

36-315: Statistical Graphics and Visualization

Homework 9 (there was no homework 8)

Date: March 12, 2003

Due: start of class March 17, 2003

1. On the last two labs, you showed that PCTCOLL4 (education) and PCTFEMHE (family type) were stronger than POPPSQMI for predicting income. This leaves open the question of whether education or family type is the strongest overall. The data and functions you need for this are the same as in lab 9.
 - (a) Determine the appropriate transformations for all three variables. Histograms or pairwise scatterplots work well for this. You should use these transformations throughout the problem.
 - (b) Construct a prediction surface for PCI as a function of PCTCOLL4 and PCTFEMHE, using a good choice of span. Show it off with a still three-dimensional plot from a well-chosen viewpoint.
 - (c) Make an unfilled contour plot of the prediction surface, using a good aspect ratio. You may want to experiment with the predictor ordering to get the best picture.
 - (d) Using any of the previous plots, describe the relative strength of PCTCOLL4 and PCTFEMHE. If their strength varies, describe how.
2. In this problem, you will investigate an outlier in the above plots.
 - (a) In your contour plot you may have noticed a few points which have high education level yet low income. To get a better look at this, make a scatterplot of PCI versus PCTCOLL4 (appropriately transformed, of course). Circle the candidate outliers (high education but low income).
 - (b) Now color the points according to PCTFEMHE. This should explain some of the outliers, but there is one in particular which is *not* explained by PCTFEMHE. Circle it. Why isn't it explained?
 - (c) Try again by coloring the points according to PCTUNEMP (percent unemployed). Describe how this helps explain the remaining outlier (though it does not completely explain it).
3. In lab 8 you found that a three-dimension plot can refute the relationships shown in pairwise scatterplots. In this problem, you will demonstrate the same phenomenon in the variable median rent MEDRENT.
 - (a) Start by making pairwise scatterplots to predict MEDRENT from POPPSQMI and PCI, with prediction lines of the appropriate span. (These variables are all included in lab9.csv.)

- (b) According to the scatterplots, does rent increase as you move from country to suburbs? What about suburbs to city?
- (c) Now make a color plot with all three variables, **MEDRENT** as the coloring variable. If income (**PCI**) is held fixed, does rent increase as you move from country to suburbs? What about suburbs to city? Are there any special cases? You can draw contour lines if it helps answer the question.
- (d) Explain how the three-dimension plot refutes the scatterplots. What caused this to happen?