**14.310x: Data Analysis for Social Scientists - Homework 7**

Welcome to your seventh 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 homeworks to find numeric solutions.

Good luck :)!

## Question 1: Inference for a Randomized Experiment

This problem is based on the following 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.

In this experiment, the researchers set out to test whether providing teachers with cameras to take photos to prove their attendance, could be effective in reducing teacher absenteeism. 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.

Let's start by thinking through how Fisher's ideas can be applied to evaluate this program in this context.

1. First, suppose we only 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

   After putting together a list of the schools, they randomly assigned four of the schools to be "treated" and the remaining half to continue with buisiness as usual. Next, after implementing the program, they collected outcome data. One of the variables they collected is the fraction of days that the school was found to be open when random visits were made by a member of the research team.

Here is the data for 8 schools from the data in this study (found in `teachers_final.csv`):

| **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 |

Assume that we define as our statistic the absolute difference in means by treatment status. To help you compute the test statistic for the observed data, we have provided you with the following R code to load in this table and generate different permutations, although it is missing some parts that you will need to fill in. We make use of the package `perm`, specifically the function `ChooseMatrix`, look it up and look up it's arguments and read the code to make sure you understand what it is doing.

```
perms <- chooseMatrix(?,?)
A <- matrix(c(0.462, 0.731, 0.571, 0.923, 0.333, 0.750, 0.893, 0.692), nrow=8, ncol=1, byrow=TRUE)
treatment_avg <- (1/4)*perms%*%A
control_avg <- (1/4)*(1-perms)%*%A
test_statistic <- abs(treatment_avg-control_avg)
rownumber <- apply(apply(perms, 1,
function(x) (x == c(0, 1, 0, 0, 0, 1, 1, 1))),
2, sum)
```

2. For this observed data, what would be the value of our statistic? *(Please round your answer to two decimal places)* We recommend you solve this problem algebraically and using R to check your answer.

3. Now, use your results to compute 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? *(Please round your answer to two decimal places)*

   *How should we interpret this? In general, we want to know whether the camera intervention had an effect, and whether the treated schools were open more frequently than the control schools. The mean* *of the treatment group is higher than the mean of the control group, indicating the teacher camera intervention may have indeed had an effect. However, under the sharp null hypothesis that there is no treatment effect in any of the schools in our sample, we have that if we randomly allocate 4 units to treatment, 23% of the time, the treatment and control groups would have looked at least as different as what we observed here, or even more different.*

Now load the data set `teachers_final.csv` in R. If 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 do the problem (at least with your laptop and less than an hour of computing time)?

(a) Yes

(b) No

5. 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 1,000 simulations, read the code, make sure you understand it and fill in the missing blanks:

```r
simul_stat <- as.vector(NULL)
schools <- read.csv('teachers_final.csv')
set.seed(1001)
for(i in 1:1000) {
print(i)
schools$rand <- runif(100,min=0,max=1)
schools$treatment_rand <- as.numeric(rank(schools$rand)<=?)
schools$control_rand = 1-schools$treatment_rand
simul_stat <-append(simul_stat,
    sum(schools$treatment_rand*schools$?)/sum(schools$treatment_rand)
    - sum(schools$control_rand*schools$open)/sum(schools$?))
}

schools$control = 1-schools$treatment
actual_stat <- sum(schools$treatment*schools$open)/sum(schools$treatment) - sum(schools$control*schools$open)/sum(schools$control)
sum(abs(simul_stat) >= actual_stat)/NROW(simul_stat)
```

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

6. 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? *(Please round your answer to three decimal places)*

$$\overline{\mathbf{Y}}_T^{obs} - \overline{\mathbf{Y}}_C^{obs}$$

$$\overline{\mathbf{Y}}_T^{obs} - \overline{\mathbf{Y}}_C^{obs} = 0.1969$$

7. What is the upper bound of the standard error of this point estimate using Neyman's method? *(Please round your answer to three decimal places)*

8. What is the t-statistic if we want to test the null hypothesis the ATE is equal to zero? *(Please round your answer to two decimal places)*

9. Is the associated p-value to this test similar to the one we found for the sharp null hypothesis in question 5?

   (a) Yes

   (b) No

10. 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.

