Wolfgang Härdle, Marlene Müller, Stefan Sperlich, Axel Werwatz

Nonparametric and Semiparametric Models

An Introduction

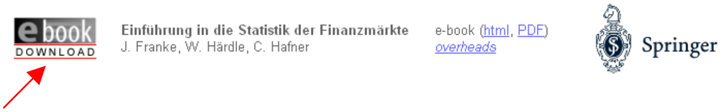
February 6, 2004

Springer

Berlin Heidelberg New York
Hong Kong London
Milan Paris Tokyo

Please note: this is only a sample of the full book. The complete book can be downloaded on the e-book page of XploRe. Just click the download logo:

http://www.xplore-stat.de/ebooks/ebooks.html



Einführung in die Statistik der Finanzmärkte
J. Franke, W. Härdle, C. Hafner

e-book (html, PDF)
overheads

download logo

For further information please contact MD*Tech at mdtech@mdtech.de

Preface

The concept of smoothing is a central idea in statistics. Its role is to extract structural elements of variable complexity from patterns of random variation. The nonparametric smoothing concept is designed to simultaneously estimate and model the underlying structure. This involves high dimensional objects, like density functions, regression surfaces or conditional quantiles. Such objects are difficult to estimate for data sets with mixed, high dimensional and partially unobservable variables. The semiparametric modeling technique compromises the two aims, flexibility and simplicity of statistical procedures, by introducing partial parametric components. These (low dimensional) components allow one to match structural conditions like for example linearity in some variables and may be used to model the influence of discrete variables. The flexibility of semiparametric modeling has made it a widely accepted statistical technique.

The aim of this monograph is to present the statistical and mathematical principles of smoothing with a focus on applicable techniques. The necessary mathematical treatment is easily understandable and a wide variety of interactive smoothing examples are given. This text is an e-book; it is a downloadable entity (http://www.i-xplore.de) which allows the reader to recalculate all arguments and applications without reference to a specific software platform. This new technique for proliferation of methods and ideas is specifically designed for the beginner in nonparametric and semiparametric statistics. It is based on the XploRe quantlet technology, developed at Humboldt-Universität zu Berlin.

The text has evolved out of the courses “Nonparametric Modeling” and “Semiparametric Modeling”, that the authors taught at Humboldt-Universität zu Berlin, ENSAE Paris, Charles University Prague, and Universidad de Cantabria, Santander. The book divides itself naturally into two parts:

- **Part I: Nonparametric Models**
  histogram, kernel density estimation, nonparametric regression
- **Part ??: Semiparametric Models**
  generalized regression, single index models, generalized partial linear models, additive and generalized additive models.

The first part (Chapters 2–4) covers the methodological aspects of nonparametric function estimation for cross-sectional data, in particular kernel smoothing methods. Although our primary focus will be on flexible regression models, a closely related topic to consider is nonparametric density estimation. Since many techniques and concepts for the estimation of probability density functions are also relevant for regression function estimation, we first consider histograms (Chapter 2) and kernel density estimates (Chapter 3) in more detail. Finally, in Chapter 4 we introduce several methods of nonparametrically estimating regression functions. The main part of this chapter is devoted to kernel regression, but other approaches such as splines, orthogonal series and nearest neighbor methods are also covered.

The first part is intended for undergraduate students majoring in mathematics, statistics, econometrics or biometrics. It is assumed that the audience has a basic knowledge of mathematics (linear algebra and analysis) and statistics (inference and regression analysis). The material is easy to utilize since the e-book character of the text allows maximum flexibility in learning (and teaching) intensity.

The second part (Chapters 5–9) is devoted to semiparametric regression models, in particular extensions of the parametric generalized linear model. In Chapter 5 we summarize the main ideas of the generalized linear model (GLM). Typical concepts are the logit and probit models. Nonparametric extensions of the GLM consider either the link function (single index models, Chapter 6) or the index argument (generalized partial linear models, additive and generalized additive models, Chapters 7–9). Single index models focus on the nonparametric error distribution in an underlying latent variable model. Partial linear models take the pragmatic point of fixing the error distribution but let the index be of non- or semiparametric structure. Generalized additive models concentrate on a (lower dimensional) additive structure of the index with fixed link function. This model class balances the difficulty of high-dimensional smoothing with the flexibility of nonparametrics.

In addition to the methodological aspects, the second part also covers computational algorithms for the considered models. As in the first part we focus on cross-sectional data. It is intended to be used by Master and PhD students or researchers.

This book would not have been possible without substantial support from many colleagues and students. It has benefited at several stages from useful remarks and suggestions of our students at Humboldt-Universität zu Berlin, ENSAE Paris and Charles University Prague. We are grateful to Lorens Helmchen, Stephanie Freese, Danilo Mercurio, Thomas Kühn, Ying Chen and Michal Benko for their support in text processing and programming, Caroline Condron for language checking and Pavel Čížek, Zdeněk Hlávka and Rainer Schulz for their assistance in teaching. We are indebted to Joel Horowitz (Northwestern University), Enno Mammen (Universität Heidelberg) and Helmut Rieder (Universität Bayreuth) for their valuable comments on earlier versions of the manuscript. Thanks go also to Clemens Heine, Springer Verlag, for being a very supportive and helpful editor.

Berlin/Kaiserslautern/Madrid, February 2004

Wolfgang Härdle
Marlene Müller
Stefan Sperlich
Axel Werwatz

Contents

Part I Nonparametric Models

Part II Semiparametric Models

List of Figures

List of Tables

Notation

Abbreviations

cdf	cumulative distribution function
df	degrees of freedom
iff	if and only if
i.i.d.	independent and identically distributed
w.r.t.	with respect to
pdf	probability density function
ADE	average derivative estimator
AM	additive model
AMISE	asymptotic MISE
AMSE	asymptotic MSE
APLM	additive partial linear model
ASE	averaged squared error
ASH	average shifted histogram
CHARN	conditional heteroscedastic autoregressive nonlinear
CV	cross-validation
DM	Deutsche Mark
GAM	generalized additive model
GAPLM	generalized additive partial linear model
GLM	generalized linear model

GPLM	generalized partial linear model
ISE	integrated squared error
IRLS	iteratively reweighted least squares
LR	likelihood ratio
LS	least squares
MASE	mean averaged squared error
MISE	mean integrated squared error
ML	maximum likelihood
MLE	maximum likelihood estimator
MSE	mean squared error
PLM	partial linear model
PMLE	pseudo maximum likelihood estimator
RSS	residual sum of squares
S.D.	standard deviation
S.E.	standard error
SIM	single index model
SLS	semiparametric least squares
USD	US Dollar
WADE	weighted average derivative estimator
WSLS	weighted semiparametric least squares

Scalars, Vectors and Matrices

X, Y	random variables
x, y	scalars (realizations of X, Y)
$X_1, \ldots, X_n$	random sample of size n
$X_{(1)}, \ldots, X_{(n)}$	ordered random sample of size n
$x_1, \ldots, x_n$	realizations of $X_1, \ldots, X_n$
$\boldsymbol{X}$	vector of variables
$\boldsymbol{x}$	vector (realizations of $\boldsymbol{X}$)
x_0	origin (of histogram)

h	binwidth or bandwidth
$\widetilde{h}$	auxiliary bandwidth in marginal integration
$\mathbf{H}$	bandwidth matrix
$\mathbf{I}$	identity matrix
$\mathbf{X}$	data or design matrix
$\boldsymbol{Y}$	vector of observations $Y_1, \ldots, Y_n$
β	parameter
$\boldsymbol{\beta}$	parameter vector
$\boldsymbol{e}_0$	first unit vector, i.e. $\boldsymbol{e}_0 = (1, 0, \ldots, 0)^\top$
$\boldsymbol{e}_j$	$(j+1)$th unit vector, i.e. $\boldsymbol{e}_j = (0, \ldots, 0, \underset{j}{1}, 0, \ldots, 0)^\top$
$\mathbb{1}_n$	vector of ones of length n
$\boldsymbol{\mu}$	vector of expectations of $Y_1, \ldots, Y_n$ in generalized models
$\boldsymbol{\eta}$	vector of index values $\boldsymbol{X}_1^\top \beta, \ldots, \boldsymbol{X}_n^\top \beta$ in generalized models
LR	likelihood ratio test statistic
$\boldsymbol{U}$	vector of variables (linear part of the model)
$\boldsymbol{T}$	vector of continuous variables (nonparametric part of the model)
$\boldsymbol{X}_{\underline{\alpha}}$	random vector of all but αth component
$\boldsymbol{X}_{\underline{\alpha j}}$	random vector of all but αth and jth component
$\mathbf{S}$, $\mathbf{S}^P$, $\mathbf{S}_\alpha$	smoother matrices
$\boldsymbol{m}$	vector of regression values $m(\boldsymbol{X}_1), \ldots, m(\boldsymbol{X}_n)$
$\boldsymbol{g}_\alpha$	vector of additive component function values $g_\alpha(\boldsymbol{X}_1),\ldots,g_\alpha(\boldsymbol{X}_n)$

Matrix algebra

$\mathrm{tr}(\mathbf{A})$	trace of matrix $\mathbf{A}$
$\mathrm{diag}(\mathbf{A})$	diagonal of matrix $\mathbf{A}$
$\det(\mathbf{A})$	determinant matrix $\mathbf{A}$
$\mathrm{rank}(\mathbf{A})$	rank of matrix $\mathbf{A}$

$\mathbf{A}^{-1}$	inverse of matrix $\mathbf{A}$
$\|\boldsymbol{u}\|$	norm of vector $\boldsymbol{u}$, i.e. $\sqrt{\boldsymbol{u}^\top \boldsymbol{u}}$

Functions

log	logarithm (base e)
φ	pdf of standard normal distribution
Φ	cdf of standard normal distribution
I	indicator function, i.e. $\mathrm{I}(A) = 1$ if A holds, 0 otherwise
K	kernel function (univariate)
K_h	scaled kernel function, i.e. $K_h(u) = K(u/h)/h$
$\mathcal{K}$	kernel function (multivariate)
$\mathcal{K}_{\mathbf{H}}$	scaled kernel function, i.e. $\mathcal{K}_{\mathbf{H}}(\boldsymbol{u}) = \mathcal{K}(\mathbf{H}^{-1}\boldsymbol{u}) / \det(\mathbf{H})$
$\mu_2(K)$	second moment of K, i.e. $\int u^2 K(u)\,du$
$\mu_p(K)$	pth moment of K, i.e. $\int u^p K(u)\,du$
$\|K\|_2^2$	squared L_2 norm of K, i.e. $\int \{K(u)\}^2\,du$
f	probability density function (pdf)
f_X	pdf of X
$f(x,y)$	joint density of X and Y
∇_f	gradient vector (partial first derivatives)
$\mathcal{H}_f$	Hessian matrix (partial second derivatives)
$K \star K$	convolution of K, i.e. $K \star K(u) = \int K(u-v)K(v)\,dv$
$w, \widetilde{w}$	weight functions
m	unknown function (to be estimated)
$m^{(\nu)}$	νth derivative (to be estimated)
ℓ, ℓ_i	log-likelihood, individual log-likelihood
G	known link function
g	unknown link function (to be estimated)
a, b, c	exponential family characteristics in generalized models
V	variance function of Y in generalized models
g_α	additive component (to be estimated)

$g_\alpha^{(\nu)}$	νth derivative (to be estimated)
f_α	pdf of X_α

Moments

EX	mean value of X
$\sigma^2 = \text{Var}(X)$	variance of X, i.e. $\text{Var}(X) = E(X - EX)^2$
$E(Y\|X)$	conditional mean Y given X (random variable)
$E(Y\|X = x)$	conditional mean Y given $X = x$ (realization of $E(Y\|X)$)
$E(Y\|x)$	same as $E(Y\|X = x)$
$\sigma^2(x)$	conditional variance of Y given $X = x$ (realization of $\text{Var}(Y\|X)$)
$E_{X_1}g(X_1, X_2)$	mean of $g(X_1, X_2)$ w.r.t. X_1 only
$\text{med}(Y\|X)$	conditional median Y given X (random variable)
μ	same as $E(Y\|X)$ in generalized models
$V(\mu)$	variance function of Y in generalized models
ψ	nuisance (dispersion) parameter in generalized models
MSE_x	MSE at the point x
$\mathcal{P}_\alpha$	conditional expectation function $E(\bullet\|X_\alpha)$

Distributions

$U[0,1]$	uniform distribution on $[0,1]$
$U[a,b]$	uniform distribution on $[a,b]$
$N(0,1)$	standard normal or Gaussian distribution
$N(\mu,\sigma^2)$	normal distribution with mean μ and variance σ^2
$N(\boldsymbol{\mu},\boldsymbol{\Sigma})$	multi-dimensional normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
χ^2_m	χ^2 distribution with m degrees of freedom
t_m	t-distribution with m degrees of freedom

Estimates

$\widehat{\beta}$	estimated coefficient
$\widehat{\boldsymbol{\beta}}$	estimated coefficient vector
$\widehat{f}_h$	estimated density function
$\widehat{f}_{h,-i}$	estimated density function when leaving out observation i
$\widehat{m}_h$	estimated regression function
$\widehat{m}_{p,h}$	estimated regression function using local polynomials of degree p and bandwidth h
$\widehat{m}_{p,\mathbf{H}}$	estimated multivariate regression function using local polynomials of degree p and bandwidth matrix $\mathbf{H}$

Convergence

$o(\bullet)$	$a = o(b)$ iff $a/b \to 0$ as $n \to \infty$ or $h \to 0$
$O(\bullet)$	$a = O(b)$ iff $a/b \to$ constant as $n \to \infty$ or $h \to 0$
$o_p(\bullet)$	$U = o_p(V)$ iff for all $\epsilon > 0$ holds $P(\|U/V\| > \epsilon) \to 0$
$O_p(\bullet)$	$U = O_p(V)$ iff for all $\epsilon > 0$ exists $c > 0$ such that $P(\|U/V\| > c) < \epsilon$ as n is sufficiently large or h is sufficiently small
$\xrightarrow{a.s.}$	almost sure convergence
$\xrightarrow{P}$	convergence in probability
$\xrightarrow{L}$	convergence in distribution
$\approx$	asymptotically equal
$\sim$	asymptotically proportional

Other

$\mathbb{N}$	natural numbers
$\mathbb{Z}$	integers
$\mathbb{R}$	real numbers

$\mathbb{R}^d$	d-dimensional real space
$\propto$	proportional
$\equiv$	constantly equal
#	number of elements of a set
B_j	jth bin, i.e. $[x_0 + (j-1)h, x_0 + jh)$
m_j	bin center of B_j, i.e. $m_j = x_0 + (j - \frac{1}{2})h$

1

Introduction

1.1 Density Estimation

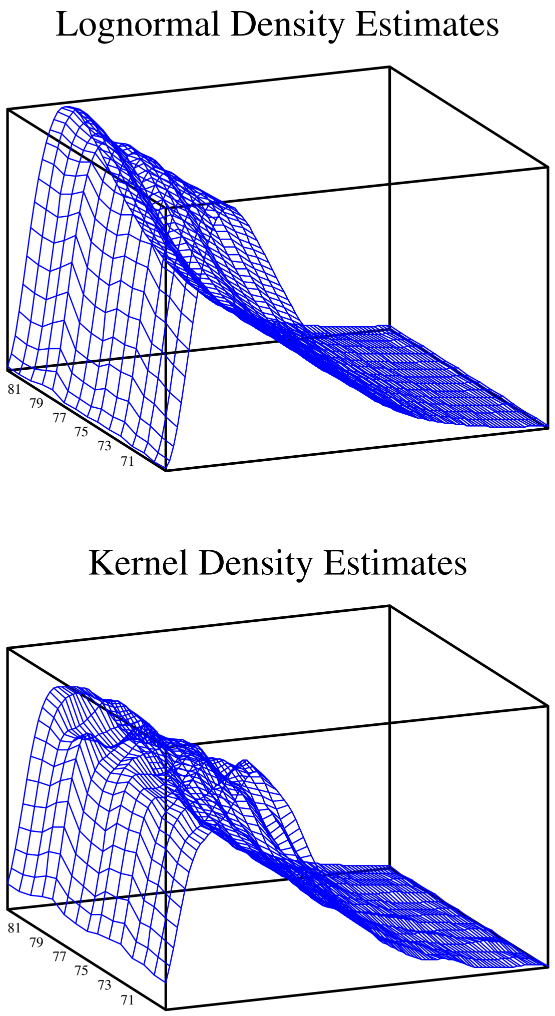
Consider a continuous random variable and its *probability density function* (pdf). The pdf tells you "how the random variable is distributed". From the pdf you cannot only calculate the statistical characteristics as mean and variance, but also the probability that this variable will take on values in a certain interval.

