \kappa_4 \sum a_{ii} b_{ii} + 2\sigma^4 \operatorname{tr}(AB), \tag{2.2}$$

where $\kappa_4 = E(\epsilon_i^4) - 3\sigma^4$ is the fourth cumulant of ϵ_i. We denote by a_{ij} and b_{ij} the elements of A and B respectively.

In this section, we consider the linear regression model $Y = X\beta + \epsilon$, where X is a fixed $n \times k$ design matrix and the residual error vector $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^t$ satisfies the conditions listed above. Suppose that $\hat{\beta} = (X^t X)^{-1} X^t Y$ is the least squares estimator of β. Using the notations of Section 1, a future value is predicted by $\hat{Y} = X\hat{\beta}$. We can easily verify that the conditional prediction error as defined in Section 1 is given by $e = n\sigma^2 + \epsilon^t P_k \epsilon$, where $P_k = X(X^t X)^{-1} X^t$ is the projection matrix.

It follows from (2.1) that $E(e) = (n + k)\sigma^2$. Thus to estimate e, a natural choice would be $\hat{e} = (n + k)\hat{\sigma}^2$, where $\hat{\sigma}^2$ is an estimate of σ^2. In this paper, we restrict ourselves to quadratic estimators $\hat{\sigma}^2 = Y^t B Y$ for some non-negative definite matrix B. In order for $\hat{e}$ to be unbiased, we must have

$$BX = 0 \quad \text{and} \quad \operatorname{tr}(B) = 1. \tag{2.3}$$

Let $A = P_k$ and B be arbitrary. Since $A^2 = A$ and $\operatorname{tr}(A) = k$, we have from (2.1) and (2.2) that

$$\rho(\hat{e}, e) = \frac{\kappa_4 \sum a_{ii} b_{ii}}{\sqrt{\kappa_4 \sum a_{ii}^2 + 2k\sigma^4} \sqrt{\kappa_4 \sum b_{ii}^2 + 2\sigma^4 \operatorname{tr}(B^2)}}. \tag{2.4}$$

Some observations are in order here. The correlation between $\hat{e}$ and e is not always positive. When $\kappa_4 = 0$, as is the case for normally distributed errors, the correlation coefficient is zero. Presumably, a good estimate $\hat{e}$ should have high correlation with e. In this sense, an estimate such that $\rho(\hat{e}, e) = 0$ hardly make any sense because it fails to capture any structure of e. When $\rho(\hat{e}, e)$ is negative, the situation is even worse because now our estimate $\hat{e}$ is totally misleading. The correlation coefficient is positive if and only if $\kappa_4 > 0$, i.e., the residual error distribution has heavier tail than normal distribution. Another observation is that for any given B, $\rho(\hat{e}, e)$ depends only on the a_{ii}'s, i.e., the leverages of the design points. Thus the robustness of the design matrix has a direct impact on

the quality of $\hat{e}$ as an estimate of e. This issue will be further discussed in the next section.

The usual estimator of σ^2 based on residual sum of squares corresponds to $B = (n-k)^{-1}P_k^{\perp}$, where $P_k^{\perp} = I - P_k$. Under the assumption of normal distribution for the residual errors, this estimator has the property that $\operatorname{var}(\hat{e})$ is minimum among all estimators satisfying (2.3). This can be seen by noting that $\operatorname{var}(\hat{e}) = 2\sigma^4 \operatorname{tr}(B^2)$. Since $BX = 0$, B has at most $n-k$ non-zero eigenvalues. Let them be $\lambda_1, \ldots, \lambda_{n-k}$. Since $\sum \lambda_i = \operatorname{tr}(B) = 1$, we have $\operatorname{tr}(B^2) = \sum \lambda_i^2 \geq (n-k)^{-1}$ due to the Cauchy-Schwarz inequality. The equality holds if and only if $\lambda_1 = \cdots = \lambda_{n-k} = (n-k)^{-1}$, which corresponds to $B = (n-k)^{-1}P_k^{\perp}$.

THEOREM 2.1. *Suppose that B satisfies* (2.3) *and that* $\max_{1\leq i\leq n} a_{ii} = O(n^{-1})$. *Then* $\rho(\hat{e}, e) = O(n^{-1/2})$. *Furthermore, if* $B = (n-k)^{-1}P_k^{\perp}$, *then*

$$\rho(\hat{e}, e) = \frac{k\kappa_4}{\sqrt{2\sigma^4(2\sigma^4 + \kappa_4)}} \cdot n^{-1/2} + o(n^{-1/2}).$$

PROOF. For a general B, the assumption implies that

$$\kappa_4 \sum a_{ii}^2 + 2k\sigma^4 = 2k\sigma^4(1 + O(n^{-1})).$$

Since $\operatorname{tr}(B^2) = \sum_{i,j} b_{ij}^2 \geq \sum b_{ii}^2$, we have

$$\kappa_4 \sum b_{ii}^2 + 2\sigma^4 \operatorname{tr}(B^2) \geq (\kappa_4 + 2\sigma^4) \sum b_{ii}^2.$$

Finally, the Cauchy-Schwarz inequality implies that $\sum a_{ii}b_{ii} \leq O(n^{-1/2}\sqrt{\sum b_{ii}^2})$. The conclusion follows by substituting the above inequalities into (2.4).

