**Session 3 and 4 (class 2)**

**December 23,2016**

1. Start thinking about term paper, please give your topics before January 7th 2017
2. From last class, we will cover hypothesis testing till end of notes, and potential outcomes. This may take about 30 minutes. Then, we can move over to the hands-on section.

**References**

Bertrand, Marianne and Sendhil Mullainathan, “Are Emily and Greg more employable than

Lakisha and Jamal? A field experiment on labor market discrimination," American Economic Review, 2004, 94 (4), 991{1013.

Washington, Ebonya, “Female Socialization: How Daughters Affect Their Legislator

Fathers," American Economic Review, 2008, 98 (1), 311-322.

**HANDS-ON MATERIAL**

In topics A, B, C below, we will explore the three research designs, namely

1. Observational data from surveys
2. Data from a randomized controlled experiment
3. Observational data based on a natural experiment

Within each setting, you can explicitly conduct checks on whether your regression is causal.

The last topic D will explore the FWL theorem, and explore more aspects of the data and ease you in with using stata.

1. **OBSERVATIONAL DATA**

Open NSS68.dta in stata. This is an extract of data from the 2011-12 National Sample Survey (employment-unemployment schedule) (NSS), a key nationally representative household survey which gives us information on labour market particulars about households in India. In addition, the data also contains socio-economic characteristics and education attainment of the household members, and consumption expenditure of the households. [[1]](#footnote-1)

1. In this section, we will estimate and explore the classical model of wage on a binary variable indicating obtaining at least a college education. Write down the regression on paper.
2. The variable gen\_edu contains information on general education status of the household. Based on this variable, create a dummy variable `college' that equals 1 for anyone with at least a college education and equals 0 for anyone with less than a college education. Hint: use *generate*, *replace* and *if* in stata (gen\_edu codes are defined in the appendix of this document).
3. Daily earnings are in the variable dwage. Use the *hist* command in stata to view the histogram of dwage of workers. What do you observe? Now, take the natural log of dwage to again view the histogram of workers. Do you see that the log transformation has affected the distribution? Comment. (hint: use *generate, histogram,* and function *ln()* in stata).
4. Use a regression to estimate the difference in log hourly wages for individuals with and without a college education. Is this difference statistically significant? Provide an interpretation of your estimates of the coefficient on the constant term and the slope coefficient.
5. Does the regression above can provide a good estimate of the causal effect of finishing college on wages? Why or why not? What is one simple way by which you can test if the regression is causal, using other characteristics of workers that you may observe in the data?
6. **RANDOMIZED EXPERIMENT**
7. We will explore how causal effects are estimated in the context of a randomized trial using the data based on a randomized experiment used in Bertrand and Mullainathan (2004). Download the paper and datasets from the American Economic Review website. This experiment essentially sends out fictious resumes in response to 1300 job advertisements in Chicago and Boston. For each ad, they send 4 resumes – two higher quality and two lower quality resumes. They randomly assign to one of higher quality and one of the lower quality resumes, black-sounding names. The other is a white sounding name. After conducting this experiment, they waited for call-back for interviews. The main idea of the paper is to examine whether call-back rates were lower for black-sounding names. This answers an important question on how the perception of race affects job prospects in the USA, a country where blacks are twice more likely to be unemployed and earn 25% less when they are employed.

Take five minutes to read the abstract and first two paragraphs of the introduction.

1. What is the key question that this study attempts to answer? Write that down in the form of a regression on paper. .
2. Run a regression of call-back on black. Interpret the coefficients in this regression, and conduct a t-test for whether the coefficient on black is statistically significant at the 5 percent level.
3. How do you know that the randomization worked? In other words, can we give the coefficients in this regression a causal interpretation? In other words, under what conditions will this simple regression yield an unbiased estimate of the true causal effect? Answers to these three questions are the same. Use several pieces of information in the data to provide evidence about whether “that” key condition is met.
4. Now run a multivariate regression of call-back on black, yearsexp, honors, computer skills, and empholes. Provide a precise interpretation of each of the coefficients in the regression.
5. How does your estimate of the effect of being black on the probability of getting a call back change in this specification? Does this regression give you more or less faith in a causal interpretation of the evidence in the previous question? Explain.
6. How does the standard error of the coefficient on black change in this new specification? What does that tell you about these newly added variables?
7. There are likely many other variables that influence whether people are called back for an interview that aren't captured in the dataset at hand. Does their omission from the models above bias the estimates from those regressions? Explain why or why not.
8. **RESEARCH DESIGN BASED ON A NATURAL EXPERIMENT**
9. This question is about a research paper where multi-variate OLS was combined with a compelling research design: Ebonya Washington's 2008 paper “Female Socialization: How Daughters Affect Their Legislator Father's Voting Behavior." The paper asks whether having daughters, conditional on overall family size causes people in the United States Congress to vote more `liberally' on bills related to “women's issues”. Download the paper and datasets from the American Economic Review website.

