Chapter 7 – Moving beyond linearity

Wenjing Liao

School of Mathematics
Georgia Institute of Technology

Math 4803
Fall 2019
Outline

1. 7.1 – Polynomial regression
2. 7.2 – Step functions
3. 7.3 – Basis functions
4. 7.4 – Splines
5. 7.5 – Smoothing splines
6. Multidimensional splines
Polynomial regression

**Linear function:**  \( y_i = \beta_0 + \beta_1 x_i + \epsilon_i \)

**Polynomial function:**

\[
y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \epsilon_i,
\]

**Logistic regression using polynomials:**

\[
\Pr(y_i > 250 | x_i) = \frac{\exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_d x_i^d)}{1 + \exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_d x_i^d)}.
\]
The **Wage** data. Left: The solid blue curve is a degree-4 polynomial of wage (in thousands of dollars) as a function of age, fit by least squares. The dotted curves indicate an estimated 95% confidence interval. Right: We model the binary event wage>250 using logistic regression, again with a degree-4 polynomial. The fitted posterior probability of wage exceeding $250,000 is shown in blue, along with an estimated 95% confidence interval.
Regression using indicator functions

Indicator functions:

\[
\begin{align*}
C_0(X) &= I(X < c_1), \\
C_1(X) &= I(c_1 \leq X < c_2), \\
C_2(X) &= I(c_2 \leq X < c_3), \\
&\vdots \\
C_{K-1}(X) &= I(c_{K-1} \leq X < c_K), \\
C_K(X) &= I(c_K \leq X),
\end{align*}
\]

Regression:

\[
y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \ldots + \beta_K C_K(x_i) + \epsilon_i.
\]

Logistic regression:

\[
\Pr(y_i > 250|x_i) = \frac{\exp(\beta_0 + \beta_1 C_1(x_i) + \ldots + \beta_K C_K(x_i))}{1 + \exp(\beta_0 + \beta_1 C_1(x_i) + \ldots + \beta_K C_K(x_i))}
\]
FIGURE 7.2. The Wage data. Left: The solid curve displays the fitted value from a least squares regression of wage (in thousands of dollars) using step functions of age. The dotted curves indicate an estimated 95% confidence interval. Right: We model the binary event wage>250 using logistic regression, again using step functions of age. The fitted posterior probability of wage exceeding $250,000 is shown, along with an estimated 95% confidence interval.
Outline

1. 7.1 – Polynomial regression
2. 7.2 – Step functions
3. 7.3 – Basis functions
4. 7.4 – Splines
5. 7.5 – Smoothing splines
6. Multidimensional splines
Regression using basis functions

Basis functions: \( b_1(\cdot), \ldots, b_K(\cdot) \)

Regression using basis functions:

\[
y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \beta_3 b_3(x_i) + \ldots + \beta_K b_K(x_i) + \epsilon_i.
\]

Popular basis:
- Polynomials
- Fourier basis
- Wavelet basis
- Splines
Outline

1. 7.1 – Polynomial regression
2. 7.2 – Step functions
3. 7.3 – Basis functions
4. 7.4 – Splines
5. 7.5 – Smoothing splines
6. Multidimensional splines
7.4.1 – Piecewise polynomials

Cubic polynomial:

\[ y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i, \]

Piecewise polynomial with a single knot at \( c \):

\[
y_i = \begin{cases} 
\beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \beta_{31} x_i^3 + \epsilon_i & \text{if } x_i < c; \\
\beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \beta_{32} x_i^3 + \epsilon_i & \text{if } x_i \geq c.
\end{cases}
\]

Constraints:

- \( \hat{f} \) is continuous
- \( \hat{f}', \hat{f}'', \ldots \) are continuous
FIGURE 7.3. Various piecewise polynomials are fit to a subset of the Wage data, with a knot at age=50. Top Left: The cubic polynomials are unconstrained. Top Right: The cubic polynomials are constrained to be continuous at age=50. Bottom Left: The cubic polynomials are constrained to be continuous, and to have continuous first and second derivatives. Bottom Right: A linear spline is shown, which is constrained to be continuous.
Figure 5. The top left panel shows a piecewise constant function fit to some artificial data. The broken vertical lines indicate the positions of the two knots $\xi_1$ and $\xi_2$. The blue curve represents the true function, from which the data were generated with Gaussian noise. The remaining two panels show piecewise linear functions fit to the same data—the top right unrestricted, and the lower left restricted to be continuous at the knots. The lower right panel shows a piecewise-linear basis function, $h_3(X) = (X - \xi_1)^+$, continuous at $\xi_1$. The black points indicate the sample evaluations $h_3(x_i), i = 1, \ldots, N$. 