The pdf is, thus, very useful as it characterizes completely the "behavior" of a random variable. This fact might provide enough motivation to study nonparametric density estimation. Moreover nonparametric density estimates can serve as a building block in nonparametric regression estimation, as regression functions are fully characterized through the distribution of two (or more) variables.

The following example, which uses data from the Family Expenditure Survey of each year from 1969 to 1983, gives some illustration of the fact that density estimation has a substantial application in its own right.

Example 1.1.
Imagine that we have to answer the following questions: Is there a change in the structure of the income distribution during the period from 1969 to 1983? (You may recall, that many people argued that the neo-liberal policies of former Prime Minister Margaret Thatcher promoted income inequality in the early 1980s.)

To answer this question, we have estimated the distribution of net-income for each year from 1969 to 1983 both parametrically and nonparametrically. In parametric estimation of the distribution of income we have followed standard practice by fitting a log-normal distribution to the data. We employed the method of kernel density estimation (a generalization of the familiar histogram, as we will soon see) to estimate the income distribution nonparametrically. In the upper graph in Figure 1.1 we have plotted the estimated log-normal densities for each of the 15 years: Note that they are all very similar. On the other hand the analogous plot of the kernel density estimates show a movement of the net-income mode (the maximum of the den-


Lognormal Density Estimates

Kernel Density Estimates

Figure 1.1. Log-normal density estimates (upper graph) versus kernel density estimates (lower graph) of net-income, U.K. Family Expenditure Survey 1969–83 SPMfesdensities

sity) to the left (Figure 1.1, lower graph). This indicates that the net-income distribution has in fact changed during this 15 year period. □

1.2 Regression

Let us now consider a typical linear regression problem. We assume that anyone of you has been exposed to the linear regression model where the mean of a dependent variable Y is related to a set of explanatory variables $X_1, X_2, \ldots, X_d$ in the following way:

$$E(Y|\boldsymbol{X}) = X_1\beta_1 + \ldots + X_d\beta_d = \boldsymbol{X}^\top \boldsymbol{\beta}. \tag{1.1}$$

Here $E(Y|\boldsymbol{X})$ denotes the expectation conditional on the vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)^\top$ and β_j, $j = 1, 2, \ldots, d$ are unknown coefficients. Defining ε as the deviation of Y from the conditional mean $E(Y|\boldsymbol{X})$:

$$\varepsilon = Y - E(Y|\boldsymbol{X}) \tag{1.2}$$

we can write

$$Y = \boldsymbol{X}^\top \boldsymbol{\beta} + \varepsilon. \tag{1.3}$$

Example 1.2.
To take a specific example, let Y be *log wages* and consider the explanatory variables *schooling* (measured in years), labor market *experience* (measured as AGE − SCHOOL − 6) and *experience squared*. If we assume that, on average, log wages are linearly related to these explanatory variables then the linear regression model applies:

$$E(Y|\text{SCHOOL}, \text{EXP}) = \beta_0 + \beta_1 \cdot \text{SCHOOL} + \beta_2 \cdot \text{EXP} + \beta_3 \cdot \text{EXP}^2. \tag{1.4}$$

Note that we have included an intercept (β_0) in the model. □

The model of equation (1.4) has played an important role in empirical labor economics and is often called *human capital earnings equation* (or *Mincer earnings equation* to honor Jacob Mincer, a pioneer of this line of research). From the perspective of this course, an important characteristic of equation (1.4) is its *parametric* form: the shape of the regression function is governed by the unknown parameters β_j, $j = 1, 2, \ldots, d$. That is, all we have to do in order to determine the linear regression function (1.4) is to estimate the unknown parameters β_j. On the other hand, the parametric regression function of equation (1.4) a priori rules out many conceivable nonlinear relationships between Y and $\boldsymbol{X}$.

Let $m(\text{SCHOOL}, \text{EXP})$ be the true, unknown regression function of log wages on schooling and experience. That is,

$$E(Y|\text{SCHOOL}, \text{EXP}) = m(\text{SCHOOL}, \text{EXP}). \tag{1.5}$$

Suppose that you were assigned the following task: estimate the regression of log wages on schooling and experience as accurately as possible in *one* trial. That is, you are not allowed to change your model if you find that the initial specification does not fit the data well. Of course, you could just go ahead and assume, as we have done above, that the regression you are supposed to estimate has the form specified in (1.4). That is, you assume that

$$m(\text{SCHOOL}, \text{EXP}) = \beta_1 + \beta_2 \cdot \text{SCHOOL} + \beta_3 \cdot \text{EXP} + \beta_4 \cdot \text{EXP}^2,$$

and estimate the unknown parameters by the method of ordinary least squares, for example. But maybe you would not fit this parametric model if we told you that there are ways of estimating the regression function without having to make *any* prior assumptions about its functional form (except that it is a smooth function). Remember that you have just one trial and if the form of $m(\text{SCHOOL}, \text{EXP})$ is very different from (1.4) then estimating the parametric model may give you very inaccurate results.

It turns out that there are indeed ways of estimating $m(\bullet)$ that merely assume that $m(\bullet)$ is a smooth function. These methods are called *nonparametric* regression estimators and part of this course will be devoted to studying nonparametric regression.

Nonparametric regression estimators are very flexible but their statistical precision decreases greatly if we include several explanatory variables in the model. The latter caveat has been appropriately termed *the curse of dimensionality*. Consequently, researchers have tried to develop models and estimators which offer more flexibility than standard parametric regression but overcome the curse of dimensionality by employing some form of *dimension reduction*. Such methods usually combine features of parametric and nonparametric techniques. As a consequence, they are usually referred to as *semiparametric* methods. Further advantages of semiparametric methods are the possible inclusion of categorical variables (which can often only be included in a parametric way), an easy (economic) interpretation of the results, and the possibility of a part specification of a model.

In the following three sections we use the earnings equation and other examples to illustrate the distinctions between parametric, nonparametric and semiparametric regression and we certainly hope that this will whet your appetite for the material covered in this course.

1.2.1 Parametric Regression

Versions of the human capital earnings equation of (1.4) have probably been estimated by more researchers than any other model of empirical economics. For a detailed nontechnical and well-written discussion see Berndt (1991, Chapter 5). Here, we want to point out that:

- Under certain simplifying assumptions, β_2 accurately measures the rate of return to schooling.
- Human capital theory suggests a concave wage-experience profile: rapid human capital accumulation in the early stage of one's labor market career, with rising wages that peak somewhere during midlife and decline thereafter as hours worked and the incentive to invest in human capital decrease. This is the reason for including both EXP and EXP^2 in the model. In order to get a profile as the one envisaged by theory, the estimated value of β_3 should be positive and that of β_4 should be negative.

Table 1.1. Results from OLS estimation for Example 1.2

Dependent Variable: Log Wages			
Variable	Coefficients	S.E.	t-values
SCHOOL	0.0898	0.0083	10.788
EXP	0.0349	0.0056	6.185
EXP^2	–0.0005	0.0001	–4.307
constant	0.5202	0.1236	4.209
$R^2 = 0.24$, sample size $n = 534$			

We have estimated the coefficients of (1.4) using ordinary least squares (OLS), using a subsample of the 1985 Current Population Survey (CPS) provided by Berndt (1991). The results are given in Table 1.1.

The estimated rate of return to schooling is roughly 9%. Note that the estimated coefficients of EXP and EXP^2 have the signs predicted by human capital theory. The shape of the wage-schooling (a plot of SCHOOL vs. $0.0898\cdot$ SCHOOL) and wage-experience (a plot of EXP vs. $0.0349\cdot$ EXP $-$ $0.0005\cdot$ EXP^2) profiles are given in the left and right graphs of Figure 1.2, respectively.

The estimated wage-schooling relation is linear "by default" since we did not include SCHOOL^2, say, to allow for some kind of curvature within the parametric framework. By looking at Figure 1.2 it is clear that the estimated coefficients of EXP and EXP^2 imply the kind of concave wage-earnings profile predicted by human capital theory.

We have also plotted a graph (Figure 1.3) of the estimated regression surface, i.e. a plot that has the values of the estimated regression function (obtained by evaluating $0.0898 \cdot \text{SCHOOL} + 0.0349 \cdot \text{EXP} - 0.0005 \cdot \text{EXP}^2$ at the observed combinations of schooling and experience) on the vertical axis and schooling and experience on the horizontal axes.

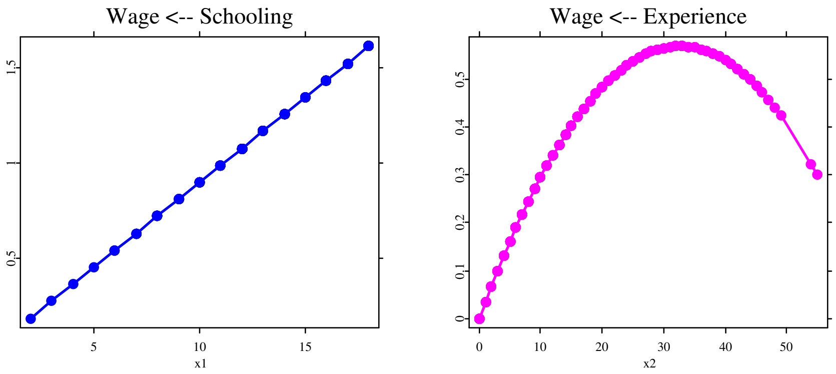


Figure 1.2. Wage-schooling and wage-experience profile SPMcps85lin

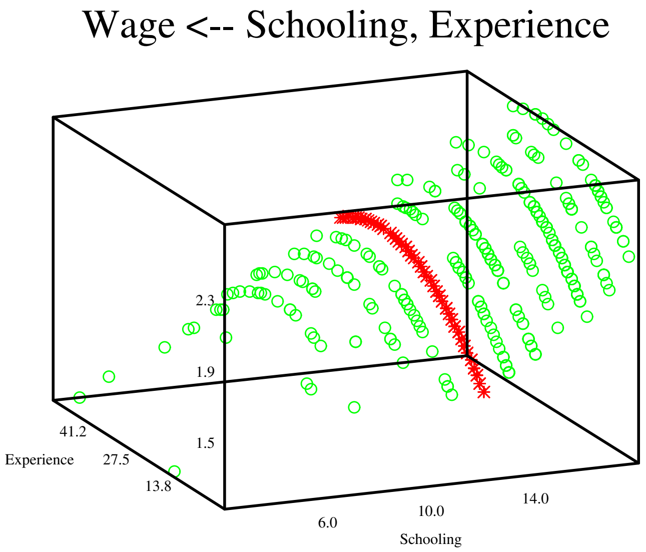


Figure 1.3. Parametrically estimated regression function SPMcps85lin

All of the element curves of the surface appear similar to Figure 1.2 (right) in the direction of experience and like Figure 1.2 (left) in the direction of schooling. To gain a better understanding of the three-dimensional picture we have plotted a single wage-experience profile in three dimensions, fixing schooling at 12 years. Hence, Figure 1.3 highlights the wage-earnings profile for high school graduates.

1.2.2 Nonparametric Regression

Suppose that we want to estimate

$$E(Y|\text{SCHOOL}, \text{EXP}) = m(\text{SCHOOL}, \text{EXP}). \tag{1.6}$$

and we are only willing to assume that $m(\bullet)$ is a smooth function. Nonparametric regression estimators produce an estimate of $m(\bullet)$ at an arbitrary point (SCHOOL $= s$, EXP $= e$) by *locally weighted averaging* over log wages (here s and e denote two arbitrary values that SCHOOL and EXP may take on, such as 12 and 15). Locally weighting means that those values of log wages will be higher weighted for which the corresponding observations of EXP and SCHOOL are close to the point (s, e). Let us illustrate this principle with an example. Let $s = 8$ and $e = 7$ and suppose you can use the four observations given in Table 1.2 to estimate $m(8, 7)$:

Table 1.2. Example observations

Observation	log(WAGES)	SCHOOL	EXP
1	7.31	8	8
2	7.6	16	1
3	7.4	8	6
4	7.8	12	2

In nonparametric regression $m(8, 7)$ is estimated by averaging over the observed values of the dependent variable log wage. But not all values will be given the same weight. In our example, observation 1 will get the most weight since it has values of schooling and experience that are very close to the point where we want to estimate. This makes a lot of sense: if we want to estimate mean log wages for individuals with 8 years of schooling and 7 years of experience then the observed log wage of a person with 8 years of schooling and 8 years of experience seems to be much more informative than the observed log wage of a person with 12 years of schooling and 2 years of experience.

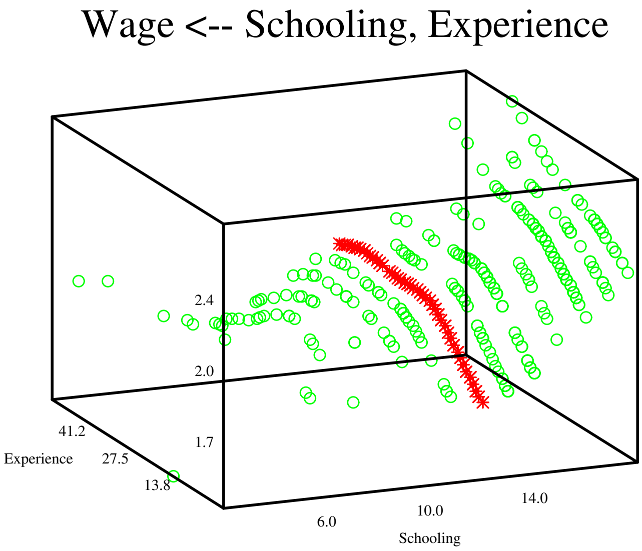


Figure 1.4. Nonparametrically estimated regression function SPMcps85reg

Consequently, any reasonable weighting scheme will give more weight to 7.31 than to 7.8 when we average over observed log wages. The exact method of weighting is determined by a weight function that makes precise the idea of weighting nearby observations more heavily. In fact, the weight function might be such that observations that are too far away get zero weight. In our example, observation 2 has values of experience and schooling that are so far away from 8 years of schooling and 7 years of experience that a weight function might assign zero value to the corresponding value of log wages (7.6). It is in this sense that the averaging is local. In Figure 1.4, the surface of nonparametrically estimated values of $m(\bullet)$ are shown. Here, a so-called kernel estimator has been used.

As long as we are dealing with only one regressor, the results of estimating a regression function nonparametrically can easily be displayed in a graph. The following example illustrates this. It relates net-income data, as we considered in Example 1.1, to a second variable that measures household expenditure.

