COMPUTING NONPARAMETRIC FUNCTIONAL ESTIMATES IN SEMIPARAMETRIC PROBLEMS

Miguel A. Delgado
Universidad Carlos III de Madrid
Getafe; Madrid 28903; SPAIN

Key Words and Phrases: nonparametric functional estimation; semiparametric models; Fortran routines.

ABSTRACT

The purpose of this note is to provide a brief account of available FORTRAN routines for computing nonparametric functional estimates, frequently used in semiparametric problems, evaluated at each data point. Then semiparametric estimates can be computed employing the user-favored econometric software.

1. INTRODUCTION

The nonparametric functionals more frequently used in semiparametric estimation of econometric models are density functions and their derivatives, and regression curves. Different semiparametric problems require different nonparametric estimation methods. Densities and their derivatives are usually estimated by the kernel method. Regression curves are estimated by either kernels or nearest neighbors.

Given n observations $\{(X_i,Y_i),\ i=1,\dots,n\}$ of a random variable (X,Y), where Y is scalar and X is a p-dimensional random vector, nonparametric estimates of the regression function $m(x)=E(Y|X=x)$ can be defined as

$$\hat{m}(x)= \sum_{i=1}^{n} Y_i\ W_i(x),$$

where $\{W_i(x),\ i=1,\dots,n\}$ is a sequence of weights. Kernel weights are defined as $W_i(x)= n^{-1}\ K_H(x- X_i)/\hat{f}_H(x)$, where $\hat{f}_H(x)= n^{-1}\sum_{i=1}^{n} K_H(H^{-1}(x- X_i))$ is the density function estimate of X evaluated at x, and $K_H(u)= det(H)^{-1} K(H^{-1}u)$ is the kernel with scale smoothing matrix H. The function K(.) integrates to one. Nearest neighbor weights are defined as $W_i(x)= c_{\tau_i}(x, k)$, where $\sum_{i=1}^{n} c_i(x, k)= 1$, $c_i(x,k)> 0$ when $i\leq k$, and τ_i is the rank of X_i according to increasing distances $\rho(X_i, x)$, where $\rho(.,.)$ is some distance function (if $\rho(X_i,x) = \rho(X_j, x)$, then we arbitrarily call X_i closer to x if $i< j$). This tie-breaking-rule was suggested by Devroye (1978) and is computationally convenient.

The routines described below compute regression estimates and many other related functionals like density estimates, robust conditional M-estimates, and conditional quantile estimates. Soft copies of the code and detailed documentation on the routines can be obtained by e-mail from DELGADO@ECO.UC3M.ES, or sending a formatted floppy disk and a self-address stamped envelope to the author.

2. ROUTINES FOR NONPARAMETRIC FUNCTIONAL ESTIMATION

The output of the routines consists of a vector containing $\hat{m}(X_i)$, $i=1,..,n$, or other related functionals (conditional robust regression and conditional quantiles estimates) when required. Kernel routines also provide $\hat{f}_H(X_i)$ and $\hat{P}_H(X_i) = \hat{m}(X_i)\hat{f}_H(X_i)$, $i=1,..,n$, by default.

The input for kernel estimation consists of the kernel function and the bandwidth matrix. Different option parameters allow to choose H diagonal or $H= h\ \hat{\Sigma}$, where h is scalar and $\hat{\Sigma}$ is the sample covariance matrix of X. The user can choose whether or not employ the own observation when computing the kernels. If instead of the kernel function, kernel derivatives are provided, the output will consist of the derivative density estimates evaluated at each data point. An efficient algorithm, respect to storage space and time, is employed when the kernel is symmetric.

The input for nearest neighbor estimation consists of the number of nearest neighbors k and the weight function. The user can choose whether or not employ the own observation when computing the weights and the distance function. The nearest neighbors are found using the algorithm proposed by Friedman et al (1985). This algorithm is pretty fast compared with a brute force method.

3. APPLICATIONS TO SEMIPARAMETRIC PROBLEMS

Many semiparametric estimates appearing in the recent econometric literature can be computed using these routines and standard econometric software. We mention only a few.

Generalized least squares (GLS) estimators in the presence of heteroskedasticity of unknown form, proposed by Carroll (1982) and Robinson (1987) among others, are straightforward to compute using standard econometric software (TSP, SAS, PC-GIVE etc). Once ordinary least squares (OLS) residuals are obtained, their squares and the set of regressors observations is the input of the routines and the output is the vector of weights in the GLS procedure.

The kernel routines are suitable for computing semiparametric estimates of the parameter vector β in the partial linear regression model $E(Y|X_1,X_2)= X_1'\beta+ \theta(X_2)$. Once $E(Y|X_2)$ and $E(X_1|X_2)$ evaluated at each data point are estimated, β can be estimated as suggested by Robinson (1988) by linear regression.

Optimal instrumental variables (IV) estimates in nonlinear equation systems, proposed by Newey (1990), are obtained by computing the optimal instruments using our routines. The input, in this case, is the vector of derivatives of the error function evaluated at some root-n-consistent preliminary estimate, and the regressors observations. The output is the vector of optimal instruments. Our program also includes a routine for computing optimal instruments by resampling, as proposed by Robinson (1990). Once optimal instruments are available, instrumental variables estimates can be computed using TSP.

Density derivative estimates are needed when computing average derivatives of regression functions as suggested by Powell et al (1989).

Some nonparametric and semiparametric test procedures, e.g. Robinson (1989), require $\hat{f}_H(X_i)$, $\hat{P}_H(X_i)$, and their derivatives.

ACKNOWLEDGEMENTS

These routines were written in the summer of 1990, when I was visiting the

London School of Economics. This work was supported by the Economic and Social Research Council (ESRC), reference number: R00023411.

REFERENCES

Carroll, R.J. (1982): "Adapting for heteroscedasticity in linear models", Annals of Statistics 10, 1224-1233.

Devroye, L. (1978), "Uniform convergence of nearest neighbor regression function estimators and their application in optimization", IEE Transactions in Information Theory, IT-24, 142- 151.

Friedman, J.H., F. Baskett and L.J. Shustek (1975), "An algorithm for finding nearest neighbors", IEE Transactions on Computers C-24, 1149-1158.

Robinson, P.M. (1988): "Root-n-consistent semiparametric regression", Econometrica 56, 931-954.

Robinson, P.M. (1989): "Hypothesis testing in semiparametric and nonparametric models for econometric time series", Review of Economic Studies 56, 511-534.

Robinson, P.M. (1990): "Best nonlinear three-stage least squares of certain econometric models", Econometrica, .

Newey, W. (1990): "Efficient instrumental variable estimation of nonlinear models", Econometrica 58, 809-837.

Powell J.L., J.H. Stock and T.M. Stoker (1989): "Semiparametric estimation of index coefficients", Econometrica 57, 1403-1430.