ST505: Fall 2012 IE projects

Here is a description of projects for the IE component of ST505. In a few cases where text listings of the variables were immediately available, I listed them. But, this is not meant to indicate that these data are necessarily more interesting than the others. I may combine items 3 and 4 into one project. There will most like be 8 projects with 7 groups of three and one group of two. The number of projects probably needs to be kept to 8 due to time constraints when projects are presented at the end of the semester.

The last two (with a *) are tentative as I'm not positive I can get access to the raw data.

- On a separate sheet, rank the items from 1 to 13 in terms of your interest. 1 = most want to work on, etc. Give this to me by Wed., Oct. 3 at the latest.

I also have the 1985 CSFII data, a national survey which records food intakes of various types (fat, protein, ...) for many individuals over two different days and some socioeconomic variables. I'm not sure if this would work as a project here but if this is something you would really like to work on, let me know.

1. MCAS math scores

   This projects involves data from the 2003 10th grade MCAS test scores for many individual high schools in Massachusetts (data courtesy of former M.S. student Eric Simoneau). The main response variable is the proportion of students passing and a vast array of socio-economic factors associated with each school are available The goal to analyze the effects of socio-economic factors on the passing percent

2. Baseball salaries

   The data consist of three files consisting of data on the regular and leading substitute hitters in 1986, the regular pitchers in 1986 and the team statistics. The salary data were taken from Sports Illustrated, April 20, 1987. The salary of any player not included in that article is listed as an NA. For example the hitter files consists of the variables listed below. The main goal is to see what seems to influence salary. We may just use hitters data or hitters and pitchers both.

   ```
   -     hitter's name,
   -     number of times at bat in 1986,
   -     number of hits in 1986,
   -     number of home runs in 1986,
   -     number of runs in 1986,
   -     number of runs batted in in 1986,
   -     number of walks in 1986,
   -     number of years in the major leagues,
   -     number of times at bat during his career,
   ```

```
-     number of hits during his career,
-     number of home runs during his career,
-     number of runs during his career,
-     number of runs batted in during his career,
-     number of walks during his career,
-     player's league at the end of 1986,
-     player's division at the end of 1986,
-     player's team at the end of 1986,
-     player's position(s) in 1986,
-     number of put outs in 1986,
-     number of assists in 1986,
-     number of errors in 1986,
-     1987 annual salary on opening day in thousands of  dol-
-     lars,
-     player's league at the beginning of 1987,
-     player's team at the beginning of 1987.
```

3. Wages and gender

   This uses data from 1985 current population survey to asses the impact of gender on wages after accounting for other variables. These data consist of a random sample of 534 persons from the CPS, with information on wages and other characteristics of the workers, including sex, number of years of education, years of work experience, occupational status, region of residence and union membership.

4. Sex discrimination data,

   Data shows starting salaries for a sample of males and females as well as measures of seniority, age, education and experience. The goal, similar to the previous item (these two may possibly be combined into one project) is to isolate the gender effect after accounting for other variables.

5. Pollution and mortality

   The data is from the U.S. Department of Labor Statistics on 60 Standard Metropolitan Statistical Areas (a standard Census Bureau designation of the region around a city) in the United States, collected from a variety of sources. The data include information on the social and economic conditions in these areas, on their climate, and some indices of air pollution potentials, as follows:

```
1.city: City name
2.JanTemp: Mean January temperature (degrees Farenheit)
3.JulyTemp: Mean July temperature (degrees Farenheit)
4.RelHum: Relative Humidity
5.Rain: Annual rainfall (inches)
```

```
 6.Mortality: Age adjusted mortality
 7.Education: Median education
 8.PopDensity: Population density
 9.%NonWhite: Percentage of non whites
10.%WC: Percentage of white collar workers
11.pop: Population
12.pop/house: Population per household
13.income: Median income
14.HCPot: HC pollution potential
15.NOxPot: Nitrous Oxide pollution potential
16.SO2Pot: Sulfur Dioxide pollution potential
17.NOx: Nitrous Oxide
```

The goal of this project is to examine how the age adjusted mortality rate potentially depends on the other (predictor) variables and in particular the effects of various pollutants.

6. Hospitals and infections. There have been increasing problems with individuals picking up infections during a hospital stay. This project will use the SENIC data from the book to examine the relationship between the probability of infection and hospital characteristics.

7. Gypsy moth, mice and acorns.

   This is from a study examining the relationships among gyspy month, mice and acorn abundances. A series of stands on the Quabbin reservoir were examined over a number of years with measures of abundance of each of the three variables. The goal is to model the mouse population as a function of the other variables (maybe current, maybe previous) and the previous years mouse population and to explore other relationships among the variables.

8. Voting patterns across counties in Florida for the 200 presidential election (and the potential outlying nature of Palm Beach county). This project will model the vote counts for candidates as a function of characteristics of the county (omitting Palm Beach) and then at the end examine whether the Palm Beach vote was outlying in some way.

9. Homeowner Insurance policies.

   This project will use data from a number of Chicago neighborhoods to examine how the voluntary issuance of insurance policies relates to race and other variables (fire rates, theft rates, income, and age).

10. Physical activity

    This project will use the baseline data from the Open Door to Health Study with physical activity as the outcome and numerous demographic variables as predictors.

11. Air pollution and traffic volume

    The data are a subsample of 500 observations from a data set that originate in a study where air pollution at a road is related to traffic volume and meteorological variables, collected by the Norwegian Public Roads Administration. The response variable consist of hourly values of the logarithm of the concentration of NO2 (particles), measured at Alnabru in Oslo, Norway, between October 2001 and August 2003. The predictor variables are the logarithm of the number of cars per hour, temperature 2 meter above ground (degree C), wind speed (meters/second), the temperature difference between 25 and 2 meters above ground (degree C), wind direction (degrees between 0 and 360), hour of day and day number from October 1. 2001.

12. ** Banks charge different interest rates. This project will use two data sets, one with 2229 loans on new automobiles and the other on 5664 indirect loans through dealers, and look at how the interest rate relates to a number of variables (13 in total including loan size , young borrower or not, income, credit rating, years at current address, etc.)

13. ** The price of wheat. This will look at data used in a legal case that suggested the cash price of wheat was manipulated. A model will be built for how wheat price relates to three supply-and-demand variables (which will then be used to assess if the price for the period in question is in line with past behavior).