Spline Regression

Suppose we are given real numbers $t_1, \ldots, t_n$ on some interval $[a, b]$, satisfying $a < t_1 < t_2 < \ldots < t_n < b$. A function g defined on $[a, b]$ is a *cubic spline* if two conditions are satisfied.

Firstly, on each of the intervals $(a, t_1), (t_1. t_2), (t_2, t_3), \ldots, (t_n, b)$, g is a cubic polynomial;

Secondly the polynomial pieces fit together at the points t_i in such a way that ***g*** **itself and its first and second derivatives are continuous at each** t_i, and hence on the whole of $[a, b]$. The points t_i are called *knots*.

$$g(t) = d_i(t- t_i)^3 + c_i(t- t_i)^2 + b_i(t- t_i) + a_i \text{ for } t_i \leq t \leq t_{i+1}$$

A cubic spline on an interval $[a, b]$ will be said to be a ***natural cubic spline*** (NCS) **if its second and third derivatives are zero at** ***a*** **and** ***b***. These conditions are called the *natural boundary conditions*.

They imply that $d_0 = c_0 = d_n = c_n = 0$, so that g is linear on the two extreme intervals $[a, t_1]$ and $[t_n, b]$. (Green, 1993)

For the illustration we will use the data set “triceps” which is included in the R package MultiKink.

The data are derived from an anthropometric study of 892 females under 50 years in three Gambian villages in West Africa.

There were 892 observations on the following 3 variables.

1. age: Age of respondents.

2. lntriceps: Log of the triceps skinfold thickness.

3. triceps: Triceps skinfold thichness

Truncated power funcation and spline

When use splines, we need to partition the range of the continuous covariate x into smaller intervals, and those partition points are called knots.

For piecewise continuous model (the lines connected between different intervals) we usually use the truncated power functions to fit the data. The truncated power function is defined as:

$$(x-k_i)_+^d = \begin{cases} 0, & \text{if } x < k_i \\ (x-k_i)^d, & \text{if } x \geq k_i \end{cases} \qquad (1)$$

where d=1,2,3,..

Using above truncated power function, we can specify a cubic spline regression model with five knots by the following formula:

$$y=\beta_0+\beta_1 x+\beta_2 x^2+\beta_3 x^3+\beta_4(x-k_1)_+^3+\beta_5(x-k_2)_+^3+\beta_6(x-k_3)_+^3+$$
$$\beta_7(x-k_4)_+^3+\beta_8(x-k_5)_+^3+\varepsilon \qquad (2)$$

$$y=\beta_0+\beta_1 x+\beta_2 x^2+\beta_3 x^3+\beta_4 (x-k_1)_+^3+\beta_5 (x-k_2)_+^3+\beta_6 (x-k_3)_+^3+$$
$$\beta_7 (x-k_4)_+^3+\beta_8 (x-k_5)_+^3+\varepsilon \qquad (2)$$

Note, there is no 'natural' or 'restricted' mentioned for above model.

It is just a cubic spline regression with continuous connection at each knot.

"A natural cubic splines adds additional constraints, namely that function is linear beyond the boundary knots."

Splines with this additional constraint are also known as "restricted" splines.

We will discuss what is exactly meaning for the 'natural' and 'restricted' mathematically later.

Conduct an “un-natural” cubic spline regression

Let us use the euqation (2) and the data set tricepts to conduct an ‘un-nutural’ cubic spline regression, i.e. there are no any restrictions on the equation (2)

Scatter plot between age and triceps. Conduct an un-natural cubic spline regression with knots at age 5, 10, 20, 30 and 40.

Note this is just a usual linear regression with a x value transformed by truncated power power function at different intervals.

We can see before the first knot (age<5) and after the last knot (age>40) the regression line is quite wiggly, and people think this is “un-natural”.

By the way, the term of ‘natural’ spline may come from beam models in construction.

