



Université d'Orléans

MASTER ECONOMETRIE ET STATISTIQUE APPLIQUEE (ESA)

Université d'Orléans

Econométrie et Statistique Non Paramétrique (Partie II)

Régressions Non Paramétriques Univariées

Christophe Hurlin

Année Universitaire 2007-2008

Master Econométrie et Statistique Appliquée (ESA)
Université d'Orléans
Faculté de Droit, d'Economie et de Gestion
Bureau A 224
Rue de Blois – BP 6739
45067 Orléans Cedex 2
www.univ-orleans.fr/deg/masters/ESA/

October 10, 2007

Contents

1	Introduction	1
2	Principe d'Estimations Non Paramétriques	2
3	Régressions Kernel	4
3.1	Régression avec lissage par moyenne mobile	4
3.1.1	Etude de la convergence en probabilité	6
3.1.2	Etude de la convergence en distribution et intervalles de confiance	8
3.2	Régression avec lissage par opérateur à noyau ou régression kernel	9
3.2.1	Etude de la convergence en probabilité	11
3.2.2	Etude de la convergence en loi et intervalles de confiance	12
3.3	Sélection du paramètre de lissage dans la régression Kernel	14
3.3.1	Critère de la MISE et approche GCV	16
3.3.2	Critère de l'AMISE	17
3.4	Estimation d'une fonction de densité	18
3.4.1	Procédure UNIVARIATE	20
3.4.2	Procédure KDE	22
3.4.3	Procédure SAS INSIGHT	26
3.5	Regressions Kernel : Applications SAS INSIGHT	28
4	Régressions locales	30
4.1	Régression locale : LOESS et LOWESS regressions	30
4.2	Procédure LOESS	33
4.2.1	Sorties graphiques	39
4.2.2	Sélection du paramètre de lissage	40
4.2.3	Autres options de la procédure LOESS	45

1. Introduction

La régression non paramétrique a longtemps été opposée à la régression économétrique usuelle. En effet, dans l'esprit des travaux de la Cowles Commission, cette dernière est généralement fondée sur une spécification issue de la théorie économique et peut se ramener à une forme explicite dépendant d'un nombre fini de paramètres que l'on cherche ensuite à estimer par la méthode la plus appropriée (maximum de vraisemblance, MCO, GMM etc.). A l'inverse, la régression non paramétrique est une régression "sans modèle", au sens où comme nous allons le voir, on cherche à estimer la fonction de lien caractérisant la relation entre deux variables économiques. C'est donc une régression "a-théorique" (au sens de la théorie économique) par opposition à la régression paramétrique usuelle, censée découler de l'estimation d'une forme réduite d'un modèle théorique.

Historiquement, le principe des régressions non paramétriques remonte au 19^{ème} siècle selon Cleveland and Loader (1995), toutefois les premiers travaux modernes sur ce sujet datent des années 50. La première application que nous verrons relève de l'**estimation de fonctions de densité** par des méthodes d'opérateur à noyau (kernel) avec les travaux fondateurs de **Rosenblatt (1956) et de Parzen (1962)**. Ces premiers travaux ont été étendus à la notion de **régression kernel**, imparfaitement traduit en français par le terme de régression avec lissage par opérateur à noyau. Dans ce domaine, on identifie deux papiers fondateurs publiés la même année : **Nadaraya (1964) et Watson (1964)**. Enfin, la **régression local polynomiale, plus récente Cleveland et Devlin (1988)** constitue une généralisation de ces méthodes.

Le présent cours ne portera que sur les régressions non paramétriques univariées. Pour les régressions multivariées, nous revoyons le lecteur à l'ouvrage de référence de Yatchew (2003). Dans ce cours nous présenterons tout d'abord les grands principes de la régression non paramétrique. Dans une seconde section, nous étudierons la régression non paramétrique kernel. Dans ce cadre nous étudierons comme un cas particulier, le principe de l'estimation kernel d'une densité. Enfin, dans une troisième section nous étudierons la régression locale polynomiale et plus spécifiquement les régressions de type LOESS et la LOWESS.

Toutes les applications se feront sur le logiciel SAS à, partir des procédures UNIVARIATE (estimation kernel de densité), KDE (estimation kernel de den-

sité), SAS INSIGHT (estimation kernel de densité, regression kernel et regression locale polynomiale) et LOESS (regression locale polynomiale).

2. Principe d'Estimations Non Paramétriques

Lorsque l'on souhaite décrire l'influence d'une variable quantitative sur un événement en faisant le moins d'hypothèse possible sur la forme de la relation, on distingue deux approches¹ :

- L'approche de la **régression paramétrique**
- L'approche de la **régression non-paramétrique**

Comme on le sait le but d'un modèle de régression consiste à déterminer la façon dont l'espérance d'une variable dépendante Y dépend d'un ensemble de variables explicatives X . Supposons pour simplifier que $X \in \mathbb{R}$. Le problème consiste donc à déterminer pour chaque réalisation de la variable x , la valeur de la fonction $f(x)$, dite fonction de lien.

Definition 2.1. *On appelle fonction de lien, la fonction $f(x)$ qui a toute réalisation x de la variable explicative X associe la quantité :*

$$E(Y | X = x) = f(x) \quad (2.1)$$

Pour caractériser cette fonction de lien, la première approche consiste à utiliser un modèle de **régression paramétrique**. On suppose que cette fonction peut s'écrire comme une **fonction explicite des valeurs** de X . Cette fonction peut être linéaire, logarithmique, non-linéaire etc. Par exemple, dans le cas linéaire on postule que :

$$E(Y | X = x) = \alpha + \beta x \quad (2.2)$$

On cherche alors à déterminer les meilleures valeurs de α et β compte tenu d'un critère, par exemple celui de la MSE.

Definition 2.2. *Dans un modèle de régression paramétrique, la fonction de lien est (i) de forme explicite et (ii) peut s'écrire en fonction d'un nombre réduit de paramètres. Exemple :*

$$E(Y | X = x) = f(x, \theta) \quad (2.3)$$

où $f(\cdot)$ est connue avec $\theta \in \mathbb{R}^K$.

¹Nous n'évoquerons pas dans ce cours l'approche dite semi-paramétrique, généralement réservée aux modèles de régression multivariée.

L'exemple typique est celui d'un modèle linéaire, où l'on postule que :

$$E(Y | X = x) = \alpha + \beta x = f(x, \alpha, \beta) \quad (2.4)$$

On sait qu'à partir de ce type de modèle, on dispose :

1. D'une mesure synthétique du lien qui lie X à Y qui peut être notamment **confrontée à une théorie économique** (tests de spécification, approche à la **Cowles Commission**).
2. **D'écarts type et d'intervalle de confiance associés aux paramètres et aux valeurs prévues de la variable Y**
3. De **tests simples (inférence)** à mettre en oeuvre sur la valeur des paramètres du modèle.

Au contraire, on peut retenir une approche non paramétrique dans laquelle on va **estimer la relation entre le niveau moyen de Y et toutes les valeurs réalisées de X** . On ne postule aucune forme spécifique sur la fonction de lien.

Definition 2.3. *Dans un modèle de régression non-paramétrique, la fonction de lien (i) n'a pas de forme explicite et (ii) ne peut pas s'écrire en fonction d'un nombre réduit de paramètres.*

$$E(Y | X = x) = f(x) \quad (2.5)$$

Le principal avantage (ou inconvénient suivant le point de vue adopté) de cette approche c'est qu'elle ne nécessite aucune hypothèse a priori sur la forme du lien entre X et Y . On a donc une approche a-théorique, encore plus générale que celle développée par exemple dans le cadre des modèles VAR de Sims (1980). Avec une approche non paramétrique, on aboutit à :

1. une **représentation graphique de la relation entre X et Y** .
2. Il n'existe pas de **forme analytique de la fonction de lien $f(x)$** .

Tout le problème consiste alors à estimer cette fonction de lien $f(x)$, qui est a priori inconnue, et non plus uniquement les paramètres de

cette fonction comme c'est le cas dans l'approche paramétrique standard². Pour cela, il existe deux grandes familles de méthodes de régression non paramétriques :

1. La **régression kernel** (Nadaraya, 1964; Watson, 1964)
2. La **régression locale polynomiale** (Cleveland, 1979; Cleveland et Devlin, 1988)

Nous commencerons par présenter le principe de la régression kernel.

3. Régressions Kernel

Le principe de la régression kernel repose en fait sur **des méthodes de lissage**. Afin de bien comprendre le principe d'une régression kernel ou régression par lissage par opérateur à Noyau, nous commencerons par exposer le principe de la régression avec lissage par moyenne mobile. Une fois que l'on aura démontré un certain nombre de résultats dans ce cas simple, nous nous contenterons d'énoncer plusieurs résultats dans le cas de la régression kernel.

3.1. Régression avec lissage par moyenne mobile

Admettons que le "vrai" modèle de l'économie s'écrive sous la forme :

$$y_i = f(x_i) + \varepsilon_i \quad i = 1, \dots, N \quad (3.1)$$

où ε_i est un bruit blanc avec $E(\varepsilon_i) = 0$ et $E(\varepsilon_i^2) = \sigma_\varepsilon^2$. On suppose que la fonction $f(\cdot)$ est inconnue et l'on se propose d'estimer cette fonction par une **méthode de lissage par moyenne mobile (MA)**. L'idée consiste tout simplement à appliquer une moyenne mobile aux valeurs de Y pour obtenir un estimateur de la fonction de lien.

Definition 3.1. *L'estimateur de la fonction de lien par moyenne mobile s'écrit sous la forme suivante :*

$$\hat{f}(x_i) = (\bar{Y}_j)_{x_j \in V_{k,x_i}} \quad (3.2)$$

où $V_{k,x}$ désigne un voisinage de x_i défini par les k individus ayant les valeurs de X les plus proches de x_i .

²On peut en effet faire le reproche aux économistes - économètres d'être prétentieux au point de prétendre connaître le modèle, c'est à dire la forme de la fonction de lien $f(x, \theta)$, et de simplement prétendre ignorer uniquement la valeur des paramètres. Mais c'est précisément tout l'intérêt d'une démarche à la Cowles Commission que de partir de la théorie économique pour aider à la spécification d'un modèle et d'une forme réduite (fonction de lien).

On peut donner une autre définition de cette fonction de lien.

Definition 3.2. *Supposons que les observations x_i sont ordonnées de façon croissante $x_1 \leq x_2 \leq \dots \leq x_N$ et que k est un entier impair, alors :*

$$\widehat{f}(x_i) = \frac{1}{k} \sum_{j=\underline{i}}^{\bar{i}} y_j \tag{3.3}$$

où l'on a $\underline{i} = i - (k - 1) / 2$ et $\bar{i} = i + (k - 1) / 2$.

Considérons l'exemple suivant. On suppose que l'on dispose d'un échantillon de $N = 5$ couples de valeurs (x, y) telles que

$\{(3, 7); (2, 4); (6, 16); (7, 19); (9, 25)\}$. Les valeurs classées sont alors définies de la façon suivante :

x_1	x_2	x_3	x_4	x_5
2	3	6	7	9
y_1	y_2	y_3	y_4	y_5
4	7	16	19	25

Si l'on suppose que la taille de la fenêtre $k = 3$, alors on peut calculer 3 estimations de la fonction $f(x)$ aux points $x = \{3, 6, 7\}$. On obtient ainsi :

$$\widehat{f}(x_2) = \widehat{f}(3) = \frac{1}{3} (4 + 7 + 16) = 9 \tag{3.4}$$

$$\widehat{f}(x_3) = \widehat{f}(6) = \frac{1}{3} (7 + 16 + 19) = 14 \tag{3.5}$$

$$\widehat{f}(x_4) = \widehat{f}(7) = \frac{1}{3} (16 + 19 + 25) = 20 \tag{3.6}$$

De façon générale, l'estimateur MA de la fonction de lien peut s'écrire sous la forme :

$$\widehat{f}(x_i) = \frac{1}{k} \sum_{j=\underline{i}}^{\bar{i}} f(x_j) + \frac{1}{k} \sum_{j=\underline{i}}^{\bar{i}} \varepsilon_j \tag{3.7}$$

Etudions la convergence de cet estimateur. Pour cela, nous étudierons successivement :

1. la **convergence en probabilité** de $\widehat{f}(x_i)$
2. la **convergence en loi** de $\widehat{f}(x_i)$ afin d'en déduire des **intervalles de confiance** sur $f(x_i)$.

3.1.1. Etude de la convergence en probabilité

Considérons l'écriture suivante :

$$\widehat{f}(x_i) = \frac{1}{k} \sum_{j=i}^{\bar{i}} f(x_j) + \frac{1}{k} \sum_{j=i}^{\bar{i}} \varepsilon_j \quad (3.8)$$

Utilisons une décomposition en séries de Taylor à l'ordre 2 de la fonction $f(x_j)$ autour du point de référence x_i . Il vient, $\forall j = 1, \dots, N$:

$$\begin{aligned} f(x_j) &= f(x_i) + f'(x_i)(x_j - x_i) + \frac{f''(x_i)}{2}(x_j - x_i)^2 + o(x_j - x_i)^2 \\ &\simeq f(x_i) + f'(x_i)(x_j - x_i) + \frac{f''(x_i)}{2}(x_j - x_i)^2 \end{aligned} \quad (3.9)$$

On obtient alors :

$$\begin{aligned} \widehat{f}(x_i) &\simeq \frac{1}{k} \sum_{j=i}^{\bar{i}} \left[f(x_i) + f'(x_i)(x_j - x_i) + \frac{f''(x_i)}{2}(x_j - x_i)^2 \right] + \frac{1}{k} \sum_{j=i}^{\bar{i}} \varepsilon_j \\ &\simeq f(x_i) + \frac{f'(x_i)}{k} \sum_{j=i}^{\bar{i}} (x_j - x_i) + \frac{f''(x_i)}{2k} \sum_{j=i}^{\bar{i}} (x_j - x_i)^2 + \frac{1}{k} \sum_{j=i}^{\bar{i}} \varepsilon_j \end{aligned}$$

Or, on sait que si les k valeurs x_j sont choisies de façon symétrique par rapport à la valeur pivotale x_i on a :

$$\sum_{j=i}^{\bar{i}} (x_j - x_i) = 0 \quad (3.10)$$

De plus, on peut montrer que si les x_i sont équi-réparties sur un intervalle unitaire :

$$\frac{1}{2k} \sum_{j=i}^{\bar{i}} (x_j - x_i)^2 = \frac{1}{24} \frac{(k^2 - 1)}{N^2} \simeq \frac{1}{24} \left(\frac{k}{N} \right)^2 \quad (3.11)$$

On en déduit donc finalement que :

$$\widehat{f}(x_i) \simeq f(x_i) + f''(x_i) \frac{1}{24} \left(\frac{k}{N} \right)^2 + \frac{1}{k} \sum_{j=i}^{\bar{i}} \varepsilon_j \quad (3.12)$$

Le dernier terme est une somme de k termes indépendants et indentiquement distribués dès lors, de variance finie σ_ε^2/k , par conséquent on obtient le résultat suivant.