Example 1.3.
Consider for instance the dependence of food expenditure on net-income. Figure 1.5 shows the so-called Engel curve (after the German Economist Engel) of net-income and food share estimated using data from the 1973 Family

Expenditure Survey of roughly 7000 British households. The figure supports the theory of Engel who postulated in 1857:

> ... je ärmer eine Familie ist, einen desto größeren Antheil von der Gesammtausgabe muß zur Beschaffung der Nahrung aufgewendet werden ... (The poorer a family, the bigger the share of total expenditure that has to be used for food.) □

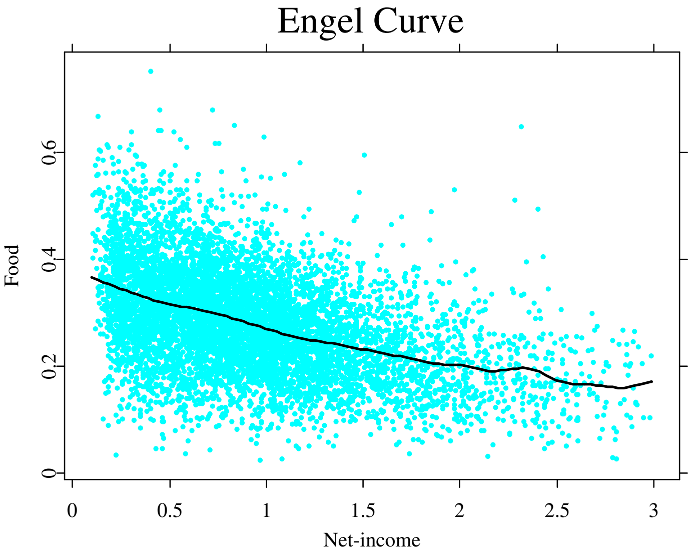


Figure 1.5. Engel curve, U.K. Family Expenditure Survey 1973 SPMengelcurve2

1.2.3 Semiparametric Regression

To illustrate semiparametric regression let us return to the human capital earnings function of Example 1.2. Suppose the regression function of log wages on schooling and experience has the following shape:

$$E(Y|\text{SCHOOL}, \text{EXP}) = \alpha + g_1(\text{SCHOOL}) + g_2(\text{EXP}). \tag{1.7}$$

Here $g_1(\bullet)$ and $g_2(\bullet)$ are two unknown, smooth functions and α is an unknown parameter. Note that this model combines the simple additive structure of the parametric regression model (referred to hereafter as the *additive* *model*) with the flexibility of the nonparametric approach. This is done by not imposing any strong shape restrictions on the functions that determine how schooling and experience influence the mean regression of log wages. The procedure employed to estimate this model will be explained in greater detail later in this course. It should be clear, however, that in order to estimate the unknown functions $g_1(\bullet)$ and $g_2(\bullet)$ nonparametric regression estimators have to be employed. That is, when estimating semiparametric models we usually have to use nonparametric techniques. Hence, we will have to spend a substantial amount of time studying nonparametric estimation if we want to understand how to estimate semiparametric models. For now, we want to focus on the results and compare them with the parametric fit.

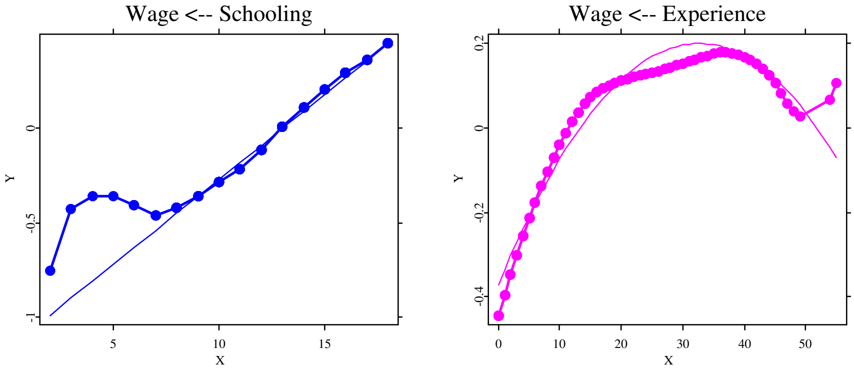


Figure 1.6. Additive model fit vs. parametric fit, wage-schooling (left) and wage-experience (right) SPMcps85add

In Figure 1.6 the parametrically estimated wage-schooling and wage-experience profiles are shown as thin lines whereas the estimates of $g_1(\bullet)$ and $g_2(\bullet)$ are displayed as thick lines with bullets. The parametrically estimated wage-school and wage-experience profiles show a good deal of similarity with the estimate of $g_1(\bullet)$ and $g_2(\bullet)$, except for the shape of the curve at extremal values. The good agreement between parametric estimates and additive model fit is also visible from the plot of the estimated regression surface, which is shown in Figure 1.7.

Hence, we may conclude that in this specific example the parametric model is supported by the more flexible nonparametric and semiparametric methods. This potential usefulness of nonparametric and semiparametric techniques for checking the adequacy of parametric models will be illustrated in several other instances in the latter part of this course.

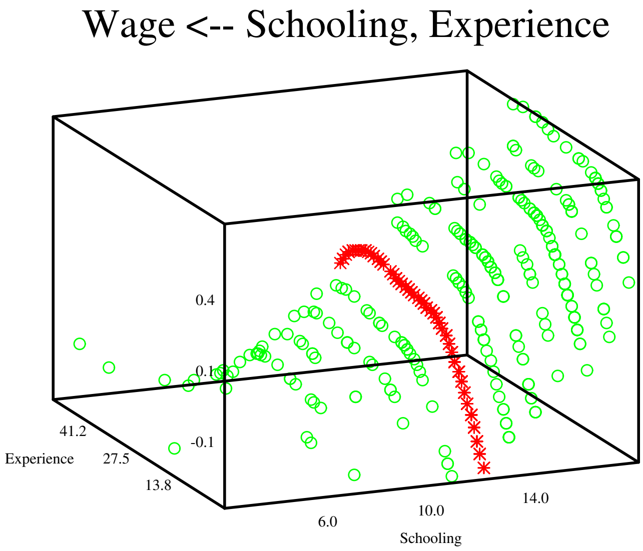


Figure 1.7. Surface plot for the additive model SPMcps85add

Take a closer look at (1.6) and (1.7). Observe that in (1.6) we have to estimate one unknown function of two variables whereas in (1.7) we have to estimate two unknown functions, each a function of one variable. It is in this sense that we have reduced the dimensionality of the estimation problem. Whereas all researchers might agree that additive models like the one in (1.7) are achieving a dimension reduction over completely nonparametric regression, they may not agree to call (1.7) a semiparametric model, as there are no parameters to estimate (except for the intercept parameter α). In the following example we confront a standard parametric model with a more flexible model that, as you will see, truly deserves to be called semiparametric.

Example 1.4.
In the earnings-function example, the dependent variable log wages can principally take on *any* positive value, i.e. the set of values Y is infinite. This may not always be the case. For example, consider the decision of an East-German resident to move to Western Germany and denote the decision variable by Y. In this case, the dependent variable can take on only *two* values,

$$Y = \begin{cases} 1 & \text{if the person can imagine moving to the west,} \\ 0 & \text{otherwise.} \end{cases}$$

We will refer to this as a *binary response* later on. □

In Example 1.2 we tried to estimate the effect of a person's education and work experience on the log wage earned. Now, say we want to find out how these two variables affect the decision of an East German resident to move west, i.e. we want to know $E(Y|\boldsymbol{x})$ where $\boldsymbol{x}$ is a $(d \times 1)$ vector containing all d variables considered to be influential to the migration decision. Since Y is a binary variable (i.e. a Bernoulli distributed variable), we have that

$$E(Y|\boldsymbol{X}) = P(Y = 1|\boldsymbol{X}). \tag{1.8}$$

Thus, the regression of Y on $\boldsymbol{X}$ can be expressed as the probability that a randomly sampled person from the East will migrate to the West, given this person's characteristics collected in the vector $\boldsymbol{X}$. Standard models for $P(Y = 1|\boldsymbol{X})$ assume that this probability depends on $\boldsymbol{X}$ as follows:

$$P(Y = 1|\boldsymbol{X}) = G(\boldsymbol{X}^\top \boldsymbol{\beta}), \tag{1.9}$$

where $\boldsymbol{X}^\top \boldsymbol{\beta}$ is a linear combination of all components of $\boldsymbol{X}$. It aggregates the multiple characteristics of a person into one number (therefore called the *index function* or simply the *index*), where $\boldsymbol{\beta}$ is an unknown vector of coefficients. $G(\bullet)$ denotes any continuous function that maps the real line to the range of $[0, 1]$. $G(\bullet)$ is also called the *link function*, since it links the index $\boldsymbol{X}^\top \boldsymbol{\beta}$ to the conditional expectation $E(Y|\boldsymbol{X})$.

In the context of this lecture, the crucial question is precisely *what* parametric form these two functions take or, more generally, whether they will take any parametric form *at all*. For now we want to compare two models: one that assumes that $G(\bullet)$ is of a known parametric form and one that allows $G(\bullet)$ to be an unknown smooth function.

One of the most widely used fully parametric models applied to the case of binary dependent variables is the *logit model*. The logit model assumes that $G(\boldsymbol{X}^\top \boldsymbol{\beta})$ is the (standard) logistic cumulative distribution function (cdf) for all $\boldsymbol{X}$. Hence, in this case

$$E(Y|\boldsymbol{X}) = P(Y = 1|\boldsymbol{X}) = \frac{1}{\exp(-\boldsymbol{X}^\top \boldsymbol{\beta})}. \tag{1.10}$$

Example 1.5.
In using a logit model, Burda (1993) estimated the effect of various explanatory variables on the migration decision of East German residents. The data for fitting this model were drawn from a panel study of approximately 4,000 East German households in spring 1991. We use a subsample of $n = 402$ observations from the German state "Mecklenburg-Vorpommern" here. Due to space constraints, we merely report the estimated coefficients of three components of the index $\boldsymbol{X}^\top \boldsymbol{\beta}$, as we will refer to these estimates below:

$$\begin{aligned} &\beta_0 + \beta_1 \cdot \text{INC} + \beta_2 \cdot \text{AGE} \\ &= -2.2905 + 0.0004971 \cdot \text{INC} - 0.45499 \cdot \text{AGE} \end{aligned} \tag{1.11}$$

INC and AGE are used to abbreviate the household income and age of the individual. □

Figure 1.8 gives a graphical presentation of the results. Each observation is represented by a "+". As mentioned above, the characteristics of each person are transformed into an index (to be read off the horizontal axis) while the dependent variable takes on one of two values, $Y = 0$ or $Y = 1$ (to be read off the vertical axis). The curve plots estimates of $P(Y = 1|\boldsymbol{X})$, the probability of $Y = 1$ as a function of $\boldsymbol{X}^\top \boldsymbol{\beta}$. Note that the estimates of $P(Y = 1|\boldsymbol{X})$, by assumption, are simply points on the cdf of a standard logistic distribution.

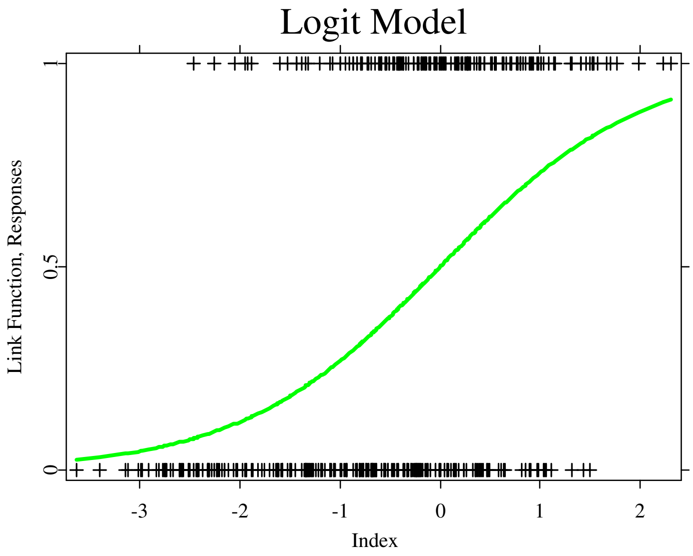


Figure 1.8. Logit fit SPMlogit

We shall continue with Example 1.4 below, but let us pause for a moment to consider the following substantial problem: the logit model, like other parametric models, is based on rather strong functional form (linear index) and distributional assumptions, neither of which are usually justified by economic theory.

The first question to ask before developing alternatives to standard models like the logit model is: what are the consequences of estimating a logit model if one or several of these assumptions are violated? Note that this is a crucial question: if our parametric estimates are largely unaffected by model violations, then there is no need to develop and apply semiparametric models and estimators. Why would anyone put time and effort into a project that promises little return?

One can employ the tools of asymptotic statistical theory to show that violating the assumptions of the logit model leads parameter estimates to being inconsistent. That is, if the sample size goes to infinity, the logit maximum-likelihood estimator (logit-MLE) does not converge to the true parameter value in probability. While it doesn't converge to the true parameter value it does, however, converge to some other value. If this "false" value is close enough to the true parameter value then we may not care very much about this inconsistency.

Consistency is an asymptotic criterion for the performance of an estimator. That is, it looks at the properties of the estimator if the sample size grows without limits. Yet, in practice, we are dealing with finite samples. Unfortunately, the finite-sample properties of the logit maximum-likelihood estimator can not be derived analytically. Hence, we have to rely on simulations to collect evidence of its small-sample performance in the presence of misspecification. We conducted a small simulation in the context of Example 1.4 to which we now return.

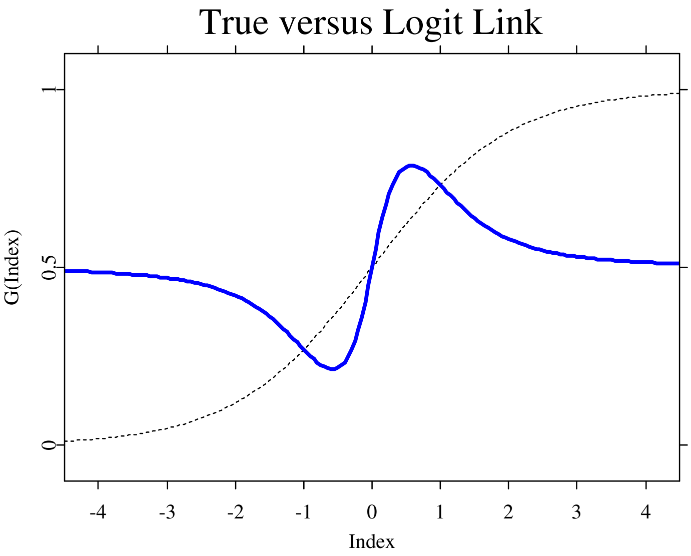


Figure 1.9. Link function of the homoscedastic logit model (thin line) versus the link function of the heteroscedastic model (solid line) SPMtruelogit

Example 1.6.
Following Horowitz (1993) we generated data according to a heteroscedastic model with two explanatory variables, INC and AGE. Here we considered heteroscedasticity of the form

$$\mathrm{Var}(\varepsilon|\boldsymbol{X}=x)=\frac{1}{4}\left\{1+(\boldsymbol{x}^\top\boldsymbol{\beta})^2\right\}^2\cdot\mathrm{Var}(\zeta),$$

where ζ has a (standard) logistic distribution. To give you an impression of how dramatically the *true* heteroscedastic model differs from the *supposed* homoscedastic logit model, we plotted the link functions of the two models as shown in Figure 1.9. □

To add a sense of realism to the simulation, we set the coefficients of these variables equal to the estimates reported in (1.11). Note that the standard logit model introduced above does not allow for heteroscedasticity. Hence, if we apply the standard logit maximum-likelihood estimator to the simulated data, we are estimating under misspecification. We performed 250 replications of this estimation experiment, using the full data set with 402 observations each time. As the estimated coefficients are only identified up to scale, we compared the ratio of the true coefficients, β_{INC}/β_{AGE}, to the ratio of their estimated logit-MLE counterparts, $\widehat{\beta}_{INC}/\widehat{\beta}_{AGE}$. Figure 1.10 shows the sampling distribution of the logit-MLE coefficients, along with the true value (vertical line).

