The dataset for today is ex1123.csv. I recommend you use PROC REG. The data step is as follows.

```
data AirPollution; length city $ 18;
   infile input1 missover DSD dlm="," firstobs=2;
   input CITY $ MORT PRECIP EDUC NONWHITE NOX SO2;
datalines;
run;
```

The variable “city” will not play a roll in this analysis, but may be useful as a graphics indicator and for locating individual observations.

1) Start by sorting the raw data by CITY and listing the raw data. The alphabetical order will make locating individual cities easier.
2) First fit the full model, “MODEL MORT = PRECIP EDUC NONWHITE NOX SO2;”. From this beginning, do the following.
3) On this full model, check the VIF (variance inflation factor). Do they indicate a problem?
4) If there are any non-significant variables (α=0.05) make a copy of the PROC REG above and remove the least significant variable. Do not delete the original full model above, leave it in your program. Repeat this step as often as necessary (one variable at a time) until all variables in the model are significant.
5) For the last model from part 4 (with all terms significant) output the values for “student, rstudent, cookd, leverage and dffits. The “keywords” are the same as those used in the preceding sentence except for leverage which has the keyword “h”. You are welcome to borrow code from my examples (see example on SAT scores).
As VREF reference values use “2.7” for student and rstudent (approximate 99% level of t for degrees of freedom around 56), use “1” for cookd and dffits and use “0.17” for leverage (approximate median level for 2*p/n where p is 4 to 6).
6) Plot these values. Again, borrowed code may save you some time.

Questions to answer in homework or email or on this page to be turned in (your choice).

Leave all models in your program and email me a copy. I will check to see which one you left as having all significant variables.

1) Did the VIF indicate problems? Indicate YES or NO

2) Were the atmospheric pollutants, NO_X and SO_2, implicated as having a possible correlation with mortality? In your answer specify which one or both. Of course, such a correlation does not prove a relationship. Give the P-value ______________________

3) Are there any problems suggested by the residual and influence diagnostics? YES or NO

Which one(s) indicate problems? ______________________________________

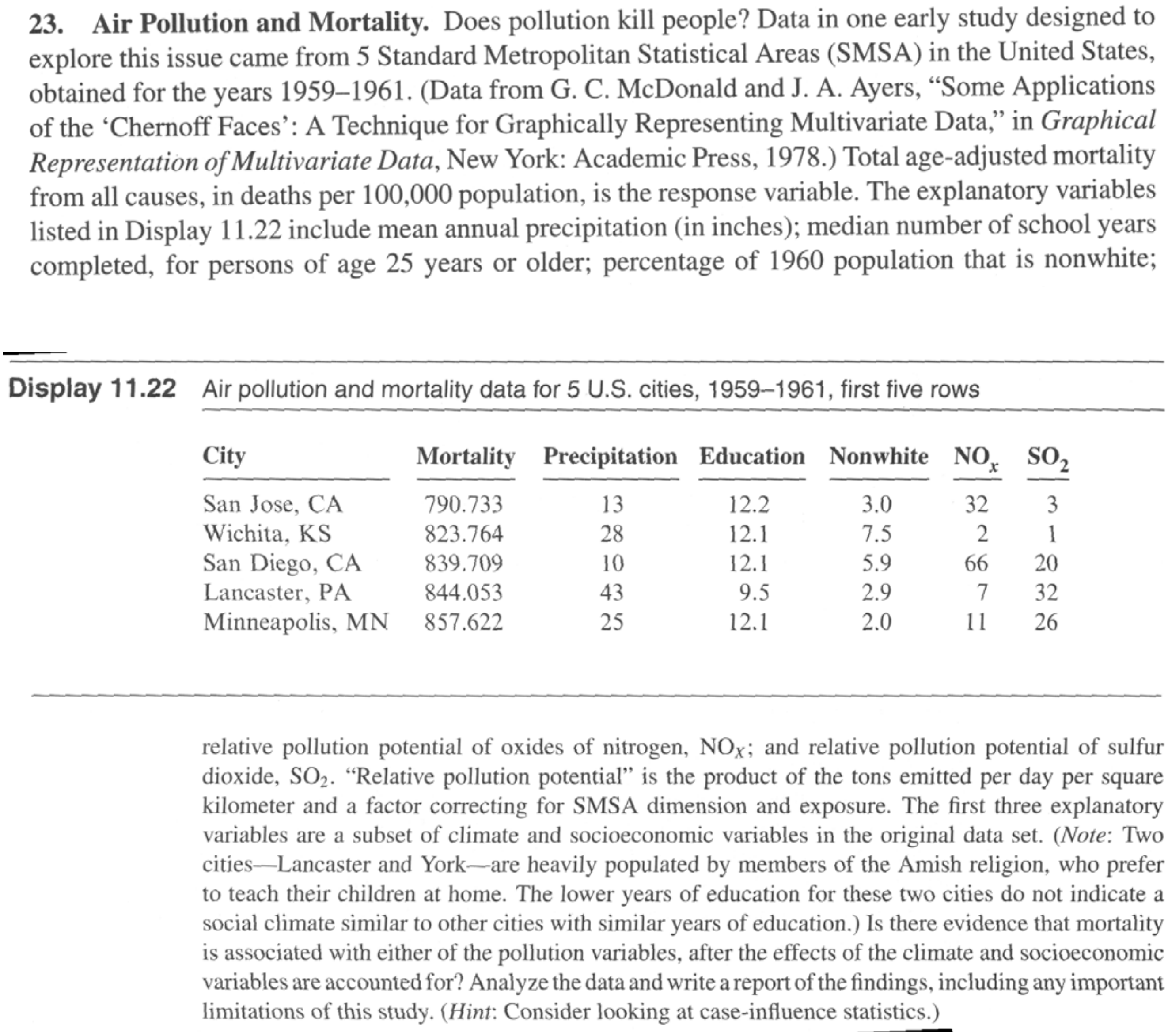
4) Which city had the largest RSTUDENT value? ______________________

