Chapter 4 Asymptotic Results for MRR2

In this chapter we develop asymptotic results for the MRR2 estimate and will again compare various schemes for selecting the mixing parameter λ. As in the MRR1 case, the mixing parameter plays a vital role in the MRR2 estimate.

The MRR2 estimate was developed by Mays (1995) and is simpler in nature than the MRR1 procedure. Instead of forming a straight mix of parametric and nonparametric fits, MRR2 uses a parametric fit regardless of the situation, then adds in a nonparametric fit as necessary via the mixing parameter λ. Recall that a further important difference between MRR1 and MRR2 is that the nonparametric fit is obtained using the residuals of the parametric fit.

Overview of Conventions

As in Chapter 3, our model is written as

$$Y_i = \theta(\mathbf{x}_i) + \varepsilon_i \text{ , for } i = 1 \text{ to } n.$$

The ε_i's are i.i.d. with $E(\varepsilon_i) = 0$ and $V(\varepsilon_i) = \sigma^2$. We will assume that the $\mathbf{x}_i$'s are fixed uniformly on a compact set C in $\Re^{p+1}$. We will also assume that $\theta(\mathbf{x}_i)$ is continuous on C. The ordered pairs $(\mathbf{x}_i, Y_i)$ form the observations from which two estimates are derived. The term "asymptotic" will have the same meaning as in Chapter 3 and we will utilize the same O_p notation used there. The two estimates used in forming the MRR2 estimate are given by $\hat{f}, \hat{g}$, which once again designate the parametric estimate and the nonparametric estimate respectively. The nonparametric estimate has the same restrictions and assumptions (to include bandwidth assumptions) as in Chapter 3. $\hat{f}$ is based on the original data, and $\hat{g}$ is based on the residual vector $\mathbf{Y} - \hat{\mathbf{f}}$.

After obtaining each, the estimates are combined to form the MRR2 estimate

$$\hat{f}(\mathbf{x}_i) + l\hat{g}(\mathbf{x}_i) = \hat{q}(\mathbf{x}_i)\text{, for } i = 1 \text{ to } n$$

where l is a constant chosen by various formulae.

Recall that for any two functions of $\mathbf{x}_i$, $\mathbf{h}_1$ and $\mathbf{h}_2$, we define the inner product as

$$<\mathbf{h}_1,\mathbf{h}_2> = n^{-1}\sum_{i=1}^{n}\mathbf{h}_1(\mathbf{x}_i)\mathbf{h}_2(\mathbf{x}_i) \ ,$$

and the norm as

$$||\mathbf{h}_1||^2 = <\mathbf{h}_1,\mathbf{h}_1>, \text{ with } ||\mathbf{h}_1|| = (<\mathbf{h}_1,\mathbf{h}_1>)^{.5}.$$

The parametric fit $\hat{f}$ is particularly important in the MRR2 estimate and so we will again denote by δ_n, the distance between the unknown regression function, q, and the parametric family of continuous (on C) regression functions under consideration, where

$$\delta_n = \inf\{\|q - f(\beta)\| : \beta \in \Re^{p+1}\}.$$

And if the infimum is attained at a unique β, we will designate this as β^* and write

$$\delta_n = \|q - f(\beta^*)\|.$$

Once more, restricting the user to a parametric family of continuous (on C) regression functions guarantees that δ_n is always finite. Similarly we again set γ_n as a distance measure for the non-parametric estimate,

$$\gamma_n^{\ 2} = E(\|\hat{g} - (q - f(\beta^{**}))\|^2).$$

Note that this time the distance is taken using the optimal parametric fit. We are not compromising our definition of δ_n. In this case we will simply say that $\beta^{**} = \beta^*$ if δ_n is obtained at a unique β, and that $\beta^{**} \in \{\beta : \|q - f(\beta)\| = \delta_n\}$ otherwise.

Since the nonparametric estimate is based on residuals, one might ask whether or not the rate of convergence of the nonparametric estimate is as fast here as it is in the MRR1 case. In fact, it is.

