STA130 Fall 2019 – T0107

Week 10: Multiple Linear Regression

(Materials used in this presentation are provided by the U of T Statistical Sciences Department.

This presentation was prepared by Vivian Ngo.)

Github.com/vivianngo97/STA130-Fall-2019

viv.ngo@mail.utoronto.ca

Agenda

- **This is the LAST tutorial!**
- **Poster expectations and reminders**
- **Material & Vocabulary Review**
- **Group work and presentations**
- **Time to work on poster project**

Poster Project Reminders

- Submission links, rubrics, schedule, etc: On QUERCUS
- Deadlines are STRICT
- Arrive on time to put up and take down your poster. Any work left behind will be discarded.
- Arrive early for your scheduled poster presentation time. **If you arrive late, you will not be marked.** This is a large class which requires a lot of coordination. As such, you have a designated presentation time for a reason.
- Only **one group member** needs to submit their poster pdf and rmd files on behalf of the group.
- You must complete the group evaluation form on Quercus before midnight on the day of the poster fair (Thursday, Dec 5th).
- Make note of your group number – you will need to know your group number to find your poster board during the poster fair (Emailed)

Group numbers

Group #	Tutorial TA	Group member 1	Group Member 2	Group Member 3	Group Member 4
107-1	Vivian	Johnson Vo	Nicole Sin	Shuyi Zuo	Stephanie Cristea
107-2	Vivian	Jiaqi Wu	Zhexi Guan	Xiao Bai	Yichun Zhang
107-3	Vivian	Yian Hu	Yuhan Chen	Haifeng Sun	Xiao Yan
107-4	Vivian	Yilun Chen	Dechen Han	Dickson Li	Miakrishna Cherthedath
107-5	Vivian	Xiaoqian Wang	Tyler Kelly	Xiaozhou Ye	

Poster Project Reminders

- **You will be reviewing one group during poster day**. The review schedule is posted on Quercus –know which group you are reviewing and when. You will need to complete the oral presentation form and write down (and ask) at least one question to the group. This form must be uploaded to Quercus before midnight on poster day. Paper copies will be available at the poster fair.
- A poster board with Velcro strips to hang up your printed sheets will be provided. You only need to bring your printed poster pages.
- **Every group member is expected to speak.** Members who do not speak will be penalized. See the poster page on Quercus for further details.
- **Your poster will be graded on content.** It should stand alone (it will be marked by your TA who will not have seen your presentation). The #1 mistake we see is that people do not **proofread** their work.
- We will have **extra TAs in the Stats Aid center** the week of the poster (Monday and Tuesday) to help you. Come prepared. It is NOT our job to work on your project for you!

Poster Project Reminders

- We can also answer questions through Piazza all week, but this is not an appropriate forum for extensive programming help. Questions asked after 3pm on Wed Dec 4th **WILL NOT** be answered until after the poster fair.

Poster Project Reminders

- GRADING:
- Reasonable scope: your research question should be **clear and answerable**
- **Data wrangling:** creative use of data wrangling to produce ***informative* variables**. Think about if your variables make sense and how they are useful for answering your question.
- Graphical display. Use **appropriate (and creative) figures** to tell your story. Remember, these need to have clear and meaningful titles (i.e., don't label with an R variable name if it doesn't make sense). **CRITICALLY, your figures need to stand alone.** Note that providing pages of data tables is not very meaningful for your reader. It's your job as a data scientist to turn this into something your audience can easily understand. REMEMBER: these are intended for TPS. (Use captions?)

Poster Project Reminders

- GRADING (continued)
- Statistical methods. The choice of methods should be **appropriate, complete and creative.** You shouldn't do EVERYTHING you've learned this term but pick the methods that make the most sense. You will be penalized if you use EVERY method because this doesn't make sense. Think about what a "method" is. Using an R command is NOT a method. Cleaning your data is NOT a method. Using linear regression to study the relationship between X and Y IS a method.
- Appropriate conclusion. Your conclusion should clearly follow the work you've done and your **results**. State any **limitations**. For example, maybe there is something you'd like to study but it is not available in the data.

