STAT 700
Homework 2 Problems
due Wed. Sept. 19

3 Problems. Show all work.
Please follow the Lab report directions off the homework web page for R Problems.
Please work in HW Groups!

1. Consider the linear model from class,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Assume that the ε_i are independent $N(0, \sigma^2)$ random variables or equivalently

$$\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n).$$

Also, assume that $\boldsymbol{X}'\boldsymbol{X}$ is invertible.

Let $\boldsymbol{e}$ be the vector of residuals (this is different from $\boldsymbol{\varepsilon}$) and $\widehat{\boldsymbol{Y}}$ be the vector of fitted values from performing OLS (ordinary least-squares). Show that $\boldsymbol{e}$ and $\widehat{\boldsymbol{Y}}$ are orthogonal to each other, in the sense that

$$\boldsymbol{e}'\widehat{\boldsymbol{Y}} = 0.$$

2. **Forbes' data**. James D. Forbes a Scottish physicist in 1857 collected data to see if the simpler measurement of the boiling point of water could be substituted for a direct reading of barometric pressure. The data are measurements in the Alps and Scotland from a barometer and a thermometer. Boiling point measurements were adjusted for the difference between ambient air and a standard temperature. The data are for $n = 17$ locations and measurements on $Temp$=boiling point (degrees Fahrenheit) and $Pressure$=corrected barometric pressure (inches of mercury).

The data is available off the class web page:
`https://edoras.sdsu.edu/~babailey/stat700/forbes.txt`

Use the R `read.table` command with the `header=T` option. You do not need to make your own labels!

(a) Make a scatter plot of $Pressure$ (Y) versus $Temp$ (X). Make an appropriate title for the plot.

(b) Would a straight line closely match the data? To answer this question fit the model $Pressure = \beta_0 + \beta_1 Temp + \varepsilon$. Include the R summary of the model fit from `lm`. Also, include the four R diagnostic plots. Do you detect any pattern in the residuals when you examine the plot of the residuals versus fitted values? Does the Normal Q-Q plot of the residuals indicate any departures from normality?

(c) Fit the model $\log(Pressure) = \beta_0 + \beta_1 Temp + \varepsilon$. Include summary and diagnostic plots from `lm`. Do you detect any pattern in the residuals when you examine the plot of the residuals versus fitted values?

(d) Construct 90% CIs for β_0 and β_1 for the model in part (c) using R commands. (Hint: there is a `confint` function in R.) Using your 90% CI for β_1, expain if the slope of the regression line is significant or not significant.

3. **GPA data**. (Ref: Graybill and Iyer (1994)) Consider the population of high school graduates who were admitted to a particular university during the a ten year time period and who completed at least the first year of course work after being admitted. We are interested in investigating how well the first year grade point average (GPA) can be predicted by using the following quantities with 20 students:

X_1 = the score on the mathematics part of the SAT (SATmath)
X_2 = the score on the verbal part of the SAT (SATverb)
X_3 = the grade point average of all high school mathematics courses (HSmath)
X_4 = the grade point average of all the high school English courses (HSenglish)

We will use data available off the class web page:
`https://edoras.sdsu.edu/~babailey/stat700/gpa.dat`

(a) Plot the data using the `pairs` function.

(b) Fit a linear model using all the predictor variables. Include summary and diagnostic plots from `lm`. How well does the linear model fit the data?

(c) Test whether the regression is significant at the 0.05 significance level. Be sure to state the null and alternative hypotheses.

(d) Use the R function `step` (or `drop1`) and the AIC model selection criteria to determine the "best" model. You should call the function with your lm object from part (b).

Examine the AIC values. If you want to drop just 1 variable from the full model (so you would have 3 variables included), which one would you drop? What is the AIC for this model?

Note: If you have trouble understanding the output, you can always fit each model and use the `extractAIC` function with each model as the argument and it will return the AIC value.

Is there an even a "better" model than the previous one, based on AIC. If so, what is that model and what is the AIC value?