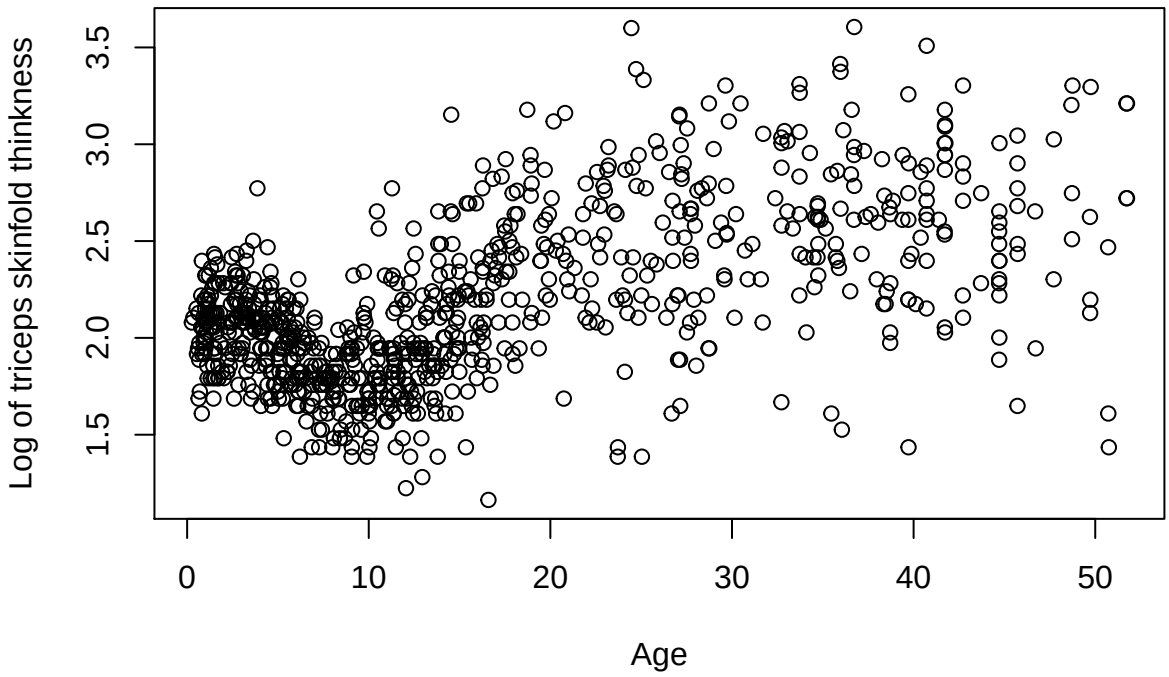
Assignment1

Tiffany Woodley

5/25/2020

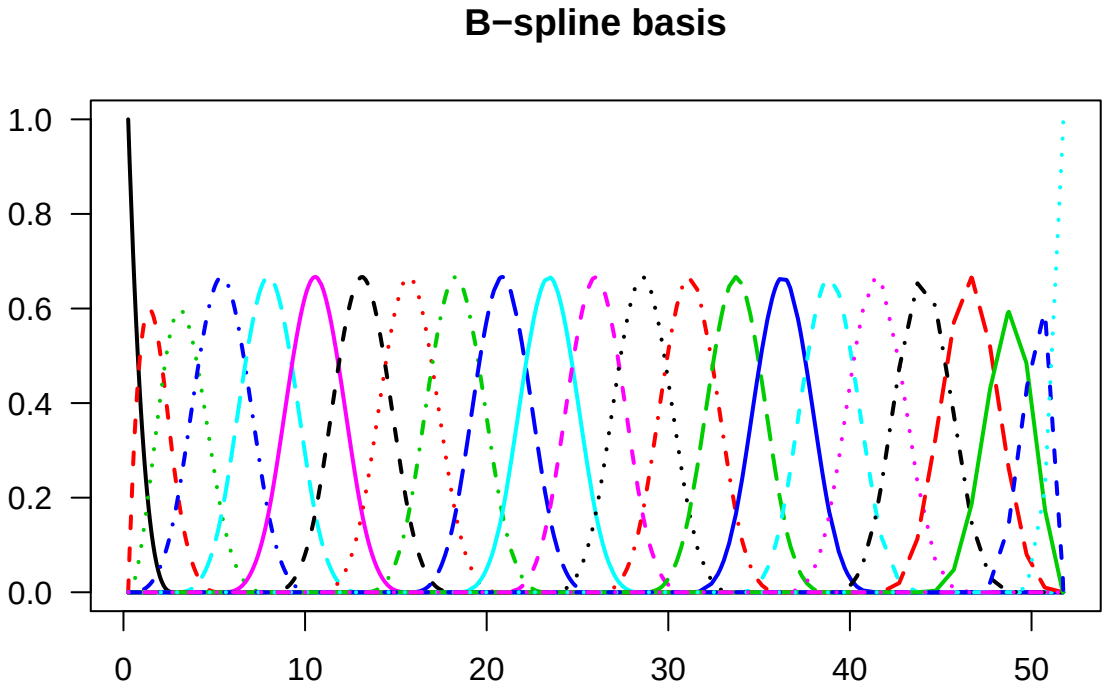
Question 1: P-Splines

The data utilised in this question comes from a dataset containing information about 892 females under 50 years of age, collected from three villages in West Africa. The below figure shows the log of their skinfold thickness vs their age. The relationship between the two is clearly non-linear. The data contains a majority of observations of females under the age of 20 