Conduct a 'half-natural' cubic spline regression

For a 'full' natural cubic spline regression, we will force the function linear beyond the boundary knots, i.e. we will force the regression line before the first knot (age=5) and after the last knot (age=40) linear.

For the 'half natural cubic spline' I mean we can just force one tail linear, for a easier part, we just force the regression line before knot (age=5) linear.

We can do this just by dropping off the x^2 and x^3 terms in the equation (2), then the regression model becomes

$$y = \beta_0 + \beta_1 x + \beta_4 (x - k_1)_+^3 + \beta_5 (x - k_2)_+^3 + \beta_6 (x - k_3)_+^3 + \beta_7 (x - k_4)_+^3 + \beta_8 (x - k_5)_+^3 + \varepsilon \quad (3)$$

We use the following R code to conduct the 'half natural' cubic spline regression, just by dropping off the x^2 and x^3 terms in the model (2).

Compare the ‘un-natural’ spline regression with the ‘half-natural’ spline regession

Here, we can see the ‘half-natural’ cubic spline regression (orange) and the ‘un-natural’ cubic spline regression (blue) are only quite different before the knot (age=5), since we have forced the function before the knot (age=5) linear for the ‘half-natural’ cubic spline regression.

Conduct a ‘full-natural’ cubic spline regression

Here, the ‘full-natural’ cubic spline regression just means the natural cubic spline regression, I give the name “full-natural’ just for comparing with”un-natural” and “half-natural” spline regressions above.

We use “rcs” function from the rms package to conduct the natural spline regression

Next, we put the “un-nature”, “half-nature” and “full-nature” cubic spline regression together.

We can see for the full-natural cubic spline regression (red), after the last knot (age=40), the line is linear however the un-natural and half-natural are not.

Next we can get the final model function form by the following r code

Finally, our final natural cubic spline regression model can be written as E(triceps)=$X\beta$, where

$X\hat{\beta}$=8.657058−0.3043198age+0.005997174 $(age-5)_+^3$ − 0.01065974 $(age-10)_+^3$+ 0.006292726$(age-20)_+^3$ −0.001596325$(age-30)_+^3$ − 3.382968×10^{-5} $(age-40)_+^3$. (4)

$(x)_+$=x if x>0, 0 otherwise

Note, the above equation is produced by the function.

Noted, our final result, the equation (4), seems quite different from the below natural cubic spline regression output except the intercept and the coefficient for the age variable. What are the age', age'' and age'''?

To understand the output of the natural cubic spline regression from the rms package, we need to do some mathematical derivations and conduct some "hand calculations"

Mathematical derivations of the natual cubic splines

First let us write down a general form of cubic splines regression model with K knots (we do not count two ends of the line as knots).

$$f(x) = \sum_{j=0}^{3} \beta_j x^j + \sum_{k=1}^{K} \theta_k (x - \xi_k)_+^3 . \quad (5)$$

This is what I called "un-natural" spline regression model, since there were no any restrictions on the formula (5).

For the natural spline we need to force the line before **the first knot** and line after **the last knot** linear (i.e. straight), these restrictions imply the following relationship:

For the first restriction, when x is less than **the first knot**, we need:

$$\beta_2 = 0 \text{ and } \beta_3 = 0. \quad (6)$$

For the second restriction, when x is greater than the **last knot**, we also need

$$\sum_{k=1}^{K} \theta_k = 0 \; and \sum_{k=1}^{K} \theta_k \xi_k = 0. \quad (7)$$

The proofs of the above relationship are not very difficult.

From the equation (5) ($f(x) = \sum_{j=0}^{3} \beta_j x^j + \sum_{k=1}^{K} \theta_k (x - \xi_k)_+^3$) when $x<\xi_1$ we need the $f(x)$ be linear, note, the 'linear' only means $f(x)$ can be expressed as some combinations of x at power 0 or 1, i.e. there were no x^2, x^3…, noted sometimes linear and non-linear can be in term of regression coefficients, here the 'linear' is just in term of x.

