Modèle de régression semi-paramétrique partiellement linéaire. Aplication à la librarie Prestige R

Article · December 2013

CITATIONS
0

READS
4,021

2 authors, including:

Mamoudou Sow
Électricité de France (EDF)
1 PUBLICATION **0** CITATIONS

SEE PROFILE

Économétrie non paramétrique

Modèle de régression semi-paramétrique partiellement linéaire avec régresseurs strictement exogènes

SOW Mamoudou [1]

Professeur : NGUYEN-VAN Phu [2]

11 décembre 2013

Résumé

Les modèles semi-paramétriques de régression sont un compromis entre la régression paramétrique (linéaire ou non linéaire) et l'approche entièrement non paramétrique. Cependant, leur principale avantage par rapport à la régression non paramétrique est la convergence plus rapide des estimateurs (réduction de dimension), mais au prix d'un risque plus élevé d'erreur de spécification. Notons que nous distinguons trois types de régression semi-paramétrique exploitant l'approche du noyau : les régressions partiellement linéaires, à coefficients variables et sur indice. Dans cet travail nous appliquerons la régression partiellement linéaire sur la base de donnée dénommée "Prestige" de la librairie car dans R. Dans un souci de robustesse des résultats, nous utiliserons les différentes techniques de régression non paramétrique (least square cross-validation,Nadaraya-Watson et leave-one-out) et utiliser enfin le calcul de l'erreur quadratique moyenne appliquée à des données hors-échantillon.

1. Étudiant Master 2-Statistique et Économétrie, Université de Strasbourg
2. Enseignat-Chercheur. Chargé de recherche CNRS-HDR, BETA, Université de Strasbourg. Thèmes de recherche : Économie de l'environnement, de l'énergie et des ressources naturelles et Économétrie appliquée. Il a reçu le "Prix Guy OURISSON" le 19 novembre 2013. Ce prix récompense un jeune chercheur (jusqu'à 40 ans) menant des recherches particulièrement prometteuses en Alsace.

1 Introduction

La régression paramétrique est de nos jours très critiquée du fait que nous sommes souvent confrontés à une mauvaise spécification de modèle. De plus l'hypothèse de normalité faite sur les erreurs du modèle est très forte pour l'obtention d'estimateurs efficaces et convergents. C'est pourquoi, les modèles non paramétriques (forme fonctionnelle non spécifiée) et semi-paramétriques qui sont composés d'une partie paramétrique et non paramétrique deviennent un puissant outil pour résoudre ce problème.
Cependant, l'estimation semi-paramétrique est beaucoup utilisée dans des situations où la régression non paramétrique ne fournit pas de résultats efficaces. Ce type d'estimation a été proposée par Robinson (1988) impliquant trois différentes étapes dans le processus de calcul du coefficient de la partie paramétrique et de la fonction de forme inconnue. Ces étapes seront détaillées dans la suite avec des exemples à l'appui.
De nombreuses revues de littérature ont abordé l'étude de ce type d'estimation. Nous pouvons citer par exemple la méthode d'Ichimura (1993) dans le cadre de la régression sur indice. Klein et Spady (1993) proposent un estimateur lorsque la variable dépendante est binaire en utilisant l'estimation par le maximum de vraisemblance. Rappelons toutefois que parmi les différents types de régressions semi-paramétriques proposées (les régressions partiellement linéaires, à coefficients variables et sur indice), l'estimation partiellement linéaire reste la méthode la plus utilisée. C'est cette méthode qui sera utilisée et comparée à quelques méthodes non paramétriques par le biais du calcul des erreurs quadratiques moyennes.

2 Spécification du modèle

Notre modèle de régression partiellement linéaire s'écrit de la forme suivante :

$$Y_i = X_i^{'}\beta + g(Z_i) + \varepsilon_i, i = 1, ..., n \tag{1}$$

Les variables (Y_i, X_i et Z_i) sont indépendantes et identiquement distribuées. $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^p$ et $Z_i \in \mathbb{R}^q$. De plus, $\mathbb{E}(\varepsilon|X_i, Z_i) = 0$ et $\mathbb{V}(\varepsilon|X_i, Z_i) = \mathbb{E}(\varepsilon^2)$ avec des ε_i normaux indépendants et identiquement distribués.
Cependant, la variable X ne doit pas contenir de constante car elle n'est pas séparément

identifiable de la fonction inconnue $g(Z_i)$. Une généralisation de ce modèle peut être le suivant :

$$\alpha + g(Z) = \alpha + c + [g(Z) - c] \equiv \tilde{\alpha} + \tilde{g}(Z) \tag{2}$$

