Application of nonparametric regression on "Relative spinal bone mineral density measurements on 261 North American adolescents" data

N. R. , Hakiim (2021)

Data is taken from loon.data package on R software. From the definition :

> From the web source: "Relative spinal bone mineral density measurements on 261 North American adolescents. Each value is the difference in spnbmd taken on two consecutive visits, divided by the average. The age is the average age over the two visits." The data are a repackaging and extension of the data of the same name from the now archived (in 2020) of the 2015 'ElemStatLearn' package of Kjetil B. Halvorsen.

For simplicity, writer will refer the data as "bone" data. This note will cover the results of applicating regression methods from parametric to nonparametric on the bone data. The methods which are used limited to linear regression (also included : polynomial regression) and smoothing regression (kernel and spline).

First of all, we'll show the structure and some of the observations in the data.

```
> str(bone)
'data.frame':   485 obs. of  5 variables:
 $ idnum  : int  1 1 1 2 2 2 3 3 3 4 ...
 $ age    : num  11.7 12.7 13.8 13.2 14.3 ...
 $ sex    : Factor w/ 2 levels "female","male": 2 2 2 2 2 2 2 2 2 1 ...
 $ rspnbmd: num  0.01808 0.06011 0.00586 0.01026 0.21053 ...
 $ ethnic : Factor w/ 4 levels "Asian","Black",..: 4 4 4 4 4 4 4 4 4 4 ...
```

```
> head(bone, 20)
   idnum   age    sex      rspnbmd ethnic
1      1 11.70   male  0.018080670  White
2      1 12.70   male  0.060109290  White
3      1 13.75   male  0.005857545  White
4      2 13.25   male  0.010263930  White
5      2 14.30   male  0.210526300  White
6      2 15.30   male  0.040843210  White
7      3 11.40   male -0.029641190  White
8      3 12.40   male -0.006430868  White
9      3 13.40   male  0.056634300  White
10     4 10.55 female  0.108043200  White
11     4 11.50 female  0.219912500  White
12     4 12.55 female  0.076640850  White
13     5 12.75 female  0.096413870  White
14     5 13.70 female  0.043816940  White
15     5 14.75 female  0.047489820  White
16     6 18.00 female  0.005836576  White
17     7 13.20 female  0.106995900  White
18     7 14.25 female  0.094626560  White
19     7 15.30 female  0.014590350  White
20     8 12.95   male  0.194464300  White
```

As it can be seen from above, there are essentially 4 variables which consists of 2 categorical and 2 numerical variable. To emphasize again, this note will explore causal effect with the target variable is the "relative spinal bone mineral density" labeled rspnbmd with regression analysis. Also since this note will not touch the multivariate topic, the model would be built with age, sex, and ethnic as predictor variables for rspnbmd. Thus, the first model that is hypothesized is :

$$\hat{y} = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 \quad \text{(model 1)}$$

With x_1, x_2, x_3 is age, sex, and ethnic consecutively. The results from fitting the data to model 1 is as follow :

```
Call:
lm(formula = rspnbmd ~ age + sex + ethnic, data = bone)

Residuals:
     Min       1Q   Median       3Q      Max
-0.12912 -0.02720 -0.00445  0.01940  0.15732

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.1350849  0.0095621  14.127   <2e-16 ***
age            -0.0061157  0.0004871 -12.555   <2e-16 ***
sexmale         0.0011336  0.0039343   0.288    0.773
ethnicBlack     0.0044515  0.0062233   0.715    0.475
ethnicHispanic -0.0026997  0.0063391  -0.426    0.670
ethnicWhite     0.0044394  0.0050694   0.876    0.382
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04274 on 477 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.2667,    Adjusted R-squared:  0.259
F-statistic: 34.69 on 5 and 477 DF,  p-value: < 2.2e-16
```

Age is the only significant predictor. This would raise one question, are sex and ethnic really not a good predictor for rspnbmd? To answer this, ANOVA is used to model rspnbmd with these 2 categoric variables. The results :

```
             Df Sum Sq  Mean Sq F value  Pr(>F)
ethnic        3 0.0287 0.009553   3.919 0.00878 **
sex           1 0.0003 0.000266   0.109 0.74115
ethnic:sex    3 0.0016 0.000541   0.222 0.88119
Residuals   475 1.1577 0.002437
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
2 observations deleted due to missingness
```

