Homework 2
Due: Thursday, September 13, 2001

1. Reading.

   - In RPD (the Rawlings, Pantula and Dickey regression text) you should read through all of chapter 1, on simple regression.
   - Work through the two Splus tutorials
     - Objects: 'Objects and Data Types in Splus"
     - Programming: "Programming in Splus"

   available in many places, for example, at `http://www.stat.cmu.edu/~brian/707/S-Tutorial`. You do not have to hand anything in for this question.

2. RPD, p. 30, Ex. 1.3.

3. (a) Show that $\sum \hat{Y}_i e_i = 0$, where $e_i = Y_i - \hat{Y}_i$.
   (b) Show that $\sum Y_i = \sum \hat{Y}_i$.
   (c) Use parts (a) and (b) to prove the ANOVA decomposition of the total sum of squares, S_{YY}. That is, demonstrate that $S_{YY} = SS_{Reg} + RSS$.

4. (a) Show that $\hat{\beta}_1 = r_{XY} \cdot \frac{S_Y}{S_X}$, where where r_{XY} is the sample correlation between the X's and the Y's, S_Y^2 is the sample variance of the Y's and S_X^2 is the sample variance of the X's.
   (b) Recall that $SS_{Reg} = \sum(\hat{Y}_i - \overline{Y})^2 = \sum(\hat{\beta}_0 + \hat{\beta}_1 X_i - \overline{Y})^2$. From this, show that $SS_{Reg} = \hat{\beta}_1^2 \sum(X_i - \overline{X})^2$, and therefore $R^2 = SS_{Reg}/SYY$ is equal to $(r_{XY})^2$.

The `Angell.dat` data set that we used last week came from a study by Angell (1951) on 43 US cities. The variables collected were:

- `Moral`: An index of "moral" or social integration, combining information on on the incidence of crime with data on social welfare expenditures.
- `Ethnic`: An index of ethnic heterogeneity, calculated from the proportions of nonwhites and foreign-born whites residing in the cities.
- `Geog`: An index of geographic mobility, constructed from the proportions of residents moving into and out of the cities.
- `Region`: Region of the US where the city is: (E)ast, (S)outh, (W)est, or (M)id(W)est.

Angell argues that the moral integration of cities should decrease with ethnic heterogeneity and geographic mobility. In the next several problems we briefly explore the relationship between `Moral` and `Ethnic`.

Homework 2

5. Regress `Moral` on `Ethnic`.

   (a) Plot the least-squares regression line on a scatterplot of the data, and comment on the results. Examine and interpret the values of $\hat{\beta}_0$, $\hat{\beta}_1$, $s = \sqrt{RSS/(n-2)}$, and r_{XY} (or R^2). You can use commands like

   ```
   myfit _ lm(Moral ~ Ethnic, data = angell)
   summary(myfit)
   plot(angell$Ethnic,angell$Moral)
   abline(myfit)
   ```

   (b) Examine and comment on the residuals for the regression of `Moral` on `Ethnic` (the residuals should be approximately normally distributed, without patterns, outliers, etc.). You can use commands like

   ```
   plot(angell$Ethnic,residuals(myfit))
   abline(h=0) # or abline(0,0)
   qqenv(residuals(myfit))
   ```

   or examine the first plot in each row of `par(mfrow=c(2,3)); plot(myfit)`. We don't know how to interpret the other diagnostic plots here yet, but you can learn a little more about them by consulting `help(plot.lm)`.

6. Examine `Ethnic`, and various simple transformations of it, with a histogram and quantile-normal plot. Among the variables `Ethnic`, `sqrt(Ethnic)` and `log(Ethnic)`, 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.*

7. Re-do problem #5 using the best transformation you found in #6.

   Note: suppose your transformed `Ethnic` variable is called `NewEthnic`. Then you can make either of two plots comparing the regression line to the data:

   ```
   plot(NewEthnic,Moral)
   abline(lm(Moral ~ NewEthnic))
   ```

   or

   ```
   plot(Ethnic,Moral)
   lines(sort(Ethnic),fitted(lm(Moral ~ NewEthnic))[order(Ethnic)])
   ```

   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.