M358K - Final project

posted on: Monday October 30th, 2017

due by: December 12th, 2017

Instructions and grading scheme

Submission instructions. Please submit three files on Canvas with the following names:

- **project-writeup**: this is a word file containing your answers. All figures, tables etc must be inside this file! For how to save R figures, see instructions on Canvas.
- **project-final.R**: this is a text file containing the R codes you used to compute the numbers, tables, and figures CONTAINED IN your report ONLY. This file IS graded, see below.
- **project-draft.R**: this is a text file containing the R codes you used to explore the data. This file should give an idea on how you explored the dataset, and how you arrived at code-final.R. It can tables/figures which are EXCLUDED from your report, comments, etc. For instance, the R history of your various sessions is enough.

Acceptable word files are: anything that can be opened by LibreOffice, OpenOffice, Microsoft Word, or equivalent software. Example file extensions are .doc, .docx, .odt.

Grading scheme. For the write up: on each question you can earn $0/1/2 \times$ the maximal number of points for that question.

- 2 = correct answer

- 1 = partially correct answer, or correct answer but with muddy or missing justification
- 0 = incorrect answer, unreasonable answer

For code-final.R: you can earn 0/10 points.

- 10 = code-final.R runs on the given dataset, gives all the tables and figures included in your report.
- 0 = no code file OR code file does not run at all OR code file does not produce the reports' tables/figures OR code file raises a plagiarism red flag, etc

For code-draft.R: you can earn 0/10 points.

- 10 = The grader is sufficiently convinced that you explored this dataset on your own.
- 0 = no code file OR code file raises a plagiarism red flag OR code file is irrelvant, does not give an indication on how you arrived at code-final.R

Bonus points for presentation:
Write-up: +5 points for nice report(grammatically correct sentences, no rambling discussions, discussions exceed expectations, extra analysis of the dataset)
R code: +5 points for neat layout. (code adequately commented, clearly laid out, easy to understand).

Bonus questions
These are extra challenging questions, and are additional opportunities to score points. On each bonus question you can get 0/5 points.

Project overview

In this project, your goal is to predict house prices based on the house attributes (such as ZIP code, number of bedrooms, square feet etc).

The data file, house.csv, is under files/finalProject on Canvas. For variable descriptions, see house-descrip.txt.

A plagiarism WARNING: this is a real public dataset. Just as it is easy for you to find public regression codes written by other people, it is easy for us to verify if you plagiarised. Everyone has a specific coding style, and we have seen plenty examples from your code in the previous homework. Therefore, I repeat, it is EASY for us to identify cheaters. Do NOT cheat.

Another WARNING: your final project is NOT judged solely on the R^2 or goodness of fit of your model. Therefore, do NOT use statistical techniques outside those introduced in the lectures. If you do choose to use it, you will need to write a one-page explanation for each such technique (what it is, when people use it, and why did you use it), and attach it to your submission.

Question 1: descriptive statistics (25 points)

Of the variables in the dataset, which variables are interesting for predict house price?
The questions you should address are:

1. (presentation point) Give a clear list of the interesting variables.
2. (10pt) For each of the interesting variable, explain by plot(s) and number(s) its relationship with price. Write a brief summary.
3. (10pt) For each of the uninteresting variable, explain by plot(s) and number(s) why it is uninteresting. Write a brief summary.
4. (5pt) What is an outlier? Are there outliers in your data? Should you exclude such outliers from the regression model below, and why?
5. (Bonus: 5pt): Are there interesting relationships amongst the input variables? (input variables are all variables except price).

Question 2: inferential statistics (30 points)

Run a linear regression model of price vs other input variables.
The questions you should address are:

1. (20pt) Detail how you did variable selection: which models did you run, why did you discard certain models or variables, any variable transformations you did and why, which diagnostic tests did you run and what they showed, justifications if you removed outliers.

2. (10pt) Call your final regression model `model.lm`. Clearly show your final regression model: the R command, and the R output summary. Write down the equation that R gives you. Interpret all the coefficients and the p-values associated with the coefficients. Report the R^2 and adjusted R^2 of your model. What are the meaning of these values? Run a diagnostic plot for your model, and explain if your model is a good fit.

Question 3: sampling, error and other topics (20 points)

When searched for products online (eg: Google shopping, Amazon), some users claim that the prices can change depending on their browsing data (such as their location, inferred through their IP addresses, how long they stay at a particular page product, how many times they rephrase their search queries, etc). You want to decide which, if any of such factors, affect the items' prices.

1. (10 points) Detail how you would design an experiment to answer the above question. Clearly list: the research question, the population, how you sample, and the variables you collect.

2. (5 points) List at least TWO significant biases your experiment may have, and how your design mitigates them.

A pollster surveyed 1500 people, and reported that 45% of them buy more than half of their household products online, while the rest buy the majority of their household products in shops.

1. (3 points) Compute a 95% confidence interval for p_0, the true proportion of Americans who buy the majority of their household products online. Do a hypothesis test for whether p_0 is significantly different from 0.5. What is the meaning of the p-value of this test?

2. (2 points) Another pollster reported that their poll gives 47%, with a 95%-confidence interval of 42% to 52%. Why is it WRONG to say that "there is a 95% chance that between 42% and 49% of Americans buy the majority of their household products online"? Give the correct interpretation of this confidence interval.