PHD-302 Introduction to Open Data Science 2022

Assignment 1

Rong Guang

06/11/2022

Contents

1 About the project

Write a short description about the course and add a link to your GitHub repository here. This is an R Markdown (.Rmd) file so you should use R Markdown syntax.

```
# This is a so-called "R chunk" where you can write R code.

date()
```

```
## [1] "Sun Nov  6 20:19:25 2022"
```

2 1. Thoughts about the course

2.1 1.1 My feeling about the course

I am feeling very good, since I finally start using GitHub from scratch. **All my colleagues** are using GitHub for collaboration, but I do not even have any idea what it is. This course is very timely. Hopefully after this course, I can join their collaboration loop effectively.

2.2 1.2 What do I expect to learn

I do not have any particular expectation in the statistical knowledge I am going to learn, because anything about statistics is interesting to me. Actually, I hope in the course I could have chance of exercising **GitHub collaboration in practical manner**. For example, my colleagues always mention the creation of some 'branches' of some "root" data sets, which I still do not understand very clearly what is it.

3 2. reflect on my learning with the R for Health Data Science book

3.1 2.1 How did it work as a "crash course" on modern R tools and using RStudio?

Luckily, I just finished taking Prof. Kimmo's another course "Quantitative Research Skills", where I got chance to walk through several chapters of **R for Health Data Science book** . This way I suppose I could make some comments on the book retrospectively. I would say this is the best book for novice R users to start off. The author did the instruction from an industry perspective (health science), which avoid getting too deep in math but focuses on the R as a statistical tool for solving industry problem itself. When I tried to start with R years ago, one of the biggest barrier is for every statistical feature I want to realize, R always present a multitude of pathways to realize it (for example, to compute a new variable for a data set, you can you ".[,c] <-", or ".$c<-", or " %>% " + "mutate"), other source of material/textbook always tried to feed you with all these possibilities, which makes things complicated. This book just doesn't. For each function to execute, it only show you one approach. This is just what most non-statistics-majors want!

3.2 2.2 Which were your favorite topics?

Data wrangling, definitely (although there is no specific chapter for it, but it scatters throughout the whole book). Most descriptive, inferential and exploratory statistics are so easy to realize in R. By the merit of R, for these purposes, all I need is first, know what I am going to do (base on my stat knowledge), and then search in Google about which package(s) are for doing that, and finally install it, read the instruction&example and then do it. Very mechanical. There is never a big challenge. To me, data wrangling (more than dozens of variable) requires some real effort, and hence more interesting. And there is never an end. Today I realize a wrangling task in three lines of codes, maybe tomorrow I could come up with or stumble over a way that only requires two, or even one. This process is super exciting.

3.3 2.3 Which topics were most difficult?

Data wrangling, of course according to my story above, followed by logistic regression. For logistic regression, I have several big unsolved problem on mind. They are: **a.** considering there is not an alternative of linear regression's r square for logistic regression, how do I evaluate the variance that my model could explain? **b.** after modeling additive effect, model modification is recommended to follow. It is not difficult to create a main effect plot to observe candidate multiplicative variables if **y** is continuous. But is it also applicable to a binary **y** ? However, I suppose this is mainly due to my lack of relevant stat knowledge, having nothing to do with R language.