# Chapter 4

# Splines

Through-out this section, the regression function $f$ will depend on a single, real-valued predictor $X$ ranging over some possibly infinite interval of the real line, $I \subset \mathbb{R}$. Therefore, the (mean) dependence of $Y$ on $X$ is given by

$$f(x) = \mathrm{E}(Y|X = x), x \in I \subset \mathbb{R}. \tag{4.1}$$

For spline models, estimate definitions and their properties are more easily characterized in the context of linear spaces.

## 4.1 Linear Spaces

In this chapter our approach to estimating $f$ involves the use of finite dimensional linear spaces.

Remember what a linear space is? Remember definitions of dimension, linear subspace, orthogonal projection, etc...

Why use linear spaces?

- Makes estimation and statistical computations easy.
- Has nice geometrical interpretation.
- It actually can specify a broad range of models given we have discrete data.

Using linear spaces we can define many families of function $f$; straight lines, polynomials, splines, functions with two continuous derivatives, and many other spaces (these are examples for the case where $\mathbf{x}$ is a scalar). The point is: we have many options.

Notice that in most practical situation we will have observations $(\mathbf{X}_i, Y_i), i = 1, \ldots, n$. In some situations we are only interested in estimating $f(\mathbf{X}_i), i = 1, \ldots, n$. In fact, in many situations it is all that matters from a statistical point of view. We will write $\mathbf{f}$ when referring to the this vector and $\hat{\mathbf{f}}$ when referring to an estimate. Think of how its different to know $f$ and know $\mathbf{f}$.

Let's say we are interested in estimating $\mathbf{f}$. A common practice in statistics is to assume that $\mathbf{f}$ lies in some *linear space*, or is well approximated by a $\mathbf{g}$ that lies in some *linear space*.

For example for simple linear regression we assume that $\mathbf{f}$ lies in the linear space of lines:

$$\alpha + \beta\mathbf{x}, (\alpha, \beta)' \in \mathbb{R}^2.$$

For linear regression in general we assume that $\mathbf{f}$ lies in the linear space of linear combinations of the covariates or rows of the design matrix. How do we write it out?

Note: Through out this chapter $f$ is used to denote the true regression function and $g$ is used to denote an arbitrary function in a particular space of functions. It isn't necessarily true that $f$ lies in this space of function. Similarly we use $\mathbf{f}$ to denote the true function evaluated at the design points or observed covariates and $\mathbf{g}$ to denote an arbitrary function evaluated at the design points or observed covariates.

Now we will see how and why it's useful to use linear models in a more general setting.

A linear model of order $p$ for the regression function (4.1) consists of a $p$-dimensional linear space $\mathcal{G}$, having as a basis the function

$$B_j(\mathbf{x}), j = 1, \ldots, p$$

defined for $\mathbf{x} \in I$. Each member $g \in \mathcal{G}$ can be written uniquely as a linear combination

$$g(\mathbf{x}) = g(\mathbf{x}; \boldsymbol{\theta}) = \theta_1 B_1(\mathbf{x}) + \ldots + \theta_p B_p(\mathbf{x})$$

for some value of the coefficient vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)' \in \mathbb{R}^p$.

Notice that $\boldsymbol{\theta}$ specifies the point $g \in \mathcal{G}$.

How would you write this out for linear regression?

Given observations $(\mathbf{X}_i, Y_i), i = 1, \ldots, n$ the least squares estimate (LSE) of $\mathbf{f}$ or equivalently $f(\mathbf{x})$ is defined by $\hat{f}(\mathbf{x}) = g(\mathbf{x}; \hat{\boldsymbol{\theta}})$, where

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \sum_{i=1}^{n} \{Y_i - g(\mathbf{X}_i, \boldsymbol{\theta})\}^2.$$

