3. Spline Regression

3.1 Introduction

One simple way of taking the non-linearity of a regression equation into account is to split the sample into a number of segments and then estimate a linear relationship for each of them.

Spline regression is based on this principle, and can be seen as an extension of the linear model.

Presented as such, this approach corresponds to piecewise linear regression, which can be estimated by Ordinary Least Squares.

Piecewise regression takes non-linearity into account but does not guarantee the differentiability of the regression function at all points.

The particularity of spline regression is to estimate polynomials rather than straight lines in each segment, chosen such that the estimated function is continuous and differentiable at the junctions of the polynomials.

These junction points are called knots.

The splines are defined such that the estimated function is smooth everywhere.

The results of spline estimation are sensitive to the choice of the number of knots and their position.

To minimize the influence of these-choices on the results, we introduce a penalizing parameter: this is the case of penalized splines, which are the subject of this chapter.

Our first application is the estimation of a Phillips curve.

The underlying idea of such a curve, which was widespread in the 1960s, was that it was possible to reduce the unemployment rate if we were prepared to accept higher inflation.

To check whether this relationship holds empirically, we consider the following non-parametric model: $y = m(x) + \varepsilon, (3.1)$, where y is the inflation rate and x is the unemployment rate.

The monthly American data run from January 1960 to March 1970, and are available on the website of the Federal Reserve Bank of St Louis.

The inflation rate here corresponds to the variable CPIAUCSL (Consumer price Index For All Urban Consumers : Al] Items ; percentage change fror7tone year ago) and unemployment corresponds to the variable UNRATE (Civilian Unemployment Rate).

Figure 3.1 presents the regression results from penalized spline estimation of the relationship between unemployment and inflation, as well as the 95% confidence intervals.

The estimation results are consistent with the results in Phillips (1958).

First, we note that there is a negative relationship between unemployment and inflation.

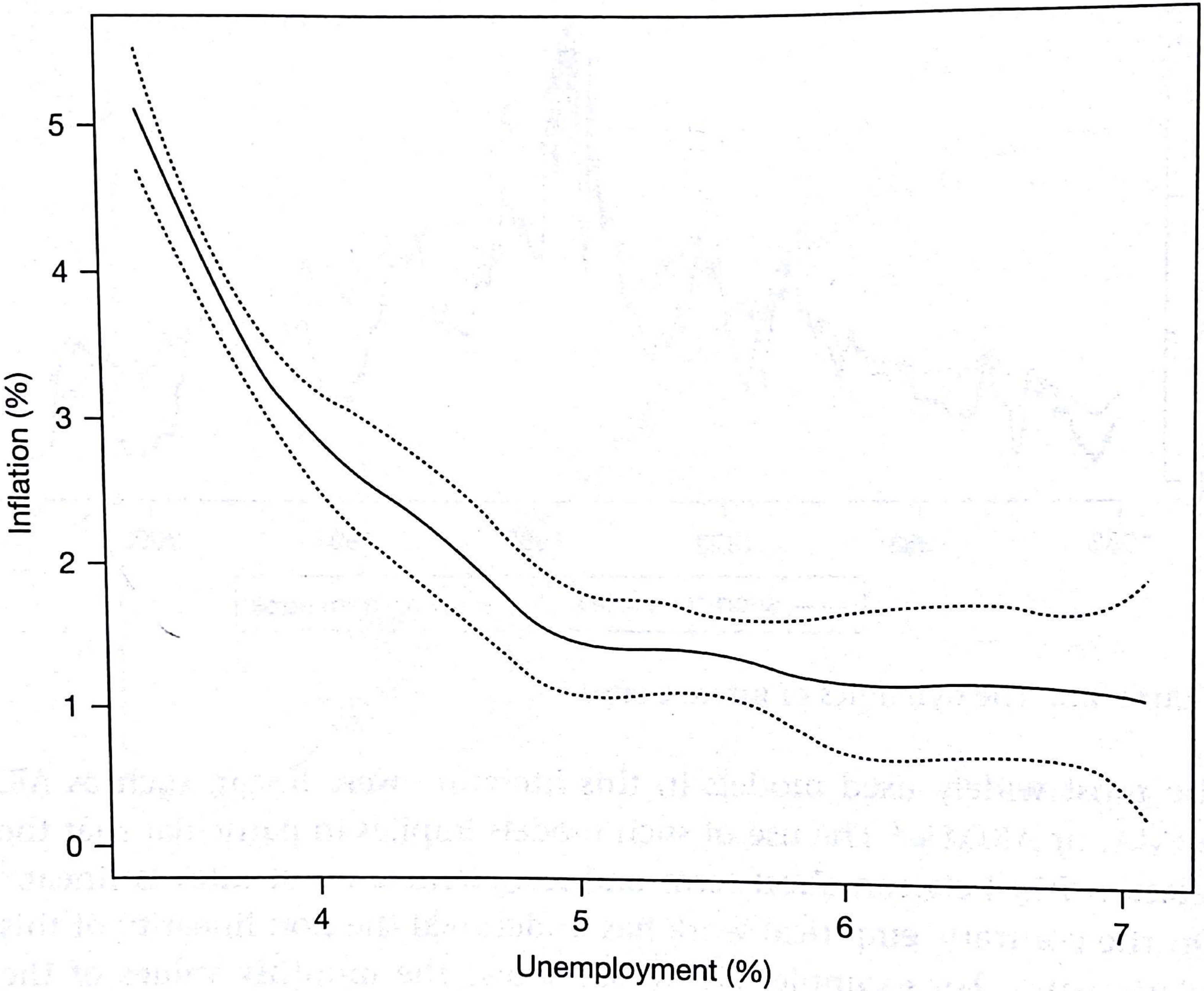


Figure 3.1: The Phillips curve: the unemployment–inflation trade-off

Moreover, this relationship is not linear, being much flatter when the unemployment rate is greater than 5% and much stronger when unemployment is under 4%.

There does thus seem to be a trade-off between unemployment and inflation, which relationship was then some what challenged in the 1970s by the joint appearance of higher rates of both unemployment and inflation.

The second application concerns the relationship between short-term and long-term interest rates.

The dynamics of interest rates of different terms has attracted a great deal of attention in finance.