allowing us to see a more definitive curve in ages below 20. In this question, P-splines are used in order to attempt to capture this relationship.

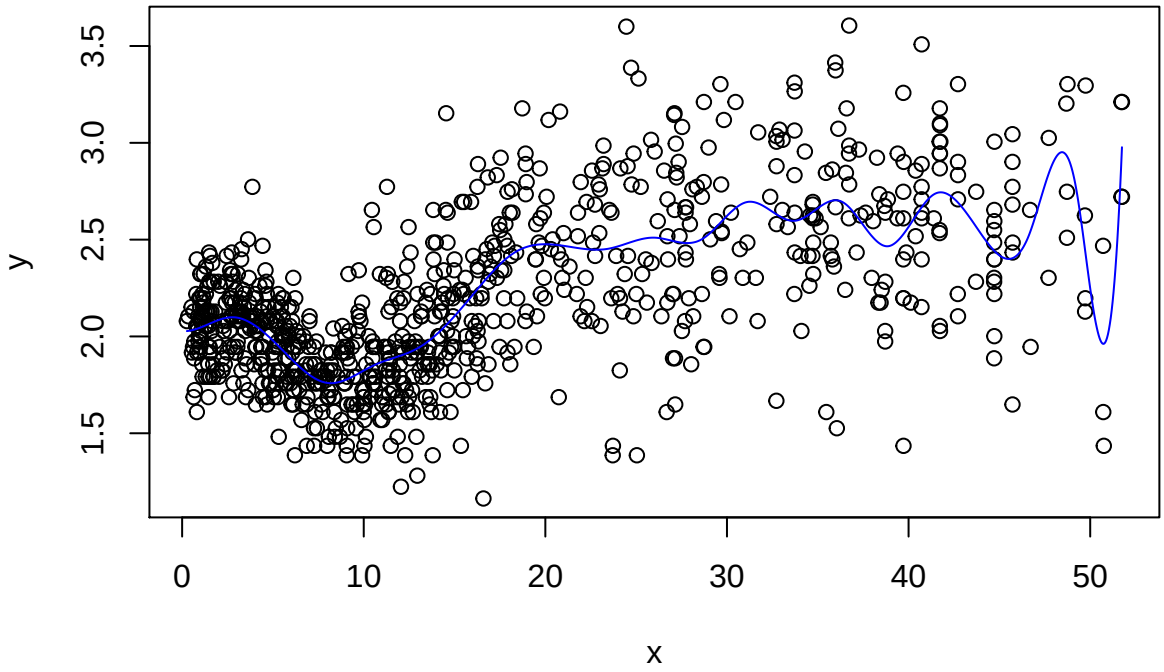


P-splines use a B-spline basis in the regression equation, this allows for the non-linear relationship to be captured without having to make assumptions about the relationship before hand. Since B-splines provide local functions, the calculated betas become more interperatbale as they are only relevent over certain intervals.

