Department of Economics
Yale University
Dr. Guillermo Noguera

ECON 117
Intro to Data Analysis & Econometrics
Spring 2022

Problem Set 2

Due February 14th by 11am. Submit through Canvas.

This problem set should be submitted as a PDF together with the .R (or .Rmd) file used to run the analysis. The PDF should include **all** your work (code, output, hand-written/typed answers) and can be prepared in Word, Latex, or any other electronic format (such as R Markdown). Hand-written math can be included as needed, though this should be done through embedded photos or scanned pages. Handwritten math that cannot be read by the teaching assistants will be marked incorrect. For more details, see Problem Set Submission Guidelines on Canvas.

In case of technical problems when submitting your problem set, please send an email to Vladimir (vladimir.smirnyagin@yale.edu) and Hanxiao (hanxiao.cui@yale.edu) detailing the problems along with your problem set. You should send it before the deadline.

1 Theoretical Questions

Remember to show all your work. If you use any probability, expectation, or variance rules, specify which ones and where they are used.

1.1 Rapid Covid Tests

Beginning this January, US households can order free at-home rapid Covid tests from the government. Rapid tests offer unique benefits by allowing one to get results quickly and conveniently before making travel or social gathering decisions, but they are less sensitive than PCR tests in detecting the virus. One study[1] reports that rapid tests have 78 percent probability of identifying the virus among those who are infected,[2] and have 3 percent probability of reporting a false positive result for those who are not infected.

For the ease of notation, let $T+$ $(T-)$ denote the event of a positive (negative) rapid test result, and let $V+$ $(V-)$ denote the event of having (not having) the virus.

1. Express the findings of this study (i.e., 78 percent and 3 percent) in terms of conditional probabilities.
2. Tim woke up one morning feeling slightly under the weather. He decided to first do a rapid self-test. Suppose the probability that his symptoms are related to a Covid infection is 0.4.
   (a) What is the probability of him having the virus AND getting a positive rapid test?
   (b) What is the probability of him getting a negative result on the rapid test?
   (c) Suppose that he tested negative on the rapid test. How confident should he be that he is not infected with Covid?
3. Since rapid tests are less sensitive in detecting infections, the FDA recommends doing serial testing over several days to improve accuracy. Tim did another rapid test 2 days later. Let $T'+$ $(T'-)$ denote the event of a positive (negative) result on the second test. Assume that the sensitivity of the test (78 percent) is constant over the course of infection.
   (a) Are the events of testing positive on the first test and testing positive on the second test independent? Explain.
   (b) Suppose Tim was indeed infected with Covid. What is the probability of him testing positive at least once in the two serial tests?

[1] Harmon A, Chang C, Salcedo N, et al. Validation of an At-Home Direct Antigen Rapid Test for COVID-19. JAMA Netw Open. 2021;4(8):e2126931. Here is a link to the paper.

[2] Within days 0 to 12 of symptom onset.

1.2 True or False

Please include a short explanation for your answer, even if it is true.

1. For two events A and B, $\mathrm{P}(A \mid B)$ is always greater than or equal to $\mathrm{P}(A)$.
2. If two events A and B are independent, then $Not\ A$ and $Not\ B$ are also independent.
3. The total area under the curve of any probability density function is the same.
4. For a discrete random variable, the value of the probability mass function (PMF) cannot exceed the value of the cumulative distribution function (CDF) evaluated at any point, whereas for a continuous random variable, the value of the probability density function (PDF) can be much larger than the value of the CDF.

1.3 Uniform Distribution

The continuous random variable X is uniformly distributed over the interval $[\alpha, \beta]$ with $\alpha < \beta$.

1. Write down the probability density function of X, for all x.
2. Suppose that $\mathrm{E}(X) = 2$ and $\mathrm{P}(X \leq 3) = \frac{5}{8}$. Find the value of α and the value of β.
3. Draw the CDF of X.

2 The Role of Colleges in Intergenerational Mobility

The growth in inequality in the developed world during the last few decades has increased public interest in the topic of inequality. The literature on this topic robustly finds that family background has a strong impact on a person's ultimate socioeconomic status. As a result, there is intergenerational persistence of inequality: The rich and educated can guarantee better inputs to their children, who then have a better chance of remaining in the upper echelons of society upon adulthood. The term intergenerational mobility is used to refer to the strength of this perpetuation mechanism. If perpetuation of inequality is intense, we say intergeneration mobility is low, and vice-versa.

Economist Raj Chetty and his co-authors are at the forefront of the study of such intergenerational inequality. In a series of papers, they use high-quality data from the IRS on the incomes of millions of American workers to examine what factors affect intergenerational mobility. This Problem Set focuses on their 2017 paper[3] examining the role played by colleges in promoting (or hampering) intergenerational mobility. We will use a variety of data sets made available by Chetty's team on their "Equality of Opportunity" website.

2.1 Income Segregation Across Colleges

Using Table II from this paper and probability rules learned in class, we first examine the degree of segregation by parental income across colleges. This problem does not require you to use R and can be completed based on statistics from Table II in the paper.

1. Calculate:

   (a) What is P(bottom 60% $\cap$ Not bottom 20% |Ivy Plus), i.e., the share of Ivy-Plus students from families between the 20th and 60th percentiles of the income distribution?