For a long time, the most widely used models in this literature were linear, such as AR, ARMA, or ARIMA.

The use of such models implies in particular that the relationship between short-term and long-term interest rates is linear.

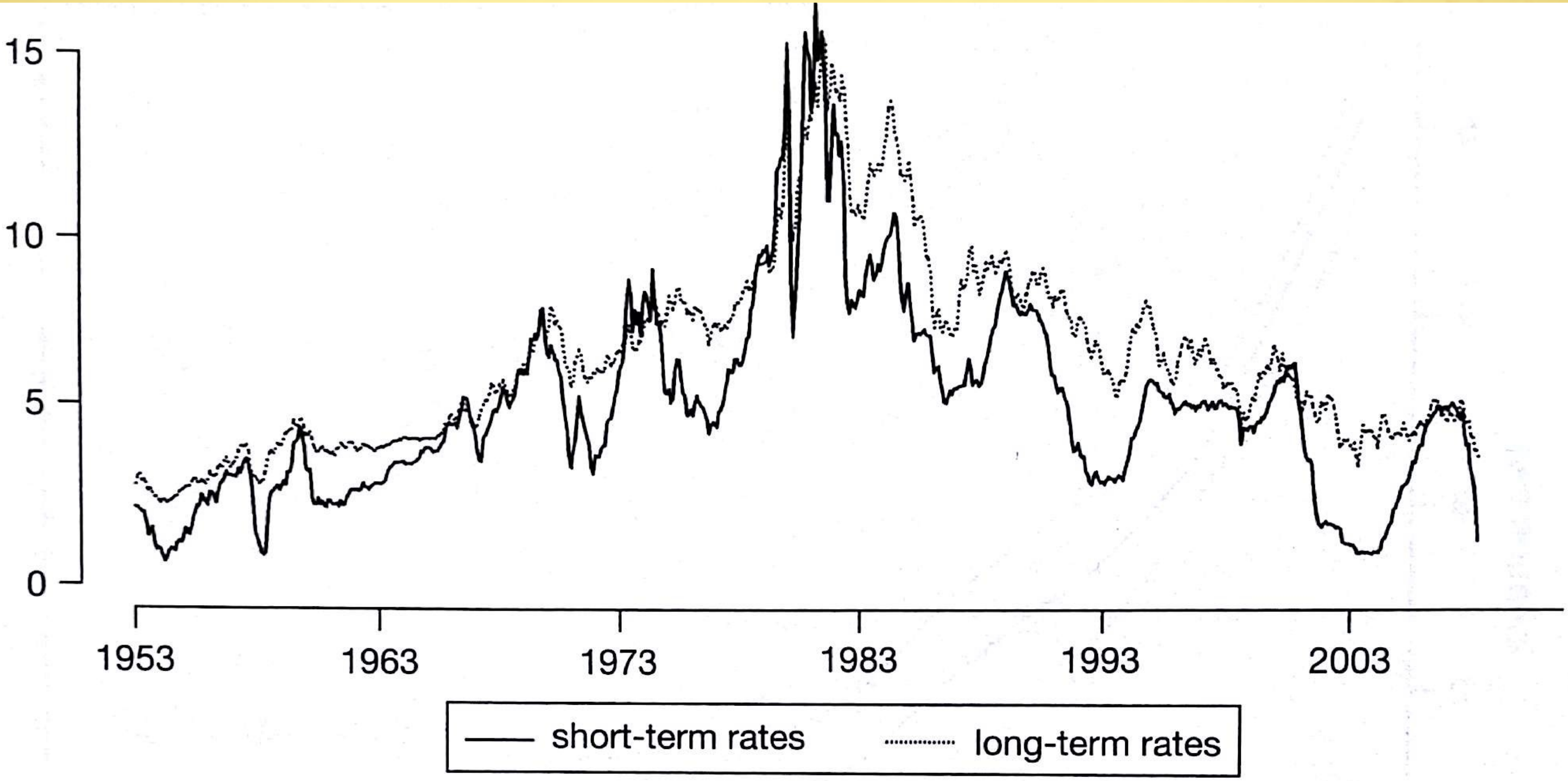


Figure 3.2: The dynamics of interest rates

On the contrary, empirical work has underlined the non-linearity of this relationship.

For example, Figure 3.2 shows the monthly values of the short-term interest rate (3 months) and the long-term interest rate (10 years) in the USA, from April 1953 to March 2008.

This figure suggests that long-term rates are less responsive to short-term rates when these latter are higher.

As noted by Pfann et al. (1996), the relation between the two rates is different when the short-term rate is higher and more volatile, that is, between 1979 and 1982.

To represent this non-linearity, we estimate (3.1), with y being the long-term interest rate-corresponding to the returns on Government Bonds and x being the short-term interest rate-reflecting the monetary policy of the Central Bank.

Figure 3.3 shows the results of this penalized spline estimation and the associated 95% confidence intervals.

The estimated relationship is non-linear, with the slope of the function being flatter for higher values of the short-term interest rate.

This is similar to the results in Pfann et al. (1996), which led them to use the SETAR (self-exciting Threshold Autoregressive) models in order to better capture the non-linear dynamics of the structure of interest rates.

In this chapter, we will present the non-parametric estimation of a simple regression model by the method of penalized splines, illustrated by the preceding examples.

The first section will discuss splines as an extension of the linear model, and we will then introduce the concept of penalized splines.

This will allow us to demonstrate that this is a projective estimation technique, which requires the use of a base.