Résultat L'estimateur MA de la fonction de lien $f(x_i) \forall i = 1, \dots, N$, noté $\hat{f}(x_i)$, est tel que :

$$\hat{f}(x_i) = f(x_i) + O\left(\frac{k}{N}\right)^2 + Op\left(\frac{1}{k^{1/2}}\right) \quad (3.13)$$

Par conséquent, l'erreur quadratique moyenne vérifie :

$$E\left[\hat{f}(x_i) - f(x_i)\right]^2 = O\left(\frac{k}{N}\right)^4 + Op\left(\frac{1}{k}\right) \quad (3.14)$$

On sait dès lors que le biais de l'estimateur est défini par :

$$\hat{f}(x_i) - f(x_i) \simeq f''(x_i) \frac{1}{24} \left(\frac{k}{N}\right)^2 + \frac{1}{k} \sum_{j=i}^{\bar{i}} \varepsilon_j \quad (3.15)$$

et que la variance de $\hat{f}(x_i)$ est approximativement égale à :

$$\text{var}\left[\hat{f}(x_i)\right] = \text{var}\left(\frac{1}{k} \sum_{j=i}^{\bar{i}} \varepsilon_j\right) = \frac{1}{k^2} \left[\sum_{j=i}^{\bar{i}} \text{var}(\varepsilon_j)\right] = \frac{k\sigma_\varepsilon^2}{k^2} = \frac{\sigma_\varepsilon^2}{k} \quad (3.16)$$

On en tire la conséquence suivante :

$$\lim_{k/N \rightarrow 0} E\left[\hat{f}(x_i) - f(x_i)\right] = 0 \quad (3.17)$$

$$\lim_{k \rightarrow \infty} \text{Var}\left[\hat{f}(x_i) - f(x_i)\right] = 0 \quad (3.18)$$

De ces deux propriétés, on déduit immédiatement que :

Résultat L'estimateur MA de la fonction de lien $f(x_i) \forall i = 1, \dots, N$, noté $\hat{f}(x_i)$, n'est convergent que si conjointement $k/N \rightarrow 0$ et $k \rightarrow \infty$:

$$\hat{f}(x_i) \xrightarrow{p} f(x_i) \quad \forall i = 1, \dots, N, \quad \frac{k}{N} \rightarrow 0 \quad \text{et} \quad k \rightarrow \infty \quad (3.19)$$

L'estimateur MA n'est donc pas un bon estimateur, puisque que pour qu'il soit convergent il faut à la fois une très large fenêtre et que cette fenêtre ne représente qu'une part infime des observations de l'échantillon. Toute l'idée de la régression kernel consistera à améliorer cette propriété afin d'obtenir un estimateur qui converge dès lors "simplement" que la taille de l'échantillon N est "grande".

3.1.2. Etude de la convergence en distribution et intervalles de confiance

Considérons l'écriture suivante :

$$\hat{f}(x_i) \simeq f(x_i) + f''(x_i) \frac{1}{24} \left(\frac{k}{N}\right)^2 + \frac{1}{k} \sum_{j=i}^{\bar{i}} \varepsilon_j \quad (3.20)$$

Si le nombre de points de la MA, c'est à dire k , augmente avec N , alors par un théorème central limite, on montre d'après le Théorème Centrale Limite (TCL) que le terme de droite est asymptotiquement distribué selon une loi normale de moyenne nulle et de variance finie telle que :

$$\sqrt{k} \left(\frac{1}{k} \sum_{j=i}^{\bar{i}} \varepsilon_j \right) \xrightarrow[N \rightarrow \infty]{D} N(0, \sigma_\varepsilon^2) \quad (3.21)$$

Par conséquent, on en déduit que la quantité :

$$\sqrt{k} \left[\hat{f}(x_i) - f(x_i) - \frac{1}{24} \left(\frac{k}{N}\right)^2 f''(x_i) \right] \quad (3.22)$$

converge asymptotiquement vers une loi normale.

Résultat *L'estimateur MA de la fonction de lien $f(x_i) \forall i = 1, \dots, N$, noté $\hat{f}(x_i)$, vérifie:*

$$\sqrt{k} \left[\hat{f}(x_i) - f(x_i) - \frac{1}{24} \left(\frac{k}{N}\right)^2 f''(x_i) \right] \xrightarrow{d} N(0, \sigma_\varepsilon^2) \quad (3.23)$$

Tout le problème est que **cette propriété ne permet pas de construire un intervalle de confiance sur $f(x_i)$, puisque par définition la quantité $f''(x_i)$ est inconnue**. On doit donc chercher une taille de fenêtre qui croît avec la taille N de l'échantillon et dont la vitesse de convergence "annule" le terme $\left(\frac{k}{N}\right)^2 f''(x_i)$.

Supposons que la taille de la fenêtre vérifie la propriété suivante :

$$k = k(N) = N^\alpha \quad (3.24)$$

Comment fixer la valeur de α de sorte à "annuler" asymptotiquement le terme $\left(\frac{k}{N}\right)^2 f''(x_i)$ qui dépend de la quantité (finie) inconnue $f''(x_i)$? Supposons que l'on fixe $\alpha = 4/5$, c'est à dire que $k = N^{4/5}$ alors

$$\sqrt{k} \left(\frac{k}{N}\right)^2 = N^{\frac{2}{5}} \left(N^{-\frac{1}{5}}\right)^2 = 1 \quad (3.25)$$

La construction d'un IC est alors rendue impossible par la présence du terme $f''(x_i)$. En revanche, si k croît plus lentement que $N^{4/5}$, c'est à dire si $\alpha < 4/5$, alors ce terme s'annule asymptotiquement. Par exemple si $K = N^{3/4}$, alors

$$\lim_{N \rightarrow \infty} \sqrt{k} \left(\frac{k}{N} \right)^2 = \lim_{N \rightarrow \infty} N^{\frac{3}{8}} (N^{-2/4}) = \lim_{N \rightarrow \infty} N^{-\frac{1}{8}} = 0$$

Dans ce cas, on montre alors immédiatement que :

$$\sqrt{k} \left[\hat{f}(x_i) - f(x_i) \right] \xrightarrow[N \rightarrow \infty]{D} N(0, \sigma_\varepsilon^2) \quad \forall i = 1, \dots, N \quad (3.26)$$

L'idée est la suivante : si l'on rajoute des observations, on en retient relativement moins dans la fenêtre pour leur permettre de se concentrer autour du point d'estimation x_i .

Résultat Si la fenêtre k est telle que $\lim kN^{4/5} = 0$, alors l'estimateur MA de la fonction de lien $f(x_i) \forall i = 1, \dots, N$, noté $\hat{f}(x_i)$, vérifie :

$$\sqrt{k} \left[\hat{f}(x_i) - f(x_i) \right] \xrightarrow[N \rightarrow \infty]{D} N(0, \sigma_\varepsilon^2) \quad \forall i = 1, \dots, N \quad (3.27)$$

De ce résultat, on déduit les intervalles de confiance sur $f(x_i)$.

Definition 3.3. Si la fenêtre k est telle que $\lim kN^{4/5} = 0$, un intervalle de confiance au seuil $\alpha\%$ sur la valeur de $f(x_i)$ pour tous les points x_1, x_2, \dots, x_N est donné par :

$$IC_\alpha = \left[\hat{f}(x_i) - C_{1-\alpha/2} \frac{\sigma_\varepsilon}{\sqrt{k}}, \hat{f}(x_i) + C_{1-\alpha/2} \frac{\sigma_\varepsilon}{\sqrt{k}} \right] \quad (3.28)$$

où $C_{1-\alpha/2}$ désigne le fractile de la loi $N(0, 1)$.

Nous allons à présent énoncer directement les résultats concernant la régression kernel qui pour l'essentiel ressemblent, dans l'esprit, à ceux que nous venons de démontrer dans le cas de la régression MA.

3.2. Régression avec lissage par opérateur à noyau ou régression kernel

Comme dans la partie précédente, on cherche à estimer la fonction de lien $f(x_i)$ en tout point x_1, x_2, \dots, x_N . Pour cela nous allons à présent utiliser le lissage par opérateur à noyau ou **kernel smoother** (Nadaraya, 1964 et Watson, 1964).

Definition 3.4. L'estimateur à noyau (kernel estimate) de la fonction de lien évaluée au point x_0 , noté $\hat{f}(x_0)$, est défini par :

$$\hat{f}(x_0) = \sum_{i=1}^N w_i(x_0) y_i \quad (3.29)$$

avec :

$$w_i(x_0) = \frac{K\left(\frac{x_i - x_0}{\lambda}\right)}{\sum_{i=1}^N K\left(\frac{x_i - x_0}{\lambda}\right)} \quad (3.30)$$

où $K(\cdot)$ désigne une fonction kernel, $\lambda > 0$ un paramètre de lissage (bandwidth parameter) et N la taille de l'échantillon utilisé pour l'estimation.

On peut faire ici plusieurs remarques :

Remarque 1 La fonction de lien évaluée au point x_0 est donc définie comme une somme pondérée des observations y_i dont les poids $w_i(x_0)$ dépendent de x_0 .

Remarque 2 La fonction $w_i(x_0)$ ou $w(x_0, x_i)$ définit le poids qui doit être attribué au couple d'observations (x_i, y_i) dans la valeur de la fonction de lien évaluée au point d'abscisse x_0 . Généralement, plus les points x_i sont proches de x_0 , plus le poids sera important : $w(x_0, x_i)$ est donc décroissante dans la distance $|x_0 - x_i|$.

Ces poids dépendent de fonction kernel (ou opérateur à noyau) qui correspondent tout simplement à des fonctions de densité de probabilité.

Definition 3.5. Une fonction kernel $K\left(\frac{x_i - x_0}{\lambda}\right) = K(u)$ vérifient les propriétés suivantes :

- (i) $K(u) \geq 0$
- (ii) $K(u)$ est normalisé de sorte que

$$\int K(u) du = 1 \quad (3.31)$$

(iii) $K(u)$ atteint son maximum en 0 lorsque $x_i = x_0$ et décroît avec la distance $|x_0 - x_i|$.

(iv) $K(u)$ est symétrique : le kernel ne dépende que de la distance $|x_0 - x_i|$ et non du signe de $x_0 - x_i$.

Différentes fonctions kernel peuvent être utilisées :

$$\text{Uniforme : } K(u) = \frac{1}{2} \quad u \in [-1, 1] \quad (3.32)$$

$$\text{Triangulaire : } K(u) = 1 - |u| \quad u \in [-1, 1] \quad (3.33)$$

$$\text{Quartic ou BiWeight : } K(u) = \frac{15}{16} (1 - u^2)^2 \quad u \in [-1, 1] \quad (3.34)$$

$$\text{Epanechnikov : } K(u) = \frac{3}{4} (1 - u^2) \quad u \in [-1, 1] \quad (3.35)$$

$$\text{Triweight : } K(u) = \frac{35}{32} (1 - u^2)^3 \quad u \in [-1, 1] \quad (3.36)$$

$$\text{Normal : } K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) \quad u \in [-\infty, \infty] \quad (3.37)$$

Remarque *On montre qu'en pratique le choix de la fonction kernel n'influence que peu les résultats d'estimation. La seule exception notable étant liée à l'utilisation d'une fonction kernel uniforme qui peut donner des résultats sensiblement différents des autres kernel.*

Enfin, les poids $w_i(x_0)$ dépendent en outre du paramètre de lissage λ qui contrôle l'amplitude des poids.

Remarque *Plus le paramètre de lissage λ (bandwidth parameter) est élevé, plus l'on attribue un poids relativement important aux observations x_i éloignées du point de référence x_0 dans la construction de $f(x_0)$.*

Nous verrons dans la section suivante comment choisir le bandwidth parameter.

3.2.1. Etude de la convergence en probabilité

On admet le résultat suivant :

Proposition 3.6. *L'estimateur à noyau de Nadaraya-Watson est convergent. Si les variables X sont distribuées selon une loi de probabilité de densité $p(x)$, le numérateur converge vers $f(x_0)p(x_0)$ et le dénominateur converge vers $p(x_0)$.*

$$\hat{f}(x_i) \xrightarrow{p} f(x_i) \quad N \rightarrow \infty \quad (3.38)$$

La grande différence avec le cas de la régression MA, c'est que l'estimateur kernel est convergent dès lors que N tend vers l'infini et non plus sous l'hypothèse d'une condition sur la taille de la fenêtre.

3.2.2. Etude de la convergence en loi et intervalles de confiance

De la même façon, on peut étudier la distribution de $\hat{f}(x_i)$ pour construire un IC sur $f(x_i)$. On admet le résultat suivant :

Proposition 3.7. *L'estimateur à noyau de Nadaraya-Watson vérifie*

$$\sqrt{\lambda}\sqrt{N} \left\{ \hat{f}(x_0) - f(x_0) - \frac{1}{2}a_K\lambda^2 \left[f''(x_0) + 2f'(x_0) \frac{p'(x_0)}{p(x_0)} \right] \right\} \xrightarrow{d} N \left(0, \frac{b_K\sigma_\varepsilon^2}{p(x_0)} \right)$$

où $p(\cdot)$ désigne la densité de x et

$$a_K = \int u^2 K(u) du \quad b_K = \int K(u)^2 du \quad (3.39)$$

Voir Wand et Jones (1995) pour les valeurs de a_K et b_K pour de nombreux kernels. On admettra en particulier que :

$$\begin{aligned} \text{Uniforme} : b_K &= \frac{1}{2} \\ \text{Triangulaire} : b_K &= \frac{2}{3} \\ \text{Quartic ou BiWeight} : b_K &= \frac{5}{7} \\ \text{Epanechnikov} : b_K &= \frac{3}{5} \\ \text{Triweight} : b_K &= \frac{350}{429} \\ \text{Normal} : b_K &= \frac{1}{2\sqrt{\pi}} \end{aligned} \quad (3.40)$$

Comme dans le cas MA, des simplifications peuvent être apportées si la valeur de λ décroît avec N plus rapidement que $\lambda = N^{-1/5}$. Dans ce cas, le terme de biais disparaît et donc on obtient le résultat suivant dans le cas d'un kernel uniforme ($b_K = 1/2$) :

$$\sqrt{\lambda}\sqrt{N} \left[\hat{f}(x_0) - f(x_0) \right] \xrightarrow{d} N \left(0, \frac{\sigma_\varepsilon^2}{2p(x_0)} \right) \quad (3.41)$$

On peut donc en déduire la manière de construire des IC sur les valeurs de $f(x_i)$.

