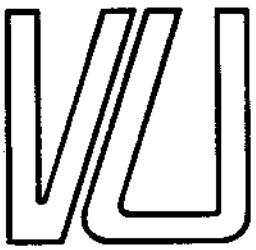
SERIE RESEARCH MEMORANDA

NONLINEAR REGRESSION WITH DISCRETE
EXPLANATORY VARIABLES

Herman J. Bierens

Research Memorandum 1988-61 Dec. '88



VRIJE UNIVERSITEIT
FACULTEIT DER ECONOMISCHE WETENSCHAPPEN
EN ECONOMETRIE
A M S T E R D A M

NONLINEAR REGRESSION WITH DISCRETE EXPLANATORY VARIABLES *

Herman J.Bierens
Department of Econometrics
Free University, Amsterdam

December 1988

---

* Chapter 8 of:
PRINCIPLES OF NONLINEAR AND NONPARAMETRIC REGRESSION ANALYSIS
To be published by Marcel Dekker Inc., New York.

PRINCIPLES OF NONLINEAR AND NONPARAMETRIC REGRESSION ANALYSIS

TABLE OF CONTENTS:

PREFACE

This book deals with statistical inference of nonlinear regression models from two opposite points of view, namely the case where the functional form of the model is completely specified as a known function of regressors and unknown parameters, and the opposite case where the functional form of the model is completely unknown. First it is assumed that the response function of the regression model under review belongs to a certain well-specified parametric family of functional forms, by which estimation of the model merely amounts to estimation of the unknown parameters. For this class of models we review the asymptotic properties of the nonlinear least squares estimator for independent data as well as for time series.

In practice assumptions on the functional form are often made on the basis of computational convenience rather than on the basis of precise a priori knowledge of the empirical phenomenon under review. Therefore the linear regression model is still the most popular model specification in applied research. However, even if the specification of the functional form is based on sound theoretical considerations there is quite often a large range of functional forms that are theoretically admissible, so that there is no guarantee that the actually chosen functional form is true. Functional specification of a parametric nonlinear regression model should therefore always be verified by conducting model misspecification tests. Various model misspecification tests will therefore be discussed, in particular consistent tests which have asymptotic power 1 against all deviations from the null hypothesis that the model is correct.

The opposite case of parametric regression is nonparametric regression. Nonparametric regression analysis is concerned with estimation of a regression model without specifying in advance its functional form. Thus the only source of information about the functional form of the model is the data set itself. In this book we shall review various nonparametric regression approaches, with special emphasis on the kernel method, under various distributional assumptions.

This book is divided into three parts. In the first part we review the elements of abstract probability theory we need in part 2. Part 2 is devoted to the asymptotic theory of para-

metric and nonparametric regression analysis in the case of independent data generating processes. In part 3 we extend the analysis involved to time series.

The selection of the topics mainly reflexes my own interest in the subject. Instead of providing an encyclopedic survey of the literature, I have chosen for a setup which aims to fill the gap between intermediate statistics (including linear time series analysis) and the level necessary to get access to the recent literature on nonlinear and nonparametric regression analysis, with emphasis on my own contributions. The ultimate goal is to provide the student with the tools for his own independent research in this area, by showing what tools I and others have used and what they have been used for. Thus, this book may be viewed as an account of my own struggle with the material involved. I think this book is particularly suitable for self-tuition (at least it aims to be), and may prove useful in a graduate course in mathematical statistics and advanced econometrics.

Acknowledgements:

The first five chapters of this book have been disseminated in draft form as working papers. I am grateful to Anil Bera, Alexander Georgiev and Jan Magnus for suggesting additional references, and in particular to Lourens Broersma, Johan Smits and Ton Steerneman who suggested various improvements.

A large body of the material in chapter 6 has been published earlier in Truman F.Bewley (ed.), *Advances in Econometrics, Fifth World Congress*, Cambridge University Press. I am indebted to Cambridge University Press for granting permission to reprint it.

8. NONLINEAR REGRESSION WITH DISCRETE EXPLANATORY VARIABLES

In this chapter we present the approach of Bierens and Hartog (1988) for specifying, estimating and testing regression models with discrete explanatory variables. Bierens and Hartog developed this approach in close harmony with an empirical application to the earnings function. The earnings function involved relates the log of individual hourly wages to job characteristics and personal characteristics, i.e. level of education, sex, age, company experience and the nature of wage bargaining. As the emphasis in this chaptere will be on the statistical theory, we will be rather casual about this earnings function and we will not try to read in the empirical results.

A typical feature of the explanatory variables in this earnings function is that they take on a finite number of values. It will be shown that such regression models take the form of a finite-order polynomial of a linear function of the regressors. Therefore we propose a two-stage estimation procedure, where in the first stage the linear function involved is estimated by ordinary linear least squares (OLS) and in the second-stage the polynomial involved is estimated by regression on orthogonal polynomials. Moreover, we propose a number of tests for testing the order of the polynomial and the redundancy of explanatory variables.

The plan of this chapter is as follows. In section 8.1 we briefly discuss the earnings function. In section 8.2 we show that regression models with discrete explanatory variables take the form of a polynomial of a linear combination of the regressors. Section 8.3 is devoted to the problem of how this linear combination should be specified and estimated. Section 8.4 deals with estimating and testing the regression function. Finally, the proofs of the main theorems are given in section 8.5.

8.1 *The earnings function*

The earnings function estimated by Bierens and Hartog relates the log of gross hourly wages (measured in 0.01 guilders) to the following five explanatory variables:

$x^{(1)}$ - level of education, ranging from 1 to 7:
1 = basic;
2 = lower vocational;
3 = intermediate general;
4 = intermediate vocational;
5 = higher general;
6 = higher vocational;
7 = university.

$x^{(2)}$ - sex:
1 = male;
2 = female.

$x^{(3)}$ - age in full years, ranging from 16 to 64.

$x^{(4)}$ - experience with the present employer, in full years, ranging from 1 to 50.

$x^{(5)}$ - collective agreement:
1 = wages set in a collective agreement;
2 = wages not set in a collective agreement.

The data were taken from the Dutch Wage Structure Survey 1979 [CBS (1979)] collected by the Dutch national statistical office, CBS. The Wage Structure Survey (WSS) is a representative wage survey, where data on individual workers are taken from the administration of firms and institutions. In comparison to individual surveys, this allows a careful observation of earnings, using well-defined concepts. The hourly earnings were found from dividing reported earnings per payment interval (week, month, etc.) by reported hours of work.

The data set Bierens and Hartog worked with consists of 2000 observations (Y_j, X_j), $j=1,2.....,n=2000$, where Y_j is the natural logarithm of gross hourly wages of individual j, and

$$X_j = (X_j^{(1)}, \ldots\ldots, X_j^{(6)}) \in \Xi$$

is the corresponding vector of explanatory variables specified above, including a constant term $X_j^{(6)} = 1$, with

$$\Xi = \{(x^{(1)},\ldots,x^{(6)})' : x^{(1)}\in\{1,2,\ldots,7\},\ x^{(2)}\in\{1,2\},\ x^{(3)}\in\{16,17,\ldots,64\},\ x^{(4)}\in\{1,\ldots,50\},\ x^{(5)}\in\{1,2\},\ x^{(6)} = 1\}$$

the space of regressors.

Now the earnings function is the response function g(x) of the nonlinear regression model

$$Y_j = g(X_j) + U_j;\ X_j \in \Xi;\ E(U_j|X_j) = 0 \text{ a.s.}, \quad j=1,2,\ldots,n,\ldots \tag{8.1.1}$$

A typical feature of this earnings function, and in fact of all empirical earnings functions considered in the literature, is that the explanatory variables are discrete. In particular in the present case the space of explanatory variables, Ξ, is finite. As will be shown in section 8.2 below, this feature will enable us to determine the *exact* functional form of the regression function g(x), namely as a finite polynomial of a linear function of the regressors.