Define the vector $\mathbf{g} = \{g(x_1), \ldots, g(x_n)\}'$. Then the distribution of the observations of $Y|X = x$ are in the family

$$\{N(\mathbf{g}, \sigma^2 \mathbf{I}_n); \mathbf{g} = [g(x_1), \ldots, g(x_n)]', \; g \in \mathcal{G}\} \tag{4.2}$$

and if we assume the errors $\varepsilon$ are IID normal and that $f \in \mathcal{G}$ we have that $\hat{\mathbf{f}} = [g(x_1; \hat{\boldsymbol{\theta}}), \ldots, g(x_n; \hat{\boldsymbol{\theta}})]$ is the maximum likelihood estimate. The estimand $\mathbf{f}$ is an $n \times 1$ vector. But how many parameters are we really estimating?

Equivalently we can think of the distribution is in the family

$$\{N(\mathbf{B}\boldsymbol{\theta}, \sigma^2); \boldsymbol{\theta} \in \mathbb{R}^p\} \tag{4.3}$$

and the maximum likelihood estimate for $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}}$. Here $\mathbf{B}$ is a matrix of basis elements defined soon...

Here we start seeing for the first time where the name *non-parametric* comes from. How are the approaches (4.2) and (4.3) different?

Notice that obtaining $\hat{\boldsymbol{\theta}}$ is easy because of the linear model set-up. The ordinary least square estimate is

$$(\mathbf{B}'\mathbf{B})\hat{\boldsymbol{\theta}} = \mathbf{B}'\mathbf{Y}$$

where $\mathbf{B}$ is is the $n \times p$ design matrix with elements $[\mathbf{B}]_{ij} = B_j(\mathbf{X}_i)$. When this solution is unique we refer to $g(x; \hat{\boldsymbol{\theta}})$ as the OLS projection of $\mathbf{Y}$ into $\mathcal{G}$ (as learned in the first term).

### 4.1.1 Parametric versus non-parametric

In some cases, we have reason to believe that the function $f$ is actually a member of some linear space $\mathcal{G}$. Traditionally, inference for regression models depends on $f$ being representable as some combination of known predictors. Under this assumption, $f$ can be written as a combination of basis elements for some value of the coefficient vector $\boldsymbol{\theta}$. This provides a *parametric* specification for $f$. No matter how many observations we collect, there is no need to look outside the fixed, finite-dimensional, linear space $\mathcal{G}$ when estimating $f$.

In practical situations, however, we would rarely believe such relationship to be exactly true. Model spaces $\mathcal{G}$ are understood to provide (at best) approximations to $f$; and as we collect more and more samples, we have the freedom to audition richer and richer classes of models. In such cases, all we might be willing to say about $f$ is that it is *smooth* in some sense, a common assumption being that $f$ have two bounded derivatives. Far from the assumption that $f$ belong to a fixed, finite-dimensional linear space, we instead posit a *nonparametric* specification for $f$. In this context, model spaces are employed mainly in our approach to inference; first in the questions we pose about an estimate, and then in the tools we apply to address them. For example, we are less interested in the actual values of the coefficient $\boldsymbol{\theta}$, e.g. whether or not an element of $\boldsymbol{\theta}$ is significantly different from zero to the 0.05 level. Instead we concern ourselves with functional properties of $g(\mathbf{x}; \hat{\boldsymbol{\theta}})$, the estimated curve or surface, e.g. whether or not a peak is real.

To ascertain the local behavior of OLS projections onto approximation spaces $\mathcal{G}$, define the pointwise, mean squared error (MSE) of $\hat{g}(\mathbf{x}) = g(\mathbf{x}; \hat{\boldsymbol{\theta}})$ as

$$\mathrm{E}\{f(\mathbf{x}) - \hat{g}(\mathbf{x})\}^2 = \mathrm{bias}^2\{\hat{g}(\mathbf{x})\} + \mathrm{var}\{\hat{g}(\mathbf{x})\}$$

where

$$\mathrm{bias}\{\hat{g}(\mathbf{x})\} = f(x) - \mathrm{E}\{\hat{g}(\mathbf{x})\} \tag{4.4}$$

and

$$\mathrm{var}\{\hat{g}(\mathbf{x})\} = \mathrm{E}\{\hat{g}(\mathbf{x}) - \mathrm{E}[\hat{g}(\mathbf{x})]\}^2$$

When the input values $\{\mathbf{X}_i\}$ are deterministic the expectations above are with respect to the noisy observation $Y_i$. In practice, MSE is defined in this way even in the random design case, so we look at expectations conditioned on $\mathbf{X}$.