Proposition 3.8. *Sous l'hypothèse que λ vérifie $N^{1/5}\lambda \rightarrow 0$, l'écart type de l'estimateur à noyau de Nadaraya-Watson $\hat{f}(x_0)$ vérifie*

$$\sqrt{\lambda}\sqrt{N} \left[\hat{f}(x_0) - f(x_0) \right] \xrightarrow{d} N \left(0, \frac{b_K \sigma_\varepsilon^2}{p(x_0)} \right)$$

où $p(\cdot)$ désigne la densité de x et

$$b_K = \int K(u)^2 du \quad (3.42)$$

Un intervalle de confiance sur $\hat{f}(x_0)$ au seuil de $\alpha\%$ est donc défini par

$$IC_\alpha = \left[\hat{f}(x_0) - C_{1-\alpha/2} s_{\hat{f}(x_0)}, \hat{f}(x_0) + C_{1-\alpha/2} s_{\hat{f}(x_0)} \right] \quad (3.43)$$

où $C_{1-\alpha/2}$ désigne le fractile de la loi $N(0, 1)$ et où

$$s_{\hat{f}(x_0)} = \sqrt{\frac{b_K \hat{\sigma}_\varepsilon^2}{\hat{p}(x_0)}} \quad (3.44)$$

avec

$$\hat{p}(x_0) = K \left(\frac{x_i - x_0}{\lambda} \right)$$

La procédure pour obtenir les IC est donc la suivante :

1. On choisit λ vérifie $N^{1/5}\lambda \rightarrow 0$ et une fonction kernel $K(u)$, d'où l'on déduit b_K .
2. On construit l'estimateur à noyau de Nadaraya-Watson $\hat{f}(x_0)$. On recommence pour toutes les valeurs x_1, \dots, x_N .
3. On calcule l'estimateur de la variance des résidus :

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N} \sum_{i=1}^N \left[y_i - \hat{f}(x_i) \right]^2$$

4. On estime la valeur de $p(x_0)$, densité de X au point x_0 (cf. section sur l'estimation des densités) par :

$$\hat{p}(x_0) = K \left(\frac{x_i - x_0}{\lambda} \right) \quad (3.45)$$

5. On calcule l'intervalle sur $\hat{f}(x_0)$ au seuil α défini par :

$$IC_\alpha = \left[\hat{f}(x_0) - C_{1-\alpha/2} \sqrt{\frac{b_K \hat{\sigma}_\varepsilon^2}{\hat{p}(x_0)}}, \hat{f}(x_0) + C_{1-\alpha/2} \sqrt{\frac{b_K \hat{\sigma}_\varepsilon^2}{\hat{p}(x_0)}} \right] \quad (3.46)$$

On recommence alors toute la procédure pour toutes les valeurs disponibles de X , x_1, \dots, x_N .

3.3. Sélection du paramètre de lissage dans la régression Kernel

Comment choisir le paramètre de lissage λ dans le cadre d'une régression kernel ? C'est sans doute le point le plus important de ce type de méthodes. Rappelons que pour certaines fonctions Kernel, les points x_i qui sont distants de plus de λ du point de référence x_0 ne sont pas pris en compte dans le calcul de $f(x_0)$.

Exemple : dans le cas d'une fonction kernel Epanechnikov, on a

$$K(u) = \begin{cases} \frac{3}{4}(1-u^2) & \text{si } u \in [-1, 1] \\ 0 & \text{sinon} \end{cases} \quad (3.47)$$

avec $u = (x_i - x_0) \setminus \lambda$. Donc si $|x_i - x_0| > \lambda$, alors $u \notin [-1, 1]$, $K(u) = 0$ et par conséquent $w_i(x_0) = 0$.

Pour les autres kernels (gaussien par exemple), le paramètre λ représente la distance au delà de laquelle les observations x_i ont un poids négligeable dans la quantité $w_i(x_0)$.

Remarque De façon générale, λ représente le radius de la fenêtre de valeurs x_i autour de x_0 prises en compte dans le calcul de $m(x_0) = f(x_0)$

Cette fenêtre a donc une amplitude 2λ .

De façon générale, il convient de retenir le principe suivant :

Proposition 3.9. *Le choix du bandwidth parameter λ correspond à un arbitrage variance / biais :*

(i) *Plus λ est élevé, plus la courbe $\hat{f}(x)$ sera lisse. La variance de l'estimation est limitée, mais l'estimateur $f(x)$ peut être fortement biaisé.*

(ii) *Plus λ est faible, plus la courbe $\hat{f}(x)$ est irrégulière. Les biais d'estimation de $f(x)$ sont faibles, mais la variance de $\hat{f}(x)$ est très importante.*

Le choix de λ résulte donc d'un arbitrage biais versus variance, mais aussi d'un arbitrage lissage / non lissage de $f(x)$.

Exemple 1 : supposons que l'on choisisse λ tel que $\lambda \rightarrow \infty$. Dès lors, on a :

$$\lim_{\lambda \rightarrow \infty} K\left(\frac{x_i - x_0}{\lambda}\right) = K(0) \quad \forall x_i \quad (3.48)$$

Ceci implique que les poids de tous les individus i dans le calcul de $\hat{f}(x_0)$ sont strictement identiques et égaux à :

$$\lim_{\lambda \rightarrow \infty} w_i(x_0) = \lim_{\lambda \rightarrow \infty} \frac{K\left(\frac{x_i - x_0}{\lambda}\right)}{\sum_{i=1}^N K\left(\frac{x_i - x_0}{\lambda}\right)} = \frac{K(0)}{N K(0)} = \frac{1}{N} \quad (3.49)$$

Ainsi, l'estimateur de $\hat{f}(x_0)$ est défini par :

$$\lim_{\lambda \rightarrow \infty} \hat{f}(x_0) = \lim_{\lambda \rightarrow \infty} \sum_{i=1}^N w_i(x_0) y_i = \frac{1}{N} \sum_{i=1}^N y_i = \bar{y} \quad (3.50)$$

Ainsi si le paramètre de lissage tend vers l'infini, pour tous les points de l'échantillon, l'estimateur kernel correspond à la moyenne empirique \bar{y} . La fonction de lien estimée correspond à une droite horizontale : la variance de $\hat{f}(x)$ est nulle, mais le biais est sans doute fort.

Exemple 2 : supposons au contraire que l'on choisisse λ tel que $\lambda \rightarrow 0$. Dès lors, pour tous les points x_j différents du point x_i de référence :

$$\lim_{\lambda \rightarrow 0} K\left(\frac{x_j - x_i}{\lambda}\right) = K(\pm\infty) = 0 \quad \forall_j \neq i \quad (3.51)$$

En revanche, pour le point de référence x_i on a :

$$K\left(\frac{x_i - x_i}{\lambda}\right) = K(0) \quad \forall \lambda \quad (3.52)$$

Dès lors, pour tous les individus autres que l'individu de référence dans le calcul de $\hat{f}(x_i)$, les poids $w_j(x_i)$ sont nuls :

$$\lim_{\lambda \rightarrow 0} w_j(x_i) = \lim_{\lambda \rightarrow 0} \frac{K\left(\frac{x_j - x_i}{\lambda}\right)}{\sum_{j=1}^N K\left(\frac{x_j - x_i}{\lambda}\right)} = 0 \quad \forall_j \neq i \quad (3.53)$$

En revanche, le poids de l'individu de référence x_i vérifie :

$$\lim_{\lambda \rightarrow 0} w_i(x_i) = \lim_{\lambda \rightarrow 0} \frac{K\left(\frac{x_i - x_i}{\lambda}\right)}{\sum_{j=1}^N K\left(\frac{x_j - x_i}{\lambda}\right)} \quad (3.54)$$

$$= \lim_{\lambda \rightarrow 0} \frac{K(0)}{\sum_{j \neq i} K\left(\frac{x_j - x_i}{\lambda}\right) + K\left(\frac{x_i - x_i}{\lambda}\right)} \quad (3.55)$$

$$= \frac{K(0)}{K(0)} = 1 \quad (3.56)$$

Ainsi, l'estimateur de $\hat{f}(x_0)$ est défini par :

$$\lim_{\lambda \rightarrow 0} \hat{f}(x_i) = \lim_{\lambda \rightarrow 0} \sum_{j \neq i} w_j(x_i) y_j + \lim_{\lambda \rightarrow 0} w_i(x_i) y_i = y_i \quad (3.57)$$

Ainsi si le paramètre de lissage tend vers zéro, pour tous les points de l'échantillon, l'estimateur kernel correspond exactement à l'observation y_i . La fonction de lien estimée passe exactement par tous les points de l'échantillon : la variance de $\hat{f}(x)$ est aussi grande que la variance de y , mais le biais est sans faille.

Toute la question est comment choisir une valeur optimale du paramètre de lissage permettant d'arbitrer au mieux entre variance et biais.

3.3.1. Critère de la MISE et approche GCV

Il existe des procédures numériques de choix d'un λ optimal. La première méthode consiste à choisir λ de sorte à minimiser la MISE (Mean Integrated Squared Error). C'est la définition même du paramètre de lissage optimal.

Definition 3.10. La MISE (Mean Integrated Squared Error) associée à un paramètre de lissage λ , correspond à la quantité :

$$MISE(\lambda) = E \left\{ \int_x [\hat{f}(x, \lambda) - f(x)]^2 dx \right\} \quad (3.58)$$

SAS considère une autre expression de la MISE faisant intervenir la variance de l'estimateur :

$$MISE(\lambda) = \int_x E [\hat{f}(x, \lambda) - f(x)]^2 dx + \int_x Var [\hat{f}(x, \lambda)] dx \quad (3.59)$$

La MISE correspond ainsi à la somme de l'intégrale des biais au carré et de la variance de l'estimateur $\hat{f}(x, \lambda)$. Dans l'absolu on cherche la valeur optimale de λ telle que :

$$\lambda_{MISE}^* = \underset{\{\lambda \in \mathbb{R}^{*+}\}}{ArgMin} MISE(\lambda)$$

Le problème c'est que l'on ne connaît pas la quantité $f(x)$ et que l'on ne peut donc directement évaluer cette MISE. Donc on utilise une approche qui asymptotiquement nous donne une valeur proche de λ_{MISE}^* : l'approche de la cross-validation function ou General Cross-Validation (GCV).

On peut faire l'analogie avec la méthode simple qui consisterait à déterminer la valeur de λ qui minimiserait la variance estimée des résidus.

$$\hat{\sigma}_\varepsilon^2(\lambda) = \frac{1}{N} \sum_{i=1}^N [y_i - \hat{f}(x, \lambda)]^2$$

Ce critère nous permettrait d'obtenir la valeur de λ telle que les données sont parfaitement ajustées. En effet, si l'on cherche :

$$\tilde{\lambda} = \underset{\{\lambda \in \mathbb{R}^{**}\}}{\text{ArgMin}} \widehat{\sigma}_\varepsilon^2(\lambda) \quad (3.60)$$

on va alors aboutir au résultat $\tilde{\lambda} \rightarrow 0$, puisque si le paramètre de lissage tend vers 0, alors nous avons vu que $y_i = \widehat{f}(x, \lambda)$ et donc $\widehat{\sigma}_\varepsilon^2(0) = 0$. Ce critère est a priori sans intérêt, mais on peut considérer une légère variation connue sous le nom de **cross validation function**.

Definition 3.11. *La Cross Validation Function est définie par la quantité :*

$$CV(\lambda) = \frac{1}{N} \sum_{i=1}^N \left[y_i - \widehat{f}_{-1}(x, \lambda) \right]^2$$

La seule différence avec le critère précédent réside dans l'utilisation de l'indice \widehat{f}_{-1} . Cet indice signifie que pour chaque $i = 1, \dots, N$, la valeur de $f(x_i)$ est obtenue en enlevant la $i^{\text{ème}}$ observation x_i . Le modèle est estimé sur toutes les autres observations x_j , $j \neq i$, puis on estime la valeur de $f(\cdot)$ au point x_i à partir de cette régression. C'est cette valeur estimée (out of sample) qui figure dans la formule $CV(\lambda)$ sous la notation $\widehat{f}_{-1}(x, \lambda)$. Pour chaque valeur de λ , $CV(\lambda)$ requiert l'estimation de N kernel.

Proposition 3.12. *Soit λ_{CV}^* la valeur de λ telle que :*

$$\lambda_{CV}^* = \underset{\{\lambda \in \mathbb{R}^{**}\}}{\text{ArgMin}} CV(\lambda) \quad (3.61)$$

On peut montrer que :

$$MISE(\lambda_{CV}^*) \xrightarrow[N \rightarrow \infty]{P} MISE(\lambda_{MISE}^*) \quad (3.62)$$

L'utilisation de la fonction CV permet ainsi d'obtenir un estimateur du paramètre optimal λ_{MISE}^* . Sous SAS, on peut utiliser cette valeur en précisant **l'option C=MISE**.

3.3.2. Critère de l'AMISE

Enfin, signalons que SAS utilise par défaut un autre critère pour dériver la valeur optimale du bandwidth parameter : il s'agit de l'AMISE.

Definition 3.13. L'AMISE (Approximate Mean Integrated Squared Error) associé à un paramètre de lissage λ , correspond à la quantité :

$$AMISE(\lambda) = \frac{1}{4}\lambda^4 \left(\int_u u^2 K(u) du \right)^2 \left[\int_x f''(x)^2 dx \right] + \frac{1}{N\lambda} \int_u (K(u))^2 du \tag{3.63}$$

ou de façon équivalent dans nos notations :

$$AMISE(\lambda) = \frac{1}{4}\lambda^4 a_K^2 \int_x f''(x)^2 dx + \frac{b_K}{N\lambda} \tag{3.64}$$

On reconnaît dans le terme de gauche un indicateur du lissage de la fonction estimée et dans le terme de droite un indicateur de la variance des résidus (cf. distribution asymptotique de l'estimateur). Encore une fois, cette quantité dépend de la fonction $f(\cdot)$ inconnue au travers de $f''(x)$. Mais on admet le résultat suivant :

$$\lambda^* = \underset{\{\lambda \in \mathbb{R}^{**+}\}}{ArgMin} AMISE(\lambda) \simeq \lambda_{MISE}^* \tag{3.65}$$

Nous verrons trois méthodes permettant de calculer l'AMISE et donc d'en déduire un paramètre de lissage optimal.