Piecewise linear

![Diagram of piecewise linear functions and knots](image.png)
Piecewise cubic polynomials

FIGURE 5.2. A series of piecewise-cubic polynomials, with increasing orders of continuity.
7.4.1 – The spline basis representation

**General model:** Fit a piecewise degree $d$ polynomial under the constraint that its first $d − 1$ derivatives are continuous

**Cubic spline:** $K$ knots at $\xi_1, \ldots, \xi_K$

\[
y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+3} b_{K+3}(x_i) + \epsilon_i,
\]

**Truncated power basis:**

\[
h(x, \xi) = (x - \xi)_+^3 = \begin{cases} 
(x - \xi)^3 & \text{if } x > \xi \\
0 & \text{otherwise},
\end{cases}
\]

**Basis functions for cubic spline:**

\[
1, X, X^2, X^3, h(x, \xi_1), h(x, \xi_2), \ldots, h(x, \xi_K)
\]

**Coefficients:** $\beta_0, \ldots, \beta_{K+3}$, degree of freedom $= K + 4$
Natural cubic spline

**Natural cubic spline**: is a regression spline with additional boundary constraints: for example, the function is linear at the boundary

**Degree of freedom**: $K$

**Example:**

![Figure 7.4](image.png)

**FIGURE 7.4.** A cubic spline and a natural cubic spline, with three knots, fit to a subset of the Wage data.
7.4.4 – Choosing the number and locations of the knots

**Questions:**
- Where should we place the knots? – Adaptive methods
- How many knots should we use, or equivalently how many degrees of freedom should our spline contain? – Cross validation

**FIGURE 7.6.** Ten-fold cross-validated mean squared errors for selecting the degrees of freedom when fitting splines to the Wage data. The response is wage and the predictor age. Left: A natural cubic spline. Right: A cubic spline.


7.4.5 – Spline and polynomial regression

- Splines are more flexible and stable
- Polynomials may have the Runge phenomenon.

**FIGURE 7.7.** On the Wage data set, a natural cubic spline with 15 degrees of freedom is compared to a degree-15 polynomial. Polynomials can show wild behavior, especially near the tails.
Outline

1. 7.1 – Polynomial regression
2. 7.2 – Step functions
3. 7.3 – Basis functions
4. 7.4 – Splines
5. 7.5 – Smoothing splines
6. Multidimensional splines
Regularization

\[
\text{RSS}(f, \lambda) = \sum_{i=1}^{N} \{y_i - f(x_i)\}^2 + \lambda \int \{f''(t)\}^2 dt,
\]

\(\lambda\) : a smoothing parameter

- \(\lambda = 0\) : \(f\) can be any function that interpolates the data
- \(\lambda = \infty\) : will obtain a line such that \(f'' = 0\)

**Solution:** See Exercise 5.7 in the book “The elements of statistical learning”
Ex. 5.7 Derivation of smoothing splines (Green and Silverman, 1994). Suppose that \( N \geq 2 \), and that \( g \) is the natural cubic spline interpolant to the pairs \( \{x_i, z_i\}_{i=1}^N \), with \( a < x_1 < \cdots < x_N < b \). This is a natural spline with a knot at every \( x_i \); being an \( N \)-dimensional space of functions, we can determine the coefficients such that it interpolates the sequence \( z_i \) exactly. Let \( \tilde{g} \) be any other differentiable function on \([a, b]\) that interpolates the \( N \) pairs.

(a) Let \( h(x) = \tilde{g}(x) - g(x) \). Use integration by parts and the fact that \( g \) is a natural cubic spline to show that

\[
\int_a^b g''(x)h''(x) \, dx = -\sum_{j=1}^{N-1} g'''(x_j^+)[h(x_{j+1}) - h(x_j)] \tag{5.72}
\]

\( = 0 \).

(b) Hence show that

\[
\int_a^b \tilde{g}''(t)^2 \, dt \geq \int_a^b g''(t)^2 \, dt,
\]

and that equality can only hold if \( h \) is identically zero in \([a, b]\).

(c) Consider the penalized least squares problem

\[
\min_f \left[ \sum_{i=1}^N (y_i - f(x_i))^2 + \lambda \int_a^b f''(t)^2 \, dt \right].
\]

Use (b) to argue that the minimizer must be a cubic spline with knots at each of the \( x_i \).
How to find the cubic spline?