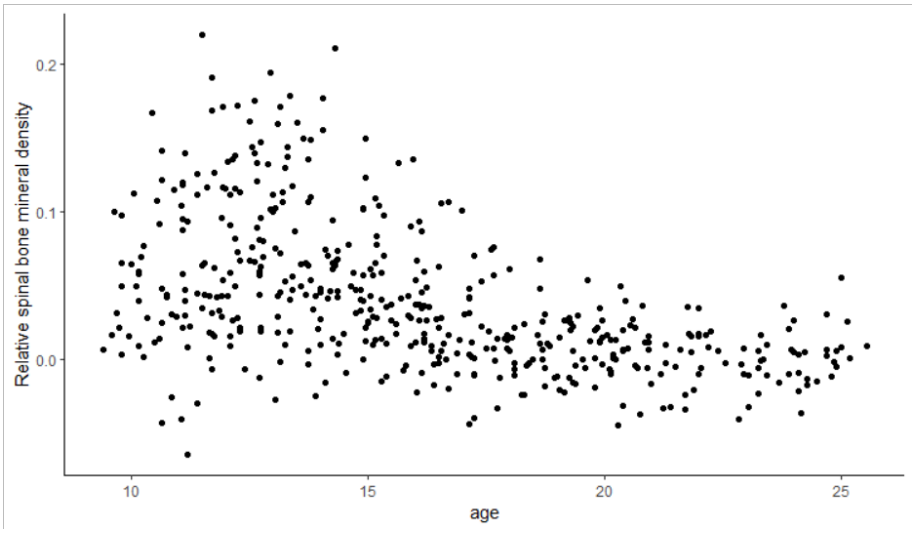
Only ethnic is significant and even after eliminating the interaction term, sex doesn’t seem to be a good predictor for rspnbmd on linear regression and ANOVA. Thus, for the categorical variable, writer concludes that Relative spinal bone mineral density is different in at least 2 ethnic groups in the bone data. Further, since in the linear regression only age is found to be significant predictor, the analysis will advance with only age as the predictor for rspnbmd.

Going back to regression model, since age is the only predictor left for rspnbmd, model 1 is reduced to

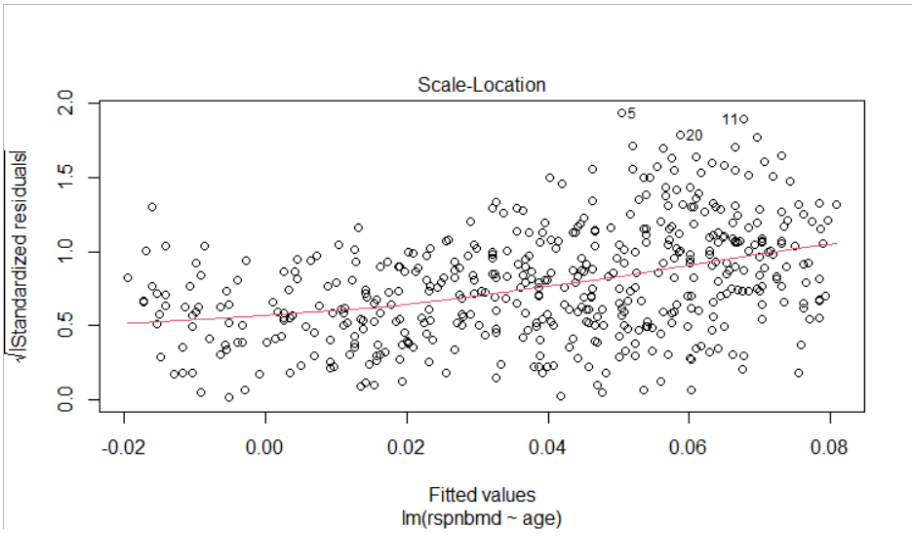
$$\hat{y} = \beta_0 + x_1\beta_1 \quad \text{(model 2)}$$

```
Call:
lm(formula = rspnbmd ~ age, data = bone)

Residuals:
      Min        1Q    Median        3Q       Max
-0.133738 -0.028337 -0.004409  0.019243  0.160132

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1391519  0.0079105   17.59   <2e-16 ***
age         -0.0062068  0.0004765  -13.03   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04275 on 483 degrees of freedom
Multiple R-squared:   0.26,     Adjusted R-squared:  0.2585
F-statistic: 169.7 on 1 and 483 DF,  p-value: < 2.2e-16
```

