1. EPA Mileage Data

The U.S. Environmental Protection Agency provides estimates of fuel efficiency that are required to appear on the window stickers of new cars.

The fact that the phrase "your actual mileage will vary" has become part of the vernacular is testament to the fact that the numbers are in fact estimates of mileage. The estimates are arrived at from labaratory measurements of vehicle performance and a mathematical model.

At one point, EPA revised its model because consumers were complaining the estimates overstated fuel efficiency.

EPA publishes the data used as input to their models. A link of the numbers for 2009 is posted in the notes and handouts section of the web page.

If you are running R and connected to the internet, you can download this to a data frame called `epa1` with the following statement:

```
epa1=read.csv("http://www.sandgquinn.org/stonehill/MA225/notes/09tstcar.csv")
```

You can discover the field names using `str(epa1)`. Alternatively, there is a link pointing to a document with more information on the data fields in the notes and handouts section.

The task for this project is to find a linear model of the form:

$$Y \;=\; X\beta + e$$

Where X is a matrix consisting of data in the columns of `09tstcar.csv`. These can be continuous measures or factors (i.e., discrete or classification variables).

The assignment is to find the combination of columns that minimizes the total squared error,

$$\sum_{i=1}^{n}(Y - \hat{Y})^2$$

subject to the restriction that the rank of X (i.e., the number of linearly independent rows and columns) is six or less.

The data contains information for cars and trucks. The class will be divided into two groups, one will analyze the car data and the other the truck data.

The groups for this project are:

- Group 1 (cars) Corey, Cortney, Kathleen, Meghan
- Group 2 (trucks) Dan, Pat, Steve