Basis expansion:

\[ f(x) = \sum_{j=1}^{N} N_j(x) \theta_j, \]

Define matrix: \( \{N\}_{ij} = N_j(x_i) \) and \( \{\Omega_N\}_{jk} = \int N_j''(t) N_k''(t) dt \)

\[ \text{RSS}(f, \lambda) = \sum_{i=1}^{N} \{y_i - f(x_i)\}^2 + \lambda \int \{f''(t)\}^2 dt, \]

is equivalent to

\[ \text{RSS}(\theta, \lambda) = (y - N\theta)^T (y - N\theta) + \lambda \theta^T \Omega_N \theta, \]

Solution:

\[ \hat{\theta} = (N^T N + \lambda \Omega_N)^{-1} N^T y, \]

Degree of freedom: \( = \) the number of coefficients
**FIGURE 5.6.** The response is the relative change in bone mineral density measured at the spine in adolescents, as a function of age. A separate smoothing spline was fit to the males and females, with $\lambda \approx 0.00022$. This choice corresponds to about 12 degrees of freedom.
7.5.2 – Choosing the smoothing parameter $\lambda$

$\hat{f}$: the $N$-vector of fitted values $\hat{f}(x_i)$ at the training points $\{x_i\}_{i=1}^N$

\[
\hat{f} = N(N^T N + \lambda \Omega_N)^{-1} N^T y
= S_\lambda y.
\]

**Smother matrix:** $S_\lambda$
FIGURE 5.8. The smoother matrix for a smoothing spline is nearly banded, indicating an equivalent kernel with local support. The left panel represents the elements of $S$ as an image. The right panel shows the equivalent kernel or weighting function in detail for the indicated rows.
How to choose $\lambda$?

Effective degree of freedom: the sum of diagonals of $S_\lambda$

$$df_\lambda = \text{trace}(S_\lambda),$$

Red: $df_\lambda = 5$; Green: $df_\lambda = 11$
Example

\[
Y = f(X) + \varepsilon,
\]
\[
f(X) = \frac{\sin(12(X + 0.2))}{X + 0.2},
\]
\[
X \sim U[0,1] \text{ and } \varepsilon \sim N(0,1)
\]
5.5 Automatic Selection of the Smoothing Parameters

$\text{EPE}(\lambda)$ and $\text{CV}(\lambda)$ curves for a realization from a nonlinear additive error model (5.22). The remaining panels show the data, the true functions (in purple), and the fitted curves (in green) with yellow shaded $\pm 2\times$ standard error bands, for three different values of $df_\lambda$.

**FIGURE 5.9.** The top left panel shows the EPE(\lambda) and CV(\lambda) curves for a realization from a nonlinear additive error model (5.22). The remaining panels show the data, the true functions (in purple), and the fitted curves (in green) with yellow shaded $\pm 2\times$ standard error bands, for three different values of $df_\lambda$. 

$df_\lambda = 5$

$df_\lambda = 9$

$df_\lambda = 15$
Outline

1. 7.1 – Polynomial regression
2. 7.2 – Step functions
3. 7.3 – Basis functions
4. 7.4 – Splines
5. 7.5 – Smoothing splines
6. Multidimensional splines
Tensor product basis

2D splines: $X \in \mathbb{R}^2$

- Basis functions of coordinate $X_1$: $h_{1,k}(X_1), k = 1, \ldots, M_1$
- Basis functions of coordinate $X_2$: $h_{2,k}(X_2), k = 1, \ldots, M_2$

Tensor product basis:

$$g_{jk}(X) = h_{1j}(X_1)h_{2k}(X_2), \ j = 1, \ldots, M_1, \ k = 1, \ldots, M_2$$

To represent functions:

$$g(X) = \sum_{j=1}^{M_1} \sum_{k=1}^{M_2} \theta_{jk} g_{jk}(X).$$
FIGURE 5.10. A tensor product basis of B-splines, showing some selected pairs. Each two-dimensional function is the tensor product of the corresponding one dimensional marginals.
Smoothing splines in two dimensions

\[
\min_f \sum_{i=1}^{N} \{ y_i - f(x_i) \}^2 + \lambda J[f],
\]

\[
J[f] = \int \int_{\mathbb{R}^2} \left[ \left( \frac{\partial^2 f(x)}{\partial x_1^2} \right)^2 + 2 \left( \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} \right)^2 + \left( \frac{\partial^2 f(x)}{\partial x_2^2} \right)^2 \right] dx_1 dx_2.
\]

- As \( \lambda \to 0 \), the solution approaches an interpolating function
- As \( \lambda \to \infty \), the solution approaches the least squares plane
Reference


**Chapter 5**: Trevor Hastie, Robert Tibshirani, The Elements of Statistical Learning, Second Edition.