Proceeding to analyze the result of model 2 : it can be seen that even though the model is significant, the value of R^2 (which represent the variance of the data that can be explained by the regression) is pretty low. This implies that the regression ignores high amount of variance (thus ignores most of the information from the data). To see where the problem may come from, here is the plot between x = age and y = rspnbmd

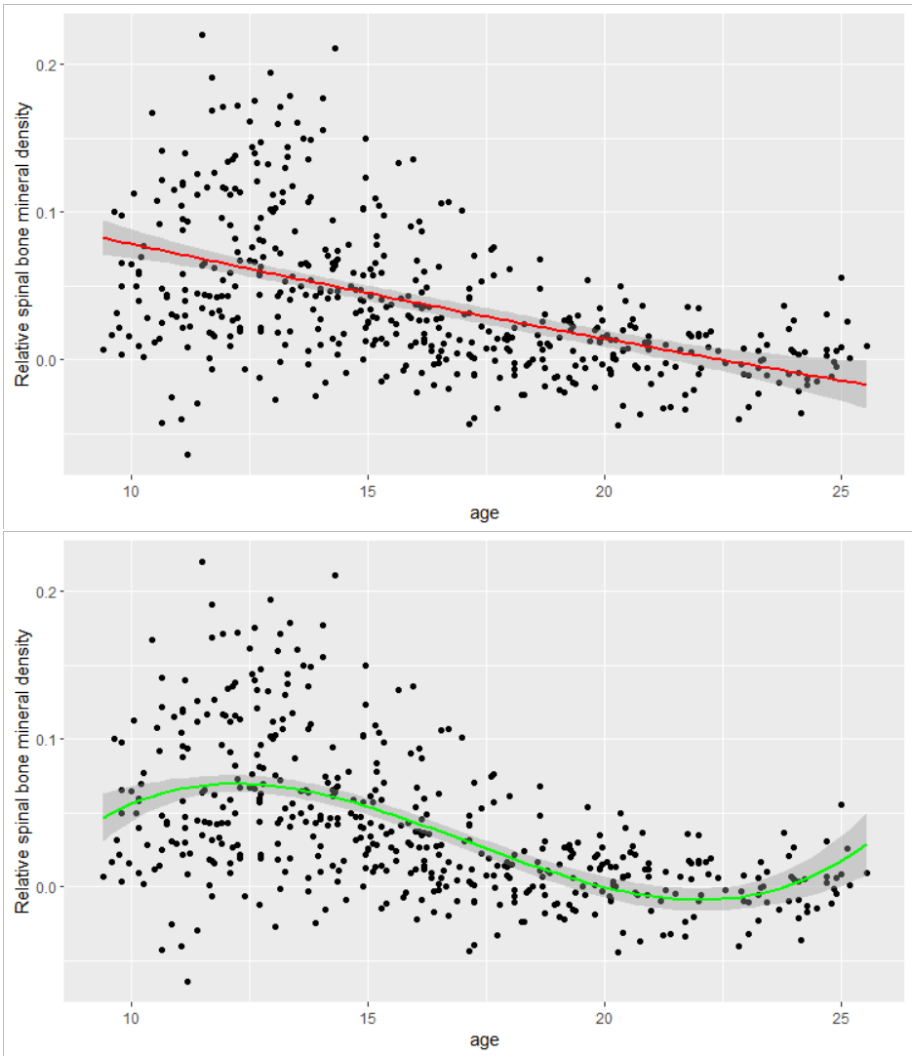


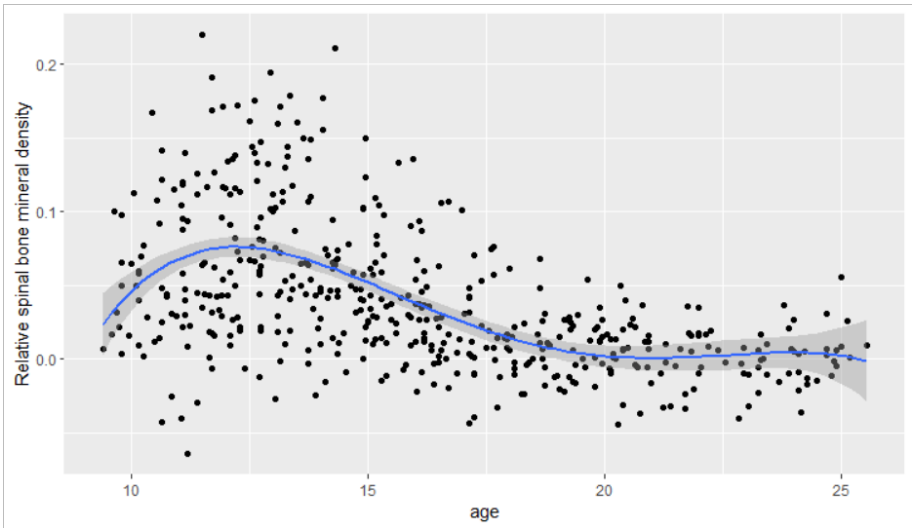
Subjectively speaking, the plot does seem to not follow a linear pattern. Also checking on homogeneity assumption :



Variance for the residual shows an increasing pattern as the fitted values increase. Thus a simple linear regression seems to underfit the bone data. While a lot of solution to this problem is available, writer focused on the plot between age and rspnbmd and decided to try fitting polynomial regression as the plot seems to show curve pattern.

After several trials, it is found that the optimal polynomial regression model for bone data is polynomial regression with degree of 4. To compare the fit for each polynomial regression model from degree 2 to 4, the following plots represent consecutively polynomial regression model degree 2, 3, and 4 applied to bone data :





