36-401 Homework #4 Due Tuesday Oct 9, 2001

This problem set has 4 problems. For problems 1–3 we return to the `prestige` data set from hw's 2 and 3. Construct new variables for `prestige` and create a new data frame `newprestige` as follows:

```
> LogInc _ log(prestige$income)
> Prof _ ifelse(prestige$occ.type=="prof",1,0)  # professionals
> WC   _ ifelse(prestige$occ.type=="wc",1,0)    # white collar workers
> BC   _ ifelse(prestige$occ.type=="bc",1,0)    # blue collar workers
> Unknown _ ifelse(prestige$occ.type=="?",1,0)  # unknown
> newprestige _ data.frame(prestige[,c(4,1,3)],
+ LogInc,Prof,WC,BC,Unknown)
```

`LogInc` is the transformation of `income` that we liked in HW #2. The variables `Prof`, `WC`, `BC` and `Unknown` are called "dummy variables" because they are 0/1 dichotomies that code different levels of a categorical variable `occ.type`. You can see what they are doing by examining the output from `newprestige[c(1,28,31,34),c("Prof","BC","WC","Unknown")]`.

1. (a) Make a scatterplot matrix of the first four variables in `newprestige`, and also a correlation matrix for these for variables. Write a few sentences describing which variables may be related to `occ.prestige`. Are there any (approximate) collinearities evident among `education`, `income` and `pct.female`?

   (b) Summarize the result of regressing `occ.prestige` on the three variables `education`, `pct.female`, and `LogInc`. Turn in (i) the fitted regression equation, with standard errors for each regression coefficient; (ii) R^2, R^2_{Adj}, and F; (iii) a small set of appropriate diagnostic plots; and (iv) one sentence summarizing the evidence you have provided for goodness of fit of the multiple regression model.

   (c) Turn in the output from

   ```
   > sum.step(stepwise(newprestige[,2:4], prestige$occ.prestige,
   + method="exhaustive", nbest=2),
   + newprestige, "occ.prestige", includeAIC=T, includeR2=T, sort=0)
   ```

   (you will need to get `sum.step.q` from the 401 web site and do `source("sum.step.q")` to do this). Does the best model listed here agree with your intuition, based on the t statistics in part (b)?

2. [Continuation.] Consider the general regression model

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

   An alternative procedure for calculating the least-squares coefficient $\hat{\beta}_1$ is as follows:

   - Regress y on $X_2, \ldots, X_k$
   - Regress X_1 on $X_2, \ldots, X_k$
   - Regress the residuals $e_{y \mid x_2 \cdots x_k}$ on the residuals $e_{x_1 \mid x_2 \cdots x_k}$

   (a) Apply this procedure to the `newprestige` data set and the regression `occ.prestige ~ education + pct.female + LogInc` to obtain the coefficient for `education`. Is it the same as in the multiple regression in part (a)? *Hint:* Remember that `residuals(myfit.lm)` gives you the residuals from a fitted regression.

(b) Illustrate the regression in part (a) with an appropriate scatter plot, and comment on the usefulnesss of X_1 in the model, based on this scatter plot (this is the "partial regression leverage plot").

(c) Notice that the intercept in your regression in part (a) is 0. Why is that?

Note: In light of this procedure it is reasonable to describe $\hat{\beta}_1$ as the "effect of X_1 on y after removing the influence of $X_2, \ldots, X_k$", and one often speaks of tests of $H_0 : \beta_1 = 0$ as testing "whether X_1 has any effect on y after removing influence of $X_2, \ldots, X_k$".

3. [Continuation.]

(a) Make parallel boxplots of `occ.prestige`, split by `occ.type`, using the `prestige` data set. Turn in the graph, with informative title and other labels, and a sentence or two describing any important features of the graph.

(b) Summarize the multiple regression `occ.prestige ~ Prof + WC + BC + Unknown - 1` , using the `newprestige` data set. Turn in (i) the fitted regression equation, with standard errors for each regression coefficient; (ii) R^2, R^2_{Adj} and F; (iii) a small set of appropriate diagnostic plots; and (iv) one sentence summarizing the evidence you have provided for goodness of fit of the multiple regression model.

(c) In part (b), give the fitted values and 95% confidence intervals for the occupational prestiges of each of the four groups: Professionals, White collar workers, Blue collar workers, and workers of unknown type. Write a sentence or two relating these confidence intervals to the plots you made in part (a).

Remarks. This is exactly the same as the "cell means" parameterization of a one-way analysis of variance of `occ.prestige` using the four-level categorical variable `occ.type`.

(d) In part (b), what does the "`- 1`" do? Try the multiple regression `occ.prestige ~ Prof + WC + BC + Unknown`, and explain why this doesn't work (give the specific relationship among the four variables `Prof`, `WC`, `BC` and `Unknown` that illustrates your explanation).

4. The fuel consumption data set described in class is available as `fuel-alr.dat` in the data area of the 401 web pages; create a data frame from it using `fuel _ read.table("fuel-alr.dat",header=T)`. The variables are

```
(State) = Two-Letter State Code
    Pop = Population in each state
    Tax = 1971 Gas tax, cents per gallon
   Nlic = 1971 thousands of drivers
    Inc = 1972 per capita income
   Road = 1971 miles of primary highways
  Fuelc = fuel consumption, millions of gallons
   Dlic = 100*Nlic/Pop = proportion of licensed drivers
   Fuel = 1000*Fuelc/Pop = per-capita fuel consumption
```

(a) Make a scatterplot matrix of all of the variables in `fuel`, and also a correlation matrix for these variables (use `round(cor(fuel),4)` to make it fit well in 80 column output). Write a few sentences describing which variables may be related to `occ.prestige`. Are there any (approximate) collinearities evident among the variables?

(b) Summarize the multiple regression of `Fuel` on all other variables in the `fuel` data frame. Turn in (i) the output of the `summary()` command on this regression; and (ii) two or three sentences indicating which variables show symptoms of collinearity, and what the evidence is.

(c) Get `collin.q` from the data area of the 401 website, and do `source("collin.q")`. Then run `collin` on `fuel[,-8]` (the `fuel` data frame without the response variable `Fuel`). Turn in the output from this command, and a sentence or two relating this output to your conclusions in part (b).

(d) Use `sum.step` and `stepwise` (using `method="exhaustive"`) to find the best model according to BIC and the best model according to AIC. Examine the `summary()` output from each model, and decide which one you like better. For the model you like better, turn in: (i) the fitted regression equation, with standard errors for each regression coefficient; (ii) R^2, R^2_{Adj} and F; (iii) a small set of appropriate diagnostic plots; and (iv) one sentence summarizing the reason(s) you like this model better.