When we do this, standard results in regression theory can be applied to derive an expression for the variance term

$$\mathrm{var}\{\hat{g}(\mathbf{x})\} = \sigma^2 \mathbf{B}(\mathbf{x})'(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}(\mathbf{x})$$

where $\mathbf{B}(\mathbf{x}) = (B_1(\mathbf{x}), \ldots, B_p(\mathbf{x}))'$, and the error variance is assumed constant.

Under the parametric specification that $f \in \mathcal{G}$, what is the bias?

This leads to classical t- and F-hypothesis tests and associated parametric confidence intervals for $\boldsymbol{\theta}$. Suppose on the other hand, that $f$ is not a member of $\mathcal{G}$, but rather can be reasonably approximated by an element in $\mathcal{G}$. The bias (4.4) now reflects the ability of functions in $\mathcal{G}$ to capture the essential features of $f$.

## 4.2 Local Polynomials

In practical situations, a statistician is rarely blessed with simple linear relationship between the predictor $X$ and the observed output $Y$. That is, as a description of the regression function $f$, the model

$$g(x; \boldsymbol{\theta}) = \theta_1 + \theta_2 x, x \in I$$

typically ignores obvious features in the data. This is certainly the case for the values of $^{87}Sr$.

The Strontium data set was collected to test several hypotheses about the catastrophic events that occurred approximately 65 million years ago. The data contains Age in million of years and the ratios described here. There is a division between two geological time periods, the Cretaceous (from 66.4 to 144 million years ago) and the Tertiary (spanning from about 1.6 to 66.4 million years ago). Earth scientist believe that the boundary between these periods is distinguished by tremendous changes in climate that accompanied a mass extension of over half of the species inhabiting the planet at the time. Recently, the compositions of Strontium (Sr) isotopes in sea water has been used to evaluate several hypotheses about the cause of these extreme events. The dependent variable of the data-set is related to the isotopic make up of Sr measured for the shells of marine organisms. The Cretaceous-Tertiary boundary is referred to as KTB. There data shows a peak is at this time and this is used as evidence that a meteor collided with earth.

The data presented in the Figure **??** represents standardized ratio of strontium-87 isotopes ($^{87}$Sr) to strontium-86 isotopes ($^{86}$Sr) contained in the shells of foraminifera fossils taken form cores collected by deep sea drilling. For each sample its time in history is computed and the standardized ratio is computed:

$$^{87}\delta\text{Sr} = \left( \frac{^{87}\text{Sr}/^{86}\text{Sr sample}}{^{87}\text{Sr}/^{86}\text{Sr sea water}} - 1 \right) \times 10^5.$$

Earth scientist expect that $^{87}\delta$Sr is a smooth-varying function of time and that deviations from smoothness are mostly measurement error.

To overcome this deficiency, we might consider a more flexible polynomial model. Let $\mathcal{P}_k$ denote the linear space of polynomials in $x$ of order at most $k$ defined as

$$g(x; \boldsymbol{\theta}) = \theta_1 + \theta_2 x + \ldots + \theta_k x^{k-1}, x \in I$$

for the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k$. Note that the space $\mathcal{P}_k$ consists of polynomials having degree at most $k - 1$.

In exceptional cases, we have reasons to believe that the regression function $f$ is in fact a high-order polynomial. This parametric assumption could be based on physical or physiological models describing how the data were generated.