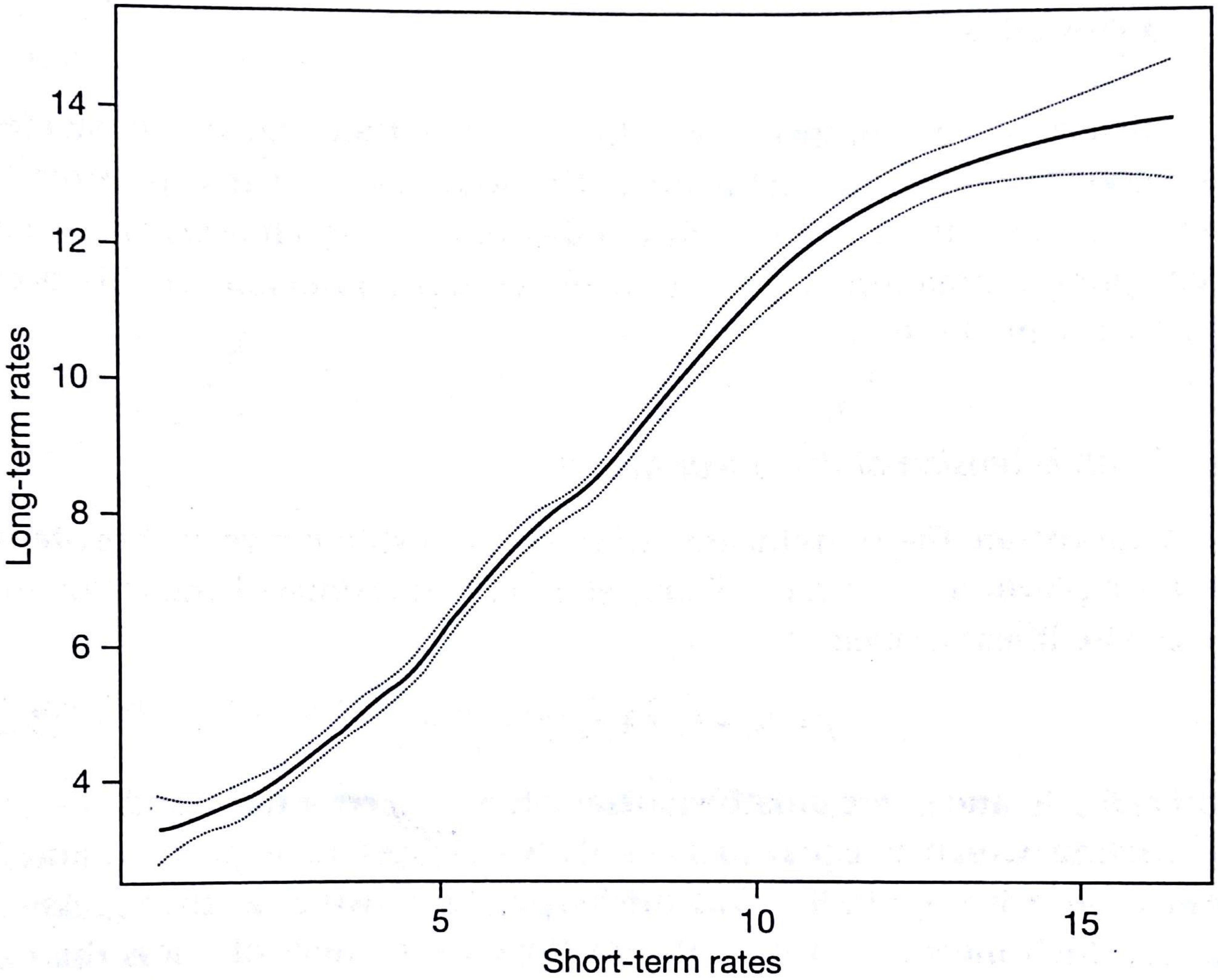


Figure 3.3: The relationship between short-term and long-term interest rates

The following section then discusses the choice of an appropriate base.

The last section presents the choice of parameters: the number and position of the knots, and the penalization parameter.

Further reading on this topic can be found in, amongst others, the work of Wahba (1990), Green and Silverman (1993), Eubank (1999), Lader (1999b), de Boor (2001), Marsh and Cormier (2001), Gu (2002), and Ruppert et al. (2003).

3.2 Principles

The first appealing property of splines is that they can be considered as piecewise polynomial estimators.

The second is that it is possible to minimize the impact of a certain number of arbitrary choices, via penalized spline estimation.

This section presents the principle of this non-parametric method.

3.2.1 An Extension of the Linear Model

To demonstrate the non-linearity of the relationship between short-term and long-term interest rates, Pfann et al. (1996) estimated the following piecewise linear model:

$$y = \beta_0 + \beta_1 x + \beta_2 (x - \kappa)_+ + \varepsilon, (3.2)$$

where β_0, β_1, and β_2 are unknown parameters, ε an error term, and $(x-\kappa)_+$ a function which is equal to $x-\kappa$ if this expression is positive and 0 otherwise, with $\kappa=10.8$.

This model produces two straight regression lines, which meet at κ, with different slopes.

For values of x less than κ, the regression line has a slope of β_1; for values of x greater than κ, this slope is $\beta_1 + \beta_2$.

Panel (a) of Figure 3.4 depicts the OLS estimation results of this model, which can be compared to those previously presented in Figure 3.3.

