Linear models for non-linear curves

Rasmus Waagepetersen
Department of Mathematics
Aalborg University
Denmark

October 10, 2016

Outline for today

what is a good model

BLUE and optimal prediction

linear models and basis functions

polynomial regression and overfitting

piecewise linear functions

What is a good statistical model ?

Desirable properties:

(a) accurate predictions

(b) good trade-off between fit of data and model complexity

(c) interpretable

(d) contains relevant variables for hypothesis testing/assessment of scientific question

(e) model assumptions valid

How to assess:
(a): cross validation.
(b): AIC/cross validation.
(c) and (d): qualitative assessment.
(d) depends on model. I.e. decomposition $Y = X\beta + \epsilon$ always valid. Assumptions to check are those imposed on $\epsilon = Y - X\beta$. E.g. $\epsilon \sim N(0, \sigma^2 I)$. For this, residuals $r = y - X\hat{\beta}$ obviously useful.

General linear model and BLUE

General linear model:

$$Y = X\beta + \epsilon$$

where X $n \times p$ fixed matrix, $\beta \in \mathbb{R}^p$ parameter vector, $\epsilon \in \mathbb{R}^n$ zero-mean random noise.

Let $L = \mathrm{col}(X)$. Assume $\mathbb{V}\mathrm{ar}\epsilon = \sigma^2 I$. Then least squares estimate $\hat{\mu} = p_L(y) = PY$ is BLUE (best linear unbiased estimate): $\mathbb{V}\mathrm{ar}(\tilde{\mu}) - \mathbb{V}\mathrm{ar}(\hat{\mu})$ is positive semi-definite for any other linear unbiased estimate $\tilde{\mu} = BY$ where $\mathbb{E}\tilde{\mu} = \mu$.

More generally, if $\psi = A\mu$ for some matrix A then $\hat{\psi} = A\hat{\mu} = APY$ is BLUE of ψ.

In particular (full rank X), $\hat{\beta} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}Y$ is BLUE.

Proof that $\hat{\psi} = APY$ is BLUE for $\psi = A\mu$:

Assume $\tilde{\psi}$ is LUE. I.e. $\tilde{\psi} = BY$ and $\mathbb{E}\tilde{\psi} = B\mu = A\mu$ for all $\mu \in L$. We also have $AP\mu = A\mu$ for all $\mu \in L$ (which implies that $\hat{\psi}$ is unbiased). Thus for all $w \in \mathbb{R}^p$

$$(B - AP)Pw = BPw - APw = APw - APw = 0$$

since $Pw \in L$. This implies $(B - AP)P = 0$ which gives

$$\mathbb{C}\mathrm{ov}(\tilde{\psi} - \hat{\psi}, \hat{\psi}) = \sigma^2(B - AP)P^T = 0$$

so that

$$\mathbb{V}\mathrm{ar}(\tilde{\psi}) = \mathbb{V}\mathrm{ar}(\tilde{\psi} - \hat{\psi}) + \mathbb{V}\mathrm{ar}\hat{\psi}$$

and the proof is completed.

Note: like Pythogoras if we say $\tilde{\psi} - \hat{\psi}$ and $\hat{\psi}$ orthogonal when their covariance is zero.

Optimal prediction revisited

X and Y random variables, g real function. General result

$$\mathbb{C}\mathrm{ov}(Y - \mathbb{E}[Y|X], g(X)) =$$
$$\mathbb{C}\mathrm{ov}(\mathbb{E}[Y - \mathbb{E}[Y|X]|X], \mathbb{E}[g(X)|X]) +$$
$$\mathbb{E}\mathbb{C}\mathrm{ov}(Y - \mathbb{E}[Y|X], g(X)|X) = 0$$

Note, since $\mathbb{E}[Y - \mathbb{E}[Y|X]] = 0$ we also have

$$\mathbb{E}(Y - \mathbb{E}[Y|X])g(X) = 0$$

In particular, for any prediction $\tilde{Y} = f(X)$ of Y:

$$\mathbb{E}\big[(Y - \mathbb{E}[Y|X])(E[Y|X] - f(X))\big] = 0$$

from which it follows that

$$\mathbb{E}(Y - \tilde{Y})^2 = \mathbb{E}(Y - \mathbb{E}[Y|X])^2 + \mathbb{E}(\mathbb{E}[Y|X] - \tilde{Y})^2$$

Let's consider EXY as inner product for random variables X and Y.

Pythagoras and conditional expectation

Space of real random variables with finite variance may be viewed as a vector space with inner product and (L_2) norm

$$< X, Y >= E(XY) \quad \|X\| = \sqrt{\mathbb{E}X^2}$$

Orthogonal decomposition (Pythagoras):

$$\|Y\|^2 = \|\mathbb{E}[Y|X]\|^2 + \|Y - \mathbb{E}[Y|X]\|^2$$

$\mathbb{E}[Y|X]$ may be viewed as projection of Y on X since it minimizes distance

$$\mathbb{E}(Y - \tilde{Y})^2$$

among all predictors $\tilde{Y} = f(X)$.

