Final Exam: Take-home Data Analysis Question

- This is an exam. Please work independently. Do not ask your classmates questions, send them to me instead. You can email (sternh@uci.edu) or phone (949-824-1568 office) me with questions.
- This data analysis is worth 30% of the final exam grade.

VEHICLE FUEL EFFICIENCY

**Background.** As gasoline prices have risen there is greater public interest in automobile fuel efficiency. The data here provide an opportunity to learn about the key factors influencing mileage of vehicles and to assess the relative fuel efficiency of different manufacturers/vehicles.

**Assignment.** Develop a model that explains and predicts city mileage for 2004 vehicles. It is expected *a priori* that heavier, larger, and more powerful vehicles will have worse mileage. It is also expected that certain vehicle types (e.g., sports cars and SUVs) and certain drives (e.g., all-wheel and rear-wheel) are less fuel efficient.

The goal is to develop a statistical model that can explain and predict the city mileage (miles per gallon) for the vehicles. Questions of interest include:

- the relative importance of different physical characteristics of the car and engine;
- any differences among vehicle types (sedans, sports car, SUV, etc.);
- any differences among front-wheel, rear-wheel, or all-wheel drive;
- specific vehicles that appear to be more fuel efficient than their like-sized counterparts.

Please write a short report that develops a model for explaining and predicting city mileage of 2004 vehicles. Your report should include a brief introduction, a discussion of your data analysis, and a summary addressing the questions above. The data analysis section should describe the steps that you took in developing your model (i.e., preliminary data examination, important regression findings, relevant diagnostics, etc.). You should include any tables/figures that you use to reach your conclusions (e.g., regression results, residual plots or diagnostics) but you do not need to include output for models that turn out to be irrelevant to your analysis. You should discuss any unusual observations and their impact on the results as well. I expect that 2-5 pages of text is enough of a report (this is referring just to the text part of your writeup). Be sure to summarize your results and indicate the model you would use for future predictions in that summary. (I have withheld 100 vehicles at random to test the models on.) Remember the report should integrate tables/output you want me to see. Additional output can be included in an appendix.

**Data Description.** The data in the file cars04.xls (cars04.txt) are from Edmunds.com, Inc (the company behind the motor vehicle website). The data are information on 307 new U.S. passenger cars and trucks for the 2004 model year; there are an additional 100 models from that year that have been withheld for test purposes. A list of variables with short descriptions is provided on the reverse. You should not use the name, price, cost, or hwyMPG variables in your analysis. (Note: The spreadsheet includes a vehicle name field that is omitted from the text file.)
VARIABLE LIST
name = name of vehicle
price = manufacturer’s suggested retail price
cost = cost to manufacturer
cityMPG = city miles per gallon (response of interest)
hwyMPG = highway miles per gallon
weight = vehicle weight in thousands of pounds
horsepower = engine horsepower in hundreds
engsize = engine size in liters
cylinders = number of cylinders
wheelbase = wheelbase measurement of vehicle (in hundreds of inches)
type = vehicle type (character variable), also available as indicators (with sedan as the reference group)
sports = indicator for sports car (1=Sports car, 0=other)
suv = indicator for sports utility vehicle (1=suv, 0=other)
wagon = indicator for station wagon (1=wagon, 0=other)
van = indicator for minivan (1=van, 0=other)
truck = indicator for pick-up truck (1=truck, 0=other)
drive = type of drive (character variable), also available as indicators
    (front-wheel drive (fwd) is the reference group)
awd = indicator for all-wheel drive (1=awd, 0=other)
rearwd = indicator for rear-wheel drive (1=rwd, 0=other)