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

36-462/662, Data Mining, Spring 2020

Due at 10:00 pm on Thursday, 23 January 2020

AGENDA: Remembering how to use linear regression to explore relationships between variables; first taste of using nearest neighbors.

The Census Bureau divides the country up into geographic regions called "tracts", each of which contains a few thousand people, and reports much of its data at the level of tracts. This week's 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.

Note 1: As you recall from earlier courses, "RMSE" standards for "root mean squared error", in symbols $\sqrt{n^{-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$.

Note 2: Whenever a problem asks you to explain something, use your own words, rather than quoting, copying or paraphrasing.

1. (10) *Online questions about the reading* are online.

2. 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.

   (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?

3. *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. Also report the RMSE.

   (b) (2) Regress median house value on mean household income. Report the intercept and the coefficient, and explain what they mean. Also report the RMSE. 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. Also give the RMSE.

   (d) (1) Which model has the best RMSE? Is that a reason to prefer one model over another?

4. (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. Why are the coefficients on income different from in the previous models? What is the RMSE?

5. One way to measure the importance of a predictor variable is to see how much the prediction changes in response to a one-standard-deviation change in the predictor.

(a) (3) Explain how to calculate this change, for a linear model, from the predictor's standard deviation and from the estimated coefficients. (You do not need to re-fit anything, or actually calculate any predictions.)

(b) (3) Explain why this way of measuring variable importance is unaffected by the units we use for the variable (e.g., recording income in dollars or thousands of dollars), unlike the size of the coefficients.

(c) (2) Explain why it is not appropriate to measure how important a variable is by the magnitude of the t statistic for its coefficient.

(d) (3) What are the three most important variables, in this model?

6. Fit the model from 4 to data from California alone, and again to data from Pennsylvania alone.

   (a) (2) Report the two sets of coefficients and standard errors.

   (b) (3) Is it plausible that the coefficients are really equal across the two states? Explain.

   (c) (1) What are the RMSEs of the Pennsylvania and California coefficients, on their own data?

   (d) (2) What is the RMSE of using each state's coefficients to predict the *other* state's data? Do the coefficients generalize well across states?

7. (10) Fit a linear regression with all the variables from problem 4, 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? What has happened to the RMSE?

8. (10) Use k-nearest neighbors regression, with $k = 3$, to predict housing prices, using:

   (a) Median income alone;

   (b) Mean income alone;

   (c) Latitude and longitude alone ;

   (d) All the variables from problem 7 ;

   (e) All the variables from problem 7, having first centered standardized them (so they have mean 0 and variance 1) .

   For each alternative, plot the predictions against the actual values, and report the root-mean-squared error.

   (I suggest using the `FNN` package, but there are lots of options.)

9. (2) Did you get different predictions in problem 8d than in problem 8e? Should you? Explain.

10. *First step at generalization*

    (a) (3) Randomly divide the data set into two equally-sized halves (with no overlap). Report the summary statistics for the following variables for each half: latitude, longitude, population, mean income, median income, median value. Explain why it is important that the summary statistics should be close for the two halves, but also why we should not expect that they will be equal.

    (b) (3) Estimate the linear model from problem 7 using *only* the data from one half. What is the RMSE on this "training" data?

    (c) (3) Using the same "training" data as in problem 10b, estimate the 3-nearest neighbor models from problems 8c, 8d and 8e, and report their RMSEs on the training data.

    (d) (5) Now take each of the models you have estimated in this problem. and find their RMSEs on the other half of the data, the "testing" data. Be careful not to re-estimate any of the models. Report your results in a figure or table.

    (e) (5) Based on your findings in this problem, which model would you recommend using?

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; there are no dangling or useless commands. All parts of all problems are answered with actual coherent sentences, and raw computer code or output are only shown when explicitly asked for.

EXTRA CREDIT (3 points): In problem 5, you ranked predictor variables by seeing how much the prediction changed in response to a one standard deviation change in the variable. Explain how you could use this idea with nearest neighbor regression (as opposed to the linear model considered in problem 5). Does it matter whether the variable increases or decreases?

EXTRA CREDIT (continued, 2 points): Implement your idea in code. What are the 3 most important variables for the models from problems 8d and 8e?