STA302H1F / 1001HF Autumn 2020 Assignment # 3

A Multiple Linear Model for Toronto and Mississauga House Prices

Posted by: Shivon Sue-Chee on Saturday, November 21, 2020

Due: In **Quercus** by 8pm on Saturday, December 5, 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 and written (full) answers for the data analysis parts.
  2. The corresponding data analysis report, in html or pdf or docx format.

  See posted templates.

- Create a video presentation of no longer than 3 minutes with you presenting your knitted R report. Your face should be shown throughout your presentation. Your oral presentation should be a summary of the main results from your data analysis. 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 3, upload the following three (3) files:

  1. A docx page with a MyMedia link to your video presentation and a portrait picture of yourself holding your T-card. *An example is posted.*
  2. An Rmd file with your RMarkdown codes and written assignment answers
  3. The corresponding knitted report, in html or pdf or docx format.

- Note that this assignment will not be subjected to peer reviews.

- 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

A grading rubric will be posted in Quercus for this assignment.

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 houses in two separate neighbourhoods- one in the city of Toronto and another in the city of Mississauga. It is available in the file "real203.csv" on the assignment 3 page. The property-based variables in the dataset are:

- `ID`: property identification
- `sale`: the actual sale price of the property in Canadian dollars
- `list`: the last list price of the property in Canadian dollars
- `bedroom`: the total number of bedrooms
- `bathroom`: the number of bathrooms
- `parking`: the total number of parking spots
- `maxsqfoot`: the maximum square footage of the property
- `taxes`: previous year's property tax
- `lotwidth`: the frontage in feet
- `lotlength`: the length in feet of one side of the property
- `location`: M - Mississauga Neighbourhood, T - Toronto Neighbourhood,

Missing values are labelled as 'NA'.

4 The Analysis

For this assignment, we extend our work in Assignment 2 to find a more complex linear model which home buyers can use to predict the sale price of single-family, detached homes in the two neighbourhoods in the Greater Toronto Area.

I. Data Manipulation section.

(a) Set the seed of your randomization to be your student number. Randomly select a sample of 150 cases. Report the IDs of the sample selected.
Use your sample data, for the following sections.

(b) Create a new variable with the name '`lotsize`' by multiplying `lotwidth` by `lotlength`. Use this new variable to replace `lotwidth` and `lotlength`.

(c) Clean the data by removing at most **eleven** cases and one predictor. Briefly explain your choices. Then use this updated data for the successive parts of this assignment.

II. Exploratory Data Analysis section.

(a) Classify each variable included in this assignment as categorical or discrete or continuous.

(b) Produce the pairwise correlations and scatterplot matrix for all pairs of quantitative variables in the data. Describe how each quantitative predictor for sale price rank, in terms their correlation coefficient, from highest to lowest.

(c) Based on the scatterplot matrix, for which single predictor of sale price would the assumption of constant variance be strongly violated? Confirm your answer by showing an appropriate plot of the (standardized) residuals.

III. Methods and Model section.

i. Fit an additive linear regression model with all available predictors variables for sale price. List the estimated regression coefficients and the p-values for the corresponding t-tests for these coefficients. Interpret the estimated model coefficient if the t-test result was significant.

ii. One commonly-used method to find a parsimonious model is stepwise regression with AIC. In backward elimination, it starts with all the potential predictors in the model, then removes the predictor with the largest p-value each time to give a smaller AIC. The forward selection method is the reverse of the backward method. It starts with no explanatory variable in the model, then adds one predictor at a time (with the smallest p-value) until no further variables can be added to produce a smaller AIC value. Stepwise regression alternates forward steps with backward steps. The idea is to end up with a model where no variables are redundant given the other variables in the model. Often, in practice, backward elimination and forward selection will produce the same 'final' model.

Start with the full model ('`fullmodel`') fitted in part i above and use backward elimination with AIC. What is the final model (write the fitted model)? Are the results consistent with those in part i?

iii. Use BIC instead of AIC and repeat part ii above. What is the final model? Are the results consistent with what you saw in parts i and ii? Explain.

Here are some R codes you may use for parts ii & iii:

```
# backward AIC
step(fullmodel, direction = "backward")
# backward BIC; n is the number of data points
step(fullmodel, direction = "backward", k=log(n))
```

IV. Discussion and Limitations section.

(a) Using a 2-by-2 layout, show the 4 diagnostic plots that are obtained in R by plotting the model obtained in part III.iii above.

(b) Interpret each of the four residual plots given in part (a) above. Discuss whether the normal error MLR assumptions are satisfied.

(c) Discuss the next steps you would take towards finding a valid 'final' model.