Exploring Data with Non- and Semiparametric Models

Marlene Müller

This version: July 12, 2010

BEUTH HOCHSCHULE FÜR TECHNIK BERLIN
University of Applied Sciences

Plan

Course Topics

target audience: master level students in economics at a university with strong quantitative focus

- 2h → introduction: overview on the course topics, linear vs. nonparametric and semiparametric regression
- 4h → histogram: concept of density estimation, counting frequencies in bins
- 10h → kernel density estimation (KDE): local binning, weighting by kernels, statistical properties, bandwidth selection
- 10h → nonparametric regression: Nadaraya-Watson and local polynomial kernel regression, other methods (in partcular: splines, k-NN), computational aspects, smoothing parameter selection
- 6h → semiparametric extensions to GLM: single index models, additive models, partial linear models, generalized partial linear models

---

32h in total (units a 45min)

What students should understand?

Objectives: Concepts rather than detailed derivation of all theoretical properties.

- nonparametric function estimates may recover more features of the data than parametric function estimates, difference to parametric estimates
- the relevant steps to derive statistical properties (bias, variance, MSE and MISE)
- smoothing parameter selection: plug-in estimation and cross-validation as rather universal concepts
- the choice of the kernel function has a rather small impact on the function estimator but may be essential for its computation (and smoothness of the resulting estimates)

What students should understand? (Cont'd)

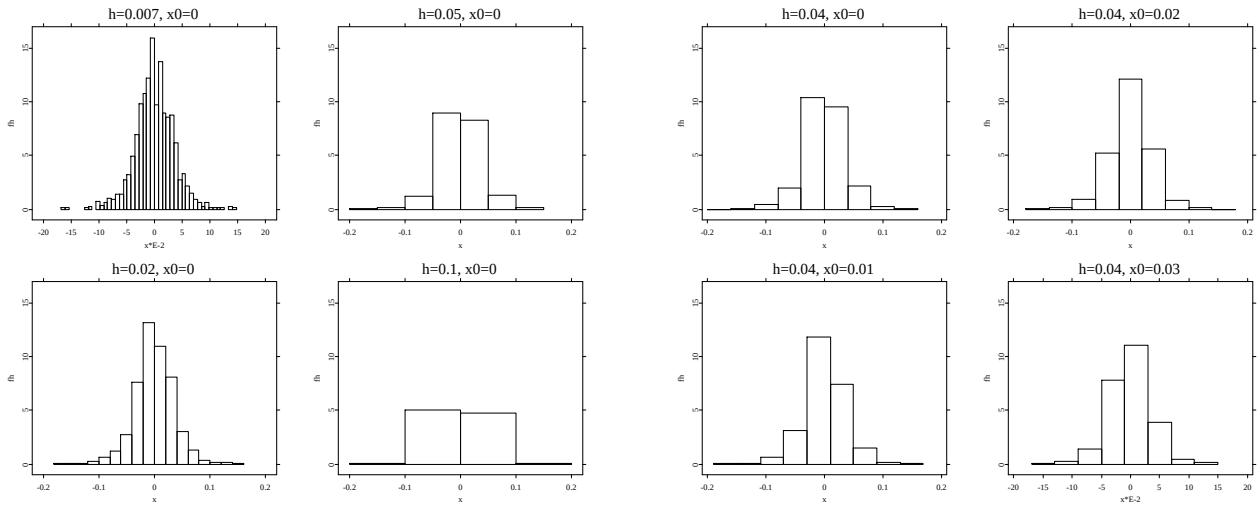
- multidimensional nonparametric function estimation leads to the "curse of dimensionality"; problems in graphical presentation and interpretation
- different smoothers have complexity of computation; in particular, spline smoothing and k-nearest neighbour regression (one-dimensional case) as competitors kernel regression
- semiparametric regression extends the concept of fully nonparametric regression; it allows for partially known or parametric components and provide an easy interpretation together with interesting statistical propoerties

Throughout the course R scripts are provided that can be altered or modified by the course participants to get a deeper insight into the methodology.

