Stat 215b (Spring 2004): Lab 1

B. M. Bolstad
bolstad@stat.berkeley.edu

Due Feb 12, 2004 by 2:30pm

Introduction

The aim of this lab is to get you familiar with fitting a linear model using R/S-Plus. It will require much more computation than lab 0. Remember that you need to prepare a formal lab report (printout of computer output is not sufficient). In this lab we will consider another automobile mileage dataset. You will find the dataset on the website. Please note that it differs from the dataset used in the previous lab.

Data Description

The data comes from the April 1993 issue of consumer reports. We are interested in which factors have an effect on gas mileage. We have data on 82 cars. For each car we have 26 variables. Table 1 gives a full description of each of these variables.

Questions to be addressed

You should not answer these questions directly (ie in the form 1, 2, 3, ...) instead make sure your report covers the material discussed in each point.

1. Write a function to compute the OLS estimates given a design matrix and a vector of responses. Similarly write a function to compute the variance covariance matrix of the OLS. Use your functions to construct a coefficient table (Estimate, Standard error, t-statistic and P-value) if you fit the following model.

$$\frac{100}{\text{City MPG}} = \beta_0 + \beta_1 \text{Weight} + \beta_2 \frac{\text{Horsepower}}{\text{Weight}} + \epsilon$$

   Be sure to state your assumptions, why they are required and if they are justified.

   After this point you no longer need to use your functions from question 1. Feel free to use `lm` or another method of your own choosing.

Column	Description
1	Manufacturer
2	Model
3	Type: Small, Sporty, Compact, Midsize, Large
4	Minimum Price (in $1,000) - Price for the base version
5	Midrange Price (in $1,000) - Average of Min and Max prices
6	Maximum Price (in $1,000) - Price for the fully loaded version
7	City MPG (miles per gallon as rated by EPA)
8	Highway MPG
9	Air Bags standard [0 = none, 1= driver only, 2= driver and passenger]
10	Drive train type [0 = rear wheel drive, 1= front wheel drive, 2 = all wheel drive]
11	Number of cylinders
12	Engine size (liters)
13	Horsepower (max)
14	RPM (revolutions per minute at maximum horsepower)
15	Engine revolutions per mile (in highest gear)
16	Manual transmission available [0 = No, 1= yes]
17	Fuel tank capacity (gallons)
18	Passenger capacity (persons)
19	Length (inches)
20	Wheelbase (inches)
21	Width (inches)
22	U-turn space (feet)
23	Rear seat room (inches)
24	Luggage capacity (cu. ft.)
25	Weight (pounds)
26	Domestic? [0= non-U.S. manufacturer, 1= U.S. Manufacturer]

Table 1: Names of variables in car dataset

2. Produce an ANOVA table for the above model. Be sure to explain and interpret it.

3. Compute the diagonal of the hat matrix and use it to standardize the residuals. Look for patterns in the standardized residuals eg plot residuals against fitted values, other variables etc. Comment carefully on anything you discover.

4. Prepare a normal probability plot of the standardized residuals. Compare this to normal probability plots created using independent normal random variables. Comment on the results. Explain why we check for normality.

5. Comment on any outliers you may have found. Do you think that they affect your results? What happens to your model if you remove these observations?

6. Feel free to explore the data by adding additional variables to the model. Report your models and results.