Course Project Information
ECN 140—Spring 2016
Due in class on May 26 (Thursday, Week 9)

Updates (5/11/2016)

Maximum length: 6 pages of writing (1.5 or double spaced) plus tables at the end

Here is the rubric for grading (out of 100 points):

1. 5pts— Clear description of research question and variables of interest
2. 10pts – Summary of main findings
3. 10pts – Summary statistics (and description)
4. 5pts – Shows/interprets a graph
5. 10pts – Explicitly state regression model, justification for model
6. 25pts – Discussion of main results (including significance levels and any formal tests; correct interpretation of variables and how they are measured)
7. 10pts – Other models—either to explain main finding or to try different functional forms
8. 15pts – Discussion of possible issues with your model—omitted variable bias, reverse causality (does y cause x?)
9. 10pts – Overall organization and readability; e.g. tables are easy to read, well labeled; text is clear

I’ve shortened the outline below to highlight the things that you must cover.

The Basic Idea

For the course project, you are to write a short paper investigating the effect of alcohol consumption on a labor market outcome. You will do this by applying the econometric techniques used in class to the data set that is available on SmartSite: **alcohol.dta**

Chapter 19 in the text discusses carrying out an empirical project and has many helpful hints (I’ve posted a pdf of this chapter on SmartSite). Read that chapter closely. Note, however, that for this project you do not need to include a literature review, since the focus is more on the process of doing empirical work than on truly pushing forward the existing literature. While I have defined the basic outline of the question for you, you need to refine it. You first need to think about what labor market outcome interests you the most. Possibilities include such things 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 ceteris paribus effect. There is no one “right answer” for what dependent variable or key explanatory variables 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 to have a causal interpretation.

For an example of a real study that tackles a related issue, consider Levine, Phillip B., Tara A. Gustafson and Ann D. Velenchik. 1997. "More Bad News for Smokers? The Effects of Cigarette Smoking on Wages," Industrial and Labor Relations Review, Vol. 50, No 3 (April) pp. 493-509 which is available SmartSite (along with the Stata program and log file which contains the main results in Table 4 of the article) and based partly on these data. Reading this article and looking at how the reported results relate to the data and the Stata program should help you to think clearly about your project.

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 outcome measure. Then look and see how each concept can be captured by the data. For example, in the above article, the precise question being asked is, "What is the effect of smoking behavior on the wages of full-time, full-year workers in 1984 and 1991?" For Column 6 of Table 4, the dependent variable is log hourly wage (lnhrwg91), and the key explanatory variable is whether the individual is a daily smoker - versus a nonsmoker or just occasional smoker (smkdly91). 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 (higrad91), years of work experience (exp91), score on an intelligence test (afqt), dummies for living in an urban area (urban91) in the south (south91), being nonwhite (nonwhite), being female (female), being married (marrd91), and the number of children (numkids91). While the effects are not reported, 9 additional variables about the parents and household at age 14 are included to capture family background. In the order listed in Table 1, these are: urban14, south14, mtreduc, ftreduc, numsibs, bothpar, mtronly, mtrwrk14, ftrwrk14.

This article is not meant to be an exact template for your project. Rather it is meant to help you in thinking about how to approach the question at hand. Nonetheless, as described below you will be using similar data so it should be useful to see what types of things are considered to affect labor market outcomes. Your paper should then discuss the theory/common sense behind your choices of variables to use in the estimated models.

After coming up with a conceptual model and a base model specification, you need to estimate the model and do some specification checking. Estimating this final specification of the model is not the end of the story, though. You will still need to consider whether there are any problems with the estimated model that interfere with your ability to make proper inferences. It will be appropriate to test for and correct any such problems. Note that it is typical to present the simplest regresssion model (e.g. Table 4 in the article) corrected for basic problems such as heteroskedasticity first, and then consider more complicated approaches (e.g. Tables 5 and 6) to more serious problems such as omitted variable bias. Different approaches may have different strengths and weaknesses or even capture a somewhat different concept. Thus, for each of your approaches you should provide an interpretation of the results and discuss the strengths and weaknesses. If there are weaknesses that you are unable to address with these data, discuss what kind of additional data would be useful and how that data would help to solve your problem.

The Data

The data set alcohol.dta 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. It is defined as follows:
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 – a 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 follows: 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 the individual has had at least 1 drink perday – the average number of drinks per day on a day when the individual drinks (this is 0 if the individual doesn't drink)
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 for the analysis) 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 – a dummy variable = 1 if the individual lived in the south when they were 14 years old

wdad14 – a dummy variable =1 if the individual lived with their father when they were 14

wmom14 – a 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 – the 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* – the 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. (note: this variable might look a little funny to you because it was created as the midpoint of a range so that the place the individual lives could not be identified)

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 regression analysis. You will definitely have to create some new variables! Additionally, not all of the available variables will make sense in the model. Be especially careful of variables that don't make sense together, but may alone. You will have to make decisions about what to include, keeping in mind the ceteris paribus interpretation of multiple regression analysis. Additional Notes: You should maintain log files with not only the results presented in the paper, but with any other Stata commands used to define variables, perform tests, etc. So that I can consult them in case of questions, these log files should be turned in with your completed paper, and should be brought with you if you come to office hours to discuss the project. Within the actual paper, you'll want to present your estimated models and also descriptive statistics (mean, standard deviation, maximum, minimum) for all of the variables used in these models. Note that you want to look at these statistics to see if they make sense! When working on your project, you will find it useful to use a Stata "do file" or at a minimum to use the cmdlog command so that you can recreate your work if necessary.

The paper must address all of the subjects in the outline below.

Outline

1. Abstract: one paragraph description of your paper and findings
2. Introduction:
   a. What question are you asking?
   b. Why is this question interesting/important?
   c. Brief overview of what you will do in your paper/your main findings
3. Data and Methods: a reader has to understand exactly how to repeat this paper
   a. Describe your data
      i. Variables you are using in estimation and how they are measured
      ii. Include a table of descriptive statistics (mean, SD, min, max) and discuss in the context of the question you are asking.

1. Label your variables in such a way that they are easy to understand (ex: write out “mother’s level of educational attainment”, not just “medu”)

iii. Include a graph of some sort-- histogram, scatter plot, etc-- that provides some suggestive evidence with respect to your research question.

b. Describe your estimation methods:
   i. Choice of variables
   ii. Functional form

c. Write out your regression model (the equation you are estimating)

d. How will the model answer your research question/test your theory?
   i. What are you testing? What is the specific question(s) you are asking?
   ii. What signs should the betas take and why?

4. Results
   a. Show table(s) of results: tables must be formatted in such a way that they are easy to read (DO NOT SIMPLY COPY YOUR RESULTS WINDOW)
      i. Label your variables in such a way that they are easy to understand (ex: write out “mother’s level of educational attainment”, not just “medu”)
   b. Discuss your findings
      i. Interpret all of the coefficients (not in a list). Discuss what the results actually mean economically/what we learned from them
   c. Discuss why you settled on this model:
      i. Why these variables?
      ii. Why not other variables?
      iii. Why the functional forms for each of the variables included?
      iv. Are there any omitted variables that can cause bias?
   d. Show the regression results of alternative models: the goal is to show that your general results hold when you make slight changes to your model
      i. Do the results look similar to your preferred model?
      ii. Why did the results change?
5. Conclusions
   a. Do your results match your theory?
      i. Recap what you did, then tell me if your answer matched your prior beliefs
   b. How did these findings contribute to the literature?
   c. What is the take-away message (think about how you would summarize your paper in a sentence or two)
6. References: list the papers you referenced in the text (if any)