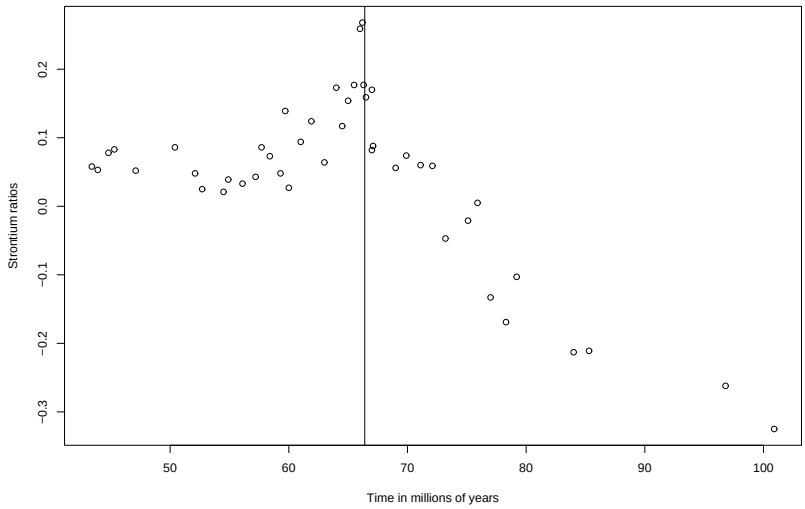


Figure 4.1: $^{87}\delta$Sr data.

For historical values of $^{87}\delta Sr$ we consider polynomials simply because our scientific intuition tells us that $f$ should be smooth.

Recall Taylor's theorem: polynomials are good at approximating well-behaved functions in reasonably tight neighborhoods. If all we can say about $f$ is that it is smooth in some sense, then either implicitly or explicitly we consider high-order polynomials because of their favorable approximation properties.

If $f$ is not in $\mathcal{P}_k$ then our estimates will be biased by an amount that reflects the approximation error incurred by a polynomial model.

Computational Issue: The basis of monomials

$$B_j(x) = x^{j-1} \text{ for } j = 1, \ldots, k$$

is not well suited for numerical calculations ($x^8$ can be VERY BIG compared to $x$). While convenient for analytical manipulations (differentiation, integration), this basis is *ill-conditioned* for $k$ larger than $8$ or $9$. Most statistical packages use the orthogonal Chebyshev polynomials (used by the R command `poly()`).

An alternative to polynomials is to consider the space $\mathcal{PP}_k(\mathbf{t})$ of piecewise polynomials with break points $\mathbf{t} = (t_0, \ldots, t_{m+1})'$. Given a sequence $a = t_0 < t_1 < \ldots < t_m < t_{m+1} = b$, construct $m + 1$ (disjoint) intervals

$$I_l = [t_{l-1}, t_l), 1 \leq l \leq m \text{ and } I_{m+1} = [t_m, t_{m+1}],$$

whose union is $I = [a, b]$. Define the piecewise polynomials of order $k$

$$g(x) = \begin{cases} g_1(x) = \theta_{1,1} + \theta_{1,2}x + \ldots + \theta_{1,k}x^{k-1}, & x \in I_1 \\ \vdots & \vdots \\ g_{m+1}(x) = \theta_{m+1,1} + \theta_{m+1,2}x + \ldots + \theta_{m+1,k}x^{k-1}, & x \in I_{k+1}. \end{cases}$$

In homework 2, we saw or will see that piecewise polynomials are a linear space that present an alternative to polynomials. However, it is hard to justify the breaks in the function $g(x; \hat{\boldsymbol{\theta}})$.

## 4.3 Splines

In many situations, breakpoints in the regression function do not make sense. Would forcing the piecewise polynomials to be continuous suffice? What about continuous first derivatives?

We start by consider the subspaces of the piecewise polynomial space. We will denote it with $\mathcal{PP}_k(\mathbf{t})$ with $\mathbf{t} = (t_1, \ldots, t_m)'$ the break-points or interior knots. Different break points define different spaces.

