Homework 2

Solution

9 February 2017

0.1 Introduction

Appraising residential real estate — predicting the price at which it could be sold, under current market conditions — is important not only for people buying and selling houses to live in, but also for real estate developers, mortgage lenders, and local tax assessors. Currently, appraisal is usually done by skilled professionals who make a good living at it, so naturally there is interest in replacing them by machines. In this report, we investigate the feasibility of real estate appraisal by means of linear statistical models.

Specific points of interest to the client include the relationship between the quality of the house's construction and its price; the relationship between age and price, and whether this changes depending on proximity to a highway; and the relationship between price, the finished area of the house, and the number of bedrooms.

0.2 Exploratory data analysis

The data, supplied by an undisclosed client, come from a selection of "arms-length" residential real estate transactions in an unnamed city in the American midwest in 2002. This records, for 522 transactions, the sale price of the house, its finished area and the area of the lot, the number of bedrooms, the number of bathrooms, the number of cars that will fit in its garage, the year it was built, whether it has air conditioning, whether it has a pool, whether it is adjacent to a highway, and the quality of construction, graded from low to medium or high. It is notable that, except for highway adjacency, we have no information about the location of the houses, though this is proverbially a very important influence on their price, through access to schools, commuting time, land value, etc.

Pairwise scatter-plots for the quantitative variables (Figure 1) show that, unsurprisingly, there is a positive relationship between price and area (stronger for finished area than the total lot size), and price and the number of bedrooms, bathrooms, or garage slots (all three of which are strongly positively related to each other). The relation between price and these three "count" variables could well be linear. There is a positive relation between price and the year of construction, i.e., newer houses cost more. Newer houses also tend to be larger, both in finished area and the number of rooms, though not to have bigger lots.

Inspection of the plots shows there is one record with 0 bedrooms, 0 bathrooms, and a three-car garage with air conditioning. This is either not a piece of residential real estate, or its data is hopelessly corrupt; either way, we drop it from the data from now on.

Box-plots, showing the conditional distribution of price for each level of the categorical predictors, suggest that houses with air-conditioning and pools are more expensive, that being next to a highway makes little difference, and that higher quality of construction implies, on average, higher prices. The mid-points of the boxes for quality don't *quite* fall on a straight line, so treating quality as a numerical variable isn't obviously compelling, but not clearly crazy either.

0.3 Initial Modeling

To answer the client's questions, our model should include quality, finished area, the number of bedrooms (and the interaction between those two), and the year the house was built and whether it is adjacent to a highway (and the interaction between those two). Based on our EDA, it also seems reasonable to include air-conditioning and pools. We deliberately left out the number of bathrooms, the size of the garage, and the