Poster Project Reminders

- GRADING (continued):
- Organization. The contents should be ordered logically, use of **sub-headers** is recommended, e.g. **Background, Methods, Results, Discussion, Conclusions.**
- WOW Factor: bonus points for doing something very creative, e.g. using an additional data source, creating a new (meaningful) variable, etc.
- Your rmd file will also be marked, so make sure it runs smoothly. We suggest **annotating** your code so that your TA can navigate it better.

Poster Project Reminders

- DO NOT LEAVE THE PROJECT UNTIL THE LAST MINUTE. It takes time to do this properly and if you run into any last-minute issues, we will **not** be there to help you. By Wednesday Dec 4th, your project should be *entirely* finished – Wednesday should only be for practicing your presentation.
- PRACTICE, PRACTICE, PRACTICE! Ideally, practice as a group as many times as you can so that your transitions seem more natural.

Material and Vocabulary Review

- Multiple linear regression
- Categorical predictor
- Dummy (or indicator) variable
- Baseline value
- Interaction
- Modification/ Modifier
- Extrapolation

Material and Vocabulary Review

- Dummy Coding/Indicator Variables
- Example: Question 1c

The regression model is:

$$totalPr = \beta_0 + \beta_1 seller\_rating\_is\_low + \beta_2 seller\_rating\_is\_medium + \beta_3 duration$$

$$+\beta_4 seller\_rating\_is\_low \times duration + \beta_5 seller\_rating\_is\_medium \times duration + \epsilon$$

```
summary(lm(totalPr ~ seller_rating*duration, data=marioKart2))$coefficients
```

```
##                                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)                  55.399199  1.9003593 29.1519610 4.101299e-60
## seller_ratinglow             -8.185758  3.6192132 -2.2617507 2.531113e-02
## seller_ratingmedium          -2.387931  3.0064661 -0.7942651 4.284351e-01
## duration                     -2.937082  0.7652626 -3.8380058 1.897989e-04
## seller_ratinglow:duration     2.620252  0.9533562  2.7484504 6.807181e-03
## seller_ratingmedium:duration  1.538756  0.8856835  1.7373654 8.460333e-02
```

ii. What is the fitted regression line for sellers with low ratings?

The regression model is:

$$totalPr = \beta_0 + \beta_1 seller\_rating\_is\_low + \beta_2 seller\_rating\_is\_medium + \beta_3 duration$$

$$+\beta_4 seller\_rating\_is\_low \times duration + \beta_5 seller\_rating\_is\_medium \times duration + \epsilon$$

```
summary(lm(totalPr ~ seller_rating*duration, data=marioKart2))$coefficients
```

```
##                                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)                  55.399199  1.9003593 29.1519610 4.101299e-60
## seller_ratinglow             -8.185758  3.6192132 -2.2617507 2.531113e-02
## seller_ratingmedium          -2.387931  3.0064661 -0.7942651 4.284351e-01
## duration                     -2.937082  0.7652626 -3.8380058 1.897989e-04
## seller_ratinglow:duration     2.620252  0.9533562  2.7484504 6.807181e-03
## seller_ratingmedium:duration  1.538756  0.8856835  1.7373654 8.460333e-02
```

ii. What is the fitted regression line for sellers with low ratings?

$$\widehat{totalPr} = 55.40 - 8.19 + (-2.94 + 2.62) \times duration$$

iii. What is the equation of the fitted regression line for sellers with medium ratings?

The regression model is:

$$totalPr = \beta_0 + \beta_1 seller\_rating\_is\_low + \beta_2 seller\_rating\_is\_medium + \beta_3 duration$$

$$+\beta_4 seller\_rating\_is\_low \times duration + \beta_5 seller\_rating\_is\_medium \times duration + \epsilon$$

```
summary(lm(totalPr ~ seller_rating*duration, data=marioKart2))$coefficients
```

```
##                                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)                  55.399199  1.9003593 29.1519610 4.101299e-60
## seller_ratinglow             -8.185758  3.6192132 -2.2617507 2.531113e-02
## seller_ratingmedium          -2.387931  3.0064661 -0.7942651 4.284351e-01
## duration                     -2.937082  0.7652626 -3.8380058 1.897989e-04
## seller_ratinglow:duration     2.620252  0.9533562  2.7484504 6.807181e-03
## seller_ratingmedium:duration  1.538756  0.8856835  1.7373654 8.460333e-02
```

ii. What is the fitted regression line for sellers with low ratings?