Histogram

- → histogram as a density estimator that students are familiar with
- → provides a step function, different settings for the histogram bins may lead to different interpretation of the estimated distributions
- → short derivation of the theoretical (asymptotic) properties is given to compare with parametric and (later on) kernel density estimates

Different histogram bin settings



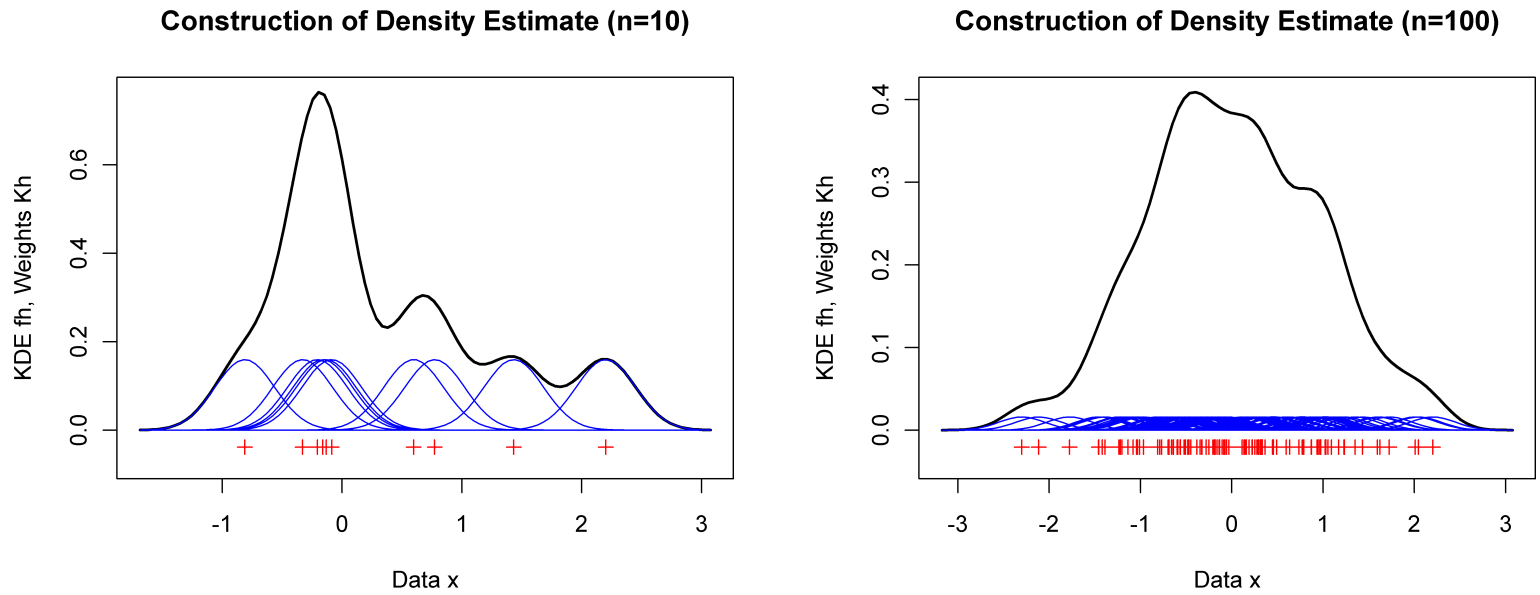
Different bin settings lead to different conclusions on the shape of the distribution

→ SPMhistogram

Kernel Density Estimation

→ concept of local averaging, different kernel functions, more detailed derivation of bias, variance and consequently MSE and MISE (Taylor expansions)

Construction of the kernel density estimator



Visualization of the kernel density estimate construction: density estimate as a sum of rescaled kernel functions (10 vs. 100 observations) → SPMkdeconstruct

Kernel Density Estimation (Cont'd)

Properties of nonparametric function estimators

→ with simulated data, theoretical propoertis of relevant terms can be also visualized (bias effects, MSE of the estimator)

