14.310x Data Analysis for Social Scientists
Causality, Analyzing Random Experiments, and Nonparametric Regression

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. 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!

Please find a glossary of R terms that will be useful for this week's homework here.

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. [View on WorldCat].

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 using the link above. You can refer back to the paper as necessary.

Note: The dataset used to generate the Lecture 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. The code has some missing parts that you have to fill in order to run it. The dataset that you will need is teachers_final.csv

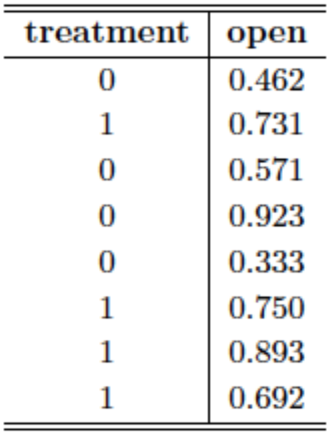
Let's start thinking through how Fisher's ideas can be applied to evaluate this program in this context.

Question 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?

- 50
- 60
- 70
- 80

Suppose that after the treatment has been assigned and the experiment has been carried out, 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.

For Questions 2-4, we will look at these 8 schools 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 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`. Be sure to look up the documentation to make sure you understand what it is doing.

Question 2
For this observed data, what would be the value of our statistic?

We recommend you compute this test statistic on your own and then check your answer using the code provided. *Please round your answer to two decimal spaces.*

Question 3
According to your results, among the test statistics computed for all treatment assignments, how many are larger than the observed test statistic?

- 11
- 16
- 21
- 26
- 31
- 36

Question 4

What would be the Fisher's Exact p-value in this case? *Please round your answer to two decimal places.*

Question 5

Now load the data set teachers_final.csv in R and name it `schools`. This is done in line 31 of the provided R code.

With 49 schools treated, what are the number of possible assignments in this case?

- $\binom{49}{8}$
- $\binom{100}{51}$
- $\binom{1001}{49}$
- $\binom{100}{49}$

Question 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 the code that performs this exercise on the data `teachers_final.csv` with 100 simulations. However, we have replaced line 46 with blanks for you to fill in (XXXX).

Fill in line 46 with the correct code. What is the result of that line?

The figure you calculated in Question 6 represents an approximation to Fisher's p-value. You can explore with changing the number of simulations and the number of schools to see if they change the p-value.

Question 7

Since we are working in a very large sample, we can now consider Neyman's methods of inference. What is the Average Treatment Effect (ATE) on the observed data set? You will need to use R to compute this answer. *Please round your answer to three decimal places.*

Question 8

What is the upper bound of the standard error of this point estimate using Neyman's method? (Hint: Use the conservative estimator of sampling standard deviation, $\sqrt{\widehat{\mathbb{V}}_{reymar}}$, as your upper bound.) *Please round your answer to three decimal places.*

Question 9

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

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

- Yes
- No

Question 11
The 95% confidence interval is given by (A, B). What are the values of A and B? *Please round to three decimal places. For instance, if your answer is .6789, please round to .679.*

Now, imagine that you are considering a randomized experiment similar to the camera experiment. The exception is that you plan to give teachers lower incentives: half the monetary amount that was given in the previous experiment.

Question 12
Imagine 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 the third decimal place, i.e. if it is 0.3414, please round to 0.341.*

Question 13
Assume that the value from Question 12 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 the following properties?

- 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 treatment group in the existing data

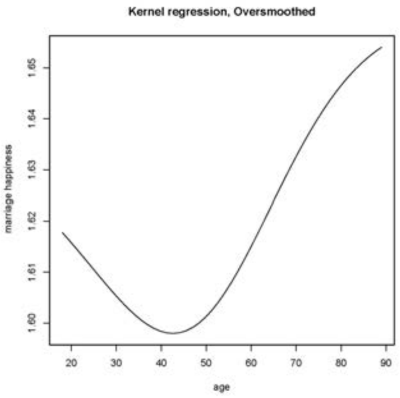
Hint: Recall that the formula for sample size is:

$$N = \frac{(\Phi^{-1}(1-\beta) + \Phi^{-1}\left(1-\frac{\alpha}{2}\right))^2}{\frac{\tau^2}{\sigma^2}\gamma(1-\gamma)}$$

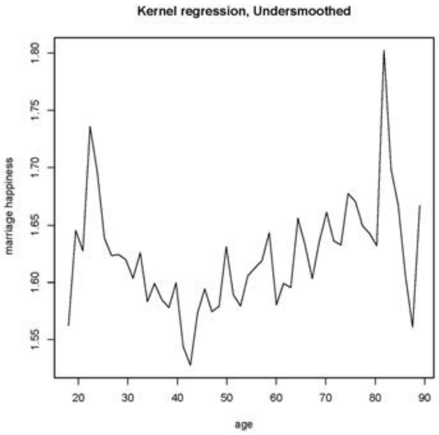
Where β is the operating characteristic and $1-\beta$ is the desired power.

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

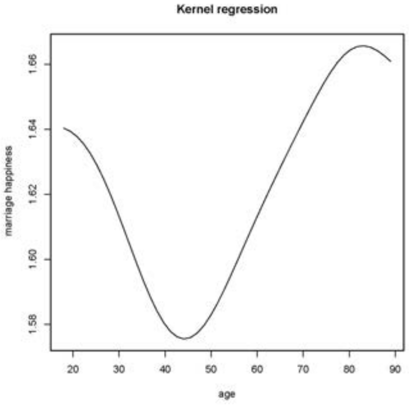
Plot A:



Plot B:



Plot C:



Question 14

Rank the three plots from the one with the narrowest to the widest bandwidth.

- a, c, b
- b, a, c

- c, b, a
- b, c, a
- c, a, b
- a, b, c

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
- `open`

We want to see what the relationship between the fraction of days the school is open and student achievement. Use the code below (from lecture) to plot the kernel regression between these two variables using the R package np:

```
attach(schools)
plot <- npreg(xdat=XXX, ydat=XXX, bws=XXX,
bandwidth.compute=FALSE)
plot(plot)
```

Question 15

Use your code to generate plots for the following bandwidths. Which of them seems most appropriate given the data?

- 0.001
- 0.04
- 1
- 20

Question 16

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 that of the control group. Which of the following would be most useful for these purposes?

- Joint density plot
- Histogram of the variable by group
- Kolmogrov-Smirnov test
- Kernel regression
- None of the above

Question 17

Let $i \in T, C$ index the cohort school i assigned. m_i denotes the sample mean of a variable (e.g. student scores) for group i, μ_i denotes the population mean of the variable, and F_i denotes the CDF for group i

For each hypothesis test below, indicate which of the following methods is most useful for testing that hypothesis. **Enter 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. $H_0: F_T > G\ vs. H_1: F_T \leq G\ where\ G \sim N(0,1)$

Question 18

Use the R command `stat_ecdf()` to generate a plot of the CDFs for each cohort to see those results visually. Does the distribution of open in the treatment group FOSD that of the control group?

- Yes
- No

Note: the command to run a KS test in R is `ks.test()`. Look up the help file for this function and use it to test whether the distribution of test scores (`pctpostwritten`) in the treatment group first order stochastically dominates the distribution of test scores in the control group. Though the test fails when you have ties (so we are unable to use it in the case of test scores), you may find it useful in other applications.

14.310x Data Analysis for Social Scientists
Spring 2023