A cubic B-spline basis was created using 20 knots exluding the intercept, resulting in a degree of freedom of $df = M - 1 + k = 22$. The local functions on the right are less smooth as they are plotted on fewer points.



Fitting a linear model using this basis without penalisation results in a curve containing a lot of wigglyness (as can be seen below). This is to be expected since the 20 knots used provides a high degree of freedom. The flexibility of the model is allowing the model to fit to the inherent noise in the data, regularization is therefore needed in order to curb the overfitting.



Due to the local nature of the B-spline basis functions, we will apply a difference penalty to the β_j in order to force the curvature of the functions to be similar in consecutive intervals in turn forcing the resultant fit to be smoother. This will in turn constrain the influence the amount of knots has on the wigglyness.

The penalty used will be in the form:

$$P = \sum_{j=1}^{k-1} (\beta_{i+1} - \beta_i)^2 = \beta_1^2 - 2\beta_1\beta_2 + 2\beta_2^2 - 2\beta_2\beta_3 + ... + \beta_k^2$$

This can be expressed in matrix form

$$P = \beta^T \begin{bmatrix} 1 & -1 & 0 & \dots \\ -1 & 2 & -1 & \dots \\ 0 & -1 & 2 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \beta$$

Using this penalisation matrix we can evaluate the penalised sum of squares by

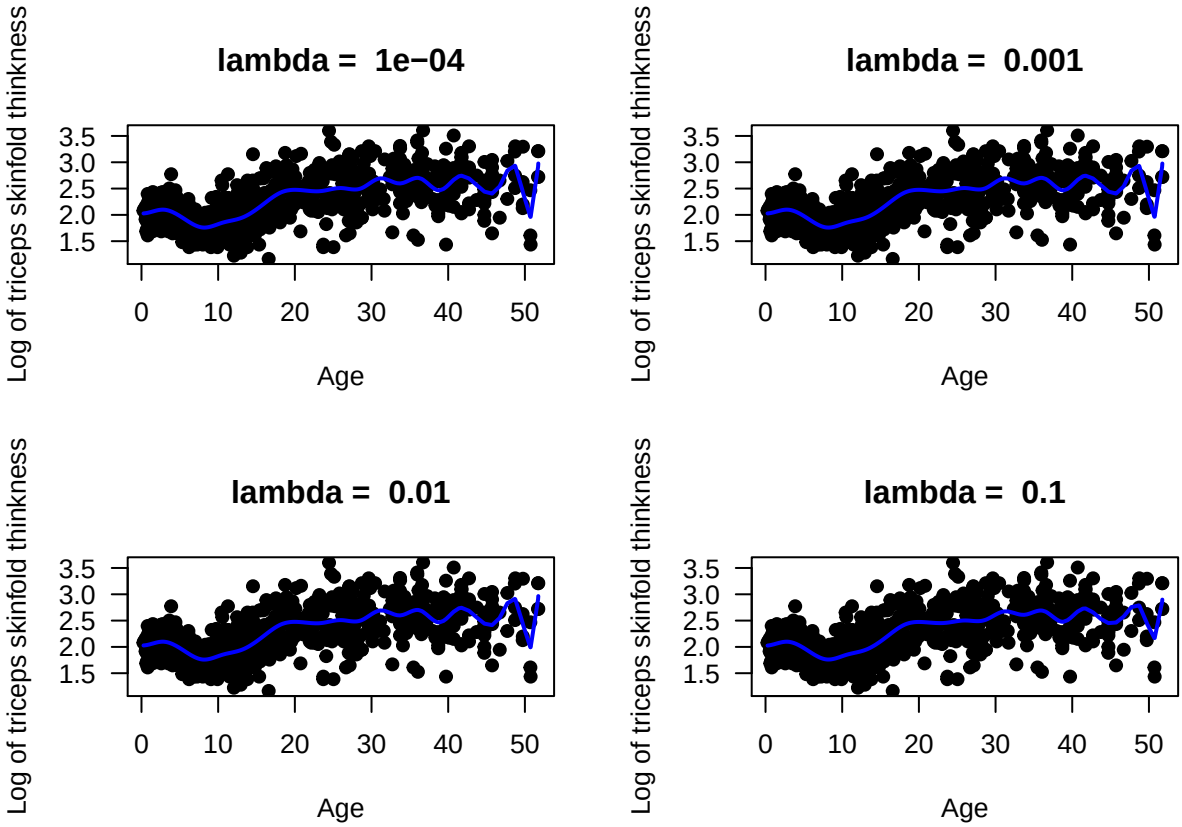
$$PSS = ||y - X\beta||^2 + \lambda\beta^T P\beta$$

