STAT40810 — Stochastic Models

Brendan Murphy

Week 3

Spline Regression

Splines

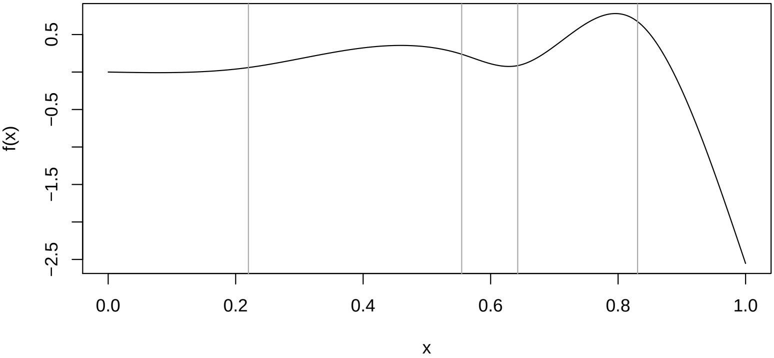
- We will look at spline regression as a precursor to doing smoothing splines. regression.
- First, we need to look at splines in general.
- **Potential Books Of Interest:**
  - Ramsay and Silverman (1997) Functional Data Analysis.
  - Hastie and Tibshirani (1990) Generalized Additive Models.
  - Hastie, Tibshirani and Friedman (2001) The Elements of Statistical Learning.
  - de Boor (1978) A Practical Guide to Splines.

Splines

- A Kth order spline with M knots located at $z_1 < z_2 < \cdots < z_M$ is a function $f(x)$ such that
  1. The function is a Kth order polynomial in the intervals (z_i, z_{i+1}).
  2. The function has $K - 1$ continuous derivatives.
- A spline of order K with M knots has $(K+1)(M+1) - KM = K + M + 1$ parameters. The $(K+1)(M+1)$ term is for the equation between the knots and the KM term is for the boundary conditions at the knots.
- We will primarily concentrate on cubic splines.

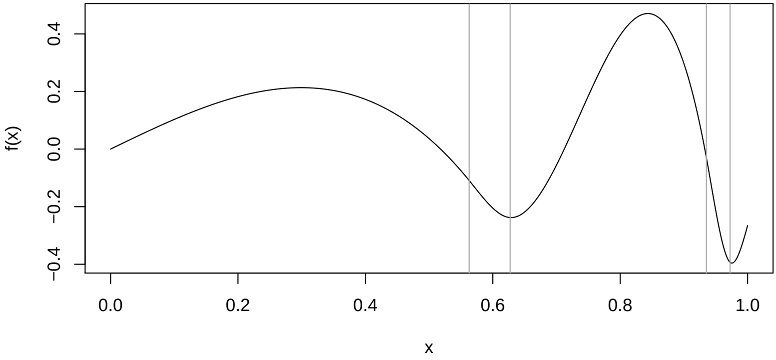
Example 1: Spline

- Here's an example of a cubic spline $f(x)$. The knots are marked in grey.



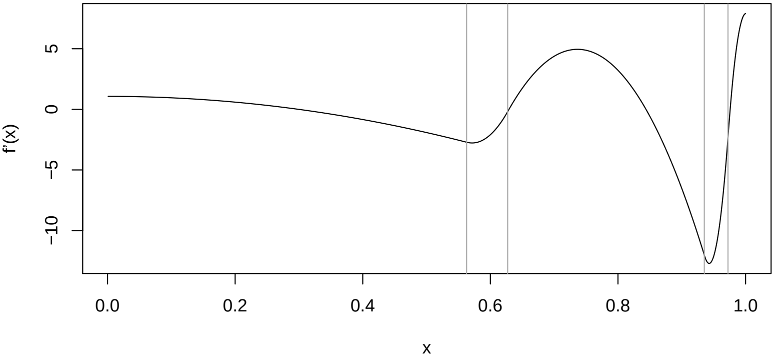
Example 2: Spline

- Here's another example of a cubic spline $f(x)$. The knots are marked in grey.