When $x<\xi_1$ we have

$$f(x) = \sum_{j=0}^{3} \beta_j x^j = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3. \quad (8)$$

therefore, by the linearity, both β_2 and β_3 have to be equal to 0.

Method 2.

Another way to prove the condition is to use some calculus properties, since $f(x)$ needs to be linear in term of x, therefore, the first derivative of $f(x)$ will be a constant and the second derivative of $f(x)$ will be 0, i.e if $f(x)=kx+b$ then $f'(x)=k$ and $f''(x)=0$

We can set $f''(x)=0$ then we obtain $2\beta_2+6\beta_3 x=0$ for any $x<\xi_1$, therefore, β_2 and β_3 are both equal to 0.

For the second restriction, when $x>\xi_K$, i.e. when x is greater than **the last knot**, we can drop the + sign for the truncated function since when $x>k_i$, $(x-k_i)_+^d=(x-k_i)^d$ then $f(x)$ can be written as:

$$f(x)=\sum_{j=0}^{3}\beta_j x^j+\sum_{k=1}^{K}\theta_k(x-\xi_k)^3\,. \quad (9)$$

For the proof of the last restriction, we can tediously expand $f(x)$ and drop x^2 and x^3 terms and show $\sum_{k=1}^{K} \theta_k = 0$ and $\sum_{k=1}^{K} \xi_k \theta_k = 0$.

Here, we will use the derivative method since it is a little bit concise.

$$f''(x) = 6 \sum_{k=1}^{K} \theta_k (x - \xi_k) = 0 \Longrightarrow$$

$$\left(\sum_{k=1}^{K} \theta_k \right) x - \sum_{k=1}^{K} \theta_k \xi_k = 0, \forall x > \xi_K$$

Since $\forall x$ that $x > \xi_K$, we have $f''(x) = 0$, therefore, both $\sum_{k=1}^{K} \theta_k$ and $\sum_{k=1}^{K} \xi_k \theta_k$ have to be 0. i.e.

$$\sum_{k=1}^{K} \theta_k = 0 \; and \sum_{k=1}^{K} \xi_k \theta_k = 0$$

Because we put restrictions (6) (β_2=0 and β_3=0) and (7) ($\sum_{k=1}^{K}\theta_k = 0$ *and* $\sum_{k=1}^{K}\xi_k\theta_k = 0$) on the model (5) ($f(x) = \sum_{j=0}^{3}\beta_j x^j + \sum_{k=1}^{K}\theta_k(x-\xi_k)_+^3$) we have to develop a new representation of the linear regression model.

Next, show to obtain these spline "basis functions" for the natural cubic spline regression.

Let us see what will happen when we put restrictions (6) and (7) on (5)

Because of (6), we can easily see the first part of (5) becomes $\beta_0+\beta_1 x$

Let see what will happen when we put the restriction (7) on (5)

$$\sum_{k=1}^{K}\theta_k(x-\xi_k)_+^3 = \sum_{k=1}^{K-1}\theta_k(x-\xi_k)_+^3 + \theta_K(x-\xi_K)_+^3. \quad (10)$$

$$\sum_{k=1}^{K} \theta_k (x - \xi_k)_+^3 = \sum_{k=1}^{K-1} \theta_k (x - \xi_k)_+^3 + \theta_K (x - \xi_K)_+^3. \quad (10)$$

From $\sum_{k=1}^{K} \theta_k = 0$ in (7) we can infer that

$$\sum_{k=1}^{K-1} \theta_k + \theta_K = 0 \Longrightarrow \theta_K = -\sum_{k=1}^{K-1} \theta_k$$

We can replace θ_K in (10) with $-\sum_{k=1}^{K-1} \theta_k$ and we get:

$$\sum_{k=1}^{K} \theta_k (x - \xi_k)_+^3 = \sum_{k=1}^{K-1} \theta_k (x - \xi_k)_+^3 - \sum_{k=1}^{K-1} \theta_k (x - \xi_K)_+^3$$

$$= \sum_{k=1}^{K-1} \theta_k [(x - \xi_k)_+^3 - (x - \xi_K)_+^3]. \quad (11)$$

Notice, for the restriction (7) we have not use the condition $\sum_{k=1}^{K}\xi_k\theta_k = 0$ yet, now we will use this condition.