Ainsi, en l'absence de forme fonctionnelle pour g, les expressions de gauche et de droite de l'équation (2) constituent un même modèle.
Dans le cadre conventionnel pour l'estimation de l'ensemble des observations, réécrivons notre modèle sous sa forme matricielle. La spécification finale que nous retiendrons sera le suivant :

$$Y = X\beta + g(Z) + \varepsilon \tag{3}$$

3 Processus d'estimation

Nous allons cette fois-ci nous intéresser à l'estimation du modèle (3). Nous estimerons donc le paramètre β et notre fonction $g(Z)$ dont la forme fonctionnelle est inconnue.
L'estimation de β et de $g(Z)$ provient d'une version généralisée de la régression partitionnée (voir théorème de Frisch-Waugh) [3]
Pour procéder à l'estimation, écrivons d'abord l'espérance conditionnelle à Z de notre modèle. Nous obtenons la relation suivante :

$$\mathbb{E}(Y|Z) = \beta\ \mathbb{E}(X|Z) + g(Z)$$

Nous allons ensuite soustraite cette expression de notre modèle semi-paramétrique (équation 3). Cette étape nous permet d'obtenir l'équation ci-dessous :

$$Y \text{ - } \mathbb{E}(Y|Z) = [X - \mathbb{E}(X|Z)]\beta$$

A partir de cette équation, nous aurons ainsi deux vecteurs de résidus caractérisés par l'équation (4) suivante :

$$Y^* = X^*\beta + \varepsilon^* \tag{4}$$

Notons toutefois que l'estimation des paramètres de l'équation (4) n'est pas réalisable, puisque les fonctions $\mathbb{E}(Y|Z)$ et $\mathbb{E}(X|Z)$ sont inconnues.
Pour cela, posons $\beta \equiv \beta_{nr}$. En appliquant les moindres carrés ordinaires (MCO) à (4) pour

3. Le théorème de Frisch-Waugh permet de récupérer les coefficients partiels de régression d'un sous-ensemble de variables explicatives dans un modèle de régression linéaire.

estimer β_{nr}, nous obtenons :

$$\hat{\beta}_{nr} = (X^*X^*)^{-1}X^*Y^* = \left[\sum_i^n X_i^* X_i^{*'}\right]^{-1} \sum_i^n X_i^* Y_i^*$$

L'application du théorème central limite (TCL) de Lindeberg-Lévy nous montre que $\sqrt{n}(\hat{\beta}_{nr} - \beta_{nr}) \longrightarrow^p N(0, V)$

L'estimateur MCO de la partie linéaire converge vers sa vraie valeur (même vitesse que celle du modèle linéaire) et possède une distribution asymptotiquement normale.

Pour rendre l'estimation de β et de g(Z) réalisable, Robinson (1988) propose une démarche en trois étapes :

1. $\forall$ i $\in$ 1,...,n, estimer le modèle suivant :

$$\hat{Y}_i^* = Y_i \text{ - } \hat{E}(Y_i|Z_i) \text{ et } \hat{X}_i^* = X_i \text{ - } \hat{E}(X_i|Z_i)$$

   Où $\hat{E}(Y_i|Z_i)$ et $\hat{E}(X_i|Z_i)$ représentent les espérances conditionnelles estimées par la méthode du noyau.

2. Obtenir β_{MCO} en régressant Y_i^* sur X_i^*
3. Ensuite, remplacer β_{MCO} dans le modèle de base (équation 3). On obtient :

$$Y_i \text{ - } X_i'\hat{\beta}_{MCO} = g(Z_i) + \varepsilon_i$$

Enfin estimer $g(Z)$ à l'aide d'une régression non paramétrique en utilisant un estimateur de noyau de notre choix (Nadaraya-Watson, linéaire locale,...).

4 Données et modèle "Prestige"

Nous cherchons à établir la relation entre le logarithme d'un indicateur du prestige social lié au métier exercé par un individu (le score de Pineo-Porter calculé sur la base d'une enquête menée dans les années 60 au Canada), son revenu (en dollars de 1971) et son éducation (en nombre d'années).

