CSCI 452 (Data Mining)
Dr. Schwartz
Linear Regression
100 pts

This assignment asks you to perform analysis of linear regression models by:

- using R to fit models
- interpreting coefficients
- reporting confidence and prediction intervals
- observing the presence or absence of statistical significance
- generating a residual plot visualization

This is an individual assignment. You will create an R Markdown file called regression.Rmd and submit both the .Rmd and resulting .html (after knitting) files to autolab. Please be sure to follow the order of the questions and identify your answers/analysis clearly so that I can easily find the individual pieces.

1. Realtor Data

a) Background

Read this paper: *Pardoe, "Modeling Home Prices Using Realtor Data", Journal of Statistics Education, 16:2, 2008.*

Note: This is an education paper; though its intention is different from most research papers, the case study presented will be very accessible to us and what we have covered in the course so far. The vast majority of linear regression papers are currently too advanced for us. Topics in this paper that are probably beyond the scope of our course due to time constraints include: the nested model F-test, and variable transformation. At the very least, this paper motivates these topics.

Look at the dataset documentation and accompanying dataset. You should download the dataset to the same directory as your working directory for this project (I don't want any path info in your load commands, just the homes76.dat.txt file).

In your submission, answer the following questions about the paper and data:

1) How are half-bathrooms represented? Why? Do you agree?
2) The model prediction intervals are very wide. What types of data are missing from the modeling?
3) Briefly explain how age is represented.
4) Lot size is discretized into categories instead of being a continuous variable. Why?

b) Analysis in R.
Use the above home values data set for the following questions. Document your work. Include R commands and output in your submission file.

Use simple regression to estimate price based on floor size alone. Answer the following questions:

1) What is the estimated regression equation?
2) What is the value of the *slope coefficient*? Interpret it.
3) What is the value of the *y-intercept*? Interpret it. Does it make sense?
4) What would be a typical *prediction error* obtained from using this model to predict price?
5) How closely does the model fit the data?
6) Find a *point estimate* for the price of a house with a floor size of 1950 square feet.
7) Find a *95% confidence interval* for the true mean price for all homes with a floor size content of 1950 square feet.
8) Find a *95% prediction interval* for a randomly chosen house with a floor size of 1950 square feet.
9) Create a *scatter plot visualization* that plots price versus floor size.

Now use multiple regression to estimate price based on floor size and number of bedrooms. Answer the following questions:

10) What is the estimated regression equation?
11) Explain the value of the coefficient for floor size. Is floor size statistically significant? Why or why not?
12) Compare the R2 values from the multiple regression with the previous simple regression. What does this mean?
13) Create and interpret a residual plot. Does the relationship between predictors and the target appear to be linear? Discuss.

Submission instructions

Present your answers to the questions above, analysis, commentary, R code and visualizations in an R Markdown file. Name your file regression.Rmd and, when you knit, create an .html file. Zip these two files together (don't zip a directory) and upload your submission to Autolab as the Linear Regression lab.