$$\sum_{k=1}^{K}\theta_k = 0 \; and \sum_{k=1}^{K}\theta_k\xi_k = 0 \Rightarrow \left(\sum_{k=1}^{K}\theta_k\right)\xi_K - \sum_{k=1}^{K}\theta_k\xi_k = 0. \quad (12)$$

note what we did here is just $0*\xi_K - 0 = 0$. From the (12)

$$\left(\sum_{k=1}^{K}\theta_k\right)\xi_K - \sum_{k=1}^{K}\theta_k\xi_k = 0 \Rightarrow \sum_{k=1}^{K}\theta_k(\xi_K - \xi_k) = 0$$

$$\Rightarrow \sum_{k=1}^{K-1}\theta_k(\xi_K - \xi_k) + \theta_K(\xi_K - \xi_K) = 0 \Rightarrow \sum_{k=1}^{K-1}\theta_k(\xi_K - \xi_k) = 0$$

$$\Rightarrow \sum_{k=1}^{K-2}\theta_k(\xi_K - \xi_k) + \theta_{K-1}(\xi_K - \xi_{K-1}) = 0$$

$$\Rightarrow \theta_{K-1} = -\sum_{k=1}^{K-2}\theta_k\frac{\xi_K - \xi_k}{\xi_K - \xi_{K-1}}. \quad (13)$$

Combine the (13) ($\theta_{K-1} = -\sum_{k=1}^{K-2} \theta_k \frac{\xi_K - \xi_k}{\xi_K - \xi_{K-1}}$) and (11)

($\sum_{k=1}^{K} \theta_k (x-\xi_k)_+^3 = \sum_{k=1}^{K-1} \theta_k [(x-\xi_k)_+^3 - (x-\xi_K)_+^3]$)

$$\sum_{k=1}^{K} \theta_k (x-\xi_k)_+^3 = \sum_{k=1}^{K-1} \theta_k [(x-\xi_k)_+^3 - (x-\xi_K)_+^3]$$

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

(replace θ_{K-1} from the equation (13))

$$= \sum_{k=1}^{K-2} \theta_k [(x-\xi_k)_+^3 - (x-\xi_K)_+^3]$$

$$- \sum_{k=1}^{K-2} \theta_k \frac{\xi_K - \xi_k}{\xi_K - \xi_{K-1}} [(x-\xi_{K-1})_+^3 - (x-\xi_K)_+^3]. \quad (14)$$

Define a function

$$d_k(x) = \frac{(x-\xi_k)_+^3 - (x-\xi_K)_+^3}{\xi_K - \xi_{K-1}}. \quad (15)$$

From the equation (15) we can get

$$d_{K-1}(x) = \frac{(x-\xi_{K-1})_+^3 - (x-\xi_K)_+^3}{\xi_K - \xi_{K-1}}. \quad (16)$$

Replace truncated functions in (14) by $d_k(x)$ and $d_{K-1}(x)$ and we get

$$\sum_{k=1}^{K} \theta_k (x-\xi_k)_+^3 = \sum_{k=1}^{K-1} \theta_k [(x-\xi_k)_+^3 - (x-\xi_K)_+^3]$$

$$= \sum_{k=1}^{K-2} \theta_k [(x-\xi_k)_+^3 - (x-\xi_K)_+^3] - \sum_{k=1}^{K-2} \theta_k \frac{\xi_K - \xi_k}{\xi_K - \xi_{K-1}} [(x-\xi_{K-1})_+^3 - (x-\xi_K)_+^3]$$

$$= \sum_{k=1}^{K-2} \theta_k (\xi_K - \xi_k) d_k(x) - \sum_{k=1}^{K-2} \theta_k \frac{\xi_K - \xi_k}{\xi_K - \xi_{K-1}} (\xi_K - \xi_{K-1}) d_{K-1}(x)$$

$$= \sum_{k=1}^{K-2} \theta_k (\xi_K - \xi_k)[d_k(x) - d_{K-1}(x)]. \quad (17)$$

Put (17) and (6) back to (5) we get the final natural cubic spline regression model.