We can put constrains on the behavior of the functions $g$ at the break points. (We can construct tests to see if these constrains are suggested by the data but, will not go into this here)

Here is a trick for forcing the constrains and keeping the linear model set-up. We can write any function $g \in \mathcal{PP}_k(\mathbf{t})$ in *the truncated basis power*:

$$g(x) = \theta_{0,1} + \theta_{0,2}x + \ldots + \theta_{0,k}x^{k-1} +$$

$$\theta_{1,1}(x-t_1)_+^0 + \theta_{1,2}(x-t_1)_+^1 + \ldots + \theta_{1,k}(x-t_1)_+^{k-1} +$$
$$\vdots$$
$$\theta_{m,1}(x-t_m)_+^0 + \theta_{m,2}(x-t_m)_+^1 + \ldots + \theta_{m,k}(x-t_m)_+^{k-1}$$

where $(\cdot)_+ = \max(\cdot, 0)$. Written in this way the coefficients $\theta_{1,1}, \ldots, \theta_{1,k}$ record the jumps in the different derivative from the first piece to the second.

Notice that the constrains reduce the number of parameters. This is in agreement with the fact that we are forcing more smoothness.

Now we can force constrains, such as continuity, by putting constrains like $\theta_{1,1} = 0$ etc...

We will concentrate on the cubic splines which are continuous and have continuous first and second derivatives. In this case we can write:

$$g(x) \;=\; \theta_{0,1} + \theta_{0,2}x + \ldots + \theta_{0,4}x^3 + \theta_{1,k}(x-t_1)^3 + \ldots + \theta_{m,k}(x-t_m)^3$$

How many "parameters" in this space?

Note: It is always possible to have less restrictions at knots where we believe the behavior is "less smooth", e.g for the Sr ratios, we may have "unsmoothness" around KTB.

We can write this as a linear space. This setting is not computationally convenient. In S-Plus there is a function `bs()` that makes a basis that is convenient for computations.

There is asymptotic theory that goes along with all this but we will not go into the details. We will just notice that

$$\mathrm{E}[f(x) - g(x)] = O(h_l^{2k} + 1/n_l)$$

where $h_l$ is the size of the interval where $x$ is in and $n_l$ is the number of points in it. What does this say?

### 4.3.1 Splines in terms of Spaces and sub-spaces

The$p$-dimensional spaces described in Section 4.1 were defined through basis function $B_j(\mathbf{x}), j = 1, \ldots, p$. So in general we defined for a given range $I \subset \mathbb{R}^k$

$$\mathcal{G} = \{g : g(\mathbf{x}) = \sum_{j=1}^{p} \theta_j \beta_j(\mathbf{x}), \mathbf{x} \in I, (\theta_1, \ldots, \theta_p) \in \mathbb{R}^p\}$$

In the previous section we concentrated on $\mathbf{x} \in \mathbb{R}$.

In practice we have design points $x_1, \ldots, x_n$ and a vector of responses $\mathbf{y} = (y_1, \ldots, y_n)$. We can think of $\mathbf{y}$ as an element in the $n$-dimensional vector space $\mathbb{R}^n$. In fact we can go a step further and define a Hilbert space with the usual inner product definition that gives us the norm

$$||\mathbf{y}|| = \sum_{i=1}^{n} y_i^2$$

Now we can think of least squares estimation as the projection of the data $\mathbf{y}$ to the sub-space $\mathbf{G} \subset \mathbb{R}^n$ defined by $\mathcal{G}$ in the following way

$$\mathbf{G} = \{\mathbf{g} \in \mathbb{R}^n : \mathbf{g} = [g(x_1), \ldots, g(x_n)]', g \in \mathcal{G}\}$$

Because this space is spanned by the vectors $[B_1(x_1), \ldots, B_p(x_n)]$ the projection of $\mathbf{y}$ onto $\mathbf{G}$ is

$$\mathbf{B}(\mathbf{B}'\mathbf{B})^{-}\mathbf{B}'\mathbf{y}$$

as learned in 751. Here $[\mathbf{B}]_{ij} = B_j(x_i)$.

