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

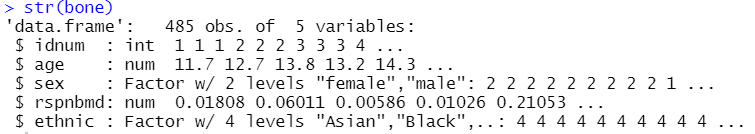
N. R. , Hakiim (2021)

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.



Table

Description automatically generated

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 :

(model 1)

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

Text

Description automatically generated

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 :

Text

Description automatically generated

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

(model 2)

Text

Description automatically generated

Proceeding to analyze the result of model 2 : it can be seen that even though the model is significant, the value of R2 (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

Chart, scatter chart

Description automatically generated

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

Chart, scatter chart

Description automatically generated

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 :

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

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 :

(model 3)

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

Text

Description automatically generated with medium confidence

Analyzing the summary, it is found that R2 after model improvement still doesn’t reach acceptable value for writer. Looking into the assumption, homogeneity is still the problem.

Chart, scatter chart

Description automatically generated

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 :

A picture containing text, watch

Description automatically generated Text

Description automatically generated

With represent the kernel function and 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 :

Text

Description automatically generated

Text

Description automatically generated with medium confidence

Generalized Cross-Validation

Text

Description automatically generated

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

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

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.

Chart, line chart

Description automatically generated

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 :

A picture containing text, watch

Description automatically generated

Where represent the smoothing parameter (equivalent to that of bandwidth in kernel) and is the penalty term. is the -th derivative of and being the function’s space which integrable at -th derivative.

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

Text, letter

Description automatically generated

Below is the summary for the spline smoothing model from above

Text, letter

Description automatically generated

Plotting the spline smoothing line on the data scatterplot

Chart, scatter chart

Description automatically generated

Lastly, the predicted line with 95% bayesian confidence interval

Chart, line chart

Description automatically generated

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 :

Chart, scatter chart

Description automatically generated

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