As we have subtracted the true value from each estimated ratio and divided this difference by the true ratio's absolute value, the true ratio is standardized to zero and differences on the horizontal axis can be interpreted as percentage deviations from the truth. In Figure 1.10, the sampling distribution of the estimated ratios is centered around -0.11 which is the percentage deviation from the truth of 11%. Hence, the logit-MLE underestimates the true value.

Now that we have seen how serious the consequences of model misspecification can be, we might want to learn about semiparametric estimators that have desirable properties under more general assumptions than their parametric counterparts. One way to generalize the logit model is the so-called *single index model* (SIM) which keeps the linear form of the index $\boldsymbol{X}^\top\boldsymbol{\beta}$ but allows the function $G(\bullet)$ in (1.9) to be an arbitrary smooth function $g(\bullet)$ (not necessarily a distribution function) that has to be estimated from the data:

$$E(Y|\boldsymbol{X})=g(\boldsymbol{X}^\top\boldsymbol{\beta}), \tag{1.12}$$

Estimation of the single index model (1.12) proceeds in two steps:

- Firstly, the coefficient vector $\boldsymbol{\beta}$ has to be estimated. Methods to calculate the coefficients for discrete and continuous variables will be covered in depth later.

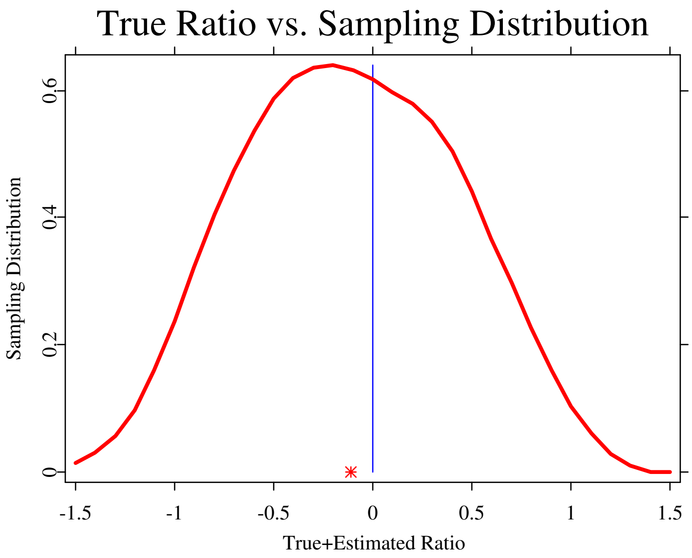


Figure 1.10. Sampling distribution of the ratio of the estimated coefficients (density estimate and mean value indicated as $*$) and the ratio's true value (vertical line) SPMsimulogit

- Secondly, we have to estimate the unknown link function $g(\bullet)$ by nonparametrically regressing the dependent variable Y on the fitted index $\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}$ where $\widehat{\boldsymbol{\beta}}$ is the coefficient vector we estimated in the first step. To do this, we use again a nonparametric estimator, the kernel estimator we mentioned briefly above.

Example 1.7.
Let us consider what happens if we use $\widehat{\boldsymbol{\beta}}$ from the logit fit and estimate the link function nonparametrically. Figure 1.11 shows this estimated link function. As before, the position of a + sign represents at the same time the values of $\boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}$ and Y of a particular observation, while the curve depicts the estimated link function. □

One additional remark should be made here: As you will soon learn, the shape of the estimated link function (the curve) varies with the so-called bandwidth, a parameter central in nonparametric function estimation. Thus, there is no unique estimate of the link function, and it is a crucial (and difficult) problem of nonparametric regression to find the "best" bandwidth and thus the optimal estimate. Fortunately, there are methods to select an ap-

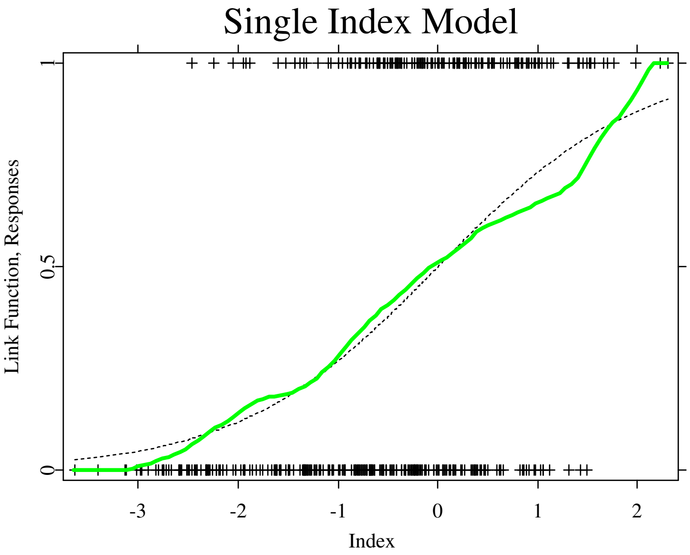


Figure 1.11. Single index versus logit model SPMsim

propriate bandwidth. Here, we have chosen $h = 0.7$ "index units" for the bandwidth. For comparison the shapes of both the single index (solid line) and the logit (dashed line) link functions are shown ins in Figure 1.8. Even though not identical they look rather similar.

Summary

- ⋆ Parametric models are fully determined up to a parameter (vector). The fitted models can easily be interpreted and estimated accurately if the underlying assumptions are correct. If, however, they are violated then parametric estimates may be inconsistent and give a misleading picture of the regression relationship.

- ⋆ Nonparametric models avoid restrictive assumptions of the functional form of the regression function m. However, they may be difficult to interpret and yield inaccurate estimates if the number of regressors is large.

- ⋆ Semiparametric models combine components of parametric and nonparametric models, keeping the easy interpretability of the former and retaining some of the flexibility of the latter.

5

Semiparametric and Generalized Regression Models

In the previous part of this book we found the curse of dimensionality to be one of the major problems that arises when using *nonparametric multivariate* regression techniques. For the practitioner, a further problem is that for more than two regressors, graphical illustration or interpretation of the results is hardly ever possible. Truly multivariate regression models are often far too flexible and general for making detailed inference.

5.1 Dimension Reduction

Researchers have looked for possible remedies, and a lot of effort has been allocated to developing methods which reduce the complexity of high dimensional regression problems. This refers to the reduction of dimensionality as well as allowance for partly parametric modeling. Not surprisingly, one follows the other. The resulting models can be grouped together as so-called *semiparametric* models.

All models that we will study in the following chapters can be motivated as generalizations of well-known parametric models, mainly of the linear model

$$E(Y|\boldsymbol{X}) = m(\boldsymbol{X}) = \boldsymbol{X}^\top \boldsymbol{\beta}$$

or its generalized version

$$E(Y|\boldsymbol{X}) = m(\boldsymbol{X}) = G\{\boldsymbol{X}^\top \boldsymbol{\beta}\}\,. \tag{5.1}$$

Here G denotes a known function, $\boldsymbol{X}$ is the d-dimensional vector of regressors and $\boldsymbol{\beta}$ is a coefficient vector that is to be estimated from observations for Y and $\boldsymbol{X}$.

Let us take a closer look at model (5.1). This model is known as the generalized linear model. Its use and estimation are extensively treated in McCullagh & Nelder (1989). Here we give only some selected motivating examples.

What is the reason for introducing this functional G, called the link? (Note that other authors call its inverse G^{-1} the link.) Clearly, if G is the identity we are back in the classical linear model. As a first alternative let us consider a quite common approach for investigating growth models. Here, the model is often assumed to be multiplicative instead of additive, i.e.

$$Y = \prod_{j=1}^{d} X_j^{\beta_j} \cdot \varepsilon, \quad E \log(\varepsilon) = 0 \tag{5.2}$$

in contrast to

$$Y = \prod_{j=1}^{d} X_j^{\beta_j} + \xi, \quad E\xi = 0. \tag{5.3}$$

Depending on whether we have multiplicative errors ε or additive errors ξ, we can transform model (5.2) to

$$E\{\log(Y)|\boldsymbol{X}\} = \sum_{j=1}^{d} \beta_j \log(X_j) \tag{5.4}$$

and model (5.3) to

$$E(Y|\boldsymbol{X}) = \exp\left\{\sum_{j=1}^{d} \beta_j \log(X_j)\right\}. \tag{5.5}$$

Considering now $\log(\boldsymbol{X})$ as the regressor instead of $\boldsymbol{X}$, equation (5.5) is equivalent to (5.1) with $G(\bullet) = \exp(\bullet)$. Equation (5.4), however, is a transformed model, see the bibliographic notes for references on this model family.

The most common cases in which link functions are used are binary responses ($Y \in \{0,1\}$) or multicategorical ($Y \in \{0,1,\ldots,J\}$) responses and count data ($Y \sim$ Poisson). For the binary case, let us introduce an example that we will study in more detail in Chapters 7 and 9.

Example 5.1.
Imagine we are interested in possible determinants of the migration decision of East Germans to leave the East for West Germany. Think of Y^* as being the net-utility from migrating from the eastern part of Germany to the western part. Utility itself is not observable but we can observe characteristics of the decision makers and the alternatives that affect utility. As Y^* is not observable it is called a latent variable. Let the observable characteristics

Table 5.1. Descriptive statistics for migration data, $n = 3235$

		Yes	No	(in %)	
Y	MIGRATION INTENTION	38.5	61.5		
X_1	FAMILY/FRIENDS IN WEST	85.6	11.2		
X_2	UNEMPLOYED/JOB LOSS CERTAIN	19.7	78.9		
X_3	CITY SIZE 10,000–100,000	29.3	64.2		
X_4	FEMALE	51.1	49.8		
		Min	Max	Mean	S.D.
X_5	AGE (in years)	18	65	39.84	12.61
X_6	HOUSEHOLD INCOME (in DM)	200	4000	2194.30	752.45

be summarized in a vector $\boldsymbol{X}$. This vector $\boldsymbol{X}$ may contain variables such as education, age, sex and other individual characteristics. A selection of such characteristics is shown in Table 5.1. □

In Example 5.1, we hope that the vector of regressors $\boldsymbol{X}$ captures the variables that systematically affect each person's utility whereas unobserved or random influences are absorbed by the term ε. Suppose further, that the components of $\boldsymbol{X}$ influence net-utility through a multivariate function $v(\bullet)$ and that the error term is additive. Then the latent-variable model is given by

$$Y^* = v(\boldsymbol{X}) - \varepsilon \quad \text{and} \quad Y = \begin{cases} 1 & \text{if } Y^* > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{5.6}$$

Hence, what we really observe is the binary variable Y that takes on the value 1 if net-utility is positive (person intends to migrate) and 0 otherwise (person intends to stay). Then some calculations lead to

$$P(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = E(Y \mid \boldsymbol{X} = \boldsymbol{x}) = G_{\varepsilon|\boldsymbol{x}}\{v(\boldsymbol{x})\} \tag{5.7}$$

with $G_{\varepsilon|\boldsymbol{x}}$ being the cdf of ε conditional on $\boldsymbol{x}$.

Recall that standard parametric models assume that ε is independently distributed of $\boldsymbol{X}$ with known distribution function $G_{\varepsilon|\boldsymbol{x}} = G$, and that the index $v(\bullet)$ has the following simple form:

$$v(\boldsymbol{x}) = \beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta}. \tag{5.8}$$

The most popular distribution assumptions regarding the error are the normal and the logistic ones, leading to the so-called *probit* or *logit* models with $G(\bullet) = \Phi(\bullet)$ (Gaussian cdf), respectively $G(\bullet) = \exp(\bullet)/\{1 + \exp(\bullet)\}$. We will learn how to estimate the coefficients β_0 and $\boldsymbol{\beta}$ in Section 5.2.

The binary choice model can be easily extended to the multicategorical case, which is usually called *discrete choice model*. We will not discuss extensions for multicategorical responses here. Some references for these models are mentioned in the bibliographic notes.

Several approaches have been proposed to reduce dimensionality or to generalize parametric regression models in order to allow for nonparametric relationships. Here, we state three different approaches:

- variable selection in nonparametric regression,
- generalization of (5.1) to a nonparametric link function,
- generalization of (5.1) to a semi- or nonparametric index,

which are discussed in more detail.

5.1.1 Variable Selection in Nonparametric Regression

The intention of variable selection is to choose an appropriate subset of variables, $\boldsymbol{X}_r = (X_{j_1}, \ldots, X_{j_r})^\top \in \boldsymbol{X} = (X_1, \ldots, X_d)^\top$, from the set of all variables that could potentially enter the regression. Of course, the selection of the variables could be determined by the particular problem at hand, i.e. we choose the variables according to insights provided by some underlying economic theory. This approach, however, does not really solve the statistical side of our modeling process. The curse of dimensionality could lead us to keep the number of variables as low as possible. On the other hand, fewer variables could in turn reduce the explanatory power of the model. Thus, after having chosen a set of variables on theoretical grounds in a first step, we still do not know how many and, more importantly, which of these variables will lead to optimal regression results. Therefore, a variable selection method is needed that uses a statistical selection criterion.

Vieu (1994) has proposed to use the integrated square error ISE to measure the quality of a given subset of variables. In theory, a subset of variables is defined to be an optimal subset if it minimizes the integrated squared error:

$$\mathrm{ISE}(\boldsymbol{X}_r^{opt}) = \min_{\boldsymbol{X}_r} \mathrm{ISE}(\boldsymbol{X}_r)$$

where $\boldsymbol{X}_r \subset \boldsymbol{X}$. In practice, the ISE is replaced by its sample analog, the multivariate analog of the cross validation function (3.38). After the variables have been selected, the conditional expectation of Y on $\boldsymbol{X}_r$ is calculated by some kind of standard nonparametric multivariate regression technique such as the kernel regression estimator.

5.1.2 Nonparametric Link Function

Index models play an important role in econometrics. An *index* is a summary of different variables into one number, e.g. the price index, the growth index, or the cost-of-living index. It is clear that by summarizing all the information contained in the variables $X_1, \dots, X_d$ into one "single index" term we will greatly reduce the dimensionality of a problem. Models based on such an index are known as *single index models* (SIM). In particular we will discuss single index models of the following form:

$$E(Y|\boldsymbol{X}) = m(\boldsymbol{X}) = g\left\{v_{\boldsymbol{\beta}}(\boldsymbol{X})\right\}, \tag{5.9}$$

where $g(\bullet)$ is an *unknown link* function and $v_{\boldsymbol{\beta}}(\bullet)$ an up to $\boldsymbol{\beta}$ specified index function. The estimation can be carried out in two steps. First, we estimate $\boldsymbol{\beta}$. Then, using the index values for our observations, we can estimate g by nonparametric regression. Note that estimating $g(\bullet)$ by regressing the Y on $v_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{X})$ is only a one-dimensional regression problem.

Obviously, (5.9) generalizes (5.7) in that we do not assume the link function G to be known. For that purpose we replaced G by g to emphasize that the link function needs to be estimated. Notice, that often the general index function $v_{\boldsymbol{\beta}}(\boldsymbol{X})$ is replaced by the linear index $\boldsymbol{X}^\top\boldsymbol{\beta}$. Equations (5.5) and (5.6) together with (5.8) give examples for such linear index functions.

