RECURSIVELY RE-WEIGHTED LEAST-SQUARES ESTIMATION IN REGRESSION MODELS WITH PARAMETERIZED VARIANCE

Luc Pronzato and Andrej Pázman

Laboratoire I3S, UNSA-CNRS
Bât. Euclide, Les Algorithmes
2000 route des Lucioles, BP 121
06903 Sophia Antipolis Cedex, France
email: pronzato@i3s.unice.fr

and

Department of Probability and Statistics
Comenius University, Mlynská Dolina
84215 Bratislava, Slovakia
email: pazman@center.fmph.uniba.sk

ABSTRACT

We consider a nonlinear regression model with parameterized variance and compare several methods of estimation: the Weighted Least-Squares (WLS) estimator; the two-stage LS (TSLS) estimator, where the LS estimator obtained at the first stage is plugged into the variance function used for WLS estimation at the second stage; and finally the recursively re-weighted LS (RWLS) estimator, where the LS estimator obtained after k observations is plugged into the variance function to compute the k-th weight for WLS estimation. We draw special attention to RWLS estimation which can be implemented *recursively* when the regression model in linear (even if the variance function is nonlinear), and is thus *particularly attractive for signal processing applications*.

1. INTRODUCTION

We consider a nonlinear regression problem, with observations

$$Y_k = y(x_k) = \eta(x_k,\bar{\theta}) + \varepsilon_k\,,\ \mathtt{E}_{x_k}\{\varepsilon_k\} = 0\,,\ \ k = 1,\ldots,N\,, \quad (1)$$

where $\bar{\theta}$ denotes the unknown true value of the model parameters. The observation errors $\varepsilon_k = \varepsilon(x_k)$ are assumed to be independently distributed. It frequently happens that the full parameterized probability distribution of the errors ε_k is not available, whereas their variance is a known function of the design variable $x \in \mathtt{X} \subset I\!R^d$ and of (some of) the parameters θ of the mean response, that is,

$$\sigma^2(x_k) = \mathtt{E}_{x_k}\{\varepsilon_k^2\} = c\,\lambda(x_k,\bar{\theta})\,,\ \ k = 1,\ldots,N\,, \quad (2)$$

with c some positive constant. The (ordinary) LS estimator is strongly consistent and asymptotically normally distributed under standard assumptions. However, it ignores the information contained in the variance function. Using this information may yield a more precise estimation, hence the importance of choosing a suitable estimation method. The asymptotic properties of the WLS estimator are recalled in Section 3: in particular, the asymptotic variance of the estimator is minimum when the weights are proportional to the inverse of the variance of the observation errors. Since these optimum weights are unknown ($\bar{\theta}$ is unknown in (2)), we consider the two-stage least-squares (TSLS) method, where the WLS estimator obtained at the first stage is plugged in the variance function used for WLS estimation at the second stage. A third approach, the recursively re-weighted LS (RWLS) estimation, is considered in Section 4: the LS estimator obtained from k observations is used to compute the k-th weight (and only this one) to be used for WLS estimation. When the model $\eta(x,\theta)$ is linear is θ (the variance function $\lambda(x,\theta)$ may be nonlinear), the estimator can be implemented *recursively*, by combining two recursive LS methods. In applications with large data sets, or when an on-line implementation is required, this is a definite advantage over the TSLS approach.

The asymptotic properties of the estimators are obtained under the assumption of a *randomized design*, introduced in Section 2, which allows rigorous proofs for the asymptotic properties of the estimators while avoiding the technical difficulties encountered in classical references such as [2] (finite tail product of the regression function and its derivatives, etc.). We show that the asymptotic performances of the TSLS and RWLS estimators are similar, and coincide with those of the WLS estimator with optimum weights. The main lines of the proofs are given in Section 6. Some simulations results are presented in Section 5.

2. RANDOMIZED DESIGNS

In order to study the asymptotic properties of estimators we need to specify how the sequence of points $x_1, x_2, \ldots$ is generated. The following definition is well adapted to situations where the sequence is not completely under control.

Definition 1 *We call* randomized design *with measure ξ on the design space $\mathtt{X} \subset I\!R^d$, $\int_{\mathtt{X}} \xi(dx) = 1$, a sequence $\{x_i\}$ of points independently sampled from the mesure ξ on* X.

The following assumptions will be used throughout the paper.