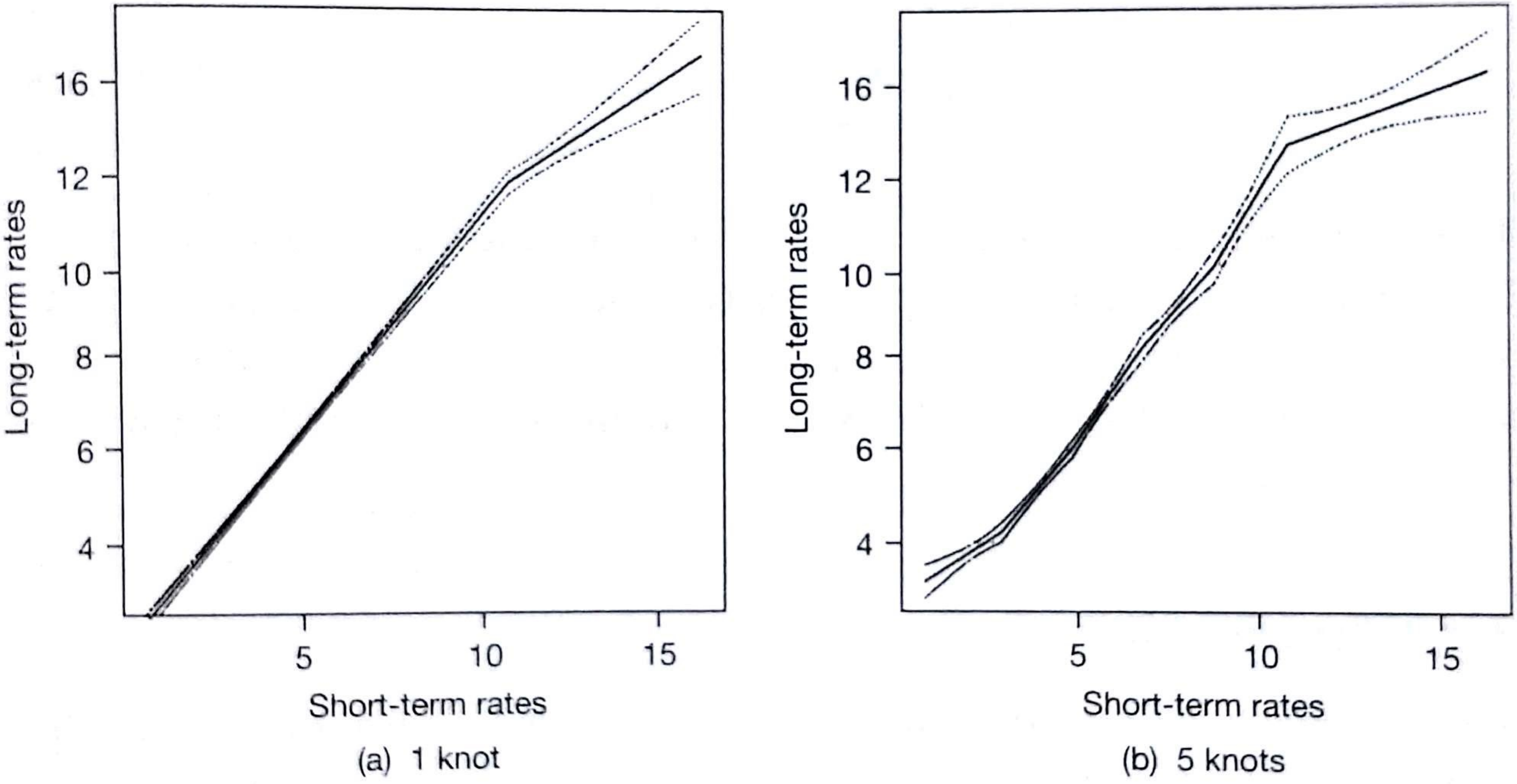


Figure 3.4: Piecewise linear models

The figure shows an estimated piecewise linear relationship.

The estimated β_2 coefficient is statistically significant ($\hat{\beta}_2$=-0.42910, standard error = 0.08958), implying that the slope of the second linear segment of the regression is significantly different from the slope of the first.

This reflects the non-linearity of the relationship between the two variables.

From a geometric point of view, OLS estimation of a linear regression model is equivalent to an orthogonal projection of the dependent variable y in the space of the regressors.

In other words, the regressors in the model $y = \beta_0 + \beta_1 x + \beta_2 (x - \kappa)_+ + \varepsilon, (3.2)$ can be seen as representing a base, defined by the following functions: ι, x, $(x\text{-}\kappa)_+$, where, ι is the unit vector.

Estimation is a projection on this base of functions, and the model parameters are the coordinates of y in this base.

We thus refer to projective estimation methods over a pre-defined base.

In their article, Pfann et al. (1996) use the same data as we used in the introduction to this chapter, but from January 1962 to June 1990.

The above regression model can be extended to include more than two segments, by choosing the number of junction points or knots before- hand.

For the prior choice of q knots, the base is defined by the following functions:

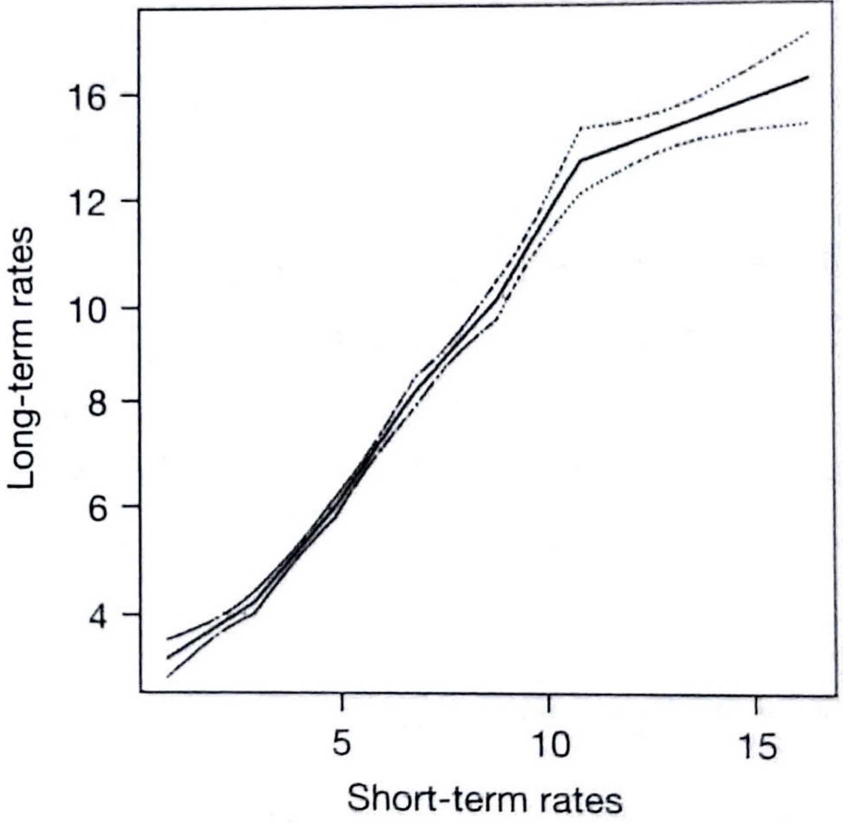
$$\iota, x, (x-\kappa_1)_+, (x-\kappa_2)_+, \ldots, (x-\kappa_q)_+, \quad (3.3)$$

The projection of y over this base of functions is equivalent to estimating the following piecewise linear model:

$$y=\beta_0+\beta_1 x+\sum_{j=1}^{q} \beta_{1j}\left(x-\kappa_j\right)+\varepsilon, \qquad (3.4)$$

Panel (b) of Figure 3.4 depicts the estimation results of this model for q=5 pre-defined knots : κ_1=2.8, κ_2=4.8, κ_3=6.8, κ_4=8.8, and κ_5=10.8.