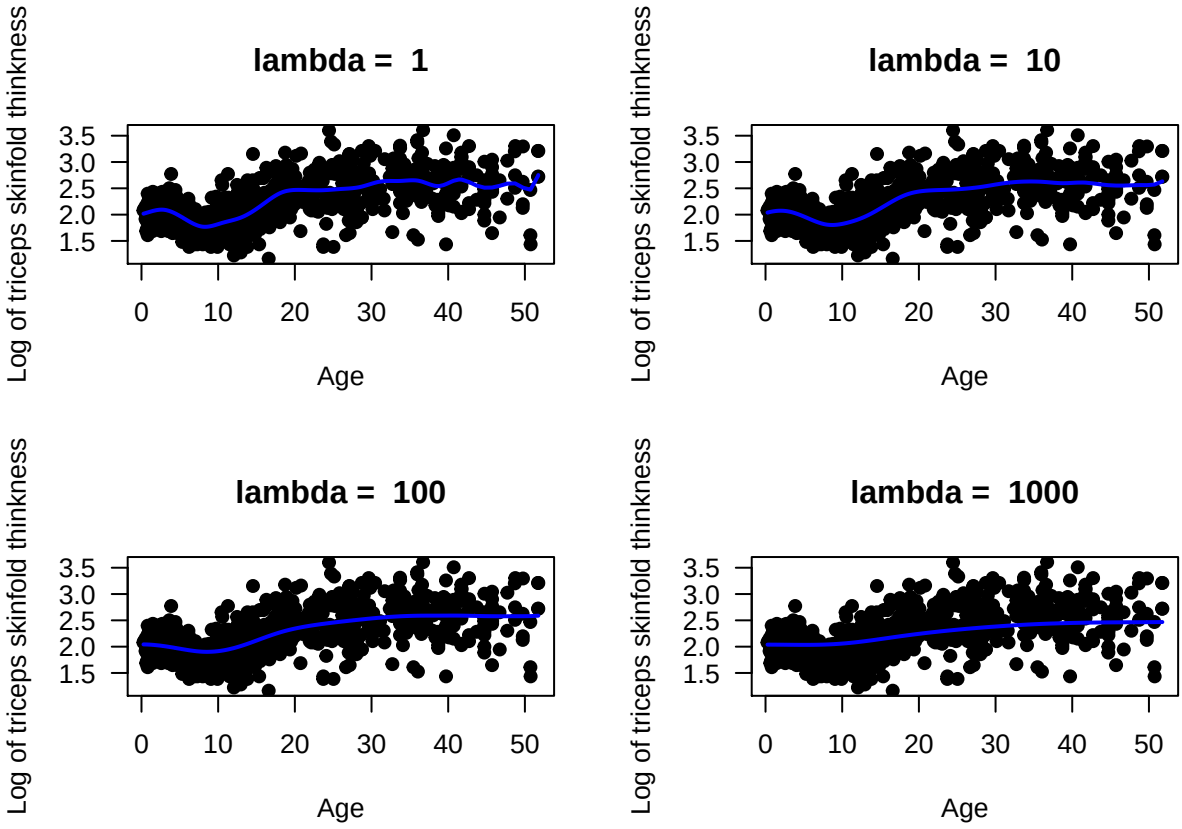
where X is the computed B-spline basis

The $\hat{\beta}$ needed in order to minimize the penalised sum of squares can be calculated using the following equation (whilst λ is held constant).

$$\hat{\beta} = (X^T X + \lambda P)^{-1} X^T y$$

P-splines are fit whilst holding different λ values constant in order to visualize the impact the λ has on the fit. As lambda increases the fitted line becomes less wiggly and starts to resemble a straight line.





Visually it is hard to identify the best fitting λ, and so Generalised Cross Validation(GCV) is used in order to evaluate the lambdas. GCV approximates leave one out cross validation and so will give us an idea of how well the model will generalise to new data i.e. how well the true underlying distribution has been captured.