La base de donnée en question comporte une liste de 102 professions (chimiste, économiste, architecte, médecin,...) classées selon leur degré de rang social (100 étant le score le plus élevé obtenu par une profession et 0 le plus faible possible). La variable dépendante est le

prestige. le revenu et le niveau d'éducation dénotent les variables explicatives.
Ces données ont été obtenues suite à une enquête menée par "*Canada (1971) Census of Canada. Vol. 3, Part 6. Statistics Canada [pp. 19-1,19-21]. Personal communication from B. Blishen, W. Carroll, and C. Moore, Departments of Sociology, York University and University of Victoria.* Nous pourrons ainsi écrire notre modèle comme suit :

$$log(Prestige) = f(income, education) + \varepsilon.$$

$$f(income, education) = \begin{cases} \alpha_0 + \alpha_1 income + \alpha_2 education \\ \beta education + g(income) \\ m(income, education) \end{cases}$$

La fonction f caractérise les différentes formes fonctionnelles prise par $g(Z)$ dans l'équation (3). Cela nous conduit donc à estimer dans un premier temps un modèle linéaire. Dans un second temps, nous aborderons la problématique étudiée, c'est-à-dire le modèle de régression semi-paramétrique partiellement linéaire.
Pour des raisons d'analyses approfondies, nous examinerons l'estimation non paramétrique afin de conclure si le thème abordé tient ou pas avec l'étude du *prestige.*

5 Résultats d'estimation

5.1 Estimation du modèle linéaire

Les résultats MCO du modèle linéaire sont détaillés ci-dessous :

```
Call:
lm(formula = lprestige ~ income + education, x = TRUE, y = TRUE)
Residuals:
     Min       1Q   Median       3Q      Max
-0.80870 -0.12450  0.03527  0.14143  0.44742
Coefficients:
               Estimate  Std. Error t value Pr(>|t|)
(Intercept) 2.630196185 0.088994625  29.555  < 2e-16 ***
income      0.000029438 0.000006199   4.749 6.91e-06 ***
```

```
education   0.087952505 0.009646324   9.118 9.18e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.2159 on 99 degrees of freedom
Multiple R-squared: 0.7024,     Adjusted R-squared: 0.6964
F-statistic: 116.8 on 2 and 99 DF,  p-value: < 2.2e-16
```

Les résultats MCO de la régression de log(Presitige) sur income et education, nous montrent que 70% de la variation totale du prestige est expliquée par le revenu et l'éducation ($R^2 \equiv 0.7$). Le R^2 ajusté est proche du R^2. Ce qui signifie que le nombre de régressseurs ne gonfle pas artificiellement le pouvoir explicatif du modèle.
De plus, au seuil de 1% les coefficients de income et education sont individuellement et conjointement différents de 0 (les t-stat sont en dehors de l'intervalle [-1.96, 1.96] et la p-value pour la stat-F est nulle).
Nous pourrons ainsi conclure que le revenu et l'éducation influencent positivement le prestige social des professions. Cependant, nous ne savons pas si les conclusions suite à cette régression linéaire sont exactes. C'est pourquoi, nous allons effectuer les tests usuels pour cette régression. A présent, testons la présence ou non d'hétéroscédascité des erreurs ainsi que la spécification de forme fonctionnelle linéaire.

Le test de Breusch-Pagan d'absence d'hétéroscédasticité, de Ramsey pour la spécification linéaire et le test non paramétrique nous donnent les résultats suivants :

```
 studentized Breusch-Pagan test

data:  lin
BP = 4.9931, df = 2, p-value = 0.08237

 RESET test

data:  lin
RESET = 9.0795, df1 = 2, df2 = 97, p-value = 0.0002429
```

```
Consistent Model Specification Test
Parametric null model: lm(formula = lprestige ~ income + education, x = TRUE, y = TRUE)
Number of regressors: 2
IID Bootstrap (399 replications)

Test Statistic 'Jn': 2.823608   P Value: 0.0025063 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Null of correct specification is rejected at the 1% level
```

En vertu des résultats obtenus, le test de Breusch-Pagan rejette l'hypothèse nulle d'homoscédasticité des erreurs au seuil de 10%.
De plus, nous rejetons au seuil de 1% la forme paramétrique postulée, que cela soit avec le test RESET ou le test non paramétrique.

Ainsi, les conclusions obtenues avec les MCO deviennent invalides. Le processus générateur de données est inadéquat. Nous allons donc recourir à notre spécification semi-paramétrique partiellement linéaire, pour assurer une meilleure compatibilité entre les données et le processus générateur de données.