$$\widehat{totalPr} = 55.40 - 8.19 + (-2.94 + 2.62) \times duration$$

iii. What is the equation of the fitted regression line for sellers with medium ratings?

$$\widehat{totalPr} = 55.40 - 2.39 + (-2.94 + 1.54) \times duration$$

iv. What is the equation of the fitted regression line for sellers with high ratings?

The regression model is:

$$totalPr = \beta_0 + \beta_1 seller\_rating\_is\_low + \beta_2 seller\_rating\_is\_medium + \beta_3 duration$$

$$+\beta_4 seller\_rating\_is\_low \times duration + \beta_5 seller\_rating\_is\_medium \times duration + \epsilon$$

```
summary(lm(totalPr ~ seller_rating*duration, data=marioKart2))$coefficients
```

```
##                                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)                  55.399199  1.9003593 29.1519610 4.101299e-60
## seller_ratinglow             -8.185758  3.6192132 -2.2617507 2.531113e-02
## seller_ratingmedium          -2.387931  3.0064661 -0.7942651 4.284351e-01
## duration                     -2.937082  0.7652626 -3.8380058 1.897989e-04
## seller_ratinglow:duration     2.620252  0.9533562  2.7484504 6.807181e-03
## seller_ratingmedium:duration  1.538756  0.8856835  1.7373654 8.460333e-02
```

ii. What is the fitted regression line for sellers with low ratings?

$$\widehat{totalPr} = 55.40 - 8.19 + (-2.94 + 2.62) \times duration$$

iii. What is the equation of the fitted regression line for sellers with medium ratings?

$$\widehat{totalPr} = 55.40 - 2.39 + (-2.94 + 1.54) \times duration$$

iv. What is the equation of the fitted regression line for sellers with high ratings?

$$\widehat{totalPr} = 55.40 - 2.94 \times duration$$

Material and Vocabulary Review

- Estimating vs predicting
  - An *estimator* uses data to guess at a parameter while a *predictor* uses the data to guess at some random value that is not part of the dataset.

Material and Vocabulary Review

- The response variable, Y, doesn't need to be normal as this is not an assumption of multiple regression. In fact, Y will rarely be normal. What must be true is that the errors around a prediction Y^ must be normal which one can check with a normal plot of the residuals. The errors must also have a constant variance which you can check with a predicted by residual plot (more on this in future stats classes).

Material and Vocabulary Review

- Regression does not assume that the regressors have any distribution. One should, however, use box plots to check for outliers in the regressors. You need a *very good* reason to remove an outlier! This decision needs to be adequately justified.

Oral Presentations

- 1. (Based on Practice Problem 1e) Does seller rating appear to modify the association between duration and total price? Justify your answer using the results from previous questions; i.e., 1a-d. Do you feel comfortable using your model to predict durations >10 days? Why or why not?
- 2. (Based on Practice Problem 1f-g) Which of the models you considered in question 1f shows the best predictive performance? Justify your answer using the results from Questions 1f and 1g. Are there any limitations to these models?
- 3. (Based on Practice Problem 3a-c) Is there a linear association between the size of the house (in square feet) and the sale price? Explain. Are there other variables that might be important?
- 4. (Based on Practice Problem 3d-e) Is there an association between presence of a fireplace and the sale price of a house? Explain. Do you think that presence of a fireplace might be associated with other characteristics of a house? E.g. it's size, design, appeal, etc. How might this influence your interpretation of these findings?
  (NOTE to TA: This question is hinting at confounding – which they will learn next week. Houses with fireplaces are probably larger and more nicely designed, which would also increase their price).
- 5. (Based on Practice Problem 3f-g) Which of the four models you considered is most suitable for prediction? How well does each model generalize to new observations? Which would you prefer, a simpler model or more complicated one? Explain.