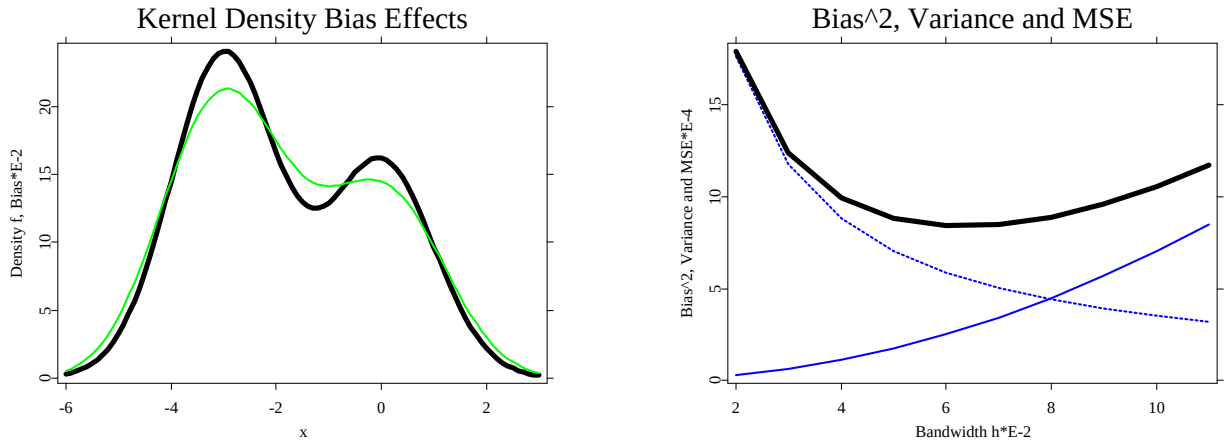
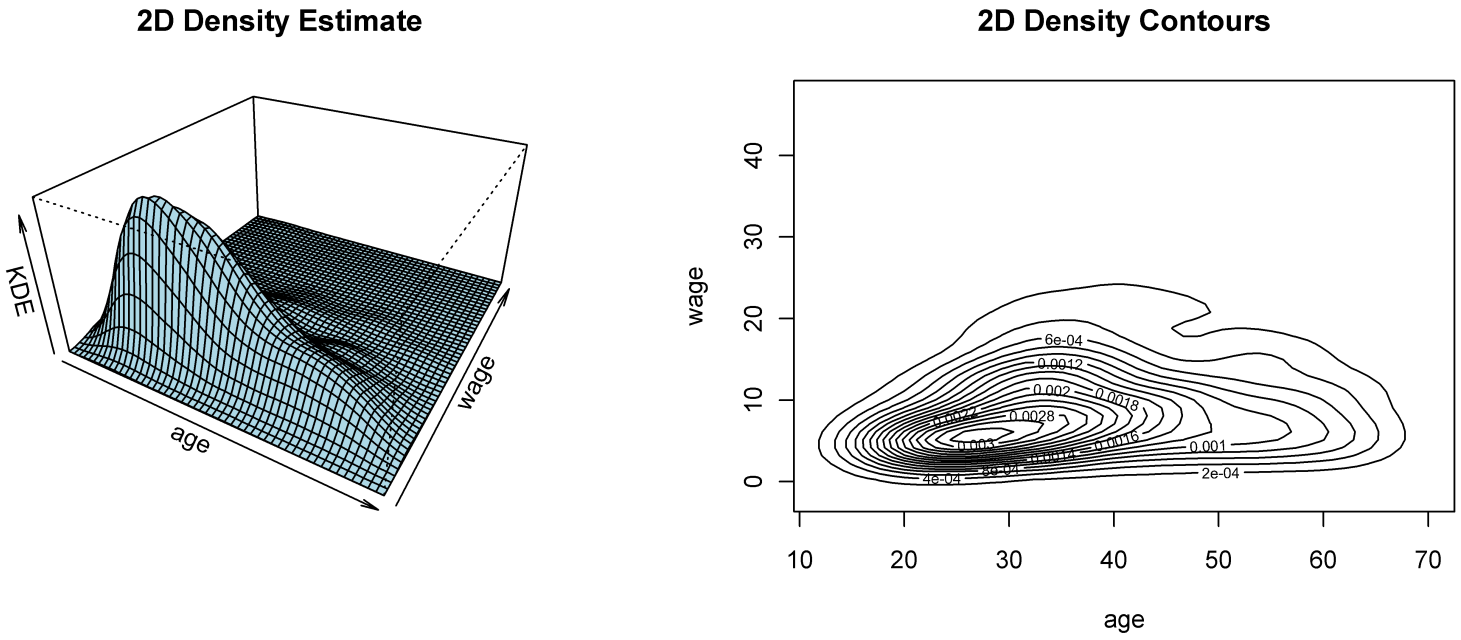


Illustration of the theoretical properties of the kernel density estimate (on the left: bias of the estimate and on the right: MSE as the sum of squared bias and variance)

