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

36-402, Advanced Data Analysis, Spring 2013

Due at 11:59 pm on Monday, 21 January 2013

PROBLEM: As a warm-up and refresher in using linear regression to explore relationships between variables, we will look at a large data set on real estate prices.

The Census Bureau divides the country up into geographic regions, smaller than counties, called "tracts" of a few thousand people each, and reports much of its data at the level of tracts. This data set, drawn from the 2011 American Community Survey, contains information on the housing stock and economic circumstances of every tract in California and Pennsylvania. For each tract, the data file records a large number of variables (not all of which will be used in this assignment):

- A geographic ID code, a code for the state, a code for the county, and a code for the tract
- The population, latitude and longitude of the tract
- Its name
- The median value of the housing units in the tract
- The total number of units and the number of vacant units
- The median number of rooms per unit
- The mean number of people per household which owns its home, the mean number of people per renting household
- The median and mean income of households (in dollars, from all sources)
- The percentage of housing units built in 2005 or later; built in 2000–2004; built in the 1990s; in the 1980s; in the 1970s; in the 1960s; in the 1950s; in the 1940s; and in 1939 or earlier
- The percentage of housing units with 0 bedrooms; with 1 bedroom; with 2; with 3; with 4; with 5 or more bedrooms

- The percentage of households which own their home, and the percentage which rent

Remember that these are not values for individual houses or families, but summaries of all of the houses and families in the tract.

The basic question here has to do with how the quality of the housing stock, the income of the people, and the geography of the tract relate to house values in the tract. We will look at several different linear models, and see if they have reasonable interpretations, and/or make systematic errors.

INSTRUCTIONS: You are allowed to discuss the assignment with others, but *all* work you turn in must be your own. (In particular, copying old solutions, if you find them, is not allowed.) If you do work with others, name them in your assignment.

Your write-up should be a single PDF file, containing your answers to all problems (including all figures, if applicable), written to be read by a human being and not a computer (so: no raw R output, no more precision in numbers than is justified, etc.). Submit a second plain text file containing all your R code.

Begin by loading the data set into a new data-frame in R.

1. (3 pts) Not all variables are available for all tracts. Remove the rows containing NA values. All subsequent problems will be done on this cleaned data set. *Hint:* Recipe 5.27.
   (a) (1) How many tracts are eliminated?
   (b) (1) How many people live in those tracts?
   (c) (1) What happens to the summary statistics for median house value and median income?
2. (7) *House value and income*
   (a) (1) Linearly regress median house value on median household income. Report the intercept and the coefficient (to reasonable precision), and explain what they mean.
   (b) (2) Regress median house value on mean household income. Report the intercept and the coefficient (to reasonable precision), and explain what they mean. Why are the coefficients for two different measure of household income different?
   (c) (4) Regress median house value on both mean and median household income. Report the estimates, and interpret the coefficients, as before. Does this interpretation seem reasonable? Explain.
3. (10) Regress median house value on median income, mean income, population, number of housing units, number of vacant units, percentage of owners, median number of rooms, mean household size of homeowners, and mean household size of renters. Report all the estimated coefficients and their standard errors to reasonable precision, and explain what they

mean. Why are the coefficients on income different from in the previous models?

4. (5) Which three variables are most important, in this model, for predicting house values? Explain your reasoning for deciding on this. *Hint:* make sure your answers wouldn't change if we changed the units of measurement for the predictor variables.

5. (20) *Checking residuals* for the model from problem 3.

   (a) (5) Make a $Q - Q$ plot of the regression residuals.

   (b) (5) Make scatter-plots of the regression residuals against each of the predictor variables, and add kernel smoother curves (as in Chapter 1). Describe any patterns you see. (A *very* rough rule of thumb is that the bandwidth should be about $\sigma n^{-1/5}$, where σ is the standard deviation of the predictor variable and n is the sample size.)

   (c) (5) Make scatter-plots of the squared residuals against each of the predictor variables, and add kernel smoother curves. Describe any patterns you see.

   (d) (5) Explain, using these plots, whether the residuals appear Gaussian and independent of the predictors.

6. (12) Fit the model from 3 to data from California alone, and again to data from Pennsylvania alone.

   (a) (5) Report the two sets of coefficients and standard errors. Explain whether or not it is plausible that the true coefficients are really equally.

   (b) (2) What are the square root of the mean squared error (RMSEs) of the Pennsylvania and California coefficients, on their own data?

   (c) (5) Use the California coefficients to predict the Pennsylvania data. What is the RMSE? What is the correlation between the California coefficients' predictions for Pennsylvania, and the Pennsylvania coefficients' predictions? *Hint:* Recipe 11.18.

7. (10) Make a map of median house values. The vertical coordinate should be latitude, the horizontal coordinate should be longitude, and the house value should be indicated either by the color of the points (*Hint:* recipe 10.23), or by using a third dimension in a perspective plot. Describe the patterns that you see.

8. (10) Make a map of the regression residuals for the model from problem 3. Are they randomly scattered over space, or are there regions where the model systematically over- or under- predicts? Are there regions where the errors are unusually large in both directions? (You might also want to make a map of the absolute value of the residuals.) — If you cannot make a map, you can still get partial credit for scatter-plots of residuals against latitude and longitude.

9. (8) Fit a linear regression with all the variables from problem 3, as well as latitude and longitude. Report the new coefficients and their standard errors. What do the coefficients on latitude and longitude mean? How important are latitude and longitude in this new model?

10. (5) Make a map of the regression residuals for the new model from problem 9. Compare and contrast it with the map of the residuals from the previous model. Are the new residuals spatially uniform, or are there patterns?

11. (10) Regress the *log* of median house value on the same variables as in problem 9. Which model more accurately predicts housing prices? How can you tell?