If $B = (n-k)^{-1}P_k^{\perp}$, notice that $b_{ii} = (n-k)^{-1}(1 - a_{ii})$ and $\operatorname{tr}(B^2) = (n-k)^{-1}$, we have $\kappa_4 \sum b_{ii}^2 + 2\sigma^4 \operatorname{tr}(B^2) = n^{-1}(2\sigma^4 + \kappa_4)(1 + o(1))$ and $\sum a_{ii}b_{ii} = n^{-1}k\kappa_4(1 + o(1))$. The conclusion follows immediately. □

On the one hand, the above theorem states that it is in general impossible to obtain good estimates for e since $\rho(\hat{e}, e)$ always tends to zero. On the other hand, one rarely has $k/n \to 0$ in practice. When k/n is bounded away from zero, Theorem 2.1 does not apply since the assumption $a_{ii} = O(n^{-1})$ is no longer valid. We shall argue below that the value of $\rho(\hat{e}, e)$ is essentially determined by the kurtosis of the residual error distribution.

3. Some small sample considerations

We consider in this section only the case where $B = (n-k)^{-1}P_k^{\perp}$. In other words, the σ^2 is to be estimated by the usual residual variance. Let $h_i = a_{ii}$ be the i-th diagonal element of P_k. Then (2.4) becomes

$$\rho(\hat{e}, e) = \frac{\kappa_4 \sum h_i(1 - h_i)}{\sqrt{\kappa_4 \sum h_i^2 + 2k\sigma^4}\sqrt{\kappa_4 \sum (1 - h_i)^2 + 2(n-k)\sigma^4}}.$$

Denote by ρ_4 the kurtosis for the distribution of the residual errors, i.e., $\rho_4 = \kappa_4/\sigma^4$. Recall that ρ_4 measures the heaviness of the corresponding distribution as compared with normal distribution. When $\rho > 0$ the tail is heavier than normal while a negative ρ value indicates the opposite. By the Cauchy-Schwarz inequality,

$$\rho(\hat{e}, e) \leq \sqrt{\frac{\rho_4 \sum h_i^2}{\rho_4 \sum h_i^2 + 2k}} \sqrt{\frac{\rho_4 \sum (1-h_i)^2}{\rho_4 \sum (1-h_i)^2 + 2(n-k)}}.$$

Thus $\rho(\hat{e}, e)$ can be very small if either $2k \gg \rho_4 \sum h_i^2$ or $2(n-k) \gg \rho_4 \sum (1-h_i)^2$. In particular, these are the case if $\rho_4 \ll 2$. We only need to verify the first inequality. Now $\rho_4 \ll 2$ implies that

$$2k \gg \rho_4 k = \rho_4 \sum h_i \geq \rho_4 \sum h_i^2.$$

The last inequality above is because $h_i \leq 1$ by the properties of projection matrices.

Next, let us consider the case when $\rho_4 \to \infty$. We then have

$$\rho(\hat{e}, e) \to \frac{\sum h_i(1-h_i)}{\sqrt{\sum h_i^2} \sqrt{\sum (1-h_i)^2}}.$$

In the special case where all the h_i are the same, the above limit equals 1. Consequently, $\rho(\hat{e}, e)$ may be very close to 1 if the h_i's are relatively homogenous. Using the language of robust statistics, a large value for h_i indicates that the i-th observation is an outlier.

The above arguments seem to suggest that the ordinary prediction error estimate is good only when the residual error disbribution has very heavy tails and that no outlier presents in the observed data. One might argue that a small correlation coefficient does not necessarily mean that $\hat{e}$ is not related to e the right way because they might have a non-linear relationship. This, however, could not happen in our case. Consider simply the case where the residual errors are normally distributed. Not only does $\rho(\hat{e}, e) = 0$, $\hat{e}$ and e are also independent of each other!

4. Biased estimates

In this section, we assume that the residual errors are normally distributed. In this case, the correlation coefficient between $\hat{e}$ and e equals to zero as long as we require $E(\hat{e}) = E(e)$. We shall show below that by relaxing the unbiasedness requirement, it is possible to keep $\rho(\hat{e}, e)$ bounded away from zero uniformly for all $n \geq 1$.