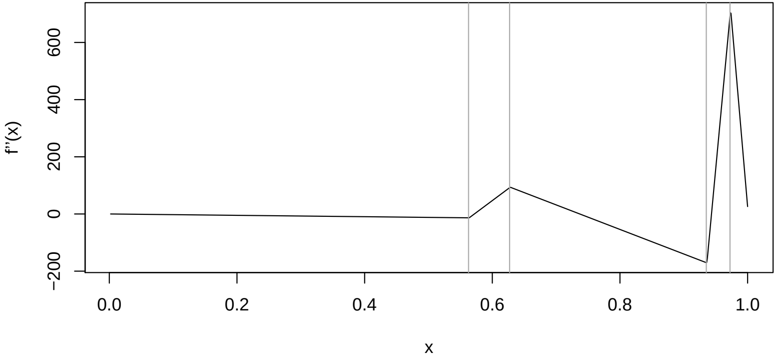
Example 2: Spline (First Derivative)

- The derivative $f'(x)$ of the spline is a spline of order 2. Notice the quadratic shape between knots.



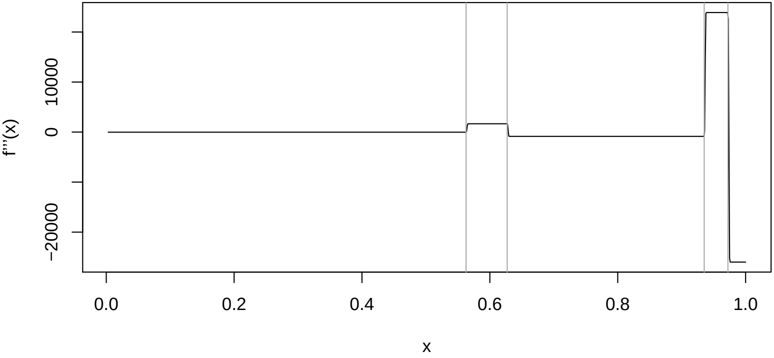
Example 2: Spline (Second Derivative)

- The second derivative $f''(x)$ of the spline is piecewise linear and continuous.



Example 2: Spline (Third Derivative)

- The third derivative is piecewise constant; that is, where it exists. It doesn't necessarily exist at the knots.



Spline Basis Functions

- A spline of order K with M knots at $z_1 < z_2 < \cdots < z_M$, can be written as a linear combination of basis functions.
- That is,

$$f(x) = b_0 + \sum_{l=1}^{M+K} b_l f_l(x).$$

- There are a number of ways of doing this.
- One standard way is to write

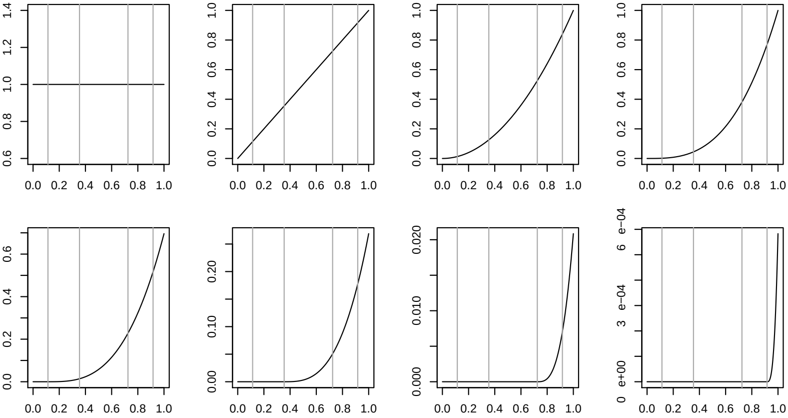
$$f(x) = b_0 + \sum_{k=1}^{K} b_k x^k + \sum_{m=1}^{M} b_{K+m}(x - z_k)_+^K,$$

where $(x - z_k)_+$ is the positive part of $(x - z_k)$.

- This is called the *truncated power series basis*.

Spline Basis Functions

- An example of the truncated power series basis is shown.



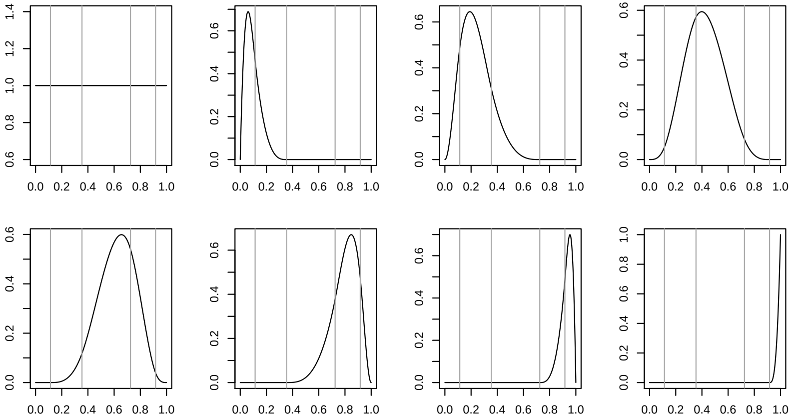
- This basis is not computationally efficient when the number of knots is large.