11. If you think that the relationship between incentives and the variable open is linear, what would be the expected ATE of this new intervention? *(Please round your answer to three decimal places)*

12. 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
    - $\sigma^2$ is the average of the variance of the control and the treatment group in the existing data

    (a) 100

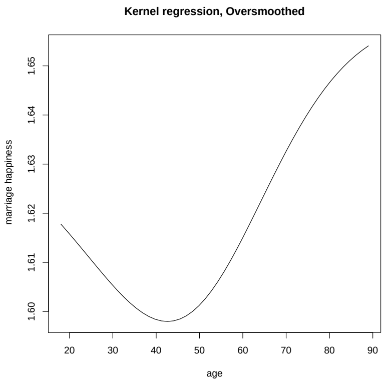
    (b) 110

    (c) 120

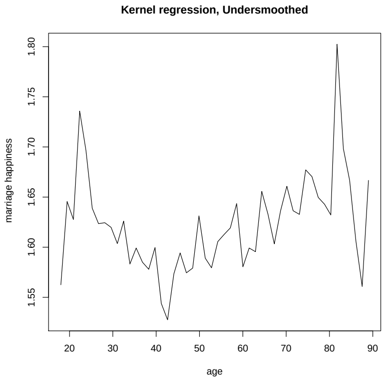
    (d) 130

## Question 2: Nonparametric Regressions

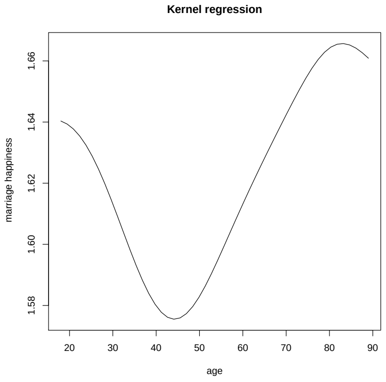
Now we are going to consider non parametric regressions. The following plots show three different non-parametric regressions that relates the level of happiness in a marriage with age (where 2 corresponds to "very happy", 1 "pretty happy", and 0 "not too happy").



(a)



(b)



(c)

13. 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

Going back to the data from `teachers_incentives.csv`, we are now going to focus on two variables: `pctpostwritten` , which denotes the mean student test scores after the intervention and `open`. We want to see what the relationship between the fraction of days the school is open and student achievment. To this end, use the kernel regression code from lecture, to plot the kernel regression between these two variables using the R package `np` which Prof.Duflo illustrated in lecture.

14. Use your code to generate plots for the following bandwiths, which of them seems most appropriate given the data? (select one)

    (a) 0.04
    (b) 0.001
    (c) 1
    (d) 20

### Question 3: First Order Stochastic Dominance

15. Suppose we are interested in testing whether or not the distribution of the share of days a school is found to be `open` in the treatment group is statistically distinguishable from the distribution for the control group. Which of the following would be most useful for this purposes? (select one)

    (a) a kernel regression
    (b) a histogram of the variable by group
    (c) a Kolmogrov-Smirnov test
    (d) a joint density plot
    (e) None of the above

16. Assume the following notation:

    - Let $i \in T, C$ index the cohort school $i$ is assigned.
    - $m_i$ denotes the sample mean for group $i$.
    - $\mu_i$ denotes the population mean.
    - $F_i$ denotes the CDF for group $i$.

    The outcome of interest is still the variable `open`.

    For each hypothesis test below, indicate which of the following methods is most useful for testing that hypotheses, by entering (N- for using Neyman's method of inference, F- for Fisher's exact test, and K- for the KS test):

    (a) $H_0 : \mu_T - \mu_C = 0$ vs. $H_1 : \mu_T - \mu_C \neq 0$
    (b) $H_0 : \mu_T - \mu_C > 0$ vs. $H_1 : \mu_T - \mu_C \leq 0$
    (c) $H_0 : m_T - m_C < 0$ vs. $H_1 : m_T - m_C \geq 0$
    (d) $H_0 : F_T = F_C$ vs. $H_1 : F_T \neq F_C$
    (e) $H_0 : F_T > G\ N(0, 1)$ vs. $H_1 : F_T \leq G$

17. Generate a plot of the CDFs for each cohort to see those results visually. Does the distribution of scores in the treatment group FOSD that of the control group?

    (a) Yes

    (b) No