5.1.3 Semi- or Nonparametric Index

In many applications a canonical partitioning of the explanatory variables exists. In particular, if there are categorical or discrete explanatory variables we may want to keep them separate from the other design variables. Note that only the continuous variables in the nonparametric part of the model cause the curse of dimensionality (Delgado & Mora, 1995). In the following chapters we will study the following models:

- *Additive Model* (AM)
  The standard additive model is a generalization of the multiple linear regression model by introducing one-dimensional nonparametric functions in the place of the linear components. Here, the conditional expectation of Y given $\boldsymbol{X} = (X_1, \dots, X_d)^\top$ is assumed to be the sum of unknown functions of the explanatory variables plus an intercept term:

  $$E(Y|\boldsymbol{X}) = c + \sum_{j=1}^{d} g_j(X_j) \tag{5.10}$$

  Observe how reduction is achieved in this model: Instead of estimating one function of several variables, as we do in completely nonparametric regression, we merely have to estimate d functions of one-dimensional variables X_j.
- *Partial Linear Model* (PLM)
  Suppose we only want to model parts of the index linearly. This could

be for analytical reasons or for reasons going back to economic theory. For instance, the impact of a dummy variable $X_1 \in \{0,1\}$ might be sufficiently explained by estimating the coefficient β_1.

For the sake of clarity, let us now separate the d-dimensional vector of explanatory variables into $\boldsymbol{U} = (U_1, \ldots, U_p)^\top$ and $\boldsymbol{T} = (T_1, \ldots, T_q)^\top$. The regression of Y on $\boldsymbol{X} = (\boldsymbol{U}, \boldsymbol{T})$ is assumed to have the form:

$$E(Y|\boldsymbol{U},\boldsymbol{T}) = \boldsymbol{U}^\top \boldsymbol{\beta} + m(\boldsymbol{T}) \tag{5.11}$$

where $m(\bullet)$ is an unknown multivariate function of the vector $\boldsymbol{T}$. Thus, a partial linear model can be interpreted as a sum of a purely parametric part, $\boldsymbol{U}^\top \boldsymbol{\beta}$, and a purely nonparametric part, $m(\boldsymbol{T})$. Not surprisingly, estimating $\boldsymbol{\beta}$ and $m(\bullet)$ involves the combination of both parametric and nonparametric regression techniques.

- *Generalized Additive Model* (GAM)
  Just like the (standard) additive model, generalized additive models are based on the sum of d nonparametric functions of the d variables $\boldsymbol{X}$ (plus an intercept term). In addition, they allow for a known parametric link function, $G(\bullet)$, that relates the sum of functions to the dependent variable:
  $$E(Y|\boldsymbol{X}) = G\left\{c + \sum_{j=1}^{d} g_j(X_j)\right\} . \tag{5.12}$$
- *Generalized Partial Linear Model* (GPLM)
  Introducing a link $G(\bullet)$ for a partial linear model $\boldsymbol{U}^\top \boldsymbol{\beta} + m(\boldsymbol{T})$ yields the *generalized partial linear model* (GPLM):
  $$E(Y|\boldsymbol{U},\boldsymbol{T}) = G\left\{\boldsymbol{U}^\top \boldsymbol{\beta} + m(\boldsymbol{T})\right\} .$$
  G denotes a known link function as in the GAM. In contrast to the GAM, $m(\bullet)$ is possibly a multivariate nonparametric function of the variable $\boldsymbol{T}$.
- *Generalized Partial Linear Partial Additive Model* (GAPLM)
  In high dimensions of $\boldsymbol{T}$ the estimate of the nonparametric function $m(\bullet)$ in the GPLM faces the same problems as the fully nonparametric multidimensional regression function estimates: the curse of dimensionality and the practical problem of interpretability. Hence, it is useful to think about a lower dimensional modeling of the nonparametric part. This leads to the GAPLM with an additive structure in the nonparametric component:
  $$E(Y|\boldsymbol{U},\boldsymbol{T}) = G\left\{\boldsymbol{U}^\top \boldsymbol{\beta} + \sum_{j=1}^{q} g_j(T_j)\right\} .$$
  Here, the $g_j(\bullet)$ will be univariate nonparametric functions of the variables T_j. In the case of an identity function G we speak of an additive partial linear model (APLM)

More discussion and motivation is given in the following chapters where the different models are discussed in detail and the specific estimation procedures are presented. Before proceeding with this task, however, we will first introduce some facts about the parametric generalized linear model (GLM). The following section is intended to give more insight into this model since its concept and the technical details of its estimation will be necessary for its semiparametric modification in Chapters 6 to 9.

5.2 Generalized Linear Models

Generalized linear models (GLM) extend the concept of the widely used linear regression model. The linear model assumes that the response Y (the dependent variable) is equal to a linear combination $\boldsymbol{X}^\top \boldsymbol{\beta}$ and a normally distributed error term:

$$Y = \boldsymbol{X}^\top \boldsymbol{\beta} + \varepsilon.$$

The least squares estimator $\widehat{\boldsymbol{\beta}}$ is adapted to these assumptions. However, the restriction of linearity is far too strict for a variety of practical situations. For example, a continuous distribution of the error term implies that the response Y has a continuous distribution as well. Hence, this standard linear regression model fails, for example, when dealing with binary data (Bernoulli Y) or with count data (Poisson Y).

Nelder & Wedderburn (1972) introduced the term *generalized linear models* (GLM). A good resource of material on this model is the monograph of McCullagh & Nelder (1989). The essential feature of the GLM is that the regression function, i.e. the expectation $\mu = E(Y|\boldsymbol{X})$ of Y is a monotone function of the index $\eta = \boldsymbol{X}^\top \boldsymbol{\beta}$. We denote the function which relates μ and η by G:

$$E(Y|\boldsymbol{X}) = G(\boldsymbol{X}^\top \boldsymbol{\beta}) \quad \Longleftrightarrow \quad \mu = G(\eta).$$

This function G is called the *link function*. (We remark that Nelder & Wedderburn (1972), McCullagh & Nelder (1989) actually denote G^{-1} as the link function.)

5.2.1 Exponential Families

In the GLM framework we assume that the distribution of Y is a member of the *exponential family*. The exponential family covers a broad range of distributions, for example discrete as the Bernoulli or Poisson distribution and continuous as the Gaussian (normal) or Gamma distribution.

A distribution is said to be a member of the exponential family if its probability function (if Y discrete) or its density function (if Y continuous) has the structure

$$f(y,\theta,\psi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\psi)} + c(y,\psi)\right\} \tag{5.13}$$

with some specific functions $a(\bullet)$, $b(\bullet)$ and $c(\bullet)$. These functions differ for the distinct Y distributions. Generally speaking, we are only interested in estimating the parameter θ. The additional parameter ψ is — as the variance σ^2 in the linear regression — a nuisance parameter. McCullagh & Nelder (1989) call θ the *canonical parameter*.

Example 5.2.
Suppose Y is normally distributed, i.e. $Y \sim N(\mu,\sigma^2)$. Hence we can write its density as

$$\varphi(y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{\frac{-1}{2\sigma^2}(y-\mu)^2\right\} = \exp\left\{y\frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{y^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma)\right\}$$

and we see that the normal distribution is a member of the exponential family with

$$a(\psi) = \sigma^2,\ b(\theta) = \frac{\mu^2}{2},\ c(y,\psi) = -\frac{y^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma),$$

where we set $\psi = \sigma$ and $\theta = \mu$. □

Example 5.3.
Suppose now Y is Bernoulli distributed, i.e. its probability function is

$$P(Y=y) = \mu^y(1-\mu)^{1-y} = \begin{cases} \mu & \text{if} \quad y = 1, \\ 1-\mu & \text{if} \quad y = 0. \end{cases}$$

This can be transformed into

$$P(Y=y) = \left(\frac{\mu}{1-\mu}\right)^y (1-\mu) = \exp\left\{y\log\left(\frac{p}{1-\mu}\right)\right\}(1-\mu)$$

using the *logit*

$$\theta = \log\left(\frac{\mu}{1-\mu}\right) \quad\Longleftrightarrow\quad \mu = \frac{e^\theta}{1+e^\theta}\,.$$

Thus we have an exponential family with

$$a(\psi) = 1,\ b(\theta) = -\log(1-\mu) = \log(1+e^\theta),\ c(y,\psi) \equiv 0.$$

This is a distribution without an additional nuisance parameter ψ. □

It is known that the least squares estimator $\widehat{\boldsymbol{\beta}}$ in the classical linear model is also the maximum-likelihood estimator for normally distributed errors. By imposing that the distribution of Y belongs to the exponential family it is possible to stay in the framework of maximum-likelihood for the GLM. Moreover, the use of the general concept of exponential families has the advantage that we can derive properties of different distributions at the same time.

To derive the maximum-likelihood algorithm in detail, we need to present some more properties of the probability function or density function $f(\bullet)$. First of all, f is a density (w.r.t. the Lebesgue measure in the continuous and w.r.t. the counting measure in the discrete case). This allows us to write

$$\int f(y, \theta, \psi)\, dy = 1.$$

Under some suitable regularity conditions (it is possible to exchange differentiation and integration) this yields

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta} \int f(y, \theta, \psi)\, dy = \int \frac{\partial}{\partial\theta} f(y, \theta, \psi)\, dy \\ &= \int \left\{ \frac{\partial}{\partial\theta} \log f(y, \theta, \psi) \right\} f(y, \theta, \psi)\, dy = E \left\{ \frac{\partial}{\partial\theta} \ell(y, \theta, \psi) \right\}, \end{aligned}$$

where $\ell(y, \theta, \psi)$ denotes the *log-likelihood*, i.e.

$$\ell(y, \theta, \psi) = \log f(y, \theta, \psi). \tag{5.14}$$

The function $\frac{\partial}{\partial\theta}\ell(y, \theta, \psi)$ is typically called *score* and it is known that

$$E\left\{ \frac{\partial^2}{\partial\theta^2} \ell(y, \theta, \psi) \right\} = -E \left\{ \frac{\partial}{\partial\theta} \ell(y, \theta, \psi) \right\}^2 .$$

This and taking first and second derivatives of (5.13) gives now

$$0 = E\left\{ \frac{Y - b'(\theta)}{a(\psi)} \right\}, \quad \text{and} \quad E\left\{ \frac{-b''(\theta)}{a(\psi)} \right\} = -E\left\{ \frac{Y - b'(\theta)}{a(\psi)} \right\}^2 .$$

We can conclude

$$\begin{aligned} E(Y) &= \mu = b'(\theta), \\ \operatorname{Var}(Y) &= V(\mu) a(\psi) = b''(\theta) a(\psi). \end{aligned}$$

We observe that the expectation of Y only depends on θ whereas the variance of Y depends on the parameter of interest θ and the nuisance parameter ψ. Typically one assumes that the factor $a(\psi)$ is identical over all observations.

5.2.2 Link Functions

Apart from the distribution of Y, the link function G is another important part of the GLM. Recall the notation

$$\eta = \boldsymbol{X}^\top \boldsymbol{\beta}, \; \mu = G(\eta).$$

In the case that

$$\boldsymbol{X}^\top \boldsymbol{\beta} = \eta = \theta$$

the link function is called *canonical link* function. For models with a canonical link, some theoretical and practical problems are easier to solve. Table 5.2 summarizes characteristics for some exponential functions together with canonical parameters θ and their canonical link functions. Note that for the binomial and the negative binomial distribution we assume the parameter k to be known. The case of binary Y is a special case of the binomial distribution ($k = 1$).

What link functions can we choose apart from the canonical? For most of the models a number of special link functions exist. For binomial Y for example, the logistic or Gaussian link functions are often used. Recall that a binomial model with the canonical logit link is called *logit* model. If the binomial distribution is combined with the Gaussian link, it is called *probit* model. A further alternative for binomial Y is the complementary log-log link

$$\eta = \log\{-\log(1-\mu)\}.$$

A very flexible class of link functions is the class of power functions which are also called Box-Cox transformations (Box & Cox, 1964). They can be defined for all models for which we have observations with positive mean. This family is usually specified as

$$\eta = \begin{cases} \mu^\lambda & \text{if} \quad \lambda \neq 0, \\ \log \mu & \text{if} \quad \lambda = 0. \end{cases}$$

5.2.3 Iteratively Reweighted Least Squares Algorithm

As already pointed out, the estimation method of choice for a GLM is maximizing the likelihood function with respect to $\boldsymbol{\beta}$. Suppose that we have the vector of observations $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ and denote their expectations (given $\boldsymbol{X}_i = \boldsymbol{x}_i$) by the vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$. More precisely, we have

$$\mu_i = G(\boldsymbol{x}_i^\top \boldsymbol{\beta}).$$

The log-likelihood of the vector $\boldsymbol{Y}$ is then

$$\ell(\boldsymbol{Y}, \boldsymbol{\mu}, \psi) = \sum_{i=1}^{n} \ell(Y_i, \theta_i, \psi), \tag{5.15}$$

where $\theta_i = \theta(\eta_i) = \theta(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ and $\ell(\bullet)$ on the right hand side of (5.15) denotes the individual log-likelihood contribution for each observation i.

Table 5.2. Characteristics of some GLM distributions

Notation	Range			Canonical	Variance	
	of y	$b(\theta)$	$\mu(\theta)$	link $\theta(\mu)$	$V(\mu)$	$a(\psi)$
Bernoulli $B(\mu)$	$\{0,1\}$	$\log(1+e^\theta)$	$\frac{e^\theta}{1+e^\theta}$	logit	$\mu(1-\mu)$	1
Binomial $B(k,\mu)$	$[0,k]$ integer	$k\log(1+e^\theta)$	$\frac{ke^\theta}{1+e^\theta}$	$\log\left(\frac{\mu}{k-\mu}\right)$	$\mu\left(1-\frac{\mu}{k}\right)$	1
Poisson $P(\mu)$	$[0,\infty)$ integer	$\exp(\theta)$	$\exp(\theta)$	log	μ	1
Negative Binomial $NB(\mu,k)$	$[0,\infty)$ integer	$-k\log(1-e^\theta)$	$\frac{ke^\theta}{1-e^\theta}$	$\log\left(\frac{\mu}{k+\mu}\right)$	$\mu+\frac{\mu^2}{k}$	1
Normal $N(\mu,\sigma^2)$	$(-\infty,\infty)$	$\theta^2/2$	θ	identity	1	σ^2
Gamma $G(\mu,\nu)$	$(0,\infty)$	$-\log(-\theta)$	$-1/\theta$	reciprocal	μ^2	$1/\nu$
Inverse Gaussian $IG(\mu,\sigma^2)$	$(0,\infty)$	$-(-2\theta)^{1/2}$	$\frac{-1}{\sqrt{(-2\theta)}}$	squared reciprocal	μ^3	σ^2

Example 5.4.
For $Y_i \sim N(\mu_i,\sigma^2)$ we have

$$\ell(Y_i,\theta_i,\psi) = \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2\sigma^2}(Y_i-\mu_i)^2.$$

This gives the sample log-likelihood

$$\ell(\boldsymbol{Y},\boldsymbol{\mu},\sigma) = n\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\mu_i)^2. \tag{5.16}$$

Obviously, maximizing the log-likelihood for $\boldsymbol{\beta}$ under normal Y is equivalent to minimizing the least squares criterion as the objective function. □

Example 5.5.
The calculation in Example 5.3 shows that the individual log-likelihood for the binary responses Y_i equals $\ell(Y_i, \theta_i, \psi) = Y_i \log(\mu_i) + (1 - Y_i) \log(1 - \mu_i)$. This leads to the sample version

$$\ell(\boldsymbol{Y}, \boldsymbol{\mu}, \psi) = \sum_{i=1}^{n} \{Y_i \log(\mu_i) + (1 - Y_i) \log(1 - \mu_i)\}\,. \tag{5.17}$$

Note that one typically defines $0 \cdot \log(0) = 0$. □

Let us remark that in the case where the distribution of Y itself is unknown, but its two first moments can be specified, then the quasi-likelihood may replace the log-likelihood (5.14). This means we assume that

$$\begin{aligned} E(Y) &= \mu, \\ \operatorname{Var}(Y) &= a(\psi)\, V(\mu). \end{aligned}$$

The quasi-likelihood is defined by

$$\ell(y, \theta, \psi) = \frac{1}{a(\psi)} \int\limits_{\mu(\theta)}^{y} \frac{(s - y)}{V(s)}\, ds\,, \tag{5.18}$$

cf. Nelder & Wedderburn (1972). If Y comes from an exponential family then the derivatives of (5.14) and (5.18) coincide. Thus, (5.18) establishes in fact a generalization of the likelihood approach.