Even though the nonparametric estimate is dependent upon the convergence of the parametric estimate to its optimal form (given above), the nonparametric rate of convergence is still based upon the sample size. So that even if the nonparametric estimate has a different target function for each sample of size n, it will reach that goal with a degree of accuracy based on the number of observations (as in the MRR1 estimate). We need only concern ourselves with whether or not $\left\| \hat{f} - f(\boldsymbol{b}^{**}) \right\|$ converges faster than $\left\| \hat{g} - (\boldsymbol{q} - f(\boldsymbol{b}^{**})) \right\|$. We are taking g_n to represent a nonparametric convergence rate (the rate of the second term in the previous sentence). So this is in fact, the case since

$$\left\| \hat{f} - f(\boldsymbol{b}^{**}) \right\| = O_P(n^{-.5}) \qquad (4.1)$$

(by Burman and Chaudhuri (1992) equation 6.1), and that by assumption A5 (section 3a), $\lim_{n\to\infty} n^{-.5} \boldsymbol{g}_n^{-1} = 0$. We shall form a very similar group of assumptions for the MRR2 estimate that will also support this result in the next section.

Part 4a Basic MRR2 Asymptotics

We turn our attention to the task of developing the theoretical optimal mixing parameter $\boldsymbol{l}^*$. Note that in the MRR2 estimate $\boldsymbol{l}$ is the coefficient of the nonparametric estimate which is opposite from the MRR1 estimate. So in this case we wish to minimize

$$\| \hat{f}(\mathbf{x}_i) + \boldsymbol{l}\hat{g}(\mathbf{x}_i) - \boldsymbol{q}(\mathbf{x}_i) \|^2 = \frac{\sum_{i=1}^{n} (\hat{f}(\mathbf{x}_i) + \boldsymbol{l}\hat{g}(\mathbf{x}_i) - \boldsymbol{q}(\mathbf{x}_i))^2}{n}$$

with respect to $\boldsymbol{l}$.

Taking the derivative and setting it equal to zero we obtain

$$\frac{d}{d\lambda}\frac{\sum_{i=1}^{n}(\hat{f}(\mathbf{x}_i)+\lambda\hat{g}(\mathbf{x}_i)-\theta(\mathbf{x}_i))^2}{n}$$

$$=\frac{\sum_{i=1}^{n}2(\hat{f}(\mathbf{x}_i)+\lambda\hat{g}(\mathbf{x}_i)-\theta(\mathbf{x}_i))\hat{g}(\mathbf{x}_i)}{n}$$

$$=\frac{\sum_{i=1}^{n}2(\hat{f}(\mathbf{x}_i)-\theta(\mathbf{x}_i))\hat{g}(\mathbf{x}_i)}{n}+\frac{\sum_{i=1}^{n}2\lambda\hat{g}^2(\mathbf{x}_i)}{n}=0.$$

Then

$$\frac{\sum_{i=1}^{n}(\theta(\mathbf{x}_i)-\hat{f}(\mathbf{x}_i))\hat{g}(\mathbf{x}_i)}{n}=\frac{\sum_{i=1}^{n}\lambda\hat{g}^2(\mathbf{x}_i)}{n}$$

and our optimal λ candidate (which we shall call λ^*) is given by

$$\lambda^*=\frac{<\hat{g},\theta-\hat{f}>}{\|\hat{g}\|^2} \qquad (4.A.1)$$

Note that the second derivative is given by

$$\frac{d^2}{d\lambda^2}\frac{\sum_{i=1}^{n}(\hat{f}(\mathbf{x}_i)+\lambda\hat{g}(\mathbf{x}_i)-\theta(\mathbf{x}_i))^2}{n}$$

$$=\frac{\sum_{i=1}^{n}2\hat{g}^2(\mathbf{x}_i)}{n}=2\|\hat{g}\|^2>0 \text{ for } \lambda\in\Re,$$

except for the degenerate case which we will not worry about here. So λ^* is optimal for minimizing $\|\hat{f}(\mathbf{x}_i)+\lambda\hat{g}(\mathbf{x}_i)-\theta(\mathbf{x}_i)\|^2$.

We will use the same assumptions given in Part 3a, except for the following changes.

A2. $$\frac{\|\hat{g}-(\theta-f(\beta^{**}))\|^2-E(\|\hat{g}-(\theta-f(\beta^{**}))\|^2)}{E(\|\hat{g}-(\theta-f(\beta^{**}))\|^2)}\Rightarrow^{P}0, \text{ as n -> } \infty.$$

A6. $$\hat{g}^{(i)}(x_i)=\frac{\sum_{j:j\neq i}W_2(x_i,x_j)e_j}{\sum_{j:j\neq i}W_2(x_i,x_j)} \text{ where } e_j=Y_j-\hat{f}(x_j).$$

Again the implications are as follows. A2 guarantees that $\left\|\hat{g} - (\boldsymbol{q} - f(\boldsymbol{b}^{**}))\right\| = O_P(\gamma_n)$. A5 again implies that the nonparametric estimate converges at a rate slower than $n^{-.5}$. And A6 is satisfied for local polynomial and spline smoothing estimates.