## 4.4 Natural Smoothing Splines

Natural splines add the constrain that the function must be linear after the knots at the end points. This forces 2 more restrictions since $f''$ must be 0 at the end points, i.e the space has $k + 4 - 2$ parameters because of this extra 2 constrains.

So where do we put the knots? How many do we use? There are some data-driven procedures for doing this. Natural Smoothing Splines provide another approach.

What happens if the knots coincide with the dependent variables $\{X_i\}$. Then there is a function $g \in \mathcal{G}$, the space of cubic splines with knots at $(x_1, \ldots, x_n)$, with $g(x_i) = y_i, i, \ldots, n$, i.e. we haven't smoothed at all.

Consider the following problem: among all functions $g$ with two continuous first two derivatives, find one that minimizes the penalized residual sum of squares

$$\sum_{i=1}^{n} \{y_i - g(x_i)\}^2 + \lambda \int_a^b \{g''(t)\}^2 \, dt$$

where $\lambda$ is a fixed constant, and $a \leq x_1 \leq \ldots \leq x_n \leq b$. It can be shown (Reinsch 1967) that the solution to this problem is a natural cubic spline with knots at the values of $x_i$ (so there are $n-2$ interior knots and $n-1$ intervals). Here $a$ and $b$ are arbitrary as long as they contain the data.

It seems that this procedure is over-parameterized since a natural cubic spline as this one will have $n$ degrees of freedom. However we will see that the penalty makes this go down.

### 4.4.1 Computational Aspects

We use the fact that the solution is a natural cubic spline and write the possible answers as

$$g(x) = \sum_{j=1}^{n} \theta_j B_j(x)$$

where $\theta_j$ are the coefficients and $B_j(x)$ are the basis functions. Notice that if these were cubic splines the functions lie in a $n+2$ dimensional space, but the natural splines are an $n$ dimensional subspace.

Let $\mathbf{B}$ be the $n \times n$ matrix defined by

$$B_{ij} = B_j(x_i)$$

and a penalty matrix $\mathbf{\Omega}$ by

$$\Omega_{ij} = \int_a^b B_i''(t) B_j''(t)\, dt$$

now we can write the penalized criterion as

$$(\mathbf{y} - \mathbf{B}\boldsymbol{\theta})'(\mathbf{y} - \mathbf{B}\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}\mathbf{\Omega}\boldsymbol{\theta}$$

It seems there are no boundary derivatives constraints but they are implicitly imposed by the penalty term.

Setting derivatives with respect to $\boldsymbol{\theta}$ equal to 0 gives the estimating equation:

$$(\mathbf{B}'\mathbf{B} + \lambda\mathbf{\Omega})\boldsymbol{\theta} = \mathbf{B}'\mathbf{y}.$$

The $\hat{\boldsymbol{\theta}}$ that solves this equation will give us the estimate $\hat{\mathbf{g}} = \mathbf{B}\hat{\boldsymbol{\theta}}$.

Is this a linear smoother?

Write:

$$\hat{\mathbf{g}} = \mathbf{B}\boldsymbol{\theta} = \mathbf{B}(\mathbf{B}'\mathbf{B} + \lambda\mathbf{\Omega})^{-1}\mathbf{B}'\mathbf{y} = (\mathbf{I} + \lambda\mathbf{K})^{-1}\mathbf{y}$$

where $\mathbf{K} = \mathbf{B} - 1'\mathbf{\Omega}\mathbf{B}^{-1}$. Notice we can write the criterion as

$$(\mathbf{y} - \mathbf{g})'(\mathbf{y} - \mathbf{g}) + \lambda \mathbf{g}'\mathbf{K}\mathbf{g}$$

If we look at the "kernel" of this linear smoother we will see that it is similar to the other smoothers presented in this class.

