Part 4b Asymptotic Results for MRR2 using PRESS

Recall that the PRESS statistic is a special type of cross validation procedure (see Allen (1971)) particular to the regression problem and involves finding $\hat{Y}_{i,-i}$, the estimate at the ith observation found by removing the ith data pair, ($\mathbf{x}_i$, Y_i) from the data set. In the MRR2 case

$$\text{PRESS}(\lambda) = \sum_{i=1}^{n} (y_i - (\hat{f}^{(i)} + \lambda \hat{g}^{(i)}))^2 .$$

Once again, we choose λ by finding the value of λ that minimizes PRESS(λ). This is done by setting $\frac{d}{d\lambda}$PRESS(λ) = PRESS'(λ) = 0 and solving for λ. We obtain PRESS'(λ) =

$$\sum_{i=1}^{n} 2(Y_i - (\hat{f}^{(i)} + \lambda \hat{g}^{(i)}))(-\hat{g}^{(i)})$$

$$= \sum_{i=1}^{n} -2(Y_i \hat{g}^{(i)} - (\hat{f}^{(i)} \hat{g}^{(i)} + \lambda \hat{g}^{(i)2})) = 0$$

Solving this equation results in $\hat{\lambda}^P$ as

$$\hat{\lambda}^P = \frac{\sum_{i=1}^{n} \hat{g}^{(i)}(Y_i - \hat{f}^{(i)})}{\sum_{i=1}^{n} (\hat{g}^{(i)})^2} = \frac{< \hat{g}^{(i)}, Y - \hat{f}^{(i)} >}{\left\| \hat{g}^{(i)} \right\|^2} .$$

Observe that this parameter estimate is similar to $\hat{\lambda}^*$ except that the parametric and nonparametric estimates have been replaced with the analogous cross validated estimates. Of course we must ensure that PRESS''(λ) > 0. This follows from

$$\text{PRESS''}(\lambda) = \sum_{i=1}^{n} 2\hat{g}^{(i)2} > 0, \text{ for } \lambda \in \Re ,$$

except for the degenerate case which we will not worry about here. Thus $\hat{\lambda}^P$ does, in fact, produce a global minimum. And it is this estimate that we will study asymptotically in the remainder of the section.

We will again obtain the difference between λ^* and $\hat{\lambda}^P$, then investigate that difference asymptotically.

$$\lambda^* - \hat{\lambda}^P = \frac{<\hat{g}, \theta - \hat{f}>}{\|\hat{g}\|^2} - \frac{<\hat{g}^{(i)}, Y_i - \hat{f}^{(i)}>}{\|\hat{g}^{(i)}\|^2} \quad (4.B.1)$$

$$= \frac{\sum(\hat{g})(\theta - \hat{f})}{n(\|\hat{g}\|^2)} - \frac{\sum(\hat{g}^{(i)})(Y_i - \hat{f}^{(i)})}{n(\|\hat{g}\|^2 + (\|\hat{g}^{(i)}\|^2 - \|\hat{g}\|^2))}. \quad (4.B.2)$$

We have the same denominator problem as in section 3b. Recall from that section that the asymptotic rates for $\|\hat{f}^{(i)} - \hat{f}\|$ and $\|\hat{g}^{(i)} - \hat{g}\|$ are $O_P(n^{-1})$, and $O_P(\gamma_n^2)$ respectively (Burman and Chaudhuri (1992) results 6.21, and 6.20 respectively). Next, set α =

$$(\|\hat{g}^{(i)}\|^2 - \|\hat{g}\|^2) = \frac{\sum(\hat{g}^{(i)2} - \hat{g})^2}{n}.$$

Recall that in using cross validated estimates this difference is very important. We will need the following lemma and its corollary. These results deal with the difference term α and a closely related term which will prove important in the results that follow. The proofs for both results are found in appendix 4b.

Lemma 4.b.1: Assuming conditions A1-A6…

$$\alpha = \begin{cases} O_P(\gamma_n^{\,2}), & \text{if } \lim_{n\to\infty} \delta_n \neq 0 \\ O_P(\gamma_n^{\,3}), & \text{if } \delta_n = 0 \end{cases}.$$

Corollary 4.b.1: Assuming conditions A1-A6…

$$\|\hat{g}^{(i)}\| - \|\hat{g}\| = \begin{cases} O_P(\gamma_n), & \text{if } \lim_{n\to\infty} \delta_n \neq 0 \\ O_P(\gamma_n^{\,1.5}), & \text{if } \delta_n = 0 \end{cases}$$

An important artifact of this lemma is that α converges to zero faster than $\|\hat{g}\|$. This implies that the denominator on the right side of 4.B.2 can be handled (asymptotically) by dealing with $\|\hat{g}\|$.