SAS ne permet pas de contrôler directement le bandwidth parameter λ , mais une constante C définie de la façon suivante.

Definition 3.14. La donnée du paramètre de lissage λ (bandwidth parameter), est équivalente à la donnée du paramètre lissé C (standardized bandwidth parameter) tel que :

$$\lambda = CQN^{-\frac{1}{5}} \iff C = \frac{\lambda}{Q}N^{\frac{1}{5}} \tag{3.66}$$

où $Q = Q_3 - Q_1$ désigne l'amplitude de l'interquartile (interquartile range).

Cette formulation de C permet de rendre la valeur du paramètre de lissage indépendante de l'unité de X .

3.4. Estimation d'une fonction de densité

Naturellement le même type de méthode peut être utilisé pour estimer une fonction de densité à partir d'un N-échantillon de réalisation. Soit $f(x)$ la fonction de densité associée à la variable aléatoire X . Soit $\{x_i\}_{i=1}^N$ un échantillon de taille N d'observations de cette variable X . On pourrait tout d'abord penser à estimer cette densité par un histogramme avec des classes très fines. Considérons

Figure 3.1: Simulation et Histogramme

```

data donnees;
  seed = 1283470;
  do i = 1 to 50000;
    x = rannor(seed);
    t=i;
    output;
  end;
  output;
  drop seed;
run;

proc univariate data=donnees;
  var x;
  histogram x /normal(noprint) cbarline=grey ;
run;

```

l'exemple suivant dans lequel on simule 50000 réalisations d'une variable aléatoire de loi $N(0, 1)$.

Dans la procédure UNIVARIATE qui nous permet de grapher l'histogramme, on sur-impose sur le graphique la fonction de densité d'une loi $N(0, 1)$. On observe que pour un découpage fin des classes, les sommets de classes peuvent constituer des estimateurs des valeurs de la densité $f(x)$ aux points correspondants.

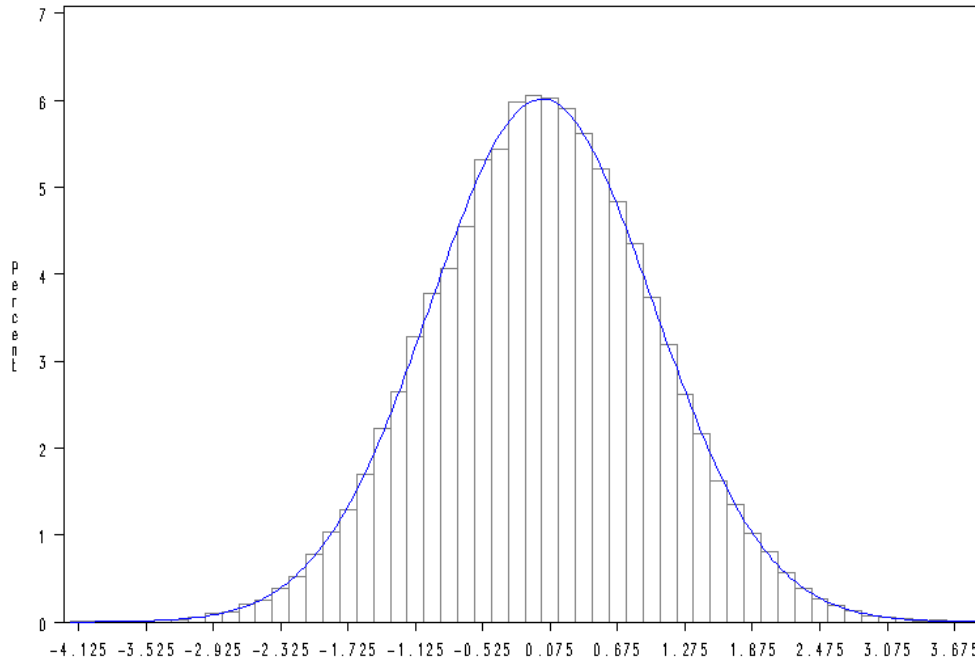
Toutefois, on préférera utiliser un estimateur de $f(\cdot)$ au point x_0 obtenu par lissage des sommets de classe définis au voisinage de x_0 . Voilà quel est le principe de l'estimation par lissage. L'estimateur à noyau (*kernel estimate*) ne conduit alors qu'à définir une forme des poids à accorder aux différents valeurs des sommets de classes obtenus pour des valeurs x_i , en fonction de la distance entre le point de référence x_0 et le point considéré dans le lissage x_i .

Definition 3.15. *L'estimateur à noyau (kernel estimate) de la fonction de densité de la variable X évaluée au point x_0 , noté $\hat{f}(x_0)$, est défini par :*

$$\hat{f}(x_0) = \sum_{i=1}^N K\left(\frac{x_i - x_0}{\lambda}\right) \quad (3.67)$$

où $K(\cdot)$ désigne une fonction kernel, λ un paramètre de lissage (*bandwidth parameter*) et N la taille de l'échantillon utilisé pour l'estimation.

Les règles pour déterminer la valeur du paramètre λ sont alors les mêmes que celles évoquées précédemment. Comme précédemment, le choix du kernel n'a que peu d'importance (cf. section précédente).

Figure 3.2: Histogramme 50000 Simulations d'une loi $N(0, 1)$ 

3.4.1. Procédure UNIVARIATE

La première façon d'obtenir l'estimateur à noyau de la densité de X aux points x_1, x_2, \dots, x_N consiste à utiliser la procédure UNIVARIATE avec l'option HISTOGRAM. Pour cela, on spécifie la syntaxe suivante :

```
HISTOGRAM [Nom Variable] K=NORMAL | QUADRATIC | TRIANGULAR
```

Le fonction kernel par défaut est la fonction normale. La procédure SAS permet alors de contrôler la valeur du bandwidth parameter λ au travers de la valeur de la constante C , (standardized bandwidth parameter) définie par :

$$C = \frac{\lambda}{Q} N^{\frac{1}{5}} \quad (3.68)$$

où Q désigne l'amplitude de l'interquartile (interquartile range). Trois solutions sont alors possibles :

1. On peut fixer une valeur pour C (et donc λ) Exemple : $C=3$.
2. On peut retenir la valeur de C qui minimise le critère $AMISE(\lambda)$ (par défaut) ou le critère $MISE(\lambda)$ en posant $C=MISE$.

3. On peut spécifier plusieurs valeurs de C : Exemple : $C=3$ 2 5 ou $C=3$ MISE 5
4. On peut utiliser différentes valeurs de C associées à différents kernel. Si l'on fixe plus de valeurs de C que de choix de kernel les valeurs de C en excès sont alors utilisées avec la dernière fonction kernel spécifiée. Exemple : `kernel(c=1 2 3 k=normal quadratic)`. The first uses a normal kernel and a bandwidth of 1, the second uses a quadratic kernel and a bandwidth of 2, and the third uses a quadratic kernel and a bandwidth of 3.

Exemple 1 : Estimation Kernel d'une densité d'une variable $N(0, 1)$

On estime la densité de la variable X à partir d'un kernel Quadratique. On utilise un paramètre de lissage standardisé C dérivé de l'optimisation du MISE et un autre éagl à 18. On observe sur le graphique de résultats que pour $C = 18$, la densité estimée est très margement éloignée de la vraie densité de la loi normale (courbe en bleue) ce qui met en lumière l'importance de ce paramètre de lissage dans l'estimation de $f(x)$.

Figure 3.3: Estimation Kernel d'une Densité

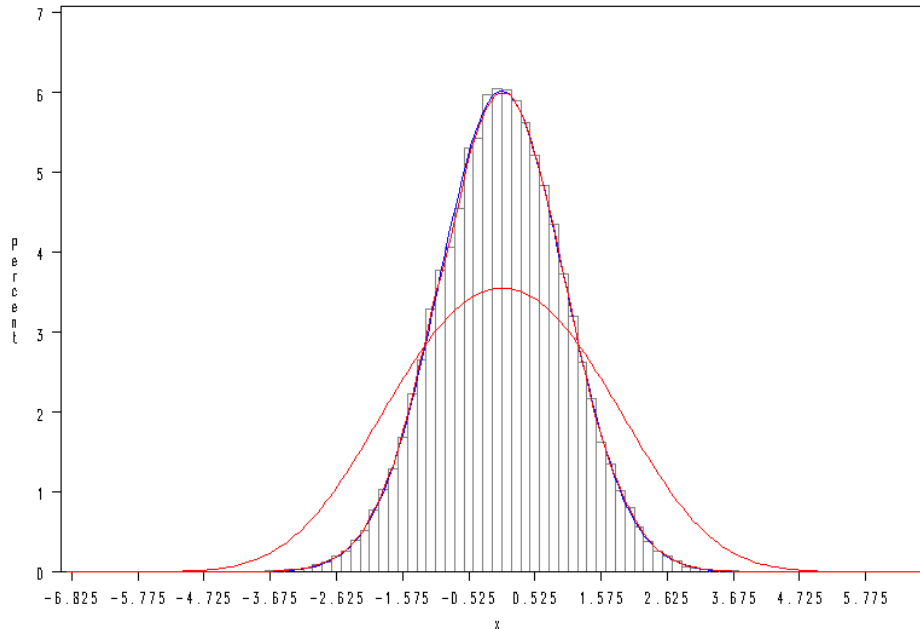
```
proc univariate data=donnees;
  var x;
  histogram x /normal(noprint) cbarline=grey kernel(c = MISE 18 k = Quadratic COLOR=red ) ;
run;
```

Exemple 2 : Estimation Kernel de la distribution des rendements de l'indice SP500

On reprend l'exemple du cours d'Econométrie pour la Finance du rendement quotidien du SP500 sur la période du 03/07/1989 au 24/11/2003 :

$$r_t = \log(p_t) - \log(p_{t-1}) = \log(1 + R_t) \quad (3.69)$$

où $R_t = (p_t - p_{t-1})/p_t$ désigne la variation relative des prix. Le programme est alors le suivant pour une fonction kernel de type Normal et un bandwidth optimal au sens du critère de l'AMISE. On vérifie sur le graphique (3.6) que l'estimateur "optimal" au sens de le l'AMISE (courbe rouge) ne correspond pas du tout à la densité d'une loi normale, ce qui confirme le rejet largement admis de l'hypothèse d'une distribution normale des rendements financiers. On observe notamment des effets leptokurtiques à partir de l'estimateur à noyau, même si il convient de se méfier de la précision des estimateurs kernel concernant les queues de distribution.

Figure 3.4: Résultats d'Estimation Kernel d'une Fonction de Densité $N(0, 1)$ 

3.4.2. Procédure KDE

Les deux principaux inconvénients de la procédure UNIVARIATE sont les suivants : d'une part elle ne permet pas d'évaluer les intervalles de confiance sur les estimateurs de la densité, d'autre part elle ne permet pas d'obtenir des valeurs estimées pour les différentes fractiles qui peuvent être utiles par exemple dans le cadre d'une application VaR. La procédure **KDE (Kernel Density Estimate)** permet de palier à ces insuffisances. Elle permet de faire une estimation d'une fonction de densité uniquement à partir d'une **fonction kernel de type normale** :

$$\hat{f}(x_0) = \frac{1}{\sqrt{2\pi}} \sum_{i=1}^N \exp -\frac{1}{2} \left(\frac{x_i - x_0}{\lambda} \right)^2 \quad (3.70)$$

Trois méthodes de détermination du bandwidth parameter sont proposées grâce à l'option METHOD.

- **METHOD=SJPI**, Sheather-Jones Plug In
- **METHOD=SNR**, Simple Normal Reference
- **METHOD=SROT**, Silverman's rule of thumb
- **METHOD=OS**, OverSmoothed

Figure 3.5: Estimation Kernel de la Distribution des Rendements du SP500

```

data donneesSP;
  infile 'C:\Chris\Cours\Econometrie_Finance\ApplisAS\bonds2.csv' dlm=';' ;
  input tc t1 cac sp;
  lsp = log(sp);
  dlsp=dif(lsp);
  t+1;
  keep dlsp lsp t;
run;

proc univariate data=donneesSP;
  var dlsp;
  histogram dlsp /normal(noprint) cbarline=grey kernel(COLOR=red ) ;
run;

```

La méthode Sheather-Jones plug in (SJPI) est la méthode par défaut pour des densités de variables univariées. Dans les cas 4, il s'agit de déterminer la valeur de λ qui permet de minimiser le critère $AMISE(\lambda)$:

$$\lambda^* = \underset{\{\lambda \in \mathbb{R}^{**}\}}{ArgMin} AMISE(\lambda) \quad (3.71)$$

Pour le cas d'un Kernel normal, la valeur de λ^* sous l'hypothèse $h \rightarrow 0$ et $Nh \rightarrow \infty$, est définie par :

$$\lambda^* = \left[\frac{1}{2\sqrt{2\pi}N \int_x (f'')^2 dx} \right]^{1/5} \quad (3.72)$$

