Data Challenge READ-ME

Alice Hsu

February 2020

1 Hello

Thank you for your patience, and thank you in advance for taking the time to read all of this. Quite a lot of time has passed since I was last in contact with you, and I want to thank you for considering me for one of the positions at the Systems Medicine Lab. I have quite enjoyed working on these problems that you have posed–spending time on these problems has been more exciting and frustrating than I've felt in a long time at my current job.

That being said, I wholly underestimated the the amount of time I had and overestimated the amount of coding skill I had when it came to this project. It was definitely a challenge for someone who has never actually used R to wrangle any kind of data. Many days were spent poring over R documentation, StackExchange, and online copies of coding textbooks on Google Scholar, trying to understand what exactly it was I was supposed to do and how to get there. Then, once I thought I was finished, I saved everything and exited out. BIG mistake, because when I tried to open up everything again, it seems that while my workspace was saved, the contents of my script were not. So, I have nothing to show, really, besides what I managed to copy over into this file.

At this point, I'm still not sure what I did, nor could I possibly tell you what went wrong, but I will do my best to walk you as clearly through my thought processes as possible. In any case, I felt that I'd dragged this out for long enough. Please take a look through this document to see what I've accomplished. I totally understand if you never talk to me again. Thank you again for the opportunity to learn quite a lot about data science and its applications to health care. You and your team have helped me realized quite a lot about myself and where I want to take my career. I do hope to continue along and eventually end up doing something akin to what your lab does, but I do believe that I also need QUITE a lot more training and education. If it turns out that this long, ramble-y letter has convinced you to take me on as a novice data scientist, that would be the best case scenario. I know I'll work long and hard at whatever you throw my way because I genuinely do enjoy the challenge of teaching myself new skills and strategies. Nevertheless, I am a realist who knows you're probably looking for someone with more skill than I.

2 Part 1: Basics

Problem 1. **Describe the data** (i.e. what information is provided in the dataset?), in a few sentences. What does it appear we've been given?

The dataset consists of patient records from Massachusetts hospitals between 1920 and 2018. There are 147,699 observations across 28 variables including allergy data, encounters, drugs prescribed, vitals, etc. It is highly redundant due to the nature of how every single hospital service is recorded with a unique code, and not all entries contain all the same information.

Problem 2. **Summary statistics.** Create a sample of patients with emergency visits between from 2008-2016. Most medical papers begin with a ”Table 1” that summarizes basic aspects of the sample (see this paper for an example). Create a similar ”Table 1” that describes basic attributes of this sample, including n (visits and patients), demographics (age, sex, race), and the medical conditions diagnosed over the year prior to the emergency visits. We don't have the exact same illness categories as in the linked paper, so please substitute rows for the most common diagnoses you find in the data. If anything needs further explanation, just put a footnote in the table.

The table is included below.

The code is titled ”data-challenge-hsu”. To make this table, I first found all the unique patients who had visited the emergency room between 2008-2016, inclusive, to build a subset that contained only the data for those who visited, extracting their zip code information so that I can merge this data with household-level income data from ACS. From there, I built a summary table that contained the top 14 diagnoses, using keywords to combine similar diagnoses, as well as demographic data.

	. (N = 59,964)
Emergency Room Visits (2008-2016)	
No. of vists	922
No. of Patients	570
Most common diagnoses in the 12 months prior to visit (2007-2016)	
Total No. of Diagnoses	5718
Pregnancy	2304
Sinusitis	630
Bronchitis	529
Fracture	329
Pharyngitis	324
Child attention deficit disorder	321
Lacerations	179
Asthma	165
Streptococcal sore throat	111
Concussions	73
Otitis Media	87
Polyps of Colon/Rectum	66
Lung Cancer	63
Emphysema	50
Demographic data	
Mean age (sd)	41.77 (24.11)
Median age (IQR)	39.00 (21.00, 59.00)
Sex	
Female	320 (56)
Male	249 (44)
Race	
White	426 (75)
Hispanic	52 (9)
Black	51 (9)
Asian	40 (7)
Income Data	
Mean Household Income (sd)	85,495.11 (30,305.39)
Median Household Income (IQR)	82,159.00 (64,544.50, 102,187.50)

Problem 3. **Run a simple model.** We're interested in the relationship between mortality after these visits, and income and race. To (roughly) measure income, we'll use zip-code level median household income in the past 12 months from the 2012-2016 ACS– oh, and please go back and add this variable to your "Table 1 above." Then, using regression, model mortality in the year after emergency visits as a function of income, race and present the model output. You can use these ugly screenshots as a general guide for what key information to present, but you don't need to replicate it exactly or present all the same information.

