Homework Assignment 6

Total points: 112

Due: Friday October 28, 2022 at 5PM

Please email your answer (compiled pdf file from R markdown) and R code to Yen-Yi Ho (hoyen@stat.sc.edu).

Instructions: feel free to discuss the homework with other students. However, each student must conduct their own analyses and write-up their own solutions. Write as if for a scientific journal. Be brief and accurate.

Use the WHO Child Growth Standards (IGROWUP) data for child age **0-6** to study the dependence of weight on age with and without adjustment for height. IGROWUP data is available http://people.stat.sc.edu/hoyen/Stat704/Data/survey.csv.
More information about IGROWUP data can be found in
http://www.who.int/childgrowth/en/

I. Interpreting Simple and Multiple Linear Regression Coefficients

1. Make an exploratory plot of weight versus age. Comment in a few words on the relationship you observe. (3 points)
2. Fit a multiple linear regression model (MLR) of weight on age and height. Compare the coefficients and confidence intervals for age from the simple linear regression model (SLR) and MLR. (3 points)
3. Compare three age slopes: (1) using all the data estimated from simple linear regression; (2) using all the data estimated from the MLR with age and height; (3) the mean of the age slopes for the ten deciles of height (from Q10 in Homework 4). Given your findings, explain in a sentence or two, the meaning of the MLR coefficient for age. Do not use statistical jargon. (5 points)
4. The MLR assumes the relationship of weight on age is the same in each decile of height. Check the table from Q10 in Homework 4 to see if this is reasonable. (5 points)

II. Modelling Non-linear Relationship with MLR

For the task below, use the WHO Child Growth Standards data for children age **0-6** years.

1. Create a set of dummy variables to represent age in **bi-month age bins**. Calculate the mean weight for each **bi-month** age bin. Plot weight against agemons; add the mean values for each **bi-month** age bin (with bold symbols and a connecting line) to highlight the trend. (10 points)

2. Regression weight on age and add the least squares line to the plot. Plot the residuals from this linear regression against age. Comment in one sentence on the adequacy of a linear assumption for "growth." (5 points)
3. Linear Splines: (a-c, 6 points total
    a. Create five new linear spline variables with knots at (6, 12, 24, 36, 48 month), for example: age_sp1= $(\text{age} - 6)^{+}$=age-6 if agemons > 6, 0 if not; age_sp2= $(\text{age} - 12)^{+}$=age-12 if agemons > 12, 0 if not; … age_sp5= $(\text{age} - 48)^{+}$=age-48 if agemons > 48, 0 if not.
    b. Regress weight on age, age_sp1, age_sp2, …age_sp5.
    c. Plot the weight data with the fitted values from this regression added. (Add fitted values from linear splines in the figure plotted in II.Q1, Q2)
    d. Interpret the meaning of the coefficients from the first 2 "linear spline" terms: age_sp1, age_sp2. (6 points)
    e. Comment in a few sentences on the evidence form this analysis for or against a linear growth curve. (5 points)
4. Cubic splines: (6 points)
    a. Create new variables for cubic splines with knots at (6, 12, 24, 36, 48 month), for example: age^2, age^3 and age_csp1= $[(\text{age} - 6)^{+}]^3$, …etc.
    b. Regress weight on age, age^2, age^3, age_csp1, …
    c. Plot the weight data with the fitted values from this "cubic regression spline" added along with the fitted values from the straight line, **bi-monthly** means, and linear model.
5. Complete all but the last column in the table below (10 points)

Model	Degrees of freedom	Residual sum of squares	Residual mean square	AIC	Cross-validated residual mean square prediction
Linear	2				
Bi-monthly means					
Linear spline					
Cubic spline					

Comment on which of these models are most faithful to these data using both the plot in 4c and the table above. (10 points)

6. Cross-validated prediction error. Divide the data into 3 (we usually use more) random subsets of roughly equal size. (3 points) To calculate a "cross-validated" mean squared error for each model, execute the following steps.
   a. Leave out the first third and use the remaining two-thirds to fit each of the four models above. Use the fitted models to predict the weights for the left-out third. Note we are using different kids to fit the model and to assess the quality of the model predictions. (5 points)
   b. Save the sum of the squared deviations between the observed and predicted children's weights for each model. (5 points)
   c. Repeat this process two more times for the remaining thirds producing a sum of squared prediction errors for each model for each the three thirds. (10 points)
   d. Sum the total prediction errors across the three subsets and divide by the total number of children to obtain the cross-validated mean squared prediction error. Add this to the table above (II Q5). (5 points)
   e. Compare the cross-validated and ordinary mean square errors with each other and with the AIC. Which model is preferred under each criterion? (10 points)

[Hint: To make R code "cleaner", it is highly recommended to write tasks into functions]