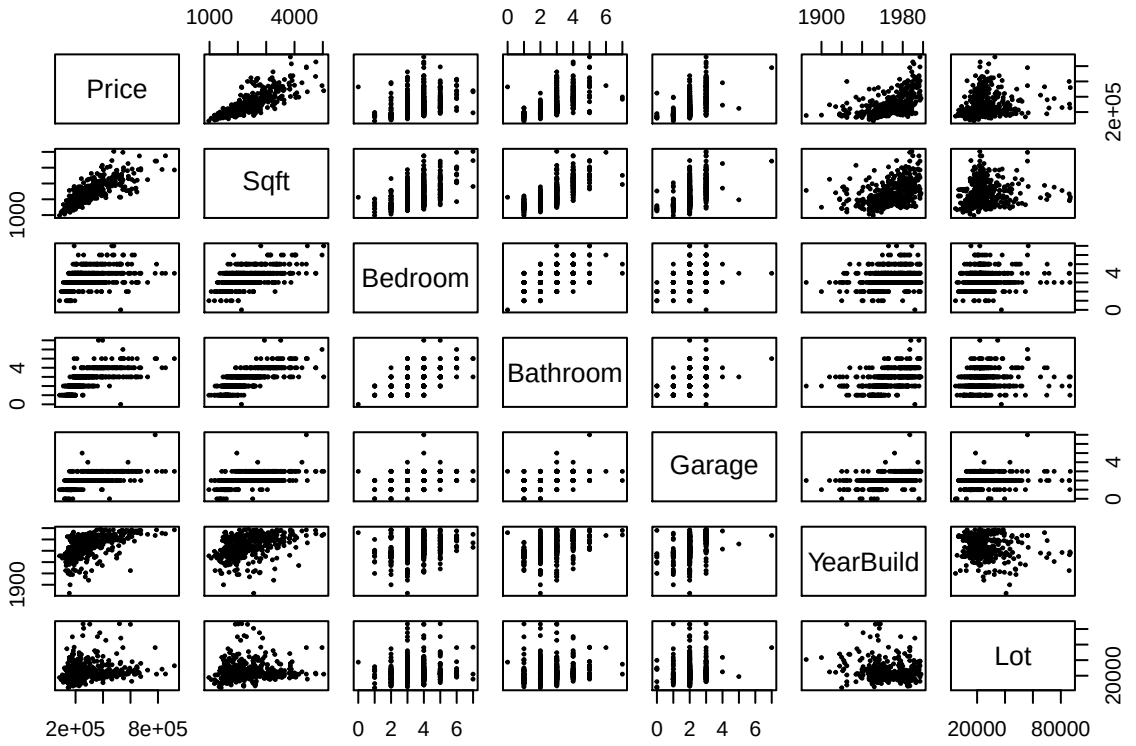

Figure 1: Pairs plot for quantitative variables

size of the lot. While price seems to be linearly related to the number of bedrooms, we include it as a factor, both to check that, and to get three distinct slopes for price on finished area as quality varies.

This initial model has a root-mean-squared error of $\$ \pm 5.94 \times 10^4$, which is not shabby when the median house price is $\$ 2.3 \times 10^5$. Before passing to issues of model selection, however, such as whether all the interactions are necessary, whether discrete variables might be usefully recoded, etc., let's look at the diagnostic plots.

The first thing to say is that the distribution of the residuals doesn't look very Gaussian, and a Box-Cox transformation suggests the un-intuitive, indeed un-interpretable, transformation $1/\sqrt[3]{Y}$.

Clients who ask for a model of prices are rarely happy with models for the inverse cubic roots of prices, so we must be doing something wrong. Examining plots of residuals versus predictors suggests that lot size matters after all, at least for big lots. The plots also suggest that houses built after ≈ 1980 are worth more than the model anticipates. The distributions of residuals conditional on discrete predictors, however, actually look mostly homogeneous.

0.4 Outliers

In addition to the house with no bedrooms or bathrooms, examination of Cook's distance shows two houses with exceptional influence over the model.

On examination, these are quite weird: small in area, fairly cheap, but heavy on bedrooms. These look more like rental properties than residences. Checking the pairs plot again shows no other such anomalies, so we delete them but leave the rest alone. Re-doing the other diagnostic plots shows little over-all change, however (figures omitted).