Residuals:				
Min	1Q	Median	3Q	Max
-0.5880	-0.3282	-0.3185	0.6613	0.7080
Coefficients:				
	Estimate	Std. Error	t value	Pr(> $\lvert t\rvert$)
(Intercept)	3.571e-01	1.926e-01	1.854	0.0688 .
HouseholdsMA:race.fasian	2.705e-06	3.802e-06	0.712	0.4796
HouseholdsMA:race.fblack	-3.731e-06	4.113e-06	-0.907	0.3681
HouseholdsMA:race.fhispanic	-3.560e-07	2.879e-06	-0.124	0.9020
HouseholdsMA:race.fwhite	-4.105e-07	2.331e-06	-0.176	0.8608

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4826 on 58 degrees of freedom (17 observations deleted due to missingness) Multiple R-squared: 0.03526, Adjusted R-squared: -0.03127 F-statistic: 0.53 on 4 and 58 DF, p-value: 0.7141

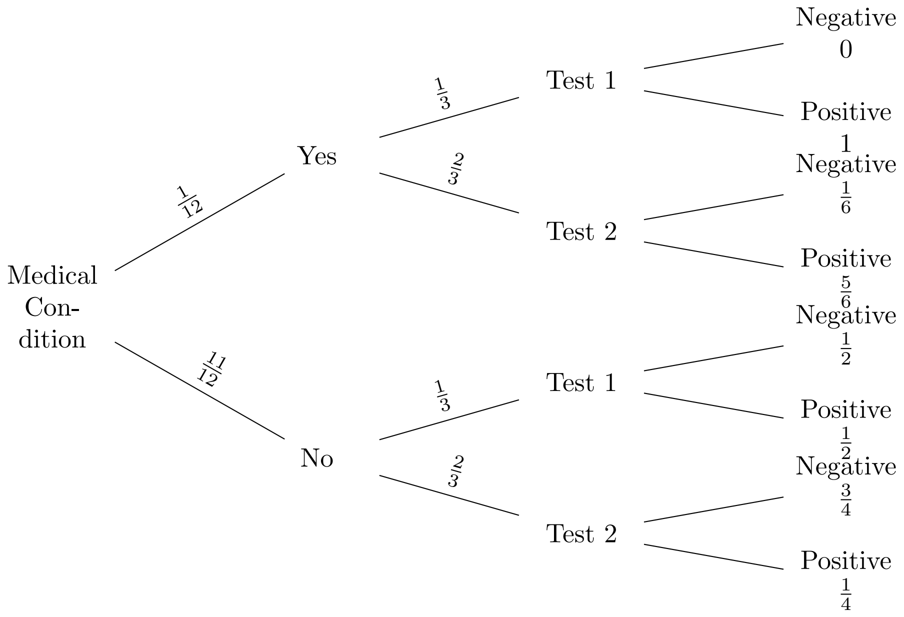
3 Part 2: $\binom{4}{2}$

Problem 1. Imagine you are partnering with a large hospital network (comprising 15 hospitals) and your task is to develop a model that predicts how many emergency visits they will have on any given day. They are giving you access to all the historical emergency department data from 3 of their hospitals spanning 2008-2016. They hope to implement your model across the entire network of hospitals for the upcoming year. In a few sentences, take us through your thoughts about setting up a process to build and evaluate your model. You do not need to talk about specifics (e.g. you don't need to talk about which kind of algorithm you would use) – just your high-level strategy for the design of the algorithm.

First, I'd probably determine how many emergency visits each day between 2008-2016 receives. I want to see if there are any clusters of dates that see a spike in emergency visits and if they correlate with national data (e.g. holidays, big events, etc). Given the regularity of these dates, we may be able to write a Fourier series or other equation that can predict the number of emergency visits for any day based on certain variables such as weather, traffic, etc. This can be followed by training, using a dataset similar to the one we used at the beginning and then implementation. As long as there are no large natural or manmade disasters, I imagine this will work for a day-to-day basis.

Problem 2. Consider a particular medical condition with $\frac{1}{12}$ prevalence in the population. There are two tests for this condition and doctors randomly choose test 1 with probability $\frac{1}{3}$ and test 2 with probability $\frac{2}{3}$. If a given individual has the condition, test 1 would correctly identify this with probability 1 and test 2 would correctly identify this with probability $\frac{5}{6}$. If a given individual does not have the condition, test 1 correctly identifies this with probability $\frac{1}{2}$ and test 2 correctly identifies this with probability $\frac{3}{4}$. Imagine an individual is chosen completely at random from the population, they receive one of the tests, and the test comes out positive. What is the probability that the individual actually has the disease?

This is an implementation of Bayes' Theorem.



We can mathematically represent the above statement as $P(diseased|tested, positive)$. By the definition of Bayes' Theorem, this can be expanded to

$$P(diseased|tested, positive) = \frac{P(diseased)P(tested, positive|diseased)}{P(tested, positive)}$$

The overall probability of this random individual having this medical condition is $P(diseased) = \frac{1}{12}$. The probability that they were tested and tested positive given that they have the medical condition is the sum of the products of the branches following "Yes...Positive". Thus, $P(tested, positive|diseased) = \frac{1}{12}(\frac{1}{3} * 1 + \frac{2}{3} * \frac{5}{6}) = \frac{2}{27}$. Finally, the probability that a random individual tested positive at all is simply the sum of the true positives and the false positives: $P(tested, positive) = P(tested, positive|diseased) + P(tested, positive|notdiseased) = \frac{2}{27} + \frac{11}{12}(\frac{1}{3} * \frac{1}{2} + \frac{2}{3} * \frac{1}{6}) = \frac{2}{27} + \frac{11}{36} = \frac{41}{108}$. Substituting in all of these numbers, we obtain

$$P(diseased|tested, positive) = \frac{\frac{1}{12} \cdot \frac{2}{27}}{\frac{41}{108}}$$

$$P(diseased|tested, positive) = \frac{2}{123}$$

Or, 1.6%.