SML 201 Project 3 Detailed Instructions and Hints

2018-12-15

Question 2

Part a

You should make several plots instead of just one so that your scatterplots will not be too small for you to study the relationships. Make sure that the correlation figures are small enough that they fit within the margins of the plots but big enough to be legible.

Note that you should not include `date` in your scatterplots since `date` has 372 levels and since the closing date of a transaction should not affect the sold price of a house; (since we have limited time for this project I investigated the relationship between the month of sale and `price` and did not see any patterns); therefore, we will not include the date-related information in our model.

Also, note that including `zipcode` in the scatterplot matrix will not give you much info as the graph will be too small to show the relationship between `price` and `zipcode`; thus, we will investigate the relationship between `price` and `zipcode` separately later in the report.

Part d

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 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

You should consider all the "best" models (i.e., "best" for each given number of predictors). 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 exercises at the end of Ch 6.3).

Also note that if you want to exclude a term in your model you can use the minus sign; e.g., `lm(y~.-1, ...)` will exclude the y-intercept from your model.

Part c

Make sure that the smoothed histograms of the distributions are transparent enough for you to see the smoothed histogram on the bottom layer. Also, since we have a lot of data points make sure that the dots on the scatterplot are small and transparent enough that you can actually see the pattern on the graph.

Question 7

You should write out the mathematical form of the model first.