Rewriting the right hand term in 4.B.2 we have

$$\frac{\sum(\hat{g}^{(i)})(Y_i-\hat{f}^{(i)})}{n(\|\hat{g}\|^2+(\|\hat{g}^{(i)}\|^2-\|\hat{g}\|^2))}$$

$$=\frac{\sum(\hat{g}^{(i)})(Y_i-\hat{f}^{(i)})}{n(\|g\|^2+\alpha)}$$

$$=\frac{\sum(\hat{g}^{(i)})(Y_i-\hat{f}^{(i)})}{n(\|\hat{g}\|^2)}\left(1-\frac{\alpha}{(\|\hat{g}\|^2+\alpha)}\right)$$

$$=\frac{\sum(\hat{g}^{(i)})(Y_i-\hat{f}^{(i)})}{n(\|\hat{g}\|^2)}-\frac{\sum(\hat{g}^{(i)})(Y_i-\hat{f}^{(i)})}{n(\|\hat{g}\|^2)}\left(\frac{\alpha}{\|\hat{g}\|^2+\alpha}\right).$$

As before, the left part of the last term is what we need to complete the problem. The right part, however, we must ultimately deal with and shall call it remainder term 1 ($R1$). We have the following lemmas that give asymptotic results for $R1$, and will ultimately provide a foundation for finding the asymptotic convergence rates for 4.B.2. The proofs for Lemmas 4.b.2 and 4.b.3 are found in appendix 4b.

Lemma 4.b.2: Assuming conditions A1-A6…

$$\frac{\sum(\hat{g}^{(i)})(Y_i-\hat{f}^{(i)})}{n(\|\hat{g}\|^2)}=\begin{cases}O_P(1), & if\ \lim_{n\to\infty}\delta_n\neq 0\\ O_P(n^{-.5}\gamma_n^{-1}), & if\ \delta_n=0\end{cases}.$$

Lemma 4.b.3: Assuming conditions A1-A6…

$$R1=\begin{cases}O_P(\gamma_n^{\,2}), & if\ \lim_{n\to\infty}\delta_n\neq 0\\ O_P(n^{-.5}), & if\ \delta_n=0\end{cases}$$

The importance of the preceding result is that it will allow us to rewrite 4.B.2 with a common denominator, which will lead to the important result of Lemma 4.b.4.

With our new notation 4.B.2 becomes

$$l^* - \hat{l}^P = \frac{\sum(\hat{g})(\theta - \hat{f})}{n\|\hat{g}\|^2} - \frac{\sum(\hat{g}^{(i)})(Y_i - \hat{f}^{(i)})}{n\|\hat{g}\|^2} + R1$$

$$= \frac{\sum(\hat{g}\theta - \hat{g}\hat{f}) - \sum(\hat{g}^{(i)}Y_i - \hat{g}^{(i)}\hat{f}^{(i)})}{n\|\hat{g}\|^2} + R1$$

$$= \frac{\sum((\hat{g}\theta - \hat{g}^{(i)}Y_i) + (\hat{g}^{(i)}\hat{f}^{(i)} - \hat{g}\hat{f}))}{n\|\hat{g}\|^2} + R1$$

$$= \frac{\sum(\hat{g}\theta - (\hat{g} + \hat{g}^{(i)} - \hat{g})Y_i) + \sum((\hat{g} + \hat{g}^{(i)} - \hat{g})(\hat{f} + \hat{f}^{(i)} - \hat{f}) - \hat{g}\hat{f})}{n\|\hat{g}\|^2} + R1$$

$$= \frac{\sum(\hat{g}\theta - (\hat{g} + \hat{g}^{(i)} - \hat{g})(\theta + e_i)) + \sum((\hat{g}^{(i)} - \hat{g})\hat{f} + \hat{g}(\hat{f}^{(i)} - \hat{f}) + (\hat{g}^{(i)} - \hat{g})(\hat{f}^{(i)} - \hat{f}))}{n\|\hat{g}\|^2} + R1$$