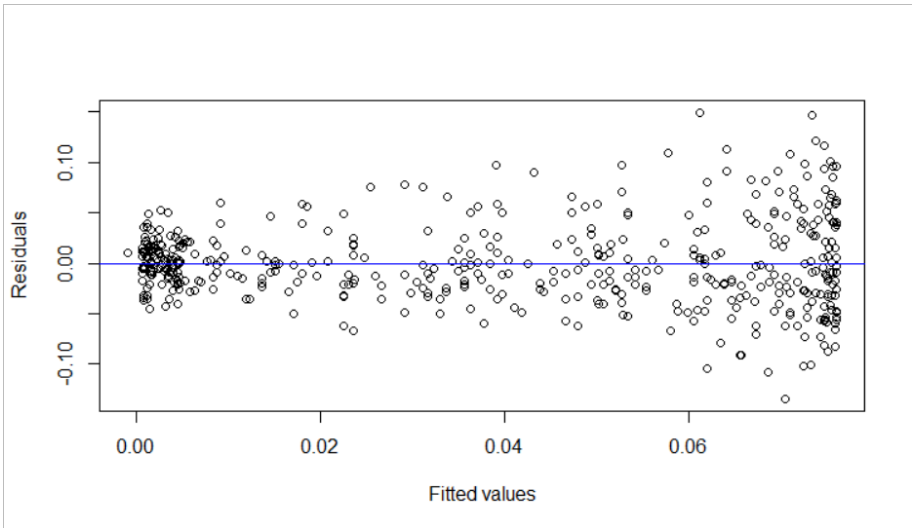
Curved line from the polynomial regression with degree of 4 seems reasonable enough to represent the relationship between age and relative spinal bone mineral density. As such, the current fitted model is :

$$\hat{y} = \beta_0 + x\beta_1 + x^2\beta_2 + x^3\beta_3 + x^4\beta_4 \qquad \text{(model 3)}$$

And the summary of the model 3 fitted into bone data :

```
Call:
lm(formula = rspnbmd ~ age + I(age^2) + I(age^3) + I(age^4),
    data = bone)

Residuals:
      Min        1Q    Median        3Q       Max
-0.134602 -0.024491 -0.004059  0.020471  0.149365

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.236e+00  5.106e-01  -4.379 1.46e-05 ***
age          5.514e-01  1.282e-01   4.301 2.06e-05 ***
I(age^2)    -4.713e-02  1.174e-02  -4.014 6.93e-05 ***
I(age^3)     1.704e-03  4.652e-04   3.663 0.000277 ***
I(age^4)    -2.231e-05  6.737e-06  -3.312 0.000996 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04081 on 480 degrees of freedom
Multiple R-squared:  0.3297,    Adjusted R-squared:  0.3241
F-statistic: 59.03 on 4 and 480 DF,  p-value: < 2.2e-16
```

Analyzing the summary, it is found that R^2 after model improvement still doesn’t reach acceptable value for writer. Looking into the assumption, homogeneity is still the problem.