The resulting relationship is non-linear, and bears a close resemblance to that in Figure 3. 3.



(b) 5 knots

This example illustrates the potential of piecewise linear estimation to recreate a distinctly non-linear function, by the simple addition of extra terms such as $(x-\kappa_j)_+$ to a linear regression.

The estimated functions in Figure 3.4 do not have continuous first derivatives: they jump discretely at the knots.

To avoid this, we appeal to functions that are different from $(x-\kappa_j)_+$ to define the base.

We then refer to spline functions, which can be linear, quadratic, cubic, or of some other form.

Let $b_0(x, \kappa),\ldots, b_r(x, \kappa)$ be the spline functions which define the base, where κ is a vector of q pre-defined knots, then the spline regression model can be written as follows:

$$y = \sum_{j=1}^{r} \beta_j b_j(x, \kappa) + \varepsilon \quad (3.5)$$

The OLS estimation of this model produces estimated parameters $\hat{\beta}_j$, for $j=0,\ldots,r$, which corresponds to the coordinates of y in the chosen spline base.

The appropriate choice of this base produces an estimator with continuous first derivatives, as in Figure 3.3.

The following section discusses the choice of the base.

3.2.2 Penalized Splines

The approach described above requires the prior choice of the number of knots and their position.

The estimation results are sensitive to this choice: the two panels of Figure 3.4 show that the estimated regression is notably different according to the choice of two or five knots.

Automatic selection methods for the number and position of these knots have been developed, but these are relatively complex and demanding in terms of computer time.

The problem is that, for a given number of knots, there is a very large number of possible regression models.

These different methods are compared in Wand (2000).

To minimize the influence of knot number and position, a different approach is to constrain the estimation of the parameters in $y = \sum_{j=1}^{r} \beta_j b_j(x, \kappa) + \varepsilon$, (3.5).

The estimation by penalized splines of the regression $y = m(x) + \varepsilon$, (3.1) consists in decomposing the function m(x) in a spline base function; that is, to rewrite the model as in (3.5), and then to choose parameter values (β_j, for j = 0,…, r) which minimize $\sum_{i=1}^{n}[y_i - m(x_i)]^2$ subject to $\lambda \int [m''(x)]^2 dx \leq C$ where C is positive.

Or minimize the following criterion:

$$\sum_{i=1}^{n}[y_i - m(x_i)]^2 + \lambda \int [m''(x)]^2 dx \, , (3.6)$$

where n is the number of observations and λ>0. This criterion corresponds to a minimization with a Lagrange multiplier.

This constraint depends on the second derivative of the function, which reveals the variation in the slope of the function.

The term $\lambda \int [m''(x)]^2 dx$ is thus a measure of the speed at which the function m(x) varies.

A function which varies only little will yield a small value for this term, while a function which varies greatly will yield a larger value.

Estimation by penalized splines corresponds to an OLS estimation where we constrain the speed at which the function varies.

A sound choice of C, or equivalently of λ, produces an estimation of the function which does not vary over much.

Spline estimation of parameters is a distinct modification of OLS estimation.

Given the parameter vector $\beta = [\beta_0, \beta_1, \dots, \beta_r]$ and the spline function base $X = [b_0(x, \kappa), b_1(x, \kappa), \dots, b_r(x, \kappa)]$, the model $y = \sum_{j=1}^{r} \beta_j b_j(x, \kappa) + \varepsilon$ (3.5) can be written as $y = X\beta + \varepsilon$.

Furthermore, we can express the penalization term as a function of the parameters and of the base : $\lambda \int [m''(x)]^2 dx = \beta^T S \beta$, where S is a matrix that depends on the components of the base X.

The estimation of the model (3.5) via the minimization of the criterion (3.6: $\sum_{i=1}^{n} [y_i - m(x_i)]^2 + \lambda \int [m''(x)]^2 dx$) is therefore the same thing as the minimization of:

$$\sum_{i=1}^{n} (y_i - X_i \beta)^2 + \lambda \beta^T S \beta$$

which is solved by: $\hat{\beta} = (X^T X + \lambda S)^{-1} X^T y$. The parameter λ in this expression is fixed.

This plays the role of a smoothing parameter, and allows us reach a compromise between the distance to the data (the first term in 3.6: $\sum_{i=1}^{n}[y_i - m(x_i)]^2 \ +\lambda \int [m''(x)]^2 dx$) and the overall evenness of the function (the second term).

In the last section, we will present various methods of choosing this parameter, according to a number of optimality criteria.

3.3 The Choice of the Spline Base

The first stage in spline estimation of the regression model $y = \sum_{j=1}^{r} \beta_j b_j(x, \kappa) + \varepsilon$, (3.5)consists of the choice of a base of functions.

The choice of this base is guided by the properties that we would like it to satisfy.

In general, we would like the variations in the m(x) function not to be too abrupt at the junction points, as they are in the case of piecewise linear estimation (Figure 3.4).

As such, we require the first derivative of the function to be continuous.

A number of different bases satisfy this property.

In this section, we present power function bases, and then B-splines.

3.3.1 Power Function Bases

One particularly simple way of moving beyond piecewise linear regression consists in adding x^2 terms to the base (3.3) (ι, x, $(x-\kappa_1)_+$, $(x-\kappa_2)_+$, …, $(x-\kappa_q)_+$,) and replacing the $(x-\kappa_j)_+$ terms by their squares $(x-\kappa_j)^2_+$.

The $(x-\kappa_j)^2_+$ functions have a continuous first derivative.

It therefore follows that any linear combination of the functions:

$\iota, x, x^2, (x-\kappa_1)^2_+, (x-\kappa_2)^2_+, \ldots, (x-\kappa_q)^2_+$

will also have a continuous first derivative.

As a result, the use of this quadratic spline base allows us to estimate (3.5) $y = \sum_{j=1}^{r} \beta_j b_j(x, \kappa) + \varepsilon$ with smooth transitions at the knots $\kappa_1, \ldots, \kappa_q$.

Another advantage of the utilization of this base is that the approximation within each segment is carried out using a polynomial of degree 2, which allows us to pick up any concavity or convexity in the function between two knots.