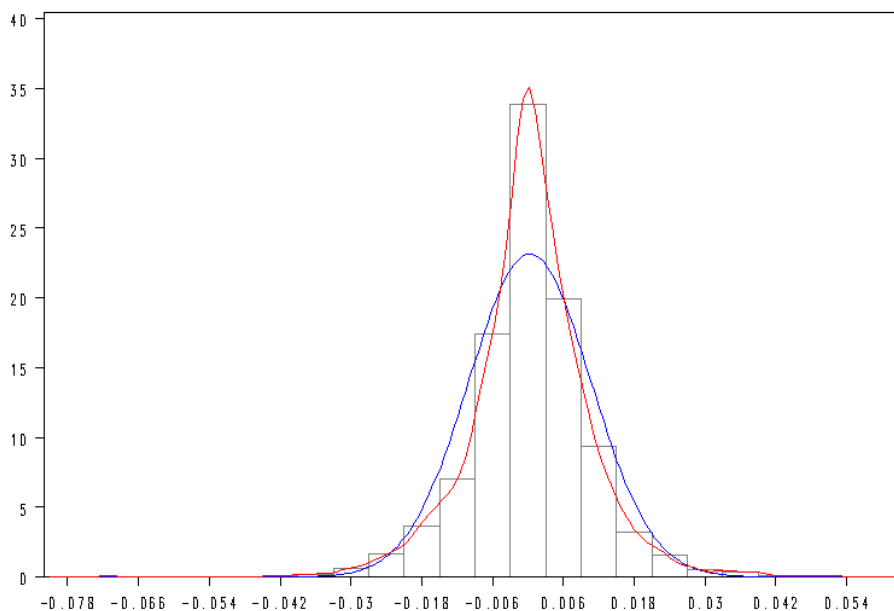
Naturellement, cette valeur est inconnue puisque la quantité $\int_x (f'')^2 dx$ est elle même inconnue. Pour approximer cette valeur optimale par rapport au critère de l'AMISE, une méthode consiste à déterminer un point fixe de l'application :

$$\lambda = \left(\frac{\int_x \phi(x) dx}{N \left(\int_x (\hat{f}'')^2 dx \right) \left(\int_x x^2 \phi(x) dx \right)^2} \right)^{1/5} \quad (3.73)$$

où $\phi(x)$ désigne la fonction de densité de la loi normale. C'est la **méthode Sheather-Jones Plug In** recommandée notamment par Jones, Marron et Sheater (1996). La méthode dite Simple Normal Reference consiste tout simplement à déterminer $f''(x)$ dans la formule (3.72) du λ^* optimal en postulant que la densité $f(x)$ correspond à la densité normale. On obtient alors :

$$\lambda_{SNR}^* = \hat{\sigma} \left(\frac{4}{3N} \right)^{1/5} \quad (3.74)$$

Figure 3.6: Estimation Kernel de la Distribution des Rendements du SP500



où $\hat{\sigma}$ désigne l'écart type de l'échantillon. La règle de Silverman, utilisée par défaut dans d'autres logiciels (comme Eviews 4.0) détermine le bandwidth parameter de la façon suivante :

$$\lambda_{SROT}^* = \frac{0.9}{N^{1/5}} \min \left[\hat{\sigma}, \left(\frac{Q_3 - Q_1}{1.34} \right) \right] \quad (3.75)$$

où Q_3 et Q_1 désignent les troisième et premier quartiles. Enfin, la méthode dite OverSmoothed, pose que :

$$\lambda_{OS}^* = 3\hat{\sigma} \left[\frac{1}{70N\sqrt{\pi}} \right]^{1/5} \quad (3.76)$$

Si l'on souhaite ajuster ces valeurs optimales, on peut utiliser un facteur multiplicatif (valeur par défaut égale à 1) en utilisant l'option **BMW= Valeur**.

$$\lambda = \lambda^* * BMW$$

La syntaxe générale de cette procédure est la suivante :

PROC KDE DATA=[Nom de Fichier Entrée] OUT=[Nom du Fichier de Sortie];

BY variables ;

FREQ variable ;

VAR variables ;

WEIGHT variable ;

Les options disponibles pour KDE sont les suivantes :

- **GRIDL**=numlist specifies the lower grid limits for the kernel density estimate. You should specify one number for univariate smoothing and two numbers separated by a comma for bivariate smoothing. The default values equal the minimum observed values of the variables.
- **GRIDU**=numlist specifies the upper grid limits for the kernel density estimate. You should specify one number for univariate smoothing and two numbers separated by a comma for bivariate smoothing. The default values equal the maximum observed values of the variables.
- **NGRID**=numlist ou **NG**=numlist specifies the number of grid points associated with the variables in the VAR statement. You should specify one number for univariate smoothing and two numbers separated by a comma for bivariate smoothing. The default values are 401 when there is a single VAR variable and 60 when there are two VAR variables.
- **PERCENTILES**=numlist lists percentiles to be computed for each VAR variable. The default percentiles are 0.5, 1, 2.5, 5, 10, 25, 50, 75, 90, 95, 97.5, 99, and 99.5.
- **OUT**=SAS-data-set specifies the output SAS data set containing the kernel density estimate. This output data set contains the following variables: variables you specify in the VAR statement, with values corresponding to grid coordinates, **density**, with values equal to kernel density estimates at the associated grid point and **count**, containing the number of original observations contained in the bin corresponding to a grid point

Exemple 1 : Estimation d'une densité d'un échantillon de variables $N(0, 1)$

Le programme suivant simule 50000 réalisations d'une variable aléatoire tirée dans une loi normale $N(0, 1)$ et estime la fonction de densité de cette variable à partir de cet échantillon. Le graphique de la densité empirique est alors reporté.

Le fichier de résultat comporte tout d'abord les informations reportées sur le graphique (3.8). On observe que par défaut la procédure KDE utilise un estimateur Kernel de type normal avec une méthode de sélection du paramètre de lissage λ de type Sheather-Jones Plug In. La densité sera évaluée par défaut

Figure 3.7: Utilisation de la Procédure KDE

```

data donnees;
  seed = 1283470;
  do i = 1 to 50000;
    x = rannor(seed);
    output;
  end;
  output;
  drop seed;
run;

proc kde data=donnees out=sortie ;
  var x;
run;

proc gplot data=sortie;
  symbol1 i=join;
  plot density*x=1;
run;

```

sur 401 points uniformément répartis entre les bornes du tirage, à savoir -4.097 et 3.8659 . La valeur du coefficient multiplicatif (égale à l'unité) implique que l'on adopte la valeur par défaut retenue selon la méthode de sélection du poaramère de lissage.

La suite des résultats est reportée sur la figure (3.9). On y trouve la liste des valeurs des fractiles pour les valeurs par défaut à savoir 0.5, 1, 2.5, 5, 10, 25, 50, 75, 90, 95, 97.5, 99, and 99.5. On vérifie que dans notre exemple, le fractile à 2.5% correspond au fractile théorique de la loi normale, à savoir -1.95 . On vérifie sur le graphique (3.10) que la densité estimée est identique à celle d'une loi normale.

3.4.3. Procédure SAS INSIGHT

Pour estimer une fonction de densité, on peut en outre utiliser la procédure SAS INSIGHT qui permet d'adopter un environnement convivial (au regard des procédures SAS...) de programmation. La commande est simple est la suivante

PROC SAS INSIGHT DATA=[Nom du Fichier]; RUN;

Appliquons cette procédure aux données tirées dans une loi normale de l'exemple précédent. On obtient alors, la sortie de la figure (??). Dans le menu ANALYZE, on choisit alors l'onglet DISTRIBUTION. Puis apparaît un écart dans lequel on choisit l'option METHOD, dans lequel on choisit DENSITY ESTIMATION.

On peut alors choisir la fonction kernel. Dans tous les cas le paramètre de lissage optimale est obtenu par la méthode de type AMISE (procédure par défaut

Figure 3.8: Résultats de la Procédure KDE : 1ère Partie

The KDE Procedure

Inputs

Data Set	WORK.DONNEES
Number of Observations Used	50001
Variable	x
Bandwidth Method	Sheather-Jones Plug In

Controls

	x
Grid Points	401
Lower Grid Limit	-4.097
Upper Grid Limit	3.8659
Bandwidth Multiplier	1

Statistics

	x
Mean	0.0018
Variance	0.99
Standard Deviation	0.99
Range	7.96
Interquartile Range	1.34
Bandwidth	0.12

de la procédure KDE). On a donc plus de choix au niveau du kernel, mais moins de choix a priori sur la méthode de détermination du bandwidth parameter. On obtient alors la valeur précise des paramètres de lissage λ et C , ainsi que les quartiles Q_3 et Q_1 , qui permettent de calculer $Q = Q_3 - Q_1$ et ainsi de passer de λ à C selon la formule :

$$\lambda = CQN^{-\frac{1}{5}} \iff C = \frac{\lambda}{Q}N^{\frac{1}{5}} \tag{3.77}$$

où $Q = Q_3 - Q_1$ désigne l'amplitude de l'interquartile (interquartile range). Ainsi dans le cas de la Kernel triangulaire le paramètre de lissage optimal au sens de l'AMLISE λ^* est égal à 0.2931 ce qui correspond à un paramètre de lissage standardisé $C^* = 1.9096$, selon la formule :

$$C = \frac{\lambda}{Q}N^{\frac{1}{5}} = \frac{0.2931}{1.3360} * 50001^{\frac{1}{5}} = 1.9906 \tag{3.78}$$

car la différence des quartiles $Q_3 - Q_1$ est dans cet exemple égale à 1.3360 et $N = 50001$ comme l'indique le tabelau en bas à droite de la sortie SAS INSIGHT (figure 3.13).

Figure 3.9: Résultats Procedure KDE (Partie II)

Percentiles			
x			
0.5			-2.57
1.0			-2.32
2.5			-1.95
5.0			-1.63
10.0			-1.28
25.0			-0.66
50.0	0.0045		
75.0			0.67
90.0			1.27
95.0			1.64
97.5			1.96
99.0			2.31
99.5			2.57

Levels			
Percent	Density	Lower	Upper
1	0.01466	-2.56	2.59
5	0.06146	-1.95	1.95
10	0.1037	-1.65	1.64
50	0.3150	-0.65	0.70
90	0.3971	-0.12	0.14
95	0.3989	-0.076	0.064
99	0.3998	-0.036	0.0039
100	0.3998	-0.016	-0.016

3.5. Regressions Kernel : Applications SAS INSIGHTH

Les applications sous SAS diffèrent suivant que l'on souhaite faire une régression ou estimer une densité. Si l'on souhaite effectuer une regression Kernel, on peut aussi utiliser la procédure SAS INSIGHTH :

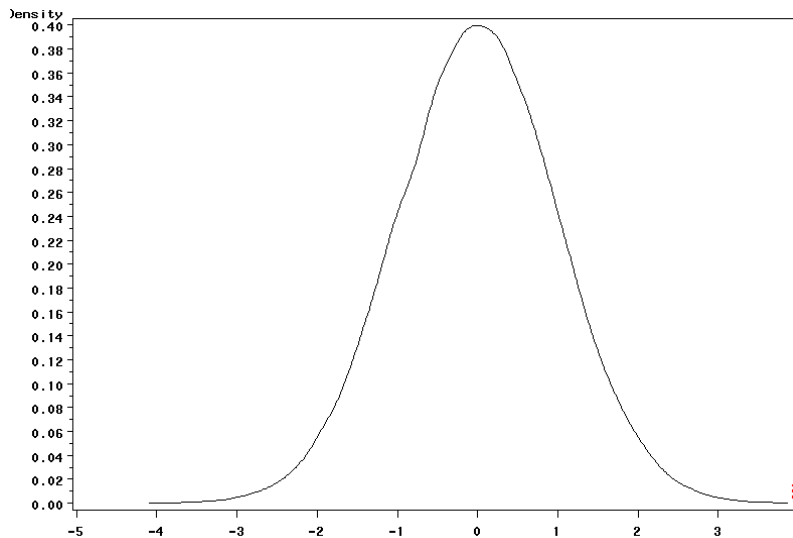
```
PROC SAS INSIGHT DATA=[Nom du Fichier]; RUN;
```

Exemple 1 : Homicide à Toronto

On considère la relation entre le nombre d'homicides dans la métropole de Toronto entre 1960-1996 (variables HOM) et la population de cette ville (variable POP). Pour cela on utilise la procédure suivante :

On obtient alors une sortie similaire à la la figure (??). Toutefois dans le menu ANALYZE, on choisit alors l'onglet FIT. Puis on choisit la variable expliquée (HOM) et la variable explicative (POP). On clique alors sur l'onglet OUTPUT pour faire apparaître une nouvelle fenêtre comme représentée sur la figure (3.15). On choisit alors l'onglet KERNEL (Normal CGV) pour Cross Validation Function (voir sections précédentes). Pour régler les options, on clique alors sur NONPARAMETRIC CURVES (CGV).

Figure 3.10: Densité Estimée



Une nouvelle fenêtre (figure 3.16) apparaît dans lequel on peut régler le choix de la fonction kernel (KERNEL SMOOTHER) pour les poids des observations dans la regression. SAS/INSIGHT permet de construire trois types de regressions :

- Locally-Weighted Mean
- A Locally-Weighted Regression Line (LOESS, LOWESS et KERNEL)
- A locally-weighted quadratic polynomial regression (LOESS, LOWESS)

On peut ainsi retrouver nos deux regressions :

1. Kernel regression avec l'option FIXED BANDWIDTH
2. Loess regression ou Lowess (voir section suivante) avec l'option LOESS.

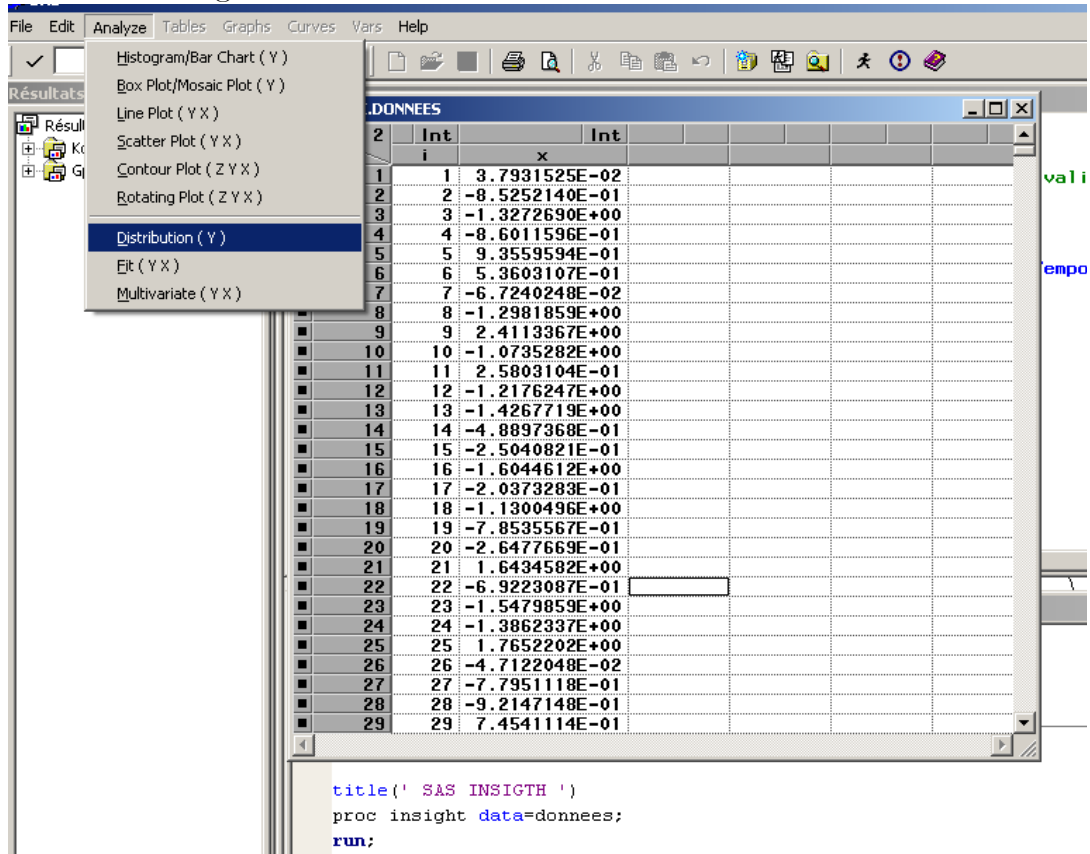
A ce niveau une remarque importante est la suivante :

Remarque *Un estimateur de regression locale avec un paramètre de lissage fixe (fixed bandwidth local mean estimator) est équivalent à un estimateur kernel.*

Les résultats d'estimation pour trois kernel et une regression loess sont reportés sur la figure (3.17).