For zero-mean random variables, orthogonal is the same as uncorrelated.

(Grimmett & Stirzaker, Prob. and Random Processes, Chapter 7.9 good source on this perspective on prediction and conditional expectation)

Basis functions representations of unknown function

Suppose we are given measurements (x_i, y_i) where y_i are observations of Y_i with $\mathbb{E}Y_i = f(x_i)$ for some unknown function f.

Idea: represent $f(\cdot)$ as a linear combination of specified basis functions

$$f(x) = \sum_{i=0}^{p-1} \beta_i B_i(x)$$

Example (linear regression): $p = 2$, $B_0(x) = 1$, $B_1(x) = x$.

Polynomial regression: $B_i(x) = x^i$, $i = 0, \ldots, p-1$

Overfitting

Suppose we are given observations (x_i, y_i) $i = 1, \ldots, n$.

Then we can always find a nth order polynomial $\hat{f}(x)$ that fits exactly these observations - i.e. $y_i - \hat{f}(x_i) = 0$ for all i (Note: if design matrix $n \times n$ and full rank then $L = \mathbb{R}^n$ and $P = I$).

However, typically such a high order polynomial fits actual data "too well" - it fits not only f but also the noise.

This means fitted $\hat{f}$ bad for prediction of new observations.

Another problem: polynomials "global" - if just one (x_i, y_i) is changed this affects the whole fitted polynomial.

Piecewise linear function

A first approximation of f might be a linear regression $f(x) = a + bx$ but this is often too crude.

A next step might be a piecewise linear function f

$$f(x) = a_l + b_l(x - c_l), \quad x \in [c_l, c_{l+1}[$$

for some 'cut'-points or 'knots' c_l, $l = 1, \ldots, p$.

However, we typically want f to be continuous !

This is ensured if we require $a_l + b_l(c_{l+1} - c_l) = a_{l+1}$.

A continous piece-wise linear curve from c_0 to c_p is obtained with the following parametrization:

$$f(x) = \begin{cases} a_0 + b_0(x - c_0) & x \in [c_0, c_1] \\ f(c_1) + b_1(x - c_1) & x \in]c_1, c_2] \\ f(c_2) + b_2(x - c_2) & x \in]c_2, c_3] \\ \text{etc.} \end{cases}$$

This still defines a linear model !

Basis functions: $B_0(x) = 1$,

$$B_1(x) = \begin{cases} (x - c_0) & x \in [c_0, c_1] \\ (c_1 - c_0) & x > c_1 \end{cases} \quad B_2(x) = \begin{cases} 0 & x \in [c_0, c_1] \\ (x - c_1) & x \in]c_1, c_2] \\ (c_2 - c_1) & x \in]c_2, c_3] \end{cases}$$

Yet another basis: $B_0(x) = 1$, $B_1(x) = x$,
$B_2(x) = 1(x > c_2)(x - c_2)$, $B_3(x) = 1[x > c_3](x - c_3), \ldots$

Alternative basis functions for piecewise linear functions

Cut-points $c_0, c_1, \ldots, c_{p-1}, c_p$.

For $i = 1, \ldots, p$:

$$B_i(x) = \begin{cases} \frac{x - c_{i-1}}{c_i - c_{i-1}} & x \in [c_{i-1}, c_i[\\ 1 - \frac{x - c_i}{c_{i+1} - c_i} & x \in [c_i, c_{i+1}[\\ 0 & \text{otherwise} \end{cases}$$

Note: $f(x) = \sum_{i=0}^{p-1} \beta_i B_i(x)$ piecewise linear and continuous.

Hence we obtain exactly same set of functions as with basis on previous slide !

Note: new set of basis functions "local" - only non-zero on intervals $[c_{i-1}, c_{i+1}[$. Thereby more sparse $X^T X$ matrix.

Disadvantage: f above is not smooth at cutpoints.