We present the following technical results and theorems which closely parallel the results found in section 3a. The first of these deals with the denominator in the difference term for the mixing parameters. Its proof is found in appendix 4a.

Lemma 4.a.1: Assuming conditions A1, A2, and A5…

$$\left\|\hat{g}\right\| = \begin{cases} O_P(1), & \text{if } \lim_{n\to\infty} \delta_n \neq 0 \\ O_P(\gamma_n), & \text{if } \delta_n = 0 \end{cases}.$$

The preceding lemma also gives information on the performance of the nonparametric estimate. As expected, the estimate will converge to 0 if the user's model is correct. Lemma 4.a.2 deals with the activity of the theoretical asymptotically optimal mixing parameter. Again, the performance is expected. The mixing parameter converges to 0 if the user's model is correct, and to 1 otherwise. The proof of Lemma 4.a.2 is found in appendix 4a.

Lemma 4.a.2: Assuming conditions A1, A2, and A5…

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

Before proceeding further we need to obtain an optimal data driven mixing parameter. Similar to section 3a, we suggest

$$\hat{\lambda}^* = \frac{< \hat{g}, Y - \hat{f} >}{\left\|\hat{g}\right\|^2} = \frac{\sum_{i=1}^{n} (\hat{g}(\mathbf{x}_i)(Y_i - \hat{f}(\mathbf{x}_i)))}{\sum_{i=1}^{n} (\hat{g}(\mathbf{x}_i))^2} \qquad (4.A.2)$$

Note the geometric interpretation. When $\hat{\lambda}^* = 0$, we have that in n-space the vectors $\hat{\boldsymbol{g}}$ and $\boldsymbol{Y} - \hat{\boldsymbol{f}}$ are orthogonal, meaning that $\hat{\boldsymbol{g}}$ does not provide a particularly good fit to the residual vector. Also observe that if we plot e_i against $\hat{g}(\mathbf{x}_i)$, $\hat{\lambda}^*$ is the slope of the linear relationship in the no intercept model. So that if there is basically no predictive relationship between e_i and $\hat{g}(\mathbf{x}_i)$, then $\hat{\lambda}^*$ is close to zero. This is precisely what we want to occur.

Lemma 4.a.3 provides a crucial part of the estimate convergence theorems that follow. It gives the convergence rates for the mixing parameters which, as illustrated in section 3a, drive the estimate convergence rates that follow. The proof is found in appendix 4a.

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

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

Notice that $\hat{\lambda}^*$ outperforms its MRR1 counterpart, $\hat{\lambda}^{*C}$ when the user's model is incorrect. Although this is a positive result, unfortunately we shall see that for this section the estimate convergence rates will not be affected. However, it will turn out that the MRR2 procedure does have some improved asymptotic properties over that of the MRR1 estimate. This will become evident in the next section. For now, we will continue presenting lemmas dealing with the asymptotic performances of parts of the MRR2 estimate.

Once again, the next two lemmas address the situation in which the parametric estimate becomes correct ($\lim_{n\to\infty} \delta_n = 0$) as the sample size increases. Note that for the most part all convergence rates are similar to those in the analogous lemma in section 3a. Lemma 4.a.4

gives asymptotic results for the expressions found in the previous three lemmas for the case in which $\lim_{n\to\infty} \delta_n = 0$. Lemma 4.a.5 also addresses this case for another term and is necessary in the proof of Theorems 4.A.2 and 4.A.4. The proofs for both lemmas are found in appendix 4a.

Lemma 4.a.4: Assuming that $\lim_{n\to\infty} \delta_n = 0$…

a) Under assumptions A1, A2, and A5 we have

$$\|\hat{g}\| = \begin{cases} O_P(\delta_n), \text{ if } \dfrac{\delta_n}{\gamma_n} > 1 \\ O_P(\gamma_n), \text{ if } \dfrac{\delta_n}{\gamma_n} < 1 \end{cases},$$

b) Under assumptions A1, A2, and A5

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

c) Assuming conditions A1-A6 …

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

Lemma 4.a.5: Assuming conditions A1-A6…

$$\left|\hat{\lambda}^* - \lambda^*\right| \|\hat{g}\| = O_P(n^{-.5}) \text{ whether } \delta_n = 0, \lim_{n\to\infty} \delta_n = 0, \text{ or } \lim_{n\to\infty} \delta_n \neq 0.$$

As in section 3a, the first three lemmas give results conditioned on the dichotomy: the parametric estimate is correct versus the parametric estimate is incorrect. Lemma 4.a.4 simply addresses the situation in which the parametric estimate may not be technically correct, but becomes correct as n -> ∞. All of the lemmas will now be utilized in proving the following important theorems dealing with estimate convergence.

