STATS 500, HOMEWORK #8, due Wednesday, December 1st, 3PM

1. In HW #7, Problem 2, we consider the `aatemp` data from 1881 to 2000 and use the following models when analyzing the data.

   - Orthogonal Polynomials to the 4-th degree.
   - Orthogonal Polynomials to the 5-th degree.
   - Orthogonal Polynomials to the 6-th degree.
   - Cubic spline with three internally evenly spaced knots (see the R-comments below)

   Using the AIC values from these models and the definition of AIC and BIC to obtain the BIC values for these models. Among these choices, what model would you recommend when using BIC as the criterion? What would you recommend when using AIC as the criterion?

2. Use the prostate data with “lpsa” as the response and the other variables as predictors. You can use

   ```
   > library("faraway")
   > library(leaps)
   > data(prostate)
   > help(prostate)
   > step(lm(lpsa~., prostate), k=log(97))
   > step(lm(lpsa~., prostate), k=2)
   > help(regsubsets)
   ```

   to learn about each variable in the data set and some of the useful R-functions.

   Implement the following variable selection methods to determine the “top” model(s) and comment on your answers:

   (a) AIC

   (b) BIC

   (c) Adjusted R2

3. Download the data file from the canvas site and use the modification of the R-codes here to add column names to the data. The data contains 80 observations on 9 variables. For this homework, please load both your answers and the R-codes (could be a Rmarkdown, a word or an ascii file) that you use to produce your outcomes.

   ```
   library(faraway)
   # use ’getwd()’ to find out where to save your file or use ’setwd’ to set your
   # directory that contains the data.
   #
   house.dat <- read.csv("house.csv", header=F)
   cName<-c("id", "price", "size","bath","halfbath","bed","age","garage","elem")
   colnames(house.dat)<-cName
   class(house.dat$bath)
   house.dat$bath<-factor(house.dat$bath)
   class(house.dat$bath)
   dim(house.dat)
   ```

The variables are:

- id = home ID number
- Price = sale price (thousands of dollars)
- Size = floor size (thousands of square feet)
- Bath = number of bathrooms
- Halfbath = number of half-bathrooms
- Bed = number of bedrooms
- Age = age (standardized)
- Garage = garage size (0, 1, 2, or 3 cars)
- Elem = nearest elementary school

Please focus on a model with Y = Price and contains the following variables to address the model selection questions. I will suggest a different model for inferences later.

```
bath*size+bed+halfbath+poly(age,2)+garage+elem
```

Would the model selection decision changed when the presence of influential points are taken into consideration? Please (a) report what are your two points with the largest cook's distances and do you consider them being influential, and (b) address the question using numerical outcomes produced with or without these two points. Please use AIC as the criterion and report the two models you selected (with and without the two data points).

To answer the following questions, please now consider a model: Y = Price and the predictors contain the main effects of "size", "bath", "bed", "halfbath", "age", "elem", and the interactions between "bath" and "size". Treat "bath" as a categorical variable. This may not be the model you select above but use it anyway.

(a) Focus on the two houses that have the largest and smallest residuals. Please provide the 95% prediction intervals for these two houses. Based on your answer, do you think the prices for these two houses follow the general regression model that applies to this data set? Please justify your answer.

(b) Someone argues that we should consider the quadratic function of "age". Do you agree with this suggestion? Please justify your answer with 0.1-level test.

(c) Are the regression slopes for "size" the same for houses with 1 bathroom and houses with 2-bathrooms? Are the regression slopes for "size" the same for houses with 2 bathrooms and houses with 3 bathrooms?
State the null and alternative hypothesis and explain your outcomes. Use α-level = 0.05 to make your conclusion. Feel free to drop the two houses with the largest cook's distances in this and the next problem. (hint: use help(relevel) to see the way to change the baseline to "2-bathroom".)

(d) If you are interested at buying a 3-bedroom house with two and a halfbath near elementary school "A".You like the size of the house to be 2.0 and age to be 0.0. (a) What should be the average price of the house you are interested? What would be the price range that are reasonable for such a house? Justify your answer.

You may find the following codes helpful to get you start.

```
#
#  Please note that you would need to modify the codes to what you need.
#  The codes are just for hints.
#
lm.aa<-lm(price~bath*size+bed+halfbath+poly(age,2)+garage+elem, data=house.dat)
summary(lm.aa)
cook <- cooks.distance(lm.aa)
Lc.ind<-order(cook, decreasing = TRUE)[1:2]
Lc.ind #find the index of the two largest values
step(lm.aa)

newh.dat<-house.dat[-Lc.ind,]
#levels(newh.dat$bath)
newh.dat$bath<-relevel(newh.dat$bath, ref="2")
lm.dd<-lm(price~bath*size+bed+halfbath+age+elem, newh.dat)
confint(lm.dd)

predict(lm.dd, data.frame(size=2, age=0 ,elem= "A", bed=3, bath="2",
halfbath=1),level=0.95,interval="prediction")
```