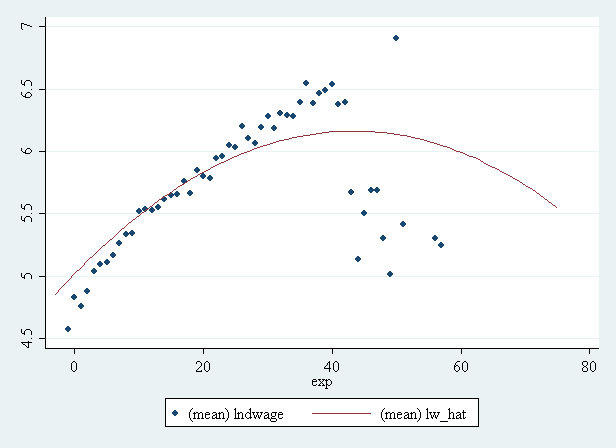
The paper estimates the effects of having a daughter (or number of daughters) on congress members’ voting behavior on liberal issues related to women. Let’s break this down. The US congress is bicameral (has an upper house- senate, and a lower house: the House of Representatives), and the members vote on bills to make them laws. The National Organization for Women (NOW) and American Association of University Women (AAUW) develop scores based on how congress members vote on bills. If a member voted more liberally, they get a higher NOW and AAUW score and vice versa. The paper uses cross section data on members of the US house of representatives (lower house) of four congresses (105th to 108th), spanning years 1997 to 2004. Remember, this is not experimental data, but observational data. The author uses a “research design”, which is a natural experiment. She uses as her identification strategy, that the gender of the child is completely random. Let’s proceed to see how the data looks like. Open wash basic.dta

Take five minutes to read the abstract and first two paragraphs of the introduction.

1. Write down the regression in column 1 of Table 2 on paper. Estimate that regression and interpret the coefficient on number of girls.
2. Succinctly describe the author's identification strategy, or why she believes that the regression results in this Table reflect the true causal impact of daughters. What assumptions are necessary for the coefficients above to consistently estimate the true causal effect?
3. Given the stated research design of the paper, what is the most parsimonious specification (the regression with fewest control variables that one could estimate) that should still yield the true causal effect of daughters on voting behavior? Run that regression and comment on how your results affect your evaluation of the validity of the identifying assumptions.
4. Conduct some extra analyses to see whether the identifying assumptions made by the paper seem valid.
5. If you were a referee (or the editor) for the AER, would you recommend the paper be published or not, or what additional analyses would you request of the author to convince you of the paper's merit.
6. If you were the author of this paper, would you have presented the analyses summarized in Table 2 differently? Explain how or why?
7. **EXPLORING MORE STATA COMMANDS AND THE FRISCH–WAUGH–LOVELL (FWL) THEOREM**

Open NSS68.dta

1. A benchmark in labour economics is the following “human capital" regression model that relates an individual's wage to their experience and education levels. The idea is that individuals' wages represent a return on particular kinds of investments (e.g., in general knowledge, S, and in work experience).
2. Create the variable ‘s’ (years of schooling) based on the variable gen\_edu and the rules in appendix (hint: use stata commands *generate, replace, if*):
3. Create a measure of ‘potential experience’: potexp = age - S - 5. This essentially assumes individuals start their schooling at age 5, and go to work directly after spending S years continuously enrolled in school. Create squared of experience variable (hint: you can use stata unary operators (##) to achieve the same without creating a separate variable)
4. Run the human capital model regression. What is the average return (in terms of increased wages) of an additional year of experience? Do this by hand (not stata). How is “average return” related to “returns at average” in this setting? (hint: *Method* *1:* *Use stata commands* *generate, \_b[] to access coefficients, and e(sample) to obtain sample of the previous regression. Method 2: Use lincom.*
5. What is the role of the experienced-squared term (why not include only experience by itself in the regression)?
6. Produce a graph like one below:



1. Generate a variable marfem = married\*female. Now run a regression of log wages on experience, experience-squared, years of schooling, female, married, and marfem. Interpret the coefficient on marfem. Do women `benefit' from marriage in the labour market? (calculate this by hand, before using stata)
2. Use the *margins* command to repeat iii and vi.
3. FWL theorem: Regress log wages on experience and experience-squared and a constant term, and compute the residual . Now regress years of schooling on the same variables and compute the residual . Regress on with no constant term in the regression. Compare your estimate (and standard error) to you answer in 2(iii) above. Hint: You will need the `predict' command for this question (consider using the option `,resid').
4. Compute the mean level of log wages for men and for women separately. After the command sum, type “return list". Either using macros, or by hand calculate the difference in mean female and male log wages.
5. Regress log wages on a dummy variable for female. Interpret the coefficient on female. How does this coefficient compare to what you found in the question ix?
6. Now create a new variable which is equal to one if male and zero otherwise. Run a regression of log wages on `male' and `female' but omitting the constant and compare this to what you got from your means command.

**Appendix**

**Code for gen\_edu**

*not literate -01,* literate without formal schooling: *EGS/ NFEC/ AEC -02, TLC -03, others -04;* literate: *below primary -05, primary -06, middle -07, secondary -08, higher secondary -10, diploma/certificate course -11, graduate -12, postgraduate and above -13.*

**Rules for creating years of education variable based on general education categories:**

* No literacy, literacy through informal means gives 0 years education
* Below primary school competition (this can anything between class 1 to 4) gives on average 2.5 years of education
* Completing primary school gives 5 years of education
* Completing middle school gives 8 years of education
* Completing secondary school gives 10 years of education
* Completing higher secondary school gives 12 years of education
* Completing a diploma or a certificate course gives 12 years of education (Assume some diplomas can be obtained by passing 10th standard, or after passing 12th class. Assume diplomas are for 1 year. Total years of education will be an average of 11 or 13, which is 12)
* Completing a graduate course (college degree) gives on average 13.5 years of education (college can be for 3 years or 4 years after class 12th)
* Completing a postgraduate course (Master’s degree) gives on average 15.5 years of education (adding 2 years after college).

1. More information about the NSS can be found in their website [here](http://www.mospi.gov.in/national-sample-survey-office-nsso). [↑](#footnote-ref-1)