On voit que la relation entre les homicides et la population est loin d'être linéaire (courbe rouge, polynôme de degré 1). En, effet, les estimateurs kernel

Figure 3.11: Procedure SAS INSIGHT : Distribution



quelle que soit le choix du kernel et la regression LOess donnent approximativement la même chose. On vérifie en particulier l'existence d'une sorte d'asymptote à droite de la relation. Aucune forme explicite de la relation n'est donnée, seuls ces graphiques sont disponibles.

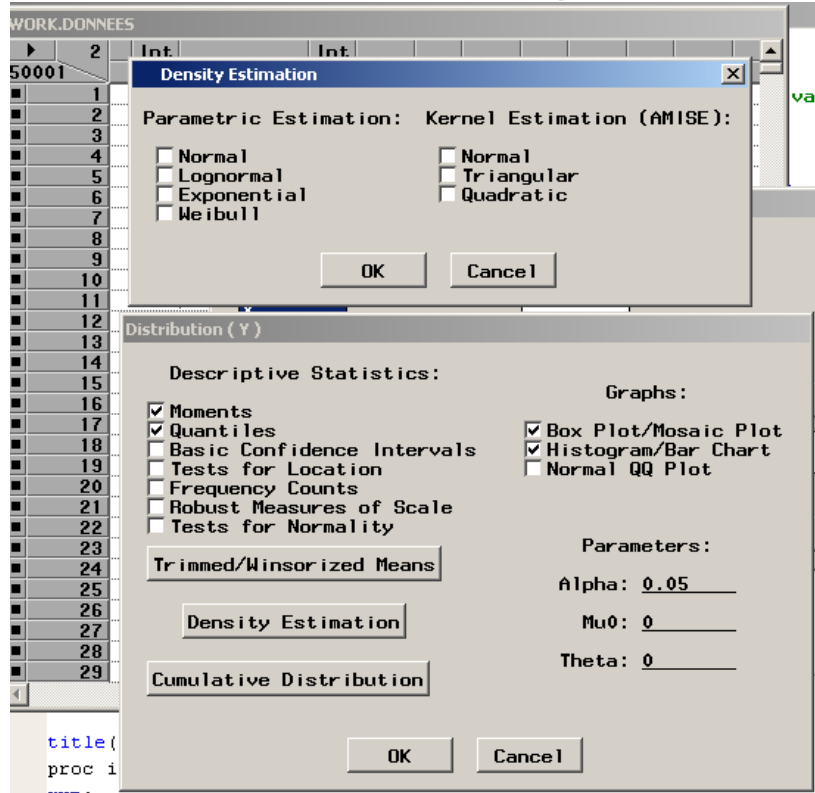
4. Régressions locales

Un des problèmes essentiels avec la régression kernel ou l'estimation de densité par noyau réside dans le manque de robustesse de ces estimateurs pour les valeurs extrêmes de X . Une solution alternative, plus robuste pour les valeurs extrêmes consiste en l'utilisation de régressions locales.

4.1. Régression locale : LOESS et LOWESS regressions

On cherche à estimer la relation $y_i = f(x_i) + \varepsilon_i$ où la fonction $f(x_i)$ est inconnue. L'idée de la régression linéaire locale consiste à utiliser un modèle de régression

Figure 3.12: Estimation d'une Densité par SAS INSIGHT



défini uniquement dans un voisinage du point x_0 d'intérêt. Notons $N(x_0)$ ce voisinage.

Definition 4.1. *Le principe général d'une régression locale est de postuler que la fonction de lien $f(x_0)$ évaluée au point x_0 peut être approximée par la valeur d'une fonction paramétrique évaluée localement au voisinage $N(x_0)$ du point de référence x_0 .*

Par exemple, on peut penser approximer $f(x_0)$ par son estimateur :

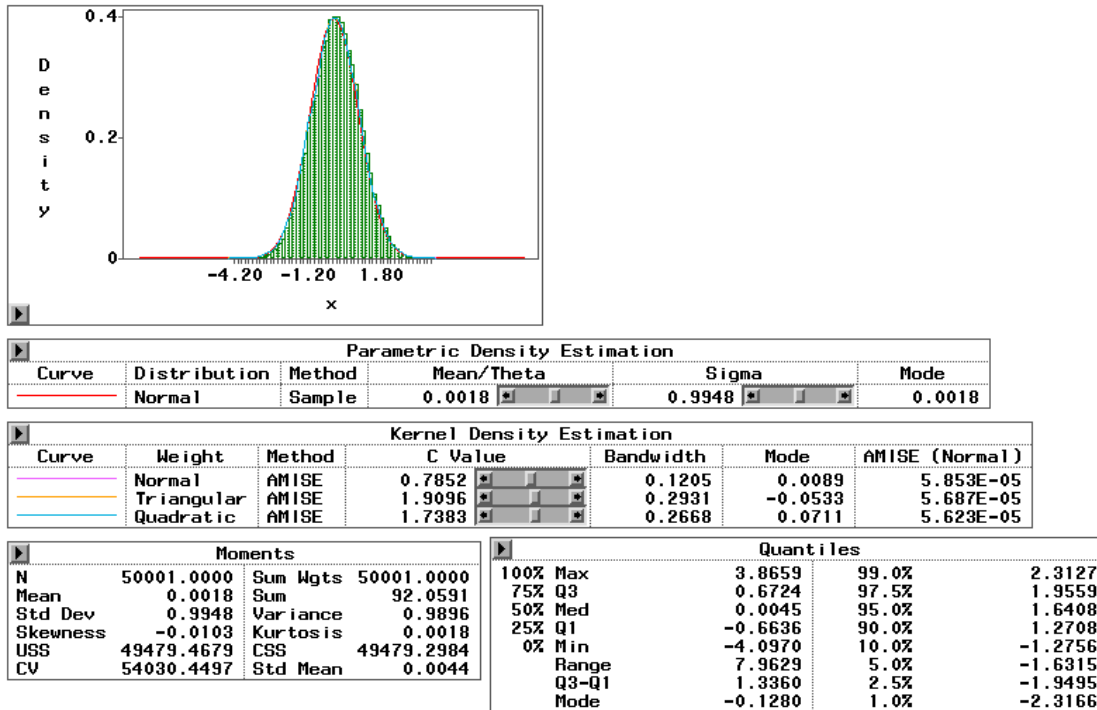
$$\hat{f}(x_0) = \hat{a}(x_0) + \hat{b}(x_0) x_0 \quad (4.1)$$

où les estimateurs des paramètres $\hat{a}(x_0)$ et $\hat{b}(x_0)$ sont déterminés par le programme suivant :

$$\{\hat{a}(x_0), \hat{b}(x_0)\} = \underset{\{a(x_0), b(x_0)\}}{\text{ArgMin}} \sum_{x_i \in N(x_0)} [y_i - a(x_0) - b(x_0) x_i]^2 \quad (4.2)$$

Les paramètres du modèle linéaire varient suivant le point de référence. Dans ce programme toutes les observations ont le même poids respectifs, mais on peut

Figure 3.13: Résultats d'Estimation SAS INSIGHT



naturellement envisager des variantes dans lesquelles les poids des observations (x_i, y_i) diminue avec la distance entre x_i et x_0 suivant par exemple une fonction kernel :

$$\left\{ \hat{a}(x_0), \hat{b}(x_0) \right\} = \underset{\{a(x_0), b(x_0)\}}{ArgMin} \sum_{x_i \in N(x_0)} [y_i - a(x_0) - b(x_0)x_i]^2 K\left(\frac{x_i - x_0}{\lambda}\right) \quad (4.3)$$

où λ désigne un paramètre de lissage. Voir Cleveland (1979) et Cleveland et Devlin (1988). Pour les constructions d'intervalle de confiance voir Fan et Gijbels (1996). Ces deux types de variantes correspondent aux deux cas :

1. La régression locale ou **LOESS (Local rEgrESSion)** de Cleveland (1979).
2. La régression locale pondérée ou **LOWESS (LOcally WEighted Scatterplot Smothing)** de Cleveland et Devlin (1988).

Considérons le cas de la LOESS regression. La principale différence avec la régression kernel c'est que la valeur de $f(x_0)$ estimé&xe n'est pas une moyenne mais une valeur prévue par une droite de régression. Par contre

Figure 3.14: Kernel Regression

```

data homicides;
  infile 'C:\Chris\Cours\Econometrie_NonParametrique\AppliSAS\homicides.txt' ;
  input date hom pop ratio;
  t+1;
run;

proc insight data=homicides;
run;

```

c'est une méthode qui requiert plus de temps de calcul : pour N observations on doit faire N régressions.

Definition 4.2. Dans le contexte de la LOESS regression, souvent on caractérise la voisinage de la variable x_0 , noté $N(x_0)$, par un rapport constant, appelé smoothing parameter, quel que soit le point considéré :

$$\lambda = \frac{\dim [N(x_0)]}{N} \in]0, 1] \quad \forall x_0 \quad (4.4)$$

Si λ est trop faible, l'estimateur des paramètres $a(x_0)$ et $b(x_0)$ manque de précision car le voisinage est trop petit, si au contraire le voisinage couvre l'ensemble des observations ($\lambda = 1$), alors on retrouve la droite d'ajustement linéaire (modèle de régression simple).

4.2. Procédure LOESS

La procédure SAS permettant de réaliser une régression de ce type est tout simplement appelé LOESS. La syntaxe générale de la procédure est la suivante :

```

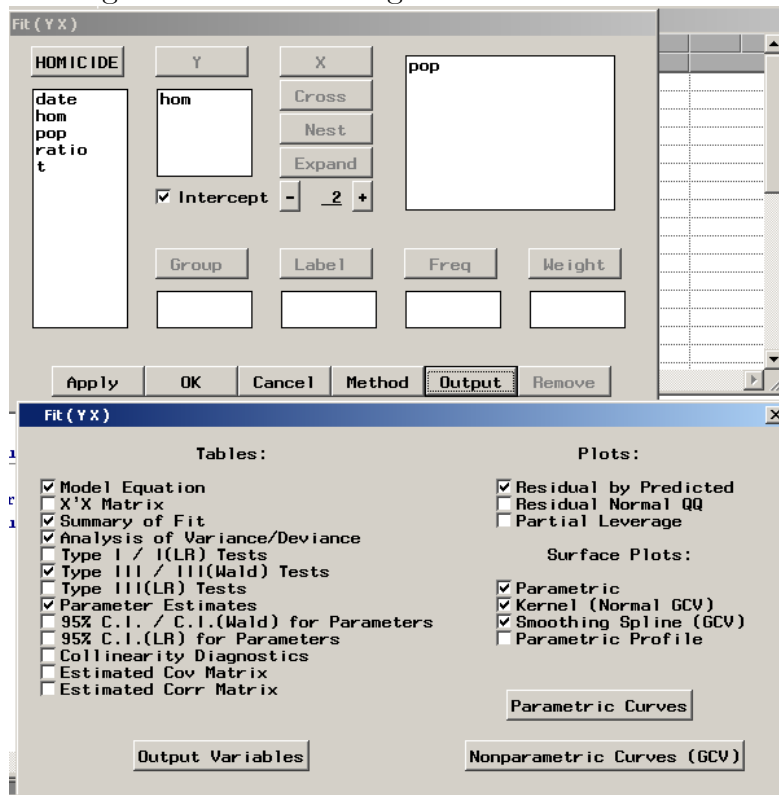
PROC LOESS <DATA=SAS-data-set> ;
MODEL dependents=regressors < / options > ;
ID variables ;
BY variables ;
WEIGHT variable ;
SCORE DATA=SAS-data

```

Exemple 1 : Application LOESS procedure.

On cherche à modéliser le lien entre le nombre d'homocides à Toronto et la population de ce centre urbain par une régression de type LOESS. Le graphique reportant les homicides en fonction de la population est reproduit sur la figure (4.1). Le résultat de la procédure LOESS est reporté sur la figure (4.2). On utilise

Figure 3.15: Kernel Regression : SAS INSIGHT



l'option details(OutputStatistics) afin d'afficher notamment les valeurs prévues par la procédure de la fonction de lien.

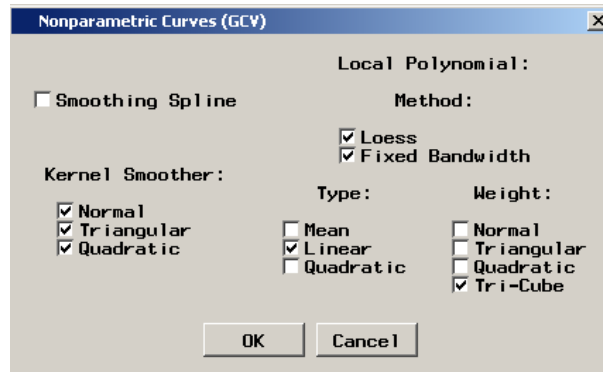
Le résultat de la procédure LOESS est alors reproduit sur la figure (4.3) et la figure (4.4). On vérifie que par défaut la procédure utilise une approximation linéaire (degree =1) et un paramètre de lissage optimale au sens d'un critère AIC égal à 0.66216. Sur la figure (4.4) sont reportées un ensemble de valeurs estimées de la fonction $hom = f(pop)$ pour l'ensemble des 37 valeurs (Number of Observations) de la variables pop . Pour chaque valeur est affichée la valeur estimée correspondante de $hom = f(pop)$ et la vraie valeur réalisée de cette variable hom . On peut ainsi caculer pour chaque valeur le résidu.

Un certain nombre de remarques doivent être faites à ce niveau sur l'utilisation de la procédure SAS.