Theorem 4.A.1: Assuming conditions A1, A2 and A5 hold...

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

Theorem 4.A.1 tells us that the theoretical asymptotically optimal mixing parameter can produce asymptotic rates of convergence no worse than the nonparametric rate. The next theorem reveals that a data driven mixing parameter will produce the same result. The proofs for both theorems are found in appendix 4a.

Theorem 4.A.2: Assume that conditions A1-A6 hold…

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

If Theorem 3.A.2 is the "Golden Result of Model Robust Regression" then Theorem 4.A.2 is another form of the same result, and, no doubt, just as valuable. Once more, we obtain the best asymptotic convergence rate available regardless of whether the model is correct or incorrect.

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 quadratic regression and nonparametrically by a Local Quadratic Regression (LQR)

using the asymptotically optimal constant bandwidth, h_{ROT}, from p. 111 of Fan and Gijbels (1996). This bandwidth selector is an estimate of the h_{OPT} selector seen earlier, and can be used under "certain conditions" which we will assume are met here since a) the authors indicate that the selector is robust to these conditions for initial bandwidth selection, b) the conditions are not specified and do not appear to be difficult to meet, and c) do not appear to affect the rate of convergence.

We will also use the Epanechnikov Kernel in the nonparametric estimate and $\hat{l}^*$ for the mixing parameter. From Ruppert and Wand (1994) we have that at any given x in C, the convergence rate of the LQR estimate is given by

$$\left|\hat{g}(x) - \boldsymbol{q}(x)\right|^2 = O_P(h_{ROT}^8) + O_P(n^{-1}h_{ROT}^{-1})$$

where

$$h_{ROT} = o_P(n^{-\frac{1}{7}}),$$

so that

$$\left|\hat{g}(x) - \boldsymbol{q}(x)\right|^2 = O_P(n^{-\frac{6}{7}}).$$

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

$$g_n^2 = O_P(n^{-\frac{6}{7}}).$$

Then asymptotically the user has an estimate such that

$$\left\| \hat{l} * \hat{g} + \hat{f} - \boldsymbol{q} \right\| = \begin{cases} O_P(n^{-\frac{3}{7}}), & \text{if } \lim_{n\to\infty} \boldsymbol{d}_n \neq 0 \\ O_P(n^{-.5}), & \text{if } \boldsymbol{d}_n = 0 \end{cases},$$

and will thus converge to the true mean function at a rate no slower than $O_P(n^{-\frac{3}{7}})$ if the model is misspecified (this faster rate of convergence is due to the slow convergence of the LQR bandwidth), and as fast as $O_P(n^{-.5})$ if $\boldsymbol{q}(\mathbf{x})$ is truly a quadratic function on C.[].

The next two theorems give estimate convergence rates in the case where $\lim_{n\to\infty} \boldsymbol{d}_n = 0$. The proofs for both theorems are found in appendix 4a. Notice that the results are similar to those given in Theorems 3.A.3 and 3.A.4.

Theorem 4.A.3: Assume that conditions A1-A6 hold, and that $\lim_{n\to\infty} \boldsymbol{d}_n = 0$.

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

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

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

Once again, Theorems 4.A.1 and 4.A.2 give the dichotomous convergence rates of the MRR2 estimate to the true mean function q, with the theoretical asymptotically optimal mixing parameter and the asymptotically optimal data driven mixing parameter respectively. Theorems 4.A.3 and 4.A.4 give the MRR2 convergence rates to q, in the cases stemming from the situation in which $\lim_{n\to\infty} d_n = 0$, with the theoretical asymptotically optimal mixing parameter and the asymptotically optimal data driven mixing parameter respectively.

Comments

We have demonstrated in this section that the MRR2 estimate possesses the same asymptotic properties as those of the MRR1 estimate. This compares positively with the strengths of the MRR1 procedure. Once again we achieve the optimal asymptotic convergence rate whether or not the user's model is correct. This section also will serve as a standard for asymptotic results for the general MRR2 estimate. In the next section we will consider asymptotic properties of MRR2 using PRESS as the mixing parameter selector. We shall see that the MRR2 estimate is more robust to the problems inherent in the mixing parameter chosen by PRESS than the MRR1 estimate with the analogous mixing parameter.