14.310x: Data Analysis for Social Scientists - Homework 8

Welcome to your eight homework assignment! You will have about one week to work through the assignment. We encourage you to get an early start, particularly if you still feel you need more experience using R. We have provided this PDF copy of the assignment so that you can print and work through the assignment offline. You can also go online directly to complete the assignment. If you choose to work on the assignment using this PDF, please go back to the online platform to submit your answers based on the output produced. Some of the questions we are asking are not easily solvable using math so we recommend you to use your R knowledge and the content of previous homework assignments to find numeric solutions.

Good luck :)!

Question 1: Inference for a Randomized Experiment

The following problems are based on the paper:

> Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." American Economic Review, 102(4): 1241-78.

First, read the abstract of the paper in the following link `http://economics.mit.edu/files/5582`. You can refer back to the paper as necessary.

You will complete the following exercise for the variable `open`, the proportion of times the school was open during a random visit.

Note: The dataset used to generate Lecture the 15 slides relating to this paper is slightly different than the dataset we have provided, so do not be alarmed if your answers are slightly different!

In order to complete this exercise we are providing you with the code *problem1.R*. The code has some missing parts that you have to fill in order to run it. The dataset that you will need its *teachers_final.csv*

1. First, consider the case where we have 8 schools. Our aim is to calculate the Fisher's exact p-value. Under the assumption that we will have the same number of treated and control units, how many potential treatment assignments across these 8 units are possible?

   (a) 50

   (b) 60

   (c) 70

   (d) 80

   Suppose that after the treatment has been assigned and the experiment has been carried on, the researcher has the following data. The variable **open** corresponds to the fraction of days that the school was opened when random visits were made.

treatment	open
0	0.462
1	0.731
0	0.571
0	0.923
0	0.333
1	0.750
1	0.893
1	0.692

2. Assume that we wish to calculate the absolute difference in means by treatment status as our statistic. For this observed data, what would be the value of our statistic?

   Now use the R code we have provided and fill in the missing information. The code calculates the value of this statistic (the difference in means for the treatment vs control group) for all the potential treatment assignments.

3. How many of these statistics are larger than the one from our observed data?

   (a) 11
   (b) 16
   (c) 21
   (d) 26
   (e) 31
   (f) 36

4. What would be the Fisher's Exact p-value in this case?

5. Now load the data set in R. Suppose we want to test the sharp null hypothesis in this data, with 49 schools treated. Is it the case that the number of possible assignments would be too large to test this sharp null hypothesis (at least with your laptop and less than an hour of computing time)?

   (a) Yes
   (b) No

6. A solution to this problem with a large number of observations is to simulate different random assignments and calculate the proportion of simulations in which the statistic exceeds the value of the observed data. We have provided you with a code that performs this exercise on the data `teachers_final.csv` with 100,000 simulations. If you run this code, is the approximate Fisher's p-value similar to the one we got with our 8 schools example?

   (a) No
   (b) Yes

7. Since we are working in a much large sample, we can now consider Neyman's methods of inference. What is the Average Treatment Effect (ATE) on the observed data set?

8. What is the upper bound of the standard error of this point estimate using Neyman's method?

9. What is the t-statistic if we want to test the null hypothesis the ATE is equal to zero?

10. Is the associated p-value to this test similar to the one we found for the sharp null hypothesis in question 6?

    (a) Yes

    (b) No

11. What is the 95% confidence interval of this test?

    (a) It is given by $(0.127, 0.267)$

    (b) It is given by $(0.147, 0.247)$

    (c) It is given by $(0.157, 0.237)$

    (d) It is given by $(0.137, 0.257)$

    Now, imagine that you are considering a similar randomized experiment as the Duflo/Hanna/Ryan camera experiment, except you plan to give teachers lower incentives - half the monetary amount as in the Duflo/Hanna/Ryan experiment.

12. If you think that the relationship between incentives and the variable open is linear, what would be the expected ATE of this new intervention?

13. Assume that this value is the minimum ATE such that the intervention is cost-effective, what is the sample size required to have a power of at least 90%?

    - with a significance level of 5%
    - an equal number of treated and control units
    - σ^2 is the average of the variance of the control and the treatment group in the existing data

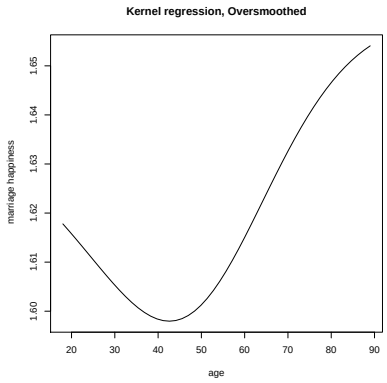
    (a) 100

    (b) 110

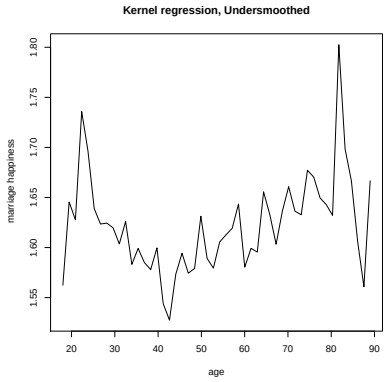
    (c) 120

    (d) 130

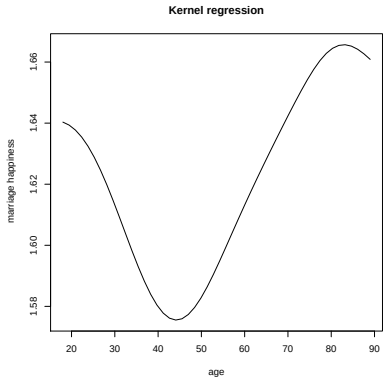
    Now we are going to consider non parametric regressions. The following plots show three different non-parametric regressions that relates the level of happiness in a marriage with age (where 2 corresponds to "very happy", 1 "pretty happy", and 0 "not too happy").



(a)



(b)



(c)

14. Rank the three plots from the one with the narrower to the wider bandwidth

(a) a, b, c
(b) a, c, b
(c) b, c, a
(d) b, c, a
(e) c, a, b
(f) c, b, a