H1 Θ is a compact subset of $I\!R^p$ such that $\Theta \subset \overline{\mathrm{int}(\Theta)}$.
H2 $\eta(x,\theta)$ and $\lambda(x,\theta)$ are continuous functions of $\theta \in \Theta$ for any $x \in \mathtt{X}$, with $\eta(x,\theta)$ and $\lambda^{-1}(x,\theta)$ bounded on $\mathtt{X} \times \Theta$ and $\lambda(x,\bar{\theta})$ bounded on X.
H3 $\eta(x,\theta)$ and $\lambda(x,\theta)$ are two times continuously differentiable with respect to $\theta \in \mathrm{int}(\Theta)$ for any $x \in \mathtt{X}$, these first two derivatives are bounded on $\mathtt{X} \times \mathrm{int}(\Theta)$.

Our proofs are based on the uniform convergence with respect to θ of the criterion function $J_N(\theta)$ defining the estimator $\hat{\theta}^N = \arg\min_\theta J_N(\theta)$. We shall thus need a uniform Strong Law of Large Numbers (SLLN). Note that the proper definition of the estimator as a random variable is ensured by Lemma 2 in [2] (see also [1], p. 16). In the following $\hat{\theta}^N$

This work was partly supported by the Network of Excellence "Pascal": Pattern Analysis, Statistical Modelling and Computational Learning, nb. 2002–506778

will refer to the measurable choice from $\arg\min_{\theta\in\Theta} J_N(\theta)$. The asymptotic results of the next sections are based on the following lemma, which is derived from Theorem 2.7.1 in [1].

Lemma 1 (Uniform SLLN) *Let $\{z_i\}$ be a sequence of i.i.d. random vectors of $\mathbb{R}^r$, and $a(z,\theta)$ be a Borel measurable real function of $(z,\theta)\in\mathbb{R}^r\times\Theta$, continuous in θ for any z, with Θ a compact subset of $\mathbb{R}^p$. Suppose that*

$$\mathrm{E}[\max_{\theta\in\Theta}|a(z,\theta)|]<\infty, \tag{3}$$

then $\mathrm{E}[a(z,\theta)]$ is continuous in $\theta\in\Theta$ and $\frac{1}{N}\sum_{i=1}^N a(z_i,\theta)\overset{\theta}{\rightrightarrows}\mathrm{E}[a(z,\theta)]$ a.s. when $N\to\infty$, where $\overset{\theta}{\rightrightarrows}$ means uniform convergence with respect to θ.

Once the almost sure uniform convergence of the criterion function $J_N(\cdot)$ is obtained, the almost sure convergence of the estimator will follow from the next lemma. The proof is a straightforward application of the continuity and uniform convergence properties.

Lemma 2 *Assume that the sequence of functions $\{J_N(\theta)\}$ converges uniformly on Θ to the function $J(\theta)$, with $J_N(\theta)$ continuous with respect to $\theta\in\Theta$ for any N, Θ a compact set of $\mathbb{R}^p$, and $J(\theta)$ such that for some $\bar\theta\in\Theta$,*

$$\forall\theta\in\Theta,\ \theta\neq\bar\theta,\ J(\theta)>J(\bar\theta).$$

Then $\lim_{N\to\infty}\hat\theta^N=\bar\theta$, where $\hat\theta^N\in\arg\min_{\theta\in\Theta}J_N(\theta)$. When the functions $J_N(\cdot)$ are random, and the uniform convergence to $J(\cdot)$ is almost sure, the convergence of $\hat\theta^N$ to $\bar\theta$ is also almost sure.

3. WEIGHTED LS AND TWO-STAGE LS

3.1 Weighted LS estimation

The WLS estimator $\hat\theta^N_{WLS}$ minimizes

$$J_N(\theta)=\frac{1}{N}\sum_{k=1}^N w(x_k)[y(x_k)-\eta(x_k,\theta)]^2 \tag{4}$$

with $w(x)\geq 0$ and bounded on $\mathtt{X}$. The following theorem is a standard property of LS estimation.

