SPI 507c, Fall 2022 Mark Watson
(This Version: 12/28/2022)

PROJECT 1

*Assignments will be graded on a (*0, ✓-, ✓) *scale. (The preceptor may occasionally assign a* ✓+ *grade for an exceptionally good Project.)* ***Assignments are due by 11:59PM on Friday December 16, 2022.*** *Late assignments will not be accepted.*

Your grade will be based on the clarity of your analysis and the clarity of your writing.

You are working for the Healthy Lives, a nonprofit organization advocating for safe and healthy lifestyles. A memo from the director of the organization is attached.

Please respond to his request with a memo that does not exceed five pages (single-spaced), and may be shorter. Carefully choose tables and graphs to communicate your results as efficiently as possible. You should assume the director knows elementary statistics and regression analysis – indeed he received an A in 507c just ten years ago, and has been reading statistical reports since leaving the School.

You analysis should be based on the data contained in the file **MLDA_507.dta**. These data are described below.

Rules on collaboration (from the syllabus): "You may work in teams of four or fewer students. For each project, you will be required to turn in a STATA *do* file and *log* file (or the R equivalent), and a write-up of your results. You must include the names of all group members at the top of the do and log files, and all of you may turn in the same do and log files. <u>But you must complete the write-up independently, and each student must turn in their own write-up.</u>"

SPI 507c, Fall 2022 Mark Watson
(This Version: 12/28/2022)

To: _____
From: Christopher J. Maghi, Director
Date: December 1, 2022

I need to prepare for next month's conference on the minimum legal drinking age (MLDA), and I'm really at a loss to know what our position should be. Indeed, I don't even know the basic facts.

The MLDA is 21 in the U.S., but there seems to be plenty of drinking by people under the MLDA. Does the MLDA make *any* difference? Does crossing the 21-year-old threshold really make it more likely that someone will drink?

I understand that we have access to some data from the National Health Interview Survey that contains self-reported drinking by young men and women, some younger and some older than 21. (We all the know the problems with self-reports of drinking, but the survey data seems like a sensible place to start.)

Can you look at these data to see if there is a "jump" in the likelihood of drinking at age 21? I'm being a bit vague here, but I guess that the likelihood of drinking is increasing as age increases from say 18 years of age to say just shy of 21 years of age, but I'm not sure what the function looks like. And from 21 onward the function may be increasing, or may be flat, or? The question I'm interested in is whether the function jumps up by a large amount at age 21? At age 21, does the likelihood of drinking jump up by 5%? By 10%? More than 10%? Less than 5%? I'm hoping that estimates of the size of this jump will help me think about what would happen if we decreased the MLDA down to 20, or increased it above 21.

A few other things:

(1) Drinking "some alcohol" is one thing, but drinking "regularly" or "heavily" are quite other matters. Can you say something about the jump at age 21 for different levels of alcohol use?

(2) Are there important differences in the effect MLDA on the likelihood of drinking for specific groups (say men vs. women, or blacks vs. Hispanics vs. whites, or …)?

(3) And of course, you should do what you can to control for other factors.

Thanks, and I look forward to seeing what you come up with.

SPI 507c, Fall 2022 Mark Watson
(This Version: 12/28/2022)

Documentation for MLDA_507

The data file contains data on over 61,000 individuals drawn from the National Health Interview Sample Audit Files from 1997-2007. The data were collected by Professor Carlos Dobkin of UC Santa Cruz, who has done important research on the minimum legal drinking age.

Variable Name	Description
Age and Drinking Variables	
days_21	Days from 21st birthday. (positive values indicate age ≥ 21; negative values indicate age < 21)
drinks alcohol	Binary variable = 1 if person reports that he/she drinks alcohol, 0 otherwise
perc_days_drink	Percent of days on which he/she reports drinking alcohol
Other Variables	
HS Diploma	Binary variable = 1 if person has a high school diploma, 0 otherwise
Hispanic	Binary variable = 1 if person is Hispanic, 0 otherwise
Black	Binary variable = 1 if person is Black, 0 otherwise
Employed	Binary variable = 1 if person is employed, 0 otherwise
Student	Binary variable = 1 if person is a student, 0 otherwise
Male	Binary variable = 1 if person is a male, 0 otherwise
Married	Binary variable = 1 if person is married, 0 otherwise

Note: I eliminated all observations with $|\text{days\_21}| \leq 21$ to eliminate alcohol consumption associated with someone's 21st birthday.

SPI 507c, Fall 2022 Mark Watson
(This Version: 12/28/2022)

Here are some exercises to help organize your thoughts:

Below is an outline of a set of exercises that should help you respond to the Director's request. My suggestion is that you carry out these exercises, see what you learn, and then draft your memo to the Director.

Exercises:

1. Construct a table of summary statistics for the variables in the data set. Use these summary statistics to think about the general features of the data set. Are there any important outliers? If so, how should you handle them?

2. Estimate the probability of drinking for those younger than 21. Estimate the probability for those 21 or older. Estimate the difference in these probabilities. Is the difference large? Is there statistically significant evidence that the probability of drinking is lower for those under the MLDA than for those over the MLDA?

3. Let p denote the probability of drinking, and $p(age)$ be a function that shows how this probability depends on *age*. Estimate the function $p(age)$ using OLS regressions of *drinks_alcohol* on *days_21* and appropriately chosen functions of *days_21*. Choose regressions so you can think about the following questions:

(a) Is the $p(age)$ function increasing, decreasing, or "flat" for $age \leq 21$ years (that is for people below the MLDA)? Is the function linear or non-linear. If nonlinear, what does it look like? Using your preferred specification, estimate the probability of drinking for someone 20.5 years old. For someone 20 years old. For someone 19.5 years old. Compute the standard error for each of these estimates. Also compute the change in this probability (and its standard error) as a person ages from 19.5 to 20 and from 20 to 20.5.

(b) Repeat (a) but for people over age 21.

(c) Does the data suggest that there is a discontinuity -- a "jump" -- in $p(age)$ at the MLDA? If so, how large is this jump? How precisely is it estimated.

4. The dataset contains a small number of demographic controls. Do the conclusions you reached in (3) change (importantly) after accounting for these controls? Do the conclusions differ for different demographic groups?

5. Questions (3) and (4) concerned the binary variable *drinks_alcohol*. Does the data on *perc_days_drink* allow you to refine your conclusions.

6. Discuss the limitation of these data and internal validity of your results for answering questions about the *causal* effect of the MLDA on drinking. Would you feel comfortable using your regression to study the effects of a policy that lowered the MLDA to 20? Why or why not?