STATS 790 Assignment 3

Yiran Zhang
400119421

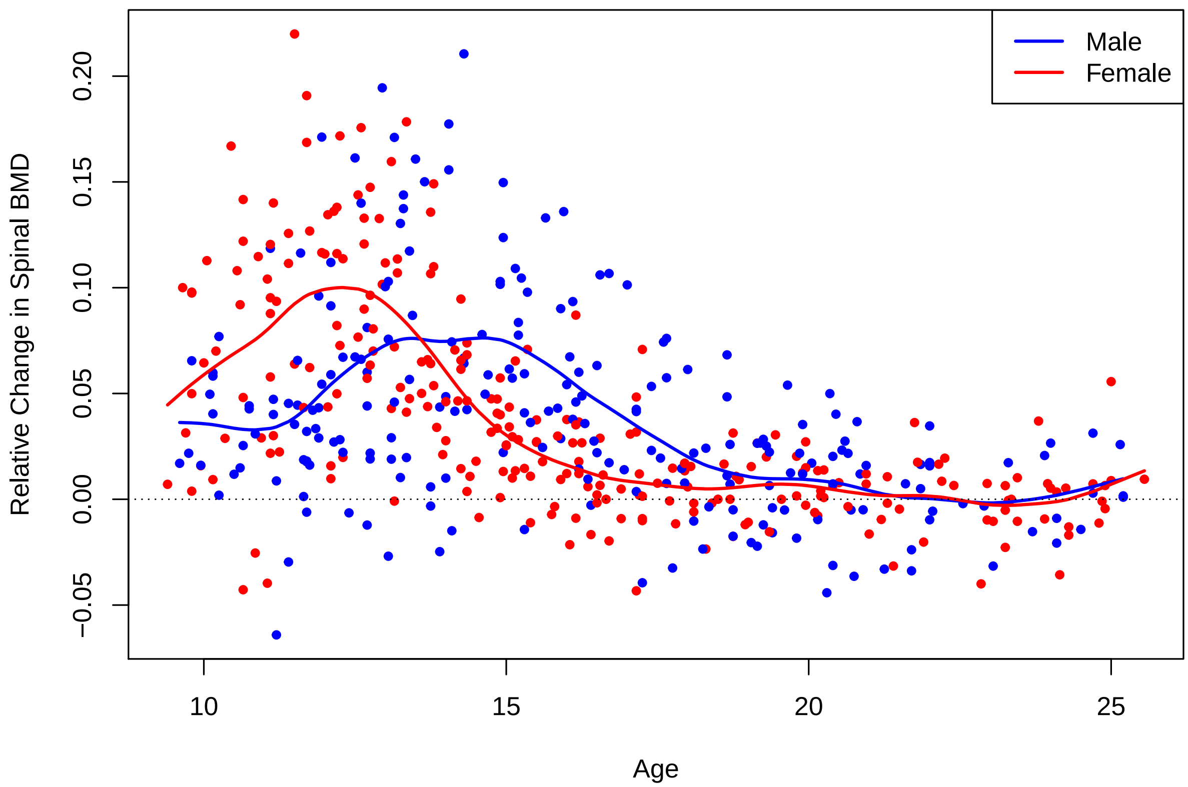
27 February, 2023

Question 1

ESL Chapter 5 Figure 5.6 is replicated, the code is shown as below.

```
# Question 1
# Replicate ESL Chapter 5 Figure 5.6

# Import dataset
bone <- read.table('https://hastie.su.domains/ElemStatLearn/datasets/bone.data',
                   header = TRUE)

# According to ESL Figure 5.6 description, spnbmd is the target variable
# and age is the predictor.

# Plot the age against relative change in spinal BMD, color separated by gender.
plot(x = bone$age, y = bone$spnbmd,
     col = ifelse(bone$gender == 'female','red','blue'),
     pch = 20, xlab = 'Age', ylab = 'Relative Change in Spinal BMD')

# Split male and female.
male_bone <- bone[bone$gender == 'male', ]
female_bone <- bone[bone$gender == 'female', ]

# Spline, using degree of freedom = 12 (given by the textbook)
male_spline <- smooth.spline(x = male_bone$age, y = male_bone$spnbmd, df = 12)
female_spline <- smooth.spline(x = female_bone$age, y = female_bone$spnbmd, df = 12)
```

```r
# Add splines to the graph with corresponding color for each gender.
lines(male_spline, col = 'blue', lwd = 2)
lines(female_spline, col = 'red', lwd = 2)

# Add a horizontal dash line at y = 0.
abline(h=0, lty=3)

# Add a legend at the top right corner.
legend(x='topright', legend=c('Male', 'Female'),
       col=c('blue', 'red'), lwd=2)
```



Question 2 (South Africa coronary heart disease data)

```r
# Import South Africa coronary heart disease data.
url <- "http://www-stat.stanford.edu/~tibs/ElemStatLearn/datasets/SAheart.data"
```

```r
heart <- read.csv(url, row.names = 1)
```

Question 3 (..)