Theorem 1 *Let $\{x_i\}$ be a randomized design with measure ξ on $\mathtt{X}\subset\mathbb{R}^d$. Assume that H1 and H2 are satisfied and that $\forall\theta,\theta'\in\Theta$,*

$$\int_{\mathtt{X}} w(x)[\eta(x,\theta)-\eta(x,\theta')]^2\xi(dx)=0\Leftrightarrow\theta=\theta'. \tag{5}$$

Then the estimator $\hat\theta^N_{WLS}$ that minimises (4) in the model (1,2) converges a.s. to $\bar\theta$. If, moreover, H3 is satisfied, $\bar\theta\in\mathrm{int}(\Theta)$ and the matrix

$$\mathbf{M}_1(\xi,\bar\theta)=\int_{\mathtt{X}} w(x)\frac{\partial\eta(x,\theta)}{\partial\theta}_{|\bar\theta}\frac{\partial\eta(x,\theta)}{\partial\theta^\top}_{|\bar\theta}\xi(dx) \tag{6}$$

is nonsingular, then $\hat\theta^N_{WLS}$ satisfies

$$\sqrt{N}(\hat\theta^N_{WLS}-\bar\theta)\overset{\mathrm{d}}{\to}z\sim\mathtt{N}(0,\mathbf{C}(w,\xi,\bar\theta))$$

as $N\to\infty$, where $\mathbf{C}(w,\xi,\bar\theta)=\mathbf{M}_1^{-1}(\xi,\bar\theta)\ \mathbf{M}_2(\xi,\bar\theta)\ \mathbf{M}_1^{-1}(\xi,\bar\theta)$ with

$$\mathbf{M}_2(\xi,\bar\theta)=\int_{\mathtt{X}} w^2(x)\sigma^2(x)\frac{\partial\eta(x,\theta)}{\partial\theta}_{|\bar\theta}\frac{\partial\eta(x,\theta)}{\partial\theta^\top}_{|\bar\theta}\xi(dx). \tag{7}$$

Moreover, $\mathbf{C}(w,\xi,\bar\theta)-\mathbf{M}^{-1}(\xi,\bar\theta)$ is non-negative definite for any choice of $w(x)$, with

$$\mathbf{M}(\xi,\bar\theta)=\int_{\mathtt{X}}\ \sigma^{-2}(x)\frac{\partial\eta(x,\theta)}{\partial\theta}_{|\bar\theta}\frac{\partial\eta(x,\theta)}{\partial\theta^\top}_{|\bar\theta}\xi(dx), \tag{8}$$

and $\mathbf{C}(w,\xi,\bar\theta)=\mathbf{M}^{-1}(\xi,\bar\theta)$ for $w(x)=\alpha\sigma^{-2}(x)$ with α any positive constant.

3.2 Two-stage LS estimation

The optimum weights $w(x)=\lambda^{-1}(x,\bar\theta)$ cannot be used for WLS estimation since $\bar\theta$ is unknown. It is therefore tempting to use the weights $\lambda^{-1}(x,\theta)$, that is, to choose $\hat\theta^N$ that minimises the criterion

$$J_N(\theta)=\frac{1}{N}\sum_{k=1}^N\frac{[y(x_k)-\eta(x_k,\theta)]^2}{\lambda(x_k,\theta)}. \tag{9}$$

However, this approach *is not recommended* since $\hat\theta^N$ is generally not even consistent.

Theorem 2 *Let $\{x_i\}$ be a randomized design with measure ξ on $\mathtt{X}\subset\mathbb{R}^d$. Assume that H1 and H2 are satisfied. Then the estimator $\hat\theta^N_{LS}$ that minimises (9) in the model (1,2) converges a.s. to the set $\bar\Theta$ of values of θ that minimise*

$$\begin{aligned}J(\theta)\ &=\ c\int_{\mathtt{X}}\lambda(x,\bar\theta)\lambda(x,\theta)^{-1}\xi(dx)\\ &\quad+\int_{\mathtt{X}}\lambda(x,\theta)^{-1}[\eta(x,\theta)-\eta(x,\bar\theta)]^2\xi(dx).\end{aligned}$$

Notice that, in general, $\bar\theta\notin\bar\Theta$.

Consider now a two-stage approach, where some estimator $\hat\theta^N_1$ is constructed at the first stage, and then plugged into the weight function $\lambda(x,\theta)$. The second-stage estimator $\hat\theta^N_{TSLS}$ is then obtained by minimizing

$$J_N(\theta,\hat\theta^N_1)=\frac{1}{N}\sum_{k=1}^N\frac{[y(x_k)-\eta(x_k,\theta)]^2}{\lambda(x_k,\hat\theta^N_1)} \tag{10}$$

with respect to $\theta\in\Theta$. We have the following.