	Price	Sqft	Bedroom	Bathroom
11	190000	2812	7	5

	Price	Sqft	Bedroom	Bathroom
383	219900	1852	6	3

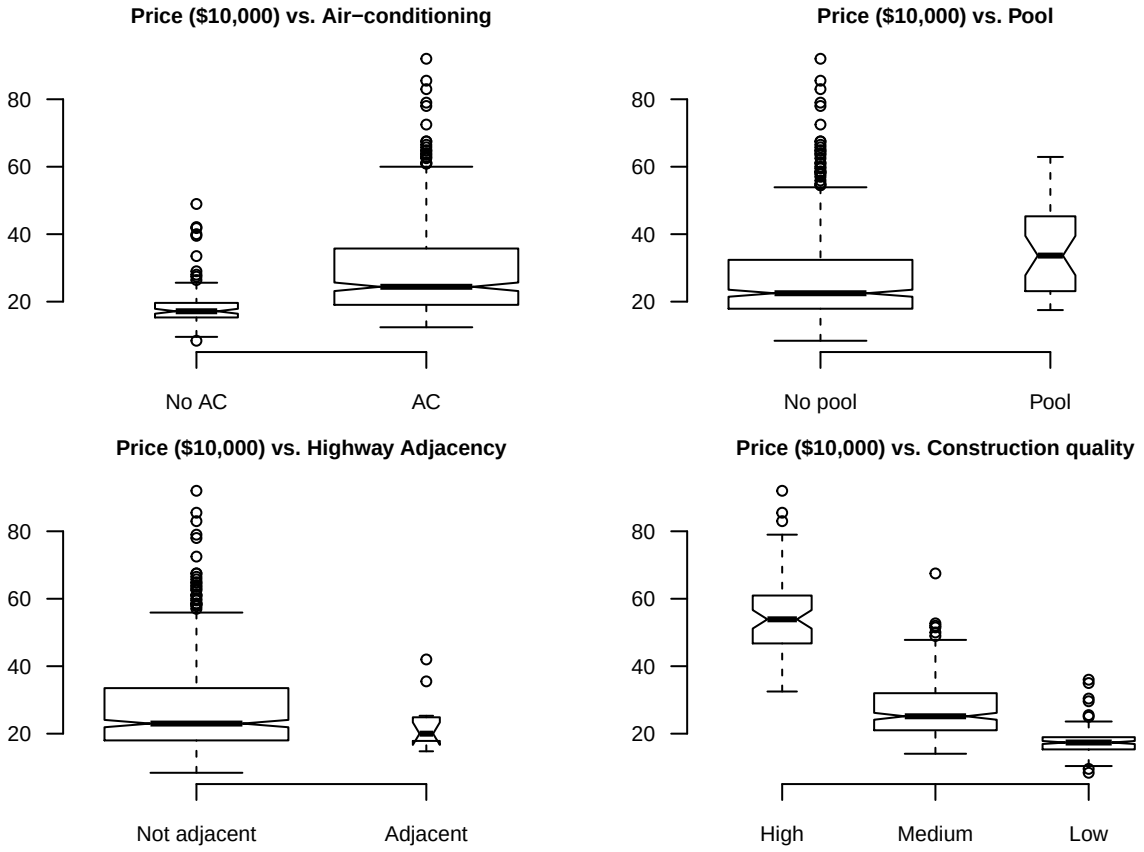

Figure 2: Conditional distributions of price given qualitative predictors. Box widths reflect the number of points in each group, notches show medians plus/minus a margin of error.

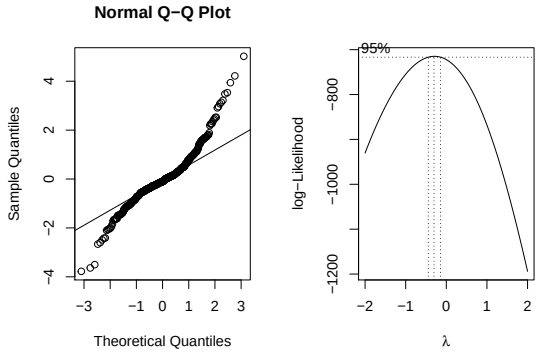

Figure 3: Q-Q plot of the standardized residuals (left) and Box-Cox plot (right)

