Moindres carrés pondérés récursifs dans les modèles de régression à variance paramétrée

Luc Pronzato[1] & Andrej Pázman[2]

Laboratoire I3S, UNSA-CNRS, Bât. Euclide, Les Algorithmes
2000 route des Lucioles, BP 121, 06903 Sophia Antipolis Cedex, France
email: pronzato@i3s.unice.fr
&
Department of Probability and Statistics, Comenius University
Mlynská Dolina, 84215 Bratislava, Slovaquie
email: pazman@center.fmph.uniba.sk

Résumé On compare différentes méthodes d'estimation pour un modèle de régression non linéaire à variance paramétrée: les moindres carrés pondérés (MCP), les moindres carrés pondérés en deux étapes (MCP2), où l'estimateur des moindres carrés obtenu à la première étape sert à calculer la variance utilisée pour une estimation par MCP à la seconde étape, et enfin un estimateur des MCP avec des pondérations déterminées récursivement (MCPR), pour lequel l'estimateur des moindres carrés obtenus à partir de k observations sert à calculer la k-ème pondération. Ce dernier peut être mis en œuvre récursivement quand le modèle de régression est de paramétrisation linéaire (même si ce n'est pas le cas pour la variance des erreurs), et est donc particulièrement intéressant pour les applications faisant intervenir un grand nombre de données.
Mots clés Régression non linéaire, moindres carrés pondérés, convergence, normalité asymptotique

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 applications with large data sets.
Keywords Nonlinear regression, weighted least squares, consistency, asymptotic normality

[1]Avec le soutien du Réseau d'Excellence PASCAL (Pattern Analysis, Statistical Modelling and Computational Learning), no. 2002–506778

[2]Avec le soutien de VEGA (Slovak Grant Agency)

1 Introduction

On considère un modèle de régression non linéaire, avec des observations

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

où $\bar{\theta}$ est la vraie valeur (inconnue) des paramètres du modèle. Les erreurs de mesure $\varepsilon_k = \varepsilon(x_k)$ sont supposées indépendantes. On suppose que leur distribution est inconnue, mais que leur variance est une fonction connue de la variable expérimentale $x \in \mathtt{X} \subset I\!R^d$ et (d'une partie) des paramètres θ apparaissant dans la moyenne $\eta(x, \theta)$, c'est-à-dire,

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

avec c une constante positive. Bien que l'estimateur des moindres carrés (MC) est fortement convergent et asymptotiquement normal sous des hypothèses classiques, il ignore l'information contenue dans la variance des erreurs. Un estimation plus précise peut être obtenue en la prenant en compte, ce qui est l'objectif de cet article. Les propriétés asymptotiques de l'estimateur des MC pondérés (MCP) sont rappelées au paragraphe 3. En particulier, l'estimation est à variance minimale quand les pondérations sont inversement proportionnelles à la variance des erreurs. Comme ces pondérations optimales sont inconnues ($\bar{\theta}$ est inconnu dans (2)), nous considérons une méthode MCP2 en deux étapes: l'estimation par MC de la première étape sert à calculer les pondérations pour une estimation par MCP à la seconde étape. Une troisième méthode d'estimation, par MCP avec pondérations déterminées récursivement (MCPR), est considérée au paragraphe 4: l'estimateur des MC obtenu à partir de k observations sert à calculer *uniquement* la k-ème pondération. Lorsque $\eta(x, \theta)$ est une fonction linéaire de θ (la variance $\lambda(x, \theta)$ peut être non linéaire), la méthode peut être mise en œuvre récursivement, en combinant deux algorithmes des moindres carrés récursifs.

Pour démontrer les propriétés asymptotiques des estimateurs nous utilisons l'hypothèse d'un plan d'expérience randomisé (paragraphe 2), ce qui permet d'éviter les difficultés techniques habituelles, voir [2]. Nous montrons que les propriétés asymptotiques des méthodes MCP2 et MCPR sont identiques, et coincident avec celle de l'estimation par MCP avec pondération optimale. Nous présenterons quelques résultats de simulation qui montrent : 1) que la baisse de performance due à l'estimation des pondérations dans la méthode MCP2 par rapport à l'utilisation des pondérations optimales est presque négligeable, 2) que la baisse supplémentaire de performance due à l'estimation *récursive* des pondérations dans la méthode MCPR est également presque négligeable, 3) qu'en revanche, le gain en précision par rapport aux moindres carrés ordinaires est tout à fait significatif.