8.2 *The functional form of a regression model with discrete explanatory variables*

In this section it will be shown that the regression function g(x) of model (8.1.1) takes the form of a finite polynomial of a linear function of the regressors. In order to illustrate our main point, consider the following nonlinear regression model:

$$Y_j = g(X_{1,j}, X_{2,j}) + U_j \quad , j = 1,2\ldots, \tag{8.2.1}$$

where $E(U_j|X_{1,j}, X_{2,j}) = 0$ a.s. , $X_j' = (X_{1,j}, X_{2,j})$ is a two-components vector contained with probability 1 in the set

$$\Xi = \{(0,0)', (1,0)', (0,1)', (1,1)'\} = \{x_1, x_2, x_3, x_4\} \tag{8.2.2}$$

and g is *any* real function defined on Ξ. Moreover, let

$$\theta' = (1,-2). \tag{8.2.3}$$

Then it is easy to verify (cf. exercise 1) that

$$g(x) = \alpha_0 + \alpha_1(\theta'x) + \alpha_2(\alpha'x)^2 + \alpha_3(\alpha'x)^3 \text{ for } x \in \Xi, \tag{8.2.4}$$

where

$$(\alpha_0, \alpha_1, \alpha_2, \alpha_3)' = \alpha = B^{-1}g$$

with

$$g = (g(x_1), g(x_2), g(x_3), g(x_4))'$$

and

$$B = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & -2 & 4 & -8 \\ 1 & -1 & 1 & -1 \end{bmatrix}.$$

This easy result is not specific for this particular θ, but it holds for all linear functions $\theta'x$ which are one-to-one mappings from Ξ into R. Such vectors θ will be called *linear separators* of Ξ. More generally:

Definition 8.2.1. A vector θ is a linear separator of a countable subset Ξ of R^k if for all pairs $(x_1, x_2) \in \Xi \times \Xi$, $\theta'x_1 = \theta'x_2$ implies $x_1 = x_2$.

The existence of a linear separator of a countable set is always guaranteed. In fact the set of all linear separators is uncountable, as is illustrated by the following theorem.

Theorem 8.2.1. Let Ξ be a countable subset of R^k and let S be the set of all linear separators of Ξ. Then the set $R^k \backslash S$ has Lebesgue measure zero.

Proof: Let

$$C = \{(x_1, x_2) : x_1 \in \Xi,\ x_2 \in \Xi,\ x_1 \neq x_2\}$$

and let

$$T(x_1, x_2) = \{\theta \in R^k : \theta'(x_1 - x_2) = 0\}.$$

For $x_1 \neq x_2$ the set $T(x_1, x_2)$ is of lower dimension than k, hence $T(x_1, x_2)$ has Lebesgue measure zero. Now

$$R^k \backslash S = \cup_{(x_1, x_2) \in C} T(x_1, x_2)$$

is a countable union of sets with Lebesgue measure zero and therefore a set with Lebesgue measure zero itself. Q.E.D.

From definition 8.2.1 it follows that for any linear separator θ of Ξ each point in the range of $\theta'x$ ($x \in \Xi$) can uniquely be associated to a point in the domain Ξ, and vice versa. Thus $\theta'x$ is a one-to-one mapping from Ξ into R. From theorem 3.2.1(VIII) it therefore follows:

$$E(Y|X) = E(Y|\theta'X) \text{ a.s.},$$

provided the left and right-hand side conditional expectations exist. We recall that $E|Y| < \infty$ is a sufficient condition for the latter. Thus:

Theorem 8.2.2. Let (Y,X) be a random vector in $R \times \Xi$, where Ξ is a countable subset of R^k and $E|Y| < \infty$. For any linear separator θ of Ξ there exists a Borel measurable real function φ_θ on R such that

$$Y = \varphi_\theta(\theta'X) + U, \text{ with } E(U|X) = 0 \text{ a.s.}$$

Note that we do not require that Ξ is finite. Thus theorem 8.2.2 also holds if Ξ is the whole space Q^k, i.e., the k-dimensional space of vectors with rational-valued components, for the set of rational numbers is countable [see Royden (1968, Proposition 6 on page 21)].

In the case that Ξ is finite it is easy to verify that similarly to (8.2.4) the function φ_θ is a polynomial of finite order, where the order involved is less than or equal to the number of elements in Ξ minus 1. Realizing that without loss of generality we may assume that Ξ only contains points with positive probability, we now have the following result.

Theorem 8.2.3. Let the conditions of theorem 8.2.2 be satisfied and let

$$\Xi^* = \{x^* \in \Xi : P(X = x^*) > 0\}.$$

If Ξ^* is finite of size m, then for every linear separator θ of Ξ^* there exist real numbers $\alpha_0(\theta)$, $\alpha_1(\theta), \ldots \alpha_{m-1}(\theta)$, depending

on θ, such that

$$\varphi_\theta(\theta'x) = \Sigma_{\ell=0}^{m-1}\alpha_\ell(\theta)(\theta'x)^\ell \text{ a.s. for } x \in \Xi^*.$$

The main results in this section may now be summarized as follows. Assume:

Assumption 8.2.1. The data generating process $\{(Y_j, X_j)\}$, $j=1,2,..$ is an i.d.d. stochastic process in $R \times \Xi$, where Ξ is an finite subset of R^k such that $x \in \Xi$ implies $P(X_j = x) > 0$. Moreover, $E|Y_j|^2 < \infty$.

Then for any linear separator θ of Ξ the regression model takes the form

$$Y_j = \Sigma_{\ell=0}^{m-1}\alpha_\ell(\theta)(\theta'X_j)^\ell + U_j, \text{ with } E(U_j|X_j) = 0 \text{ a.s.}, \tag{8.2.5}$$

where m is less than or equal to the size of Ξ.

Note that assumption 8.2.1 is more restrictive than needed for the result (8.2.5), for (8.2.5) also holds if the data generating process is identically distributed but not independent, and if $E|Y_j| < \infty$. However, the additional conditions in assumption 8.2.1 will be needed in the sequel. Moreover, note that the polynomial representation (8.2.5) is only one way of modelling regression models with discrete explanatory variables. For example we may replace $\theta'X_j$ in (8.2.5) by $\psi(\theta'X_j)$, where ψ is an arbitrary one-to-one mapping from R into R. Model (8.2.5) is, however, easier to handle than the latter class of models, and therefore we shall proceed on the basis of model (8.2.5).

Remarks:
Model (8.2.5) is a special case of a projection pursuit regression (PPR) model. See Friedman and Stuetzle (1981), Huber (1985) and the references therein. In PPR the unknown regression function $g(x)$, $x \in R^k$, is approximated by a sum of univariate functions s_ℓ of linear combinations $\theta_\ell'x$:

$$\tilde{g}(x) = \Sigma_{\ell=0}^{m} s_\ell(\theta_\ell'x).$$

Gallant's (1981) Fourier flexible functional form is another special case of PPR, where the θ_ℓ are so-called multi-indices and the functions $s_\ell(.)$ are of the form $\alpha_\ell\cos(.) + \beta_\ell\sin(.)$, with α_ℓ and β_ℓ parameters to be estimated (by say least squares). Friedman and Stuetzle propose a recursive algorithm for estimating θ_ℓ and s_ℓ. A disadvantage of this approach, however, is that (quoting Huber) "the sampling theory of PPR is practically non-existent". The sampling theory of Gallant's Fourier flexible form is more developed, as consistency results are available for m increasing with the sample size n in some order. See Gallant (1985) for a general consistency proof. However, asymptotic distribution theory is yet absent, due to the fact that in general regression functions can only be represented exactly by a Fourier flexible functional form if $m = \infty$. In our case where the x are finite-valued it is possible te represent the regression function g(x) by a finite Fourier expansion and even by a univariate finite Fourier expansion in $\theta'x$ with θ a linear separator of Ξ. Again, however, model (8.2.5) seems easier to handle than such a Fourier expansion in $\theta'x$, in particular because the polynomial representation allows estimation by using orthogonal polynomials.

Thus, the general PPR method is not yet a practical alternative to our approach. On the other hand, Gallant's Fourier flexible functional form is a serious competitor. Which method is better depends on the nature of the data generating process, i.e., it is conceivable that in some cases our approach will do with a lower m than the Fourier flexible functional form, and vice versa in other cases. Therefore one should view the linear separator approximation in this chapter as an addition to the menu of possible ways of accounting for nonlinearity.

Exersises:

1. Prove (8.2.4).
2. Prove theorem 8.2.1 by showing that $P[\theta'(x_1-x_2)] = 0$ for all $x_1, x_2 \in S$, $x_1 \neq x_2$ and θ an absolutely continuously distributed random vector in R^k.

8.3 *The choice of the linear separator*

At first sight one might think of estimating the parameters θ and $\alpha_0,\ldots,\alpha_{m-1}$ of model (8.2.5) by nonlinear least squares. One of the problems is, however, that the parametric representation involved is far form unique, as (8.2.5) holds for any linear separator. Therefore these parameters cannot be estimated consistenly by nonlinear least squares (cf. theorem 4.2.1). Only if we choose in advance a linear separator θ the remaining parameters $\alpha_1,\ldots,\alpha_{m-1}$ can in principle be consistently estimated by polynomial least squares. However, the latter procedure will be hampered by numerical problems if m is large [see Seber (1977, sections 8.1 and 8.2)]. For example in the case of the earnings function under review the maximum size of the space of regressors is $7\times2\times49\times50\times2 = 68600$. Although not all points in this set Ξ will have positive probability mass, it is clear that for an arbitrary chosen linear separator the necessary order of the polynomial will probably be unmanageably large.

