Homework Assignment 1: What's That Got to Do with the Price of Condos in California?

36-402, Advanced Data Analysis

Due at the start of class, Tuesday, 18 January 2011

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.

Do not give raw computer output as your main answer to any question; do include R code in an appendix. Do not report numbers to more significant digits than is warranted. Remember that providing a clear and reasonable justification of your answers is at least as important as getting the answer right.

1. (10 points) Make a plot of median house prices over space (i.e., a map). The x axis should be longitude and the y axis latitude; 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 printed out before turning the assignment in.

2. (20 points) Use the `lm` command to linearly regress price on all the other variables. Report the coefficients, R^2, and the mean squared error. Explain, in words, what the estimated coefficients mean.

3. (20 points) Which variables seem most important? Why? Do they make sense?

4. (10 points) Use a $Q-Q$ plot to check if your residuals have a Gaussian distribution. Do they?

   Extra credit (10 points): find and apply an appropriate formal test of normality to the residuals.

5. (10 points) Make a map of your residuals. Do they look uniform over space? Should they? If they are not uniform, where, geographically, does the regression tend to over-predict prices, and where does it under-predict?

6. (20 points) For each input variable, make a scatter plot of the features vs. residuals. Add a kernel smoothing curve to each scatter plot. (See the

handout to lecture 1 for a code example of doing that.) Do the residuals seem to be independent of the features? If not, what patterns are there?

7. (10 points) Your regression gave standard errors for each coefficient; find the corresponding 95% confidence intervals, under the usual assumptions. Should you trust those confidence intervals here? (Explain.)

8. (10 points) Regress the *log* of the housing prices on the other features, and report the results as in part (2). Which model do you prefer, and why?