Alternatively to the log-likelihood the *deviance* is used often. The deviance function is defined as

$$D(\boldsymbol{Y}, \boldsymbol{\mu}, \psi) = 2\,\{\ell(\boldsymbol{Y}, \boldsymbol{\mu}^{max}, \psi) - \ell(\boldsymbol{Y}, \boldsymbol{\mu}, \psi)\}\,, \tag{5.19}$$

where $\boldsymbol{\mu}^{max}$ (typically $\boldsymbol{Y}$) is the non-restricted vector maximizing $\ell(\boldsymbol{Y}, \bullet, \psi)$. The deviance (up to the factor $a(\psi)$) is the GLM analog of the *residual sum of squares* (RSS) in linear regression and compares the log-likelihood ℓ for the "model" μ with the maximal achievable value of ℓ. Since the first term in (5.19) is not dependent on the model and therefore not on $\boldsymbol{\beta}$, minimization of the deviance corresponds exactly to maximization of the log-likelihood.

Before deriving the algorithm to determine $\boldsymbol{\beta}$, let us have a look at (5.15) again. From $\ell(Y_i, \theta_i, \psi) = \log f(Y_i, \theta_i, \psi)$ and (5.13) we see

$$\ell(\boldsymbol{Y}, \boldsymbol{\mu}, \psi) = \sum_{i=1}^{n} \left\{ \frac{Y_i \theta_i - b(\theta_i)}{a(\psi)} - c(Y_i, \psi) \right\}. \tag{5.20}$$

Obviously, neither $a(\psi)$ nor $c(Y_i, \psi)$ have an influence on the maximization, hence it is sufficient to consider

$$\widetilde{\ell}(\boldsymbol{Y}, \boldsymbol{\mu}) = \sum_{i=1}^{n} \left\{ Y_i \theta_i - b(\theta_i) \right\}. \tag{5.21}$$

We will now maximize (5.21) w.r.t. $\boldsymbol{\beta}$. For that purpose take the first derivative of (5.21). This yields the gradient

$$\mathcal{D}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \widetilde{\ell}(\boldsymbol{Y}, \boldsymbol{\mu}) = \sum_{i=1}^{n} \left\{ Y_i - b'(\theta_i) \right\} \frac{\partial}{\partial \boldsymbol{\beta}} \theta_i \tag{5.22}$$

and our optimization problem is to solve

$$\mathcal{D}(\boldsymbol{\beta}) = 0,$$

a (in general) nonlinear system of equations in $\boldsymbol{\beta}$. For that reason, an iterative method is needed. One possible solution is the *Newton-Raphson algorithm*, a generalization of the Newton algorithm for the multidimensional parameter. Denote $\mathcal{H}(\boldsymbol{\beta})$ the *Hessian* of the log-likelihood, i.e. the matrix of second derivatives with respect to all components of $\boldsymbol{\beta}$. Then, one Newton-Raphson iteration step for $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}^{new} = \widehat{\boldsymbol{\beta}}^{old} - \left\{ \mathcal{H}(\widehat{\boldsymbol{\beta}}^{old}) \right\}^{-1} \mathcal{D}(\widehat{\boldsymbol{\beta}}^{old}).$$

A variant of the Newton-Raphson is the *Fisher scoring algorithm* which replaces the Hessian by its expectation (w.r.t. the observations Y_i)

$$\widehat{\boldsymbol{\beta}}^{new} = \widehat{\boldsymbol{\beta}}^{old} - \left\{ E\mathcal{H}(\widehat{\boldsymbol{\beta}}^{old}) \right\}^{-1} \mathcal{D}(\widehat{\boldsymbol{\beta}}^{old}).$$

To present both algorithms in a more detailed way, we need again some additional notation. Recall that we have $\mu_i = G(\boldsymbol{x}_i^\top \boldsymbol{\beta}) = b'(\theta_i)$, $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ and $b'(\theta_i) = \mu_i = G(\eta_i)$. For the first and second derivatives of θ_i we obtain (after some calculation)

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}} \theta_i &= \frac{G'(\eta_i)}{V(\mu_i)} \boldsymbol{x}_i \\ \frac{\partial^2}{\partial \boldsymbol{\beta} \boldsymbol{\beta}^\top} \theta_i &= \frac{G''(\eta_i) V(\mu_i) - G'(\eta_i)^2 V'(\mu_i)}{V(\mu_i)^2} \boldsymbol{x}_i \boldsymbol{x}_i^\top. \end{aligned}$$

Using this, we can express the gradient of the log-likelihood as

$$\mathcal{D}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ Y_i - \mu_i \right\} \frac{G'(\eta_i)}{V(\mu_i)} \boldsymbol{x}_i.$$

For the Hessian we get

$$\mathcal{H}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ -b''(\theta_i) \left(\frac{\partial}{\partial \boldsymbol{\beta}} \theta_i \right) \left(\frac{\partial}{\partial \boldsymbol{\beta}} \theta_i \right)^{\top} - \{Y_i - b'(\theta_i)\} \frac{\partial^2}{\partial \boldsymbol{\beta} \boldsymbol{\beta}^{\top}} \theta_i \right\}$$

$$= \sum_{i=1}^{n} \left\{ \frac{G'(\eta_i)^2}{V(\mu_i)} - \{Y_i - \mu_i\} \frac{G''(\eta_i)V(\mu_i) - G'(\eta_i)^2 V'(\mu_i)}{V(\mu_i)^2} \right\} \boldsymbol{x}_i \boldsymbol{x}_i^{\top}.$$

Since $EY_i = \mu_i$ it turns out that the Fisher scoring algorithm is easier: We replace $\mathcal{H}(\boldsymbol{\beta})$ by

$$E\mathcal{H}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ \frac{G'(\eta_i)^2}{V(\mu_i)} \right\} \boldsymbol{x}_i \boldsymbol{x}_i^{\top}.$$

For the sake of simplicity let us concentrate on the Fisher scoring for the moment. Define the weight matrix

$$\mathbf{W} = \operatorname{diag}\left(\frac{G'(\eta_1)^2}{V(\mu_1)}, \ldots, \frac{G'(\eta_n)^2}{V(\mu_n)} \right).$$

Additionally, define

$$\widetilde{\mathbf{Y}} = \left(\frac{Y_1 - \mu_1}{G'(\eta_1)}, \ldots, \frac{Y_n - \mu_n}{G'(\eta_n)} \right)^{\top}$$

and the design matrix

$$\mathbf{X} = \begin{pmatrix} \boldsymbol{x}_1^{\top} \\ \vdots \\ \boldsymbol{x}_1^{\top} \end{pmatrix}.$$

Then one iteration step for $\boldsymbol{\beta}$ can be rewritten as

$$\begin{aligned} \boldsymbol{\beta}^{new} &= \boldsymbol{\beta}^{old} + (\mathbf{X}^{\top}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\top}\mathbf{W}\widetilde{\mathbf{Y}} \\ &= (\mathbf{X}^{\top}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\top}\mathbf{W}\mathbf{Z} \end{aligned} \tag{5.23}$$

where $\mathbf{Z} = (Z_1, \ldots, Z_n)^{\top}$ is the vector of *adjusted dependent variables*

$$Z_i = \boldsymbol{x}_i^{\top}\boldsymbol{\beta}^{old} + (Y_i - \mu_i)\{G'(\eta_i)\}^{-1}. \tag{5.24}$$

The iteration stops when the parameter estimate or the log-likelihood (or both) do not change significantly any more. We denote the resulting parameter estimate by $\widehat{\boldsymbol{\beta}}$.

We see that each iteration step (5.23) is the result of a weighted least squares regression on the adjusted variables Z_i on $\boldsymbol{x}_i$. Hence, a GLM can be estimated by *iteratively reweighted least squares* (IRLS). Note further that in the linear regression model, where we have $G' \equiv 1$ and $\mu_i = \eta_i = \boldsymbol{x}_i^{\top}\boldsymbol{\beta}$, no iteration is necessary. The Newton-Raphson algorithm can be given in a similar way (with the more complicated weights and a different formula for the adjusted variables). There are several remarks on the algorithm:

- In the case of a canonical link function, the Newton-Raphson and the Fisher scoring algorithm coincide. Here the second derivative of θ_i is zero. Additionally we have

  $$b'(\theta_i) = G(\theta_i) \quad \Longrightarrow \quad b''(\theta_i) = G'(\theta_i) = V(\mu_i).$$

  This also simplifies the weight matrix $\mathbf{W}$.
- We still have to address the problem of starting values. A naive way would be just to start with some arbitrary $\boldsymbol{\beta}_0$, as e.g. $\boldsymbol{\beta}_0 = 0$. It turns out that we do not in fact need a starting value for $\boldsymbol{\beta}$ since the adjusted dependent variable can be equivalently initialized by appropriate $\eta_{i,0}$ and $\mu_{i,0}$. Typically the following choices are made, we refer here to McCullagh & Nelder (1989).
  - For all but binomial models:
    $\mu_{i,0} = Y_i$ and $\eta_{i,0} = G^{-1}(\mu_{i,0})$
  - For binomial models:
    $\mu_{i,0} = (Y_i + \frac{1}{2})/(k+1)$ and $\eta_{i,0} = G^{-1}(\mu_{i,0})$.
    (k denotes the binomial weights, i.e. $k = 1$ in the Bernoulli case.)
- An estimate $\widehat{\psi}$ for the dispersion parameter ψ can be obtained from

  $$\widehat{a}(\psi) = \frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}, \tag{5.25}$$

  when $\widehat{\mu}_i$ denotes the estimated regression function for the ith observation.

The resulting estimator $\widehat{\boldsymbol{\beta}}$ has an asymptotic normal distribution, except of course for the standard linear regression case with normal errors where $\widehat{\boldsymbol{\beta}}$ has an exact normal distribution.

Theorem 5.1.
Under regularity conditions and as $n \to \infty$ we have for the estimated coefficient vector

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{L} N(0, \boldsymbol{\Sigma}).$$

Denote further by $\widehat{\boldsymbol{\mu}}$ the estimator of $\boldsymbol{\mu}$. Then, for deviance and log-likelihood it holds approximately: $D(\boldsymbol{Y}, \widehat{\boldsymbol{\mu}}, \psi) \sim \chi^2_{n-d}$ and $2\{\ell(\boldsymbol{Y}, \widehat{\boldsymbol{\mu}}, \psi) - \ell(\boldsymbol{Y}, \boldsymbol{\mu}, \psi)\} \sim \chi^2_d$.

The asymptotic covariance of the coefficient $\widehat{\boldsymbol{\beta}}$ can be estimated by

$$\widehat{\boldsymbol{\Sigma}} = a(\widehat{\psi})\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{G'(\eta_{i,last})^2}{V(\mu_{i,last})}\right\}\boldsymbol{X}_i\boldsymbol{X}_i^\top\right]^{-1} = a(\widehat{\psi})\cdot n \cdot \left(\mathbf{X}^\top\mathbf{W}\mathbf{X}\right)^{-1},$$

with the subscript *last* denoting the values from the last iteration step. Using this estimated covariance we can make inference about the components of $\boldsymbol{\beta}$ such as tests of significance. For selection between two nested models, typically a likelihood ratio test (LR test) is used.

Example 5.6.
Let us illustrate the GLM using the data on East-West German migration from Table 5.1. This is a sample of East Germans who have been surveyed in 1991 in the German Socio-Economic Panel, see GSOEP (1991). Among other questions the participants have been asked if they can imagine moving to the Western part of Germany or West Berlin. We give the value 1 for those who responded positively and 0 if not.

Recall that the economic model is based on the idea that a person will migrate if the utility (wage differential) exceeds the costs of migration. Of course neither one of the variables, wage differential and costs, are directly available. It is obvious that age has an important influence on migration intention. Younger people will have a higher wage differential. A currently low household income and unemployment will also increase a possible gain in wage after migration. On the other hand, the presence of friends or family members in the Western part of Germany will reduce the costs of migration. We also consider a city size indicator and gender as interesting variables (Table 5.1).

Table 5.3. Logit coefficients for migration data

	Coefficients	t-value
constant	0.512	2.39
FAMILY/FRIENDS	0.599	5.20
UNEMPLOYED	0.221	2.31
CITY SIZE	0.311	3.77
FEMALE	-0.240	-3.15
AGE	$-4.69 \cdot 10^{-2}$	-14.56
INCOME	$1.42 \cdot 10^{-4}$	2.73

Now, we are interested in estimating the probability of migration in dependence of the explanatory variables $\boldsymbol{x}$. Recall, that

$$P(Y = 1|\boldsymbol{X}) = E(Y|\boldsymbol{X}).$$

A useful model is a GLM with a binary (Bernoulli) Y and the logit link for example:

$$P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = G(\boldsymbol{x}^\top \boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}^\top \boldsymbol{\beta})}.$$

Table 5.3 shows in the middle column the results of this logit fit. The migration intention is definitely determined by age. However, also the unemployment, city size and household income variables are highly significant, which is indicated by their high *t*-values ($\widehat{\beta}_j / \sqrt{\widehat{\Sigma}_{jj}}$). □

Bibliographic Notes

For general aspects on semiparametric regression we refer to the textbooks of Pagan & Ullah (1999), Yatchew (2003), Ruppert, Wand & Carroll (1990). Comprehensive presentations of the generalized linear model can be found in Dobson (2001), McCullagh & Nelder (1989) and Hardin & Hilbe (2001). For a more compact introduction see Müller (2004), Venables & Ripley (2002, Chapter 7) and Gill (2000).

In the following notes, we give some references for topics we consider related to the considered models. References for specific models are listed in the relevant chapters later on.

The transformation model in (5.4) was first introduced in an econometric context by Box & Cox (1964). The discussion was revised many years later by Bickel & Doksum (1981). In a more recent paper, Horowitz (1996) estimates this model by considering a nonparametric transformation.

For a further reference of dimension reduction in nonparametric estimation we mention projection pursuit and sliced inverse regression. The projection pursuit algorithm is introduced and investigated in detail in Friedman & Stuetzle (1981) and Friedman (1987). Sliced inverse regression means the estimation of $Y = m\left(\boldsymbol{X}^\top\boldsymbol{\beta}_1, \boldsymbol{X}^\top\boldsymbol{\beta}_2, \ldots, \boldsymbol{X}^\top\boldsymbol{\beta}_k, \varepsilon\right)$, where ε is the disturbance term and k the unknown dimension of the model. Introduction and theory can be found e.g. in Duan & Li (1991), Li (1991) or Hsing & Carroll (1992).