2 Plan d'expérience randomisé

L'étude des propriétés asymptotiques d'un estimateur demande de préciser comment sont obtenus les points d'expérience $x_1, x_2, \ldots$ La notion de plan d'expérience randomisé est particulièrement bien adaptée aux situations où la suite des x_i n'est pas contrôlée. Dans toute la suite, *on dira qu'un plan d'expérience est randomisé avec une mesure ξ sur le domaine expérimental $\mathtt{X} \subset \mathbb{R}^d$, $\int_{\mathtt{X}} \xi(dx) = 1$, si les x_i forment une suite de v.a.i. de mesure de probabilité ξ sur $\mathtt{X}$.*

On fera régulièrement appel aux hypothèses suivantes.
H1 Θ est un compact de $\mathbb{R}^p$ tel que $\Theta \subset \overline{\mathrm{int}(\Theta)}$.
H2 $\eta(x,\theta)$ et $\lambda(x,\theta)$ sont des fonctions continues de $\theta \in \Theta$ pour tout $x \in \mathtt{X}$, avec $\eta(x,\theta)$ et $\lambda^{-1}(x,\theta)$ bornées sur $\mathtt{X} \times \Theta$ et $\lambda(x,\bar{\theta})$ bornée sur $\mathtt{X}$ avec $\bar{\theta} \in \Theta$.
H3 $\bar{\theta} \in \mathrm{int}(\Theta)$, $\eta(x,\theta)$ et $\lambda(x,\theta)$ sont deux fois continûment dérivables en $\theta \in \mathrm{int}(\Theta)$ pour tout $x \in \mathtt{X}$, les deux dérivées sont bornées sur $\mathtt{X} \times \mathrm{int}(\Theta)$.

Les démonstrations reposent sur la convergence uniforme p.s. en θ du critère à minimiser $J_N(\cdot)$ qui définit l'estimateur[3]: $\hat{\theta}^N = \arg\min_\theta J_N(\theta)$. Il nous faut donc une loi forte uniforme des grands nombres ; nous utiliserons pour cela le lemme suivant, qui s'inspire du Théorème 2.7.1 de [1].

Lemme 1 (Loi forte uniforme des grands nombres) *Soit $\{z_i\}$ une suite de vecteurs aléatoires i.i.d. de $\mathbb{R}^r$ et $a(z,\theta)$ une fonction réelle mesurable de $(z,\theta) \in \mathbb{R}^r \times \Theta$, continue en θ pour tout z, avec Θ un compact de $\mathbb{R}^p$. Supposons que*

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

alors $\mathtt{E}[a(z,\theta)]$ est continue en $\theta \in \Theta$ et $\frac{1}{N}\sum_{i=1}^N a(z_i,\theta) \overset{\theta}{\to} \mathtt{E}[a(z,\theta)]$ p.s. when $N \to \infty$, où $\overset{\theta}{\to}$ signifie que la convergence est uniforme en θ.

Une fois obtenue la convergence uniforme p.s. de $J_N(\cdot)$, la convergence p.s. de l'estimateur découle du lemme suivant (qui est une simple conséquence de la continuité et de la convergence uniforme).

Lemme 2 *Supposons que la suite de fonctions $\{J_N(\theta)\}$ converge vers $J(\theta)$ uniformément en $\theta \in \Theta$, avec $J_N(\theta)$ continue en $\theta \in \Theta$ pour tout N et Θ un compact de $\mathbb{R}^p$. Alors si $J(\theta)$ satisfait pour une valeur $\bar{\theta} \in \Theta$,*

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

on a $\lim_{N\to\infty} \hat{\theta}^N = \bar{\theta}$ pour tout $\hat{\theta}^N \in \arg\min_{\theta\in\Theta} J_N(\theta)$. Si les fonctions $J_N(\cdot)$ sont aléatoires et la convergence uniforme vers $J(\cdot)$ p.s., la convergence de $\hat{\theta}^N$ vers $\bar{\theta}$ est p.s.

[3]La définition de l'estimateur comme v.a. est assurée par le Lemme 2 de [2], voir aussi [1], p. 16 ; dans ce qui suit on note $\hat{\theta}^N$ le choix mesurable de $\arg\min_{\theta\in\Theta} J_N(\theta)$.

3 Moindres carrés pondérés en deux étapes

L'estimateur des moindres carrés pondérés $\hat{\theta}^N_{MCP}$ minimise

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

avec $w(x) \geq 0$ et borné sur $\mathtt{X}$. Le théorème suivant rappelle des propriétés bien connues de $\hat{\theta}^N_{MCP}$.