Residual plot above shows similar pattern to that of residual plot from simple linear regression/model 2. Of course, same as before, there are several methods to overcome this problem. But, for the next part, this note is going to focus on the nonparametric solution.

Skipping details and theories, writer decide to fit a smoothed regression with 2 estimation methods : (1) kernel and (2) spline.

For the first part of nonparametric regression : kernel-smoothed regression. Kernel regression is an extension from kernel density estimation. It uses a kernel estimation to estimate the functional relationship between predictor and target. Estimation works based on weighted observation :

$$\hat{f}(x) = \sum_{i=1}^{n} y_i w_i(x) \qquad w_i(x) = \frac{K([x - x_i]/h)}{\sum_{i=1}^{n} K([x - x_i]/h)}$$

With $K(.)$ represent the kernel function and $h > 0$ is known as the bandwidth parameter. Bandwidth is used to control the information to use from nearby observations.

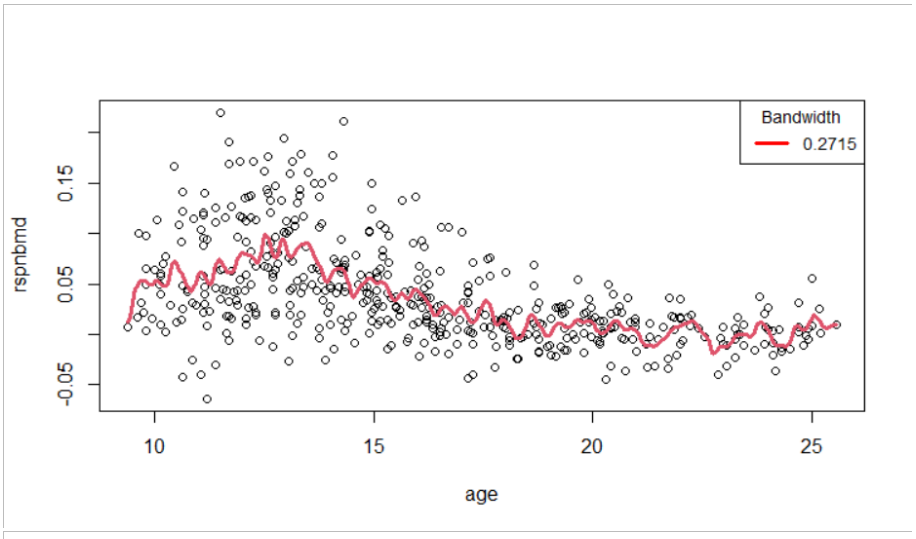
Most commonly used kernel function is gaussian/normal. Method on choosing the bandwidth used is based on cross-validation (CV). Writer uses 3 formula : maximum-likelihood, unbiased, and generalized. The report from each formula for the value of bandwidth is as follow :

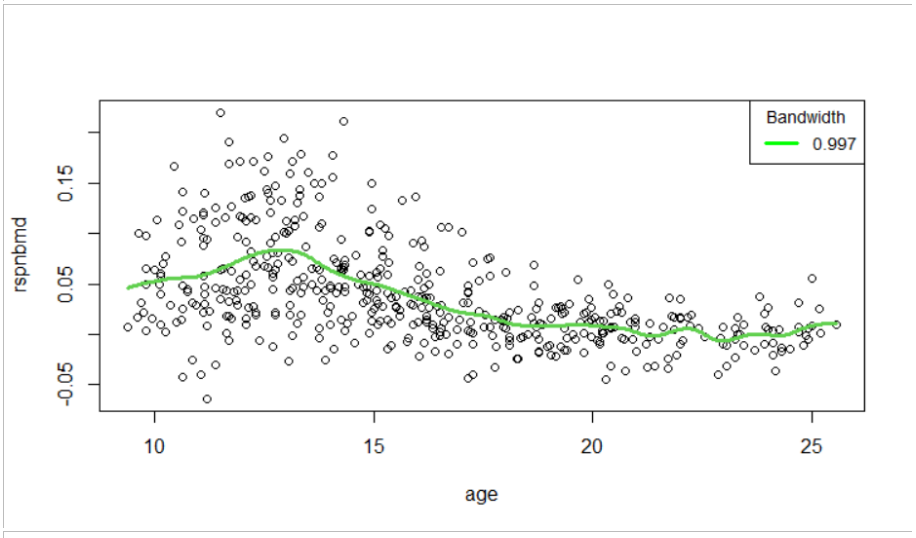
```
Call:           Maximum-Likelihood Cross-Validation

Data: bone$age (485 obs.);      Kernel: gaussian
Max CV = -2.732;          Bandwidth 'h' = 0.2715
Call:            Unbiased Cross-Validation

Derivative order = 0
Data: bone$age (485 obs.);      Kernel: gaussian
Min UCV = -0.06945042;  Bandwidth 'h' = 0.9972252
```