Kernel Density Estimation (Cont'd)

Properties of nonparametric function estimators

→ two- and multidimensional data can be fitted by straightforward generalization, however there is the "curse of dimensionality" and the issue of graphical representation



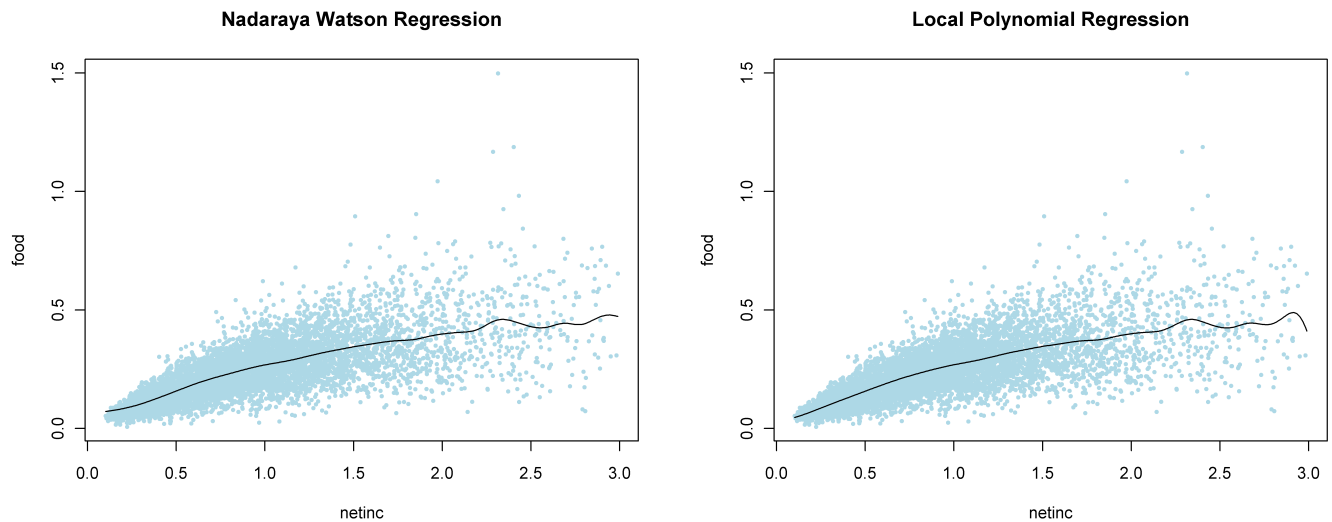
Graphical representation of 2D density estimates, perspective plot and contour curves

→ SPMdensity2D

Nonparametric Regression

→ Nadaraya-Watson regression as directly based on kernel density estimates, local polynomial estimators as generalizing the local constant Nadaraya-Watson estimator, comparison of different smoothing methods (regressograms, spline smoothing, k-nearest neignbor regression)

Construction of the kernel density estimator



Nadaraya-Watson regression vs. local linear regression → SPMkernelregression

Semiparametric Extensions to Generalized Linear Models

- → overview on semiparametric models: in particular additive models and generalized additive models, single index models, partial linear and generalized partial linear models
- → an introduction to the topics, discuss some applications and modelling assumptions (identification issues)
- → key model: the binary choice model makes it easy to understand why all these semiparametric models are useful to explore regression data

$$E(Y|X) = P(Y = 1|X) = F(\beta^\top X)$$

for example

$$P(Y = 1|X) = F(\beta_0 + \beta_{11}X_1 + \beta_{12}X_1^2 + \beta_{21}X_2 + \beta_{22}X_2^2 + \ldots)$$

using the logistic link function

$$F(u) = \frac{1}{1 + exp(-u)}$$

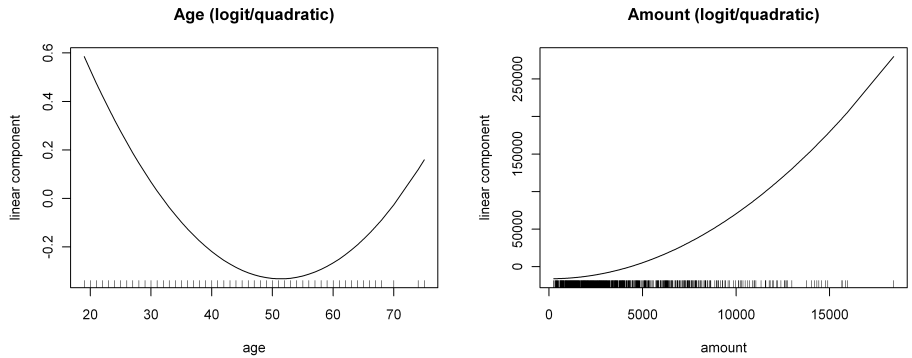
Semiparametric Extensions to Generalized Linear Models (Cont'd)