$$= \frac{\sum(-(\hat{g}^{(i)} - \hat{g})(\theta + e_i) - \hat{g}e_i) + \sum((\hat{g}^{(i)} - \hat{g})\hat{f} + \hat{g}(\hat{f}^{(i)} - \hat{f}) + (\hat{g}^{(i)} - \hat{g})(\hat{f}^{(i)} - \hat{f}))}{n\|\hat{g}\|^2} + R1$$

$$= \frac{-<(\hat{g}^{(i)} - \hat{g}), e_i> - <\hat{g}, e_i> + <(\hat{g}^{(i)} - \hat{g}), (\hat{f} - \theta)> + <\hat{g}, (\hat{f}^{(i)} - \hat{f})> + <(\hat{g}^{(i)} - \hat{g})(\hat{f}^{(i)} - \hat{f})>}{\|\hat{g}\|^2} + R1$$

$$\leq \frac{\|(\hat{g}^{(i)} - \hat{g})\|\|e_i\| + \|\hat{g}\|\|e_i\| + \|(\hat{g}^{(i)} - \hat{g})\|\|\hat{f} - \theta\| + \|\hat{g}\|\|\hat{f}^{(i)} - \hat{f}\| + \|(\hat{g}^{(i)} - \hat{g})\|\|\hat{f}^{(i)} - \hat{f}\|}{\|\hat{g}\|^2} + R1$$

(by the Cauchy-Schwartz and Triangle inequalities)

$$= \frac{O_P(g_n^2 n^{-.5}) + O_P(g_n^2)\|\hat{f} - f(\beta^{**}) + f(\beta^{**}) - \theta\|}{\|\hat{g}\|^2} + \frac{O_P(n^{-.5})}{\|\hat{g}\|} + R1$$

(by Burman and Chaudhuri (1992) results 6.20 and 6.21, and A4.A.3 (with pursuant comments))

$$= \frac{O_P(g_n^2 n^{-.5}) + O_P(g_n^2 d_n)}{\|\hat{g}\|^2} + \frac{O_P(n^{-.5})}{\|\hat{g}\|} + R1 \qquad (4.B.3)$$

by the Triangle inequality, 4.1, and the definition of d_n. With this result for $l^* - \hat{l}^P$ in hand

we may proceed with the following lemma dealing with convergence rates for the PRESS selected mixing parameter to the theoretically optimal mixing parameter. Lemma 4.b.4 is the most important lemma leading up to the estimate convergence theorems in this section. It is analogous to Lemma 4.a.3 in the previous section and its proof is in appendix 4b.

Lemma 4.b.4: Assuming conditions A1-A6…

$$\lambda^* - \hat{\lambda}^P = \begin{cases} O_P(g_n^2) + O_P(n^{-.5}), & \text{if } \lim_{n\to\infty} \delta_n \neq 0 \\ O_P(n^{-.5} g_n^{-1}), & \text{if } \delta_n = 0 \end{cases}$$

The next lemma gives asymptotic converges rates for all of the previous quantities in the instance in which the parametric estimate becomes correct as the sample size increases ($\lim_{n\to\infty} \delta_n = 0$). It is analogous to Lemma 4.a.4 in the previous section and its proof can be found in appendix 4b.

Lemma 4.b.5: Now assume that $\lim_{n\to\infty} \delta_n = 0$. Under assumptions A1-A6…

a) $$\alpha = \begin{cases} O_P(g_n^2 \delta_n), & \text{if } \frac{\delta_n}{g_n} > 1 \\ O_P(g_n^3), & \text{if } \frac{\delta_n}{g_n} < 1 \end{cases}$$

b) $$R1 = \begin{cases} O_P(g_n^2 \delta_n^{-1}), & \text{if } \frac{\delta_n}{g_n} > 1 \\ O_P(\delta_n), & \text{if } \frac{n^{-.5}}{g_n} < \frac{\delta_n}{g_n} < 1 \\ O_P(n^{-.5}), & \text{if } \frac{\delta_n}{g_n} < \frac{n^{-.5}}{g_n} \end{cases}$$

Lemma 4.b.5(cont.):

$$\text{c)} \quad \lambda^* - \hat{\lambda}^P = \begin{cases} O_P(\gamma_n^2 \delta_n^{-1}) + O_P(\delta_n^{-1} n^{-.5}), & \text{if } \dfrac{\delta_n}{\gamma_n} > 1 \\ O_P(\delta_n) + O_P(\gamma_n^{-1} n^{-.5}), & \text{if } \dfrac{n^{-.5}}{\gamma_n} < \dfrac{\delta_n}{\gamma_n} < 1 \\ O_P(\gamma_n^{-1} n^{-.5}), & \text{if } \dfrac{\delta_n}{\gamma_n} < \dfrac{n^{-.5}}{\gamma_n} \end{cases}$$

Before taking on any of the theorems dealing with estimate convergence, we need to do a little algebra similar to that done in Part 3a (particularly the proof of Theorem 3.A.2).

