Biostatistics 140.653
Third Term, 2021
Problem Set 1

Instructions: Feel free to work with other students on interpreting the questions posed in the problem set and analysis strategy and implementation (i.e. coding and model fitting). However, each student must write-up their own solutions. Write as if for a scientific journal. Be brief, precise, and jargon free. Create and submit a pdf with your report, please include your code. R users: use an Rmd file so we can easily see what code you used to generate your findings. Stata users: add your do-file to the end of your report.

Due in CoursePlus drop box date: February 11 by 5:00pm

For this problem set, use the Nepal Children's Anthropometry data to study the dependence of weight on age with and without "control for" height. Use only the first observation for each child and only those children with complete data on age, height and weight.

I. Interpreting Simple and Multiple Linear Regression Coefficients

Upon successful completion of this problem, a student will be able to:

- Make a publication-quality scatterplot to display the dependence of a response on an explanatory (predictor) variable using colors/symbols to represent subgroups
- Estimate and graphically display the results of a simple linear regression (SLR) and smooth nonparametric curve for reference
- Explain the objectives, assumptions and limitations of simple linear regression (SLR)
- Interpret the coefficients and residual standard deviation for a SLR
- Explain how the interpretation of a SLR coefficient changes when another predictor is added to the model
- Graphically represent the dependence of a response on a predictor variable after "controlling" for other explanatory variables
- Estimate and correctly interpret the coefficients from a multiple linear regression (MLR)

1. Using only the data from the first measurement time for each child, plot weight against age as if for an international nutrition journal. Label the axes clearly and make sure that all observations can be seen. Jitter the data or use different levels of transparency as necessary. Use different colors for the plotting symbols for boys and girls. Add a smooth curve (e.g. natural spline with ~3 degrees of freedom or loess with span =0.5 or kernel smoother with bandwidth 20 months) to the plot to emphasize the relationship of the observed mean weight at each age without making a stronger parametric assumption (e.g. linearity). Familiarize yourself with

how each of these smoothers works. Now make the curves separately for boys and girls.

2. Fit the simple linear regression of weight on age. In a few sentences, as if for a public health audience, interpret the: intercept, slope, and residual standard deviation in **anthropometric** terms. Include the estimates and confidence intervals in your sentences to be quantitative but use no statistical jargon (e.g. “intercept”, “slope”). For example, use “difference in average weight among children one year older” rather than “slope”.

3. Now display the three variables age, weight, and height so that you can better understand their joint distribution.

```
install.packages("scatterplot3d")
install.packages("rgl")
library(rgl)
library(scatterplot3d)
#
plot3d(d$age,d$ht,d$wt)
scatterplot3d(d$age,d$ht,d$wt,pch=16,type="h",highlight.3d=TRUE,xlab="age
(months)",ylab="height (cm)",zlab="weight (grams)",main="Nepal Children's Study")
pairs(d)
```

4. Conduct a multiple linear regression of weight on age and height. In a few sentences, as if for a public health audience, interpret the intercept, age coefficient, and residual standard deviation in **anthropometric** terms. Include the estimates and confidence intervals in your sentences to be quantitative but use no statistical jargon (e.g. “intercept”, “slope”).

5. In a few sentences, compare the coefficients and confidence intervals for age from the SLR and MLR and explain differences in their interpretations and estimated values.

6. Draw a directed acyclic graph (DAG) showing the likely causal relationships of aging on height and weight.

II. Modeling Non-linear Relationships with MLR

Upon successful completion of this problem, a student will be able to:

- Use linear, cubic and natural cubic regression splines to describe a non-linear relationship between a response and continuous predictor variable.

1. Linear splines:
    a. create three new variables:
        age_c = age - 6
        age_sp6 = $(age - 6)^+$ = age – 6 if age > 6, 0 if not
        age_sp12 = $(age - 12)^+$ = age – 12 if age > 12, 0 if not
    b. Regress weight on age_c, age_sp6 and age_sp12
    c. Plot the raw weight against age data; add the fitted values from this regression.
    d. Interpret the meaning of the coefficients for the three terms: age_c, age_sp6 and age_sp12 as if for a growth journal.
    e. Comment in a few sentences on the evidence from this analysis for or against a linear growth curve

2. Cubic regression splines:
    a. create three new variables:
        age2 = $(age - 6)^2$
        age3= $(age-6)^3$
        age_csp1 = $[(age-6)^+]^3$
    b. Regress weight on age_c, age2, age3 and age_csp1.
    c. Plot the weight data with the fitted values from this "cubic regression spline" added along with the fitted values from the linear spline.
    d. Contrast your estimated curves using linear and cubic splines.

3. Natural cubic splines
    a. Read about natural splines (*ns(x,df)*) to learn how they differ from regression splines. Both are linear regressions.
    b. Regress weight on the natural spline *ns(age,df=3)*.
    c. Obtain the design matrix (call it X) for this linear regression. (Use the R command *model.matrix*). Calculate the "hat" matrix $H= X(X'X)^{-1}X'$ that takes its name because the vector of predicted values in the regression ("Y-hat") is given by the matrix product HY where Y is the vector of observed responses. That is, the j^{th} predicted value is a linear combination of all the responses Y with weights given by j^{th} row of H. Choose three children from the data with different ages. On a single graph, plot each child's row of H against age. Comment on patterns you observe; i.e. what values of Y are most informative for each child's predicted value?
    d. Plot the weight data as above in 2c. Add the fitted values from this "natural cubic spline" along with the fitted values from the linear spline and cubic regression spline. Contrast your estimated curves.

III. Selecting Among Competing Models Based Upon Cross-validated Prediction Error

Upon successful completion of this problem, a student will be able to:

- Understand the importance of estimating a model and assessing its prediction error using different sets of data.
- Explain cross-validation as a method for unbiased estimation of prediction error

For each of the models above, we used 3 or 4 degrees of freedom for age, thereby allowing the relationship of average weight to be a non-linear function of age. The question is how many degrees of freedom are optimal to predict weight given any one of these methods? In this question, we use cross-validation to choose the degrees of freedom, df=1,…,8 within the natural spline family.

1. Randomly split the observations into 10 categories
2. For each df value, obtain the total cross-validated prediction error by regressing weight on ns(age, df), df=1,..,8, leaving out $1/10^{th}$ of the observations and summing the squared prediction errors for the left out values across the 10 "leave-out" iterations.
3. Plot the total cross-validated prediction error against the degrees of freedom to see which of the df values results in the best predictions of data, not also used to fit the model.
4. Compare the cross-validated prediction error to the non-CV prediction error for each df where the latter uses the same data to fit the model as assess its prediction error.
5. Fit this optimal model to all of the data; plot weight data against age, and add this optimal curve to the display
6. In a paragraph or two, summarize your findings as if for a public health journal. Explain the method you used and the results you found.