Cogs 109: Modeling and Data Analysis

Homework 1

Note: Each part of each problem was worth 2 points, unless noted otherwise.

1. In a short paragraph (3-5 sentences), identify one problem or challenge that could be addressed, at least partially, through the following techniques. 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.
    a. Predictive modeling

What is our climate future? Climate modelling to predict features of the global climate in 5, 10 or 50 years from now. Predictors could be current and historical data about atmospheric CO_2 concentration, latent water vapor, sea salt content and solar energy data, while the response variable could be average global temperature.

    b. Inference

What is the relationship between parent's income level and the future earnings of children? In attempting to understand questions of social mobility, one could investigate the relationship between parent's household income, years of education, access to welfare programs and future earning capacity of individuals. In this case the goal is to determine (1) whether there is a statistically significant effect of parents' income on children's earnings (yes or no), and (2) how strong is this effect?

    c. Clustering (unsupervised learning)

Are there subsets of patients with a particular disease (e.g. Autism Spectrum Disorder, ASD) who may have different underlying biological pathologies? Identifying subgroups of patients has been very helpful for cancer research and treatment, where different types of tumors are known to respond to different treatments. If the same could be done for complex neuropsychiatric conditions, it might help improve our understanding of these diseases.

Are there subsets of consumers with distinct characteristics who may respond differently to advertising? For example, can we identify subsets of individuals based on demographics (age, zip code, gender) or other available information (number of Twitter followers…) who respond best to different types of ads.

2. ISLR problem 2.1

a) Flexible – Large number of data points (n) and small number of predictors (p) means that models are less susceptible to overfitting/variance. A more flexible could capture more complex relationships in the data and reduce bias.

b) Inflexible – Large number of predictors (p) and a small amount of data points (n) means models are more susceptible to overfitting, so to guard against this you would want an inflexible model. This will reduce variance.

c) Flexible – A highly non-linear model is unlikely to be described well by an inflexible (more linear) function, so it is more suitable to use a flexible model. Of course, this requires having sufficient data to be able to reliably fit the parameters of the flexible model.

d) Inflexible – If the variance in the noise is large you want to guard against using a flexible model that could overfit the data, as such one should use a more inflexible model.

3. ISLR problem 2.7

a) Note that Euclidean distance between a point (x_1, x_2, x_3) and the test point $(0, 0, 0)$ is

$$d = \sqrt{(x_1 - 0)^2 + (x_2 - 0)^2 + (x_3 - 0)^2} = \sqrt{x_1^2 + x_2^2 + x_3^2}$$

Obs.	Euclidean Distance
1	3
2	2
3	$\sqrt{10} = 3.16$
4	$\sqrt{5} = 2.24$
5	$\sqrt{2} = 1.41$
6	$\sqrt{3} = 1.73$

b) Green - as Obs. 5 is the closest (smallest Euclidean Distance) and Y=Green for that observation.

c) Red - as Obs. 6 and 2 each vote red and only 5 votes green

d) Small values for K -> can create a less smooth decision boundary

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`.
    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`/

In [12]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

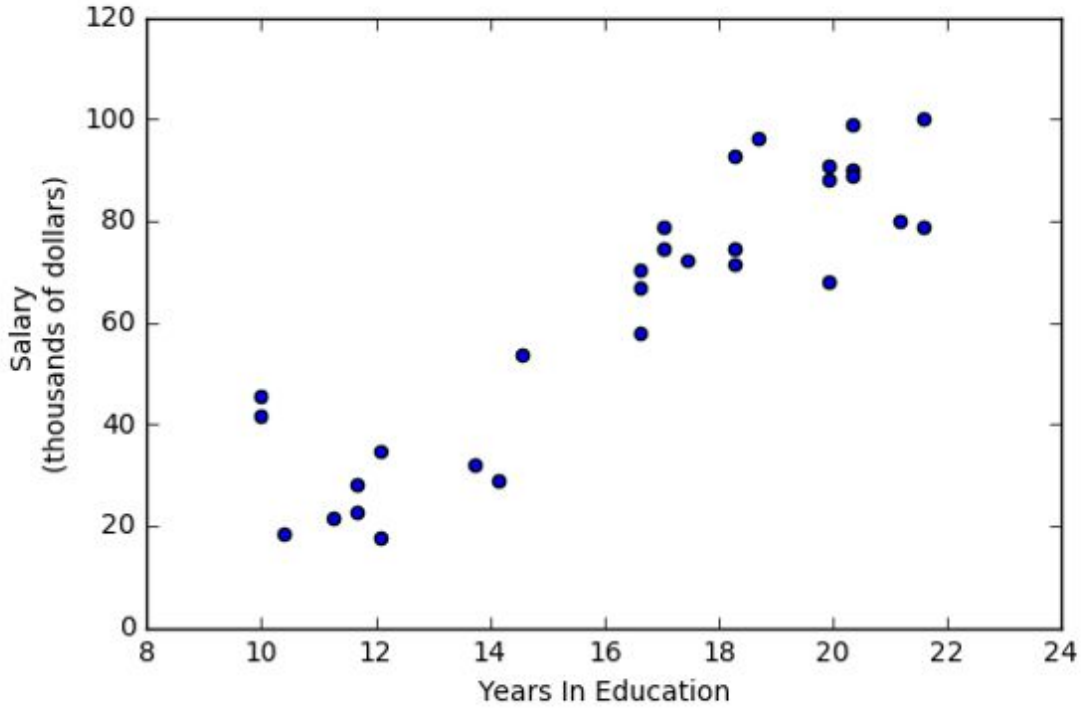
In [13]:

```
income = pd.read_csv('./Income2.csv', index_col=0)
```

[2 Points] a)

In [16]:

```
plt.scatter(income.Education, income.Income)
plt.xlabel('Years In Education')
plt.ylabel('Salary \n(thousands of dollars)')
plt.show()
```



[3 Points] b), c) and d)

In [7]:

```
print('Mean income (thousands of dollars) = %.1f' % np.mean(income.Income))
print('Standard deviation of income (thousands of dollars) = %.1f' % np.std(in
come.Income))
print('Standard error of the mean (thousands of dollars) = %.1f' % (np.std(inc
ome.Income)/np.sqrt(len(income))))
```

```
Mean income (thousands of dollars) = 62.7
Standard deviation of income (thousands of dollars) = 26.6
Standard error of the mean (thousands of dollars) = 4.8
```

[2 Points] e)

In [8]:

```
higher_ed = income.Education.values >= 16
plt.boxplot([income.Income[~higher_ed], income.Income[higher_ed]], labels = ['
Not Higher Education', 'Higher Education'])
plt.ylabel('Income \n(thousands of dollars)')
plt.show()
```