The above base can be generalized by using truncated power functions of order p. A spline base of degree p is written as:

$$\iota, x, \ldots, x^p, (\mathrm{x}-\kappa_1)_+^p, (\mathrm{x}-\kappa_2)_+^p, \ldots, (\mathrm{x}-\kappa_\mathrm{q})_+^p, (3.7)$$

The choice of this base to estimate the model (3.1) leads us to write the spline regression model of order p as follows:

$$y = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{j=1}^{q} \beta_{pj} (\mathrm{x}-\kappa_j)_+^p + \epsilon, (3.8)$$

The function $(\mathrm{x}-\kappa_j)_+^p$ has a continuous derivative of order p-1.

As such, the choice of this spline base of degree p ensures that the derivative of order p-1 of the m(x) function in (3.1) is continuous.

In practical terms, cubic spline functions are often used (p=3).

These guarantee the continuity of the first derivative $m'(x)$ and the second derivative $m''(x)$ and permit change of concavity of m(x) between two knots.

As an example, suppose that the estimation of the model (3.1) with the base of cubic spline functions (p=3) produces the following results:

$$\widehat{m}(x) = 2 + x - 2x^2 + x^3 + (\mathrm{x}-0.4)_+^3 - (\mathrm{x}-0.8)_+^3$$

This shows that the estimated function $\widehat{m}(x)$ has coordinates of (2, 1,-2, 1, 1,-1) in the cubic spline base, with knots of 0.4 and 0.8.

The estimated function can be rewritten as :

$$\widehat{m}(x) = \begin{cases} 2 + x - 2x^2 + x^3, if\ x < 0.4 \\ 2 + x - 2x^2 + x^3 + (\mathrm{x}-0.4)_+^3, if\ 0.4 \leq x < 0.8 \\ 2 + x - 2x^2 + x^3 + (\mathrm{x}-0.4)_+^3 - (\mathrm{x}-0.8)_+^3, if\ x \geq 0.8 \end{cases}$$

It is easy to check that the first and second derivatives of this function are continuous at the knots.

This way of writing the results underlines that the function is an approximation via three third-order polynomials over adjacent segments, with the passage from one segment to another being smooth; that is, practically imperceptible at the junction points, as shown in Figure 3.5.

3.3.2 B-spline Bases

A second commonly used solution is that of B-spline bases.

Contrary to those described above, these bases have the advantage of being defined over a compact support.

In other words, the functions that define this base are strictly local: they take on values of zero outside of a number of adjacent knots.

This property is useful inasmuch as it allows local solutions to the estimation and simplifies the numerical calculations.

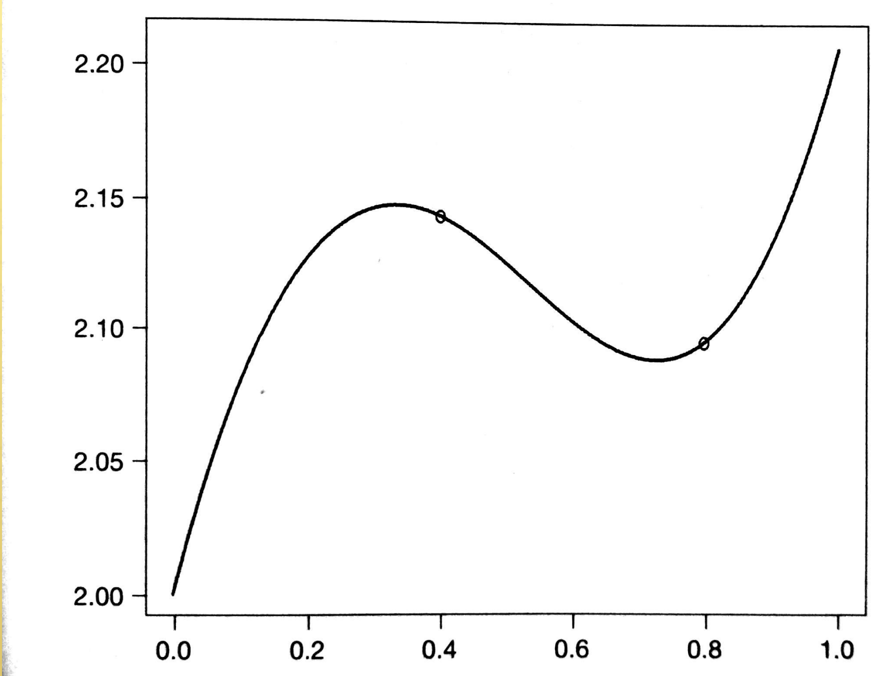


Figure 3.5: An example of estimation by cubic splines

However, the formulation of such a base is somewhat more complicated as it is defined recursively.

To define a B-spline base of dimension r+1 :

$$B_0^p(x), B_1^p(x), \dots, B_r^p(x), (3.9)$$

allowing us to decompose in the spline base any curve that is locally similar to a polynomial of degree *p* at each of the points, we first have to choose r+p+2 knots, denoted by $\kappa_0,\kappa_1,\dots,\kappa_{r+p+1}$.

The B_j^p components of the base are defined recursively as follows:

$$B_j^p(x) = \frac{x-\kappa_j}{\kappa_{j+p}-\kappa_j} B_j^{p-1}(x) + \frac{\kappa_{j+p+1}-x}{\kappa_{j+p+1}-\kappa_{j+1}} B_{j+1}^{p-1}(x)$$

for j=0,…,r and p>0.

For p=0, the $B_j^0(x)$ component is equal to 1 if $\kappa_j \leq x < \kappa_{j+1}$ and to 0 otherwise.

With the B-spline base defined as above, the regression model (3.1)($y = m(x) + \varepsilon$) can be rewritten as follows:

$$y = \sum_{j=0}^{r} \beta_j \beta_j^p(x), (3.10)$$

Note that the number of knots $r+p+2$ is greater than the number of parameters to be estimated $r+1$.

