My Solution of Data Analysis 3

Shaohao Chen, shaohao@bu.edu

April 5, 2018

1 Smoothing splines

In this session, I use smoothing splines method to fit the BMD data. The implementation is as following.

The training set is $\{x_i, y_i\}_{i=1}^N$, where $x = (x_1, ..., x_i, ..., x_N)^T$ and $y = (y_1, ..., y_i, ..., y_N)^T$. First, compute the the natural cubic splines basis functions $N_k(x)$ using Eq. (5.4) and Eq. (5.5) in the book, choosing $K = N$ knots at the training points $\{x_i\}_{i=1}^N$. Obtain the regression coefficients using Eq. (5.12),

$$\hat{\theta} = (H^T H + \lambda \Omega_N)^{-1} H^T y \tag{1}$$

where the *k-th* column of H is $N_k(x)$, and $\{\Omega\}_{jk} = \int N_j^{''}(t) N_j^{''}(t) dt$. Predict the response at original trianing points $\{x_i\}_{i=1}^N$,

$$\hat{y} = \hat{f}(x) = H\hat{\theta} \tag{2}$$

and predict the response at new test points $\{x_i^*\}_{i=1}^N$ chosen as uniformly distributed values in the range $[x_{min}\ x_{max}]$,

$$\hat{y}^* = \hat{f}(x^*) = H^* \hat{\theta} \tag{3}$$

where $x^* = (x_1^*, ..., x_i^*, ..., x_N^*)^T$, and the *k-th* column of H^* is $N_k(x^*)$. To obtain the confidence band at test points $\{x_i^*\}_{i=1}^N$, we first consider the matrix S_λ^* that transforms original y to the predicted $\hat{y}^*$ at test points,

$$\hat{y}^* = S_\lambda^* y \tag{4}$$

From Eq. (1), (3) and (4), we obtain,

$$S_\lambda^* = H^* (H^T H + \lambda \Omega_N)^{-1} H^T \tag{5}$$

Using the property of covariance, we have,

$$Var(\hat{y}^*) = S_\lambda^* \, Var(y) \, S_\lambda^{*T} = S_\lambda^* \, (\hat{\sigma}^2 I) \, S_\lambda^{*T} \tag{6}$$

where $\hat{\sigma}^2$ can be estimated using Eq. (8.21),

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{f}(x_i))^2 \tag{7}$$

It has the same form as the least-square estimated training error. Finally the estimated standard deviation at test point is calculated as,

$$\hat{se} = z^* * \sqrt{v_d^*}, \tag{8}$$

where v_d^* is a vector that contains the diagonal elements of $Var(\hat{y}^*)$. I choose $z^* = 1.645$ to be consistent with 90% confidence as required.

Note that if $\lambda = 0$, it becomes a non-smoothing cubic splines model, in which case the training error (estimated by Eq. (7)) is smaller than the training errors of any nonzero λ. But the predicted curve is not smooth (there are large wigglings) as $\lambda = 0$. I choose $\lambda = 0.25$ to obtain a relatively small training error and smoothness at the same time. The training error is $\hat{\sigma}^2 = 0.00162$ for male data and $\hat{\sigma}^2 = 0.00116$ for female data.

The predicted curves $\{x^*, \hat{y}^*\}$ as well as the original data points are plotted in Figs. 1 and 2 respectively for male and female data. As we can see, the result in Fig. 5.6 in the book is reproduced, and the error bands are added. Due to the penalty term in smoothing splines method, the predicted curves are smooth, that is, there is no large wiggling, which is often seen in normal (non-smoothing) cubic splines.

Note: the cross validation of this problem will be done in section 4.

2 Bayesian method

I use Eq. (8.28) to calculate the posterior mean E and covariance V. I choose $\Sigma = I$. I use the smoothing splines basis functions, i.e. use the same H as in section 1. I estimate σ^2 using Eq. (7). The estimated standard deviation is $\hat{se} = z^* * \sqrt{v_d}$, where v_d is a vector that contains the diagonal elements of V. The results for $\tau = 0.1, 1$, and 10 are plotted in Figs. 3, 4, 5 for male data and Figs. 6, 7, 8 for female data respectively. As τ decreases, the curve becomes smoother. The predicted curves with $\tau = 0.1$ is close to those in section 1. The error intervals are in general a little larger those in smoothing splines method.

Here are more discussions on smoothness of the predicted curves. In the smoothing splines model, the second derivative is explicitly taken into account in the penalty term, so it is guananteed that there is no large wiggling. In the Bayesian model, the term $\sigma^2\Sigma/\tau$ roughly plays a role of penalty, which also makes the curve smooth, but not as efficient as the smoothing splines model. As a result, we can see some wigglings in Figs. 3, 4, 5, 6, 7, 8. If we decrease τ in the Bayesian model, i.e. increase the weight of the penalty term, the predicted curve beomce smoother, and they are very close to the curves predicted by the smoothing splines model (in Figs 1 and 2).