Question 4 (..)

Question 5

ESL 5.4 The natural boundary conditions for natural cubic splines is that "the function is linear beyond the boundary knots". When X $< \xi_1$, f(x) $= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$, to make it linear, we need β_2 and β_3 equal to 0. Alternatively, we can prove by taking the second derivative of f(x) and set it to 0. $f''(x) = 2\beta_2 + 6\beta_3 x = 0$, hence $\beta_2 = \beta_3 = 0$.

Similarly, the function f(x) should be linear when X $> \xi_k$ where f(x) $= \beta_0 + \beta_1 x + \sum_{k=1}^{K} \theta_k (X - \xi_k)^3$, we take second derivative and set it to 0. $f''(x) = 6\sum_{k=1}^{K} \theta_k (X - \xi_k) = 6(\sum_{k=1}^{K} \theta_k X - \sum_{k=1}^{K} \theta_k \xi_k) = 0$, then we have $\sum_{k=1}^{K} \theta_k X - \sum_{k=1}^{K} \theta_k \xi_k = 0$, which implies that $\sum_{k=1}^{K} \theta_k = 0$ and $\sum_{k=1}^{K} \theta_k \xi_k = 0$, as required.

Now we derive (5.4) and (5.5). The function we have is $f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} \theta_k (X - \xi_k)_+^3 = 0$, by observing the function, we get that $N_1(X) = 1, N_2(X) = X$, we now want to prove that $\sum_{k=1}^{K} \theta_k (X - \xi_k)_+^3$ can be written is the form of $N_{k+2}(X) = d_k(X) - d_{k-1}(X)$.

Given that $\sum_{k=1}^{K} \theta_k = 0$ and $\sum_{k=1}^{K} \theta_k \xi_k = 0$, we know that $\sum_{k=1}^{K-2} \theta_k = -\theta_K - \theta_{K-1}$ and $\sum_{k=1}^{K-2} \theta_k \xi_k = -\theta_K \xi_K - \theta_{K-1} \xi_{K-1}$.

$\sum_{k=1}^{K} \theta_k (X - \xi_k)_+^3 = \sum_{k=1}^{K-2} \theta_k (X - \xi_k)_+^3 + \theta_{K-1} (X - \xi_{K-1})_+^3 + \theta_K (X - \xi_K)_+^3$

$$
\begin{aligned}
\theta_{K-1}(X - \xi_{K-1})_+^3 &= \theta_{K-1}(X - \xi_{K-1})_+^3 \frac{\xi_{K-1} - \xi_K}{\xi_{K-1} - \xi_K} \\
&= \frac{(X - \xi_{K-1})_+^3}{\xi_{K-1} - \xi_K} (\theta_{K-1}\xi_{K-1} - \theta_{K-1}\xi_K) \\
&= \frac{(X - \xi_{K-1})_+^3}{\xi_{K-1} - \xi_K} (\theta_{K-1}\xi_{K-1} - \theta_{K-1}\xi_K + \theta_K \xi_K - \theta_K \xi_K) \\
&= \frac{(X - \xi_{K-1})_+^3}{\xi_{K-1} - \xi_K} (\theta_{K-1}\xi_{K-1} + \theta_K \xi_K - \theta_{K-1}\xi_K - \theta_K \xi_K) \\
&= \frac{(X - \xi_{K-1})_+^3}{\xi_{K-1} - \xi_K} (-\sum_{k=1}^{K-2} \theta_k \xi_k + \xi_K \sum_{k=1}^{K-1} \theta_k) \\
&= \frac{(X - \xi_{K-1})_+^3}{\xi_{K-1} - \xi_K} \sum_{k=1}^{K-2} \theta_k (\xi_K - \xi_k) \\
&= \sum_{k=1}^{K-2} \theta_k (\xi_K - \xi_k) \frac{(X - \xi_{K-1})_+^3}{\xi_{K-1} - \xi_K}
\end{aligned}
$$