Observe that

$$\left\| \hat{\lambda}^P \hat{g} + \hat{f} - \theta \right\|^2 - \left\| \lambda^* \hat{g} + \hat{f} - \theta \right\|^2$$

$$= \frac{\sum (\hat{\lambda}^P \hat{g} + \hat{f} - \theta)^2}{n} - \frac{\sum (\lambda^* \hat{g} + \hat{f} - \theta)^2}{n}$$

$$= \frac{\sum (t_1 - \theta)^2}{n} - \frac{\sum (t_2 - \theta)^2}{n} \text{ (say)}$$

$$= \frac{\sum (t_1 - t_2)^2}{n} - \frac{\sum (t_1 - t_2) 2 (t_2 - \theta)}{n}$$

(following the proof of Theorem 3.A.2)

$$= \frac{\sum ((\hat{\lambda}^P - \lambda^*) \hat{g})^2}{n} - \frac{\sum (\hat{\lambda}^P - \lambda^*) \hat{g} 2 (\lambda^* \hat{g} + \hat{f} - \theta)}{n}$$

So that

$$\left\| \hat{\lambda}^P \hat{g} + \hat{f} - \theta \right\|^2 \le \left(\left| \hat{\lambda}^P - \lambda^* \right| \left\| \hat{g} \right\| \right)^2 + 2 \left| \hat{\lambda}^P - \lambda^* \right| \left\| \hat{g} \right\| \left\| \lambda^* \hat{g} + \hat{f} - \theta \right\| + \left\| \lambda^* \hat{g} + \hat{f} - \theta \right\|^2$$

Then

$$\left\| \hat{\lambda}^P \hat{g} + \hat{f} - \theta \right\| \le \left| \hat{\lambda}^P - \lambda^* \right| \left\| \hat{g} \right\| + 2 \left(\left| \hat{\lambda}^P - \lambda^* \right| \left\| \hat{g} \right\| \left\| \lambda^* \hat{g} + \hat{f} - \theta \right\| \right)^{.5} + \left\| \lambda^* \hat{g} + \hat{f} - \theta \right\| \quad \text{(4.B.4)}$$

We can now obtain the following two theorems dealing with estimate convergence rates. The proofs of Theorems 4.B.2 and 4.B.4 are found in appendix 4b. The numbering is such that they can be compared with their counterparts in the previous sections. Will the MRR2 estimate using the PRESS selected mixing parameter yield results that are comparable?

Theorem 4.B.2: Assuming conditions A1-A6...

$$\left\| \hat{l}^{P}\hat{g} + \hat{f} - \boldsymbol{q} \right\| = \begin{cases} O_P(g_n), & \text{if } \lim_{n\to\infty} \boldsymbol{d}_n \neq 0 \\ O_P(n^{-.5}), & \text{if } \boldsymbol{d}_n = 0 \end{cases}$$

Theorem 4.B.2 gives us an affirmative response (to the previous question) in the form of a third "Golden Result of Model Robust Regression". This time the result demonstrates the flexibility of the MRR2 procedure to handle a mixing parameter estimate that involves cross validation, and is the first result of this type in MRR. We will later discuss the reasons for this.

We will demonstrate the convergence rates of MRR2 with an example. Suppose a user is estimating a function $\boldsymbol{q}$ by using MRR2 and attempting to model the function parametrically with an OLS quartic regression and nonparametrically by a Local Linear Regression (LLR) using the asymptotically optimal constant bandwidth, h_{ROT}, from p. 111 of Fan and Gijbels (1996). We will once again use the Epanechnikov Kernel in the nonparametric estimate and $\hat{l}^{P}$ for the mixing parameter. From Ruppert and Wand (1994) we have that at any given $\mathbf{x}$ in C, the convergence rate of the LLR estimate is given by