Remarque 1 *Comme on le voit la fonction de lien $f(x)$ n'est pas estimée pour les $N = 37$ observations de la variable x , mais sur un sous ensemble de valeurs, ici $n = 14$ observations.*

En effet, SAS ne met pas en place l'estimation pour tous les N points sauf si

Figure 3.16: Régression Kernel : SAS INSIGHT (suite)



on le requiert avec l'option **DIRECT**. Dans le cas contraire (par défaut) SAS n'estime le polynôme que sur un nombre restreint de points, **puis réutilise le même polynôme au voisinage de ce point**. SAS utilise alors une procédure de type kd tree pour diviser les valeurs de x en segments de sorte à ce que les valeurs de $f(x)$ correspondantes soient comprises dans des segments de taille identique (rectangular cells). Le point médian des segments de x détermine alors le point x_0 autour duquel la fonction de lien $f(x_0)$ sera estimée par le polynôme $a(x_0) + b(x_0)x_0$. Ce polynôme sera utilisé pour toutes les valeurs x_i de ce segment, en postulant :

$$\hat{f}(x_i) = a(x_0) + b(x_0)x_i$$

Le nombre de valeurs de $f(x_i)$ comprises dans chaque segment est réglé par le paramètre dénommé **bucket size**, via l'option **BUCKET=**. Par défaut la valeur de ce paramètre est égale à :

$$\text{Bucket size} = \text{floor} \left(\frac{N\lambda}{5} \right) \quad (4.5)$$

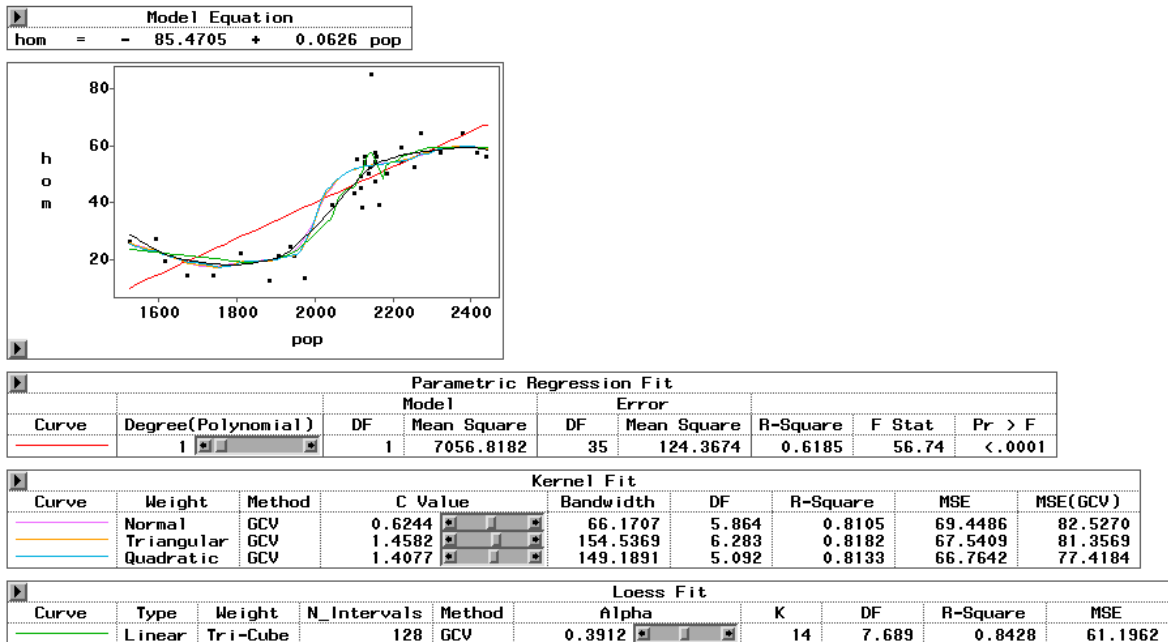
Par exemple, dans notre exercice on a $\lambda = 0.66216$ et $N = 37$, d'où :

$$\frac{N\lambda}{5} = \frac{37 * 0.66216}{5} = 4.9 \implies \text{Bucket size} = 4$$

Donc on regroupe les valeurs de la variable explicative pop dans des segments tels que les valeurs ajustées de $f(x) = \text{hom}$ correspondantes soient réparties dans des segments de taille identiques comprenant au plus 4 valeurs. Il peut y avoir plus de $N/4$ segments, dès lors que certains segments sur x peuvent contenir moins de 4 valeurs. Dans le cas présent, la procédure identifie 14 segments (**Number of Fitting Points**) pour lesquels on va considérer les 14 valeurs correspondantes du point moyen :

$$pop_i \in C_i, i = 1, \dots, 14$$

Figure 3.17: Résultats d'Estimation Kernel et Loess : SAS INSIGHT



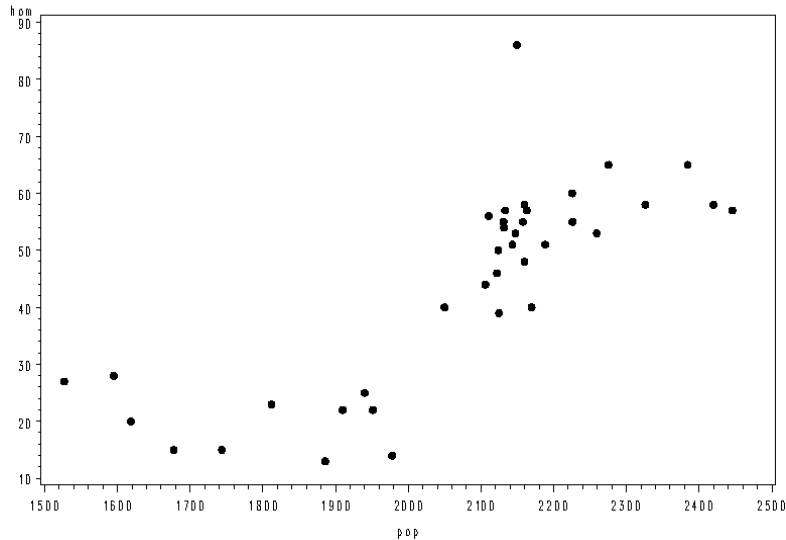
Pour une classe donnée C_i , on estime un polynôme de degré un au point moyen noté x_0

$$\left\{ \hat{a}(pop_0), \hat{b}(pop_0) \right\} = \underset{\{a(pop_0), b(pop_0)\}}{ArgMin} \sum_{pop_i \in C_i} [y_i - a(pop_0) - b(pop_0) pop_i]^2 \quad (4.6)$$

et l'on utilise ce polynôme pour estimer $f(pop_i)$ pour n'importe quel point pop_i appartenant à ce segment C_i par interpolation des valeurs $f(pop_j)$ des valeurs de pop_j connues sur ce segment. On peut alors choisir entre options pour la méthode d'interpolation via l'option **INTERP=** : soit une interpolation linéaire (par défaut), soit une interpolation cubique (**CUBIC**). On a donc les options suivantes :

- **DEGREE= 1 | 2** sets the degree of the local polynomials to use for each local regression. The valid values are 1 for local linear fitting or 2 for local quadratic fitting, with 1 being the default.
- **DROPSQUARE=(variables)** specifies the quadratic monomials to exclude from the local quadratic fits. This option is ignored unless the **DEGREE=2** option has been specified. For example, model `z=x y / degree=2 dropsquare=(y)` uses the monomials 1, x, y, x², and x y in performing the local fitting.

Figure 4.1: Relation Population - Homicides
Scatter Plot of Homicides



- **DIRECT** specifies that local least squares fits are to be done at every point in the input data set. When the direct option is not specified, a computationally faster method is used. This faster method performs local fitting at vertices of a kd tree decomposition of the predictor space followed by blending of the local polynomials to obtain a regression surface.
- **BUCKET=number** specifies the maximum number of points in the leaf nodes of the kd tree. The default value used is $s*n/5$, where s is a smoothing parameter specified using the **SMOOTH=** option and n is the number of observations being used in the current **BY** group. The **BUCKET=** option is ignored if the **DIRECT** option is specified.
- **INTERP= LINEAR | CUBIC** The **INTERP=** option specifies the degree of the interpolating polynomials used for blending local polynomial fits at the kd tree vertices. This option is ignored if the **DIRECT** option is specified in the model statement. **INTERP=CUBIC** is not supported for models with more than two regressors. The default is **INTERP=LINEAR**.

Exemple 2 : On estime la relation entre les homicides et la population en mode **DIRECT** c'est à dire pour chacun des points (figure 4.5). Les résultats sont reportés sur la figure (4.6). On vérifie que la mention Number of Fitting Points a disparu compartivement à la figure (4.3), puisque la régression est évaluée pour

Figure 4.2: LOESS Procedure

```

data homicides;
  infile 'C:\Chris\Cours\Econometrie_NonParametrique\ApplISAS\homicides.txt' ;
  input date hom pop ratio;
  t+1;
run;

symbol1 color=black value=dot ;
proc gplot data=homicides;
  title1 'Scatter Plot of Homicides';
  plot hom*pop;
run;

proc loess data=homicides;
  model hom = pop /details(OutputStatistics);;
  ods output OutputStatistics=Results;
run;

```

Figure 4.3: LOESS Regression

The LOESS Procedure
 Selected Smoothing Parameter: 0.662
 Dependent Variable: hom

Fit Summary

Fit Method	kd Tree
Blending	Linear
Number of Observations	37
Number of Fitting Points	14
kd Tree Bucket Size	4
Degree of Local Polynomials	1
Smoothing Parameter	0.66216
Points in Local Neighborhood	24
Residual Sum of Squares	2200.10242
Trace[L]	4.70105
GCV	2.10895
AICC	5.46166

chacune des $N = 37$ valeurs de la variable pop . De plus le paramètre de lissage optimal est passé de 0.6616 à 0.7162. De même la mention BLENDING LINEAR a disparu puisque qu'il n'y a pas d'interpolation entre les valeurs de pop .

Remarque 2 La procédure LOESS attribue des poids selon une fonction de type tri-cubique en fonction de la distance au centre de classe.

Supposons que l'on ait q points au voisinage d'un point x_0 et que l'on note d_1, d_2, \dots, d_q les distances par ordre croissant des q points du segment par rapport au point de référence. Chaque observation x_i se verra attribué un poids en fonction de sa distance :

$$\left\{ \hat{a}(x_0), \hat{b}(x_0) \right\} = \underset{\{a(x_0), b(x_0)\}}{ArgMin} \sum_{x_i \in N(x_0)} [y_i - a(x_0) - b(x_0)x_i]^2 w_i \quad (4.7)$$

Figure 4.4: LOESS Regression

The LOESS Procedure
Selected Smoothing Parameter: 0.662
Dependent Variable: hom

Output Statistics

Obs	pop	hom	Predicted hom
1	1527.00000	27.00000	22.90147
2	1595.00000	28.00000	22.13322
3	1619.00000	20.00000	21.86207
4	1677.70000	15.00000	21.41078
5	1744.00000	15.00000	20.90106
6	1812.20000	23.00000	21.22181
7	1886.00000	13.00000	21.56889
8	1910.00000	22.00000	21.68177
9	1940.30000	25.00000	23.67978
10	1951.70000	22.00000	24.43151
11	1978.10000	14.00000	28.48375
12	2050.20000	40.00000	39.55068
13	2106.20000	44.00000	48.14635
14	2160.10000	48.00000	53.68661
15	2125.10000	39.00000	50.89722
16	2157.90000	55.00000	53.56198
17	2188.70000	51.00000	55.05137
18	2226.20000	55.00000	56.66622
19	2259.60000	53.00000	57.83602
20	2131.20000	55.00000	51.47859
21	2143.50000	51.00000	52.53049
22	2110.90000	56.00000	48.87150
23	2122.30000	46.00000	50.63037
24	2124.00000	50.00000	50.79239
25	2131.90000	54.00000	51.54530
26	2163.30000	57.00000	53.86789
27	2169.90000	40.00000	54.24179
28	2225.60000	60.00000	56.64038
29	2133.50000	57.00000	51.68119
30	2160.00000	58.00000	53.68094
31	2147.60000	53.00000	52.87871
32	2149.90000	86.00000	53.03128

Figure 4.5: Estimation LOESS REGRESSION en mode Direct

```
proc loess data=homicides;
  model hom = pop /details(OutputStatistics) direct ;
  ods output OutputStatistics=Results;
run;
```

$$w_i = \left(\frac{32}{5}\right) \left[1 - \left(\frac{d_i}{d_q}\right)^3\right]^3 \quad d_i = |x_i - x_0| \tag{4.8}$$

Si le paramètre de lissage $\lambda > 1$, tous les points sont pris en compte dans la régression et le poids est alors défini par :

$$w_i = d_q \lambda^{1/p} \tag{4.9}$$

4.2.1. Sorties graphiques

La sortie des résultats se fait en utilisant une procédure de type ODS. On utilise pour cela l'option **ODS OUTPUT**. On peut sortir deux types de résultats :

Figure 4.6: Estimation Regression LOESS en mode DIRECT

```

The LOESS Procedure
Selected Smoothing Parameter: 0.716
Dependent Variable: hom

Fit Summary

Fit Method          Direct
Number of Observations      37
Degree of Local Polynomials    1
Smoothing Parameter      0.71622
Points in Local Neighborhood    26
Residual Sum of Squares      2227.25935
Trace[L]                  4.19050
GCV                        2.06905
AICC                       5.43455

```

- **OutputStatistics** = [Nom de fichier]
- **FitSummary** = [Nom de fichier] contient les éléments de la table Fit Summary
- **PredAtVertices**= [Nom de fichier] contient les valeurs prévues aux points d'estimation

Remarque *Si l'on veut sortir les valeurs des résidus et des intervalles de confiance sur les valeurs estimées, les options RESIDUAL et CLM sont nécessaires dans la spécification MODEL.*

Exemple 2 : On estime la relation entre les homicides et la population pour la valeur optimale du paramètre de lissage (figure 4.5). On cherche ensuite à grapher la relation estimée entre les variables homicides et populations pour ces deux valeurs. Les résultats sont reportés sur la figure (4.6).

Exemple 3 : On estime la relation entre les homicides et la population pour des valeurs de 0.3, 0.4, 0.5 et de 0.6 du paramètre de lissage (figure 4.9). On cherche ensuite à grapher la relation estimée entre les variables homicides et populations pour ces deux valeurs. Les résultats sont reportés sur la figure (4.10) avec 4 graphiques pour les quatre valeurs du paramètre de lissage λ . Dans ce cas précis, l'allure générale de la relation est relativement peu sensible au choix de λ , mais cette observation est très loin d'être générale.

