Homework Assignment 6: What's That Got to Do with the Price of Condos in California?

36-350, Data Mining

Due at the start of class, Friday, 30 October 2009

The Census Bureau divides the country up into "tracts" of approximately equal population. For the 1990 Census, California was divided into 20640 tracts. One of the standard data sets (`housing` on `lib.stat.cmu.edu`; accompanying this problem set) records the following for each tract in California: Median house price, median house age, total number of rooms, total number of bedrooms, total number of occupants, total number of houses, median income (in thousands of dollars), latitude and longitude.

1. Make a map of median house prices. The x axis should be longitude and the y axis latitude; you can show price either as a z axis in a 3D plot or by color or grey-scale intensity; make sure the results are legible when you turn them in. Include your code.
2. Use the `lm` command to perform a linear regression of price on all the other variables. Report the coefficients, R^2, and the mean squared error. Include your code. (Do not report any numbers to more than three significant digits.)
3. Which variables seem most important? Why? Do they make sense?
4. Use a $Q - Q$ plot to check if your residuals have a Gaussian distribution. Do they? (EXTRA CREDIT: find and apply an appropriate formal test of normality to the residuals.)
5. Make a map of your residuals. Do they look uniform over space? Should they? If they are not uniform, where does the regression tend to over-predict prices, and where does it under-predict?
6. For each input variable, make a scatter plot of the features vs. residuals. Add a `lowess` smoothing line to each scatter plot. (See the handout to the 23 October lecture for a code example of doing that.) Do the residuals seem to be independent of the features? If not, what patterns are there?
7. Regress the *log* of the housing prices on the other features, and report the results as in part (2). Is there any reason to prefer this regression over that one?