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K-2} \theta_k (\xi_K - \xi_k)[d_k(x) - d_{K-1}(x)]. \quad (18)$$

Notice, in (18)$(\xi_K - \xi_k)[d_k(x) - d_{K-1}(x)]$ is a function of x we may write it as

$$s_k(x) = (\xi_K - \xi_k)[d_k(x) - d_{K-1}(x)]. \quad (19)$$

It is the basis function of the natural spline regression, now $f(x)$ can be treated as a regular linear regression model, we can calculate the coefficients by usual methods such as ordinary least square methods or maximum likelihood methods.

Conduct the natural spline regression 'by hand'

Since we know that a natural cubic spline regression model can be represented by (18) ($f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K-2} \theta_k (\xi_K - \xi_k)[d_k(x) - d_{K-1}(x)]$), now we can perform the regression 'by hand'.

'By hand' means we can conduct the regression without using special functions from R packages such as "splines" and 'rms' or functions from other statistical software and we even can conduct the regression model by using Excel or paper and pencils.

Next, show how to conduct the regression model step by step without using special R or other software's build in functions, and compare the results from our calculations with results from R packages such as "splines" and "rms".

The followings are the first 100 data records of the triceps data set. We will do calculations on data number 93 and 95 manually, then we will write a function to calculate basis function values for each data record.

From the equation (18) and (19) we can see except x we have $K-2$ basis functions in the model, i.e. from $k=1$ to $K-2$ basis functions. For triceps data set, we set five knots ($K=5$) at age 5, 10, 20, 30, 40, therefore we have 3 (5-2) spline basis. We write them as:

$$s_1=(\xi_5-\xi_1)[d_1(x)-d_4(x)]. \quad (20)$$

$$s2=(\xi_5-\xi_2)[d_2(x)-d_4(x)]. \quad (21)$$

$$s3=(\xi_5-\xi_3)[d_3(x)-d_4(x)]. \quad (22)$$

From the equation (15)

$$d_1(x) = \frac{(x-\xi_1)_+^3 - (x-\xi_5)_+^3}{\xi_5 - \xi_1}$$

$$d_2(x) = \frac{(x-\xi_2)_+^3 - (x-\xi_5)_+^3}{\xi_5 - \xi_1}$$

$$d_3(x) = \frac{(x-\xi_3)_+^3 - (x-\xi_5)_+^3}{\xi_5 - \xi_1}$$

$$d_4(x) = \frac{(x-\xi_4)_+^3 - (x-\xi_5)_+^3}{\xi_5 - \xi_1}$$

Now, let us look at the data number 93 whose age=40.72, from the above equations we get:

$$d_1 = \frac{(40.72-5)_+^3 - (40.72-40)_+^3}{40-5} = 1302.155$$

$$d_2 = \frac{(40.72-10)_+^3 - (40.72-40)_+^3}{40-5} = 966.3552$$

$$d_3 = \frac{(40.72-20)_+^3 - (40.72-40)_+^3}{40-5} = 444.755$$

$$d_4 = \frac{(40.72-30)_+^3 - (40.72-40)_+^3}{40-5} = 123.1552$$

Put d_1, d_2, d_3, d_4 back to (20), (21), (22) and we get:

s_1=(40−5)∗(1302.155−123.1552)=41265

s_2=(40−10)∗(966.3552−123.1552)=25296

s_3=(40−20)∗(444.755−123.1552)=6432

Notice the values of these basis functions are quite large, usually we can scale these values smaller, we scale these values by dividing them by $(\xi_5-\xi_1)^2$, as has been done by rms package in R.