3 Bootstrapping

I obtain 100 bootstrapping samples from the original training data,

$$y^b = y + \varepsilon;\ \varepsilon \sim N(0, \sigma^2) \tag{9}$$

where $b = 1, 2, ..., 100$, and $N(0, \sigma^2)$ is normal distribution with mean at 0 and σ^2 is calculated by Eq. (7). For each b, use $\{x_i, y_i^b\}_{i=1}^N$ as training data set, and do the smoothing splines procedure in section 1 to obtain the predicted responses $\hat{f}^b(x^*)$ at test points and the training error $\hat{\sigma}_b^2$. The predicted curves in five selected samples are plotted in Fig. 9 for male data and Fig. 11 for female data (similar to Fig. 8.2 in the book). Then calculate the average of predicted responses and the average of the training error over all bootstrapping samples,

$$\overline{\hat{y}^*} = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x^*) \tag{10}$$

$$\overline{\hat{\sigma}^2} = \frac{1}{B}\sum_{b=1}^{B}\hat{\sigma}_b^2 = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}^b(x_i))^2 \tag{11}$$

where $B = 100$ is the number of samples. The standard deviation is calculated by Eq. (8). The average training error for $\lambda = 0.25$ is 0.00318 and 0.00221 for male data and female data respectively. The average predicted curves $\{x^*, \overline{\hat{y}^*}\}$ and the error band as well as the original data points are plotted in Fig. 10 for male data and Fig. 12 for female data. The predictive curves are very close to that in section 1, since the average over a large number of (say 100) samples bootstrapped from normal distribution is close to its theoretical mean value (that is the result in section 1). The error intervals are larger than those in section 1, since additional variance is introduced from the bootstrapping samples.

4 Cross validation

I do a 5-fold cross validation here. First evenly split the original data set into five sub sets $\{x_k^j, y_k^j\}_{k=\kappa_j(i)}$, where $j = 1, 2, 3, 4, 5$, and $\kappa_j(i)$ is the index of the *i-th* original data point that falls into the *j-th* one-fifth subset. Note that the data points in each one-fifth subset are chosen in an unsorted order to make the subset of data distribute in most of the original range $[x_{min}\ x_{max}]$. Then use the *j-th* one-fifth sub set $\{x_i^j, y_i^j\}_{k=\kappa_j(i)}$ as validation data and use the other four one-fifth subsets $\{x_k^j, y_k^j\}_{k=-\kappa_j(i)}$ as training data, where $-\kappa_j(i)$ means all indexes that are not in the *j-th* one-fifth subset. Then repeat the smoothing splines procedure in section 1 to obtain the predicted responses

$$\hat{y}^{*j} = \hat{f}^{-\kappa_j(i)}(x^*), \tag{12}$$

and calculate the error as,

$$\hat{\sigma}_j^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}^{-\kappa_j(i)}(x_i))^2. \tag{13}$$

where $\hat{f}^{-\kappa_j(i)}$ is the function predicted on the training data without the points indexed $\kappa_j(i)$. Note that different from Eq. (7), out-of-sample data are used to estimate the error in Eq. (13). The j is varied from 1 to 5, so that all data points have been used as training data for four times (i.e. experience four epoches). Then calculate the average of predicted responses and the average of the training error over all j,

$$\overline{\hat{y}^*} = \frac{1}{J}\sum_{j=1}^{J}\hat{y}^{*j} \tag{14}$$

$$\overline{\hat{\sigma}^2} = \frac{1}{J}\sum_{j=1}^{J}\hat{\sigma}_j^2 \tag{15}$$

where $J = 5$ is the number of training data sets. The standard deviation is calculated by Eq. (8). The training error for $\lambda = 0.25$ is 0.00165 and 0.00118 for male data and female data respectively. The predicted curves on the five training sets are plotted in Fig. 13 for male data and Fig. 15 for female data. The average predicted curves $\{x^*, \overline{\hat{y}^*}\}$ and the error band as well as the original data points are plotted in in Fig. 14 for male data and Fig. 16 for female data. The predictive curves are close to that in section 1, since the average of all subsets of data is close the original data. The error intervals are larger than those in section 1, since out-of-sample data is included in the calculation.

5 All together

The predicted curves in section 1, 2, 3, and 4 are plotted together in Fig. 17 for male data and Fig. 18 for female data. The bootstrapping and cross validation curves are very close to that in its original smoothing spline model. There are more wigglings in the curve predicted by the Bayesian model due to its rough estimation of the penalty. The bootstrapping error is larger because it brings in the additional variance of different samples. The cross validation error is larger because out-of-sample data is included in the calculation. The original smoothing splines model (without bootstrapping or cross validation) underestimates the error, via the errors in bootstrapping or cross validation models are more close to the true error.