Mini-project 2

S681

Upload your project through the Assignments tab on Canvas by 11:59 pm, Sunday 24th March.

You may (but are not obligated to) submit this project in pairs. If you're working in a pair, email me to let me know.

Part One: Housing prices

The file `housetrain.txt` contains measurements on 372 recent house sales in one city. The variables are:

- `ID`: A random ID I assigned to the sales.
- `Price` in dollars
- `Sqft`: finished square feet
- `Bedroom`: number of bedrooms
- `Bathroom`: number of bathrooms
- `Airconditioning`: 1 if the house has air conditioning, 0 otherwise
- `Garage`: number of cars the garage holds (if any)
- `Pool`: 1 if the house has a pool, 0 otherwise
- `YearBuild`: the year the house was built
- `Quality`: 1 = high, 2 = medium, 3 = low
- `Lot`: lot size in square feet
- `AdjHighway`: 1 if adjacent to a highway, 0 otherwise

Questions

1. **Prediction (10 points.)** Build a regression model to predict *log* house price from the other covariates. You must be able to write down an equation for the model (i.e. it should be parametric and not `loess` or `gam`.) Put code in a .R (or .Rmd) file that fits the model and creates a vector called `predictions` that contains predictions for log house prices for a test set of data in a file (in the working directory) called `housetest.txt`.

   You don't have `housetest.txt`; we do. We'll apply your code and compare it to a baseline model that just throws all of the other variables into `lm()`.

   Grading: 5 points for beating the baseline model in terms of root mean squared error; 5 points for your code running without us having to edit it.

2. **Displaying an interaction (10 points.)** Find TWO variables that have a significant interaction when used to predict log house price. Graphically show how the predicted log house price varies with these two variables, taking care to only display predictions for ranges of values covered by your data set. Describe in full what your graph(s) tell you about the interaction.

3. **Do bedrooms matter? (10 points.)** Number of bedrooms is positively correlated with log house price, but is the apparent effect of bedrooms better explained by other covariates? One way of addressing this is to fit two models: the best model you can find that includes bedrooms, and the same model with all terms with bedrooms excluded. Fit and compare these models. What does this analysis tell you about the relationship between number of bedrooms and house price?

Part Two: Build your own bootstrap (20 points)

We've seen examples where the residual bootstrap for confidence intervals does meaningfully better than the percentile bootstrap, in terms of coverage. What's an example where 95% residual bootstrap confidence intervals perform meaningfully better than 95% percentile bootstrap confidence intervals. (By "meaningfully better," let's say that the residual intervals comes closer to the nominal level of coverage by at least half a percent.) To justify your claim, run a simulation and explain its results. Upload a .R (or .Rmd) file with your code so we can reproduce your results (set the random seed to ensure we get the same results as you.)

What to submit

- A PDF or other file containing your report.
- A .R (or .Rmd) file containing code to produce your predictions for Part 1, question 1.
- A .R (or .Rmd) file containing code to reproduce your simulation for Part 2.
- A .Rmd or other file containing the rest of your code.
- Any other supplementary files required to reproduce your work.