5.2 Résultats du modèle semi-paramétrique

Les estimations sont ici basées sur le modèle (3) en appliquant la méthode de Robinson qui requiert le calcul de régressions auxiliaires. La variable dépendante (y = lprestige) et la variable explicative de la portion linéaire (x = education) font l'objet d'une régression non paramétrique sur la variable explicative de la portion non paramétrique (z = income). Les fenêtres de lissage correspondent à ces deux régressions non paramétriques auxiliaires. En appliquant cette méthode d'estimation, nous obtenons les résultats ci-dessous :

```
Partially Linear Model
Regression data: 102 training points, in 2 variable(s)
```

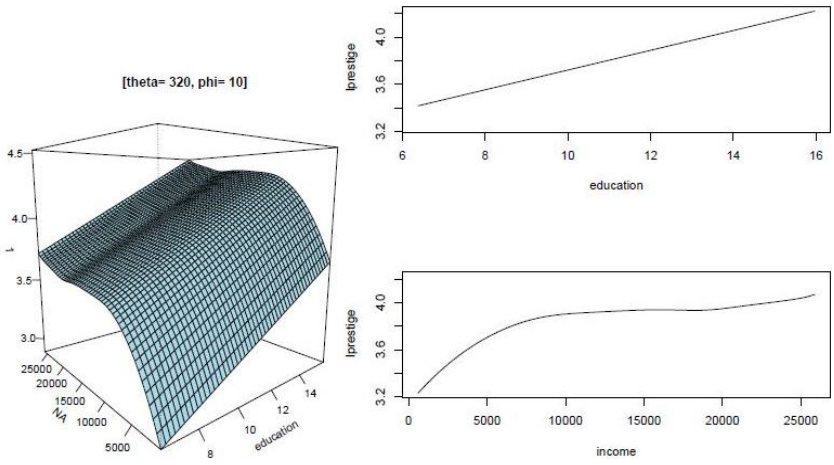
```
With 1 linear parametric regressor(s), 1 nonparametric regressor(s)
                  y(z)
Bandwidth(s): 3484.499

                  x(z)
Bandwidth(s): 2751.956
                education
Coefficient(s): 0.08377204
Kernel Regression Estimator: Local-Linear
Bandwidth Type: Fixed
Residual standard error: 0.1808075
R-squared: 0.7852653
Continuous Kernel Type: Second-Order Gaussian
No. Continuous Explanatory Vars.: 1
```

Nous pouvons remarquer que ce modèle a une meilleure performance prédictive intra-échantillon que le modèle linéaire (Son $R^2 = 0.786$ est plus élevé que celui du modèle linéaire).

5.3 Analyse graphique

Le premier graphique à gauche est une représentation en 3 dimensions (3D) de la relation. Le second graphique montre les effets partiels associés à chaque variable explicative en gardant les autres régresseurs à leur niveau médian. La composante paramétrique de la fonction est linéaire et la portion non paramétrique ressemble à un polynôme de degré 3. Ce modèle est additif et aucune interaction n'a été spécifiée entre l'éducation et le revenu. Par conséquent un graphique bivarié des relations partielles suffit à caractériser les relations entre la variable dépendante et chaque régresseur et le graphique 3D n'apporte pas d'information supplémentaire pertinente. Le prestige professionnel s'accroit fortement avec le revenu lorsque le revenu est faible, il stagne à partir de 10000$ puis repart légèrement à la hausse à partir de 18000$. Ceci reste valable pour tout niveau de la variable education. La relation entre le prestige et l'éducation est positive et de pente 0.083 (proche de la

valeur obtenue avec le modèle linéaire). Ceci reste valable pour tout niveau de la variable income.



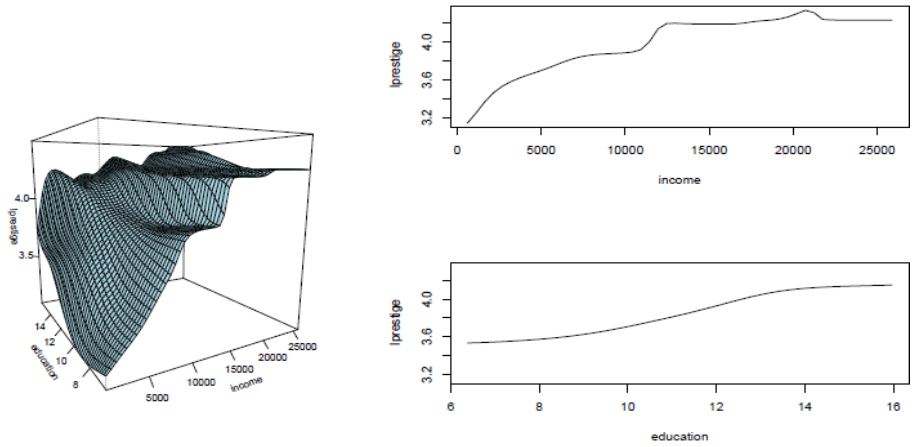
5.4 Résultats du modèle non paramétrique

