Tutorial 8: suggested solutions

1. suppose we have sample $(\mathbf{x}_i, y_i), i = 1, ..., n$. we estimate the regression model $y_i = g(\mathbf{x}_i) + \varepsilon_i$, where $g(x)$ is a spline function of the form

$$g(x) = \sum_{j=1}^{J+4} \beta_j B_j(x)$$

(a) Estimate the derivative $g'(x)$ of $g(x)$

(b) find the 95% confidence band for $g'(x)$.

Let

$$\mathbf{X} = \begin{pmatrix} B_1(\mathbf{x}_1) & B_2(\mathbf{x}_1) & ... & B_{J+4}(\mathbf{x}_1) \\ B_1(\mathbf{x}_2) & B_2(\mathbf{x}_2) & ... & B_{J+4}(\mathbf{x}_2) \\ ... \\ B_1(\mathbf{x}_1) & B_2(\mathbf{x}_1) & ... & B_{J+4}(\mathbf{x}_1) \end{pmatrix} \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ ... \\ Y_n \end{pmatrix}$$

and

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ ... \\ \hat{\beta}_n \end{pmatrix} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

The estimated funciton is

$$\hat{g}(x) = \sum_{j=1}^{J+4} \hat{\beta}_j B_j(x)$$

(a) $g'(x)$ can be estimated by

$$\hat{g}'(x) = \sum_{j=1}^{J+4} \hat{\beta}_j B_j'(x) = \sum_{j=2}^{J+4} \hat{\beta}_j B_j'(x)$$

where

$$B_1(x) = 1, B_2(x) = x, B_3(x) = x^2, B_4(x) = x^3, B_{j+4}'(x) = (x - t_j)_+^3, j = 1, ..., J$$

thus

$$B_2'(x) = 1, B_3'(x) = 2x, B_4'(x) = 3x^2, B_{j+4}'(x) = 3(x - t_j)_+^2, j = 1, ..., J$$

(b) Because

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ ... \\ \hat{\beta}_{J+4} \end{pmatrix} - \beta \sim N(0, (\mathbf{X}^\top \mathbf{X})^{-1} \sigma^2)$$

we have

$$\hat{g}'(x) - g'(x) \sim N(0, \ell^\top(x) (\mathbf{X}^\top \mathbf{X})^{-1} \sigma^2 \ell(x)$$

where $\ell(x) = [0, 1, 2x, 3x^2, 3(x - t_1)_+^2, ..., 3(x - t_1)_+^2]$. The 95% confidence band is

$$\hat{g}'(x) \pm 1.96 \ell(x) (\mathbf{X}^\top \mathbf{X})^{-1} \sigma^2 \ell(x).$$

2. *Suppose we have sample* $(\mathbf{x}_i, \mathbf{z}_i, y_i), i = 1, ..., n$. *we estimate the regression model* $y_i = \alpha_0 + \alpha_1 \mathbf{z}_i + g(\mathbf{x}_i) + \varepsilon_i$ *and* $g(x)$ *is a spline function of the form*

$$g(x) = \sum_{j=1}^{J+4} \beta_j B_j(x)$$

   (a) *write the expression for the estimator of* $g(x)$

   (b) *find the 95% confidence band for* $g(x)$.

   (a) let

$$\mathbf{X} = \begin{pmatrix} 1 & \mathbf{z}_1 & B_2(\mathbf{x}_1) & ... & B_{J+4}(\mathbf{x}_1) \\ 1 & \mathbf{z}_2 & B_2(\mathbf{x}_2) & ... & B_{J+4}(\mathbf{x}_2) \\ ... \\ 1 & \mathbf{z}_n & B_2(\mathbf{x}_n) & ... & B_{J+4}(\mathbf{x}_n) \end{pmatrix} \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ ... \\ Y_{J+4} \end{pmatrix}$$

   and

$$\begin{pmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \\ \hat{\beta}_2 \\ ... \\ \hat{\beta}_n \end{pmatrix} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

   The estimated g is

$$g(x) = \sum_{j=2}^{J+4} \hat{\beta}_j B_j(x)$$

   (with a constant difference)

   (b) the 95% confidence band for $g(x)$.

$$\hat{g}(x) \pm 1.96 \ell(x)^\top (\mathbf{X}^\top \mathbf{X})^{-1} \sigma^2 \ell(x)$$

   where $\ell(x) = [0, 0, B_2(x), ..., B_{J+4}(x)]^\top$.

3. In the first question, if $\mathbf{x}_i$ just takes m ($< n$) values. what is the maximum number of knots you can use to estimate the regression function g.

   suggested $J < m - 4$

4. Consider **data A** (the first column is Y and the others X). fit a "best" generalized additive model to the data and plot the estimated functions with 2 times standard error confidence bands. predict **data B**.

   (codeDM7t1.R)