Example: Credit scoring

→ dependent variable Y: credit default (Y=1) or non-default (Y=0)

→ explanatory variables: describe loan characteristics (amount, maturity, purpose) and socio-economic chararacteristics of the credit applicants (age, employment and wealth variables, information on previous loans etc.)

Visualization of the effects of Age and Amount in the classical logit fit



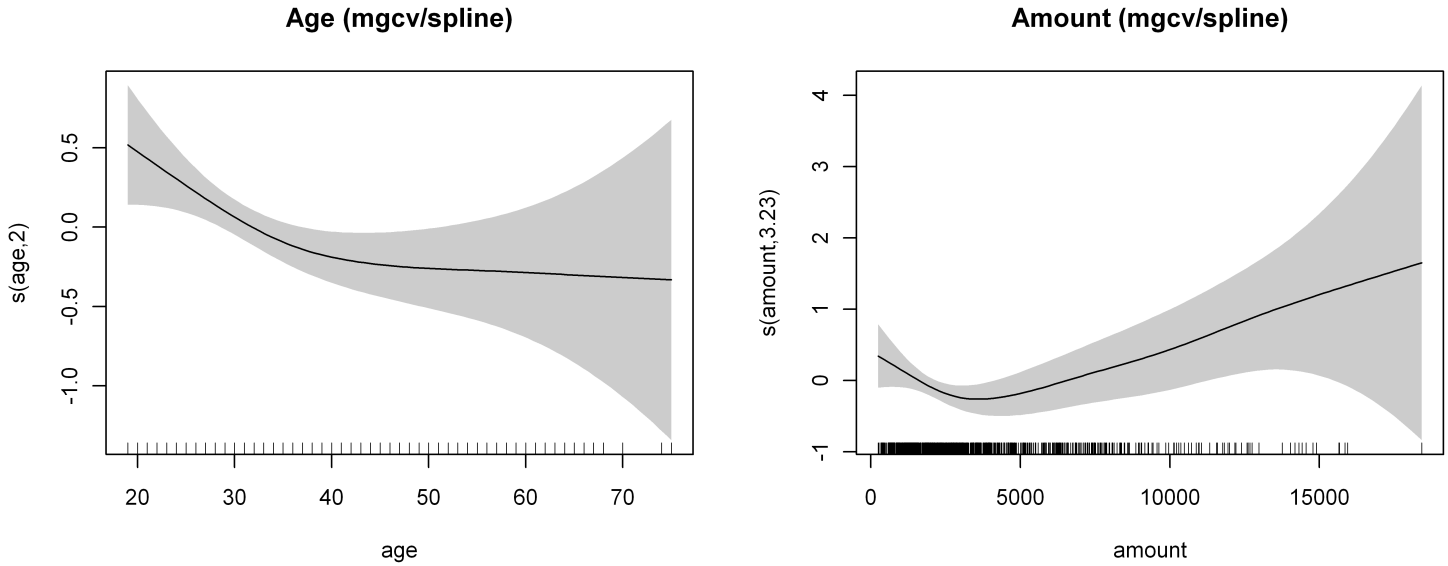
Effect of the variables Age and Amount in the linear predictor of the logit credit scoring model (both variables with up to quadratic terms) → SPMlogitkredit

Semiparametric Extensions to Generalized Linear Models (Cont'd)

Semiparametric alternative: GAM

→ generalized additive model with logistic link from the R package mgcv

$$P(Y = 1) = F(\beta_0 + m_1(Age) + m_2(Amount) + \ldots)$$



Effects of Age and Amount in the predictor of the semiparametric logit credit scoring model (both variables included as spline fitted components) → SPMgamkredit

Semiparametric Extensions to Generalized Linear Models (Cont'd)