Theorem 3 *Let $\{x_i\}$ be a randomized design with measure ξ on $\mathtt{X}\subset\mathbb{R}^d$. Assume that H1 and H2 are satisfied, that $\hat\theta^N_1$ converges a.s. to some $\bar\theta_1\in\Theta$ and that for any $\theta,\theta'\in\Theta$,*

$$\int_{\mathtt{X}}\lambda^{-1}(x,\bar\theta_1)[\eta(x,\theta)-\eta(x,\theta')]^2\xi(dx)=0\Leftrightarrow\theta=\theta'.$$

Then the estimator $\hat\theta^N_{TSLS}$ that minimises (10) in the model (1,2) converges a.s. to $\bar\theta$. If, moreover, H3 is satisfied, the matrix $\mathbf{M}(\xi,\bar\theta)$ given by (8) is nonsingular and the first-stage estimator $\hat\theta^N_1$ plugged in (10) is $\sqrt{N}$-consistent[1], with $\bar\theta\in\mathrm{int}(\Theta)$, then $\hat\theta^N_{TSLS}$ satisfies

$$\sqrt{N}(\hat\theta^N_{TSLS}-\bar\theta)\overset{\mathrm{d}}{\to}z\sim\mathtt{N}(0,\mathbf{M}^{-1}(\xi,\bar\theta)),\ N\to\infty.$$

[1] that is, when $\sqrt{N}(\hat\theta^N_1-\bar\theta)$ is bounded in probability: $\forall\varepsilon>0\ \exists A$ and N_0 such that $\forall N>N_0$, $\mathrm{Prob}\{\sqrt{N}(\hat\theta^N_1-\bar\theta)>A\}<\varepsilon$

Notice that $\mathbf{M}^{-1}(\xi,\bar{\theta})$ is the asymptotic covariance matrix of the WLS estimator in the ideal case where *the variance function (2) is known*, see Theorem 1.

Also note that a natural candidate for the first-stage estimator $\hat{\theta}_1^N$ is the WLS estimator $\hat{\theta}_{WLS}^N$ that minimises $J_N(\theta)$ given by (4) with *arbitrary weights*: under the assumptions of Theorem 1 $\hat{\theta}_{WLS}^N$ is $\sqrt{N}$-consistent since $\sqrt{N}(\hat{\theta}_{WLS}^N-\bar{\theta}) \xrightarrow{\mathrm{d}} z \sim \mathtt{N}(0,\mathbf{C}(w,\xi,\bar{\theta}))$. In particular, one may choose $w(x)=1$ for any x, which corresponds to the ordinary LS estimator $\hat{\theta}_{LS}^N$.

Increasing the number of stages leads to *iteratively re-weighted LS estimation*, which relies on sequence of estimators constructed as follows:

$$\hat{\theta}_k^N = \arg\min_{\theta\in\Theta} J_N(\theta,\hat{\theta}_{k-1}^N)\,,\ k=2,3\ldots \tag{11}$$

where $J_N(\theta,\theta')$ is defined by (10) and where $\hat{\theta}_1^N$ can be taken equal to $\hat{\theta}_{LS}^N$. A simple induction shows that, for any *fixed* k, $\hat{\theta}_k^N$ has the same asymptotic properties as $\hat{\theta}_{TSLS}^N$. Although there is apparently no gain in pushing the recursion (11) to its limit rather than simply using $\hat{\theta}_{TSLS}^N$, these are only asymptotic results and the finite sample behaviors of both methods may differ.

4. RECURSIVELY RE-WEIGHTED LS

We define the recursively re-weighted LS estimator $\hat{\theta}_{RWLS}^N$ as the value of $\theta\in\Theta$ that minimises the criterion

$$J_N(\theta) = \frac{1}{N}\sum_{k=1}^{N} \frac{[y(x_k)-\eta(x_k,\theta)]^2}{\lambda(x_k,\hat{\theta}_{WLS}^k)}\,, \tag{12}$$

where the auxiliary estimate $\hat{\theta}_{WLS}^k$ uses arbitrary weights $w(x)$ and is based on the *first k observations $Y_1,\ldots,Y_k$ and design points $x_1,\ldots,x_k$ only*. Using Lemmas 1 and 2 we can show the following.