Intuitively one may feel that the order of the polynomial is not independent of the choice of the linear separator θ. If the linear separator is such that the shape of the regression function $g(x) = E(Y_j|X_j = x)$ is very different from that of the linear function $\theta'x$, it will be necessary to compensate this by a large m. For example, if the regression function g of model (8.2.1) is

$$g(X_{1,j},X_{2,j}) = X_{1,j} + 2\cdot X_{2,j}$$

and if we choose the linear separator θ as in (8.2.3) then (8.2.4) becomes

$$g(x) = (-7/3)(\theta'x) + 2(\theta'x)^2 + (4/3)(\theta'x)^3$$

for $x \in \Xi$ [defined by (8.2.2)]. Thus if we choose the linear separator $\theta' = (1,-2)$ then the necessary order of the polynomial equals the size of Ξ minus 1, whereas using the linear separator $\theta' = (1,-2)$ would give us the true model at once.

In view of the above argument the best choice of the linear separator seems therefore be the one for which the order of the polynomial is as small as possible. However, we have not succeeded in finding a practical criterion for classifying the

linear separators according to the order of the corresponding polynomials. Therefore we propose as a 'second best' procedure to choose the vector of OLS estimators,

$$\hat{\theta} = [(1/n)\Sigma_{j=1}^{n} X_j X_j']^{-1}[(1/n)\Sigma_{j=1}^{n} X_j Y_j]$$

as the linear separator, for $\hat{\theta}'X$ will probably be close to the true regression function, especially if one of the components of X_j is a constant term. Choosing $\hat{\theta}$ as a linear separator will likely have a favourable influence on the order of the polynomial. The question remains, of course, whether $\hat{\theta}$ converges to a linear separator. This point will be dealt with in the next section, together with the problem of testing whether a specified m is sufficiently large.

The linear regression model

$$\hat{Y}_j = \hat{\theta}'X_j$$

is, in our approach, nothing more than the best linear approximation of the true regression model. Nevertheless the asymptotic properties of $\hat{\theta}$ are quite similar to those of the OLS estimator of the classical linear regression model, as has been shown by White (1980). Employing the additional condition

Assumption 8.3.1. The matrix $E\, X_j X_j'$ is nonsingular,

these asymptotic properties can be summarized as follows.

Theorem 8.3.1. Let assumptions 8.2.1 and 8.3.1 be satisfied and let

$$\theta_0 = (E\, X_1 X_1')^{-1} E\, X_1 Y_1,$$

$$\Omega = (E\, X_1 X_1')^{-1}[E(Y_1 - \theta_0' X_1)^2 X_1 X_1'](E\, X_1 X_1')^{-1},$$

$$\hat{\Omega} = ((1/n)\Sigma_{j=1}^{n} X_j X_j')^{-1}[(1/n)\Sigma_{j=1}^{n}(Y_j - \hat{\theta}' X_j)^2 X_j X_j'] \times (1/n)\Sigma_{j=1}^{n} X_j X_j')^{-1}, \qquad (8.3.1)$$

Then for $n \to \infty$,

$\hat{\theta} \to \theta_0$ a.s., $\sqrt{n}(\hat{\theta} - \theta_0) \to N_k(0,\Omega)$ in distr. and

$\hat{\Omega} \to \Omega$ a.s.

Exercise:

1. Prove theorem 8.3.1.

8.4 *Estimating and testing the regression function*

8.4.1 *Estimation*

Using the OLS statistic $\hat{\theta}$ as a linear separator the model becomes a linear model in the parameters α_0, $\alpha_1,\ldots,\alpha_{m-1}$ and the variables

$$\hat{\theta}'X_j,\ (\hat{\theta}'X_j)^2,\ldots,(\hat{\theta}'X_j)^{m-1}.$$

However, applying OLS to estimate the α may be hampered by numerical instability if m is large, due to the fact that then the matrix with elements

$$(1/n)\Sigma_{j=1}^{n}(\hat{\theta}'X_j)^{\ell_1+\ell_2},\ \ell_1,\ell_2=0,1,\ldots,m-1$$

will probably be nearly singular (see Seber 1977, Section 8.1). A neat cure for this problem is suggested by Forsythe (1957) and others, namely to use orthogonal polynomials. Moreover, numerical stability of polynomial regressions can be further improved by standardizing the variables into the interval [+1,1]. In our case the variables involved are $\theta'X_j$ with θ a linear separator, which will be standardized into the interval [-1,1] by using the transformation

$$z(x,\theta) = [2\cdot\theta'x - M_1(\theta) - M_2(\theta)]/[M_1(\theta) - M_2(\theta)] \quad (8.4.1)$$

where

$$M_1(\theta) = \Sigma_{i=1}^{k}\max_{x\in\Xi}\theta_i x^{(i)} \text{ and } M_2(\theta) = \Sigma_{i=1}^{k}\min_{x\in\Xi}\theta_i x^{(i)},$$

with

$$(\theta_1,\ldots,\theta_k)' = \theta \in R^k \text{ and } (x^{(1)},\ldots,x^{(k)})' = x \in \Xi.$$

Thus we now propose to rewrite model (8.2.5) for j=1,2,...,n as

$$Y_j = \Sigma_{\ell=0}^{m-1}\gamma_\ell(\theta)\psi_\ell[z(X_j,\theta)|\theta] + U_j, \text{ with } E(U_j|X_j) = 0 \text{ a.s.} \tag{8.4.2}$$

where the $\psi_\ell(z|\theta)$ are orthogonal polynomials of order ℓ, respectively.

Forsythe (1957) proposes to generate orthogonal polynomials $P_\ell(.)$ on basis of a data set $\{z_1,..,z_n\}$ by the following recurrence relation:

$$\begin{aligned} P_0(z_j) &= 1/\sqrt{n}\ ,\ P_1(z_j) = (z_j-\bar{z})/\sqrt{[\Sigma_{j=1}^n(z_j-\bar{z})^2]}, \\ P_r^*(z_j) &= (P_1(z_j) - d_{r,r-1})P_{r-1}(z_j) - d_{r,r-2}P_{r-2}(z_j),\ r \geq 2 \\ P_r(z_j) &= P_r^*(z_j)/d_{r,r}, \end{aligned} \tag{8.4.3}$$

where

$$\bar{z} = (1/n)\Sigma_{j=1}^n z_j,$$

$$d_{r,r-2} = \Sigma_{j=1}^n P_1(z_j)P_{r-1}(z_j)P_{r-2}(z_j),$$

$$d_{r,r-1} = \Sigma_{j=1}^n P_1(z_j)P_{r-1}^2(z_j),$$

$$d_{r,r} = \{\Sigma_{j=1}^n P_r^*(z_j)^2\}^{\frac{1}{2}}.$$

By construction these polynomials satisfy

$$\begin{aligned} \Sigma_{j=1}^n P_{r_1}(z_j)P_{r_2}(z_j) &= 1 \text{ if } r_1 = r_2, \\ \Sigma_{j=1}^n P_{r_1}(z_j)P_{r_2}(z_j) &= 0 \text{ if } r_1 \neq r_2. \end{aligned} \tag{8.4.4}$$

In this paper we shall use orthogonal polynomials of the type

$$\psi_r(.|\theta) = \sqrt{n}\ P_r(.|\theta),$$

where $P_r(.|\theta)$ is generated by (8.4.3) with $z_j = z(X_j,\theta)$, so that

$$(1/n)\Sigma_{j=1}^{n}\psi_{r_1}(z(X_j,\theta)|\theta)\psi_{r_2}(z(X_j,\theta)|\theta) = 1 \quad \text{if} \quad r_1 = r_2 \,,$$
$$(1/n)\Sigma_{j=1}^{n}\psi_{r_1}(z(X_j,\theta)|\theta)\psi_{r_2}(z(X_j,\theta)|\theta) = 0 \quad \text{if} \quad r_1 \neq r_2 \,. \tag{8.4.5}$$

This will prove to be more convenient than (8.4.4).

The difference between the two models (8.2.5) and (8.4.2) is that (8.2.5) is true for all $j \geq 1$, whereas (8.4.2) is only true for $j=1,2,\ldots,n$, given that θ is a linear separator and m is sufficiently large.