Malgré que notre analyse ne soit pas fondée sur la régression non paramétrique, pour des mesures de résultats et de conclusions solides sur le thème étudié, nous allons l'aborder et établir une comparaison avec le modèle semi-paramétrique partiellement linéaire.
Les premiers résultats de régression non paramétrique par noyau avec l'estimateur Nadaraya-Watson sont les suivants :

```
Regression Data: 102 training points, in 2 variable(s)
                income education
Bandwidth(s): 1031.859  1.217422
Kernel Regression Estimator: Local-Constant
Bandwidth Type: Fixed
Residual standard error: 0.1646645
R-squared: 0.8259631
Continuous Kernel Type: Second-Order Gaussian
No. Continuous Explanatory Vars.: 2
```

La combinaison de l'estimateur Nadaraya-Watson et la technique de validation croisée par les moindres carrés permet d'interpréter la taille des fenêtres de lissage de chacun des

régresseurs comme un indicateur d'influence sur la moyenne conditionnelle : si le paramètre de lissage est grand, le régresseur n'a pas d'effet sur l'espérance conditionnelle, il est 'surlissé' ou 'smoothed out'. Etant donné les unités des variables explicatives (le dollar et les années d'éducation), les deux fenêtres sont de taille raisonnable. Les variables explicatives ne sont pas surlissées, leur présence comme régresseurs se justifie selon ce critère.
Comme fait dans le cadre semi-paramétrique, nous allons également étudier les effets partiels des variables income et education.