Theorem 4 *Let $\{x_i\}$ be a randomized design with measure ξ on $\mathtt{X}$ a compact set of $\mathbb{R}^d$. Assume that H1 and H2 are satisfied, that $\lambda(x,\theta)$ is continuous on $\mathtt{X}\times\Theta$ with $\mathtt{X}$ compact, that for any $\theta,\theta'\in\Theta$,*

$$\int_{\mathtt{X}} \lambda^{-1}(x,\bar{\theta})[\eta(x,\theta)-\eta(x,\theta')]^2\,\xi(dx) = 0 \Leftrightarrow \theta=\theta'$$

and that $w(x)$ is such that (5) is satisfied. Then the estimator $\hat{\theta}_{RWLS}^N$ that minimises (12) in the model (1,2) converges a.s. to $\bar{\theta}$. If, moreover, H3 is satisfied, the matrix $\mathbf{M}(\xi,\bar{\theta})$ given by (8) is nonsingular and $\bar{\theta}\in\mathrm{int}(\Theta)$, then $\hat{\theta}_{RWLS}^N$ satisfies

$$\sqrt{N}(\hat{\theta}_{RWLS}^N-\bar{\theta}) \xrightarrow{\mathrm{d}} z\sim\mathtt{N}(0,\mathbf{M}^{-1}(\xi,\bar{\theta}))\,,\ N\to\infty.$$

The two estimators $\hat{\theta}_{TSLS}^N$ and $\hat{\theta}_{RWLS}^N$ have therefore the same asymptotic performance (in terms of covariance matrix) as the WLS estimator with optimum weights (note that it does not imply that their finite sample behaviors are similar). This makes $\hat{\theta}_{RWLS}^N$ particularly attractive when $\eta(x,\theta)$ is linear in θ, that is, when

$$\eta(x,\theta) = f^\top(x)\theta\,. \tag{13}$$

Indeed, the auxiliary WLS estimator $\hat{\theta}_{WLS}^k$ can be constructed recursively through

$$\begin{aligned}
\mathbf{P}_{k+1} &= \mathbf{P}_k - \frac{\mathbf{P}_k f(x_{k+1}) f^\top(x_{k+1})\mathbf{P}_k}{w^{-1}(x_{k+1})+f^\top(x_{k+1})\mathbf{P}_k f(x_{k+1})}\,,\\
\hat{\theta}_{WLS}^{k+1} &= \hat{\theta}_{LS}^k + \frac{\mathbf{P}_k f(x_{k+1})}{w^{-1}(x_{k+1})+f^\top(x_{k+1})\mathbf{P}_k f(x_{k+1})}\\
&\quad\times[y(x_{k+1})-f^\top(x_{k+1}\hat{\theta}_{WLS}^k]\,.
\end{aligned}$$

Let k_0 be the first integer such that $f(x_1),\ldots,f(x_{k_0})$ span $\mathbb{R}^p$. The recursion can be initialized at $k=k_0$ by

$$\mathbf{P}_{k_0} = \left[\sum_{i=1}^{k_0} w(x_i) f(x_i) f^\top(x_i)\right]^{-1}.$$

A similar recursion can be used to compute $\hat{\theta}_{RWLS}^k$ simultaneously,

$$\begin{aligned}
\mathbf{P}'_{k+1} &= \mathbf{P}'_k - \frac{\mathbf{P}'_k f(x_{k+1}) f^\top(x_{k+1})\mathbf{P}'_k}{\lambda(x_{k+1},\hat{\theta}_{WLS}^{k+1})+f^\top(x_{k+1})\mathbf{P}'_k f(x_{k+1})}\,,\\
\hat{\theta}_{RWLS}^{k+1} &= \hat{\theta}_{RWLS}^k + \frac{\mathbf{P}'_k f(x_{k+1})}{\lambda(x_{k+1},\hat{\theta}_{WLS}^{k+1})+f^\top(x_{k+1})\mathbf{P}'_k f(x_{k+1})}\\
&\quad\times[y(x_{k+1})-f^\top(x_{k+1}\hat{\theta}_{RWLS}^k]\,,
\end{aligned}$$

with the initialisation $\mathbf{P}'_{k_0}=\mathbf{P}_{k_0}$ and $\hat{\theta}_{RWLS}^{k_0}=\hat{\theta}_{WLS}^{k_0}$. Notice that $\hat{\theta}_{WLS}^k$ is linear with respect to the observations $Y_1,\ldots,Y_k$ but $\hat{\theta}_{RWLS}^k$ *is not*.