   (b) What is P(bottom 60% $\cap$ Not bottom 20% |Nonsel. 4-yr public), i.e., the share of nonselective 4-year public college students from families between the 20th and 60th percentiles of the income distribution?

2. How many times more likely are children from the top 1% families to be in Ivy-plus colleges than children from the bottom 20% families? To answer this question, let's take the following steps:

   (a) Find the expressions for P(Ivy Plus|top 1%) and P(Ivy Plus|bottom 20%) using Bayes' Rule. No need to plug in the numbers yet.

   (b) What is the marginal probability of being from the top 1% distribution? What is the marginal probability of being from the bottom 20% distribution? (Hint: these numbers are not in the table, but should be straightforward to answer.)

[3] Chetty, R., Friedman, J. N., Saez, E., Turner, N., & Yagan, D. (2017). Mobility Report Cards: The Role of Colleges in Intergenerational Mobility. National Bureau of Economic Research. Here is a link to the paper.

(c) Now take the ratio of P(Ivy Plus|top 1%) and P(Ivy Plus|bottom 20%) using your expressions in (a). You should have all the numbers needed to calculate this ratio. Are children from the top 1% families more likely to be in Ivy-plus colleges as compared with children from the bottom 20% families?

3. From Table I of the paper, we know that there are $10,757,269$ children in total in the cohorts they study (including those who never attended college). Using this number, plus numbers in column (14) of Table II, find the marginal probabilities of (i) attending ANY college, (ii) being at an Ivy-Plus institution.

4. Use your calculations in the previous question and numbers from Table II to calculate P(Any college|Top 40%) and P(Ivy plus|Top 40%). Note the second probability is **not** conditional on college enrollment.

The following questions will require the use of R. You must provide the code for your solutions and the corresponding output.

2.2 College Mobility Rates

1. Load data "`college_mobility_rates.csv`". In this data set, each observation is a college. The data includes students' family income, the income they eventually earn in adulthood, as well as the geographic location and selectivity tiers of the college.

   (a) How many colleges are there in the data set?

   (b) How many variables are there in the data? Display the names of the variables.

2. For each college, the mobility rate is defined by the following formula:

$$\text{Mobility rate} = \text{Access rate} \times \text{Success rate}$$

   where the *access rate* is the share of students coming from the bottom 20% of the parental income distribution, and the *success rate* is the share of those students from the bottom 20% families who make it to the top 20% of the earnings distribution themselves.

   (a) Variable `par_q1` reports the *access rate* (i.e., the share of students coming from the lowest income quintile). Sort colleges according to `par_q1`, then find and report colleges with the 10 lowest and 10 highest access rates.[4]

   (b) Variable `kq5_cond_parq1` reports the *success rate* (i.e., the share of students making it to the highest quintile conditional on coming from the lowest quintile). What is Yale's success rate? Compare this to the success rates of Cornell University, Princeton University, and Quinnipiac University.

[4] To sort a data frame in R, use the `order()` function. By default, sorting is ascending. For instance, `sorted.chetty <-chetty[order(par_q1),]` creates a new sorted data frame `sorted.chetty` sorted by `par_q1`. To obtain the first n observations of a data frame, use `head(data frame, n)`. To obtain the last n observations of a data frame, use `tail(data frame, n)`.

(c) Make a scatter plot of `kq5_cond_parq1` against `par_q1`.[5] Make sure the plot has proper title and axis labels. What does the plot suggest about the relationship between *access rate* and *success rate*? Do more selective colleges have higher value-added for low-income students? Are there colleges that do reasonably well on both metrics?

(d) Create a new variable named `mr_kq5_pq1` that represents the *mobility rate* calculated in accordance with the formula above.[6] What is the lowest value, the highest value, the mean, and the median of mobility rates among colleges?

(e) Sort colleges with respect to the *mobility rate*, then find and report colleges with 10 lowest and 10 highest mobility rates.

(f) Look for Yale in the rankings of the *mobility rate*. How does Yale's rank compare to that of Cornell University, Princeton University, and Quinnipiac University?[7]

2.3 College Characteristics and Mobility Rates

1. In the data, variable `type` takes on the value of 1 if the college is public, 2 if private non-profit, and 3 if for-profit. Use `table()` function to find out how many colleges in the data are public, private non-profit, and for-profit.

2. Calculate the average mobility rate conditional on being a public college, a private non-profit college, and a for-profit college, respectively.

3. In the data, variable `tier_name` represents the selectivity tiers of colleges. Find out the selectivity tiers of colleges with the 10 highest mobility rates.

4. Based on your analysis above, what types of schools do you think have the greatest contribution to upward mobility? (Max 3 sentences.)

[5] To make a scatterplot for two variables `x` and `y`, one can use the following command:

```
plot(x, y, xlab = "x-axis title", ylab = "y-axis title", main = "title")
```

[6] Make sure the new variable is added as an additional column to the existing data frame rather than a separate object in the environment.

[7] The `rank()` function can be used to get the rank of colleges according to the mobility rate in ascending order. For instance, `chetty$mr_rk=rank(chetty$mr_kq5_pq1)` ranks colleges according to the mobility rate and assign it to a new variable `mr_rk`.