A slight modification of (2.2) yields that

$$\rho(\hat{e}, e) = \frac{2\sigma^4 \operatorname{tr}(BP_k)}{\sqrt{2k\sigma^4} \sqrt{\sigma^2 \|BX\beta\|^2 + 2\sigma^4 \operatorname{tr}(B^2)}}.$$

Throughout this section, we assume that $\hat{e}$ takes the form Y^tBY and $B = aP_k + bP_k^{\perp}$ where a and b are scaler constants. It is easy to verify that $\mathrm{tr}(BP_k) = ak$, $\|BX\beta\|^2 = a^2\|X\beta\|^2$ and $\mathrm{tr}(B^2) = a^2k + b^2(n-k)$. Hence

$$\rho(\hat{e}, e) = \frac{2\sigma^4 ak}{\sqrt{2k\sigma^4}\sqrt{\sigma^2 a^2\|X\beta\|^2 + 2\sigma^4(a^2k + b^2(n-k))}}. \tag{4.1}$$

Without loss of generality, assume that $a > 0$ is bounded. Thus in order for (4.1) to be bounded away from zero, one must require $\|X\beta\| \leq C$ and $b \leq Cn^{-1/2}$ for some $C < \infty$. We state the conclusion in the following theorem.

THEOREM 4.1. *Let ϵ_i's be independent identically distributed with distribution $N(0, \sigma^2)$ and $\hat{e} = Y^tBY$ with $B = aP_k + bP_k^{\perp}$ as above. Suppose that there exists constant $C > 0$ such that $0 < a < C$, $|b| < Cn^{-1/2}$ and $\|X\beta\|^2 < C$. Then we can find $\rho_0 > 0$ such that $\rho(\hat{e}, e) \geq \rho_0$ for all $n \geq 1$.*

In general, a condition such as $\|X\beta\| \leq C$ is not regarded as a reasonable one. When $\|X\beta\| \to \infty$, (4.1) implies that

$$\rho(\hat{e}, e) = \frac{\sigma\sqrt{2k}}{\|X\beta\|} \cdot (1 + o(1)).$$

This suggests, as also observed earlier in Section 2, that we consider the kind of asymptotics where k/n does not tend to zero. If this is case, then in order for (4.1) to be bounded away from zero, one must have $\|X\beta\| = O(n^{1/2})$ and $|b/a| = O(1)$. Combining this with the situation described in Theorem 4.1, we see that whatever the case, in order for (4.1) to be non-zero asymptotically, b/a must be bounded. In particular, a must not equal to zero. This means that the estimate $\hat{e}$ must be biased.

To assess the bias, let us write

$$E(\hat{e}) - E(e) = a^2\|X\beta\|^2 + \sigma^2[ak + b(n-k)] - (n+k)\sigma^2.$$

It is clear that under the set up of Theorem 4.1, $\hat{e}$ is severely biased downward. In fact, since $\|X\beta\|$ is bounded, $n^{-1}[E(\hat{e}) - E(e)] \to -\sigma^2$. For the case when k/n does not tend to zero, we can actually solve the equation $E(\hat{e}) = E(e)$ to find the corresponding a and b. Notice that we need $a > 0$ because one wants $\rho(\hat{e}, e)$ to be positive. While such solutions are available if we set $b < (n+k)/(n-k)$, the solutions depend on the unknown parameters β and σ^2. We can always substitute the parameters by their corresponding estimates. Asymptotically, it is reasonable to expect that this will lead to an estimate that is approximately unbiased yet having positive correlation with e. It is still unclear how such a procedure works in practice.

Acknowledgements

The author wishes to thank J. S. Marron and I. Johnstone for bringing to his attention some important references. This work is supported by the Research Foundation of the University of Pennsylvania.

References

Bickel, P. and Zhang, P. (1991). Variable selection in non-parametric regression with categorical covariates, *J. Amer. Statist. Assoc.*, **87**, 90–97.

Breiman, L. and Freedman, D. (1983). How many variables should be entered in a regression equation, *J. Amer. Statist. Assoc.*, **78**, 131–136.

Chiu, S. T. and Marron, J. S. (1990). The negative correlations between data-determined bandwidths and the optimal bandwidth, *Statist. Probab. Lett.*, **10**, 173–180.

Härdle, W., Hall, P. and Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum?, *J. Amer. Statist. Assoc.*, **83**, 86–95.

Johnstone, I. (1988). On inadmissibility of some unbiased estimates of loss, *Statistical Decision Theory and Related Topics IV* (eds. S. S. Gupta and J. O. Berger), Vol. 1, 361–379, Springer, New York.

Johnstone, I. and Hall, P. (1991). Empirical functionals and efficient smoothing parameter selection, Tech. Report No. 373, Department of Statistics, Stanford University, California.

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

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects (with discussions), *Statist. Sci.*, **6**, 15–51.