More sophisticated models like censored or truncated dependent variables, models with endogenous variables or simultaneous equation systems (Maddala, 1983) will not be dealt with in this book. There are two reasons: On one hand the non- or semiparametric estimation of those models is much more complicated and technical than most of what we aim to introduce in this book. Here we only prepare the basics enabling the reader to consider more special problems. On the other hand, most of these estimation problems are rather particular and the treatment of them presupposes good knowledge of the considered problem and its solution in the parametric world. Instead of extending the book considerably by setting out this topic, we limit ourselves here to some more detailed bibliographic notes.

The non- and semiparametric literature on this is mainly separated into two directions, parametric modeling with unknown error distribution or modeling non-/semiparametrically the functional forms. In the second case a principal question is the identifiability of the model.

For an introduction to the problem of truncation, sample selection and limited dependent data, see Heckman (1976) and Heckman (1979). See also the survey of Amemiya (1984). An interesting approach was presented by Ahn & Powell (1993) for parametric censored selection models with nonparametric selection mechanism. This idea has been extended to general pairwise difference estimators for censored and truncated models in Honoré & Powell (1994). A mostly comprehensive survey about parametric and semiparametric methods for parametric models with non- or semiparametric selection bias can be found in Vella (1998). Even though implementation of and theory on these methods is often quite complicated, some of them turned out to perform reasonably well.

The second approach, i.e. relaxing the functional forms of the functions of interest, turned out to be much more complicated. To our knowledge, the first articles on the estimation of triangular simultaneous equation systems have been Newey, Powell & Vella (1999) and Rodríguez-Póo, Sperlich & Fernández (1999), from which the former is purely nonparametric, whereas the latter considers nested simultaneous equation systems and needs to specify the error distribution for identifiability reasons. Finally, Lewbel & Linton (2002) found a smart way to identify nonparametric censored and truncated regression functions; however, their estimation procedure is quite technical. Note that so far neither their estimator nor the one of Newey, Powell & Vella (1999) have been proved to perform well in practice.

Exercises

Exercise 5.1. Assume model (5.6) and consider $\boldsymbol{X}$ and ε to be independent. Show that

$$P(Y = 1|\boldsymbol{X}) = E(Y|\boldsymbol{X}) = G_\varepsilon\{v(\boldsymbol{X})\}$$

where G_ε denotes the cdf of ε. Explain that (5.7) holds if we do not assume independence of $\boldsymbol{X}$ and ε.

Exercise 5.2. Recall the paragraph about partial linear models. Why may it be sufficient to include $\beta_1 X_1$ in the model when X_1 is binary? What would you do if X_1 were categorical?

Exercise 5.3. Compute $\mathcal{H}(\boldsymbol{\beta})$ and $E\mathcal{H}(\boldsymbol{\beta})$ for the logit and probit models.

Exercise 5.4. Verify the canonical link functions for the logit and Poisson model.

Exercise 5.5. Recall that in Example 5.6 we have fitted the model

$$E(Y|\boldsymbol{X}) = P(Y = 1|\boldsymbol{X}) = G(\boldsymbol{X}^\top \boldsymbol{\beta}),$$

where G is the standard logistic cdf. We motivated this model through the latent-variable model $Y^* = \boldsymbol{X}^\top \boldsymbol{\beta} - \varepsilon$ with ε having cdf G. How does the logit model change if the latent-variable model is multiplied by a factor c? What does this imply for the identification of the coefficient vector $\boldsymbol{\beta}$?

Summary

- ⋆ The basis for many semiparametric regression models is the generalized linear model (GLM), which is given by

$$E(Y|\boldsymbol{X}) = G\{\boldsymbol{X}^\top \boldsymbol{\beta}\}\,.$$

Here, $\boldsymbol{\beta}$ denotes the parameter vector to be estimated and G denotes a known link function. Prominent examples of this type of regression are binary choice models (logit or probit) or count data models (Poisson regression).

- ⋆ The GLM can be generalized in several ways: Considering an unknown smooth link function (instead of G) leads to the single index model (SIM). Assuming a nonparametric additive argument of G leads to the generalized additive model (GAM), whereas a combination of additive linear and nonparametric components in the argument of G give a generalized partial linear model (GPLM) or generalized partial linear partial additive model (GAPLM). If there is no link function (or G is the identity function) then we speak of additive models (AM) or partial linear models (PLM) or additive partial linear models (APLM).

- ⋆ The estimation of the GLM is performed through an interactive algorithm. This algorithm, the iteratively reweighted least squares (IRLS) algorithm, applies weighted least squares to the adjusted dependent variable Z in each iteration step:

$$\boldsymbol{\beta}^{new} = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{Z}$$

This numerical approach needs to be appropriately modified for estimating the semiparametric modifications of the GLM.

References

Achmus, S. (2000). Nichtparametrische additive Modelle, *Doctoral Thesis*, Technical University of Braunschweig, Germany.

Ahn, H. & Powell, J. L. (1993). Semiparametric selection of censored selection models with a nonparametric selection mechanism, *Econometrica* **58**: 3–29.

Amemiya, T. (1984). Tobit models: A survey, *Journal of Econometrics* **24**: 3–61.

Andrews, D. W. K. & Whang, Y.-J. (1990). Additive interactive regression models: circumvention of the curse of dimensionality, *Econometric Theory* **6**: 466–479.

Begun, J., Hall, W., Huang, W. & Wellner, J. (1983). Information and asymptotic efficiency in parametric–nonparametric models, *Annals of Statistics* **11**: 432–452.

Berndt, E. (1991). *The Practice of Econometrics*, Addison–Wesley.

Bickel, P. & Doksum, K. (1981). An analysis of transformations revisited, *Journal of the American Statistical Association* **76**: 296–311.

Bickel, P., Klaassen, C., Ritov, Y. & Wellner, J. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*, The Johns Hopkins University Press.

Bickel, P. & Rosenblatt, M. (1973). On some global measures of the deviations of density function estimators, *Annals of Statistics* **1**: 1071–1095.

Bierens, H. (1990). A consistent conditional moment test of functional form, *Econometrica* **58**: 1443–1458.

Bierens, H. & Ploberger, W. (1997). Asymptotic theory of integrated conditional moment tests, *Econometrica* **65**: 1129–1151.

Bonneu, M. & Delecroix, M. (1992). Estimation semiparamétrique dans les modèles explicatifs conditionnels à indice simple, *Cahier de gremaq, 92.09.256*, GREMAQ, Université Toulouse I.

Bonneu, M., Delecroix, M. & Malin, E. (1993). Semiparametric versus nonparametric estimation in single index regression model: A computational approach, *Computational Statistics* **8**: 207–222.

Bossaerts, P., Hafner, C. & Härdle, W. (1996). Foreign exchange rates have surprising volatility, *in* P. M. Robinson & M. Rosenblatt (eds), *Athens Conference on Applied Probability and Time Series Analysis. Volume II: Time Series Analysis. In Memory of E.J. Hannan*, Lecture Notes in Statistics, Springer, pp. 55–72.

Boularan, J., Ferré, L. & Vieu, P. (1994). Growth curves: a two-stage nonparametric approach, *Journal of Statistical Planning and Inference* **38**: 327–350.

Bowman, A. & Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis*, Oxford University Press, Oxford, UK.

Box, G. & Cox, D. (1964). An analysis of transformations, *Journal of the Royal Statistical Society, Series B* **26**: 211–243.

Breiman, L. & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlations (with discussion), *Journal of the American Statistical Association* **80**(391): 580–619.

Buja, A., Hastie, T. J. & Tibshirani, R. J. (1989). Linear smoothers and additive models (with discussion), *Annals of Statistics* **17**: 453–555.

Burda, M. (1993). The determinants of East–West German migration, *European Economic Review* **37**: 452–461.

Cao, R., Cuevas, A. & González Manteiga, W. (1994). A comparative study of several smoothing methods in density estimation, *Computational Statistics & Data Analysis* **17**(2): 153–176.

Carroll, R. J., Fan, J., Gijbels, I. & Wand, M. P. (1997). Generalized partially linear single–index models, *Journal of the American Statistical Association* **92**: 477–489.

Carroll, R. J., Härdle, W. & Mammen, E. (2002). Estimation in an additive model when the components are linked parametrically, *Econometric Theory* **18**(4): 886–912.

Chaudhuri, P. & Marron, J. S. (1999). SiZer for exploration of structures in curves, *Journal of the American Statistical Association* **94**: 807–823.

Chen, R., Liu, J. S. & Tsay, R. S. (1995). Additivity tests for nonlinear autoregression, *Biometrika* **82**: 369–383.

Cleveland, W. S. (1979). Robust locally-weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* **74**: 829–836.

Collomb, G. (1985). Nonparametric regression – an up-to-date bibliography, *Statistics* **2**: 309–324.

Cosslett, S. (1983). Distribution–free maximum likelihood estimation of the binary choice model, *Econometrica* **51**: 765–782.

Cosslett, S. (1987). Efficiency bounds for distribution–free estimators of the binary choice model, *Econometrica* **55**: 559–586.

Dalelane, C. (1999). Bootstrap confidence bands for the integration estimator in additive models, *Diploma thesis*, Department of Mathematics, Humboldt-Universität zu Berlin.

Daubechies, I. (1992). *Ten Lectures on Wavelets*, SIAM, Philadelphia, Pennsylvania.

Deaton, A. & Muellbauer, J. (1980). *Economics and Consumer Behavior*, Cambridge University Press, Cambridge.

Delecroix, M., Härdle, W. & Hristache, M. (2003). Efficient estimation in conditional single-index regression, *Journal of Multivariate Analysis* **86**(2): 213–226.

Delgado, M. A. & Mora, J. (1995). Nonparametric and semiparametric estimation with discrete regressors, *Econometrica* **63**(6): 1477–1484.

Denby, L. (1986). Smooth regression functions, *Statistical report 26*, AT&T Bell Laboratories.

Dette, H. (1999). A consistent test for the functional form of a regression based on a difference of variance estimators, *Annals of Statistics* **27**: 1012–1040.

Dette, H., von Lieres und Wilkau, C. & Sperlich, S. (2004). A comparison of different nonparametric methods for inference on additive models, *Nonparametric Statistics* **16**. forthcoming.

Dobson, A. J. (2001). *An Introduction to Generalized Linear Models*, second edn, Chapman and Hall, London.

Donoho, D. L. & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage, *Biometrika* **81**: 425–455.

Donoho, D. L. & Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage, *Journal of the American Statistical Association* **90**: 1200–1224.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. & Picard, D. (1995). Wavelet shrinkage: Asymptopia? (with discussion), *Journal of the Royal Statistical Society, Series B* **57**: 301–369.

Duan, N. & Li, K.-C. (1991). Slicing regression: A link-free regression method, *Annals of Statistics* **19**(2): 505–530.

Duin, R. P. W. (1976). On the choice of smoothing parameters of Parzen estimators of probability density functions, *IEEE Transactions on Computers* **25**: 1175–1179.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing with b-splines and penalties (with discussion, *Statistical Science* **11**: 89–121.

Epanechnikov, V. (1969). Nonparametric estimation of a multidimensional probability density, *Teoriya Veroyatnostej i Ee Primeneniya* **14**: 156–162.

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing*, Marcel Dekker, New York.

Eubank, R. L., Hart, J. D., Simpson, D. G. & Stefanski, L. A. (1995). Testing for additivity in nonparametric regression, *Annals of Statistics* **23**: 1896–1920.

Eubank, R. L., Kambour, E. L., Kim, J. T., Klipple, K., Reese, C. S. & Schimek, M. G. (1998). Estimation in partially linear models, *Computational Statistics & Data Analysis* **29**: 27–34.

Fahrmeir, L. & Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer.

Fan, J. & Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, Vol. 66 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, New York.

Fan, J., Härdle, W. & Mammen, E. (1998). Direct estimation of low-dimensional components in additive models, *Annals of Statistics* **26**: 943–971.

Fan, J. & Li, Q. (1996). Consistent model specification test: Omitted variables and semiparametric forms, *Econometrica* **64**: 865–890.

Fan, J. & Marron, J. S. (1992). Best possible constant for bandwidth selection, *Annals of Statistics* **20**: 2057–2070.

Fan, J. & Marron, J. S. (1994). Fast implementations of nonparametric curve estimators, *Journal of Computational and Graphical Statistics* **3**(1): 35–56.

Fan, J. & Müller, M. (1995). Density and regression smoothing, *in* W. Härdle, S. Klinke & B. A. Turlach (eds), *XploRe – an interactive statistical computing environment*, Springer, pp. 77–99.

Friedman, J. H. (1987). Exploratory projection pursuit, *Journal of the American Statistical Association* **82**: 249–266.

Friedman, J. H. & Stuetzle, W. (1981). Projection pursuit regression, *Journal of the American Statistical Association* **76**(376): 817–823.

Friedman, J. H. & Stuetzle, W. (1982). Smoothing of scatterplots, *Technical report*, Department of Statistics, Stanford.

Fuss, M., McFadden, D. & Mundlak, Y. (1978). A survey of functional forms in the economic analysis of production, *in* M. Fuss & D. McFadden (eds), *Production Economics: A Dual Approach to Theory and Applications*, North-Holland, Amsterdam, pp. 219–268.

Gallant, A. & Nychka, D. (1987). Semi-nonparametric maximum likelihood estimation, *Econometrica* **55**(2): 363–390.

Gasser, T. & Müller, H. G. (1984). Estimating regression functions and their derivatives by the kernel method, *Scandinavian Journal of Statistics* **11**: 171–185.

Gill, J. (2000). *Generalized Linear Models: A Unified Approach*, Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-134, Thousand Oaks, CA.

Gill, R. D. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method (Part I), *Scandinavian Journal of Statistics* **16**: 97–128.

Gill, R. D. & van der Vaart, A. W. (1993). Non- and semi-parametric maximum likelihood estimators and the von Mises method (Part II), *Scandinavian Journal of Statistics* **20**: 271–288.

González Manteiga, W. & Cao, R. (1993). Testing hypothesis of general linear model using nonparametric regression estimation, *Test* **2**: 161–189.

Gozalo, P. L. & Linton, O. (2001). A nonparametric test of additivity in generalized nonparametric regression with estimated parameters, *Journal of Econometrics* **104**: 1–48.

Grasshoff, U., Schwalbach, J. & Sperlich, S. (1999). Executive pay and corporate financial performance. an explorative data analysis, *Working paper 99-84 (33)*, Universidad Carlos III de Madrid.

Green, P. J. & Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*, Vol. 58 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Green, P. J. & Yandell, B. S. (1985). Semi-parametric generalized linear models, *Proceedings 2nd International GLIM Conference*, Vol. 32 of *Lecture Notes in Statistics 32*, Springer, New York, pp. 44–55.

GSOEP (1991). *Das Sozio-ökonomische Panel (SOEP) im Jahre 1990/91*, Projektgruppe "Das Sozio-ökonomische Panel", Deutsches Institut für Wirtschaftsforschung. Vierteljahreshefte zur Wirtschaftsforschung, pp. 146–155.

Habbema, J. D. F., Hermans, J. & van den Broek, K. (1974). A stepwise discrimination analysis program using density estimation, *COMPSTAT '74. Proceedings in Computational Statistics*, Physica, Vienna.

Hall, P. & Marron, J. S. (1991). Local minima in cross–validation functions, *Journal of the Royal Statistical Society, Series B* **53**: 245–252.

Hall, P., Marron, J. S. & Park, B. U. (1992). Smoothed cross-validation, *Probability Theory and Related Fields* **92**: 1–20.

Hall, P., Sheather, S. J., Jones, M. C. & Marron, J. S. (1991). On optimal data–based bandwidth selection in kernel density estimation, *Biometrika* **78**: 263–269.

