CS 476/676 (Spring 2021): Homework 1

Due: Feb 3, 2021 at 11:59pm EST

Name: Josh Popp

Instructions: This homework requires some open-ended questions, short proofs, and programming. This is an individual assignment, not group work. Though you may discuss the problems with your classmates, you must solve the problems and write the solutions independently. As stated in the syllabus, copying code from a classmate or the internet (even with minor changes) constitutes plagiarism. You are required to submit your answers in pdf form (use LaTeX) in a file called `<your-JHED>-hw1.pdf` to Gradescope. (Note: for those who not familiar with LaTeX, you can visit `https://www.overleaf.com` to quickly start by using the provided .tex template file.) The code for the programming assignment will be handed in separately on github; see the instructions under Problem 8. Late submissions will be penalized, except in extenuating circumstances such as medical or family emergency. Submissions submitted 0-24 hours late will be penalized 10%, 24-48 hours late by 20%, 48-72 hours late by 30%, and later than 72 hours by 100%. Late days may be used (if available) to avoid these penalties. Problems 1 and 2 are worth 5 points each, Problem 3 is worth 10 points, and Problem 4 is worth 15 points, for a total of 35 points.

Problem 1

This is an open-ended question to get you a head start on thinking about what you would like to do for your final project. Write a brief description of the kinds of data you have worked with (if you have not worked with data before, it is ok to say so) and would like to work with in this course. What questions did you/would you like to answer with such data? There are no right or wrong answers; a paragraph or two is sufficient.

I've worked primarily with genetic and gene expression data, and would like to continue doing so. I've especially focused on single cell RNA sequencing, and would be interested in doing more with this, but would also like to gain experience with different modalities or even biological imaging data. Some of the questions I'm interested in are:

- How do gene expression patterns change during cellular differentiation?
- How can we use gene expression or other modalities (chromatin accessibility, spatial transcriptomics) to robustly characterize cellular differentiation?
- What technical and biological (especially genetic) factors impact cellular differentiation, particularly of induced pluripotent stem cells?
- How can we leverage molecular perturbations of the genome to disentangle correlation from causal effects?

Problem 2

Prove that $X \perp\!\!\!\perp Y \mid Z$ implies that $\mathrm{OR}(X, Y \mid Z) = 1$ for all x, y, z.

$$\begin{aligned}
OR(X \perp\!\!\!\perp Y \mid Z) &= \frac{p(X = x \mid Y = y, Z)}{p(X = x_0 \mid Y = y, Z)} * \frac{p(X = x_0 \mid Y = y_0, Z)}{p(X = x \mid Y = y_0, Z)} \\
&= \frac{p(X = x \mid Y = y, Z)}{p(X = x \mid Y = y_0, Z)} * \frac{p(X = x_0 \mid Y = y_0, Z)}{p(X = x_0 \mid Y = y, Z)} \\
&= \frac{p(X = x \mid Z)}{p(X = x \mid Z)} * \frac{p(X = x_0 \mid Z)}{p(X = x_0 \mid Z)} \\
&= 1
\end{aligned}$$

Problem 3

Consider three sets of random variables X, Y, and Z. Assume that X and Y are binary random variables and the relations between X, Y, and Z can be modeled via a linear logistic regression (i.e., no interaction terms.) That is,

$$p(X = 1 \mid Y, Z) = \frac{1}{1 + \exp\big(- (\beta_0 + \beta_1 Y + \beta_2 Z) \big)}.$$

Let the reference values of X and Y be $x_0 = 0$ and $y_0 = 0$ respectively. Prove

$$\mathrm{OR}(X, Y \mid Z) = \begin{cases} \exp(\beta_1), & \text{if } X = Y = 1 \\ 1, & \text{otherwise.} \end{cases}$$

$$\begin{aligned}\mathrm{OR}(X,Y\mid Z) &= \frac{p(X=x|Y,Z)}{p(X=x_0|Y,Z)} * \frac{p(X=x_0|Y=y_0,Z)}{p(X=x|Y=y_0,Z)}\\ &= \frac{p(X=x|Y,Z)}{p(X=0|Y,Z)} * \frac{p(X=0|Y=0,Z)}{p(X=x|Y=0,Z)}\end{aligned}$$

Note that

$$\begin{aligned}
p(X=1\mid Y=0,Z) = p_1 &= \frac{1}{1+\exp\big(-(\beta_0+\beta_2 Z)\big)}\\
p(X=0\mid Y=0,Z) = p_2 &= \frac{\exp\big(-(\beta_0+\beta_2 Z)\big)}{1+\exp\big(-(\beta_0+\beta_2 Z)\big)}\\
p(X=1\mid Y=1,Z) = p_3 &= \frac{1}{1+\exp\big(-(\beta_0+\beta_1+\beta_2 Z)\big)}\\
p(X=0\mid Y=1,Z) = p_4 &= \frac{\exp\big(-(\beta_0+\beta_1+\beta_2 Z)\big)}{1+\exp\big(-(\beta_0+\beta_1+\beta_2 Z)\big)}
\end{aligned}$$