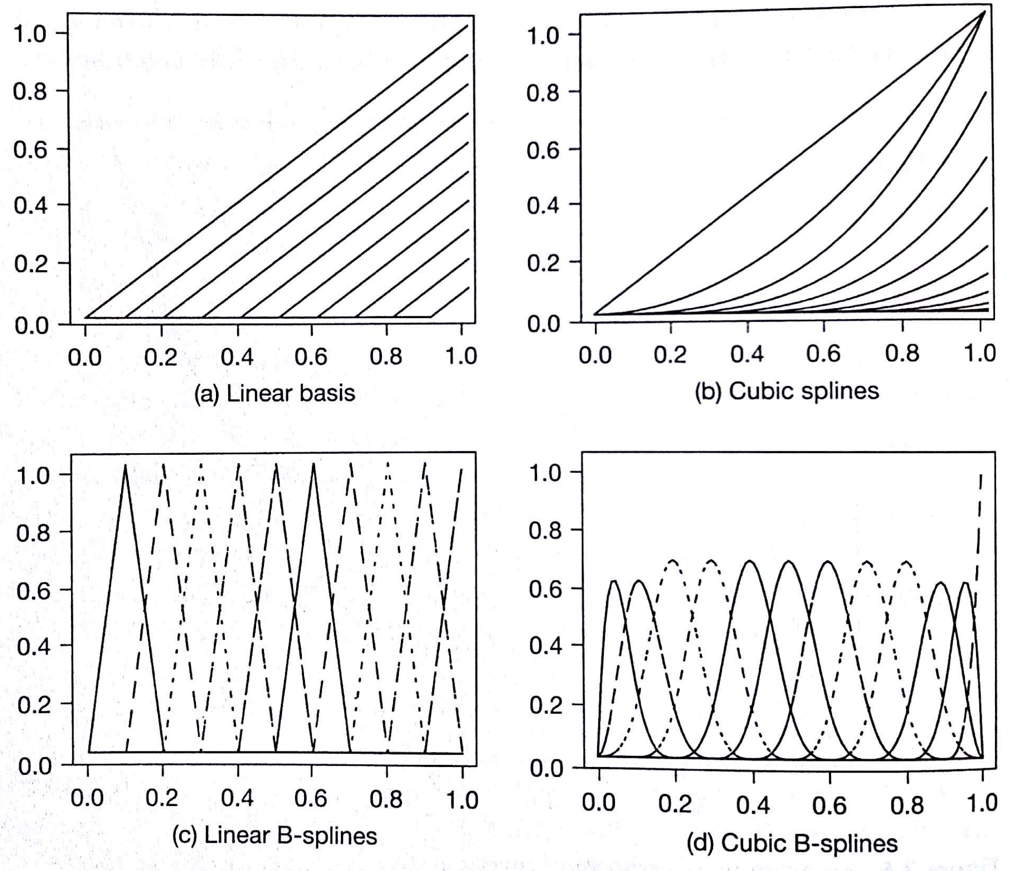
For a given number of knots, a power function base (3.7) $\iota, x, \ldots, x^p, (x-\kappa_1)_+^p, (x-\kappa_2)_+^p, \ldots, (x-\kappa q)_+^p$ uses $2(p+1)$ parameters more than does a B-spline base (3.9) $B_0^p(x), B_1^p(x), \ldots, B_r^p(x)$.

Panel (a) depicts the power function base (3.7) with p=1.The use of this base allows the estimation of piecewise linear models.

Panel (b) shows the power function base (3.7) with p=3; that is, with cubic splines.

Panel (c) shows the B-spline base (3.9) with p=1, and panel (d) represents the same base with p=3; that is, the cubic B-spline base.

We can see from this figure that the two B-spline bases have the following property they are composed of functions which take values of zero outside of the p+2 adjacent knots.



(a) Linear basis

(b) Cubic splines

(c) Linear B-splines

(d) Cubic B-splines

Figure 3.6: Spline bases

3.4 The Choice of the Parameters

The application of spline estimation requires prior choice of the knots, and of a smoothing parameter in the case of penalized splines.

The choice of a set of knots raises two central questions: the number of knots and their position.

In this section we consider the choice of the knots, and then turn to optimal smoothing.

3.4.1 The Number and Position of Knots

Comparing the two panels of Figure 3.4, we can see that the estimation results are sensitive to the number of knots.

The greater the number of knots, the more the estimated function varies: more knots corresponds to a greater number of different regimes in the different segments of the estimation of the regression function.

Knots can have an economic interpretation, if they are associated with a particular date or some observed change of regime.

If we have information on all of the dates at which such regime changes took place, we can simply place the knots at these dates.

In practice, our information is rarely complete enough for this to be a viable option.

The choice of the position of the knots can be carried out in two different ways.

The first consists in positioning the knots at the various quantiles of the explanatory variable x (Ruppert et al., 2003).

For example, if we fix the number of knots to be equal to three, then we place them at the quartiles of x, such that each segment contains 25% of the observations.

The second approach is to position the knots equidistantly, so that the segments are of the same width (Eilers and Marx, 1996).

Eilers and Marx (2004) compare these two approaches.

Ruppert (2002) has developed an algorithm for the choice of the number of knots.

He uses simulations to show that if the number of knots is small relative to a critical level, the estimated function will be biased and will contain relatively sizeable errors.

If the number of knots is greater than this critical level, then the estimation results are globally satisfactory.

In practice, we tend to favor a relatively large number of knots, in order to be above the critical value.

It is, however, obvious that the choice of the number of knots may be limited by the sample size.

If the number of observations is relatively small, it is possible to choose as many knots as there are observations, by making each knot correspond to an observation.

We refer to this as the special case of smoothing by splines.

To control for the evenness of the function, only the smoothing Parameter λ is used.

One remarkable property is that when the number of knots rises with the number of observations at a suitable rate, the estimation of the regression function by penalized splines converges on the result from smoothing by splines (Faraway, 2006).

3.4.2 The Smoothing parameter

To attenuate the effects of the number and position of the knots, a penalization parameter was introduced into the minimization criterion (3.6) describing the regression model.

As we saw, this penalization parameter constrains the extent to which the estimated function fluctuates.

The question which remains hanging is how the parameter associated with this term, λ , which plays the role of a smoothing parameter, should be chosen.

The optimal smoothing parameter is that which produces an estimation which is as close as possible to the true function.

In other words, we want to select the value of λ which minimizes the distance between $\hat{m}(x)$ and $m\,(x)$.

One solution consists in choosing the value of λ which minimizes the mean squared error:

$$MSE(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[\hat{m}(x_i) - m(x_i)]^2$$

But the function m(x) is unknown, and it is not possible to actually calculate the MSE.