Théorème 1 *Soit $\{x_i\}$ un plan d'expérience randomisé de mesure ξ sur $\mathtt{X} \subset I\!R^d$. Supposons que H1 et H2 sont satisfaites et que*

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

Alors $\hat{\theta}^N_{MCP}$ qui minimise (4) dans le modèle (1,2) converge p.s. vers $\bar{\theta}$. Si de plus H3 est satisfaite et la matrice

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

est de rang plein, alors $\hat{\theta}^N_{MCP}$ satisfait $\sqrt{N}(\hat{\theta}^N_{MCP} - \bar{\theta}) \xrightarrow{\mathrm{d}} z \sim \mathtt{N}(0, \mathbf{C}(w,\xi,\bar{\theta}))$ quand $N \to \infty$, avec $\mathbf{C}(w,\xi,\bar{\theta}) = \mathbf{M}_1^{-1}(\xi,\bar{\theta})\ \mathbf{M}_2(\xi,\bar{\theta})\ \mathbf{M}_1^{-1}(\xi,\bar{\theta})$ où

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

La matrice $\mathbf{C}(w,\xi,\bar{\theta}) - \mathbf{M}^{-1}(\xi,\bar{\theta})$ est définie non-négative pour tout choix de $w(x)$, avec

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

et $\mathbf{C}(w,\xi,\bar{\theta}) = \mathbf{M}^{-1}(\xi,\bar{\theta})$ quand $w(x) = \alpha\,\sigma^{-2}(x)$ avec α une constante positive.

Les pondérations optimales $w(x) = \lambda^{-1}(x,\bar{\theta})$ ne peuvent pas être utilisées puisque $\bar{\theta}$ est inconnu. On pourrait alors être tenté d'utiliser des pondérations $\lambda^{-1}(x,\theta)$, c'est-à-dire choisir $\hat{\theta}^N$ qui minimise $J_N(\theta) = (1/N)\sum_{k=1}^N [y(x_k) - \eta(x_k,\theta)]^2 \lambda^{-1}(x_k,\theta)$. Cette approche est cependant à proscrire car en général $\hat{\theta}^N$ ne converge pas vers $\bar{\theta}$.

Considérons alors une méthode en deux étapes, où l'estimateur $\hat{\theta}^N_1$ construit à la première étape sert à calculer les pondérations $\lambda^{-1}(x,\theta)$. L'estimateur de la seconde étape $\hat{\theta}^N_{MCP2}$ est obtenu par minimisation de

$$J_N(\theta, \hat{\theta}^N_1) = \frac{1}{N}\sum_{k=1}^{N} \frac{[y(x_k) - \eta(x_k,\theta)]^2}{\lambda(x_k,\hat{\theta}^N_1)} \tag{9}$$

par rapport à $\theta \in \Theta$. Cette méthode a les propriétés asymptotiques suivantes.

Théorème 2 *Soit $\{x_i\}$ un plan d'expérience randomisé de mesure ξ sur $\mathtt{X} \subset I\!R^d$. Supposons que H1 et H2 sont satisfaites, que $\hat{\theta}_1^N$ converge p.s. vers un vecteur $\bar{\theta}_1 \in \Theta$ et que*

$$\forall \theta, \theta' \in \Theta\,, \ \int_{\mathtt{X}} \lambda^{-1}(x, \bar{\theta}_1)[\eta(x,\theta) - \eta(x,\theta')]^2\, \xi(dx) = 0 \Leftrightarrow \theta = \theta'\,.$$

Alors l'estimateur $\hat{\theta}_{MCP2}^N$ qui minimise (9) dans le modèle (1,2) converge p.s. vers $\bar{\theta}$. Si de plus H3 est satisfaite, la matrice $\mathbf{M}(\xi, \bar{\theta})$ donnée par (8) est de rang plein et l'estimateur auxiliaire $\hat{\theta}_1^N$ utilisé dans (9) converge en $\sqrt{N}$, alors $\hat{\theta}_{MCP2}^N$ satisfait

$$\sqrt{N}(\hat{\theta}_{MCP2}^N - \bar{\theta}) \xrightarrow{\mathrm{d}} z \sim \mathtt{N}(0, \mathbf{M}^{-1}(\xi, \bar{\theta}))\,, \ N \to \infty\,.$$

On peut noter que $\mathbf{M}^{-1}(\xi, \bar{\theta})$ est la matrice de covariance asymptotique de l'estimateur des MCP dans le cas idéal où les pondérations optimales sont connues, voir le théorème 1.