By virtue of (8.4.5) the least squares estimators of the parameters $\gamma_\ell(\theta)$ of model (8.4.2) can now simply be calculated by

$$\hat{\gamma}_\ell(\theta) = (1/n)\Sigma_{j=1}^{n} Y_j \psi_\ell(z(X_j,\theta)|\theta), \quad \ell = 0,1,2,\ldots \tag{8.4.6}$$

In practice we will use $\hat{\theta}$ instead of θ as a linear separator. We then have:

Theorem 8.4.1. Let the conditions of theorem 8.3.1 be satisfied and assume that θ_0 is a linear separator of Ξ. Then for $n \to \infty$ and fixed $\ell = 0,1,2\ldots$,

$$|\hat{\gamma}_\ell(\hat{\theta}_0) - \gamma_\ell(\theta_0)| \to 0 \quad \text{a.s}$$

$$\sup_{|z|\leq 1}|\psi_\ell(z|\hat{\theta}) - \psi_\ell(z|\theta_0)| \to 0 \quad \text{a.s.}$$

Moreover, denoting

$$\hat{g}_m(x|\hat{\theta}) = \Sigma_{\ell=0}^{m-1}\hat{\gamma}_\ell(\hat{\theta})\psi_\ell(z(x,\hat{\theta})|\hat{\theta})$$

we have

$$\sup_{x\in\Xi}|\hat{g}_m(x|\hat{\theta}) - g(x)| \to 0 \quad \text{a.s.},$$

provided m is large enough.

Proof: Exercise 1.

Thus $\hat{g}_m(x)$ is a uniformly consistent estimator of the true regression function g(x) if θ_0 is a linear separator and m is large enough.

8.4.2 *Model specification testing*

Next we consider the problem of how to test whether m is sufficiently large and θ_0 is a linear separator. Let for $\theta \in R^k$ and $\tau \in R$,

$$\hat{\rho}_{m,j}(\tau|\theta) = \exp[\tau z(X_j,\theta)] - \Sigma_{\ell=0}^{m-1}\{\psi_\ell(z(X_j,\hat{\theta})|\hat{\theta}) \times (1/n)\Sigma_{j_0=1}^{n}\psi_\ell(z(X_{j_0},\hat{\theta})|\hat{\theta})\exp[\tau z(X_{j_0},\theta)]\} \quad (8.4.7)$$

$$\hat{\xi}_m(\tau|\theta) = (1/n)\Sigma_{j=1}^{n}(\partial/\partial\theta')\hat{g}_m(X_j|\hat{\theta})\exp[\tau z(X_j,\theta)], \quad (8.4.8)$$

$$\hat{s}_m^2(\tau|\theta) = (1/n)\Sigma_{j=1}^{n}(Y_j-\hat{g}_m(X_j|\hat{\theta}))^2\hat{\rho}_{m,j}(\tau|\theta)^2 -2[(1/n)\Sigma_{j=1}^{n}(Y_j-\hat{g}_m(X_j|\hat{\theta}))(Y_j-\hat{\theta}'X_j)\hat{\rho}_{m,j}(\tau|\theta)X_j'] \times[(1/n)\Sigma_{j=1}^{n}X_jX_j']^{-1}\hat{\xi}_m(\tau|\theta) + \hat{\xi}_m(\tau|\theta)'\hat{\Omega}\,\hat{\xi}_m(\tau|\theta), \quad (8.4.9)$$

and

$$\hat{\eta}_m(\tau|\theta) = ((1/\sqrt{n})\Sigma_{j=1}^{n}(Y_j-\hat{g}_m(X_j|\hat{\theta}))\exp[\tau z(X_j,\theta)]) /\sqrt{(\hat{s}_m^2(\tau|\theta))} \quad (8.4.10)$$

where $\hat{\Omega}$ is defined by (8.3.1). Then we have:

Theorem 8.4.2. Let assumptions 8.2.1 and 8.3.1 be satisfied and let θ^* be an arbitrary linear separator of Ξ. If

H_0 : m and $\theta=\theta_0$ are such that model (8.2.5) is true (8.4.11)

then for $n \to \infty$ and $\tau \neq 0$

$$\hat{\eta}_m(\tau|\theta^*) \to N(0,1) \text{ in distr.} \quad (8.4.12)$$

If the null hypothesis (8.4.11) is false then there exists a countable subset T of R (depending on θ^*) such that for every $\tau \notin T$,

$$|\hat{\eta}_m(\tau|\theta^*)| \to \infty \quad \text{a.s.} \quad (8.4.13)$$

Moreover, under the maintained hypothesis that θ_0 is a linear separator the above conclusions also hold for $\theta^* = \hat{\theta}$. Furthermore, there exists a subset S of R^{k+1} with Lebesgue measure zero such that under the alternative that (8.4.11) is false, (8.4.13) holds for every $(\tau,\theta^*) \notin S$.

Proof: Section 8.5.1.

Remark: The result (8.4.13) is due to the fact that under H_1 there exists a $c_m(\tau|\theta^*) \neq 0$ and a $\sigma_m^2(\tau|\theta^*)$ such that

$$\hat{\eta}_m(\tau|\theta^*) - c_m(\tau|\theta^*)\sqrt{n} \to N[0,\sigma_m^2(\tau|\theta^*)] \quad \text{in distr.} \quad (8.4.14)$$

We see from theorem 8.4.2 that the power of the test depends on the appropriate choice of τ and θ^*. Although we may substitute $\theta^* = \theta$, the test might then no longer be consistent, as we then only test

$$\Sigma_{\ell=0}^{m-1}\gamma_\ell(\theta_0)\psi_\ell(z(X_j,\theta_0)|\theta_0) = E(Y_j|\theta_0'X_j) \text{ a.s.}$$

However, if θ_0 is not a linear separator this conditional expectation need no longer be equal to $E(Y_j|X_j)$. Thus, in the first instance, we should not use $\theta^* = \hat{\theta}$.

Theorem 8.2.1 suggests that we may choose θ^* randomly from a continuous distribution. But how does this random choice of θ^* affect the asymptotic properties of the test statistc involved? The following corollary of theorem 8.4.2 shows that it does not, and we may even choose τ randomly from a continuous distribution.

Theorem 8.4.3. Let (θ^*,τ) be a random drawing from a continuous (k+1)-variate distribution. Then (8.4.12) holds if (8.4.11) is true and (8.4.13) holds if (8.4.11) is false.

Proof: Section 8.5.2.

Thus we may for example draw τ and the components of θ^* independently from a uniform distribution, and then conduct the test in the same way as before.

Having tested and accepted a particular m in the above way, we might reduce m to m_*, say, where $0 \leq m_* < m$, by testing whether

$$H_0: \gamma_{m_*}(\theta_0) = \ldots = \gamma_{m-1}(\theta_0) = 0. \qquad (8.4.15)$$

This test can be conducted by using the Wald test on basis of the following result. Denote

$$\hat{\Gamma}_m = [(\partial/\partial\theta')\hat{\gamma}_0(\hat{\theta}), \ldots, (\partial/\partial\theta')\hat{\gamma}_{m-1}(\hat{\theta})]'$$

$$\hat{\Psi}_{j,m} = [\psi_0(z(X_j, \hat{\theta})|\hat{\theta}), \ldots, \psi_{m-1}(z(X_j, \hat{\theta})|\hat{\theta})]'$$

$$\hat{\Delta}_m = (1/n)\Sigma_{j=1}^n (Y_j - \hat{g}_m(X_j|\hat{\theta}))^2 \hat{\Psi}_{j,m}\hat{\Psi}_{j,m}'$$

$$\hat{\Sigma}_m = (1/n)\Sigma_{j=1}^n \{(Y_j - \hat{g}_m(X_j|\hat{\theta}))^2 \hat{\Psi}_{j,m}X_j'\}\{(1/n)\Sigma_{j=1}^n X_j X_j'\}^{-1}$$

and

$$\hat{\Lambda}_m = \hat{\Gamma}_m \hat{\Omega} \hat{\Gamma}_m' + \hat{\Gamma}_m \hat{\Sigma}_m' + \hat{\Sigma}_m \hat{\Gamma}_m' + \hat{\Delta}_m$$

where $\hat{\Omega}$ is defined by (8.3.1). Then we have:

Theorem 8.4.4. Let the conditions of theorem 8.3.1 be satisfied and assume that the hypothesis (8.4.11) holds. Then there exists a positive definite matrix Λ_m such that for $n \to \infty$,

$$\sqrt{n}[\hat{\gamma}_0(\hat{\theta}) - \gamma_0(\theta_0), \ldots, \hat{\gamma}_{m-1}(\hat{\theta}) - \gamma_{m-1}(\theta_0)]' \to N_m(0, \Lambda_m) \text{ in distr.}$$

and

$$\hat{\Lambda}_m \to \Lambda_m \quad \text{a.s.}$$

Proof: Exercise 2.

