Nonparametric regression with missing data

González Manteiga, Wenceslao
University of Santiago, Dep. Statistics and O.R.
South Campus, 15706
Santiago de Compostela, Spain
wences@zmat.usc.es

Pérez González, Ana
University of Vigo, Dep. Statistics and O.R.
University Campus of As Lagoas, 32004
Orense, Spain
anapg@uvigo.es

1. Introduction

In this work our objective is to consider a local polynomial non-parametric regression estimator adapted to the case where the dependent variable Y has missing observations and the covariables are totally observable. In this way we extend the works of Ruppert and Wand (1994), who studied the local polynomial estimator with a complete sample, and Chu and Cheng (1995) who applied the local polynomial estimator to the case of missing observations with a single covariable.We assume a heterocedastic regression model of the type $Y = m(X) + v^{\frac{1}{2}}(X)\varepsilon$, where ε is the error term, having mean zero and variance one, and where $v(x) = Var[Y/X = x]$. It may be possible that Y_i is not observed for any index i, and so we find that: $(X_i^t, Y_i) \in \mathbb{R}^{d+1}$ if Y_i is observed and $(X_i^t, ?) \in \mathbb{R}^d$ for the opposite case. In order to check whether an observation is complete or not, a new variable δ is introduced into the model thus $\delta_i = 1$ if Y_i is observed and zero if Y_i is missing for $i = 1, ..., n$. In this paper we suppose that the data are missing at random (MAR), i.e. $P(\delta = 1/Y, X) = P(\delta = 1/X) = p(X)$.

The first estimator proposed is the *Simplified Multivariate Local Linear Smoother* (*SMLLS*) which consist of using only complete observations, (those where $\delta_i = 1$.) Thus the estimator has this expression: $\widehat{m}_{S,H}(x) = \widehat{\alpha} = e_1^t \left(\mathbf{X}^t W^\delta \mathbf{X}\right)^{-1} X^t W^\delta \mathbf{Y}$, where $\mathbf{X}$ has the same expression as for the complete data (see Ruppert and Wand 1994), and $W^\delta = diag\{K_H(X_i - x)\delta_i\}_{i=1}^n$.

The second estimator, the *Imputed Multivariate Local Linear Smoother* (*IMLLS*) which consists of initially employing the *SMLLS* to estimate the missing observations $\widehat{Y}_i = \delta_i Y_i + (1 - \delta_i)\widehat{m}_{S,G,(L)}(X_i)$, with $\widehat{m}_{S,G,(L)}(X_i)$ as the *SMLLS,* using a bandwidth matrix G and a kernel function L, and then applying the *Multivariate Local Linear Smoother* to the completed sample: $\widehat{m}_{I,H,G}(x) = \widehat{\alpha} = e_1^t \left(\mathbf{X}^t W \mathbf{X}\right)^{-1} \mathbf{X}^t W \widehat{\mathbf{Y}}$.

2. Results

Let there be a point x from the interior of supp(f) (f is the density of X) Under certain regularity conditions we obtain that (conditional to $\mathbb{X} = (X_1, ..., X_n)$)

$$AMSE\left[\widehat{m}_{S,H}(x)/\mathbb{X}\right] = \left(\frac{1}{2} tr\left[H\mathcal{H}_m(x)\right]\mu_2(K)\right)^2 + \frac{v(x)}{n|H|^{\frac{1}{2}} f(x) p(x)} \int K(z)^2 dz \left\{1 + o_p(1)\right\},$$

$$AMSE\left[\widehat{m}_{I,H,G}\left(x\right)/\mathbb{X}\right] = \left(\frac{1}{2}tr\left[H\mathcal{H}_m\left(x\right)\right]\mu_2\left(K\right)+\frac{1}{2}q\left(x\right)tr\left[G\mathcal{H}_m\left(x\right)\right]\mu_2\left(L\right)\right)^2 + \\ +\frac{\left|H\right|^{\frac{-1}{2}}v\left(x\right)}{nf\left(x\right)p\left(x\right)}\int\left(p\left(x\right)K\left(z\right)+q\left(x\right)\frac{\left|H\right|^{\frac{1}{2}}}{\left|G\right|^{\frac{1}{2}}}A\left(z\right)\right)^2 dz\left\{1+o_p\left(1\right)\right\}$$

where $A\left(z\right)=\int K\left(v\right)L\left(G^{\frac{-1}{2}}H^{\frac{1}{2}}\left(z-v\right)\right)dv$, $q\left(x\right)=1-p\left(x\right)$, and $\left|G\right|^{\frac{-1}{2}}\left|H\right|^{\frac{1}{2}}=O\left(1\right)$.

