

Research Project, ECON 30331
Due by noon, in hard copy, Wednesday May, 2 2018

Bill Evans
Spring 2018

The research project for this course is a short empirical paper that answers one of the seven general questions outlined below. You are to utilize the techniques we've outline in class such as the multivariate regression model, interpreting coefficients, using t and F-tests, the use of dummy variables, etc. when addressing your question. Select one of seven topics listed below and use the data provided on the class web page. Work in groups of three and you can start to sign up after class today. Anyone not in a group by Monday, March 21st will be randomly assigned to a group and a topic. The number of groups who can work on any given topic is restricted to 3. The seven topics for the paper are the following:

- **What is the impact of alcohol consumption on labor market outcomes?** There are some suggestive results indicating that heavy alcohol consumption may adversely impact labor market outcomes like employment and wages. There is also some evidence indicating that moderate alcohol consumption might be beneficial. You are to investigate this question using data from the National Health Interview Survey. The data set is called `nhis_alcohol.dta`. The data set has some other measures of health like how much one exercises and the body mass index so I am open to papers that examine these activities on labor market outcomes as well. **Most difficult aspect of project:** Deciding how to measure alcohol use/abuse. There are many different ways.
- **Has the Affordable Care Act (ACA) increased insurance coverage for young adults?** In an effort increase insurance health insurance coverage rates for young adults, the ACA required health insurance companies to provide insurance for children under the age of 26 through their parent's coverage. Using data from the March Current Population Survey for people aged 19-29, you are to estimate the impact of the ACA on insurance rates for people under 26. This project will require that students read ahead about "difference-in-difference" models but this is a fun and great project. This data set is named `aca_data_march_cps.dta`. **Most difficult aspect of project:** Will require some work learning about the ACA and reading ahead about "difference-in- difference" models which we will start to study in two weeks.
- **How does obesity and being overweight impact health care utilization spending among the elderly?** Obesity rates have doubled in the past 30 years. Medicare, the federal health insurance program for the elderly, is the fastest growing expense in the federal budget. Knowledge of how obesity and being overweight impacts health will help identify the impact of these two conditions on federal budget outlays over time. You are to use data from the 2005 Medical Expenditures Panel Survey to examine this question. The data set for this project is named `meps_senior.dta`. **Most difficult aspect of project:** Deciding what covariates to use and interpreting the results.
- **What evidence is there of stock market anomalies in recent periods?** A variety of authors have documented a number of arbitrage opportunities in stock market prices. For example, average daily returns tend to be rise in January, are lower on Mondays, tend to be higher right before Holidays, etc. These anomalies run counter to the notion that the stock market is efficiently priced. If there are periods when profits are persistently higher or lower, then there is an arbitrage opportunity to earn excess returns. You are to use the daily return (defined as the percentage change in price between today and yesterday) for the Dow Jones Industrial Average from 1950 though 2007 to examine whether some of the anomalous patterns exist in your data and maybe even find a few anomalies yourself. A nice article that outlines this basic topic is: Richard Thaler, "Anomalies: Weekend, Holiday, Turn of the Month, and Intraday Effects," *Journal of Economic Perspectives*, 1, Fall 1987, 169-78. This article is available electronically. The data set for this project is called `djia_daily_data.dta`. **Most difficult aspect of project:** Depending on what anomaly you consider. The programming on this project is somewhat difficult.
- **What are the determinants of test scores for 5th grade students?** A frequent measure of academic success is a score on a standardized exam. Authors have demonstrated that many characteristics of the student, their family, their home environment and their school are predictive of test scores. In this case, you will use data

from the Early Childhood Longitudinal Survey which is a nationally-representative survey. The original survey respondents were from the Kindergarten class of 1998-99 and children were re-surveyed in the 1st, 3rd, and 5th grade. The data for this project comes from the 5th grade follow-up. The data contains test scores in reading, math, and science, plus detailed information about the child, their family, and their school. For this data set, consider asking a very specific question, such as whether school type (private versus public) impacts test scores? Does a negative school environment generated by crime, violence or gangs impacts test scores? Does participation in extracurricular activities impact learning? Does TV rot your brain? Are children negatively impacted by divorce? There are lots of variables in this data set so there are many possible paper topics. The data set for this topic is named `ecls_5th_grade`. **Most difficult aspect of project:** Deciding on what to focus on as the covariate of interest and what to control for in the regressions.