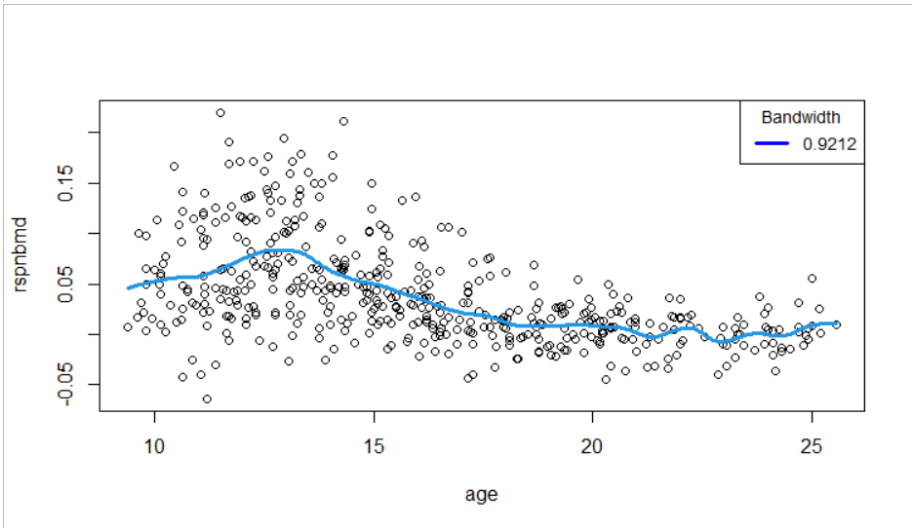
Generalized Cross-Validation

```
$h
[1] 0.9211855
```

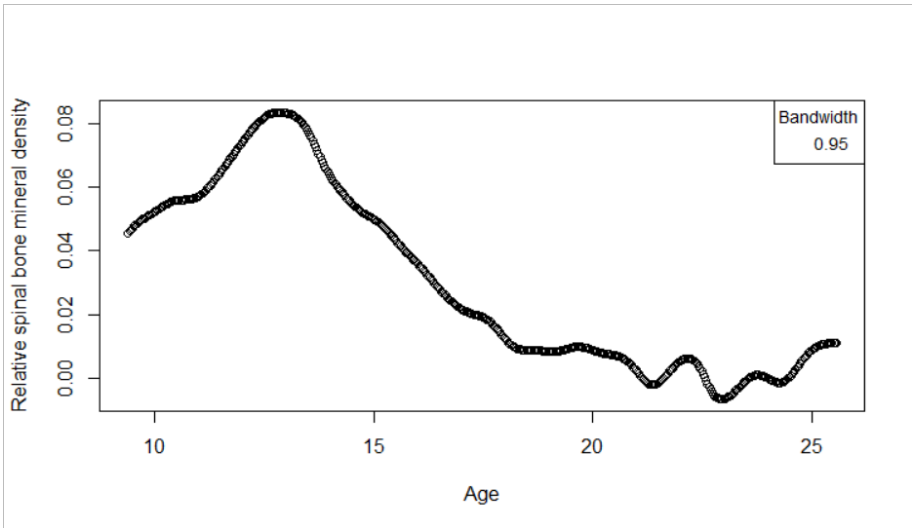
After having bandwidth estimation based on CV, writer plot the smoothed kernel regression into the data plot.







