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

36-402, Spring 2016

Due at 11:59 pm on Wednesday, 20 January 2016

AGENDA: A refresher in using linear regression to explore relationships between variables; also, fun with linear smoothers.

INSTRUCTIONS: See the general instructions on the class website for formats, collaboration, etc., and the rubric at the end of this assignment.

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 `http://www.stat.cmu.edu/~cshalizi/uADA/16/hw/01/CAPA.csv` records lots 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

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.

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

1. (2) Not all variables are available for all tracts. Remove the rows containing NA values. How many tracts are eliminated? How many people live in those tracts? What happens to the summary statistics for median house value and median income? (*Hint:* Recipe 5.27 in *The R Cookbook*.)

   All subsequent problems will be done on this cleaned data set.

2. *House value and income*

   (a) (1) Linearly regress median house value on median household income. Report the intercept and the coefficient, and explain what they mean.

   (b) (2) Linearly regress median house value on mean household income. Report the intercept and the coefficient and explain their meanings. Why are the coefficients for two different measure of household income different?

   (c) (3) 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. (5) 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. *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.

5. 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 the same.

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

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

6. (10) Make a map of the regression residuals for the model from problem 3. The vertical coordinate should be latitude, the horizontal coordinate should be longitude, and the size of the residual should be shown by the color of the points (*Hint:* Recipe 10.23). Are the residuals 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.

7. (5) 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?

8. (5) Make a map of the regression residuals for the new model from problem 7. Compare and contrast it with the map from problem 6. Which one looks better, and why?

9. *Degrees of freedom* Suppose $Y_i = \mu(X_i) + \epsilon_i$, where X_i is p-dimensional, and ϵ_i is a random variable, uncorrelated with the Xs and the other ϵs, with expectation 0 and constant variance σ^2. Our data consists of n pairs $(X_1, Y_1), \ldots (X_n, Y_n)$.

   (a) (5) Consider the intercept-only model which always predicts the sample average of the Y_i. Find the influence matrix $\mathbf{w}$, and show that it has 1 degree of freedom, using the definition in §1.5.3.

   (b) (5) Consider predicting Y using k nearest neighbors. Explain the form of the influence matrix $\mathbf{w}$ (it may be inconvenient to give an exact formula for it), and find a formula for the number of degrees of freedom in terms of k and n. (*Hint:* Your formula should reduce to the answer of the previous problem when $k = n$; why?) Why doesn't p matter?

10. *More Freedom, More Optimism* We're in the same mathematical set-up as in the previous problem. We use some linear smoother (not necessarily linear regression) to get an estimate of the regression function $\widehat{\mu}$. The "optimism" of the estimate is

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y'_i - \widehat{\mu}(x_i))^2\right] - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{\mu}(x_i))^2\right] \quad (1)$$

where Y'_i is an independent copy of Y_i. That is, the optimism is the difference between the in-sample MSE, and how well the model would predict on new data taken at exactly the same x_i values.

    (a) (10) Find a formula for the optimism in terms of n, σ^2, and the number of degrees of freedom. *Hints:* Re-write Eq. 1 as a sum of differences of expectations. What is $\mathrm{Cov}(Y_i, \widehat{\mu}_i)$? What is $\mathrm{Cov}(Y'_i, \widehat{\mu}_i)$?

    (b) (2) Find the optimism of a linear regression for Y, in terms of n, σ^2, and p. What happens as $n \rightarrow \infty$?

    (c) (3) Find the optimism of a k-nearest neighbor regression for Y, in terms of n, σ^2 and k. What happens as $n \rightarrow \infty$? What if k changes with n, as $k = \sqrt{n}$?

RUBRIC (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Questions which ask for a plot or table are answered with both the figure itself and the command (or commands) use to make the plot. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is properly integrated with a tool like R Markdown or knitr, and both the knitted file and the source file are submitted. The code is indented, commented, and uses meaningful names. All code is relevant to the text; there are no dangling or useless commands. All parts of all problems are answered with actual coherent sentences, and never with raw computer code or its output.

EXTRA CREDIT (10): Using the function `knn.reg` from the `FNN` package, as in chapter 1, do a five-nearest-neighbor regression for the house values, using latitude and longitude as the only predictor variables. Find the RMSE and make a map of the residuals. How does this compare to the linear models you estimated? (You will need to calculate distance between locations as a function of latitude and longitude.)