Name _____

Kuehn
Economics 339
Fall 2022

Final Paper Guidelines

DUE DATE: Friday, December 2

The basic idea of this project is to write a short paper investigating the effect of alcohol consumption on labor market outcomes. You will do this by applying the econometric techniques used in this class to the data set **alcohol.xlsx** found on the class website.

DATA

The data set *alcohol.xlsx*, comes from the National Longitudinal Survey of Youth (NLSY) and includes information on labor market outcomes, alcohol consumption, and assorted demographics for individuals in each of 2 years – 1989 and 1994. The data are restricted to young adults who are between the ages of 24 and 32 in 1989 (and hence 29-37 in 1994). Each individual has a unique identifier (variable named *id*) and the year is indicated by the variable named *year.*

The following labor market variables are available:
wgsal: Total wage and salary income in the past calendar year, in dollars
hrswrk: Total number of hours worked in the past calendar year
wkswrk: Total number of weeks worked in the past calendar year
wksue: Total number of weeks spent unemployed in the past calendar year
wksolf: Total number of weeks spent out of the labor force in the past calendar year
empst: A categorical variable indicating the individual's current employment status where:

1 = Employed
2 = Unemployed
3 = Out of Labor Force
4 = In Active Armed Forces

numjob: Total number of jobs the individual has ever held in their lifetime

The following alcohol consumption variables are available:
drinkev: A dummy variable = 1 if the individual has ever had a drink, 0 otherwise
drnkmo: Dummy variable =1 if the individual has had a drink in the last month, 0 otherwise
drnk6m: A categorical variable indicating the number of times in the past month the individual has had 6 or more drinks in one sitting. It is defined as:

0 = Never
1 = Once
2 = 2 or 3 Times
3 = 4 or 5 Times
4 = 6 or 7 Times
5 = 8 or 9 Times
6 = 10 or More Times

days: The number of days in the last month that the individual has had at least 1 drink
perday: The average number of drinks per day on a day when the individual drinks
gtint: A categorical variable that answers the question of whether the individual has ever drunk more than intended. It is defined as follows

0 = Don't Drink
1 = Happened 3+ Times in Past Year
2 = Happened 2 Times in Past Year
3 = Happened 1 Time in Past Year
4 = Happened in Lifetime Other than Past Year
5 = Never Happened

The following demographic variables (which may or may not be useful) are available:
age: Age of the individual in years
sex: A categorical variable = 1 if the individual is a man and =2 if a woman
race: A categorical variable = 1 if the individual is Hispanic, =2 if the individual is Black and =3 otherwise
south14: Dummy variable = 1 if the individual lived in the south when they were 14 years ol
wdad14: A dummy variable =1 if the individual lived with their father when they were 14
wmom14: Dummy variable = 1 if the individual lived with their mother when they were 14
dadwork: A dummy variable = 1 if the individual's father worked when they were 14. This is set to 0 if they didn't know, which often happens if they didn't live with dad, so this variable should always be used along with wdad14.
momwork: A dummy variable = 1 if the individual's mother worked when they were 14. This is set to 0 if they didn't know, which often happens if they didn't live with mom, so this variable should always be used along with wmom14
dadhgc: The number of years of education the individual's father has. This is set to 0 if they didn't know, which often happens if they didn't live with dad, so this variable should always be used along with wdad14
momhgc: The number of years of education the individual's mother has. This is set to 0 if they didn't know, which often happens if they didn't live with mom, so this variable should always be used along with wmom14
numsib: The number of siblings the individual has
hvsib: A dummy variable =1 if the individual has a sibling in the data set
sibid1: the value of the variable id for the individual's sibling in the data set. This is missing if there is no sibling in the data set.
religkid: A categorical variable reporting what religion the individual was at age 14. It is defined as follows:

0 = None, No Religion
1 = Protestant, unspecified
2 = Baptist
3 = Episcopalian
4 = Lutheran
5 = Methodist
6 = Presbyterian
7 = Roman Catholic
8 = Jewish
9 = Other

relignow: A categorical variable reporting what religion the individual is now. It is defined the same as religkid
afqtrev: Percentile in which the individual scored on an intelligence test given in 1979

height: The individual's height, measured in inches
weight: The individual's weight, measured in pounds
health: A dummy variable = 1 if the individual has a health problem that limits the amount or kind of work that can be done
higrad: The number of years of education the individual has completed
numkid: The number of children the individual has
urbrur: A dummy variable = 1 if the individual lives in an urban error
*famsz: T*he number of people in the individual's family (i.e. self, plus spouse, plus dependent children)
faminc: Net income for the family in the past year, measured in dollars
povst: A dummy variable =1 if the individual's family was below the poverty line last year
region: A categorical variable for the region the individual lives in. It is defined as follows:

