36-401 Homework #5 Due Tuesday Oct 25, 2001

1. [20 pts] RwG problem 15, Ch 3, on page 104. The data is in `salamander.dat` in the 401 web pages. **Turn in** the table of coefficients and p-values, the graph, **and** an interpretation of the results in terms of the *substantive meaning* of the variables. The graph should mark the points separately for lunged vs lungless as well as showing the two regression lines. An Splus code example for this type of graphing is given in `spruce.q`.

   Note: Here are some ways of expressing models with dummy variables in Splus. Let X and Y be continuous, and D is a dummy (coded as 0 or 1).

```
(1) lm(Y ~ D + X)       # D used for an intercept dummy


(2) lm(Y ~ X + D:X)     # D used for a slope dummy, i.e. include
                        #    D times X as a new variable in the model

(3) Z <- D*Z            # same as model (2) but not using Splus's
    lm(Y ~ X + Z)       #    ``:'' notation

(4) lm(Y ~ D + X + D:X) # D used for both intercept and slope dummy


(5) lm(Y ~ D*X)         # Splus shorthand for model (4)
```

2. [30 pts] Return to the prestige data set. Construct a `newprestige` data set as follows:

```
> attach(prestige)
> newprestige _ data.frame(occ.prestige,occ.type,education,
+ LogInc=log(income),pct.female)
> detach(prestige)
> newprestige$Prof _ ifelse(prestige$occ.type=="prof",1,0)
> newprestige$WC _ ifelse(prestige$occ.type=="wc",1,0)
> newprestige$BC _ ifelse(prestige$occ.type=="bc",1,0)
> newprestige$Unknown _ ifelse(prestige$occ.type=="?",1,0)
```

   We want to review and examine the role of all these variables in predicting occupational prestige. Some suggestions follow (what to turn in is at the end of the problem):

   - Make a scatterplot matrix ("pairs()" or "mypairs()") of the first five columns of `newprestige`, and examine the correlations between these variables [Note: For the correlations, you may want to convert `occ.type` to numeric values, using `as.numeric(occ.type)`].
   - Review the regression of `occ.prestige` on the other three continuous variables in the data frame.
   - Review boxplots and regression of `occ.prestige` on the four dummy variables `Prof`, `WC`, `BC`, and `Unknown`.

- Find the best model for `occ.prestige`. (Note: you shouldn't use both `occ.type` and the dummies `Prof, WC, BC, Unknown`, since they are different ways of expressing the same categorical information).
- Consider adding any interactions that would make sense. You can either make interactions as you need them using ":" in the `lm` statement (e.g. `lm(occ.prestige ~ education + Prof:education)`), or by constructing new variables in the data frame (e.g. `newprestige$ProfEduc _ newprestige$Prof * newprestige$education`), depending on what is useful for what you want to do in Splus.
- For residual plots it may be useful to keep track of which points belong to which professions. One way to do this is, e.g.

  ```
  > plot(fitted(my.lm),residuals(my.lm),type="n") # axes only, no points
  > text(fitted(my.lm),residuals(my.lm),labels=newprestige$occ.type)
  ```

  and you can repeat this idea for other scatter plots that are useful for your analyses.

Turn in:

(a) A description of your final model or models including (i) the prediction equation with numeric coefficients (with an appropriate number of decimal places) and (ii) one to four paragraphs giving your interpretation of what the model tells us about the determinants of occupational prestige in this data set.

(b) A multi-panel plot or plots that can be used for assumption checking of your best model(s). Include a few sentences with your interpretation of the plots. Be sure to be clear which plots go with what.

(c) An appendix that is a diary or log of the work you performed, E.g. "I looked at a histogram of — ", "Variable selection using the BIC criterion was performed", "A — by — interaction was tried, but found not to be significant", etc., etc. You may (should!) intersperse raw Splus output and graphs with this to help justify what you've done. See the handout on Fuel Consumption Data from 9/29 for an example (your diary does not have to be this long; it should be as long as is needed to describe what you did).