Problem Set 1

Predictive Analytics for Business Strategy

Spring 2022

Due: Sunday, January 15th, at **11:59pm (exactly, please submit early)**.

McDermott

Include everyone’s name on the submission and paste any code used into the problem set. Also paste the code used to generate the results you arrived at (if there’s an error, this may help explain it).

Add your solutions to this file, putting your solutions in **bold**.

1. Suppose you work for a firm that supplies electricity from renewable sources and you operate in a market that allows prices to vary completely according to market conditions. Your firm of course wishes to price optimally, so they want to know the causal impact of price on quantity demanded. We have no data for this question but instead want to practice the initial steps of the thought process we’ll need for identifying causal relationships (when we are working with observational data). What are some other variables that might correlate with price (name 2+)? Discuss each and whether it directly impacts the quantity sold.
2. Many times when you have two populations/groups, you are interested in knowing whether one population’s mean is significantly different from the other’s (something like a treated group vs. an untreated group). Open the **googleplaystore.xlsx** in Stata and do a difference in means t test for ratings of paid and Free apps. Since Stata sees Rating as a string, use the following code and then using ratingNum instead of rating for calculations and regressions: **gen ratingNum = real(Rating)**
   1. Display and explain the output of this test.
   2. What does the result tell you and does it imply the pricing ***caused*** a significant difference in the groups (if significant)? Explain why you do/don’t think so.
3. Again using the googleplaystore.xlsx file, do a single population t test to test whether the ratings of paid apps is significantly different from zero. Assuming you used the extra code I added in question 2, use this code and then use the variable paidRating while working on this question: **gen paidRating = ratingNum if Type == "Paid"**
   1. Display and explain the output of this test.
   2. Does your result imply the pricing (pricing meaning free or not) ***caused*** this significant difference (if it is significant)? Explain why you do/don’t think so.
4. Use the PS1.xlsx to answer the following questions.
   1. Run a two population t test to test whether there is a difference in the likelihood of severe Covid cases between the vaccinated and unvaccinated in this data set. Display and explain the results.
   2. Regress severe covid case on vaccine – display and discuss the results.
   3. Assuming this is a randomized, controlled experiment, do you think we can attribute the difference in likelihood of severe Covid cases to being caused by the vaccination status? Explain why you do/don’t think so.
5. Considering the previous question about having a severe case of covid and whether or not someone was vaccinated, suppose you have separate data that is not from an RCT. Considering the impact of vaccination status (X) on the likelihood of getting a severe case of Covid, what are other variables that might be correlated with vaccination status? For each, discuss whether it is positively correlated or negatively correlated?