Similarly,

$$\begin{aligned}
\theta_K(X-\xi_K)^3_+ &= \frac{(X-\xi_K)^3_+}{\xi_{K-1}-\xi_K}\theta_K(\xi_{K-1}-\xi_K) \\
&= \frac{(X-\xi_K)^3_+}{\xi_{K-1}-\xi_K}(\theta_K\xi_{K-1}-\theta_K\xi_K+\theta_{K-1}\xi_{K-1}-\theta_{K-1}\xi_{K-1}) \\
&= \frac{(X-\xi_K)^3_+}{\xi_{K-1}-\xi_K}(-\xi_{K-1}\sum_{k=1}^{K-2}\theta_K+\sum_{k=1}^{K-2}\theta_k\xi_k) \\
&= \sum_{k=1}^{K-2}\theta_k(\xi_k-\xi_{K-1})\frac{(X-\xi_K)^3_+}{\xi_{K-1}-\xi_K}
\end{aligned}$$

Put them back together and get:

$$\begin{aligned}
\sum_{k=1}^{K}\theta_k(X-\xi_k)^3_+ &= \sum_{k=1}^{K-2}\theta_k(X-\xi_k)^3_+ + \theta_{K-1}(X-\xi_{K-1})^3_+ + \theta_K(X-\xi_K)^3_+ \\
&= \sum_{k=1}^{K-2}\theta_k(X-\xi_k)^3_+ + \sum_{k=1}^{K-2}\theta_k(\xi_K-\xi_k)\frac{(X-\xi_{K-1})^3_+}{\xi_{K-1}-\xi_K} + \sum_{k=1}^{K-2}\theta_k(\xi_k-\xi_{K-1})\frac{(X-\xi_K)^3_+}{\xi_{K-1}-\xi_K} \\
&= \sum_{k=1}^{K-2}\theta_k[(X-\xi_k)^3_+ + (\xi_K-\xi_k)\frac{(X-\xi_{K-1})^3_+}{\xi_{K-1}-\xi_K} + (\xi_k-\xi_{K-1})\frac{(X-\xi_K)^3_+}{\xi_{K-1}-\xi_K}] \\
&= \sum_{k=1}^{K-2}\theta_k[(X-\xi_k)^3_+\frac{\xi_K-\xi_k}{\xi_K-\xi_k} + (\xi_K-\xi_k)\frac{(X-\xi_{K-1})^3_+}{\xi_{K-1}-\xi_K} + (\xi_k-\xi_{K-1})\frac{(X-\xi_K)^3_+}{\xi_{K-1}-\xi_K}\frac{\xi_K-\xi_k}{\xi_K-\xi_k}] \\
&= \sum_{k=1}^{K-2}\theta_k(\xi_K-\xi_k)[\frac{(X-\xi_k)^3_+}{\xi_K-\xi_k} + \frac{(X-\xi_{K-1})^3_+}{\xi_{K-1}-\xi_K} + (\xi_k-\xi_{K-1})\frac{(X-\xi_K)^3_+}{(\xi_{K-1}-\xi_K)(\xi_K-\xi_k)}] \\
&= \sum_{k=1}^{K-2}\theta_k(\xi_K-\xi_k)[\frac{(X-\xi_k)^3_+}{\xi_K-\xi_k} + \frac{(X-\xi_{K-1})^3_+}{\xi_{K-1}-\xi_K} + (X-\xi_K)^3_+(\frac{-1}{\xi_{K-1}-\xi_K} - \frac{1}{\xi_K-\xi_k})] \\
&= \sum_{k=1}^{K-2}\theta_k(\xi_K-\xi_k)[\frac{(X-\xi_k)^3_+-(X-\xi_K)^3_+}{\xi_K-\xi_k} + \frac{(X-\xi_{K-1})^3_+-(X-\xi_K)^3_+}{\xi_{K-1}-\xi_K}] \\
&= \sum_{k=1}^{K-2}\theta_k(\xi_K-\xi_k)[\frac{(X-\xi_k)^3_+-(X-\xi_K)^3_+}{\xi_K-\xi_k} - \frac{(X-\xi_{K-1})^3_+-(X-\xi_K)^3_+}{\xi_K-\xi_{K-1}}] \\
&= \sum_{k=1}^{K-2}\theta_k(\xi_K-\xi_k)(d_k(X)-d_{K-1}(X)) \\
&= \sum_{k=1}^{K-2}\theta_k(\xi_K-\xi_k)N_{k+2}(X)
\end{aligned}$$

Therefore, to conclude, we get $N_1(X)=1, N_2(X)=X, N_{k+2}(X)=d_k(X)-d_{K-1}(X)$ as desired in (5.4) and (5.5).

ESL 5.13