Han, A. (1987). Non–parametric analysis of a generalized regression model, *Journal of Econometrics* **35**: 303–316.

Hardin, J. & Hilbe, J. (2001). *Generalized Linear Models and Extensions*, Stata Press.

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., Huet, S., Mammen, E. & Sperlich, S. (2004). Bootstrap inference in semiparametric generalized additive models, *Econometric Theory* **20**: to appear.

Härdle, W., Kerkyacharian, G., Picard, D. & Tsybakov, A. B. (1998). *Wavelets, Approximation, and Statistical Applications*, Springer, New York.

Härdle, W. & Mammen, E. (1993). Testing parametric versus nonparametric regression, *Annals of Statistics* **21**: 1926–1947.

Härdle, W., Mammen, E. & Müller, M. (1998). Testing parametric versus semiparametric modelling in generalized linear models, *Journal of the American Statistical Association* **93**: 1461–1474.

Härdle, W. & Müller, M. (2000). Multivariate and semiparametric kernel regression, *in* M. Schimek (ed.), *Smoothing and Regression*, Wiley, New York, pp. 357–391.

Härdle, W. & Scott, D. W. (1992). Smoothing in by weighted averaging using rounded points, *Computational Statistics* **7**: 97–128.

Härdle, W., Sperlich, S. & Spokoiny, V. (2001). Structural tests in additive regression, *Journal of the American Statistical Association* **96**(456): 1333–1347.

Härdle, W. & Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives, *Journal of the American Statistical Association* **84**: 986–995.

Härdle, W. & Tsybakov, A. B. (1997). Local polynomial estimators of the volatility function in nonparametric autoregression, *Journal of Econometrics* **81**(1): 223–242.

Harrison, D. & Rubinfeld, D. L. (1978). Hedonic prices and the demand for clean air, *J. Environ. Economics and Management* **5**: 81–102.

Hart, J. D. (1997). *onparametric Smoothing and Lack-of-Fit Tests*, Springer, New York.

Hastie, T. J. & Tibshirani, R. J. (1986). Generalized additive models (with discussion), *Statistical Science* **1**(2): 297–318.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized Additive Models*, Vol. 43 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such model, *Annals of Economic and Social Measurement* **5**: 475–492.

Heckman, J. (1979). Sample selection bias as a specification error, *Econometrica* **47**: 153–161.

Hengartner, N., Kim, W. & Linton, O. (1999). A computationally efficient oracle estimator for additive nonparametric regression with bootstrap confidence intervals, *Journal of Computational and Graphical Statistics* **8**: 1–20.

Honoré, B. E. & Powell, J. L. (1994). Pairwise difference estimators of censored and truncated regression models, *Journal of Econometrics* **64**: 241–278.

Horowitz, J. L. (1993). Semiparametric and nonparametric estimation of quantal response models, *in* G. S. Maddala, C. R. Rao & H. D. Vinod (eds), *Handbook of Statistics*, Elsevier Science Publishers, pp. 45–72.

Horowitz, J. L. (1996). Semiparametric estimation of a regression model with an unknown transformation of the dependent variable, *Econometrica* **64**: 103–137.

Horowitz, J. L. (1998a). Nonparametric estimation of a generalized additive model with an unknown link function, *Technical report*, University of Iowa.

Horowitz, J. L. (1998b). *Semiparametric Methods in Econometrics*, Springer.

Horowitz, J. L. & Härdle, W. (1994). Testing a parametric model against a semiparametric alternative, *Econometric Theory* **10**: 821–848.

Horowitz, J. L. & Härdle, W. (1996). Direct semiparametric estimation of single-index models with discrete covariates, *Journal of the American Statistical Association* **91**(436): 1632–1640.

Hsing, T. & Carroll, R. J. (1992). An asymptotic theory for sliced inverse regression, *Annals of Statistics* **20**(2): 1040–1061.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single–index models, *Journal of Econometrics* **58**: 71–120.

Ingster, Y. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. I - III, *Math. Methods of Statist.* **2**: 85 – 114, 171 – 189, 249 – 268.

Jones, M. C., Marron, J. S. & Sheather, S. J. (1996). Progress in data-based bandwidth selection for kernel density estimation, *Computational Statistics* **11**(3): 337–381.

Kallenberg, W. C. M. & Ledwina, T. (1995). Consistency and Monte-Carlo simulations of a data driven version of smooth goodness-of-fit tests, *Annals of Statistics* **23**: 1594–1608.

Klein, R. & Spady, R. (1993). An efficient semiparametric estimator for binary response models, *Econometrica* **61**: 387–421.

Korostelev, A. & Müller, M. (1995). Single index models with mixed discrete-continuous explanatory variables, *Discussion Paper 26*, Sonderforschungsbereich 373, Humboldt-Universität zu Berlin.

Ledwina, T. (1994). Data-driven version of Neyman's smooth test of fit, *Journal of the American Statistical Association* **89**: 1000–1005.

Lejeune, M. (1985). Estimation non-paramétrique par noyaux: régression polynomiale mobile, *Revue de Statistique Appliqueés* **33**: 43–67.

Leontief, W. (1947a). Introduction to a theory of the internal structure of functional relationships, *Econometrica* **15**: 361–373.

Leontief, W. (1947b). A note on the interrelation of subsets of independent variables of a continuous function with continuous first derivatives., *Bulletin of the American Mathematical Society* **53**: 343–350.

Lewbel, A. & Linton, O. (2002). Nonparametric censored and truncated regression, *Econometrica* **70**: 765–780.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion), *Journal of the American Statistical Association* **86**(414): 316–342.

Linton, O. (1997). Efficient estimation of additive nonparametric regression models, *Biometrika* **84**: 469–473.

Linton, O. (2000). Efficient estimation of generalized additive nonparametric regression models, *Econometric Theory* **16**(4): 502–523.

Linton, O. & Härdle, W. (1996). Estimation of additive regression models with known links, *Biometrika* **83**(3): 529–540.

Linton, O. & Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration, *Biometrika* **82**: 93–101.

Loader, C. (1999). *Local Regression and Likelihood*, Springer, New York.

Mack, Y. P. (1981). Local properties of k-nn regression estimates, *SIAM J. Alg. Disc. Math.* **2**: 311–323.

Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*, Econometric Society Monographs No. 4, Cambridge University Press.

Mammen, E., Linton, O. & Nielsen, J. P. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions, *Annals of Statistics* **27**: 1443–1490.

Mammen, E. & Nielsen, J. P. (2003). Generalised structured models, *Biometrika* **90**: 551–566.

Manski, C. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator, *Journal of Econometrics* **3**: 205–228.

Marron, J. S. (1989). Comments on a data based bandwidth selector, *Computational Statistics & Data Analysis* **8**: 155–170.

Marron, J. S. & Härdle, W. (1986). Random approximations to some measures of accuracy in nonparametric curve estimation, *J. Multivariate Anal.* **20**: 91–113.

Marron, J. S. & Nolan, D. (1988). Canonical kernels for density estimation, *Statistics & Probability Letters* **7**(3): 195–199.

Masry, E. & Tjøstheim, D. (1995). Non-parametric estimation and identification of nonlinear arch time series: strong convergence properties and asymptotic normality, *Econometric Theory* **11**: 258–289.

Masry, E. & Tjøstheim, D. (1997). Additive nonlinear ARX time series and projection estimates, *Econometric Theory* **13**: 214–252.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*, Vol. 37 of *Monographs on Statistics and Applied Probability*, 2 edn, Chapman and Hall, London.

Müller, M. (2001). Estimation and testing in generalized partial linear models — a comparative study, *Statistics and Computing* **11**: 299–309.

Müller, M. (2004). Generalized linear models, *in* J. Gentle, W. Härdle & Y. Mori (eds), *Handbook of Computational Statistics (Volume I). Concepts and Fundamentals*, Springer, Heidelberg.

Nadaraya, E. A. (1964). On estimating regression, *Theory of Probability and its Applications* **10**: 186–190.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models, *Journal of the Royal Statistical Society, Series A* **135**(3): 370–384.

Newey, W. K. (1990). Semiparametric efficiency bounds, *Journal of Applied Econometrics* **5**: 99–135.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimation, *Econometrica* **62**: 1349–1382.

Newey, W. K. (1995). Convergence rates for series estimators, *in* G. Maddala, P. Phillips & T. Srinavsan (eds), *Statistical Methods of Economics and Quantitative Economics: Essays in Honor of C.R. Rao*, Blackwell, Cambridge, pp. 254–275.

Newey, W. K., Powell, J. L. & Vella, F. (1999). Nonparametric estimation of triangular simultaneous equation models, *Econometrica* **67**: 565–603.

Nielsen, J. P. & Linton, O. (1998). An optimization interpretation of integration and backfitting estimators for separable nonparametric models, *Journal of the Royal Statistical Society, Series B* **60**: 217–222.

Nielsen, J. P. & Sperlich, S. (2002). Smooth backfitting in practice, *Working paper 02-59*, Universidad Carlos III de Madrid.

Opsomer, J. & Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression, *Annals of Statistics* **25**: 186–211.

Pagan, A. & Schwert, W. (1990). Alternative models for conditional stock volatility, *Journal of Econometrics* **45**: 267–290.

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

Park, B. U. & Marron, J. S. (1990). Comparison of data–driven bandwidth selectors, *Journal of the American Statistical Association* **85**: 66–72.

Park, B. U. & Turlach, B. A. (1992). Practical performance of several data driven bandwidth selectors, *Computational Statistics* **7**: 251–270.

Powell, J. L., Stock, J. H. & Stoker, T. M. (1989). Semiparametric estimation of index coefficients, *Econometrica* **57**(6): 1403–1430.

Proença, I. & Werwatz, A. (1995). Comparing parametric and semiparametric binary response models, *in* W. Härdle, S. Klinke & B. Turlach (eds), *XploRe: An Interactive Statistical Computing Environment*, Springer, pp. 251–274.

Robinson, P. M. (1988a). Root *n*–consistent semiparametric regression, *Econometrica* **56**: 931–954.

Robinson, P. M. (1988b). Semiparametric econometrics: A survey, *Journal of Applied Econometrics* **3**: 35–51.

Rodríguez-Póo, J. M., Sperlich, S. & Fernández, A. I. (1999). Semiparametric three step estimation methods for simultaneous equation systems, *Working paper 99-83 (32)*, Universidad Carlos III de Madrid.

Rodríguez-Póo, J. M., Sperlich, S. & Vieu, P. (2003). Semiparametric estimation of weak and strong separable models, *Econometric Theory* **19**: 1008–1039.

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

Ruppert, D., Wand, M. P. & Carroll, R. J. (1990). *Semiparametric Regression*, Cambridge University Press.

Schimek, M. G. (2000a). Estimation and inference in partially linear models with smoothing splines, *Journal of Statistical Planning and Inference* **91**: 525–540.

Schimek, M. G. (ed.) (2000b). *Smoothing and Regression*, Wiley, New York.

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

Scott, D. W. & Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation, *Journal of the American Statistical Association* **82**(400): 1131–1146.

Scott, D. W. & Wand, M. P. (1991). Feasibility of multivariate density estimates, *Biometrika* **78**: 197–205.

Severance-Lossin, E. & Sperlich, S. (1999). Estimation of derivatives for additive separable models, *Statistics* **33**: 241–265.

Severini, T. A. & Staniswalis, J. G. (1994). Quasi-likelihood estimation in semiparametric models, *Journal of the American Statistical Association* **89**: 501–511.

Severini, T. A. & Wong, W. H. (1992). Generalized profile likelihood and conditionally parametric models, *Annals of Statistics* **20**: 1768–1802.

Sheather, S. J. & Jones, M. C. (1991). A reliable data–based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society, Series B* **53**: 683–690.

Silverman, B. W. (1984). Spline smoothing: the equivalent variable kernel method, *Annals of Statistics* **12**: 898–916.

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

Simonoff, J. (1996). *Smoothing Methods in Statistics*, Springer, New York.

Speckman, P. E. (1988). Regression analysis for partially linear models, *Journal of the Royal Statistical Society, Series B* **50**: 413–436.

Sperlich, S. (1998). *Additive Modelling and Testing Model Specification*, Shaker Verlag.

Sperlich, S., Linton, O. & Härdle, W. (1999). Integration and backfitting methods in additive models: Finite sample properties and comparison, *Test* **8**: 419–458.

Sperlich, S., Tjøstheim, D. & Yang, L. (2002). Nonparametric estimation and testing of interaction in additive models, *Econometric Theory* **18**(2): 197–251.

Spokoiny, V. (1996). Adaptive hypothesis testing using wavelets, *Annals of Statistics* **24**: 2477–2498.

Spokoiny, V. (1998). Adaptive and spatially adaptive testing of a nonparametric hypothesis, *Math. Methods of Statist.* **7**: 245–273.

Staniswalis, J. G. & Thall, P. F. (2001). An explanation of generalized profile likelihoods, *Statistics and Computing* **11**: 293–298.

Stone, C. J. (1977). Consistent nonparametric regression, *Applied Statistics* **5**: 595–635.

Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates, *Annals of Statistics* **12**(4): 1285–1297.

Stone, C. J. (1985). Additive regression and other nonparametric models, *Annals of Statistics* **13**(2): 689–705.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models, *Annals of Statistics* **14**(2): 590–606.

Stone, C. J., Hansen, M. H., Kooperberg, C. & Truong, Y. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion), *Annals of Statistics* **25**: 1371–1470.

Stute, W. (1997). Nonparametric model checks for regression, *Annals of Statistics* **25**: 613–641.

Stute, W., González Manteiga, W. & Presedo-Quindimi, M. (1998). Bootstrap approximations in model checks for regression, *Journal of the American Statistical Association* **93**: 141–149.

Tjøstheim, D. & Auestad, B. (1994a). Nonparametric identification of nonlinear time series: Projections, *Journal of the American Statistical Association* **89**: 1398–1409.

Tjøstheim, D. & Auestad, B. (1994b). Nonparametric identification of nonlinear time series: Selecting significant lags, *Journal of the American Statistical Association* **89**: 1410–1430.

Treiman, D. J. (1975). Problems of concept and measurement in the comparative study of occupational mobility, *Social Science Research* **4**: 183–230.

Vella, F. (1998). Estimating models with sample selection bias: A survey, *The Journal of Human Resources* **33**: 127–169.

Venables, W. N. & Ripley, B. (2002). *Modern Applied Statistics with S*, fourth edn, Springer, New York.

Vieu, P. (1994). Choice of regressors in nonparametric estimation, *Computational Statistics & Data Analysis* **17**: 575–594.

Wahba, G. (1990). *Spline models for observational data*, 2 edn, SIAM, Philadelphia, Pennsylvania.

Wand, M. P. & Jones, M. C. (1995). *Kernel Smoothing*, Vol. 60 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Watson, G. S. (1964). Smooth regression analysis, *Sankhyā, Series A* **26**: 359–372.

Wecker, W. & Ansley, C. (1983). The signal extraction approach to nonlinear regression and spline smoothing, *Journal of the American Statistical Association* **78**: 351–365.

Weisberg, S. & Welsh, A. H. (1994). Adapting for the missing link, *Annals of Statistics* **22**: 1674–1700.

Wu, C. (1986). Jackknife, bootstrap and other resampling methods in regression analysis (with discussion), *Annals of Statistics* **14**: 1261–1350.

Yang, L., Sperlich, S. & Härdle, W. (2003). Derivative estimation and testing in generalized additive models, *Journal of Statistical Planning and Inference* **115**(2): 521–542.

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

Zheng, J. (1996). A consistent test of a functional form via nonparametric estimation techniques, *Journal of Econometrics* **75**: 263–289.

Author Index

Subject Index