36-315: Statistical Graphics and Visualization

Final Project
Date: April 2, 2002 **Due: May 6, 2002**

The project as a whole is due at the end of the semester, but pieces of it will be due in weekly assignments, the first of which is due next Monday and described on the next page.

The project is to describe how a variable of interest (the 'response') varies with respect to other census variables in Pennsylvania. Ideally, you would like to be able to predict the response in a census tract given any subset of the other census variables. For example, you would like to be able to predict rent from population density, rent from the percentage of homeowners, or from both. This requires knowing which variables are relevant to the response, how they interact, and special cases. Your project should provide useful information toward this goal, describing how the response varies with several of the census variables, with special attention to results that are non-obvious or surprising. The results should be understandable to a layperson, so you should not merely report the coefficients of a multivariate regression.

Pick a response from one of the eight variable groups listed below, and explain its behavior using variables from at least four of the other seven groups. You want to pick variables and analyses that show something non-obvious. An increase in income with education would be obvious, but an increase in income with family type (for a fixed education level) would be non-obvious. Outliers and changes in predictor strength (as seen on contour plots) are often surprising.

You can study a state other than Pennsylvania, or you can use another dataset entirely. In either case, get our approval first (in the first assignment).

The report should follow a logical sequence and be reasonably concise (less than twenty pages). Throughout the report, you will probably have to make assumptions that cannot be verified from the available data. Make it clear which of your statements are assumptions and which come from the data.

You will probably find it useful to define your own tract categories, such as "desirable neighborhood", based on the values of several variables. Then your results might simplify into rules like "if the neighborhood is desirable and vacancies are low, then rent is high, barring the following exceptions ..."

The report will be graded on the following criteria:

Technique Are your graphs well-designed? This includes, for example, variable transformation, smoothing parameters, symbols, aspect ratio, and color palette. These are the same criteria that were used to grade homeworks.

Choice Are your visualizations the right ones for supporting your argument? Does each tell a different story, or are they redundant?

Interpretation Is each graph interpreted completely and correctly?

Depth The variety and amount of information conveyed by your graphs (excluding redudant graphs). You want to go beyond simple trends found in scatterplots.

Time requirements Your classes should not require more time per week than listed in the catalog, including and especially the last week of classes. You should plan to spend about 5 hours per week on the project. There are five weeks until the end of the semester, giving a total of 25 hours on the project. If you find yourself needing more time than this, let us know about it. Projects are to be done individually.

Other things to keep in mind:

1. The data is at the census tract level, not the individual level.
2. Percentages are usually more informative than raw counts.
3. Know what the variables are really measuring. For example, `PCTELEM` means elementary education *only*, and 'household income' increases with the number of working people in the household.
4. What looks like a simple trend may actually be several distinct clusters.
5. Outliers can be just as interesting as the main trend, and should be investigated.
6. Many associations that appear significant can be explained by a lurking variable.
7. Let the data guide you to a conclusion, not the other way around.

First project assignment
Date: April 2, 2002 **Due: April 7, 2002**

The first step of the project is to choose the dataset that you want to use, the response variable that you want to study, a set of relevant predictor variables, and questions to answer.

If you want to use a state other than Pennsylvania, note that the ten states with smallest populations are not allowed. A map of allowed states is posted on the course web page.

On the class web page, under "project information", you will find links to short and long descriptions of the census variables available. The 200+ census variables are arranged into eight basic groups:

1. Location
2. Population density
3. Ethnic composition
4. Age
5. Households/Families
6. Income
7. Education/Employment
8. Housing

Your task is to first pick a variable that interests you, e.g. `PCTVACNT`. Call this the 'response'. Then identify useful predictor variables from *four* of the other seven groups. Give a short list of the questions you could answer using these variables. You are welcome to make plots to help in your search, but no plots need to be handed in.

For example, some questions you might try to answer in the project are: is the response related only to income, or does ethnic composition play a role? Is ethnic composition irrelevant once you consider income? Do ethnic composition and income interact in predicting the response? Does the response correlate with geographical features of your state, like military bases and national parks?

There are various places where you can find inspiration. Throughout the class, various outliers and unusual trends were explored in Pennsylvania. You can choose to pursue one of these in more depth. It was shown that Kentucky had an East-West dichotomy on several variables, and that this was partly explainable by geography. Does Pennsylvania have similar dichotomies, and are they also explainable by geography?

Some questions that people tend to be interested in: Are educated people more likely to be married? What demographic groups are most likely to be married? Do Hispanics have larger families? Do larger households make more money? What if many of them work?

For more ideas, check out the interesting article "Marriage, Motherhood and Money," at `http://www.stls.frb.org/publications/re/2003/b/pages/marriage.html`.