B-Spline Basis

- The truncated power series basis for splines is not suitable when the number of knots is large.
- An alternative basis called the *B-spline basis* offers a computational advantage.
- Each basis function in the *B*-spline basis has bounded support (ie. positive over a finite interval).
- This simplifies some calculations needed when using splines.

B-Spline Basis

- An example of the *B*-spline basis for the same knots as the previous truncated power series basis is shown.



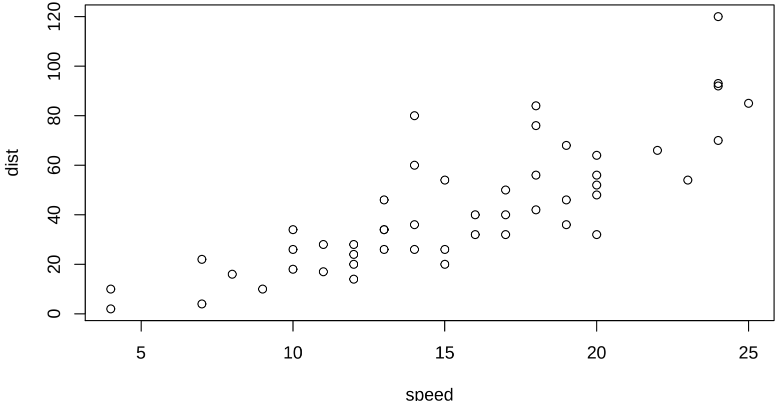
Spline Regression

- Suppose that we have data values $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- We could model the relationship between the variables using a spline function.
- If we specify the knot locations $z_1 < z_2 < \ldots < z_M$ and the order of the spline, then the problem is simply a regression problem.
- We can use regression techniques to find the coefficients $(b_0, b_1, \ldots, b_{K+M})$ that minimize

$$\sum_{i=1}^{n} \left[y_i - b_0 - \sum_{l=1}^{K+M} b_l f_l(x_i) \right]^2 .$$

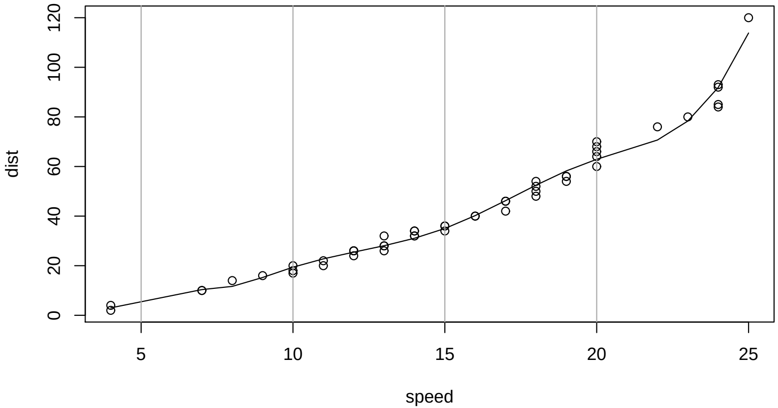
Example: Braking Distances

- The data were collected, recording the speed of cars and the distances taken to stop. Note that the data were recorded in the 1920s.
- There are 50 observations on 2 variables: Speed (mph) and Stopping Distance (ft).



Spline Regression

- Suppose we fit a cubic spline with knots at 10, 15 and 20ft.



Spline Regression

Here's code to model the motorcycle data.

```
#Load data and plot it
data(cars)
plot(cars)

# Load splines library
library(splines)

#Form B-spline basis functions for data
basis <- bs(cars$speed,knots=c(5,10,15,20))

# Fit the model
fit <- lm(cars$dist~basis)

#Show fit on the plot
plot(cars)
points(cars$speed,predict(fit),type="l",col="red")
```