Now the Wald test for testing (8.4.15) can be conducted as follows. Let $\hat{\Lambda}_{m_*}^*$ be the submatrix of $\hat{\Lambda}_m$ corresponding to $(\hat{\gamma}_{m_*}(\hat{\theta}), \ldots, \hat{\gamma}_{m-1}(\hat{\theta}))$. Then under the conditions of theorem 8.4.4,

$$\hat{w}_{m_*} = n[\hat{\gamma}_{m_*}(\hat{\theta}),\ldots,\hat{\gamma}_{m-1}(\hat{\theta})]\hat{\Lambda}_{m_*}^{-1}[\hat{\gamma}_{m_*}(\hat{\theta}),\ldots,\hat{\gamma}_{m-1}(\hat{\theta})]'$$

$$\to \chi^2_{m-m_*} \text{ in distr. if (8.4.15) is true} \qquad (8.4.16)$$

$$\to \infty \text{ a.s. if (8.4.15) is false}$$

The ultimate purpose of most empirical econometric analysis is to determine which explanatory variables are important in the model under review and which are not. However, since the models (8.2.5) and (8.4.2) are true for any linear separator, provided m is sufficiently large, it is clear that OLS approximation results are not conclusive with respect to this question. For example, consider an i.d.d. sample $\{(Y_1,X_1),\ldots(Y_n,X_n)\}$ of random vectors in $R \times \Xi$, where

$$\Xi = \{(1,0)',(2,0)',(3,1)',(4,1)'\} \subset R^2 .$$

Suppose

$$P[X_j=(1,0)'] = P[X_j=(2,0)']$$

$$= P[X_j=(3,1)'] = P[X_j=(4,1)'] = 1/4,$$

$$Y_j = 0.44X_{1,j}^2 + U_j \ , \ E\,U_j = 0 \ , \ E\,U_j^2 = \sigma^2 < \infty \ ,$$

where $X_{1,j}$ is the first component of X_j and U_j and X_j are mutually independent. Then the OLS statistic $\hat{\theta}$ converges a.s. to $\theta_0 = (1,2)'$, which is obviously a linear separator of Ξ. Since both components of θ_0 are non-zero, we will find that both components of $\hat{\theta}$ are significant, while only the first component of X_j plays a role in the true model.

So the question arises how to test whether one or more components of X_j are redundant. To answer this question, we return to the general case and assume that the components $p_1,\ldots,p_r$ $(1 \le p_1 < \ldots < p_r \le k)$ of X_j are redundant. Let X_j^* be the vector of components of X_j with the components $p_1,\ldots,p_r$ fixed on say their minimum values. Now conduct the test in theorems 8.4.2 and 8.4.3 as before on the basis of the data set

$$\{(Y_1,X_1^*),\ldots,(Y_n,X_n^*)\}.$$

This yields a statistic $\hat{\eta}_m(\tau|\theta,p_1,\ldots,p_r)$ for which the follow-

ing variant of theorem 8.4.3 holds.

Theorem 8.4.5. Let assumptions 8.2.1 and 8.3.1 be satisfied and suppose that the hypothesis (8.4.11) is true. Let τ be an independent random drawing from a continuous distribution on R. Then

$$\hat{\eta}_m(\tau|\hat{\theta},p_1,\ldots,p_r) \to N(0,1) \quad \text{in distr.}$$

if the components $p_1, p_2,\ldots,p_r$ of X_j are redundant, and

$$|\hat{\eta}_m(\tau|\hat{\theta},p_1,\ldots,p_r)| \to \infty \quad \text{a.s.}$$

if at least one of these components is not redundant.

8.4.3 *The selection of the polynomial order*

In our empirical application we shall apply the above approach in the following way. First we choose a priori an m and we apply theorem 8.4.3 to test whether m is large enough and θ_0 is a linear separator. If the test accepts the null hypothesis involved we then apply the Wald test in theorem 8.4.4 to reduce m. Finally we apply theorem 8.4.5 in order to test the redundancy of the explanatory variables.

This procedure involves a number of sequential tests depending on the outcome of the first test. If the test in theorem 8.4.3 would reject the specification of m and if this is due to too low a value of m and not due to θ_0 failing to be a linear separator of Ξ, then it is possible to increase m stepwise until it is accepted. This procedure yields an m which is determined by the data, and is therefore an integer valued random variable. Now the question arises how this random m affects the tests in theorem 8.4.4 and 8.4.5.

The answer depends on the way the tests in theorem 8.4.3 and (8.4.16) are conducted. In particular it makes a big difference whether we use a fixed critical value or a critical value growing at order $o(\sqrt{n})$ with the sample size n. In the first case the type I error is fixed, and in the latter case the type I error vanishes as $n\to\infty$. Since the test in theorem 8.4.3 is consistent, in both cases the type II errors vanish. Thus if we use a fixed critical value, the hazard exists of

overshooting the true order of the polynomial, due to the persisting type I error. In that case the resulting m remains a random variable in the limit. In the case we use the increasing critical value the final m emerging from the sequential test procedure is a consistent estimator of the true m.

Thus the sequential test procedure may now be summarized as follows. Select an initial $m=m_0$ and a sequence (L_n) of critical values converging to infinity at order $o(\sqrt{n})$.

Step 1: Apply the test in theorem 8.4.3. If

$$|\hat{\eta}_{m_0}(r|\theta)| < L_n$$

then go to Step 2, else increase m with m_0 (thus $m = m+m_0$) and repeat Step 1.

Step 2: Apply the Wald test (8.4.16), i.e. specify w_{m_*} such that

$$P(\chi^2_{m-m_*} \leq w_{m_*}) = P(|U| \leq L_n) \text{ for } U \sim N(0,1),$$

and select the maximum m_* for which $|\hat{w}_{m_*}| \leq w_{m_*}$. Put $\hat{m} = m_*$ and stop.

Now let $\bar{m}$ be the minimum m for which model (8.2.5) with $\theta = \theta_0$ holds, given that θ_0 is a linear separator. (Otherwise no such $\bar{m}$ exists.) Then

$$\lim_{n\to\infty} P(\hat{m} = \bar{m}) = 1, \text{ whenever } L_n/\sqrt{n} \to 0 \text{ and } L_n \to \infty. \quad (8.4.17)$$

Using result (8.4.17) it is easy to show that theorems 8.4.1 and 8.4.5 go through with $m = \hat{m}$. Also theorem 8.4.4 goes through after an appropriate modification

If θ_0 is not a linear separator, model (8.2.5) is invalid for every m. In that case $\hat{m}$ will increase beyond the support of θ_0'x minus 1. For such an $\hat{m}$, however, the parameters $\alpha_0, \alpha_1, \ldots, \alpha_{m-1}$ in (8.2.5) are no longer indentifiable, which renders $\hat{m}$ indeterminate. Thus the above sequential test procedure for determining the order of the polynomial only works if θ_0 is a linear separator of Ξ. In view of theorem 8.2.1, however, the hazard of θ_0 is a linear separator of Ξ seems not too great.

Summarizing, we have

Theorem 8.4.6. Suppose that θ_0 is a linear separator of Ξ. Estimating the m of model (8.2.5) by repeated application of theorem 8.4.3, according to the procedure given by steps 1 and 2, and replacing m by $\hat{m}$, theorems 8.4.1 and 8.4.5 go through. If in addition we replace the m at righthand side of '$\rightarrow$' in theorem 8.4.4 by $\bar{m}$ then theorem 8.4.4 goes through as well.

Proof: Exercise 3.

The importance of this result lies especially in the conclusion that theorem 8.4.5 carries over. For, consistent estimation of an unknown regression is not of considerable intrinsic interest. What really matters is the possibility to test hypotheses about which variables should be included in the model and which not, without worrying about possible model specification errors.

Moreover we note that the problem of estimating m is similar to the problem of choosing the length of Gallant's (1981, 1982) Fourier series expansion and that the proposed approach regarding an increasing significance level L_n may be considered as a Bayesian approach. See, e.g., Rubin and Sethuraman(1965).