We can get round this problem by appealing to cross-validation; that is, by choosing λ so as to minimize the following criterion:

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{m}_{-i}(x_i)]^2$$

where $\hat{m}_{-i}(x)$ is the estimated regression function from the sample of all observations apart from x_i .

This criterion is similar to that used to choose the smoothing parameter in kernel (2.8).

The drawback here is that we are required to estimate the same model n times, over n different samples of size n-1.

In practice, we often therefore use an approximation to this criterion, known as generalized cross-validation:

$$GCV(\lambda) = \frac{n\sum_{i=1}^{n}[y_i - \widehat{m}(x_i)]^2}{[tr(I - A)]^2}$$

where $A=X(X^TX + \lambda S)^{-1}X^T$ is the hat matrix of the regression, X is the column of the observed values $x_1,\ldots,x_n$, I is the identity matrix, and tr is the trace.

The evaluation of GCV criterion requires only one estimation of the regression function, over the sample containing all of the observations.

Moreover, for a sufficiently large number of observations, this criterion chooses a value of λ similar to that which pertains from the minimization of the MISE.

The application discussed in the introduction, regarding the Phillips curve (Figure 3.1), resulted from penalized spline estimation, with a cubic base, ten knots, and bandwidth that was determined according to the GCV criterion ($\lambda = 2.19$).

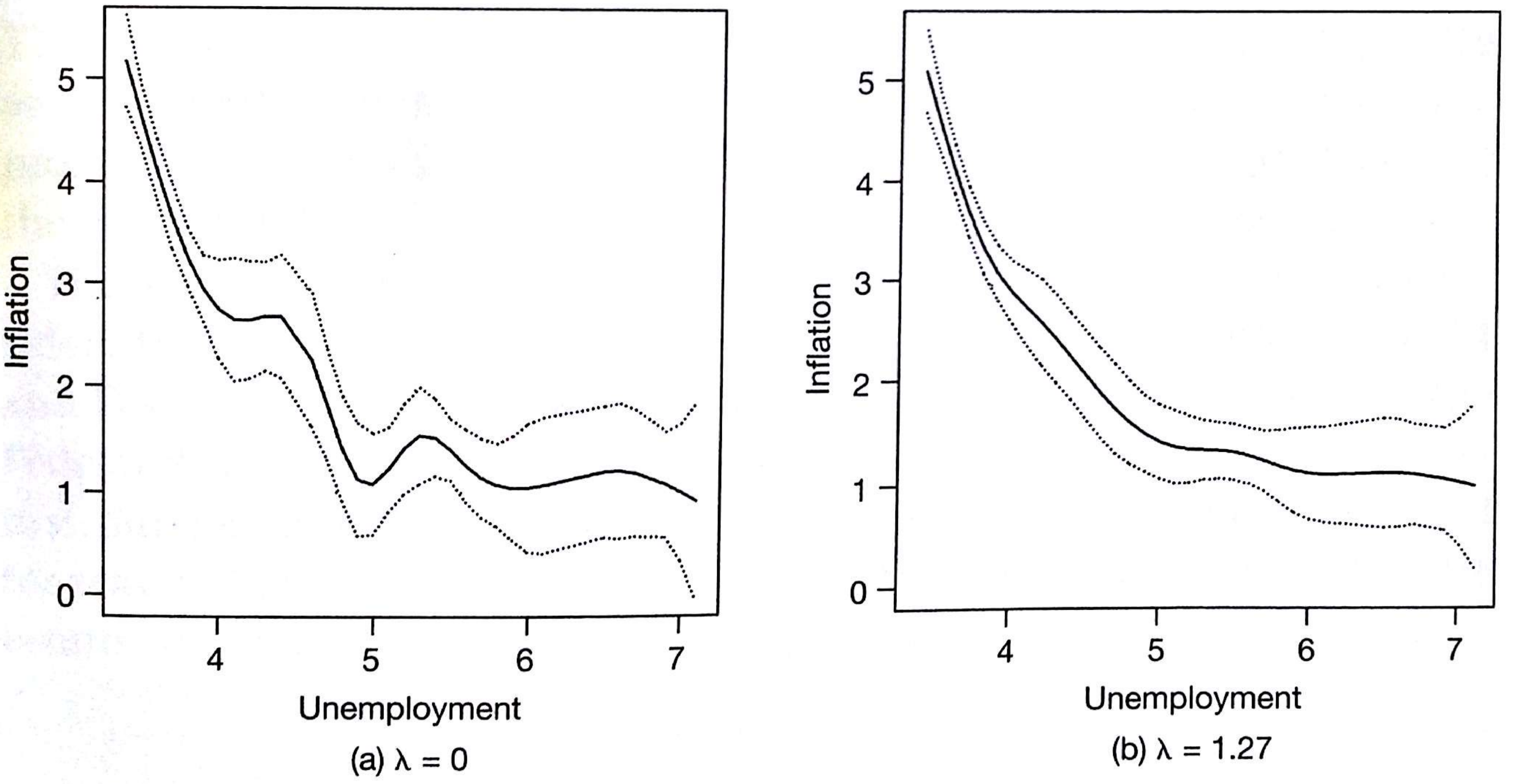


Figure 3.7: The smoothing parameter

Panel (a) of Figure 3.7 depicts the results from the estimation of the Phillips curve using the same data and with the same choices as in Figure 3.1, except for the smoothing parameter, which is here set so that λ equals 0.

In other words, the estimation here comes from non-penalized splines.

Compared to Figure 3. 1, the estimated function is much more variable.

A comparison of these two graphs emphasizes that the penalization of the estimation leads to a greater smoothing of the estimated function.

Panel (b) of Figure 3.7 presents the estimation of the Phillips curve using the same data and with the same choices as in Figure 3.1, except for the number of knots, which is here set equal to 9.

The smoothing parameter, obtained from the GCV criterion, now produces λ =1.27.

The comparison to Figure 3.1, which has ten knots and $\lambda=2.19$, reveals that the two figures are in fact very similar: different knot choices can make only little difference to the estimation results.

Here it is the smoothing parameter which is adjusted, so as to penalize the estimation more severely when the number of knots is greater.