5. EXAMPLE

We take $f(x)=(1\ x\ x^2)^\top$ in (13) and $\lambda(x,\theta)=|f^\top(x)\theta|$ in (2). We suppose that the true (unknown) value for θ is $\bar{\theta}=(0\ 0\ 1)^\top$, so that the variance of the observation error at x is $\sigma^2(x)=c\,\lambda(x,\bar{\theta})=cx^2$. This gives the value

$$\mathbf{M}(\xi,\bar{\theta}) = \frac{1}{c}\begin{pmatrix} \mu_{-2} & \mu_{-1} & 1\\ \mu_{-1} & 1 & \mu_1\\ 1 & \mu_1 & \mu_2\end{pmatrix}$$

for the matrix (8), with $\mu_p=\int_{\mathtt{X}} x^p\xi(dx)$. For TSLS and RWLS estimation we use the ordinary LS ($w(x)\equiv 1$) as auxiliary estimator. Similar calculation can be done for the matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ involved in the asymptotic covariance matrix of $\hat{\theta}_{LS}^N$, see Theorem 1. When ξ is such that x is uniformly distributed in the interval $[0.1,1.1]$ we obtain the asymptotic covariance matrices

$$\mathbf{M}^{-1}(\xi,\bar{\theta}) = c\begin{pmatrix} 0.9698 & -5.3891 & 5.1059\\ -5.3891 & 35.2660 & -35.5726\\ 5.1059 & -35.5726 & 38.8819\end{pmatrix}$$

for $\hat{\theta}_{TSLS}^N$ and $\hat{\theta}_{RWLS}^N$ and

$$\mathbf{C}(\xi,\bar{\theta}) = c\begin{pmatrix} 3.7434 & -18.6593 & 16.8094\\ -18.6593 & 98.8149 & -91.6457\\ 16.8094 & -91.6457 & 88.3714\end{pmatrix}$$

for the ordinary LS estimator $\hat{\theta}_{LS}^N$.

We take $c = 0.01$ and repeat 2,000 experiments with $N = 100$ observations each, the errors ε_k are independently and normally distributed. The empirical mean-squared error (MSE) matrix obtained for $\hat{\theta}_{LS}^N$ is

$$\begin{pmatrix} 0.0396 & -0.1962 & 0.1775 \\ -0.1962 & 1.0343 & -0.9645 \\ 0.1775 & -0.9645 & 0.9358 \end{pmatrix}$$

which is close to $\mathbf{C}(\xi, \bar{\theta})$. For the weighted LS estimator using the *true weight function*, we obtain the empirical MSE matrix

$$\begin{pmatrix} 0.0107 & -0.0589 & 0.0563 \\ -0.0589 & 0.3798 & -0.3849 \\ 0.0563 & -0.3849 & 0.4217 \end{pmatrix},$$

close to $\mathbf{M}^{-1}(\xi, \bar{\theta})$. Finally, we obtain the matrices

$$\begin{pmatrix} 0.0131 & -0.0694 & 0.0653 \\ -0.0694 & 0.4277 & -0.4269 \\ 0.0653 & -0.4269 & 0.4587 \end{pmatrix}$$

and

$$\begin{pmatrix} 0.0158 & -0.0795 & 0.0728 \\ -0.0795 & 0.4813 & -0.4729 \\ 0.0728 & -0.4729 & 0.5002 \end{pmatrix}$$

for the estimators $\hat{\theta}_{TSLS}^N$ and $\hat{\theta}_{RWLS}^N$ respectively, showing that (i) the decrease of performance due to the estimation of the weight function in $\hat{\theta}_{TSLS}^N$ is almost negligible and (ii) the additional decrease of performance due to the *recursive* estimation of the weight function in $\hat{\theta}_{RWLS}^N$ is almost negligible too. On the other hand, the gain in precision compared to ordinary LS is quite significative.

6. INDICATIONS OF PROOFS

Consistency. In Theorem 1 we simply apply Lemma 1 to show that $J_N(\theta) \overset{\theta}{\to} J(\theta) = \int_{\mathbb{X}} w(x)\sigma^2(x)\xi(dx) + \int_{\mathbb{X}} w(x)[\eta(x,\theta) - \eta(x,\bar{\theta})]^2\xi(dx)$ a.s. as $N \to \infty$ and then Lemma 2 to get $\hat{\theta}_{WLS}^N \overset{\text{a.s.}}{\to} \bar{\theta}$. The method is similar for Theorem 2 with a different $J(\theta)$.

In Theorem 3, we show that $J_N(\theta,\theta') \overset{\text{a.s.}}{\to} J(\theta,\theta') = \int_{\mathbb{X}} \lambda^{-1}(x,\theta')\sigma^2(x)\xi(dx) + \int_{\mathbb{X}} \lambda^{-1}(x,\theta')[\eta(x,\theta) - \eta(x,\bar{\theta})]^2\xi(dx)$ and the convergence is uniform in θ and θ'. Therefore, $J_N(\theta,\hat{\theta}_1^N) \overset{\theta}{\to} J(\theta,\bar{\theta}_1)$ a.s. and Lemma 2 gives $\hat{\theta}_{TSLS}^N \overset{\text{a.s.}}{\to} \bar{\theta}$.