From an asymptotic point of view the choice of L_n is not critical as long as $L_n/\sqrt{n} \rightarrow 0$ and $L_n \rightarrow \infty$. Cf. (8.4.17). In finite samples, however, the estimate $\hat{m}$ may vary substantially with L_n. Obviously the best choice of L_n would be such that

$$P(\hat{m} \neq \bar{m})$$

is minimal for a finite sample size n. The finite sample distribution of $\hat{m}$ is unknown, however, due to the fact that the results of theorem 8.4.2 and 8.4.4 only hold asymptotically and that

$$P(\hat{m} < \bar{m})$$

depends on the extent of the misspecification of g(x). On the other hand, at least we know that

$P(\hat{m} < \bar{m})$ is an increasing function of L_n

and that

$$P(\hat{m} < \bar{m}) \text{ is a decreasing function of } L_n .$$

Since

$$\hat{m} < \bar{m} \text{ is worse than } \hat{m} > \bar{m}$$

we should therefore choose L_n not too large. But how large is too large? In order to answer this question we focus on the outcome of step 1. Let $\hat{\ell}$ be the number of time step 1 is applied and let $\bar{\ell}$ be such that

$$(\bar{\ell}-1)m_0 < \bar{m} \leq \bar{\ell}m_0 .$$

Moreover, assume that step 1 has been conducted for the 'ideal' situation that the asymptotic normality results in theorem 8.4.2 holds for finite n as well. Thus, let

$$\hat{\eta}_m = \hat{\eta}_m(\tau|\theta_*) ,$$

where θ_* is a fixed linear separator of Ξ and τ is a fixed number such that (8.4.14) holds if (8.4.11) is false. Then the 'ideal' version of theorem 8.4.2 is:

$$\hat{\eta}_m \sim N(0,1) \quad \text{if} \quad m \geq \bar{m} ; \quad [\text{cf.}(8.4.12)]$$

$$\hat{\eta}_m \sim N(c_m\sqrt{n},\sigma^2{}_m) \quad \text{if} \quad m < \bar{m}, \text{ where } c_m \neq 0. \quad [\text{cf.}(8.4.14)]$$

Then

$$P(\hat{\ell} < \bar{\ell}) = P[\min_{1\leq j\leq \bar{\ell}-1}|\hat{\eta}_{jm_0}| < L_n]$$

$$\leq \Sigma_{j=1}^{\bar{\ell}-1} \int_{a_n}^{b_n} [\exp(-\tfrac{1}{2}u^2)/\sqrt{(2\pi)}]du$$

$$(\text{with} \quad a_n = (-L_n - c_{jm_0}\sqrt{n})/\sigma_{jm_0}, \; b_n = (L_n - c_{jm_0}\sqrt{n})/\sigma_{jm_0})$$

$$\leq (\bar{\ell}-1)\int_{-\infty}^{L_n/\sigma-\lambda\sqrt{n}} [\exp(-\tfrac{1}{2}u^2)/\sqrt{(2\pi)}]du,$$

where

$$\lambda = \min_{1\le j\le \bar{\ell}-1} |c_{jm_0}|/\sigma_{jm_0}, \quad \sigma = \min_{1\le j\le \bar{\ell}-1} \sigma_{jm_0}.$$

Moreover, if

$$\bar{\ell}m_0 < n \tag{8.4.18}$$

(which is not unreasonable to assume) and

$$L_n/\sqrt{n} < \sigma\lambda/(1+\sigma) \tag{8.4.19}$$

then $\bar{\ell} + 1 < n/m_0 + 1$ and $L_n/\sigma - \lambda\sqrt{n} < -L_n$, hence

$$P(\hat{\ell} < \bar{\ell}) \le ((n/m_0) - 1) \int_{L_n}^{\infty} [\exp(-\tfrac{1}{2}u^2)/\sqrt{(2\pi)}]du.$$

Furthermore, we have

$$P(\hat{\ell} > \bar{\ell}) = P[\min_{1\le j\le \bar{\ell}} |\hat{\eta}_{jm_0}| \ge L_n]$$

$$\le P(|\hat{\eta}_{\bar{\ell}m}| \ge L_n) = 2\int_{L_n}^{\infty} [\exp(-\tfrac{1}{2}u^2)/\sqrt{(2\pi)}]du,$$

hence

$$P(\hat{\ell} \ne \bar{\ell}) \le ((n/m_0) + 1)\int_{L_n}^{\infty} [\exp(-\tfrac{1}{2}u^2)/\sqrt{(2\pi)}]du,$$

provided (8.4.18) and (8.4.19) hold. This result suggests to choose the maximal L_n for which

$$L_n/\sqrt{n} \le \mu, \text{ where } \mu = \sigma\lambda/(1+\sigma).$$

However, since μ is unknown this is not feasible. Therefore we turn to a Bayesian approach. Assume for μ a uniform $[0,\sqrt{n}]$ prior distribution, say. This prior reflexes the fact that we only know that $0 < \mu < \infty$. Then

$$P(\hat{\ell} \ne \bar{\ell}) = P[\hat{\ell} \ne \bar{\ell} | \mu < L_n/\sqrt{n}]P[\mu < L_n/\sqrt{n}]$$

$$+ P[\hat{\ell} \ne \bar{\ell} | \mu \ge L_n/\sqrt{n}]P[\mu \ge L_n/\sqrt{n}]$$

$$\le (L_n/n) + ((n/m_0) + 1)\int_{L_n}^{\infty} [\exp(-\tfrac{1}{2}u^2)/\sqrt{(2\pi)}]du \tag{8.4.20}$$

Minimizing the righthand side of (8.4.20) to L_n yields (cf.

exercise 4)

$$L_n = \{2 \cdot \ln[n(n/m_0+1)/\sqrt{(2\pi)}]\}^{\frac{1}{2}}. \qquad (8.4.21)$$

For example, with $m_0 = 50$ and $n = 2000$ as in Bierens and Hartog (1988) we have

$$L_n = 4.5597; \quad 2\int_{L_n}^{\infty} [\exp(-\tfrac{1}{2}u^2) / \sqrt{(2\pi)}]du = 0.51227E\text{-}5$$

(cf.step 2),

$P(\hat{\ell} \neq \bar{\ell}) \leq 0.0025$ (in the 'ideal' case).

Of course, this critical value L_n heavily depends on the choice of the prior distribution for μ. For example, if we would choose the uniform [0,n] prior then

$$L_n = \{2 \cdot \ln[n^{3/2}(n/m_0+1)/\sqrt{(2\pi)}]\}^{\frac{1}{2}},$$

which, with $m_0 = 50$ and $n = 2000$, yields $L_n = 5.3284$.

It seems possible to extend this approach also to step 2. Step 1, however, is the most important as step 1 aims to result in a correct (but possibly too large) polynomial order.

The problem of selecting the model size addressed to above has been considered in various ways in the literature. See the reviews by Hocking (1976) and Thompson (1978) and in particular Geweke and Meese (1981). Most selection criteria are based on minimizing the error variance subject to a penalty for including an additional variable. Given a true standard normal linear regression model

$$Y_j = \Sigma_{i=1}^{\bar{m}} \beta_j X_{ij} + U_j \quad , \quad j=1,\ldots,n,$$

where $\bar{m}$ is only known to be finite, and a specified model

$$Y_j = \Sigma_{i=1}^{m} \beta_i X_{ij} + U_j$$

with corresponding estimated residual variance $\hat{\sigma}_m^2$, Geweke and Meese propose to select m by minimizing

$$\hat{\sigma}_m^2 + m\varphi(n)$$

subject to

$$m \in \{1,2,\ldots,m_n^*\}, \quad \text{with } m_n^* \to \infty \text{ as } n \to \infty,$$

where $\varphi(n)$ is a nonnegative penalty function. Denoting the resulting estimate of $\bar{m}$ by $\hat{m}$, they show that under some mild conditions on φ

$$P(\hat{m} = \bar{m}) \to 1 \quad \text{as } n \to \infty .$$

This result is the same as (8.4.17). Although the result is the same, it is not clear to us how our approach relates to those of Geweke and Meese. The comparison is hampered by the fact that our model is highly nonlinear and the estimation is carried out by a two-stage procedure. On the other hand, the approach of Geweke and Meese seems to be applicable to our model selection problem, after some modification accounting for the typical structure of our model. Which method is better, however, is an open problem.