$$\left|\hat{g}(x) - \boldsymbol{q}(x)\right|^2 = O_P(h_{ROT}^4) + O_P(n^{-1}h_{ROT}^{-1})$$

where for LLR,

$$h_{ROT} = o_P(n^{-\frac{1}{5}}).$$

Then

$$\left|\hat{g}(x) - \boldsymbol{q}(x)\right|^2 = O_P(n^{-\frac{4}{5}}).$$

Next, we extend this result to the n dimensional nonparametric vector estimate. For a rigorous presentation of this extension see the proof of Lemma 5.a.1 in appendix 5a. The extension results in

$$g_n^2 = O_P(n^{-\frac{4}{5}}),$$

so that asymptotically the user has an estimate such that

$$\left\| \hat{\boldsymbol{l}}^P \hat{g} + \hat{f} - \boldsymbol{q} \right\| = \begin{cases} O_P(n^{-.4}), & \text{if } \lim_{n\to\infty} \boldsymbol{d}_n \neq 0 \\ O_P(n^{-.5}), & \text{if } \boldsymbol{d}_n = 0 \end{cases}.$$

This MRR2 estimate will converge to the true mean function at a rate no slower than $O_P(n^{-.4})$ if the model is misspecified, and as fast as $O_P(n^{-.5})$ if $\boldsymbol{q}(\mathbf{x})$ is truly a quartic function on C.[].

We present one final theorem in this section for the case where $\lim_{n\to\infty} \boldsymbol{d}_n = 0$. Once more, MRR2 proves to be a capable alternative to MRR1.

Theorem 4.B.4: Assuming conditions A1-A6 hold, and that $\lim_{n\to\infty} \boldsymbol{d}_n = 0$.

$$\left\| \hat{\boldsymbol{l}}^P \hat{g} + \hat{f} - \boldsymbol{q} \right\| = \begin{cases} O_P(g_n), & \text{if } \dfrac{\boldsymbol{d}_n}{g_n} > 1 \\ O_P(\boldsymbol{d}_n), & \text{if } \dfrac{n^{-.5}}{g_n} < \dfrac{\boldsymbol{d}_n}{g_n} < 1 \\ O_P(n^{-.5}), & \text{if } \dfrac{\boldsymbol{d}_n}{g_n} < \dfrac{n^{-.5}}{g_n} \end{cases}$$

Theorem 4.B.4 is comparable to Theorem 3.A.4 even though this theorem deals with the MRR2 estimate using the PRESS selected mixing parameter. Thus, this result is as striking as that of Theorem 4.B.2. We will discuss the reasons for this in the next part of this section.

Comments

In the MRR2 case the mixing parameter $\hat{\lambda}^P$ outperforms its MRR1 counterpart for the most part. In comparing Theorem 3.B.4 to Theorem 4.B.4 it is evident that the MRR2 estimate with $\hat{\lambda}^P$ has the capability of converging more rapidly in either of the last two cases $\left(\frac{n^{-.5}}{g_n} < \frac{\delta_n}{g_n} < 1, or\ \frac{\delta_n}{g_n} < \frac{n^{-.5}}{g_n}\right)$, and is equal in the first $\left(\frac{\delta_n}{g_n} > 1\right)$. In fact, in the same context, its asymptotic performance is equal to that of the MRR2 estimate using the asymptotically optimal mixing parameter $\hat{\lambda}^*$. Observe the results in the cases in which $\lim_{n\to\infty}\delta_n \neq 0$, or $\delta_n = 0$, i.e. compare Theorems 3.B.2 and 4.B.2. The MRR2 estimate with $\hat{\lambda}^P$ equals its MRR1 counterpart estimate in the first instance and betters it (asymptotically) in the second. This is most likely attributable to the robustness of the MRR2 estimate, particularly the limited role of the nonparametric estimate (even if we allow λ to be larger than one). Note that the MRR2 estimate retains all the advantages of the parametric estimate. That is, it is never "slowed down" by the mixing parameter as in the MRR1 case. This is a desirable quality and has been demonstrated mathematically in this section.

In conclusion, our work would indicate that MRR2 is more robust when $\hat{\lambda}^P$ is used to select the mixing parameter. The MRR2 estimate retains all of the positive asymptotic properties of the MRR1 estimate and does not lose those capabilities when using $\hat{\lambda}^P$, the mixing parameter selected using PRESS.

We now turn our attention to the application of MRR (particularly MRR1) to quantal regression.