Concerning Theorem 4, we define $J_N(\theta,\lambda_1^N) = (1/N)\sum_{k=1}^N \lambda_k^{-1}[y(x_k) - \eta(x_k,\theta)]^2$ with $\lambda_1^N = (\lambda_1,\ldots,\lambda_N)$ a bounded random sequence of N weights. Using H1 and H2 we show that $|\lambda_k - \bar{\lambda}_k| \overset{\text{a.s.}}{\to} 0$ as $k \to \infty$, with $\bar{\lambda}_k$ a deterministic bounded sequence, implies $\max_{\theta\in\Theta} |J_N(\theta,\lambda_1^N) - J_N(\theta,\bar{\lambda}_1^N)| \overset{\text{a.s.}}{\to} 0$, $N \to \infty$. When $\lambda_k = \lambda(x_k,\hat{\theta}_{WLS}^k)$ and $\hat{\theta}_{WLS}^k \overset{\text{a.s.}}{\to} \bar{\theta}$, under the assumptions of the theorem $\max_{x\in\mathbb{X}} |\lambda(x,\hat{\theta}_{WLS}^k) - \lambda(x,\bar{\theta})| \overset{\text{a.s.}}{\to} 0$ so that, using Lemma 1, $J_N(\theta,\lambda_1^N) \overset{\theta}{\to} \int_{\mathbb{X}} \lambda^{-1}(x,\bar{\theta})\sigma^2(x)\xi(dx) + \int_{\mathbb{X}} \lambda^{-1}(x,\bar{\theta})[\eta(x,\theta) - \eta(x,\bar{\theta})]^2\xi(dx)$ a.s. and Lemma 2 implies $\hat{\theta}_{RWLS}^N \overset{\text{a.s.}}{\to} \bar{\theta}$.

Asymptotic normality. The generic idea goes as follows. Since $\bar{\theta} \in \text{int}(\Theta)$, when the estimator $\hat{\theta}^N$ satisfies $\hat{\theta}^N \overset{\text{a.s.}}{\to} \bar{\theta}$, it implies that exists N_0 (with prob. 1) such that $\hat{\theta}^N \in \text{int}(\Theta)$ for all $N > N_0$. We can thus differentiate the criterion J_N with respect to θ, and its derivative $\nabla J_N(\theta)$ is zero at $\theta = \hat{\theta}^N$. We then perform a Taylor series development of $\nabla J_N(\theta)$ at $\theta = \bar{\theta}$, $\nabla J_N(\hat{\theta}^N) = \nabla J_N(\bar{\theta}) + \nabla^2 J_N(\beta^N)(\hat{\theta}^N - \bar{\theta}) = 0$ for some $\beta^N = (1-\alpha_N)\bar{\theta} + \alpha_N\hat{\theta}^N$, $\alpha_N \in (0,1)$ (and β^N is measurable, see Lemma 3 of [2]), and consider the convergence of the different terms in $\nabla^2 J_N(\beta^N)[\sqrt{N}(\hat{\theta}^N - \bar{\theta})] = -\sqrt{N}\nabla J_N(\bar{\theta})$.

In Theorem 1, we use Lemma 1 to show that $\nabla^2 J_N(\theta) \overset{\theta}{\to} 2\mathbf{M}_1(\xi,\theta) - 2\int_{\mathbb{X}} w(x)[\eta(x,\bar{\theta}) - \eta(x,\theta)]\partial^2\eta(x,\theta)/\partial\theta\partial\theta^\top\xi(dx)$ a.s. Since $\hat{\theta}_{WLS}^N \overset{\text{a.s.}}{\to} \bar{\theta}$, $\beta^N \overset{\text{a.s.}}{\to} \bar{\theta}$ and $\nabla^2 J_N(\beta^N) \overset{\text{a.s.}}{\to} 2\mathbf{M}_1(\xi,\bar{\theta})$, which is nonsingular by assumption. The Central Limit Theorem (CLT) then shows that $-\sqrt{N}\nabla J_N(\bar{\theta})$ is asymptotically normal $\mathtt{N}(0, 4\mathbf{M}_2(\xi,\bar{\theta}))$, which completes the proof.