Exercises:

1. Prove theorem 8.4.1.

Hint: Prove first that the functions $M_1(\theta)$ and $M_2(\theta)$ are continuous in a neighborhood Θ_0 of θ_0, and so is $z(x,\theta)$ for each $x \in \Xi$. Then use one of the uniform strong laws in chapter 2 to prove

$$\sup_{|z|\le 1} \sup_{\theta\in\Theta_0} |\psi_\ell(z|\theta) - \bar{\psi}_\ell(z|\theta)| \to 0 \text{ a.s.}$$

for $\ell=0,1,2,\ldots$, where similarly to (8.4.4)

$$\begin{aligned} &E\ \bar{\psi}_{r_1}(z(X_j,\theta)|\theta)\bar{\psi}_{r_2}(X_j,\theta)|\theta) = 1 \quad \text{if } r_1=r_2, \\ &E\ \bar{\psi}_{r_1}(z(X_j,\theta)|\theta)\bar{\psi}_{r_2}(X_j,\theta)|\theta) = 0 \quad \text{if } r_1\neq r_2. \end{aligned} \tag{8.4.22}$$

Next, prove that

$$\sup_{\theta\in\Theta_0} |\hat{\gamma}_\ell(\theta) - \bar{\gamma}_\ell(\theta)| \to 0 \text{ a.s.}$$

where

$$\bar{\gamma}_\ell(\theta) = E\ Y_j\bar{\psi}_\ell(z(X_j,\theta)|\theta).$$

2. Prove theorem 8.4.4.

3. Prove theorem 8.4.6, using (8.4.17).
4. Prove (8.4.21).

8.5 *Proofs*

8.5.1 *Proof of theorem 8.4.2*

Assume that (8.4.11) holds. Then

$$E(Y_j|X_j) = g(X_j) = \Sigma_{\ell=0}^{m-1}\gamma_\ell(\theta_0)\psi_\ell(z(X_j,\theta_0)|\theta_0) \text{ a.s.} \quad (8.5.1)$$

so that

$$\begin{aligned}
&(1/\sqrt{n})\Sigma_{j=1}^{n}(Y_j - \hat{g}_m(X_j|\hat{\theta}))\exp(\tau z(X_j,\theta^*)) \\
&= (1/\sqrt{n})\Sigma_{j=1}^{n}(U_j + g(X_j) - \hat{g}_m(X_j|\theta_0) + \hat{g}_m(X_j|\theta_0) - \hat{g}_m(X_j|\hat{\theta})) \\
&\qquad \times \exp(\tau z(X_j,\theta^*)) \\
&\quad - (1/\sqrt{n})\Sigma_{j=1}^{n}U_j\exp(\tau z(X_j,\theta^*)) \\
&\quad - (1/\sqrt{n})\Sigma_{j=1}^{n}\Sigma_{\ell=0}^{m-1}(\hat{\gamma}_\ell(\theta_0) - \gamma_\ell(\theta_0))\psi_\ell(z(X_j,\theta_0)|\theta_0) \\
&\qquad \times \exp(\tau z(X_j,\theta^*)) \\
&\quad - (1/\sqrt{n})\Sigma_{j=1}^{n}(\hat{g}_m(X_j|\hat{\theta}) - \hat{g}_m(X_j|\theta_0))\exp(\tau z(X_j,\theta^*)) \\
&= \tilde{c}_1(\tau,\theta^*) - \tilde{c}_2(\tau,\theta^*) - \tilde{c}_3(\tau,\theta^*), \text{ say.} \quad (8.5.2)
\end{aligned}$$

Observe from (8.4.5), (8.4.6) and (8.5.1) that for $\ell = 0,1,2,\ldots$

$$\hat{\gamma}_\ell(\theta_0) - \gamma_\ell(\theta_0) = (1/n)\Sigma_{j=1}^{n}U_j\psi_\ell(z(X_j,\theta_0)|\theta_0),$$

hence

$$\begin{aligned}
\tilde{c}_2(\tau,\theta^*) = \Sigma_{\ell=0}^{m-1}\{&((1/\sqrt{n})\Sigma_{j=1}^{n}U_j\psi_\ell(z(X_j,\theta_0)|\theta_0) \\
&\times \{(1/n)\Sigma_{j=1}^{n}\psi_\ell(z(X_j,\theta_0)|\theta_0)\exp(\tau z(X_j,\theta^*))\}.
\end{aligned}$$

Denoting

$$\tilde{\tilde{c}}_2(\tau,\theta^*) =$$

$$= \Sigma_{\ell=0}^{m-1}\{(1/\sqrt{n})\Sigma_{j=1}^{n} U_j \psi_\ell(z(X_j,\theta_0)|\theta_0)E[\bar{\psi}_\ell(z(X_j,\theta_0)|\theta_0) \times \exp(\tau z(X_j,\theta^*))]\},$$

where $\bar{\psi}_\ell(z|\theta_0)$ is the probability limit of $\psi_\ell(z|\theta_0)$, it is easy to verify that

$$\text{plim}_{n\to\infty}\{\tilde{\tilde{c}}_2(\tau,\theta^*) - \tilde{c}_2(\tau,\theta^*)\} = 0. \tag{8.5.3}$$

Next, observe that by the mean value theorem there exists a random vector $\bar{\theta}(\tau,\theta^*)$ satisfying

$$|\bar{\theta}(\tau,\theta^*) - \theta_0| \le |\hat{\theta} - \theta_0| \text{ a.s.} \tag{8.5.4}$$

such that

$$\tilde{c}_3(\tau,\theta^*) = (1/\sqrt{n})\Sigma_{j=1}^{n}(\hat{g}_m(X_j|\hat{\theta}) - \hat{g}_m(X_j|\theta_0)) \times \exp(\tau z(X_j,\theta^*))$$

$$- \sqrt{n}(\hat{\theta}-\theta_0)'(1/n)\Sigma_{j=1}^{n}(\partial/\partial\theta')\hat{g}_m(X_j|\bar{\theta}(\tau,\theta^*)) \times \exp(\tau z(X_j,\theta^*)).$$

Moreover, it follows from (8.5.4) and the fact that for fixed x,

$(\partial/\partial\theta')\hat{g}_m(x|\theta)$ converges in prob., uniformly on a neighborhood of θ_0,

that

$$\text{plim}_{n\to\infty}(1/n)\Sigma_{j=1}^{n}(\partial/\partial\theta')\hat{g}_m(X_j|\bar{\theta}(\tau,\theta^*))\exp(\tau z(X_j,\theta^*))$$

$$= \text{plim}_{n\to\infty}(1/n)\Sigma_{j=1}^{n}(\partial/\partial\theta')\hat{g}_m(X_j|\hat{\theta})\exp(\tau z(X_j,\theta^*))$$

$$= \text{plim}_{n\to\infty}(1/n)\Sigma_{j=1}^{n}(\partial/\partial\theta')\hat{g}_m(X_j|\theta_0)\exp(\tau z(X_j,\theta^*))$$

$$= \bar{\xi}_m(\tau|\theta^*), \text{ say.}$$

Denoting

$$\tilde{\tilde{c}}_3(\tau,\theta^*) = \sqrt{n}(\hat{\theta}-\theta_0)'\bar{\xi}_m(\tau|\theta^*)$$

we thus have

$$\text{plim}_{n\to\infty}\{\tilde{c}_3(\tau,\theta_0) - \tilde{\tilde{c}}_3(\tau,\theta^*)\} = 0. \qquad (8.5.5)$$

Furthermore, observe that

$$\sqrt{n}(\hat{\theta}-\theta_0) = [(1/n)\Sigma_{j=1}^n X_j X_j']^{-1}(1/\sqrt{n})\Sigma_{j=1}^n X_j(Y_j - X_j'\theta_0)$$

and that

$$\text{plim}_{n\to\infty}(\sqrt{n}(\hat{\theta}-\theta_0) - (E\ X_1X_1')^{-1}(1/\sqrt{n})\Sigma_{j=1}^n X_j(Y_j - X_j'\theta_0)) = 0.$$

Thus denoting

$$\tilde{\tilde{\tilde{c}}}_3(\tau,\theta^*) = \bar{\xi}_m(\tau,\theta^*)'(E\ X_1X_1')^{-1}(1/\sqrt{n})\Sigma_{j=1}^n X_j(Y_j - X_j'\theta_0)$$

we have

$$\text{plim}_{n\to\infty}\{\tilde{\tilde{c}}_3(\tau,\theta_0) - \tilde{\tilde{\tilde{c}}}_3(\tau,\theta^*)\} = 0. \qquad (8.5.7)$$

