STA 141A
Fall 2016
Homework 2

Due: November 2 (Wednesday)

Submit the assignment both electronically (through smartsite) and by submitting the printed copy. The electronic submission must be in the form of a zip folder (with extension .zip, .7z, etc.) containing two files: (i) data analysis and report; (ii) codes used to perform the data analysis.

Honor Code: *"The codes and results derived by using these codes constitute my own work. I have consulted the following resources regarding this assignment:" (ADD: names of persons or web resources, if any, excluding the instructor, TAs, and materials posted on course website)*

Description of the data NHANES Data (available in the smartsite folder "Homework/Data for Homework/" as `NHANES.Rdata`): National Health and Nutrition Examination Survey. It is part of `hexbin` package in R. This data frame contains the following columns:

1. **Cancer.Incidence**: binary factor with levels No and Yes.
2. **Cancer.Death**: binary factor with levels No and Yes.
3. **Age**: numeric vector giving age of the person in years.
4. **Smoke**: a factor with levels Current, Past, Nonsmoker, and Unknown.
5. **Ed**: numeric vector of $\{0,1\}$ codes giving the education level.
6. **Race**: numeric vector of $\{0,1\}$ codes giving the person's race.
7. **Weight**: numeric vector giving the weight in kilograms
8. **BMI**: numeric vector giving Body Mass Index, i.e., Weight/Height^2 where Height is in meters, and missing values (61% !) are coded as 0 originally.
9. **Diet.Iron**: numeric giving Dietary iron.
10. **Albumin**: numeric giving albumin level in g/l.
11. **Serum.Iron**: numeric giving Serum iron in μg/l.
12. **TIBC**: numeric giving Total Iron Binding Capacity in μg/l.
13. **Transferin**: numeric giving Transferin Saturation which is just 100 × serum.iron/TIBC.
14. **Hemoglobin**: numeric giving Hemoglobin level.
15. **Sex**: a factor with levels F (female) and M (male).

You need to use both graphical and analytical methods (such as linear regression and scatterplot smoother) to find interesting patterns in the data that may help in build predictive statistical models. The data analysis needs to be summarized in the form of a report (a maximum of 500 words). As a possible guideline, you may consider the following questions (you should try to come up with your own questions relevant to the data analysis). Also, the data has many missing values, which need to be properly accounted for while carrying out any analysis. Also, you may need to consider transforming certain variables and fitting linear regression models using original or transformed variables to build predictive models involving these variables.

1. How do the various continuous variables (including Age) relate to each other ?
2. Are their natural clusters in the data ? Are these clusters related to different natural strata determined by the values of the categorical variables like Ed, Race, Sex, Smoke ?
3. What kind of pattern is present in terms of the effect of Smoking on incidences of cancer ?
4. How do such patterns change across age groups, gender, race ?
5. How do the relationship of the variables change across strata as determined by factors such as Ed, Race, Sex, Smoke ?
6. Which subsets of variables are effective while predicting BMI ? Are there differential effects across educational level, gender or race ?

References: (Online versions are available through UC Davis Library.)

1. Dalgaard, P. (2008). *Introductory Statistics with R, 2nd Edition*. Springer.
2. Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer.