36-401 Homework #2 Due Thursday Sept 20

1. (**10 pts**) The data set "Prestige.dat" (available from the data section of the 401 website; see also "Prestige.cbk") is extracted from the 1971 Census of Canada, and reports various attributes (average income, education, prestige, etc.) of 102 occupations. There are no column headings, so either add them to the file by hand, or do something like

```
prestige _ read.table("Prestige.dat")
names(prestige) _ c("education", "income", "pct.female",
   "occ.prestige", "occ.code", "occ.type")
```

   Learn about the individual variables present in the file using `summary()` and other simple statistical commands plus appropriate graphs (e.g. `rwgbox()`, `symplot()`, and either `qqn()` or `qqenv()`). *Turn in, for each variable: (a) numerical summaries; (b) the most informative plot(s); and (c) a two or three sentence description of each variable (audience: your classmates).* [Use information from "Prestige.cbk" as appropriate.]

2. (**10 pts**) Using the same data as in #1, perform a simple regression of the outcome `occ.prestige` on the predictor `income`. Use commands like

```
attach(prestige)
prestige.lm _ lm(occ.prestige ~ income)
```

   Turn in a scatter plot of x =`income`, y =`occ.prestige`, with the fitted regression line overlaid; and a residual plot (y =residuals, x =fitted values from the regression). You can use abline(prestige.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$\sim$x for lm, but x,y for plot.) Use identify() to label the 3 most extreme residuals in the residual plot. Use something like

```
plot(fitted(prestige.lm),residuals(prestige.lm))
abline(h=0)
identify(fitted(prestige.lm),residuals(prestige.lm),
     row.names(prestige))
```

   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: a labor historian who is not too familiar with regression analysis, but is interested in how various occupations stacked up in Canada in the late 60's and early 70's).

[PLEASE TURN OVER]

3. (**10 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 problem #2. 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**) Return to the `prestige` data set. Examine `income`, and various simple transformations of it, with a histogram and quantile-normal plot. Among the variables `income`, `sqrt(income)` and `log(income)`, which exhibits the best symmetry or normality properties? *Note: The X variable does not have to be normally distributed in a regression analysis, but skewing in either X or Y may suggest a transformation that improves the regression model.*

5. (**10 pts**) Re-do problem #2 using the best transformation you found in #4.

   Note: suppose your transformed `income` variable is called `NewIncome`. Then you can make either of two plots that overlays the regression line on the raw scatter plot:

   ```
   plot(NewIncome,occ.prestige)
   abline(lm(occ.prestige ~ NewIncome))
   ```

   or

   ```
   plot(income,occ.prestige)
   lines(sort(income),fitted(lm(pcc.prestige ~ NewIncome))[order(income)])
   ```

   Which do you prefer for examining the regression? Which do you prefer for communicating your results to a non-statistical collaborator? A similar decision arises for examining the residuals from this regression.

6. (**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.