When we scale continuous predictor it will not affect model fit and prediction, however it does affect the coefficient of the predictor variable.

$$s_1 = \frac{41265}{(40-5)^2} = 33.68571 \quad (23)$$

$$s_2 = \frac{25296}{(40-5)^2} = 20.6498 \quad (24)$$

$$s_3 = \frac{6432}{(40-5)^2} = 5.250612 \quad (25)$$

We continue the calculations for the data number 95, i.e. for age=17.92.

Note, for age=17.92, there are only two extra basis functions, s_1 and s_2 need to be calculated, since s_3 will be 0.

$$d_1 = \frac{(17.92-5)_+^3 - (17.92-40)_+^3}{40-5} = 61.61969$$

Note, $(17.92-40)_+^3 = 0$

$$d_2 = \frac{(17.92-10)_+^3 - (17.92-40)_+^3}{40-5} = 16.55977$$

d_3 and d_4 are all zeros for age=17.92.

$s_1=(40-5)*(61.61969-0)=2156.689$

$s_2=(40-10)*(16.55977-0)=496.7931$

Scale s_1 and s_2 by $(\xi_5-\xi_1)^2$

$$s_1 = \frac{2156.689}{(40-5)^2} = 1.760563$$

$$s_2 = \frac{496.7931}{(40-5)^2} = 0.4055454$$

I will stop manual calculations here, next I will write a R function to calculate the values of basis functions for each data record in the triceps data set.

Now we can understand that age' stands for s_1 basis function, age'' stands for s_2 function, and age''' stands for s_3 function. Next, I will show how to get the equation (4), i.e.

$$X\hat{\beta}=8.657058-0.3043198\text{age}+0.005997174\,(age-5)_+^3-0.01065974\,(age-10)_+^3+0.006292726(age-20)_+^3-0.001596325(age-30)_+^3-3.382968\times10^{-5}\,(age-40)_+^3. \quad (4)$$

To understand the equation (4) we need to look at the equation (17), when we restrict two tails of the regression to be linear, we have

$$\sum_{k=1}^{K}\theta_k(x-\xi_k)_+^3=\sum_{k=1}^{K-2}\theta_k(\xi_K-\xi_k)[d_k(x)-d_{K-1}(x)]. \quad (17)$$

From the equation (17) we can see from k to $K-2$ the coefficients of $(x-\xi_k)_+^3$ and $(\xi_K-\xi_k)[d_k(x)-d_{K-1}(x)]$ are the same, however, remember when we do the calculations for the equations (23), (24) and (25) we scale the basis functions (divided by $(\xi_K-\xi_1)^2$ (i.e. divided by 35^2).

We know for a linear regression model if we divide a continuous predictor by a constant c then the coefficient of the predictor will be increased by c times.

Since we divided the basis function values by 35^2 in our calculations for the regression model, we deliberately increased the coefficient by 35^2 times, therefore, for the original age scale, we need to divide the coefficients from the regression model by 35^2. Therefore,

$$\theta_1 = \frac{7.3465378}{35^2} = 0.005997174$$

$$\theta_2 = \frac{-13.05819}{35^2} = -0.01065974$$

$$\theta_3 = \frac{7.708589}{35^2} = 0.006292726$$

Now we obtained three coefficients by hand for the natural cubic spline regression, next we need to calculate θ_4 and θ_5, these two coefficients can be calculated by the equation (13) and the equation (7).

From the equation (13)

$$\theta_4 = -(\theta_1 * \frac{40-5}{40-30} + \theta_2 * \frac{40-10}{40-30} + \theta_2 * \frac{40-20}{40-30}) = -0.001596325$$

From the equation (7)

$$\theta_5 = -(\theta_1+\theta_2+\theta_3+\theta_4) = -3.382968\times10^{-5}$$

Now, we finished our hand calculations for the natural cubic spline regression model.

Hopefully, we can understand the meaning of the statistical analysis results.