We also considered in our work the other cases $\left|G\right|^{\frac{-1}{2}}\left|H\right|^{\frac{1}{2}}\to 0$ and $\left|H\right|^{\frac{-1}{2}}\left|G\right|^{\frac{1}{2}}\to 0$ as $n\to\infty$. In general the IMLLS is better than the SMLLS in the first case of the three considered.

In the following example, the quotients between the AMSEs for the IMLLS and for the SMLLS are described for the three cases before commented, for spherically bandwidth matrices: $H=h^2I$ and $G=g^2I$. The regression function considered is $m(x_1,x_2)=\frac{5}{\pi}\exp\left\{-5\left(x_1\right)^2/8\right\}$, with $x_1\in(-2,2)$ and $x_2\in(0,1)$. The sample size is 100 and $v(x)=0.5$. The bidimensional kernel functions K and L are the product of two Epanechnikov unidimensional functions, and the model of missing data considered is $p(x_1,x_2)=1/\left(1+\exp\left\{-x_1^2\right\}\right)$.

In the figures 1, 2 and 3 is showed in the example the behaviour of the estimators for cases corresponding to $g\sim h$, $g>>h$ and $h>>g$. As it can be seen in the Figure 1 the first case is the best situation for the imputation.

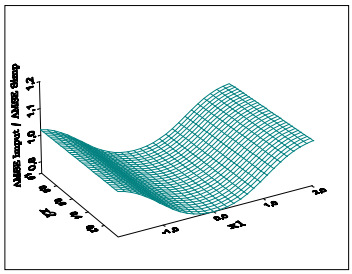

Fig.1: h=0.3 and g=0.25

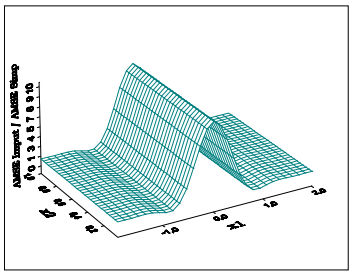

Fig. 2: h=0.3 and g=3

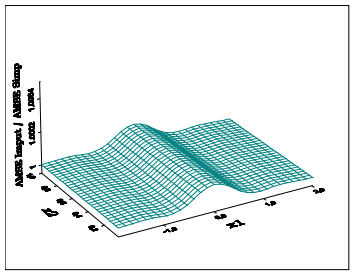

Fig. 3: h=0.3 and g=0.05

A Bootstrap method also has been studied to obtain optimal estimated bandwidth matrices.

REFERENCES

Chu, C.K. and Cheng. P.E. (1995). Nonparametric regression estimation with missing data. *Journal of Statistical Planning and Inference*, **48**, 85-99.

Ruppert, D. and Wand, M. P.(1994). Multivariate locally weighted least squares regression. *The Annals of Statistics*, **22**, 3, 1346-1370.

RESUME

TITRE: Regression non-paramétrique avec des données manquantes.

Plusieurs fois, il nous faut estimer la fonction de regression lorsque quelques observations sont manquantes. Dans ce travail on étudie l'effet des observations manquantes de la variable réponse, dans l'estimation de la fonction de regression multivariante. On propose deux estimateurs non-paramétriques basés dans l'estimation "local linéal multivariante" (Ruppert et Wand 1994). Le premier n'utilise que des observations complètes; et le deuxième impute les observations manquantes et après, il réalise l'estimation avec l'échantillon déjà complété. On étudie son erreur quadratique moyenne et on propose une méthode Bootstrap pour estimer la largeur de bande optimale.