and

$$\begin{aligned}
\frac{p_3}{p_4} * \frac{p_2}{p_1} &= \frac{\frac{1}{1+\exp\left(-(\beta_0+\beta_1+\beta_2 Z)\right)}}{\frac{\exp\left(-(\beta_0+\beta_1+\beta_2 Z)\right)}{1+\exp\left(-(\beta_0+\beta_1+\beta_2 Z)\right)}} * \frac{\frac{\exp\left(-(\beta_0+\beta_2 Z)\right)}{1+\exp\left(-(\beta_0+\beta_2 Z)\right)}}{\frac{1}{1+\exp\left(-(\beta_0+\beta_2 Z)\right)}}\\
&= \frac{\exp\big(-(\beta_0+\beta_2 Z)\big)}{\exp\big(-(\beta_0+\beta_1+\beta_2 Z)\big)}\\
&= \exp(\beta_1)
\end{aligned}$$

Therefore

$$\mathrm{OR}(X,Y\mid Z) = \begin{cases} \frac{p_3}{p_4} * \frac{p_2}{p_1} = \exp(\beta_1), & \text{if } X=Y=1\\ \frac{p_4}{p_4} * \frac{p_2}{p_2} = 1, & \text{if } X=0, Y=1\\ \frac{p_1}{p_2} * \frac{p_2}{p_1} = 1, & \text{if } X=1, Y=0\\ \frac{p_2}{p_2} * \frac{p_2}{p_2} = 1, & \text{if } X=0, Y=0 \end{cases}$$

Problem 4

Implement a method/function which takes data in the form of a comma separated file (csv) and returns a point estimate for the odds ratio via logistic regression, and confidence intervals via bootstrap for the same. We recommend using the `statsmodels` package in Python for fitting any regression models in this course – the reason for this is because `sklearn` does not allow you to (easily) turn off regularization. Regularization is useful when considering out-of-sample prediction, but will bias your estimate of the statistical parameters we are interested in computing.

Use your function to compute point estimates and confidence intervals for $\text{OR}(\text{Mortality}, \text{Opera})$ and $\text{OR}(\text{Mortality}, \text{Opera} \mid \text{Income})$ from the data provided in `data.txt`. Set the number of bootstraps to 200 when computing your confidence intervals. Based on the numbers you have computed, would you say

a) Mortality $\perp\!\!\!\perp$ Opera?
b) Mortality $\perp\!\!\!\perp$ Opera $\mid$ Income?

Provide a brief comment on how seriously the above numbers you computed and the implied dependence/independence relations should be interpreted causally.

Report the numbers you computed and answers to the above questions in the PDF that you turn in. Hand in your Python code separately. Please make sure your variables and functions are aptly named, and your code is clear and readable, with comments where appropriate. We may try to execute your code on a new dataset as a test case, but don't worry too much about making your code robust in terms of error checking and adversarial inputs – this new dataset will look very similar to the one we have provided.

A brief explanation of the data and the variables. The data is simulated, but heavily inspired by a longitudinal study on the health benefits of engaging in the arts – going to the opera, theatre, museum etc.

- `age`: individual's age at beginning of the study
- `income`: individual's income at beginning of the study
- `college_degree`: did the individual hold a college degree at the beginning of the study (1 = yes, 0 = no)
- `opera`: was the individual attending the opera regularly (once a month or more) at the beginning of the study (1 = yes, 0 = no)

- `mortality`: did the individual die during the 5-year period of the study (1 = yes, 0 = no)

The point estimates and 95% confidence intervals (CI), computed over 200 bootstrap iterations, are as follows:

$\text{OR}(\text{Mortality}, \text{Opera}) = 0.33, \text{CI} = (0.28, 0.38)$
$\text{OR}(\text{Mortality}, \text{Opera} \mid \text{Income}) = 0.74, \text{CI} = (0.63, 0.86)$

Based on these results, we cannot say that Mortality $\perp\!\!\!\perp$ Opera or even that Mortality $\perp\!\!\!\perp$ Opera $\mid$ Income. However, this is not to say that opera attendance is responsible for an increased lifespan. We see that controlling for income decreases the significance of this relationship notably. What remains of the dependence of mortality on opera attendance is likely driven by other unobserved confounders. For example, if we additionally control for education, we find

$\text{OR}(\text{Mortality}, \text{Opera} \mid \text{Income}, \text{Degree}) = 0.93, \text{CI} = (0.77, 1.11)$

We note that $\text{OR}(\text{Mortality}, \text{Opera} \mid \text{Income}, \text{Degree}) = 1$ (corresponding to conditional independence of mortality and opera attendance) falls within the 95% confidence interval here. Additionally, pulling in prior knowledge, it seems likely that the causal relationship is not between opera attendance and mortality but instead stems from other factors such as income and educational attainment. We expect these to be causally related to increased lifespan/ opera attendance through known mechanisms such as differential access to health care and different behaviors.