Stat 20 Evaluating and Improving Predictions

Today's data is on house pricing in Los Angeles! We have access to data on all of the houses that sold in four cities around west LA in the course of a single month. Our goal is to predict the price of a home based on a few attributes of that home. Check Ed for a link to the data.

Where applicable, answer the following questions using R code and write the code you used in the space below.

1. Create a scatterplot examining the relationship between square footage and price. Write the code you used below.

2. Add two columns onto the `LA` dataset which measure *log* price and *log* square footage. Save the new dataset back into itself.

3. Create another scatterplot examining the relationship between *log* square footage and *log* price. Rather than writing the code you used, explain in at least two sentences whether the logged variables or the regular variables are more suitable for a linear model.

4. Fit two linear models

- one which predicts price with square footage
- one which predicts log price with log square footage

Report the R^2s for both models, and in a sentence, state whether the results you have line up with your explanation in the last question.

Recently the University of California purchased a new house to serve as the residence of the university President and to host university functions. The address of the house is 2821 Claremont Blvd in Berkeley.

5. Use your linear model to predict the sale price in log USD of this house. (hint: the internet is helpful!). Then, find this price in regular USD.

6. Was your model an under- or over-estimate? Why do you think this is?