Midterm project

June 24, 2021

1 Midterm project

Congratulations! You've been hired as a data scientist at the hottest new social media startup.

Your company produces an app via which users can post short videos for anyone to view. They can also like, repost, and comment on the videos they view. The key data product is a recommendation engine that determines the order in which videos are shown to a user.

The recommendation engine has a parameter, *theta*, that affects the ordering of the videos. Recently the team of engineers that works on the recommendation engine ran it with different settings of *theta* and, for each setting, measured the amount of time users spent on the app. They have collected these measurements into a data set of 20 samples of (*theta*, *time_spent*) pairs.

Additionally, they have identified two auxiliary features (*aux1* and *aux2*) that they hypoithesize should correlate with *time_spent*. These two features are measures of time spent by users in the recent past. The engineers have not verified that the features explain *time_spent*.

(The engineers call these two features "auxiliary" because, while they might help explain *time_spent*, the engineers' ultimate interest lies in the dependence of *time_spent* on *theta*.)

Your first project at your new company is to tell the engineers which setting you think they should use for *theta*, based on the data.

1.1 1. Prepare the data

- Inspect the data. Identifiy and remove any suspicious or unusable samples.
- Put the samples in a data structure that you can work with.

```
In [2]: theta = [0.03906292, 0.05119367, 0.06004468, 0.06790036, 0.19152079,
                 0.28298816, 0.294665  , 0.3578136 , 0.48352862, 0.53058676,
                 0.55175137, 0.57560289, 0.59751325, 0.6375209 , 0.65241862,
                 0.65633352, 0.78698546, 0.8640421 , 0.87729053, 0.94568319]
        aux1 = [ 0.53983961, -1.77528229,  1.31487654, -0.47344805, -1.0922299 ,
                -0.25002744, -0.9822943 ,  1.03126909,  0.49133378, -0.4466466 ,
                -0.80636008,  0.13126776, -1.21256024,  0.15999085, -0.75522304,
                 0.34989599,  0.97754176, -0.13858525,  0.10385631,  0.30059104]
        aux2 = [ 0.9682053 ,  0.86962384,  0.56778309,  0.46528234, -1.16537308,
                -2.03599479, -1.15541329,  3.34515739,  0.12672721, -0.6941789 ,
                 0.55767443,  0.0991466 ,  0.63792617,  0.70311068, -0.91609315,
                -0.78601423,  1.1191818 , -0.98339611,  0.24452002, -0.58140974]
        time_spent = [10.79768391, 10.87648065, 10.29274937, 10.78756647,  9.51844772,
                9.18078781,  9.90063639, 12.84823357, 10.92743478,  9.88927608,
```

```
11.3373709 , 11.43996915, 11.88392171, -11.88135476, 11.73452467,
11.18844425, 12.19144316, 11.35294826, 12.2385441 , 11.98428985]
```

1.2 2. Build a model

Write functions to run a regression, calculate the regression statistics listed below, and print a report. - B (regressor coefficients plus one for an intercept, if appropriate) - R2 - RSS - RegSS - TSS - t statistic for each regressor coefficient

I found it useful to decompose the problem into three functions: `regress_calc()`, `regress_tstat()`, and `regress_report()`. You may write it however you see fit.

You may include either, both, or neither of *aux1* and *aux2* in your final model. Experiment. What works best? Justify your decision.

1.3 3. Propose a setting for *theta*

Now that you have a model built, you should be able to plot estimated *time_spent* vs. *theta* over a reasonable range of *theta*. By inspecting that plot – and knowning that the company wants to maximize the time users spend on the app – which value of *theta* would you propose the engineers use? Explain how the data and your model support your decision.

The engineer's have capacity to take another set of measurements. Which settings of *theta* do you suggest they measure? Why?

1.4 4. Experiment or observation?

Is this data set experimental or observational? Explain clearly. Consider how the effect of *theta* on *time_spent* differs from the effect of *aux1* or *aux2*.

1.5 Presentation

You will give a 5 minute presentation in class. Your presentation should focus on step 3, above: Summarize your results, display any relevant visualization, and provide guidance to the engineers in the form of recommended setting for *theta*. Also please let the engineers know which value(s) of *theta* you think they should measure next.