5) Which city had the largest LEVERAGE value? ______________________

6) Which city had the largest DIFFITS value? ______________________

7) Which city had the largest COOKD value? ______________________

OVER



23. Air Pollution and Mortality. Does pollution kill people? Data in one early study designed to explore this issue came from 5 Standard Metropolitan Statistical Areas (SMSA) in the United States, obtained for the years 1959–1961. (Data from G. C. McDonald and J. A. Ayers, "Some Applications of the 'Chernoff Faces': A Technique for Graphically Representing Multivariate Data," in *Graphical Representation of Multivariate Data*, New York: Academic Press, 1978.) Total age-adjusted mortality from all causes, in deaths per 100,000 population, is the response variable. The explanatory variables listed in Display 11.22 include mean annual precipitation (in inches); median number of school years completed, for persons of age 25 years or older; percentage of 1960 population that is nonwhite; relative pollution potential of oxides of nitrogen, NO_X; and relative pollution potential of sulfur dioxide, SO_2. "Relative pollution potential" is the product of the tons emitted per day per square kilometer and a factor correcting for SMSA dimension and exposure. The first three explanatory variables are a subset of climate and socioeconomic variables in the original data set. (*Note*: Two cities—Lancaster and York—are heavily populated by members of the Amish religion, who prefer to teach their children at home. The lower years of education for these two cities do not indicate a social climate similar to other cities with similar years of education.) Is there evidence that mortality is associated with either of the pollution variables, after the effects of the climate and socioeconomic variables are accounted for? Analyze the data and write a report of the findings, including any important limitations of this study. (*Hint*: Consider looking at case-influence statistics.)

Display 11.22 Air pollution and mortality data for 5 U.S. cities, 1959–1961, first five rows

City	Mortality	Precipitation	Education	Nonwhite	NO_x	SO_2
San Jose, CA	790.733	13	12.2	3.0	32	3
Wichita, KS	823.764	28	12.1	7.5	2	1
San Diego, CA	839.709	10	12.1	5.9	66	20
Lancaster, PA	844.053	43	9.5	2.9	7	32
Minneapolis, MN	857.622	25	12.1	2.0	11	26

OVER