From (8.5.2), (8.5.3), (8.5.5) en (8.5.7) we now obtain

$$\text{plim}_{n\to\infty}\{(1/\sqrt{n})\Sigma_{j=1}^n (Y_j - \hat{g}_m(X_j|\hat{\theta}))\exp(\tau z(X_j,\theta^*)) - \hat{d}(\tau|\theta^*)\} = 0, \qquad (8.5.8)$$

where

$$\hat{d}(\tau|\theta^*) = \tilde{c}_1(\tau,\theta^*) - \tilde{\tilde{c}}_2(\tau,\theta^*) - \tilde{\tilde{\tilde{c}}}_3(\tau,\theta^*)$$

$$= (1/\sqrt{n})\Sigma_{j=1}^n U_j \exp(\tau z(X_j,\theta^*))$$

$$- (1/\sqrt{n})\Sigma_{j=1}^n U_j \Sigma_{\ell=0}^{m-1}\psi_\ell(z(X_j,\theta_0)|\theta_0)E(\bar{\psi}_\ell(z(X_j,\theta_0)|\theta_0)$$

$$\times \exp(\tau z(X_j,\theta^*))$$

$$- \bar{\xi}_m(\tau|\theta^*)'(E\ X_1X_1')^{-1}(1/\sqrt{n})\Sigma_{j=1}^n x_j(Y_j - X_j'\theta_0)$$

$$= (1/\sqrt{n})\Sigma_{j=1}^{n} U_j \rho_{j,m}(\tau|\theta^*) - \bar{\xi}_m(\tau|\theta^*)'(E\ X_1X_1')$$

$$\times (1/\sqrt{n})\Sigma_{j=1}^{n} X_j(Y_j - X_j'\theta) \qquad (8.5.9)$$

with [cf. (8.4.22)]

$$\rho_{j,m}(\tau|\theta^*) = \exp(\tau z(X_j,\theta^*)) - \Sigma_{\ell=0}^{m-1}\psi_\ell(z(X_j,\theta_0)|\theta_0)$$

$$\times E[\bar{\psi}_\ell(z(X_j,\theta_0)|\theta_0)\exp(\tau z(X_j,\theta^*))].$$

Realizing that the terms in (8.5.9) are i.i.d. with zero mean and variance

$$\bar{s}_m^2(\tau|\theta^*) = E\{U_1\rho_{1,m}(\tau|\theta^*) - \bar{\xi}_m(\tau|\theta^*)'(E\ X_1X_1')^{-1}X_j(Y_j - X_j'\theta_0)\}^2$$

$$= E\{U_j^2\rho_{1,m}^2(\tau|\theta^*)\} - 2\ E\{U_j\rho_{j,m}(\tau|\theta^*)\bar{\xi}_m(\tau|\theta^*)'$$

$$\times (E\ X_jX_j')^{-1}X_j(Y_j - X_j'\theta_0)$$

$$+ \bar{\xi}_m(\tau|\theta^*)'(E\ X_1X_1')^{-1}(E(Y_1 - \theta_0'X_1)^2X_1X_1')(E\ X_1X_1')^{-1}\bar{\xi}_m(\tau|\theta^*)$$

$$= E\ U_1^2\rho_{1,m}^2(\tau|\theta^*)$$

$$- 2\ E\ U_1(Y_1 - X_1'\theta_0)\rho_{1,m}(\tau|\theta^*)X_1'(E\ X_1X_1')^{-1}\bar{\xi}_m(\tau|\theta^*)$$

$$+ \bar{\xi}_m(\tau|\theta^*)'\Omega\bar{\xi}_m(\tau|\theta^*),$$

where Ω is defined in theorem 8.3.1, we have by the central limit theorem

$$\hat{d}_m(\tau|\theta^*) \to N(0, \bar{s}_m^2(\tau|\theta^*)) \quad \text{in distr.} \qquad (8.5.11)$$

We leave it to the reader to verify that

$$\hat{s}_m^2(\tau|\theta^*) \to \bar{s}_m^2(\tau|\theta^*) \quad \text{a.s.}$$

$$\bar{s}_m^2(\tau|\theta^*) > 0 \text{ for } \tau \neq 0. \qquad (8.5.12)$$

Combining (8.5.8), (8.5.11) and (8.5.12), part (8.4.12) of theorem 8.4.2 follows.

Next, assume that (8.4.11) fails to hold. Then

$$(1/n)\Sigma_{j=1}^{n}(Y_j - \hat{g}_m(X_j|\hat{\theta}))\exp(\tau z(X_j,\theta^*))$$

$$\to E(Y_j - \Sigma_{\ell=0}^{m-1}\gamma_\ell(\theta_0)\psi_\ell(z(X_j,\theta_0)|\theta_0))\exp(\tau z(X_j,\theta^*)) \text{ a.s.} \qquad (8.5.13)$$

Moreover, it is not hard to verify that also now

$$\hat{s}_m^2(\tau|\theta^*) \to s_*^2(\tau|\theta^*) \text{ a.s.},$$

say, where the limit is positive for $\tau \neq 0$. Thus part (8.4.13) follows straightforwardly from (8.5.13) and lemma 3.3.1.

Finally, the conclusion that we may substitute $\hat{\theta}$ for θ^* follows from the fact that by theorem 8.3.1

$$\operatorname{plim}_{n\to\infty}((1/\sqrt{n})\Sigma_{j=1}^{n}(Y_j - \hat{g}_m(X_j,\hat{\theta}))\exp(\tau z(X_j,\hat{\theta}))$$

$$- (1/\sqrt{n})\Sigma_{j=1}^{n}(Y_j - \hat{g}_m(X_j,\hat{\theta}))\exp(\tau z(X_j,\theta_0))) = 0 \qquad (8.5.14)$$

provided (8.4.11) is satisfied. Proving (8.5.14) is not too hard and therefore left to the reader. Q.E.D.

8.5.2 *Proof of theorem 8.4.3.*

The result (8.4.12) is equivalent with

$$\lim_{n\to\infty} E\exp(i\cdot t\hat{\eta}_m(\tau|\theta^*)) = \exp(-\tfrac{1}{2}t^2) \text{ for every } t\in R$$

If τ and θ^* are random and independent from the data-generating process then similarly we have

$$E(\exp(i\cdot t\hat{\eta}_m(\tau|\theta^*))|\tau,\theta^*) \to \exp(-\tfrac{1}{2}t^2) \text{ a.s.}$$

Hence by bounded convergence,

$$E\exp(i\cdot t\hat{\eta}_m(\tau|\theta^*)) = E[E\exp(i\cdot t\hat{\eta}_m(\tau|\theta^*))|\tau,\theta^*] \to \exp(-\tfrac{1}{2}t^2)$$

which proves that (8.4.12) carries over if θ^* and τ are random. Finally, suppose that (8.4.11) fails to hold. Lemma 3.3.1 implies that (8.4.13) hold for $\tau \in R\backslash T$, where T is a countable subset of R. But since τ is now continuously distributed we have

$P(\tau \in R\backslash T) = 1.$

Moreover, theorem 8.2.1 implies that θ^* is a.s. a linear separator. Therefore (8.4.13) also holds for the random τ and θ^* involved. Q.E.D.

References

Bierens,H.J. and J.Hartog (1988), "Non-Linear Regression with Discrete Explanatory Variables, with an Application to the Earnings Function", *Journal of Econometrics* 38, 269-299.

Forsythe,G.F. (1957), "Generation and Use of Orthogonal Polynomials for Data-Fitting with Digital Computer, *J. Soc. Indust. Appl. Math.* 5, 74-87.

Friedman,J.H. and W.Stuetzle (1981), "Projection Pursuit Regression", *Journal of the American Statistical Association* 76, 817-823.

Gallant,A.R. (1981), "On the Bias in Flexible Functional Forms and an Essentially Unbiased Form: The Fourier Flexible Form", *Journal of Econometrics* 15, 211-245.

Gallant,A.R. (1982), "Unbiased Determination of Production Technologies", *Journal of Econometrics* 20, 285-323.

Gallant,A.R. (1985), "Identification and Consistency in Semi-nonparametric regression", Invited paper presented to the World Congress of the Econometric Society, Cambridge, MA.

Geweke,J. and R.Meese (1981), "Estimating Regression Models of Finite But Unknown Order", *International Econometric Review* 22, 55-70.

Hocking,R.R. (1976), "The Analysis and Selection of Variables in Linear Regression", *Biometrics* 32, 1-49.

Huber,P. (1985) "Projection Pursuit" (with discussion), *Annals of Statistics* 13, 435-475.

Royden,H.L. (1968), "Real Analysis", London: MacMillan.

Rubin,H. and J.Sethuraman (1965), "Bayes Risk Efficiency, *Sankhya A* 27, 347-356.

Seber,G.A.F. (1977), "Linear Regression Analysis", New York: Wiley.

Thompson,M. (1978), "Selection of Variables in Multiple Regression: Parts I and II", *International Statistical Review* 46, 1-19 and 129-146.

White,H. (1980), "Using Least Squares to Approximate Unknown Regression Functions", *International Economic Review* 21, 149-170.