36-401 Homework #2 Due 9/14/99

1. (**10 pts**) "fuel.dat" contains information on 60 cars from Consumer Reports, 1990. There are no column headings, so either add them to the file by hand, or do fuel_read.table("fuel.dat") followed by names(fuel)_c("Weight", "Disp", "Mileage", "Fuel", "Type"). Learn about the individual variables present in the file using summary() and other simple statistical commands plus appropriate graphs. Give two or three sentence descriptions of each variable (audience: your classmates).

2. (**15 pts**) Using the same data as problem 1, perform a simple regression of outcome variable Mileage on covariate Weight. Use the command "fuel.lm_lm(Mileage~Weight, data=fuel)". Turn in plots of Mileage on Weight, observed vs. fitted and residual vs. fitted. You will want to use fuel.lm$fit and fuel.lm$res. You can use abline(fuel.lm) to add the regression line to a raw data plot, abline(h=0) to add a horizontal line through 0 to a plot, and abline(0,1) to add a line with intercept 0 and slope 1. (Note that the argument order is y~x for lm, but x,y for plot.) Use identify() to label the 3 most extreme residuals in the residual vs. fitted plot: enter "identify(fuel.lm$fit, fuel.lm$resid, row.names(fuel))", then left click on the 3 most extreme points, then middle click anywhere to return to the Splus command prompt. Write a couple of sentences telling what the plots show (audience: the average consumer).

3. (**15 pts**) Do problem 7 in RwG p. 60 using "lead.dat" **but instead** of using X=gas, use X=air (concentration of lead in room air). The file contains observations with missing data (coded as NA). Use "lead.lm_lm(blood~air, data=lead, na.action=na.omit)" and "summary(lead.lm)" to do this problem. Turn in the same plots as for the previous problem. Write a paragraph telling what the plots show and interpreting the coefficient values, the p-values, the residual standard error and the multiple r-squared value (audience: lead researcher who understands the basic concepts of regression).

4. (**10 pts**) Consider the error-free, non-linear model $Y = 5 + X^2/2$. Calculate the Y values for X=1,2,3,4. Plot the data and the regression line for the model $Y = \beta_0 + \beta_1 X + \epsilon$ using S-plus. Make a table with one row per observation, and columns labeled, X, Y, $X - \bar{X}$, $Y - \bar{Y}$, $(X - \bar{X})^2$, and $(X - \bar{X})(Y - \bar{Y})$. Total the columns. Turn in the graphs and the table, and a sentence explaining why the regression line is where it is. Also show that the linear regression slope can be obtained as

$$b_1 = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2}$$

and the intercept is

$$b_0 = \bar{Y} - b_1\bar{X}.$$

Explain why the numerator of this b_1 formula makes sense.