Cogs 109: Modeling and Data Analysis

Homework 1

Due Wednesday 10/4 in class

1. In a short paragraph (3-5 sentences), identify one problem or challenge that could be addressed, at least partially, through:
    a. Predictive modeling
    b. Inference
    c. Clustering (unsupervised learning)

    For example, these might be a scientific problem from one of your previous classes, a social or political challenge, or even a situation arising in sports. Explain (briefly) how statistical analysis or data modeling might be helpful.
2. ISLR problem 2.1
3. ISLR problem 2.7
4. Applied exercise: Download the data set **Income2.csv** from the textbook's website (http://www-bcf.usc.edu/~gareth/ISL/data.html). Load this data set into your favorite data analysis software environment (MATLAB, Python or R). In MATLAB, you could use the commands `readtable` or `csvread`. NOTE: Please include your code
    a. Make a scatter plot showing years of education on the x-axis vs. income (in thousands of dollars) on the y-axis. Make sure to label the x and y axes (in MATLAB, use the functions `xlabel` and `ylabel`).
    b. Calculate the mean income level for this data set
    c. Calculate the standard deviation of the income level
    d. Calculate the standard error of the mean (SEM)
    e. Create a new categorical variable called `HigherEd`. This variable is defined to be 1 if the subject has ≥16 years of education, and 0 otherwise. Make a box plot comparing the income level of subjects with `HigherEd=0` vs. `HigherEd=1`.

    Hint: In MATLAB, you can create a binary categorical variable from a continuous variable. For example:

```
>> x=[0:10]
x =
     0     1     2     3     4     5     6     7     8     9    10
>> x_categorical = (x>=5)
x_categorical =
     0     0     0     0     0     1     1     1     1     1     1
```