From the plot above, writer deducted that optimal bandwidth of kernel smoothing regression for bone data is around 0.9-1. Therefore, writer decided to choose bandwidth 0.95.



Next, is spline smoothing. Similar concept to that of kernel regression, spline weight each of the observation and build the functional relationship with spline. Smoothing spline choose to minimize the penalized least square functional :

$$f_\lambda = \min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda J_m(f)$$

Where $\lambda > 0$ represent the smoothing parameter (equivalent to that of bandwidth in kernel) and $J_m(f) = \int |f^{(m)}(z)|^2 dz$ is the penalty term. $f^{(m)}(.)$ is the m-th derivative of $f(.)$ and $\mathcal{H} = \{f : J_m(f) < \infty)$ being the function's space which integrable at m-th derivative.

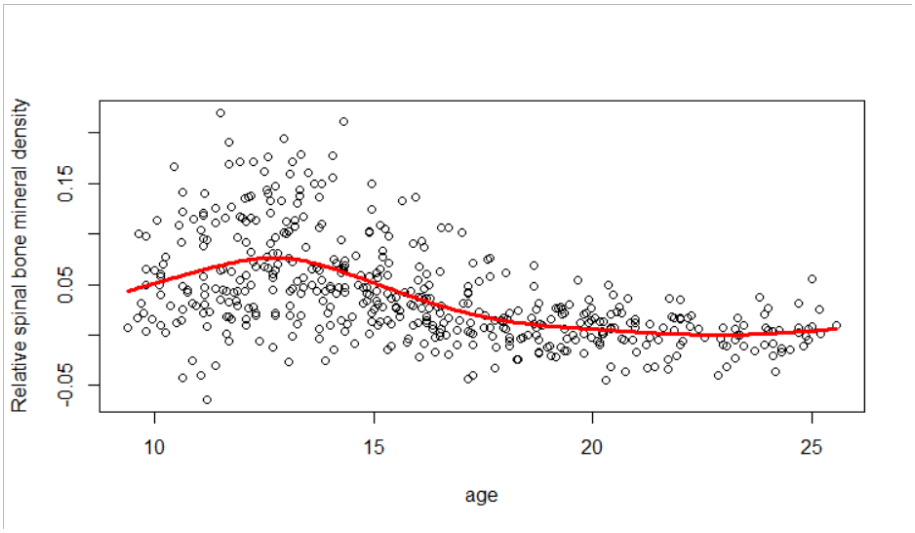
Applying the function ss function from npreg package would automatically gives all the necessary result for spline smoothing.

```
Call:
ss(x = age, y = rspnbmd)

Smoothing Parameter  spar = 0.3331017   lambda = 1.520011e-05
Equivalent Degrees of Freedom (Df) 6.489693
Penalized Criterion (RSS) 0.789812
Generalized Cross-Validation (GCV) 0.00167295
```

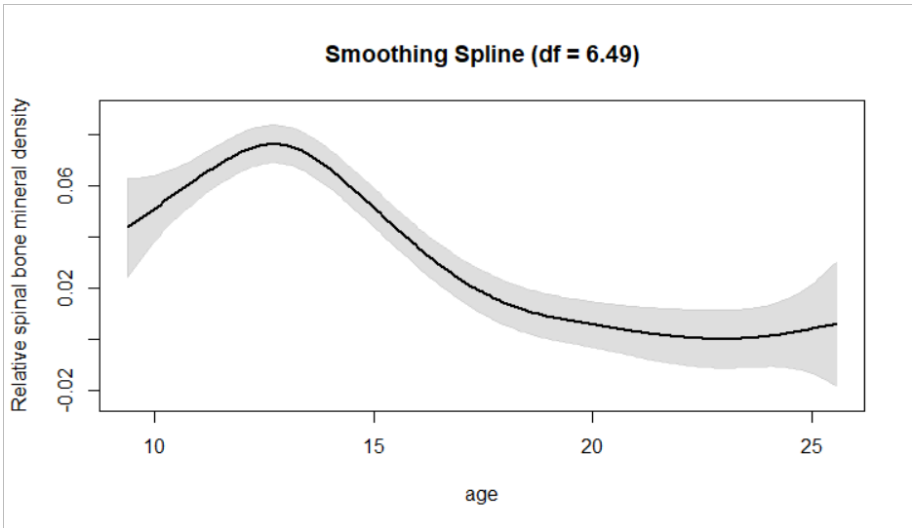
Below is the summary for the spline smoothing model from above

```
Call:
ss(x = age, y = rspnbmd)

Residuals:
      Min        1Q    Median        3Q       Max
-0.129536 -0.024683 -0.004216  0.018857  0.151170

Approx. Signif. of Parametric Effects:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02995    0.002017  14.852  0.00000 ***
x           -0.03772    0.015867  -2.377  0.01784   *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approx. Signif. of Nonparametric Effects:
             Df  Sum Sq  Mean Sq F value    Pr(>F)
s(x)       4.49 0.08294 0.018474   11.19 1.922e-09 ***
Residuals 478.51 0.78981 0.001651
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04063 on 478.5 degrees of freedom
Multiple R-squared:  0.338,    Adjusted R-squared:  0.3302
F-statistic: 43.38 on 5.49 and 478.5 DF,  p-value: <2e-16
```

