STA302H1F / 1001HF Autumn 2020 Assignment # 2

A Simple Linear Model for Toronto and Mississauga House Prices

Posted by: Dr. Shivon Sue-Chee on Saturday, October 10, 2020

Due: In **Quercus** by 8pm on Saturday, October 24, 2020.
Late assignments will be subjected to a penalty of 20% per day late. Submissions will not be accepted beyond 48 hours of the due date. Email submissions are not allowed.

1 Instructions

- Use R Studio to create two files:
  1. An Rmarkdown file with your codes according to the standard format in the A2_a20_RMForm.Rmd file.
  2. The corresponding report in html or pdf format.
- Create a video presentation of no longer than 5 minutes with you presenting your written report. Your face should be shown at least once during your presentation. Save your video presentation as an MP4 file and upload it into your UofT MyMedia account.

  I suggest that you use Zoom to create your video. Here are two demonstration videos of how to record and download your presentation in Zoom:
  - `https://www.youtube.com/watch?v=P6cTbnUPwfY`
  - `https://kb.siue.edu/61721`

  Here is documentation on using UofT MyMedia:
  - `https://www.oise.utoronto.ca/online/Instructors/Video_server_-_MyMedia/index.html`
- Into Quercus Assignment 2, submit the following three items:
  1. A MyMedia link to your video presentation
  2. An Rmd file with your RMarkdown codes
  3. A Portrait picture of yourself with your T-card
- Note that for a separate participation activity, which would be announced later, you would be asked to upload your video presentation to peer Scholar for peer reviews. For the sake of privacy, please avoid revealing your full identity in your video presentation. You could use your initials and up to the last four digits of your student number, if you like.
- Presentation of your report is very important. Do not show R codes unless it is required for your solutions. Only required numbers and plots should be shown. Extraneous output should be hidden. Use options, include=FALSE, echo=FALSE, message=FALSE, where necessary.
- Write and present **your own work**. For instance, personalized your code as much as possible, using your initials. **All plots produced must be given a title with the last 4 digits of your student number**.
- Use a benchmark significance level of 5%. Report p-values to 4 decimal places.

2 Grading Scheme

Grading rubrics will be posted in Quercus for the video presentation and the RMarkdown file.

Note that if a portrait picture of yourself with a clear view of your T-card is not received by the due date or if your picture and T-card do not correspond to our other records, a mark of zero will be given for the entire assignment.

3 The Data

First-time home buying is currently a major federal issue. Prices for detached houses have been at an all-time high during the current COVID-19 period. Data for this assignment was obtained from the Toronto Real Estate Board (TREB) on detached hourses in two separate neighbourhoods- one in the city of Toronto and another in the city of Mississauga. Data is contained in the file "real20.csv" on the assignment 2 page. The variables in the dataset are:

- `ID`: property identification
- `sold`: the actual sale price of the property in millions of Canadian dollars
- `list`: the last list price of the property in millions of Canadian dollars
- `taxes`: previous year's property tax in Canadian dollars
- `location`: M- Mississauga Neighbourhood, T- Toronto Neighbourhood

For this assignment, we are interested in establishing a simple linear model that home buyers can use to determine the expected sale price of detached, single family homes in the two neighbourhoods in the Greater Toronto Area.

4 The Analysis

Set the seed of your randomization to be the last 4 digits of your student number. Randomly select a sample of 200 cases. Based on your sample data, complete an RMarkdown file and the corresponding report with the following sections.

I. Exploratory Data Analysis section.

- Use a single plot to describe your data. Create a subset of your data by removing at most two cases and briefly explain your choice.
- *Then using the data subset for this and the remaining parts of this assignment*, draw two scatterplots of the response variable- sale price by (i) list price and then (ii) taxes. In each plot, distinguish between properties in neighbourhood M and those in neighbourhood T, and include a legend/key.
- Interpret each of the three plots produced in this part, that is, describe at least one major highlight from each plot. Each highlight should differ from the other.

II. Methods and Model section.

- Carry out three simple linear regressions (SLR) for sale price from list price, one for all data, one for properties of neighbourhood M and another for properties of neighbourhood T. In a table, give the values of the following for each of these regressions:

  - R^2
  - the estimated intercept
  - the estimated slope
  - the estimate of the variance of the error term
  - the p-value for the test with null hypothesis that the slope is 0
  - a 95% confidence interval for the slope parameter
- Interpret and compare the three R^2 values. Give a brief explanation why they appear similar or different.
- Briefly discuss whether a pooled two-sample t-test can be used to determine if there is a statistically significant difference between the slopes of the simple linear models for the two neighbourhoods. (Note: You do not need to carry out a pooled two-sample t- test here.)

III. Discussion and Limitations section.
A sensible data science approach is to base inferences or conclusions only on valid models.

- Select one of the three fitted models in part II and give a brief explanation for your choice.
- Discuss whether there are any violations of the normal error SLR assumptions for your selected model. Use at most two plots.
- Identify two potential numeric predictors (other than those given in the data set) that could be used to fit a multiple linear regression for sale price.