The proof for Theorem 3 is based on a Taylor development of $\nabla_\theta J_N(\theta,\theta')$, the first derivative of $J_N(\theta,\theta')$ with respect to θ, at $\theta = \theta' = \bar{\theta}$. It gives $\nabla^2_{\theta,\theta} J_N(\beta^N,\hat{\theta}_1^N)[\sqrt{N}(\hat{\theta}_{TSLS}^N - \bar{\theta})] = -\sqrt{N}\nabla_\theta J_N(\bar{\theta},\bar{\theta}) - \nabla^2_{\theta,\theta'} J_N(\bar{\theta},\gamma^N)[\sqrt{N}(\hat{\theta}_1^N - \bar{\theta})]$, with β^N, γ^N measurable and tending a.s. to $\bar{\theta}$, and $\nabla^2_{\theta,\theta} J_N(\theta,\theta')$, $\nabla^2_{\theta,\theta'} J_N(\theta,\theta')$ the derivatives of $\nabla_\theta J_N(\theta,\theta')$ with respect to θ and θ' respectively. Lemma 1 shows that $\nabla^2_{\theta,\theta} J_N(\beta^N,\hat{\theta}_1^N) \overset{\text{a.s.}}{\to} 2\mathbf{M}(\xi,\theta)$ and $\nabla^2_{\theta,\theta'} J_N(\bar{\theta},\gamma^N) \overset{\text{a.s.}}{\to} 0$. Since $\hat{\theta}_1^N$ is $\sqrt{N}$-consistent, $\sqrt{N}\nabla^2_{\theta,\theta'} J_N(\bar{\theta},\gamma^N)(\hat{\theta}_1^N - \bar{\theta}) \overset{\text{p}}{\to} 0$. Finally, the CLT shows that $-\sqrt{N}\nabla_\theta J_N(\bar{\theta},\bar{\theta})$ is asymptotically normal $\mathtt{N}(0, 4\mathbf{M}(\xi,\bar{\theta}))$, which completes the proof.

In Theorem 4, let $\hat{\theta}_\lambda^N$ be the value of $\theta \in \Theta$ that minimises $J_N(\theta,\lambda_1^N)$. We obtain $\nabla^2 J_N(\beta^N,\lambda_1^N)[\sqrt{N}(\hat{\theta}_\lambda^N - \bar{\theta})] = -\sqrt{N}\nabla J_N(\bar{\theta},\lambda_1^N)$ for some measurable β^N tending to $\bar{\theta}$ a.s. Define $\bar{\lambda}_k = \lambda(x_k,\bar{\theta})$ for all k. As in the proof of consistency, we show that $|\lambda_k - \bar{\lambda}_k| \overset{\text{a.s.}}{\to} 0$ implies $\nabla^2 J_N(\theta,\lambda_1^N) - \nabla^2 J_N(\theta,\bar{\lambda}_1^N) \overset{\theta}{\to} 0$ a.s. and then use Lemma 1 to show that $\nabla^2 J_N(\beta^N,\lambda_1^N) \overset{\text{a.s.}}{\to} 2\mathbf{M}(\xi,\bar{\theta})$. Finally, we write $\sqrt{N}\nabla J_N(\bar{\theta},\lambda_1^N) = \sqrt{N}\nabla J_N(\bar{\theta},\bar{\lambda}_1^N) + \sqrt{N}[\nabla J_N(\bar{\theta},\lambda_1^N) - \nabla J_N(\bar{\theta},\bar{\lambda}_1^N)]$. The first term on the right-hand side is asymptotically normal $\mathtt{N}(0, 4\mathbf{M}(\xi,\bar{\theta}))$. The second equals $\Delta_N = (2/\sqrt{N})\sum_{k=1}^N [\lambda_k^{-1} - \bar{\lambda}_k^{-1}]\varepsilon_k\partial\eta(x_k,\theta)/\partial\theta_{|\bar{\theta}}$. We compute $\mathtt{E}\{\Delta_N^2\}$ for $\lambda_k = \lambda(x_k,\hat{\theta}_{WLS}^k)$. The expectation of any cross-product term in Δ_N^2 equals zero, and we get $\mathtt{E}\{\Delta_N^2\} \to 0$ as $N \to \infty$. Chebyshev's inequality then implies $\Delta_N \overset{\text{p}}{\to} 0$ and the proof is complete.

REFERENCES

[1] H.J. Bierens. *Topics in Advanced Econometrics*. Cambridge University Press, Cambridge, 1994.

[2] R.I. Jennrich. Asymptotic properties of nonlinear least squares estimation. *Annals of Math. Stat.*, 40:633–643, 1969.

[3] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis (2nd Edition)*. Springer, Heidelberg, 1993.