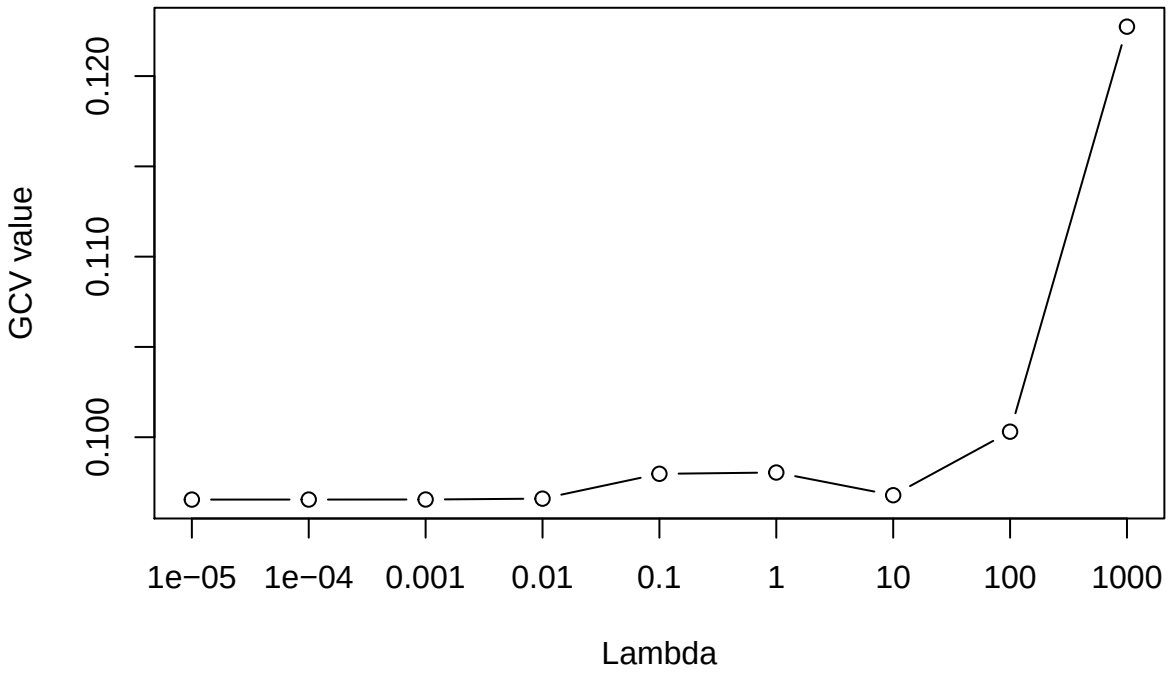
The GCV can be calculated using

$$GCV(f) = \frac{1}{N}\sum_{i=1}^{N}(\frac{y_i - f(x_i)}{1 - trace(S)/N})^2$$

with

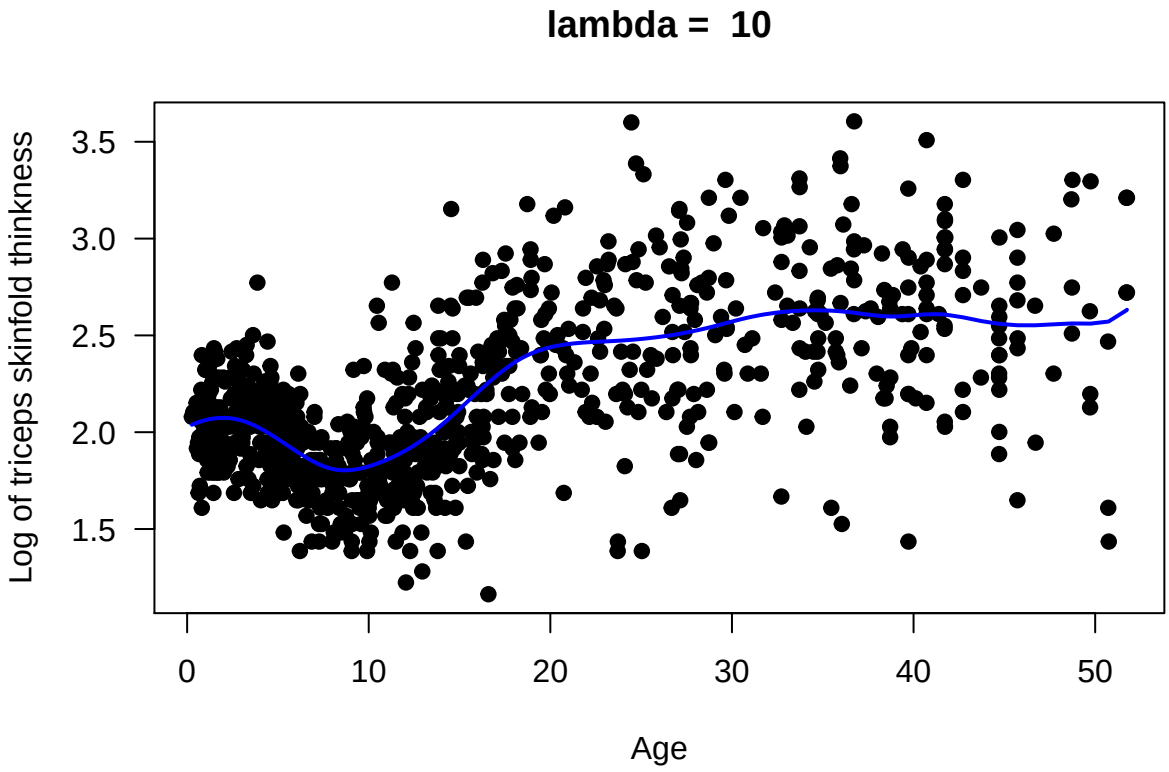
$$S = X(X^T X - \lambda P)^{-1} X^T$$

where S estimates the expected degree of freedom.



With the majority of the points being located on the left side of the x axis, the fit of the curve on the left will affect the GCV value more than the fit of the right hand side of the axis. Because of this extremelly small lambda values are favored in order to capture the small curve for ages below 5 as this is were most points are saturated. This leaves the right side of the function extremelly wiggly and likely to be overfitting to the data, this can be visualised in the figure showing the unpenalised fit.

After $\lambda = 10$ the GCV starts to increase rapidly as the model becomes to smooth and begins to underfit. $\lambda = 10$ was chosen for the final fit having a very similar GCV value to the smaller lambdas, the very slight increase in the GCV value for greater smoothness gives the model a better chance at generalising.



REFERENCES

1. Erni, Birgit, 2020, *Chapter 1: Splines*, lecture notes, Advanced Analytics STA5057W, University of Cape Town, May 2020