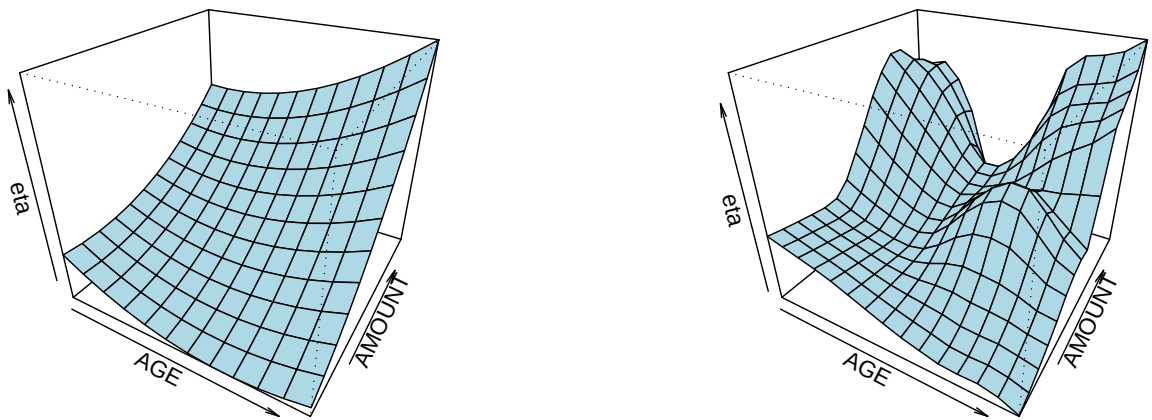
Semiparametric alternative: GPLM

$\rightarrow$ logit model with up to quadratic terms

$$P(Y = 1|X) = F(\beta_0 + \beta_{11}\text{Age} + \beta_{12}\text{Age}^2 + \beta_{21}\text{Amount} + \beta_{22}\text{Amount}^2 + \beta_{i12}\text{AgeAmount} + \ldots)$$

vs. generalized partial linear model

$$P(Y = 1) = F(\beta_0 + m_1(Age, Amount) + \ldots)$$



Joint effect of Age and Amount in parametric and semiparametric

Summary

- we propose a one-semester course that introduces to the most important concepts of nonparametric function estimation, target audience: master level students in economics at a university with strong quantitative focus
- objective is that students of applied sciences can apply these techniques (freely available in R, see www.R-project.org) to explore their data and to be able to assess these methods for their applicability
- the course is complemented by R scripts which can be individually altered or modified in order to see more features of the proposed estimators as well as to apply the methods on other data sets; interactive elements (sliders, 3D) are useful for demonstration and exploration but are not over-emphasized

  $\rightarrow$ http://www.marlenemueller.de/nspm.html

References I

Fan, J. & Gijbels, I. (1996). Local Polynomial Modelling and Its Applications, Chapman and Hall, New York.

Härdle, W. (1990): Applied Nonparametric Regression. Econometric Society Monographs No. 19, Cambridge University Press.

Härdle, W. (1991). Smoothing Techniques, With Implementations in S. Springer, New York.

Härdle, W., Müller, M., Sperlich, S. & Werwatz, A. (2004). Nonparametric and Semiparametric Models. Springer Series in Statistics, Springer, New York.

Hastie, T.J., Tibshirani, R.J. (1990). Generalized Additive Models. Chapman and Hall, London.

Müller, M. (2009). R material for “Nonparametric and Semiparametric Models”. Available at http://www.marlenemueller.de/nspm.html.

Pagan, A., Ullah, A. (1999). Nonparametric Econometrics. Cambridge University Press.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

References II

Ruppert, D., Wand, M.P. & Carroll, R.J. (2003). Semiparametric Regression. Cambridge University Press.

Scott, D. W. (1992). Multivariate Density Estimation: Theory, Practice, and Visualization. John Wiley & Sons, New York, Chichester.

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. Vol. 26 of Monographs on Statistics and Applied Probability, Chapman and Hall, London.

Wand, M.P. & Jones, M.C. (1995). Kernel Smoothing. Chapman and Hall, London.

Wood, S.N. (2006).Generalized Additive Models: An Introduction with R. Chapman and Hall, London.

Yatchew, A., (2003). Semiparametric Regression for Applied Econometrician. Cambridge University Press, Cambridge.