Un candidat naturel pour l'estimateur auxiliaire de la première étape est $\hat{\theta}_{MCP}^N$ qui minimise $J_N(\theta)$ donné par (4) avec des pondérations *arbitraires*: sous les hypothèses du théorème 1 $\hat{\theta}_{MCP}^N$ converge en $\sqrt{N}$ puisque $\sqrt{N}(\hat{\theta}_{MCP}^N - \bar{\theta}) \xrightarrow{\mathrm{d}} z \sim \mathtt{N}(0, \mathbf{C}(w, \xi, \bar{\theta}))$. On peut choisir en particulier $w(x) = 1$ pour tout x, ce qui correspond à l'estimateur des moindres carrés (ordinaires) $\hat{\theta}_{MC}^N$.

On peut accroître le nombre d'étapes, ce qui conduit à un estimateur *des moindres carrés re-pondérés itérativement*, une méthode qui repose sur une suite d'estimations:

$$\hat{\theta}_k^N = \arg\min_{\theta \in \Theta} J_N(\theta, \hat{\theta}_{k-1}^N)\,, \ k = 2, 3 \ldots \tag{10}$$

avec $J_N(\theta, \theta')$ défini par (9) et $\hat{\theta}_1^N$ pouvant être pris égal à $\hat{\theta}_{LS}^N$. On montre alors par une simple récurrence que pour k *fixé* $\hat{\theta}_k^N$ à les mêmes propriétés asymptotiques[4] que $\hat{\theta}_{MCP2}^N$.

4 Moindres carrés pondérés récursivement

On définit l'estimateur des MCP à pondérations déterminées récursivement $\hat{\theta}_{MCPR}^N$ comme la valeur de $\theta \in \Theta$ qui minimise

$$J_N(\theta) = \frac{1}{N} \sum_{k=1}^{N} \frac{[y(x_k) - \eta(x_k, \theta)]^2}{\lambda(x_k, \hat{\theta}_{MCP}^k)}\,, \tag{11}$$

où l'estimateur auxiliaire $\hat{\theta}_{MCP}^k$ utilise des pondérations $w(x)$ arbitraires et est construit *seulement à partir des k premières observations* $Y_1, \ldots, Y_k$ et des points expérimentaux $x_1, \ldots, x_k$. Les lemmes 1 et 2 permettent de montrer la propriété suivante.

[4] Il n'y a donc aucun intérêt *du point de vue asymptotique* à utiliser plus de deux étapes dans (10). Les choses peuvent être différentes pour N fixé.

Théorème 3 *Soit $\{x_i\}$ un plan d'expérience randomisé de mesure ξ sur $\mathtt{X} \subset I\!R^d$. Supposons que H1 et H2 sont satisfaites, que $\lambda(x,\theta)$ est continue sur $\mathtt{X} \times \Theta$ avec $\mathtt{X}$ compact, que*

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

et que la fonction $w(x)$ est telle que (5) est satisfaite. Alors $\hat{\theta}^N_{MCPR}$ qui minimise (11) dans le modèle (1,2) converge p.s. vers $\bar{\theta}$. Si de plus H3 est satisfaite et la matrice $\mathbf{M}(\xi,\bar{\theta})$ donnée par (8) est de rang plein, alors $\hat{\theta}^N_{MCPR}$ satisfait

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

Les deux estimateurs $\hat{\theta}^N_{MCP2}$ et $\hat{\theta}^N_{MCPR}$ ont les mêmes performances asymptotiques (en termes de matrice de covariance) que l'estimateur des MCP avec des pondérations optimales. ceci rend $\hat{\theta}^N_{MCPR}$ particulièrement intéressant quand $\eta(x,\theta)$ est linéaire en θ, c'est-à-dire quand

$$\eta(x,\theta) = f^\top(x)\theta\,. \tag{12}$$

En effet, l'estimateur auxiliaire $\hat{\theta}^k_{MCP}$ se calcule alors de façon récursive selon

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

Soit k_0 le premier entier tel que $f(x_1), \ldots, f(x_{k_0})$ engendrent $I\!R^p$. La récursion peut être initialisée à $k = k_0$ par

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

L'estimateur $\hat{\theta}^k_{MCPR}$ s'obtient par une récursion similaire, effectuée simultanément, avec $w^{-1}(x_{k+1})$ remplacé par $\lambda(x_{k+1}, \hat{\theta}^{k+1}_{MCP})$. Notons cependant que $\hat{\theta}^k_{MCP}$ est linéaire par rapport aux observations $Y_1, \ldots, Y_k$ tandis que $\hat{\theta}^k_{MCPR}$ *ne l'est pas.*

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.