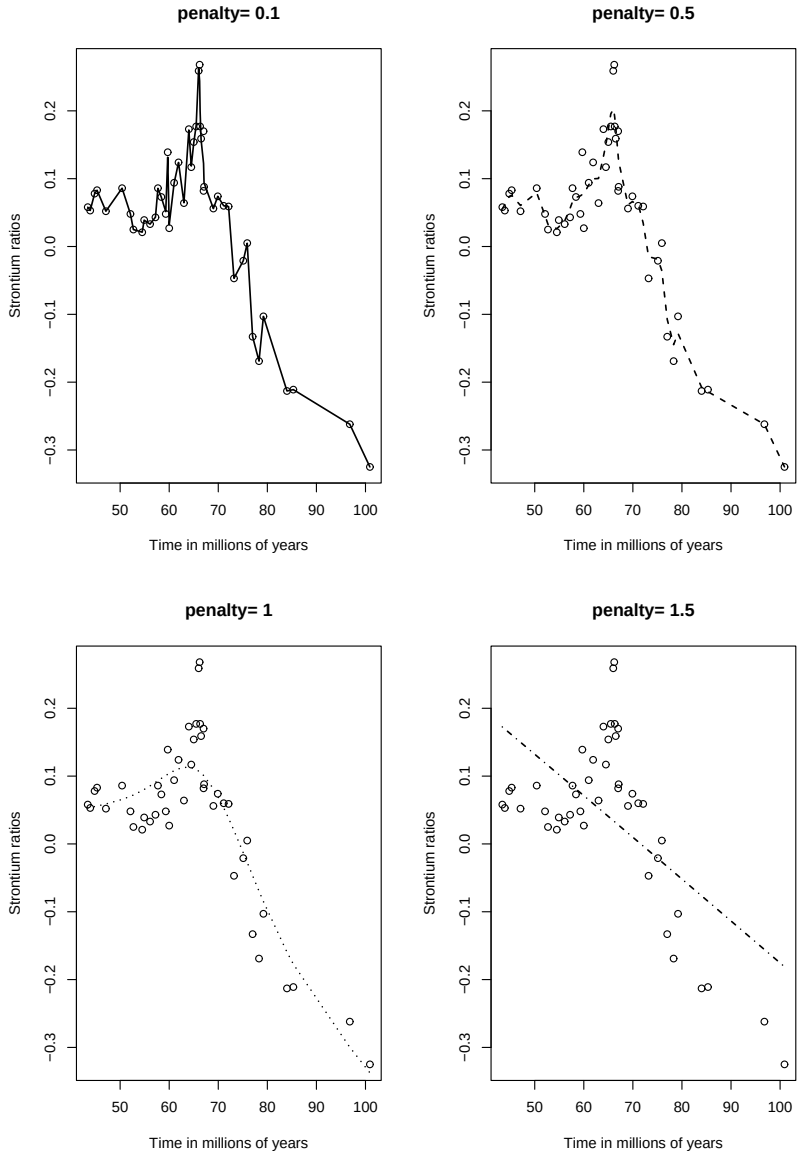


Figure 4.2: Smoothing spline fitted using different penalties.

# Bibliography

[1] Eubank, R.L. (1988), *Smoothing Splines and Nonparametric Regression,* New York: Marcel Decker.

[2] Reinsch, C.H. (1967) Smoothing by Splins Functions. *Numerische Mathematik,* 10: 177–183

[3] Schoenberg, I.J. (1964), "Spline functions and the problem of graduation," *Proceedings of the National Academy of Science,* USA 52, 947–950.

[4] Silverman (1985) "Some Aspects of the spline smoothing approach to non-parametric regression curve fitting". *Journal of the Royal Statistical Society B* 47: 1–52.

[5] Wahba, G. (1990), *Spline Models for Observational Data,* CBMS-NSF Regional Conference Series, Philadelphia: SIAM.