B-spline basis functions

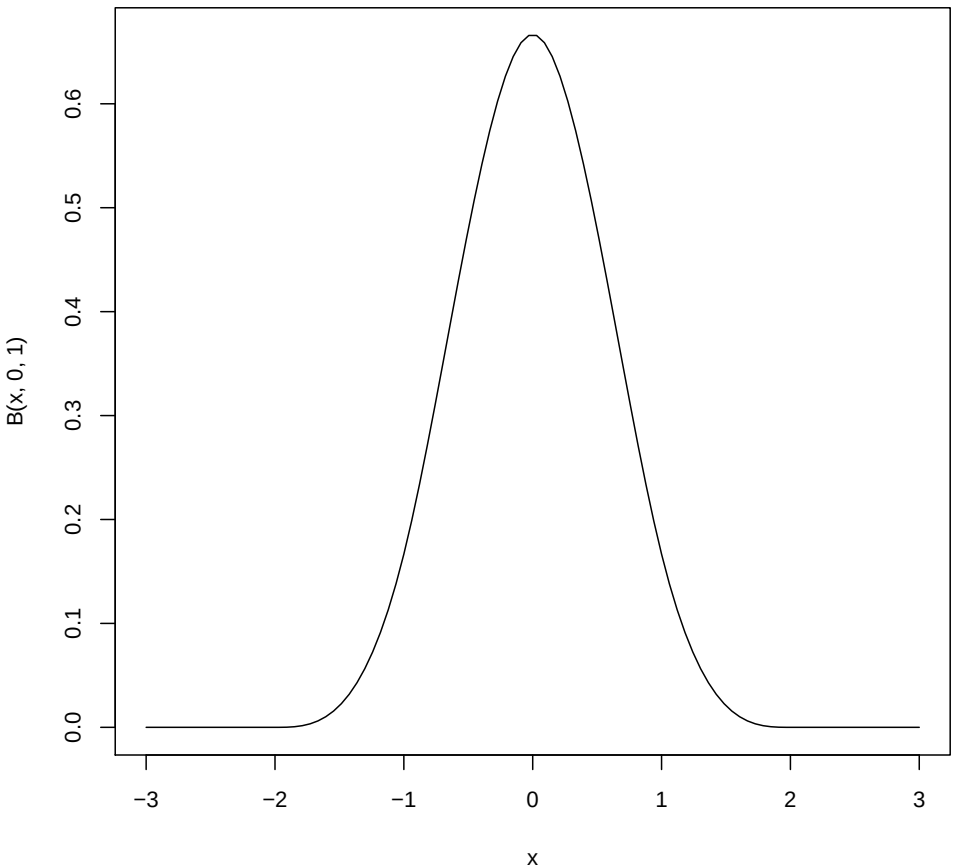
Consider doubly infinite sequence of equi-distant cut points $\ldots, c_{-1}, c_0, c_1, c_2, \ldots$ with (wlog) $c_{i+1} - c_i = 1$.

Define $B_i(x) = B(x - c_i)$ where

$$B(x) = \begin{cases} \frac{1}{6}(x+2)^3 & x \in [-2, -1[\\ \frac{1}{6}(1 + 3(x+1) + 3(x+1)^2 - 3(x+1)^3) & x \in [-1, 0[\\ \frac{1}{6}(4 - 6x^2 + 3x^3) & x \in [0, 1[\\ \frac{1}{6}(1 - 3(x-1) + 3(x-1)^2 - (x-1)^3)) & x \in [1, 2[\\ 0 \text{ otherwise} \end{cases}$$

$B(x)$ is a cubic spline: composed of the constant function $g(x) = 0$ and 4 third-order polynomials such that it is everywhere continuous and twice-differentiable.

The B-spline basis function:



Cubic spline

$$f(x) = f_i(x) = a_{i0} + a_{i1}(x - c_i) + a_{i2}(x - c_i)^2 + a_{i3}(x - c_i)^3 \quad x \in [c_i, c_{i+1}[$$

Require continuity

$$f_i(c_{i+1}) = f_{i+1}(c_{i+1})$$

and twice differentiability:

$$f_i'(c_{i+1}) = f_{i+1}'(c_{i+1}) \quad f_i''(c_{i+1}) = f_{i+1}''(c_{i+1})$$

Again possible to compute basis functions and fit model in `R`.

`R` Function `bs()` can be used to generate required basis functions for linear model.

Suppose we use cut-points/knots $c_1, \ldots, c_q$ and the $q - 1$ associated cubic polynomials. Then we have
$p = (q - 1) * 4 - 3 * (q - 2) = q + 2$ free parameters.

Equivalence of bases for cubic splines

Fitting a cubic spline with knots $0, 1, \ldots, q-1$ (starting at $c_1 = 0$ and ending at $c_q = q-1$) is equivalent to fitting the linear model based on the B-spline basis functions $B(x-i)$, $i = -1, \ldots, q$.

Note: same number of free parameters.

Intuitively makes sense, since both models generate continous piecewise cubic splines with continuous first and second derivatives.

Exercises

1. Write down the design matrix for a piece-wise linear regression model with cut-points c_1 and c_2 (i.e. the curve is composed of three segments).

2. Implement in `R` the above piece-wise model for the wind/power data. Try also the 'local' basis functions.

3. Write down the equations for a cubic spline with knots $0, 1, 2$ starting at 0 and ending at 2. Write down the associated design matrix.

4. Fit a cubic spline to the wind/power data (use your own basis or the `R`-function `bs()` with different choices of knots).