- **How has the college premium changed over time?** The college premium is defined as the difference in $\ln(\text{weekly earnings})$ or $\ln(\text{hourly wages})$ between college and high school graduates. Data from a number of sources indicates that this gap has increased over time. Your job is to document the change in this premium. I have put together data from 1975 through 2010 for full-time/full year employed high school and college graduates aged 23-50 from the March Current Population Survey. This survey asks people about their earnings in the previous year so data from 1990 is for the 1989 calendar year. This data set is relatively large so I have put together two files: one for males, one for females. You can answer any number of questions for this broad topic? How has the premium changed for males? Is there a differential change in the value for females? **Most difficult aspect of project:** The programming is more difficult than average and since you will have many years of results, you will need to display results in a series of graphs rather than tables.
- **What is the impact of higher minimum wage rates on teen employment?** Over the past 30 years the real wages for people in the lower 50% of the wage distribution have fallen while the wages for those at the top of the distribution have skyrocketed. This has led to increased calls for much higher minimum wage rates. Currently there are efforts at the federal, state and local level to raise the minimum wage to \$15/hour. Opponents of higher minimum wages argue that a higher minimum wage will reduce employment among the least skills. In this project you are to estimate the impact of a higher minimum wage on employment of teenagers aged 16-19. This is a low-skilled group and one most likely impacted by the minimum wage. The data set is measures of teen employment at the state level from 1990 through 2013 (24 years) for all states and the District of Columbia (51 cross-sectional groups) for a total of $24 \times 51 = 1224$ observations. The data is from the March CPS, described above in the ACA example. The variation used to identify the model is the fact that some states have increased the minimum wage over time – state can raise their wage above the federal minimum level but not reduce it. **Most difficult aspect of project:** The model you will estimate is a “difference-in- difference” specification and we will not start to talk about these models for about two weeks.

The paper should be about 15 pages, double spaced, 1 inch margins, 12 point type, including all figures, tables and references. Your paper should include an introduction that outlines why this is an important question, a review of the relevant literature, a brief outline of the econometric model, a description of the data, and a discussion of the econometric results. Your paper should include a table of descriptive statistics for your data set. When possible, you should also contain a detailed discussion of problems in your econometric specification, e.g. are the results potentially biased by omitted variables that are not in your sample. Grades for the paper will be based on the completeness of the paper (how thorough is the literature review, description of the data, etc.), grammar, the soundness of the empirical model, and your accuracy of the interpretation of the econometric model. All papers must demonstrate understanding of t and f-tests.

The data for the projects are in the *Assignments* portion of the class web page and I provide detailed descriptions of the data in PDF files as well. For the first two projects above, the data descriptions also provide a link to web pages that describe in more detail the data set where your samples came.

Some hints/comments about the research process

- Before you start playing with the data set, do some reading. See how other people have addressed the particular problem in the past.
- Please read chapter 19 of the textbook. It has a number of useful suggestions about how to write an empirical paper.
- Put some thought into the model. Think about what variables should be used as outcomes (y's), what should be used as control variables (x's).
- Think about how the variables should be measured? Should they be logged or linear?
- Just because a variable is in the data set does not mean it should be used. Think about what is an appropriate control variable and what is not.
- Before running a regression, take a look at the data set in a systematic way. Get the descriptive statistics for important variables. Look at the means by important subgroups. For example, in project 1 above, compare wages of heavy drinkers, light drinkers and non-drinkers. For project 3, look at daily returns by the day of the week, by month of the year.
- When looking for a template of what a paper should look like, both in structure, context, and visually, use the academic readings from class as a guide. For example, last semester, most of the papers had poorly crafted tables. Look at the papers we've read in class for a guide about how to construct a table.

Some helpful hints when writing up your paper

- Please, please, please add page numbers
- Footnotes come after punctuation, ¹ instead of ¹.
- It is et al. (period after al) – not et. al. or et al (no periods)
- References should be noted within the paper with last names only and date, then the full bibliographic reference listed in the references at the end of the paper. So you would have in the text. “Results from White and Black (2015) suggest that a 10 percent hike in the minimum wage reduces teen employment by 2 percent.” For reference style, use the Chicago manual of style
http://www.chicagomanualofstyle.org/16/ch15/ch15_sec009.html?sessionId=6475e28a-4aea-4d9b-84a7-7f2d59bc63a6
- STATA requires short variables names and underscores, mom_educ, ln_weekly_earn, but word processing packages do not. Use words and phrases in tables, so use Mother's education instead of mom_educ.
- Most which's should be that. Here is a helpful guide. <http://www.dailywritingtips.com/that-vs-which/>
- DO NOT just cut and paste STATA output into your paper, create your own tables to fit your needs. If you look at any paper we have read for class, you will note that there are NO vertical lines. Look at your favorite paper from class and structure your tables like the tables in the article.
- The order of the paper should be: text, references, figures, then tables. DO NOT try to put tables and figures into the text – this makes editing hard and the pagination is almost always screwed up. You are not professional type setters.
- Include a title page that has the paper title, names of students on the projects, the current date and a 100-200 word abstract
- Please think hard about how to express your equation of interest in mathematical notation. Variable should be subscripted to note the dimensions of the data – “i” subscript for cross sectional data and “t” subscript for time series. In general, we write variable then parameter – $x\beta$ – NOT βx . This has more to do with linear algebra than anything else. Use math to free up space. For example, in the text score topic, you have three test scores, reading, math and science. In most cases, people will estimate the same model for each test score, so it is easier to write

$$y_i^j = \beta_0^j + x_i \beta_1^j + \varepsilon_i^j$$

Where j is the test (math, reading or science) and “i” represents the individual than to write three equations – one for each test score