SML 201 Project 4 Detailed Instructions and Hints

2019-12-13

Some shorthand notations for the formulas argument in lm() and regsubsets():

Include all the columns in the data as the x-variables in the model, except the column for y, in the model: y ~ .

Include all the columns in the data as the x-variables in the model, except the columns for y and x1, in the model: y ~ . -x1

Include all the columns in the data as the x-variables in the model, except the column for y, in the model, and exclude the y-intercept for the model: y ~ . -1

Include the y-intercept only (i.e., no predictors in the model) for the model: y ~ 1

Question 2

Part b

Recall that in the framework for linear models, if we divide the x-axis into multiple intervals and calculate the mean of the y-values for each vertical strip that correspond to the x-subinterval, then y-means should form a line. However, our data do not show this pattern, so we might want to transform our variable(s) to make the relationships more linear.

Note that from the scatterplot for mean `price` v.s. unique numbers of `bedrooms` in question 2.d we see that it will be good to have different slopes for the lines depending on whether the value of `bedrooms` is greater than 8 or not. Thus, we should consider the interaction $(bedrooms <= 8) : bedrooms$ for our model. Please keep this in mind when building the model later.

Question 3 Zipcode variable

Part a

Make sure that your variable `zipcode` is of the correct data type for making the boxplots and make sure that all the zip codes are legible on your graph (you might want to refresh your memory about what the input argument `las` in `par()` controls).

Part b

Hint: the y-intercept is the average effect (of being among a certain subset of houses) on `price`.

Question 4

Part a

A more generalized form of the relationship between Y and X would have been

$$Y \approx a + (X)^b$$

However, from the scatterplot in question 2.c we see that it is reasonable to assume that a = 0. Setting a = 0 will make the procedure to estimate b easier.

Part c

We decided not to keep the variable `sqft_lot` since it does not have much linear relationship with `log.price`. `sqft_above` is also dropped since it is highly correlated with `sqft_living`.

Question 5

Part a

Since you have a large number of predictors to consider, it will not be practical to consider all possible subsets of the predictors. However, if you use the backward selection algorithm you must compare the result with the one produced by the forward selection algorithm as the two do not always give you the same result and there is no guarantee that the two algorithm outputs the true best model–they will only try. (You can skip the sequential replacement algorithm for this project–I already checked and the result does not give you extra information.)

For each of the criteria (BIC and Adjusted R^2) you should plot the results from backward and forward selection algorithms on the same graph so that you and your readers can compare the results easily.

If you are not sure about how many terms you have in the mathematical form of the full model the function `lm()` will be helpful (see exercise in Precept 12).