4.2.2. Sélection du paramètre de lissage

Sous SAS on peut tout d'abord utiliser l'option SMOOTH pour spécifier un ensemble de valeurs pour le paramètre de lissage pour lesquelles on estimera la

Figure 4.7: Estimation LOESS Regression : λ optimal

```

data homicides;
  infile 'C:\Chris\Cours\Econometrie_NonParametrique\AppliSAS\homicides.txt' ;
  input date hom pop ratio;
  t+1;
run;

proc loess data=homicides;
  model hom = pop /details(OutputStatistics) all ;
  ods output OutputStatistics=Results;
run;

symbol1 color=black value=dot;
symbol2 color=black interp=join value=none;

proc gplot data=Results;
  title1 'Loess Regression Homicides ' ;
  plot DepVar*pop Pred*pop/ &opts;
run;

```

régression. En l'absence de critère de sélection le modèle sera estimé pour chacune des valeurs retenues. C'est ce que nous avons vu dans la section précédente. On peut en outre utiliser un critère de sélection du paramètre de lissage optimal. Ces critères d'information sont toujours de la forme suivante :

$$\text{Critère} = \log(\hat{\sigma}_\varepsilon^2) + \phi(L)$$

où $\hat{\sigma}_\varepsilon^2$ désigne un estimateur de la variance moyenne des résidus et où $\phi(\cdot)$ désigne une fonction de pénalité construite de sorte à être croissante avec l'aspect lisse de la composante ajustée $\hat{f}(x)$. On a donc un arbitrage traditionnel entre la variance des résidus et la variance de la composante ajustée ou fit. Soit la matrice de lissage L telle que :

$$\hat{y} = \hat{f}(x) = Ly \quad (4.10)$$

Un de ces critères est le critère d'Akaike corrigé pour les petits échantillons qui satisfait alors la définition suivante :

Definition 4.3. *Le critère d'information d'Akaike corrigé AIC_C est défini par la relation :*

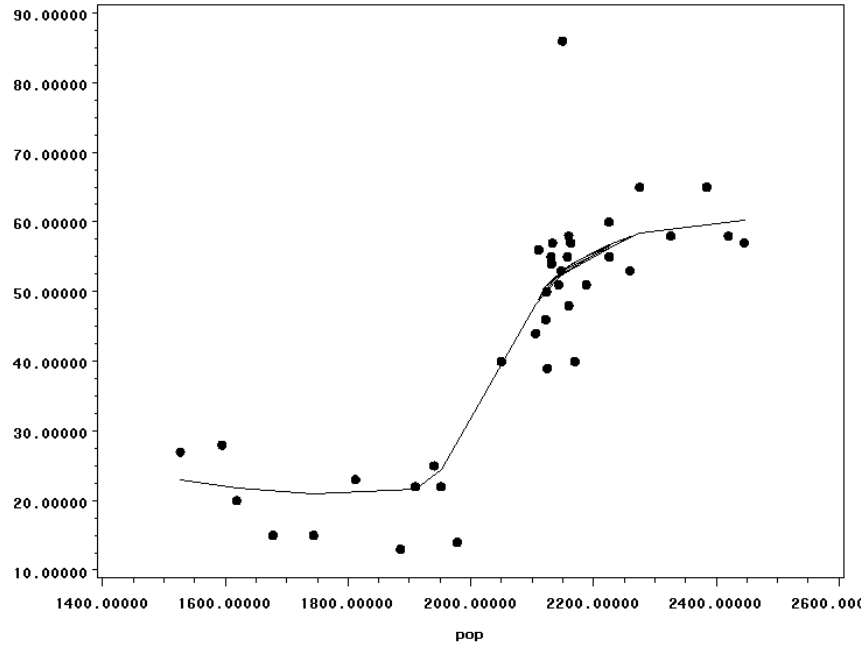
$$AIC_C = N \log(\hat{\sigma}_\varepsilon^2) + \frac{2N}{(N-1)} \quad (4.11)$$

Un autre critère d'Akaike corrigé est en outre proposé :

Definition 4.4. *Le critère d'information d'Akaike corrigé ou AIC_{C_1} d'Hurvich, Simonoff, et Tsai (1998) est défini par la relation :*

$$AIC_{C_1} = N \log(\hat{\sigma}_\varepsilon^2) + N \frac{(\delta_1/\delta_2)(N+v_1)}{(\delta_1^2/\delta_2) - 2} \quad (4.12)$$

Figure 4.8: Graphique Estimation LOESS Regression
Loess Regression Homicides



où N est le nombre d'observations,

$$\delta_1 = \text{Trace} [(I - L)'(I - L)] \tag{4.13}$$

$$\delta_2 = \text{Trace} \left\{ [(I - L)'(I - L)]^2 \right\} \tag{4.14}$$

$$v_1 = \text{Trace}(L'L) \tag{4.15}$$

Ce critère tend à corriger la tendance qu'avait le critère d'Akaike usuel à ne pas assez lisser la composante ajustée. La valeur de v_1 correspond ainsi à ce que l'on appelle le nombre équivalent de paramètre (**Equivalent Number of Parameter**) affiché par SAS lorsque l'on spécifie l'option **ALL** comme l'indique la figure (4.11).

Exemple 1 : Calcul de critère AIC_C et AIC_{C_1} . Calculons le critère AIC_C à partir des éléments de la figure (4.11). On a donc

$$AIC_C = N \log (\hat{\sigma}_\varepsilon^2) + \frac{2N}{(N - 1)} \tag{4.16}$$

$$= 37 * \log (8.33065^2) + \frac{2 * 37}{37 - 1} = 158.93 \tag{4.17}$$

Figure 4.9: LOESS Regression avec plusieurs Valeurs de λ

```

/* macro used in subsequent examples */
%let opts=vaxis=axis1 hm=3 vm=3 overlay;

proc loess data=homicides;
  model hom = pop /details(OutputStatistics) all smooth= 0.3 0.4 0.5 0.6;
  ods output OutputStatistics=Results;
run;

proc print data=Results(obs=5);
  id obs;
run;

goptions display;
proc gplot data=Results;
  by SmoothingParameter;
  plot DepVar*pop=1 Pred*pop / &opts name='fit';
run; quit;

goptions display;
proc greplay nofs tc=sashelp.templt template=l2r2;
  igout gseg;
  treplay 1:fit 2:fit2 3:fit1 4:fit3;
run; quit;

```

En ce qui concerne le critère AIC_1 , on a immédiatement que :

$$v_1 = 4.1040$$

$$\delta_1 = 31.70190$$

$$\delta_2 = 31.26808$$

On en déduit alors, la valeur du critère AIC_{C_1} égale à 202.31423.

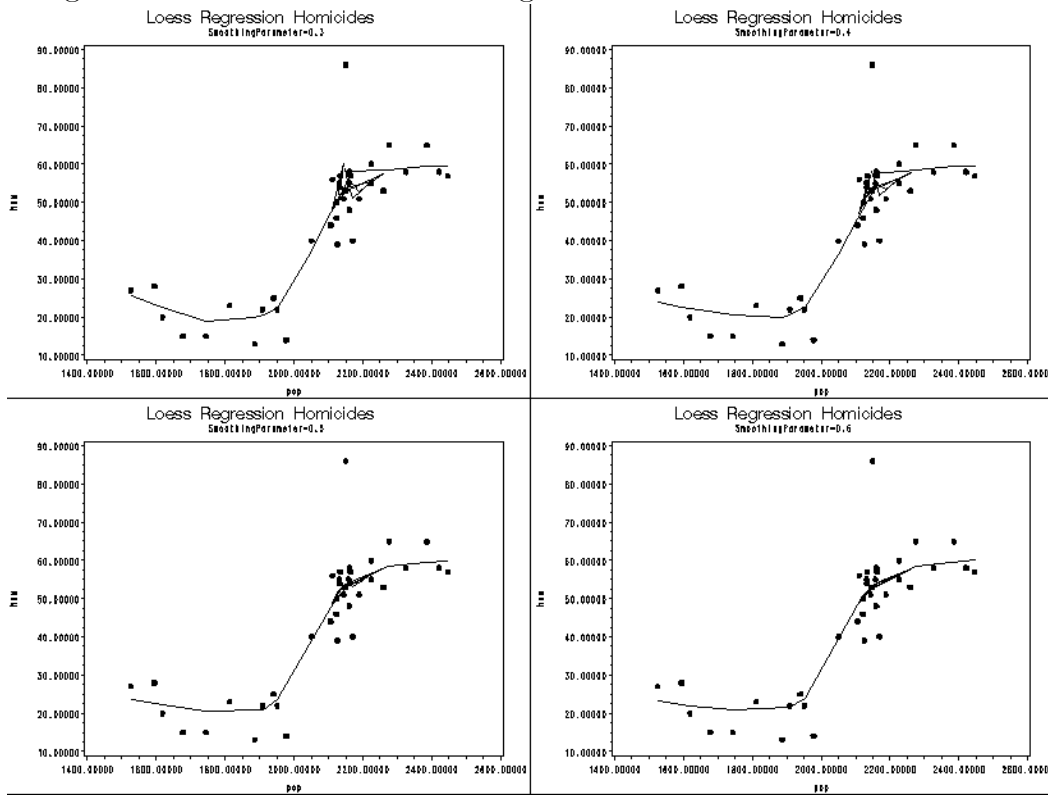
Si l'on ne spécifie pas de liste de valeurs pour le paramètre de lissage avec l'option **SMOOTH=** et si l'on ne spécifie pas de critère de sélection avec l'option **SELECT=**, la procédure LOESS détermine par défaut le paramètre de lissage λ par **minimisation du critère d'information d'Akaike corrigé AIC_C** (et non le critère AIC_{C_1}) :

$$\lambda^* = \underset{\lambda \in [0,1]}{\text{ArgMin}} AIC_C(\lambda) \quad (4.18)$$

Au contraire on peut spécifier la méthode de sélection grâce à l'option **SELECT=** :

1. **AICC** pour le **Biais Corrected Akaike Information Criteria**.
2. **AICC1** pour le **Biais Corrected Akaike Information Criteria AIC_{C_1}** (Hurvich, Simonoff, and Tsai 1998).

Figure 4.10: Estimation LOESS Regression Pour différentes valeurs de λ .



3. GCV Generalized Cross-Validation criterion (Craven and Wahba 1979).

On peut enfin croiser différents critères : chercher par un critère la valeur optimale de λ sur une liste de paramètres en utilisant de façon simultanée les options SELECT et SMOOTH. On peut en outre soit rechercher un optimum global sur le segment définies par les valeurs retenues dans l'option SMOOTH avec l'option

SELECT=Criterion(GLOBAL)

Ou alors on peut restreindre le domaine en utilisant l'option :

SELECT=Criterion(RANGE(lower, upper))

Dans ce cas on limite la recherche sur un segment défini entre deux valeurs. Pare xemple, on peut utiliser les syntaxes suivantes :

SELECT= GCV

Figure 4.11: LOESS Procedure avec Option ALL

The LOESS Procedure
 Selected Smoothing Parameter: 0.662
 Dependent Variable: hom

Fit Summary

Fit Method	kd Tree
Blending	Linear
Number of Observations	37
Number of Fitting Points	14
kd Tree Bucket Size	4
Degree of Local Polynomials	1
Smoothing Parameter	0.66216
Points in Local Neighborhood	24
Residual Sum of Squares	2200.10242
Trace[L]	4.70105
GCV	2.10895
AICC	5.46166
AICC1	202.31423
Delta1	31.70190
Delta2	31.26808
Equivalent Number of Parameters	4.10400
Lookup Degrees of Freedom	32.14173
Residual Standard Error	8.33065

SELECT= AICC(GLOBAL)

SELECT= AICC1(RANGE(0.2,0.6))

Ainsi, la procédure LOESS admet différentes options au niveau de l'instruction MODEL.

- **SMOOTH=**value-list specifies a list of positive smoothing parameter values. A separate fit is obtained for each smoothing value specified.
- **TRACEL** option specifies that the trace of the prediction matrix as well as the GCV and AICC statistics are to be included in the "FIT Summary" table. The use of any of the MODEL statement options ALL, CLM, DFMETHOD=EXACT, DIRECT, SELECT=, or T implicitly selects the TRACEL option.
- **DFMETHOD=** option specifies the method used to calculate the "lookup" degrees of freedom used in performing statistical inference. The default is DFMETHOD=NONE. Approximate methods for computing the "lookup" degrees of freedom are not currently supported. The use of any of the MODEL statement options ALL, CLM or T or any SCORE statement CLM option implicitly selects the DFMETHOD=EXACT option.

4.2.3. Autres options de la procédure LOESS

Enfin la procédure LOESS présente un certain nombre d'otptions permettant notamment de calculer des intervalles de confiance sur les estimateurs de $f(x)$:

- **ALL** requests all these options: CLM, RESIDUAL, SCALEDINDEP, STD, and T.
- **ALPHA=number** sets the significance level used for the construction of confidence intervals for the current MODEL statement. The value must be between 0 and 1; the default value of 0.05 results in 95% intervals.
- **CLM** requests that confidence limits on the mean predicted value be added to the "Output Statistics" table. By default, 95% limits are computed; the ALPHA= option in the MODEL statement can be used to change the -level. The use of this option implicitly selects the model option DFMETHOD=EXACT if the DFMETHOD= option has not been explicitly used.
- **DETAILS < (tables) >** selects which tables to display, where tables is one or more of kdTree (or TREE), PredAtVertices (or FITPOINTS), and OutputStatistics (or STATOUT). A specification of kdTree outputs the kd tree structure, PredAtVertices outputs fitted values and coordinates of the kd tree vertices where the local least squares fitting is done, and OutputStatistics outputs the predicted values and other requested statistics at the points in the input data set. The kdTree and PredAtVertices specifications are ignored if the DIRECT option is specified in the MODEL statement. Specifying the option DETAILS with no qualifying list outputs all tables.
- **ITERATIONS=number** specifies the number of iterative reweighting steps to be done. Such iterations are appropriate when there are outliers in the data or when the error distribution is a symmetric long-tailed distribution. The default number of iterations is 1.
- **RESIDUAL | R** specifies that residuals are to be included in the "Output Statistics" table.
- **SCALE= NONE | SD < (number) >** specifies the scaling method to be applied to scale the regressors. The default is NONE, in which case no scaling is applied. A specification of SD(number) indicates that a trimmed standard deviation is to be used as a measure of scale, where number is the trimming fraction. A specification of SD with no qualification defaults to 10% trimmed standard deviation.
- **SCALEDINDEP** specifies that scaled regressor coordinates be included in the output tables. This option is ignored if the SCALE= model option is not used or if SCALE=NONE is specified.

- STD specifies that standardized errors are to be included in the "Output Statistics" table.
- T specifies that t statistics are to be included in the "Output Statistics" table.