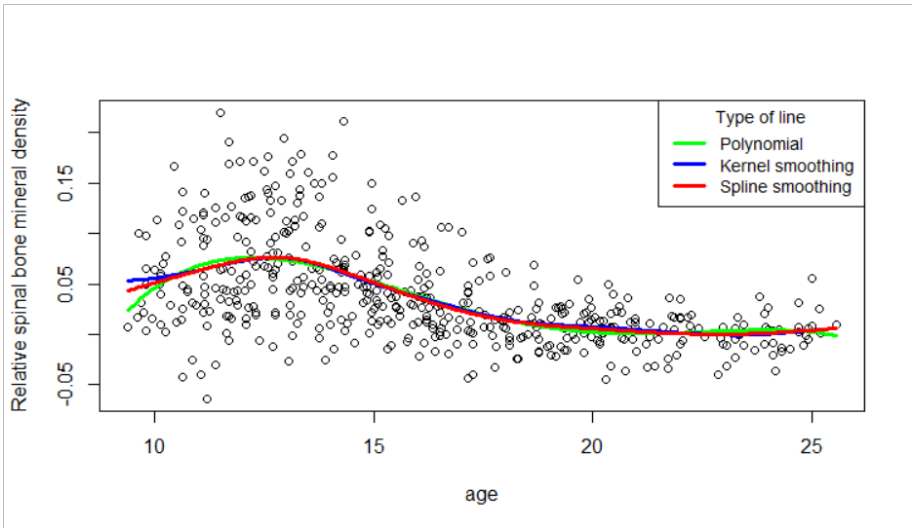
Plotting the spline smoothing line on the data scatterplot



Lastly, the predicted line with 95% bayesian confidence interval



It can be seen that spline smoothing produce a finer line than kernel smoothing. To compare the results from polynomial to smoothing regression in a plot :



From the comparison plot above, it can be seen that the difference between three type of regressions fitted into bone data seems to be very little. While one may think the polynomial regression is enough to model the bone data and thus keeping the model on parametric realm, writer would like to point out that “adjustment” to avoid violation of assumption is necessary. One of the solution is to try the data transformation such as stated in Mendenhall, W. (2012) book of regression analysis.

Lastly, for both of the nonparametric regression, writer would also like to remind that the interpretation is not as straightforward as in linear regression. Therefore, as of current writer’s knowledge, sticking to graphical explanation seems to be the safest route to use the smoothing regression results.

Reference :

García-Portugués, E. (2021). *Notes for Predictive Modeling*. Version 5.9.3. ISBN 978-84-09-29679-8. Available at https://bookdown.org/egarpor/PM-UC3M/. accessed on December 1, 2021

Helwig, N. E. (2021), *Nonparametric Regression (Smoothers) in R*, accessed accessed: http://users.stat.umn.edu/~helwig/notes/smooth-notes.html on December 1, 2021

Helwig, N. E. (2021), *Smoothing Spline Regression in R*, accessed at: http://users.stat.umn.edu/~helwig/notes/smooth-notes.html on December 1, 2021

Mendenhall, W. et. Al., 2012, *A Second Course in Statistics : Regression Analysis 7th edition*, Pearson Education, inc., Boston : United States

Prabhakaran, S., 2017, *LOESS Regression with R*, accessed at http://r-statistics.co/Loess-Regression-With-R.html on December 1, 2021