1 = Northeast
2 = North Central
3 = South
4 = West

urate: The unemployment rate for the local labor market of the individual
marst: A categorical variable for the individual's marital status. It is defined as follows:

1 = never married
2 = married with a spouse present
3 = other

It is important to note that not all of the variables are in the most appropriate form for the regression analysis. You will definitely have to create some new variables. Additionally, not all the available variables will make sense in the model. You will have to make decisions about what to include, keeping in mind the *ceteris paribus* interpretation of multiple regression analysis.

STEPS TO COMPLETE PROJECT

1) CHOOSE YOUR REGRESSION MODEL

While the basic question that you are asking is what is the effect of alcohol consumption on labor market outcomes, the way in which you answer that question can vary. You first need to think about what labor market outcome interests you the most. Possibilities include things such as wages, earnings, weeks worked, unemployment status, etc. Similarly you need to decide what kind of alcohol use to investigate. Possible choices include any consumption, any heavy drinking, amount consumed, frequency of consumption, etc. You will also need to decide what else you think should be controlled for to obtain an appropriate causal effect. There is no one "right answer" for what dependent variable or key explanatory variable to choose. What is important is to think about what question the model you choose is answering, and to properly interpret your results in that context. You'll want to consider whether it makes sense to consider your model as having a causal interpretation.

Before running regressions, you want to put some thought into your model. A good way to start is to come up with a conceptual model. That is, think about what types of things you think affect your labor outcome measure. Then look and see how each concept can be captured in the data. For example suppose you were asking the question, "What is the effect of smoking behavior on the wages of full-time, full-year workers in 1984 and 1991?" Then you might choose log hourly wage as the dependent variable, and the key explanatory variable is whether the individual is a daily smoker or instead a nonsmoker or just an occasional smoker. Additional things that might affect the wage include education, work experience, ability, geography, race, sex, current family status, and family background. These concepts are captured using variables for years of education, years of work experience, score on an intelligence test, dummies for living in an urban area, in the south, being nonwhite, being female, being married, and the number of children.

Summary:

a) Choose your dependent variable
b) Choose your key explanatory (or key independent) variable
c) Choose additional variables to add to the regression

2) ESTIMATE YOUR REGRESSION MODEL

After coming up with a conceptual model, you need to estimate the model using R (or some other programming language) and then make inference on the estimates. This includes testing whether or not the regression coefficients are statistically different from zero (i.e. you can reject the null hypothesis that the regression coefficient is equal to zero), and constructing 95% confidence intervals for your results.

3) INTERPRET YOUR RESULTS

Interpret what your results mean for answering your question of the effect of alcohol consumption on labor market outcomes.

4) WRITE PAPER

The written paper that you turn in should be between 3-4 pages (not including up to 2 pages of tables and/or graphs). The project must be word processed and double-spaced using 12 point font and 1 inch margins on the top/bottom/left/right. The paper should include the following:

a) Introduction

Begin with a brief introduction that briefly motivates and describes the issue that you are studying and briefly summarize the main empirical findings. The introduction should be written in simple non-technical language, with statistical and economic jargon kept to a minimum. A reader who knows nothing about econometrics should be able to read the introduction and understand the general issues and findings of your paper.

b) Description of the Data

In this section you should briefly describe your data by providing summary statistics of the key variables of interest. This can include a table of summary statistics (such as the mean, median, standard deviation, min, or max), or a graph of a statistical relationship between the variables (e.g. a scatterplot).

c) Empirical Approach

In this section you should outline what question you are asking and how you intend to answer than question. You should outline the model you are planning to estimate by letting the reader know your dependent variable, your key independent variable, and the control variables you will use in estimation. You should also briefly discuss why you chose these variables to answer your question.

d) Estimation Results how do i implement

This section should contain a table of your results which includes:

- estimated coefficients
- standard errors of the coefficients
- whether the coefficients are statistically significant or not
- 95% confidence interval for the coefficients
- number of observations where from?
- R-squared of the regression

all in same table?

This section should also include a brief discussion of your results and an interpretation of the results in your own words. This should include how you believe the results of your regression answer the question you initially posed about the effects of alcohol on labor market outcomes.

e) Conclusion

Briefly (no more than a few sentences) conclude by summarizing the question you addressed in the paper and the answer you arrived at from your regression analysis.