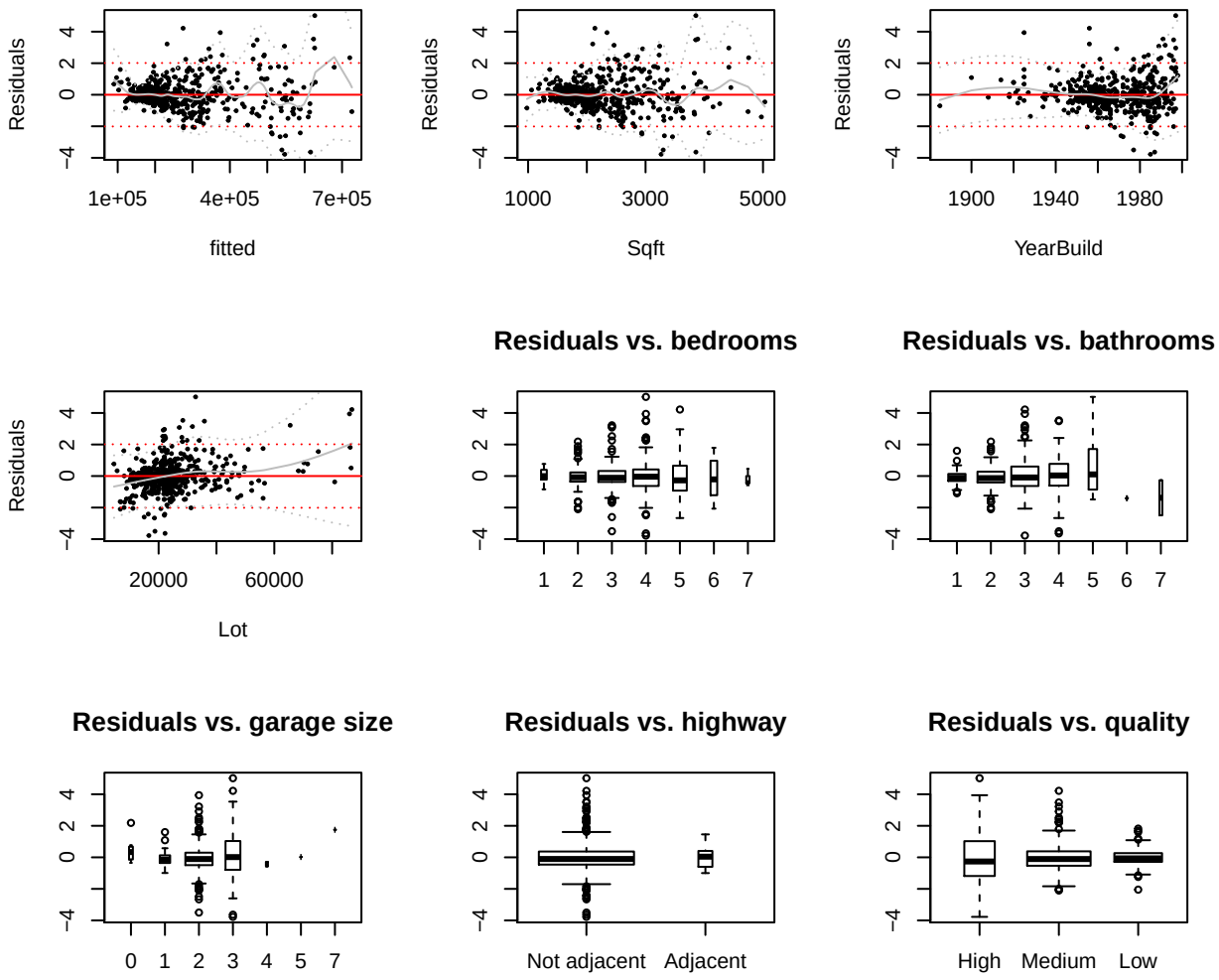

Figure 4: Residuals versus fitted values and continuous predictors, and versus the discrete predictors. Grey lines are smoothing splines; dotted lines indicate plus/minus 2 standard deviations, either constant (red) or from a spline smoothing of the squared residuals (grey).

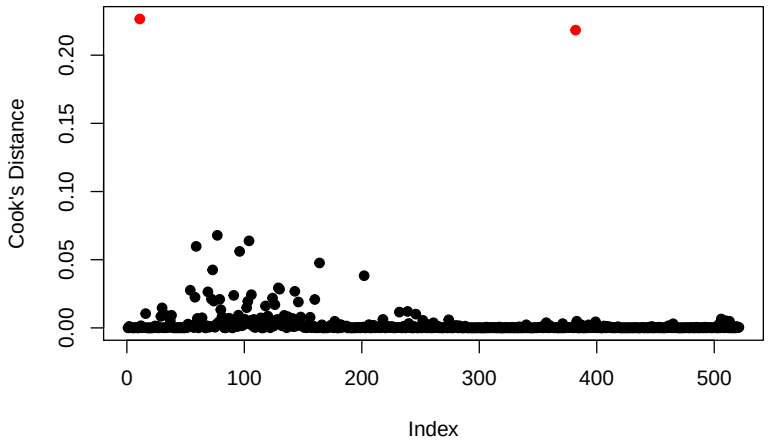

Figure 5: Cook's distance for each data point: extremely influential points are flagged in red.