Les portions plates des graphiques ci-dessus coïncident avec une région du support qui contient peu d'observations. En effet, nous n'avons qu'une quinzaine de valeurs supérieures à un revenu de 10000$. Par conséquent, cette zone est estimée avec moins de précision, surtout avec l'estimateur de Nadaraya-Watson dont le biais est influencé par la densité des points sur le support.
A cause de ce problème, nous allons tester la significativité des variables explicatives avec la fonction npsigtest de R.

```
Kernel Regression Significance Test
Type I Test with IID Bootstrap (399 replications)
Explanatory variables tested for significance:
income (1), education (2)
                income education
Bandwidth(s): 1031.859  1.217422
Significance Tests
```

```
P Value:
income    0.218045
education 0.030075 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ce test ne confirme pas que le régresseur income a un effet significatif puisque nous ne pouvons rejeter H0 : $E[lprestige|income; education] = E[lprestige|education]$: Rappelons que Racine et al. (2006) montrent que ce test est équivalent à tester H0 :

$$\delta \frac{E(lprestige|income,education)}{\delta income} = 0 \text{ presque partout,}$$

contre l'aternative 'différent de zéro'. La fonction npsigtest() utilise une distribution bootstrap pour évaluer la statistique de test en échantillon fini. Seul l'effet marginal de la variable education est significativement différent de 0 au seuil de 5%.

6 Performance hors échantillons

Dans ce qui suit, nous considérons trois méthodes de validation croisée pour juger la performance prédictive hors-échantillon des différents modèles. Nous utilisons la fonction de perte donnée par l'erreur quadratique moyenne (ou 'mean squared error').

$$\sum_i^n (Y_i - \hat{m}(X_i))^2.$$

Appliquée à des données hors-échantillon $(Y_{new,i}, X_{new,i})$, cette mesure est appelée erreur quadratique moyenne de la prédiction (ou 'predicted mean squared error' abrégée pmse).

Intéressons nous d'abord à la technique appelée 'leave-one-out cross-validation'.
Le calcul de pmse s'effectue ici à l'aide d'une boucle qui porte sur les n observations. Le principe est le suivant : estimer à chaque itération un modèle sur n-1 observations en omettant l'observation i ; utiliser ce modèle pour prédire l'observation manquante i ; calculer la moyenne des erreurs de prédiction pour les n points. On répète ensuite la procédure pour

les modèles concurrents et on compare les pmse entre modèles.
Les pmse obtenus sont de 0.045, 0.031 et 0.034 respectivement pour le modèle linéaire, semi-paramétrique et non paramétrique. C'est donc le modèle partiellement linéaire qui a la plus petite erreur de prédiction selon ce critère et performe donc mieux.

Nous appliquons maintenant la technique appelée 'K-fold cross-validation' avec K = 3, soit avec 3 sous-échantillons de taille identique (34 observations).
On commence par partager (aléatoirement) l'échantillon global en $K = 3$ sous-échantillons E1, E2 et E3 de taille identique (n1 = n2 = n3). On estime K fois le modèle en utilisant les observations des sous-échantillons $E1; E2$, $E1; E3$ et $E2; E3$. On se sert de chacun des trois modèles estimés pour calculer l'erreur moyenne de prévision hors échantillon sur, respectivement, les points des sous-échnatillons E3, E2 et E1 ; on calcule la moyenne de ces K erreurs moyennes de prédiction. On répète ensuite la procédure pour les modèles concurrents et on compare les erreurs quadratiques moyennes de prédiction. Nous obtenons des "pmse" de 0.114, 0.052 et 0.088 respectivement pour le modèle linéaire, semi-paramétrique et non paramétrique. Une fois de plus, c'est le "pmse" du modèle semi-paramétrique qui est le plus faible.

Appliquons la technique de validation croisée randomisée 'leave-d-out' pour un échantillon de calibrage (training dataset) de 80 observations, un échantillon d'évaluation de taille $d = 22$ et B = 100 tirages aléatoires.
Ici, on effectue B tirages aléatoires sans remise d'un sous-échantillon de taille fixe $n_2 << n$; on estime B fois le modèle sur les $n_1 = n - n_2$ observations restantes ; on calcule B erreurs quadratiques moyennes de prédiction sur les B échantillons de taille n_2 ; on calcule ensuite l'indicateur d'erreur globale de prédiction avec les B erreurs moyennes de prédiction. On répète la procédure pour les modèles concurrents et on compare les erreurs quadratiques moyennes de prédiction.
Nous obtenons des "pmse" de 0.052, 0.039 et 0.037 respectivement pour le modèle linéaire, semi-paramétrique et non paramétrique. C'est le modèle non-paramétrique qui dispose cette fois-ci du plus faible score et performe mieux.

7 Conclusion

L'étude de la relation entre le prestige social, le revenu et le niveau d'éducation fait recours à des techniques d'estimations assez variées. Malgré que notre but était d'utiliser un modèle de régression semi-paramétrique partiellement linéaire, nous avons étendu notre étude en comparant notre modèle à une régression non paramétrique. Ceci nous permet de souligner des biais éventuels cachés par notre modèle de régression semi-paramétrique. Si l'objectif de la modélisation est la prévision, le modèle partiellement linéaire est un bon candidat car il présente la plus petite erreur moyenne de prévision hors échantillon avec la technique "leave-one-out cross validation" et "K-fold". De plus, il présente une performance relativement similaire au modèle non paramétrique avec la validation croisée randomisée. Pour l'analyse économique/sociologique, le modèle non paramétrique fait ressortir un effet d'interaction entre les variables explicatives, qui est ignoré dans l'estimation partiellement linéaire. On pourra alors privilégier le modèle non paramétrique, car il ne performe pas mal en hors échantillon par rapport au modèle partiellement linéaire. Si l'on choisit le modèle non paramétrique, l'estimateur localement linéaire devrait être privilégié au détriment de l'estimateur Nadaraya-Watson, car son biais en échantillon fini ne dépend pas de la densité des points du support.

Références

[1] Ibrahim Ahamada et Emmanuel Flachaire "Économétrie non paramétrique (2008)". *Econometrica.*

[2] Li, Q. et J. Racine. "Nonparametric Econometrics," *Princeton University Press.*

[3] Ichimura, H. (1993), "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models", *Journal of Econometrics, 58, 71-120.*

[4] Klein, R.W. et Spady, R.H. (1993), "An efficient semiparametric estimator for binary response models", *Econometrica, 61, 387-421.*

[5] Logiciel R version 2.15.3, library lmtest, library np-Kernel Consistent Model Specification Test with Mixed Data Types (npcmstest).